MEDIZINISCHE UNIVERSITÄT WIEN

Identification and characterization of long non-protein-coding RNAs in the human genome

Doctoral thesis at the Medical University of Vienna for obtaining the academic degree

Doctor of Philosophy

Submitted by
Aleksandra E. Kornienko

Supervisor: Denise P. Barlow, Ph.D.,
CeMM Principal Investigator,
Honorary Professor of Genetics at the University of Vienna.
CeMM
Research Center for Molecular Medicine of the Austrian Academy of Sciences,
Lazarettgasse 14, AKH BT25.3, 1090 Vienna, Austria

Vienna, February 2016


MEDIZINISCHE
UNIVERSITÄT
WIEN

DECLARATION

This Doctoral Thesis is written in the cumulative format. Three publications, where the author of the thesis is the first author, are included. For all three publications the author of this thesis performed the majority of the experimental/bioinformatic work and writing of the manuscript. All the three publications included in the thesis are preceded with cover pages giving a brief description and specifying all the authors' contributions.

Publication 1 – a review "**Gene regulation by the act of long non-coding RNA transcription**" – is included into the INTRODUCTION in the Chapter "*1.3 Functions and mechanisms of lncRNAs*" and discusses this topic, overviewing and illustrating various modes of lncRNA (long non-coding RNA) action, including the newly discovered mechanism of transcription interference.

Publication 2 – a research article "**Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans**" – is included into the RESULTS as Chapter 1 and constitutes the main research project and the major novel scientific outcome of the thesis. Supplemental Information (referred to as Additional File 1 in the publication) is included in the APPENDIX at the end of the thesis and contains 35 Supplemental Figures with Legends and Supplemental Methods. Supplemental Tables 2 and 11 (referred to as Additional Files 2 and 11 in the publication) are also included in the APPENDIX. Additional Files 3-10 containing annotations of genes identified in the project are not included because of their size, but are available at the Journal webpage (http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0873-8) in the Additional Files section. Datasets produced in this project were submitted to GEO (accession number GSE70390).

Publication 3 – a research article "**A human haploid gene trap collection to study lncRNAs with unusual RNA biology**" – is included in the RESULTS as Chapter 3. Supplemental Figures and Supplemental Tables accompanying this publication are included in the APPENDIX part of the thesis under the titles referred to in the publication. Supplemental Tables 1C and 1E are not included into the thesis because of their size, but they are available at the publication page in the Supplemental Information

section on the Journal website (http://www.tandfonline.com/doi/abs/10.1080/15476286.2015.1110676?journalCode=krnb20). Datasets produced in this project were submitted to GEO (accession number GSE71284).

Chapter 2 of the RESULTS includes unpublished research and describes the analysis of histone modification of granulocyte lncRNA and mRNA genes – data obtained for the main project (Publication 2), but not included in the manuscript.

N.B.: MATERIALS AND METHODS Chapter of this Doctoral Thesis only includes methods relevant for Chapter 2, while methods relevant for other results (namely Publication 2 and Publication 3) obtained in this Doctoral Thesis are included in Publication 2 and Publication 3 and comprehensively describe all the experimental and bioinformatic methods applied during the completion of this Doctoral Thesis: see Publication 2 (Methods section and Supplemental Methods in the APPENDIX) and Publication 3 (Methods section).

All the work was performed at the laboratory of Denise P. Barlow at CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences in Vienna, Austria.

Publication 1 was published in BMC Biology (impact factor as of 1/2016 – 7.984) on 30.05.2013 with Open Access: PMID: 23721193.
Link: http://bmcbiol.biomedcentral.com/articles/10.1186/1741-7007-11-59

Publication 2 was published in Genome Biology (impact factor as of 1/2016 – 10.8) on 29.01.2016 with Open Access: PMID: 26821746.
Link: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0873-8

Publication 3 was published in RNA Biology (impact factor as of 1/2016 – 4.974) on 15.12.2015 (Epub ahead of print): PMID: 26670263.
Link:
http://www.tandfonline.com/doi/abs/10.1080/15476286.2015.1110676?journalCode=krnb20

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Long non-coding RNAs (lncRNAs) are a relatively recently emerged new class of genes, that are becoming increasingly appreciated as gene regulators and disease players. It has recently become clear that lncRNA genes are inherent to the genomes of most organisms, and that they are unexpectedly numerous - likely even more numerous, in some organisms, than classical protein-coding genes. LncRNA genes resemble protein-coding genes at the first glance, but it has become increasingly clear that lncRNAs are a more diverse gene and transcript class with a set of non-mRNA-like features. Two major features of lncRNAs is their low abundance and extreme tissue-specificity. These features make lncRNA identification challenging and deep coverage analysis of pure tissues and cell types is required to comprehensively annotate lncRNAs. LncRNA natural expression variation is a feature that has not been examined in comparison to protein-coding genes. We used human primary granulocytes obtained from healthy volunteers to fill this knowledge gap. Granulocyte-specific transcriptome (and particularly lncRNA) annotation was unavailable and we used PolyA+ RNA-seq data from 10 individuals to create *de novo* lncRNA and mRNA annotation in granulocytes, identifying numerous novel lncRNAs. We then used ribosomal depleted RNA-seq of granulocytes from 7 individuals sampled at >=1-month intervals to calculate expression variation of the annotated lncRNAs and mRNAs. We discovered that lncRNA expression was notably more variable than mRNAs, even when controlling for general lncRNA low expression level. We confirmed the generality of the discovered phenomenon by analyzing publicly available data from 9 human tissues (20 individuals each) from GTEx project and lymphoblastoid cell lines (462 individuals). Further analysis of the latter dataset allowed us to show that high expression variability influences the process of lncRNA identification and the number of identified lncRNA loci increases steadily with the number of healthy donors used for the identification. These findings provide important novel insight into lncRNA biology and also identify a new non-mRNA-like feature of lncRNAs that together give new guidelines for lncRNA identification and their functional characterization strategy. In addition, these results influence potential prospects of the use of lncRNAs as biomarkers and their implication in personalized medicine.

The ever-increasing number of lncRNAs annotated in the human genome raises concerns about meaningfulness of their transcription and questions their functionality. Thus, a convenient large-scale lncRNA functional assessment method is of high importance. We used a previously uncharacterized lncRNA *SLC38A4-AS* as a model to propose a rapid RNA-biology feature characterization pipeline followed by genetic truncation leading to the functional knockout, using ready-made Human Haploid Gene Trap Collection. We show that *SLC38A4-AS* lncRNA is a lncRNA possessing unusual RNA-biology features, including inefficient splicing, dramatically distinguishing it from a typical mRNA or many lncRNAs. However, we show that Human Haploid Gene Trap Collection is an efficient tool for genetic manipulation and functional study of such a lncRNA. The results showed that the *SLC38A4-AS* lncRNA is a functional regulator and they provide a list of 6 stringently filtered potential targets of *SLC38A4-AS*, which included *CD9* and *RORB* protein-coding genes.

ZUSAMMENFASSUNG

Lange nicht-kodierende RNAs (lncRNAs) sind eine relativ neue Klasse an Genen die zunehmend als Genregulatoren und als krankheitsrelevant betrachtet werden. Genomuntersuchungen haben gezeigt, dass lncRNAs in fast allen Organismen vorkommen und in manchen wahrscheinlich sogar zahlenmäßig häufiger als klassische Protein-kodierende Gene. Auf den ersten Blick ähneln lncRNAs Protein-kodierenden Genen aber es hat sich gezeigt, dass lncRNAs sehr unterschiedlich sein können und viele lncRNAs Eigenschaften haben die sie deutlich von mRNAs unterscheiden. Zwei der Haupteigenschaften von lncRNAs sind ihre geringes Vorkommen und ihre extreme Gewebsspezifität. Diese Eigenschaften machen die Identifizierung und Annotation von lncRNAs sehr schwierig und erfordern eine sehr tiefe Sequenzierung von sehr reinen Geweben und Zelltypen. Natürliche Expressionsunterschiede von lncRNAs im direkten Vergleich zu Protein-kodierenden Genen sind bisher nur unzureichend untersucht. Wir haben daher primäre humane Granulozyten von gesunden Freiwilligen isoliert um diese Wissenslücke zu schließen. Da ein Granulozyten-spezifisches lncRNA Transkriptom nicht verfügbar war, verwendeten wir PolyA+ Granulozyten RNA-Seq Daten von 10 Personen um eine *de novo* Annotation von lncRNAs und mRNAs zu generieren und identifizierten dabei zahlreiche bisher unbekannte lncRNAs. Dann verwendeten wir ribosomenlose RNA-seq Daten von Granulozyten, die wir von 7 Personen in >=1-Monats-Intervallen isoliert haben und analysierten damit die Expressionsunterschiede der annotierten lncRNAs und mRNAs. Wir fanden heraus, dass in den einzelnen Personen die Expression der lncRNAs deutlich variabler ist als die der mRNAs, sogar wenn für die geringeren Expressionslevels der lncRNAs kontrolliert wurde. Wir bestätigten auch, dass dieses Phänomen nicht nur bei unserem Datensatz auftritt, sondern auch bei veröffentlichten RNA-Seq Daten von 9 humanen Geweben (20 Personen pro Gewebe) des GTEx Projekts und bei lymphoblastoiden Zelllinien von 462 Personen. Weitere Analysen des letzten Datensatzes zeigten uns auch, dass die variable Expression der lncRNAs einen deutlichen Einfluss auf die Identifizierung der lncRNAs hat und dass die Anzahl der neu annotierten lncRNAs steigt, je mehr Personen untersucht werden. Diese Resultate bieten wertvolle neue Einblicke in die Biologie der lncRNAs und besonders interessant sind einige neu identifizierte Eigenschaften die besonders unterschiedlich sind zwischen lncRNAs und mRNAs. Die hier präsentierten Daten werden als Orientierungshilfe bei der weiteren Identifizierung von lncRNAs

dienen und in weiterer Folge auch entscheidend beeinflussen, wie lncRNAs als Biomarker in der personalisierten Medizin funktionieren können.

Die stetig steigende Zahl an annotierten lncRNAs im humanen Genom hat Bedenken ausgelöst, ob all diese Transkripte auch wirklich eine biologische Funktion haben. Daher ist eine Methode zur großangelegten Untersuchung der Funktionen von lncRNAs von größter Wichtigkeit. Wir haben die bisher uncharakterisierte lncRNA *SLC38A4-AS* als Modell verwendet um eine schnelle Serie an Experimenten aufzusetzen mit dem Ziel die spezifischen Eigenschaften dieser lncRNA zu untersuchen. Zusätzlich verwendeten wir eine fertige Kollektion von humanen haploiden Gene-Traps um eine genetische Verkürzung und damit einen Knock-out dieser lncRNAs zu erreichen. Wir konnten zeigen, dass die lncRNA *SLC38A4-AS* eine ungewöhnliche RNA Biologie hat da sie unter anderem ineffizient gespleißt ist, was sie sehr von mRNAs und auch vielen anderen lncRNAs unterscheidet. Weiters zeigen wir, dass die Kollektion von humanen haploiden Gene-Traps ein effizientes Werkzeug für genetische Manipulationen ist und ausgezeichnet für die Charakterisierung von lncRNAs funktioniert. Die Ergebnisse zeigen, dass die lncRNA *SLC38A4-AS* ein funktioneller Regulator ist und 6 potentielle Zielgene (inkl. die Protein-kodierende Gene *CD9* und *RORB*) hat.

PUBLICATIONS ARISING FROM THIS THESIS

1. Kornienko AE, Dotter CP, Guenzl PM, Gisslinger H, Gisslinger B, Cleary C, Kralovics R, Pauler FM, Barlow DP (2016) Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans. **Genome Biol**: Jan 29;17(1):14. doi: 10.1186/s13059-016-0873-8.
2. Kornienko AE, Vlatkovic I, Neesen J, Barlow DP, Pauler FM (2015) A human haploid gene trap collection to study lncRNAs with unusual RNA biology. **RNA Biol**: Dec 15:0. [Epub ahead of print]
3. Kornienko AE, Guenzl PM, Barlow DP, Pauler FM (2013) Gene regulation by the act of long non-coding RNA transcription. **BMC biology**: 2013 May 30;11:59. doi: 10.1186/1741-7007-11-59. Review.

ABBREVIATIONS

Abbreviation	Meaning
100PE	100bp paired-end RNA-seq
100SE	100bp single-end RNA-seq
50PE	50bp paired-end RNA-seq
ActD	Actinomycin D
Airn	Antisense Igf2r RNA non-coding
ANRIL	Antisense noncoding RNA in the INK4 locus
BACE1-AS	'Beta-site APP-cleaving enzyme 1 isoform A' gene antisense lncRNA
BOK-AS	BOK (BCL2-Related Ovarian Killer) gene antisense lncRNA
bp	Base pairs
CAGE tag	Cap Analysis Gene Expression tag
CHART	Capture Hybridization Analysis of RNA Targets
ChIRP	Chromatin Isolation by RNA Purification
CLASH	Crosslinking Analysis of Synthetic Hybrids
CLIP	Crosslinking Immunoprecipitation
CTD	Carboxy-terminal domain (of RNAPII)
DISC2	Disrupted-In-Schizophrenia 2 (lncRNA)
DNA	Deoxyribonucleic acid
eQTL	Expression Quantitative Trait Loci
EZH2	Enhancer of zeste homolog 2
FISH	Fluorescent In Situ Hybridization
GWAS	Genome-Wide Association Study
H3K27ac	Acetylation of H3 lysine 27
H3K27me3	Trimethylation of H3 lysine 27
H3K36me3	Trimethylation of H3 lysine 36
H3K4me1	Monomethylation of H3 lysine 4
H3K4me3	Trimethylation of H3 lysine 4
H3K9me3	Trimethylation of H3 lysine 4
HELLP syndrome	"HELLP" abbreviates the three major features of the disease – Hemolysis, Elevated Liver enzymes, Low Platelet count

hESC	Human Embryonic Stem Cells
HOTAIR	HOX transcript antisense RNA
HOTTIP	HOXA transcript at the distal tip
lincRNA	Long intergenic non-protein-coding RNA
lncRNA	Long non-protein-coding RNA
LSD1	lysine (K)-specific demethylase 1A (also known as Lysine-specific histone demethylase 1A (KDM1A))
MNC	Mononuclear cells
mRNA	Messenger RNA
NAT-RAD18	Natural Antisense Transcript Against Rad18 Gene
NET-seq	Native elongating transcript sequencing
NR1F2	Nuclear Receptor subfamily 1, group F, member 2
nt	Nucleotides
PINK1-AS	PINK1 (PTEN Induced Putative Kinase 1) gene antisense lncRNA
PRC1/PRC2	Polycomb repressive complex 1/2
RNA	Ribonucleic acid
RNAPII	RNA polymerase II
RNA-seq	RNA sequencing
RORB	RAR-related Orphan Receptor B
rRNA	Ribosomal RNA
siRNA	Small interfering RNA
snoRNA	Small nucleolar RNA
SNP	Single nucleotide polymorphism
TE	Transposable elements
tRNA	Transfer RNA
TSS	Transcription start site
Ube3a-AS (Ube3a-ATS)	Ube3a (Ubiquitin Protein Ligase E3A) antisense lncRNA
XIST	X-inactive specific transcript

1 INTRODUCTION

What makes humans who they are is, to my mind, the question that motivates people around the globe to study biology. The consequent global question is – once we have understood how and what we are made of in good enough detail, how can we use this knowledge to help humanity to live longer and suffer less?

With this in mind and arising from my particular interest in the biology of the human genome, a very important, though less general, question is – what is encoded in our genome? How does an enormously complex, exceptionally robust and brilliant biological system, such as the human body develop from a single fertilized egg given just a stretch of 3.2 billion "letters"? How does a chemical molecule, namely DNA, accompanied by other molecules and molecular complexes, perform an information transfer that allows formation of all the various tissue types, what is this information and how multilayered is it?

1.1 Human genome composition

The view on what information the human genome contains has been evolving exponentially over the last two centuries. The word "genome" is generally defined as the set of all the genes of a particular organism. However, the question of what the word "gene" should refer to is currently being disputed as never before, since the unexpected complexity of eukaryotic genome composition has been continuously unraveling and bringing surprises to the genome research community (Gerstein et al, 2007; Mudge et al, 2013; Raabe & Brosius, 2015).

The word "gene" emerged in 1909, articulated by Wilhelm Johannsen, based on the studies and the concept of heredity created by Gregor Mendel in the 19th century (see (Gerstein et al, 2007)). Johannsen's definition of the gene was: "special conditions, foundations and determiners which are present [in the gametes] in unique, separate and thereby independent ways [by which] many characteristics of the organism are specified" (Johannsen 1909, p124, cited in (Gerstein et al, 2007)). A few years later, studies in *Drosophila Melanogaster* led geneticist Thomas Hunt Morgan to the view of genes as some physical units on chromosomes, like beads on a string, with distance between them affecting their ability of crossing-over (Gerstein et al, 2007; Morgan, 1915). Thus, a gene

was now viewed as a distinct locus. In 1913, Alfred Sturtevant showed that genes are linearly arranged on chromosomes and created the first genetic map (Sturtevant, 1913). However, both latter studies still held a rather abstract notion of a gene studying gene-phenotype connection, i.e. they studied genes through observing the resulting phenotypes, without any insight into the mechanism of how genes might affect the phenotype (Gerstein et al, 2007). Only in 1941 was it found that genes correspond to proteins (Beadle & Tatum, 1941) and then, surprisingly quickly, the role of DNA in maintaining the organism's genetic information was discovered and the genetic code determining protein sequences was solved (Avery et al, 1944; Watson & Crick, 1953). These studies and discoveries strongly bound our view of "genes" to protein synthesis which resulted in the Central Dogma of molecular biology, formulated by Francis Crick in 1958 (Crick, 1958). The Central Dogma postulated that information is transferred from genes (DNA) to protein synthesis machinery via RNA (Figure 1). While by that time some non-protein-coding RNAs, such as tRNAs and rRNAs, were already known, they were considered exceptions and the major role allocated to RNA molecules was merely that of information transfer for protein synthesis. Only with the discovery more than six decades later, of various classes of genome-encoded non-coding RNA it became clear, that information transfer, as performed by mRNAs, was only one of the numerous roles RNA can perform in the cell (Clark et al, 2013).

The protein-centric definition of a "gene" as the DNA sequence encoding a certain protein, that had been stable for several decades, has recently been dramatically challenged (Gerstein et al, 2007). First, by the discovery of the complexity of protein-coding gene organization - mammalian genomes contain a myriad of *cis*-regulatory elements, such as promoters and enhancers, that reside mainly outside of the classically defined protein-coding genes, but are crucial for protein-coding gene function and establishment of tissue and cell type identity (Shlyueva et al, 2014) and, in principle, could be considered integral to the gene but do not directly encode for proteins (Gerstein et al, 2007). Moreover, with the discovery of alternative splicing, often tissue-specific, being inherent to the majority of mammalian protein-coding genes (Wang et al, 2008), it became clear that what we call a gene is rather biased to the deepness of our knowledge about a particular genomic region. For example, unknown alternative isoforms could add new exons to the gene or extend it by identifying an alternative transcription start site (TSS) (Forrest et al, 2014) or a polyadenylation (polyA) site (Ozsolak et al, 2010) in

previously not analyzed tissues, all of which can change not only the sequence, but also a function of the protein produced (Figure 1). These and other findings led to a skewing of gene definition towards its genomic component making the previous gene-phenotype connection, in a classic Mendelian definition, more subtle.

Table 1

General statistics of the human genome: GENCODE v 23 (March 2015 freeze, GRCh38)		
GENES		
Type of genes:	Number	**% total number of genes**
Total Number of Genes	**60498**	**100%**
Protein-coding genes	19797	32.7%
Long non-coding RNA genes	15931	26.3%
Small non-coding RNA genes	9882	16.3%
Pseudogenes	14477	23.9%
- Processed pseudogenes:	10727	- 17.7%
- Unprocessed pseudogenes:	3271	- 5.4%
- Unitary pseudogenes:	172	- 0.3%
- Polymorphic pseudogenes:	59	- 0.1%
- Pseudogenes:	21	- 0.03%
Immunoglobulin/T-cell receptor gene segments		
- Protein coding segments:	411	- 0.7%
- Pseudogenes:	227	- 0.4%
TRANSCRIPTS		
Type of transcripts:	Number	% total number of transcripts
Total number of transcripts	**198619**	**100%**
Protein-coding transcripts	79795	40.2%
- Full length protein-coding:	54775	- 27.6%
- Partial length protein-coding:	25020	- 12.6%
Nonsense mediated decay transcripts (i.e., erroneously transcribed mRNAs)	13307	6.7%
Long non-coding RNA loci transcripts	27817	14.0%
Total number of distinct translations	59774	
Genes that have more than one distinct translations	13556	

Table 1. Number of the main gene types in the human genome (Data taken from GENCODE v23 gene annotation release http://www.gencodegenes.org/stats/current.html). See (Djebali et al, 2012) for details.

The second thing challenging the protein-centric definition of a "gene" is the finding that in the last two decades it became clear that "protein-coding" is just one type of genes in the human genome. In 2003, the Human Genome Project (http://web.ornl.gov/sci/techresources/Human_Genome/index.shtml) revealed that only a small part of human DNA sequence actually encodes protein sequences. It was shown that while approximately 40% of the human genome is covered by protein-coding genes (introns and exons) only as little as 1.2% code for exons (Lander et al, 2001) and this trend is persistent in other mammals, e.g. mouse (Waterston et al, 2002). While only these 40% would be predicted to be transcribed (converted into RNA) from the human genome by the Central Dogma of Molecular biology, it has been shown that the majority of our genome is transcribed in one or another cell type (Djebali et al, 2012). Pervasive transcription of the human genome has been confirmed by a variety of techniques, including tiling array profiling, cloning, CAGE tag analysis that captures the start of a transcript and direct sequencing of cDNA fragments reverse transcribed from the whole RNA population (known as RNA-seq) (Clark et al, 2013).

Extensive gene discovery has now revealed tens of thousands of non-protein-coding genes in human (Table 1) and other mammals (mouse: http://www.gencodegenes.org/mouse_stats/current.html), and also the presence of regulatory non-protein-coding genes in all organisms down to bacteria (Clark et al, 2013; Kornienko et al, 2013; Ulitsky & Bartel, 2013). Among non-protein-coding genes the most prominent are non-coding RNAs, or ncRNAs, (42.6% of all genes, Table 1) and pseudogenes (23.9% of all genes, Table 1). NcRNAs and pseudogenes are numerous and largely contribute to transcriptome complexity (Table 1, Figure 1) and many of them were shown to possess an important function and to have played important roles in the process of evolution (Mattick et al, 2010).

With the discovery of ubiquitous non-coding transcription and a mosaic of genes filling our genome, our view of the genome changed from relatively simple to significantly complex (Figure 1, bottom). While, as discussed above, the Central Dogma of Molecular Biology only talked about a conventional protein-synthesis view on the gene, summarized in "DNA->RNA->Protein", the current view on the genome, adjusted by a multiplicity of ground-breaking discoveries (reviewed in (Rinn & Chang, 2012)), has gained several dimensions of complexity with nearly the whole genome sequence giving rise to a myriad

of various transcripts inside and outside, sense and antisense, to protein-coding genes, or "genes" as they were known before (Clark et al, 2013). Among these transcripts are various non-coding RNAs (Figure 1) that can be split into two major classes according to their size. These are small ncRNAs (reviewed in (Clark et al, 2013)), shorter than 200nt, and long non-protein-coding RNAs (lncRNAs), longer than 200nt (Quinn & Chang, 2015), that are believed to outnumber protein-coding genes (Clark et al, 2013), display various forms and functions (Quinn & Chang, 2015) (Rinn & Chang, 2012) and are the main focus of this Doctoral Thesis.

Figure 1

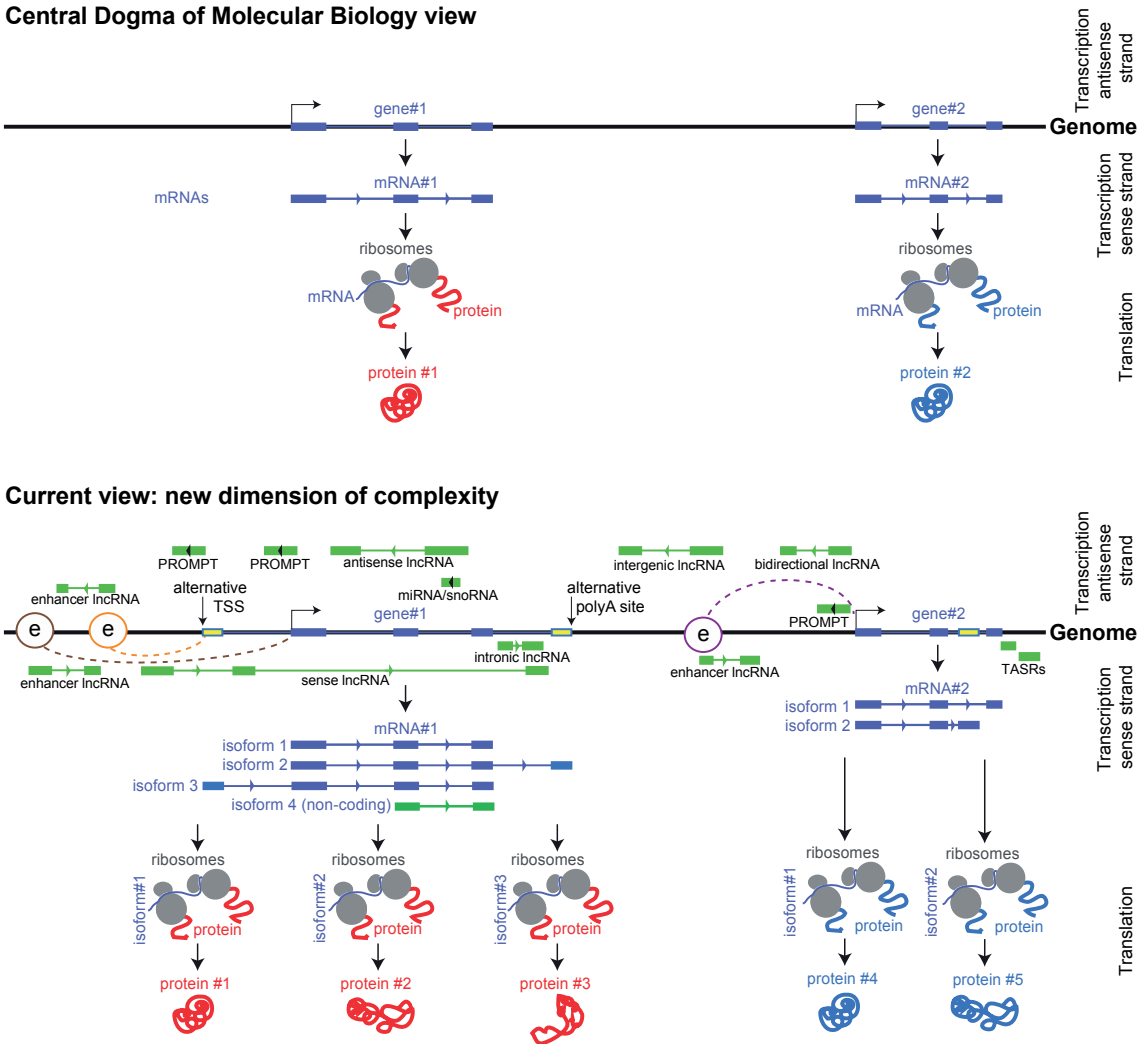


Figure 1. Difference between the classical (Central Dogma based) and the current view on mammalian genome. **Top:** Classical view on the content of the mammalian genome, dominated by the Central Dogma of Molecular Biology. Genome (black line) consists of protein-coding genes (blue bars) – stretches of DNA encoding a certain functional protein sequence. A gene is transcribed into a messenger RNA (blue bars and lines, mRNAs) in order to be transported to the cytoplasm and used by ribosomes (gray balls) as a template for protein (red and blue blobs)

production. One gene encodes one protein. **Bottom:** Current view on the content of the mammalian genome, complicated by the discovery of pervasive non-coding transcription, *cis*-regulatory genomic sequences and alternative isoforms. Today we know that genome (black line) consists of protein-coding genes (blue bars) and genes giving rise to lncRNAs and small ncRNAs (green bars). Protein-coding genes are supplemented with additional exons (yellow bars), used by alternative isoforms, which include alternative exons inside the initial gene body, but also alternative exons arising from the usage of an alternative TSS or polyadenylation site, which extends the initial gene annotation (Mudge et al, 2013). In addition, protein-coding genes are only active when an enhancer (a circle with "e" inside) comes in close proximity to the gene promoter/TSS allowing transcription machinery to assemble and function (Shlyueva et al, 2014). Alternative TSS may use different tissue-specific enhancers (Shlyueva et al, 2014). Enhancers also give rise to enhancer lncRNAs (or eRNAs (Orom et al, 2010)), that might be expressed sense or antisense to the protein-coding gene. Intergenic regions give rise to intergenic lncRNAs (Cabili et al, 2011). LncRNAs are also transcribed antisense to protein-coding genes (antisense lncRNA), or initiate from the same promoter, being transcribed divergently to the protein-coding gene (bidirectional lncRNAs). Similarly, PROMPTs (PROMoter uPstream Transcripts, (Preker et al, 2011)) are transcribed divergently to the protein-coding gene from its promoter, but differ to lncRNAs in being so unstable, depleted by the exosome (Preker et al, 2008), that they are usually not detected in the steady state RNA-seq data. Some lncRNAs can be transcribed sense to protein-coding genes and overlap them (sense lncRNAs) or be contained inside their introns (intronic lncRNAs), although the latter is disputed of being an independent transcript. Small lncRNAs, such as miRNAs (microRNAs), snoRNAs (small nucleolar RNAs), TASRs (termini-associated small RNAs) (Kapranov et al, 2007) and tiRNAs (transcription initiation (tiny) RNA) (Preker et al, 2008) are transcribed from within or next to protein-coding genes and contribute to the complexity of the genome and transcriptome. RNA classes are reviewed in (Clark et al, 2013).

1.2 Long non-coding RNAs: a new layer of information in the genome

LncRNAs are generally defined by their length (>200nt – a technical cutoff related to the RNA isolation method) and the absence of a protein-coding potential. It is important to note, however, that some well- and long-known ncRNAs, such as rRNAs and snoRNAs, though fulfilling these criteria, are usually not referred to as "lncRNAs" (Rinn & Chang, 2012), however, the nomenclature guidelines for lncRNAs are still being developed (Mattick & Rinn, 2015; Wright, 2014).

LncRNAs have only recently become appreciated as an important class of genes in the human genome (Morris & Mattick, 2014; Rinn & Chang, 2012). Today, just 28 years after the discovery of the first mammalian lncRNA in mouse (H19 lncRNA (Pachnis et al, 1988)), the lncRNA field is expanding exponentially (Figure 2) with over 1000 scientific publications on lncRNAs emerged within last year (year 2015, gray bars, Figure 2), while only approximately 100 publications on lncRNAs were emerging per year in 2000s (years 2000-2009, gray bars, Figure2). Moreover, lncRNAs are increasingly implicated in various diseases, such as cancer (Figure 2, orange bars) and a successful

use of a specific lncRNA as a therapeutic target in mouse disease models has already been reported (Meng et al, 2015).

Figure 2

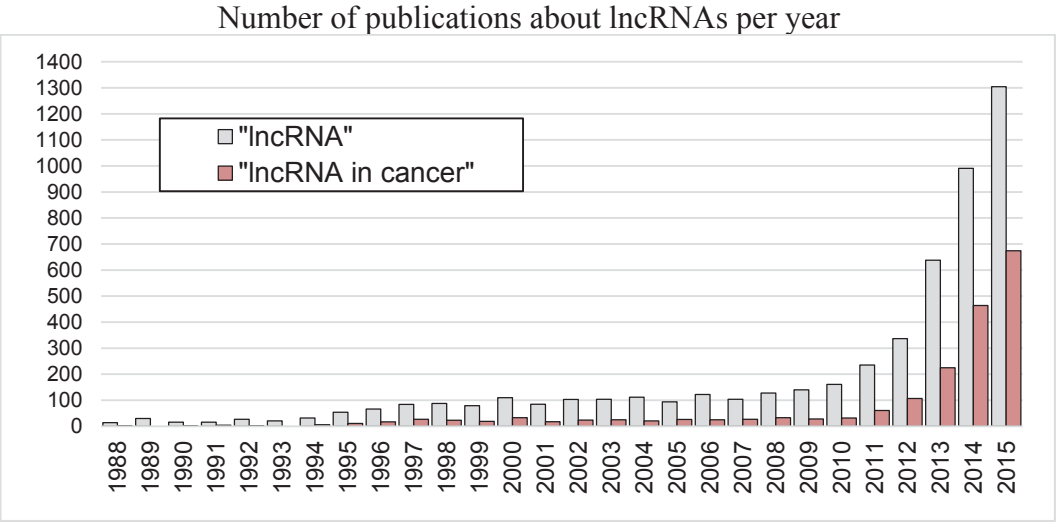


Figure 2. The number of studies on lncRNAs increased exponentially in the last decades. The bar plot shows the number of publications about lncRNAs in general (search word "lncRNA", gray bars) and lncRNAs involved in cancer (search words "lncRNA in cancer", orange bars). The search for the word combination in the publication title and/or abstract was performed in PubMed (http://www.ncbi.nlm.nih.gov/pubmed) and the number of publications per year was obtained from the "Results by year" section. Statistics for years 1988-2015 is shown.

Three decades after the discovery of *H19* and just a decade after the first indications of widespread lncRNA transcription were noted (Carninci et al, 2005), it is now clear, that lncRNAs are numerous in the genomes of all organisms (Ulitsky & Bartel, 2013), constitute the largest and the most diverse class of ncRNAs and likely outnumber protein-coding genes. Figure 3 briefly overviews milestones in the lncRNA research and gives an impression of its history and evolution.

Figure 3

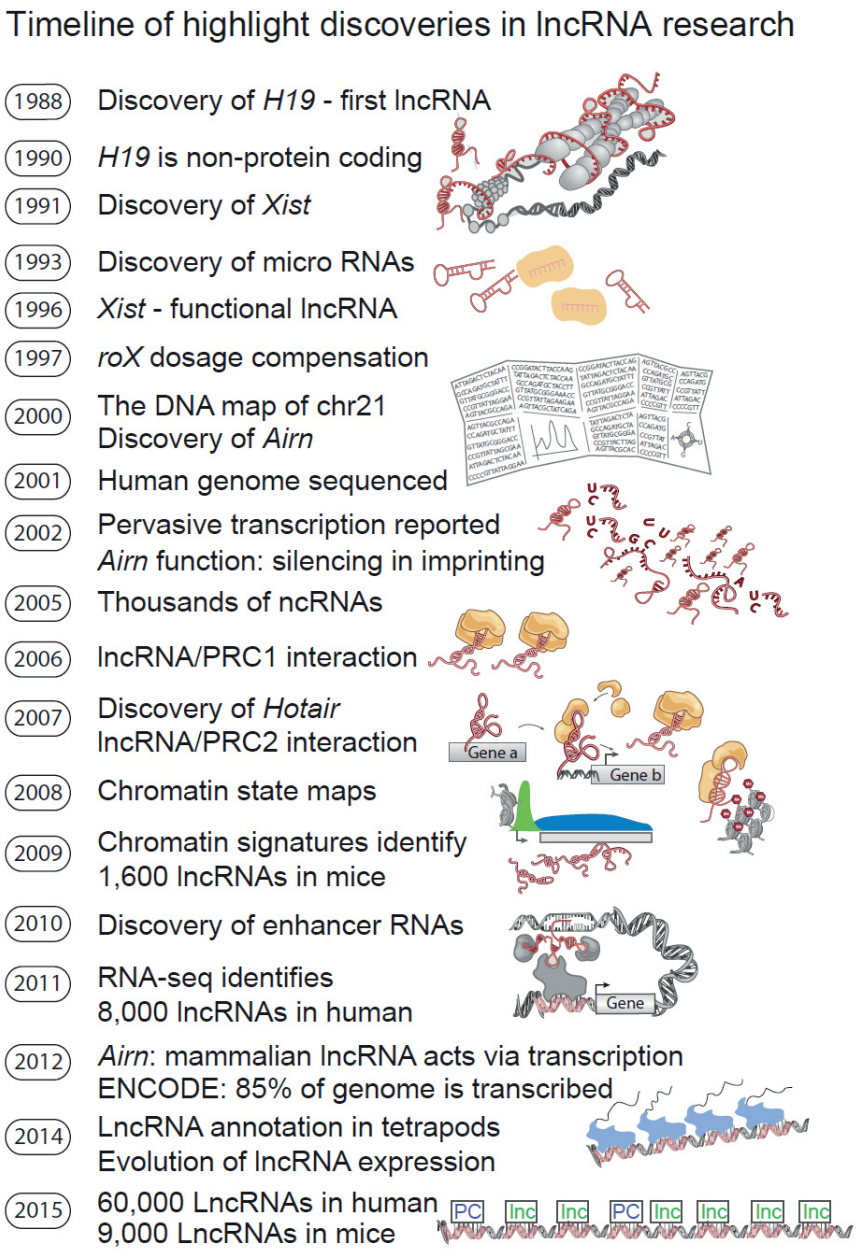


Figure 3. "Timeline of highlight discoveries in lncRNA research" (modified, supplemented and updated from Figure 1 from (Rinn & Chang, 2012)). **1988**: First lncRNA (first long non-protein-coding RNA that is not a part of the protein synthesis machinery, such as snRNAs and rRNAs) – *H19* – is discovered in mice while being first mistaken for an mRNA (Pachnis et al, 1988). **1990**: *H19* is reported to not encode a protein (Brannan et al, 1990). **1991**: *Xist* lncRNA is discovered in mice (Brown et al, 1991). **1993**: discovery of micro RNAs in *C. elegans* (Lee et al, 1993; Wightman et al, 1993). **1996**: *Xist* is shown to be required for chromosome X inactivation – first report of the functionality of a lncRNA (Penny et al, 1996). **1997**: discovery of *roX* lncRNA responsible for dosage compensation in drosophila (Amrein & Axel, 1997; Meller et al, 1997). **2000**: chromosome 21 is fully DNA-sequenced as a part of the Human Genome Project (Hattori et al, 2000); imprinted lncRNA *Airn* (*Air*) is discovered in mouse (Lyle et al, 2000). **2001**: The Human Genome Project is complete providing the first reference human genome sequence (Lander et al, 2001). **2002**: Unexpectedly broad transcriptional activity is reported throughout chromosomes 21 and 22 (Kapranov et al, 2002); *Airn* lncRNA is identified as a silencer of 3

imprinted genes in *Igf2r* imprinted cluster in mice (Sleutels et al, 2002). **2005**: thousands of non-coding RNAs are shown to be transcribed from the mouse genome (Carninci et al, 2005). **2006**: lncRNAs are shown to interact with Polycomb repressive complex 1 (PRC1) (Bernstein et al, 2006). **2007**: discovery of the lncRNA *Hotair* in mice: the first report of a lncRNA acting via binding to Polycomb repressive complex 2 (PRC2) (Rinn et al, 2007). **2008:** chromatin state maps are characterized for different cell types (Mendenhall & Bernstein, 2008). **2009:** chromatin signatures of active promoters and transcribed gene bodies identify ~ 1,600 active lncRNA in mice (Guttman et al, 2009). **2010**: lncRNAs transcribed from active enhancers are discovered (Orom et al, 2010). **2011**: RNA-seq analysis and *de novo* transcriptome assembly identifies ~6,000 active lncRNA genes in various human tissues (Cabili et al, 2011). **2012**: ENCODE project reports that the vast majority of the human genome (almost 90%) is transcribed in one or other cell type (Djebali et al, 2012). **2014**: identification of thousands of lncRNAs in several tetrapod species and analysis of their sequence and expression conservation (Necsulea et al, 2014). **2015**: RNA-seq analysis and *de novo* transcriptome assembly in thousands of malignant and normal samples from numerous donors identifies ~ 60,000 lncRNA genes in human (Iyer et al, 2015).

1.3 Functions and mechanisms of lncRNAs

The discovery of more and more lncRNAs in human and mice provided us with the knowledge that tens of thousands of lncRNAs reside in the genome, however, their function has been difficult to determine (see Chapters *1.8 Debate on lncRNA transcription meaningfulness* and *1.9 Assigning functionality to lncRNAs* below). Only a small number of lncRNAs have been assigned with a particular function and in much fewer cases has the mechanism underlying the function been identified (Kornienko et al, 2013; Quek et al, 2015; Rinn & Chang, 2012; Wang & Chang, 2011). Interestingly, thousands of mRNA genes in all organisms have precisely the same function at the RNA level – namely, their function is to transfer genomic information necessary for protein production from the nucleus to the cytoplasm where it is used as a template for ribosomes. At the same time, with just a few hundreds of functional lncRNAs known (Quek et al, 2015), it has become clear that lncRNAs may display an outstanding variety of functions and mechanisms (Nakagawa & Kageyama, 2014; Quek et al, 2015; Rinn & Chang, 2012) (Figure 4). Some LncRNAs were shown to organize functional nuclear domains by acting as a nuclear scaffold (Quinodoz & Guttman, 2014). Several others were shown to act as a guide for chromatin modifiers, and this mode of action is currently the most appreciated and well-known lncRNA function. For example, a well-known lncRNA *HOTAIR* guides Polycomb repressive complex 2 (PRC2) to target genes in the *HOXD* cluster on another chromosome, which causes H3 lysine-27 trimethylation (H3K27me3) of their promoters and consequent silencing (Gupta et al, 2010).

Another well-known lncRNA *HOTTIP* also acts as a guide, however, its action is only restricted to the nearby genes on the same chromosome, i.e. *in cis*, and it brings an activating complex to its target genes in *HOXA* cluster (Wang et al, 2011b). This shows that even within a similar basic function of guiding proteins to target genes – the functional outcome (i.e., activation or silencing of the target gene) will depend on which protein type the lncRNA interacts with. Another supposed function for lncRNAs is enabling and facilitating the action of genomic enhancers (Orom et al, 2010). Moreover, several lncRNAs have been shown to be involved in the regulation at post-transcriptional and post-translational level, yet again, by a variety of mechanisms, such as sponging miRNA, i.e., binding miRNAs via antisense homology and reducing their concentration in the cell, thus impeding the cleavage of a corresponding mRNA and upregulating its level in the cell (Poliseno et al, 2010). LncRNAs were shown to perform posttranslational regulation by, for example, binding proteins and thereby changing their conformation, which in turn can regulate transcription if this protein is a transcription factor, such as in case of *7SK* lncRNA binding to and inactivating PTEFβ (Peterlin et al, 2012). LncRNAs can also act post-translationally as protein decoys, binding target proteins and preventing their function (Duss et al, 2014; Kino et al, 2010) or as translation regulators (Carrieri et al, 2012).

Figure 4

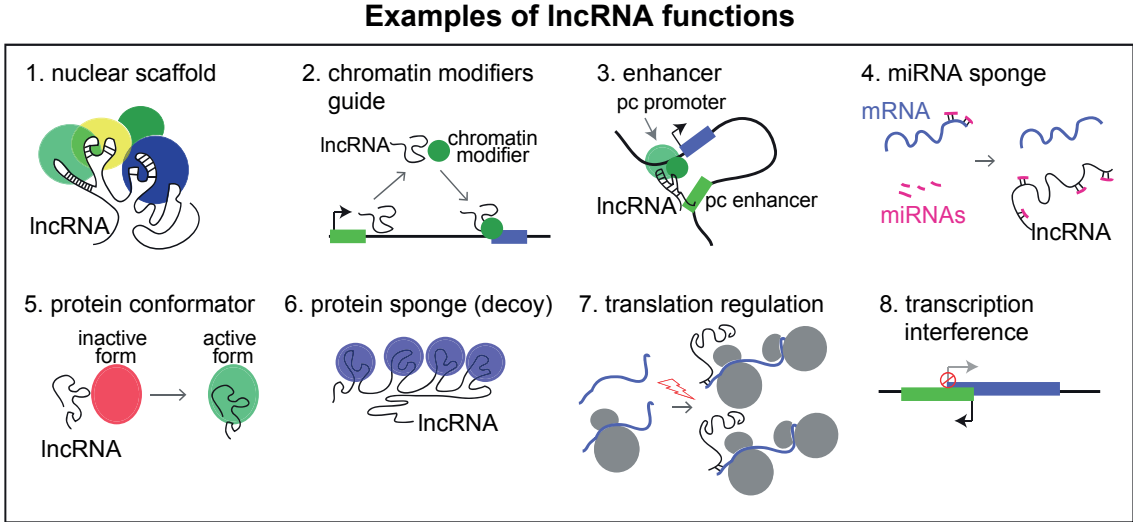


Figure 4. Various modes of lncRNA action in gene regulation. Example lncRNAs acting via the displayed mechanism: 1. *FIRRE* (Hacisuleyman et al, 2014), 2. *Hotair* (Gupta et al, 2010), 3. eRNAs (Orom et al, 2010), 4. *PTENP1* (Poliseno et al, 2010), 5. *7sk* (Peterlin et al, 2012), 6. RsmZ (Duss et al, 2014), 7. Antisense-*Uchl1* (Carrieri et al, 2012), 8. *Airn* (Latos et al, 2012).

Interestingly, while all the above examples involve the product of lncRNA gene transcription, i.e. the RNA molecules produced and active in the regulation, the transcription processes itself was also shown to be potentially functional. While examples of lncRNA acting through the mere act of their transcription and producing the lncRNA transcript as a by-product were mainly identified in yeast (reviewed in (Kornienko et al, 2013)), only one example of a mammalian lncRNA acting through transcription has recently been reported in mouse (Latos et al, 2012). However, it is likely that many lncRNAs in mammals function similarly, since many lncRNAs show features (such as lack of conservation, repeat richness, inefficient splicing – see Chapter *1.5 Understanding lncRNA biology* below) that together indicate a low probability of the lncRNA product being functional. The following Chapter 1.3.1 considers this transcription-centered mode of lncRNA action that is independent of the product in further detail.

1.3.1 Publication 1: "Gene regulation by the act of long non-coding RNA transcription" (Review)

Authors: Aleksandra E. Kornienko, Philipp M. Guenzl, Denise P. Barlow and Florian M. Pauler

Published in **BMC Biology** (Impact factor 7.98) on 30.05.2013.
Article's web page: http://www.biomedcentral.com/1741-7007/11/59
As of 01.02.2106 the article has a "Highly accessed" tag with 26,385 accesses and 96 citations.

A variety of lncRNAs have been identified in the genomes of human and other organisms, however, only a few hundreds of lncRNAs were assigned a function. Of these even fewer were studied mechanistically. The mammalian lncRNA field is dominated by the notion of lncRNAs acting via their transcript and bringing chromatin modifiers to the target genes thus regulating their expression. However, some studies in unicellular organisms (Houseley et al, 2008; van Werven et al, 2012), a few studies of gene stop signal mutations in human disease (Ligtenberg et al, 2009; Tufarelli et al, 2003) and a study of the mammalian lncRNA *Airn* (Latos et al, 2012), indicate that the mere act of lncRNA transcription possesses a potential for gene expression regulation. Our review overviews all the known modes of lncRNA action and summarizes the evidence for transcriptional interference being an important mechanism of lncRNA action.

In this review I created all the Figures and wrote all the text, except the Chapters "lncRNA transcription creating a permissive chromatin environment" and "lncRNA transcription and locus activation", which were contributed by Philipp M. Guenzl. Denise P. Barlow and Florian M. Pauler supervised the writing and revised and edited the manuscript prior to submission.

Kornienko et al. BMC Biology 2013, 11:59
http://www.biomedcentral.com/1741-7007/11/59

BMC Biology

REVIEW Open Access

Gene regulation by the act of long non-coding RNA transcription

Aleksandra E Kornienko, Philipp M Guenzl, Denise P Barlow and Florian M Pauler*

Abstract

Long non-protein-coding RNAs (lncRNAs) are proposed to be the largest transcript class in the mouse and human transcriptomes. Two important questions are whether all lncRNAs are functional and how they could exert a function. Several lncRNAs have been shown to function through their product, but this is not the only possible mode of action. In this review we focus on a role for the process of lncRNA transcription, independent of the lncRNA product, in regulating protein-coding-gene activity *in cis*. We discuss examples where lncRNA transcription leads to gene silencing or activation, and describe strategies to determine if the lncRNA product or its transcription causes the regulatory effect.

Keywords: Gene expression regulation, Histone modifications, lincRNA, lncRNA, Silencing, Transcriptional interference

LncRNAs - a new layer of genome regulatory information

It is now well appreciated that less than two percent of the human genome codes for proteins and the majority of the genome gives rise to non-protein-coding RNAs (ncRNAs) [1], which are predicted to play essential roles in a variety of biological processes [2,3].

The focus of this review is long ncRNAs (known as lncRNAs), which constitute the biggest class of ncRNAs with approximately 10,000 lncRNA genes so far annotated in humans [4]. lncRNAs are RNA polymerase II (RNAPII) transcripts that lack an open reading frame and are longer than 200 nucleotides. This size cut-off distinguishes lncRNAs from small RNAs such as microRNAs, piwi-interacting RNAs (piRNAs), small nucleolar RNAs (snoRNAs) and small interfering RNAs (siRNAs) and arises from RNA preparation methods that capture RNA molecules above this size. Although the function of most lncRNAs is unknown, the number of characterized lncRNAs is growing and many publications suggest they play roles in negatively or positively regulating gene expression in development, differentiation and human disease [2,5-10]. lncRNAs may regulate protein-coding (pc) gene expression at both the post-transcriptional and transcriptional level. Posttranscriptional regulation could occur by lncRNAs acting as competing endogenous RNAs to regulate microRNA levels as well as by modulating mRNA stability and translation by homologous base pairing, or as in the example of NEAT1 that is involved in nuclear retention of mRNAs [11]. In this review we focus on the regulation at the transcriptional level.

Modes of transcriptional regulation by lncRNAs

Regulation of transcription is considered to be an interplay of tissue and developmental-specific transcription factors (TFs) and chromatin modifying factors acting on enhancer and promoter sequences to facilitate the assembly of the transcription machinery at gene promoters. With a growing number of lncRNAs implicated in transcriptional gene regulation, this view may need refinement to include networks of tissue and developmental-stage specific lncRNAs that complement known regulators to tightly control gene expression and thereby organism complexity [12,13]. Transcriptional regulation by lncRNAs could work either in *cis* or in *trans*, and could negatively or positively control pc gene expression. lncRNAs work in *cis* when their effects are restricted to the chromosome from which they are transcribed, and work in *trans* when they affect genes on other chromosomes.

Regulation in *trans*

Some significant examples of lncRNAs that act in *trans* are those that can influence the general transcriptional output of a cell by directly affecting RNAPII activity (Figure 1a,b). One example is the 331 nucleotide 7SK

* Correspondence: fpauler@cemm.oeaw.ac.at
CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Lazarettgasse 14, AKH-BT25.3, 1090, Vienna, Austria

lncRNA, which represses transcription elongation by preventing the PTEFβ transcription factor from phosphorylating the RNAPII carboxy-terminal domain (CTD) [14] (Figure 1a). Another example is the 178 nucleotide B2 lncRNA, a general repressor of RNAPII activity upon heat shock [15]. The B2 lncRNA acts by binding RNAPII and inhibiting phosphorylation of its CTD by TFIIH, thus disturbing the ability of RNAPII to bind DNA [16,17].

Regulation in *trans* can also act locus-specifically. While the ability of lncRNAs to act locus-specifically to regulate a set of genes was first demonstrated for imprinted genes where lncRNA expression was shown to silence from one to ten flanking genes in *cis* [18-20], lncRNAs that lie outside imprinted gene clusters, such as the *HOTAIR* lncRNA, were later found also to have locus-specific action. *HOTAIR* is expressed from the *HOXC* cluster and was shown to repress transcription in *trans* across 40 kb

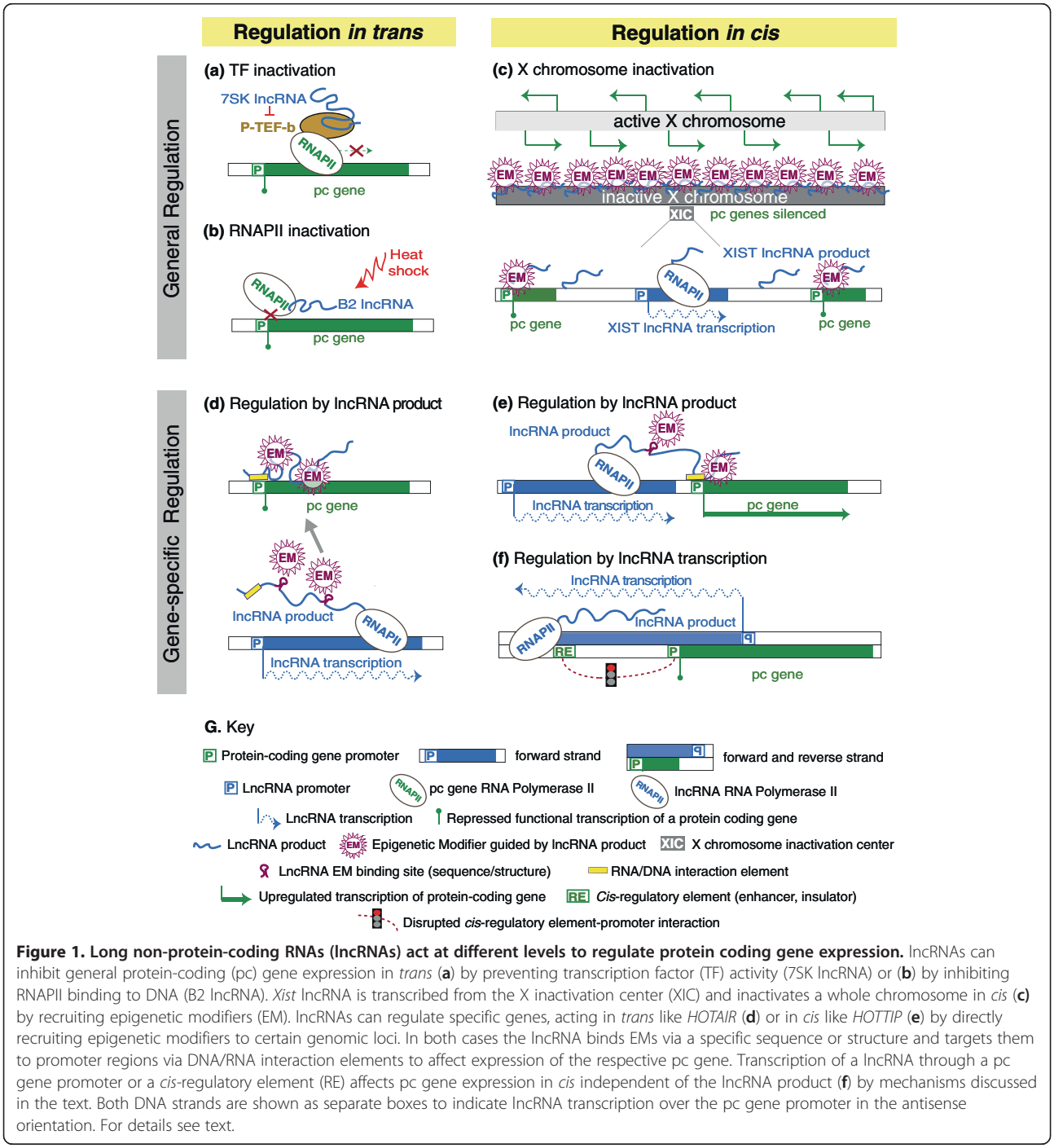


Figure 1. Long non-protein-coding RNAs (lncRNAs) act at different levels to regulate protein coding gene expression. lncRNAs can inhibit general protein-coding (pc) gene expression in *trans* (**a**) by preventing transcription factor (TF) activity (7SK lncRNA) or (**b**) by inhibiting RNAPII binding to DNA (B2 lncRNA). *Xist* lncRNA is transcribed from the X inactivation center (XIC) and inactivates a whole chromosome in *cis* (**c**) by recruiting epigenetic modifiers (EM). lncRNAs can regulate specific genes, acting in *trans* like *HOTAIR* (**d**) or in *cis* like *HOTTIP* (**e**) by directly recruiting epigenetic modifiers to certain genomic loci. In both cases the lncRNA binds EMs via a specific sequence or structure and targets them to promoter regions via DNA/RNA interaction elements to affect expression of the respective pc gene. Transcription of a lncRNA through a pc gene promoter or a *cis*-regulatory element (RE) affects pc gene expression in *cis* independent of the lncRNA product (**f**) by mechanisms discussed in the text. Both DNA strands are shown as separate boxes to indicate lncRNA transcription over the pc gene promoter in the antisense orientation. For details see text.

of the *HOXD* cluster [21]. *HOTAIR* interacts with Polycomb repressive complex 2 (PRC2) and is required for repressive histone H3 lysine-27 trimethylation (H3K27me3) of the *HOXD* cluster. Targeting of epigenetic modifiers (EMs) by lncRNAs provided a much sought after model to explain how EMs gain locus specificity (Figure 1d), and has since been suggested as a general mechanism for *trans*-acting lncRNAs [22,23].

Regulation in *cis*

In contrast to *trans*-acting lncRNAs, which act via their RNA product, *cis*-acting lncRNAs have the possibility to act in two fundamentally different modes. The first mode depends on a lncRNA product. The major example of general *cis*-regulation is induction of X inactivation by the *Xist* lncRNA in female mammals. *Xist* is expressed from one of the two X chromosomes and induces silencing of the whole chromosome [24] (Figure 1c). As an example of locus-specific regulation it has been proposed that enhancer RNAs activate corresponding genes in *cis* via their product [25]. A well-studied *cis*-acting lncRNA acting through its product is the human HOTTIP lncRNA that is expressed in the HOXA cluster and activates transcription of flanking genes. HOTTIP was shown to act by binding WDR5 in the MLL histone modifier complex, thereby bringing histone H3 lysine-4 trimethylation (H3K4me3) to promoters of the flanking genes [26]. Such a mechanism in which a nascent lncRNA transcript binds and delivers epigenetic modifiers to its target genes while still attached to the elongating RNAPII is generally termed 'tethering' and is often used to explain *cis*-regulation by lncRNAs [23,27] (Figure 1e). It was also proposed to act in plants. In *Arabidopsis thaliana*, the *COLDAIR* lncRNA is initiated from an intron of the *FLC* pc gene and silences it by targeting repressive chromatin marks to the locus to control flowering time [28].

In contrast, the second mode of *cis* regulation by lncRNAs involves the process of transcription itself, which is *a priori cis*-acting (Figure 1f). Several lines of evidence suggest that the mere process of lncRNA transcription can affect gene expression if RNAPII traverses a regulatory element or changes general chromatin organization of the locus. In this review we discuss this underestimated role for lncRNA transcription in inducing protein-coding gene silencing or activation in *cis*, and overview possible mechanisms for this action in mammalian and non-mammalian organisms. Finally, we describe experimental strategies to distinguish lncRNAs acting as a transcript from those acting through transcription.

Mechanisms by which lncRNA transcription silences gene expression

Transcription-mediated silencing, also referred to as 'transcriptional interference' (TI), is defined here as a case in which the act of transcription of one gene can repress in *cis* the functional transcription of another gene [29,30]. TI has been reported in unicellular and multicellular organisms [30]. Mechanistic details are still largely unclear, but TI could theoretically act at several stages in transcription: by influencing enhancer or promoter activity or by blocking RNAPII elongation, splicing or polyadenylation. All that would be required is that the RNA polymerase (RNAPII) initiated from an 'interfering' promoter traverses a 'sensitive' DNA regulatory sequence. TI has mainly been reported at overlapped promoters [31-35], but there are also examples where TI acts downstream of the promoter. In mouse, overlapping transcription controls polyadenylation choice of two imprinted genes [36,37]. In *Saccharomyces cerevisiae*, collisions between elongating antisense RNAPIIs can lead to stalling of both polymerases that is resolved by ubiquitylation-directed proteolysis, and this has been proposed to be a regulatory mechanism [38]. However, it is unknown if RNAPII collisions occur sufficiently frequently *in vivo* in yeast or other organisms to offer a means of regulating convergent genes, or if this mechanism could lead to an interfering RNAPII eliminating its sensitive collision partner. Despite these examples, the most common reports of TI concern an overlapped promoter, and in the following sections we describe studies investigating the molecular mechanisms underlying interference at the promoter.

Transcriptional interference acting by promoter nucleosome repositioning

DNA in the nucleus is organized into chromatin with the organizational scaffold consisting of nucleosomes, each with two copies of H3, H4, H2A and H2B histones [39]. Nucleosomes can be densely packed, interfering with protein-DNA interactions, or relaxed, facilitating these interactions [40]. The transcription process, which generates single-stranded DNA as RNAPII progresses along a gene locus, can directly affect nucleosome positioning [41-43] (reviewed in [44,45]). Thus, lncRNA transcription could cause TI by depositing nucleosomes in a manner unfavorable for TF binding on promoters or enhancers. An example of this mechanism is the silencing of the yeast *SER3* pc gene by transcriptional overlap by the *SRG1* lncRNA (Figure 2a) [46]. *SRG1* transcription increases nucleosome density at the overlapped *SER3* promoter. Deletion of three transcription elongation factors that are associated with the elongating polymerase and are necessary for nucleosome repositioning (*SPT16*, *SPT6*, *SPT2*) [47-49] abolished the silencing effect without stopping transcription of the overlapping lncRNA *SRG1* [50,51], indicating the necessity of chromatin reorganization for silencing. In contrast, deletion of epigenetic modifiers (such as SET1/2

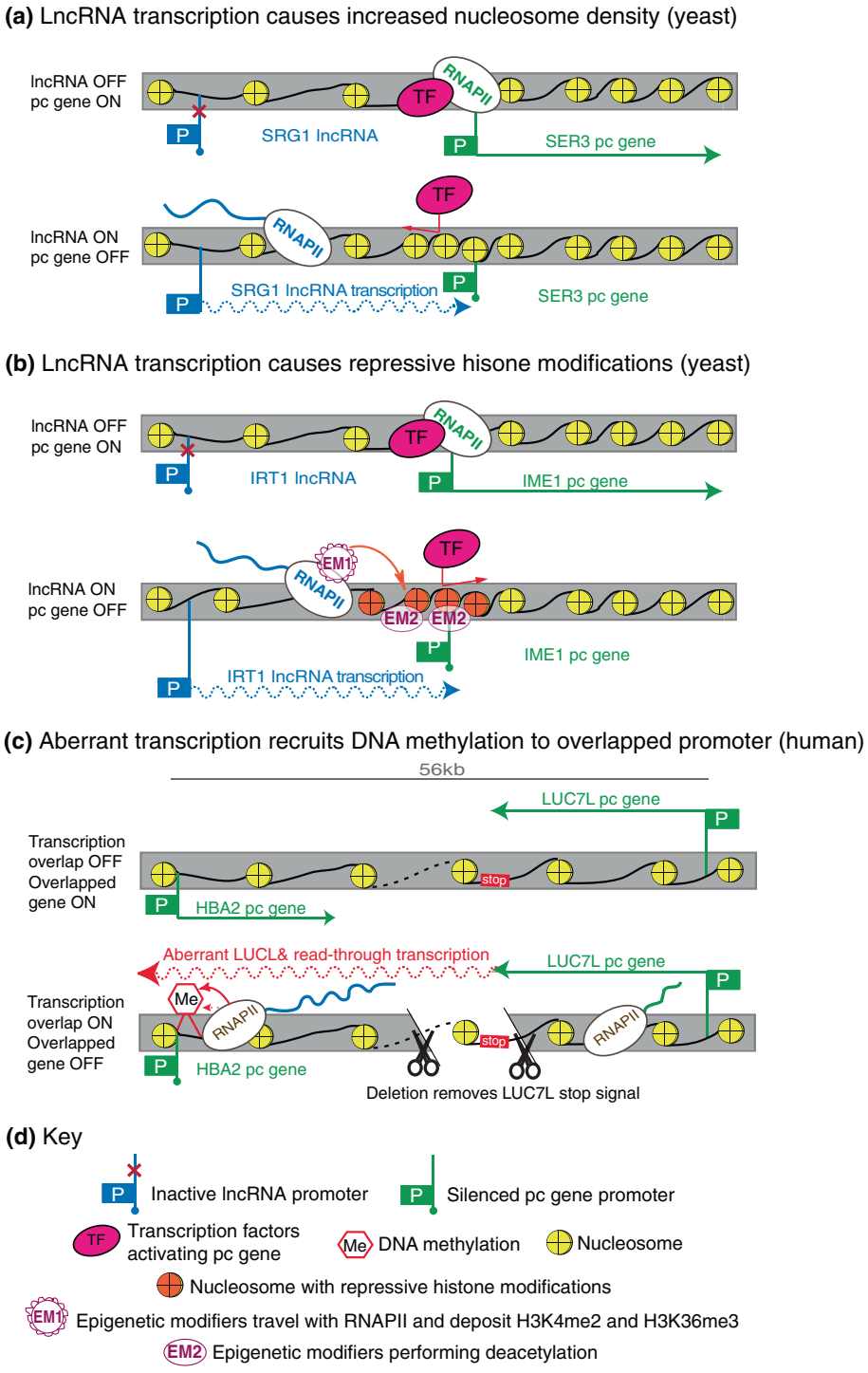


Figure 2. Transcription interference-mediated silencing by chromatin changes. (**a**) Top: in yeast the absence of *SRG1* lncRNA allows transcription machinery assembly at the *SER3* protein coding gene promoter. Bottom: *SRG1* lncRNA transcription causes dense nucleosome packing over the downstream *SER3* pc gene promoter that blocks TF binding and pc gene expression. (**b**) Top: in yeast the absence of *IRT1* lncRNA allows *IME1* pc gene expression. Bottom: RNAPII transcribing the *IRT1* lncRNA carries EMs that deposit repressive histone modifications at the *IME1* promoter (EM1 - methyltransferases). These modifications allow the binding of other EMs that remove active histone modifications (EM2 - deacetylases) and cause a repressive chromatin environment that blocks TF binding leading to silencing. (**c**) Top: in a healthy human, *LUC7L* and *HBA2* pc genes do not overlap and are both expressed. Bottom: a chromosomal deletion of the *LUC7L* transcriptional stop signal (red 'stop' box) causes transcription of the *LUC7L* pc gene through the promoter of the *HBA2* pc gene. By an unknown mechanism this aberrant transcription causes DNA methylation and silencing of the *HBA2* promoter. For details see Figure 1g and text.

histone methyltransferases and SET3C/RPD3S deacetylases described later) did not affect silencing, showing that nucleosome positioning, but not changes in histone modifications, is responsible for repression. The experiments did not directly exclude a role for the *SRG1* lncRNA product, but the silencing can be explained solely by the process of transcription [44,45]. TI by nucleosome repositioning may be a general mechanism in yeast, as the RNAPII elongation and chromatin organization factors responsible for *SER3* silencing are also known to be involved in the suppression of transcription initiation from cryptic promoters within the body of actively transcribed genes [52,53]. Since genes controlling RNAPII elongation and chromatin organization are largely conserved, it is possible that lncRNAs could use similar nucleosome repositioning silencing in mammals. This is supported by the example that chromatin reassembly factors are necessary for silencing an HIV provirus when integrated into an actively transcribed host gene in a human cell system [54].

Transcriptional interference acting by promoter histone modifications

Promoter associated nucleosomes carry post-translational histone tail modifications that reflect the activity state of the promoter and also influence accessibility of DNA binding factors involved in transcription [55]. Active gene promoters correlate with H3 and H4 acetylation and with H3K4me3, while inactive promoters do not and, in mammals, they also gain repressive histone marks such as H3K9me3 or H3K27me3. Some histone modifying enzymes have been shown to bind and travel with elongating RNAPII [56,57], so it is possible that lncRNA transcription can induce TI by affecting histone modifications at the promoter of an overlapped target gene. For example, in yeast the SET1/2 methyltransferases, which induce H3K4me2 and H3K36me3 in the body of transcribed genes, bind and travel with elongating RNAPII [58-60]. These modifications in turn recruit the SET3C/RPD3S histone deacetylase complexes to create a chromatin environment repressive for transcription initiation [61-63].

Two studies indicate that this is a mechanism used by lncRNAs to induce TI in yeast. In the first study the *IME1* pc gene, which induces gametogenesis in diploid *S. cerevisiae* cells but is repressed in haploid cells, was shown to be silenced by the *IRT1* lncRNA that overlaps its promoter [64]. Genetic experiments repositioning the *IRT1* lncRNA distant from *IME1* on the same chromosome showed that *IRT1* transcriptional overlap of the *IME1* promoter is necessary for silencing. Interestingly, the instability of the *IRT1* lncRNA product and its non-specific cellular localization indicated the lncRNA product is unlikely to play a role in the silencing mechanism. Instead, *IRT1* lncRNA transcription through the *IME1* promoter reduced recruitment of the essential POG1 transcription factor, increased nucleosome density and induced the SET1/2 mediated cascades of histone modifications, which were shown to be necessary for silencing [64] (Figure 2b). In the second study lncRNA transcription was shown to be causative for silencing of the *GAL1* and *GAL10* genes, involved in galactose metabolism in *S. cerevisiae*. *GAL10* and *GAL1* are divergently transcribed from a bidirectional promoter. The 4 kb lncRNA, called *GAL10*-ncRNA, initiates in the body of the *GAL10* gene, and is transcribed through the *GAL10/GAL1* promoter antisense to the *GAL10* gene. *GAL10*-ncRNA transcription induces SET2-mediated establishment of H3K36me3 along its gene body, thereby recruiting RPD3S-dependent deacetylation that resulted in reduced transcription factor binding and repression of the *GAL1/GAL10* promoter [65]. Both SET3C and RPD3S are proposed to have a general role in repressing cryptic promoters within gene bodies [61,66] and a genome-wide study implied a role for SET3C in overlapping lncRNA-mediated silencing of a set of pc genes in yeast [66]. This indicates that the mechanism described above might be widely used to control gene expression in yeast. Although similar studies have not been described for the mammalian genome, H3K36me3 marks the body of transcribed genes in mammals, raising the possibility that such TI mechanisms could be conserved [56,57].

Transcriptional interference acting by promoter DNA methylation

In mammalian genomes DNA methylation is generally associated with silent CpG island promoters, but the majority of CpG island promoters remain methylation free independent of their expression status [67-69]. The process of *de novo* methylation depends on the DNMT3A/3B methyltransferases and the catalytically inactive DNMT3L homologue and requires histones lacking H3K4me3, ensuring that active promoters remain methylation-free [70]. Notably, while DNA methylation at the promoter blocks transcription initiation, methylation in the gene body does not. Two important examples in humans based on genetic analyses indicate that DNA methylation can be involved in TI-induced silencing, although the causality between DNA methylation and silencing is still a matter of discussion [67]. One study of a patient with inherited α-thalassemia identified a deletion of the *LUC7L* 3' end that allowed aberrant transcription of *LUC7L* through the downstream *HBA2* gene, causing its silencing and the disease phenotype [71] (Figure 2c). Mouse models that mimicked the deleted genomic locus showed that the main cause of silencing was the acquisition of DNA methylation at the *HBA2* promoter. Notably, DNA methylation acquisition was not simply the consequence of an inactive promoter, as removal of *HBA2* transcription by deleting its TATA box did not induce methylation. The sequence of the

LUC7L gene and thus the aberrant RNA product was also not essential for *HBA2* silencing, as replacing the *LUC7L* gene body with another protein-coding gene did not remove the repressive effect. In a second example, a subset of Lynch syndrome patients display DNA methylation and inactivation of the mismatch repair *MSH2* gene that correlates with aberrant transcription from the flanking *EPCAM* gene that carries a 3' deletion [72].

In both these examples, the molecular details of methylation establishment and the mechanism by which the methylation machinery targets the overlapped promoter are yet unknown. However, the data so far show that it is a *cis*-acting mechanism as only the allele carrying the deletion silences the overlapped protein-coding gene. In addition, although a role for the aberrant RNA product was not excluded, it appears unlikely that mutation-induced transcription of two independent intergenic chromosomal regions in the described diseases produces lncRNA products with similar repressive functions. Interestingly, the silencing of imprinted pc genes by lncRNAs is also often correlated with the gain of DNA methylation on the silent pc gene promoter [73]. In the case of the *Igf2r* gene, this DNA methylation mark is not necessary for initiation or maintenance of the silent state but seems to play a role in re-enforcing the silent state [35,74].

Transcriptional interference in the absence of chromatin changes at the silenced promoter

In addition to RNAPII acting as a carrier of chromatin modifying enzymes, other TI models predict that RNAPII from one promoter traversing across another promoter can interfere with its activity without introducing chromatin changes [30,75,76]. An indication that such a mechanism can be used by lncRNAs in mammals comes from a study that used a genetic approach to dissect the silencing function of the imprinted mouse *Airn* lncRNA [77,78]. *Airn* is an inefficiently spliced 118 kb lncRNA expressed on paternally inherited chromosomes that overlaps and silences the promoter of the *Igf2r* pc gene - a dose-sensitive and essential embryonic growth suppressor [18,79] (Figure 3a). To determine if *Airn* transcription or its lncRNA product were required for silencing, homologous recombination in embryonic stem cells was used to shorten the length of *Airn*, either before or after the *Igf2r* promoter, by insertion of a polyadenylation cassette [35]. Notably, only shortened *Airn* variants that traversed the *Igf2r* promoter induced silencing. Furthermore, while *Igf2r* silencing is normally accompanied by DNA methylation, repressive histone marks and chromatin compaction of the silent *Igf2r* promoter [80,81], *Igf2r* silencing was not dependent on DNA methylation - in contrast to the silencing of *HBA2* by aberrant *LUC7L* transcription described above. Instead, *Airn* transcriptional overlap interfered with the accumulation of functional RNAPII on the *Igf2r* promoter in the presence of open chromatin [35]. Additional support for *Igf2r* silencing by *Airn* transcriptional interference is provided by genetic experiments that used an inducible *Airn* promoter to silence *Igf2r* at different stages of embryonic stem cell differentiation [74]. The demonstration that *Airn* transcription is continuously required for *Igf2r* silencing and that its silencing efficiency decreases when the *Igf2r* promoter is strongly expressed provides support for a model whereby RNAPII initiated from an 'interfering' promoter interferes with transcription initiation from a 'sensitive' promoter.

To date, other examples of lncRNAs acting by this mechanism in mammals are lacking. It has been suggested that silencing of an alternative promoter of the mouse *fpgs* pc gene is an example of transcription inducing silencing without introducing chromatin changes [82], but this system has not been subject to a similar genetic analysis and alternative explanations remain possible. How RNAPII from an interfering promoter is able to suppress functional transcription of the overlapped promoter remains to be determined, but stalling of the interfering RNAPII elongating over the sensitive promoter has been suggested to block access of essential TFs [30,83]. This mechanism should not be confused with the phenomenon of genome-wide RNAPII pausing at promoters, which represents an intermediate step between RNAPII initiation and elongation phases and might be a common mechanism regulating differential gene expression in metazoans [84,85].

The above examples describe repressive effects from RNAPII transcribing lncRNAs through promoters of silenced genes. However, transcriptional interference might also disrupt enhancer function when RNAPII traverses an enhancer, and this is an attractive model to explain the repression of a cluster of genes by a lncRNA in a tissue-specific manner [75] (Figure 3b). This situation arises in two imprinted gene clusters where the *Airn* and *Kcnq1ot1* lncRNAs each overlap one gene, but silence multiple genes in *cis* in a tissue-specific manner. The repressive histone EHMT2 methyltransferase has been shown to be necessary in the placenta to silence one of the three genes controlled by *Airn* [86]. The *Kcnq1ot1* lncRNA has been shown to silence multiple genes in placental cells by the action of repressive POLYCOMB histone modifying enzymes [87,88]. In both cases, a direct role for the lncRNA in targeting the histone modifying complexes was proposed, based on the findings that the lncRNAs interact with the respective histone modifying complex. This correlation-based evidence is, however, not sufficient to rule out the possibility that both lncRNAs silence distant genes by transcription alone (reviewed in [75,76]). In support of a transcription-based

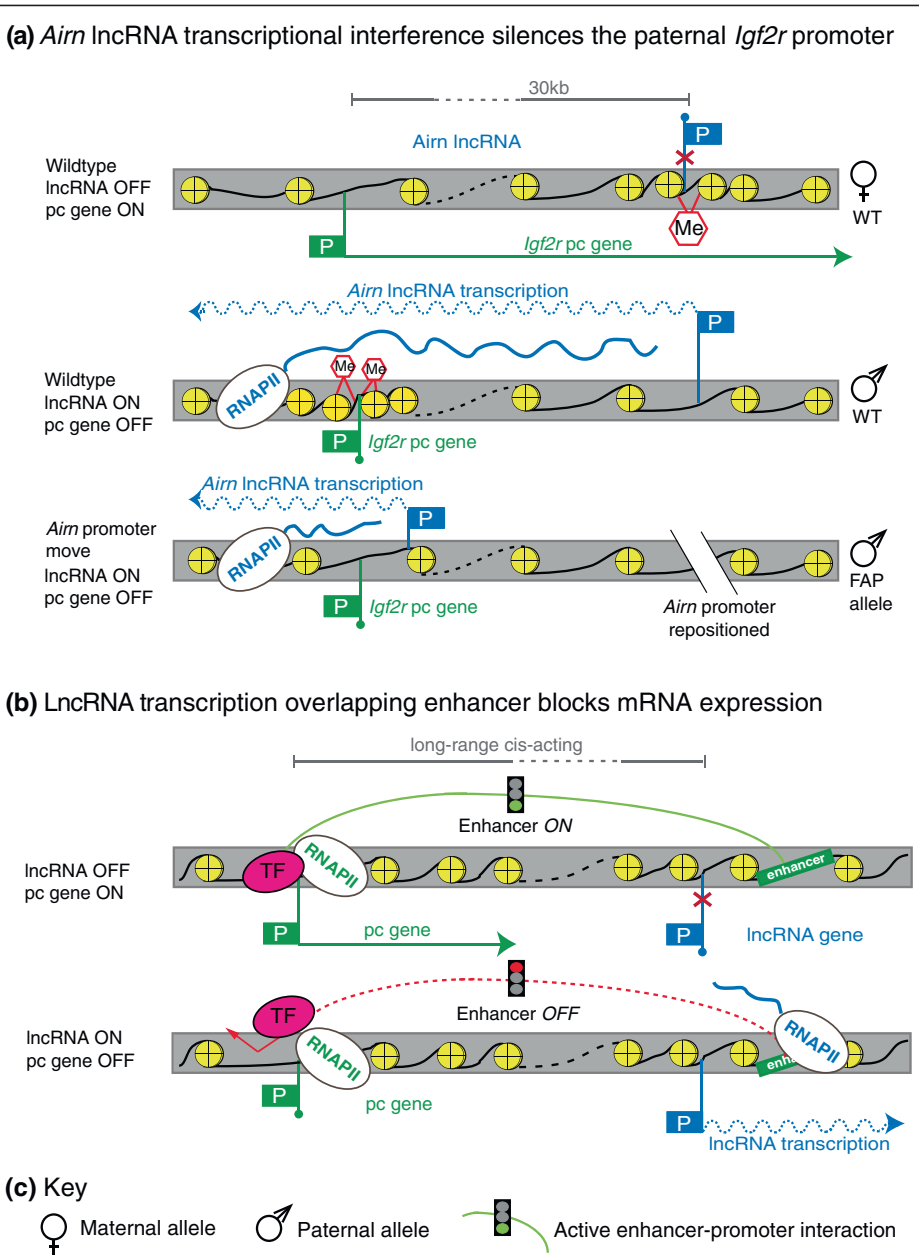


Figure 3. Transcription interference-mediated silencing without chromatin changes. (**a**) Top: a wild-type maternal allele does not express *Airn* lncRNA as its promoter is repressed by a DNA methylation imprint, thus allowing the *Igf2r* gene to be active. Middle: on the wild-type paternal allele *Airn* transcription overlaps with and silences the *Igf2r* pc gene promoter, independent of the *Airn* lncRNA product. The silent *Igf2r* promoter is marked by increased nucleosome density and DNA methylation in the absence of active histone modifications. Bottom: increased nucleosome density, loss of active histone marks and DNA methylation are not necessary for *Igf2r* repression as demonstrated by the FAP allele that moved the *Airn* promoter close to the *Igf2r* promoter and silenced *Igf2r* in the absence of repressive chromatin features. (**b**) Top: a hypothetical enhancer activates a pc gene by direct long-range DNA interactions. Bottom: transcription of a lncRNA overlapping the enhancer interferes with the DNA interaction and thereby silences the pc gene. For details see Figure 1g, Figure 2d and text.

model, it was shown that *Kcnq1ot1* silences at least one gene by regulating chromatin flexibility and access to enhancers [89]. This is consistent with a two-step model whereby lncRNA transcription initiates silencing of non-overlapped genes by enhancer interference, then repressive histone modifying enzymes maintain that silencing.

lncRNA transcription creating a permissive chromatin environment

Enhancers are genetic elements that bind transcription factors facilitating transcription machinery assembly at nearby promoters [90,91]. RNAPII transcripts up to 2 kb long are transcribed bi-directionally from some neuronal

enhancers (termed enhancer or eRNAs) [91,92]. Transcription of eRNAs positively correlated with expression of nearby mRNAs and a model was proposed, but not yet experimentally tested, in which their transcription establishes a chromatin landscape that supports enhancer function (Figure 4a). lncRNA transcription, either by opening chromatin or inhibiting repressor protein binding, could similarly result in gene or locus activation. One example of this is the process of V(D)J recombination, which joins elements of the V, D and J multigene family by chromosomal rearrangements to create functional B cell immunoglobulins and T cell receptors [93] (Figure 4b). The V, D and J genes lie next to each other on the same chromosome and antisense intergenic transcription through these genes is detected prior to the recombination process [94]. Genetic experiments have shown that intergenic lncRNA transcription is required for both B and T cell V(D)J recombination [95,96]. Similar correlations between intergenic transcription and gene expression were observed for the mouse *β-globin* locus [97] where promoter deletion experiments showed that lncRNA transcription was responsible for stable, active and hyper-accessible chromatin [98].

lncRNA transcription and locus activation

Other examples indicate that lncRNA transcription activates gene expression by blocking access of repressor complexes to chromatin. In *Drosophila*, intergenic noncoding transcription at the *BITHORAX* complex (*BX-C*) is implicated in reversing *POLYCOMB* group (*PCG*)-mediated gene silencing and is correlated with an active chromatin state [99]. This mode of action was later suggested to be a general mechanism where the act of transcription serves as an epigenetic switch that relieves *PCG*-mediated gene silencing by recruiting epigenetic modifiers to induce gene expression and generate stable and heritable active chromatin [100]. In line with this hypothesis, intergenic transcription through *PCG* response elements (PREs) in the *BX-C* cluster is not only found during embryogenesis but also in late stage larvae, indicating that continuous transcription is required to keep genes active [101]. In mouse and human, a similar role for PRE transcription has been proposed. An analysis of lncRNA transcription in the human *HOXA* cluster revealed a positive correlation between lncRNA transcription and the loss of PCG/chromatin interactions that precedes *HOXA* gene activation [102]. Additionally, lncRNAs have been identified at promoter regions of PCG-regulated genes in mouse cells; while their role is not yet clear, it has been suggested that they either promote or interfere with PCG binding at target genes [103,104].

A further example of a lncRNA mediating chromatin opening was described at the *S. cerevisiae PHO5* gene. Transcription of an antisense lncRNA that initiates near the 3'end of *PHO5* and overlaps its gene body and promoter is associated with rapid activation of *PHO5* by enabling nucleosome eviction. Biochemical inhibition of RNAPII elongation as well as genetic disruption of lncRNA elongation demonstrated a direct role in *PHO5* activation [105]. The association of lncRNA transcription with gene activation needs, however, to be considered within the framework that most protein-coding gene promoters in yeast and mammalian cells give rise to a bidirectional antisense lncRNA transcript [106,107]. To date it is unclear if promoter-associated bidirectional lncRNAs represent spurious transcription in the context of open chromatin [108,109] or is required to maintain open chromatin. In the latter case enhanced TF binding ensures accessible chromatin that allows more constant pc gene expression within a cell population [110] (Figure 4c).

Strategies for distinguishing a role for the lncRNA product from that of its transcription

Following genome-wide lncRNA mapping, functional studies so far have mainly focused on lncRNA products [7,111]. As it becomes clear that lncRNAs can act through their transcription, it is important to identify strategies to determine the function and mode of action of each particular lncRNA. One common starting point to determine lncRNA function has been RNA interference (RNAi)-mediated knockdown, despite long-standing observations that the RNAi machinery in mammalian cells is located in the cytoplasm [112]. While there is evidence that some RNA-induced silencing complex (RISC) components are found in the nucleus, functional complexes are specifically loaded in the cytoplasm, prohibiting the application of RNAi strategies for nuclear localized lncRNAs [113]. In contrast, antisense oligonucleotides (ASO) that work via an RNaseH-dependent pathway will deplete nuclear-localized lncRNAs [114,115]. However, three additional points of caution should be noted. First, non-specific effects arising from nuclear transfection reagents [116] have confused some observations. One critical validation step for knockdown studies would be a rescue experiment in which the lncRNA, modified to be invulnerable to the knockdown, is expressed as a transgene under the same transfection conditions [111]. Second, some results have highlighted major differences when functional studies used post-transcriptional depletion strategies in cell lines in contrast to genetic studies in the organism. Notable examples are *Neat1* [117], *Malat1* [116,118,119] and *Hotair* [120] where studies of mice carrying genetically disrupted alleles of these three lncRNAs failed to reproduce phenotypes deduced from cell lines following RNAi, ASO or over-expression studies. Third, while knockdown experiments may elucidate the function of lncRNAs acting through their product, the function of *cis*-acting lncRNAs that depend only on transcription will not be disturbed.

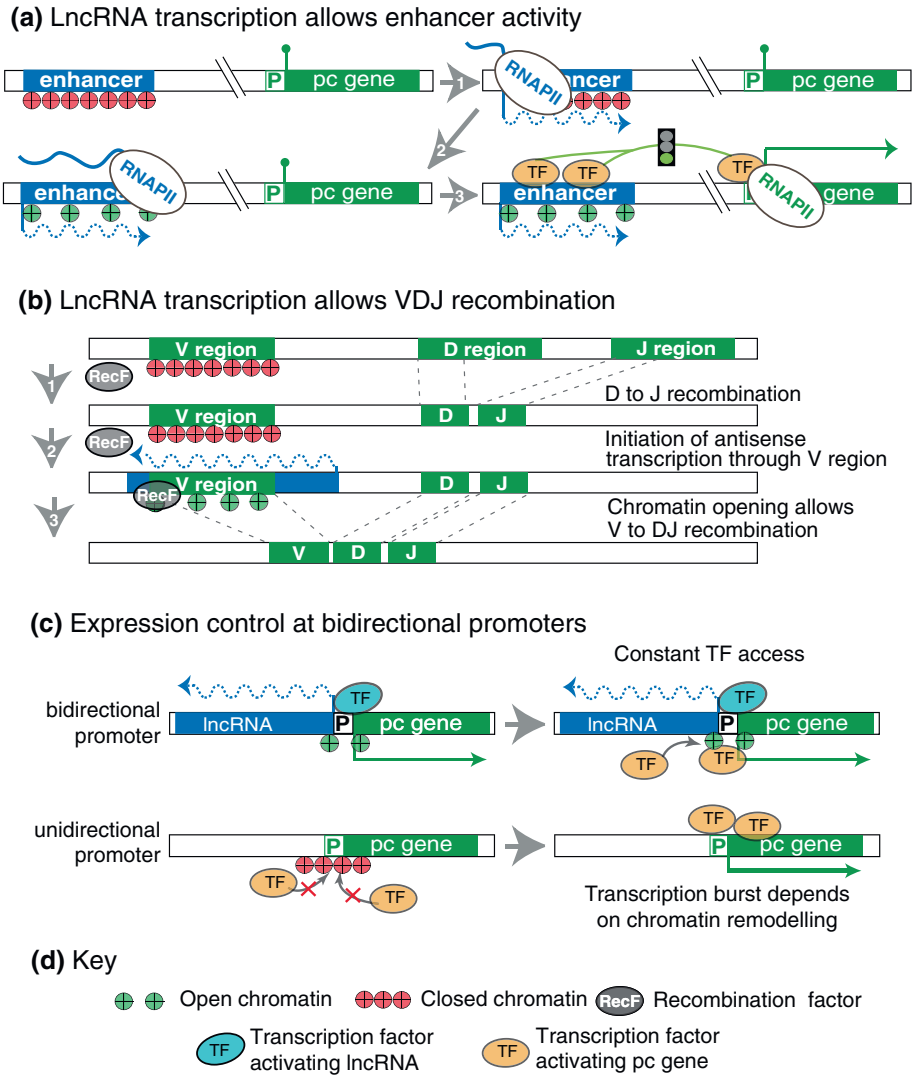


Figure 4. Transcription of lncRNA creates permissive chromatin environment. (**a**) Top left: an inactive enhancer with closed chromatin cannot activate the pc gene. Top right, bottom left: transcription of the enhancer opens chromatin. Bottom right: open chromatin at the enhancer allows TF binding and interaction with and activation of the pc gene promoter. (**b**) VDJ recombination. From top to bottom: 1, D and J segments are joined and the V region has closed chromatin; 2, antisense transcription through the V region opens the chromatin and allows recombination factors to bind; 3, a V segment is joined to the DJ segment. (**c**) Top: at a bidirectional promoter a lncRNA and a pc gene are transcribed in opposite directions. The promoter is always in an open chromatin conformation as either the lncRNA or the pc gene is transcribed, which is thought to reduce transcriptional noise. Bottom: a unidirectional pc gene promoter can acquire a closed chromatin conformation due to stochastic TF binding, which is thought to increase transcriptional noise. Noise defines the variation of expression of a transcript between genetically identical cells caused by the stochastic binding of TFs regulated by the local chromatin environment. For details see Figure 1g, Figure 2d, Figure 3c and text.

Features such as subcellular localization, half-life and steady-state abundance would form a good basis to allow functional tests to be designed. In addition, knowledge of the lncRNA splicing efficiency, conservation of splicing pattern in multiple tissues and species, an estimation of transcript repeat content and, finally, an accurate mapping of lncRNA 5' and 3' ends are essential preliminary steps. We have previously proposed that a subclass of lncRNAs, 'macro' lncRNAs, show RNA biology hallmarks such as inefficient splicing, extreme length, high repeat content, lack of conservation and a short half-life. These features are also indicators that the lncRNA product is less important than the act of transcription [121]. Once RNA biology features are known, experiments can be designed to distinguish between a role for the lncRNA product or its transcription.

From the caveats of posttranscriptional knockdown experiments described above, it becomes clear that genetic

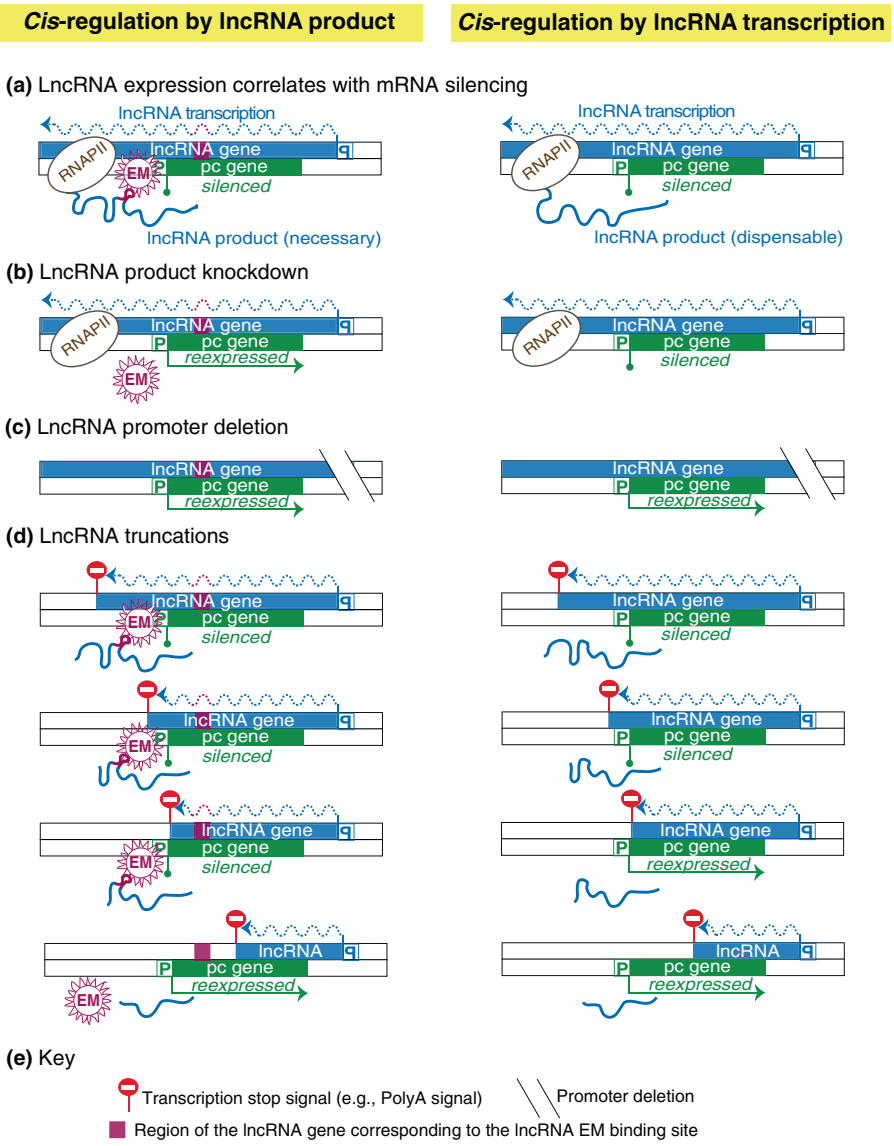


Figure 5. Strategies to distinguish between the function of a lncRNA product and its transcription. Both DNA strands are shown as separate boxes to indicate a lncRNA transcribed from the top reverse strand, overlapping a pc gene transcribed from the bottom forward strand in antisense orientation. A silencing function of the lncRNA can be predicted by an anti-correlating expression pattern. (**a**) Left: the lncRNA silencing effect is mediated by tethering of the lncRNA product at the site of transcription, sequence-specific binding of an EM to the lncRNA and guidance of the EM to the pc gene promoter. Right: silencing is mediated by a transcription process independent of the lncRNA product. (**b**) Posttranscriptional knockdown removes the lncRNA product, thus reversing a lncRNA product-mediated effect (left) but not the transcription-mediated effect (right). (**c**) lncRNA promoter deletion removes both lncRNA product- (left) and transcription-mediated (right) effects. (**d**) Truncation experiments inserting transcriptional stop signals at different positions within the lncRNA gene identify the functional region of the lncRNA gene (RNAPII is not shown). Left: lncRNA is only functional when the region corresponding to EM binding site is present. Right: lncRNA is only functional when it crosses the promoter of the overlapped pc gene. For details see Figure 1g, Figure 2e, Figure 3c, Figure 4d and text.

strategies are optimal for testing lncRNA function. These strategies include manipulating the endogenous locus to delete the promoter or the whole gene or to shorten its length using inserted polyadenylation signals, as described for several examples above. This may appear a formidable task with the appreciation that lncRNAs in the human genome may outnumber protein-coding genes [4]; however, suitable cell systems already exist. These include the use of haploid cell lines with transcriptional stop signal insertions in most human genes that are screened by RNA sequencing [122], gene targeting by engineered zinc-finger nucleases [123] or CRISPR systems [124] or the use of mouse embryonic stem cells that have efficient rates of homologous targeting [125,126].

These genetic strategies could be applied to determine if the lncRNA is functional and if its function requires the lncRNA product or only depends on the act of transcription (Figure 5). Once these answers are obtained, it will be useful to test whether additional chromatin features are involved. This could include chromatin accessibility assays to address nucleosome density in the regulated gene; and mapping of histone modifications and DNA methylation, and of the presence of RNAPII and other transcription machinery components. These studies have been made easier in the mouse and human genome due to the publicly available ENCODE data [127]. As lncRNA identification becomes easier due to improved sequencing and bioinformatics tools, the number of annotated lncRNA transcripts is rising sharply [4,128]. It is therefore a high priority to determine which lncRNAs are functional and which represent spurious transcription [109,129]. To date only a relatively small number of mammalian lncRNAs have clearly been shown to regulate gene expression and most attention has centered on lncRNAs that act through their transcription product [23]. With the recent demonstration that for some mammalian lncRNAs the act of their transcription is sufficient for function [35], it becomes clear that there can be a number of lncRNAs acting in a similar way. If the above described findings and approaches are used as guidelines, many new lncRNAs regulating genes by the act of transcription are likely to be discovered.

Acknowledgements

We thank Quanah Hudson and Federica Santoro for comments on the manuscript. The authors are partly supported by the Austrian Science Fund: FWF SFB-F43 and FWF W1207-BO9. PG is recipient of a DOC Fellowship of the Austrian Academy of Sciences.

Published: 30 May 2013

References

1. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Bar NS, Batut P, Bell K, Bell I, Chakrabortty S, Chen X, Chrast J, Curado J, *et al*: **Landscape of transcription in human cells.** *Nature* 2012, **489:**101–108.
2. Wilusz JE, Sunwoo H, Spector DL: **Long noncoding RNAs: functional surprises from the RNA world.** *Genes Dev* 2009, **23:**1494–1504.
3. Pauli A, Rinn JL, Schier AF: **Non-coding RNAs as regulators of embryogenesis.** *Nat Rev Genet* 2011, **12:**136–149.
4. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, Lagarde J, Veeravalli L, Ruan X, Ruan Y, Lassmann T, Carninci P, Brown JB, Lipovich L, Gonzalez JM, Thomas M, Davis CA, Shiekhattar R, Gingeras TR, Hubbard TJ, Notredame C, Harrow J, Guigó R: **The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression.** *Genome Res* 2012, **22:**1775–1789.
5. Taft RJ, Pang KC, Mercer TR, Dinger M, Mattick JS: **Non-coding RNAs: regulators of disease.** *J Pathol* 2010, **220:**126–139.
6. Huarte M, Rinn JL: **Large non-coding RNAs: missing links in cancer?** *Hum Mol Genet* 2010, **19:**R152–R161.
7. Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, Munson G, Young G, Lucas AB, Ach R, Bruhn L, Yang X, Amit I, Meissner A, Regev A, Rinn JL, Root DE: **Lander ES: lincRNAs act in the circuitry controlling pluripotency and differentiation.** *Nature* 2011, **477:**295–300.
8. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai MC, Hung T, Argani P, Rinn JL, Wang Y, Brzoska P, Kong B, Li R, West RB, van de Vijver MJ, Sukumar S, Chang HY: **Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis.** *Nature* 2010, **464:**1071–1076.
9. Prensner JR, Iyer MK, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC, Laxman B, Asangani IA, Grasso CS, Kominsky HD, Cao X, Jing X, Wang X, Siddiqui J, Wei JT, Robinson D, Iyer HK, Palanisamy N, Maher CA, Chinnaiyan AM: **Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression.** *Nat Biotechnol* 2011, **29:**742–749.
10. Yap KL, Li S, Munoz-Cabello AM, Raguz S, Zeng L, Mujtaba S, Gil J, Walsh MJ, Zhou MM: **Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a.** *Mol Cell* 2010, **38:**662–674.
11. Yoon JH, Abdelmohsen K, Gorospe M: **Posttranscriptional gene regulation by long noncoding RNA.** *J Mol Biol* 2012. **pii:**S0022-2836(12)00896-0.
12. Mattick JS: **Deconstructing the dogma: a new view of the evolution and genetic programming of complex organisms.** *Ann N Y Acad Sci* 2009, **1178:**29–46.
13. Mattick JS, Taft RJ, Faulkner GJ: **A global view of genomic information–moving beyond the gene and the master regulator.** *Trends Genet* 2010, **26:**21–28.
14. Peterlin BM, Brogie JE, Price DH: **7SK snRNA: a noncoding RNA that plays a major role in regulating eukaryotic transcription.** *Wiley Interdiscip Rev RNA* 2012, **3:**92–103.
15. Espinoza CA, Allen TA, Hieb AR, Kugel JF, Goodrich JA: **B2 RNA binds directly to RNA polymerase II to repress transcript synthesis.** *Nat Struct Mol Biol* 2004, **11:**822–829.
16. Espinoza CA, Goodrich JA, Kugel JF: **Characterization of the structure, function, and mechanism of B2 RNA, an ncRNA repressor of RNA polymerase II transcription.** *RNA* 2007, **13:**583–596.
17. Yakovchuk P, Goodrich JA, Kugel JF: **B2 RNA represses TFIIH phosphorylation of RNA polymerase II.** *Transcription* 2011, **2:**45–49.
18. Sleutels F, Zwart R, Barlow DP: **The non-coding Air RNA is required for silencing autosomal imprinted genes.** *Nature* 2002, **415:**810–813.
19. Mancini-Dinardo D, Steele SJ, Levorse JM, Ingram RS, Tilghman SM: **Elongation of the Kcnq1ot1 transcript is required for genomic imprinting of neighboring genes.** *Genes Dev* 2006, **20:**1268–1282.
20. Williamson CM, Ball ST, Dawson C, Mehta S, Beechey CV, Fray M, Teboul L, Dear TN, Kelsey G, Peters J: **Uncoupling antisense-mediated silencing and DNA methylation in the imprinted Gnas cluster.** *PLoS Genet* 2011, **7:**e1001347.
21. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E, Chang HY: **Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs.** *Cell* 2007, **129:**1311–1323.
22. Ng SY, Johnson R, Stanton LW: **Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors.** *EMBO J* 2011, **31:**522–533.
23. Guttman M, Rinn JL: **Modular regulatory principles of large non-coding RNAs.** *Nature* 2012, **482:**339–346.
24. Wutz A: **Gene silencing in X-chromosome inactivation: advances in understanding facultative heterochromatin formation.** *Nat Rev Genet* 2011, **12:**542–553.
25. Ørom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, Lai F, Zytnicki M, Notredame C, Huang Q, Guigo R, Shiekhattar R: **Long noncoding RNAs with enhancer-like function in human cells.** *Cell* 2010, **143:**46–58.
26. Wang KC, Yang YW, Liu B, Sanyal A, Corces-Zimmerman R, Chen Y, Lajoie BR, Protacio A, Flynn RA, Gupta RA, Wysocka J, Lei M, Dekker J, Helms JA, Chang HY: **A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression.** *Nature* 2011, **472:**120–124.
27. Magistri M, Faghihi MA, St Laurent G 3rd, Wahlestedt C: **Regulation of chromatin structure by long noncoding RNAs: focus on natural antisense transcripts.** *Trends Genet* 2012, **28:**389–396.
28. Heo JB, Sung S: **Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA.** *Science* 2011, **331:**76–79.
29. Shearwin KE, Callen BP, Egan JB: **Transcriptional interference–a crash course.** *Trends Genet* 2005, **21:**339–345.

30. Palmer AC, Egan JB, Shearwin KE: **Transcriptional interference by RNA polymerase pausing and dislodgement of transcription factors.** *Transcription* 2011, **2:**9–14.
31. Bird AJ, Gordon M, Eide DJ, Winge DR: **Repression of ADH1 and ADH3 during zinc deficiency by Zap1-induced intergenic RNA transcripts.** *EMBO J* 2006, **25:**5726–5734.
32. Bumgarner SL, Dowell RD, Grisafi P, Gifford DK, Fink GR: **Toggle involving cis-interfering noncoding RNAs controls variegated gene expression in yeast.** *Proc Natl Acad Sci U S A* 2009, **106:**18321–18326.
33. Petruk S, Sedkov Y, Riley KM, Hodgson J, Schweisguth F, Hirose S, Jaynes JB, Brock HW, Mazo A: **Transcription of bxd noncoding RNAs promoted by trithorax represses Ubx in cis by transcriptional interference.** *Cell* 2006, **127:**1209–1221.
34. Gummalla M, Maeda RK, Castro Alvarez JJ, Gyurkovics H, Singari S, Edwards KA, Karch F, Bender W: **abd-A regulation by the iab-8 noncoding RNA.** *PLoS Genet* 2012, **8:**e1002720.
35. Latos PA, Pauler FM, Koerner MV, Şenergin HB, Hudson QJ, Stocsits RR, Allhoff W, Stricker SH, Klement RM, Warczok KE, Aumayr K, Pasierbek P, Barlow DP: **Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing.** *Science* 2012, **338:**1469–1472.
36. MacIsaac JL, Bogutz AB, Morrissy AS, Lefebvre L: **Tissue-specific alternative polyadenylation at the imprinted gene Mest regulates allelic usage at Copg2.** *Nucleic Acids Res* 2012, **40:**1523–1535.
37. Wood AJ, Schulz R, Woodfine K, Koltowska K, Beechey CV, Peters J, Bourc'his D, Oakey RJ: **Regulation of alternative polyadenylation by genomic imprinting.** *Genes Dev* 2008, **22:**1141–1146.
38. Hobson DJ, Wei W, Steinmetz LM, Svejstrup JQ: **RNA polymerase II collision interrupts convergent transcription.** *Mol Cell* 2012, **48:**365–374.
39. Kornberg RD, Lorch Y: **Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome.** *Cell* 1999, **98:**285–294.
40. Li B, Carey M, Workman JL: **The role of chromatin during transcription.** *Cell* 2007, **128:**707–719.
41. Weiner A, Hughes A, Yassour M, Rando OJ, Friedman N: **High-resolution nucleosome mapping reveals transcription-dependent promoter packaging.** *Genome Res* 2010, **20:**90–100.
42. Hughes AL, Jin Y, Rando OJ, Struhl K: **A functional evolutionary approach to identify determinants of nucleosome positioning: a unifying model for establishing the genome-wide pattern.** *Mol Cell* 2012, **48:**5–15.
43. Valouev A, Johnson SM, Boyd SD, Smith CL, Fire AZ, Sidow A: **Determinants of nucleosome organization in primary human cells.** *Nature* 2011, **474:**516–520.
44. Segal E, Widom J: **What controls nucleosome positions?** *Trends Genet* 2009, **25:**335–343.
45. Radman-Livaja M, Rando OJ: **Nucleosome positioning: how is it established, and why does it matter?** *Dev Biol* 2010, **339:**258–266.
46. Martens JA, Laprade L, Winston F: **Intergenic transcription is required to repress the Saccharomyces cerevisiae SER3 gene.** *Nature* 2004, **429:**571–574.
47. Belotserkovskaya R, Oh S, Bondarenko VA, Orphanides G, Studitsky VM, Reinberg D: **FACT facilitates transcription-dependent nucleosome alteration.** *Science* 2003, **301:**1090–1093.
48. Reinberg D, Sims RJ 3rd: **de FACTo nucleosome dynamics.** *J Biol Chem* 2006, **281:**23297–23301.
49. Nourani A, Robert F, Winston F: **Evidence that Spt2/Sin1, an HMG-like factor, plays roles in transcription elongation, chromatin structure, and genome stability in Saccharomyces cerevisiae.** *Mol Cell Biol* 2006, **26:**1496–1509.
50. Hainer SJ, Pruneski JA, Mitchell RD, Monteverde RM, Martens JA: **Intergenic transcription causes repression by directing nucleosome assembly.** *Genes Dev* 2011, **25:**29–40.
51. Thebault P, Boutin G, Bhat W, Rufiange A, Martens J, Nourani A: **Transcription regulation by the noncoding RNA SRG1 requires Spt2-dependent chromatin deposition in the wake of RNA polymerase II.** *Mol Cell Biol* 2011, **31:**1288–1300.
52. Kaplan CD, Laprade L, Winston F: **Transcription elongation factors repress transcription initiation from cryptic sites.** *Science* 2003, **301:**1096–1099.
53. Cheung V, Chua G, Batada NN, Landry CR, Michnick SW, Hughes TR, Winston F: **Chromatin- and transcription-related factors repress transcription from within coding regions throughout the Saccharomyces cerevisiae genome.** *PLoS Biol* 2008, **6:**e277.
54. Gallastegui E, Millan-Zambrano G, Terme JM, Chavez S, Jordan A: **Chromatin reassembly factors are involved in transcriptional interference promoting HIV latency.** *J Virol* 2011, **85:**3187–3202.
55. Bannister AJ, Kouzarides T: **Regulation of chromatin by histone modifications.** *Cell Res* 2011, **21:**381–395.
56. Brookes E, Pombo A: **Modifications of RNA polymerase II are pivotal in regulating gene expression states.** *EMBO Rep* 2009, **10:**1213–1219.
57. Ehrensberger AH, Svejstrup JQ: **Reprogramming chromatin.** *Crit Rev Biochem Mol Biol* 2012, **47:**464–482.
58. Ng HH, Robert F, Young RA, Struhl K: **Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity.** *Mol Cell* 2003, **11:**709–719.
59. Krogan NJ, Kim M, Tong A, Golshani A, Cagney G, Canadien V, Richards DP, Beattie BK, Emili A, Boone C, Shilatifard A, Buratowski S, Greenblatt J: **Methylation of histone H3 by Set2 in Saccharomyces cerevisiae is linked to transcriptional elongation by RNA polymerase II.** *Mol Cell Biol* 2003, **23:**4207–4218.
60. Schneider R, Bannister AJ, Myers FA, Thorne AW, Crane-Robinson C, Kouzarides T: **Histone H3 lysine 4 methylation patterns in higher eukaryotic genes.** *Nat Cell Biol* 2004, **6:**73–77.
61. Carrozza MJ, Li B, Florens L, Suganuma T, Swanson SK, Lee KK, Shia WJ, Anderson S, Yates J, Washburn MP, Workman JL: **Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription.** *Cell* 2005, **123:**581–592.
62. Kim T, Buratowski S: **Dimethylation of H3K4 by Set1 recruits the Set3 histone deacetylase complex to 5' transcribed regions.** *Cell* 2009, **137:**259–272.
63. Keogh MC, Kurdistani SK, Morris SA, Ahn SH, Podolny V, Collins SR, Schuldiner M, Chin K, Punna T, Thompson NJ, Boone C, Emili A, Weissman JS, Hughes TR, Strahl BD, Grunstein M, Greenblatt JF, Buratowski S, Krogan NJ: **Cotranscriptional set2 methylation of histone H3 lysine 36 recruits a repressive Rpd3 complex.** *Cell* 2005, **123:**593–605.
64. van Werven FJ, Neuert G, Hendrick N, Lardenois A, Buratowski S, van Oudenaarden A, Primig M, Amon A: **Transcription of two long noncoding RNAs mediates mating-type control of gametogenesis in budding yeast.** *Cell* 2012, **150:**1170–1181.
65. Houseley J, Rubbi L, Grunstein M, Tollervey D, Vogelauer M: **A ncRNA modulates histone modification and mRNA induction in the yeast GAL gene cluster.** *Mol Cell* 2008, **32:**685–695.
66. Kim T, Xu Z, Clauder-Munster S, Steinmetz LM, Buratowski S: **Set3 HDAC mediates effects of overlapping noncoding transcription on gene induction kinetics.** *Cell* 2012, **150:**1158–1169.
67. Jones PA: **Functions of DNA methylation: islands, start sites, gene bodies and beyond.** *Nat Rev Genet* 2012, **13:**484–492.
68. Deaton AM, Bird A: **CpG islands and the regulation of transcription.** *Genes Dev* 2011, **25:**1010–1022.
69. Ooi SK, O'Donnell AH, Bestor TH: **Mammalian cytosine methylation at a glance.** *J Cell Sci* 2009, **122:**2787–2791.
70. Ooi SK, Qiu C, Bernstein E, Li K, Jia D, Yang Z, Erdjument-Bromage H, Tempst P, Lin SP, Allis CD, Cheng X, Bestor TH: **DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA.** *Nature* 2007, **448:**714–717.
71. Tufarelli C, Stanley JA, Garrick D, Sharpe JA, Ayyub H, Wood WG, Higgs DR: **Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease.** *Nat Genet* 2003, **34:**157–165.
72. Ligtenberg MJ, Kuiper RP, Chan TL, Goossens M, Hebeda KM, Voorendt M, Lee TY, Bodmer D, Hoenselaar E, Hendriks-Cornelissen SJ, Tsui WY, Kong CK, Brunner HG, van Kessel AG, Yuen ST, van Krieken JH, Leung SY, Hoogerbrugge N: **Heritable somatic methylation and inactivation of MSH2 in families with Lynch syndrome due to deletion of the 3' exons of TACSTD1.** *Nat Genet* 2009, **41:**112–117.
73. Santoro F, Barlow DP: **Developmental control of imprinted expression by macro non-coding RNAs.** *Semin Cell Dev Biol* 2011, **22:**328–335.
74. Santoro F, Mayer D, Klement RM, Warczok KE, Stukalov A, Barlow DP, Pauler FM: **Imprinted Igf2r silencing depends on continuous Airn lncRNA expression and is not restricted to a developmental window.** *Development* 2013, **140:**1184–1195.
75. Pauler FM, Barlow DP, Hudson QJ: **Mechanisms of long range silencing by imprinted macro non-coding RNAs.** *Curr Opin Genet Dev* 2012, **22:**283–289.
76. Pauler FM, Koerner MV, Barlow DP: **Silencing by imprinted noncoding RNAs: is transcription the answer?** *Trends Genet* 2007, **23:**284–292.

77. Barlow DP: **Genomic imprinting: a mammalian epigenetic discovery model.** *Annu Rev Genet* 2011, **45:**379–403.
78. Koerner MV, Pauler FM, Huang R, Barlow DP: **The function of non-coding RNAs in genomic imprinting.** *Development* 2009, **136:**1771–1783.
79. Wang ZQ, Fung MR, Barlow DP, Wagner EF: **Regulation of embryonic growth and lysosomal targeting by the imprinted Igf2/Mpr gene.** *Nature* 1994, **372:**464–467.
80. Pauler FM, Stricker SH, Warczok KE, Barlow DP: **Long-range DNase I hypersensitivity mapping reveals the imprinted Igf2r and Air promoters share cis-regulatory elements.** *Genome Res* 2005, **15:**1379–1387.
81. Stoger R, Kubicka P, Liu CG, Kafri T, Razin A, Cedar H, Barlow DP: **Maternal-specific methylation of the imprinted mouse Igf2r locus identifies the expressed locus as carrying the imprinting signal.** *Cell* 1993, **73:**61–71.
82. Racanelli AC, Turner FB, Xie LY, Taylor SM, Moran RG: **A mouse gene that coordinates epigenetic controls and transcriptional interference to achieve tissue-specific expression.** *Mol Cell Biol* 2008, **28:**836–848.
83. Palmer AC, Ahlgren-Berg A, Egan JB, Dodd IB, Shearwin KE: **Potent transcriptional interference by pausing of RNA polymerases over a downstream promoter.** *Mol Cell* 2009, **34:**545–555.
84. Adelman K, Lis JT: **Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans.** *Nat Rev Genet* 2012, **13:**720–731.
85. Levine M: **Paused RNA polymerase II as a developmental checkpoint.** *Cell* 2011, **145:**502–511.
86. Nagano T, Mitchell JA, Sanz LA, Pauler FM, Ferguson-Smith AC, Feil R, Fraser P: **The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin.** *Science* 2008, **322:**1717–1720.
87. Mager J, Montgomery ND, de Villena FP, Magnuson T: **Genome imprinting regulated by the mouse Polycomb group protein Eed.** *Nat Genet* 2003, **33:**502–507.
88. Terranova R, Yokobayashi S, Stadler MB, Otte AP, van Lohuizen M, Orkin SH, Peters AH: **Polycomb group proteins Ezh2 and Rnf2 direct genomic contraction and imprinted repression in early mouse embryos.** *Dev Cell* 2008, **15:**668–679.
89. Korostowski L, Sedlak N, Engel N: **The Kcnq1ot1 long non-coding RNA affects chromatin conformation and expression of Kcnq1, but does not regulate its imprinting in the developing heart.** *PLoS Genet* 2012, **8:**e1002956.
90. Visel A, Rubin EM, Pennacchio LA: **Genomic views of distant-acting enhancers.** *Nature* 2009, **461:**199–205.
91. Ong CT, Corces VG: **Enhancers: emerging roles in cell fate specification.** *EMBO Rep* 2012, **13:**423–430.
92. Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, Markenscoff-Papadimitriou E, Kuhl D, Bito H, Worley PF, Kreiman G, Greenberg ME: **Widespread transcription at neuronal activity-regulated enhancers.** *Nature* 2010, **465:**182–187.
93. Schatz DG, Swanson PC: **V(D)J recombination: mechanisms of initiation.** *Annu Rev Genet* 2011, **45:**167–202.
94. Bolland DJ, Wood AL, Afshar R, Featherstone K, Oltz EM, Corcoran AE: **Antisense intergenic transcription precedes Igh D-to-J recombination and is controlled by the intronic enhancer Emu.** *Mol Cell Biol* 2007, **27:**5523–5533.
95. Giallourakis CC, Franklin A, Guo C, Cheng HL, Yoon HS, Gallagher M, Perlot T, Andzelm M, Murphy AJ, Macdonald LE, Yancopoulos GD, Alt FW: **Elements between the IgH variable (V) and diversity (D) clusters influence antisense transcription and lineage-specific V(D)J recombination.** *Proc Natl Acad Sci U S A* 2010, **107:**22207–22212.
96. Abarrategui I, Krangel MS: **Noncoding transcription controls downstream promoters to regulate T-cell receptor alpha recombination.** *EMBO J* 2007, **26:**4380–4390.
97. Ashe HL, Monks J, Wijgerde M, Fraser P, Proudfoot NJ: **Intergenic transcription and transinduction of the human beta-globin locus.** *Genes Dev* 1997, **11:**2494–2509.
98. Gribnau J, Diderich K, Pruzina S, Calzolari R, Fraser P: **Intergenic transcription and developmental remodeling of chromatin subdomains in the human beta-globin locus.** *Mol Cell* 2000, **5:**377–386.
99. Cumberledge S, Zaratzian A, Sakonju S: **Characterization of two RNAs transcribed from the cis-regulatory region of the abd-A domain within the Drosophila bithorax complex.** *Proc Natl Acad Sci U S A* 1990, **87:**3259–3263.
100. Beisel C, Paro R: **Silencing chromatin: comparing modes and mechanisms.** *Nat Rev Genet* 2011, **12:**123–135.
101. Schmitt S, Prestel M, Paro R: **Intergenic transcription through a polycomb group response element counteracts silencing.** *Genes Dev* 2005, **19:**697–708.
102. Sessa L, Breiling A, Lavorgna G, Silvestri L, Casari G, Orlando V: **Noncoding RNA synthesis and loss of Polycomb group repression accompanies the colinear activation of the human HOXA cluster.** *RNA* 2007, **13:**223–239.
103. Kanhere A, Viiri K, Araújo CC, Rasaiyaah J, Bouwman RD, Whyte WA, Pereira CF, Brookes E, Walker K, Bell GW, Pombo A, Fisher AG, Young RA, Jenner RG: **Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2.** *Mol Cell* 2010, **38:**675–688.
104. Hekimoglu-Balkan B, Aszodi A, Heinen R, Jaritz M, Ringrose L: **Intergenic Polycomb target sites are dynamically marked by non-coding transcription during lineage commitment.** *RNA Biol*, **9:**314–325.
105. Uhler JP, Hertel C, Svejstrup JQ: **A role for noncoding transcription in activation of the yeast PHO5 gene.** *Proc Natl Acad Sci U S A* 2007, **104:**8011–8016.
106. Neil H, Malabat C, d'Aubenton-Carafa Y, Xu Z, Steinmetz LM, Jacquier A: **Widespread bidirectional promoters are the major source of cryptic transcripts in yeast.** *Nature* 2009, **457:**1038–1042.
107. Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, Flynn RA, Young RA, Sharp PA: **Divergent transcription from active promoters.** *Science* 2008, **322:**1849–1851.
108. Brosius J: **Waste not, want not–transcript excess in multicellular eukaryotes.** *Trends Genet* 2005, **21:**287–288.
109. Kowalczyk MS, Higgs DR, Gingeras TR: **Molecular biology: RNA discrimination.** *Nature* 2012, **482:**310–311.
110. Wang GZ, Lercher MJ, Hurst LD: **Transcriptional coupling of neighboring genes and gene expression noise: evidence that gene orientation and noncoding transcripts are modulators of noise.** *Genome Biol Evol* 2011, **3:**320–331.
111. Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP: **Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution.** *Cell* 2011, **147:**1537–1550.
112. Zeng Y, Cullen BR: **RNA interference in human cells is restricted to the cytoplasm.** *RNA* 2002, **8:**855–860.
113. Ohrt T, Muetze J, Svoboda P, Schwille P: **Intracellular localization and routing of miRNA and RNAi pathway components.** *Curr Top Med Chem* 2012, **12:**79–88.
114. Ideue T, Hino K, Kitao S, Yokoi T, Hirose T: **Efficient oligonucleotide-mediated degradation of nuclear noncoding RNAs in mammalian cultured cells.** *RNA* 2009, **15:**1578–1587.
115. Tse MT: **Antisense therapeutics: Nuclear RNA more susceptible to knockdown.** *Nat Rev Drug Discov* 2012, **11:**674.
116. Zhang B, Arun G, Mao YS, Lazar Z, Hung G, Bhattacharjee G, Xiao X, Booth CJ, Wu J, Zhang C, Spector DL: **The lncRNA Malat1 is dispensable for mouse development but its transcription plays a cis-regulatory role in the adult.** *Cell Rep* 2012, **2:**111–123.
117. Nakagawa S, Naganuma T, Shioi G, Hirose T: **Paraspeckles are subpopulation-specific nuclear bodies that are not essential in mice.** *J Cell Biol*, **193:**31–39.
118. Eißmann M, Gutschner T, Hämmerle M, Günther S, Caudron-Herger M, Groß M, Schirmacher P, Rippe K, Braun T, Zörnig M, Diederichs S: **Loss of the abundant nuclear non-coding RNA MALAT1 is compatible with life and development.** *RNA Biol* 2012, **9:**1076–1087.
119. Nakagawa S, Ip JY, Shioi G, Tripathi V, Zong X, Hirose T, Prasanth KV: **Malat1 is not an essential component of nuclear speckles in mice.** *RNA* 2012, **18:**1487–1499.
120. Schorderet P, Duboule D: **Structural and functional differences in the long non-coding RNA hotair in mouse and human.** *PLoS Genet* 2011, **7:**e1002071.
121. Guenzl PM, Barlow DP: **Macro lncRNAs: A new layer of cis-regulatory information in the mammalian genome.** *RNA Biol* 2012, **9:**731–741.
122. Carette JE, Guimaraes CP, Wuethrich I, Blomen VA, Varadarajan M, Sun C, Bell G, Yuan B, Muellner MK, Nijman SM, Ploegh HL, Brummelkamp TR: **Global gene disruption in human cells to assign genes to phenotypes by deep sequencing.** *Nat Biotechnol* 2011, **29:**542–546.
123. Wirt SE, Porteus MH: **Development of nuclease-mediated site-specific genome modification.** *Curr Opin Immunol* 2012, **24:**609–616.
124. Mali P, Yang L, Esvelt KM, Aach J, Guell M, Dicarlo JE, Norville JE, Church GM: **RNA-Guided Human Genome Engineering via Cas9.** *Science* 2013, **339:**823–826.
125. Latos PA, Stricker SH, Steenpass L, Pauler FM, Huang R, Senergin BH, Regha K, Koerner MV, Warczok KE, Unger C, Barlow DP: **An in vitro ES cell imprinting model shows that imprinted expression of the Igf2r gene arises from an allele-specific expression bias.** *Development* 2009, **136:**437–448.

126. Kohama C, Kato H, Numata K, Hirose M, Takemasa T, Ogura A, Kiyosawa H: **ES cell differentiation system recapitulates the establishment of imprinted gene expression in a cell-type-specific manner.** *Hum Mol Genet* 2012, **21:**1391–1401.
127. Rosenbloom KR, Sloan CA, Malladi VS, Dreszer TR, Learned K, Kirkup VM, Wong MC, Maddren M, Fang R, Heitner SG, Lee BT, Barber GP, Harte RA, Diekhans M, Long JC, Wilder SP, Zweig AS, Karolchik D, Kuhn RM, Haussler D, Kent WJ: **ENCODE Data in the UCSC Genome Browser: year 5 update.** *Nucleic Acids Res* 2013, **41:**D56–63.
128. Loven J, Orlando DA, Sigova AA, Lin CY, Rahl PB, Burge CB, Levens DL, Lee TI, Young RA: **Revisiting global gene expression analysis.** *Cell* 2012, **151:**476–482.
129. Clark MB, Mattick JS: **Long noncoding RNAs in cell biology.** *Semin Cell Dev Biol* 2011, **22:**366–376.

doi:10.1186/1741-7007-11-59
Cite this article as: Kornienko AE *et al*: Gene regulation by the act of long non-coding RNA transcription. *BMC Biology* 2013 **11**:59.

1.4 LncRNAs in disease

In parallel to the discovery of more and more functions performed by lncRNAs at various levels in the cell, lncRNAs have also been increasingly implicated in disease (reviewed in (Batista & Chang, 2013; Wapinski & Chang, 2011)). Underlining the relevance of lncRNAs for human health, extensive studies revealed that lncRNAs play important roles in normal function, development and disease of various physiological systems, such as brain/nervous system (Briggs et al, 2015; Ng et al, 2013; Qureshi et al, 2010), immunity (Atianand & Fitzgerald, 2014; Heward & Lindsay, 2014) and heart (Scheuermann & Boyer, 2013).

That lncRNAs actively play roles in disease, seems to be one of the major keys (St Laurent et al, 2014) to the riddle arising from GWAS studies which show that 85% of disease-associated genetic variants reside in non-(protein)-coding regions of the human genome (Hindorff et al, 2009). For example, numerous GWAS reported SNPs lie inside a lncRNA called *ANRIL* (Pasmant et al, 2011), which has now been shown to be a functional regulator and an important player in numerous human diseases (Congrains et al, 2013). The list of human diseases associated with lncRNAs is constantly growing (Chen et al, 2013). An extensive literature describes roles of certain lncRNAs in various diseases, such as cancer (*HOTAIR* (Gupta et al, 2010; Kogo et al, 2011; Niinuma et al, 2012), *lincRNA-p21* (Huarte et al, 2010)), *MALAT1* (Schmidt et al, 2011), *ANRIL* (Pasmant et al, 2007), *BOK-AS* (Zhang et al, 2009), etc.), Schizophrenia (*DISC2* (Millar et al, 2000)), Alzheimer's disease (*BACE1-AS* (Faghihi et al, 2008), *NAT-RAD18* (Parenti et al, 2007)), Parkinson disease (*PINK1-AS* (Scheele et al, 2007)), Angelman syndrome (UBE3A-AS (Meng et al, 2013)) and HELLP syndrome (*HELLP* lincRNA (van Dijk et al, 2012)). LncRNAs playing roles in diseases are reviewed in (Wahlestedt, 2013), (Cheetham et al, 2013; Fitzgerald & Caffrey, 2014; Martens-Uzunova et al, 2013; Roth & Diederichs, 2015; Wapinski & Chang, 2011). Various kinds of lncRNA perturbations can cause diseases, since different studies show associations of the disease state with either mature lncRNA product expression level (Gupta et al, 2010), copy number variation inside a lncRNA gene (Cabianca et al, 2012), chromosomal translocations affecting lncRNAs (Maass et al, 2012) or SNPs/mutations in lncRNA genes (Cartault et al, 2012; Pan et al, 2015).

Besides investigating roles of particular lncRNAs in disease, intensive studies have been performed that examined correlations between perturbations in the overall lncRNA expression landscape in a certain tissue with the disease/healthy status of this tissue (Iyer et al, 2015; Mirza et al, 2015; Tsoi et al, 2015; White et al, 2014). Multiple diseases were shown to be well-distinguishable based on analyzing lncRNA expression either by microarray technology (Li et al, 2013a; Luo et al, 2014; Yang et al, 2015) or RNA-seq (Iyer et al, 2015; Ounzain et al, 2015; Prensner et al, 2011). Interestingly, apart from the deregulated lncRNA expression, such studies also identify novel lncRNAs in the diseased tissues, such as in psoriatic skin (Tsoi et al, 2015) or post-myocardial-infarction heart (Ounzain et al, 2015). Expression of these novel lncRNAs notably contributes to the overall change in the transcriptional landscape of the disease tissue. Discovery of novel lncRNAs expressed upon diseased condition, first, contributes to the notion of lncRNA tissue-specificity, and second – might give a meaning to the previously unexplained disease GWAS hits. A massive study by Iyer et al., analyzed thousands of cancer samples and found thousands of novel previously undescribed lncRNAs (Iyer et al, 2015), some of which might be induced by a disease state of a certain tissue. Malignant tissues might possess genome mutations or rearrangements that can cause emergence of lncRNA promoters absent in normal situation, and these lncRNAs may in turn promote or cause the disease (Cheetham et al, 2013).

With the appreciation of lncRNAs playing roles and being dysregulated in disease, multiple studies and reviews proposed their use as biomarkers and prognostic factors (Iyer et al, 2015; Martens-Uzunova et al, 2013; Roth & Diederichs, 2015; Schmidt et al, 2011). For example, it was shown that overexpression of the *ANRIL* lncRNA can predict poor gastric cancer prognosis, indicating a role in the acceleration of tumor growth (Zhang et al, 2014a). Moreover, since many lncRNAs were shown to actively promote or even cause disease, therapeutic targeting of lncRNAs was proposed and is seen as a very promising future medical direction (Li & Chen, 2013; Qureshi & Mehler, 2013; Wahlestedt, 2013). A successful recovery of a healthy phenotype by targeting a lncRNA *Ube3a-AS* in Angelman syndrome model mice was reported (Meng et al, 2015).

LncRNAs appear to be mainly gene-specific regulators which makes them an appealing therapeutic target, while targeting, for example, histone modifying proteins, an approach that is currently massively investigated and proposed to the clinics (West & Johnstone,

2014), causes multiple side effects since the activity of these proteins is very general and, incidentally, chromatin modifiers only gain gene specificity when guided by lncRNAs (Rinn & Chang, 2012). The current main way of targeting lncRNAs *in vivo* is by introducing antisense oligonucleotides, which would target the lncRNA transcripts via RNA interference (Guttman et al, 2011) or RNase H mediated decay (Tripathi et al, 2010). However, if a lncRNA functions through the act of its transcription, as described above (reviewed in (Kornienko et al, 2013)), targeting its product would not rescue the healthy phenotype and, potentially, gene therapy has to be applied, which makes the therapy approach notably more complicated.

Overall, with their considerable involvement in the various diseases, including many cancer types, and the high potential of the clinical use as biomarkers, prognostic factors and even therapy targets, lncRNAs gain outstanding attention as a gene class that unquestionably requires further investigation.

1.5 Understanding lncRNA biology

1.5.1 LncRNA evolution

To date the precise history of how lncRNAs emerged and developed in the course of evolution is unclear. For example, the well-studied lncRNA *Xist* was shown to have evolved from a pseudogene (Duret et al, 2006). It was also suggested that transposable elements (TE) make a major contribution in the emergence of lncRNA genes in vertebrates (Kapusta et al, 2013). Lineage-specific genome mobility of TEs was shown to shape lncRNA landscapes of human, mouse and zebrafish (Kapusta et al, 2013). Importantly, lncRNAs as a class of genes seem to be present in all organisms, from yeast to human (Figure 5) (Kapusta & Feschotte, 2014; Ulitsky & Bartel, 2013), and studies in multiple organisms identified at least several lncRNAs as functional regulators (Amaral et al, 2011; Jin et al, 2013). This indicates that in every organism lncRNAs, at least some of them, represent meaningful functional genes rather than just purely spurious transcription activity. However, the evolutionary "purpose" of massive transcription of lncRNAs as well as their origin is being debated (Ulitsky & Bartel, 2013). One theory suggests that lncRNAs, being very evolutionary dynamic (Johnsson et al, 2014), may serve as a pool for evolution to safely "experiment" by mutating these pre-genes

"searching" for a beneficial functional element without a risk of knocking out a vital gene (Clark et al, 2013).

Figure 5

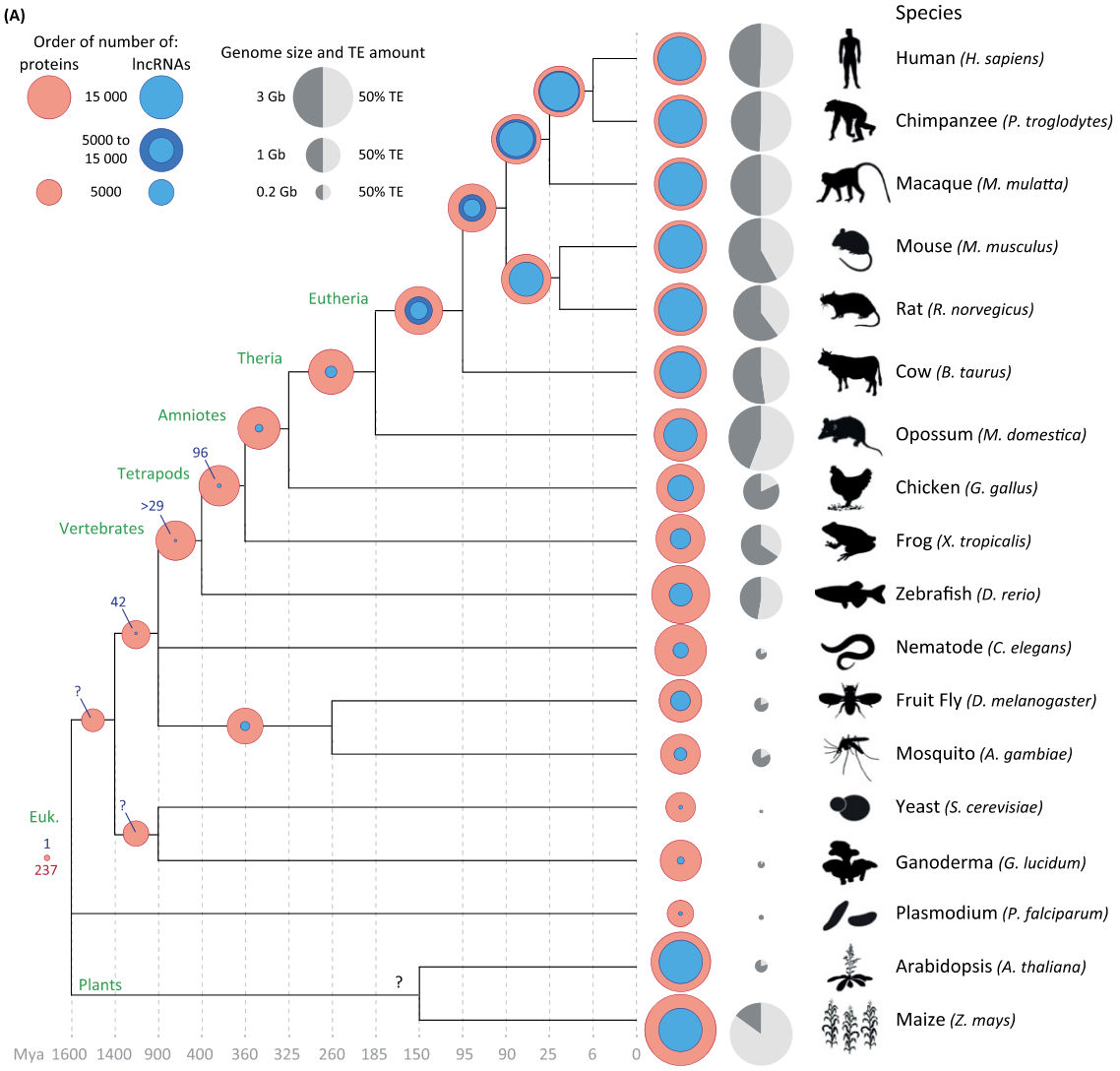


Figure 5. LncRNAs are present in the genomes of all organisms. (Figure taken unchanged from (Kapusta & Feschotte, 2014) (Figure 1(A)) after permission for reprint was obtained). Figure shows the numbers of lncRNAs (blue circles) and protein-coding genes (red circles) together to make comparison easier. Percentage of transposable elements (TE) is shown for all organisms as gray pie-charts: light gray indicates TE, while dark grey indicates the proportion of non-TE regions in the genome (N.B., Plasmodium has 0% TE in its genome). The size of the grey circle reflects the size of the genome of the corresponding organism. The number of genes conserved between the evolutionary branches is indicated at each branch node. The number of lncRNAs conserved among tetrapods is taken from (Necsulea et al, 2014) and among pan-vertebrates - from (Ulitsky et al, 2011). In placental mammals the number of shared lncRNAs is inferred from (Washietl et al, 2014) and (Kutter et al, 2012) with variations between the studies shown as dark-blue circles. The number of lncRNAs shared between mosquitos and Drosophila are inferred from (Inagaki et al, 2005), while 42 lncRNAs between vertebrates and Drosophila are taken from (Young et al, 2012). Only one lncRNA is known to be conserved across most eukaryotes – telomeric TERRA lncRNA (Luke & Lingner, 2009; Schoeftner & Blasco, 2008; Vrbsky et al, 2010). lncRNA numbers in different organisms displayed in this figure: human (GENCODE v19,

GRCh37, Dec 2013), chimpanzee, macaque (Necsulea et al, 2014), mouse (Gencode v2, GRCm38, Dec 2013), rat and cow lncRNA numbers were extrapolated by (Kapusta & Feschotte, 2014) from lncRNA identifications in single tissues (Billerey et al, 2014; Kutter et al, 2012; Weikard et al, 2013; Xie et al, 2014) and data for other mammals (Necsulea et al, 2014), opossum (Necsulea et al, 2014), chicken (Necsulea et al, 2014; Xie et al, 2014), frog (Necsulea et al, 2014), zebrafish (Pauli et al, 2012; Qu & Adelson, 2012; Ulitsky et al, 2011; Xie et al, 2014), nematode (Nam & Bartel, 2012; Xie et al, 2014), Drosophila (Brown et al, 2014; Young et al, 2012), mosquito (http://biorxiv.org/content/early/2014/07/26/007484.full-text.pdf+html), yeast (Xie et al, 2014), Ganoderma lucidum (Li et al, 2014a), Plasmodium (Broadbent et al, 2011), Arabidopsis (Liu et al, 2012), maize (Boerner & McGinnis, 2012; Li et al, 2014b).

While, as mentioned above, lncRNAs as a gene class are likely to be present in the genomes of all organisms, their gene body sequence conservation is generally very low in comparison to protein-coding genes (Marques & Ponting, 2009; Marques & Ponting, 2014). While some well-studied lncRNAs are well conserved (Chodroff et al, 2010), most lncRNAs evolve very rapidly. It was estimated that only <6% of lincRNAs in the zebrafish genome show visible conservation of their sequence with human or mouse (Ulitsky et al, 2011), and generally only about 12% of human and mouse lncRNAs show conservation with other species (Cabili et al, 2011; Church et al, 2009). Even the conservation between human and primates is significantly lower for lncRNAs compared to protein-coding genes (Necsulea & Kaessmann, 2014). While protein-coding genes show little variation across primates (Khaitovich et al, 2005) with 98% of protein-coding genes being conserved across all primates, only 92% of human lncRNAs have a detectable homologue in chimpanzee and just 72% - in the more evolutionary distant macaque (Necsulea et al, 2014). Interestingly, in addition to high protein-coding gene conservation, it was shown that if mRNA levels differ between species, it does not always result in a proportional protein level change and mechanisms for buffering transcript level changes were reported, indicating that protein abundance evolves under a stronger constraint than mRNA abundance for a certain gene (Khan et al, 2013), which brings even stronger contrast to the seemingly unconstrained evolution of lncRNAs.

Low lncRNA conservation leads some researches to argue against the possibility of lncRNA genes being functional (Palazzo & Lee, 2015; Ulitsky & Bartel, 2013). However, there are several strong arguments advocating functions for lncRNAs (Briggs et al, 2015). First, lncRNA functionality, given that they may function as described above (see *1.3. Functions and mechanisms of lncRNAs*), is not as sensitive to sequence change as is the functionality of protein-coding genes (mRNAs), where a point mutation can disrupt an

ORF or a non-synonymous substitution can severely affect the structure and function of the encoded protein. If a lncRNA does contain a functional domain, which could be inactivated upon sequence change, this domain is usually notably smaller than the whole lncRNA gene (Kutter et al, 2012; Ulitsky et al, 2011), thus making the gene conservation appear to be very low. It has indeed been shown that only a small piece of the full length lncRNA product can be essential for its function (Ulitsky et al, 2011). Moreover, lncRNAs often function via forming particular secondary structures (Rinn & Chang, 2012), which could be more resistant to mutations than mRNA translation into a functional protein. Supporting this argumentation, it was shown that some well-studied lncRNAs, such as *Megamind* (Ulitsky et al, 2011), *Xist* and *HOTAIR* (Johnsson et al, 2014), preserve their function in vertebrates by conserving just a small region inside the gene essential for their function (Figure 6, see Figure legend (Kapusta & Feschotte, 2014)).

Figure 6

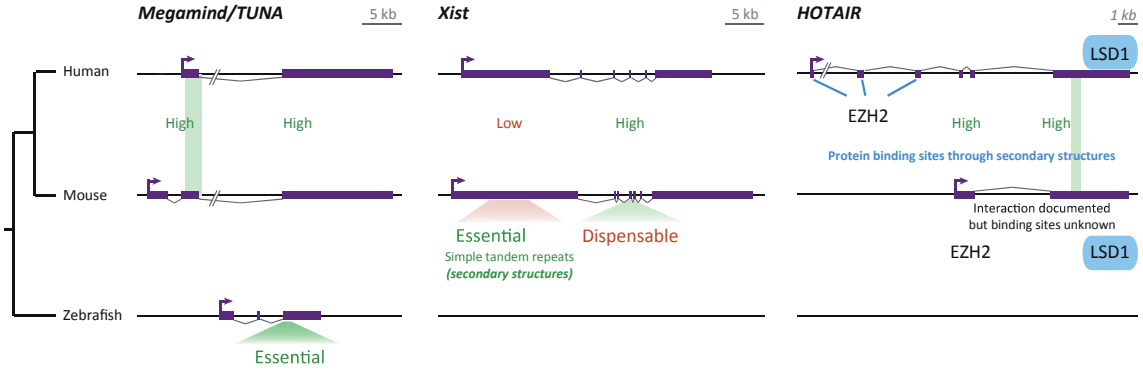


Figure 6. Examples of lncRNAs that display conserved function in different organisms, but show little sequence conservation (figure taken from (Kapusta & Feschotte, 2014), reprint with permission). *Megamind*, *Xist* and *HOTAIR* lncRNAs serve as illustrations to the three following points respectively: 1) it is possible to conserve biological function with low sequence conservation, 2) it is possible that the main active parts of a lncRNAs do not show highest sequence conservation and 3) secondary structures, in contrast to overall sequence, are essential for the function of some lncRNAs. **1) *Megamind***. *Megamind* lncRNA is expressed in human, mouse and zebrafish brains and is essential for their development (Lin et al, 2014; Ulitsky et al, 2011). Functional conservation of *Megamind* is very high since an injection of a human or mouse *Megamind* transcript into a *Megamind* knockdown zebrafish allows to rescue the wild type phenotype (Ulitsky et al, 2011). The exon structure of *Megamind* shows low conservation across vertebrates and the only conserved piece of the gene body is 200nt long, but was shown to be essential for *Megamind* function (Lin et al, 2014; Ulitsky et al, 2011). **2) *Xist.*** Xist is a lncRNA responsible for X-inactivation. First exon of Xist lncRNA contains the majority of its functional elements, including tandem repeats that were shown to be crucial for Xist function in human and mice (Duszczyk et al, 2011; Maenner et al, 2010), but shows low conservation across mammals. Exon 4 shows high conservation, but is dispensible (Caparros et al, 2002). **3) *HOTAIR***. *HOTAIR* recruits EZH2, a subunit of PRC2, and LSD1, an H3K4me3 demethylase, to silence genes in HOXD cluster (Gupta et al, 2010). Mouse and human show little *HOTAIR* sequence or exon

structure similarity (Schorderet & Duboule, 2011). Notably, small pieces of the sequence necessary and sufficient for EZH2 binding in human are poorly conserved (Wu et al, 2013).

Second advocacy of lncRNA functionality is that, in contrast to gene body sequence conservation, lncRNA promoters were shown to be nearly as well-conserved as protein-coding gene promoters (Derrien et al, 2012). Moreover, lncRNA splice sites were shown to be better conserved than the whole gene body (Chodroff et al, 2010) with some lncRNA splice junction motifs being highly conserved (Nitsche et al, 2015), which indicates importance of splicing for function of a fraction of lncRNAs. Third, analysis of single nucleotide polymorphisms (SNPs) identified in primate-specific lncRNAs (those with evolutionary age lower than 25 million years) (Necsulea et al, 2014) and reports of significant constraint of lncRNAs humans (Ward & Kellis, 2012) indicate recent purifying selection occurring for lncRNAs. A purifying selection might argue towards functionality rather than spuriousness of lncRNA expression.

Apart from sequence conservation, expression conservation of lncRNAs and mRNAs was investigated (Necsulea et al, 2014). It was revealed that both lncRNA sequence and expression levels evolve faster than that of protein-coding genes, and hundreds of lncRNAs are highly conserved both inter-species and inter-tissue (Necsulea & Kaessmann, 2014). However, even when a lncRNA is conserved sufficiently to find its homology loci in different species, its expression is notably less conserved than expression of a protein-coding gene (Necsulea & Kaessmann, 2014). This study underlined technical challenges associated with the identification of lncRNA homologs in different species, caused by the low sequence conservation and low expression level, such that the ability to identify a very lowly abundant lncRNA consistently expressed in another species may be masked by a technical detection bias. Necsulea et al., being aware of the biases, still concluded that lncRNA expression is genuinely notably less conserved than that of mRNAs (Necsulea et al, 2014).

Overall, it seems clear that lncRNAs do not follow the same rules of conservation as protein-coding genes. This highlights fundamental differences between these two classes of genes. As discussed above, lncRNAs might function via a small piece of their sequence with all the rest of its gene body being indifferent to mutations. Similarly, if a lncRNA functions through the mere act of its transcription, conservation inside the gene body

would be unnecessary. Moreover, if a lncRNA acts through forming a certain structure, that can be achieved by various RNA sequences, the conservation becomes unnecessary too. Importantly, in contrast to protein-coding genes, neither the degree of lncRNA sequence conservation, nor the lncRNA gene sequence itself can give any hint about lncRNA function or even about a mere presence/absence of function.

1.5.2 LncRNA features compared to mRNA features

The discovery of the first mammalian lncRNA *H19* in 1989 was in fact initially reported as a characterization of an unusual protein-coding gene involved in embryogenesis and *H19* was called an "mRNA" (Pachnis et al, 1988). Two years later it was suggested that while *H19* is transcribed by RNAPII, spliced and polyadenylated like other mRNAs, it might function as an RNA and encode no protein (Brannan et al, 1990). Later, with the discovery of a few more lncRNAs in mouse, such as spliced and polyadenylated *Xist* lncRNA, unspliced *Kcnq1ot1* (Redrup et al, 2009; Smilinich et al, 1999) and inefficiently spliced Airn (Lyle et al, 2000; Seidl et al, 2006; Wutz et al, 1997), it became clear that lncRNAs can be dramatically diverse in their features and functions. However, the few lncRNAs known at that time did not allow solid conclusions about what lncRNAs are like, as a class. The main commonality between lncRNAs and mRNAs is that they are produced by RNAPII, and were reported to possess similar active/inactive histone marks and be spliced, capped and polyadenylated (Rinn & Chang, 2012), and the first genome-wide attempts to map and characterize lncRNAs were based on the presumption of their similarity to mRNAs (Khalil et al, 2009). However, it is clear that apart from the *a priori* difference in protein-coding function and thus absence of an open reading frame (ORF) in lncRNA gene and absence of lncRNA translation (Guttman et al, 2013), the fundamental difference in the function of mRNAs and lncRNAs might result in different biology of these two classes of genes. Several studies analyzed various aspects of lncRNA biology and genomic features using mRNA as a reference for comparison (Mercer & Mattick, 2013; Ulitsky & Bartel, 2013). To date, it is widely appreciated that lncRNAs differ notably from mRNAs in many ways (Quinn & Chang, 2015).

First, the initial genome-wide lncRNA identification studies that used multiple human and mouse tissues reported that lncRNAs display low expression levels and outstanding tissue specificity compared to protein-coding genes (Cabili et al, 2011; Guttman et al,

2009), with testes showing the highest number of lncRNAs expressed specifically in this tissue (Soumillon et al, 2013). Importantly, lncRNAs show distinct expression profiles not only among different tissues, but also among different cell types, even those in close proximity within the differentiation tree (Amin et al, 2015; Ranzani et al, 2015). For example, identification and expression characterization of lncRNAs in 13 subtypes of T and B lymphocytes revealed that these cell types can be clearly distinguished based on their lncRNA expression profiles with 73% lncRNA identified in the study showing cell type specific expression in one of the 13 cell types (Ranzani et al, 2015). Thus, it may be suggested that every cell type in the human, and likely other mammals, inherently possesses a certain unique lncRNA landscape, which has to be defined prior to studying expression and other features of lncRNAs in a particular cell type. While it can be hypothesized that tight tissue-specific lncRNA expression has a functional goal, it is to date unclear by what means the outstanding tissue-specificity is achieved in contrast to relatively lower mRNA tissue-specificity (Tsoi et al, 2015).

Relatively low expression level of lncRNAs has been noted before, for example, *Airn* lncRNA is ~300-fold less abundant than its overlapped *Igf2r* protein-coding gene (in mouse embryonic fibroblasts RPKM of *Igf2r* is 337 and RPKM of *Airn* is 1, see (Andergassen et al, 2015)). Other examples of pioneer lncRNAs showed high, e.g. *Xist* lncRNA (Mercer et al, 2012), or even outstandingly high, e.g. *MALAT1* (Gutschner et al, 2013) and *NEAT1* (Nakagawa et al, 2011) lncRNAs, expression levels. However, with a genome-wide discovery and characterization it became clear that the majority of lncRNAs show very low expression level in different organisms including human (Cabili et al, 2011; Guttman et al, 2009; Nam & Bartel, 2012), with some lncRNAs being present in the cell just as a single molecule (Mercer et al, 2012). It was hypothesized that lncRNAs might be not robustly expressed in different single cells in the population thus resulting in a seemingly lower expression level when analyzing RNA collected from multiple cells (Dinger et al, 2009; Shalek et al, 2013). However, this hypothesis was disproven as it was shown that lncRNA cell-to-cell variability is similar to that of mRNAs (Cabili et al, 2015). Another potential cause of reduced lncRNA abundance might be their reduced stability. It was shown that lncRNAs are less stable than mRNAs (Clark et al, 2012) and thus, an unstable lncRNA whose expression strength is similar to an mRNA's will appear less expressed. This case might be exemplified by *Airn* lncRNA whose promoter is strong, however, the predominant unspliced isoforms are unstable with a half-life of 90 minutes

resulting in a reduction in the overall steady-state *Airn* expression level (Seidl et al, 2006). It is to date unclear what is the actual cause of predominantly low expression level in the lncRNA population.

Another feature that distinguishes lncRNAs from mRNAs is their cellular localization. While all mRNAs must be exported to the cytoplasm in order to perform their function, lncRNAs are notably more nuclear (Cabili et al, 2015; Derrien et al, 2012). Some lncRNAs were shown to be exported to the cytoplasm (Cabili et al, 2015), however the vast majority are retained in the nucleus (Derrien et al, 2012). Cytoplasmic localization of some lncRNAs indicates that they might represent misclassified protein-coding genes coding for a functional small protein, as was shown in (Anderson et al, 2015; Kondo et al, 2010). It was also surprisingly reported that the majority of cytoplasmic lncRNAs associate with ribosomes (Ingolia et al, 2011). However, misclassification of a protein-coding gene into a lncRNA appears to be exceptional cases since it was later clearly shown that the majority of cytoplasmic lncRNAs do not produce a protein even when associated with ribosomes (Banfai et al, 2012; Gascoigne et al, 2012; Guttman et al, 2013). Several cytoplasmic lncRNAs are known (Brannan et al, 1990; Coccia et al, 1992; Klattenhoff et al, 2013). However, as discussed above (see *1.3 Functions and mechanisms of lncRNAs*), the functions known to be most commonly performed by lncRNAs, such as chromatin modifier guidance (Wang & Chang, 2011) are inherently nuclear. Nuclear functions of lncRNAs are reviewed in (Nakagawa & Kageyama, 2014), (Vance & Ponting, 2014) and (Quinodoz & Guttman, 2014). Nuclear retention of strictly nuclear-functional lncRNAs can be exemplified with pioneer lncRNAs, such as *Airn*, *Xist* and *NEAT1*, as well as with recently discovered lncRNAs, such as *FIRRE* (Hacisuleyman et al, 2014) and *Peril* (Sauvageau et al, 2013).

Adding more points to the differences between lncRNAs and mRNAs, lncRNAs were also shown to be less efficiently spliced (Tilgner et al, 2012) and less stable than mRNAs (Clark et al, 2012). Inefficient splicing was a highlight distinguishing several well-known functional imprinted lncRNAs, such as *Airn* (Seidl et al, 2006), *Kcnq1ot1* and *Nespas* (Kanduri, 2015). Another stage of RNA processing that appears to be less efficient for lncRNAs compared to mRNAs, is their polyadenylation. It has been shown that a large portion of lncRNA transcripts is not polyadenylated and can only be identified in PolyA- but not in PolyA+ enriched RNA-seq (Yang et al, 2011).

Table 2

Feature	Similarities	Differences		Related references
		lncRNAs	mRNAs	
Function	None	Various, mostly unknown	Template for protein synthesis	(Rinn & Chang, 2012)
Basic transcription	Transcribed by RNAPII	Differences not reported		(Rinn & Chang, 2012)
Promoter features	Transcribed from classical promoters	Can also be transcribed from enhancers	Always transcribed from a classical promoter	(Marques et al, 2013)
Expression level	Expression levels range from low to very high	Majority – lowly/marginally expressed (RPKM<1)	Majority – highly expressed (RPKM>1)	(Cabili et al, 2011)
Tissue-specificity	Tissue specific expression of some genes	Majority of genes – tissue-specific	Majority of genes – not tissue-specific	(Cabili et al, 2011)
Splicing	Can be spliced	Inefficient splicing of a part of the population	Efficiently spliced	(Tilgner et al, 2012)
Polyadeny-lation	Can be polyadenylated	Large proportion is not polyadenylated	Most are efficiently polyadenylated	(Yang et al, 2011)
Sequence conservation	Promoter conservation, some genes highly conserved	Majority - rapid evolution, low conservation, gene body conservation – extremely low	Many genes - well conserved over exons – introns not conserved	(Johnsson et al, 2014; Necsulea et al, 2014)
Expression conservation	Show some level of tissue specific expression conservation	Low expression conservation	High expression conservation	(Necsulea et al, 2014)
Repeat content	Introns mostly contain repeats	High repeat content in exons, many initiate from retrotransposons	Coding exons lack repeats, 3' non-coding exons may contain repeats	(Kapusta et al, 2013; Kelley & Rinn, 2012)
Stability	Stability ranges from low to high	Reduced stability	Majority – highly stable	(Clark et al, 2012)
Chromatin marks	Similar chromatin marks marking active and inactive genes	Can be transcribed from enhancers		(Orom et al, 2010; Rinn & Chang, 2012)
Natural expression variation	Expression variation mostly controlled by genetic variation	Differences not reported		(Lappalainen et al, 2013)

Table 2. Summary of similarities and differences between lncRNAs and mRNAs.

1.5.3 Natural variation of gene expression

While several features of lncRNAs were investigated and compared to mRNAs, one important feature of lncRNAs, namely natural expression variation, has not been assessed (Table 2). This feature is of great importance, especially when studying humans, since different individuals possess unique genomes distinguished by millions of SNPs. Natural variation of mRNAs has been extensively studied and implicated in health (Dumeaux et al, 2010), underlining the importance of expression variation characterization. Any difference between lncRNA and mRNA expression variation is thus also of particular interest, but has not yet been systematically studied. One study that focused on natural variability of protein-coding gene splicing using lymphoblastoid cell lines (LCL) from healthy donors, reported increased expression variation compared to mRNAs for a small number (183) of GENCODE v12 annotated lncRNAs (Gonzalez-Porta et al, 2012). However, this study did not support this conclusion with control analyses for lncRNAs.

When studying gene expression variability between healthy humans, both cell lines established from samples collected from healthy donors and primary tissues and cell types can be used. While cell lines have the advantage of standardized and controlled culturing procedures and provide enough material to perform any follow-up experiments, primary tissues reflect the health state of the donor more directly and the gene expression captured from primary tissues is a more 'natural' phenomenon. This might be more relevant for medical implications than analysis of gene expression in a cell line, which can be affected by immortalization and culturing procedures. Thus, while cell lines provide an important tool for basic and functional research on natural gene expression variation, primary cell types and tissues have to be analyzed in order to give gene expression variation studies more translational value.

An invaluable resource initially created by the 1000 Genomes Project http://www.1000genomes.org/) comprises a collection of lymphoblastoid cell lines (LCL) from hundreds of individuals from various populations. Fresh blood was collected from these individuals and LCL were established by Epstein-Barr virus transformation of white blood cells, which transforms and immortalizes B cells. This LCL collection has since been actively used for natural gene expression variation studies (Cheung et al, 2003; Gonzalez-Porta et al, 2012; Lappalainen et al, 2013; Spielman et al, 2007; Storey et al,

2007), which mainly focused on studying protein-coding genes. Availability of genome sequence data along with extensive transcriptome profiling of hundreds of LCL samples from the 1000 Genomes collection allowed another major project called GEUVADIS to conclude that genetic variation is the main cause of the expression variation between donors (Lappalainen et al, 2013). LncRNAs and mRNAs were analyzed together and for the both classes of genes it was shown that the observed expression variation between different populations is comprised of two types of variation - expression strength and isoform structure variation that contributed differently to expression variation in different human population pairwise comparisons (Lappalainen et al, 2013).

While the majority of gene expression variation studies have been performed in cell lines, some studies analyzed protein-coding gene expression variation in primary tissues (Chowers et al, 2003; Whitney et al, 2003). Analysis of gene expression variation using primary tissues has been mainly performed on blood (Dumeaux et al, 2010; Whitney et al, 2003) – the tissue most easily available from healthy individuals, but also the tissue with the highest routine diagnostic value. These studies identified that protein-coding gene expression changes may depend on various individual (BMI, age, lifestyle) as well as technical factors (time of the blood collection). These studies, however, did not include lncRNAs in their analyses. Thus the degree of lncRNA expression variation in human primary tissues or cell lines had been unclear, as well as the difference between lncRNAs and mRNAs in that regard. Accessing lncRNA variability in primary tissues, such as blood, is clearly of high importance, given the diagnostic value of blood together with the suggestions of lncRNAs serving as disease and prognostic biomarkers.

1.6 LncRNA discovery and annotation

As described above, the first mammalian lncRNA called *H19* was discovered in 1988 (Pachnis et al, 1988). After that it took almost two decades to realize that genomes of mammals and other organisms are full of lncRNA genes (Carninci et al, 2005; Guttman et al, 2009). The first genome-wide identification of non-coding RNAs, including lncRNAs, was performed by means of CAGE-tag analysis that traps the 5' end of a transcript (Carninci et al, 2005). Consequently, tiling arrays that allow hybridization of the whole transcript were also use to map lncRNAs (Kapranov et al, 2007; Vlatkovic, 2010b). While tiling arrays allow unprecedented sensitivity of transcript identification

(Mercer et al, 2012), they do not allow to assess exon structure or strandness of the identified transcripts and also cannot distinguish multicopy genes or pseudogenes (Vlatkovic, 2010b).

An indirect approach used genome-wide analysis of chromatin marks associated with active transcription (H3K4me3 – promoter mark, H3K36me3 – gene body mark) to identify thousands of lncRNAs in mouse (Guttman et al, 2009). However, out of the variety of methods capable of identifying lncRNAs expressed from the genome, RNA-seq has become the most preferable and commonly used. RNA-seq allows reconstruction of the exon structure of a lncRNA gene by mapping spliced transcripts to distant places with the confidence the split pieces are part of an intact transcript being provided by the way sequencing technology works (Trapnell et al, 2012). Moreover, RNA-seq allows inferring the strand from which a lncRNA is transcribed from both stranded (Iyer et al, 2015) and unstranded, by using information on canonical splice junctions (Cabili et al, 2011), RNA-seq data. Various sophisticated software has been developed in order to more robustly and sensitively assemble lncRNA gene structures and calculate their abundance (Cufflinks (Trapnell et al, 2012), Trinity (Tan et al, 2013)). Interestingly, such software allows *de novo*, or *ab initio*, transcriptome assembly that does not require information on any existing gene annotation.

Due to technological innovation facilitating lncRNA discovery, thousands of lncRNAs have been mapped in the genomes of various organisms (Ulitsky & Bartel, 2013). However, a comprehensive lncRNA annotation is still missing for any organism. Although lncRNAs are present in the genomes or nearly all organisms, the main focus of this thesis is human lncRNAs. Numerous studies made significant effort to annotate lncRNA genes in the human genome (Table 3), however the human lncRNA annotation is still likely incomplete. The precise reasons for incomplete lncRNA annotation in human (or other organisms) have not been explained, however, unusual lncRNA features that distinguish them from mRNAs might contribute to the difficulties in lncRNA identification.

Table 3

Reference	Tissues analyzed	Data for Transcript Reconstruction	Genomic Features and Filters	Coding-Potential Filters	Number of lincRNAs
(Khalil et al, 2009)	6 human cell types: hESC, hematopoietic stem cells (CD133+ and CD36+), T-cells, hLFs and normal embryonic kidney (hEK)	Chromatin marks, tiling arrays	Collection of approximate exonic regions, chromatin domain ≥ 5 kb	CSF	3,289 loci
(Jia et al, 2010)	Multiple tissues and cell lines	cDNAs from public sources	Overlap with mRNAs allowed	(ORF)-Predictor/ BLASTP pipeline	5,446 transcripts
(Orom et al, 2010)	Multiple tissues and cell lines	cDNAs from public sources	Restricted to loci >1 kb away from known protein-coding genes, ≥200 nt mature length	Manual curation based on length, conservation and other characteristics of the ORFs	3,019 transcripts from 2,286 loci
(Cabili et al, 2011)	24 human tissues and cell types	RNA-seq	Multi-exon only, ≥200 nt mature length	PhyloCSF, Pfam	8,195 transcripts (4,662 in the stringent set)
(Derrien et al, 2012) GENCODE	15 human malignant and normal cell lines	cDNAs	Overlap with mRNAs allowed (intergenic transcripts reported separately), ≥200 nt mature length	Manual curation based on length, conservation and other characteristics of the ORFs	14,880 transcripts from 9,277 loci, including 9,518 intergenic transcripts
(Sigova et al, 2013)	embryonic stem cells and endodermal cells	RNA-seq, cDNAs, chromatin marks,	Antisense overlap with mRNA introns allowed, ≥100 nt mature length	CPC	3,548 loci (hESC) 3,986 loci (Endoderm cells)
(Hangauer et al, 2013)	23 human tissues under multiple conditions + public data	Public and own RNA-seq	Single and multi-exon >200nt, FPKM>1	Filter ORF>100 amino acids	53,864 loci
(Xie et al, 2014) NONCODE	Multiple various tissues and cell types used by other datasets	Public lncRNA annotations combined and filtered	Multiple databases merged – redundancy elimination	CNCI software (Sun et al, 2013b)	95,135 transcripts from 56,018 loci
(Ranzani et al, 2015)	13 lymphocyte cell subsets	RNA-seq	>200nt, multi-exon, intergenic	HMMER-3, PhyloCSF	4,764 lncRNA loci (4,201 annotated + 562 novel)
(Iyer et al, 2015)	>18 organ systems (5,298 datasets from primary tumors, 281- from metastases, and 701 - from normal or benign adjacent tissues)	RNA-seq	>250nt, exons>15bp,	Coding Potential Assessment Tool (CPAT) (Wang et al, 2013) + presence of a known Pfam domain within ORF	58,648 lncRNA loci

Table 3. Genome-wide human lncRNA identification efforts (modified, supplemented and updated from (Ulitsky & Bartel, 2013))

Multiple studies that analyzed different cell lines and primary cells and tissues immediately recognized the outstanding tissues specificity and generally low level of lncRNA gene expression (Cabili et al, 2011), a major characteristic feature of lncRNAs described above (see *1.5.2 LncRNA features compared to mRNA features*). High tissue-specificity results in a significant challenge in lncRNA identification: analyzing just one tissue will reveal only a small portion of all lncRNAs present in the human genome – the portion that is expressed in this tissue and can therefore be detected by RNA-seq. Thus, the first commonly accepted guideline for lncRNA identification strategy is to analyze multiple tissues and cell types. Moreover, low lncRNA expression level makes it necessary to analyze the most pure cell types/tissues possible in order not to mask lowly expressed lncRNAs in the cell type present in the minority. Additionally, generally high coverage RNA-seq is needed to detect all the lncRNAs, especially the outstandingly lowly abundant transcripts. Identification of some extremely lowly abundant lncRNAs sometimes may require additional techniques to be applied. These might include, for example, the use of tiling arrays (Mercer et al, 2012) or FISH (Cabili et al, 2015) that are capable of capturing even single transcripts per cell, or exosome knockout, that stabilizes and allows detection of exosome depleted promoter-associated lncRNAs or PROMPTs (Preker et al, 2011). Usually RNA-seq of the PolyA+ enriched RNA fraction is used for lncRNA identification, however, due to other non-mRNA-like features of lncRNAs, such as low splicing efficiency and inefficient polyadenylation discussed above, a complete representation of all lncRNAs in the analyzed PolyA+ RNA fraction will be reduced. Many lncRNAs were reported to be repeat rich (Kelley & Rinn, 2012) and increased repeat content might also confound annotation of lncRNA through confusing the mapping of RNA-seq reads from repeat-rich regions back to the genome.

It is known that the process of lncRNA identification requires rigorous filtering of the initial transcriptome assembly obtained from the RNA-seq (see, for example, Supplemental Methods in (Necsulea et al, 2014)). Filtering is important to remove assembly artifacts and, most importantly, potential protein-coding transcripts. Several software tools accessing protein-coding potential have been developed (Kong et al, 2007; Washietl et al, 2011). Additionally, it is possible to mistake a fragment of a *de novo* assembled protein-coding gene extension for a lncRNA, since this piece can have no protein-coding potential, and thus it is crucial to assemble mRNAs *de novo* together with

lncRNAs in order to identify potential extensions in the analyzed system (Hangauer et al, 2013).

1.7 Classification of lncRNAs

With the annotation of numerous lncRNA genes (Table 3) the need for clear and useful classification arises. However, the extreme diversity of the lncRNA population and the ever-emerging new knowledge on the features and functions of various lncRNAs make the classification notably challenging (Kapusta & Feschotte, 2014; St Laurent et al, 2015). There is a dozen of various types of classification that may be applied to lncRNAs (St Laurent et al, 2015). Figure 7 overviews 7 of them (see Figure legend, Figure 7).

LncRNAs can be classified by their genomic context, for example by their relative position to protein-coding genes, or, alternatively, other genomic elements, such as enhancers, promoters, LTRs, etc. (Figure 1, Figure 7A). Next, lncRNAs were proposed to be classified by the chromatin characteristics of their TSS as two big distinct classes were shown to either initiate from promoter-like (H4K3me3) or enhancer-like (H3K4me1) TSSs (Marques et al, 2013) (Figure 7B). Classification can be considered successful and appears meaningful if the defined classes show consistent differences not only in the feature, that the separation into classes was based on, but also in other features (Marques et al, 2013). LncRNA can also be classified by their cellular localization, such as nuclear/cytoplasmic distinction (Guenzl & Barlow, 2012) (Figure 7C). Alternatively, lncRNAs can be classified by the end product of their processing – while the majority stay unprocessed, some are processed to small RNAs such as miRNAs, snoRNAs and piRNAs (Figure 7D). Once the function of a certain number of lncRNAs is defined, those lncRNAs can be classified by their function or the mechanism that is used to perform the function (Figure 7E). For example, HOTAIR (Gupta et al, 2010) and HOTTIP (Wang et al, 2011b) lncRNAs might join the class of chromatin modifier guides, while PTEN (Poliseno et al, 2010) can join the class of miRNA sponges, also known as competing endogenous RNAs (ceRNAs), and Airn (Latos et al, 2012) – the class of lncRNAs acting through transcription interference (Figure 7E).

Figure 7

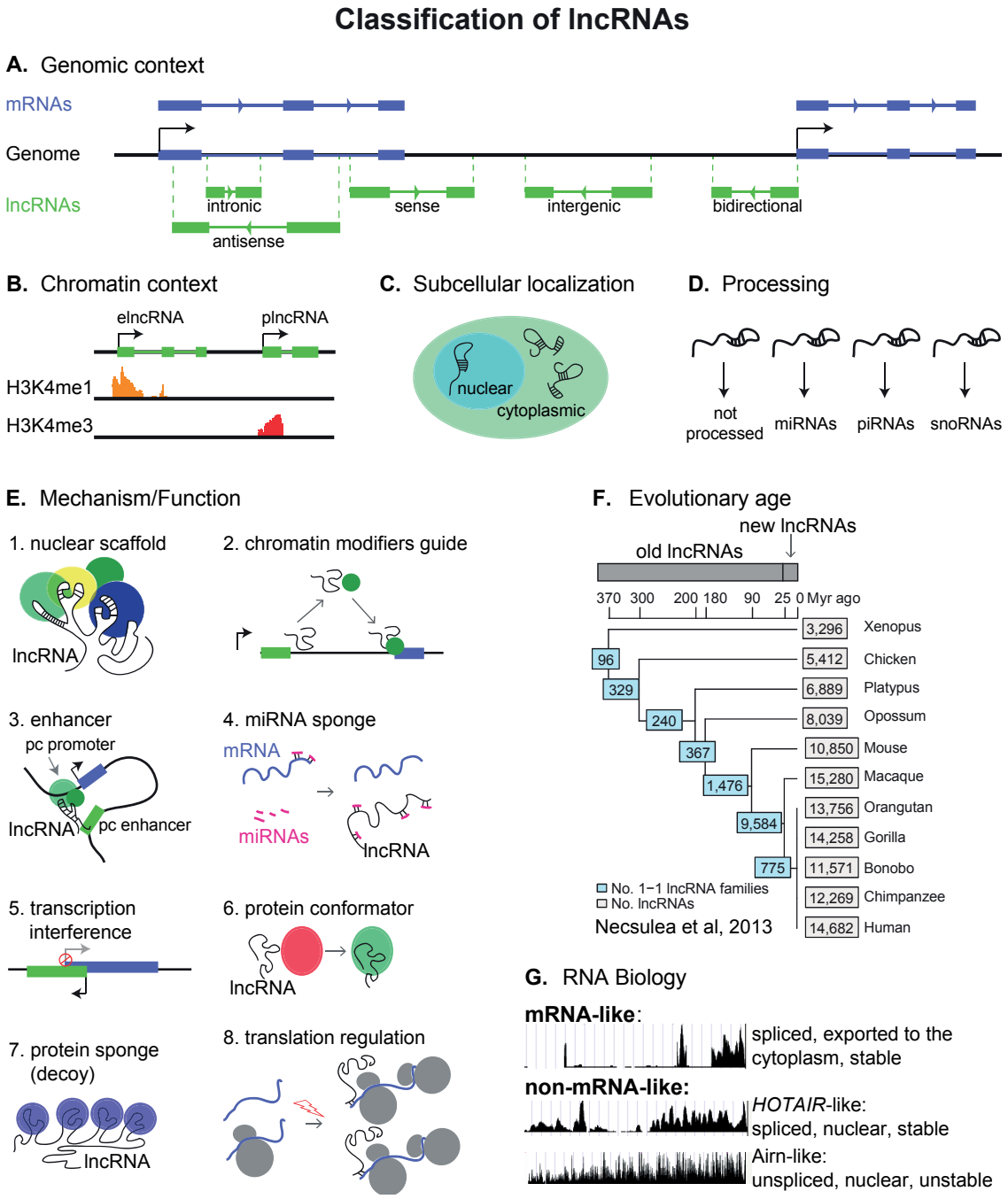


Figure 7. Various lncRNA classification approaches. **A. Genomic context.** LncRNAs can be classified based on their genomic context, such as, mainly, relative position to a nearby protein-coding gene – intronic, antisense, sense, intergenic and bidirectional lncRNAs (Derrien et al, 2012; Guttman et al, 2009). **B. Chromatin context.** LncRNAs can be initiated either from classical promoters (H3K4me3/H3K27ac) or classical enhancers (H3K4me1/H3K27ac) (Marques et al, 2013). **C. Subcellular localization.** LncRNAs can be classified by their localization in the cell. Broadly – if they reside in the nucleus or can be exported to the cytoplasm (Derrien et al, 2012), but also more precise localization can be assigned, such as localization inside paraspeckles (Nakagawa et al, 2011) or speckles (Tripathi et al, 2010). **D. Processing.** While the majority of lncRNAs are the end product of transcription, some lncRNAs may serve as precursors for small ncRNAs – for example, miRNAs, piRNAs and snoRNAs (Affymetrix_ENCODE_Transcriptome_Project, 2009; Mattick & Makunin, 2005) **E.**

Mechanism/Function. LncRNAs perform a variety of functions via a variety of mechanisms and can be classified based on that (see Figure 4 and (Kornienko et al, 2013)). **F. Evolutionary age.** LncRNAs vary from being highly conserved to human-specific, thus they can be classified according to their evolutionary age (Necsulea et al, 2014) (illustration taken from (Necsulea et al, 2014) with reprint permission). **G. RNA Biology.** As described above some parts of lncRNA population may display a various set of features distinguishing them from mRNAs and within the lncRNA population. Based on these features they can be classified in the three classes illustrated, or, alternatively stratified more finely, by including more features, such as repeat content, into the stratification (Guenzl & Barlow, 2012; Quinn & Chang, 2015).

It is also possible to classify lncRNAs based on their inferred evolutionary age, and such classification allows to create clearly distinct lncRNA classes (Necsulea et al, 2014) (Figure 7F). For example, Necsulea et al., annotated lncRNAs in various tetrapod species, searched for the homologs of human lncRNAs throughout the analyzed evolutionary tree and thus were able to assign an age to every human lncRNA and classify them into “new” (”young”), with the minimum evolutionary age of up to 25 million years, and “old”, with the minimum evolutionary age of 90-370 million years (Necsulea et al, 2014). RNA biology features also appear to be useful for lncRNA classification (Figure 7G). Most of these features, such as expression level, splicing efficiency and the degree of cytoplasmic export, are parametric values, thus a stratification of a various degree of detail is possible (Guenzl & Barlow, 2012).

Overall, the lncRNA field does not, to date, have enough knowledge on a large enough sample of lncRNAs to create a standardized classification method, which would maybe have to be notably complex, potentially including all the above mentioned classification approaches and features.

1.8 Debate on lncRNA transcription meaningfulness

While the number of lncRNA genes identified in the human genome steadily continues to increase, the argument on the meaningfulness of this vast lncRNA transcription persists (Kowalczyk et al, 2012; Palazzo & Lee, 2015). This argument is a part of a more general discussion on the pervasiveness of the transcription in the human genome that had not been expected in the earlier times (Raabe & Brosius, 2015) and even recently, 1.5 decades ago when the human genome sequence has just been solved (Lander et al, 2001). The development of the RNA-seq technology and analysis of multiple tissues and cell types allowed the ENCODE project to claim that 83.7% of the human genome can be

transcribed (Djebali et al, 2012). Moreover, the analysis of the histone and DNA modification patterns together with transcription factor binding sites throughout the genome in various tissues led the ENCODE project to the conclusion that 80% of the human genome is likely to be a subject to active regulation (ENCODE_Project_Consortium, 2012), which led some researches to even stronger claims, namely that the genome is widely functional (Djebali et al, 2012; Kellis et al, 2014). However, these claims and conclusions collided with the established view on the human genome being mostly populated by "junk" DNA and provoked massive criticism mainly coming from the evolutionary field (Doolittle, 2013; Graur et al, 2013). In contrast, other researchers suggest that in light of new discoveries and the certitude of the existence of the pervasive transcription we need to adjust the definition of the gene (Raabe & Brosius, 2015), as discussed above. LncRNAs constitute the major part of the pervasive transcription in the human genome (Djebali et al, 2012). Thus, the whole scope of criticism falls mostly on this gene type and stirs up the debate on whether the thousands of reported lncRNAs in the human genome are functional (Kowalczyk et al, 2012). In light of decades of using conservation level for inferring protein-coding gene functionality, low conservation is the main weak-point of lncRNAs, giving some researches no reason to assume any functionality of lncRNAs (Ulitsky & Bartel, 2013). Even the apparent tight tissue-specific regulation, which serves as a strong meaningful-transcription-argument for some (Mattick, 2011), is argued away by reasoning that this is a by-product of a tissue-specific chromatin landscape allowing this or that extent of meaningless lncRNA transcription from open chromatin sites (Ulitsky & Bartel, 2013). LncRNA 'sceptics' claim that lncRNAs should be *a priori* considered as "junk" and non-functional unless the contrary is solidly proven (Palazzo & Lee, 2015).

In the light of the above discussions, proof of the functionality/non-functionality of all the discovered lncRNA genes in human and other organisms is necessary. However, it may take decades of experimental research and the development of high-throughput functionality assays optimized for lncRNAs, as well as a considerable effort to attract more researchers to study lncRNAs, such that doubt against lncRNA meaningfulness is defeated.

1.9 Assigning functionality to lncRNAs

Assigning functionality or absence of functionality to a lncRNA is an outstanding challenge and notably less straightforward compared to investigating functions of small RNAs, such as piwi RNAs or miRNAs (Clark et al, 2013). If we consider pioneer lncRNAs, such as *H19*, *Xist* and *Airn*, it took decades to establish their functions and identify mechanisms, while *H19* is, after nearly 3 decades after its discovery, still disputed of being functional (Gabory et al, 2010; Keniry et al, 2012).

The goal of assigning functionality to lncRNAs is complicated for multiple reasons. The first challenge, given tens of thousands of identified lncRNAs, is how to choose a target for a functional study. Importantly, unlike in the case of mRNAs, function of a lncRNA cannot, to date, by any means be inferred from its genomic sequence (Ponjavic et al, 2009). Moreover, while mRNA conservation is used as a clear indicator of functionality, the conservation level of a lncRNA is not indicative of functionality, as discussed above. Neither the dynamic pattern nor the tissue-specificity of expression of a lncRNA is predictive of functionality (Ulitsky & Bartel, 2013). Nevertheless, several studies suggested pipelines for pinpointing a relatively small set of lncRNAs with a high likelihood of functionality from thousands of annotated genes (Nielsen et al, 2014; Sauvageau et al, 2013). These pipelines contain rigorous protein-coding potential filtering, which includes sequence analysis to detect potential encoding for short peptides, codon-substitution-frequency analysis, ribosomal profiling analysis as well as mass spectrometry data mining excluding any lncRNAs matching identified peptides (Sauvageau et al, 2013). Additional filtering steps might be undertaken to further filter the candidate list. For example, it seems rational and practical to study mice lncRNAs that have orthologues in human. Next, it is plausible that lncRNAs genes showing clear and canonical histone modification as well as indication of enhancer regulation might be preferential for functional analysis (Sauvageau et al, 2013). Filtering for lncRNAs that were previously found to be bound by regulatory proteins, such as PRC2, is another reasonable step. However, even such seemingly logical and rigorous filtering approaches might not result in identifying essential lncRNA. For example, Sauvageau et al., applied this pipeline to select 18 mouse lncRNAs for *in vivo* functional assessment and only three of them were found to be essential (Sauvageau et al, 2013). It is however, unclear, if the lncRNAs whose knock-outs did not result in a lethal phenotype do not perform a more

subtle, difficult to detect function, for example in adaptive behavior or in resistance towards certain diseases.

Another approach for choosing a lncRNA as a functional study target would be to pick those identified as hits in a drug resistance/vulnerability screening approaches (Fatemi et al, 2014; Mohr et al, 2014), or a small set of lncRNAs that show highest deregulation upon a certain disease condition (Sun et al, 2013a). Another useful hint towards functionality allowing to filter the list of lncRNAs undergoing functional investigation is a reported presence of disease/trait associated SNPs, identified by relevant GWAS studies, in the lncRNA gene body (Pasmant et al, 2011).

Several techniques, most of which have been developed or are being developed for genome-wide lncRNA study, allow more closely studying lncRNAs and getting information on which lncRNAs might be involved in the regulation in the cell. Such techniques are, for example, - FISH (Cabili et al, 2015), which is the least high-throughput one in the hereby mentioned techniques. More high-throughput methods aimed at investigating lncRNA interaction with proteins (CLIP, (Huppertz et al, 2014)), other RNAs (CLASH, (Helwak et al, 2013)) or DNA (CHART (Vance & Ponting, 2014) or ChIRP (Chu et al, 2011)) can be applied. However, special care has to be taken to eliminate non-specific binding in these assays (Brockdorff, 2013).

1.9.1 Approaches to study lncRNA function

Hints on how a lncRNA can potentially function are crucial for an appropriate design of loss-of-function (knock-out/knock-down) experiments. For example, the binding assays listed above might be useful. Assessing lncRNA binding partners might show it binds to a promoter of a gene and at the same time to a transcription inhibitor, such as a histone demethylase, which would strongly argue for a function through the lncRNA product, rather than transcription. However, there are studies that indicate that both mechanisms can coincide in one lncRNA, as was shown for the *Airn* lncRNA in mouse (Latos et al, 2012; Nagano et al, 2008).

Airn is an example of how pioneer lncRNAs may serve as a paradigm for lncRNA research. This lncRNA was initially studied *in vivo* in mouse models, where its function

in the silencing of three imprinted genes in the *Igf2r* cluster was shown (Sleutels et al, 2002). However, later on, an appropriate *in vitro* cell line system was used, which sped up the performing of sophisticated genetic manipulations to elucidate the *Igf2r*-silencing mechanism (Latos et al, 2012). Using the cell line system also facilitated conditional knock-out/knock-in experiments that elucidated the dynamics of *Igf2r*-repression (Santoro et al, 2013).

Cellular localization of a lncRNA is important when deciding on a certain knock-down approach (Bassett et al, 2014). While cytoplasm-localized transcripts can be efficiently targeted using RNAi approaches, e.g., administration of corresponding si- or shRNAs into the cell, nuclearly localized lncRNA might be targeted inefficiently (Lennox & Behlke, 2015) and, as discussed above, the majority of lncRNAs are nuclear (Derrien et al, 2012). Nuclear, or even chromatin-, localized lncRNAs would be preferentially knocked down using antisense chemically modified oligonucleotides that sequence-specifically bind the lncRNA of interest and initiate its depletion by RNase H (Tripathi et al, 2010). However, if a lncRNA acts through its transcription, such assays would not remove the functionality of this lncRNA. In this case, a genetic manipulation of the locus of interest is required. Genetic manipulations, such as lncRNA gene body or promoter deletion, as well as genetic truncation (insertion of a transcription terminator in the lncRNA gene body) appear to be more universal than knockdown approaches since they remove both transcript and transcription and avoid unspecific RNAi effects.

Creating genetic deletions of lncRNA promoter/gene body has become notably easier with the invention of targeted genome-editing tools such as TALENs and CRISPR/Cas9 system (Bassett et al, 2014). Importantly, such tools not only revolutionized genome research in general, but also allow *in vitro* manipulation of the human genome facilitating human lncRNA research. However, lncRNA gene body/promoter deletion can potentially result in the deletion or disruption of unknown *cis*-regulatory elements affecting distant genes, thus complicating interpretation of the knock-out phenotype both *in vitro* and *in vivo* (Bassett et al, 2014; Sauvageau et al, 2013). Alternatively, insertion of a transcription terminator might provide a more beneficial approach. This approach allows for creating important controls, such as insertion of a terminator sequence in different locations within the lncRNA gene body, or insertion of this sequence in an irrelevant genomic location thus providing an insertion procedure control. Several well-known lncRNAs, such as

Airn, *MALAT1* and *Ube3a-as*, were studied using transcription termination (Gutschner et al, 2011; Latos et al, 2012; Meng et al, 2013) and provided important insights into lncRNA function. With the availability of several different techniques for abolishing a lncRNA of interest for investigating its fuction, and with the potential flaws associated with every one of them, it appears optimal to use multiple types of approaches to study lncRNA function.

Overall, cell lines are convenient for lncRNA function investigation and give high degree of freedom when designing knock-out/knock-down and rescue approaches, as well as the follow-up experiments for closer mechanism investigation (Tripathi et al, 2010). However, results initially obtained in cell lines are not always reproduced when investigating the same lncRNA *in vivo* (Zhang et al, 2012). Bassett et al., in their review in 2014 overviewed actual cases of genetic disruption of a lncRNA *in vivo*, the number of which appeared to be surprisingly small (Bassett et al, 2014), and the resulting phenotypes (Table 4).

Table 4

lncRNA name	Organism	Mutation strategy	Reported animal phenotype	RNA-based rescue?	Reference
Xist	*Mus musculus*	~15 kb replaced with a *neo* expression cassette	Females inheriting paternal allele were embryonic lethal; males fully viable	No	(Marahrens et al, 1997)
Xist	*Mus musculus*	Inversion of Exon 1 to intron 5	Embryonic lethality of paternally inherited allele	No	(Senner et al, 2011)
H19	*Mus musculus*	Replacement by *neo* cassette	Slightly increased growth	No	(Ripoche et al, 1997)
roX	*Drosophila melanogaster*	Deletions of *roX1* or *roX2*	None, except when in combination: male-specific reduction in viability	Yes	(Meller & Rattner, 2002)
Kcnq1ot1	*Mus musculus*	Promoter deletion	Growth deficiency for paternally inherited mutation	No	(Fitzpatrick et al, 2002)
Airn	*Mus musculus*	Premature transcriptional termination	Growth deficiency for paternally inherited mutation	No	(Sleutels et al, 2002)
Evf2	*Mus musculus*	Premature transcriptional termination	None	N/A	(Bond et al, 2009)
BC1	*Mus musculus*	Replacement of promoter and exon by *PgkNeo* cassette	Vulnerable to epileptic fits after auditory stimulation	No	(Zhong et al, 2009)

lncRNA name	Organism	Mutation strategy	Reported animal phenotype	RNA-based rescue?	Reference
Neat1	*Mus musculus*	3 kb Promoter and 5' deletion	None	N/A	(Nakagawa et al, 2011)
Tsx	*Mus musculus*	2 kb Promoter and exon 1 deletion	Smaller testes and less fearful (males)	No	(Anguera et al, 2011)
Malat1	*Mus musculus*	Deletion	None	N/A	(Eissmann et al, 2012)
Malat1	*Mus musculus*	*lacZ* insertion and premature transcriptional termination	None	N/A	(Nakagawa et al, 2012)
Malat1	*Mus musculus*	3 kb Promoter and 5' deletion	None	N/A	(Zhang et al, 2012)
Hotair	*Mus musculus*	Deletion	Spine and wrist malformations	No	(Li et al, 2013b)
Hotdog* and *Twin of Hotdog	*Mus musculus*	Large (28 Mb) translocation by inversion	Loss of *Hoxd* expression in the cecum	N/A	{Delpretti, 2013 #1072
Fendrr	*Mus musculus*	Replacement of exon 1 with transcriptional stop signal	Embryonic lethal around E13.75	Yes (majority of embryos)	(Grote et al, 2013)
Fendrr	*Mus musculus*	Locus replacement with *lacZ*cassette	Perinatal lethality	No	(Sauvageau et al, 2013)
Peril	*Mus musculus*	Locus replacement with *lacZ*cassette	Perinatal lethality	No	(Sauvageau et al, 2013)
Mdgt	*Mus musculus*	Locus replacement with *lacZ*cassette	Reduced viability and reduced growth	No	(Sauvageau et al, 2013)
15 other lncRNA loci	*Mus musculus*	Locus replacement with *lacZ*cassette	None	N/A	(Sauvageau et al, 2013)

Table 4. List of studies disrupting lncRNAs in vivo (table taken from (Bassett et al, 2014))

It is worth mentioning, however, that human lncRNAs appear to be of the highest interest, because they can be implicated in clinics. It is convenient to study function of such lncRNAs in mice and cell lines. Well-known examples of disease relevant lncRNAs studied in mice are *MALAT1* (Tripathi et al, 2013; Zhang et al, 2012), *NEAT1* (Nakagawa et al, 2014), *XIST* (Johnsson et al, 2014) and *UBE3A-AS* (Meng et al, 2015). However, as described above, lncRNAs are poorly conserved, thus just a small number of human lncRNAs can be investigated in model organisms. Thus an appropriate human system is of the highest importance.

1.9.2 Human Haploid Gene Trap Collection

Studying lncRNA function in human cell lines has become notably more convenient in the last several years due to significant technological progress. Moreover, to date it has become possible to design functional assays accessing multiple lncRNAs, which is of crucial importance, as discussed above. CRISPR/Cas9 genome editing has been widely proposed for simultaneously studying functions of multiple lncRNAs (Han et al, 2014) (Shechner et al, 2015). Additionally, gene trap technology potentially allows creating functional reversible knockouts of multiple protein-coding genes (Burckstummer et al, 2013), and, potentially, lncRNAs (Gutschner et al, 2011).

Gene trap technology includes inserting a transcription terminator sequence, which usually contains a strong splice acceptor followed by a polyA signal, into the gene body of the target gene with a goal of disrupting its functional transcription. Thus, in order to create a functional knockout, the gene trap has to be inserted closer to the 5'-end of the gene. The gene trap cassette is introduced using retroviral vectors randomly integrating into the genome and its orientation must coincide with the orientation of the gene to stop transcription (Stanford et al, 2001). Gene traps were extensively used to study protein-coding genes in mouse (Skarnes et al, 2004), but also, for example, allowed discovery of a mouse imprinted lncRNAs *Meg3* (Schuster-Gossler et al, 1996).

Recently The Human Haploid Gene Trap Collection – a useful tool for studying function of multiple genes based on gene trap technology – has been established and proposed to the research community (Burckstummer et al, 2013). The Human Haploid Gene Trap Collection comprises an extensive library of monoclonal cell lines with gene traps inside various genes providing numerous knockout cell lines (Burckstummer et al, 2013) (see Figure 8 for the overview of the collection creation steps). The cell line used for the creation of this collection is called KBM7 (Andersson et al, 1987) and is a malignant myeloid cell line with a haploid genome, that is it carries only single copies of each chromosomes (except for chromosome 8). Haploidy of KBM7 allows achieving full knock out as soon as a gene trap cassette is integrated in the correct orientation inside the body of the gene. Importantly, the gene trap cassette contains a reporter GFP gene which enables detection of successfully infected/targeted cell lines (see Figure 8 for the algorithm of GFP-positive cell selection and Figure 9 for the cassette scheme).

Figure 8

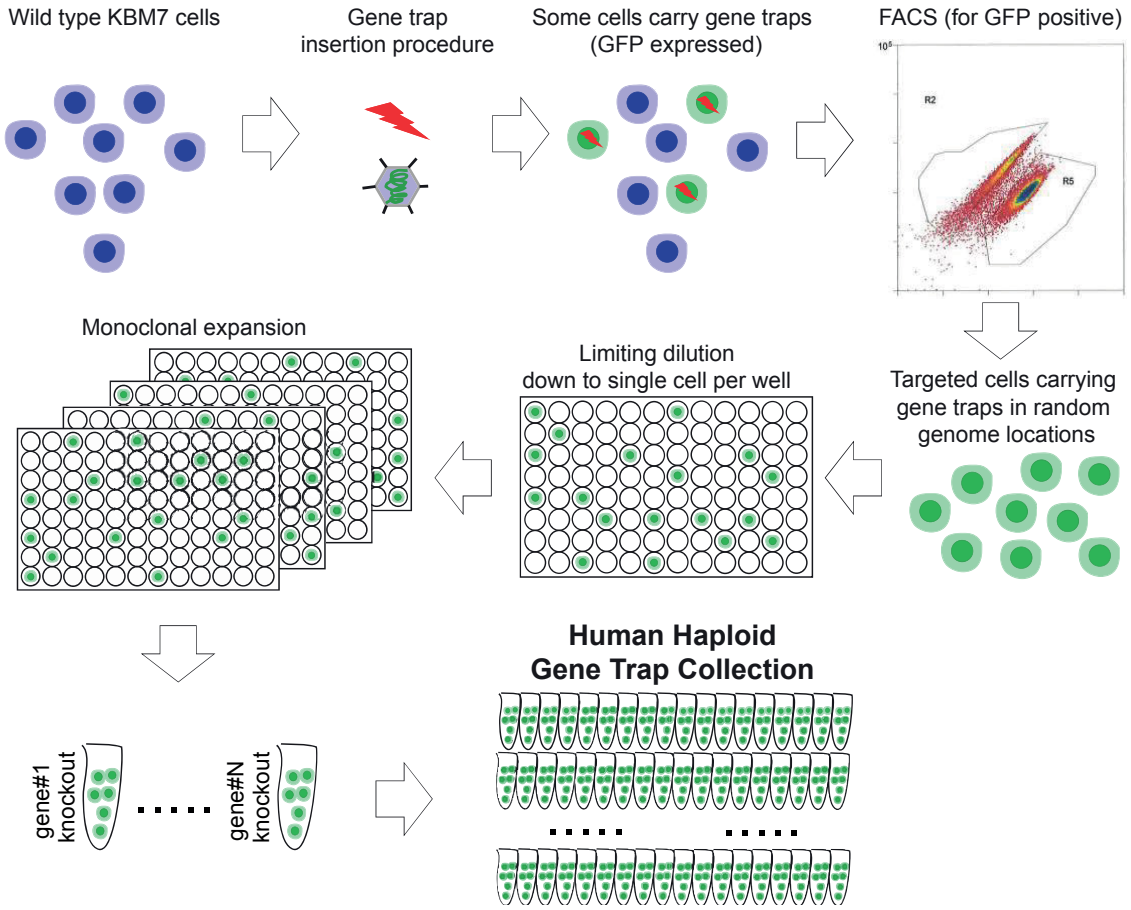


Figure 8. Scheme of The Human Haploid Gene Trap Collection creation. Wild type KBM7 cells undergo retroviral vector infection. Virus introduces a gene trap cassette (see Figure 9 for the scheme of the cassette) into random locations in the genome. Not every cell is successfully targeted, however, the presence of the GFP gene in the gene trap cassette allows distinguishing cells where the gene trap cassette integration was successful. FACS sorting allows extracting those cells from the pool of cells. Limiting dilutions are then performed in order to achieve 1-cell per well culture and expand this cell monoclonally. Afterwards, the cassette integration genomic site is identified for every colony and the targeted-gene/genomic-position is catalogued into a collection of ready-to-use knock-out monoclonal cell lines – the Human Haploid Gene Trap Collection (Burckstummer et al, 2013).

Importantly, since the integration of the retroviral vector into the genome is random, it can also target non-protein-coding regions. In fact, 45% of gene trap cassettes that produced GFP positive cells in the Human Haploid Gene Trap Collection were mapped to intergenic regions, and 23% were antisense to introns of protein-coding genes (Burckstummer et al, 2013). These gene traps most likely target lncRNAs expressed in KBM7 cells. Thus Human Haploid Gene Trap Collection appears as an invaluable tool to study numerous lncRNAs in a convenient, easy-to-handle cell system. Many genes have

multiple knockout KBM7 cell lines with gene trap cassettes inserted into different positions within the gene body, thus representing highly useful independent replicates.

Figure 9

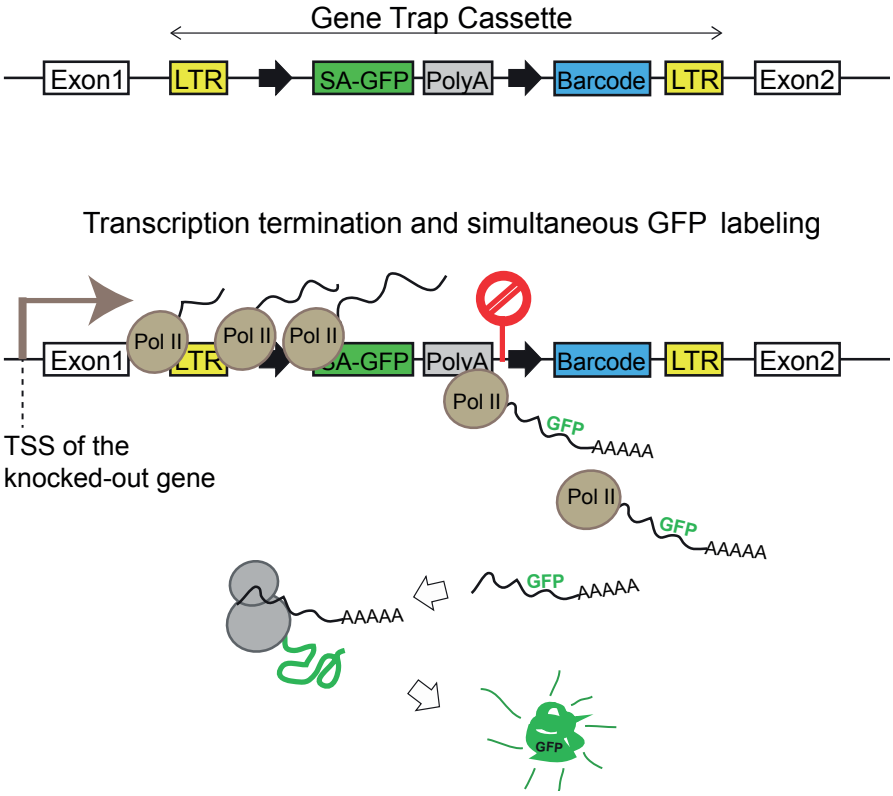


Figure 9. Scheme of the gene trap cassette and simultaneous GFP labeling of successfully targeted cell lines.

While Human Haploid Gene Trap Collection has been shown to be a useful tool to study protein-coding genes (Burckstummer et al, 2013; Carette et al, 2009), it has never been tested for lncRNA investigations. Importantly, since the gene trap cassette contains a strong splice acceptor which is supposed to 'hijack' RNAPII (Figure 9), it is unclear how efficiently such a cassette would truncate an inefficiently spliced lncRNA. As described above, inefficient splicing is one of the main non-mRNA-like features of lncRNAs. In addition, other differences between lncRNAs and mRNAs, such as nuclear localization, could contribute to difficulties in studying lncRNAs using the KBM7 collection.

1.10 Aims of this thesis

The overall aim of this thesis was to extend human lncRNA annotation and achieve deeper knowledge about human lncRNAs by accessing their features on a genome-wide level. I annotated lncRNAs in human primary granulocytes and in lymphoblastoid cell lines, thereby creating more comprehensive annotations than previously available, which contained numerous novel lncRNA loci. I discovered that high natural expression variability is a new important general feature of lncRNAs as a class that further distinguishes them from mRNAs and confounds their identification. This discovery provides valuable guidelines for lncRNA annotation, functional characterization and medical use. As a part of my Doctoral Thesis I aimed to functionally characterize a previously unstudied lncRNA *SLC38A4-AS* and found it to be a functional lncRNA with unusual RNA biology and a new regulatory lncRNA. By using the Human Haploid Gene Trap Collection for *SLC38A4-AS* functional analysis I provided a pipeline and a guidelines for the use of this valuable model system resource for studying hundreds of lncRNAs targeted in this collection.

2 RESULTS

2.1 Publication 2: "Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans" (Research Article)

Authors: Aleksandra E. Kornienko*, Christoph P. Dotter, Philipp M. Guenzl, Heinz Gisslinger, Bettina Gisslinger, Ciara Cleary, Robert Kralovics, Florian M. Pauler, Denise P. Barlow*

* corresponding authors

Published in Genome Biology (Impact factor 10.8) on 29.01.2016. Open access article, no reprint permission required.

Article webpage: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0873-8

As described in the INTRODUCTION, lncRNAs are an important class of genes, apparently vastly numerous in the human genome. While being transcribed by RNAPII and largely resembling mRNA genes, lncRNAs also display a variety of features that make them more challenging to identify and to approach functionally. We aimed to investigate the natural expression variability feature of lncRNAs that had not been in focus before, particularly in relation to mRNA variability. We used granulocytes, a relatively pure cell type, which has a potential to serve non-invasive diagnostic purposes, to access the variability of lncRNAs in healthy individuals. A granulocyte lncRNA landscape had not been defined before and, thus, we first annotated granulocyte lncRNAs as well as mRNAs for control purposes using PolyA+ RNA-seq data from ten healthy donors, thereby identifying numerous lncRNA loci absent from reference lncRNA annotations. We used this annotation to analyze expression variability in 7 healthy donors sampled 3 times under controlled conditions. We found that lncRNAs are significantly more variable than mRNAs. We confirmed this result using an independent lncRNA annotation and also using an independent RNA-seq data set from multiple human tissues. Thus we show that increased lncRNA variability is a general phenomenon. We also found that increased expression variability may contribute to incomplete representation of lncRNA genes in reference annotations by impeding their identification. We could demonstrate that including more donors into lncRNA identification pipeline allows the identification of more reference annotated lncRNA genes expressed in one cell type, but,

importantly, also allows extension of human lncRNA annotation by identifying novel lncRNA genes.

Overall, I, with the help of the co-authors, and supervised by Denise P. Barlow, have performed a massive study, which was commenced with a notable amount of experimental optimization and bioinformatic pipeline development effort, followed by the experimental work and bioinformatic analysis leading to the results presented in the manuscript attached below.

Authors' contributions:

"A.E.K. and D.P.B. conceived the study and wrote the manuscript. A.E.K. performed blood sample processing, library preparation, experimental work, *de novo* lncRNA and mRNA identification and other bioinformatic analyses. C.C. prepared the majority of PolyA enriched RNA-seq libraries. P.M.G. established RNA-seq protocols and contributed to the splicing calculation method. F.M.P. and C.P.D. assembled the protein-coding potential estimation pipeline, wrote some custom scripts used in the study and helped with the bioinformatic analysis. Blood samples were collected in collaboration with H.G., B.G. and R.K. All authors read and approved this manuscript."

(N.B. Authors' contributions are copied from the manuscript attached below and thus are enclosed in quotes)

Kornienko et al. Genome Biology (2016) 17:14
DOI 10.1186/s13059-016-0873-8

Genome Biology

RESEARCH Open Access

Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans

Aleksandra E. Kornienko[1*], Christoph P. Dotter[2], Philipp M. Guenzl[1], Heinz Gisslinger[3], Bettina Gisslinger[3], Ciara Cleary[4], Robert Kralovics[1], Florian M. Pauler[1] and Denise P. Barlow[1*]

Abstract

Background: Long non-coding RNAs (lncRNAs) are increasingly implicated as gene regulators and may ultimately be more numerous than protein-coding genes in the human genome. Despite large numbers of reported lncRNAs, reference annotations are likely incomplete due to their lower and tighter tissue-specific expression compared to mRNAs. An unexplored factor potentially confounding lncRNA identification is inter-individual expression variability. Here, we characterize lncRNA natural expression variability in human primary granulocytes.

Results: We annotate granulocyte lncRNAs and mRNAs in RNA-seq data from 10 healthy individuals, identifying multiple lncRNAs absent from reference annotations, and use this to investigate three known features (higher tissue-specificity, lower expression, and reduced splicing efficiency) of lncRNAs relative to mRNAs. Expression variability was examined in seven individuals sampled three times at 1- or more than 1-month intervals. We show that lncRNAs display significantly more inter-individual expression variability compared to mRNAs. We confirm this finding in two independent human datasets by analyzing multiple tissues from the GTEx project and lymphoblastoid cell lines from the GEUVADIS project. Using the latter dataset we also show that including more human donors into the transcriptome annotation pipeline allows identification of an increasing number of lncRNAs, but minimally affects mRNA gene number.

Conclusions: A comprehensive annotation of lncRNAs is known to require an approach that is sensitive to low and tight tissue-specific expression. Here we show that increased inter-individual expression variability is an additional general lncRNA feature to consider when creating a comprehensive annotation of human lncRNAs or proposing their use as prognostic or disease markers.

Keywords: lncRNAs, expression variation, lncRNA identification, human genome annotation, granulocytes, transcriptome, natural variation, lncRNA features

Background

Long non-protein coding RNAs (lncRNAs) have emerged as a fundamental new layer of genomic information in diverse species [1]. They are considered to participate primarily in mRNA gene regulation [2–5] and to play roles in development and disease [6–8]. LncRNAs may be medically relevant as prognostic factors, disease markers, and drug targets [9–13]. To date, it is known that lncRNA genes are abundant in the genomes of human ([14], http://www.gencodegenes.org/stats.html), mouse ([15, 16], http://www.gencodegenes.org/mouse_stats.html), other vertebrates [17–20], plants [21], and simple model organisms such as *C. elegans* [22] and yeast [23, 24]. Although large numbers of lncRNAs have been identified, they have not yet been completely annotated in any organism. Human lncRNAs annotated by the GENCODE project comprise the largest public dataset containing 15,877 lncRNA genes (version 21: http://www.gencodegenes.org/stats/archive.html#a21). Many human annotation projects use cell lines [25], however, some also use primary tissues

* Correspondence: akornienko@cemm.oeaw.ac.at; dbarlow@cemm.oeaw.ac.at
[1]CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Lazarettgasse 14, AKH BT 25.3, 1090 Vienna, Austria
Full list of author information is available at the end of the article

[14, 26]. An incomplete annotation may arise from two known features of lncRNAs - low abundance and tight tissue-specificity [14, 25]. Notably, lncRNA annotations differ not just between tissues, but also between closely related cell types [27, 28]. Thus, a comprehensive map of all lncRNA genes in the human genome would require systematic and deep analysis of all human body cell types. A recent attempt to define the human lncRNA landscape used several thousand normal and malignant samples and identified almost 47,000 new lncRNA genes [29], supporting earlier predictions that lncRNAs may outnumber protein-coding genes in human [30].

Relatively small numbers of mammalian lncRNAs have been assigned a function. A new functional lncRNA database lists only 181 human transcripts (http://www.lncrnadb.org/, [31]). While it is possible that some lncRNA transcription is a consequence of the local chromatin state [32–34], the gap between annotation and proven functionality reflects the considerable challenges in the analysis of non-coding compared to coding transcripts [35–39]. A deeper knowledge of lncRNAs as a transcript class has followed from genome-wide characterizations of their biology and genomic features with mRNAs as a reference point (reviewed in [30, 34, 40]). Both types of transcripts are transcribed by RNAPII, possess histone modifications typical of active or inactive genes and can be spliced, capped, and polyadenylated (reviewed in [41]). However, in addition to the basic lack of an open reading frame and functional translation [42], some studies have identified characteristics that differentiate lncRNAs from mRNAs. In comparison to mRNAs, lncRNAs are generally found to be more lowly-expressed, show higher tissue-specificity and be enriched in the nucleus [14, 25]. Many lncRNAs initiate from enhancer-like promoters that lack H3K4me3 histone modifications typical of standard mRNA promoters [28, 43], or from repetitive transposable elements normally absent from standard mRNA promoters [44]. In terms of genome and biology features, lncRNAs are usually shorter with fewer exons and show inefficient co-transcriptional splicing [45] and reduced stability [46]. They also show low sequence conservation and evolve faster than mRNAs [47–49].

One lncRNA feature not yet fully investigated in comparison to mRNAs that may influence identification and functional characterization is their natural expression variation. Protein-coding and lncRNA expression and transcript structure have been shown to be dependent on genetic variation in the human lymphoblastoid cell line (LCL) collection [50–52]. Analysis of protein-coding gene expression in whole human blood shows expression variation attributable to inter-individual (for example, age, BMI) and lifestyle (fasting status, smoking) differences, and technical issues such as sampling time, collection and preparation [53, 54]. In this study we use human primary granulocytes, a relatively pure cell type routinely obtained in clinics from healthy individuals and potentially useful diagnostically, to assess natural variability of lncRNA expression. We first prepared an RNA-seq dataset from 10 healthy individuals to define a human granulocyte transcriptome, not previously available. From this we annotated 6,249 lncRNA transcripts arising from 1,323 previously reported and 268 novel lncRNA loci. We show that examining granulocytes from multiple donors allows the identification of less well expressed, less efficiently spliced, and more granulocyte-specific lncRNAs. We then estimated lncRNA expression reproducibility and variability in granulocyte RNA-seq data from seven healthy individuals sampled in three replicates with approximately 1-month intervals. This inter- and intra-individual comparison demonstrated that although lncRNA expression is reproducible between replicates from the same individual, it is significantly more variable between individuals compared to mRNAs. Analysis of multiple tissues from the GTEx project [55] and lymphoblastoid cell lines from the GEUVADIS project [50] supports this conclusion and also shows that higher natural expression variability compared to protein-coding genes is a general feature of lncRNAs. Using the latter dataset we show that natural expression variability markedly influences lncRNA identification as the number of identified lncRNAs increases with the number of donors analyzed and does not reach saturation even with 120 donors. Together, the data show that high expression variability of lncRNAs is an important general feature, which not only additionally distinguishes them from mRNAs, but also will make it necessary to consider the number of individuals in strategies to comprehensively annotate and assign putative functions to lncRNAs in the human genome.

Results

Defining the human granulocyte lncRNA transcriptome

To annotate lncRNAs in human granulocytes we collected samples from five male and five female healthy individuals of varying ages under standardized sampling conditions and sequenced polyadenylated (PolyA+) RNA (Fig. 1a, b, Additional file 1: Figure S1A and Supplemental Methods, Additional file 2A). Ribosome-depleted RNA-seq, used for expression and splicing efficiency analysis, was performed for seven donors (4 male donors, 3 female donors) at three time points. To annotate lncRNAs we aligned the PolyA+ RNA-seq data with STAR [56] to obtain 757 million uniquely-mapped reads of which 187.6 million were spliced (Additional file 2B, C) and performed *de novo* transcriptome assembly using Cufflinks and Cuffmerge [57]. The term '*de novo*' is used for transcripts/loci identified in this transcriptome

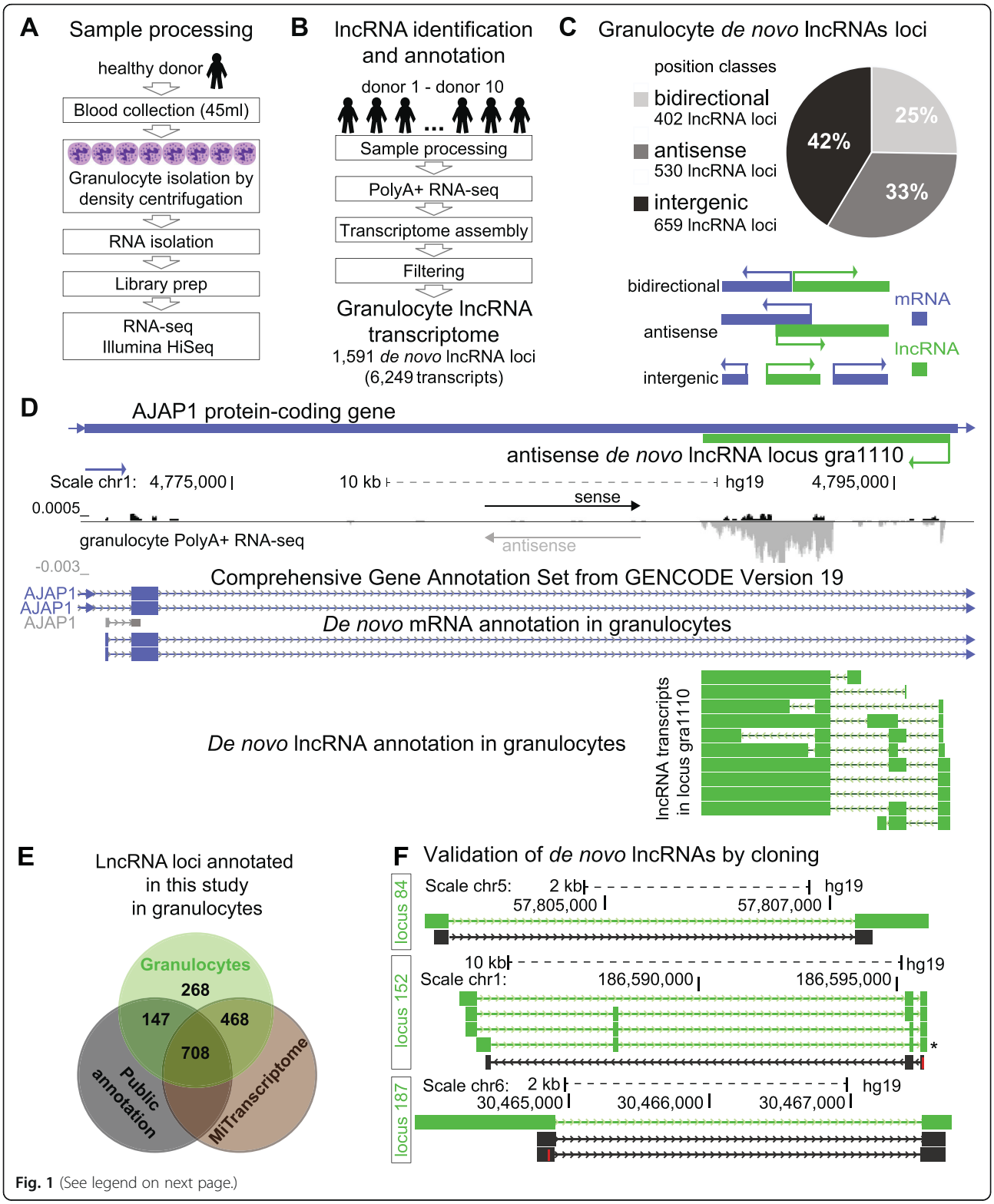


Fig. 1 (See legend on next page.)

(See figure on previous page.)
Fig. 1 Defining the lncRNA transcriptome of human primary granulocytes. **a** Sample processing overview. **b** LncRNA identification overview. Granulocyte PolyA+ RNA-seq data from 10 donors was used for transcriptome assembly and filtered to create an annotation with 1,591 lncRNA loci containing 6,249 lncRNA transcripts (Additional file 1: Figures S1-3). **c** Positional classification of lncRNA loci relative to the nearest protein-coding gene. Twenty-five percent (402) are bidirectional (light gray), 33 % (530) are antisense (medium gray), and 42 % (659) are intergenic (dark gray). Positional classes are illustrated underneath (blue: protein-coding gene, green: lncRNA). **d** Example of a novel granulocyte antisense lncRNA locus. Top: 3' part of *AJAP1* protein-coding gene (blue) and the novel antisense *gra1110* lncRNA locus (green). Underneath: normalized to read number RNA-seq signal from sample D2-2_pa_100ss (Additional file 2B); GENCODE-v19 protein-coding genes (blue lines) and *de novo* annotated mRNAs (blue) and lncRNAs (green) showing lncRNA transcripts in locus *gra1110* (Additional files 3, 4, and 6). **e** Overlap of granulocyte *de novo* lncRNA annotations (green) with commonly used public lncRNA annotations (gray) (RefSeq: 8,236 lncRNA transcripts, GENCODE-v19: 23,898 lncRNA transcripts, Cabili [14]: 21,630 lncRNA transcripts) and the 'MiTranscriptome' annotation (brown) [29]. **f** Validation of granulocyte *de novo* lncRNAs by cloning. Three *de novo* lncRNA loci (84, 152, 187) are shown (see also Additional file 1: Figures S4-S8). Top to bottom for each: scale and chromosome, *de novo* lncRNA transcript annotation in each locus (green isoforms), cloning result (black lines) showing BLAT alignment of the Sanger sequenced cloned cDNA

assembly pipeline. Only multi-exonic transcripts longer than 200 base pairs (bp) were retained and several filtering steps applied to remove potential assembly artifacts (Additional file 1: Figure S1). We next extracted multi-exonic transcripts overlapping exons annotated as protein-coding in GENCODE-v19 [58] and RefSeq [59] and used them later to generate a *de novo* protein-coding granulocyte mRNA annotation. We discarded annotated GENCODE-v19 pseudogene transcripts. To remove potential protein/peptide-coding transcripts, we estimated transcript protein-coding capability using RNAcode [60] and CPC [61]. We adjusted the criteria for the output of the protein-coding potential estimation pipeline (RNAcode score <18, CPC score <1.6) by analyzing well-known lncRNAs (Additional file 2D). We validated these criteria by applying the pipeline to the above public annotations; this identified the majority of annotated lncRNAs as non-protein-coding, whereas the majority of mRNAs were identified as protein-coding (Additional file 1:Figure S1E). To avoid confusion in later expression analysis we removed all lncRNAs overlapping a protein-coding gene in sense direction (for example, intronic lncRNAs) from our analysis. The final *de novo* lncRNA granulocyte annotation comprised 1,591 lncRNA loci (Additional file 3) expressing 6,249 lncRNA transcripts (Additional file 4) with a mean of 3.9 transcript isoforms per locus, consistent with previous observations [14]. *De novo* lncRNA transcripts contained 13,058 unique exons from 5,612 non-overlapping exonic regions. Protein-coding mRNAs were *de novo* annotated in preference to using the public annotations to avoid technical bias when comparing lncRNAs to mRNAs and to assess the quality of our annotation (Additional file 1: Figure S2). The *de novo* granulocyte mRNA annotation comprised 10,092 mRNA loci (Additional file 5) expressing 132,864 transcripts (Additional file 6) with a mean of 13.2 transcripts per locus, consistent with previous observations [62]. We assigned *de novo* annotated lncRNAs into three position-based classes relative to the nearest protein-coding gene (Fig. 1c). The majority of lncRNA loci (42 % comprising 659 loci) are intergenic, while 33 % (530 loci) are antisense and 25 % (402) are bidirectional. Figure 1d shows an example of a *de novo* annotated antisense lncRNA locus (green lines) absent from public databases.

Identification of new lncRNA loci and isoforms

We compared our granulocyte *de novo* lncRNA annotation to the most commonly used public annotations: GENCODE-v19 (23,898 lncRNA transcripts) [58], RefSeq (8,236 lncRNA transcripts) [59], and Cabili *et al.* (21,630 lncRNA transcripts) [14] and found that 46 % (736) of granulocyte *de novo* lncRNA loci were not present in public annotations, while 54 % (855) had a full or partial sense overlap with a publicly annotated lncRNA. Exon comparison with the three public annotations showed that we identified 5,694 new unique exons from 2,986 non-overlapping exonic regions. This shows that granulocytes have a specific lncRNA landscape that needs to be defined prior to granulocyte transcriptome analysis. To further assess the novelty of the annotated granulocyte *de novo* lncRNA loci we examined the MiTranscriptome lncRNA annotation based on 7,256 RNA-seq libraries from different human tissues, tumors, and cell lines [29]. Together, this shows that while 83 % of the lncRNA loci identified in this study can be found in one of the four above lncRNA annotations, 268 (17 %) are not found (Fig. 1e). To test the reliability of our granulocyte *de novo* lncRNA annotation we first determined that over 80 % of transcripts were supported by at least one exonic overlap with a spliced EST (human ESTs, UCSC table browser) (Additional file 1: Figure S3A). Second, the MiTranscriptome lncRNA annotation [29] provided an additional validation as 78 % of our granulocyte *de novo* annotated lncRNAs were supported by an exonic overlap with a spliced MiTranscriptome lncRNA (Additional file 1: Figure S3B) with a median of 51 % exonic coverage of granulocyte *de novo* lncRNAs by MiTranscriptome lncRNAs (Additional file 1: Figure S3C). Public lncRNAs annotations had less overlap

with our annotation (Additional file 1: Figure S3B) and showed poorer exonic coverage (Additional file 1: Figure S3C) and thus provided support for fewer of our granulocyte *de novo* lncRNA transcripts. In contrast, *de novo* mRNAs were well covered by public mRNA annotations and MiTranscriptome (Additional file 1: Figure S3B, D), indicating that the poor lncRNA coverage may arise from incomplete annotation of this transcript type in public annotations. Last, we used exon-spanning RT-PCR to test granulocyte *de novo* annotated lncRNA splice junctions (Additional file 2E). We confirmed 42 out of 46 tested junctions from 22 granulocyte lncRNA loci. We also cloned lncRNA transcripts from 18 granulocyte *de novo* lncRNA loci not present in public annotations, to confirm their full-length exon structure, continuity, and chromosome position (Additional file 1: Figures 1F, S4-S8 and Additional file 2F). Cloned sequences were deposited in GENBANK (Additional file 2G). In summary, we created a reliable lncRNA transcriptome annotation in healthy human granulocytes that identifies 1,591 lncRNA loci of which 17 % had not previously been described. Furthermore, we demonstrate that granulocyte *de novo* lncRNAs in contrast to mRNAs are incompletely represented in public annotations.

Non-mRNA-like features that may confound lncRNA annotation

As a basis to investigate why our granulocyte *de novo* annotation identified novel lncRNAs we classified them (Fig. 2a) according to existing public annotations (PA) as new lncRNA loci formed by 'not in PA' transcripts, or, as 'known lncRNA loci' formed by transcripts sharing all exons (PA transcripts) or sharing at least one exon (isoform not in PA, see example in Fig. 2b representing a novel isoform inside a publicly annotated lncRNA locus). The distribution was uniform with 32 % (2,003) 'PA transcripts', 37 % (2,235) 'isoform not in PA' and 31 % (1,921) 'not in PA transcripts'. We examined these three lncRNA classes for four known lncRNA features (tissue-specific expression, low expression level, PolyA+ enrichment, and splicing efficiency), which could reduce their identification in RNA-seq data compared to mRNAs.

To examine tissue-specificity we used publicly available RNA-seq data from 34 human cell types (ENCODE project (https://www.encodeproject.org), Illumina Human Body Map Project (http://www.ebi.ac.uk/gxa/experiments/E-MTAB-513) (Additional file 2H). These data were aligned as in Fig. 1b and expression levels calculated for *de novo* annotated granulocyte transcripts. A transcript was considered granulocyte-specific if its expression in granulocytes was at least three-fold higher than in all other cell types. We found granulocyte-specific expression of 32.5 % (1,927) *de novo* annotated lncRNA transcripts and 4 % of *de novo* annotated mRNA transcripts (Fig. 2c, Additional file 1: Figure S9A). This trend was also observed for granulocyte-specific expression over the whole locus, indicating it is not an artifact of the greater number of mRNA isoforms in the *de novo* annotation (Additional file 1: Figures S9B and S10). The same analysis performed for GENCODE-v19 transcripts that are annotated from multiple sample types shows a decreased percentage of lncRNAs (9.0 %) and mRNAs (1.5 %) identified as granulocyte-specific, but a similarly large difference (six-fold) between the two transcript types (Additional file 1: Figure S9C). Analysis of tissue-specific expression performed separately for the three lncRNA transcript classes described above, shows that 'in PA' lncRNAs were more similar to GENCODE-v19 transcripts being depleted for granulocyte-specific transcripts compared to the bulk population (dashed green line, Fig. 2c), while 'not in PA' and 'isoform not in PA' transcripts showed equal or increased granulocyte-specificity.

Expression level is another feature strongly differentiating lncRNAs and mRNAs. We calculated RPKMs of granulocyte *de novo* lncRNA and mRNA transcripts in the PolyA+ data used for the *de novo* annotation, which showed that lncRNA transcripts are 10-fold less abundant than mRNAs (0.65/6.14, respectively; Fig. 2d). We noted that lncRNA/mRNA expression difference was slightly reduced (seven-fold median difference) when analyzing ribosomal-depleted datasets, indicating lncRNA under-representation in PolyA+ RNA (Additional file 1: Figure S11A). Comparing the three lncRNA transcript classes showed that 'in PA' transcripts display highest expression and 'not in PA' have lowest expression among the three classes in both PolyA+ (see inset, Fig. 2d) and ribosomal-depleted (Additional file 1: Figure S11F) data.

The third feature that may influence lncRNA identification is their reduced polyadenylation efficiency, as this would lower abundance in the PolyA+ fraction usually used for transcript identification. Given our above observation of poorer lncRNA representation in PolyA+ versus ribosome-depleted datasets, we compared transcript abundance in these granulocyte datasets to estimate the enrichment of lncRNAs and mRNAs in the PolyA+ fraction (Fig. 2e). While mRNAs showed a median 2.6-fold enrichment, lncRNAs showed a significantly lower median 1.6-fold enrichment (dashed green line, Fig. 2e). We tested if this difference was influenced by low lncRNA expression levels by splitting transcripts into expression bins (Additional file 1: Figure S12A). This showed that independently of absolute expression levels, lncRNAs show significantly lower PolyA+ enrichment compared to mRNAs. Comparing the three lncRNA transcript classes demonstrated that 'not in PA' and

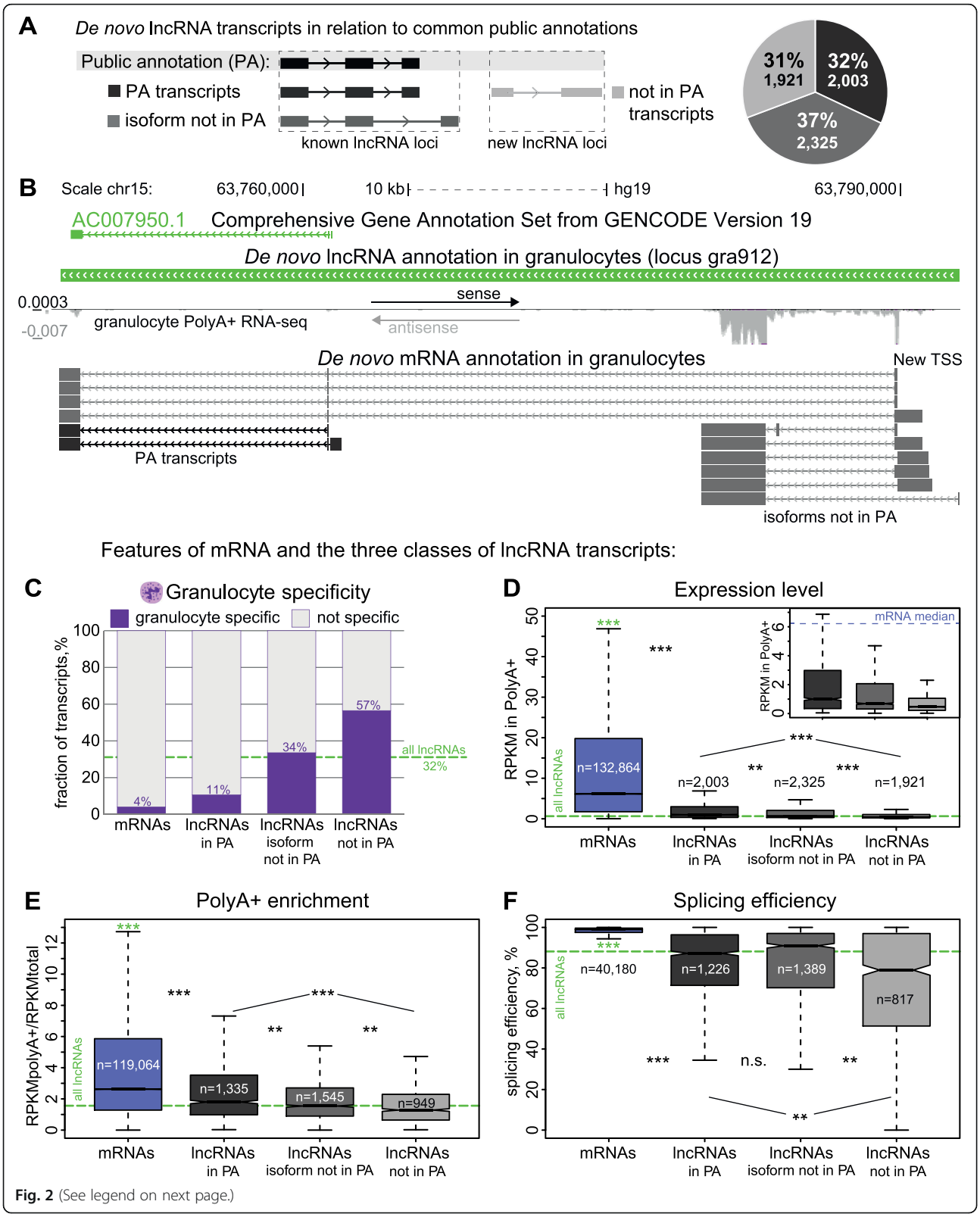


Fig. 2 (See legend on next page.)

(See figure on previous page.)
Fig. 2 LncRNAs not in public annotations show less mRNA-like features. **a** Distribution of 6,249 granulocyte *de novo* annotated lncRNA transcripts according to coverage by three commonly used public annotations (PA): RefSeq, GENCODE-v19, Cabili [14, 58, 59]. Known lncRNA loci contain two transcript types: 'PA transcripts' that show full exonic overlap with an annotated lncRNA transcript (32 %, 2,003 transcripts, dark gray), or 'isoform not in PA' transcripts, that can share exons but contain one or more additional exons not present in public annotation (37 %, 2,331 transcripts medium gray). New lncRNA loci: contain 1,921 'not in PA' transcripts (31 % of lncRNA transcripts identified in granulocytes, light gray). **b** An example of a publicly-annotated lncRNA locus (GENCODE-v19 AC007950.1) that contains additional upstream exons not in PA, from sample D2-2_pa_100ss (Additional file 2B). The annotation identifies locus *gra912* (thick green bar). The annotated lncRNA isoforms of locus *gra912* with alternative transcription start sites (TSS) are shown underneath as gray lines (the shorter PA transcript is shown in black for comparison). **c** Granulocyte-specificity analysis. Bar plot shows the percentage of granulocyte-specific (purple) and not-specific (light gray) transcripts *de novo* annotated in granulocytes. Each bar shows the percentage of granulocyte-specific transcripts for each transcript class while the dashed green line shows the percentage for all lncRNAs together. **d** Average expression level (RPKM) in granulocyte PolyA+ RNA-seq samples used for annotation. The median values are: all mRNA transcripts (blue): 6.14, all lncRNA transcripts (green dashed line): 0.65, lncRNA transcripts 'in PA' (dark gray): 1.00, lncRNA transcripts 'isoform not in PA' (medium gray): 0.68, lncRNA transcripts 'not in PA' (light gray): 0.47. **e** PolyA+ enrichment of *de novo* granulocyte annotated transcripts calculated as a ratio between abundance of a transcript in PolyA+ RNA and abundance in total ribosome-depleted RNA. Transcript abundance (RPKM) is averaged among all PolyA+ RNA-seq samples or all total RNA-Ribosomal depleted RNA-seq samples. Transcripts not detected in total RNA-seq data (average RPKM <0.2) were not analyzed. The median values are: all mRNA transcripts (blue): 2.62, all lncRNA transcripts (dashed green line): 1.56, lncRNA transcripts 'in PA' (dark gray): 1.80, lncRNA transcripts 'isoform not in PA' (medium gray): 1.54, lncRNA transcripts 'not in PA' (light gray): 1.29. **f** Splicing efficiency of *de novo* granulocyte annotated transcripts. Only transcripts with average RPKM >0.2 in 21 ribosomal-depleted RNA-seq samples were analyzed and the efficiency of the most efficiently-spliced site in each transcript is plotted. The median values are: all mRNA transcripts: 99.02 %, all lncRNA transcripts: 88.13 %, lncRNA transcripts 'in PA': 87.18 %, lncRNA transcripts 'isoform not in PA': 90.90 %, lncRNA transcripts 'not in PA': 77.97 %. Remarks to boxplots **d**, **e**, and **f**: the box plot displays the full population but *P* values are calculated using Mann–Whitney U test on equalized population sizes. *$0.001 < P < 10^{-5}$, **$10^{-5} < P < 10^{-10}$, ***$P < 10^{-16}$. Green asterisks indicate the significance of the difference between mRNAs and all lncRNAs (only the median level is plotted as a dashed green line). Outliers are not displayed

'isoform not in PA' transcripts showed significantly lower PolyA+ enrichment than 'in PA' transcripts (Fig. 2e).

Inefficient splicing is a fourth feature likely to reduce full-length lncRNA transcripts in the PolyA+ fraction. We used granulocyte ribosomal-depleted RNA-seq to calculate the splicing efficiency of every splice site in lncRNA and mRNA transcripts and defined transcript splicing efficiency as that of its most efficiently processed splice site (Additional file 1: Figure S13A, B). This shows that splicing is significantly less efficient for lncRNAs compared to mRNAs with a median splicing efficiency of 88.13 % (dashed green line, Fig. 2f) and 99.02 %, respectively. This splicing efficiency difference is independent of expression level and also persists at the locus level, that is, independently of the transcripts number per locus (Additional file 1: Figures S12B and S13C). The inefficient splicing of lncRNAs is supported by the experimental validation of lncRNA spliced products described above, which identified abundant unspliced isoforms together with spliced isoforms (see examples in Additional file 1: Figures S5B, S5C, S7A, and S13B, E). Comparing the three lncRNA transcript classes showed that 'not in PA' transcripts have lower splicing efficiency than the bulk population analysis (Fig. 2f). The similar splicing efficiency in classes 'isoform not in PA' and 'in PA' arises from transcripts sharing some splice sites. The reduced splicing of lncRNAs 'not in PA' was confirmed by analysis on the locus level (Additional file 1: Figure S13D).

In addition to these four RNA biology features, we examined four genomic features. This showed that compared to mRNAs, lncRNAs transcripts have significantly fewer exons, their transcription starts are less CG-rich but more repeat-rich, and their exons contain more repeats (Additional file 1: Figures S11B-E and S12C). With the exception of the median exon number, these features were more extreme in 'not in PA' and 'isoform not in PA' lncRNAs than in the class of 'in PA' lncRNAs. Together this shows that new granulocyte lncRNAs identified in this study have less mRNA-like features that further distinguish them from mRNAs compared to the bulk lncRNA population. To support this claim we performed the same analysis for MiTranscriptome mRNAs and lncRNAs [29], which also shows that lncRNAs not in public annotations have less mRNA-like features (Additional file 1: Figures S14 and S15). Thus we show that features such as tight tissue-specificity and low expression, reduced enrichment in PolyA+ selected RNA and reduced splicing efficiency, not only distinguish lncRNAs from mRNAs, but by reducing their representation in the analyzed transcriptome make their identification more challenging.

LncRNAs are reproducibly expressed within one donor but vary between donors

We next investigated reproducibility of lncRNA expression in healthy individuals to assess if this could also influence the lncRNA discovery. To estimate expression reproducibility within or between donors, we examined expression in granulocytes from seven donors sampled at three time points spaced by at least 1 month (Fig. 3a, Additional file 2A). These 21 samples were subject to ribosome-depleted RNA-seq (Additional file 2B) aligned with STAR and expression levels were determined of all

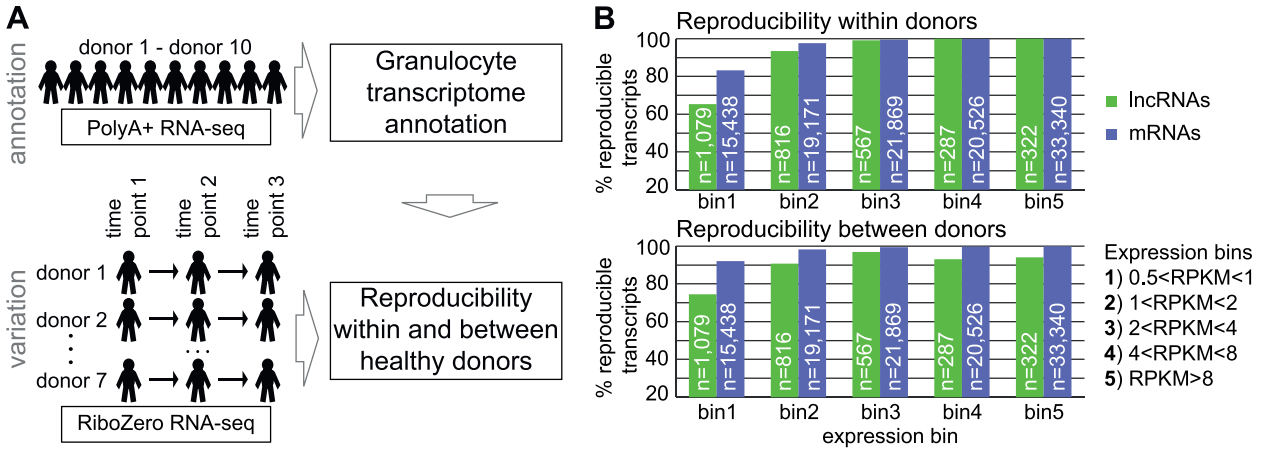


Fig. 3 Reproducibility of *de novo* lncRNA and mRNA expression. **a** Study overview. Top: the granulocyte *de novo* transcriptome annotation was generated from 10 healthy donors. Bottom: seven donors were sampled at three time points spaced by ≥1 month (Additional file 2A) and RNA was sequenced to assess intra-individual (using three time points from one donor) and inter-individual (using samples from seven different donors) expression reproducibility. **b** Granulocyte intra-individual (top) and inter-individual (bottom) expression reproducibility for *de novo* annotated lncRNAs (green) and mRNAs (blue). Transcripts detectable (RPKM >0.2) at each of three time points or not detected (RPKM <0.2) at any time point in all seven donors show intra-individual reproducibility. Transcripts detectable in each of seven donors (average RPKM of three replicates >0.2) show inter-individual reproducibility. Five expression bins were used: (1) 0.5 < RPKM ≤1; (2) 1 < RPKM ≤2; (3) 2 < RPKM ≤4; (4) 4 < RPKM ≤8; and (5) RPKM >8 (n = transcript number per bin). Chromosomes X, Y were discarded

de novo annotated lncRNAs and mRNAs. We first tested if annotated transcripts were reproducibly expressed within one donor, that is, the three time points for each donor should show consistent lncRNA expression (RPKM >0.2) or absence (RPKM ≤0.2) of expression (Fig. 3b top). This analysis was performed separately for transcripts with different expression levels. Expression levels for each donor were calculated by averaging RPKMs from the three time points and a transcript was placed into a bin according to its maximal expression level among the seven donors. We counted the number of reproducibly-expressed transcripts and found that lncRNAs are less reproducible in bins 1 and 2, but above RPKM >2 almost all *de novo* annotated lncRNAs and mRNAs (99–100 %) were reproducibly detected within one donor. In contrast, comparing expression between the seven donors showed consistent lower expression reproducibility of lncRNAs compared to mRNAs (Fig. 3b, bottom). In the three highest expression bins, mRNAs showed 100 % reproducibility while lncRNA transcripts only reached 95 %. In summary, this qualitative analysis shows that, above an expression threshold (RPKM >2), lncRNAs are as reproducibly expressed as mRNAs within replicates from one donor. However, lncRNAs show less reproducible expression than mRNAs between different donors.

LncRNAs show high expression variability between donors

We quantitated the extent of expression variability between the seven donors by calculating the standard deviation of granulocyte *de novo* lncRNA and mRNA expression (Fig. 4a, b). As RPKM is a parametric value and ranges from 0.2 (the used expression cutoff) to several thousand, we normalized standard deviation of expression for each gene between donors by the mean of expression among the seven donors (thus calculating the value also known as the coefficient of variation). We performed this analysis calculating variability of expression for each transcript separately (Fig. 4a), and expression of the whole locus disregarding identified exon structures (Fig. 4b, Additional file 1: Figure S10). Both analyses showed that lncRNAs display significantly higher variability than mRNAs (P $<10^{-16}$). LncRNA and mRNA expression variability between donors (inter-individual) was significantly higher than between the replicates from one donor (intra-individual). In addition, both inter- and intra-individual expression variability of lncRNAs exceeded that of mRNAs (Additional file 1: Figure S16). The high inter-individual variability of lncRNA expression allowed unsupervised clustering of the three time point samples according to each of the seven donors (Fig. 4c), that validates their use as replicates.

LncRNA expression is generally lower than that of mRNAs (Fig. 2d, Additional file 1: Figure S11A), which could bias the expression variability analysis, as lower expression will correlate with higher normalized standard deviation values. We controlled for this by distributing transcripts and loci into expression bins (Additional file 1: Figure S17). This showed that while variability anti-correlates with expression level for both lncRNAs and mRNAs, lncRNAs analyzed at the transcript or loci level show consistently more expression variability than

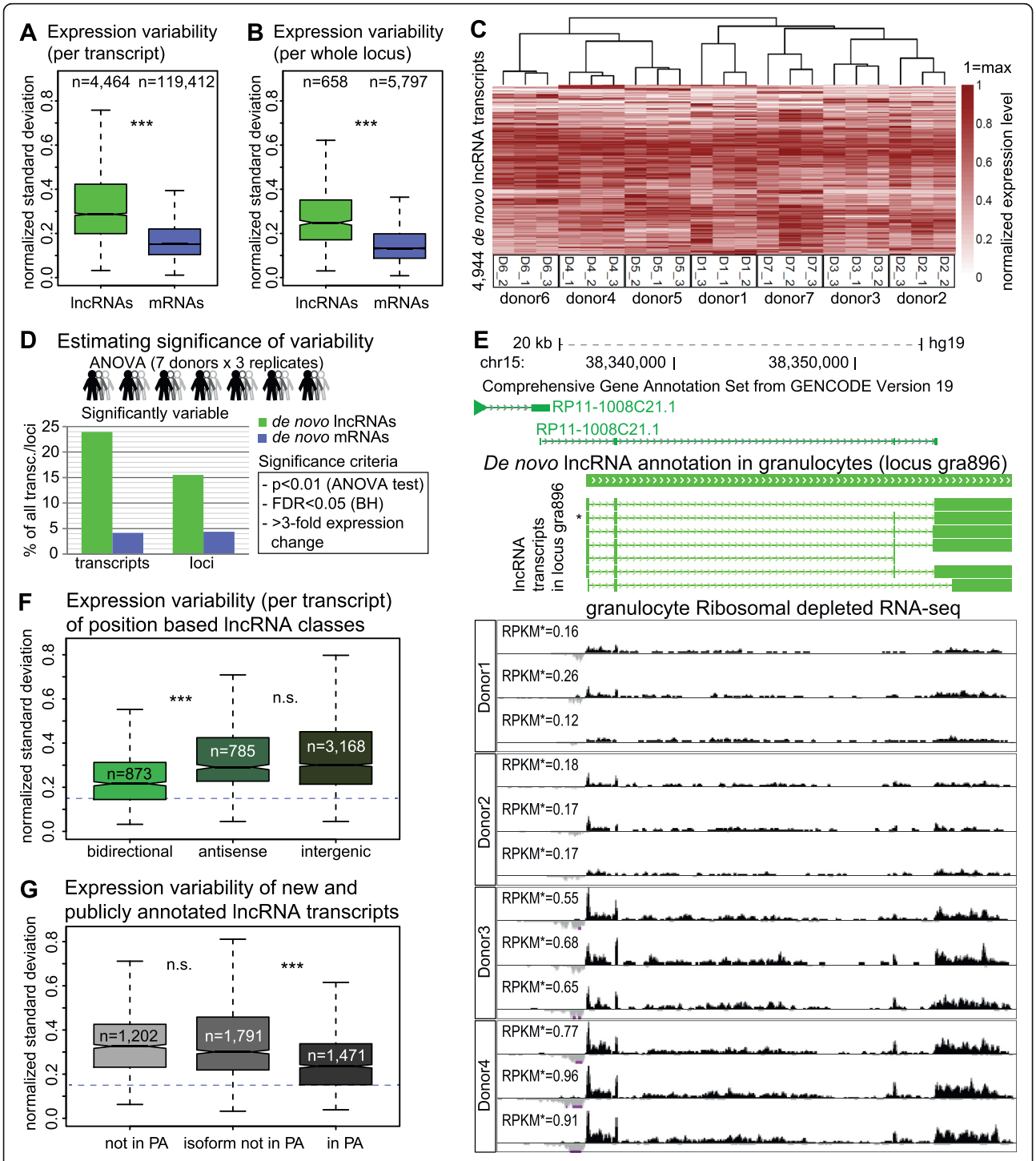


Fig. 4 (See legend on next page.)

(See figure on previous page.)
Fig. 4 LncRNAs are more variably expressed than mRNAs. **a**, **b** Genome wide inter-individual variability (normalized standard deviation between expression of each transcript/locus in granulocytes from seven donors) of *de novo* granulocyte lncRNA (green) and mRNA (blue) transcripts (**a**) and loci (**b**). Donor expression level is averaged from three replicates (***P <10^{-16}). Median values: lncRNA transcripts: 0.29, mRNA transcripts: 0.15, lncRNA loci: 0.26, mRNA loci: 0.15. **c** LncRNA inter-individual expression variability allows correct clustering (normalized level among seven donors) of three time points per donor. Maximum transcript expression among all 21 samples is set to 1 (red), minimum is 0 (white). Clustering was performed using *pheatmap* function in R (clustering_distance_rows = 'euclidean', clustering_distance_cols = 'correlation'). Only transcripts detected (RPKM >0.2) in at least one of the total RNA-seq samples were analyzed. Chromosomes X, Y were discarded. **d** Significance of granulocyte *de novo* lncRNA and mRNA expression variability in seven donors assessed by ANOVA test (the three time points are used as replicates). Bars show the percentage of significantly variable lncRNA (green) and mRNA (blue) transcripts (left) and loci (right). Criteria for calling a transcript/locus 'significantly variable': ANOVA test P value <0.01, FDR (Benjamini-Hochberg correction) <0.05, fold change between highest and lowest expression in seven donors >3. Only transcripts/loci with RPKM >0.2 in at least one donor are included. Chromosomes X and Y were discarded from the analysis. Total number analyzed: lncRNA transcripts 4,464, mRNA transcripts 119,412, lncRNA loci 658, mRNA loci 5,797. **e** Example of a significantly variable transcript from lncRNA locus *gra896*. Top: an alternative *gra896* TSS overlaps the publicly-annotated lncRNA RP11-1008C21.1 locus. Underneath: normalized total RNA-seq signal for three replicates of four donors scaling from -0.001 (reverse strand, light gray) to 0.004 (forward strand, black). Calculated expression level of the annotated lncRNA transcript marked with * is shown for each RNA-seq track. Significance result for this transcript among seven donors: ANOVA test $P = 10^{-7}$, FDR (Benjamini-Hochberg) = 10^{-6}, expression fold change = 5.2). **f** Bidirectional lncRNA transcripts show reduced expression variability. Boxplots show inter-individual variability of lncRNA transcripts split according to their position relative to protein-coding genes as in Fig. 1c. Median normalized standard deviation values: bidirectional: 0.22, antisense: 0.29, intergenic: 0.30. Dashed blue line indicates median expression variability of all *de novo* mRNA transcripts. **g** Inter-individual expression variability is lower for known 'in PA' lncRNA transcripts compared to those newly annotated in granulocytes ('not in PA' and 'isoform not in PA'). Median normalized standard deviation values: 'not in PA': 0.33, 'isoform not in PA': 0.30, 'in PA': 0.24. Dashed blue line indicates median expression variability of all *de novo* mRNA transcripts. Remarks to boxplots **a**, **b**, **c**, **g**: Transcripts/loci not expressed (RPKM <0.2) in any of seven donors (total RNA-seq data) and data from chromosomes X, Y were discarded and outliers are not displayed. The box plot displays the full population but P value is calculated using Mann–Whitney U test on equalized sample size. n.s. not significant, ***P <10^{-16}

mRNAs, independent of absolute expression level. We additionally plotted expression variability against mean expression between all donors for lncRNA and mRNA transcripts and loci (Additional file 1: Figure S18A, B). This showed that the trend lines of the anti-correlation between variability and expression level are clearly distinct for lncRNAs and mRNAs at both transcript and loci level, with lncRNAs displaying higher variability. Thus, high natural expression variability is not an artifact of the general low expression of lncRNAs. To identify the number of lncRNA and mRNA transcripts and loci significantly variable between donors we applied an ANOVA test (*aov* function in R [63]) to expression values in all the 21 (that is, seven donors sampled three times) ribosomal depleted RNA-seq samples. We find that 23.9 % (1,069) of lncRNA transcripts but only 4.2 % of mRNA transcripts are differentially expressed between the seven donors (transcripts RPKM >0.2, Fig. 4d). This trend persisted when applying an ANOVA test to expression over whole loci (Fig. 4d, 15.5 % and 4.4 % for lncRNA and mRNA loci, respectively). Importantly, this difference between lncRNAs and mRNAs was persistent when analyzing different expression bins (Additional file 1: Figure S19A). Figure 4e shows an example of a significantly variable lncRNA expressed from chromosome 15. Among the four displayed tracks donors 3 and 4 show higher expression, consistent among three replicates, while donors 1 and 2 show low expression consistent among replicates. Since 25 % of *de novo* annotated lncRNAs are bidirectional and likely share a promoter with an mRNA (Fig. 1c), we examined if this class resemble mRNAs in having less expression variability. Figure 4f shows that bidirectional lncRNA transcripts more closely resembled mRNAs and were significantly less variable than antisense or intergenic lncRNAs and this trend was also observed in all expression bins and over the whole locus (Additional file 1: Figure S20A-C).

Publicly annotated lncRNAs show less expression variability

To further confirm high lncRNA expression variability and to investigate its impact on lncRNA identification, we analyzed expression variability of publicly annotated (Additional file 1: Figure S21A, B) and of MiTranscriptome (Additional file 1: Figure S22A, B) lncRNAs and mRNAs in our granulocyte RNA-seq data. All annotations confirmed high lncRNA expression variability compared to mRNAs. However, the extent of the lncRNA/mRNA difference was reduced when analyzing public annotations compared to the MiTranscriptome annotation and our *de novo* granulocyte annotation, which both identified numerous novel lncRNAs. We then estimated expression variability separately for the three lncRNA classes described in Fig. 2a, and found that transcript types 'not in PA' and 'isoform not in PA' showed significantly higher variability between the seven donors, compared to 'in PA' transcripts (Fig. 4g) and this trend was observed in all expression bins (Additional file 1: Figure S23A) and also when analyzing expression over whole locus for 'new' and 'known' lncRNA loci (Additional file 1: Figure S23B, C). To test this further, we analyzed expression variability of MiTranscriptome lncRNAs classified according to their presence in public annotations (as described in Additional file 1: Figure S14D). This showed that

'not in PA' and 'isoform not in PA' MiTranscriptome lncRNAs displayed higher expression variability (Additional file 1: Figure S22C), consistent with results for the *de novo* granulocyte lncRNA annotation. Together this supports our arguments above, that lncRNAs not in public annotations have less mRNA-like features.

A list of robustly or variably expressed lncRNAs in human primary granulocytes

Following the discovery of high intra- and inter-individual expression variability of lncRNAs we sought to generate a list of robustly expressed and variably expressed granulocyte lncRNAs as a resource. To generate the robustly expressed list we filtered 6,249 lncRNA transcripts in our annotation (that is, the set of transcripts that 'can be' expressed in granulocytes) to identify those detected (RPKM >0.2) in all replicate samples from seven donors. This gave a robustly expressed annotation of 2,490 transcripts from 393 lncRNA loci (Additional file 7A). We applied stricter criteria and required a higher level of expression (RPKM >1) in every sample to produce another annotation of 'well-expressed robust' lncRNAs in granulocytes with 817 transcripts from 115 lncRNA loci (Additional file 7B). A list of significantly variably expressed (defined as in Fig. 4d) lncRNAs with 1,069 transcripts from 214 lncRNA loci is provided in Additional file 8.

LncRNAs expression variability in lymphoblastoid cell lines (LCL)

To test our finding of high lncRNA expression variability in an independent cell type and with larger donor numbers, we analyzed GEUVADIS project data (http://www.geuvadis.org/web/geuvadis/rnaseq-project [50]) consisting of PolyA+ non-stranded RNA-seq of lymphoblastoid cell lines (LCL) sampled once from 462 healthy individuals of various ages from five population groups (Fig. 5a) (ENA http://www.ebi.ac.uk/ena/data/view/ERR188021-ERR188482). Since LCL are a different cell type to granulocytes, we created a *de novo* LCL annotation via our pipeline (Additional file 1: Figure S24A). From the list of 462 samples, we used RNA-seq data from 20 unrelated donors (2 female donors and 2 male donors from each population with a total of 522 (26.1 million reads/donor) million uniquely mapped reads and 177.8 million spliced reads) grouped into five pools (Additional file 2I). The resulting LCL lncRNA transcriptome consisted of 2,611 lncRNA loci (Additional file 9) formed by 8,560 lncRNA transcripts (Additional file 10) with a mean of 3.3 transcripts per locus (Fig. 5b). The lncRNA transcripts contained 17,009 unique exons from 9,379 non-overlapping regions. We also annotated 12,241 *de novo* mRNA loci formed by 124,799 transcripts, with a mean of 10.1 transcript per locus. The overlap of LCL and granulocyte *de novo* lncRNA transcriptomes comprised only 536 loci (21 %) whereas the *de novo* mRNA transcriptomes overlapped by 9,357 loci (76 %), which is consistent with lncRNA high tissue-specificity (Fig. 5c). The increase in lncRNA loci number from 1,591 in granulocytes, to 2,611 in LCL may reflect increased transcriptional activity of LCL compared to primary granulocytes or the two-fold increase in donor number used for annotation (see data below). Comparison of the LCL *de novo* lncRNA annotation to public annotations and MiTranscriptome showed that 2,316 (89 %) of LCL lncRNA loci are covered by the four lncRNA annotations while 295 (11 %) are not found (Fig. 5c). The LCL annotation quality was verified in a similar manner as for the granulocyte annotation (Additional file 1: Figure S24B-G). LncRNA classification by coverage from public annotations shows that 1,536 are known loci containing 3,363 (39 %) 'in PA' while 3,111 (36 %) are 'isoform not in PA' transcripts, and 1,075 are new loci formed by 2,086 (25 %) 'not in PA' transcripts (Additional file 1: Figure S25). Exon comparison showed that *de novo* lncRNA annotation in LCL contained 6,113 unique exons not present in public annotations from 4,150 non-overlapping exonic regions. Similar to granulocytes, LCL lncRNA transcripts not in public annotations show less mRNA-like features (Additional file 1: Figure S26).

We used the LCL *de novo* annotation to calculate the RPKM of every transcript and locus in all 462 donors. An ANOVA test could not be applied due to the absence of donor replicates, but inter-individual variability was calculated from the normalized standard deviation of expression between all donors. Comparing lncRNAs to mRNAs showed that lncRNAs are significantly more variable both when calculating expression of transcripts or over whole loci (Fig. 5d, e). We controlled for expression level by distributing transcripts or loci to expression bins as described above and found that except for bin1 transcripts, lncRNAs were significantly more variable in expression than mRNAs (Additional file 1: Figure S27). To complete the comparison with the granulocyte data, we found LCL bidirectional lncRNAs to be significantly less variable than intergenic lncRNAs in all expression bins (Additional file 1: Figure S28). In addition, LCL *de novo* lncRNAs not covered by public annotations ('not in PA' transcripts) show significantly more expression variation than publicly annotated transcripts (Fig. 5f, Additional file 1: Figure S29). This analysis of an independent cell type with an independent sample collection and processing method from a larger number of donors supports our finding of high inter-individual lncRNA expression variability.

LncRNA expression variability is increased in multiple human tissues

The above analysis shows high lncRNA expression variability relative to mRNAs in a primary human cell type

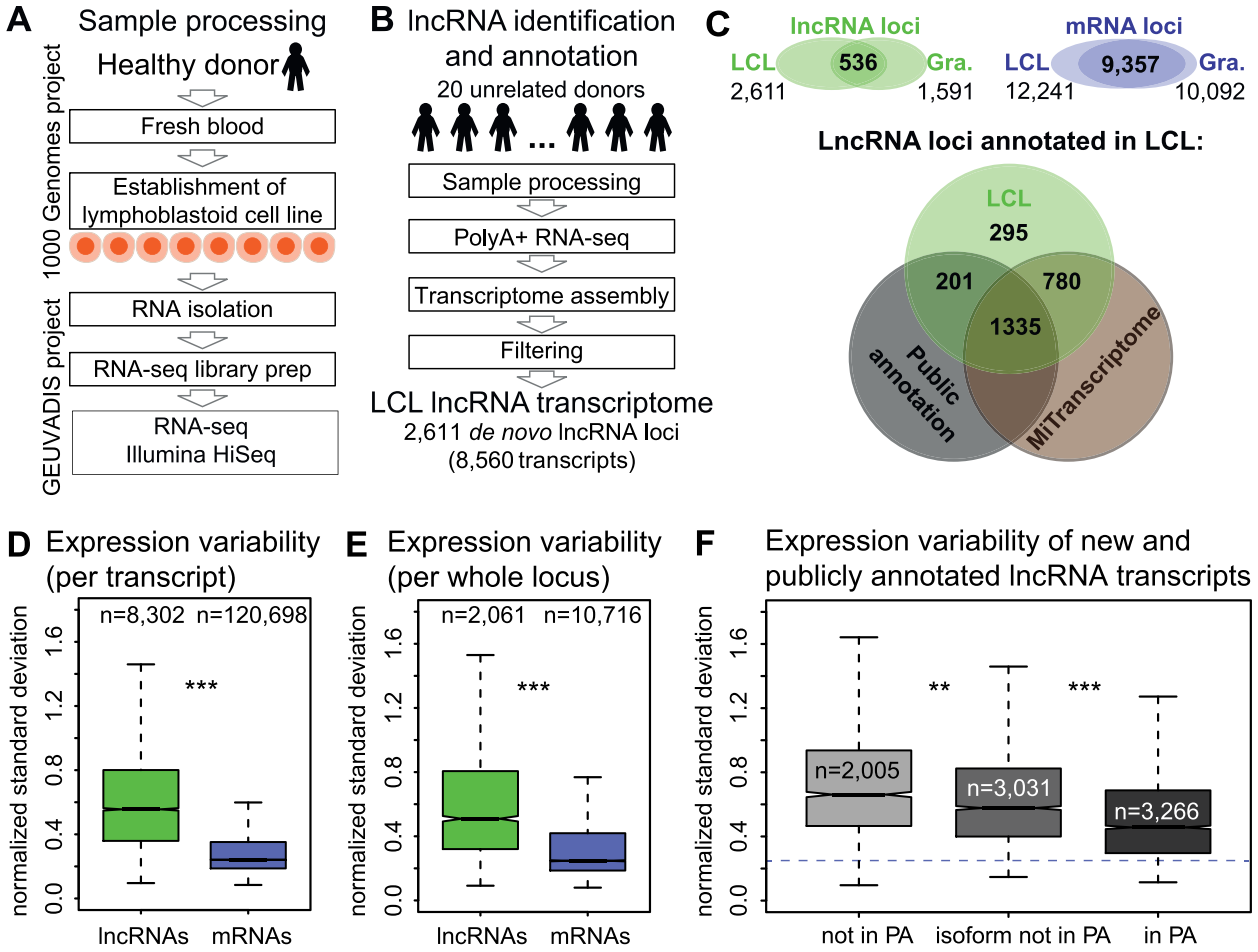


Fig. 5 GEUVADIS RNA-seq data confirm increased lncRNA expression variability. **a** Sample processing overview: 462 lymphoblastoid cell lines (LCL) established from healthy donors by EBV transformation were processed by the GEUVADIS RNA-seq Project [50]. **b** LncRNA identification overview. We picked 20 unrelated donors (total of 522 million uniquely mapped reads) from 462 donors and processed the raw RNA-seq data through the same pipeline used to annotate lncRNAs in granulocytes (Additional file 1: Figure S24). The resulting LCL lncRNA transcriptome contained 2,611 lncRNA loci formed by 8,560 lncRNA transcripts. **c** Top: overlap between LCL and granulocyte *de novo* transcriptome annotations created in the study. A total of 536 of 2,611 LCL lncRNA loci overlap granulocyte loci. A total of 9,357 of 12,241 LCL *de novo* mRNA loci overlap granulocyte loci. Bottom: overlap of *de novo* lncRNA annotation in LCL with commonly used public annotations (PA): RefSeq, GENCODE-v19, and Cabili [14, 58, 59], and the MiTranscriptome annotation [29] identifies 295 new lncRNA loci. Of these, only 18 loci overlap the *de novo* lncRNA granulocyte annotation. **d, e** LncRNAs show higher expression variability than mRNAs in LCL. The boxplots show inter-individual variability of LCL lncRNA (green) and mRNA (blue) transcripts (**d**) and loci (**e**). Inter-individual variability is estimated by calculating standard deviation between expression of each transcript/locus in 462 donors normalized to the mean expression. Both transcripts and loci variability is significantly (***$P < 10^{-16}$) different between lncRNAs and mRNA. Median values: lncRNA transcripts: 0.56, mRNA transcripts: 0.24, lncRNA loci: 0.51, mRNA loci: 0.25. **f** Inter-individual expression variability is higher for newly annotated lncRNA transcripts in LCL. Boxplot shows inter-individual expression variability of LCL lncRNA transcripts split according to coverage by public annotations (PA), which is higher for 'not in PA' and 'isoform not in PA' lncRNA transcripts compared to 'in PA'. Median normalized standard deviation values: not in PA: 0.66, isoform not in PA: 0.58, in PA: 0.46. Blue dashed line indicates median expression variability of all *de novo* mRNA transcripts in (**d**). Remarks to boxplots **d**, **e**, **f**: transcripts or loci not expressed (RPKM <0.2) in any of the 462 donors were discarded. The box plot displays the full population but *P* value is calculated using Mann–Whitney U test on equalized sample size (**$P < 10^{-10}$, ***$P < 10^{-16}$). Data from chromosomes X, Y were discarded and outliers are not displayed

(granulocytes) as well as in cell lines immortalized from lymphocytes. To test if this is a general phenomenon in human tissues, we obtained access to the GTEx project RNA-seq data [55, 64]. We downloaded RNA-seq data for nine human tissues: LCL, adipose, artery, cerebellum, heart, lung, muscle, nerve, and thyroid from 20 individuals per tissue (Additional file 2J). We used the MiTranscriptome transcript annotation derived from multiple tissue types [29], to calculate lncRNA and mRNA expression in GTEx samples and then estimated expression variability as described above using 20 donors per tissue (Fig. 6). This shows that lncRNAs are significantly more variable than mRNAs in all the analyzed tissues. We performed a binned expression control as described above

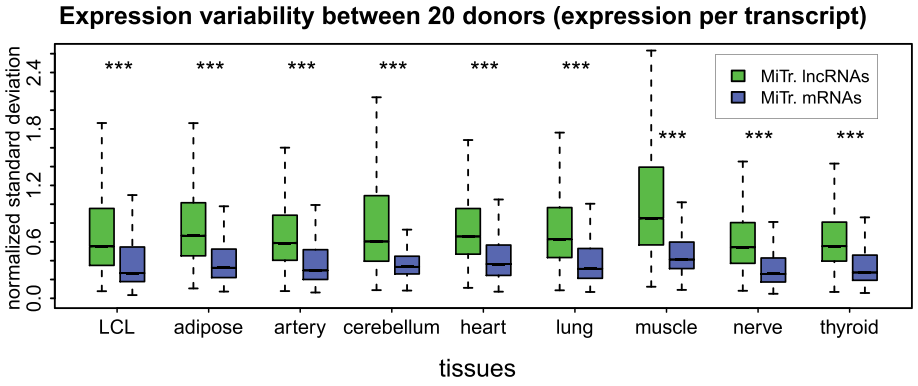


Fig. 6 GTEx RNA-seq data show increased lncRNA expression variability in multiple human tissues. Inter-individual variability of multi-exonic MiTranscriptome lncRNA (green) and mRNA (blue) transcripts analyzed in GTEx RNA-seq dataset [64]. Twenty donors per tissue are analyzed (Additional file 2J). Standard deviation is normalized to the mean expression among all 20 analyzed donors for each tissue. Only transcripts expressed in the given tissue in at least one donor (RPKM >0.2) are displayed. Number of transcripts in each box from left to right: LCL (lncRNAs: 28,571; mRNAs: 102,449), adipose (lncRNAs: 38,060; mRNAs: 113,688), artery (lncRNAs: 29,965; mRNAs: 108,082), cerebellum (lncRNAs: 44,912; mRNAs: 115,039), heart (lncRNAs: 32,827; mRNAs: 111,564), lung (lncRNAs: 39,909; mRNAs: 117,901), muscle (lncRNAs: 31,507; mRNAs: 106,099), nerve (lncRNAs: 39,167; mRNAs: 115,038), and thyroid (lncRNAs: 40,099; mRNAs: 116,206). Median expression values from left to right: LCL: 0.55, 0.27, adipose: 0.66, 0.32, artery: 0.59, 0.30, cerebellum: 0.60, 0.33, heart: 0.66, 0.36, lung: 0.63, 0.31, muscle: 0.85, 0.41, nerve: 0.54, 0.26, and thyroid: 0.56, 0.27. The box plots display the full population but *P* values are calculated using Mann–Whitney U test on equalized sample size (***$P < 10^{-16}$). Data from chromosomes X, Y were discarded and outliers are not displayed

and found that, apart from bin 1 that showed inconsistent results in two tissues, all nine tissues showed a significant increase of lncRNA expression variability independent of expression level (Additional file 1: Figure S30). Together with the above data on granulocytes and LCLs, this demonstration of increased lncRNA expression variability relative to mRNAs in multiple human tissues indicates that it is a general phenomenon inherent to all human tissues and a new lncRNA feature.

Increased expression variability affects lncRNA identification

We demonstrated above the high lncRNA inter-individual expression variability in diverse human tissues (Figs. 4a, b, 5d, e and 6) as well as the increased expression variability of novel compared to known lncRNAs (Figs. 4g and 5f, Additional file 1: Figure S22C). We asked if this expression variability feature could influence lncRNA identification. Figure 7a shows an example of a highly variably expressed *de novo* annotated LCL lncRNA not covered by public annotations (but identified with different exon models in [29]) that is well expressed (RPKM >1) in one out of 462 donors in the GEUVADIS project dataset, expressed at a low level (RPKM >0.2) in 93 donors and not detected (RPKM <0.2) in the remaining 368 donors. It is likely that such a lncRNA has a low chance of discovery when analyzing few individuals. We hypothesized that adding more individuals to the identification pipeline may increase the chance of identifying highly variably expressed lncRNAs. At the same time, given the relatively low inter-individual expression variability of mRNAs, we would expect to identify a relatively constant number of mRNA loci.

We tested this by *de novo* annotating lncRNAs and mRNAs from a variable number of individuals. We picked 120 GEUVADIS LCL donors (Fig. 7b, Additional file 11A), unified the data by sampling 25 million paired-end reads from each donor and created 30 pools, each with four donors (two male and two female donors) with a total of 100 (25 × 4) million reads. From the 30 pools we created 30 LCL *de novo* transcriptome assemblies using Cufflinks. We randomly picked 1, 2, 3, 4, 5, 6, 8, 10, 15, 20, 25 (using three replicates of random picking), and 30 assemblies, which corresponded to 4, 8, 12, 16, 20, 24, 32, 40, 60, 80, 100, and 120 donors, respectively, and applied Cuffmerge and the *de novo* transcriptome annotation pipeline to each group of assemblies (Additional file 1: Figure S31A, B and Additional file 11B). Only one pool (100 million reads) was fed at a time into the assembly pipeline, thus the sensitivity of Cufflinks was unchanged. In addition, assemblies but not reads were merged at this stage. Different number of assemblies fed into our annotation pipeline produced multiple lncRNA/mRNA annotations with different numbers of loci and transcripts. We plotted the number of mRNA and lncRNA loci (averaged from the three replicates described above) versus the number of donors used (Fig. 7c, Additional file 11C). This showed that while lncRNA loci number (green lines) grew three-fold with increasing donor number, from 1,382 loci obtained from four donors to 4,166 obtained from 120 donors,

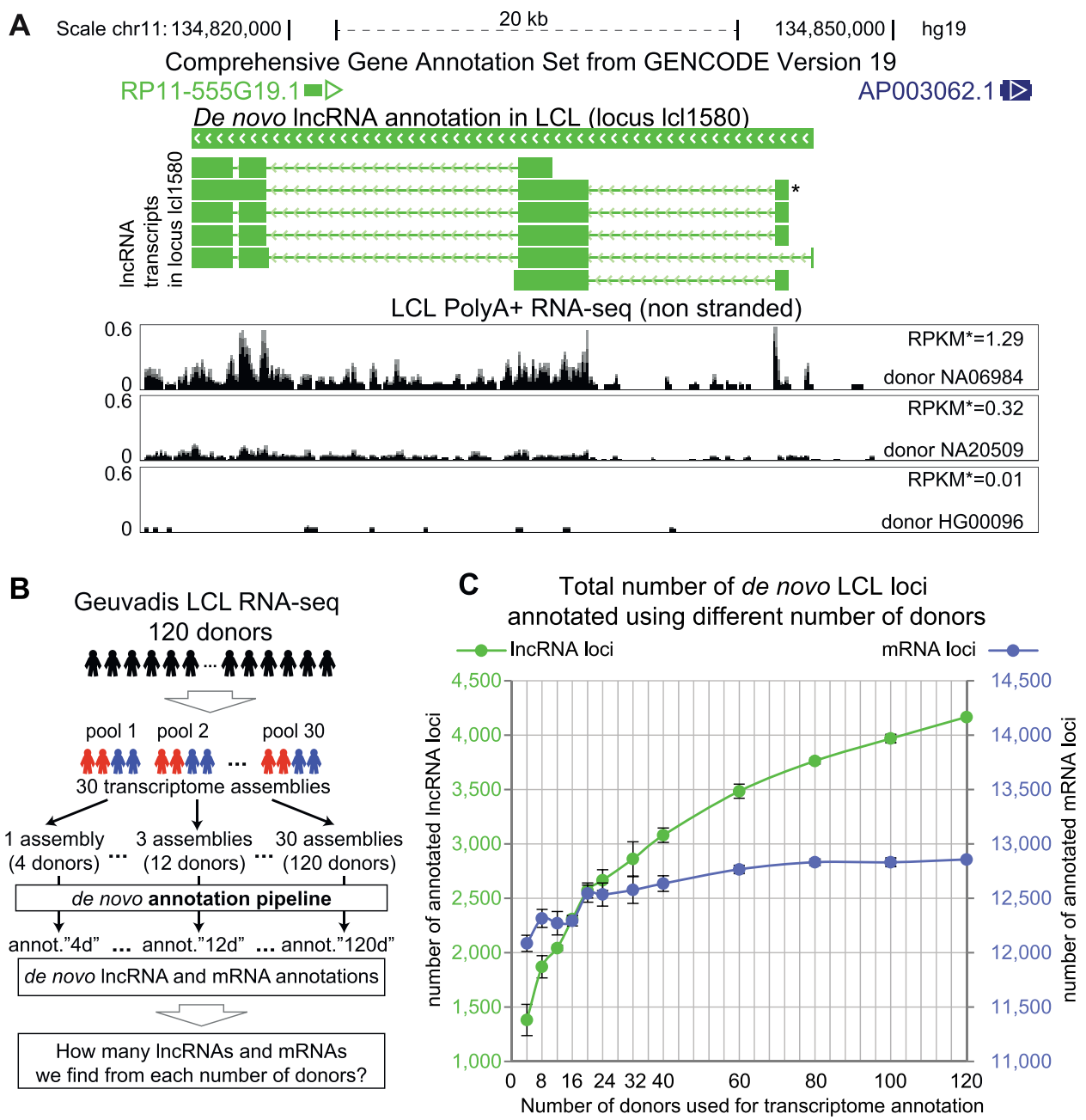


Fig. 7 Increasing donor number identifies more lncRNA loci. **a** Example of a highly variable LCL lncRNA locus lcl1580 not in public annotations. GENCODE-v19 annotates lncRNA RP11-555G19.1 and protein coding gene AP003062.1 transcribed in antisense direction to lcl1580 (top). Normalized non-strand-specific PolyA+ RNA-seq signal for three donors is displayed (scaling from 0 to 0.6). RPKM of the *transcript isoform is shown for each sample. **b** Analysis overview. GEUVADIS project LCL RNA-seq data from 120 donors was used to create 30 data pools (each with 100 million reads from two female (red) and two male (blue) donors) and to assemble 30 transcriptomes (Methods). An increasing number of assemblies (corresponding to from 4 to up to 120 donors) was merged to serve as input into the *de novo* lncRNA and mRNA identification pipeline (Additional file 1: Figure S1A). This created a series of LCL *de novo* lncRNA and mRNA annotations from an increasing number of donors. **c** LCL *de novo* lncRNA (green) and mRNA (blue) loci number annotated using increased donor number. Left: Y-axis for lncRNA loci (green). Right: Y-axis for mRNA loci (blue). The range of values is set to 3,500 on both Y-axes. Maximum number of lncRNA / mRNA loci annotated (at 120 donors): 4,166 / 12,857. Error bars: standard deviation of loci number between three replicates of random picking for each number of assemblies used (Additional file 11C)

the number of mRNA loci (blue lines) shows a much lower level of increase from 12,085 (four donors) to 12,857 loci (120 donors). This supports the hypothesis that adding more individuals to the identification pipeline increases the number of lncRNA loci but not the number of mRNA loci.

In contrast to the loci analysis, the number of transcript isoforms increased with similar kinetics for both lncRNAs and mRNAs (approximately seven-fold increase from four to 120 donors; Additional file 1: Figure S31C). The difference between lncRNAs and mRNAs is that an increasing donor number allows identification of an increasing number of transcript isoforms inside a stable number of mRNA loci, while lncRNAs retain a low median number of transcripts per locus and increase the number of loci annotated in the genome (Additional file 1: Figure S31D). Note that we did not expect to find non-annotated mRNAs loci since the mRNA *de novo* identification pipeline was limited to annotated mRNA genes. If the analysis did identify non-annotated mRNA loci they would be recognized among lncRNA candidates that were filtered by the pipeline step that estimated transcript protein-coding capability. However, this step only removed a low-level increase from 83 (four donors) to 198 (120 donors) loci (Additional file 1: Figure S31E). The slight increase in mRNA loci number with increasing donor number (Fig. 7c) likely arises from high inter-individual expression variability of a small number of mRNAs in LCLs. The larger increase in lncRNA loci number also arises from identifying more highly variable annotated lncRNAs when analyzing more donors, but also potentially by identifying novel lncRNA loci.

Assembling transcriptomes from pools of 100 million paired-end read does not increase Cufflinks sensitivity (Additional file 1: Figure S31A), but including more donors into the identification pipeline naturally increased the number of transcriptome assemblies merged and therefore the total amount of the RNA-seq data analyzed (from 1 to 30×10^8 sequencing reads). To control that this strategy did not only lead to the identification of marginally-expressed lncRNAs we plotted the RPKM of lncRNAs added to annotation with the addition of more donors (Additional file 1: Figure S32). This shows that median level of expression (in at least one donor used for identification) of newly-identified lncRNAs is RPKM of approximately 1, which means that 50 % of the newly-identified lncRNA transcripts are well-expressed (RPKM >1). This median level also does not decrease for transcripts that are only found with large donor numbers. In addition, we analyzed the dynamics of lncRNA identification with increasing the donor number in different expression bins (Additional file 1: Figure S33). This shows that lncRNAs from high-expression bins contribute substantially to the overall increase in lncRNA loci and transcript number. For example, four donor annotations identified 314 lncRNA transcripts initiating from 152 different loci in bin4 (that is, at least one donor used for identification expresses the transcript with $4 < \mathrm{RPKM} < 8$), while annotating from 120 donors identified 3,518 bin4-lncRNA transcripts initiating from 610 loci. Thus, while marginally-expressed lncRNAs are identified by adding more donors to the analysis, they only constitute a fraction of the newly-identified transcripts. Both controls show that identification of an increasing number of lncRNAs cannot be solely attributed to stochastic sampling sensitivity and identification of lowly-expressed transcripts, but likely arises from genuine expression variability between individuals.

We next asked if the lncRNA loci identified with increased donor numbers were new or known loci (as defined in Fig. 2a) and what were the dynamics of their identification. To do this we plotted the normalized number (the number of loci at 120 donors set to 100 %) of known (dark gray) and new (light gray) lncRNA loci versus donor number (Figure S34 in Additional file 1: Figure S34 and Additional file 11C). For comparison the same plot shows the dynamics of mRNA (dashed blue line) and all lncRNA (dashed green line) identification from the data in Fig. 7b. This shows that although the number of known lncRNA loci increases with donor number from 948.5 (four donors,) to 2,103 (120 donors), the number of novel lncRNA loci shows a more striking increase from 433.7 to 2063 loci (2.2-fold and 4.8-fold, respectively; Additional file 1: Figure S34) (note that non-integer loci numbers arise from averaging three replicates). While mRNA loci identification plateaued with four donors, the known lncRNA loci identification curve starts to plateau with >80 donors, but the new lncRNA identification curve does not plateau up to 120 donors.

Finally, we used the most comprehensive *de novo* annotation from 120 donors as a reference transcriptome to build a 'donor saturation curve' to test how well this annotation can be recreated using fewer individuals. We counted the number of reference 120 donor lncRNA and mRNA loci identified (defined by >50 % coverage, Additional file 1: Figure S35A, top, Additional file 11D) using a reduced number of donors. The resulting curve saturates for mRNAs, but does not saturate for lncRNAs even with 120 individuals. Only 27 % of lncRNA loci identified with 120 donors were identified using four donors, this increased to 50 % at 20 donors and thereafter continuing to rise. The difference between known and new lncRNA loci was consistent with observations in Additional file 1: Figure S34. We also assessed how well the exon structure of mRNAs and lncRNAs from the reference 120 donor annotation was recreated by annotations obtained using fewer donors (Additional file 1: Figure S35B). Median exonic coverage of mRNAs was above 90 % just using four donors, whereas lncRNAs require 80 donors to reach similar levels of exonic coverage. In summary, these analyses show that increasing the donor number will identify more lncRNA loci, however, the donor number required is vastly in excess of that required for mRNAs.

Discussion

An appreciation of the need to define the lncRNA landscape of the whole human genome is increasing with the number of known lncRNAs genes and with an understanding of the unique qualities of their biology. Although the GENCODE annotation comprises the largest public dataset with 15,877 lncRNA genes (version 21: http://www.gencodegenes.org/stats/archive.html#a21), later studies that used several thousand normal and malignant samples from numerous individuals identified four-fold more lncRNA genes [29]. Why the number of lncRNAs continues to rise apparently in excess of protein-coding gene number, is not yet clear. In this study we set out to annotate the lncRNA transcriptome of freshly harvested human granulocytes with the goal of investigating lncRNA inter-individual expression variability and determining how this influences lncRNA identification.

The resulting human granulocyte transcriptome obtained from 10 healthy individuals identified 1,591 lncRNA loci with a mean of 3.9 transcripts per locus. The same granulocytes express approximately six-fold more mRNA loci each with approximately three-fold more transcripts. The reduced activity of lncRNA loci relative to protein-coding loci has been noted [14, 62]. Comparing the granulocyte *de novo* annotation to the most commonly used public annotations (GENCODEv19: 23,898 lncRNA transcripts [58], RefSeq: 8,236 lncRNA transcripts [59], Cabili: 21,630 lncRNA transcripts [14]) that together contain 19,762 non-overlapping lncRNA loci, shows that one-third of granulocyte *de novo* lncRNA transcripts are not present and one-third added a new isoform to public-annotated loci. A comparison with the recent massive MiTranscriptome lncRNA annotation containing 46,331 new lncRNA loci [29], showed that 268 granulocyte lncRNA loci (17 % of the annotated granulocyte lncRNA transcriptome) were not previously reported. With the caveat that different annotation pipelines may influence identification, this shows that human granulocytes have a specific lncRNA landscape that needs to be defined prior to transcriptome analysis, rather than relying on integrative lncRNA landscapes from multiple cell types.

The identification of numerous new human granulocyte lncRNA loci is surprising in view of the extremely large numbers present in public annotations or datasets. Because of this we investigated if specific lncRNA biology features contribute to their under-representation in public databases by assessing if they were more prominent in new loci or isoforms. We first investigated four known features, that is, very tight tissue-specific expression, lower expression level, inefficient enrichment in PolyA+ selected fractions, and inefficient splicing (reviewed in [30, 34, 40]). In each case we demonstrated a significant difference for these features between lncRNAs and mRNAs and, in addition, demonstrated that these features are more prominent in new lncRNA loci and transcript isoforms. For example, reports from different species show that lncRNAs compared to mRNAs have tight tissue-specific expression and also are generally more lowly expressed [14, 15, 17, 18, 25, 65]. We found that while only 4 % of mRNA transcripts display granulocyte-specific expression, 32 % of lncRNA transcripts, and 57 % of novel lncRNA transcripts were granulocyte-specific. Similarly, lncRNA expression levels were 10-fold less abundant than mRNAs, as reported in many species (see above references), however expression of novel 'not in PA' lncRNA transcripts was 13-fold less abundant. We could also show that lncRNA enrichment in the PolyA+ fraction relative to total ribosomal-depleted fraction was reduced compared to mRNAs (respective median enrichments of 1.6-fold and 2.6-fold) in agreement with findings that a proportion of lncRNAs are not polyadenylated [66] and that this reduction was 1.6-fold greater for novel 'not in PA' lncRNA transcripts. A relatively new feature reported for imprinted *cis*-repressor lncRNAs such as *Airn* and *Ube3a-ats* [67, 68] and for some lncRNAs in human K562 cells [45] that could also affect the abundance of full-length transcripts in PolyA+ RNA fractions, is inefficient splicing. We accessed splicing efficiency of lncRNAs and mRNAs in our granulocyte data and showed that compared to mRNAs, lncRNAs are less efficiently spliced with a broad distribution of splicing efficiency. Median lncRNA splicing efficiency was reduced by 10.9 % compared to mRNAs, however, novel lncRNA transcripts showed 22.9 % reduction. We confirmed the inefficient splicing of lncRNAs and the greater reduction in novel lncRNA using the independent MiTranscriptome annotation [29]. Together this analysis shows not only that lncRNAs share several non-mRNA-like biology features, but also that these features are more prominent in new lncRNA loci and transcript isoforms and thus are likely to reduce lncRNA representation in public annotations.

The last feature examined that could influence the incomplete representation of lncRNAs in public databases is that of natural expression variation. We used the granulocyte annotation with seven donors sampled at three time points separated by at least 1 month, to estimate the natural expression variability of lncRNAs relative to mRNAs. This analysis shows that lncRNA expression is unexpectedly highly variable among a population and, while relatively stable over time within an individual, lncRNA expression variation is significantly larger than that of mRNAs independent of expression level. We find that when considering all the 6,249 *de novo* annotated granulocyte lncRNA transcripts only 40 % (2,490) are robustly expressed, while 17 % (1,069) display significant inter-individual variable expression even within the small

sample size of seven donors. Importantly, we show that high natural expression variability is not a consequence of the generally low expression of lncRNAs, as lncRNA transcripts/loci in all expression bins were more variable than mRNAs and also displayed higher percentage of significant inter-individual variable expression assessed by ANOVA test. The high inter-individual variability of lncRNA expression was unique enough to allow unsupervised grouping of replicates sampled over several months according to each of the seven donors. We verified high lncRNA inter-individual expression variability by demonstrating a similar difference for MiTranscriptome annotated transcripts expressed in granulocytes. We also analyzed an independent public RNA-seq lymphoblastoid cell dataset from GEUVADIS [50]. This LCL dataset derived from 462 donors displayed an overall higher median expression variability for both mRNAs and lncRNAs than the granulocyte dataset consisting of seven donors; however, the relative two-fold difference between lncRNAs and mRNAs loci and transcripts was similar. In each of the three above analyses we could show that novel lncRNA transcripts display higher expression variability than known lncRNA transcripts. Lastly, we demonstrated that high lncRNA inter-individual expression variability relative to mRNAs is a general phenomenon in human tissues, by analyzing multiple tissues from the GTEx project [64]. Interestingly, although we analyzed the same number of donors per tissue we found different absolute levels of lncRNA and mRNA expression variability, with skeletal muscle displaying the highest and LCL, nerve, and thyroid displaying the lowest variability level. As an important control, analyzing LCL in the GTEx dataset using the MiTranscriptome annotation showed similar levels of expression variability as that obtained by analyzing the GEUVADIS LCL dataset using our *de novo* LCL annotation. Overall, these expression variability analyses of public datasets, in additional to our granulocyte analysis presented here, confirm our conclusions and support the general nature of increased lncRNA natural expression variability compared to mRNAs.

Comparison of lncRNA and mRNA expression variability was performed as a small part of two previous studies. One LCL study analyzing splicing variability of protein-coding genes found a small number (183) of GENCODE lncRNAs with consistent higher expression variability than mRNAs, even in the absence of replicates [69]. The second study [55] reported a similar relative impact of inter-tissue and inter-individual variability to total variance in gene expression for highly expressed (median RPKM >2.5 among 1,641 analyzed samples comprising 43 body sites from 175 individuals) GENCODE-v12 lncRNAs and mRNAs. This implies, given the known increased inter-tissue variability of lncRNAs, that inter-individual variability of lncRNAs is also greater in its absolute value than that of mRNAs. This study additionally reported enrichment of lncRNAs among genes showing differential expression between individuals of different populations. Thus, the findings from both these studies are consistent with our demonstration here of higher natural expression variation of lncRNAs compared to mRNAs.

High lncRNA inter-individual expression variability highlights another striking biology feature that distinguishes lncRNAs from mRNAs. The finding that expression variability is more prominent in new lncRNA loci and reduced in reference lncRNA annotations also indicates it can influence identification. Thus public annotations based on limited numbers of human donors or derived from single animal or plant inbred strains, may have reduced representation of variably expressed lncRNAs. We demonstrate this with the GEUVADIS LCL RNA-seq data derived from one cell type, by showing that adding more donors to the analysis identifies more lncRNA genes in the human genome. The number of lncRNA loci increased continuously, with novel lncRNA showing a more striking increase than known lncRNAs. The MiTranscriptome study that used a donor number per tissue comparable to our LCL analysis [29] identified three-fold more novel lncRNAs than present in the three commonly used public databases (see above references). Our results also indicate that a granulocyte lncRNA annotation based on 10 donors, is most likely at the lower part of the donor saturation curve for this cell type. Moreover, our finding that the identification of novel lncRNA loci does not plateau even with 120 donors indicates that comprehensive annotation of lncRNAs in the human genome requires as many individuals as possible. The identification of high lncRNA intra- and inter-individual expression variability has implications for identifying lncRNAs and assessing their function and potential medical use. LncRNAs that lack consistent expression in some individuals are unlikely to be necessary for normal cell function, but may be functional in an age, environment, lifestyle, or disease related manner as shown for some protein-coding genes [54, 70]. At the same time, it cannot be assumed that a robustly expressed lncRNA has an important function in the cell type in which it is expressed. For example, the developmentally important *Airn* lncRNA retains robust expression after performing its silencing function [71]. Our results support the view that functional studies require an understanding of basic lncRNA biology in different individuals before they can be interpreted [36, 72].

The basis of increased inter-individual expression variation of lncRNAs relative to mRNAs is unknown. It may be relevant that, together with a lower conservation and faster evolution rate, human lncRNAs are recently

evolved loci, harboring more SNPs than protein-coding genes [49, 73]. LncRNAs may also be more susceptible to environmental and lifestyle factors that contribute to mRNA expression variation [54]. Studies of protein-coding genes and lncRNAs in LCLs prepared from different population groups conclude that both expression strength and alternative splicing contribute to expression variability [50, 69, 74, 75]. How this contributes to differences in lncRNA and mRNA expression variability is not known. Bidirectional lncRNAs that likely share a promoter with a neighboring protein-coding gene are regulated similarly to neighboring protein-coding genes [76] and we show that compared to intergenic or antisense lncRNAs, expression variability of bidirectional lncRNAs is more similar but still greater, than that of mRNAs. Inter-individual alternative splicing may contribute as some lncRNA loci display unusually high alternative splicing and variable exon structures [77]. However, this is not supported by our observation that expression variation over the whole locus is similar to that of transcript isoforms. LncRNA genes are considered to be similar to mRNA genes as both are transcribed by RNAPII (reviewed in [30, 34]). However, details of their promoters or enhancers that could explain the five non-mRNA-like features highlighted here (tight tissue-specificity, low expression, inefficient PolyA+ selection, inefficient splicing, and high inter-individual expression variation) have not yet been investigated. Some potential gene regulatory features (chromatin-modification patterns, splicing signals) are similar for lncRNAs and mRNAs [14, 18, 25, 78]. Some publications identified non-mRNA-like features in lncRNAs while others stress mRNA-like features, particularly of intergenic lncRNAs [15, 46, 79–81] (reviewed in [30, 34, 40]). The analysis of healthy granulocytes presented here supports the view that a lncRNA subpopulation shows distinct non-mRNA-like features, which now includes high inter- and intra-individual expression variability. Non-mRNA-like features of lncRNAs may have use in their classification, as it is likely to be relevant for their function [82, 83]. We show here that in healthy granulocytes only 40 % (2,490) of lncRNA transcripts are robustly expressed, while 17 % (1,069) of lncRNA transcripts show significant variable expression. The biological significance of robust or variable expression is not yet clear and both classes of lncRNAs may be useful for some studies. However, explanations of lncRNAs in terms of their evolution and function or proposals of their use as biomarkers or therapeutic targets first require an understanding of the robustness of their expression in healthy tissues.

Conclusions

We demonstrate here by analysis of human granulocyte RNA-seq data from multiple individuals that lncRNAs show unusually high natural expression variability compared to mRNAs. We use this dataset to generate a list of robustly and variably expressed granulocyte lncRNAs that will be of use in future applications. We also show that higher expression variability of lncRNAs is a general phenomenon inherent to diverse human tissues and cell lines that is of yet, unknown biological significance. High natural expression variability of lncRNAs, in addition to their tight tissue-specificity, low expression, inefficient PolyA+ selection, and inefficient splicing, identifies a set of five non-mRNA-like features that distinguish part of the lncRNA population from mRNAs and, also reduces their representation in reference annotations. We show that high inter-individual expression variability offers one explanation for the incomplete annotation of lncRNAs in many genomes. Our analysis shows that increasing the number of individuals analyzed will identify more lncRNA loci in the human genome, however, the donor number required is vastly in excess of that required for mRNAs. The finding of high expression variability of lncRNAs and its effect on identification provides novel guidelines for lncRNA annotation and additional considerations for design of functional studies and personalized medicine approaches.

Methods

Sample collection from healthy donors

Ten volunteers (five men, five women; age range: 27–62 years) without obvious disease were recruited to donate blood. Seven volunteers donated blood three times with gaps of 5 to 21 weeks (Additional file 2A). The remainder donated once only. Donors abstained from eating on the morning of donation; 45 mL of venous blood was collected between 10:00 and 11:00 into VACUETTE® Sodium Citrate Coagulation Tubes and processed immediately. Granulocytes were isolated using density gradient centrifugation and immediately used for RNA preparation either depleted for ribosomal RNA using the RiboZero rRNA removal kit Human/Mouse/Rat (Epicentre) or a polyA enriched using the TruSeq RNA Sample Prep Kit v2 (Illumina) (details in Additional file 1: Supplemental Methods).

RNA-seq library preparation and read alignment

(a) Non-strand-specific libraries were prepared using the TruSeq RNA Sample Prep Kit v2 (Illumina) following the manufacturer's protocol. (b) Strand-specific library preparation used same kit with modifications [84]. Equal concentrations of barcoded libraries were pooled for 50 bp or 100 bp paired-end sequencing by Illumina HiSeq 2000 (Biomedical Sequencing Facility http://biomedical-sequencing.at/). After base-calling and sample de-multiplexing, the RNA-seq data were provided as archived .fastq or unmapped .bam files. RNA sequencing

reads were aligned using STAR aligner with adjusted default parameters [56] (details in Additional file 1: Supplemental Methods).

RNA-seq read number

Three stranded samples were sequenced per flow cell lane generating 22 to 79 million 100 bp PE reads per sample. Unstranded PolyA+ RNA-seq samples varied from 24 to 38 million 100 bp PE and 64 to 91 million 50 bp PE reads. In total we obtained 17 PolyA+ RNA-seq datasets and 21 total RNA-seq dataset totaling 2.13 billion reads (Additional file 2B).

Annotating mRNAs and lncRNAs in primary granulocytes

A total of 784 million PolyA+ RNA-seq reads from 10 donors were used to *de novo* annotate lncRNA and mRNA transcriptomes in granulocytes (see details in Additional file 1: Supplemental Methods). The final *de novo* annotation of human primary granulocytes was 132,864 mRNAs forming 10,092 genomic loci (average 13.2 transcripts per locus) and 6,249 lncRNAs forming 1,591 genomic loci (average 3.9 transcripts per locus). Assembly quality was assessed by inspecting *de novo* annotation of well-known lncRNAs like XIST (Additional file 1: Figure S2A) and by analyzing completeness of assembly of RefSeq (Additional file 1: Figure S2B) and GENCODE-v19 (Additional file 1: Figure S2C) annotated mRNAs.

Positional classification of lncRNAs

lncRNA loci and transcripts were divided into three classes based on their relative position to protein-coding genes. We combined *de novo* mRNA annotation with public protein-coding gene annotations by GENCODE-v19 and RefSeq to obtain the most comprehensive annotation of protein-coding genes in granulocytes. We then called lncRNA loci/transcripts bidirectional if they shared or overlapped a promoter (defined as TSS +/- 1.5 kb) with a protein-coding gene. LncRNA loci/transcripts overlapping a protein-coding gene in the antisense direction were called 'antisense' (sense direction overlaps were removed from the annotation). The third position-based class 'intergenic', had no overlap with a protein-coding gene.

Cloning of full-length lncRNA transcripts

RT-PCR was performed on granulocyte cDNA to amplify full-length lncRNA transcripts prior to cloning. PCR primers (http://biotools.umassmed.edu/bioapps/primer3_www.cgi) spanned the transcript from first to the last exon and the PCR product length limited to 1.5 kb (Additional file 2F). Isolated plasmid DNA was Sanger Sequenced and aligned to the human genome using BLAT. Cloned sequences are displayed as a UCSC screen shot with the *de novo* lncRNA annotation, primers, and BLAT alignment (Additional file 1: Figures S4-S8). Seventy-five cloned sequences were submitted to GENBANK (Additional file 2G).

Public RNA-seq data mining

We downloaded publicly available raw strand-specific RNA-seq data (fastq files) from various cell types/tissues produced by the ENCODE project and Illumina Human Body Map Project (see list in Additional file 2H), processed it as for other sequencing data in the study (see: RNA-seq read alignment).

RPKM

This was calculated using RPKM_count.py (RSEQC package). Expression of a transcript is the RPKM of exons of a one transcript, expression over a locus is RPKM of the whole locus including intronic signal.

Splicing efficiency analysis

We estimated splicing efficiency for each splice site of each multiexonic transcript in our ribosomal-depleted granulocyte RNA-seq from seven donors with three time points pooled at the alignment stage to increase coverage. Splicing efficiency of each splice site was calculated separately in each donor. We calculated RPKM of the exonic and intronic boundaries of the splice site (45 bp each, leaving out 5 bp directly at the splice site to allow for imprecision of splice site identification), calculated the ratio of intronic to exonic signal, and by that estimated how efficiently this splice site was used (Additional file 1: Figure S13A). A splice site was discarded if exonic RPKM was below the cutoff (RPKM = 0.2) in any of the seven donors. We then introduced a value 'Splicing efficiency' (S), ranging from 0 for completely unused splice sites (intronic signal equal or higher than exonic signal) to 100 for optimally used spliced splice sites (no intronic signal detected). S = 100*(1-RPKMintronic/RPKMexonic). We replaced all the negative S values (when intronic signal was higher than exonic signal) with 0, defining such cases as full absence of splicing. We averaged the splicing efficiency value calculated from seven donors for each splice site. Splicing efficiency of a transcript was then defined as the maximal splicing efficiency achieved by the most efficiently spliced site of that transcript. Splicing efficiency of a locus was similarly defined by the maximal splicing efficiency among all transcripts (all splice sites) in the locus.

Assigning *P* value to boxplot comparisons

Every boxplot was plotted using values for all the transcripts/loci analyzed (number of transcripts/loci indicated in the boxplot). The difference in population sizes of compared transcript/loci types was accounted for by

performing statistical tests on equalized population sizes. Namely, the larger population was randomly subsampled to match the size of smaller population and Mann–Whitney U test was applied to estimate significance of the difference between the populations with equalized sizes. Subsampling and statistical tests were performed three times for each comparison and the three *P* values obtained were averaged to give the resulting *P* value to be indicated on the boxplot.

Inter-individual expression variability analysis

Inter-individual expression variability was estimated by calculating standard deviation of expression between analyzed donors then normalizing it to the mean expression of the locus/transcript among all analyzed donors. For granulocytes we assessed variability between seven donors (expression of a locus/transcript in each donor was calculated as a mean of expression of the three time points of this donor). For LCL we assessed variability between 462 donors.

GEUVADIS project RNA-seq data analysis

We downloaded and aligned using a common pipeline all 462 PolyA+ 75 bp paired end RNA-seq raw sequencing datasets provided by GEUVADIS RNA-seq project (http://www.ebi.ac.uk/ena/data/view/ERR188021-ERR188482). The data contained donors from five populations (http://www.1000genomes.org/category/frequently-asked-questions/population). We picked two female and two male unrelated donors from each population and used RNA-seq from these 20 donors to assemble the LCL *de novo* lncRNA and mRNA transcriptome. We pooled the samples into five groups with a similar number of aligned spliced reads (Additional file 2I) and performed transcriptome assembly following the pipeline described for granulocytes. As the RNA-seq datasets were not strand-specific we used strand-specific PolyA+ RNA-seq of GM12878 from the ENCODE project (Additional file 2H) in the pipeline where needed. Quality assembly (Additional file 1: Figure S24B) was assessed as for granulocytes.

GTEx RNA-seq data analysis

Aligned (as described in [55]) RNA-seq data from the GTEx project (http://www.gtexportal.org/home/) were downloaded from dbGaP (https://dbgap.ncbi.nlm.nih.gov/) as described in (http://www.gtexportal.org/static/misc/GTEx_Poster_CommunityMeeting_TY.pdf) after we applied and were granted data access. We downloaded RNA-seq data for nine tissues (namely lymphoblastoid cell line (LCL), adipose, artery, cerebellum, heart, lung, muscle, nerve, and thyroid), from 10 male and 10 female individuals each (Additional file 2J). The aligned RNA-seq datasets were unstranded and ranged from 14.8 to 85.4 (average 52.1) million paired-end reads each. We calculated RPKM of MiTranscriptome annotated multi-exonic lncRNAs and mRNAs in all samples and performed variability analysis between 20 individuals per tissue.

Donor saturation curve

One hundred and twenty out of 462 GEUVADIS RNA-seq samples containing more than 25 million reads were picked for the analysis from 12 unrelated women and men from each of the five population groups. A total of 25 million reads were randomly sampled from each RNA-seq sample using DownsampleSam.jar (Picard tools http://broadinstitute.github.io/picard/command-line-overview.html#DownsampleSam). Donors were grouped into 30 groups each with two women plus two men from the same population and the reads from the four donors were pooled using MergeSamFiles.jar (Picard tools http://broadinstitute.github.io/picard/command-line-overview.html#MergeSamFiles) to produce 30×100 million read pools. Cufflinks was used to assemble a transcriptome from each pool (Additional file 1: Supplemental Methods) resulting in 30 transcriptome assemblies. Of these 30 assemblies, 1, 2, 3, 4, 5, 6, 8, 10, 15, 20, 25, or 30 assemblies were used to annotate *de novo* LCL transcriptomes from different number of donors (4, 8, 12, 16, 20, 24, 32, 40, 60, 80, 100, and 120, respectively) and to define the relation between the number of loci (Y axis) and the number of donors/assemblies (X axis). We randomly picked the needed number of assemblies from the list of 30. The random picking was performed three times for each number of assemblies (Additional file 1: Figure S31B), except when all 30 assemblies were used for the last point. The picked assemblies were then merged with Cuffmerge and underwent the previously established *de novo* annotation pipeline (Additional file 1: Supplemental Methods).

Ethics statement

Peripheral blood samples were collected from healthy volunteers after written informed consent at the Vienna General Hospital (Allgemeines Krankenhaus der Stadt Wien, Klinische Abteilung für Hämatologie und Hämostaseologie). The study was approved by the local Ethics committee of the Medical University of Vienna ('Ethik Kommission der Medizinischen Universität Wien') and experimental methods comply with the Helsinki Declaration.

Availability of data

Raw granulocyte RNA-seq data, RPKM, and variability values for granulocyte *de novo* lncRNAs and mRNAs as well as their BED12 annotation files were deposited in NCBI's Gene Expression Omnibus [85] and are accessible through GEO Series accession number GSE70390 (http://www.ncbi.nlm.nih.gov/geo/

query/acc.cgi?acc=GSE70390). LncRNA annotations in granulocytes and LCL created in the study are available to directly download as Additional files in bed12 format. Genbank accession numbers for sequenced lncRNAs are listed in Additional file 2G.

Additional files

Additional file 1: Supplemental Figures (S1-S35) with legends and Supplemental Methods. (PDF 8255 kb)

Additional file 2: A Human granulocyte samples sequenced in this study. B List of human granulocyte RNA-seq datasets produced in the study. **C** Pools used for human granulocyte transcriptome assembly. **D** Well-known lncRNAs used to adjust RNAcode and CPC pipeline output. **E** Validation of *de novo* granulocyte lncRNA splice junctions by means of exon spanning RT-PCR. **F** Validation of *de novo* granulocyte lncRNA transcripts not supported by public annotations by means of cloning and Sanger Sequencing: overview. **G** Validation of *de novo* granulocyte lncRNA transcripts not supported by public annotations by means of cloning and Sanger Sequencing: sequencing results and Genbank accession numbers. **H** Overview of the publicly available RNA-seq datasets used in the study. **I** Pools used for human LCL transcriptome assembly (GEUVADIS raw RNA-seq data used - [50]). **J** Overview of the GTEx RNA-seq datasets used in the study. (XLSX 127 kb)

Additional file 3: Granulocyte *de novo* lncRNA loci annotation (1,561 loci): BED12 formatted file can be directly uploaded into UCSC browser. Column 5 indicates number of transcripts in the locus. (BED 119 kb)

Additional file 4: Granulocyte *de novo* lncRNA transcript annotation (6,249 transcripts): BED12 formatted file can be directly uploaded into UCSC browser. (BED 655 kb)

Additional file 5: Granulocyte *de novo* mRNA loci annotation (10,092 loci): BED12 formatted file can be directly uploaded into UCSC browser. Column 5 indicates number of transcripts in the locus. (BED 765 kb)

Additional file 6: Granulocyte *de novo* mRNA transcript annotation (132,864 transcripts): BED12 formatted file can be directly uploaded into UCSC browser. (BED 23458 kb)

Additional file 7: A List of robust lncRNA transcripts in granulocytes (2,825 transcripts): columns are formatted as a BED12 file. B List of robust well expressed (RPKM >1) lncRNA transcripts in granulocytes (931 transcripts): columns are formatted as a BED12 file. (XLSX 250 kb)

Additional file 8: Annotation of granulocyte *de novo* lncRNA transcripts showing significantly variable expression (1,069 transcripts): BED12 formatted file can be directly uploaded into UCSC browser. (BED 117 kb)

Additional file 9: LCL *de novo* lncRNA loci annotation (2,611 loci): BED12 formatted file can be directly uploaded into UCSC browser. Column 5 indicates number of transcripts in the locus. (BED 197 kb)

Additional file 10: LCL *de novo* lncRNA transcript annotation (8,560 transcripts): BED12 formatted file can be directly uploaded into UCSC browser. (BED 884 kb)

Additional file 11: Donor saturation curve samples and pools: overview with list of donors, assemblies, and number of loci identified using different number of donors. A List of 120 donors used in the donor saturation study with corresponding population and pool it was grouped into. **B** List of randomly picked pools for each data point. **C** Number of *de novo* lncRNA and mRNA loci annotated using different number of transcriptome assemblies (donors) – data for plotting Fig. 7b, S32C-E and S34. **D** Number of *de novo* lncRNA and mRNA loci from '120 donors' annotation identified using less transcriptome assemblies (donors) - data for plotting donor saturation curve - Figure S35A. **E** Number of *de novo* lncRNAs from different expression bins identified from increasing number of donors - data for plotting Figure S33. (XLSX 34 kb)

Competing interests
The authors (AEK, CPD, PMG, HG, BG, CC, RK, FMP, and DPB) declare no conflict of interest.

Authors' contributions
AEK and DPB conceived the study and wrote the manuscript. AEK performed blood sample processing, library preparation, experimental work, *de novo* lncRNA and mRNA identification, and other bioinformatic analyses. CC prepared the majority of PolyA enriched RNA-seq libraries. PMG established RNA-seq protocols and contributed to the splicing calculation method. FMP and CPD assembled the protein-coding potential estimation pipeline, wrote some custom scripts used in the study, and helped with the bioinformatic analysis. Blood samples were collected in collaboration with HG, BG, and RK. All authors read and approved this manuscript.

Acknowledgments
We thank Ruth Klement, Tomasz Kulinski, Elisangela Valente, Elisabeth Salzer, and Roland Jäger for technical/bioinformatic assistance and advice, the CeMM IT department and José Manuel Molero for help and advice on software usage, the Biomedical Sequencing Facility (http://biomedical-sequencing.at/) for sequencing and advice, Jacques Colinge, Daniel Andergassen, and Tomasz Kulinski for discussions, Quanah Hudson and Jörg Menche for reading and commenting on the manuscript.

Funding
This study was partly funded by the Austrian Science Fund (FWF F43-B09, FWF W1207-B09). PMG is a recipient of a DOC Fellowship of the Austrian Academy of Sciences.

Author details
[1]CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Lazarettgasse 14, AKH BT 25.3, 1090 Vienna, Austria. [2]Present Address: Institute of Science and Technology Austria, Lab Building East, Am Campus 1, A-3400 Klosterneuburg, Austria. [3]Department of Internal Medicine I, Division of Hematology and Blood Coagulation, Medical University of Vienna, Vienna, Austria. [4]Present Address: Piso 23, Av. Santa Fe No 481, Lomas de Santa Fe 05349, D.F., Mexico.

Received: 27 October 2015 Accepted: 6 January 2016
Published online: 29 January 2016

References
1. Morris KV, Mattick JS. The rise of regulatory RNA. Nat Rev Genet. 2014;15(6):423–37.
2. Bonasio R, Shiekhattar R. Regulation of transcription by long noncoding RNAs. Annu Rev Genet. 2014;48:433–55.
3. Geisler S, Coller J. RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. Nat Rev Mol Cell Biol. 2013;14(11):699–712.
4. Quinodoz S, Guttman M. Long noncoding RNAs: an emerging link between gene regulation and nuclear organization. Trends Cell Biol. 2014;24(11):651–63.
5. Bergmann JH, Spector DL. Long non-coding RNAs: modulators of nuclear structure and function. Curr Opin Cell Biol. 2014;26:10–8.
6. Batista PJ, Chang HY. Long noncoding RNAs: cellular address codes in development and disease. Cell. 2013;152(6):1298–307.
7. Ng SY, Lin L, Soh BS, Stanton LW. Long noncoding RNAs in development and disease of the central nervous system. Trends Genet. 2013;29(8):461–8.
8. Lee JT, Bartolomei MS. X-inactivation, imprinting, and long noncoding RNAs in health and disease. Cell. 2013;152(6):1308–23.
9. Wahlestedt C. Targeting long non-coding RNA to therapeutically upregulate gene expression. Nat Rev Drug Discov. 2013;12(6):433–46.
10. Cheetham SW, Gruhl F, Mattick JS, Dinger ME. Long noncoding RNAs and the genetics of cancer. Br J Cancer. 2013;108(12):2419–25.
11. Thum T. Noncoding RNAs and myocardial fibrosis. Nat Rev Cardiol. 2014;11(11):655–63.
12. Quagliata L, Matter MS, Piscuoglio S, Arabi L, Ruiz C, Procino A, et al. Long noncoding RNA HOTTIP/HOXA13 expression is associated with disease progression and predicts outcome in hepatocellular carcinoma patients. Hepatology. 2014;59(3):911–23.

13. Malik R, Patel L, Prensner JR, Shi Y, Iyer MK, Subramaniyan S, et al. The lncRNA PCAT29 inhibits oncogenic phenotypes in prostate cancer. Mol Cancer Res. 2014;12(8):1081–7.
14. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 2011;25(18):1915–27.
15. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 2009;458(7235):223–7.
16. Pervouchine DD, Djebali S, Breschi A, Davis CA, Barja PP, Dobin A, et al. Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. Nat Commun. 2015;6:5903.
17. Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell. 2011;147(7):1537–50.
18. Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, Levin JZ, et al. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res. 2012;22(3):577–91.
19. Li T, Wang S, Wu R, Zhou X, Zhu D, Zhang Y. Identification of long non-protein coding RNAs in chicken skeletal muscle using next generation sequencing. Genomics. 2012;99(5):292–8.
20. Tan MH, Au KF, Yablonovitch AL, Wills AE, Chuang J, Baker JC, et al. RNA sequencing reveals a diverse and dynamic repertoire of the Xenopus tropicalis transcriptome over development. Genome Res. 2013;23(1):201–16.
21. Liu J, Wang H, Chua NH. Long noncoding RNA transcriptome of plants. Plant Biotechnol J. 2015;13(3):319–28.
22. Nam JW, Bartel DP. Long noncoding RNAs in C. elegans. Genome Res. 2012;22(12):2529–40.
23. van Werven FJ, Neuert G, Hendrick N, Lardenois A, Buratowski S, van Oudenaarden A, et al. Transcription of two long noncoding RNAs mediates mating-type control of gametogenesis in budding yeast. Cell. 2012;150(6):1170–81.
24. Kim T, Xu Z, Clauder-Munster S, Steinmetz LM, Buratowski S. Set3 HDAC mediates effects of overlapping noncoding transcription on gene induction kinetics. Cell. 2012;150(6):1158–69.
25. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. 2012;22(9):1775–89.
26. Hangauer MJ, Vaughn IW, McManus MT. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. PLoS Genet. 2013;9(6):e1003569.
27. Ranzani V, Rossetti G, Panzeri I, Arrigoni A, Bonnal RJ, Curti S, et al. The long intergenic noncoding RNA landscape of human lymphocytes highlights the regulation of T cell differentiation by linc-MAF-4. Nat Immunol. 2015;16(3):318–25.
28. Amin V, Harris RA, Onuchic V, Jackson AR, Charnecki T, Paithankar S, et al. Epigenomic footprints across 111 reference epigenomes reveal tissue-specific epigenetic regulation of lincRNAs. Nat Commun. 2015;6:6370.
29. Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, et al. The landscape of long noncoding RNAs in the human transcriptome. Nat Genet. 2015;47(3):199–208.
30. Mercer TR, Mattick JS. Structure and function of long noncoding RNAs in epigenetic regulation. Nat Struct Mol Biol. 2013;20(3):300–7.
31. Quek XC, Thomson DW, Maag JL, Bartonicek N, Signal B, Clark MB, et al. lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. Nucleic Acids Res. 2015;43(Database issue):D168–73.
32. Kowalczyk MS, Higgs DR, Gingeras TR. Molecular biology: RNA discrimination. Nature. 2012;482(7385):310–1.
33. Raabe CA, Brosius J. Does every transcript originate from a gene? Ann N Y Acad Sci. 2015;1341:136–48.
34. Ulitsky I, Bartel DP. lincRNAs: genomics, evolution, and mechanisms. Cell. 2013;154(1):26–46.
35. Chu C, Spitale RC, Chang HY. Technologies to probe functions and mechanisms of long noncoding RNAs. Nat Struct Mol Biol. 2015;22(1):29–35.
36. Bassett AR, Akhtar A, Barlow DP, Bird AP, Brockdorff N, Duboule D, et al. Considerations when investigating lncRNA function in vivo. Elife. 2014;3:e03058.
37. Mudge JM, Frankish A, Harrow J. Functional transcriptomics in the post-ENCODE era. Genome Res. 2013;23(12):1961–73.
38. Mattick JS, Rinn JL. Discovery and annotation of long noncoding RNAs. Nat Struct Mol Biol. 2015;22(1):5–7.
39. Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP, Ulitsky I. Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. Cell Rep. 2015;11(7):1110–22.
40. Kornienko AE, Guenzl PM, Barlow DP, Pauler FM. Gene regulation by the act of long non-coding RNA transcription. BMC Biol. 2013;11:59.
41. Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. Annu Rev Biochem. 2012;81:145–66.
42. Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. Cell. 2013;154(1):240–51.
43. Orom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, et al. Long noncoding RNAs with enhancer-like function in human cells. Cell. 2010;143(1):46–58.
44. Kelley D, Rinn J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biol. 2012;13(11):R107.
45. Tilgner H, Knowles DG, Johnson R, Davis CA, Chakrabortty S, Djebali S, et al. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. Genome Res. 2012;22(9):1616–25.
46. Clark MB, Johnston RL, Inostroza-Ponta M, Fox AH, Fortini E, Moscato P, et al. Genome-wide analysis of long noncoding RNA stability. Genome Res. 2012;22(5):885–98.
47. Kutter C, Watt S, Stefflova K, Wilson MD, Goncalves A, Ponting CP, et al. Rapid turnover of long noncoding RNAs and the evolution of gene expression. PLoS Genet. 2012;8(7):e1002841.
48. Johnsson P, Lipovich L, Grander D, Morris KV. Evolutionary conservation of long non-coding RNAs; sequence, structure, function. Biochim Biophys Acta. 2014;1840(3):1063–71.
49. Necsulea A, Kaessmann H. Evolutionary dynamics of coding and non-coding transcriptomes. Nat Rev Genet. 2014;15(11):734–48.
50. Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013;501(7468):506–11.
51. Spielman RS, Bastone LA, Burdick JT, Morley M, Ewens WJ, Cheung VG. Common genetic variants account for differences in gene expression among ethnic groups. Nat Genet. 2007;39(2):226–31.
52. Storey JD, Madeoy J, Strout JL, Wurfel M, Ronald J, Akey JM. Gene-expression variation within and among human populations. Am J Hum Genet. 2007;80(3):502–9.
53. Whitney AR, Diehn M, Popper SJ, Alizadeh AA, Boldrick JC, Relman DA, et al. Individuality and variation in gene expression patterns in human blood. Proc Natl Acad Sci U S A. 2003;100(4):1896–901.
54. Dumeaux V, Olsen KS, Nuel G, Paulssen RH, Borresen-Dale AL, Lund E. Deciphering normal blood gene expression variation–The NOWAC postgenome study. PLoS Genet. 2010;6(3):e1000873.
55. Mele M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, et al. Human genomics. The human transcriptome across tissues and individuals. Science. 2015;348(6235):660–5.
56. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.
57. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7(3):562–78.
58. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012;22(9):1760–74.
59. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, et al. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. 2014;42(Database issue):D756–763.
60. Washietl S, Findeiss S, Muller SA, Kalkhof S, von Bergen M, Hofacker IL, et al. RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. RNA. 2011;17(4):578–94.
61. Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res. 2007;35(Web Server issue):W345–349.
62. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. Nature. 2012;489(7414):101–8.
63. Chambers JM, Freeny A, Heiberger RM. Analysis of variance; designed experiments. Pacific Grove, CA: Wadsworth & Brooks/Cole; 1992.

64. Consortium TG. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015;348(6235):648–60.
65. Ravasi T, Suzuki H, Pang KC, Katayama S, Furuno M, Okunishi R, et al. Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. Genome Res. 2006;16(1):11–9.
66. Wilusz JE. Long noncoding RNAs: Re-writing dogmas of RNA processing and stability. Biochim Biophys Acta. 1859;2016:128–38.
67. Seidl CI, Stricker SH, Barlow DP. The imprinted Air ncRNA is an atypical RNAPII transcript that evades splicing and escapes nuclear export. Embo J. 2006;25(15):3565–75.
68. Meng L, Person RE, Beaudet AL. Ube3a-ATS is an atypical RNA polymerase II transcript that represses the paternal expression of Ube3a. Hum Mol Genet. 2012;21(13):3001–12.
69. Gonzalez-Porta M, Calvo M, Sammeth M, Guigo R. Estimation of alternative splicing variability in human populations. Genome Res. 2012;22(3):528–38.
70. Glass D, Vinuela A, Davies MN, Ramasamy A, Parts L, Knowles D, et al. Gene expression changes with age in skin, adipose tissue, blood and brain. Genome Biol. 2013;14(7):R75.
71. Santoro F, Mayer D, Klement RM, Warczok KE, Stukalov A, Barlow DP, et al. Imprinted Igf2r silencing depends on continuous Airn lncRNA expression and is not restricted to a developmental window. Development. 2013;140(6):1184–95.
72. Goff LA, Rinn JL. Linking RNA biology to lncRNAs. Genome Res. 2015;25(10):1456–65.
73. Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, et al. The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature. 2014;505(7485):635–40.
74. Kwan T, Benovoy D, Dias C, Gurd S, Provencher C, Beaulieu P, et al. Genome-wide analysis of transcript isoform variation in humans. Nat Genet. 2008;40(2):225–31.
75. Martin AR, Costa HA, Lappalainen T, Henn BM, Kidd JM, Yee MC, et al. Transcriptome sequencing from diverse human populations reveals differentiated regulatory architecture. PLoS Genet. 2014;10(8):e1004549.
76. Sigova AA, Mullen AC, Molinie B, Gupta S, Orlando DA, Guenther MG, et al. Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. Proc Natl Acad Sci U S A. 2013;110(8):2876–81.
77. Hacisuleyman E, Goff LA, Trapnell C, Williams A, Henao-Mejia J, Sun L, et al. Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. Nat Struct Mol Biol. 2014;21(2):198–206.
78. Ponjavic J, Ponting CP, Lunter G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Res. 2007;17(5):556–65.
79. Kapranov P, St Laurent G, Raz T, Ozsolak F, Reynolds CP, Sorensen PH, et al. The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' un-annotated RNA. BMC Biol. 2010;8:149.
80. St Laurent G, Shtokalo D, Tackett MR, Yang Z, Eremina T, Wahlestedt C, et al. Intronic RNAs constitute the major fraction of the non-coding RNA in mammalian cells. BMC Genomics. 2012;13:504.
81. Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci U S A. 2009;106(28):11667–72.
82. Wright MW. A short guide to long non-coding RNA gene nomenclature. Hum Genomics. 2014;8:7.
83. St Laurent G, Wahlestedt C, Kapranov P. The Landscape of long noncoding RNA classification. Trends Genet. 2015;31(5):239–51.
84. Sultan M, Dokel S, Amstislavskiy V, Wuttig D, Sultmann H, Lehrach H, et al. A simple strand-specific RNA-Seq library preparation protocol combining the Illumina TruSeq RNA and the dUTP methods. Biochem Biophys Res Commun. 2012;422(4):643–6.
85. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30(1):207–10.

2.2 Research that was not included in Publication 2

2.2.1 Analysis of Blueprint neutrophil ChIP-seq data reveals difference in histone modifications on granulocyte lncRNAs and mRNAs.

2.2.1.1 Blueprint neutrophil ChIP-seq data analysis

In order to more comprehensively analyze the granulocyte lnRNAs annotated in the study (see Publication 2) I used publicly available ChIP-seq data from the BLUEPRINT project (http://www.blueprint-epigenome.eu/). The BLUEPRINT project has performed FACS-sorting of various primary blood cell types (http://www.blueprint-epigenome.eu/index.cfm?p=7BCEDA45-EC73-3496-2C823D929DD423DB), including neutrophils (that constitute the vast majority of the granulocyte population in healthy donors), followed by various epigenetic analyses, including ChIP-seq of various histone marks. I made use of these publicly available (under restricted use, application for the access needed) data to analyze histone modifications occurring in the gene body, promoter and exons of lncRNAs and mRNAs annotated in granulocytes from Publication 2 and to find if these show any significant difference as other features described above.

ChIP-seq data for 6 different marks were available: H3K4me3 (a classical active promoter mark), H3K27ac (a classical open chromatin mark common for enhancers and promoters), H3K4me1 (a classical active enhancer mark), H3K36me3 (a classical active transcription through the gene body mark), H3K27me3 (a mark indicative of a facultative repressed inactive promoter), and H3K9me3 (a mark indicative of a constituently repressed promoter) (Zhang et al, 2015). ChIP for each mark, as well as the Input control, was performed by the BLUEPRINT project for 6 healthy individuals (with the exception of H3K27ac that was performed for 5 individuals) and I analyzed each individual sample for unspecific binding with an Input control (Materials and Methods, Table1). I downloaded the raw sequencing data, aligned it with STAR (Dobin et al, 2013) and calculated read coverage over promoters, exons and loci of *de novo* granulocyte lncRNAs/mRNAs using *coverageBed* software (Materials and Methods). Coverage values had then to be normalized by the number of reads in each sample and by the length of the analyzed promoter/exons/locus. Next, for each histone mark in each donor, the coverage value in the corresponding Input sample was subtracted from the coverage value of the mark in order to account for unspecific binding. The resulting histone mark coverage values for each promoter/exons/locus were averaged among the available

donors and plotted as boxpots comparing granulocyte *de novo* lncRNA (green) and mRNA (blue) populations (Figure 10: H3K27ac, H3K27me3, H3K36me3; Figure 11: H3K4me1, H3K4me3, H3K9me3).

Figure 10

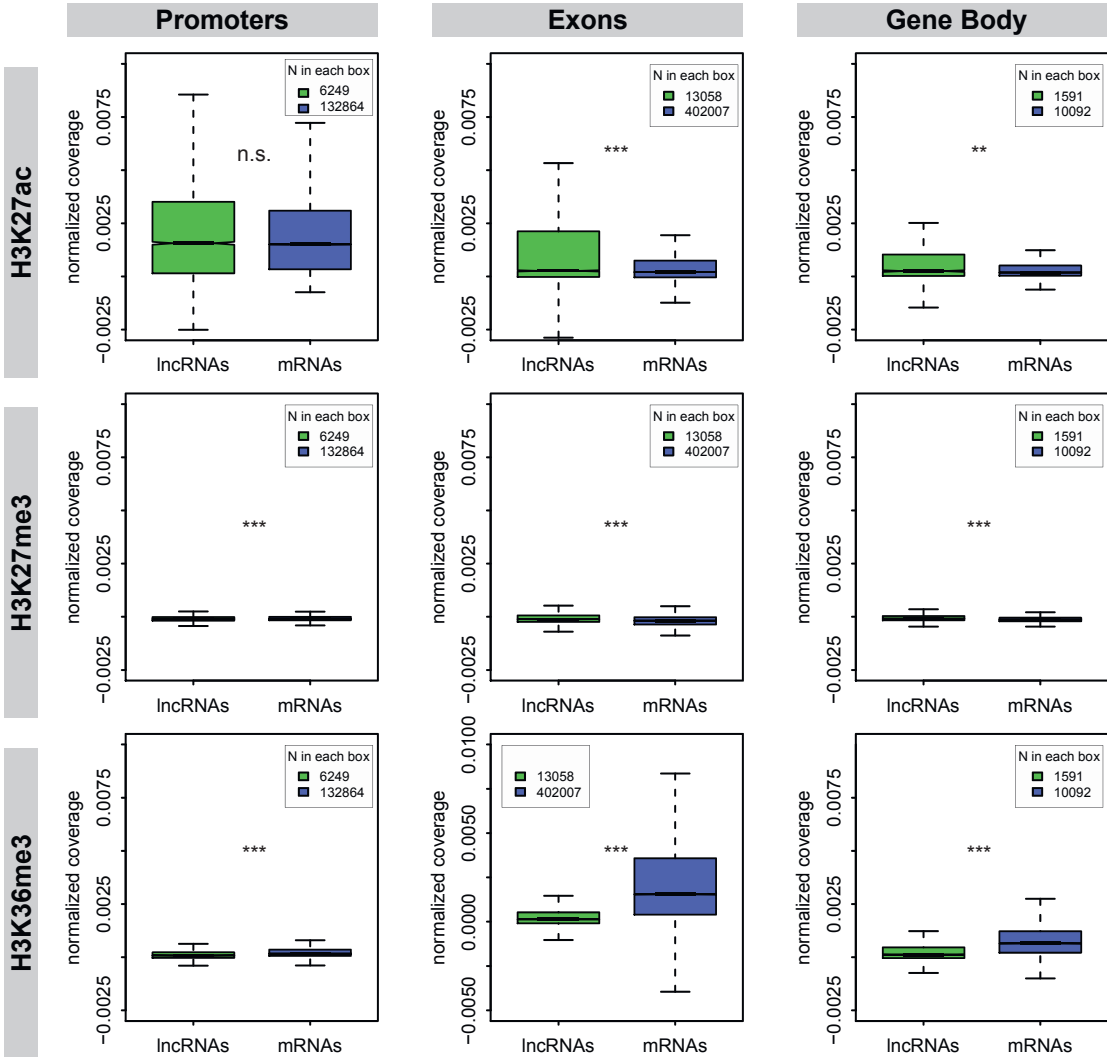


Figure 10. Histone mark coverage of granulocyte lncRNAs (green) and mRNAs (blue): H3K27ac, H3K27me3 and H3K36me3. Remarks to boxplots: Numbers on the right indicate the numbers of transcripts/loci analyzed in each boxplot. The box plots display the full population but p-values are calculated using Mann–Whitney U test with equalized sample size (see Materials and Methods). n.s. not significant ($p>0.01$), $*10^{-5}<p<0.01$, $**10^{-5}<p<10^{-10}$, $***p<10^{-16}$. Median normalized coverage values ($x10^{-3}$) from left to right (lncRNA, mRNA): H3K27ac: promoters: 1.56, 1.51, exons: 0.26, 0.21, loci: 0.26, 0.19, H3K27me3: promoters: -0.11, -0.09, exons: 0.14, 0.15, loci: -0.08, -0.14; H3K36me3: promoters: 0.08, 0.16, exons: 0.01, 0.15, loci: 0.11, 0.65. Outliers are not displayed in the box plots.

Figure 11

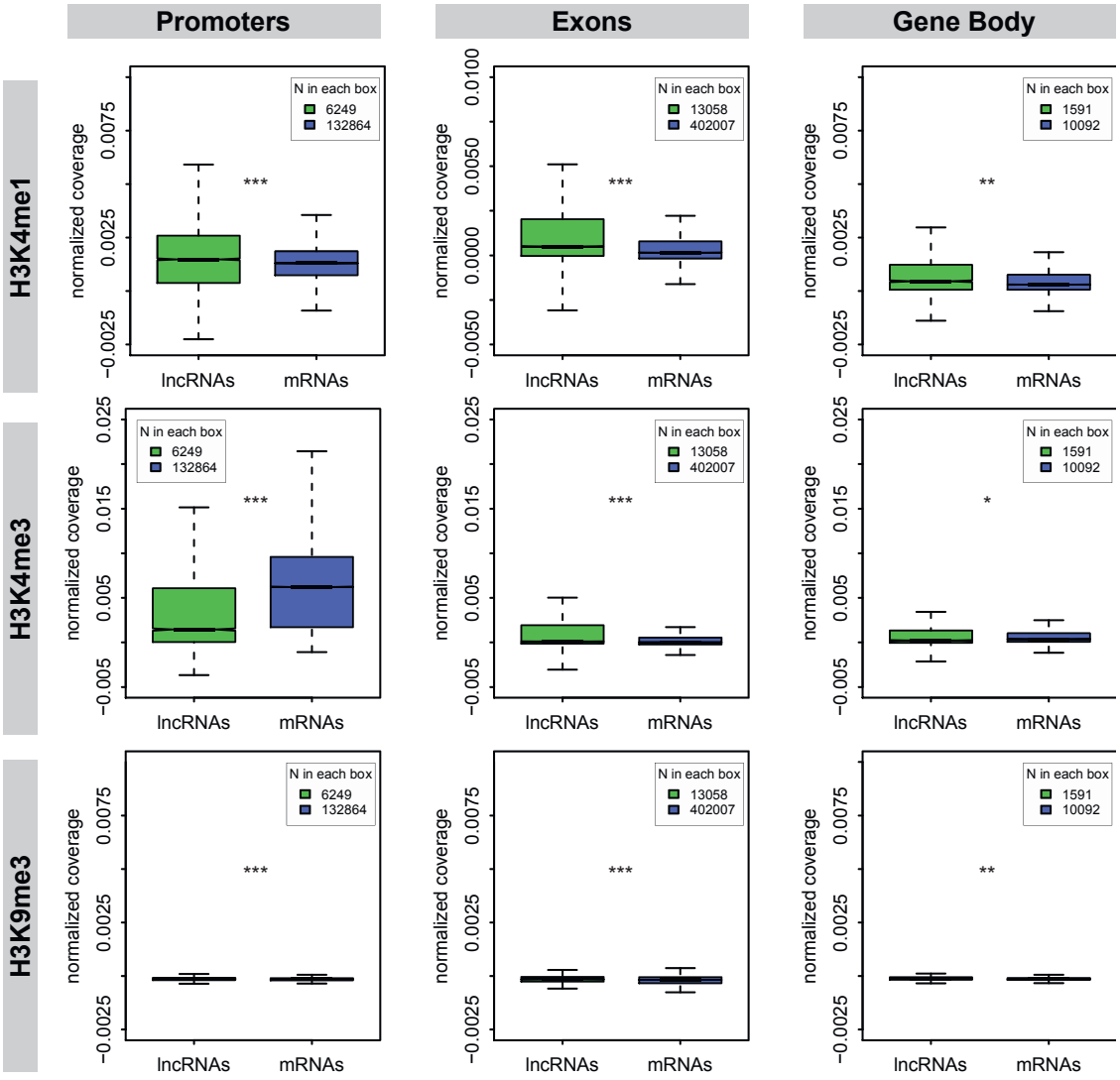


Figure 11. Histone mark coverage of granulocyte lncRNAs (green) and mRNAs (blue): H3K4me1, H3K4me3 and H3K9me3. Remarks to boxplots: Numbers on the right indicate the numbers of transcripts/loci analyzed in each boxplot. The box plots display the full population but p-values are calculated using Mann–Whitney U test with equalized sample size (see Materials and Methods). n.s. not significant ($p>0.01$), $*10^{-5}<p<0.01$, $**10^{-5}<p<10^{-10}$, $***p<10^{-16}$. Median normalized coverage values ($x10^{-3}$) from left to right (lncRNA, mRNA): H3K4me1: promoters: 1.48, 1.30, exons: 0.50, 0.15, loci: 0.46, 0.30; H3K4me3: promoters: 1.45, 6.23, exons: 0.10, 0.01, loci: 0.22, 0.37; H3K9me3: promoters: -0.14, -0.15, exons: -0.15, -0.18, loci: -0.12, -0.14. Outliers are not displayed in the box plots.

2.2.1.2 Granulocyte de novo lncRNAs display different histone modification pattern compared to mRNAs.

Comparing the histone mark coverage of different parts of granulocyte *de novo* annotated lncRNAs and mRNAs revealed notable difference in H3K36me3 mark level on exons

(15-fold median coverage difference), and in the gene body, i.e. locus, (6-fold median coverage difference) of granulocyte lncRNAs and mRNAs (Figure 10). While most comparisons showed highly significant differences, in spite of the efforts in equalizing population sample sizes (see Materials and Methods), the nominal difference in the median histone mark coverage level was in fact neglectable for most of the comparisons. Thus, in order to outline the most significant ones, I also paid attention to the absolute coverage level and the degree of median level differences. The coverage of repressive marks, such as H3K27me3 and H3K9me3, was neglectably low (see median levels in Figure 10 and 11), which was expected because only transcripts *de novo* annotated and therefore expressed in granulocytes, were analyzed. In addition to H3K36me3, I found that the active promoter mark H3K4me3 was dramatically more present on the promoters of mRNAs, than lncRNA promoters (4-fold median coverage difference). In contrast, the enhancer H3K4me1 mark was slightly more prominent on lncRNA promoters (1.14-fold median coverage difference), exons (3.3-fold median coverage difference) and over whole loci (1.5-fold median coverage difference). The higher abundance of H3K4me1 on lncRNA promoters can be explained by the fact that many lncRNAs can initiate from enhancer-like promoters (Marques et al, 2013; Orom et al, 2010). The increased H3K4me1 on lncRNA exons and gene bodies is most probably explained by the fact that lncRNA genes are usually short with few but long exons (Derrien et al, 2012; Kornienko et al, 2016), thus the promoter H3K4 monomethylation overlaps a significant part of the whole exon/gene body. Dramatic decrease of the classical active gene marks, such as H3K4me3 on promoters (Figure 11) and H3K36me3 on gene bodies (Figure 10), for lncRNAs appears surprising, because it was reported that lncRNAs and mRNAs show similar histone modification patterns (reviewed in (Rinn & Chang, 2012).

In order to test if the observed differences are biased to the generally lower lncRNA level, such as, for example, that the extremely lowly expressed lncRNAs would have an undetectable active histone modification marks on their promoters, I split all the analyzed lncRNAs and mRNAs into 5 bins according to their average expression level in our ribosomal depleted RNA-seq data from 7 healthy donors (Figure 12, 13 and 14). This confirmed that differences between lncRNAs and mRNAs are persistent in all bins and are, thus, independent of the lncRNA low expression.

Figure 12

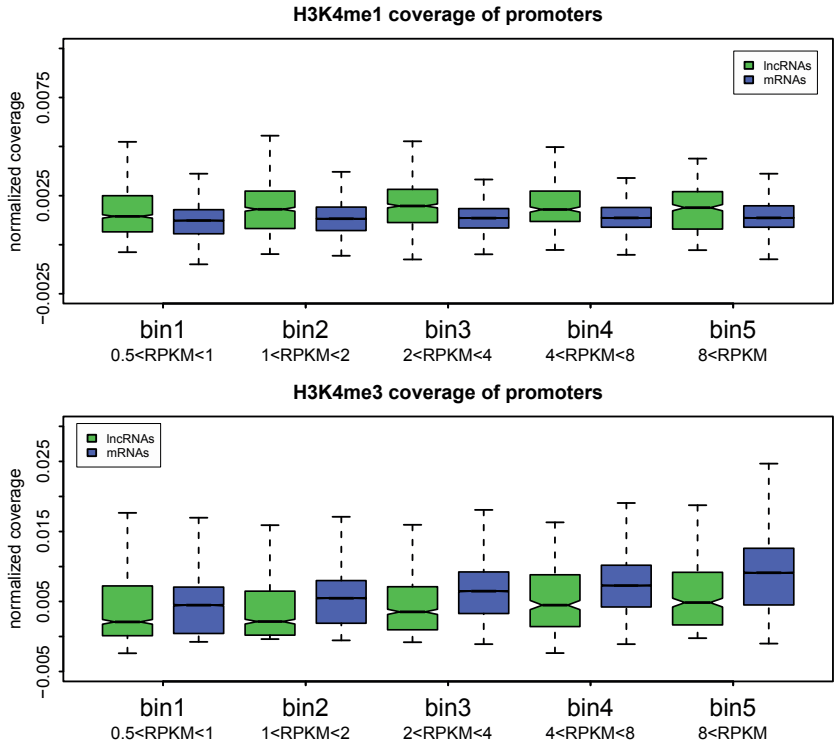


Figure 12. Binned analysis of granulocyte *de novo* lncRNA and mRNA promoter coverage by H3K4me1 and H3K4me3 histone modifications.

Figure 13

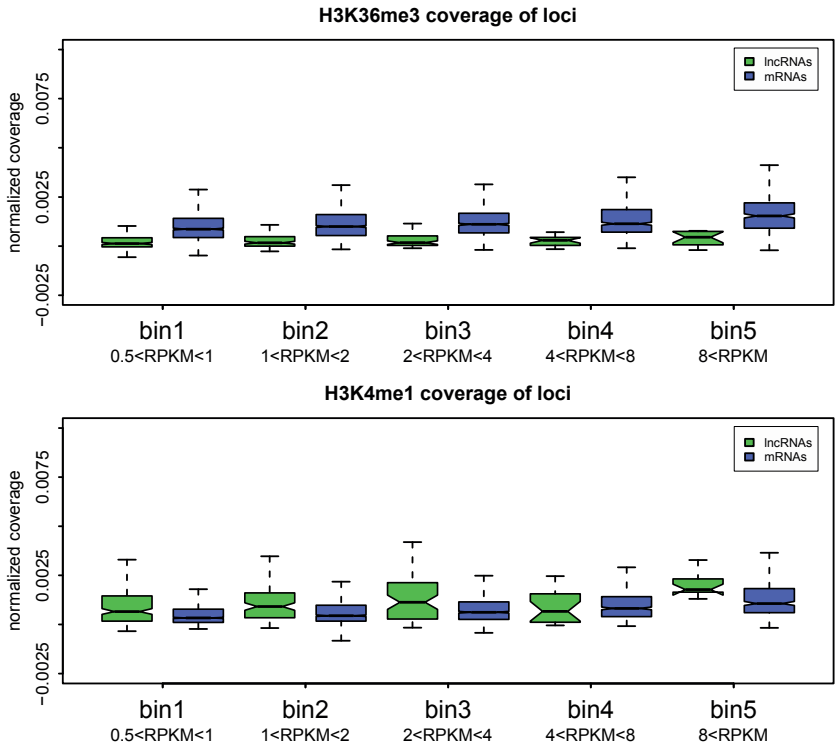


Figure 13. Binned analysis of granulocyte *de novo* lncRNA and mRNA loci coverage by H3K36me3 and H3K4me1 histone modifications.

Figure 14

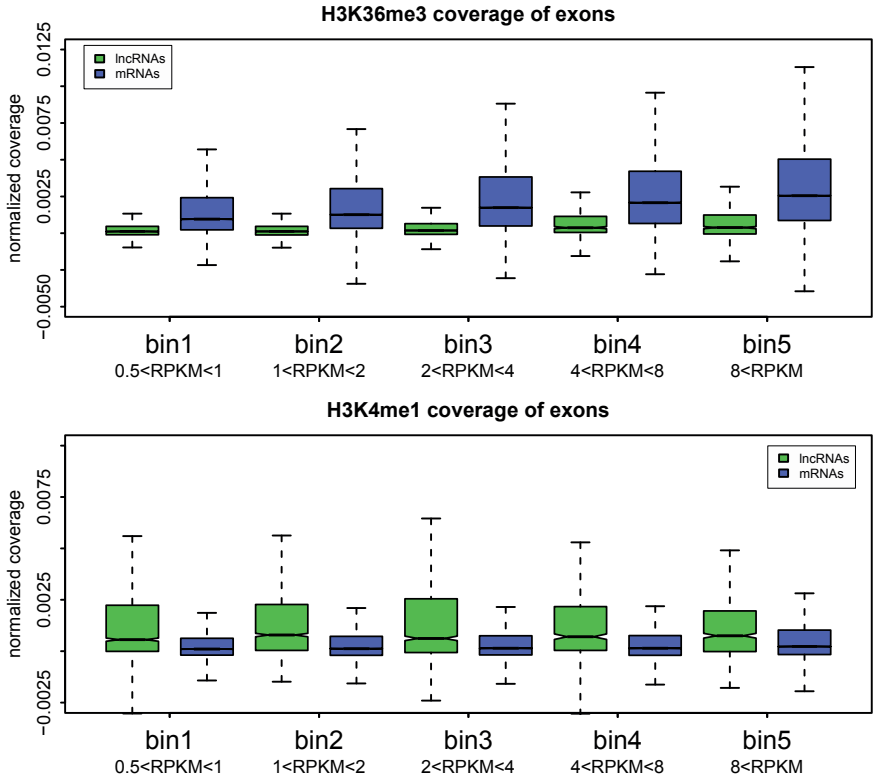


Figure 14. Binned analysis of granulocyte *de novo* lncRNA and mRNA exon coverage by H3K36me3 and H3K4me1 histone modifications.

2.3 Publication 3: "A human haploid gene trap collection to study lncRNAs with unusual RNA biology" (Research Article)

Authors: Aleksandra E. Kornienko, Irena Vlatkovic, Neesen Jürgen, Denise P. Barlow and Florian M. Pauler

Published in RNA Biology (Impact factor 4.974) on 15.12.2015 (online publication).

Article's web page:
http://www.tandfonline.com/doi/abs/10.1080/15476286.2015.1110676?journalCode=krnb20

While thousands of lncRNAs have been discovered in the human genome, only a miniscule part of them has been assigned with a function. LncRNAs display a variety of features that distinguish them from protein-coding genes and make their identification and functional characterization more challenging. New systematic approaches have to be established that preferably allow rapid and massive functional assessment of multiple lncRNAs. We investigated the use of the human haploid knock-out collection (Burckstummer et al, 2013) for studying lncRNAs. This collection consists of several thousand KBM7 clones with gene trap cassettes able to arrest transcription elongation that were inserted in various genomic locations. This collection was shown to be a convenient and ready-to-use tool for studying function of a variety of protein-coding genes (Burckstummer et al, 2013; Carette et al, 2009). We aimed to validate the use of this collection for studying lncRNAs, particularly those with distinct non-mRNA-like features. We focused on a lncRNA "*SLC38A4 down*" previously described in the Ph.D. Thesis of Irena Vlatkovic (Vlatkovic, 2010a). The preliminary analysis of this lncRNA showed that it was an inefficiently spliced, nuclear lncRNA. We used public data from multiple tissues to characterize RNA-biology of this lncRNA (also annotated by RefSeq as *LOC100288798*) and to show that it exceeds its RefSeq annotated length 2-fold, overlaps in antisense orientation downstream *SLC38A4* protein-coding gene and thus should be renamed as the *SLC38A4-AS* lncRNA. We obtained several clones from the human haploid knock-out collection, which harbored gene trap insertion cassettes in the body of *SLC38A4-AS* and thus displayed transcription termination of this lncRNA approximately 3 and 100 kb downstream the transcription start site. We performed RNA-

seq of a set of truncated and control cell lines and showed that truncation of the *SLC38A4-AS* lncRNA results in deregulation of multiple genes *in cis* and *in trans* indicating that this lncRNA is a functional regulator. One of the most strikingly affected genes – *CD9* – is known to be a pluripotency regulator and thus our results indicate that *SLC38A4-AS* lncRNA may play a role in the regulation of differentiation state of the analyzed cell line.

Overall, I, with the help of the co-authors, and supervised by Florian M. Pauler and Denise P. Barlow, have performed this project, which was a branch of my main PhD project (described in Publication 2) and was aimed at studying the function of a lncRNA which resembled the *Airn* lncRNA thoroughly studied in the Barlow laboratory for many years. The manuscript was published in RNA Biology on 15.12.2015 (online publication ahead of print).

Authors' contributions:

"A.E.K., D.P.B. and F.M.P. conceived the study and wrote the manuscript. I.V. discovered the *SLC38A4-AS* lncRNA and performed preliminary experiments charactering this lncRNA. J.N. performed karyotype analysis and FISH. A.E.K and F.M.P performed DNA blots and PCR analyses. A.E.K. performed bioinformatic analysis, cell culture and RNA-seq."

(N.B. Authors' contributions are copied from the manuscript attached below and thus are enclosed in quotes)

RNA BIOLOGY
2016, VOL. 0, NO. 0, 1–25
http://dx.doi.org/10.1080/15476286.2015.1110676

Taylor & Francis
Taylor & Francis Group

RESEARCH PAPER

OPEN ACCESS

A human haploid gene trap collection to study lncRNAs with unusual RNA biology

Aleksandra E Kornienko[a], Irena Vlatkovic[a,b,#], Jürgen Neesen[b], Denise P Barlow[a], and Florian M Pauler[a]

[a]CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Lazarettgasse 14, AKH BT 25.3, 1090 Vienna, Austria; [b]Institute of Medical Genetics, Medical University of Vienna, Währingerstrasse 10, 1090 Vienna, Austria

ABSTRACT

Many thousand long non-coding (lnc) RNAs are mapped in the human genome. Time consuming studies using reverse genetic approaches by post-transcriptional knock-down or genetic modification of the locus demonstrated diverse biological functions for a few of these transcripts. The Human Gene Trap Mutant Collection in haploid KBM7 cells is a ready-to-use tool for studying protein-coding gene function. As lncRNAs show remarkable differences in RNA biology compared to protein-coding genes, it is unclear if this gene trap collection is useful for functional analysis of lncRNAs. Here we use the uncharacterized *LOC100288798* lncRNA as a model to answer this question. Using public RNA-seq data we show that *LOC100288798* is ubiquitously expressed, but inefficiently spliced. The minor spliced *LOC100288798* isoforms are exported to the cytoplasm, whereas the major unspliced isoform is nuclear localized. This shows that *LOC100288798* RNA biology differs markedly from typical mRNAs. *De novo* assembly from RNA-seq data suggests that *LOC100288798* extends 289kb beyond its annotated 3′ end and overlaps the downstream *SLC38A4* gene. Three cell lines with independent gene trap insertions in *LOC100288798* were available from the KBM7 gene trap collection. RT-qPCR and RNA-seq confirmed successful lncRNA truncation and its extended length. Expression analysis from RNA-seq data shows significant deregulation of 41 protein-coding genes upon *LOC100288798* truncation. Our data shows that gene trap collections in human haploid cell lines are useful tools to study lncRNAs, and identifies the previously uncharacterized *LOC100288798* as a potential gene regulator.

Abbreviations: lncRNA, Long non-coding RNA, mRNAs, mRNA (protein coding); RNA-Seq, RNA-sequencing, high throughput sequencing of cDNA ends,

ARTICLE HISTORY
Received 07 August 2015
Revised 13 October 2015
Accepted 16 October 2015

KEYWORDS
Gene trap insertion; genetic truncation; human haploid cell line; lncRNA splicing; KBM7; *LOC100288798*; RNA-seq; RNA biology; *SLC38A4-AS*

Introduction

Long non-coding (lnc) RNAs can regulate gene expression and are abundant in the genomes of various organisms.[1] The human genome has been reported to contain about 60,000 lncRNA genes[2] and an increasing number is suggested to play important roles in cancer and other diseases.[3,4] Moreover, several lncRNAs were reported to serve as disease biomarkers[5,6] and potential drug targets.[7-9] LncRNAs display a wide range of functions from nuclear scaffolding[10] to post-transcriptional mRNA regulation by "sponging" regulatory miRNAs,[11] transcriptional gene activation or repression by binding and guiding histone modifiers to target genes[12,13] and silencing by transcription interference[14] (reviewed in[15]). Apart from the basic difference between the functions of lncRNAs and mRNAs, lncRNAs also display a number of RNA biology features that make their identification and functional studies more challenging than that of protein-coding genes.[16] These features include: low, tissue-specific expression,[17] nuclear localization[18] and inefficient co-transcriptional splicing,[19,20] transcription initiation from repeat rich regions[21] and unusually high isoform heterogeneity.[22]

To date, the majority of functional lncRNA studies have depleted the lncRNA of interest via post-transcriptional knock-down approaches using shRNAs,[23] morpholinos[24] or modified DNA antisense oligos that target nuclear localized transcripts.[25] Based on the atypical RNA biology features described above, these approaches might not be generally suited to study a wide range of lncRNAs. For example, shRNAs are unlikely to target lncRNAs in the nucleus,[26] while morpholinos or antisense oligos might be difficult to design for targeting complex lncRNA loci expressing multiple lncRNA

CONTACT Aleksandra E Kornienko akornienko@cemm.oeaw.ac.at; Florian M Pauler fpauler@cemm.oeaw.ac.at
[#]present address: Max Planck Institute for Brain Research, Max-von-Laue-Strasse 4, 60438 Frankfurt am Main, Germany
GEO accession number: GSE71284
Supplemental data for this article can be accessed on the publishers website.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/krnb.

isoforms. Importantly, lncRNAs that act solely by their transcription will not be affected by post-transcriptional knockdowns.[14] Genetic manipulations might be a more universal approach to interfere with lncRNA function independent of RNA-biology features. These manipulations have become more feasible due to the emergence of fast and simple genome editing technologies such as CRISPR/Cas9.[27] One strategy is the genetic deletion of the whole gene body or the promoter of the lncRNA of interest.[28-31] While this approach is appealing due to its relative simplicity, there is a risk of simultaneous deletion of potential genomic regulatory elements that could be located in the gene body of the targeted lncRNA, which can make the interpretation of the resulting phenotype problematic.[16,32] Therefore genetic insertion of transcriptional terminator sequences, or "gene traps" may be preferable to gene deletions as they are less likely to disrupt regulatory elements.

Gene trap technology is based on the insertion of "truncation cassettes," typically containing polyA signals, shortly after the transcriptional start site (TSS) of the lncRNA to stop RNA Polymerase II transcription and create functional lncRNA "knock-outs". Gene trap mutagenesis has been used extensively in the mouse to identify and study protein-coding genes.[33] Classical gene trap cassettes carry a strong splice acceptor and a reporter protein terminated by a strong polyA signal. This cassette is introduced into the cell line using retroviral vectors that cause random integration into the genome. If the cassette integrates into the gene body of a transcribed gene in the correct transcriptional orientation, transcription will be stopped.[34] An analysis of mouse lines carrying gene trap insertions that had the goal to identify key genes expressed during embryonic development, led to the isolation of the lncRNA called *gene trap locus 2 (Gtl2)* gene.[35] It is also known as *maternally expressed 3 (Meg3)*, since it is exclusively expressed from the maternally inherited allele, a phenomenon known as genomic imprinting.[36] *Gtl2/Meg3* was shown to be functional in mouse development[37,38] and human disease.[39] Subsequently a targeted approach was used to introduce polyA signals from rabbit β globin or simian virus 40 to truncate the imprinted *Airn, Kcnq1ot1* and *Ube3a-as* lncRNAs in mice, as occurs in gene trap truncations. These approaches successfully stopped lncRNA transcription and identified these lncRNAs as transcriptional regulators of developmentally important protein-coding genes.[40-43] The advent of genome editing tools such as zinc finger nucleases opened the possibility to use similar approaches also for human cells. In this way polyA containing truncation cassettes were targeted at the abundantly expressed *MALAT1* lncRNA causing efficient truncation in a number of human cell lines.[44]

Insertion of a truncation cassette may interrupt *cis*-acting genetic elements, and although this is notably less likely than with gene body deletions, it should be controlled for. Such controls include insertion of the truncation cassette at different sites, creating lncRNA truncations of different lengths, or the use of non-functional truncation cassette insertions.[32] An important advantage of the gene trap approach is the possibility to restore lncRNA transcription by removing the stop cassette.[45] However, restoration of lncRNA function will only be possible if continuous expression is required for function.[32,46] Taken together, this indicates that the truncation of lncRNAs is a useful tool to study their function in both mouse and human, and in particular gene trap insertion is a well-controlled high-throughput method to achieve this.

While tools to perform genetic manipulations in mouse and human systems are becoming faster and simpler, the creation of a human cell line carrying a lncRNA truncation may still require optimization and thus is time consuming and resource intensive. Therefore it would be beneficial to use existing lncRNA knockout resources to rapidly investigate a lncRNA of interest. Such a resource was reported for protein-coding genes as the "Human Gene Trap Mutant Collection".[45] This library is comprised of a collection of monoclonal cell lines that carry an insertion of a gene trap cassette in the gene body of a large number of genes.[45] The cell line used to establish this resource is a nearly haploid (except for chromosome 8) malignant myeloid lineage cell line called KBM7.[47] As most chromosomes are present in only one copy, the integration of a gene trap cassette results in a full knock-out in KBM7 cells. Since the creation of this gene trap collection did not select for a particular type of genomic locus, it contains cell lines with gene trap cassettes inserted into protein-coding genes, as well as into transcribed non-coding regions, including various annotated lncRNAs (visit https://opendata.cemm.at/barlowlab/ for the location of all cassettes). Thus, the KBM7 "Human Gene Trap Mutant Collection" could represent a massive ready-to-use collection of lncRNA knockouts that may be useful for rapidly assessing human lncRNA function. Importantly, efficiency of a gene trap depends on splicing from a neighboring exon of the "trapped" gene to the gene trap cassette.[34] In the above described case of *Gtl2/Meg3* efficient splicing was expected as this lncRNA produces a number of spliced isoforms.[48] While "Human Gene Trap Mutant Collection" has been proven to efficiently stop transcription of

protein-coding genes, the usefulness of this approach to study lncRNAs is unclear, since it was shown that many of them are inefficiently spliced or completely unspliced.[19]

In this study we aimed to close this knowledge gap and test if "Human Gene Trap Mutant Collection" can be successfully used for studying lncRNAs, even the inefficiently spliced ones. For this purpose we focused on a lncRNA, that was identified in a tiling array based study to be close to the *SLC38A4* protein-coding gene and named "*SLC38A4-down*".[49] It is noteworthy that mouse *Slc38a4* shows imprinted expression in extra-embryonic, embryonic and adult tissues[50] as well as in cell culture cells.[51] No lncRNA has been reported to be involved in regulating *Slc38a4* imprinted expression which is, to date, considered a solo imprinted gene (http://igc.otago.ac.nz). Although *SLC38A4* was not reported to show imprinted expression in human, the identification of *SLC38A4-down* lncRNA close to the *SLC38A4* gene allowed the possibility that this lncRNA might be involved in transcriptional regulation of *SLC38A4*. *SLC38A4-down* lncRNA was predicted from its expression profile, that lacked exon peaks, to be mainly unspliced and was also shown to be nuclear-localized.[49] These features make it an unsuitable target for a post-transcriptional knock-down approach. Importantly, we identified a number of gene trap insertions in the gene body of this lncRNA in the "Human Gene Trap Mutant Collection" in the correct transcriptional orientation, which allowed us to use this lncRNA as a model in our study. We first identified that *SLC38A4-down* corresponds to the *LOC100288798* lncRNA annotated by NCBI RNA reference sequences collection (RefSeq[52]). Using publicly available RNA-seq data from various tissues and cellular fractions we found the *LOC100288798* lncRNA to be ubiquitously expressed, inefficiently spliced and polyadenylated. Unspliced isoforms are retained in the nucleus, while minor spliced isoforms are exported to the cytoplasm. We also extended the annotation of this lncRNA by showing that it is twice as long as the annotated version, as it is transcribed over 500 kilobases (kb) and overlaps the *SLC38A4* protein-coding gene in multiple tissues. Thus we suggest renaming it *SLC38A4-AS* lncRNA in accordance with recent lncRNA nomenclature guidelines.[53] We then obtained three independent KBM7 clones harboring gene trap cassettes in the body of *SLC38A4-AS* predicted to stop transcription 3kb and 100kb downstream of its transcription start. RNA sequencing (RNA-seq) of control and *SLC38A4-AS* truncated cell lines showed that *SLC38A4-AS* was efficiently truncated, which resulted in genome-wide gene expression changes. We applied further stringent filtering to identify a small list of the most plausible *SLC38A4-AS* targets. Based on this data we conclude that lncRNA truncations available in the "Human Gene Trap Mutant Collection" are useful to study lncRNAs, making this resource a valuable tool for studying lncRNA function in a human system. In order to maximize the usefulness of this data for the scientific community we provide a UCSC genome browser hub to display all the RNA-Seq data as well as the information on gene trap insertion sites presented in this paper (https://opendata.cemm.at/barlowlab/).

Results

LOC100288798 is a ubiquitously expressed, inefficiently processed lncRNA

LOC100288798 lncRNA is annotated by several reference gene databases including RefSeq[52] and GENCODE v19 (http://www.gencodegenes.org/releases/19.html,[54]) as a 269kb lncRNA on human chromosome 12 (Fig. 1A). *LOC100288798* lncRNA was also identified by RNA-seq based human lncRNA annotation studies such as Cabili et al[17] and MiTranscriptome[2] (Fig. 1A). It is an intergenic lncRNA that initiates from its own CpG island (CpG: 106) and is located between the *SLC38A2* and *SLC38A4* protein-coding genes (Fig. 1A). Despite the 35 spliced expressed sequence tags (ESTs) mapped to this locus (Human ESTs That Have Been Spliced public track at UCSC Genome Browser), *LOC100288798* remains an uncharacterized lncRNA.

We characterized this lncRNA using publicly available human RNA-seq data. We first asked which tissues and cell types express *LOC100288798* lncRNA using polyA+ enriched and total (rRNA depleted) RNA-seq data from 34 healthy primary tissues and cell types as well as 4 normal and 3 malignant cell lines originating from different studies (total of 41 different cell types, 5 of which were replicated twice giving the total of 46 samples, Table S1A, Methods). We downloaded the raw RNA-seq data, aligned it with STAR[55] and obtained an average of 186 million uniquely mapped reads per sample (ranging from 16 to 371 million reads, Table S1A). We next calculated expression levels of *LOC100288798* lncRNA and its neighboring *SLC38A2* and *SLC38A4* genes by calculating average RPKMs of RefSeq annotated spliced isoforms (Methods). Fig. 1B shows the obtained expression profile in the 46 analyzed samples. This shows that *SLC38A2* is highly expressed (RPKM>9) in

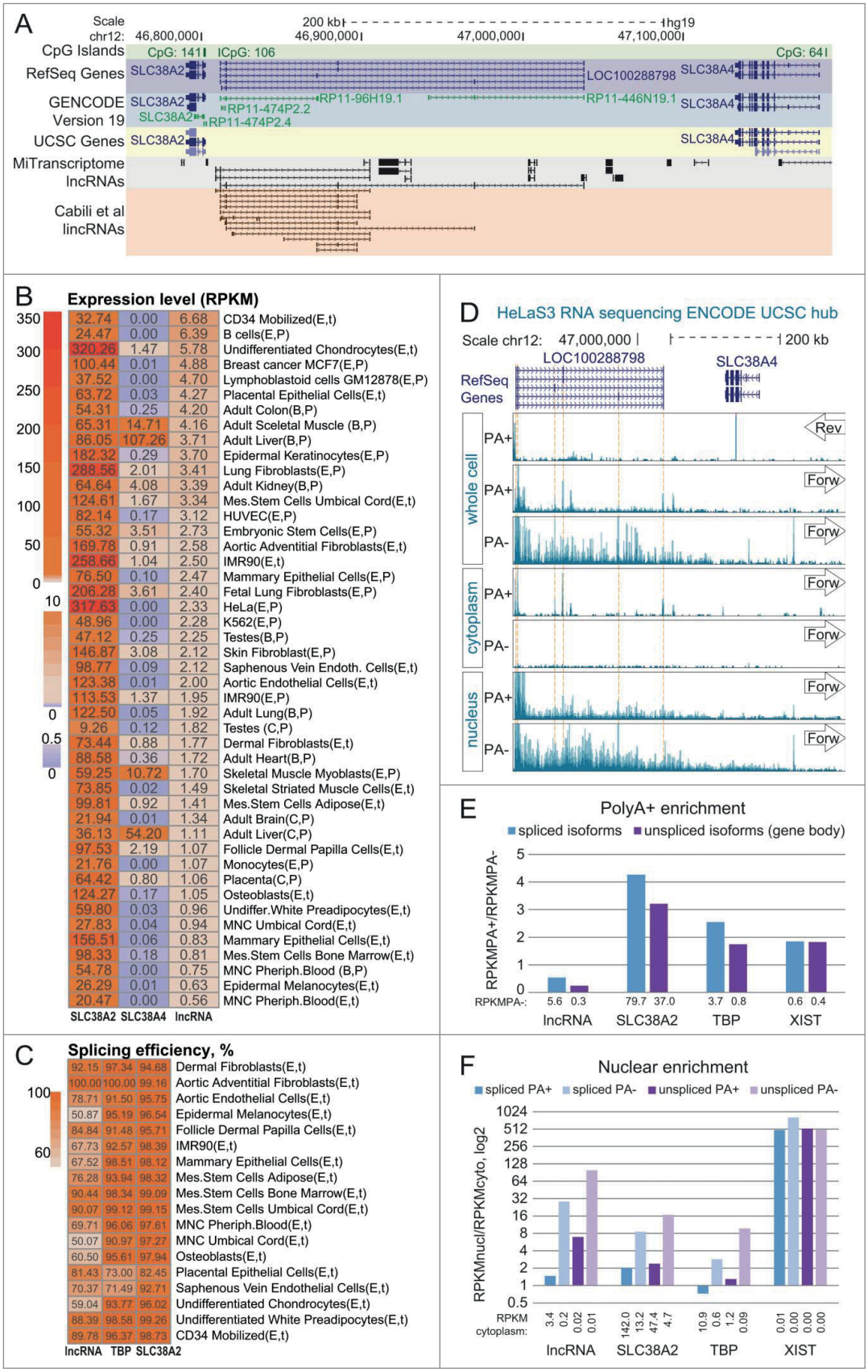


Figure 1. (For figure legend, see page 5.)

every analyzed sample and its ubiquitous expression is known (http://www.proteinatlas.org/ENSG00000134294-SLC38A2/tissue). In contrast, *SLC38A4* is expressed (RPKM > 0.5) in just 18/46 samples (which corresponds to 15/41 different cell/tissue types) with highest expression in liver and skeletal muscle, consistent with previous observations (The Human Protein Atlas: http://www.proteinatlas.
org/ENSG00000139209-SLC38A4/tissue, Expression Atlas: http://www.ebi.ac.uk/gxa/genes/ENSG00000139209). Similar to *SLC38A2*, the *LOC100288798* lncRNA is expressed (RPKM>0.5) in all analyzed samples. Notably, the highest *LOC100288798* lncRNA expression level, achieved in CD34 cells, is 48 fold lower than the highest expression level of *SLC38A2* and 16 fold lower than that of *SLC38A4*, consistent with previous observations that lncRNAs are generally lower expressed than protein-coding genes.[17] We next asked if *LOC100288798* lncRNA expression showed any correlation with the 2 nearby genes, since it is known that some lncRNAs can regulate their nearby protein-coding genes.[13,40] Although *LOC100288798* lncRNA and its closest gene *SLC38A2* were both ubiquitously expressed, they did not show correlation in expression level (Pearson correlation = 0.17, 46 samples). This, together with the fact that their transcription start sites are separated by 11kb and located in 2 separate CpG islands, indicates that these 2 genes initiate from independent promoters, and while they seem to belong to the same transcription network, the regulation of their expression level may be independent. *LOC100288798* lncRNA and *SLC38A4* showed a striking difference in cell type expression profile and no correlation in expression among the tested tissues and cell types (Pearson correlation = 0.07, 46 samples), which indicates independent transcriptional regulation. When we analyzed correlation only in tissues that express both *LOC100288798* lncRNA and *SLC38A4*, correlation between these 2 genes was still negligible (Pearson correlation = 0.11, 18 samples), although the small number of samples may impede the correlation analysis. In summary, we found that *LOC100288798* is a ubiquitously, but lowly expressed lncRNA displaying no striking correlation with the expression of its neighboring protein-coding genes.

We next characterized the efficiency of *LOC100288798* lncRNA splicing as it was previously reported that lncRNAs show reduced co-transcriptional splicing when compared to mRNAs.[19] We used publicly available total RNA-seq data (Table S1A) from 18/41 of the above described different cell types and estimated splicing efficiency for *LOC100288798* lncRNA and 2 protein-coding genes *TBP* and *SLC38A2* that were expressed in the same cell types. We calculated the average splicing efficiency of all

Figure 1. (see previous page) RefSeq *LOC100288798* is a ubiquitously expressed, inefficiently processed lncRNA (A) Overview of the genomic locus. UCSC Genome Browser screenshot – from top to bottom: CpG island annotation, RefSeq Genes annotation, GENCODE v19 annotation, UCSC Genes annotation, MiTranscriptome lncRNA transcripts,[2] Cabili et al lincRNA transcripts[17].(B) *LOC100288798* is a ubiquitously expressed lncRNA. Heat map shows expression level of *SLC38A2*, *SLC38A4* and *LOC100288798* (marked as "lncRNA" throughout the figure) in multiple tissues and cell types. Letters in brackets after the name of each sample indicate the source and the type of RNA-seq (see Table S1A for details of abbreviations). Expression levels of *SLC38A4* and *LOC100288798* were calculated as average RPKMs of RefSeq isoforms (*SLC38A2* – 1 isoform: NM_018976, *SLC38A4* – 2 isoforms: NM_018018 and NM_001143824, *LOC100288798* – 5 isoforms: NR_125377, NR_125378, NR_125379, NR_125380, and NR_125381), values are displayed inside each cell. Heat map color legend is displayed on the left. (C) *LOC100288798* lncRNA is variably spliced in different tissues. Heat map shows splicing efficiency (Methods) of *LOC100288798* and 2 protein-coding genes *TPB*, *SLC38A2* (well-spliced ubiquitously expressed protein coding gene controls) in publicly available total RNA-seq data (Table S1A). Calculated splicing efficiency is displayed inside each cell. Heat map color legend is displayed on the left. (D) Visual inspection of ENCODE HeLa RNA-seq of various cell and RNA fractions suggests that *LOC100288798* is an inefficiently processed lncRNA. From top to bottom: Chromosome position; RefSeq annotation; ENCODE HeLa RNA-seq sequencing data. RNA-seq data is displayed using the public ENCODE RNA-seq (CSHL) hub in the UCSC browser (only Replicate 2 from 2 replicates available at ENCODE RNA-seq (CSHL) hub is displayed). From top to bottom: PolyA+ RNA-seq of the whole cell Reverse and Forward strand show absence of *SLC38A4* expression from the reverse strand and visible expression from the forward strand corresponding to *LOC100288798*. Dashed orange lines indicate chromosome positions of RefSeq annotated exons of *LOC100288798*. Comparison of signal intensities between polyA+ and polyA- indicates *LOC100288798* is inefficiently spliced as it appears more abundant in polyA- fraction. Cytoplasm RNA-seq indicates that only spliced and polyadenylated *LOC100288798* transcripts can be exported to the cytoplasm (compare peaks in polyA+ and no peaks in polyA-). Nuclear RNA-seq indicates nuclear enrichment of *LOC100288798* unspliced form (compare nucleus polyA- to cytoplasm polyA-). RNA-seq tracks are displayed with the default ENCODE RNA-seq (CSHL) hub scale (range - from 0 to 100). (E) PolyA+ enrichment. Bar plot shows PolyA+ enrichment (calculated as the ratio between RPKM in PolyA+ and PolyA- RNA fractions) of the 4 indicated genes in HeLa cells (ENCODE RNA-seq data). RPKMs and consequently PolyA+ enrichment were calculated for spliced isoforms (RPKM over exons, blue bars) and unspliced isoforms (RPKM over whole gene body, purple bars) of the 4 genes. PolyA+ enrichment is a relative value, therefore we indicated the absolute RPKM values of spliced and unspliced isoforms in PolyA- fraction below each respective bar. (F) Nuclear enrichment. Bar plot shows nuclear enrichment (calculated as the ratio between RPKM in nuclear and cytoplasmic fractions) of the 4 indicated genes in HeLa cells (ENCODE RNA-seq data). RPKMs and consequently nuclear enrichment were calculated for spliced isoforms (RPKM over exons, blue bars) and unspliced isoforms (RPKM over whole gene body, purple bars) of the 4 genes in PolyA+ (darker bars) and PolyA- (lighter bars) fractions. Nuclear enrichment is a relative value, therefore we indicated the absolute RPKM values in cytoplasmic fraction below each respective bar.

unique splice sites from all isoforms of the analyzed gene (Fig. 1C) by calculating RPKMs of exonic and intronic 45bp regions surrounding the splice site (Methods). As expected, both protein-coding genes showed high splicing efficiency with an average of 93.0% (*TBP*) and 96.5% (*SLC38A2*) among analyzed cell types. Importantly only 2 (for *TBP*) and one (for *SLC38A2*) cell types showed splicing efficiencies of less than 90%. The result was different for the *LOC100288798* lncRNA. Here average splicing efficiency was 76.0%, with 14/18 cell types showing splicing efficiency of less than 90% and 7 - lower than 70%. It is noteworthy that low splicing efficiencies are not restricted to low expression levels. For example undifferentiated chondrocytes (59% splicing efficiency) and IMR90 cells (68% splicing efficiency) are in the top 25% and top 50% highest expressing tissues for the *LOC100288798* lncRNA (Fig. 1B). This indicates that *LOC100288798* lncRNA is less well spliced compared to protein-coding genes, and that splicing is variable in different cell types.

It has been reported that lncRNAs tend to be nuclear localized,[18,56] and that nuclear export depends on the addition of a 3' polyA tail, which is connected to splicing.[57] To investigate the processing of *LOC100288798* lncRNA we used publicly available ENCODE RNA-seq data from nuclear, cytoplasmic, as well as whole cell fractions (Table S1B). Importantly, the RNA from each cell fraction was further divided into polyA enriched (polyA+) and polyA depleted (polyA-), thus providing a source of information about the polyadenylation and cellular localization of *LOC100288798* lncRNA spliced/polyadenylated as well as unspliced isoforms. We first visually inspected the RNA-seq signal obtained from HeLa cells in the *LOC100288798/SLC38A4* region using the ENCODE (CSHL) RNA-seq hub in the UCSC browser (Fig. 1D). The *SLC38A4* protein-coding gene is not expressed in whole cell polyA+ RNA-seq as indicated by the absence of RNA-Seq signal over exons on the reverse strand (Fig. 1D, whole cell, top box, Arrow marked 'Rev'), consistent with our expression calculation (Fig. 1B, RPKM of *SLC38A4* = 0.00). In contrast, the forward strand showed abundant RNA-seq signals over *LOC100288798* lncRNA exons in polyA+ and over the whole gene body in polyA- RNA-seq data. Interestingly, the signal intensities in polyA+ and polyA- data were comparable confirming inefficient splicing of *LOC100288798* lncRNA (Fig. 1D, whole cell, middle and bottom box, Arrow marked 'Forw'). In the cytoplasmic fraction, only spliced and polyadenylated isoforms of *LOC100288798* lncRNA were detectable as RNA-seq signal over exons in the polyA+, but not in the polyA- fraction (Fig. 1D, cytoplasm). In the nuclear fraction, stronger RNA-seq signals were detectable over the *LOC100288798* lncRNA gene body in polyA- than in the polyA+ faction, and no clear enrichment of exonic signals was visible. This indicated that spliced isoforms of *LOC100288798* lncRNA were exported to the cytoplasm, whereas mainly unspliced isoforms were retained in the nucleus.

To quantify this visual analysis we calculated RPKM values for *LOC100288798* lncRNA and 2 control protein-coding genes, *SLC38A2* and *TBP,* as well as for the *XIST* lncRNA, which is known to be polyadenylated, nuclear localized and well spliced.[58] We first estimated the efficiency of polyadenylation by calculating the ratio of RNA-seq signal in the PolyA+ fraction over the PolyA- fraction (RPKMPA+/RPKMPA-, Fig. 1E). We observed that all the 3 control genes, which are known to be polyadenylated, show ratios of ~2-4 for both unspliced (whole gene body, purple bars) and spliced (blue bars) isoforms, indicating efficient polyadenylation of these transcripts. Spliced and unspliced isoforms of *LOC100288798* lncRNA showed ratios smaller than 1, indicating inefficient polyadenylation of *LOC100288798* lncRNA (Fig. 1E, lncRNA). We next assessed the efficiency of cytoplasmic export by calculating the ratio of RNA-seq signals in the nuclear over the cytoplasmic cell fraction for both PolyA+ and PolyA- RNA-seq datasets (Fig. 1F). As expected, PolyA- fraction showed high ratios for both spliced and unspliced isoforms of the 4 tested genes, indicating nuclear enrichment of unprocessed isoforms (Fig. 1F, light blue and light purple bars). In contrast, the pattern of nuclear enrichment of polyadenylated spliced and unspliced isoforms differed notably between the analyzed genes (Fig. 1F, blue and purple bars). While spliced and polyadenylated *XIST* isoforms were almost exclusively present in the nucleus (ratio: ~500), similar processed isoforms of the protein-coding genes *SLC38A2* and *TBP* showed low ratios, indicating no nuclear enrichment (Fig. 1F). Consistent with our conclusions from visual inspection, spliced isoforms of *LOC100288798* lncRNA were exported to the cytoplasm and showed low ratios similar to the analyzed protein-coding genes (RPKM of spliced isoforms in the polyadenylated cytoplasmic fraction = 3.4, while RPKM of spliced isoforms in the polyadenylated whole cell fraction = 2.3, Fig. 1B). Interestingly, unspliced isoforms of *LOC100288798* lncRNA showed high ratios, indicating nuclear enrichment. Similar profiles were observed for *LOC100288798* lncRNA in 4 other analyzed cell lines (Fig. S1, Table S1B). In summary, this analysis showed that *LOC100288798* lncRNA is inefficiently polyadenylated in comparison to *SLC38A2*, *TBP* and *XIST*. Whereas the small fraction of polyadenylated *LOC100288798* lncRNA isoforms is exported to the cytoplasm, the major fraction consisting of unspliced

isoforms is highly enriched in the nucleus. Therefore we show that *LOC100288798* lncRNA polyadenylation and nuclear enrichment profiles are distinct from both *XIST* lncRNA and protein-coding genes.

De novo assembly of LOC100288798 exon structure identifies overlap with SLC38A4

Visual inspection of the RNA-seq data indicated that *LOC100288798* transcription extends over the downstream *SLC38A4* gene (see continuous RNA-seq signal in Fig. 1D), in spite of RefSeq annotating the 3' end of *LOC100288798* 112kb upstream from *SLC38A4* (Fig. 2 top). Interestingly, human spliced ESTs annotated continuous spliced transcripts overlapping *SLC34A4* (Fig. 2). We next aimed to fully annotate *LOC100288798* using publicly available RNA-seq data from multiple cell types. We limited this analysis to reads aligned to a 1 Mega base pairs (Mb) region (chr12:46,500,000-

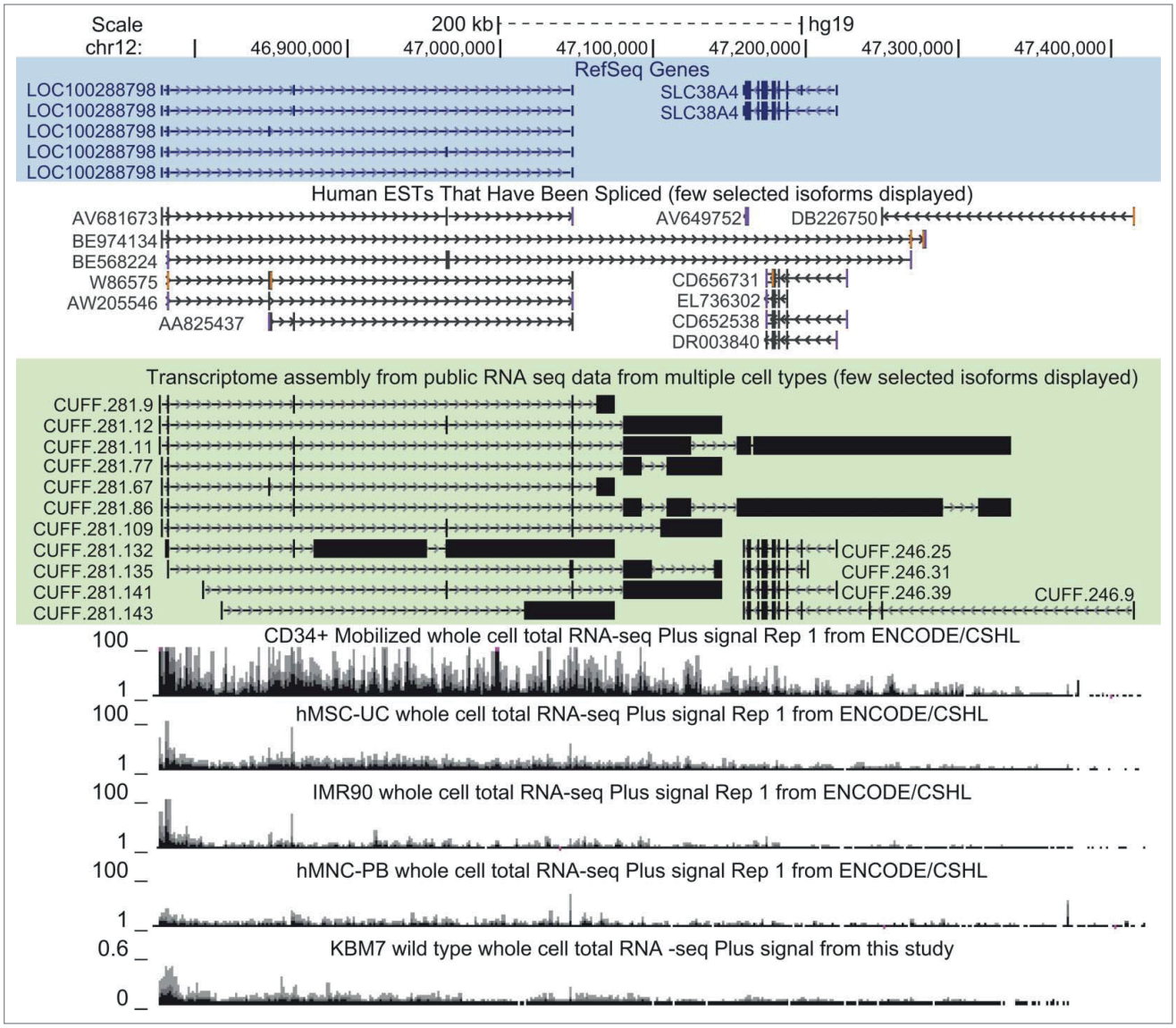


Figure 2. *LOC100288798* exon structure assembly from various tissues extends its annotation to over 500kb overlapping *SLC38A4*.UCSC Genome Browser screen shot of the studied locus (chr12:46,772,500-47,422,500). From top to bottom: Chromosome position and the scale; RefSeq gene annotation (all annotated isoforms are displayed), spliced human ESTs (12/35 ESTs displayed), transcriptome assembly of the locus obtained in this study (Results, Methods). Note that only selected transcripts are shown (11/167 *de novo* isoforms of *LOC100288798* and 4/43 *de novo* isoforms of *SLC38A4*), and that both EST and transcriptome assembly data reveal extension of *LOC100288798* to over 500kb in length. RNA-seq tracks from ENCODE/CSHL UCSC hub with the titles containing cell type name, RNA-seq type and transcriptional orientation are displayed below. Only total whole cell RNA-seq is displayed. Bottom: normalized RNA-seq signal from wild type human haploid KBM7 cell lines (merged data from 2 wild type clones sequenced in this study, Methods). For all RNA-seq tracks: only forward strand (Plus Signal) is displayed.

47,500,000) around *LOC100288798*. We extracted reads from each of the 46 aligned RNA-seq samples used in Fig. 1B (polyA+ as well as ribosomal depleted total RNA-seq) and performed *de novo* assembly using the Cufflinks software.[59] Thus, we obtained 46 assemblies, which we merged using Cuffmerge software[59] to create an integrative *de novo* annotation of the investigated region (see Fig. 2 for selected isoforms and Table S1Cfor all the isoforms annotated in the region). Importantly, we identified exon models that share exons with *LOC100288798* lncRNA and overlap the *SLC38A4* protein coding gene, indicating that *LOC100288798* is a 558kb long lncRNA (chr12:46777455-47335067, see CUFF.281.86 in Fig. 2 and Table S1C). Visual inspection of the *LOC100288798* RNA-seq signal in cell types ranging from the highest expressing (CD34 cells, RPKM=6.68) to lowest expressing (MNC Peripheral blood, RPKM=0.56), showed that extended transcription persists independently of expression level (Fig. 2). Therefore *LOC100288798* lncRNA is consistently overlapping the *SLC38A4* protein-coding gene and should be renamed as *SLC38A4-AS* according to the recently suggested nomenclature.[53] As this nomenclature also appears more intuitive we have used it for the remainder of this study.

Gene trap insertion in the haploid human KBM7 efficiently truncates SLC38A4-AS lncRNA

Although visual inspection of RNA-seq and exon model assembly suggested that *SLC38A4-AS* lncRNA is a single lncRNA gene it is possible that this was an artifact resulting from multiple short overlapping lncRNAs. To address this issue we used the haploid KBM7 cell line for which a collection of gene trap insertion clones was readily available.[45] We first confirmed that *SLC38A4-AS* was expressed in wildtype KBM7 cells and found it well expressed over the predicted length by visual inspection of RNA-Seq data performed in this study (Fig. 2 bottom). Next, we identified 3 cell lines from the publicly available KBM7 gene trap collection where independent insertion events inserted gene trap cassettes in the correct orientation into the gene body of *SLC38A4-AS* (Table 1). Two of these cell lines were predicted to stop *SLC38A4-AS* transcription at 2,904bp (3kb1 and 3kb2, Fig. 3A), and one cell line at 103,958bp (100kb) downstream of the RefSeq annotated transcription start. To create biological replicates of the single 100kb insertion cell line we recovered 2 batches of this cell line from frozen stocks and cultured them in parallel (100kb1, 100kb2, Methods, Fig. 3A). The production of KBM7 gene trap insertion cell lines is a multi-step procedure including infection of cells with the gene trap cassette, fluorescent activated cell sorting (FACS) and clonal expansion to obtain monoclonal cultures. Also different people may have handled different cell lines. These factors are possible sources of gene expression differences, so we controlled for these factors using multiple control cell lines. First, we obtained 3 different KBM7 cell lines that had not undergone the gene trap insertion procedure but were handled by different people and had different passage numbers (wild type: WT1, WT2, WT3, Fig. 3A). Second, to control for potential effects of the gene trap insertion procedure, we obtained 2 cell lines with gene trap insertions not in *SLC38A4-AS*, but in the *HOTTIP* lncRNA gene body of which one was predicted to stop *HOTTIP* lncRNA and one was not, based on mapping cassette insertion orientation (C1 and C2, Table 1, Fig. 3A). To eliminate further batch effects from handling cells and preparing RNA and RNA-Seq libraries, all cell lines were obtained as frozen stocks and recovered, cultured and harvested at the same time by one person. Similarly one person performed RNA extraction and library preparation.

After recovery we cultured the cell lines for 8 days and 2 passages. We measured the cell size prior to splitting and harvesting (Methods) and noticed that the C1 and 3kb2 cell lines showed increased peak cell size (Fig. 3B). It has been reported previously that cell size increases with ploidy[60] and therefore this result indicated that these KBM7 cell lines were not haploid. We then harvested the cells using 20 million cells for DNA isolation and 100 million cells for RNA isolation. As a further test for ploidy we measured the DNA amount obtained from the 20 million cells. Consistent with the cell size

Table 1. Stop cassette insertions overview.

Control cell lines that underwent cassette insertion				
name of the sample		position of the insertion (hg19)		strand of the gene trap
C2	chr7	27240807	27240808	−
C1	chr7	27244000	27244001	+
SLC38A4-AS truncation cell lines				
name of the sample		position of the insertion (hg19)		strand of the gene trap
3kb1	chr12	46780363	46780364	+
3kb2	chr12	46780363	46780364	+
100kb1 and 100kb2	chr12	46881417	46881418	+

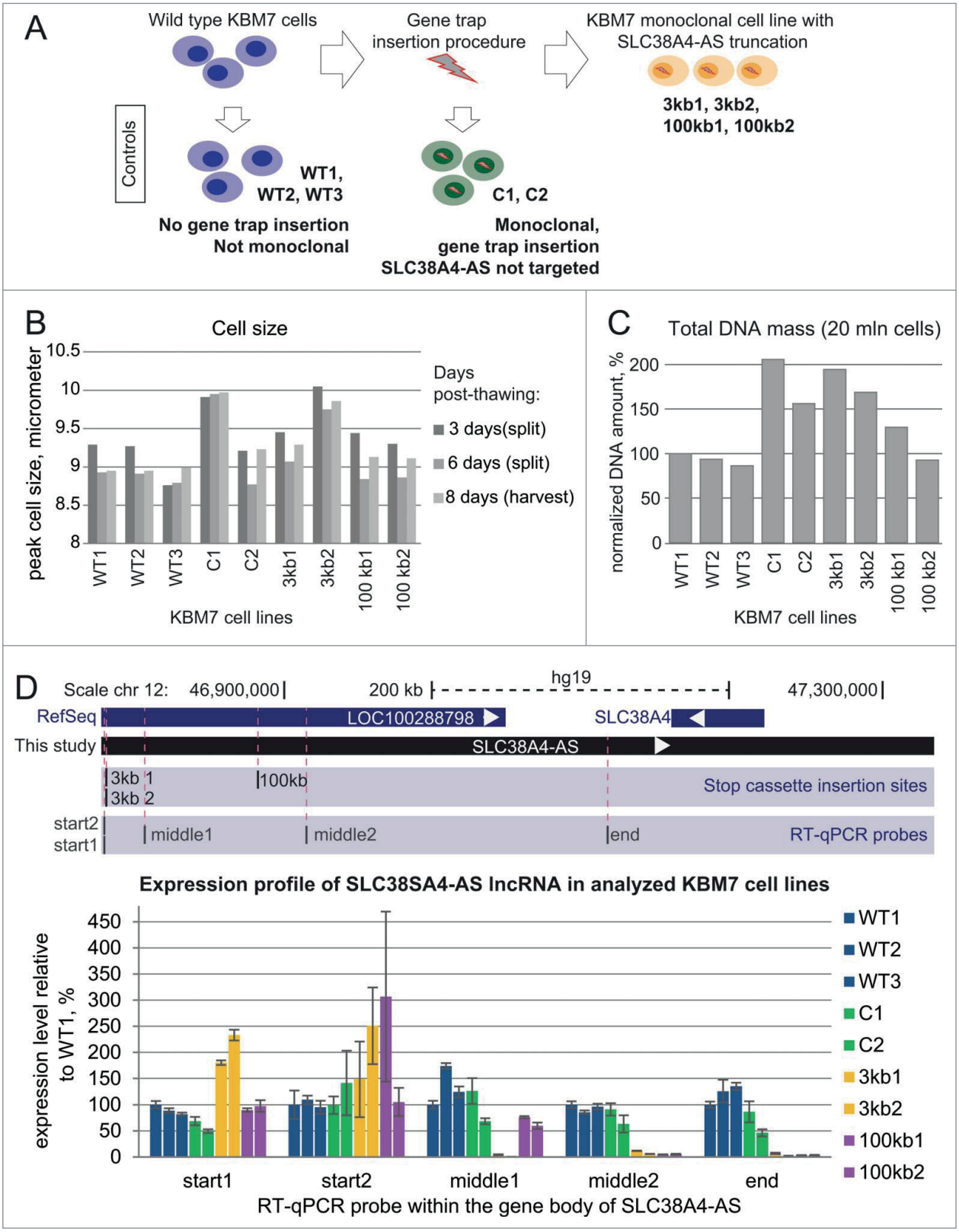


Figure 3. (For figure legend, see page 10.)

measurements we found that C1 and 3kb2 cells displayed 2 and 1.5 fold increase in DNA amount compared to wild type controls. Additionally we found that 3kb1 and C2 also showed 2 and 1.5 fold increase in DNA amount (Fig. 3C). As both cell size and DNA content are indirect measures of ploidy we performed karyotyping of selected cell lines (3kb2, 100kb, C1, WT2, Supplemental Figs. 2–5). This confirmed the haploid state of the 100kb and WT2 cell lines and the diploid state of the 3kb2 and C1 cell lines. Also we did not detect large scale chromosomal aberrations in addition to the known t(9;22) translocation.[45] This indicated that most cell lines that underwent gene trap insertion and clonal expansion procedure either gained diploidy, or were a mixture of haploid and diploid cells. Note that KBM7 cell ploidy does not interfere with any downstream analyses, as RNA-seq expression analyses are performed on normalized values that correct for increased RNA amount in diploid versus haploid cells. To confirm that both alleles carry the gene trap insertion and to validate the integrity of the genomic locus after the gene trap insertion we performed 2 DNA blotting assays for the 2 3kb truncation cell lines (see Supplemental Figure. 6A-B for maps of restriction enzymes and probes). First, we identified the expected 2.8kb (size of the gene trap cassette) increase in size of a genomic EcoRV fragment including the gene trap insertion site in 3kb1 and 3kb2 cell lines compared to wild-type (Fig. S6C–E). Second, we identified the expected size reduction of a genomic EcoRI/BamHI fragment due to the insertion of a BamHI site with the gene trap cassette (Fig. S6D–F). Importantly, we did not detect any wildtype fragment in the 3kb1 and 3kb2 cell lines indicating that gene trap insertion occurred in sorted haploid cells and that diploidy arose after cassette insertion. Therefore it can be concluded that both chromosomes in diploid cells carry the gene trap.

We next tested if gene trap cassette insertions 3kb and 100kb downstream of the *SLC38A4-AS* transcription start indeed stopped transcription elongation. We designed 5 RT-qPCR probes inside the body of the *SLC38A4-AS* gene (Table 2, Fig. 3D). We placed 2 probes (start1 and start2) upstream of the 3kb stop cassette insertion site, one probe (middle1) downstream of the 3kb, but upstream of the 100kb stop cassette, and 2 probes (middle2 and end) downstream of the 100kb stop cassette insertion site. Note, that the "end" RT-qPCR probe lies outside of the gene body of RefSeq annotated *LOC100288798*. We used all these probes to define the profile of *SLC38A4-AS* transcription in 3 wild type (blue, WT1-3), 2 control (green, C1, C2), 2 3kb (yellow, 3kb1, 3kb2) and 2 100kb (purple, 100kb1, 100kb2) *SLC38A4-AS* truncation cell lines (Fig. 3D bar plot). Since *SLC38A4-AS* RNA-Seq signals decreased from 5' to the 3' end (see Fig. 2), we normalized expression levels to WT1 for each RT-qPCR probe. All cell lines displayed transcription of *SLC38A4-AS* upstream of the 3kb gene trap insertion site, with increased expression in the 2 3kb truncation cell lines (Fig. 3D, start1 and start2). Consistent with expectations, the 2 3kb truncation cell lines displayed dramatic reduction of *SLC38A4-AS* transcription 28kb downstream of the transcription start (25kb downstream the truncation site, middle 1), while the 100kb truncation cell lines displayed continuous *SLC38A4-AS* transcription since these cell lines carried the stop

Figure 3. (see previous page) Gene trap technology allows truncation of *SLC38A4-AS* lncRNA in human haploid KBM7 cell line (A) Overview of the experimental design: *SLC38A4-AS* truncation and control cell lines used in the study. Top row: Wild type KBM7 cells underwent the gene trap insertion procedure and single clones were selected and expanded to a monoclonal population. Three independently obtained clones with gene trap cassettes mapping within the gene body of *SLC38A4-AS* lncRNA were available (see Table 1). Two monoclonal cell lines with independent insertion events that integrated a gene trap cassette 3kb downstream of *SLC38A4-AS* transcription start site (TSS) were available (3kb1 and 3kb2). Only one monoclonal cell line had a gene trap insertion 100kb downstream of the downstream of *SLC38A4-AS* TSS. Therefore we prepared biological replicates by performing independent thawing and culturing procedures (100kb1 and 100kb2). Left column: We obtained 3 wild type KBM7 control cell lines, which did not undergo any gene trap insertion procedure, were not monoclonal and were cultured by different people at different times prior to culturing for this analysis (WT1, WT2 and WT3). Middle column: To control for changes during gene trap insertion and selection procedure we obtained 2 KBM7 cell lines that did undergo gene trap insertion within the body of *HOTTIP* lncRNA and were monoclonally expanded (C1 and C2) (see Table 1). **(B)** Ploidy of KBM7 cell lines assessed by cell size. Bar plot shows peak cell size measured for 9 cultured KBM7 cell lines (Methods). All the cell lines were thawn and processed in one batch by the same person. Cell size was measured at the first splitting (3 days post-thawing, dark gray bars), second splitting (6 days post-thawing, medium gray bars), and prior to harvesting (8 days post-thawing, light gray bars). **(C)** Ploidy of KBM7 cell lines assessed by total DNA amount. Bar plot shows total DNA mass isolated from 20 million cells. DNA mass in the plot is normalized to WT1 sample (absolute value for WT1 is 109 μg). **(D)** Confirmation of successful *SLC38A4-AS* truncation by RT-qPCR. Top: schematic representation of the locus (drawn to scale). Blue bars show RefSeq annotation of *LOC100288798* and *SLC38A4* genes. Black bar underneath shows the extended annotation of *LOC100288798* (*SLC38A4-AS*) obtained in this study (Fig. 2). White arrows inside the bars indicate transcriptional orientation of the gene. Below the positions of stop cassette insertions (Table 1) and RT-qPCR probes are displayed (Table 2). Bottom: Expression profiling of *SLC38A4-AS* in the KBM7 cell lines (described in A). Error bars represent standard deviation from 3 RT-qPCR technical replicates. Bars are ordered from left to right as listed (top to bottom) in the legend on the right. For each RT-qPCR probe the expression level in WT1 is set to 100%.

Table 2. RT-qPCR probes for analyzing expression profile of *SLC38A4-AS* lncRNA.

RT-qPCR probe	forward primer, 5′-3′	reverse primer, 5′-3′	distance from TSS, bp
start1	CCCCGAGCAAATGGTGAATC	GGCATTATGTCATCGTCCTTTCA	1,560
start2	CATTCCAAGGCAGTGTTACATTTT	TCGGGGCTAAAGGTGTATGA	1,452
middle1	TGGGGCTGAAACATTTAGGC	TCAGGCTCCATGTTCCTACC	28,415
middle2	GGAACTAACAACGTCACAGGTAAT	ACCACATTCAACAGGAGAGAATAG	136,322
end	GTCCCTTCAAAGGAGGGTTT	GAAGGTGCCAAGTTTGAGGT	338,946

cassette downstream of this RT-qPCR probe (Fig. 3D, middle1). Expression levels downstream from the 100kb stop cassette were dramatically reduced in both the 3kb and 100kb truncation cells, but largely unchanged in the wild type and the control cells (Fig. 3D, middle2 and end). Thus, RT-qPCR confirmed that the *SLC38A4-AS* lncRNA was successfully truncated in KBM7 cells at the gene trap cassette insertion sites. Importantly, lack of transcription at multiple positions downstream of the gene trap cassette insertion sites in all tested cell lines further indicates that the *SLC38A4-AS* gene generates a single 558kb long transcript.

RNA-seq of KBM7 cell lines with truncated SLC38A4-AS lncRNA confirms a single transcription unit overlapping SLC38A4

As RT-qPCR only detects transcripts in a very narrow window at the chosen primer position, we performed RNA-seq to obtain a global picture of *SLC38A4-AS* truncation. We chose 2 cell line replicates per group: wild type (WT2 and WT3), control (C1 and C2), 3kb (3kb1 and 3kb2) and 100kb (100kb1 and 100kb2). 50bp single-end RNA-seq and alignment using STAR[55] produced an average of 35 million uniquely mapped reads per sample (standard deviation – 1.0 million reads) (Table S1D). Visual inspection showed similar *SLC38A4-AS* RNA-seq profiles in wild type and control cells with a similar decrease in signal from 5' to 3' end as seen before (compare Fig. 2 and Fig. 4A wild type). While the 3kb2 cell line showed a clear reduction of RNA-seq signal downstream the 3kb stop cassette insertion site, 3kb1 seemed to have residual transcription and thus truncation might be less efficient. Both the 100kb1 and 100kb2 replicates displayed a similar *SLC38A4-AS* expression profile with a clear reduction in RNA-seq signal after the gene trap cassette insertion point. We next quantified the RNA-seq signal strength to confirm the conclusions made from visual inspection. To obtain a transcription profile of *SLC38A4-AS* in each cell line we calculated RPKM of 5 regions (relative to the transcription start): 0-3kb, 3kb-50kb, 50kb-100kb, 100kb-300kb and 300kb-600kb (Fig. 4B). WT, C and 100kb cell lines showed a 3-fold RPKM drop from 0-3kb to 3kb-50kb regions with detectable expression in the 3kb-50kb window (RPKM $>$ 0.2), which is consistent with the reported RNA-seq signal decrease from 5' to the 3'end for lncRNAs.[61] In the 3kb cell lines the gene trap cassette stopped *SLC38A4-AS* and removed this pattern, and therefore all windows downstream of the gene trap cassette insertion site showed very low expression (RPKM $<=$ 0.05). WT and C cell lines showed a further 1.8- and 1.7-fold signal drop between 50-100kb and 100kb-200kb regions confirming the visual impression that the RNA-Seq signal decreases from 5' to 3' end in WT and C cell lines. The 100kb cell lines follow the expression pattern of the WT and C cell lines but the signal drops to very low expression levels (RPKM $<=$ 0.02) after the gene trap insertion site.

To allow a direct comparison between cell lines we plotted the expression of each window relative to WT (set to 100%, Fig. 4C). The first window (0-3kb) showed similar expression in WT, C and 100kb cell lines but was ~3-fold lower in 3kb cell lines. The following window (3-50 kb) showed a further ~3-fold reduction in expression for the 3kb cell lines whereas all other cell lines showed similar expression of *SLC38A4-AS*. At the 50-100kb window the expression of the 100kb truncation cell lines started to drop ~2-fold but were still ~2-fold higher than 3kb truncation cell lines. In the last 2 windows (100-300kb, 300kb-600kb) the 100kb truncation cell lines showed a low residual expression level (~10-fold less compared to WT, 6-8 fold less than C) whereas 3kb truncation cell lines showed a 2-3 fold higher residual expression likely due to the inefficient truncation of the 3kb1 cell line identified by visual inspection. We observed that while difference between 100kb replicates was low for every analyzed *SLC38A4-AS* region (maximal difference between 100kb1 and 100kb2 constituted 37% of the mean, at 100-300kb, Fig. 4C), the difference between 3kb1 and 3kb2, which resulted from different integration events, was more notable (maximal difference between 3kb1 and 3kb2 constituted 126% of the mean, at 100-300kb, Fig. 4C). 3kb1 showed 2.5- to 4.4-fold higher expression compared to 3kb2 in the 4 windows downstream the 3kb gene trap insertion (Fig. 4B). In spite of increased RNA-seq signal compared to the 3kb2 and 100kb truncations, the 3kb1 cell line did not reach the wild type and control levels of *SLC38A4-AS* transcription (Fig. 4C). It was possible that the difference

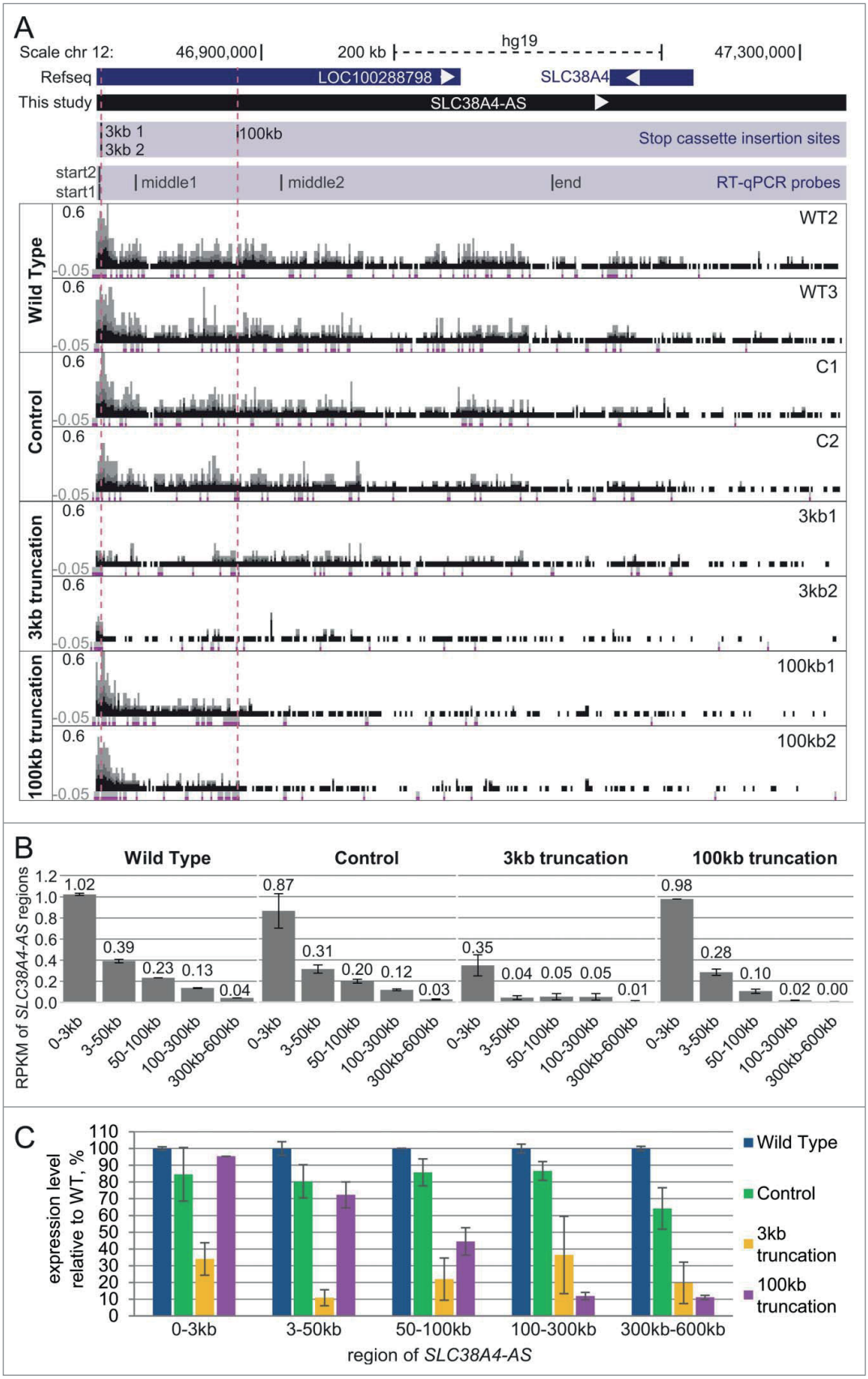


Figure 4. (For figure legend, see page 13.)

in truncation efficiency between the 3kb1 and the 3kb2 cell lines was due to sequence aberrations in the splice acceptor sequence in the gene trap cassette. Therefore we amplified and sequenced this region of the gene trap cassette and found it to be identical in the 3kb1, 3kb2 and C1 cell lines (Supplemental Fig. 7A–B). In order to discriminate inefficient truncation of *SLC38A4-AS* from a contamination of the 3kb1 cell line with wildtype cells we performed a PCR assay with primers directly flanking the cassette insertion site. We identified the correct wildtype PCR fragment in all tested cell lines, except for 3kb1 and 3kb2 cell lines, where the cassette insertion separates the primers by 2.8kb, which is not amplified in our settings (Supplemental Fig. 7C). Importantly this indicates that the 3kb1 cell line is not contaminated with wildtype cells to a detectable level. In summary, RNA-seq confirms efficient truncation of *SLC38A4-AS* in both 100kb truncation cell lines and the 3kb2 cell line. Interestingly, the global transcriptional analysis of 3kb1 truncation revealed reduced truncation efficiency in this cell line.

SLC38A4-AS truncation causes deregulation of several genes in trans

To investigate if *SLC38A4-AS* truncation had an effect on gene expression *in cis* or *in trans*, we calculated expression level of RefSeq annotated protein-coding genes and performed differential gene expression analysis using Cuffdiff software.[62] We compared WT2, WT3, C1 and C2 (4 control replicates) with 3kb1, 3kb2, 100kb1 and 100kb2 (4 targeted cell line replicates). This analysis produced a list of 120 significantly differentially expressed genes (excluding chromosomes X and Y, Table S1E) that we further filtered by requiring a 3-fold expression change between the 2 conditions, which resulted in a list of 41 protein-coding genes (Table S1 Elines in bold). This number of genes was 5-fold higher than the average number of genes differentially expressed (3-fold expression change) in 11 mock comparisons (Table S1F). Interestingly, the 41 genes were distributed across almost all chromosomes (Table S1 Elines in bold). One gene (*CD163L1*) was down-regulated and 3 (*CD9*, *EMP1* and *CRY1*) were upregulated on chromosome 12, the same chromosome that contains *SLC38A4-AS*. However, these genes were located 33-61 million bp distant from *SLC38A4-AS* and therefore their regulation is more likely to arise from *trans* effects. We then calculated expression levels (FPKM, Methods) of the 41 significantly deregulated genes reported above by Cuffdiff for each of the 8 samples separately to allow unsupervised clustering to be performed (Methods). This analysis correctly grouped the 2 biological replicas of the 3kb truncation, 100kb truncation replicates and wild type replicates (Fig. 5A). Interestingly, C1 and C2, although in the same branch, did not group together, which may relate to the fact that C1 carries a truncated *HOTTIP* lncRNA (gene trap insertion in sense to *HOTTIP*, Table 1), while C2 had an antisense insertion in the *HOTTIP* gene body, and therefore should not truncate (Table 1).

We then performed further filtering to create a small stringent list of the deregulated genes. To increase the stringency of the list of differentially expressed genes we performed 3 filtering steps. First, we filtered out genes that showed significant differential expression between wild type (WT2, WT3) and control (C1, C2) samples and thus might be differentially expressed due to the effect of the gene trap cassette insertion procedure (3/41 genes). Second, we removed the genes that showed differential expression between 3kb and 100kb truncation thus restricting our list to the genes that are regulated by the part of *SLC38A4-AS* lncRNA downstream of the 100kb cassette insertion site (18/41 genes). Third, we only retained the genes that were differentially expressed in both pairwise comparisons of control to 3kb (3kb1, 3kb2 vs C1, C2, 12 genes) and control to 100kb samples (100kb1, 100kb2 vs C1, C2, 24 genes). These filtering steps resulted in a stringent list of 6 protein-coding genes (Table 3). Three of these genes, including *CD9* (Fig. 5B) were upregulated upon *SLC38A4-AS* truncation, and 3, including *RORB* (Fig. 5C), were downregulated. In summary, these data show that genetic truncation of *SLC38A4-AS* lncRNA results in genome-wide gene expression changes and provides a stringent list of 6 potential *SLC38A4-AS* target genes.

Figure 4. (see previous page) RNA-seq confirms truncation and continuity of the *SLC38A4-AS* lncRNA gene. (A) *SLC38A4-AS* RNA-seq signal of the 8 clones analyzed in Fig. 3D. Top: schematic representation of the locus (as described for Fig. 3D). Bottom: RNA-seq signal, normalized to sample read number, pink dots indicate RNA-seq signal that exceeds the range presented inside the box. Type of the cell line is indicated on the left, name of the cell line is indicated on the right. Vertical dashed red lines indicate position of the 3kb and 100kb stop cassettes. Low density of RNA-seq signal piles indicate low expression and the smallest size corresponds to 1 read. (B) Expression profile of different regions of *SLC38A4-AS* lncRNA in the RNA-Seq data shown in (A). Bar plots show RPKM of the regions of *SLC38A4-AS* indicated on the X axis for 4 types of cell lines (as grouped on A). RPKM value for each clone type is averaged from 2 cell lines, error bars show the RPKM values of the 2 samples. Numbers above the bars show the plotted value. Note that this analysis allows the comparison of regions within one cell line but not between cell lines. (C) Expression profile comparison of *SLC38A4-AS* between analyzed clones. Bar plot shows RPKM of the regions of *SLC38A4-AS* indicated on the X axis for each cell line type normalized to the value for "Wild type". Normalized RPKM values are the average of 2 cell lines of each type, indicated by the error bars.

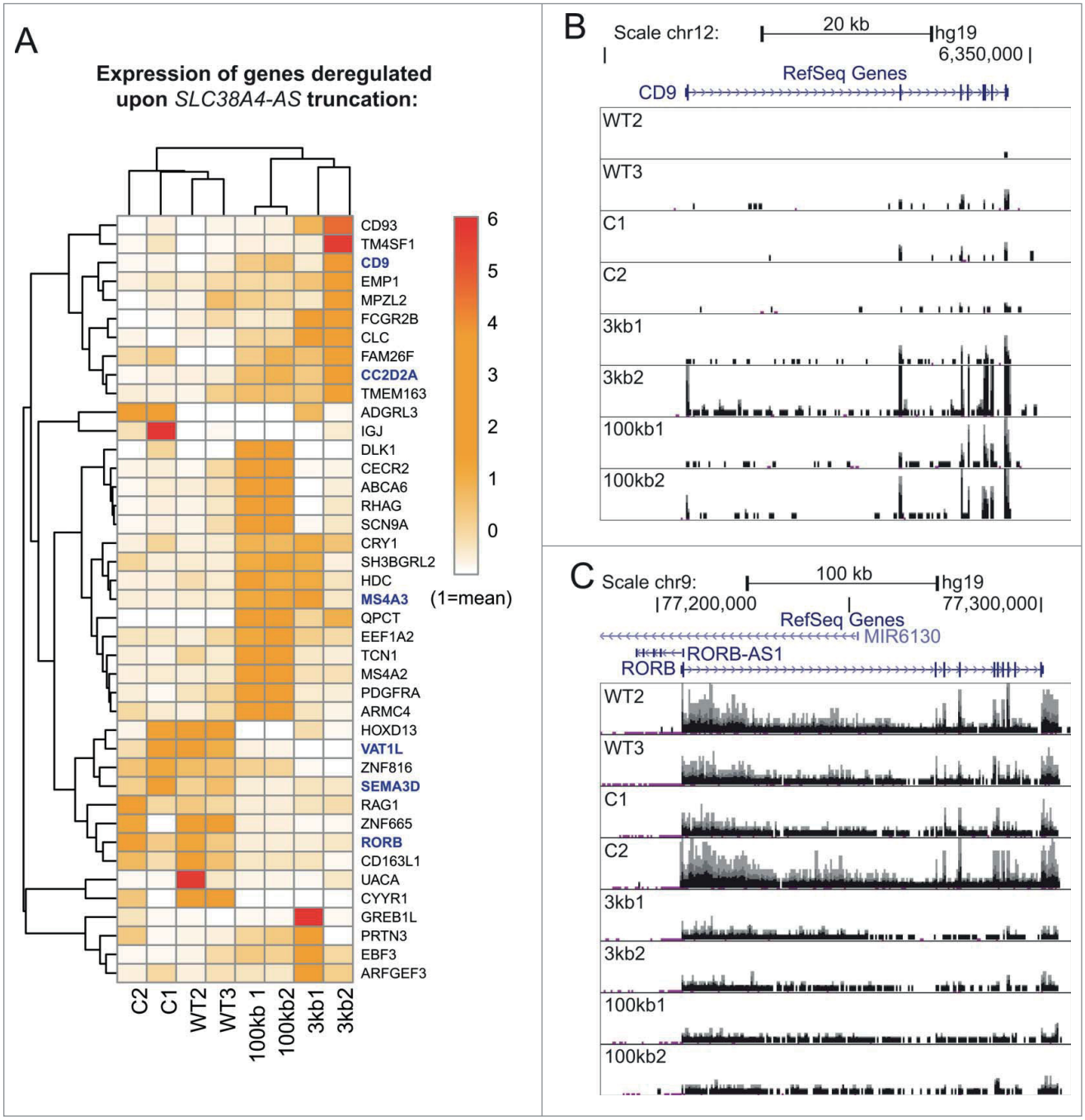


Figure 5. Genome-wide differential expression analysis reveals deregulation of protein-coding genes *in trans* upon *SLC38A4-AS* lncRNA truncation (A) Expression level of genes differentially expressed between *SLC38A4-AS* truncation cell lines and the 4 control cell lines allows unsupervised clustering of the cell lines that resembles the different cell groups. Heat map shows expression level (FPKM, Methods) of genes (name indicated on the right) with significant differential expression ($p < 0.01$, >3 fold expression change, Methods) between 2 conditions: no *SLC38A4-AS* truncation (WT2, WT3, C1, C2) and genetic truncation of *SLC38A4-AS* (3kb1, 3kb2, 100kb1, 100kb2). Expression values are normalized to the mean FPKM among all 8 samples. Mean is set to 1. Names of genes that form the filtered stringent list of deregulated genes (Table 3, Methods) are displayed in bold blue font. Heat map color legend is displayed on the right. (B) and (C) Examples of up- and downregulated protein coding genes from the stringent list (Table 3). *CD9* is markedly upregulated (B) and *RORB* is markedly downregulated (C) upon truncation of *SLC38A4-AS*. UCSC Genome Browser screen shots show normalized RNA-seq signal. Top to bottom: Chromosome position, RefSeq gene annotation, RNA-seq signal, normalized to sample read number, from eight sequenced cell lines. Each box shows the same range from 0 to 0.6, only forward strand is shown. Pink dots indicate RNA-seq signal that exceeds the range presented inside the box. Name of cell line is indicated on the left.

Table 3. Stringent list of genes affected by SLC38A4-AS lncRNA truncation.

Gene	RefSeq ID	Full name of the gene	expression fold change upon SLC38A4-AS truncation	genomic position		
CD9	NM_001769	CD9 molecule	14,3	chr12	6309481	6347437
CC2D2A	NM_001080522	coiled-coil and C2 domain containing 2A	8,4	chr4	15471488	15603180
MS4A3	NM_006138	membrane-spanning 4-domains, subfamily A, member 3 (hematopoietic cell-specific)	5,4	chr11	59824100	59838588
SEMA3D	NM_152754	sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3D	−4,2	chr7	84624871	84751247
RORB	NM_006914	RAR-related orphan receptor B	−4,8	chr9	77112251	77302117
VAT1L	NM_020927	vesicle amine transport protein 1 homolog (T. californica)-like	−17,8	chr16	77822482	78014001

As these results provide clear evidence for the use of the "Human Gene Trap Mutant Collection" to study lncRNAs, we investigated how many lncRNAs can be potentially studied using this collection in its current form. First, we calculated expression for all GENCODE v19 lncRNAs in the 2 wild type cell lines investigated in this study (WT1, WT2) and found 2,307 non-overlapping lncRNA loci to be expressed (i.e. to express at least one lncRNA isoform with RPKM>0.2). Next, we investigated how many GENCODE v19 lncRNAs contained a gene trap insertion on the same strand and found that 938 lncRNAs are likely to be truncated in one of the available cell lines (Fig. 6A left bar). Overlapping these 2 data sets revealed 409 expressed lncRNAs carrying a gene trap insertion in the current collection (Fig. 6A middle bar). If we set a higher expression cut off of RPKM>0.5, we find 266 lncRNAs carrying a gene trap (Fig. 6A right bar). We investigated the position of gene trap insertions relative to the transcriptional start site of lncRNAs and found enrichment at the 5' end (Fig. 6B). Finally we examined the well-studied lncRNA *MALAT1* and identified 5 gene trap insertions close to the 5' end corresponding to potential knock-out cell lines.(Fig. 6**C**)

Discussion

Here we report the first use of the "Human Gene Trap Mutant Collection"[45] to study the function of a human lncRNA. To demonstrate the utility of this collection we analyzed cell clones that successfully truncated the *SLC38A4-AS* lncRNA (renamed from *LOC10028879*) that displays RNA biology features distinct from protein-coding genes, including low expression and inefficient splicing. We also investigated this gene trap collection as a whole for its suitability for the study of lncRNAs, and identified hundreds of lncRNAs with gene trap insertions including the well-studied *MALAT1* lncRNA. Therefore we demonstrate here the utility of the "Human Gene Trap Mutant Collection" for studying lncRNAs and also identify *SLC38A4-AS* as a very long and novel functional regulatory lncRNA.

Prior to analyzing gene trap efficiency we examined the RNA biology of the *SLC38A4-AS* lncRNA that has not previously been characterized. We showed that *SLC38A4-AS*, unlike many lncRNAs, does not show tissue-specific expression. While tissue-specificity is often considered as an indication of functionality,[63] several ubiquitously expressed lncRNAs have been proven to play important gene regulatory roles.[40,64] We used a set of public RNA-seq data to show that *SLC38A4-AS* lncRNA is inefficiently spliced and that the major unspliced isoform is nuclear localized. Importantly, by comparing *SLC38A4-AS* to 2 control protein-coding genes, we show that the unspliced isoforms we detect for *SLC38A4-AS* are not just an intronic signal. We conclude this from the finding that the polyadenylation and localization profiles for unspliced isoforms of the protein-coding genes, which are notably highly expressed, differ dramatically from that of *SLC38A4-AS*. Minor spliced isoforms of *SLC38A4-AS* lncRNA are well detectable in the cytoplasm and thus are exported and likely stable. *SLC38A4-AS* lncRNA is thus a transcript with unusual RNA biology features different from protein-coding genes. We performed *de novo* transcriptome assembly in the region and were able to show that transcription of *SLC38A4-AS* extends 289kb downstream the RefSeq annotated 3' end and overlaps the downstream *SLC38A4* gene.

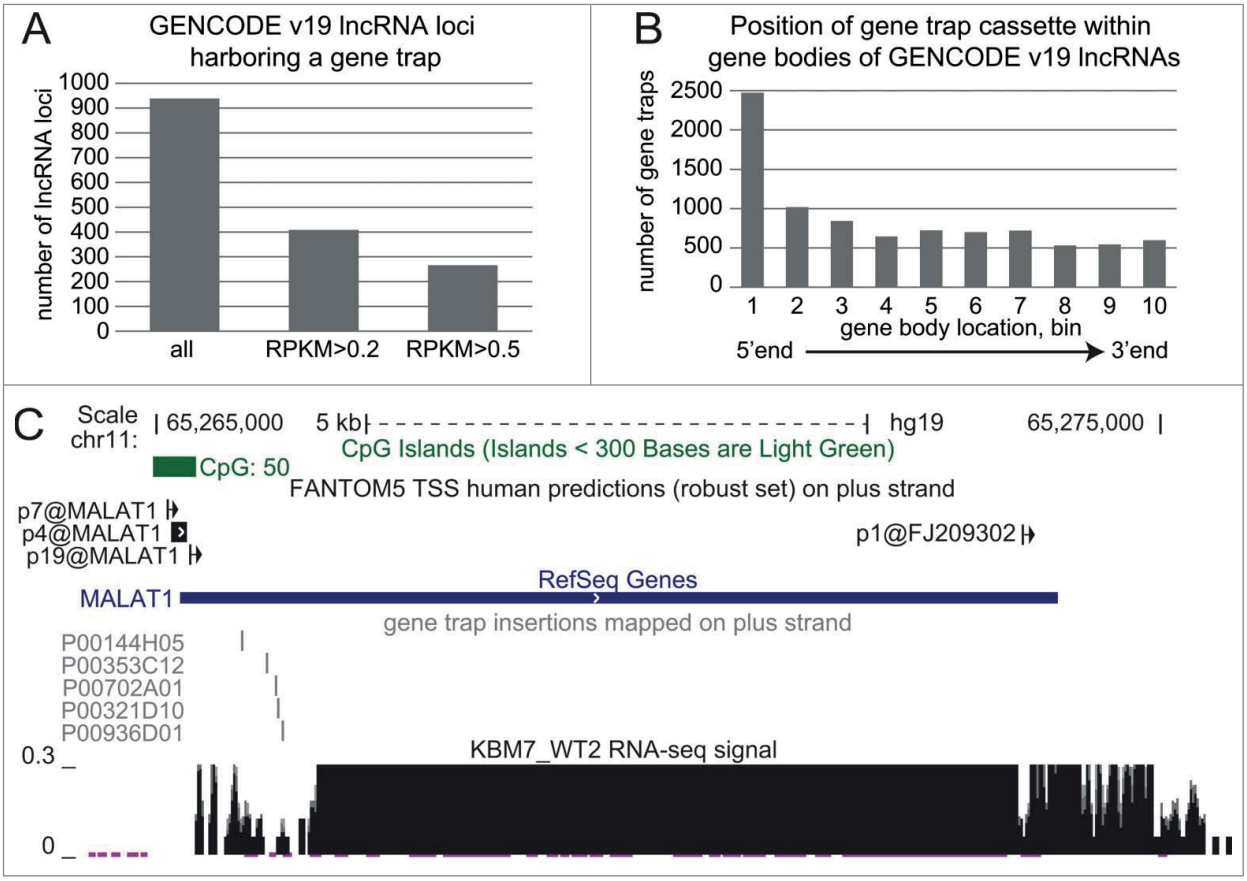


Figure 6. Haploid gene trap collection represents a rich resource for quick functional assessment of hundreds of lncRNAs. (A) Hundreds of GENCODE v19 lncRNAs expressed in KBM7 cell line are targeted by a gene trap insertion. Bar plot shows number of non-overlapping GENCODE v19 lncRNA loci that contain a gene trap cassette in the same transcriptional orientation in KBM7 clones within the "Human Gene Trap Mutant Collection" (left bar, Methods), and the number of these lncRNA loci that are expressed (middle bar, loci that contain lncRNA transcripts expressed with RPKM > 0.2) and well expressed (right bar, loci that contain lncRNA transcripts expressed with RPKM > 0.5) in wild type KBM7 cells. (B) Gene trap cassettes are preferentially inserted at the 5′ end of lncRNAs. Bar plot shows the number of gene trap cassettes inserted into different regions in the gene bodies of GENCODE v19 lncRNA. Numbers correspond to 10 equally sized, non-overlapping regions investigated for each gene. (C) Five genetic truncations of the well-known lncRNA *MALAT1* are available within the "Human Gene Trap Mutant Collection." Shown is the UCSC browser screen shot of the *MALAT1* gene region. From top to bottom: chromosome scale, CpG island annotation (UCSC track), FANTOM5 TSS predictions (robust set)[82] on the plus strand, RefSeq gene annotation, position of gene trap insertion cassettes available (plus strand), normalized RNA-seq signal from WT2 KBM7 cell line showing wild type expression of *MALAT1*.

We then obtained KBM7 cells from the "Human Gene Trap Mutant Collection" with gene trap insertions at 2 different locations (3kb and 100kb downstream of the transcription start) in the gene body of *SLC38A4-AS* lncRNA to test whether the unusual RNA biology features interfered with efficient truncation by the gene trap cassette. By using qRT-PCR as well as RNA-seq we identified one cell line with efficient truncation at both insertion sites. This data not only verifies that gene trap insertions in KBM7 cells efficiently truncate *SLC38A4-AS* lncRNA, but also confirms our prediction of the extended *SLC38A4-AS* lncRNA length. Detailed RNA-seq analysis identifies that the 3kb1 cell line shows less efficient truncation compared to 3kb2 cell line despite these cell lines sharing same gene trap insertion site. Differences in the efficiency of truncation between different insertion sites have been documented for one truncation of the *Airn* lncRNA. In this case a truncation cassette insertion at 3 different genomic loci caused successful truncation of the lncRNA whereas the same cassette was highly inefficient when inserted into a CpG island.[14] Also differences in the gene trap efficiency of protein-coding genes were noted for different cassette integration sites.[45] However, a difference between similar insertion sites as shown for 3kb1 and 3kb2, was surprising. DNA gel blotting experiments did not detect a large scale rearrangement of the chromosomal locus with the gene trap insertion nor did they identify a contamination of the 3kb1 cell line with wildtype cells. As DNA blotting might not be sensitive enough to detect a low level of

wildtype cell contamination we validated these results by a PCR assay. We also validated that the splice acceptor sequence was unchanged in the 3kb1 cell line. Taken this together, an aberration of the genetic sequence in 3kb1 is unlikely to be the cause for the reduced efficiency of transcription termination in this cell line. A connection between chromatin structure and transcription termination has been made in yeast[65] and it has been suggested that local chromatin changes influence splicing.[66] It is therefore possible that cell line specific local chromatin changes result in differences in truncation efficiency at identical cassette integration points. As global gene-expression analysis showed high similarity between both 3kb truncation cell lines, it is highly likely that the residual level of *SLC38A4-AS* expression seen in 3kb1 cell line is not sufficient to maintain a wildtype gene expression pattern. We therefore conclude that gene trap approach used for the "Human Gene Trap Mutant Collection" is a useful tool to truncate inefficiently spliced lncRNAs.

We noted that 2 qRT-PCR primers that are close to the 3kb truncation cassette insertion site, showed elevated qRT-PCR signals specifically in 3kb truncation cell lines. Interestingly RNA-seq did not support this elevated transcription on the forward strand, which corresponds to *SLC38A4-AS* lncRNA, but identified strong transcription from the reverse strand directly at the gene trap insertion site that was absent in the control cell lines. Similar transcription on the reverse strand at the gene trap insertion point was visible albeit at lower levels for the 100kb truncation cell lines (Fig. S8). Thus, we provide evidence that the gene trap cassette used for the "Human Gene Trap Mutant Collection" can drive transcriptional activity, which was suggested earlier.[45] Additionally, we also show that this activity can be strong (2-fold higher than *SLC38A4-AS*) and therefore has to be carefully considered when expression of genes in close proximity is affected, as transactivation of protein-coding genes by the transcriptionally active viral LTRs was reported in gene therapy patients.[67]

Interestingly, *SLC38A4-AS* lncRNA shares several unusual RNA biology features with the imprinted mouse lncRNA *Airn* that also overlaps in antisense orientation and silences the protein-coding *Igf2r* gene. Although *Airn* lncRNA is inefficiently spliced, 5% of its nascent transcripts are spliced and give rise to stable lncRNAs that are exported to the cytoplasm.[20] These spliced *Airn* lncRNA isoforms are, however, not connected to the silencing mechanism.[14] Interestingly, truncation experiments identified that *Airn* silences *Igf2r* due to its transcriptional overlap, a phenomenon called transcriptional interference.[14,40] The *Airn* lncRNA also silences 2 protein-coding genes that it does not overlap in a tissue-specific manner, likely by targeting repressive chromatin to the promoters of these genes.[68,69] We tested if the *SLC38A4-AS* lncRNA silences the *SLC38A4* protein-coding gene that it overlaps and/or the *SLC38A2*, which is located 10kb away in a similar manner. We were surprised to find that neither *SLC38A4* nor *SLC38A2* protein-coding genes were affected by the truncation of *SLC38A4-AS* lncRNA. In addition, expression analysis of multiple tissues did not show anti-correlating expression patterns of the 2 protein-coding genes with the lncRNA. In the case of imprinted expression involving a repressor lncRNA, such a pattern would not be expected as one allele expresses the protein-coding gene whereas the other allele expresses the lncRNA. Therefore we conclude that *SLC38A4-AS* lncRNA most likely does not share functional similarities with the imprinted *Airn* lncRNA and does not control *SLC38A4* or *SLC38A2* protein-coding gene expression. This data supports the hypothesis that imprinted expression of *Slc38a4* in the mouse, is rodent-specific as it is also absent from the pig and cow.[70,71]

In order to test the functional importance of *SLC38A4-AS* lncRNA as a gene regulator *in trans,* we tested whether the truncation of the lncRNA resulted in gene expression changes in KBM7 cells. In accordance with recent guidelines established for the correct analysis of lncRNA knockout experiments, we included a number of controls in this analysis.[32] First, we excluded batch effects from the handling of cells by having all cell lines cultured in parallel by one person. Second, it is possible that the gene trap insertion disrupts an important genetic element that causes gene expression changes of protein coding genes that are not dependent on the lncRNA. Therefore we analyzed 3 independently derived *SLC38A4-AS* lncRNA truncation cell lines: 3kb1, 3kb2 with an identical insertion site and 100kb. As controls we used 2 batches of wild type KBM7 cell lines. In order to identify genes that are specifically deregulated upon truncation we performed differential gene expression analysis between *SLC38A4-AS* lncRNA truncation cell lines (3kb1, 3kb2, 100kb1, 100kb2), and all control cell lines (C1, C2 that carried gene traps at unrelated loci, WT1, WT2 that lacked gene traps). This analysis resulted in 120 differentially expressed genes, 41 of which were more that 3-fold up/downregulated in the truncation cell lines. Importantly, none of the differentially expressed genes were located in close proximity to the *SLC38A4-AS* lncRNA, as reported for well-known *cis*-regulating lncRNAs, such as *Airn* or *KCNQ1OT1*.[36] Whereas clustering based on the 41 differentially expressed genes allowed correct grouping of the replicates, performing a similar analysis using the expression of genes in the 10Mbp region around *SLC38A4-AS* resulted in sporadic clusters. This indicates a lack of consistent changes of

these genes between control and truncation cell lines and thus further supports a lack of *cis*-acting regulatory function of *SLC38A4-AS* lncRNA (Supplemental Fig. 9). We plotted expression values of the 41 significantly deregulated genes in all the 8 cell lines as a heat map and found that a number of genes seemed to be specifically expressed in one control cell type (C1/C2 or WT1/WT2) or in one of the truncation cell types (3kb1, 3kb2 or 100kb1, 100kb2) rather than in all control vs. all truncation cell types. Therefore, we also performed pairwise comparisons to remove these genes. We do note that this approach limits the part of the lncRNA examined for function to regions downstream of the 100kb truncation cassette (*i.e.,* spanning ~400kb of the *SLC38A4-AS* gene body). Additionally we note that the function of the first 3kb of *SLC38A4-AS* lncRNA (upstream 3kb gene trap cassette position) was not assessed in our study while it is possible that this region may possess a function.

Of the 6 genes that pass the most stringent filters for deregulation in *SLC38A4-AS* lncRNA truncation cell lines 2 are of special interest. The first is the *clusters of differentiation proteins 9 (CD9)* that belongs to the superfamily of tetraspanins, integral membrane proteins that play a role in multiple biological processes by interacting with membrane proteins like other tetraspanins, growth factors and cytokine receptors. Clinical data suggests that CD9 is a suppressor of metastasis and modulates tyrosine kinase receptor signaling in cancer.[72] CD9 is also a marker for haematopoietic stem cells[73] and was found to be up-regulated upon induction of pluripotent stem cells (iPS) from KBM7 cells,[74] although it is not necessary for pluripotency in mice[75]. The second gene is *RAR-related orphan receptor B (RORB or RORβ)*, which encodes the nuclear receptor subfamily 1, group F, member 2 (NR1F2) protein that binds to DNA and inhibits transcription.[76] *RORB* has not been implicated in cancer,[77] but was associated with the mammalian circadian clock,[76] and was found to be a member of a gene hub that discriminates human iPS from stem cells.[78] Little is known about the importance of *RORB* in KBM7 cells, however it is unlikely to be essential for this cell line as an unbiased mapping of gene trap insertions in this cell line identified 7 gene trap insertion events in this gene with 4 predicted to stop *RORB* transcription.[79]

As mentioned above, gene trap cassette removal could provide a valuable rescue control. Human Haploid Gene Trap Collection contains cell lines with gene trap cassettes flanked by loxP sites that thus can be removed by Cre recombinase expression and the expression of the targeted genes might be restored. Among the analyzed *SLC38A4-AS* truncation cell lines, 3kb1 and 3kb2 did have loxP sites flanking the gene trap cassette, while 100kb truncation cell lines did not. However, while removal of the truncation cassette by expressing the Cre recombinase and subsequent re-expression of full-length *SLC38A4-AS* lncRNA could restore its wildtype gene expression pattern, it is possible that the gene expression changes initiated by *SLC38A4-AS* lncRNA are accompanied by changes in secondary gene expression or in the epigenetic landscape that may not be immediately reversible. Such an example was reported for the *Airn* lncRNA that silences the *Igf2r* protein coding gene in early development. After silencing, by *Airn* transcription, *Igf2r* acquires repressive epigenetic marks on its promoter and silencing is stably maintained in the absence of *Airn* lncRNA expression.[46] Therefore we conclude that the use of multiple control cell lines may prove a more efficient way to study lncRNA function in comparison to multiple targeted cell lines.

In summary, this report shows that the "Human Gene Trap Mutant Collection" is a useful tool to study lncRNA function. Importantly, we identified 857 GENCODE v19 lncRNAs (http://www.gencodegenes.org/releases/19.html) for which KBM7 gene trap insertions cell lines are available (Methods and https://opendata.cemm.at/barlowlab/). Similar to protein-coding genes, the gene trap cassette preferentially inserts close to the 5' end of lncRNAs, which is useful for functional studies as the bulk of the lncRNA will not be produced.[45] We found that 409 lncRNA loci with a gene trap insertion show an RPKM >0.2 (RPKM of at least one isoform in the locus) and 266 have an RPKM>0.5, which constitutes respectively 44% and 28% of all GENCODE v19 lncRNA gene trap insertion clones. It is to date unclear, which expression cutoff can be used to indicate functional importance, and it is therefore possible that also lncRNAs expressed to a lower level have a functional importance. The "Human Gene Trap Mutant Collection" could be a useful tool to study this question. Also KBM7 cells can be converted to iPS cells and have the potential to be differentiated into different lineages.[74] Therefore it is possible that lncRNAs that are lowly expressed in wild-type KBM7 cells are highly expressed in a different lineage, which can also be studied using KBM7 iPS cells. Gene trap KBM7 cells from the "Human Gene Trap Mutant Collection" are simple to obtain and culture and therefore offer a rich resource that allows analysis of lncRNA function in a human system. This is illustrated by the example of the *MALAT1* lncRNA. This lncRNA was previously studied using a truncation cassette,[44] an experiment that includes (1) cloning of the truncation cassette for homologous recombination (2) optimizing endonuclease to cleave genomic DNA at the desired position (3) selection, screening, expansion and testing of correctly targeted clones.[44] This effort linearly increases for the production of cell lines with different

truncation cassette insertion sites. In contrast to this time-consuming approach, 5 KBM7 gene trap clones are readily available truncating the *MALAT1* lncRNA at different positions close to the 5' end that are ready to be analyzed.

According to our results, the unusual RNA biology inherent to many lncRNAs does not influence the ability of the gene trap cassette to stop lncRNA transcription, and gene trap truncations are therefore a universal tool for studying a wide range of lncRNAs. The availability of multiple control cell lines is an additional advantage and allows thorough artifact control. Using *SLC38A4-AS* lncRNA as an example, we also show that gene trap resource together with the already available RNA-seq resources from the ENCODE consortium allow fast characterization of a lncRNA of interest. We anticipate that similar integrated approaches that make efficient use of these publicly available resources will allow the fast functional characterization of the many lncRNAs found in the human genome.

Methods

RPKM calculation

RPKMs were calculated using RPKM_count.py from RSeQC package (https://code.google.com/p/rseqc/) using –skip-multi-hits option.

Estimating expression of LOC100288798 and SLC38A4 in various tissues and cell types

Various public raw RNA-seq datasets (See Table S1A) were downloaded as fastq files and aligned with STAR using the following command: *STAR_2.3 –genomeDir hg19genome_no_splice_junction_database_provided –readFilesIn [read1.fastq] [read2.fastq] –outFilterMultimapNmax 10 –outFilterMismatchNmax 10 –outSJfilterOverhangMin 30 16 16 16 –alignSJDBoverhangMin 3 –alignSJoverhangMin 6 –outFilterType BySJout –outSJfilterCountUniqueMin 3 1 1 1 –outSJfilterCountTotalMin 4 2 2 2 –outSAMstrandField intronMotif –outFilterIntronMotifs RemoveNoncanonical –alignIntronMax 300000 –alignMatesGapMax 500000 –outFileNamePrefix [output] –outStd SAM –outSAMmode Full.* SAM output was converted to BAM and sorted by position using samtools software. Expression levels (RPKM) were estimated for RefSeq annotated isoforms of *SLC38A2*, *SLC38A4* and *LOC100288798* (*SLC38A2* – 1 isoform: NM_018976, *SLC38A4* – 2 isoforms: NM_018018 and NM_001143824, *LOC100288798* – 5 isoforms: NR_125377, NR_125378, NR_125379, NR_125380, and NR_125381). The average RPKM of all isoforms was displayed inside each cell of the heat map (Fig. 1B), which was built in R using the *pheatmap* function without clustering rows and columns. Rows were sorted according to expression level of *LOC100288798*. Heat map color scale was skewed toward lower values to highlight non-expressed genes (shades of blue – $0<RPKM<0.5$) and display the range of *LOC100288798* expression (shades of orange – $0.5<RPKM<10$).

Splicing efficiency calculation

Splicing efficiency was calculated using public total (ribosomal depleted) RNA-seq datasets of high depth (135-371 million reads, Table S1A). Splicing efficiency of each RefSeq annotated splice site was estimated by calculating RPKM of exonic and intronic 45bp regions surrounding the splice site starting 5bp away from the precise splice site position to allow for potentially imprecise annotation of the splice site. For each splice site, which passed the coverage cutoff (exonic RPKM > 0.2), "Splicing efficiency" (S), $S = 100^{*}(1\text{-}RPKM_{intronic}/RPKM_{exonic})$, was calculated. Splicing efficiency was within the range from 0 for fully unprocessed splice sites ($RPKM_{intronic} >= RPKM_{exonic}$, S was set to 0, when it was calculated to be <0) to 100 for perfectly processed splice sites ($RPKM_{intronic}=0$). We then calculated the average splicing efficiency of all the unique splice sites for each gene and assigned the splicing efficiency of the gene with this value.

Estimation of PolyA+ and nuclear enrichment

Publicly available cellular/PolyA fractionation RNA-seq data for 5 cell lines (HeLa, Lymphoblastoid cell line GM12878, Embryonic stem cells, HUVEC and K562) produced by the ENCODE project were downloaded as raw fastq files, aligned with STAR using default parameters. Expression of spliced products was calculated for: *LOC100288798*: averaged from NR_125379 NR_125380 NR_125378 NR_125377 NR_125381, *SLC32A2*: NM_018976, *TBP*: NM_003194, *XIST*: NR_001564 (RefSeq identifiers). Expression over the whole gene body was calculated for *LOC100288798*: over chr12:46777889-47046362 (gene body of NR_125381) and chr12:46777458-47046362 (gene body of NR_125379 NR_125380 NR_125378 NR_125377), *SLC38A2*: over chr12:46751970-4676664, *TBP*: over chr6:170863420-170881958, *XIST*: over chrX:73040485-73072588.

Assembly of SLC38A4-AS exon structure using publicly available RNA-seq data from multiple cell types

Exon structure assembly was performed for each of 46 public RNA-seq data only in the region of interest: *samtools view -b [position sorted STAR alignment] chr12:46,500,000-47,500,000 > tissue.1Mb.bam . De novo* transcriptome assembly was performed for each one of 1Mb regions in all the samples separately using

Cufflinks version 2.2.1 with the following command: *cufflinks –multi-read-correct –output-dir [output] -F 0.01 -p 7 –library-type fr-firststrand (if RNA-seq is stranded) –mask-file pseudogenes.gtf tissue.1Mb.bam* . Pseudogene annotation was obtained from GENCDOE v19. The resulting transcript assemblies were then merged using Cuffmerge with the following command: *cuffmerge -s hg19_fasta –keep-tmp -p 8 –min-isoform-fraction 0 [list of all gtf files from 46 cufflinks assemblies]*. Single exon transcripts were discarded.

KBM7 cell culture

All gene trap KBM7 cell lines were obtained frozen from Horizon Genomics GmbH (http://www.horizon-genomics.com/). WT KBM7 cell lines were from Horizon Genomics GmbH or from Sebastian Nijman lab. All cell lines were cultured in filter cap flasks in IMDM (Sigma) medium (with L-Glutamine, supplemented with Penicillin/Streptomycin and 10% Fetal Bovine Serum (PAA Laboratories (GE Healthcare)) at 37°C with 5% CO_2. KBM7 are suspension cells. Cell concentration and cell size were measured using Casy cell counter (Schärfe System GmbH).

RNA preparation

RNA was isolated from pelleted KBM7 cells using TRIreagent (Sigma), dissolved in RNA Storage Solution (RSS, Ambion) and stored at −20°C. RNA was DNAse I treated (DNAfree kit, Ambion). Quality control was performed by accessing RNA integrity using Agilent RNA 6000 Nano Kit.

RT-qPCR

RNA was converted to cDNA using RevertAid First Strand cDNA Kit (Fermentas) with –RT (no reverse transcriptase) control reaction for each RNA sample according to manufacturer's protocol. RT-qPCR was performed using MESA GREEN qPCR MasterMix Plus for SYBR Assay I dTTP (Eurogentec). Primers (Table 2) were designed using Primer3. RT-qPCR was performed using standard curves in 3 technical replicates for each sample and standard deviation between the replicates was used to define the error and plot the error bars.

DNA-blot

DNA extraction, restriction enzyme digestion and DNA gel blots were performed using standard methods. The hybridization probe was amplified by PCR, cloned and gel purified. Membranes were exposed to an imaging plate (FujiFilm) that was scanned (Typhoon TRIO, GE Healthcare). Levels were adjusted on the whole image to increase the visibility of all bands on the image.

Chromosome analysis

Metaphase preparation and FISH were carried out by standard methods. Dividing cells were locked in metaphase by adding colcemid (0.1μg/ml final concentration) (Gibco, ThermoFisher) for 60 minutes. After fixation cells were dropped onto slides, dried at 42°C for 30 minutes and then incubated at 60°C over night. One slide was used for Giemsa-trypsin banding of chromosomes. For FISH analyses a Cy3 labeled probe mix (Kreatech) was used which detects the centromeric regions of chromosomes 1, 5 and 19.

Strand-specific RNA-seq library preparation and RNA sequencing

4 μg of DNase I treated RNA underwent Ribosomal depletion using RiboZero rRNA removal kit Human/Mouse/Rat (Epicentre) following manufacturer's protocol. RNA-seq library was prepared with ribosomal depleted RNA using TruSeq RNA Sample Prep Kit v2 (Illumina) with modifications to preserve strand information as described.[80] Quality and size distribution of the prepared libraries was assessed with Experion™ DNA 1K Analysis Chips, and was used for molarity calculation. 8 RNA-seq libraries were barcoded using TruSeq RNA Sample Prep Kit v2 provided barcodes and pooled in equal molarities. 50bp single-end RNA-sequencing was performed at the Biomedical Sequencing Facility (http://biomedical-sequencing.at/BSF/) using Illumina HiSeq 2000.

KBM7 RNA-seq alignment

Raw RNA-seq data from each sample in fastq format was aligned using STAR[55] with default parameters: *STAR_2.3 –genomeDir hg19genome_no_splice_junction_database_-provided –readFilesIn [sample.fastq] –runThreadN 8 –genomeLoad NoSharedMemory –outFileNamePrefix [sample] –outStd SAM –outSAMmode Full.* Output was converted to BAM and sorted using *samtools* software. This resulted in average 35 million of uniquely mapped reads per sample with low standard deviation of 1.0 million reads.(Table S1D).

Differential gene expression analysis

RefSeq annotation downloaded from UCSC table browser on 27th January 2014 was used (filter: "name does match NM*," 36,734 isoforms, RefSeq_NM.gtf). Cuffdiff[59] (version 2.2.1) was used for expression level (FPKM) estimation and differential expression analysis with the following command: *cuffdiff RefSeq_NM.gtf -p 7 [replicates group1] [replicates group2] –labels [label group1], [label group2] –library-type fr-firststrand –mask-file pseudogenes.gtf.* The outputted list of significantly differentially expressed genes was additionally

filtered and only genes showing at least 3-fold change between non-truncated controls (WT2, WT3, C1, C2, replicate group1) and truncated cell lines (3kb1, 3kb2, 100kb1, 100kb2, replicate group2) were kept resulting in the list of 41 genes. Pairwise comparisons performed for further filtering: WT2, WT3 (replicate group1) versus C1, C2 (replicate group2) and 3kb1, 3kb2 (replicate group1) 100kb1, 100kb2 (replicate group2).

KBM7 cell lines clustering based on differential gene expression

Expression level (FPKM) of RefSeq protein coding genes was calculated in each of 8 samples separately using Cuffdiff (same command as above, no replicates). Expression of 41 significantly differentially expressed genes (Fig. 5A) or was used to perform unsupervised clustering of the samples. Heat map was built in R using *pheatmap* function with options *clustering_distance_cols= "canberra," clustering_distance_rows= "euclidean."*

Expression calculation and gene trap insertion analysis

GENCODE v19 lncRNA expression was calculated as RPKM (described above) separately for WT2 and WT3 cell lines. The average RPKM from both calculations was used in the figure. To determine the number of lncRNAs with gene trap insertion sites we downloaded cassette insertion sites from http://kbm7.genomebrowser.cemm.at/ in July 2015. Insertion sites can be updated and gene trap insertion sites used in this publication are available from http://opendata.cemm.at/barlowlab. Overlaps on the same strand with lncRNA annotations from GENCODE v19 were identified and overlapping annotations merged with *bedtools* software. GENCODE v19 lncRNA annotation was obtained at ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/release_19/gencode.v19.long_noncoding_RNAs.gtf.gz. To calculate position of gene trap insertions within the gene body we divided each GENCODE v19 lncRNA into 10 equally sized regions (numbered 1-10 starting at 5' end). Then we calculated the overlap of mapped gene trap insertion sites with these regions (bedtools) and created a sum of all insertions mapped to similar numbered regions.

Author contributions

A.E.K., D.P.B. and F.M.P. conceived the study and wrote the manuscript. I.V. discovered the *SLC38A4-AS* lncRNA and performed preliminary experiments characterizing this lncRNA. J.N. performed karyotype analysis and FISH. A.E.K and F.M.P performed DNA blots and PCR analyses. A.E.K. performed bioinformatic analysis, cell culture and RNA-seq.

Data access

Raw RNA-seq data from 8 KBM7 cell lines and the differential expression analysis output of Cuffdiff (Results, Fig. 5A) were deposited in NCBI's Gene Expression Omnibus [81] and are accessible through GEO Series accession number GSE71284 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE71284). Full *de novo* assembly in the 1Mb region around *SLC38A4-AS* lncRNA, RNA-seq signal in 8 sequenced KBM7 cell lines as well as location of gene trap insertion cassettes used in the study can be viewed in the related UCSC genome browser hub via https://opendata.cemm.at/barlowlab/.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

We thank Horizon Genomics GmbH for providing the KBM7 cell lines, Claudia Kerzendorfer and Sebastian Nijman for handling KBM7 cells and the Biomedical Sequencing Facility (http://biomedical-sequencing.at/) for advice and performing RNA sequencing. We thank Sara Sdelci, Quanah Hudson and Daniel Andergassen for technical assistance and useful discussions. We thank Quanah Hudson and Tilmann Bürckstümmer for advice and reading the manuscript. The study was partially supported by Austrian Science Fund [FWF F43-B09, FWF W1207-B09].

References

1. Ulitsky I, Bartel DP. lincRNAs: genomics, evolution, and mechanisms. Cell 2013; 154:26–46; PMID:23827673; http://dx.doi.org/10.1016/j.cell.2013.06.020
2. Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, Barrette TR, Prensner JR, Evans JR, Zhao S, et al. The landscape of long noncoding RNAs in the human transcriptome. Nat Genet 2015; 47(3):199–208; PMID:25599403; http://dx.doi.org/10.1007/82_2015_444
3. Cheetham SW, Gruhl F, Mattick JS, Dinger ME. Long noncoding RNAs and the genetics of cancer. Br J Cancer 2013; 108:2419–25; PMID:23660942; http://dx.doi.org/10.1038/bjc.2013.233
4. Batista PJ, Chang HY. Long noncoding RNAs: cellular address codes in development and disease. Cell 2013; 152:1298–307; PMID:23498938; http://dx.doi.org/10.1016/j.cell.2013.02.012
5. Prensner JR, Iyer MK, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC, Laxman B, Asangani IA, Grasso CS, Kominsky HD, et al. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nat Biotechnol 2011; 29:742–9; PMID:21804560; http://dx.doi.org/10.1038/nbt.1914
6. Roth A, Diederichs S. Long Noncoding RNAs in Lung Cancer. Curr Top Microbiol Immunol 2015; PMID:26037047

7. Meng L, Ward AJ, Chun S, Bennett CF, Beaudet AL, Rigo F. Towards a therapy for Angelman syndrome by targeting a long non-coding RNA. Nature 2015; 518:409–12; PMID:25470045; http://dx.doi.org/10.1038/nature13975
8. Wahlestedt C. Targeting long non-coding RNA to therapeutically upregulate gene expression. Nat Rev Drug Discov 2013; 12:433–46; PMID:23722346; http://dx.doi.org/10.1038/nrd4018
9. Roberts TC, Wood MJ. Therapeutic targeting of non-coding RNAs. Essays Biochem 2013; 54:127–45; PMID:23829532; http://dx.doi.org/10.1042/bse0540127
10. Quinodoz S, Guttman M. Long noncoding RNAs: an emerging link between gene regulation and nuclear organization. Trends Cell Biol 2014; 24:651–63; PMID:25441720; http://dx.doi.org/10.1016/j.tcb.2014.08.009
11. Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature 2010; 465:1033–8; PMID:20577206; http://dx.doi.org/10.1038/nature09144
12. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai MC, Hung T, Argani P, Rinn JL, et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature 2010; 464:1071–6; PMID:20393566; http://dx.doi.org/10.1038/nature08975
13. Wang KC, Yang YW, Liu B, Sanyal A, Corces-Zimmerman R, Chen Y, Lajoie BR, Protacio A, Flynn RA, Gupta RA, et al. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature 2011; 472:120–4; PMID:21423168; http://dx.doi.org/10.1038/nature09819
14. Latos PA, Pauler FM, Koerner MV, Senergin HB, Hudson QJ, Stocsits RR, Allhoff W, Stricker SH, Klement RM, Warczok KE, et al. Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. Science 2012; 338:1469–72; PMID:23239737; http://dx.doi.org/10.1126/science.1228110
15. Kornienko AE, Guenzl PM, Barlow DP, Pauler FM. Gene regulation by the act of long non-coding RNA transcription. BMC Biol 2013; 11:59; PMID:23721193; http://dx.doi.org/10.1186/1741-7007-11-59
16. Chu C, Spitale RC, Chang HY. Technologies to probe functions and mechanisms of long noncoding RNAs. Nat Struct Mol Biol 2015; 22:29–35; PMID:25565030; http://dx.doi.org/10.1038/nsmb.2921
17. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 2011; 25:1915–27; PMID:21890647; http://dx.doi.org/10.1101/gad.17446611
18. Cabili MN, Dunagin MC, McClanahan PD, Biaesch A, Padovan-Merhar O, Regev A, Rinn JL, Raj A. Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. Genome Biol 2015; 16:20; PMID:25630241; http://dx.doi.org/10.1186/s13059-015-0586-4
19. Tilgner H, Knowles DG, Johnson R, Davis CA, Chakrabortty S, Djebali S, Curado J, Snyder M, Gingeras TR, Guigó R. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. Genome Res 2012; 22:1616–25; PMID:22955974; http://dx.doi.org/10.1101/gr.134445.111
20. Seidl CI, Stricker SH, Barlow DP. The imprinted Air ncRNA is an atypical RNAPII transcript that evades splicing and escapes nuclear export. Embo J 2006; 25:3565–75; PMID:16874305; http://dx.doi.org/10.1038/sj.emboj.7601245
21. Kelley D, Rinn J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biol 2012; 13:R107; PMID:23181609; http://dx.doi.org/10.1186/gb-2012-13-11-r107
22. Hacisuleyman E, Goff LA, Trapnell C, Williams A, Henao-Mejia J, Sun L, McClanahan P, Hendrickson DG, Sauvageau M, Kelley DR, et al. Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. Nat Struct Mol Biol 2014; 21:198–206; PMID:24463464; http://dx.doi.org/10.1038/nsmb.2764
23. Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, Munson G, Young G, Lucas AB, Ach R, Bruhn L, et al. lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature 2011; 477:295–300; PMID:21874018; http://dx.doi.org/10.1038/nature10398
24. Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell 2011; 147:1537–50; PMID:22196729; http://dx.doi.org/10.1016/j.cell.2011.11.055
25. Tripathi V, Ellis JD, Shen Z, Song DY, Pan Q, Watt AT, Freier SM, Bennett CF, Sharma A, Bubulya PA, et al. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. Mol Cell 2010; 39:925–38; PMID:20797886; http://dx.doi.org/10.1016/j.molcel.2010.08.011
26. Ohrt T, Muetze J, Svoboda P, Schwille P. Intracellular localization and routing of miRNA and RNAi pathway components. Curr Top Med Chem 2012; 12:79–88; PMID:22196276; http://dx.doi.org/10.2174/156802612798919132
27. Sander JD, Joung JK. CRISPR-Cas systems for editing, regulating and targeting genomes. Nat Biotechnol 2014; 32:347–55; PMID:24584096; http://dx.doi.org/10.1038/nbt.2842
28. Sauvageau M, Goff LA, Lodato S, Bonev B, Groff AF, Gerhardinger C, Sanchez-Gomez DB, Hacisuleyman E, Li E, Spence M, et al. Multiple knockout mouse models reveal lincRNAs are required for life and brain development. Elife 2013; 2:e01749; PMID:24381249; http://dx.doi.org/10.7554/eLife.01749
29. Yin Y, Yan P, Lu J, Song G, Zhu Y, Li Z, Zhao Y, Shen B, Huang X, Zhu H, et al. Opposing roles for the lncRNA haunt and its genomic locus in regulating hoxa gene activation during embryonic stem cell differentiation. Cell Stem Cell 2015; 16:504–16; PMID:25891907; http://dx.doi.org/10.1016/j.stem.2015.03.007
30. Zhang B, Arun G, Mao YS, Lazar Z, Hung G, Bhattacharjee G, Xiao X, Booth CJ, Wu J, Zhang C, et al. The lncRNA Malat1 is dispensable for mouse development but its transcription plays a cis-regulatory role in the adult. Cell Reports 2012; 2:111–23; PMID:22840402; http://dx.doi.org/10.1016/j.celrep.2012.06.003

31. Han J, Zhang J, Chen L, Shen B, Zhou J, Hu B, Du Y, Tate PH, Huang X, Zhang W. Efficient in vivo deletion of a large imprinted lncRNA by CRISPR/Cas9. RNA Biol 2014; 11:829–35; PMID:25137067; http://dx.doi.org/10.4161/rna.29624
32. Bassett AR, Akhtar A, Barlow DP, Bird AP, Brockdorff N, Duboule D, Ephrussi A, Ferguson-Smith AC, Gingeras TR, Haerty W, et al. Considerations when investigating lncRNA function in vivo. Elife 2014; 3:e03058; PMID:25124674; http://dx.doi.org/10.7554/eLife.03058
33. Skarnes WC, von Melchner H, Wurst W, Hicks G, Nord AS, Cox T, Young SG, Ruiz P, Soriano P, Tessier-Lavigne M, et al. A public gene trap resource for mouse functional genomics. Nat Genet 2004; 36:543–4; PMID:15167922; http://dx.doi.org/10.1038/ng0604-543
34. Stanford WL, Cohn JB, Cordes SP. Gene-trap mutagenesis: past, present and beyond. Nat Rev Genet 2001; 2:756–68; PMID:11584292; http://dx.doi.org/10.1038/35093548
35. Schuster-Gossler K, Simon-Chazottes D, Guenet JL, Zachgo J, Gossler A. Gtl2lacZ, an insertional mutation on mouse chromosome 12 with parental origin-dependent phenotype. Mamm Genome 1996; 7:20–4; PMID:8903723; http://dx.doi.org/10.1007/s003359900006
36. Barlow DP, Bartolomei MS. Genomic imprinting in mammals. Cold Spring Harb Perspect Biol 2014; 6:a018382; PMID:24492710; http://dx.doi.org/10.1101/cshperspect.a018382
37. Kanduri C. Long noncoding RNAs: Lessons from genomic imprinting. Biochim Biophys Acta 2015; PMID:26004516; http://dx.doi.org/10.1016/j.bbagrm.2015.05.006
38. da Rocha ST, Edwards CA, Ito M, Ogata T, Ferguson-Smith AC. Genomic imprinting at the mammalian Dlk1-Dio3 domain. Trends Genet 2008; 24:306–16; PMID:18471925; http://dx.doi.org/10.1016/j.tig.2008.03.011
39. Benetatos L, Vartholomatos G, Hatzimichael E. DLK1-DIO3 imprinted cluster in induced pluripotency: landscape in the mist. Cell Mol Life Sci 2014; 71:4421–30; PMID:25098353; http://dx.doi.org/10.1007/s00018-014-1698-9
40. Sleutels F, Zwart R, Barlow DP. The non-coding Air RNA is required for silencing autosomal imprinted genes. Nature 2002; 415:810–3; PMID:11845212; http://dx.doi.org/10.1038/415810a
41. Mancini-Dinardo D, Steele SJ, Levorse JM, Ingram RS, Tilghman SM. Elongation of the Kcnq1ot1 transcript is required for genomic imprinting of neighboring genes. Genes Dev 2006; 20:1268–82; PMID:16702402; http://dx.doi.org/10.1101/gad.1416906
42. Meng L, Person RE, Huang W, Zhu PJ, Costa-Mattioli M, Beaudet AL. Truncation of Ube3a-ATS unsilences paternal Ube3a and ameliorates behavioral defects in the Angelman syndrome mouse model. PLoS Genet 2013; 9:e1004039; PMID:24385930; http://dx.doi.org/10.1371/journal.pgen.1004039
43. Shin JY, Fitzpatrick GV, Higgins MJ. Two distinct mechanisms of silencing by the KvDMR1 imprinting control region. Embo J 2008; 27:168–78; PMID:18079696; http://dx.doi.org/10.1038/sj.emboj.7601960
44. Gutschner T, Baas M, Diederichs S. Noncoding RNA gene silencing through genomic integration of RNA destabilizing elements using zinc finger nucleases. Genome Res 2011; 21:1944–54; PMID:21844124; http://dx.doi.org/10.1101/gr.122358.111
45. Burckstummer T, Banning C, Hainzl P, Schobesberger R, Kerzendorfer C, Pauler FM, Chen D, Them N, Schischlik F, Rebsamen M, et al. A reversible gene trap collection empowers haploid genetics in human cells. Nat Methods 2013; 10:965–71; PMID:24161985; http://dx.doi.org/10.1038/nmeth.2609
46. Santoro F, Mayer D, Klement RM, Warczok KE, Stukalov A, Barlow DP, Pauler FM. Imprinted Igf2r silencing depends on continuous Airn lncRNA expression and is not restricted to a developmental window. Development 2013; 140:1184–95; PMID:23444351; http://dx.doi.org/10.1242/dev.088849
47. Andersson BS, Beran M, Pathak S, Goodacre A, Barlogie B, McCredie KB. Ph-positive chronic myeloid leukemia with near-haploid conversion in vivo and establishment of a continuously growing cell line with similar cytogenetic pattern. Cancer Genet Cytogenet 1987; 24:335–43; PMID:3466682; http://dx.doi.org/10.1016/0165-4608(87)90116-6
48. Schuster-Gossler K, Bilinski P, Sado T, Ferguson-Smith A, Gossler A. The mouse Gtl2 gene is differentially expressed during embryonic development, encodes multiple alternatively spliced transcripts, and may act as an RNA. Dev Dyn 1998; 212:214–28; PMID:9626496; http://dx.doi.org/10.1002/(SICI)1097-0177(199806)212:2%3c214::AID-AJA6%3e3.0.CO;2-K
49. Vlatkovic I. PhD thesis: Mapping and characterization of macro non-protein coding RNAs in human imprinted gene regions, University of Vienna; available for download at http://othes.univie.ac.at/12494/1/2010-09-01_0642621.pdf 2010
50. Smith RJ, Dean W, Konfortova G, Kelsey G. Identification of novel imprinted genes in a genome-wide screen for maternal methylation. Genome Res 2003; 13:558–69; PMID:12670997; http://dx.doi.org/10.1101/gr.781503
51. Andergassen D, Dotter CP, Kulinski TM, Guenzl PM, Bammer PC, Barlow DP, Pauler FM, Hudson QJ. Allelome.PRO, a pipeline to define allele-specific genomic features from high-throughput sequencing data. Nucleic Acids Res 2015; 43(21): e146; PMID:26202974
52. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM, et al. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res 2014; 42:D756–63; PMID:24259432; http://dx.doi.org/10.1093/nar/gkt1114
53. Mattick JS, Rinn JL. Discovery and annotation of long noncoding RNAs. Nat Struct Mol Biol 2015; 22:5–7; PMID:25565026; http://dx.doi.org/10.1038/nsmb.2942
54. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res 2012; 22:1760–74; PMID:22955987; http://dx.doi.org/10.1101/gr.135350.111
55. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013; 29:15–21; PMID:23104886; http://dx.doi.org/10.1093/bioinformatics/bts635
56. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, et al. The GENCODE v7 catalog of human long noncoding

RNAs: analysis of their gene structure, evolution, and expression. Genome Res 2012; 22:1775–89; PMID:22955988; http://dx.doi.org/10.1101/gr.132159.111
57. Curinha A, Oliveira Braz S, Pereira-Castro I, Cruz A, Moreira A. Implications of polyadenylation in health and disease. Nucleus 2014; 5:508–19; PMID:25484187; http://dx.doi.org/10.4161/nucl.36360
58. Brockdorff N, Ashworth A, Kay GF, McCabe VM, Norris DP, Cooper PJ, Swift S, Rastan S. The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. Cell 1992; 71:515–26; PMID:1423610; http://dx.doi.org/10.1016/0092-8674(92)90519-I
59. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 2012; 7:562–78; PMID:22383036; http://dx.doi.org/10.1038/nprot.2012.016
60. Gregory TR. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biol Rev Camb Philos Soc 2001; 76:65–101; PMID:11325054; http://dx.doi.org/10.1017/S1464793100005595
61. Huang R, Jaritz M, Guenzl P, Vlatkovic I, Sommer A, Tamir IM, Marks H, Klampfl T, Kralovics R, Stunnenberg HG, et al. An RNA-Seq strategy to detect the complete coding and non-coding transcriptome including full-length imprinted macro ncRNAs. PLoS One 2011; 6:e27288; PMID:22102886; http://dx.doi.org/10.1371/journal.pone.0027288
62. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol 2013; 31:46–53; PMID:23222703; http://dx.doi.org/10.1038/nbt.2450
63. Kowalczyk MS, Higgs DR, Gingeras TR. Molecular biology: RNA discrimination. Nature 2012; 482:310–1; PMID:22337043; http://dx.doi.org/10.1038/482310a
64. Wutz A. Gene silencing in X-chromosome inactivation: advances in understanding facultative heterochromatin formation. Nat Rev Genet 2011; 12:542–53; PMID:21765457; http://dx.doi.org/10.1038/nrg3035
65. Alen C, Kent NA, Jones HS, O'Sullivan J, Aranda A, Proudfoot NJ. A role for chromatin remodeling in transcriptional termination by RNA polymerase II. Mol Cell 2002; 10:1441–52; PMID:12504018; http://dx.doi.org/10.1016/S1097-2765(02)00778-5
66. Luco RF, Misteli T. More than a splicing code: integrating the role of RNA, chromatin and non-coding RNA in alternative splicing regulation. Curr Opin Genet Dev 2011; 21:366–72; PMID:21497503; http://dx.doi.org/10.1016/j.gde.2011.03.004
67. Weber EL, Cannon PM. Promoter choice for retroviral vectors: transcriptional strength vs. trans-activation potential. Hum Gene Ther 2007; 18:849–60; PMID:17767401; http://dx.doi.org/10.1089/hum.2007.067
68. Zwart R, Sleutels F, Wutz A, Schinkel AH, Barlow DP. Bidirectional action of the Igf2r imprint control element on upstream and downstream imprinted genes. Genes Dev 2001; 15:2361–6; PMID:11562346; http://dx.doi.org/10.1101/gad.206201
69. Nagano T, Mitchell JA, Sanz LA, Pauler FM, Ferguson-Smith AC, Feil R, Fraser P. The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. Science 2008; 322:1717–20; PMID:18988810; http://dx.doi.org/10.1126/science.1163802
70. Babak T, DeVeale B, Tsang EK, Zhou Y, Li X, Smith KS, Kukurba KR, Zhang R, Li JB, van der Kooy D, et al. Genetic conflict reflected in tissue-specific maps of genomic imprinting in human and mouse. Nat Genet 2015; 47:544–9; PMID:25848752; http://dx.doi.org/10.1038/ng.3274
71. Baran Y, Subramaniam M, Biton A, Tukiainen T, Tsang EK, Rivas MA, Pirinen M, Gutierrez-Arcelus M, Smith KS, Kukurba KR, et al. The landscape of genomic imprinting across diverse adult human tissues. Genome Res 2015; 25:927–36; PMID:25953952; http://dx.doi.org/10.1101/gr.192278.115
72. Charrin S, Jouannet S, Boucheix C, Rubinstein E. Tetraspanins at a glance. J Cell Sci 2014; 127:3641–8; PMID:25128561; http://dx.doi.org/10.1242/jcs.154906
73. Karlsson G, Rorby E, Pina C, Soneji S, Reckzeh K, Miharada K, Karlsson C, Guo Y, Fugazza C, Gupta R, et al. The tetraspanin CD9 affords high-purity capture of all murine hematopoietic stem cells. Cell Rep 2013; 4:642–8; PMID:23954783; http://dx.doi.org/10.1016/j.celrep.2013.07.020
74. Carette JE, Pruszak J, Varadarajan M, Blomen VA, Gokhale S, Camargo FD, Wernig M, Jaenisch R, Brummelkamp TR. Generation of iPSCs from cultured human malignant cells. Blood 2010; 115:4039–42; PMID:20233975; http://dx.doi.org/10.1182/blood-2009-07-231845
75. Akutsu H, Miura T, Machida M, Birumachi J, Hamada A, Yamada M, Sullivan S, Miyado K, Umezawa A. Maintenance of pluripotency and self-renewal ability of mouse embryonic stem cells in the absence of tetraspanin CD9. Differentiation 2009; 78:137–42; PMID:19716222; http://dx.doi.org/10.1016/j.diff.2009.08.005
76. Kennaway DJ. Clock genes at the heart of depression. J Psychopharmacol 2010; 24:5–14; PMID:20663803; http://dx.doi.org/10.1177/1359786810372980
77. Baek SH, Kim KI. Emerging roles of orphan nuclear receptors in cancer. Annu Rev Physiol 2014; 76:177–95; PMID:24215441; http://dx.doi.org/10.1146/annurev-physiol-030212-183758
78. Wang A, Huang K, Shen Y, Xue Z, Cai C, Horvath S, Fan G. Functional modules distinguish human induced pluripotent stem cells from embryonic stem cells. Stem Cells Dev 2011; 20:1937–50; PMID:21542696; http://dx.doi.org/10.1089/scd.2010.0574
79. Carette JE, Guimaraes CP, Wuethrich I, Blomen VA, Varadarajan M, Sun C, Bell G, Yuan B, Muellner MK, Nijman SM, et al. Global gene disruption in human cells to assign genes to phenotypes by deep sequencing. Nat Biotechnol 2011; 29:542–6; PMID:21623355; http://dx.doi.org/10.1038/nbt.1857
80. Sultan M, Dokel S, Amstislavskiy V, Wuttig D, Sultmann H, Lehrach H, Yaspo ML. A simple strand-specific RNA-Seq library preparation protocol combining the Illumina TruSeq RNA and the dUTP methods. Biochem Biophysical Res Commun 2012; 422:643–6; PMID:22609201; http://dx.doi.org/10.1016/j.bbrc.2012.05.043

81. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002; 30:207-10; PMID:11752295; http://dx.doi.org/10.1093/nar/30.1.207

82. Forrest AR, Kawaji H, Rehli M, Baillie JK, de Hoon MJ, Haberle V, Lassmann T, Kulakovskiy IV, Lizio M, Itoh M, et al. A promoter-level mammalian expression atlas. Nature 2014; 507:462–70; PMID:24670764; http://dx.doi.org/10.1038/nature13182

3 DISCUSSION

3.1 General discussion

3.1.1 Overview

With the increasing number of reports on lncRNAs playing important roles in disease, the appreciation of the long-standing goal to annotate all lncRNAs in the human genome is increasing. The ultimate broad goal of the field is to first define all the lncRNA genes in the human genome and, in parallel, to define lncRNA landscapes of all the cell types, differentiation states and disease states. Since an altered lncRNA landscape might be indicative of the disease state, mapping of all the above mentioned lncRNA landscapes is of high importance to assess if lncRNAs might serve as successful biomarkers. This, plus a deep understanding of lncRNA biology, function and underlying molecular mechanisms, would form the basis for the successful therapeutic application of lncRNAs – the second ultimate goal of the lncRNA research. However, overall, the young field of lncRNA research is only at the foundation of these ambitious goals.

In this Doctoral Thesis I created a ready-to-use human primary granulocyte lncRNA annotation by an annotation pipeline that I established and validated. Using granulocytes as a model system, I characterized their transcriptome by analyzing several known features of lncRNAs, such as high tissue-specificity, low expression, inefficient splicing and polyadenylation, which distinguish them from mRNAs. My data showed that these features negatively influence lncRNA identification efficiency and reduce their representation in public lncRNA reference annotations. My analysis of a granulocyte transcriptome from several individuals replicated three times, allowed discovery of a novel non-mRNA-like feature – an unexpectedly high natural variation of lncRNA expression. I found that high expression variation, as well as other non-mRNA-like features, confounds lncRNA identification and reduces their representation in public annotations. I confirmed this novel finding in LCL and eight human tissues, showing that this was a general phenomenon in human body tissues. Further investigation showed that high lncRNA expression variability makes it necessary to analyze a vast number of healthy individuals, which exceeds the 120 donors analyzed in Publication 2, to find all the lncRNA genes in the human genome.

I additionally performed a functional assessment of one of lncRNAs expressed in granulocytes. This lncRNA was previously discovered by Irena Vlatkovic (Vlatkovic, 2010b) in the lab and later also annotated by RefSeq (Pruitt et al, 2014), but remained an uncharacterized lncRNA. I first characterized this lncRNA expression and non-mRNA features including nuclear localization of its unspliced isoforms showing that this lncRNA has an unusual RNA biology. I used publically available RNA-seq data from various tissues to show that this lncRNA is twice as long as annotated and is transcribed over more than 500kb and thus, represents one of the longest lncRNAs so far identified. Since this lncRNA overlaps in antisense orientation the *SLC38A4* gene, we have renamed it as *SLC38A4-AS* in agreement with current guidelines. To assess the function of *SLC38A4-AS* I used three independent cell lines from the human haploid knock-out collection, established from the malignant myeloid KBM7 cell line (Burckstummer et al, 2013), that contained stop signals within the body of *SLC38A4-AS* lncRNA. This shows that *SLC38A4-AS*, in contrast to *Airn*, did not affect genes *in cis*, but instead robustly regulated a stringent set of 6 genes on different chromosomes, including the differentiation relevant *CD9* and *RORB* protein-coding genes.

Thus, this Doctoral Thesis provides multiple novel insights into RNA biology of lncRNAs in the human genome. It identifies a novel important feature of high inter-individual variability of lncRNA expression, which provides crucial guidelines for lncRNA gene identification and medical applications, and also identifies a novel functional lncRNA – *SLC38A4-AS*. This part of the Doctoral Thesis also provides guidelines for efficient RNA biology and functional lncRNA characterization using a powerful ready-to-use haploid KBM7 knock-out collection containing knock-out cells for hundreds of uncharacterized lncRNAs.

3.1.2 The first annotation of a human primary granulocyte transcriptome

We used granulocyte PolyA+ RNA-seq from ten healthy adult individuals of various ages to establish a previously undefined human primary granulocyte transcriptome. Transcriptome assembly and lncRNA identification required optimization and resulted in a well-established procedure, which was later automated to allow running numerous *de novo* lncRNA and mRNA identification processes in parallel using different numbers of LCL donors as an input (See Figure 7 in Publication 2). We took special care to remove

potential artifacts as well as potential non-annotated protein-coding genes from our granulocyte lncRNA annotation by performing multiple filtering steps, which included expression, repeat content and exon length cutoffs, elimination of potential leak-through transcripts annotated on a wrong strand, and protein-coding potential calculation. We annotated 6,249 lncRNA transcripts that all together formed 1,529 lncRNA loci of which 46% were, surprisingly, not present in the three most commonly used public annotations. Moreover, along with identification of new lncRNA loci, we also identified new isoforms within known lncRNA loci which constituted about a third of all our lncRNA transcripts. Thus, only a third of our granulocyte lncRNA transcripts were in the three public annotations. This means, importantly, that granulocytes express a unique set of lncRNAs both in terms of genomic regions and isoform structure. Our observations, together with the known tissue specificity of lncRNA expression and exon structure, suggest that using integrative lncRNA annotations such as those previously described (Cabili et al, 2011; Iyer et al, 2015) might not be beneficial when analyzing a particular cell type. Instead, our results indicate that a *de novo* lncRNA annotation in the cell type of interest might be required to obtain full awareness of the transcriptional landscape in the studied system.

The granulocyte *de novo* lncRNA, as well as mRNA, transcriptomes we created in this study are openly available (http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0873-8) and might be of immediate use and interest to the wide granulocyte research community.

Having identified numerous novel lncRNAs, we validated our granulocyte annotation first by demonstrating that 80% of the 1,591 *de novo* annotated lncRNA loci were overlapped by ESTs or RNAs in public databases. Second, by experimentally verifying 42 splice sites from 22 of the 736 not annotated lncRNA loci. And third, by cloning transcripts from 18 new lncRNA loci. This demonstrated the reliability of our annotation and showed that the newly identified loci are unlikely to be artifacts. With the ever-increasing number of annotated lncRNA genes, it was important for us to gain confidence in the annotation prior to making conclusions about the features of new lncRNA loci, transcripts and isoforms. In addition, our validation of the granulocyte lncRNA annotation provides further confidence to other potential users of this gene annotation.

We aimed to further validate our *de novo* annotation by assessing how well its exons are covered by the three public annotations and MiTranscriptome. We first showed that *de novo* granulocyte mRNAs were equally well covered by both public and the recently published MiTranscriptome mRNAs that were annotated from thousands of RNA-seq samples from various tissues and donors (>77% exonic coverage with MiTranscriptome showing 93% coverage). The ~15% difference between MiTranscriptome and public annotations might indicate identification of protein-coding gene extensions in both our *de novo* and MiTranscriptome annotations. In contrast, exonic coverage of lncRNAs by RefSeq, GENCODE v19 or Cabili *et al.,* annotations was dramatically poorer than that of MiTranscriptome (median of 3.2%, 0%, 0% and 48.7% respectively). This indicated that public annotations missed many lncRNAs or annotated them incompletely, while MiTranscriptome contained more granulocyte-like lncRNA isoforms. We then investigated the reasons for the under-representation of lncRNAs in public annotations and showed (see below) that it is the non-mRNA-like nature of lncRNAs that makes their identification challenging.

3.1.3 Non-mRNA-like features of lncRNAs confound their annotation

We confirmed the four known non-mRNA-like features of lncRNAs – high tissue specificity and low expression level (Cabili et al, 2011), inefficient splicing (Tilgner et al, 2012) and low polyadenylation (Wilusz, 2015) – for *de novo* annotated granulocyte lncRNAs. This additionally validated our lncRNA annotation and provided important RNA biology characteristics of the granulocyte lncRNA transcriptome, which will be useful for further studies of granulocyte lncRNAs. We also showed that the four non-mRNA-like features are more prominent in lncRNA loci or transcripts newly identified in our study and absent from public annotations, suggesting these features might confound identification. Importantly, we confirmed these trends in granulocyte RNA-seq data using an independent lncRNA and mRNA annotation from MiTranscriptome (Iyer et al, 2015) and in LCL RNA-seq dataset using *de novo* LCL lncRNA and mRNA annotation, which both contained multiple novel lncRNAs. High tissue specificity, low expression level, low splicing efficiency and low polyadenylation likely confound lncRNA identification by, overall, reducing the number of spliced lncRNA transcripts in the PolyA+ RNA-seq date commonly used for identification and thus lowering the probability of the detection of a lncRNA that displays these non-mRNA-like features.

3.1.3.1 *Low expression*

Granulocyte lncRNAs displayed notably lower median expression (RPKM) level in granulocyte RNA-seq data than mRNAs: 10-fold lower in the PolyA+ fraction (0.65 vs 6.14 respectively) and 7-fold lower in the total fraction (0.31 vs. 2.18 respectively). While not focusing on granulocytes as we did, the public annotations did use whole blood for their analyses, thus they could, in principle, identify and annotate granulocyte lncRNAs. However, if a lncRNA is very lowly expressed, the chance of identifying it decreases. We showed that novel granulocyte lncRNA transcripts not present in public annotations were more lowly expressed than those in public annotations with a 2-fold median expression difference (0.47 vs 1.00 respectively). We observed the same trend for LCL *de novo* lncRNAs that showed ~13-fold lower median expression in LCL PolyA+ RNA-seq than mRNAs (0.66 vs. 9.05), with 'not in public annotation' transcripts showing 2-fold lower median expression than those in public annotations transcripts (0.38 vs. 0.83). Identification of novel lncRNAs in LCL is very interesting by itself, because LCL were one of the main cell lines analyzed by the ENCODE project. In spite of this fact, we found 1,075 LCL lncRNA loci (41%) that were not covered by public annotations. Moreover, 36% of LCL lncRNA transcripts added a new isoform to an annotated locus. This might be explained by high coverage RNA-seq data we used for the LCL analysis (total of 522 million uniquely mapped reads) allowing us to identify lowly expressed novel lncRNAs, however, it seems likely that the fact that we used 20 donors for the LCL lncRNA identification, while ENCODE focused on the GM12878 cell line (LCL established from one donor), played a significant role in our discovery of multiple novel lncRNA genes (see discussion below – 3.1.7 *Implications of increased variation of lncRNAs*).

While it is well-known and widely accepted that lncRNAs are generally more lowly expressed than mRNAs, the reason for this difference is to date unclear. It is important to note, that while lncRNAs generally show low expression level, the difference within this class of genes is outstanding and lncRNA expression ranges from marginal to the highest expressing transcripts in the cell (such as *MALAT1* and *NEAT1*). It is possible that low expression level of lncRNAs is defined by some unknown characteristics of their promoters that might affect transcription machinery assembly or modification of carboxy-terminal (CTD) domain of RNAPII, known to be responsible for modulating most RNAPII functions (Brookes & Pombo, 2012). It is also possible that lncRNAs just appear

to be generally lowly abundant because of their reduced stability (Clark et al, 2012), while being transcribed at a similar rate as mRNAs. Another possible explanation of low lncRNA abundance is that many of them are transcribed from enhancers (Arner et al, 2015; Orom et al, 2010) and thus *a priori* have a different "promoter" structure and characteristics. Importantly, while low expression might argue for "transcriptional noise" criticism of lncRNAs (Kowalczyk et al, 2012; Palazzo & Lee, 2015) and might be argued to be caused by no evolutionary selection and the absence of regulation and function, it is known that many functional lncRNAs are very lowly expressed (Quinn & Chang, 2015). Moreover, lncRNAs that act *in cis* might be meaningfully void of high expression in order to only act at the site of their transcription (Kornienko et al, 2013; Wang et al, 2011b), which is utterly different for mRNAs that have to be expressed at a sufficient level to reach the cytoplasm and be engaged by ribosomes.

3.1.3.2 Tissue specificity

The tissue specificity analysis showed that approximately every third *de novo* granulocyte lncRNA, but only every 25th mRNA, was not expressed or was dramatically lower (>3-fold) expressed in all the 34 analyzed cell types (for both transcript and locus expression level). Importantly, we only used high coverage (average 242 million reads) public RNA-seq data from various cell types thus minimizing the chance of not detecting a lowly expressed lncRNA that is not granulocyte-specific. Not unexpectedly, cell types showing highest number of *de novo* granulocyte lncRNAs expressed were MNC and B cells (Figure S9A in Additional File 1, Publication 2), i.e. blood cell types. Additional support that was provided in our study concerning lncRNA tissue specificity, comes from *de novo* lncRNA annotation in LCL, which shows that the overlap between granulocytes and LCL *de novo* annotations is notably smaller for lncRNAs than for mRNAs: only 21% of LCL lncRNA, but 76% of LCL mRNA loci are also found in granulocytes.

Similarly to low expression level, granulocyte specificity of the *de novo* granulocyte lncRNAs identified in the study could prevent them from being identified in annotation projects that did not focus on granulocytes, or did not achieve high enough coverage to assemble very lowly expressed and at the same time tissue-specific lncRNAs. We separately analyzed tissue specificity of the three novelty classes of lncRNA transcripts and found that they differed dramatically – with 'in public annotations' transcripts

containing 5-fold less granulocyte-specific transcripts than ‘not public annotations’ (11% vs 57%). This showed that our annotation, and in particular novel lncRNAs, is highly enriched in granulocyte-specific lncRNAs, while public annotations are depleted for what we identify to be granulocyte-specific lncRNAs (GENCODE v19 – 9% granulocyte-specific transcripts).

Why lncRNAs are so much more tissue-specific than protein-coding genes, as well as how this tissue specificity is achieved, is to date unknown. It is likely, that if lncRNAs are functional gene regulators contributing to differentiation and the establishment of cell identity (Fatica & Bozzoni, 2014; Qureshi & Mehler, 2013), their expression has to be more restricted to a certain tissue/cell type, than mRNA expression. It is known that enhancers act tissue-specifically (Shlyueva et al, 2014) and thus lncRNAs transcribed from active enhances should also be tissue-specific (Orom et al, 2010). However, some researchers argue that high lncRNA tissue specificity might not be meaningful and tissue-specifically expressed lncRNAs be just by-products of tissue-specific chromatin organization (Ulitsky & Bartel, 2013).

3.1.3.3 Inefficient processing

Using our PolyA+ and ribosomal-depleted granulocyte RNA-seq data and *de novo* lncRNA/mRNA annotation in granulocytes we showed that processing, such as splicing and polyadenylation, is significantly less efficient for lncRNA genes compared to mRNA genes.

We estimated polyadenylation efficiency by calculating PolyA+ enrichment as a ratio between transcript abundance in PolyA+ and ribosomal-depleted RNA fractions as assessed by RNA-seq. If a transcript is efficiently polyadenylated then PolyA tail selection would pull these transcripts into the PolyA+ RNA fraction, while non-polyadenylated transcripts would not be selected, unless they had abundant PolyA stretches in their gene body (Furuno et al, 2006). Thus, if a transcript is efficiently polyadenylated, its relative abundance (RPKM) in the PolyA+ fraction would be higher than that in the ribosomal-depleted fraction. In contrast, an inefficiently polyadenylated transcript would not get into the PolyA+ fraction and its RPKM in the ribosomal-depleted fraction would be higher. We show that lncRNAs are notably less efficiently

polyadenylated with a median PolyA+ enrichment level of 1.6 compared to median 2.6 for mRNAs. Notably, we also show that novel lncRNAs are even more non-mRNA-like with 1.3 median PolyA+ enrichment level. Thus polyadenylation efficiency directly affects the abundance of a transcript in the PolyA+ RNA-seq used for most transcriptome assembly and identification strategies would reduce the efficiency of lncRNA identification.

We calculated the splicing efficiency of every lncRNA and mRNA splice site and then defined transcript splicing efficiency by the splicing efficiency of its best processed splice site. LncRNAs showed a median of 88.13% splicing efficiency albeit with a broad distribution, while mRNAs were uniformly well-spliced with a median splicing efficiency of 99.02%. Inefficient lncRNA splicing was first reported for the imprinted lncRNAs *Airn* and *Ube3a-ats* (Meng et al, 2012; Seidl et al, 2006). Reduced efficiency of co-transcriptional splicing for lncRNAs compared to mRNAs was reported on a genome-wide level in human K562 cell line (Tilgner et al, 2012). However, steady state lncRNA splicing efficiency in a human primary tissue, as assessed in this Doctoral Thesis, was not previously studied. We confirm our observation of low lncRNA splicing efficiency showing that the lncRNA-mRNA difference is more pronounced when analyzing splicing on a locus level, *i.e.,* disregarding the number of transcripts in the locus. As with other non-mRNA-like features, reduced splicing makes lncRNA identification more challenging by reducing the abundance of spliced isoforms in the PolyA+ fraction – the isoforms that are the main material for transcriptome *de novo* assembly in our and other (Cabili et al, 2011; Iyer et al, 2015) pipelines. We show that novel lncRNAs are less efficiently spliced than those present in the public lncRNA annotations in both transcript and locus level analyses. While all lncRNAs showed 10.9% median splicing efficiency reduction compared to mRNAs, novel (i.e., 'not in public annotations') granulocyte lncRNA transcripts showed 22.9% reduction. We also analyzed splicing efficiency of MiTranscriptome mRNA and lncRNA transcripts and confirmed the observations above.

It is not clear why lncRNAs are inefficiently processed and what defines this inefficiency. Inefficient processing includes splicing and polyadenylation analyzed in Publication 2, but may also include the efficiency and consistency of transcript termination. For example, well-known imprinted lncRNAs such as *Airn* and *Kcnq1ot1* that were shown to be terminated at different positions, and thus be of a different length, in different tissues

(Huang et al, 2011). All these processing features might be controlled by the modification of CTD of the RNAPII (Brookes & Pombo, 2012; Darnell, 2013) and thus the difference in the processing of lncRNAs might be indicative of lncRNA possibly being transcribed by the RNAPII with modifications of CTD different from those of the mRNAs. However, no report of such difference has emerged so far and this is an interesting topic for future investigations.

Overall, we showed that granulocyte lncRNAs display several known features distinguishing them from mRNAs. We show that our new lncRNAs identified in a very high coverage dataset from a relatively pure primary cell type display these non-mRNA-like features to a higher extent than granulocyte lncRNAs that have already been identified. The consistent finding that lncRNAs not in public annotations show increased non-mRNA-like features supports our claim in Publication 2 that the unique RNA biology features of lncRNAs make their identification more challenging. As discussed above, in spite of these features being known and generally accepted to distinguish a bulk of lncRNAs from mRNAs, the molecular and genetic mechanisms of lncRNAs being different from mRNAs in these features is unknown.

3.1.4 LncRNAs show lower histone mark coverage than mRNAs

Analysis of the histone modifications on granulocyte lncRNAs and mRNAs revealed, surprisingly, differences in the histone modification patterns of these two classes of genes. LncRNA promoters are notably poorer in H3K4me3 mark, which classically defines active promoters, than mRNA promoters. In spite of an intuitive assumption that low lncRNA expression level might be responsible for this difference, we found that even highly expressed lncRNAs (such as in bin 4 and 5) show significantly less H3K4me3. Similarly, we found that lncRNA exons and full gene bodies (loci) are less covered, compared to mRNAs, with H3K36me3, which classically covers the gene body of an actively transcribed gene, and this difference is also persistent in all expression bins. Additionally, we found that lncRNAs, as a population, show generally higher level of H3K4me1, which classically marks active enhancers and is absent from active promoters, over their promoters, exons and gene bodies than do mRNAs. The latter is likely to arise from the fact that many lncRNAs are transcribed from enhancers (Orom et al, 2010). This might explain the increased H3K4me1 on lncRNA promoters, while the increased

H3K4me1 on lncRNA exons and gene bodies is more likely to arise from the fact that most lncRNAs, especially enhancer RNAs (Orom et al, 2010), are relatively short with longer exons than mRNAs, thus the mark might occupy a significant part of the lncRNA body/exon length. The reduced coverage of active transcription marks, such as H3K4me3 at promoters and H3K36me3 on gene bodies and exons, for lncRNAs compared to mRNAs appears to be a surprising finding. It is unclear what features of lncRNA genes might be responsible for this difference and has to be investigated further.

3.1.5 High expression variability is a novel non-mRNA like general feature of lncRNAs

The major novel finding in this Doctoral Thesis, described in Publication 2, arose from an analysis of granulocyte lncRNA expression variability in different healthy individuals and comparison to that of mRNAs, which revealed an unexpectedly high lncRNA variability. The median level of lncRNA expression variability in granulocytes (defined as normalized standard deviation, also known as the coefficient of variance) was twice higher than that of mRNAs (0.29 and 0.15 respectively). This result was surprising, particularly when we further analyzed the significance of variability using the ANOVA test and found that approximately every 4th granulocyte lncRNA transcript is differentially expressed even among seven donors, i.e., in a small number of Caucasian donors. In contrast, approximately every 24th mRNA was differentially expressed in these seven donors, which is consistent with previous reports on protein-coding gene variation in primary tissues (Chowers et al, 2003). While normalized standard deviation calculation provided us with an estimation of the range of expression variability for mRNAs and lncRNAs, assigning significance to this variability was essential and also provided us with a clear number of differentially expressed (highly variable) granulocyte lncRNAs and mRNAs.

3.1.5.1 Usage of replicates

Importantly, our experimental design, unlike the experimental designs of many previous gene expression variation studies (Gonzalez-Porta et al, 2012; Lappalainen et al, 2013), included replicate samples for every individual analyzed, which provided two crucial advantages. First, it allowed us to account for intra-individual variability, which may be attributed to various factors independent of genuine variability between the donors

(Whitney et al, 2003). By calculating expression for each donor as a mean of that for three replicates we reduced intra-individual variability influencing our results. Interestingly, both the public datasets we used for validation of the increase lncRNA expression variability finding, showed higher absolute expression variability values than granulocyte analysis, for both lncRNAs and mRNAs. This might be caused by both higher number of individuals analyzed or by the absence of replicates. The second advantage of our experimental design including replicates is that it enabled an ANOVA test and the assignment of statistical significance to the variation as described above.

3.1.5.2 Confirmation of high lncRNA expression variability in other tissues

Importantly, we confirmed our finding of increased lncRNA expression variability in several human cell and tissue types other than the initially analyzed granulocytes. While we chose human primary granulocytes as a model system for our study because of their clinical and diagnostic relevance, not confirming our main finding in other human cell types would leave some space for speculations about other potential reasons causing the observed lncRNA variability, rather than the genuine inter-individual difference in lncRNA expression between the analyzed donors. This could arise because granulocytes are rather a class of cells than a particular cell type, consisting of three cell types: basophil, eosinophil and neutrophil granulocytes. These cell types account for ~0.5-1%, 1-4% and 40-60% of all white blood cells in normal adult male and female blood (U.S. National Library of Medicine, https://www.nlm.nih.gov/medlineplus/ency/article/003657.htm). It is known that lncRNA landscapes, in contrast to mRNA landscapes, notably differ between even closely related cell types (Ranzani et al, 2015). Thus, differences in granulocyte cell type composition could cause dramatic up- or down-regulation of lncRNAs expressed specifically from eosinophils or basophils, since in the normal blood these two cell types are known to vary up to 3-fold in their absolute number. Our granulocyte isolation method (see Methods, Publication 2) did not distinguish between the three granulocyte types, and neither does the common granulocyte isolation procedure in the clinics (which primarily uses the gradient density centrifugation method used in our protocol).

Thus it was necessary to confirm the finding of increased lncRNA variability in an independent cell type. For this purpose we first made use of the publicly available

Geuvadis LCL RNA-seq dataset (Lappalainen et al, 2013), which provided a valuable resource of independently collected, processed and RNA-sequenced LCL samples from a large number of donors (462). Moreover, LCL being a pure cell type provides a necessary control for the above concerns. We first created a *de novo* LCL lncRNA and mRNA annotation using the pipeline established before for granulocytes. Analysis of expression variability of lncRNAs and mRNAs between the 462 donors confirmed that lncRNAs are more variable. Interestingly, the absolute level of variability was approximately twice higher than that in the granulocyte analysis for both lncRNAs and mRNAs. This contrasts with the expected lower variability between cell line samples cultured in standardized conditions, compared to primary granulocytes freshly collected from individuals and thus potentially displaying a transcriptome landscape reflecting different environmental exposure of the donors. The increased variability in the LCL data can arise from several potential factors. First, the number of LCL donors analyzed was dramatically larger than the granulocyte donor number (462 vs. 7). Second, the LCL dataset was obtained from individuals of five distinct populations, while our granulocyte dataset was collected from Caucasians. Third, the LCL dataset did not include any replicates, while we replicated each donor three times and averaged the expression value, thus lowering the impact of intra-individual variability on the output inter-individual variability value. However, regardless of the higher absolute expression variability values in LCL over the granulocyte analyses, the relative increase of lncRNA over mRNA expression variability was consistent at two fold.

We then aimed to extend the confirmation of our main finding to several human tissues. For that we made use of the GTEx dataset comprising RNA extracted from postmortem tissues (Baran et al, 2015). We analyzed MiTranscriptome lncRNA and mRNA expression in 9 tissues, 20 donors each, and showed that every tissue, though displaying slightly various absolute expression variability, showed the same notable and highly significant difference between lncRNAs and mRNAs. In the light of the above discussion on data validation, it is worth noting that absolute level of LCL lncRNA and mRNA expression variability in the GTEx analysis was nearly precisely recapitulating the levels we obtained when analyzing LCL in the Geuvadis dataset. This increased the confidence in our results, since two independently created RNA-seq datasets analyzed for expression of lncRNAs and mRNAs from two independently created transcriptome annotations, showed reproducible results. Interestingly, the similar variability level obtained for

analysis of 462 Geuvadis and 20 GTEx donors, suggests that variability might not be highly dependent of donor numbers, indicating that 20 donors are sufficient to estimate expression variability. Although we expected tissues, that always contain multiple cell types, to show higher expression variability values than LCL, we found that nerve and thyroid tissues showed nearly precisely the same median level of lncRNA expression variability as LCL (median normalized standard deviation (lncRNAs)/(mRNAs): LCL – 0.55/0.27, nerve – 0.54/0.26, thyroid – 0.56/0.27), and other tissues, such as heart or cerebellum, – just a slightly higher level (median normalized standard deviation (lncRNAs)/(mRNAs): heart – 0.66/0.0.36, cerebellum – 0.60/0.33). Interestingly, skeletal muscle showed the highest lncRNA and mRNA expression variability levels among all tissues (median normalized standard deviation (lncRNAs)/(mRNAs): 0.85/0.41), the reasons for which might be a topic of further investigations.

Overall, the analysis of granulocytes and 9 additional tissue/cell types provided the confidence of the general nature of the novel phenomenon discovered in this thesis.

3.1.5.3 High lncRNA expression variability confounds their identification

We found increased expression variability to be a new non-mRNA-like feature of lncRNAs, in addition to the previously described high tissue specificity, low expression level, low splicing and polyadenylation efficiency. We then showed, similarly to these four previously known features, that high expression variability not only distinguishes lncRNAs from mRNAs, but also confounds lncRNA identification. First, we showed that novel lncRNAs from our granulocyte and LCL lncRNA annotations, as well as from the MiTranscriptome lncRNA annotation, that all used RNA-seq from numerous donors for lncRNA identification, show increased expression variability when compared to lncRNAs already annotated by reference lncRNA annotations based on low donor numbers. Second, we performed a massive lncRNA identification bioinformatic experiment, to test if including more donors into the identification pipeline can identify more lncRNA genes. We discovered that the more donors we included, the more lncRNAs expressed in LCL could be annotated, and the number of genes identified increased dramatically and steadily. Thus, using four donors allowed identification of approximately 1,400 lncRNA loci, while using 120 donors identified approximately 4,200 lncRNA loci, i.e. 3-fold more lncRNA genes expressed. This increase was due to

identification of more known lncRNA loci being expressed in LCL, and also due to identification of new lncRNA loci in the human genome.

3.1.6 The potential causes of increased lncRNA expression variability

The reason for the increased inter-individual variability of lncRNAs over mRNAs is to date unclear. It has been shown that expression variability of both lncRNAs and mRNAs is notably affected by SNPs in either promoters, enhancers or other unknown *cis*-acting genomic regulatory elements (Lappalainen et al, 2013). LncRNA and mRNA expression variability was also shown to be strongly genetically regulated by analyzing allelic expression variability performed in the same study (Lappalainen et al, 2013). However, these studies did not provide any differential analyses between lncRNAs and mRNAs and thus did not investigate why lncRNA vary more between individuals.

Further lncRNA and mRNA expression studies in twins would be of high interest in order to shed light on the contribution of environmental and genetic factors to lncRNA and mRNA expression variability. Although it seems reasonable that lncRNAs, which are less conserved and evolve faster than mRNAs (Johnsson et al, 2014), and were also reported to harbor more SNPs in their promoter regions (Necsulea et al, 2014), are more variable on genetic basis, it is also possible that lncRNA expression is more sensitive to life-style and environmental differences between the individuals.

Additionally, it is worth noting that our experiments only analyzed steady state levels of mRNA and lncRNA expression. It is possible, however, that the variation of lncRNA expression might be to some extent contributed to the inter-individual differences in lncRNA processing and degradation. Therefore, in order to fully explain our finding of increased lncRNA expression variability, it would be important to analyze the natural variation of lncRNA vs. mRNA turnover (Clark et al, 2012), as well as to assess the variability of lncRNA and mRNA transcription rate, using nascent transcription sequencing techniques, such as NET-seq (Churchman & Weissman, 2011; Nojima et al, 2016).

3.1.7 Implications of increased variation of lncRNAs

3.1.7.1 New insight into lncRNA biology

High inter-individual variation of lncRNA expression is a finding of a substantial future significance. First, this new non-mRNA-like feature provides a deeper insight in lncRNA biology and allows more versatile comparison with well-understood mRNA biology. We found the median level of lncRNA expression variability to be twice higher than that of mRNAs in all analyzed human tissues. Apart from that, the distribution of expression variability value among the lncRNA population is notably broad (Fig. 4A, 5D and 6 in Publication 2) and is clearly broader than that of mRNAs, which highlights the heterogeneity of lncRNAs as a class of transcripts with varying features, structure and function, in contrast to mRNAs representing a consistent class of transcripts with a unified function. Interestingly, we observed that splicing efficiency also shows a very broad distribution in the population of lncRNAs, whereas all mRNAs show highly efficient splicing. Importantly, broad distribution of lncRNA expression variability indicates that high variability is inherent to only a part of lncRNA population, whereas another part shows consistent expression between individuals. Based on our granulocyte data analysis we created a list of lncRNAs robustly expressed in granulocytes, which comprised 2,490 lncRNA transcripts (40% of all our *de novo* annotated granulocyte lncRNA transcripts) expressed from 393 lncRNA loci (25% of all loci). This shows that, in spite of the overall increased variability of lncRNA population, a substantial part of lncRNAs is robustly expressed. Significantly variable lncRNAs, 1,069 *de novo* granulocyte lncRNA transcripts from 214 loci, defined by ANOVA test, constituted 17% of all transcripts (13% of loci). Thus, while some lncRNAs are notably variable between healthy humans, the majority are not dramatically variable and a misleading conclusion of lncRNAs being sporadically expressed among individuals must be avoided.

The biological significance of high inter-individual expression variability of some lncRNAs, as well as the significant non-variable expression of some lncRNAs, is yet to be defined. Although intuitively likely, it is currently unknown if highly variable lncRNAs are less functional than robust lncRNAs. Complete absence of expression of a particular lncRNA in some, but not all, healthy individuals (such as seen in Figure 7A, Publication 2) clearly indicates that this lncRNA is not crucial for life. However, it cannot be excluded that such a lncRNA showing “black and white” expression pattern, as well

as very highly variable lncRNAs, might play significant roles in phenotypic variability between healthy humans, which should be a topic of further investigations.

Several GWAS studies reported that human phenotypic traits, such as hair, skin and eye color, are associated with SNPs in either protein-coding genes or intergenic regions that might represent promoters or enhancers of lncRNA or functional regions inside lncRNA genes (Sturm, 2009 1004). At the same time the most comprehensive lncRNA annotation provided by MiTranscriptome is shown to overlap several thousands of GWAS-hit SNPs (2,586 SNPs overlapped by exons and 9,770 - by transcript gene bodies) (Iyer et al, 2015), many of which, apart from disease association, are associated with personal traits, such as, for example, height, hair, eye and skin color. Thus, just as some protein-coding genes, such as *TYR*, *OCA2* and *MC1R* (Sturm, 2009), are functionally responsible for human pigmentation, lncRNAs might also participate in the molecular pathways involved in melanin production and pigmentation. Indeed, it has been reported that some lncRNAs play roles in melanin synthesis (Zeng et al, 2016). Human pigmentation is just one example of phenotypic trait variation that is majorly caused by natural genetic, thus gene expression, variation, which is inherent to healthy humans. It is possible to expect that inter-individually variably expressed lncRNAs might contribute significantly to the formation of non-essential features, such as pigmentation, height and predisposition to obesity, while robustly expressed lncRNAs might contribute to essential molecular cell pathways.

Another interesting topic of further investigation is to obtain a lncRNA-population-wide estimate on how the level of lncRNA expression / abundance influences their ability to perform their function. It is possible that some types of functions require a certain level of lncRNA abundance, while others can be performed at a wide range of abundances and only require a handful of lncRNA transcripts.

An outstandingly interesting future direction of potential follow-up studies on high lncRNA variability between healthy individuals is investigation of its relation to brain function and personality traits. It is known that approximately 40% of human lncRNAs are expressed in brain (Derrien et al, 2012) and their expression is developmentally regulated (Aprea et al, 2013; Mercer et al, 2010) and is changed upon neuronal activity (Barry, 2014; Barry et al, 2013; Lipovich et al, 2012). Apart from the descriptive reports,

many brain lncRNAs have been studied in detail and their functions were revealed in developing and adult brain, as well as in several neurological disorders, including schizophrenia and Alzheimer's disease (reviewed in (Briggs et al, 2015)). This evidence makes the importance of lncRNAs in brain function clear and suggests high expression variability of some lncRNAs could potentially contribute to several important brain-related traits, such as, for example, intelligence, aggression, predisposition to Alzheimer's disease or anxiety disorders, and even temperament.

Although high expression variability could be attributed to an absence of regulation of a functionless lncRNAs, it is worth noting that high variability might simply indicate that, while being functional, they do not require strict abundance regulation to perform their function.

3.1.7.2 lncRNA identification and annotation

The discovery in this thesis of high lncRNA expression variability and its confounding effect on lncRNA identification brings important implications to lncRNA annotation strategies. The fact that public annotations rarely analyze large cohorts of donors might be a significant cause for under-identification of lncRNA genes in the human and in other genomes.

The MiTranscriptome annotation was based on thousands of donors and discovered approximately 60,000 lncRNA genes in human identifying over 40,000 novel genes (Iyer et al, 2015). We showed that number of identified lncRNA genes grows with the number of healthy unrelated individuals analyzed. Importantly, this holds true even when analyzing just one cell type, as we performed our analysis in LCL (Figure 7, Publication 2). The number of lncRNA loci increased as much as 3-fold: from 1,382 to 4,166, when raising the number of donors from 4 to 120 (30-fold). We showed that while marginally expressed lncRNAs and stochastic identification might contribute to this increase, the majority of loci identified when adding more donors are due to genuine variation in lncRNA expression between the donors. Thus, we conclude that when aiming to create a comprehensive annotation of lncRNAs in human it is necessary to not only achieve high RNA-seq coverage and include all the human body cell types to the analysis, as described

in the INTRODUCTION, but also to include as many individuals as possible per cell type, preferably also individuals from diverse populations.

It is worth noting that to date genomes of all the non-human organisms were reported to contain dramatically less lncRNA genes than that of human (Ulitsky & Bartel, 2013). For many organisms it might be caused by an anthropocentric bias, *i.e.,* that human is a subject of a more thorough research than other organisms. However, even mouse, the most studied model organism, is now reported to only contain 8,793 lncRNA genes by Mouse GENCODE version M8 (http://www.gencodegenes.org/mouse_stats/current.html), while the same project lists 15,941 human lncRNA genes (GENCODE v24, http://www.gencodegenes.org/stats/current.html), i.e. nearly 2-fold more genes. If we take into account the results obtained by MiTranscriptome (Iyer et al, 2015), this difference raises to 7-fold. Several factors can contribute to this dramatic difference. First, a genuinely higher abundance of lncRNAs in the human genome, caused, for example, by higher complexity of the human brain molecular organization, with brain expressing ~40% of all human lncRNAs as described above (Briggs et al, 2015). However, it seems highly unlikely, that a difference between two mammals can be so great, especially considering that the majority of human lncRNAs are still expressed in tissues that do not differ greatly in their function between human and mouse. A second reason for the difference in the lncRNA gene number might be residing in the cell type and the number of cell types of human and mouse analyzed by the lncRNA identification projects. Human projects contain thorough analyses of multiple cancer cell lines that have deviant transcriptomes potentially providing new lncRNA genes to the overall lncRNA genome. MiTranscriptome (Iyer et al, 2015) analyzed thousands of various tumor samples, which all represent a new cell type and thus might add new lncRNAs due to their extreme tissue-specificity. And the third possible reason, in line with the findings of this Doctoral Thesis, is that human lncRNA identification efforts included numerous outbred individuals with diverse genetic backgrounds while mouse projects normally rely on few inbred mouse strains (most commonly it is C57 Bl6, also known as “black 6” mice). We managed to identify 3-fold more lncRNA loci in LCL by increasing the donor number by 30-fold. It is possible that analysis of multiple mice strains, as well as wild mice can raise the number of mouse lncRNA genes several fold. It would also hold true for all the other model organisms that are usually highly inbred.

3.1.7.3 lncRNAs in medicine and personalized health

LncRNAs are increasingly implicated in disease (Batista & Chang, 2013) and proposed as biomarkers (Prensner et al, 2011) and therapeutic targets (Meng et al, 2015). The implications of lncRNAs in personalized medicine have been discussed (Vitiello et al, 2015) (Figure 15). The discovery that lncRNAs vary notably even between healthy humans most likely has significant implications on further investigations of lncRNA roles in disease or on designing the strategies of using lncRNA in clinics in the near future. Investigation of natural variation of any disease relevant lncRNAs has to precede a medical application and would give an additional insight into the function and features of this particular lncRNA.

Figure 15

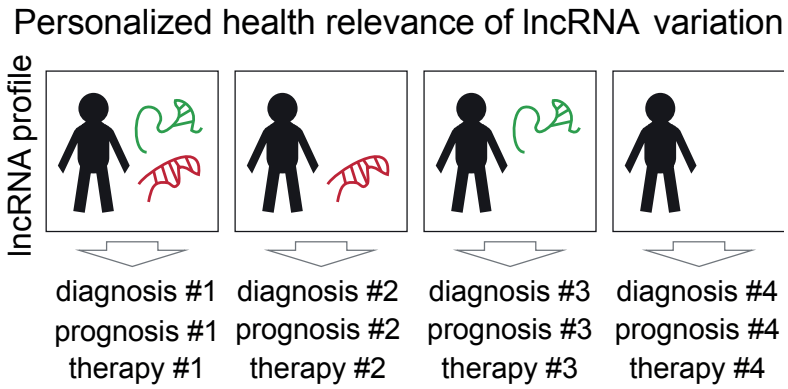


Figure 15. Personalized health relevance of lncRNA variation. Schematic representation of personalized health strategy that takes into account variability of disease related lncRNAs, whose expression might affect diagnosis (biomarker lncRNAs), prognosis (prognostic factor lncRNAs) and therapeutic strategy (therapeutic target lncRNA). In some cases just some of these medical stages are affected, whereas in others – all of them can be influenced by inter-individual lncRNA variability.

Importantly, disease-related functional studies are usually performed on inbred mice, which artificially reduces expression variation of the lncRNA of interest. This may cause clear correlation between disease state and lncRNA expression level and make this RNA appear as a plausible biomarker, while, when translated into the actual clinical use, high inter-individual variability of this lncRNA may mask this correlation even if it was meaningful and the lncRNA indeed participated in the disease molecular pathways.

Genetic and expression variability of lncRNAs may significantly affect individual susceptibility (Pan et al, 2015) (Zhang et al, 2014b) (Pasmant et al, 2011) to certain

diseases where a lncRNA performs a significant role; and it is known that many GWAS studies point out intergenic SNPs associated with disease predispositions (Hindorff et al, 2009). Therapies targeting lncRNAs are increasingly proposed and developed (Wahlestedt, 2013). The effect of high lncRNA expression variability may appear at the therapy application stage if a certain therapy targeting a lncRNA has been developed, but fails to be successful on a patient expressing this lncRNA to a very high level. Recently, personalized health has become a focus and a major goal of the medicine of the future. Inter-individual differences in genetic sequence, in methylation and histone modifications have been shown to significantly affect disease progression, treatment outcomes and, generally, treatment strategies (Rasool et al, 2015). With the discovery of unexpectedly high lncRNA expression variability, personalized health approaches, both pre-disease in individual healthy life style guidance and predisposition to diseases, and post-disease in individual medical treatment strategies, will have to account for this, potentially, major player in inter-individual variation.

3.1.8 Functional vs. non-functional lncRNAs – the meaningful transcription debate

As discussed in the INTRODUCTION, the massive identification of numerous lncRNA genes in the genome of human and other animals bewilders the scientific community. Development, optimization and price reduction of sequencing technologies, *de novo* transcriptome assembly and gene annotation pipelines make the gap between identification and functional characterization of lncRNAs immense and growing steadily. Thus, in spite of the few hundred lncRNAs shown to play important functions in various processes (as discussed in chapter *1.3 Functions and mechanisms of lncRNAs*), a part of the community is skeptical about the majority of the lncRNA population discovered (Palazzo & Lee, 2015; Raabe & Brosius, 2015). It is worth noting, that before the development of the genome-wide lncRNA identification methods, a few known lncRNAs, such as *Airn, XIST* and *KCNQ1OT1*, were thoroughly investigated for years before a substantial insight into their mechanism and function had been obtained (reviewed in (Barlow & Bartolomei, 2014) and (Lee & Bartolomei, 2013)), while the tens of thousands of lncRNAs identified over the last decade did not have a chance yet to undergo a detailed examination.

The existence of large-scale human resource projects, such as ENCODE, GTEx and Geuvadis, providing reliable high quality and high coverage RNA-seq data from more and more tissue and cell types from multiple donors to the scientific community worldwide, allows collaborative and more efficient effort on mapping lncRNAs. Moreover, RNA-seq widely entering medical practice and RNA-seq datasets becoming available when published together provide massive data on lncRNA expression in various diseases, tissues and, importantly, in the light of the discussion above, thousands of individuals to date and potentially millions in the coming decades. The study of Iyer *et al.,* (Iyer et al, 2015) that used 5,847 RNA-seq samples from The Cancer Genome Atlas and 928 from the Michigan Center for Translational Pathology along with some other public datasets and identified approximately 60,000 lncRNA human genes, underlines the power of this approach. Additionally, some countries have launched national programs of massive genome, and potentially in future transcriptome, sequencing of the population (see for Finland: http://www.sisuproject.fi/, for Netherlands: http://www.nlgenome.nl/).

Overall, it becomes clear that annotating new and refining annotations of known lncRNAs will continue and the challenge of assigning functionality to lncRNA will persist. In this light, robust technologies to efficiently assess functionality of as many lncRNA as possible are necessary. As discussed in the INTRODUCTION, CRISPR/Cas9 large-scale lncRNA knock-out approaches appear to be a promising future direction for in vitro as well as in vitro studies. At the same time, the Human Haploid Gene Trap Collection provides a ready-to-use library of lncRNA truncations described in Publication 3. As discussed above, genetic truncation of a lncRNA of interest is a complimentary and possibly a more beneficial and careful approach than promoter or whole gene body deletion. Several truncations stopping transcription at the different points inside the lncRNA gene body allow identification of regions crucial for function as well as providing important independent replicates. Moreover, gene traps introduced outside the region of interest provide an essential control for interpreting a knock-out phenotype. An important benefit of a gene trap approach is the possibility of studying a lncRNA without a prior knowledge on the mode of its action, since it removes both the transcript and transcription, without notably perturbing the genomic region, as done by promoter/gene body deletion.

However, the Human Haploid Gene Trap Collection had never been examined for its efficiency for studying lncRNAs. As discussed in detail above, lncRNAs are dramatically different from mRNAs in terms of genomic and RNA biology features. Thus, an investigation of gene trap efficiency targeting a notably non-mRNA-like lncRNA was essential. We aimed to perform such a study and have chosen a lncRNA described in the Ph.D. thesis of Irena Vlatkovic (Vlatkovic, 2010b) and named *SLC38A4down* because of its proximity to the *SLC38A4* gene. A non-coding RNA was later also mapped in this region by RefSeq (Pruitt et al, 2014) – and named *LOC100288798*. *SLC38A4down* lncRNA had preliminary been shown to be very long, unspliced and nuclearly localized, thus presenting us with an attractive non-mRNA-like lncRNA target for a functional investigation using the Human Haploid Gene Trap Collection.

3.1.9 *SLC38A4-AS* – a novel functional regulator lncRNA in human

Before analyzing the effect of *SLC38A4-AS* truncation by gene trap cassettes in KBM7 cell lines, we used public RNA-seq data from ENCODE project to gain substantial knowledge about this previously uncharacterized lncRNA. We found this lncRNA to be expressed in all the 41 human cell types analyzed and thus showed that it is, in contrast to the majority of lncRNAs, a ubiquitously expressed lncRNA. Tissue specificity is often considered a sign of functionality (Briggs et al, 2015; Kowalczyk et al, 2012), however several well-known examples of functional and even medically relevant ubiquitously expressed lncRNAs are known (Gutschner et al, 2013; Sleutels et al, 2002). We used total RNA-seq datasets to show that *SLC38A4-AS* is variably spliced among different cell types with an overall reduced splicing efficiency in comparison to control protein-coding genes. We then analyzed RNA-seq from cellular and PolyA+/- fractions to obtain a better resolution in characterizing the processing of *SLC38A4-AS*. We found that *SLC38A4-AS* shows an unusual processing pattern different from a common mRNA, but also different from *XIST*, i.e. a nuclear well-spliced lncRNA. *SLC38A4-AS* does express spliced isoforms that are exported to the cytoplasm, however, these isoforms are notably less abundant than the unspliced isoforms retained in the nucleus. These unspliced isoforms show low polyadenylation efficiency with ~3-fold higher abundance of unspliced isoforms in PolyA- fraction compared to PolyA+ fraction. Interestingly, these processing pattern resembles that of a well-studied mouse *Airn* lncRNA (Sleutels et al, 2002), that was similarly shown to be extremely long, mainly unspliced and unstable with minor

spliced isoforms being exported to the cytoplasm and stabilized (Seidl et al, 2006). *Airn* has been the main focus of our laboratory's research over the last two decades. *Airn* is an imprinted lncRNA that is known to silence three protein-coding genes *in cis*, including an antisense overlapped *Igf2r* gene that it silences by transcription interference (Latos et al, 2012).

Using the public RNA-seq data from multiple cell types expressing *SLC38A4-AS,* we assembled the transcriptome in the region around *SLC38A4-AS* and discovered that this lncRNA it nearly twice as long than had been annotated by RefSeq (as *LOC100288798*) and that it overlaps in antisense orientation the downstream *SLC38A4* protein-coding gene. This brings an additional striking similarity to *Airn* lncRNA that overlaps and silences the *Igf2*r gene. The similarities between *Airn* and *SLC38A4-AS* made us hypothesize that it was possible that *SLC38A4-AS* was involved in repressing the overlapped *SLC38A4* gene, or the nearby *SLC38A2* gene. However we did not observe any correlation between expression of *SLC38A4-AS* and these two genes, neither did our knock-out experiments reveal any potential regulation.

We used KBM7 cells whose genomes harbored a gene trap cassette truncating *SLC38A4-AS* at ~3kb and ~100kb downstream its start site, to assess if this lncRNA might be a functional regulator of any genes. We included two important types of control cell lines into the analysis: wild type cells and cells with a gene trap outside *SLC38A4-AS*. We obtained two cell line replicates per cell type (three for wild type cells). We also took special care of getting rid of batch effect and cultured all the analyzed cell lines in parallel. Additionally, all the cell culture, as well as RNA/DNA isolation and library preparation, were performed by the author of this Doctoral Thesis thus minimizing any batch effects.

We performed 50bp single-end RNA-seq of all the cell lines to estimate genome-wide gene expression changes upon *SLC38A4-AS* truncation. However, in spite of all the similarities in the RNA biology of *SLC38A4-AS* and *Airn* lncRNAs, we could not find a similarity in their function. We were surprised to find none of the genes in the 10 Mbp region around *SLC38A4-AS* consistently changed their expression upon *SLC38A4-AS* truncation. In contrast, we found 41 protein-coding gene on different chromosomes that were significantly (accessed by Cuffdiff (Trapnell et al, 2012)) and notably (>=3-fold up- or down-) deregulated upon *SLC38A4-AS* truncation. All the four control cell lines

clustered together based on the expression of these 41 genes, while the truncation cell lines clustered separately together. Moreover, 3kb replicates clustered in one branch, while 100kb replicates in another, indicating differences between these two truncation cell lines. Additional rigorous filtering (described in Results, Publication 3) shortened the gene list to just 6 genes, 3 of which were down- and 3 upregulated. The filtering steps were important to take since we wanted to remove any potential bias introduced by the gene-trap insertion procedure (for example: we filtered out all the genes that were differentially expressed between the wild type and control HOTTIP-sense and antisense gene trap insertions) and also we aimed to restrict our list to the effects of the absence of transcription 100kb downstream *SLC38A4-AS* TSS, that was shared by all 4 (2x2) truncation cell lines.

The two genes most dramatically influenced by *SLC38A4-AS* truncation, *CD9* and *RORB* (14.3-fold up- and 17.8-fold downregulation respectively), both appear to be of a particular relevance for the cell system studied. CD9 is a tetraspin, a membrane protein that plays various roles in cell function by interacting with other tetraspins, cytokine receptors and growth factors. *CD9* is a medically relevant gene since it was shown to play important roles in cancer and development (Charrin et al, 2014). *CD9* was also shown to be a marker of hematopoietic stem cells (Karlsson et al, 2013) and to be upregulated when KBM7 cells were induced to pluripotency (iPS) (Carette et al, 2010). Thus, 14.3-fold upregulation of *CD9* upon *SLC38A4-AS* truncation might indicate that *SLC38A4-AS* might be involved in sustaining differentiated state of KBM7 cells. *RORB* encodes NR1F2 protein that is capable of binding DNA and inhibiting transcription (Kennaway, 2010). Interestingly, *RORB* was also reported to be associated with induced pluripotency by showing that it is a part of a gene module discriminating human iPS cells from stem cells (Wang et al, 2011a). No studies investigated the role of *RORB* in KBM7 cells, but it is unlikely that *RORB* is an essential gene for KBM7 since the Human Haploid Gene Trap Collection contains several cell lines with *RORB* transcription eliminated, which are viable.

Thus, we show that *SLC38A4-AS* lncRNA is a functional lncRNA potentially directly, or indirectly strongly regulating expression of numerous genes in KBM7, with 6 genes being most plausible targets. Further investigation is necessary to examine if the regulation is direct or mediated by *SLC38A4-AS* interacting with a certain regulatory gene/protein,

such as transcription factors. Moreover, the mechanism of *SLC38A4-AS* regulation needs to be studied. While spliced isoforms of *SLC38A4-AS* are exported into the cytoplasm, its major unspliced isoforms are strictly nuclear and thus most likely perform their function there.

3.1.10 KBM7 gene trap collection for massive functional assessment of uncharacterized lncRNAs

We validated the use of the Human Haploid Gene Trap Collection for studying lncRNAs. The gene trap collection was initially described as a resource to study protein-coding genes (Burckstummer et al, 2013), however, a lncRNA has never been studied using this system before. In the light of the discussion above, our main goal, apart from investigating the function of *SLC38A4-AS*, was to promote this system to the lncRNA community and thus promote an acceleration of functional investigation on multiple lncRNAs. Such massive functional assessment studies on hundreds of lncRNAs might provide an important leap towards understanding of the lncRNA world and a more statistically valid estimation of the number of functional lncRNAs. We show that KBM7 gene trap collection contains clones with truncations of hundreds of GENCODE annotated lncRNAs, including several well-studied ones such as *MALAT1*. Interestingly, while *MALAT1* truncation analysis has been performed before, it was a tedious and time-consuming experiment (Gutschner et al, 2011). In contrast, the KBM7 gene trap collection contains 5 ready-to-use clones with *MALAT1* truncated. Thus many researchers might be interested in the collection if their particular lncRNA of interest is expressed in KBM7 and truncated in one of the clones. Our study delivers a valuable resource for such cases since we provide an open-access browser displaying our RNA-seq data of 8 cell lines and gene trap insertion positions, where one can search a genomic region of interest for lncRNA expression in the wild type KBM7 as well as for the presence of gene trap cassette insertion sites (https://opendata.cemm.at/barlowlab/).

Since lncRNAs are a greatly diverse class of transcripts (Quinn & Chang, 2015), it was important to prove the efficiency of gene traps acting on a notably non-mRNA-like lncRNA, such as *SLC38A4-AS*. The biggest concern might be the inefficient splicing inherent to this lncRNA, since the gene trap technology in the KBM7 collection is based on “hijacking” the RNAPII by incorporation of a strong splice acceptor site followed by a transcription stop signal. However, we showed that *SLC38A4-AS* transcription was successfully stopped using this system. While one of the 3kb truncation cell lines showed

some “leakage” transcription through the gene trap cassette (for which we investigated, but could not find the reason), the other three truncation cell lines showed excellent truncation efficiency with nearly undetectable transcription (RPKM<0.05) downstream the truncation cassette.

In summary, the Human Haploid Gene Trap Collection in KBM7 cells is an invaluable resource for studying lncRNAs despite them being a diverse class of transcripts/genes. This study performed during the completion of this Doctoral Thesis and published in the Journal RNA Biology (Publication 3) should encourage lncRNA researchers to use the Human Haploid Gene Trap Collection for more lncRNA functional studies, and provide them with the guidelines to interpret the results.

3.2 Conclusions and future prospects

LncRNAs have been gaining an exponentially growing attention in the last three decades. While they are already proposed for use in the clinics, lncRNAs as a class of genes/transcripts is only vaguely understood, as well as incompletely annotated in all organisms. This Doctoral Thesis contributed to the identification and characterization of human lncRNAs. I annotated and characterized lncRNAs of human primary granulocytes – intensively investigated and diagnostically relevant tissue, which will help future medical studies to include lncRNAs in their analyses. Importantly, I discovered a significant new feature of lncRNAs – high natural expression variation. This not only further distinguishes lncRNAs from mRNAs, but also indicates a new lncRNA classification strategy based on expression variation. We found that high natural expression variation of lncRNAs confounds their identification and that a vast number of healthy individuals must be analyzed for comprehensive lncRNA identification, which provides an important new guideline to the worldwide efforts of lncRNA identification in various species. These findings are just the groundwork for the necessary follow-up studies that need to investigate the cause of increased lncRNA expression variation and its potential significance in lncRNA evolution and function. Further investigation is also needed to assess if highly variable lncRNAs are more likely to have no function, or, rather, to participate in phenotypic variation and, relevant to medicine, disease predisposition and progression.

In the light of the urgent need for the functional characterization of numerous annotated lncRNAs, the study of *SLC38A4-AS* lncRNA performed in this Doctoral Thesis provides guidelines for efficient functional assessment of uncharacterized lncRNAs. I have found that *SLC38A4-AS* lncRNA is twice as long as was annotated in reference annotations and it is a previously unknown functional lncRNAs, regulating the *CD9* and *RORB* genes. However, the mechanism of *SLC38A4-AS* lncRNA action as well as assessment if the genes deregulated upon its truncation, are directly regulated by this lncRNA or change their expression through an action of series of factors, is unknown.

Overall, this Doctoral Thesis contributes new knowledge to the fast evolving lncRNA field and provides several important guidelines, as well as highlighting new directions for further lncRNA studies.

4 MATERIALS AND METHODS

4.1 Blood collection and granulocyte isolation from healthy donors

The study was performed under approval of the Ethics committee of the Medical University of Vienna ('Ethik Kommission der Medizinischen Universität Wien'). Blood was collected under standardized conditions in the morning before breakfast from ten healthy volunteers (five men and five women of ages ranging from 27 to 62 years) after they had signed a written informed consent (APPENDIX, Additional File 2A Table). 45 ml of blood was taken into VACUETTE® Sodium Citrate Coagulation Tubes and primary granulocytes were isolated immediately after blood collection. Granulocyte isolation was performed by means of gradient density centrifugation using Ficoll-Plaque PREMIUM (1.078 g/ml, GE Healthcare Life Sciences) with an optimized protocol (see Supplemental Methods for Publication 2 in the APPENDIX, Additional File 1, page 50: *1. Granulocyte isolation*). Granulocytes were depleted from erythrocytes in the bottom layer obtained by Ficoll centrifugation using Cell Lysis Solution (Promega) via two 5-minute incubation steps followed by centrifugation.

4.2 Granulocyte RNA-seq library preparation

After the granulocyte isolation, granulocytes were immediately lysed using 1 ml of TRI reagent (Sigma-Aldrich T9424) per 10 million cells. RNA was isolated following the optimized manufacturer's protocol (see Supplemental Methods for Publication 2 in the APPENDIX, Additional File 1, page 50: *2. RNA isolation using TRI reagent*). After the isolation, RNA was DNase I treated using DNA-free kit (Ambion) following manufacturer's protocol to remove any potential DNA contamination. DNase I treated RNA was converted into cDNA by reverse transcription (RT) using RevertAid First Strand cDNA Kit (Fermentas) following manufacturer's protocol. Each 20 µl RT reaction was performed on 0.6 µl DNA-seq treated RNA and –RT (no Reverse Transcriptase) control was made for every RT reaction set. DNAse I treated RNA was then subjected to either ribosomal depletion using RiboZero rRNA removal kit Human/Mouse/Rat (Epicentre) (see Supplemental Methods for Publication 2 in the APPENDIX, Additional File 1, page 51: *4. Ribosomal RNA depletion*) or polyA+ enriched using TruSeq RNA Sample Prep Kit v2 (Illumina) following manufacturer's protocol (see Supplemental Methods for Publication 2 in the APPENDIX, Additional File 1, page 51: *5. Polyadenylated RNA enrichment*). Strand-specific libraries were prepared following the

protocol described in Sultan et al (Sultan et al, 2012) (see Supplemental Methods for Publication 2 in the APPENDIX, Additional File 1, page 51: *6. Preparation of strand-specific RNA-seq libraries*) and non-strand-specific libraries were prepared using TruSeq RNA Sample Prep Kit v2 (Illumina) following manufacturer's protocol.

4.3 RNA-sequencing

RNA-seq libraries were pooled in equal 2nM concentrations and 50bp and 100bp PE RNA-seq was performed at the Biomedical Sequencing Facility (http://biomedical-sequencing.at/) on Illumina HiSeq 2000. The Biomedical Sequencing Facility pre-analyzed the sequencing results (base calling and demultiplexing) providing us with archived .fastq/unmapped .bam files. We obtained 22 to 79 million 100bp PE reads per from the ribosomal depleted RNA-seq samples and 24 to 38 million 100bp PE and 64 to 91 million 50bp PE reads from PolyA+ RNA-seq samples (APPENDIX, Additional File 2B Table). RNA-seq data was aligned with STAR (Dobin et al, 2013) using optimized parameters (see details and commands in Supplemental Methods for Publication 2 in the APPENDIX, Additional File 1, page 53: *7. RNA-seq read alignment*).

4.4 *De novo* granulocyte lncRNA and mRNA annotation

PolyA+ RNA-seq data obtained from primary granulocytes from ten donors was used to annotate granulocyte lncRNA and mRNA transcriptome (see detailed description and commands in Supplemental Methods for Publication 2 in the APPENDIX, Additional File 1, page 53: *11. Annotating mRNAs and lncRNAs in primary granulocytes*). We annotated granulocyte mRNAs *de novo* in preference to using publicly available protein-coding genes annotations in order to prevent our lncRNA/mRNA comparison being influenced by biases introduced by the annotation pipeline. Additionally, since the granulocyte specific gene annotation was not available at the time of the study, the mRNA annotation reveals non-annotated isoforms as well as extensions. Briefly, 784 million reads from PolyA+ RNA-seq of granulocytes from 10 donors were used for transcriptome assembly. The samples were pooled into 6 pools (APPENDIX, Additional File 2C Table) and then transcriptome was assembled *de novo* from each pool using Cufflinks (Trapnell et al, 2012). The resulting assemblies were merged using Cuffmerge (Trapnell et al, 2012) and a series of rigorous filtering steps, including protein-coding potential calculation, were performed to filter for lncRNAs (see Figure S1 in Additional File 1 (Supplemental

Figures for Publication 2) in the APPENDIX). The resulting granulocyte *de novo* lncRNA annotation consisted of 6,249 lncRNA transcripts that formed 1,591 *de novo* lncRNA loci with a mean of 3.9 transcripts per locus. *De novo* mRNAs were filtered for based on their overlap with reference protein-coding gene annotations (RefSeq (Pruitt et al, 2014) and GENCODE v19 (Harrow et al, 2012)). Granulocyte *de novo* mRNA annotation consisted of 132,864 *de novo* mRNA transcripts that formed 10,092 *de novo* mRNA loci with an average of 13.2 transcripts per locus. We assessed quality of assembly by examining several *de novo* annotation of well-known lncRNAs, such as *XIST*, and estimating how well reference mRNA annotations were covered by our granulocyte *de novo* mRNA annotation (see Figure S2 in Additional File 1 (Supplemental Figures for Publication 2) in the APPENDIX).

4.5 BLUEPRINT ChIP-seq data mining

The BLUEPRINT project is aimed at epigenetically characterizing normal and malignant human blood cell types (http://www.blueprint-epigenome.eu/), including granulocytes, namely FACS-sorted neutrophils. BLUEPRINT data is of restricted use. We applied for and were granted access to the ChIP-seq and RNA-seq data and downloaded neutrophil ChIP-seq data for 6 histone marks – H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3 and H3K9me3. Each histone mark ChIP-seq was obtained from 6 healthy donors (with the exception of H3K27ac, that was only available from 5 donors) (Table 5).

Table 5

Sample name > Modification:	C000S5H1	C0010KH2	C0011IH2	C00184H2	C001UYH1	C004GDH1
H3K27ac	V	V	V	V	X	V
H3K27me3	V	V	V	V	V	V
H3K36me3	V	V	V	V	V	V
H3K4me1	V	V	V	V	V	V
H3K4me3	V	V	V	V	V	V
H3K9me3	V	V	V	V	V	V
Input	V	V	V	V	V	V

Table 5. BLUEPRINT neutrophil samples obtained: name of the histone modification analyzed vs healthy donor code. V – sample obtained, X – sample unavailable.

4.6 ChIP-seq alignment

Downloaded raw ChIP-seq data was aligned using STAR (version 2.3.1) (Dobin et al, 2013) using the following commands:

#align#: STAR --genomeDir [hg19genome_for_STAR] --readFilesIn [fastq.gz] --readFilesCommand zcat --runThreadN 6 --genomeLoad NoSharedMemory --outStd SAM --outSAMmode Full --alignIntronMax 1 --alignEndsType EndToEnd --outFilterIntronMotifs RemoveNoncanonical --outSAMstrandField intronMotif --outFileNamePrefix [outprefix] --outFilterMultimapNmax 10 | samtools view -bS -> outprefix.bam
#sort the bam file#: samtools sort outprefix.bam outprefix.sorted
#create indexed bam file#: samtools index outprefix.sorted.bam

Bam files corresponding to the same histone mark and individuals (independently sequenced samples) were merged using *bamtools merge* (https://github.com/pezmaster31/bamtools/wiki/Tutorial_Toolkit_BamTools-1.0.pdf). This resulted in 7 aligned ChIP-seq bam files (6 marks + Input) with an average of 46.3 million mapped reads (42.1 uniquely mapped reads) ranging from 28.0 to 76.6 million reads (Table 6).

Table 6

DonorName_HistoneMark	Number of input reads	Uniquely mapped reads number	Number of reads mapped to multiple loci	Number of mapped reads
C000S5H1_H3K27ac	84,167,246	52,636,520	4,514,014	57,150,534
C000S5H1_H3K27me3	74,281,363	54,564,564	5,516,148	60,080,712
C000S5H1_H3K36me3	74,837,746	53,368,026	4,563,818	57,931,844
C000S5H1_H3K4me1	38,863,280	27,295,506	2,451,006	29,746,512
C000S5H1_H3K4me3	55,977,024	40,722,361	3,266,025	43,988,386
C000S5H1_H3K9me3	61,360,284	43,403,428	8,751,515	52,154,943
C000S5H1_Input	44,758,851	38,117,564	3,760,581	41,878,145
C0011IH2_H3K27ac	47,311,071	37,866,537	2,920,017	40,786,554
C0011IH2_H3K27me3	50,589,123	39,029,965	3,536,550	42,566,515
C0011IH2_H3K36me3	91,315,996	61,941,702	5,193,061	67,134,763
C0011IH2_H3K4me1	48,883,595	33,589,002	2,515,245	36,104,247
C0011IH2_H3K4me3	38,955,396	30,430,374	2,199,871	32,630,245
C0011IH2_H3K9me3	67,126,291	49,509,193	10,115,153	59,624,346
C0011IH2_Input	48,062,676	41,503,602	3,908,312	45,411,914
C0010KH2_H3K27ac	42,615,122	36,506,178	3,272,780	39,778,958
C0010KH2_H3K27me3	95,526,328	69,952,258	6,675,749	76,628,007
C0010KH2_H3K36me3	35,129,989	25,775,355	2,218,591	27,993,946
C0010KH2_H3K4me1	78,803,361	39,515,416	3,065,847	42,581,263
C0010KH2_H3K4me3	64,785,443	49,497,277	3,324,680	52,821,957
C0010KH2_H3K9me3	59,626,512	42,639,023	9,211,273	51,850,296
C0010KH2_Input	33,124,565	28,501,727	2,691,925	31,193,652
C001UYH1_H3K27me3	67,731,163	57,451,072	5,465,859	62,916,931
C001UYH1_H3K36me3	44,125,497	35,977,373	3,079,779	39,057,152
C001UYH1_H3K4me1	58,120,975	50,074,491	3,840,889	53,915,380
C001UYH1_H3K4me3	61,860,613	56,351,044	2,179,417	58,530,461

DonorName_HistoneMark	Number of input reads	Uniquely mapped reads number	Number of reads mapped to multiple loci	Number of mapped reads
C001UYH1_H3K9me3	60,846,221	45,723,446	8,752,510	54,475,956
C001UYH1_Input	40,215,727	34,314,753	3,369,639	37,684,392
C00184H2_H3K27ac	43,197,208	34,715,020	2,794,597	37,509,617
C00184H2_H3K27me3	45,711,549	38,356,166	3,792,770	42,148,936
C00184H2_H3K36me3	53,541,387	45,324,333	3,986,952	49,311,285
C00184H2_H3K4me1	46,493,042	39,544,241	3,312,692	42,856,933
C00184H2_H3K4me3	50,525,563	44,057,813	3,223,006	47,280,819
C00184H2_H3K9me3	57,562,378	44,132,545	7,357,344	51,489,889
C00184H2_Input	33,090,690	28,381,202	2,698,995	31,080,197
C004GDH1_H3K27ac	45,507,776	38,235,758	3,360,338	41,596,096
C004GDH1_H3K27me3	59,632,465	50,485,561	4,946,023	55,431,584
C004GDH1_H3K36me3	46,245,564	39,433,955	3,361,212	42,795,167
C004GDH1_H3K4me1	49,708,639	43,036,814	3,407,672	46,444,486
C004GDH1_H3K4me3	34,058,858	28,928,260	2,420,904	31,349,164
C004GDH1_H3K9me3	53,324,823	39,824,557	7,819,309	47,643,866
C004GDH1_Input	39,208,862	33,697,378	3,137,130	36,834,508

Table 6. BLUEPRINT neutrophil ChIP-seq alignment number of read statistics.

4.7 Histone mark coverage calculation

Histone mark coverage was calculated for *de novo* granulocyte lncRNA/mRNA 1) transcript promoters (defined as 3 kb genomic region around the annotated transcript TSS (TSS+/- 1.5kb)), 2) exons and 3) loci using coverageBed tool from bedtools package with the following command: *coverageBed -counts –abam [ChIPseq.bam] -b [bed12_annotation_file.bed] > coverage.bed*. Thus, sequencing reads mapping to the investigated were calculated. The coverage was then normalized to the total number of reads in the ChIP-seq bam file (Table 6, number of mapped reads). Afterwards, the coverage was also normalized to the length of the genomic fragment analyzed, since longer fragments would naturally contain more mapped reads. Note, that the coverage obtained in the Input samples was also analyzed and normalized by the number of reads and the genomic fragment length. Next, we accounted for the unspecific signal by subtracting the normalized Input coverage from the normalized coverage obtained for each histone mark. We averaged the coverage for each histone mark among the available donors.

4.8 Assigning significance to boxplot comparisons

Boxplots are plotted using ChIP-seq coverage values for all the *de novo* lncRNA/mRNA transcripts (numbers indicated in boxplots). The difference between lncRNA and mRNA

population sizes was taken into account in order to avoid artificial inflation of significance: the larger population was always randomly down-sampled to the size of the smaller population prior to performing statistical tests. Statistical significance of the difference between two populations was then assessed by Mann-Whitney U test. Random sampling followed by Mann-Whitney U test was performed three times for each comparison and the three p values were averaged. This average p value is indicated in the boxplot.

REFERENCES

Affymetrix_ENCODE_Transcriptome_Project (2009) Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature* **457:** 1028-1032

Amaral PP, Clark MB, Gascoigne DK, Dinger ME, Mattick JS (2011) lncRNAdb: a reference database for long noncoding RNAs. *Nucleic acids research* **39:** D146-151

Amin V, Harris RA, Onuchic V, Jackson AR, Charnecki T, Paithankar S, Lakshmi Subramanian S, Riehle K, Coarfa C, Milosavljevic A (2015) Epigenomic footprints across 111 reference epigenomes reveal tissue-specific epigenetic regulation of lincRNAs. *Nature communications* **6:** 6370

Amrein H, Axel R (1997) Genes expressed in neurons of adult male Drosophila. *Cell* **88:** 459-469

Andergassen D, Dotter CP, Kulinski TM, Guenzl PM, Bammer PC, Barlow DP, Pauler FM, Hudson QJ (2015) Allelome.PRO, a pipeline to define allele-specific genomic features from high-throughput sequencing data. *Nucleic Acids Res* **43:** e146

Anderson DM, Anderson KM, Chang CL, Makarewich CA, Nelson BR, McAnally JR, Kasaragod P, Shelton JM, Liou J, Bassel-Duby R, Olson EN (2015) A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell* **160:** 595-606

Andersson BS, Beran M, Pathak S, Goodacre A, Barlogie B, McCredie KB (1987) Ph-positive chronic myeloid leukemia with near-haploid conversion in vivo and establishment of a continuously growing cell line with similar cytogenetic pattern. *Cancer Genet Cytogenet* **24:** 335-343

Anguera MC, Ma W, Clift D, Namekawa S, Kelleher RJ, 3rd, Lee JT (2011) Tsx produces a long noncoding RNA and has general functions in the germline, stem cells, and brain. *PLoS genetics* **7:** e1002248

Aprea J, Prenninger S, Dori M, Ghosh T, Monasor LS, Wessendorf E, Zocher S, Massalini S, Alexopoulou D, Lesche M, Dahl A, Groszer M, Hiller M, Calegari F (2013) Transcriptome sequencing during mouse brain development identifies long non-coding RNAs functionally involved in neurogenic commitment. *The EMBO journal* **32:** 3145-3160

Arner E, Daub CO, Vitting-Seerup K, Andersson R, Lilje B, Drablos F, Lennartsson A, Ronnerblad M, Hrydziuszko O, Vitezic M, Freeman TC, Alhendi AM, Arner P, Axton R, Baillie JK, Beckhouse A, Bodega B, Briggs J, Brombacher F, Davis M, Detmar M, Ehrlund A, Endoh M, Eslami A, Fagiolini M, Fairbairn L, Faulkner GJ, Ferrai C, Fisher ME, Forrester L, Goldowitz D, Guler R, Ha T, Hara M, Herlyn M, Ikawa T, Kai C, Kawamoto H, Khachigian LM, Klinken SP, Kojima S, Koseki H, Klein S, Mejhert N, Miyaguchi K, Mizuno Y, Morimoto M, Morris KJ, Mummery C, Nakachi Y, Ogishima S, Okada-Hatakeyama M, Okazaki Y, Orlando V, Ovchinnikov D, Passier R, Patrikakis M, Pombo A, Qin XY, Roy S, Sato H, Savvi S, Saxena A, Schwegmann A, Sugiyama D,

Swoboda R, Tanaka H, Tomoiu A, Winteringham LN, Wolvetang E, Yanagi-Mizuochi C, Yoneda M, Zabierowski S, Zhang P, Abugessaisa I, Bertin N, Diehl AD, Fukuda S, Furuno M, Harshbarger J, Hasegawa A, Hori F, Ishikawa-Kato S, Ishizu Y, Itoh M, Kawashima T, Kojima M, Kondo N, Lizio M, Meehan TF, Mungall CJ, Murata M, Nishiyori-Sueki H, Sahin S, Nagao-Sato S, Severin J, de Hoon MJ, Kawai J, Kasukawa T, Lassmann T, Suzuki H, Kawaji H, Summers KM, Wells C, Hume DA, Forrest AR, Sandelin A, Carninci P, Hayashizaki Y (2015) Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science* **347:** 1010-1014

Atianand MK, Fitzgerald KA (2014) Long non-coding RNAs and control of gene expression in the immune system. *Trends in molecular medicine* **20:** 623-631

Avery OT, Macleod CM, McCarty M (1944) Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii. *J Exp Med* **79:** 137-158

Banfai B, Jia H, Khatun J, Wood E, Risk B, Gundling WE, Jr., Kundaje A, Gunawardena HP, Yu Y, Xie L, Krajewski K, Strahl BD, Chen X, Bickel P, Giddings MC, Brown JB, Lipovich L (2012) Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res* **22:** 1646-1657

Baran Y, Subramaniam M, Biton A, Tukiainen T, Tsang EK, Rivas MA, Pirinen M, Gutierrez-Arcelus M, Smith KS, Kukurba KR, Zhang R, Eng C, Torgerson DG, Urbanek C, Li JB, Rodriguez-Santana JR, Burchard EG, Seibold MA, MacArthur DG, Montgomery SB, Zaitlen NA, Lappalainen T (2015) The landscape of genomic imprinting across diverse adult human tissues. *Genome Res* **25:** 927-936

Barlow DP, Bartolomei MS (2014) Genomic imprinting in mammals. *Cold Spring Harbor perspectives in biology* **6**

Barry G (2014) Integrating the roles of long and small non-coding RNA in brain function and disease. *Molecular psychiatry* **19:** 410-416

Barry G, Briggs JA, Vanichkina DP, Poth EM, Beveridge NJ, Ratnu VS, Nayler SP, Nones K, Hu J, Bredy TW, Nakagawa S, Rigo F, Taft RJ, Cairns MJ, Blackshaw S, Wolvetang EJ, Mattick JS (2013) The long non-coding RNA Gomafu is acutely regulated in response to neuronal activation and involved in schizophrenia-associated alternative splicing. *Molecular psychiatry*

Bassett AR, Akhtar A, Barlow DP, Bird AP, Brockdorff N, Duboule D, Ephrussi A, Ferguson-Smith AC, Gingeras TR, Haerty W, Higgs DR, Miska EA, Ponting CP (2014) Considerations when investigating lncRNA function in vivo. *eLife* **3:** e03058

Batista PJ, Chang HY (2013) Long noncoding RNAs: cellular address codes in development and disease. *Cell* **152:** 1298-1307

Beadle GW, Tatum EL (1941) Genetic Control of Biochemical Reactions in Neurospora. *Proc Natl Acad Sci U S A* **27:** 499-506

Bernstein E, Duncan EM, Masui O, Gil J, Heard E, Allis CD (2006) Mouse polycomb proteins bind differentially to methylated histone H3 and RNA and are enriched in facultative heterochromatin. *Molecular and cellular biology* **26:** 2560-2569

Billerey C, Boussaha M, Esquerre D, Rebours E, Djari A, Meersseman C, Klopp C, Gautheret D, Rocha D (2014) Identification of large intergenic non-coding RNAs in bovine muscle using next-generation transcriptomic sequencing. *BMC Genomics* **15:** 499

Boerner S, McGinnis KM (2012) Computational identification and functional predictions of long noncoding RNA in Zea mays. *PloS one* **7:** e43047

Bond AM, Vangompel MJ, Sametsky EA, Clark MF, Savage JC, Disterhoft JF, Kohtz JD (2009) Balanced gene regulation by an embryonic brain ncRNA is critical for adult hippocampal GABA circuitry. *Nat Neurosci* **12:** 1020-1027

Brannan CI, Dees EC, Ingram RS, Tilghman SM (1990) The product of the H19 gene may function as an RNA. *Molecular and cellular biology* **10:** 28-36

Briggs JA, Wolvetang EJ, Mattick JS, Rinn JL, Barry G (2015) Mechanisms of Long Non-coding RNAs in Mammalian Nervous System Development, Plasticity, Disease, and Evolution. *Neuron* **88:** 861-877

Broadbent KM, Park D, Wolf AR, Van Tyne D, Sims JS, Ribacke U, Volkman S, Duraisingh M, Wirth D, Sabeti PC, Rinn JL (2011) A global transcriptional analysis of Plasmodium falciparum malaria reveals a novel family of telomere-associated lncRNAs. *Genome biology* **12:** R56

Brockdorff N (2013) Noncoding RNA and Polycomb recruitment. *RNA* **19:** 429-442

Brookes E, Pombo A (2012) Code breaking: the RNAPII modification code in pluripotency. *Cell Cycle* **11:** 1267-1268

Brown CJ, Ballabio A, Rupert JL, Lafreniere RG, Grompe M, Tonlorenzi R, Willard HF (1991) A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature* **349:** 38-44

Brown JB, Boley N, Eisman R, May GE, Stoiber MH, Duff MO, Booth BW, Wen J, Park S, Suzuki AM, Wan KH, Yu C, Zhang D, Carlson JW, Cherbas L, Eads BD, Miller D, Mockaitis K, Roberts J, Davis CA, Frise E, Hammonds AS, Olson S, Shenker S, Sturgill D, Samsonova AA, Weiszmann R, Robinson G, Hernandez J, Andrews J, Bickel PJ, Carninci P, Cherbas P, Gingeras TR, Hoskins RA, Kaufman TC, Lai EC, Oliver B, Perrimon N, Graveley BR, Celniker SE (2014) Diversity and dynamics of the Drosophila transcriptome. *Nature* **512:** 393-399

Burckstummer T, Banning C, Hainzl P, Schobesberger R, Kerzendorfer C, Pauler FM, Chen D, Them N, Schischlik F, Rebsamen M, Smida M, Fece de la Cruz F, Lapao A, Liszt M, Eizinger B, Guenzl PM, Blomen VA, Konopka T, Gapp B, Parapatics K, Maier B, Stockl J, Fischl W, Salic S, Taba Casari MR, Knapp S, Bennett KL, Bock C, Colinge J, Kralovics R, Ammerer G, Casari G, Brummelkamp TR, Superti-Furga G, Nijman SM

(2013) A reversible gene trap collection empowers haploid genetics in human cells. *Nature methods* **10:** 965-971

Cabianca DS, Casa V, Bodega B, Xynos A, Ginelli E, Tanaka Y, Gabellini D (2012) A long ncRNA links copy number variation to a polycomb/trithorax epigenetic switch in FSHD muscular dystrophy. *Cell* **149:** 819-831

Cabili MN, Dunagin MC, McClanahan PD, Biaesch A, Padovan-Merhar O, Regev A, Rinn JL, Raj A (2015) Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome biology* **16:** 20

Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & development* **25:** 1915-1927

Caparros ML, Alexiou M, Webster Z, Brockdorff N (2002) Functional analysis of the highly conserved exon IV of XIST RNA. *Cytogenet Genome Res* **99:** 99-105

Carette JE, Guimaraes CP, Varadarajan M, Park AS, Wuethrich I, Godarova A, Kotecki M, Cochran BH, Spooner E, Ploegh HL, Brummelkamp TR (2009) Haploid genetic screens in human cells identify host factors used by pathogens. *Science* **326:** 1231-1235

Carette JE, Pruszak J, Varadarajan M, Blomen VA, Gokhale S, Camargo FD, Wernig M, Jaenisch R, Brummelkamp TR (2010) Generation of iPSCs from cultured human malignant cells. *Blood* **115:** 4039-4042

Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K, Bajic VB, Brenner SE, Batalov S, Forrest AR, Zavolan M, Davis MJ, Wilming LG, Aidinis V, Allen JE, Ambesi-Impiombato A, Apweiler R, Aturaliya RN, Bailey TL, Bansal M, Baxter L, Beisel KW, Bersano T, Bono H, Chalk AM, Chiu KP, Choudhary V, Christoffels A, Clutterbuck DR, Crowe ML, Dalla E, Dalrymple BP, de Bono B, Della Gatta G, di Bernardo D, Down T, Engstrom P, Fagiolini M, Faulkner G, Fletcher CF, Fukushima T, Furuno M, Futaki S, Gariboldi M, Georgii-Hemming P, Gingeras TR, Gojobori T, Green RE, Gustincich S, Harbers M, Hayashi Y, Hensch TK, Hirokawa N, Hill D, Huminiecki L, Iacono M, Ikeo K, Iwama A, Ishikawa T, Jakt M, Kanapin A, Katoh M, Kawasawa Y, Kelso J, Kitamura H, Kitano H, Kollias G, Krishnan SP, Kruger A, Kummerfeld SK, Kurochkin IV, Lareau LF, Lazarevic D, Lipovich L, Liu J, Liuni S, McWilliam S, Madan Babu M, Madera M, Marchionni L, Matsuda H, Matsuzawa S, Miki H, Mignone F, Miyake S, Morris K, Mottagui-Tabar S, Mulder N, Nakano N, Nakauchi H, Ng P, Nilsson R, Nishiguchi S, Nishikawa S, Nori F, Ohara O, Okazaki Y, Orlando V, Pang KC, Pavan WJ, Pavesi G, Pesole G, Petrovsky N, Piazza S, Reed J, Reid JF, Ring BZ, Ringwald M, Rost B, Ruan Y, Salzberg SL, Sandelin A, Schneider C, Schonbach C, Sekiguchi K, Semple CA, Seno S, Sessa L, Sheng Y, Shibata Y, Shimada H, Shimada K, Silva D, Sinclair B, Sperling S, Stupka E, Sugiura K, Sultana R, Takenaka Y, Taki K, Tammoja K, Tan SL, Tang S, Taylor MS, Tegner J, Teichmann SA, Ueda HR, van Nimwegen E, Verardo R, Wei CL, Yagi K, Yamanishi H, Zabarovsky E, Zhu S, Zimmer A, Hide W, Bult C, Grimmond SM, Teasdale RD, Liu ET, Brusic V, Quackenbush J, Wahlestedt C, Mattick JS, Hume DA, Kai C, Sasaki D, Tomaru Y, Fukuda S, Kanamori-Katayama M, Suzuki M, Aoki J, Arakawa T, Iida J, Imamura K, Itoh M, Kato T, Kawaji H, Kawagashira N, Kawashima

T, Kojima M, Kondo S, Konno H, Nakano K, Ninomiya N, Nishio T, Okada M, Plessy C, Shibata K, Shiraki T, Suzuki S, Tagami M, Waki K, Watahiki A, Okamura-Oho Y, Suzuki H, Kawai J, Hayashizaki Y (2005) The transcriptional landscape of the mammalian genome. *Science* **309:** 1559-1563

Carrieri C, Cimatti L, Biagioli M, Beugnet A, Zucchelli S, Fedele S, Pesce E, Ferrer I, Collavin L, Santoro C, Forrest AR, Carninci P, Biffo S, Stupka E, Gustincich S (2012) Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature* **491:** 454-457

Cartault F, Munier P, Benko E, Desguerre I, Hanein S, Boddaert N, Bandiera S, Vellayoudom J, Krejbich-Trotot P, Bintner M, Hoarau JJ, Girard M, Genin E, de Lonlay P, Fourmaintraux A, Naville M, Rodriguez D, Feingold J, Renouil M, Munnich A, Westhof E, Fahling M, Lyonnet S, Henrion-Caude A (2012) Mutation in a primate-conserved retrotransposon reveals a noncoding RNA as a mediator of infantile encephalopathy. *Proc Natl Acad Sci U S A* **109:** 4980-4985

Charrin S, Jouannet S, Boucheix C, Rubinstein E (2014) Tetraspanins at a glance. *J Cell Sci* **127:** 3641-3648

Cheetham SW, Gruhl F, Mattick JS, Dinger ME (2013) Long noncoding RNAs and the genetics of cancer. *British journal of cancer* **108:** 2419-2425

Chen G, Wang Z, Wang D, Qiu C, Liu M, Chen X, Zhang Q, Yan G, Cui Q (2013) LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic acids research* **41:** D983-986

Cheung VG, Conlin LK, Weber TM, Arcaro M, Jen KY, Morley M, Spielman RS (2003) Natural variation in human gene expression assessed in lymphoblastoid cells. *Nature genetics* **33:** 422-425

Chodroff RA, Goodstadt L, Sirey TM, Oliver PL, Davies KE, Green ED, Molnar Z, Ponting CP (2010) Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. *Genome biology* **11:** R72

Chowers I, Liu D, Farkas RH, Gunatilaka TL, Hackam AS, Bernstein SL, Campochiaro PA, Parmigiani G, Zack DJ (2003) Gene expression variation in the adult human retina. *Human molecular genetics* **12:** 2881-2893

Chu C, Qu K, Zhong FL, Artandi SE, Chang HY (2011) Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Mol Cell* **44:** 667-678

Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S, She X, Bult CJ, Agarwala R, Cherry JL, DiCuccio M, Hlavina W, Kapustin Y, Meric P, Maglott D, Birtle Z, Marques AC, Graves T, Zhou S, Teague B, Potamousis K, Churas C, Place M, Herschleb J, Runnheim R, Forrest D, Amos-Landgraf J, Schwartz DC, Cheng Z, Lindblad-Toh K, Eichler EE, Ponting CP (2009) Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol* **7:** e1000112

Churchman LS, Weissman JS (2011) Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469:** 368-373

Clark MB, Choudhary A, Smith MA, Taft RJ, Mattick JS (2013) The dark matter rises: the expanding world of regulatory RNAs. *Essays in biochemistry* **54:** 1-16

Clark MB, Johnston RL, Inostroza-Ponta M, Fox AH, Fortini E, Moscato P, Dinger ME, Mattick JS (2012) Genome-wide analysis of long noncoding RNA stability. *Genome research* **22:** 885-898

Coccia EM, Cicala C, Charlesworth A, Ciccarelli C, Rossi GB, Philipson L, Sorrentino V (1992) Regulation and expression of a growth arrest-specific gene (gas5) during growth, differentiation, and development. *Molecular and cellular biology* **12:** 3514-3521

Congrains A, Kamide K, Ohishi M, Rakugi H (2013) ANRIL: molecular mechanisms and implications in human health. *Int J Mol Sci* **14:** 1278-1292

Crick FH (1958) On protein synthesis. *Symp Soc Exp Biol* **12:** 138-163

Darnell JE, Jr. (2013) Reflections on the history of pre-mRNA processing and highlights of current knowledge: a unified picture. *RNA* **19:** 443-460

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, Lagarde J, Veeravalli L, Ruan X, Ruan Y, Lassmann T, Carninci P, Brown JB, Lipovich L, Gonzalez JM, Thomas M, Davis CA, Shiekhattar R, Gingeras TR, Hubbard TJ, Notredame C, Harrow J, Guigo R (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome research* **22:** 1775-1789

Dinger ME, Amaral PP, Mercer TR, Mattick JS (2009) Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications. *Brief Funct Genomic Proteomic* **8:** 407-423

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Roder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Bar NS, Batut P, Bell K, Bell I, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Falconnet E, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena H, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Luo OJ, Park E, Persaud K, Preall JB, Ribeca P, Risk B, Robyr D, Sammeth M, Schaffer L, See LH, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Wrobel J, Yu Y, Ruan X, Hayashizaki Y, Harrow J, Gerstein M, Hubbard T, Reymond A, Antonarakis SE, Hannon G, Giddings MC, Ruan Y, Wold B, Carninci P, Guigo R, Gingeras TR (2012) Landscape of transcription in human cells. *Nature* **489:** 101-108

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29:** 15-21

Doolittle WF (2013) Is junk DNA bunk? A critique of ENCODE. *Proc Natl Acad Sci U S A* **110:** 5294-5300

Dumeaux V, Olsen KS, Nuel G, Paulssen RH, Borresen-Dale AL, Lund E (2010) Deciphering normal blood gene expression variation--The NOWAC postgenome study. *PLoS genetics* **6:** e1000873

Duret L, Chureau C, Samain S, Weissenbach J, Avner P (2006) The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science* **312:** 1653-1655

Duss O, Michel E, Yulikov M, Schubert M, Jeschke G, Allain FH (2014) Structural basis of the non-coding RNA RsmZ acting as a protein sponge. *Nature* **509:** 588-592

Duszczyk MM, Wutz A, Rybin V, Sattler M (2011) The Xist RNA A-repeat comprises a novel AUCG tetraloop fold and a platform for multimerization. *RNA* **17:** 1973-1982

Eissmann M, Gutschner T, Hammerle M, Gunther S, Caudron-Herger M, Gross M, Schirmacher P, Rippe K, Braun T, Zornig M, Diederichs S (2012) Loss of the abundant nuclear non-coding RNA MALAT1 is compatible with life and development. *RNA biology* **9:** 1076-1087

ENCODE_Project_Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57-74

Faghihi MA, Modarresi F, Khalil AM, Wood DE, Sahagan BG, Morgan TE, Finch CE, St Laurent G, 3rd, Kenny PJ, Wahlestedt C (2008) Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. *Nat Med* **14:** 723-730

Fatemi RP, Velmeshev D, Faghihi MA (2014) De-repressing LncRNA-Targeted Genes to Upregulate Gene Expression: Focus on Small Molecule Therapeutics. *Mol Ther Nucleic Acids* **3:** e196

Fatica A, Bozzoni I (2014) Long non-coding RNAs: new players in cell differentiation and development. *Nature reviews Genetics* **15:** 7-21

Fitzgerald KA, Caffrey DR (2014) Long noncoding RNAs in innate and adaptive immunity. *Current opinion in immunology* **26:** 140-146

Fitzpatrick GV, Soloway PD, Higgins MJ (2002) Regional loss of imprinting and growth deficiency in mice with a targeted deletion of KvDMR1. *Nature genetics* **32:** 426-431

Forrest AR, Kawaji H, Rehli M, Baillie JK, de Hoon MJ, Haberle V, Lassmann T, Kulakovskiy IV, Lizio M, Itoh M, Andersson R, Mungall CJ, Meehan TF, Schmeier S, Bertin N, Jorgensen M, Dimont E, Arner E, Schmidl C, Schaefer U, Medvedeva YA, Plessy C, Vitezic M, Severin J, Semple C, Ishizu Y, Young RS, Francescatto M, Alam I, Albanese D, Altschuler GM, Arakawa T, Archer JA, Arner P, Babina M, Rennie S, Balwierz PJ, Beckhouse AG, Pradhan-Bhatt S, Blake JA, Blumenthal A, Bodega B, Bonetti A, Briggs J, Brombacher F, Burroughs AM, Califano A, Cannistraci CV, Carbajo

D, Chen Y, Chierici M, Ciani Y, Clevers HC, Dalla E, Davis CA, Detmar M, Diehl AD, Dohi T, Drablos F, Edge AS, Edinger M, Ekwall K, Endoh M, Enomoto H, Fagiolini M, Fairbairn L, Fang H, Farach-Carson MC, Faulkner GJ, Favorov AV, Fisher ME, Frith MC, Fujita R, Fukuda S, Furlanello C, Furino M, Furusawa J, Geijtenbeek TB, Gibson AP, Gingeras T, Goldowitz D, Gough J, Guhl S, Guler R, Gustincich S, Ha TJ, Hamaguchi M, Hara M, Harbers M, Harshbarger J, Hasegawa A, Hasegawa Y, Hashimoto T, Herlyn M, Hitchens KJ, Ho Sui SJ, Hofmann OM, Hoof I, Hori F, Huminiecki L, Iida K, Ikawa T, Jankovic BR, Jia H, Joshi A, Jurman G, Kaczkowski B, Kai C, Kaida K, Kaiho A, Kajiyama K, Kanamori-Katayama M, Kasianov AS, Kasukawa T, Katayama S, Kato S, Kawaguchi S, Kawamoto H, Kawamura YI, Kawashima T, Kempfle JS, Kenna TJ, Kere J, Khachigian LM, Kitamura T, Klinken SP, Knox AJ, Kojima M, Kojima S, Kondo N, Koseki H, Koyasu S, Krampitz S, Kubosaki A, Kwon AT, Laros JF, Lee W, Lennartsson A, Li K, Lilje B, Lipovich L, Mackay-Sim A, Manabe R, Mar JC, Marchand B, Mathelier A, Mejhert N, Meynert A, Mizuno Y, de Lima Morais DA, Morikawa H, Morimoto M, Moro K, Motakis E, Motohashi H, Mummery CL, Murata M, Nagao-Sato S, Nakachi Y, Nakahara F, Nakamura T, Nakamura Y, Nakazato K, van Nimwegen E, Ninomiya N, Nishiyori H, Noma S, Noazaki T, Ogishima S, Ohkura N, Ohimiya H, Ohno H, Ohshima M, Okada-Hatakeyama M, Okazaki Y, Orlando V, Ovchinnikov DA, Pain A, Passier R, Patrikakis M, Persson H, Piazza S, Prendergast JG, Rackham OJ, Ramilowski JA, Rashid M, Ravasi T, Rizzu P, Roncador M, Roy S, Rye MB, Saijyo E, Sajantila A, Saka A, Sakaguchi S, Sakai M, Sato H, Savvi S, Saxena A, Schneider C, Schultes EA, Schulze-Tanzil GG, Schwegmann A, Sengstag T, Sheng G, Shimoji H, Shimoni Y, Shin JW, Simon C, Sugiyama D, Sugiyama T, Suzuki M, Suzuki N, Swoboda RK, t Hoen PA, Tagami M, Takahashi N, Takai J, Tanaka H, Tatsukawa H, Tatum Z, Thompson M, Toyodo H, Toyoda T, Valen E, van de Wetering M, van den Berg LM, Verado R, Vijayan D, Vorontsov IE, Wasserman WW, Watanabe S, Wells CA, Winteringham LN, Wolvetang E, Wood EJ, Yamaguchi Y, Yamamoto M, Yoneda M, Yonekura Y, Yoshida S, Zabierowski SE, Zhang PG, Zhao X, Zucchelli S, Summers KM, Suzuki H, Daub CO, Kawai J, Heutink P, Hide W, Freeman TC, Lenhard B, Bajic VB, Taylor MS, Makeev VJ, Sandelin A, Hume DA, Carninci P, Hayashizaki Y (2014) A promoter-level mammalian expression atlas. *Nature* **507:** 462-470

Furuno M, Pang KC, Ninomiya N, Fukuda S, Frith MC, Bult C, Kai C, Kawai J, Carninci P, Hayashizaki Y, Mattick JS, Suzuki H (2006) Clusters of internally primed transcripts reveal novel long noncoding RNAs. *PLoS Genet* **2:** e37

Gabory A, Jammes H, Dandolo L (2010) The H19 locus: role of an imprinted non-coding RNA in growth and development. *Bioessays* **32:** 473-480

Gascoigne DK, Cheetham SW, Cattenoz PB, Clark MB, Amaral PP, Taft RJ, Wilhelm D, Dinger ME, Mattick JS (2012) Pinstripe: a suite of programs for integrating transcriptomic and proteomic datasets identifies novel proteins and improves differentiation of protein-coding and non-coding genes. *Bioinformatics* **28:** 3042-3050

Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M (2007) What is a gene, post-ENCODE? History and updated definition. *Genome Res* **17:** 669-681

Gonzalez-Porta M, Calvo M, Sammeth M, Guigo R (2012) Estimation of alternative splicing variability in human populations. *Genome research* **22:** 528-538

Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, Elhaik E (2013) On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol Evol* **5:** 578-590

Grote P, Wittler L, Hendrix D, Koch F, Wahrisch S, Beisaw A, Macura K, Blass G, Kellis M, Werber M, Herrmann BG (2013) The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. *Developmental cell* **24:** 206-214

Guenzl PM, Barlow DP (2012) Macro lncRNAs: a new layer of cis-regulatory information in the mammalian genome. *RNA biology* **9:** 731-741

Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai MC, Hung T, Argani P, Rinn JL, Wang Y, Brzoska P, Kong B, Li R, West RB, van de Vijver MJ, Sukumar S, Chang HY (2010) Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464:** 1071-1076

Gutschner T, Baas M, Diederichs S (2011) Noncoding RNA gene silencing through genomic integration of RNA destabilizing elements using zinc finger nucleases. *Genome Res* **21:** 1944-1954

Gutschner T, Hammerle M, Diederichs S (2013) MALAT1 -- a paradigm for long noncoding RNA function in cancer. *J Mol Med (Berl)* **91:** 791-801

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, Cabili MN, Jaenisch R, Mikkelsen TS, Jacks T, Hacohen N, Bernstein BE, Kellis M, Regev A, Rinn JL, Lander ES (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458:** 223-227

Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, Munson G, Young G, Lucas AB, Ach R, Bruhn L, Yang X, Amit I, Meissner A, Regev A, Rinn JL, Root DE, Lander ES (2011) lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477:** 295-300

Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES (2013) Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* **154:** 240-251

Hacisuleyman E, Goff LA, Trapnell C, Williams A, Henao-Mejia J, Sun L, McClanahan P, Hendrickson DG, Sauvageau M, Kelley DR, Morse M, Engreitz J, Lander ES, Guttman M, Lodish HF, Flavell R, Raj A, Rinn JL (2014) Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. *Nature structural & molecular biology* **21:** 198-206

Han J, Zhang J, Chen L, Shen B, Zhou J, Hu B, Du Y, Tate PH, Huang X, Zhang W (2014) Efficient in vivo deletion of a large imprinted lncRNA by CRISPR/Cas9. *RNA Biol* **11:** 829-835

Hangauer MJ, Vaughn IW, McManus MT (2013) Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS genetics* **9:** e1003569

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, Ezkurdia I, van Baren J, Brent M, Haussler D, Kellis M, Valencia A, Reymond A, Gerstein M, Guigo R, Hubbard TJ (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research* **22:** 1760-1774

Hattori M, Fujiyama A, Taylor TD, Watanabe H, Yada T, Park HS, Toyoda A, Ishii K, Totoki Y, Choi DK, Groner Y, Soeda E, Ohki M, Takagi T, Sakaki Y, Taudien S, Blechschmidt K, Polley A, Menzel U, Delabar J, Kumpf K, Lehmann R, Patterson D, Reichwald K, Rump A, Schillhabel M, Schudy A, Zimmermann W, Rosenthal A, Kudoh J, Schibuya K, Kawasaki K, Asakawa S, Shintani A, Sasaki T, Nagamine K, Mitsuyama S, Antonarakis SE, Minoshima S, Shimizu N, Nordsiek G, Hornischer K, Brant P, Scharfe M, Schon O, Desario A, Reichelt J, Kauer G, Blocker H, Ramser J, Beck A, Klages S, Hennig S, Riesselmann L, Dagand E, Haaf T, Wehrmeyer S, Borzym K, Gardiner K, Nizetic D, Francis F, Lehrach H, Reinhardt R, Yaspo ML (2000) The DNA sequence of human chromosome 21. *Nature* **405:** 311-319

Helwak A, Kudla G, Dudnakova T, Tollervey D (2013) Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* **153:** 654-665

Heward JA, Lindsay MA (2014) Long non-coding RNAs in the regulation of the immune response. *Trends Immunol* **35:** 408-419

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106:** 9362-9367

Houseley J, Rubbi L, Grunstein M, Tollervey D, Vogelauer M (2008) A ncRNA modulates histone modification and mRNA induction in the yeast GAL gene cluster. *Mol Cell* **32:** 685-695

Huang R, Jaritz M, Guenzl P, Vlatkovic I, Sommer A, Tamir IM, Marks H, Klampfl T, Kralovics R, Stunnenberg HG, Barlow DP, Pauler FM (2011) An RNA-Seq strategy to detect the complete coding and non-coding transcriptome including full-length imprinted macro ncRNAs. *PLoS One* **6:** e27288

Huarte M, Guttman M, Feldser D, Garber M, Koziol MJ, Kenzelmann-Broz D, Khalil AM, Zuk O, Amit I, Rabani M, Attardi LD, Regev A, Lander ES, Jacks T, Rinn JL (2010) A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* **142:** 409-419

Huppertz I, Attig J, D'Ambrogio A, Easton LE, Sibley CR, Sugimoto Y, Tajnik M, Konig J, Ule J (2014) iCLIP: protein-RNA interactions at nucleotide resolution. *Methods* **65:** 274-287

Inagaki S, Numata K, Kondo T, Tomita M, Yasuda K, Kanai A, Kageyama Y (2005) Identification and expression analysis of putative mRNA-like non-coding RNA in Drosophila. *Genes Cells* **10:** 1163-1173

Ingolia NT, Lareau LF, Weissman JS (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147:** 789-802

Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, Barrette TR, Prensner JR, Evans JR, Zhao S, Poliakov A, Cao X, Dhanasekaran SM, Wu YM, Robinson DR, Beer DG, Feng FY, Iyer HK, Chinnaiyan AM (2015) The landscape of long noncoding RNAs in the human transcriptome. *Nature genetics*

Jia H, Osak M, Bogu GK, Stanton LW, Johnson R, Lipovich L (2010) Genome-wide computational identification and manual annotation of human long noncoding RNA genes. *RNA* **16:** 1478-1487

Jin J, Liu J, Wang H, Wong L, Chua NH (2013) PLncDB: plant long non-coding RNA database. *Bioinformatics* **29:** 1068-1071

Johnsson P, Lipovich L, Grander D, Morris KV (2014) Evolutionary conservation of long non-coding RNAs; sequence, structure, function. *Biochimica et biophysica acta* **1840:** 1063-1071

Kanduri C (2015) Long noncoding RNAs: Lessons from genomic imprinting. *Biochim Biophys Acta*

Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, Gingeras TR (2002) Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296:** 916-919

Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermuller J, Hofacker IL, Bell I, Cheung E, Drenkow J, Dumais E, Patel S, Helt G, Ganesh M, Ghosh S, Piccolboni A, Sementchenko V, Tammana H, Gingeras TR (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316:** 1484-1488

Kapusta A, Feschotte C (2014) Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. *Trends in genetics : TIG* **30:** 439-452

Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, Yandell M, Feschotte C (2013) Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS genetics* **9:** e1003470

Karlsson G, Rorby E, Pina C, Soneji S, Reckzeh K, Miharada K, Karlsson C, Guo Y, Fugazza C, Gupta R, Martens JH, Stunnenberg HG, Karlsson S, Enver T (2013) The tetraspanin CD9 affords high-purity capture of all murine hematopoietic stem cells. *Cell Rep* **4:** 642-648

Kelley D, Rinn J (2012) Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome biology* **13:** R107

Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J, Dunham I, Elnitski LL, Farnham PJ, Feingold EA, Gerstein M, Giddings MC, Gilbert DM, Gingeras TR, Green ED, Guigo R, Hubbard T, Kent J, Lieb JD, Myers RM, Pazin MJ, Ren B, Stamatoyannopoulos JA, Weng Z, White KP, Hardison RC (2014) Defining functional DNA elements in the human genome. *Proc Natl Acad Sci U S A* **111:** 6131-6138

Keniry A, Oxley D, Monnier P, Kyba M, Dandolo L, Smits G, Reik W (2012) The H19 lincRNA is a developmental reservoir of miR-675 that suppresses growth and Igf1r. *Nat Cell Biol* **14:** 659-665

Kennaway DJ (2010) Clock genes at the heart of depression. *J Psychopharmacol* **24:** 5-14

Khaitovich P, Hellmann I, Enard W, Nowick K, Leinweber M, Franz H, Weiss G, Lachmann M, Paabo S (2005) Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* **309:** 1850-1854

Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, Thomas K, Presser A, Bernstein BE, van Oudenaarden A, Regev A, Lander ES, Rinn JL (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **106:** 11667-11672

Khan Z, Ford MJ, Cusanovich DA, Mitrano A, Pritchard JK, Gilad Y (2013) Primate transcript and protein expression levels evolve under compensatory selection pressures. *Science* **342:** 1100-1104

Kino T, Hurt DE, Ichijo T, Nader N, Chrousos GP (2010) Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. *Sci Signal* **3:** ra8

Klattenhoff CA, Scheuermann JC, Surface LE, Bradley RK, Fields PA, Steinhauser ML, Ding H, Butty VL, Torrey L, Haas S, Abo R, Tabebordbar M, Lee RT, Burge CB, Boyer LA (2013) Braveheart, a long noncoding RNA required for cardiovascular lineage commitment. *Cell* **152:** 570-583

Kogo R, Shimamura T, Mimori K, Kawahara K, Imoto S, Sudo T, Tanaka F, Shibata K, Suzuki A, Komune S, Miyano S, Mori M (2011) Long noncoding RNA HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers. *Cancer Res* **71:** 6320-6326

Kondo T, Plaza S, Zanet J, Benrabah E, Valenti P, Hashimoto Y, Kobayashi S, Payre F, Kageyama Y (2010) Small peptides switch the transcriptional activity of Shavenbaby during Drosophila embryogenesis. *Science* **329:** 336-339

Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, Gao G (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic acids research* **35:** W345-349

Kornienko AE, Dotter CP, Guenzl PM, Gisslinger H, Gisslinger B, Cleary C, Kralovics R, Pauler FM, Barlow DP (2016) Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans. *Genome Biol* **17:** 14

Kornienko AE, Guenzl PM, Barlow DP, Pauler FM (2013) Gene regulation by the act of long non-coding RNA transcription. *BMC biology* **11:** 59

Kowalczyk MS, Higgs DR, Gingeras TR (2012) Molecular biology: RNA discrimination. *Nature* **482:** 310-311

Kutter C, Watt S, Stefflova K, Wilson MD, Goncalves A, Ponting CP, Odom DT, Marques AC (2012) Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS genetics* **8:** e1002841

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann Y, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowki J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ (2001) Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921

Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA, Gonzalez-Porta M, Kurbatova N, Griebel T, Ferreira PG, Barann M, Wieland T, Greger L, van Iterson M, Almlof J, Ribeca P, Pulyakhina I, Esser D, Giger T, Tikhonov A, Sultan M, Bertier G, MacArthur DG, Lek M, Lizano E, Buermans HP, Padioleau I, Schwarzmayr T, Karlberg O, Ongen H, Kilpinen H, Beltran S, Gut M, Kahlem K, Amstislavskiy V, Stegle O, Pirinen M, Montgomery SB, Donnelly P, McCarthy MI, Flicek P, Strom TM, Lehrach H, Schreiber S, Sudbrak R, Carracedo A, Antonarakis SE, Hasler R, Syvanen AC, van Ommen GJ, Brazma A, Meitinger T, Rosenstiel P, Guigo R, Gut IG, Estivill X, Dermitzakis ET (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501:** 506-511

Latos PA, Pauler FM, Koerner MV, Senergin HB, Hudson QJ, Stocsits RR, Allhoff W, Stricker SH, Klement RM, Warczok KE, Aumayr K, Pasierbek P, Barlow DP (2012) Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. *Science* **338:** 1469-1472

Lee JT, Bartolomei MS (2013) X-inactivation, imprinting, and long noncoding RNAs in health and disease. *Cell* **152:** 1308-1323

Lee RC, Feinbaum RL, Ambros V (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75:** 843-854

Lennox KA, Behlke MA (2015) Cellular localization of long non-coding RNAs affects silencing by RNAi more than by antisense oligonucleotides. *Nucleic acids research*

Li CH, Chen Y (2013) Targeting long non-coding RNAs in cancers: progress and prospects. *Int J Biochem Cell Biol* **45:** 1895-1910

Li J, Wu B, Xu J, Liu C (2014a) Genome-wide identification and characterization of long intergenic non-coding RNAs in Ganoderma lucidum. *PloS one* **9:** e99442

Li JP, Liu LH, Li J, Chen Y, Jiang XW, Ouyang YR, Liu YQ, Zhong H, Li H, Xiao T (2013a) Microarray expression profile of long noncoding RNAs in human osteosarcoma. *Biochem Biophys Res Commun* **433:** 200-206

Li L, Eichten SR, Shimizu R, Petsch K, Yeh CT, Wu W, Chettoor AM, Givan SA, Cole RA, Fowler JE, Evans MM, Scanlon MJ, Yu J, Schnable PS, Timmermans MC, Springer NM, Muehlbauer GJ (2014b) Genome-wide discovery and characterization of maize long non-coding RNAs. *Genome biology* **15:** R40

Li L, Liu B, Wapinski OL, Tsai MC, Qu K, Zhang J, Carlson JC, Lin M, Fang F, Gupta RA, Helms JA, Chang HY (2013b) Targeted disruption of Hotair leads to homeotic transformation and gene derepression. *Cell Rep* **5:** 3-12

Ligtenberg MJ, Kuiper RP, Chan TL, Goossens M, Hebeda KM, Voorendt M, Lee TY, Bodmer D, Hoenselaar E, Hendriks-Cornelissen SJ, Tsui WY, Kong CK, Brunner HG, van Kessel AG, Yuen ST, van Krieken JH, Leung SY, Hoogerbrugge N (2009) Heritable somatic methylation and inactivation of MSH2 in families with Lynch syndrome due to deletion of the 3' exons of TACSTD1. *Nature genetics* **41:** 112-117

Lin N, Chang KY, Li Z, Gates K, Rana ZA, Dang J, Zhang D, Han T, Yang CS, Cunningham TJ, Head SR, Duester G, Dong PD, Rana TM (2014) An evolutionarily conserved long noncoding RNA TUNA controls pluripotency and neural lineage commitment. *Mol Cell* **53:** 1005-1019

Lipovich L, Dachet F, Cai J, Bagla S, Balan K, Jia H, Loeb JA (2012) Activity-dependent human brain coding/noncoding gene regulatory networks. *Genetics* **192:** 1133-1148

Liu J, Jung C, Xu J, Wang H, Deng S, Bernad L, Arenas-Huertero C, Chua NH (2012) Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis. *Plant Cell* **24:** 4333-4345

Luke B, Lingner J (2009) TERRA: telomeric repeat-containing RNA. *The EMBO journal* **28:** 2503-2510

Luo H, Zhao X, Wan X, Huang S, Wu D (2014) Gene microarray analysis of the lncRNA expression profile in human urothelial carcinoma of the bladder. *Int J Clin Exp Med* **7:** 1244-1254

Lyle R, Watanabe D, te Vruchte D, Lerchner W, Smrzka OW, Wutz A, Schageman J, Hahner L, Davies C, Barlow DP (2000) The imprinted antisense RNA at the Igf2r locus overlaps but does not imprint Mas1. *Nature genetics* **25:** 19-21

Maass PG, Rump A, Schulz H, Stricker S, Schulze L, Platzer K, Aydin A, Tinschert S, Goldring MB, Luft FC, Bahring S (2012) A misplaced lncRNA causes brachydactyly in humans. *J Clin Invest* **122:** 3990-4002

Maenner S, Blaud M, Fouillen L, Savoye A, Marchand V, Dubois A, Sanglier-Cianferani S, Van Dorsselaer A, Clerc P, Avner P, Visvikis A, Branlant C (2010) 2-D structure of the A region of Xist RNA and its implication for PRC2 association. *PLoS Biol* **8:** e1000276

Marahrens Y, Panning B, Dausman J, Strauss W, Jaenisch R (1997) Xist-deficient mice are defective in dosage compensation but not spermatogenesis. *Genes Dev* **11:** 156-166

Marques AC, Hughes J, Graham B, Kowalczyk MS, Higgs DR, Ponting CP (2013) Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs. *Genome biology* **14:** R131

Marques AC, Ponting CP (2009) Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome biology* **10:** R124

Marques AC, Ponting CP (2014) Intergenic lncRNAs and the evolution of gene expression. *Current opinion in genetics & development* **27:** 48-53

Martens-Uzunova ES, Bottcher R, Croce CM, Jenster G, Visakorpi T, Calin GA (2013) Long Noncoding RNA in Prostate, Bladder, and Kidney Cancer. *European urology*

Mattick JS (2011) Long noncoding RNAs in cell and developmental biology. *Seminars in cell & developmental biology* **22:** 327

Mattick JS, Makunin IV (2005) Small regulatory RNAs in mammals. *Hum Mol Genet* **14 Spec No 1:** R121-132

Mattick JS, Rinn JL (2015) Discovery and annotation of long noncoding RNAs. *Nature structural & molecular biology* **22:** 5-7

Mattick JS, Taft RJ, Faulkner GJ (2010) A global view of genomic information--moving beyond the gene and the master regulator. *Trends in genetics : TIG* **26:** 21-28

Meller VH, Rattner BP (2002) The roX genes encode redundant male-specific lethal transcripts required for targeting of the MSL complex. *The EMBO journal* **21:** 1084-1091

Meller VH, Wu KH, Roman G, Kuroda MI, Davis RL (1997) roX1 RNA paints the X chromosome of male Drosophila and is regulated by the dosage compensation system. *Cell* **88:** 445-457

Mendenhall EM, Bernstein BE (2008) Chromatin state maps: new technologies, new insights. *Current opinion in genetics & development* **18:** 109-115

Meng L, Person RE, Beaudet AL (2012) Ube3a-ATS is an atypical RNA polymerase II transcript that represses the paternal expression of Ube3a. *Hum Mol Genet* **21:** 3001-3012

Meng L, Person RE, Huang W, Zhu PJ, Costa-Mattioli M, Beaudet AL (2013) Truncation of Ube3a-ATS unsilences paternal Ube3a and ameliorates behavioral defects in the Angelman syndrome mouse model. *PLoS genetics* **9:** e1004039

Meng L, Ward AJ, Chun S, Bennett CF, Beaudet AL, Rigo F (2015) Towards a therapy for Angelman syndrome by targeting a long non-coding RNA. *Nature* **518:** 409-412

Mercer TR, Gerhardt DJ, Dinger ME, Crawford J, Trapnell C, Jeddeloh JA, Mattick JS, Rinn JL (2012) Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nature biotechnology* **30:** 99-104

Mercer TR, Mattick JS (2013) Structure and function of long noncoding RNAs in epigenetic regulation. *Nature structural & molecular biology* **20:** 300-307

Mercer TR, Qureshi IA, Gokhan S, Dinger ME, Li G, Mattick JS, Mehler MF (2010) Long noncoding RNAs in neuronal-glial fate specification and oligodendrocyte lineage maturation. *BMC neuroscience* **11:** 14

Millar JK, Wilson-Annan JC, Anderson S, Christie S, Taylor MS, Semple CA, Devon RS, St Clair DM, Muir WJ, Blackwood DH, Porteous DJ (2000) Disruption of two novel genes by a translocation co-segregating with schizophrenia. *Hum Mol Genet* **9:** 1415-1423

Mirza AH, Berthelsen CH, Seemann SE, Pan X, Frederiksen KS, Vilien M, Gorodkin J, Pociot F (2015) Transcriptomic landscape of lncRNAs in inflammatory bowel disease. *Genome medicine* **7:** 39

Mohr SE, Smith JA, Shamu CE, Neumuller RA, Perrimon N (2014) RNAi screening comes of age: improved techniques and complementary approaches. *Nat Rev Mol Cell Biol* **15:** 591-600

Morgan TH, Sturtevant, A.H., Muller, H.J., Bridges, C.B. (1915) *The mechanism of Mendelian heredity*, New York: Holt Rinehart & Winston.

Morris KV, Mattick JS (2014) The rise of regulatory RNA. *Nature reviews Genetics* **15:** 423-437

Mudge JM, Frankish A, Harrow J (2013) Functional transcriptomics in the post-ENCODE era. *Genome research* **23:** 1961-1973

Nagano T, Mitchell JA, Sanz LA, Pauler FM, Ferguson-Smith AC, Feil R, Fraser P (2008) The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. *Science* **322:** 1717-1720

Nakagawa S, Ip JY, Shioi G, Tripathi V, Zong X, Hirose T, Prasanth KV (2012) Malat1 is not an essential component of nuclear speckles in mice. *RNA* **18:** 1487-1499

Nakagawa S, Kageyama Y (2014) Nuclear lncRNAs as epigenetic regulators-beyond skepticism. *Biochim Biophys Acta* **1839:** 215-222

Nakagawa S, Naganuma T, Shioi G, Hirose T (2011) Paraspeckles are subpopulation-specific nuclear bodies that are not essential in mice. *The Journal of cell biology* **193:** 31-39

Nakagawa S, Shimada M, Yanaka K, Mito M, Arai T, Takahashi E, Fujita Y, Fujimori T, Standaert L, Marine JC, Hirose T (2014) The lncRNA Neat1 is required for corpus luteum formation and the establishment of pregnancy in a subpopulation of mice. *Development* **141:** 4618-4627

Nam JW, Bartel DP (2012) Long noncoding RNAs in C. elegans. *Genome research* **22:** 2529-2540

Necsulea A, Kaessmann H (2014) Evolutionary dynamics of coding and non-coding transcriptomes. *Nature reviews Genetics* **15:** 734-748

Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grutzner F, Kaessmann H (2014) The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505:** 635-640

Ng SY, Lin L, Soh BS, Stanton LW (2013) Long noncoding RNAs in development and disease of the central nervous system. *Trends in genetics : TIG* **29:** 461-468

Nielsen MM, Tehler D, Vang S, Sudzina F, Hedegaard J, Nordentoft I, Orntoft TF, Lund AH, Pedersen JS (2014) Identification of expressed and conserved human noncoding RNAs. *RNA* **20:** 236-251

Niinuma T, Suzuki H, Nojima M, Nosho K, Yamamoto H, Takamaru H, Yamamoto E, Maruyama R, Nobuoka T, Miyazaki Y, Nishida T, Bamba T, Kanda T, Ajioka Y, Taguchi T, Okahara S, Takahashi H, Nishida Y, Hosokawa M, Hasegawa T, Tokino T, Hirata K, Imai K, Toyota M, Shinomura Y (2012) Upregulation of miR-196a and HOTAIR drive malignant character in gastrointestinal stromal tumors. *Cancer Res* **72:** 1126-1136

Nitsche A, Rose D, Fasold M, Reiche K, Stadler PF (2015) Comparison of splice sites reveals that long noncoding RNAs are evolutionarily well conserved. *RNA* **21:** 801-812

Nojima T, Gomes T, Carmo-Fonseca M, Proudfoot NJ (2016) Mammalian NET-seq analysis defines nascent RNA profiles and associated RNA processing genome-wide. *Nat Protoc* **11:** 413-428

Orom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, Lai F, Zytnicki M, Notredame C, Huang Q, Guigo R, Shiekhattar R (2010) Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143:** 46-58

Ounzain S, Micheletti R, Beckmann T, Schroen B, Alexanian M, Pezzuto I, Crippa S, Nemir M, Sarre A, Johnson R, Dauvillier J, Burdet F, Ibberson M, Guigo R, Xenarios I, Heymans S, Pedrazzini T (2015) Genome-wide profiling of the cardiac transcriptome after myocardial infarction identifies novel heart-specific long non-coding RNAs. *European heart journal* **36:** 353-368a

Ozsolak F, Kapranov P, Foissac S, Kim SW, Fishilevich E, Monaghan AP, John B, Milos PM (2010) Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell* **143:** 1018-1029

Pachnis V, Brannan CI, Tilghman SM (1988) The structure and expression of a novel gene activated in early mouse embryogenesis. *The EMBO journal* **7:** 673-681

Palazzo AF, Lee ES (2015) Non-coding RNA: what is functional and what is junk? *Front Genet* **6:** 2

Pan W, Liu L, Wei J, Ge Y, Zhang J, Chen H, Zhou L, Yuan Q, Zhou C, Yang M (2015) A functional lncRNA HOTAIR genetic variant contributes to gastric cancer susceptibility. *Molecular carcinogenesis*

Parenti R, Paratore S, Torrisi A, Cavallaro S (2007) A natural antisense transcript against Rad18, specifically expressed in neurons and upregulated during beta-amyloid-induced apoptosis. *Eur J Neurosci* **26:** 2444-2457

Pasmant E, Laurendeau I, Heron D, Vidaud M, Vidaud D, Bieche I (2007) Characterization of a germ-line deletion, including the entire INK4/ARF locus, in a melanoma-neural system tumor family: identification of ANRIL, an antisense noncoding RNA whose expression coclusters with ARF. *Cancer Res* **67:** 3963-3969

Pasmant E, Sabbagh A, Vidaud M, Bieche I (2011) ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. *FASEB J* **25:** 444-448

Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, Levin JZ, Fan L, Sandelin A, Rinn JL, Regev A, Schier AF (2012) Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome research* **22:** 577-591

Penny GD, Kay GF, Sheardown SA, Rastan S, Brockdorff N (1996) Requirement for Xist in X chromosome inactivation. *Nature* **379:** 131-137

Peterlin BM, Brogie JE, Price DH (2012) 7SK snRNA: a noncoding RNA that plays a major role in regulating eukaryotic transcription. *Wiley interdisciplinary reviews RNA* **3:** 92-103

Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP (2010) A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465:** 1033-1038

Ponjavic J, Oliver PL, Lunter G, Ponting CP (2009) Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. *PLoS genetics* **5:** e1000617

Preker P, Almvig K, Christensen MS, Valen E, Mapendano CK, Sandelin A, Jensen TH (2011) PROMoter uPstream Transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters. *Nucleic acids research* **39:** 7179-7193

Preker P, Nielsen J, Kammler S, Lykke-Andersen S, Christensen MS, Mapendano CK, Schierup MH, Jensen TH (2008) RNA exosome depletion reveals transcription upstream of active human promoters. *Science* **322:** 1851-1854

Prensner JR, Iyer MK, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC, Laxman B, Asangani IA, Grasso CS, Kominsky HD, Cao X, Jing X, Wang X, Siddiqui J, Wei JT, Robinson D, Iyer HK, Palanisamy N, Maher CA, Chinnaiyan AM (2011) Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nature biotechnology* **29:** 742-749

Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM, Murphy MR, O'Leary NA, Pujar S, Rajput B, Rangwala SH, Riddick LD, Shkeda A, Sun H, Tamez P, Tully RE, Wallin C, Webb D, Weber J, Wu W, DiCuccio M, Kitts P, Maglott DR, Murphy TD, Ostell JM (2014) RefSeq: an update on mammalian reference sequences. *Nucleic acids research* **42:** D756-763

Qu Z, Adelson DL (2012) Identification and comparative analysis of ncRNAs in human, mouse and zebrafish indicate a conserved role in regulation of genes expressed in brain. *PloS one* **7:** e52275

Quek XC, Thomson DW, Maag JL, Bartonicek N, Signal B, Clark MB, Gloss BS, Dinger ME (2015) lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic acids research* **43:** D168-173

Quinn JJ, Chang HY (2015) Unique features of long non-coding RNA biogenesis and function. *Nature reviews Genetics* **17:** 47-62

Quinodoz S, Guttman M (2014) Long noncoding RNAs: an emerging link between gene regulation and nuclear organization. *Trends in cell biology* **24:** 651-663

Qureshi IA, Mattick JS, Mehler MF (2010) Long non-coding RNAs in nervous system function and disease. *Brain research* **1338:** 20-35

Qureshi IA, Mehler MF (2013) Long non-coding RNAs: novel targets for nervous system disease diagnosis and therapy. *Neurotherapeutics* **10:** 632-646

Raabe CA, Brosius J (2015) Does every transcript originate from a gene? *Annals of the New York Academy of Sciences* **1341:** 136-148

Ranzani V, Rossetti G, Panzeri I, Arrigoni A, Bonnal RJ, Curti S, Gruarin P, Provasi E, Sugliano E, Marconi M, De Francesco R, Geginat J, Bodega B, Abrignani S, Pagani M (2015) The long intergenic noncoding RNA landscape of human lymphocytes highlights the regulation of T cell differentiation by linc-MAF-4. *Nature immunology*

Rasool M, Malik A, Naseer MI, Manan A, Ansari S, Begum I, Qazi MH, Pushparaj P, Abuzenadah AM, Al-Qahtani MH, Kamal MA, Gan S (2015) The role of epigenetics in personalized medicine: challenges and opportunities. *BMC Med Genomics* **8 Suppl 1:** S5

Redrup L, Branco MR, Perdeaux ER, Krueger C, Lewis A, Santos F, Nagano T, Cobb BS, Fraser P, Reik W (2009) The long noncoding RNA Kcnq1ot1 organises a lineage-specific nuclear domain for epigenetic gene silencing. *Development* **136:** 525-530

Rinn JL, Chang HY (2012) Genome regulation by long noncoding RNAs. *Annual review of biochemistry* **81:** 145-166

Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E, Chang HY (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129:** 1311-1323

Ripoche MA, Kress C, Poirier F, Dandolo L (1997) Deletion of the H19 transcription unit reveals the existence of a putative imprinting control element. *Genes Dev* **11:** 1596-1604

Roth A, Diederichs S (2015) Long Noncoding RNAs in Lung Cancer. *Curr Top Microbiol Immunol*

Santoro F, Mayer D, Klement RM, Warczok KE, Stukalov A, Barlow DP, Pauler FM (2013) Imprinted Igf2r silencing depends on continuous Airn lncRNA expression and is not restricted to a developmental window. *Development* **140:** 1184-1195

Sauvageau M, Goff LA, Lodato S, Bonev B, Groff AF, Gerhardinger C, Sanchez-Gomez DB, Hacisuleyman E, Li E, Spence M, Liapis SC, Mallard W, Morse M, Swerdel MR, D'Ecclessis MF, Moore JC, Lai V, Gong G, Yancopoulos GD, Frendewey D, Kellis M, Hart RP, Valenzuela DM, Arlotta P, Rinn JL (2013) Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *eLife* **2:** e01749

Scheele C, Petrovic N, Faghihi MA, Lassmann T, Fredriksson K, Rooyackers O, Wahlestedt C, Good L, Timmons JA (2007) The human PINK1 locus is regulated in vivo by a non-coding natural antisense RNA during modulation of mitochondrial function. *BMC Genomics* **8:** 74

Scheuermann JC, Boyer LA (2013) Getting to the heart of the matter: long non-coding RNAs in cardiac development and disease. *The EMBO journal* **32:** 1805-1816

Schmidt LH, Spieker T, Koschmieder S, Schaffers S, Humberg J, Jungen D, Bulk E, Hascher A, Wittmer D, Marra A, Hillejan L, Wiebe K, Berdel WE, Wiewrodt R, Muller-Tidow C (2011) The long noncoding MALAT-1 RNA indicates a poor prognosis in non-small cell lung cancer and induces migration and tumor growth. *J Thorac Oncol* **6:** 1984-1992

Schoeftner S, Blasco MA (2008) Developmentally regulated transcription of mammalian telomeres by DNA-dependent RNA polymerase II. *Nat Cell Biol* **10:** 228-236

Schorderet P, Duboule D (2011) Structural and functional differences in the long non-coding RNA hotair in mouse and human. *PLoS genetics* **7:** e1002071

Schuster-Gossler K, Simon-Chazottes D, Guenet JL, Zachgo J, Gossler A (1996) Gtl2lacZ, an insertional mutation on mouse chromosome 12 with parental origin-dependent phenotype. *Mamm Genome* **7:** 20-24

Seidl CI, Stricker SH, Barlow DP (2006) The imprinted Air ncRNA is an atypical RNAPII transcript that evades splicing and escapes nuclear export. *The EMBO journal* **25:** 3565-3575

Senner CE, Nesterova TB, Norton S, Dewchand H, Godwin J, Mak W, Brockdorff N (2011) Disruption of a conserved region of Xist exon 1 impairs Xist RNA localisation and X-linked gene silencing during random and imprinted X chromosome inactivation. *Development* **138:** 1541-1550

Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, Raychowdhury R, Schwartz S, Yosef N, Malboeuf C, Lu D, Trombetta JJ, Gennert D, Gnirke A, Goren A, Hacohen N, Levin JZ, Park H, Regev A (2013) Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498:** 236-240

Shechner DM, Hacisuleyman E, Younger ST, Rinn JL (2015) Multiplexable, locus-specific targeting of long RNAs with CRISPR-Display. *Nat Methods* **12:** 664-670

Shlyueva D, Stampfel G, Stark A (2014) Transcriptional enhancers: from properties to genome-wide predictions. *Nature reviews Genetics* **15:** 272-286

Sigova AA, Mullen AC, Molinie B, Gupta S, Orlando DA, Guenther MG, Almada AE, Lin C, Sharp PA, Giallourakis CC, Young RA (2013) Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. *Proc Natl Acad Sci U S A* **110:** 2876-2881

Skarnes WC, von Melchner H, Wurst W, Hicks G, Nord AS, Cox T, Young SG, Ruiz P, Soriano P, Tessier-Lavigne M, Conklin BR, Stanford WL, Rossant J (2004) A public gene trap resource for mouse functional genomics. *Nature genetics* **36:** 543-544

Sleutels F, Zwart R, Barlow DP (2002) The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature* **415:** 810-813

Smilinich NJ, Day CD, Fitzpatrick GV, Caldwell GM, Lossie AC, Cooper PR, Smallwood AC, Joyce JA, Schofield PN, Reik W, Nicholls RD, Weksberg R, Driscoll DJ, Maher ER, Shows TB, Higgins MJ (1999) A maternally methylated CpG island in KvLQT1 is associated with an antisense paternal transcript and loss of imprinting in Beckwith-Wiedemann syndrome. *Proc Natl Acad Sci U S A* **96:** 8064-8069

Soumillon M, Necsulea A, Weier M, Brawand D, Zhang X, Gu H, Barthes P, Kokkinaki M, Nef S, Gnirke A, Dym M, de Massy B, Mikkelsen TS, Kaessmann H (2013) Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Rep* **3:** 2179-2190

Spielman RS, Bastone LA, Burdick JT, Morley M, Ewens WJ, Cheung VG (2007) Common genetic variants account for differences in gene expression among ethnic groups. *Nature genetics* **39:** 226-231

St Laurent G, Vyatkin Y, Kapranov P (2014) Dark matter RNA illuminates the puzzle of genome-wide association studies. *BMC Med* **12:** 97

St Laurent G, Wahlestedt C, Kapranov P (2015) The Landscape of long noncoding RNA classification. *Trends in genetics : TIG* **31:** 239-251

Stanford WL, Cohn JB, Cordes SP (2001) Gene-trap mutagenesis: past, present and beyond. *Nature reviews Genetics* **2:** 756-768

Storey JD, Madeoy J, Strout JL, Wurfel M, Ronald J, Akey JM (2007) Gene-expression variation within and among human populations. *American journal of human genetics* **80:** 502-509

Sturm RA (2009) Molecular genetics of human pigmentation diversity. *Hum Mol Genet* **18:** R9-17

Sturtevant AH (1913) A Third Group of Linked Genes in Drosophila Ampelophila. *Science* **37:** 990-992

Sultan M, Dokel S, Amstislavskiy V, Wuttig D, Sultmann H, Lehrach H, Yaspo ML (2012) A simple strand-specific RNA-Seq library preparation protocol combining the Illumina TruSeq RNA and the dUTP methods. *Biochemical and biophysical research communications* **422:** 643-646

Sun L, Goff LA, Trapnell C, Alexander R, Lo KA, Hacisuleyman E, Sauvageau M, Tazon-Vega B, Kelley DR, Hendrickson DG, Yuan B, Kellis M, Lodish HF, Rinn JL (2013a) Long noncoding RNAs regulate adipogenesis. *Proc Natl Acad Sci U S A* **110:** 3387-3392

Sun L, Luo H, Bu D, Zhao G, Yu K, Zhang C, Liu Y, Chen R, Zhao Y (2013b) Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic acids research* **41:** e166

Tan MH, Au KF, Yablonovitch AL, Wills AE, Chuang J, Baker JC, Wong WH, Li JB (2013) RNA sequencing reveals a diverse and dynamic repertoire of the Xenopus tropicalis transcriptome over development. *Genome research* **23:** 201-216

Tilgner H, Knowles DG, Johnson R, Davis CA, Chakrabortty S, Djebali S, Curado J, Snyder M, Gingeras TR, Guigo R (2012) Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome research* **22:** 1616-1625

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols* **7:** 562-578

Tripathi V, Ellis JD, Shen Z, Song DY, Pan Q, Watt AT, Freier SM, Bennett CF, Sharma A, Bubulya PA, Blencowe BJ, Prasanth SG, Prasanth KV (2010) The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Molecular cell* **39:** 925-938

Tripathi V, Shen Z, Chakraborty A, Giri S, Freier SM, Wu X, Zhang Y, Gorospe M, Prasanth SG, Lal A, Prasanth KV (2013) Long noncoding RNA MALAT1 controls cell cycle progression by regulating the expression of oncogenic transcription factor B-MYB. *PLoS genetics* **9:** e1003368

Tsoi LC, Iyer MK, Stuart PE, Swindell WR, Gudjonsson JE, Tejasvi T, Sarkar MK, Li B, Ding J, Voorhees JJ, Kang HM, Nair RP, Chinnaiyan AM, Abecasis GR, Elder JT (2015) Analysis of long non-coding RNAs highlights tissue-specific expression patterns and epigenetic profiles in normal and psoriatic skin. *Genome biology* **16:** 24

Tufarelli C, Stanley JA, Garrick D, Sharpe JA, Ayyub H, Wood WG, Higgs DR (2003) Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. *Nature genetics* **34:** 157-165

Ulitsky I, Bartel DP (2013) lincRNAs: genomics, evolution, and mechanisms. *Cell* **154:** 26-46

Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP (2011) Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147:** 1537-1550

van Dijk M, Thulluru HK, Mulders J, Michel OJ, Poutsma A, Windhorst S, Kleiverda G, Sie D, Lachmeijer AM, Oudejans CB (2012) HELLP babies link a novel lincRNA to the trophoblast cell cycle. *J Clin Invest* **122:** 4003-4011

van Werven FJ, Neuert G, Hendrick N, Lardenois A, Buratowski S, van Oudenaarden A, Primig M, Amon A (2012) Transcription of two long noncoding RNAs mediates mating-type control of gametogenesis in budding yeast. *Cell* **150:** 1170-1181

Vance KW, Ponting CP (2014) Transcriptional regulatory functions of nuclear long noncoding RNAs. *Trends in genetics : TIG* **30:** 348-355

Vitiello M, Tuccoli A, Poliseno L (2015) Long non-coding RNAs in cancer: implications for personalized therapy. *Cell Oncol (Dordr)* **38:** 17-28

Vlatkovic I. (2010a) PhD thesis: Mapping and characterization of macro non-protein coding RNAs in human imprinted gene regions, University of Vienna.

Vlatkovic I (2010b) PhD thesis: Mapping and characterization of macro non-protein coding RNAs in human imprinted gene regions, University of Vienna; available for download at http://othes.univie.ac.at/12494/1/2010-09-01_0642621.pdf

Vrbsky J, Akimcheva S, Watson JM, Turner TL, Daxinger L, Vyskot B, Aufsatz W, Riha K (2010) siRNA-mediated methylation of Arabidopsis telomeres. *PLoS genetics* **6:** e1000986

Wahlestedt C (2013) Targeting long non-coding RNA to therapeutically upregulate gene expression. *Nature reviews Drug discovery* **12:** 433-446

Wang A, Huang K, Shen Y, Xue Z, Cai C, Horvath S, Fan G (2011a) Functional modules distinguish human induced pluripotent stem cells from embryonic stem cells. *Stem Cells Dev* **20:** 1937-1950

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* **456:** 470-476

Wang KC, Chang HY (2011) Molecular mechanisms of long noncoding RNAs. *Molecular cell* **43:** 904-914

Wang KC, Yang YW, Liu B, Sanyal A, Corces-Zimmerman R, Chen Y, Lajoie BR, Protacio A, Flynn RA, Gupta RA, Wysocka J, Lei M, Dekker J, Helms JA, Chang HY (2011b) A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472:** 120-124

Wang L, Park HJ, Dasari S, Wang S, Kocher JP, Li W (2013) CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic acids research* **41:** e74

Wapinski O, Chang HY (2011) Long noncoding RNAs and human disease. *Trends in cell biology* **21:** 354-361

Ward LD, Kellis M (2012) Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* **337:** 1675-1678

Washietl S, Findeiss S, Muller SA, Kalkhof S, von Bergen M, Hofacker IL, Stadler PF, Goldman N (2011) RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. *RNA* **17:** 578-594

Washietl S, Kellis M, Garber M (2014) Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res* **24:** 616-628

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korf I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Reymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhausern AC, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang SP, Zdobnov EM, Zody MC, Lander ES (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520-562

Watson JD, Crick FH (1953) Genetical implications of the structure of deoxyribonucleic acid. *Nature* **171:** 964-967

Weikard R, Hadlich F, Kuehn C (2013) Identification of novel transcripts and noncoding RNAs in bovine skin by deep next generation sequencing. *BMC Genomics* **14:** 789

West AC, Johnstone RW (2014) New and emerging HDAC inhibitors for cancer treatment. *J Clin Invest* **124:** 30-39

White NM, Cabanski CR, Silva-Fisher JM, Dang HX, Govindan R, Maher CA (2014) Transcriptome sequencing reveals altered long intergenic non-coding RNAs in lung cancer. *Genome biology* **15:** 429

Whitney AR, Diehn M, Popper SJ, Alizadeh AA, Boldrick JC, Relman DA, Brown PO (2003) Individuality and variation in gene expression patterns in human blood. *Proceedings of the National Academy of Sciences of the United States of America* **100:** 1896-1901

Wightman B, Ha I, Ruvkun G (1993) Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell* **75:** 855-862

Wilusz JE (2015) Long noncoding RNAs: Re-writing dogmas of RNA processing and stability. *Biochim Biophys Acta*

Wright MW (2014) A short guide to long non-coding RNA gene nomenclature. *Hum Genomics* **8:** 7

Wu L, Murat P, Matak-Vinkovic D, Murrell A, Balasubramanian S (2013) Binding interactions between long noncoding RNA HOTAIR and PRC2 proteins. *Biochemistry* **52:** 9519-9527

Wutz A, Smrzka OW, Schweifer N, Schellander K, Wagner EF, Barlow DP (1997) Imprinted expression of the Igf2r gene depends on an intronic CpG island. *Nature* **389:** 745-749

Xie C, Yuan J, Li H, Li M, Zhao G, Bu D, Zhu W, Wu W, Chen R, Zhao Y (2014) NONCODEv4: exploring the world of long non-coding RNA genes. *Nucleic acids research* **42:** D98-103

Yang L, Duff MO, Graveley BR, Carmichael GG, Chen LL (2011) Genomewide characterization of non-polyadenylated RNAs. *Genome biology* **12:** R16

Yang W, Li Y, He F, Wu H (2015) Microarray profiling of long non-coding RNA (lncRNA) associated with hypertrophic cardiomyopathy. *BMC Cardiovasc Disord* **15:** 62

Young RS, Marques AC, Tibbit C, Haerty W, Bassett AR, Liu JL, Ponting CP (2012) Identification and properties of 1,119 candidate lincRNA loci in the Drosophila melanogaster genome. *Genome Biol Evol* **4:** 427-442

Zeng Q, Wang Q, Chen X, Xia K, Tang J, Zhou X, Cheng Y, Chen Y, Huang L, Xiang H, Cao K, Zhou J (2016) Analysis of lncRNAs expression in UVB-induced stress responses of melanocytes. *J Dermatol Sci* **81:** 53-60

Zhang B, Arun G, Mao YS, Lazar Z, Hung G, Bhattacharjee G, Xiao X, Booth CJ, Wu J, Zhang C, Spector DL (2012) The lncRNA Malat1 is dispensable for mouse development but its transcription plays a cis-regulatory role in the adult. *Cell reports* **2:** 111-123

Zhang EB, Kong R, Yin DD, You LH, Sun M, Han L, Xu TP, Xia R, Yang JS, De W, Chen J (2014a) Long noncoding RNA ANRIL indicates a poor prognosis of gastric cancer

and promotes tumor growth by epigenetically silencing of miR-99a/miR-449a. *Oncotarget* **5:** 2276-2292

Zhang H, Gao S, De Geyter C (2009) A natural antisense transcript, BOKAS, regulates the pro-apoptotic activity of human Bok. *Int J Oncol* **34:** 1135-1138

Zhang T, Cooper S, Brockdorff N (2015) The interplay of histone modifications - writers that read. *EMBO Rep* **16:** 1467-1481

Zhang X, Zhou L, Fu G, Sun F, Shi J, Wei J, Lu C, Zhou C, Yuan Q, Yang M (2014b) The identification of an ESCC susceptibility SNP rs920778 that regulates the expression of lncRNA HOTAIR via a novel intronic enhancer. *Carcinogenesis* **35:** 2062-2067

Zhong J, Chuang SC, Bianchi R, Zhao W, Lee H, Fenton AA, Wong RK, Tiedge H (2009) BC1 regulation of metabotropic glutamate receptor-mediated neuronal excitability. *J Neurosci* **29:** 9977-9986

CURRICULUM VITAE

Name: Aleksandra E. Kornienko
Date of birth: January 20th, 1987
Place of Birth: Smolensk, Russia
Nationality: Russian
E-mail: akornienko@cemm.oeaw.ac.at

Higher education:

Diploma degree in Medical Physics
Medical Physics Department, Faculty of Physics, Lomonosov Moscow State University, Moscow, Russia
Graduation date: 02/2009
Academic average: 4.7
(Russian rating system: 5=excellent, 4=good, 3=satisfactory, 2=unsatisfactory)

Research experience:

09/2010 – present, PhD student at the lab of Denise P. Barlow, CeMM - Research Center for Molecular Medicine of the Austrian Academy of Sciences
PhD project: Identification and characterization of long non-protein-coding RNAs in the human genome

02/2009 – 09/2010, Research assistant at the Laboratory of Biological Microchips, Engelhardt Institute of Molecular Biology of the Russian Academy of Sciences, Moscow, Russia
Research project: Forensic personal identification by SNP genotyping using DNA microchips.

10/2007 – 01/2009, Diploma student at the Laboratory of Physical Biochemistry, Research Center for Hematology of the Russian Academy of Medical Sciences, Moscow, Russia
Diploma thesis: Research into the photosensitized human erythrocytes hemolysis.

10/2006 – 05/2007, Internship student at the Cell Biology Group of the Institute of Protein Research of the Russian Academy of Sciences, Moscow, Russia
Course project: Study of mitochondrial motion in living mammalian cells by fluorescence microscopy technique.

Language skills:
English (fluent), German (intermediate), French (basic)

Wet lab skills:
- Cell culture: Standard cell culture maintaining skills, microscopy
- Cell differentiation
- Molecular biology: Standard molecular biology skills (PCR, RNA/DNA isolation, reverse transcription, etc), Cloning in E.coli, Polysome fractionation, Antisense oligonucleotide RNase H mediated knock-down of lncRNAs in the nucleus.
- NGS library preparation
- Handling human blood and isolation of specific cell types

Bioinformatic skills
- RNA-seq and ChIP-seq data processing and analysis
- Handling massive RNA-seq data from multiple samples and analysis automation
- *De novo* lncRNA/mRNA identification and annotation
- R scripting
- Basic bash scripting, working with SLURM scripts.
- UCSC browser, hub creation

Publications:
1. Kornienko AE, Dotter CP, Guenzl PM, Gisslinger H, Gisslinger B, Cleary C, Kralovics R, Pauler FM, Barlow DP (2016) Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans. *Genome Biol* **17:** 14

2. Kornienko AE, Vlatkovic I, Neesen J, Barlow DP, Pauler FM (2015) A human haploid gene trap collection to study lncRNAs with unusual RNA biology. *RNA Biol***:** 0

3. Kornienko AE, Guenzl PM, Barlow DP, Pauler FM (2013) Gene regulation by the act of long non-coding RNA transcription. *BMC biology* **11:** 59

4. Fesenko DO, Ivanov PL, Kornienko AE, Zemskova E, Zasedatelev AS, Nasedkina TV (2011) [Optimization of the biological microchip for genotyping the AB0 locus: analytical aspects]. *Sud Med Ekspert* **54:** 30-33

5. Ivanov PL, Fesenko DO, Kornienko AE, Roi OV, Zemskova E, Zasedatelev AS, Nasedkina TV (2011) [Optimization of the biological microchip for genotyping of the AB0 locus: correction of DNA probes]. *Sud Med Ekspert* **54:** 32-35

Courses taken:

1) Neuroepigenetics course, Neuroscience School of Advanced Studies, Florence, Italy, May 2015
2) Applied Statistics in Biological Research course, CSF (The Campus Science Support Facilities), Vienna Biocenter, Vienna, Austria, June 2014
3) Adobe Illustrator course, MFPL, Vienna Austria, November 2013
4) Next Generation Sequencing Workshop, European Bioinformatics Institute, Hinxton, UK, March 2013
5) Introduction to R course, MFPL, Vienna, Austria, August 2012

Conferences attended and presentations given:

1) Personalized Health, EMBL|Stanford conference, Heidelberg, Germany, November 2015, poster presentation, Flash Talk oral presentation
2) Emerging Biotechnologies, EMBL Science and Society Conference, Heidelberg, Germany, November 2015
3) Vienna RNA Meeting, Vienna, Austria, October 2015, poster presentation
4) SFB "RNA Regulation of the Transcriptome"/DK RNA Biology annual retreat, Retz, Austria, April 2015, oral presentation
5) Brain Forum, Lausanne, Switzerland, March 2015
6) Long Regulatory RNAs, ESF conference, Pultusk, Poland, September 2014, poster selected for an oral presentation.
7) Vienna RNA Meeting, IMBA, Vienna, Austria, October 2013, poster presentation
8) The Non-Coding Genome, EMBL conference, Heidelberg, Germany, October 2013, poster presentation

9) SFB “RNA Regulation of the Transcriptome”/DK RNA Biology annual retreat, Aflenz, Austria, April 2013, oral presentation
10) SFB “RNA Regulation of the Transcriptome”/DK RNA Biology annual retreat, Aflenz, Austria, April 2012, poster presentation
11) Non-Coding RNA, Keystone Meeting, Salt Lake City, USA, March 2012, poster presentation
12) EMBO meeting, Vienna, Austria, September 2011

ACKNOWLEDGEMENTS

I would like to thank my supervisor Denise Barlow for, first of all, accepting me to her lab and giving me a chance to do a PhD with her. I would like to thank her for believing in me from the moment of the interviews and throughout my whole PhD. I want to thank her for her constant support, guidance and encouragement. For her admirable discipline, knowledge and attention to detail, all of which I tried to learn from her. I would like to thank her for our lab – the best organized and tidy lab I have ever seen and will ever see, and for teaching us this organization. I would also like to thank her for giving me freedom when working on my PhD project, while always giving me input and guidance. I would like to also thank her for teaching me presentation and writing skills. In general, I have learned so much during my PhD that I literally became a different person, and I am grateful to Denise for that. I would also like to thank her for always choosing great people for working in her lab, which provided me with a wonderful team and atmosphere to work in. And last, but not least I would like to thank her for all the supervision, help and guidance while working on the papers - a huge amount of work that, as I know, very few supervisors invest into their PhD students. I would like to thank her for making it possible that I was able to finish my PhD after her retirement. Overall, I would like to thank Denise for everything.

I would like to thank my PhD committee members – Robert Kralovics and Ivo Hofacker – for their support and thoughtful insights and suggestions. I would like to thank Giulio Superti-Furga for leading CeMM so efficiently for all the years I was part of it and spreading his enthusiasm and the big picture of us “doing something big” to the whole CeMM crew. I would like to thank CeMM PhD program for support and my 2010 PhD student team – Elisangela, Lisi, Branka, Jelena, Ana, Ferran, Rui, Marco, Sabrina, Johannes and Astrid – for being a group of very different, but all brilliant, cool and inspiring people I was happy to meet and share my PhD times with.

I would like to thank Austria for being a nice and welcoming country that is wisely investing into science, which allowed me to become a scientist. I could never do science of that level in my home country.

I would like to thank DK RNA Biology program for lots of support in travelling and learning and the whole Vienna RNA community, especially Renée Schroeder and Andrea Barta, for being a cozy Vienna RNA family and for the wonderful DK RNA Biology retreats.

I would like to thank our lab – I was lucky to be a part of the best lab I could wish for, with lots of great and brilliant people and the best coffee breaks. Thanks to Federica Santoro for being a never-bleaching sunshine, for her amazing cakes and her outstanding kindness. Thanks to Philipp Guenzl for being my desk/office neighbor for all these years, I could never wish for a better neighbor, I will never forget all our discussions and arguments about everything on earth. I learned a lot from him. I would like to thank Florian Pauler for his special never-ending humor, for giving me challenges throughout my PhD and building my character through it. I would like to thank Tomek Kulinski for our endless discussions about life, politics and literature, for his kindness, thoughtfulness, deepness and for being the most gentlemen I know. I would like to thank Christoph Dotter for helping me so much with bioinformatics and for being a great friend in our lab times. I would like to thank Philipp Bammer, Christoph Dotter and Markus Muckenhuber for being a great lab Master student team, it was a lot of fun. I would like to thank Daniel Andergassen for his endless questions and his constant fear that he is annoying me, for being so passionate about imprinting that it is infectious. I would like to thank Irena Vlatkovic for being my friend when I just joined the lab, for being very kind and gentle, for teaching me many things and helping me so much when I was starting my PhD. I would like to thank Quanah Hudson for his kindness and support and for telling us cool stories about Australia and New Zeeland. I would like to thank Ruth Klement and Kasia Warczok – two best technicians in the world and great, kind and smart ladies.

I would like to thank my friend Xi who supported me throughout the last years with all my ups and downs an opened me the world of drawing that sparkled my life. I would like to thank my friend Igor Smirnov, who showed me that going abroad to do a PhD is not just an impossible dream, but it can actually come true. Without him I would never dare to apply anywhere.

Finally, I would like to thank my parents and my brother – my family. I would like to thank my parents for their constant support throughout my whole life. For their kindness, understanding and bearing with my ups and downs. I wish they had a chance do a PhD abroad, they would both become greatest scientist if they had the chance I was lucky to have. I thank them dearly, without them I could not do anything I have done.

APPENDIX

Supplemental Figures and Methods for **Publication 2** (referred to as Additional File 1)

Supplemental Tables for **Publication 2** (referred to as Additional Files 2 and 11)

Supplemental Figures for **Publication 3**

Supplemental Tables for **Publication 3**

ADDITIONAL FILE 1 (Kornienko et al.)

Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans

Aleksandra E. Kornienko*, [1]Christoph P. Dotter, Philipp M. Guenzl, [2]Heinz Gisslinger, [2]Bettina Gisslinger, [3]Ciara Cleary, Robert Kralovics, Florian M. Pauler, Denise P. Barlow*

CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Lazarettgasse 14, AKH BT 25.3, 1090 Vienna, Austria

[1]Present address: Institute of Science and Technology Austria, Lab Building East, Am Campus 1, A-3400 Klosterneuburg, Austria

[2]Department of Hematology, Medical University of Vienna, Vienna, Austria

[3]Present address: Piso 23, Av. Santa Fe No 481, Lomas de Santa Fe, 05349, Mexico D.F.

*Corresponding authors (akornienko@cemm.oeaw.ac.at, dbarlow@cemm.oeaw.ac.at)

TABLE OF CONTENTS

SUPPLEMENTAL FIGURES & LEGENDS (S1 – S35)

Figure S1

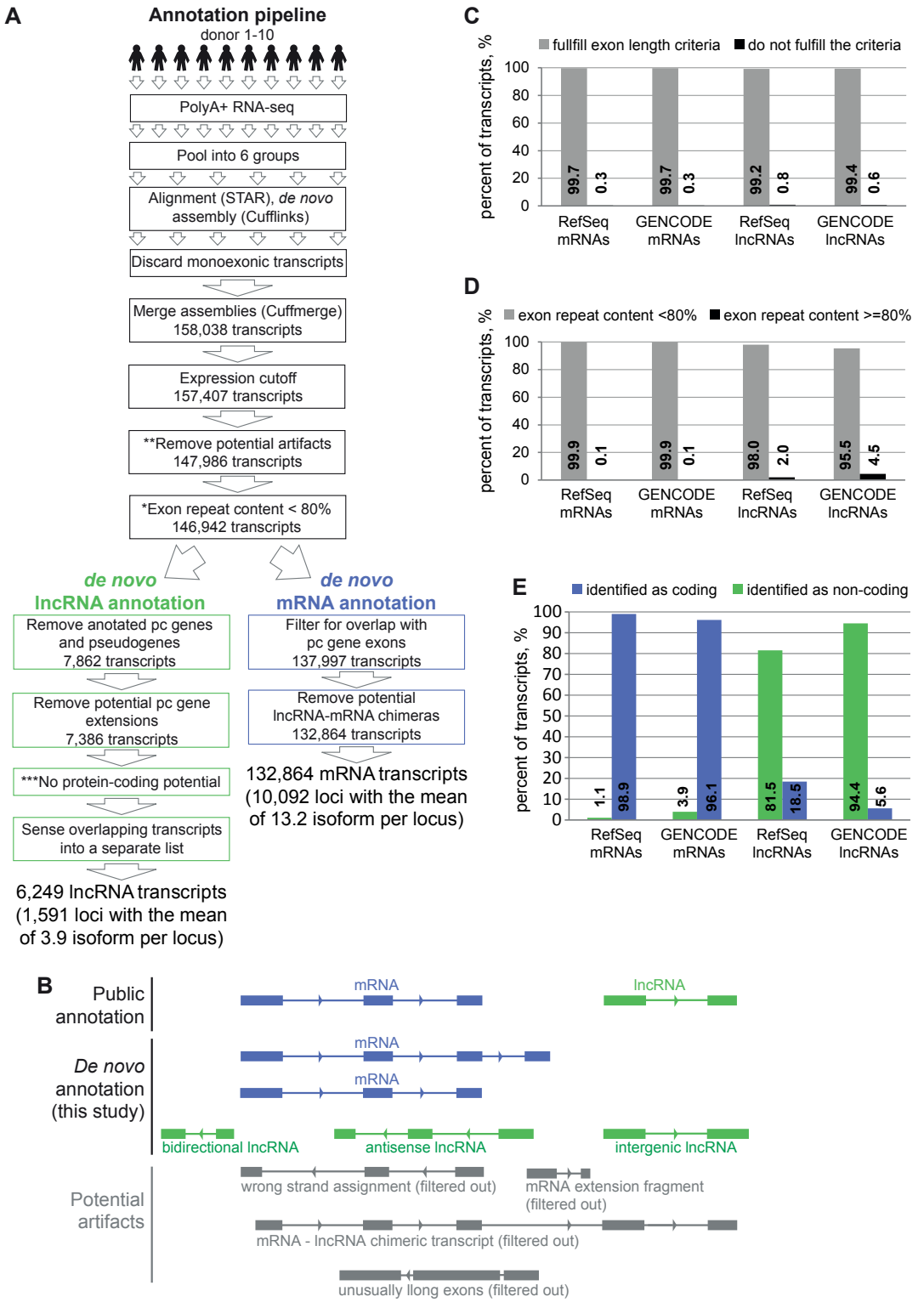


Figure S1. *De novo* lncRNA and mRNA annotation in granulocytes

A. Algorithm for *de novo* lncRNA and mRNA identification showing the number of transcripts identified at different steps (Supplemental Methods). At early steps filtering is performed for all

transcripts (lncRNA and mRNA) together. The filtering splits after the step removing transcripts with high exon repeat content*. The filtering pipeline identifies 6,249 lncRNA transcripts (corresponding to 1,591 lncRNA loci) and 132,864 mRNA transcripts (corresponding to 10,092 loci) in granulocytes.

B. Schematic representation of potential artifacts to be removed from the transcriptome assembly during filtering (step **). Top shows an annotated hypothetical mRNA (blue) and lncRNA (green) gene. Underneath is shown mRNA and lncRNA transcripts typically annotated in this study (alternative start or end sites of annotated transcripts was a frequent finding). Bottom shows transcripts (gray) annotated in this study that were filtered out as potential artifacts. These include wrong strand assignment of mRNAs (arising from poor strand-specificity of the RNA-seq), mRNA extension fragments, mRNA - lncRNA chimeric transcripts, and transcripts with unusually long exons (Supplemental Methods).

C. Validation of the filtering step that removes unusually long exons (part of step**). The bar plot shows the percentage of RefSeq and GENCODE v19 multi-exonic mRNAs (left) and lncRNAs (right) that fulfill (grey) or do not fulfill (black) the exon length filtering criteria. The vast majority of annotated multi-exonic mRNA and lncRNA transcripts pass this filtering step.

D. Validation of the step* to filter out repeat-rich transcripts. The bar plot shows the percentage of RefSeq and GENCODE v19 multi-exonic mRNAs (left) and lncRNAs (right) with <80% (pass the cut off - grey) or >=80% (do not pass the cut off - black) exonic content of repeats (http://www.repeatmasker.org/ [1]). The vast majority of annotated multi-exonic transcripts pass this filtering step.

E. Validation of the protein-coding potential estimation step***. The bar plot shows the percentage of RefSeq and GENCODE v19 multi-exonic mRNAs (left) and lncRNAs (right) identified as protein-coding (blue) and non-protein-coding (green). Nearly all mRNAs are identified as protein-coding, while most but not all, lncRNAs are identified as non-protein-coding.

Figure S2

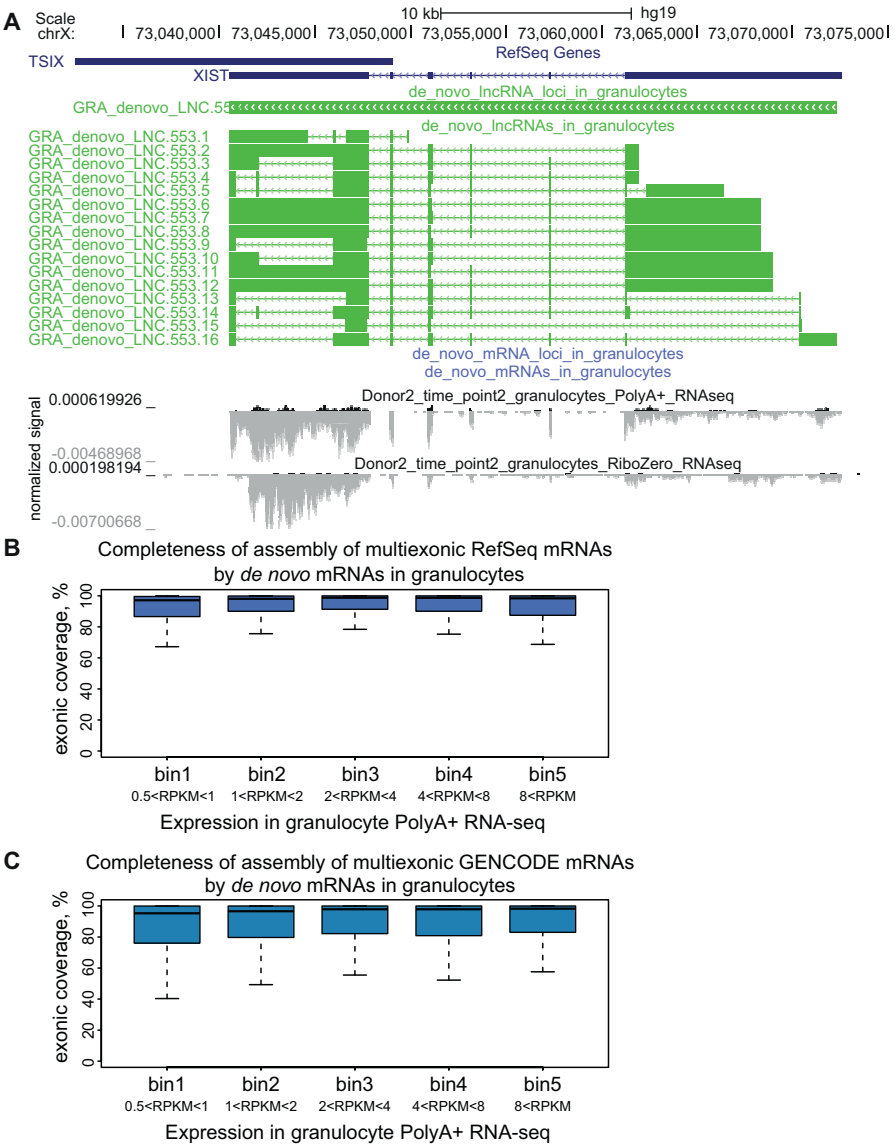


Figure S2. Quality of *de novo* transcriptome annotation in granulocytes

A. Assembling known lncRNAs - UCSC browser screen shot of an example of complete coverage of a well-known *XIST* lncRNA by *de novo* lncRNA annotation in granulocytes. From top to bottom: RefSeq gene annotation (the antisense *TSIX* lncRNA is cropped at the 5' end), *de novo* lncRNA loci annotation, *de novo* lncRNA transcript annotation (the pipeline annotates various isoforms of *XIST), de novo* mRNA loci and transcript annotation (empty in this genomic region), normalized PolyA+ RNA-seq signal for granulocytes from donor n2 (time point 2) and ribosomal depleted (Ribo Zero) RNA-seq of the same granulocyte sample (Additional File 2B). Both RNA-seq tracks show the presence of *XIST* and an absence of *TSIX* in granulocytes of Donor 2 (Additional File 2A).

B and C. Completeness of *de novo* assembly. Exonic coverage (Methods) of RefSeq (**B**) and GENCODE v19 (**C**) annotated mRNAs by *de novo* mRNA annotation in granulocytes. Genes are split into 5 bins, according to their average expression level in PolyA+ granulocyte RNA-seq samples used for transcriptome assembly (Additional File 2B), in order to account for the bias between expression level and assembly success. Median levels from left to right: **B**: 97.2, 98.1, 98.8, 98.7, 98.4; **C**: 95.3, 96.6, 97.9, 97.9, 98.2. Outliers are not displayed in the boxplots.

Figure S3

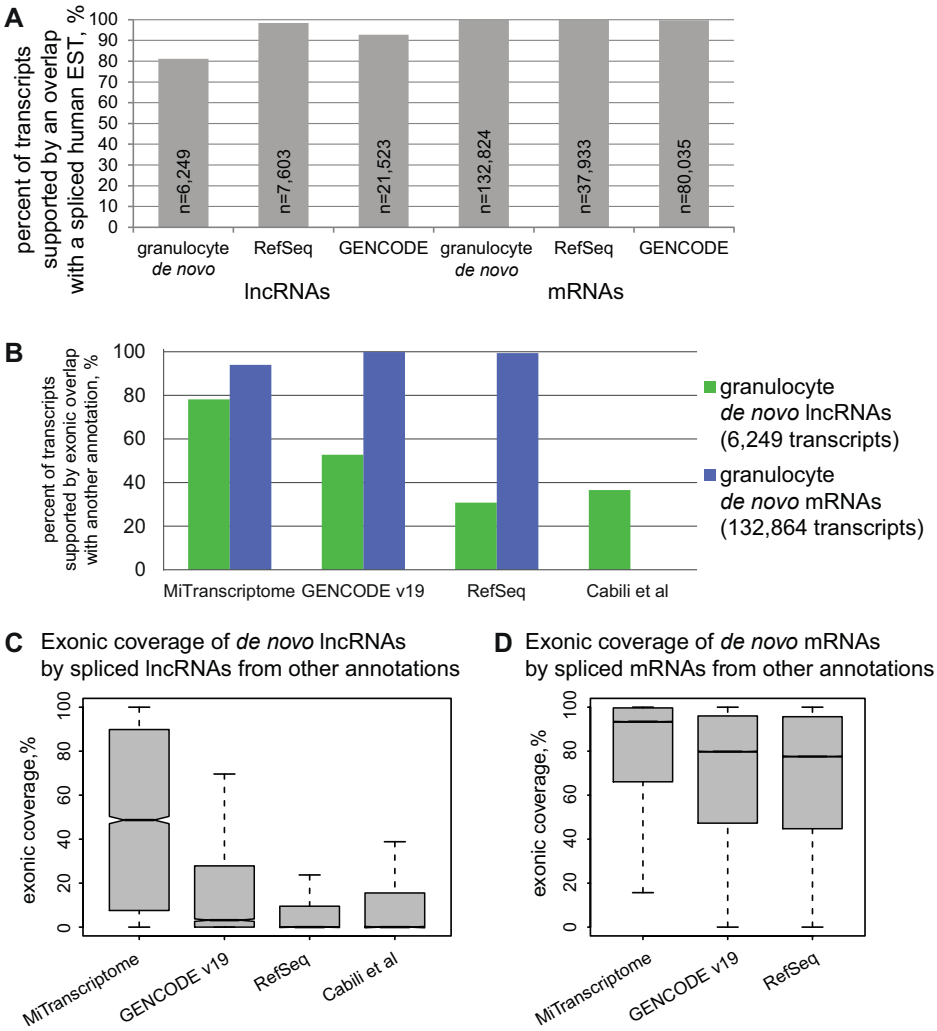


Figure S3. Validation of granulocyte *de novo* lncRNAs by overlap with other annotations

A. The majority of *de novo* lncRNAs annotated in granulocytes are supported by an overlap with an EST. The bar plot shows the percentage of transcripts that have a sense exonic overlap with a spliced EST (human EST database downloaded from UCSC) for (from left to right): *de novo* lncRNAs annotated in this study in granulocytes, lncRNAs annotated by RefSeq and GENCODE v19, *de novo* mRNAs annotated in this study in granulocytes, mRNAs annotated by RefSeq and GENCODE v19. Numbers inside bars indicate total number of transcripts in each annotation.

B. The majority of *de novo* lncRNAs in granulocytes are supported by an overlap with MiTranscriptome lncRNAs. The bar plot shows percentage of granulocyte *de novo* lncRNA (green) and mRNA (blue) transcripts supported by an exonic overlap with a multi-exonic lncRNA or mRNA respectively from MiTranscriptome, GENCODE v19, RefSeq and Cabili et al (only lncRNAs) annotations [2-5].

C. MiTranscriptome provides best exonic coverage for granulocyte *de novo* lncRNAs. The box plot shows percent exonic coverage (Methods) of granulocyte *de novo* lncRNAs by multi-exonic lncRNAs from MiTranscriptome, GENCODE v19, RefSeq and Cabili et al annotations. Number of granulocyte *de novo* lncRNA transcripts examined in each box – 6,249. Median values left to right: 48.7, 3.2, 0, 0.

D. Granulocyte *de novo* mRNAs are nearly fully exonically covered by MiTranscriptome and public annotations. The box plot shows percent exonic coverage of granulocyte *de novo* mRNAs by multi-exonic mRNAs from MiTranscriptome, GENCODE v19 and RefSeq annotations. Number of granulocyte *de novo* mRNA transcripts examined in each box – 132,864. Median values left to right: 93.4, 79.8, 77.6. Remarks: Outliers are not displayed in the box plots.

Figures S4-S8. Validation of granulocyte *de novo* lncRNAs by cloning

(N.B., Figures S4-S8 have a common legend)

Cloning result for each lncRNA locus *de novo* annotated in granulocytes is represented by a UCSC browser screen shot. Each screen shot contains (from top to bottom): chromosome scale, chromosome coordinates, RefSeq gene annotation, granulocyte *de novo* mRNA loci annotation obtained in the study (blue), granulocyte *de novo* lncRNA loci annotation obtained in the study (green), *de novo* lncRNA transcripts constituting the given lncRNA locus (green), RT-PCR primers used to amplify the targeted full length lncRNA transcript (Additional File 2F), BLAT alignments of Sanger sequences obtained after cloning the targeted lncRNA transcript (black). For the loci containing several isoforms, those used for primer design are marked with (*). RT-PCR and cloning procedure do not preserve strandness of the transcript and, thus, the BLAT alignments' strands do not necessarily match the corresponding lncRNA strand. Sanger sequencing results ("_T7" tag for T7 primer used) from some cloned products could not cover the full-length transcript and these products were sequenced from the other end ("_SP6" tag for SP6 primer used). Red lines in BLAT alignments indicate mismatches between the UCSC reference genome sequence and the sequence of the aligned Sanger sequences, the width of a red line does not scale with the size of the mismatch.

Figure S4 (legend – p.8)

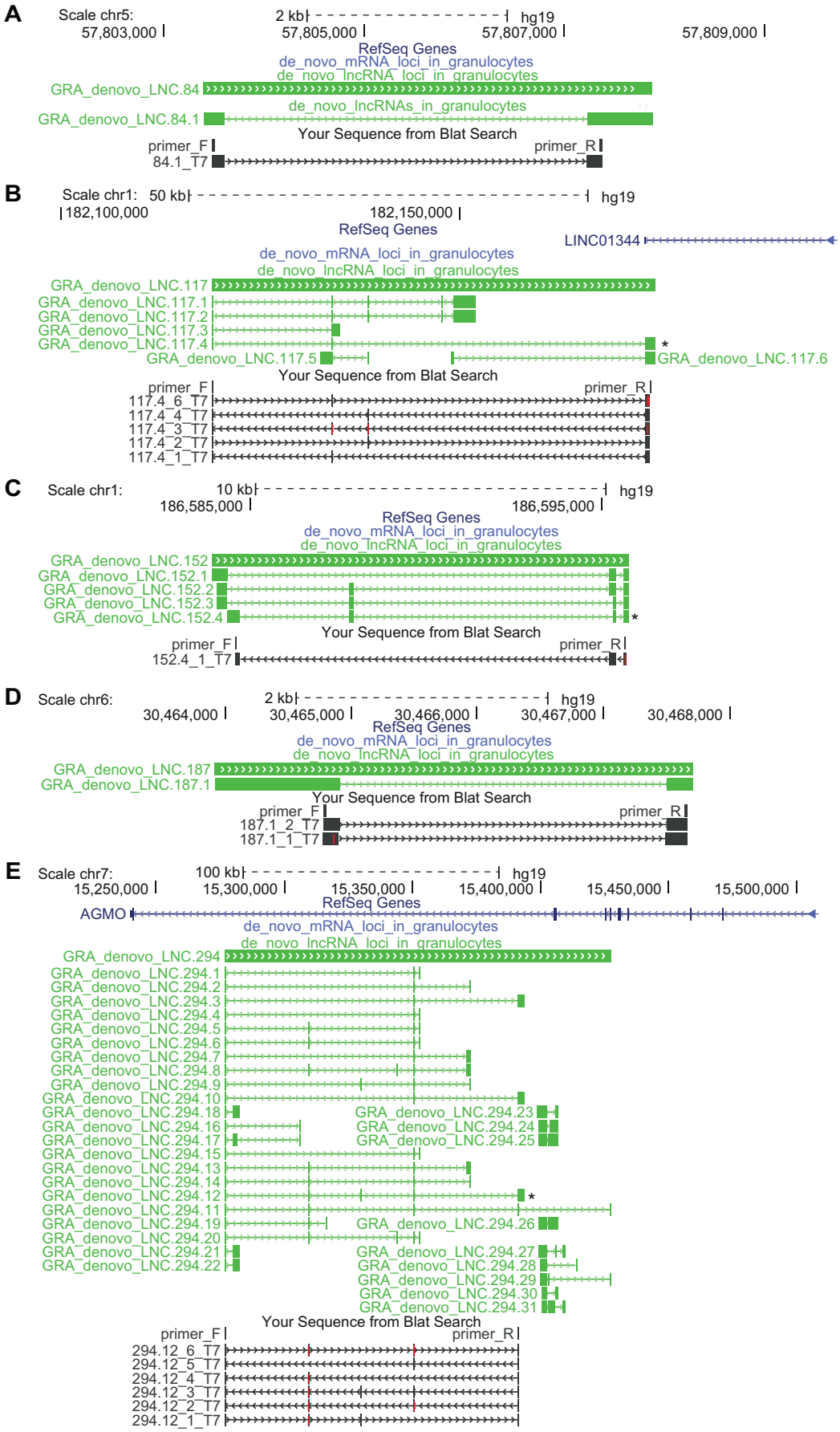


Figure S5 (legend – p.8)

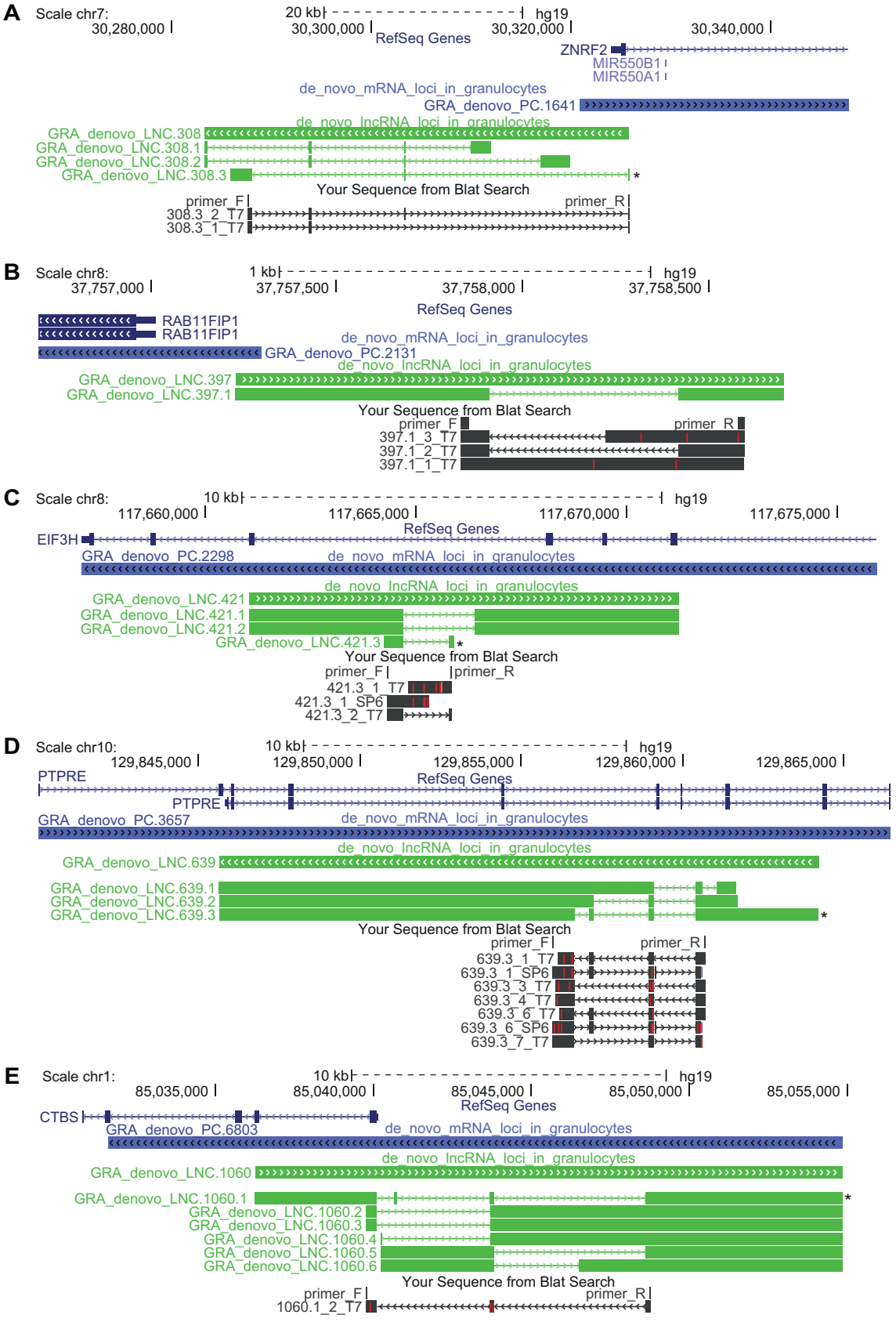


Figure S6 (legend – p.8)

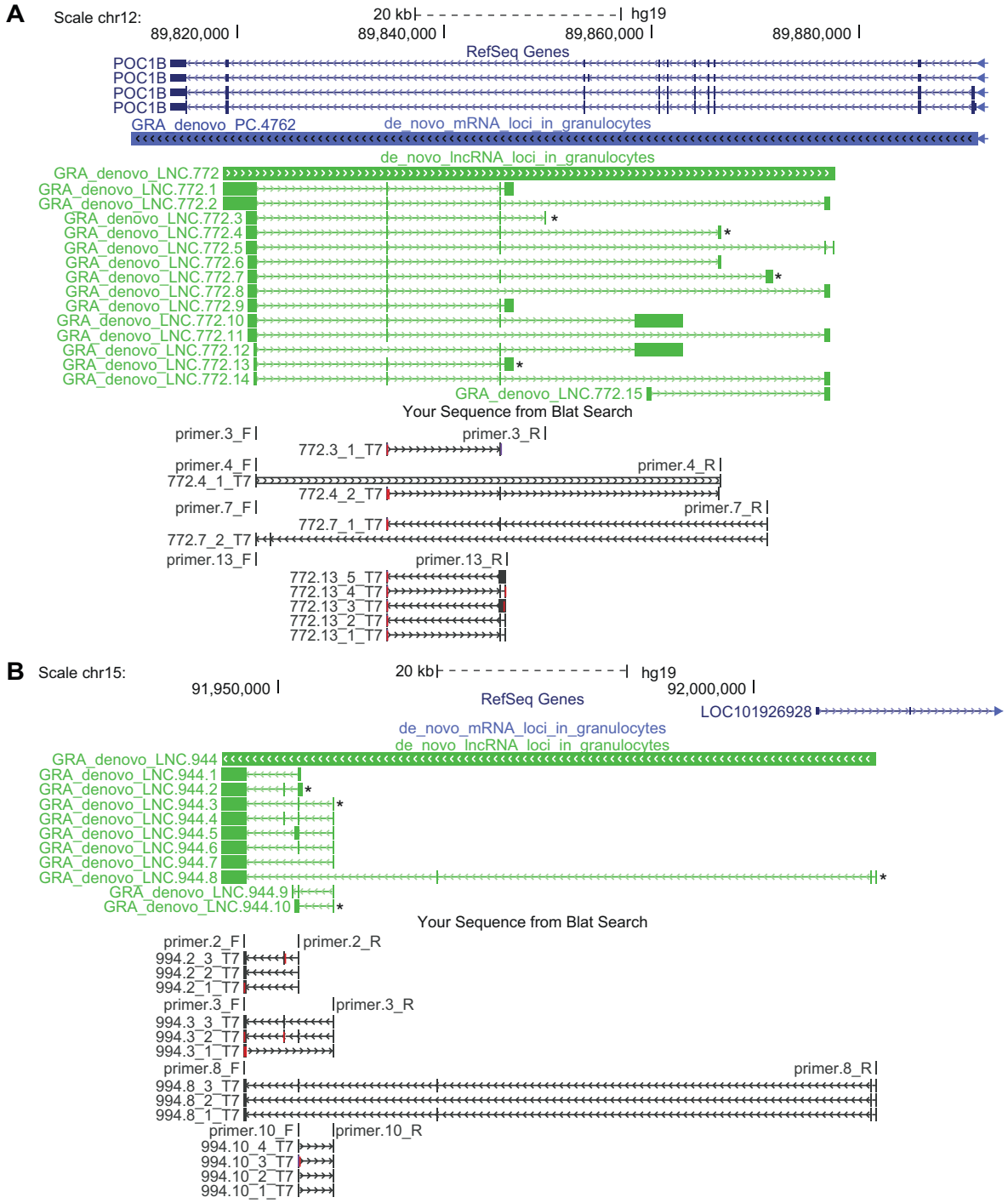


Figure S7 (legend – p.8)

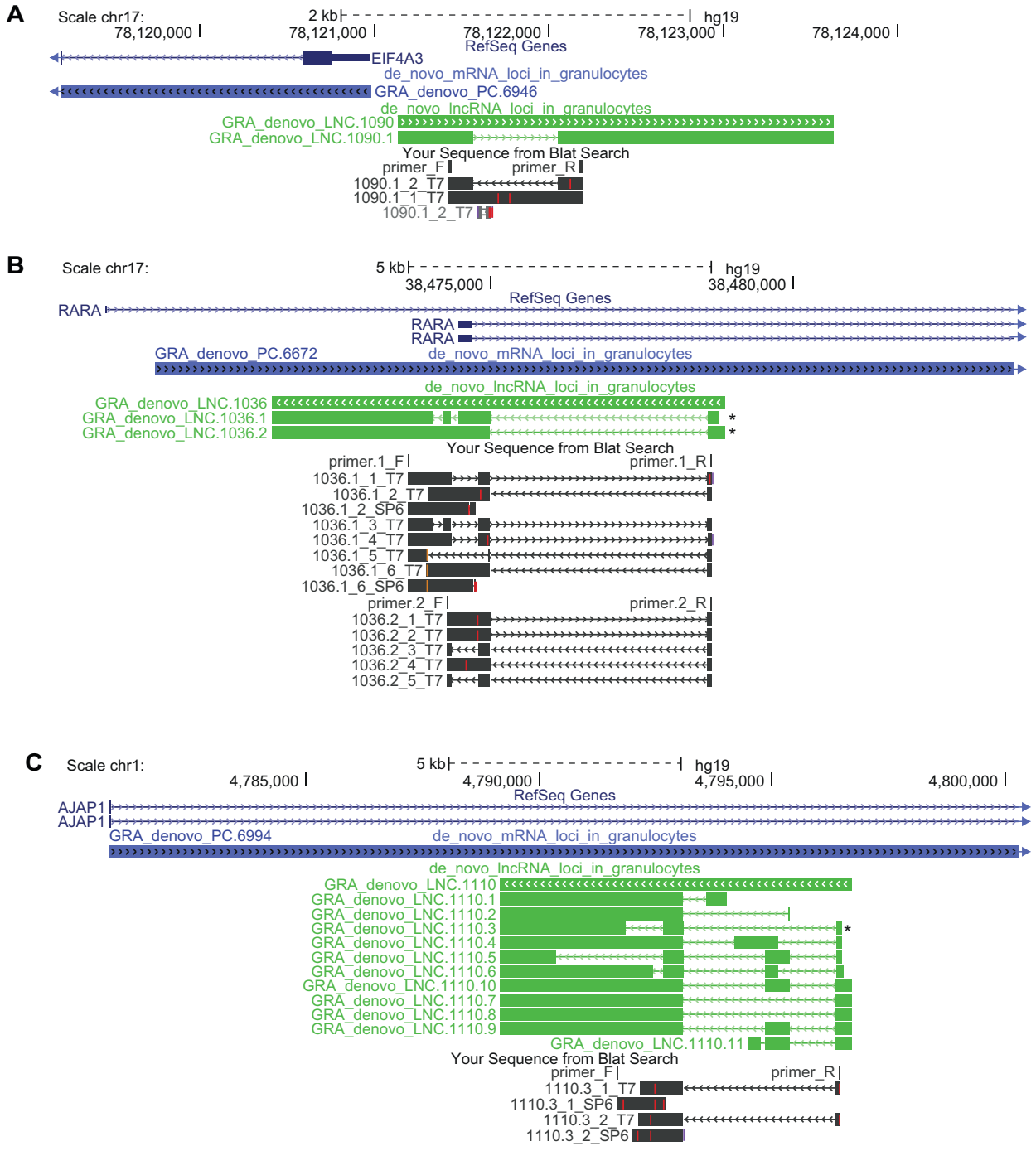


Figure S8 (legend – p.8)

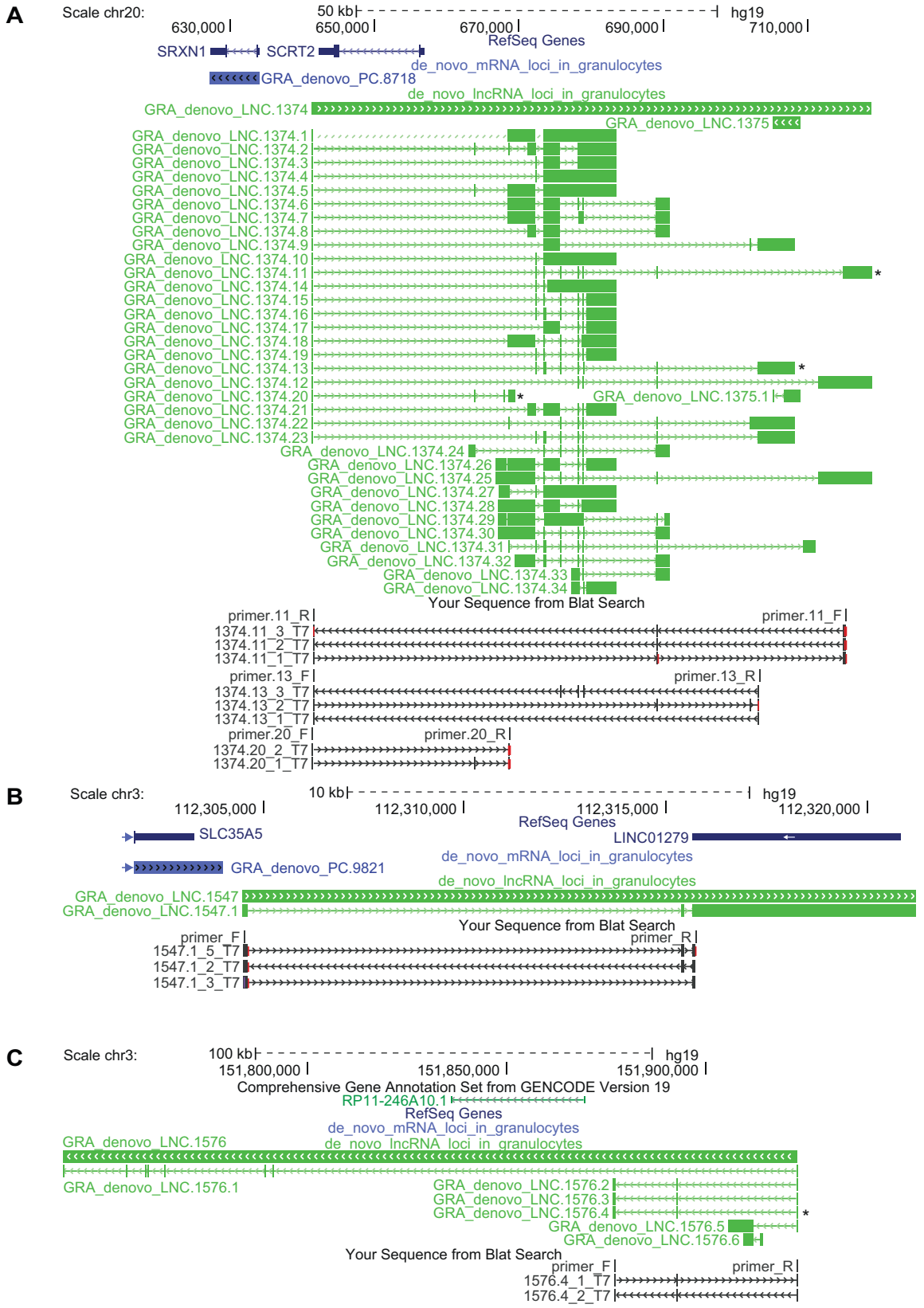


Figure S9

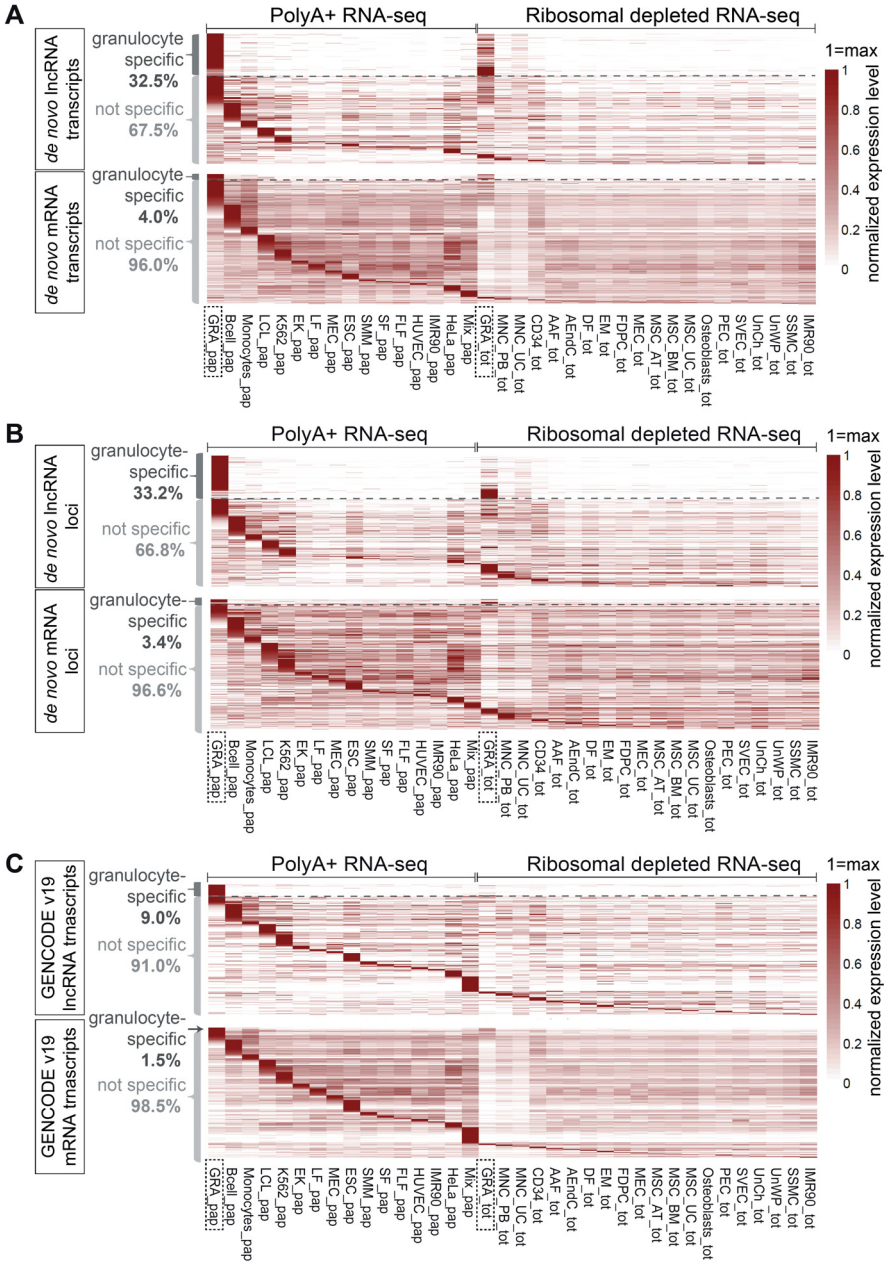


Figure S9. Granulocyte specificity of granulocyte *de novo* and GENCODE annotated lncRNAs and mRNAs

Each heat map represents expression levels of **(A)** *de novo* granulocyte transcripts - top: lncRNAs (5,936 transcripts) in the heat map), bottom: mRNAs (132,813 transcripts in the heat map) **(B)** *de novo* granulocyte loci - top: lncRNA loci (1,218 loci in the heat map), bottom: mRNA loci (9,946 loci in the heat map), **(C)** GENCODE v19 multi-exonic transcripts - top: lncRNAs (13,489 transcripts in the heat map), bottom: mRNAs (77,529 transcripts in the heat map). Heat maps show expression level RPKM (reads per kilobase of transcript per million reads mapped) of each transcript/locus in 34 public strand-specific human RNA-seq samples (Additional File 2H) normalized by maximal expression among all samples (maximum is set to 1). GRA_pap (dotted box): average expression level among 17 PolyA+ RNA-seq samples from 10-donors, GRA_tot (dotted box): average expression level among 21 ribosomal depleted RNA-seq samples from 7-donors (Additional File 2B). pap: polyA+ RNA (15 samples), tot: ribosomal depleted RNA (19 samples). Above dashed line: transcripts/loci defined as "granulocyte-specific" (dark grey brace) show maximal expression in the GRA_tot or GRA_pap samples and ≥3-fold lower expression in all other samples. Below dashed line: transcripts/loci not meeting these criteria and called not specifically expressed in granulocytes (light grey brace). Only transcripts/loci expressed (RPKM>0.2) in at least one of the samples were analyzed.

Figure S10

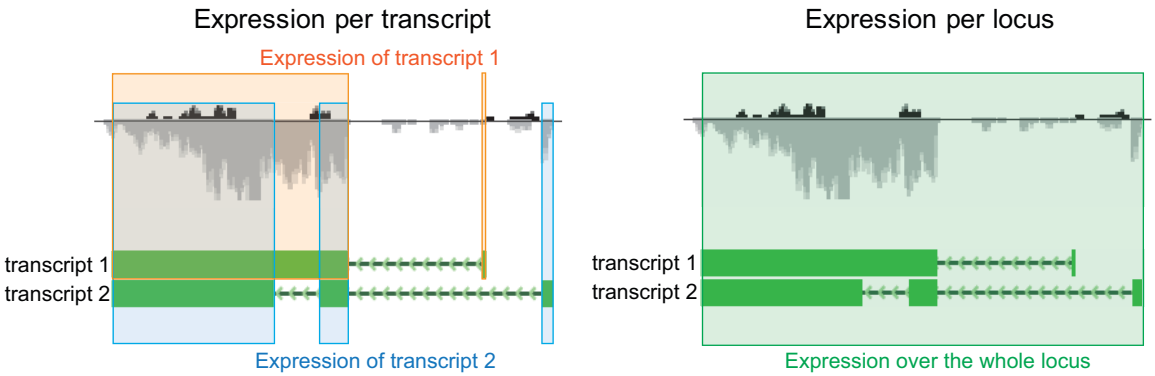


Figure S10. Illustration of expression calculation per transcript and over whole locus

Throughout the study expression of a transcript is calculated as RPKM of only its exons using BED12 transcript annotation which provides information on position of exons (left), thus only the reads that map to exons are taken into account. Locus expression is calculated for the whole locus, thus all the reads, both exonic and intronic, are counted (right).

Figure S11

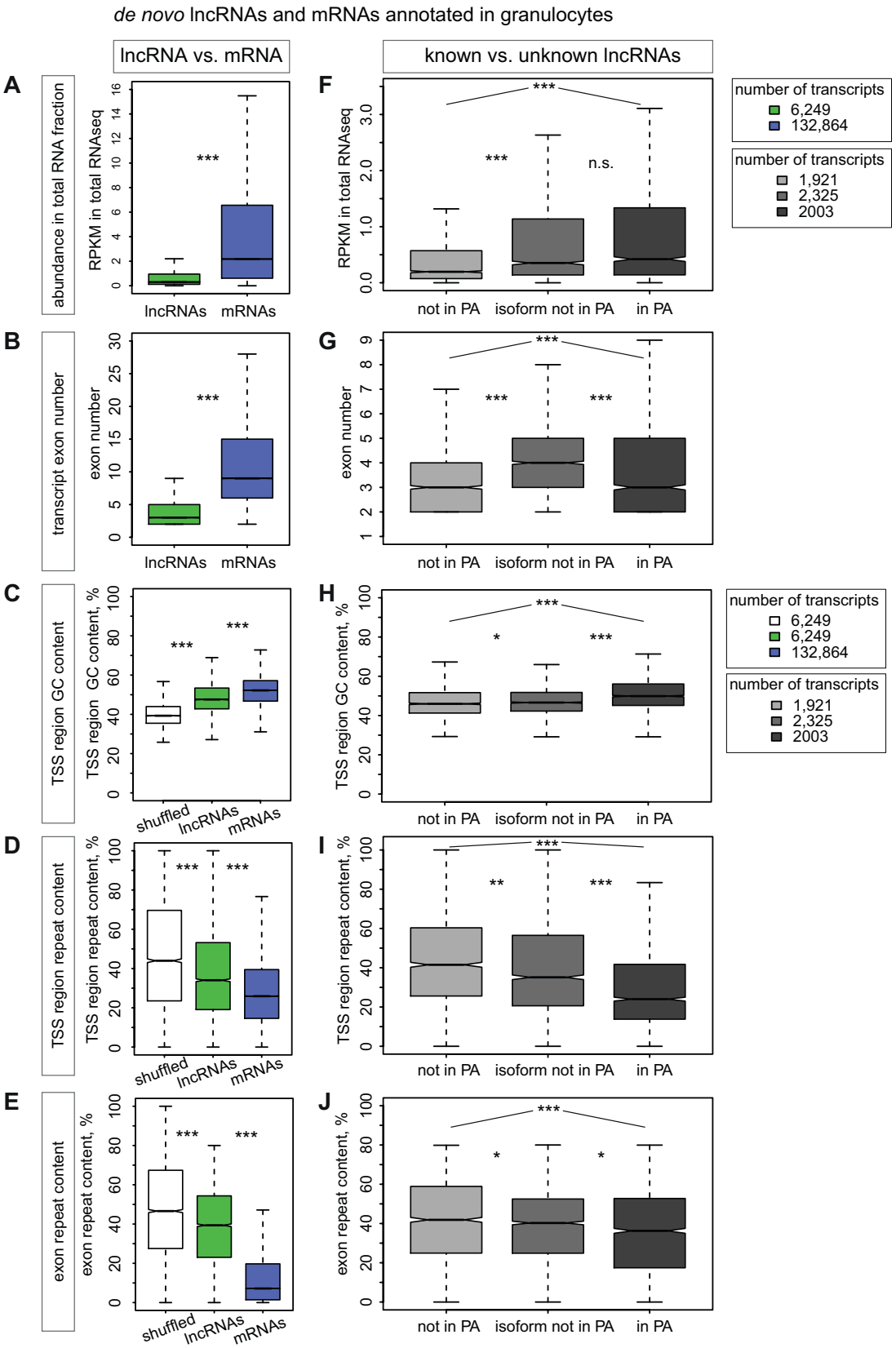


Figure S11. Features that distinguish granulocyte lncRNAs from mRNAs also distinguish novel granulocyte lncRNAs from publicly annotated lncRNAs

A. Abundance in total RNA fraction of granulocyte *de novo* lncRNA and mRNA transcripts was calculated as an average from all 21 available Ribosomal Depleted RNA-seq samples. Exon number (**B**), percent GC content of transcription start site (TSS) region (TSS +/- 1.5 kb) (**C**), percent repeat coverage of TSS region (**D**) and exons (**E**) of *de novo* annotated lncRNAs (green) and mRNA (blue) transcripts. Abundance in total RNA fraction (**F**), exon number (**G**), percent GC content of TSS region (**H**), percent repeat coverage of TSS region (**I**) and exons (**J**) of the three classes of *de novo* lncRNAs (described in Fig. 2A). **C**, **D** and **E**: "shuffled" control (white) is added to the box plots. Shuffled control represents random regions in the genome using *bedtools shuffle*. C and D: granulocyte *de novo* lncRNA TSS regions (n=6,249) were shuffled across the genome, E: granulocyte *de novo* lncRNA transcripts (n=6,249) were shuffled across the genome.
Remarks to boxplots: green: all *de novo* lncRNAs, blue: all de novo mRNAs, light gray: 'not in PA' lncRNA transcripts, medium gray: 'isoform not in PA' lncRNA transcripts, dark gray: 'in PA' lncRNA transcripts (PA: public annotations). The numbers of transcripts in each box are indicated top right. *** - $p<10^{-10}$, ** - $p<10^{-5}$, * - $p<0.01$, n.s. – $p>0.01$. The box plots display the full population but p-values are calculated using Mann–Whitney U test with equalized sample size (Methods). Median values from left to right: **A**: 0.31, 2.18; **B**: 3, 9; **C**: 47.6, 52.2; **D**: 34.0, 25.9; **E**: 39.4, 7.2; **F**: 0.20, 0.35, 0.42; **G**: 3, 4, 3; **H**: 46.0, 46.6, 49.9; **I**: 41.5, 35.1, 24.0; **J**: 41.9, 40.3, 36.3. Outliers are not displayed in the box plots.

Figure S12

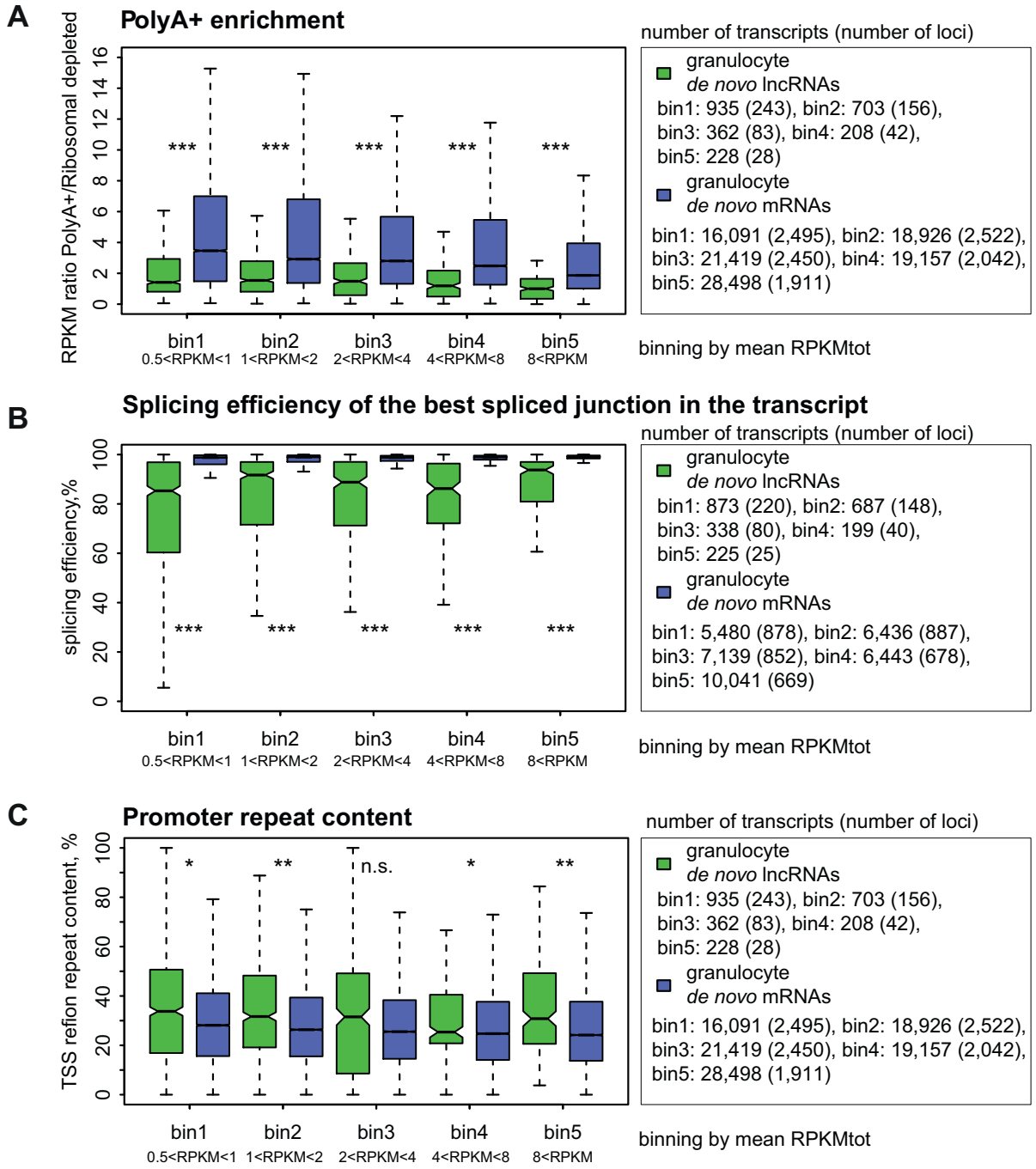


Figure S12. Difference between mRNAs and lncRNAs is persistent independently of expression level

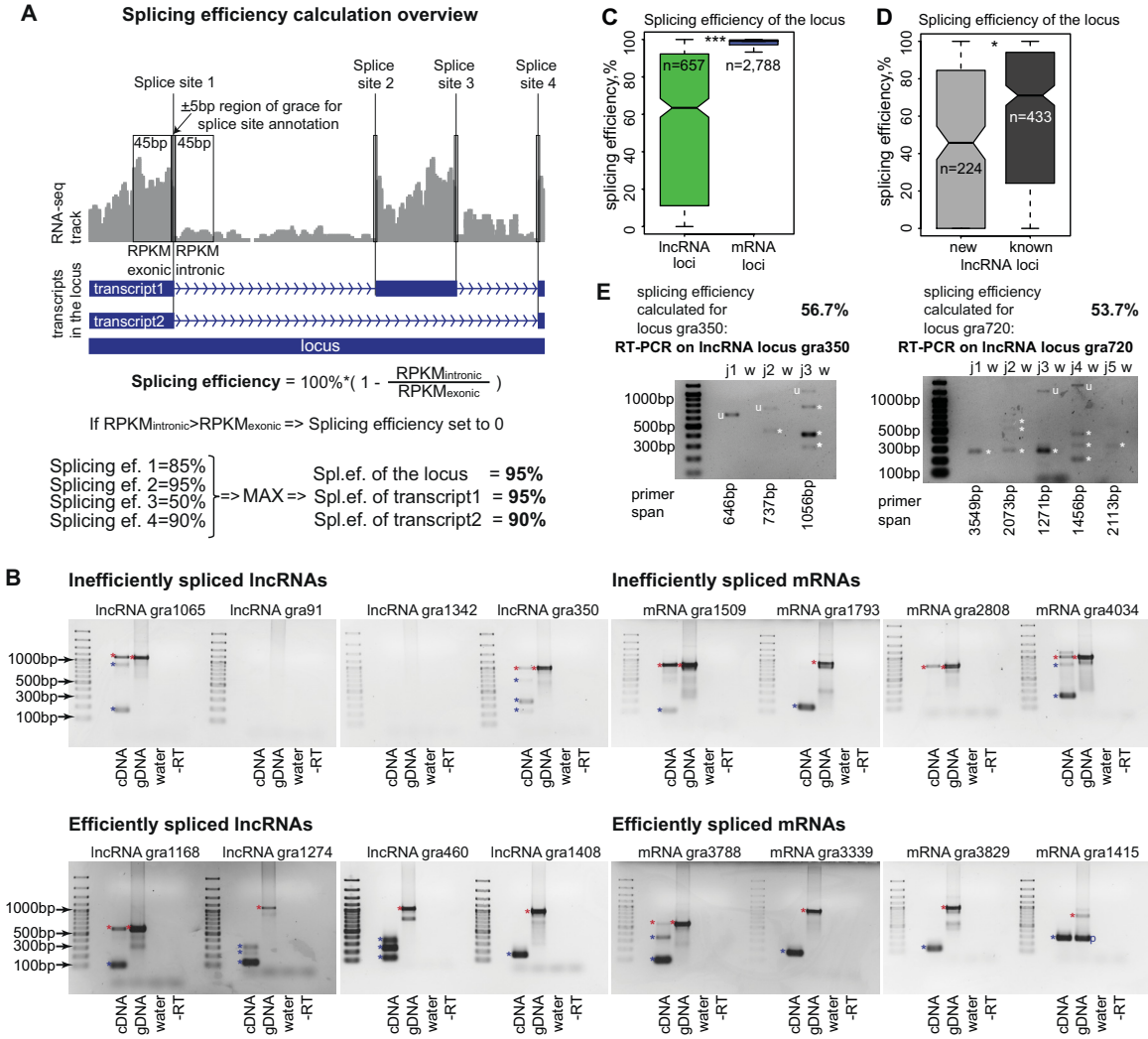
A. PolyA+ enrichment (calculated as in Figure 2E) of granulocyte *de novo* lncRNA and mRNA transcripts split into expression bins. Median values in boxes from left to right: 1.4, 3.5, 1.5, 2.9, 1.5, 2.8, 1.2, 2.5, 1.0, 1.9.

B. Splicing efficiency (calculated as in Figure 2F) of granulocyte *de novo* lncRNA and mRNA transcripts split into expression bins. Median values in boxes from left to right: 85.2%, 98.8%, 91.7%, 99.0%, 88.8%, 98.8%, 86.2%, 99.0%, 93.8%, 99.2%.

C. Percent repeat coverage of TSS region (TSS +/- 1.5 kb) of granulocyte *de novo* lncRNAs and mRNA transcripts split into expression bins. Median values in boxes from left to right: 33.8%, 28.1%, 31.7%, 26.3%, 31.5%, 25.5%, 25.4%, 24.7%, 30.8%, 24.2%.

Remarks to boxplots: The numbers of transcripts in each box are indicated on the right. Number in brackets indicates number of loci the transcripts in each box initiate from. *** - $p<10^{-10}$, ** - $p<10^{-5}$, * - $p<0.01$, n.s. – $p>0.01$. The box plots display the full population but p-values are calculated using Mann–Whitney U test with equalized sample size (Methods). Outliers are not displayed in the box plots. “mean RPKMtot” – average transcript RPKM in 21 ribosomal depleted granulocyte RNA-seq samples.

Figure S13



spliced	type	locus	Forward primer seq 5'-3'	Reverse primer seq 5'-3'	unspliced product size
inefficiently	lncRNA	GRA_denovo_LNC.1065	TTCTCATGGGAGCAGCTGTA	AGGGGGAGATGTCACCTTTT	1098bp
inefficiently	lncRNA	GRA_denovo_LNC.91	TTCCGTGATGTTTGCATTTC	AAAGGCAAACTGCTGAATCC	1014bp
inefficiently	lncRNA	GRA_denovo_LNC.1342	CAGTGTTTCCCATTTCCTGA	TTGCAATACATTCCACTCTGTG	883bp
inefficiently	lncRNA	GRA_denovo_LNC.350	CCAACCGAGGATGTAAGAGG	TGGAAGATTCAACTGGCAGA	1002bp
inefficiently	mRNA	GRA_denovo_PC.1509	ACGGCCTGGATTCAGAGACT	ACAGACACCCTTTGGACCAC	821bp
inefficiently	mRNA	GRA_denovo_PC.1793	GTCCTCGAAAGTCACCATCC	AGCAGTTCCTCACCATCCTG	944bp
inefficiently	mRNA	GRA_denovo_PC.2808	CTCTGCAGCAAGAAATGAGG	CTCTTTCTTTGAGGGAACACC	767bp
inefficiently	mRNA	GRA_denovo_PC.4034	CAGCCTACCTGTCCATGTCC	ACGGTAATCCAGCGAGGACT	1024bp
efficiently	lncRNA	GRA_denovo_LNC.1168	TCAAAAGCCGTTGAGACACA	GGAGGCATCCTCTGTCAAGA	572bp
efficiently	lncRNA	GRA_denovo_LNC.1274	CGCAGGGAAACAGTTCTCAT	TCCAGCTGACTTGCATTCTG	1099bp
efficiently	lncRNA	GRA_denovo_LNC.460	AACTCAAGCAGAAGAGCAGTTG	GCACCGGAGACTGAAAGAAG	1043bp
efficiently	lncRNA	GRA_denovo_LNC.1408	GAGGCTGAGCAAGGGACA	TCTCAGAAATCTGTGAGCTGTG	993bp
efficiently	mRNA	GRA_denovo_PC.3788	CTCGCCAAAAACTTCACCTT	AGGAGAGCAAGACACACTTCC	562bp
efficiently	mRNA	GRA_denovo_PC.3339	GATACTGCTGCTTGCCTTCC	GAAGGATTGACACCAACCTCA	902bp
efficiently	mRNA	GRA_denovo_PC.3829	ACTGCTCCAAGCACCTCCT	TAGACCCAAGGCTGTCACAC	973bp
efficiently	mRNA	GRA_denovo_PC.1415	CCTCTTCGGGGAGATGTAGC	TCCACTGCCTTACGTTTTGG	772bp

Figure S13. Splicing efficiency calculation

A. Overview of splicing efficiency calculation (Methods). Shown is an illustrative RNA-seq signal track of a genomic locus. Splicing efficiency of each splice site in the locus is calculated based on the ratio between intronic and exonic RNA-seq signal at this splice site. RNA-seq signal is calculated as RPKM of 45bp exonic and intronic regions surrounding the splice site. 45bp are positioned 5bp away from the precise splice site position to accommodate for potential imprecise splice site annotation. Splicing efficiency is calculated with the given formula. In case of intronic signal exceeding exonic

signal, splicing efficiency is set to 0. Maximal splice site splicing efficiency is taken to estimate splicing efficiency of each transcript, as well as of the whole locus.

B. RT-PCRs to test splicing efficiency calculation (see Supplemental Methods). RT-PCR result for one junction (junction length<1500bp) of eight inefficiently spliced (upper row gel pictures, see Supplemental Methods) and efficiently spliced (lower row gel pictures, see Supplemental Methods) *de novo* granulocyte transcripts (4xlncRNAss and 4xmRNAs). List of assayed lncRNA and mRNA transcripts and primers to amplify their short junction is given in the table below. Size of the expected PCR product amplifying the unspliced junction is given in the right-most column of the table. Each primer pair was used with granulocyte cDNA sample and three control samples: genomic DNA (from the same granulocyte sample) - to test efficiency of the primers when amplifying the long unspliced isoform, water - to test PCR reaction contamination, -RT (no reverse transcriptase added when preparing cDNA) control – to test for genomic DNA contamination in the cDNA sample. Red stars on the left from the band indicate the expected band corresponding to the unspliced product. Blue stars on the left from the band indicate spliced products. lncRNA gra91 and lncRNA gra1342 – absence of a band in gDNA indicates inefficiency of primers. mRNA gra1415 – band marked "p" might indicate the presence of a pseudogene corresponding to the assayed protein-coding gene somewhere in the genome. Overall, the RT-PCR test validates our splicing efficiency calculation with transcripts identified as unspliced showing an abundant unspliced isoform (5 out of 6 with one (mRNA gra1793) exception) and transcripts identified as spliced showing no unspliced isoform (6 out of 8 with two exceptions (lncRNA gra1168 and mRNA mRNA gra3788) showing some unspliced product signal, however it is much fainter than the band formed by the spliced products). Note that, as described in (**A**), we assign a transcript with the splicing efficiency of its best spliced splice site, whereas the RT-PCR assay allows to assess splicing efficiency of only the short (<1.5kb) splice sites, which might explain the slight discrepancy in the results presented in (**B**).

C. Analysis of splicing efficiency of the whole locus confirms reduced splicing efficiency of granulocyte de novo lncRNAs compared to mRNAs. Boxplot shows splicing efficiency (as described in A, Methods) of *de novo* lncRNA (green) and mRNA (blue) loci annotated in granulocytes. Splicing efficiency of each splice site was calculated (Methods) and the efficiency of the most efficiently spliced transcript (i.e. most efficiently spliced site of all the transcripts) in each locus is plotted. Median values: lncRNAs: 63.46%, mRNAs: 99.07%.

D. New lncRNA loci are less efficiently spliced than known lncRNA loci. Boxplot shows splicing efficiency (as described in A, Methods) of new (light grey) and known (dark grey) (as described in Fig. 2A) granulocyte *de novo* lncRNA loci. Splicing efficiency calculated as in (**C**). Median values: new loci: 45.75%, known loci: 71.06%.

E. Two illustrative examples of exon spanning RT-PCR amplifying a continuous unspliced isoform. Gel electrophoresis of RT-PCR over splice junctions in granulocyte *de novo* lncRNA loci 350 and 720 (See Additional File 2E for primer sequence and junction genomic position and Additional File 3 for locus annotation). Above the gel picture, splicing efficiency calculated from RNA-seq data as described in (**A**) is shown. **j**: junction number, **w**: PCR water control. *****: bands corresponding to spliced products, **u**: bands corresponding to the unspliced product. Primer span: the genomic span of PCR primers corresponding to the length of the unspliced PCR product.

Remarks to boxplots: Numbers inside boxes indicate number of loci displayed in each box. The box plots display the full population but p-values are calculated using Mann–Whitney U test with equalized sample size (Methods). *- $p<0.01$, *** - $p<10^{-16}$. Outliers are not displayed.

Figure S14

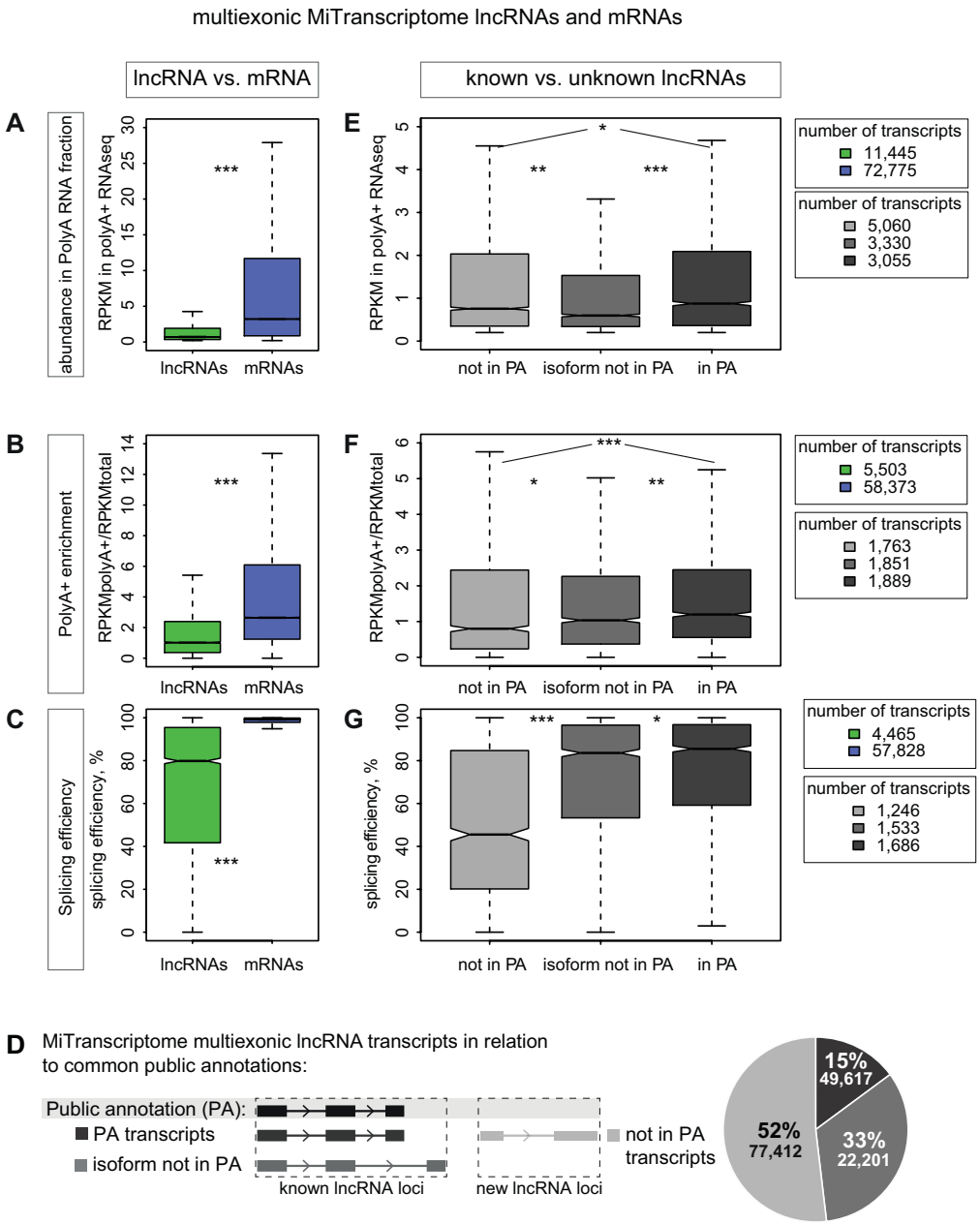


Figure S14. Analyzing features of MiTranscriptome lncRNAs and mRNAs confirms that difficult to identify lncRNAs are more different from mRNAs than publicly annotated lncRNAs

A. Average expression level (RPKM) of multi-exonic MiTranscriptome lncRNA (green) and mRNA (blue) transcripts in granulocyte PolyA+ RNA-seq samples produced in the study. Only transcripts with detectable expression are plotted (average RPKM>0.2).

B. PolyA+ enrichment of multi-exonic MiTranscriptome lncRNA (green) and mRNA (blue) transcripts as described in Fig. 2E. Only transcripts detected in total granulocyte RNA-seq data (average RPKM among 21 samples >0.2) are analyzed.

C. Splicing efficiency of multi-exonic MiTranscriptome lncRNA (green) and mRNA (blue) transcripts expressed in granulocytes. Splicing efficiency was calculated using ribosomal-depleted

RNA-seq from 7 donors (time points pooled to increase the coverage) (Methods). The splicing efficiency of the most efficiently spliced site in each transcript is plotted.

D. Distribution of multi-exonic MiTranscriptome lncRNA transcripts according to their coverage in the 3 commonly used public annotations as described in Fig. 2A for granulocyte *de novo* lncRNAs.

E. Expression level of the 3 types of multi-exonic MiTranscriptome lncRNA transcripts in granulocyte PolyA+ RNA-seq. Publicly annotated transcripts show the highest expression level.

F. PolyA+ enrichment of the 3 types of multi-exonic MiTranscriptome lncRNA transcripts.

G. Splicing efficiency of 3 types of multi-exonic MiTranscriptome lncRNA transcripts described in (**D**). The splicing efficiency of the most efficiently spliced site in each transcript is plotted.

Remarks to boxplots A, B, C, E, F and G: Numbers on the right indicate the numbers of transcripts analyzed in each boxplot. The box plots display the full population but p-values are calculated using Mann–Whitney U test with equalized sample size (Methods). *** - $p<10^{-10}$, ** - $p<10^{-5}$, * - $p<0.01$, n.s. – $p>0.01$. Median values from left to right: **A**: 0.71, 3.21; **B**: 1.02, 2.64; **C**: 79.9, 99.2; **E**: 0.76, 0.60, 0.87; **F**: 0.80, 1.04, 1.20; **G**: 45.5, 83.6, 85.5. Outliers are not displayed in the box plots. Numbers on the right represent the number of transcripts in each box.

Figure S15

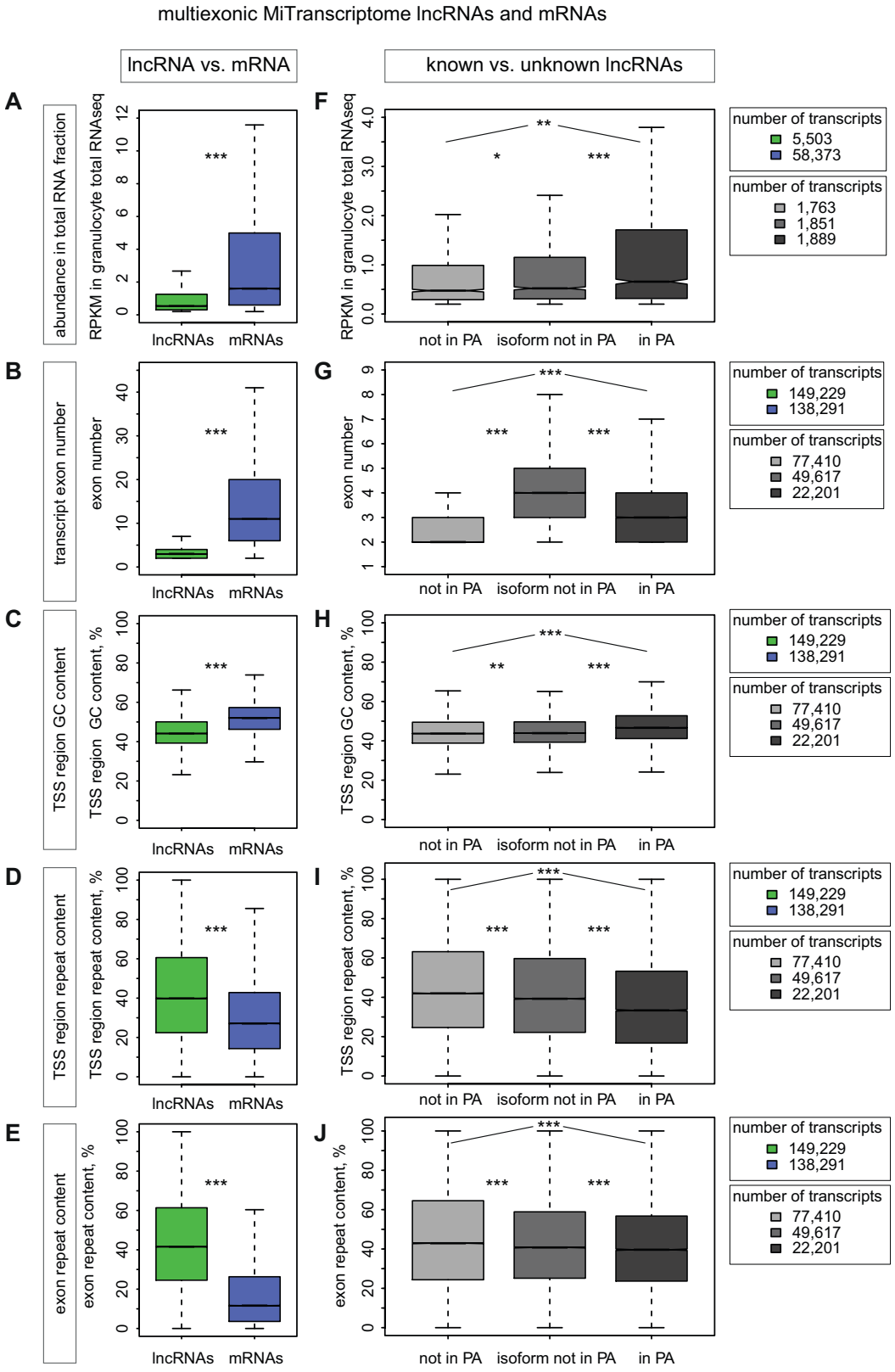


Figure S15. Additional features that distinguish MiTranscriptome lncRNAs from mRNAs, and newly identified from publicly annotated lncRNAs

A. Abundance in total RNA fraction of multi-exonic MiTranscriptome lncRNA and mRNA transcripts was calculated as average from all 21 available Ribosomal Depleted RNA-seq samples.

Only MiTranscriptome transcripts expressed in granulocyte total RNA-seq dataset (average RPKM>0.2) are plotted. Exon number (**B**), percent GC content of TSS region (TSS +/- 1.5 kb) (**C**), percent repeat coverage of TSS region (**D**) and exons (**E**) of multi-exonic MiTranscriptome lncRNAs and mRNA transcripts. Abundance in total RNA fraction in granulocytes (**F**), exon number (**G**), percent GC content of TSS region (**H**), percent repeat coverage of TSS region (**I**) and exons (**J**) of the three classes of MiTranscriptome lncRNA as described in Figure S14D. Remarks: green - multi-exonic MiTranscriptome lncRNAs, blue – multi-exonic MiTranscriptome mRNAs, light gray – "not in PA" lncRNA transcripts, medium gray – "isoform not in PA" lncRNA transcripts, dark gray – "PA" lncRNA transcripts. Numbers of transcripts in each box of the boxplots are indicated on the right. *** - $p<10^{-10}$, ** - $p<10^{-5}$, * - $p<0.01$, n.s. – $p>0.01$. The box plots display the full population but p-values are calculated using Mann–Whitney U test with equalized sample size (Methods). Median values from left to right: **A**: 0.54, 1.60; **B**: 3, 11; **C**: 44.13, 52.00; **D**: 39.87, 27.10; **E**: 41.55, 11.63; **F**: 0.48, 0.52, 0.66; **G**: 2, 4, 3; **H**: 43.70, 43.90, 46.63; **I**: 42.00, 39.27, 33.37; **J**: 42.91, 40.77, 39.63. Outliers are not displayed in the box plots.

Figure S16

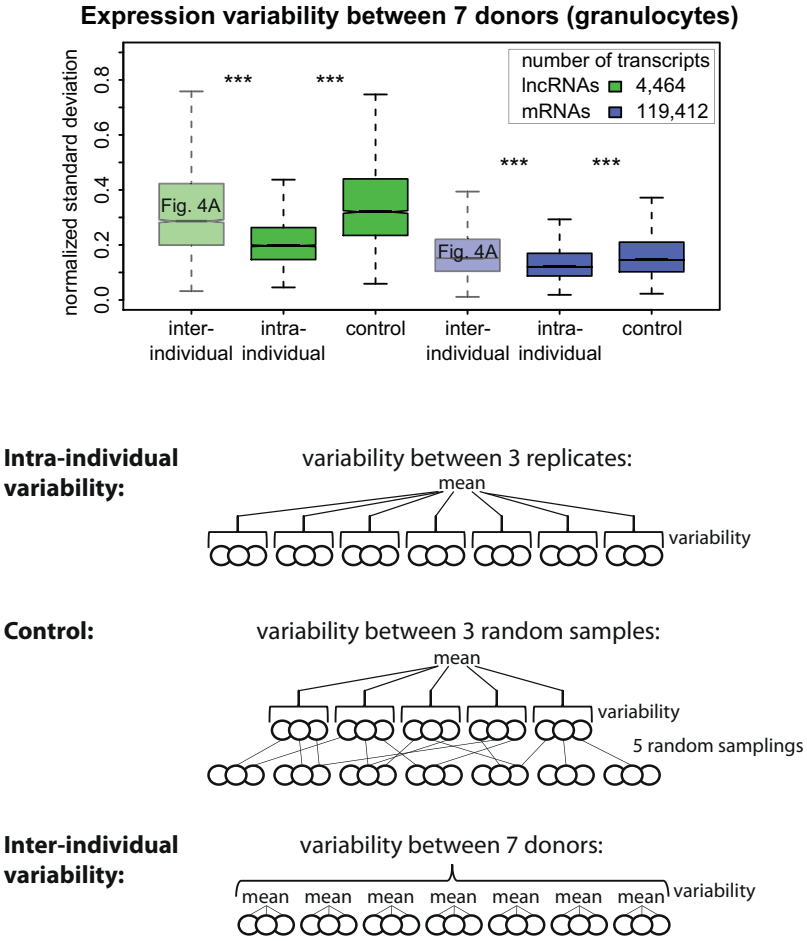


Figure S16. Intra-individual variability is significantly lower than inter-individual variability for both lncRNAs and mRNAs in granulocytes

The boxplot shows values of different types of expression variability of granulocyte *de novo* annotated lncRNA (green) and mRNA (blue) transcripts in granulocyte total RNA-seq dataset. We calculated intra-individual variability (between 3 replicates of each donor), compared it to inter-individual variability (between 7 donors, data displayed in Fig. 4A, here indicated as transparent boxes). We then controlled for the reduced sample size in intra-individual variability calculation (3 samples for intra-individual vs. 7 samples for inter-individual variability) by randomly sampling 3 replicates and asking if the inter-individual variability observed is reduced with the reduced sample size. Graphical representation of calculation of the three variability types plotted is displayed below the boxplot. From top to bottom: Intra-individual variability (variability "between 3 replicates"): standard deviation was calculated for each donor by calculating standard deviation of 3 replicates and normalized by the mean of the 3 replicates, 7 normalized standard deviation values from 7 donors were then averaged to give intra-individual variability for each transcript. Control (variability between 3 random replicates): 3 replicates were randomly picked from the 21 (7 donors x 3 replicates) samples, normalized standard deviation was calculated for the 3 samples, the random sampling was performed 5 times and the average of 5 normalized standard deviations was calculated. Inter-individual variability (variability between 7 donors): calculated as described for Fig. 4A (Results).
Remarks: Chr. X and Y were discarded from the analysis. *** - $p<10^{-16}$. P-values are calculated using Mann–Whitney U test. Numbers in the boxplot legend indicate number of transcripts analyzed. Median values of boxes left to right: lncRNAs: 0.29, 0.20, 0.32, mRNAs: 0.15, 0.12, 0.15. Outliers are not displayed in the box plot.

Figure S17

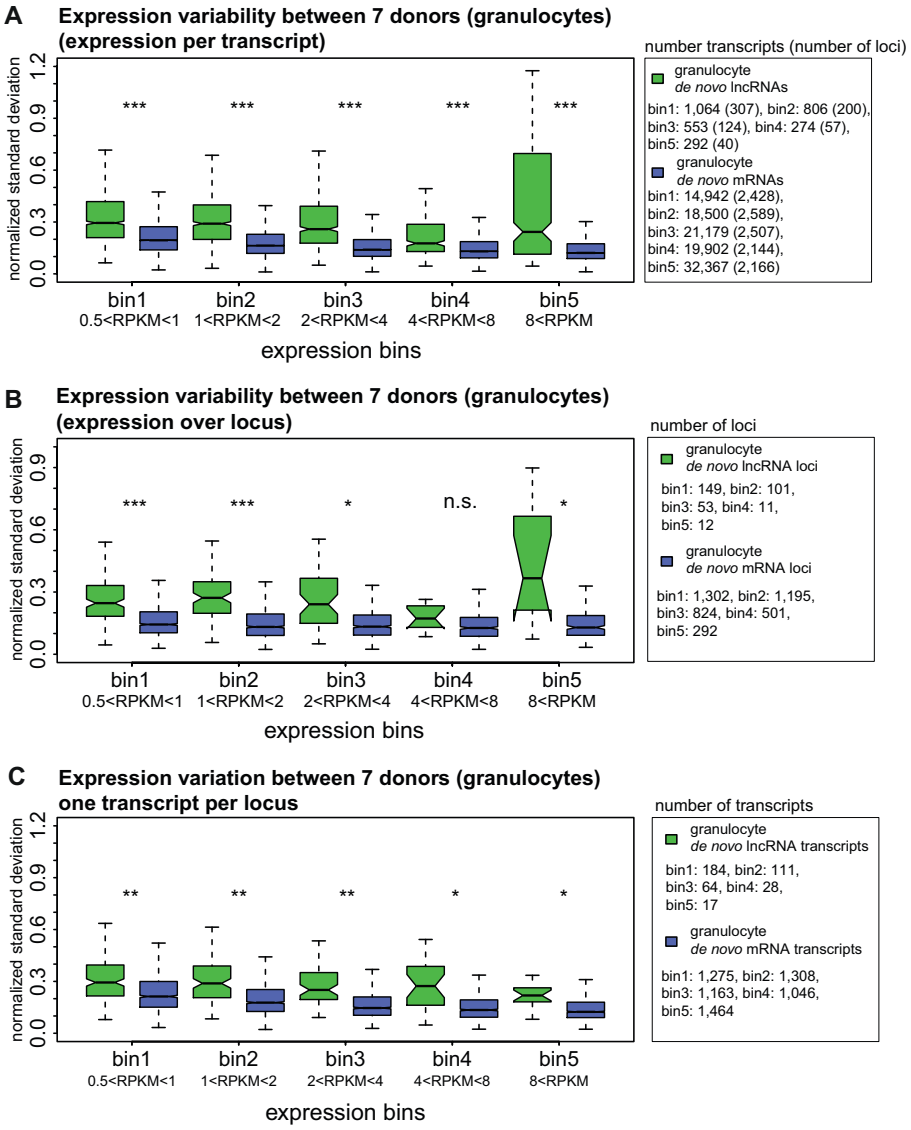


Figure S17. *De novo* lncRNAs are more variable than mRNAs independently of generally lower lncRNA expression level

The phenomenon of increased expression variability of lncRNAs compared to mRNAs is not biased to the difference in absolute expression level between lncRNAs and mRNAs.

A. and **B.** Normalized standard deviation of *de novo* granulocyte lncRNA (green) and mRNA (blue) transcripts **(A)** or loci **(B)** expression between granulocytes from 7 donors, split into 5 expression bins according to their maximal expression level (RPKM) among 7 donors. Median values from left to right: **A**: 0.29, 0.19; 0.29, 0.16; 0.26, 0.14; 0.18, 0.13; 0.24, 0.12; **B**: 0.25, 0.14; 0.27, 0.13; 0.24, 0.13; 0.17, 0.13; 0.37, 0.13. **C.** Normalized standard deviation of *de novo* granulocyte lncRNA (green) and mRNA (blue) transcripts – one transcript per locus was picked for the analysis (using *!duplicated* function in R on locus name). Median values from left to right: 0.29, 0.21, 0.29, 0.18, 0.25, 0.15, 0.27, 0.13, 0.22, 0.12.

Remarks to the boxplots: *** - $p<10^{-10}$, ** - $p<10^{-5}$, * - $p<0.01$, n.s. – $p>0.01$. The box plots display the full population but p-values are calculated using Mann–Whitney U test with equalized sample size (Methods). Numbers on the right show number of transcripts/loci in each box of the boxplots. Number in brackets (in **A**) indicate the number of loci the transcripts in each box initiate from. Outliers are not displayed in the box plots. Chr. X and Y were discarded from the variability analysis.

Figure S18

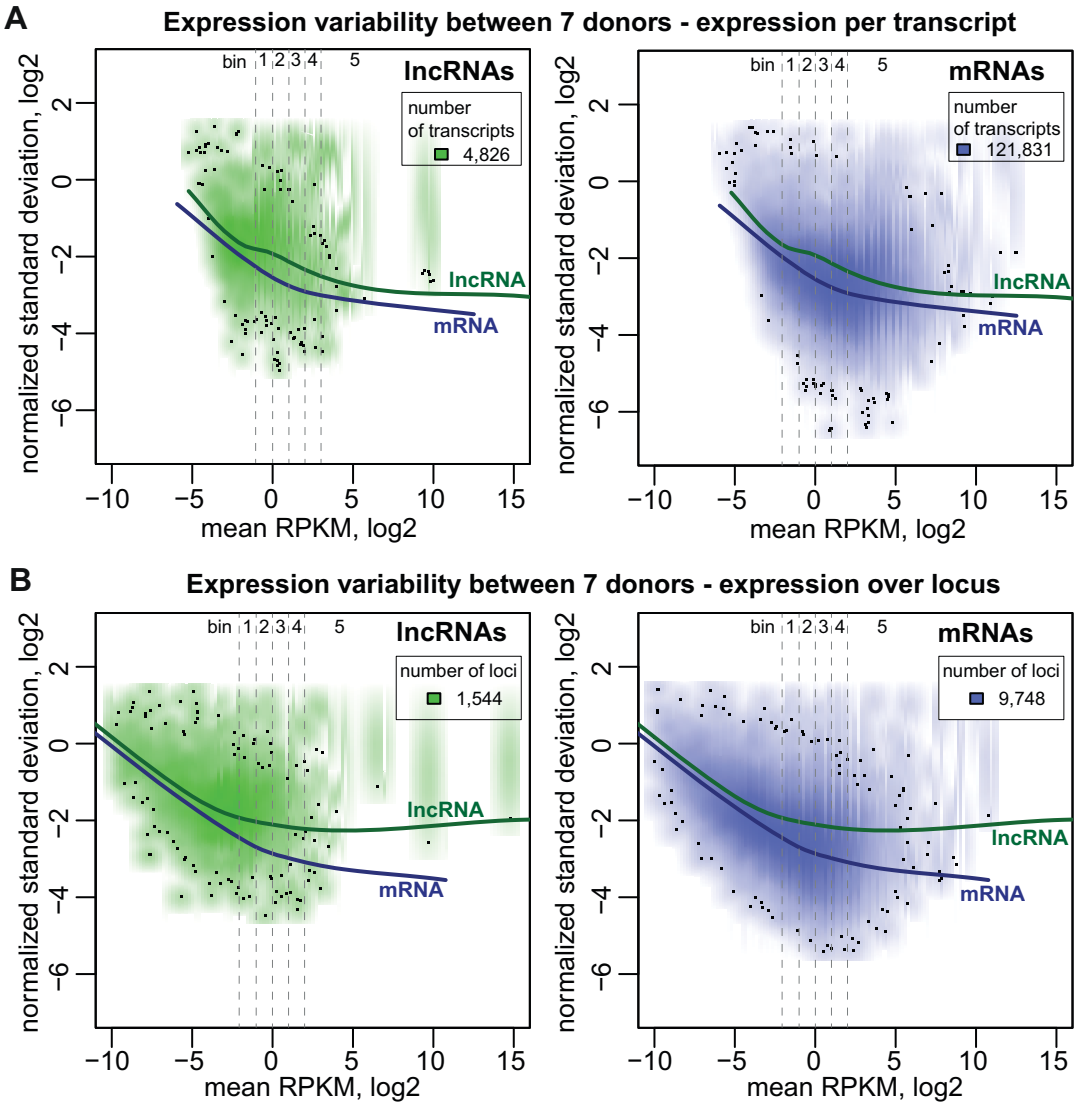


Figure S18. Plotting variability against mean RPKM level confirms increased lncRNA expression variability compared to mRNAs independent of expression level.

A. and **B.** Normalized standard deviation of *de novo* granulocyte lncRNA (green) and mRNA (blue) transcripts (**A**) and loci (**B**) expression between granulocytes from 7 donors plotted against mean expression level (RPKM) among the 7 donors. Scatter plots were built using *smoothScatter* function in R. Fitted curves were built using *loess.smooth* function in R. Both lncRNA (dark green) and mRNA (dark blue) fitted curves are displayed on each scatter plot for facilitating comparison. Dashed lines indicate the expression bins used in Figure S17. Logged (log2) values are plotted. Chr. X and Y were discarded from the variability analysis.

Figure S19

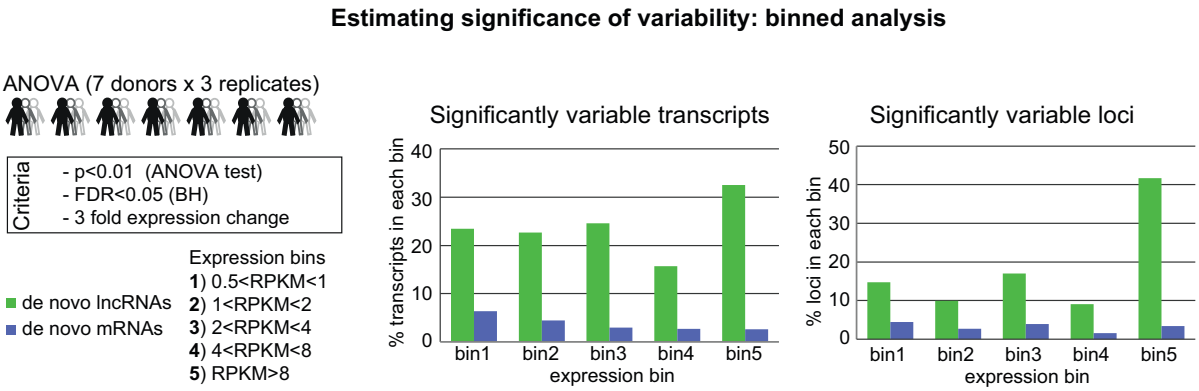


Figure S19. Percentage of *de novo* granulocyte lncRNA transcripts/loci significantly variable among seven donors is higher than that of mRNAs in all expression bins

Significance of expression variability by ANOVA test of expression variability of lncRNA (green) and mRNA (blue) transcripts (middle bar plot), and loci (right bar plot) in granulocytes from 7 donors (the 3 time points are used as replicates). Criteria for calling a transcript/locus "significantly variable": ANOVA test p value <0.01, FDR (Benjamini-Hochberg correction) <0.05, fold change between highest and lowest expression in 7 donors >3. Bar plots show percentage of significantly variable transcripts within each expression bin. LncRNA/mRNA transcripts per bin: bin1:1,064/14,942, bin2:806/18,500, bin3:553/21,179, bin4:274/19,902, bin5:292/32,367. LncRNA/mRNA locus per bin: bin1:149/1,302, bin2:101/1,195, bin3:53/824, bin4:11/501, bin5:12;/292 Chromosomes X and Y were discarded.

Figure S20

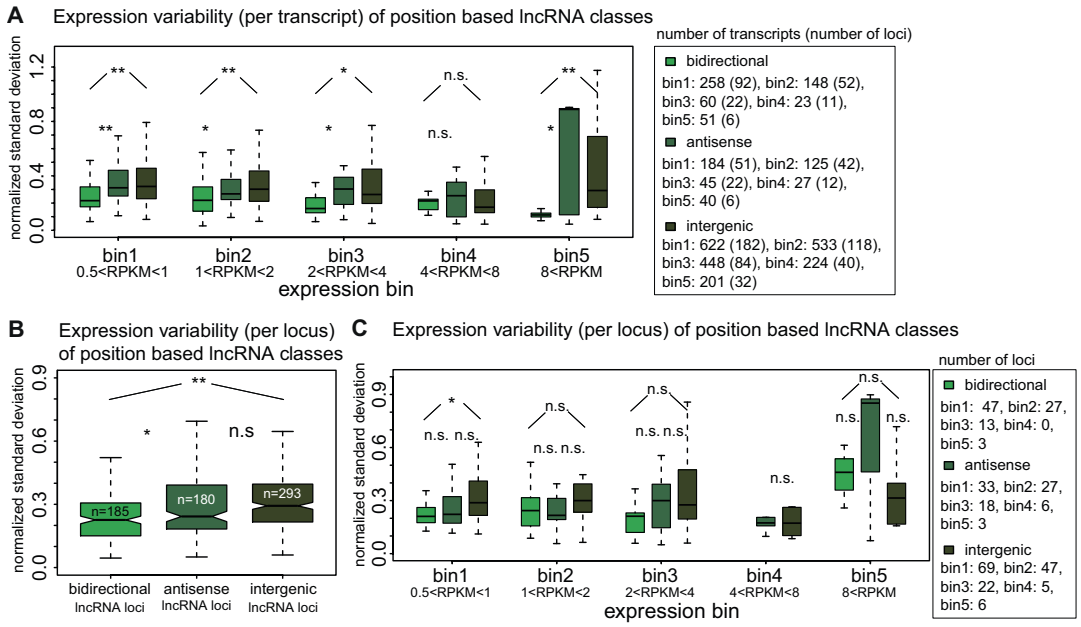


Figure S20. Bidirectional lncRNAs show reduced variability: controls

The boxplots show expression variability (normalized standard deviation) of 3 position-based classes of *de novo* annotated granulocyte lncRNAs (shades of green) between granulocyte samples from 7 donors.

A. Reduced variability of bidirectional lncRNA transcripts is persistent in all expression bins. Shown is expression variability for bidirectional, antisense and intergenic lncRNA transcripts. Median values from left to right: 0.22, 0.31, 0.32; 0.22, 0.27, 0.30; 0.16, 0.30, 0.26; 0.21, 0.25, 0.17; 0.11, 0.89, 0.29. Numbers on the right show number of transcripts in each box of the boxplot. Number in brackets indicate the number of loci the transcripts in each box initiate from. The lack of significance in bin 4 most likely arises from the low number of bidirectional and antisense transcripts in this bin.

B. Expression variability analysis over loci confirms reduced variability of bidirectional lncRNAs compared to antisense and intergenic lncRNAs. Shown is expression variability of bidirectional, antisense and intergenic lncRNA loci (shades of green). Median values from left to right: 0.23, 0.24, 0.29. The difference between antisense and bidirectional lncRNA variability is reduced compared to analysis per transcript (Fig. 4F) likely because of the bias of antisense lncRNA locus expression calculation to the bias of its highly expressed antisense protein-coding pair. This bias is not present when calculating transcript expression over exons.

C. Binned analysis of expression variability per locus does not give a meaningful confirmation for (**B**) caused by very low numbers of loci in each bin. Shown is expression variability of variability of bidirectional, antisense and intergenic lncRNA loci (shades of green). Median values from left to right: 0.21, 0.22, 0.29, 0.24, 0.22, 0.30, 0.21, 0.30, 0.28, -, 0.17, 0.17, 0.46, 0.85. 0.31. Numbers on the right show number of loci in each box of the boxplot. Absent bidirectional box in bin 4 means there were no bidirectional loci in this expression bin.

Remarks: Transcripts/loci were split into 5 bins according to their maximal expression of among 7 donors. chr X and Y were discarded from the analysis. *** - $p<10^{-10}$, ** - $p<10^{-5}$, * - $p<0.01$, n.s. – $p>0.01$. The box plots display the full population but p-values are calculated using Mann–Whitney U test with equalized sample size (Methods). Outliers are not displayed in the box plot.

Figure S21

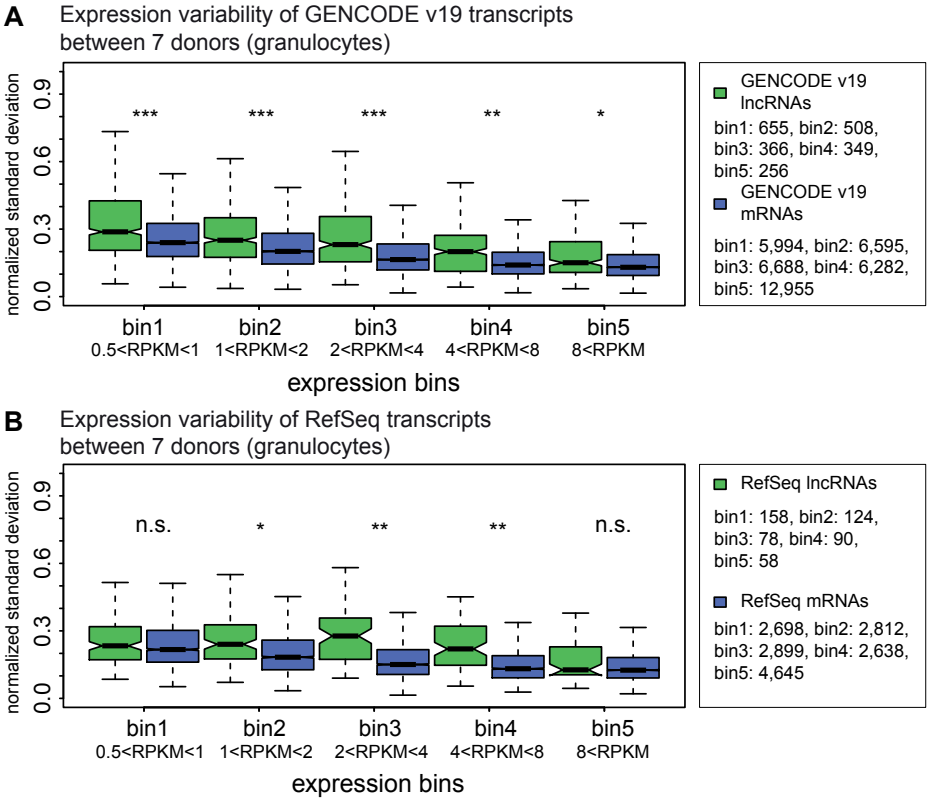


Figure S21. Increased expression variability is observed for RefSeq and GENCODE lncRNAs, however to a lesser extent

Normalized standard deviation of expression level of multi-exonic GENCODE v19 (**A**) and RefSeq (**B**) lncRNA (green) and mRNA (blue) transcripts in granulocytes between 7 donors. Transcripts are split into 5 expression bins according to their maximal expression among 7 donors. Remarks: Chr. X and Y were discarded from the analysis. *** - $p<10^{-10}$, ** - $p<10^{-5}$, * - $p<0.01$, n.s. – $p>0.01$. The box plots display the full population but p-values are calculated using Mann–Whitney U test with equalized sample size (Methods). Numbers on the right show number of transcripts in each box of the boxplot. Median values from left to right: **A**: 0.29, 0.24, 0.25, 0.20, 0.23, 0.16, 0.20, 0.14, 0.15, 0.13; **B**: 0.23, 0.22, 0.24, 0.18, 0.28, 0.15, 0.22, 0.13, 0.13, 0.13. Outliers are not displayed in the box plots.

Figure S22

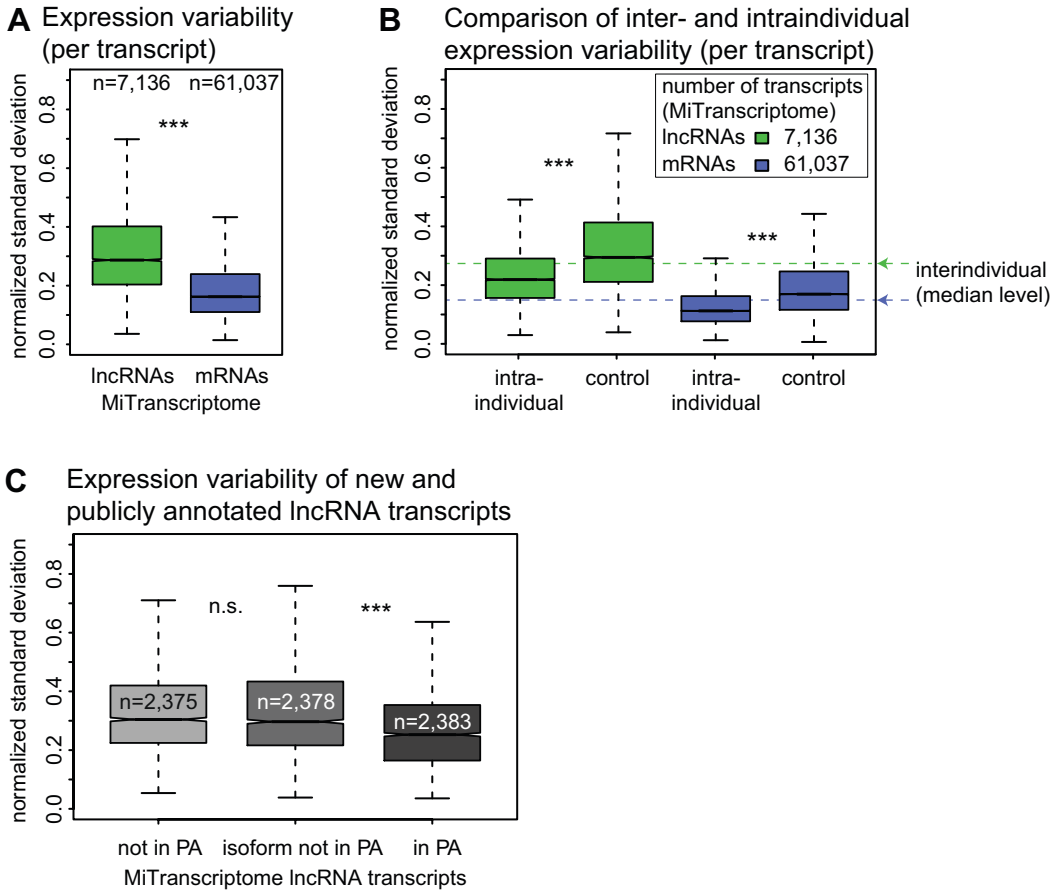


Figure S22. MiTranscriptome analysis confirms increased lncRNA expression variation

A. Genome wide inter-individual variability of multi-exonic MiTranscriptome lncRNA (green) and mRNA (blue) transcripts analyzed in the granulocyte RNA-seq data set obtained in the study. Inter-individual variability is estimated by calculating normalized (by mean) standard deviation between expression of each transcript in granulocytes from 7 donors. Expression level in each donor is averaged from three replicates. Numbers above boxes indicate number of transcripts analyzed.
B. The boxplot shows expression variability (normalized standard deviation) of MiTranscriptome lncRNA (green) and mRNA (blue) transcripts (as done for *de novo* granulocyte lncRNAs and mRNAs in Figure S16). The level of inter-individual variability between 7 donors (**A**) is indicated with green and blue dashed lines. The boxes show intra-individual variability between 3 replicates and 3 random replicates as described for Figure S16. Numbers in the legend indicate number of transcripts analyzed.
C. Novel MiTranscriptome lncRNA transcripts are more variable in our granulocyte dataset than MiTranscriptome lncRNA transcripts already present in public annotations. The boxplot shows inter-individual variability of 3 classes of multi-exonic MiTranscriptome lncRNA transcripts split according to their coverage by public annotations (Figure S14D). Numbers inside boxes indicate number of transcripts analyzed.
Remarks to boxplots: Only multi-exonic MiTranscriptome lncRNA and mRNA transcripts were analyzed. Transcripts not expressed (RPKM<0.2) in any of the 7 donors (total RNA-seq data) and transcripts from chromosomes X and Y were discarded from the analysis. The box plots display the full population but p-values are calculated using Mann–Whitney U test with equalized sample size (Methods). *** - $p<10^{-10}$, ** - $p<10^{-5}$, * - $p<0.01$, n.s. – $p>0.01$. Median values from left to right: **A**: 0.29, 0.16; **B**: 0.23, 0.31, 0.13, 018; **C**: 0.30, 0.30, 0.25. Outliers are not displayed in the boxplots.

Figure S23

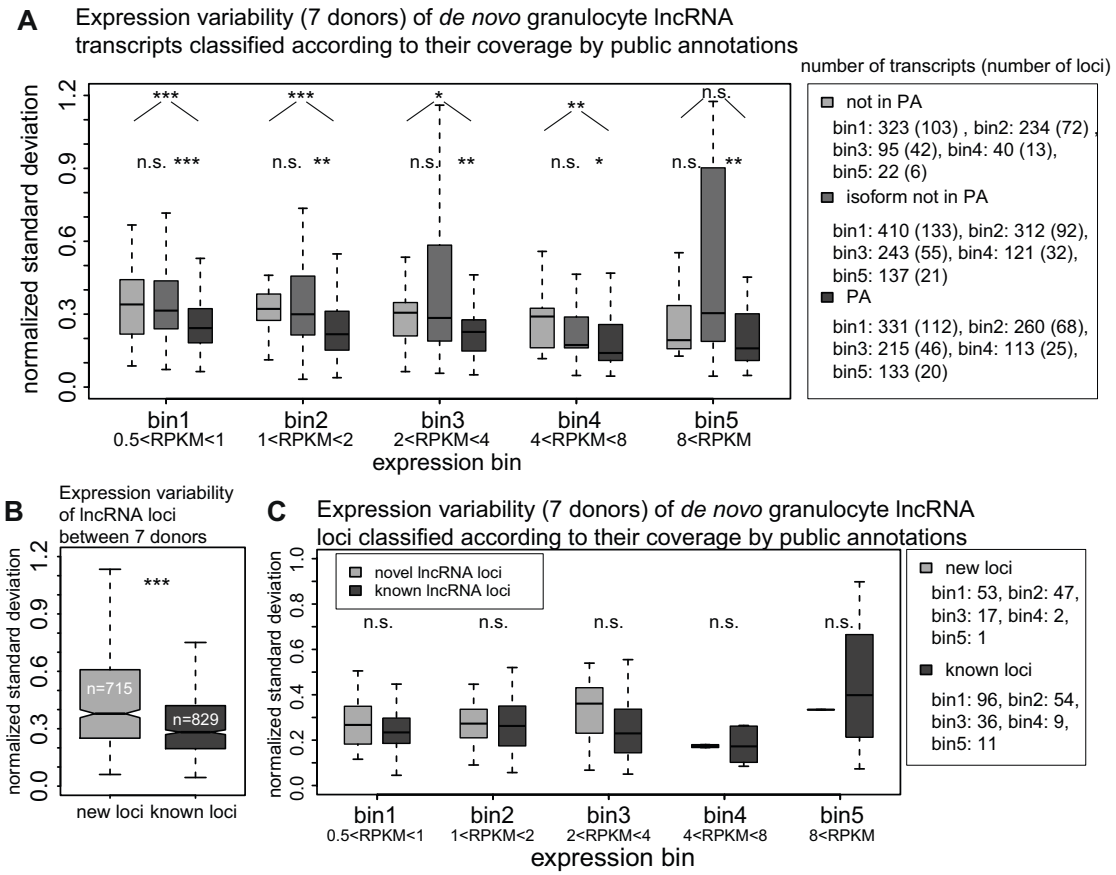


Figure S23. Granulocyte lncRNA transcripts not present public annotations (PA) show increased variability: controls

The boxplots show expression variability (normalized standard deviation) between 7 donors in granulocytes of *de novo* granulocyte lncRNA transcripts/loci classified according to their presence in public annotations.

A. Increased variability of "not in PA" and "isoform not in PA" lncRNA transcripts is persistent in all expression bins. Shown is expression variability for "not in PA" transcripts (light gray), "isoform not in PA" transcripts (medium gray) and "PA" transcripts (dark gray). Median values from left to right: 0.34, 0.31, 0.24; 0.32, 0.30, 0.22; 0.31, 0.28, 0.23; 0.29, 0.17, 0.14; 0.19, 0.30, 0.16. Numbers on the right show number of transcripts in each box of the boxplot. Number in brackets indicate the number of loci the transcripts in each box initiate from.

B. Expression variability analysis over loci confirms increased variability of lncRNAs not covered by public annotations. Shown is expression variability of "new" (light gray) and "known" loci. Median values from left to right: 0.38, 0.28.

C. Binned analysis of expression variability per locus does not give a meaningful confirmation for (**B**) caused by very low numbers of loci in each bin. Shown is expression variability of "new" (light gray) and "known" (dark grey) loci. Median values from left to right: 0.27, 0.23, 0.27, 0.26, 0.36, 0.23, 0.17, 0.17, 0.33, 0.40. Numbers on the right show number of loci in each box of the boxplot.

Remarks: Transcripts/loci were split into 5 bins according to their maximal expression of among 7 donors. chr X and Y were discarded from the analysis. *** - $p<10^{-10}$, ** - $p<10^{-5}$, * - $p<0.01$, n.s. – $p>0.01$. The box plots display the full population but p-values are calculated using Mann–Whitney U test with equalized sample size (Methods) Outliers are not displayed in the box plot.

Figure S24

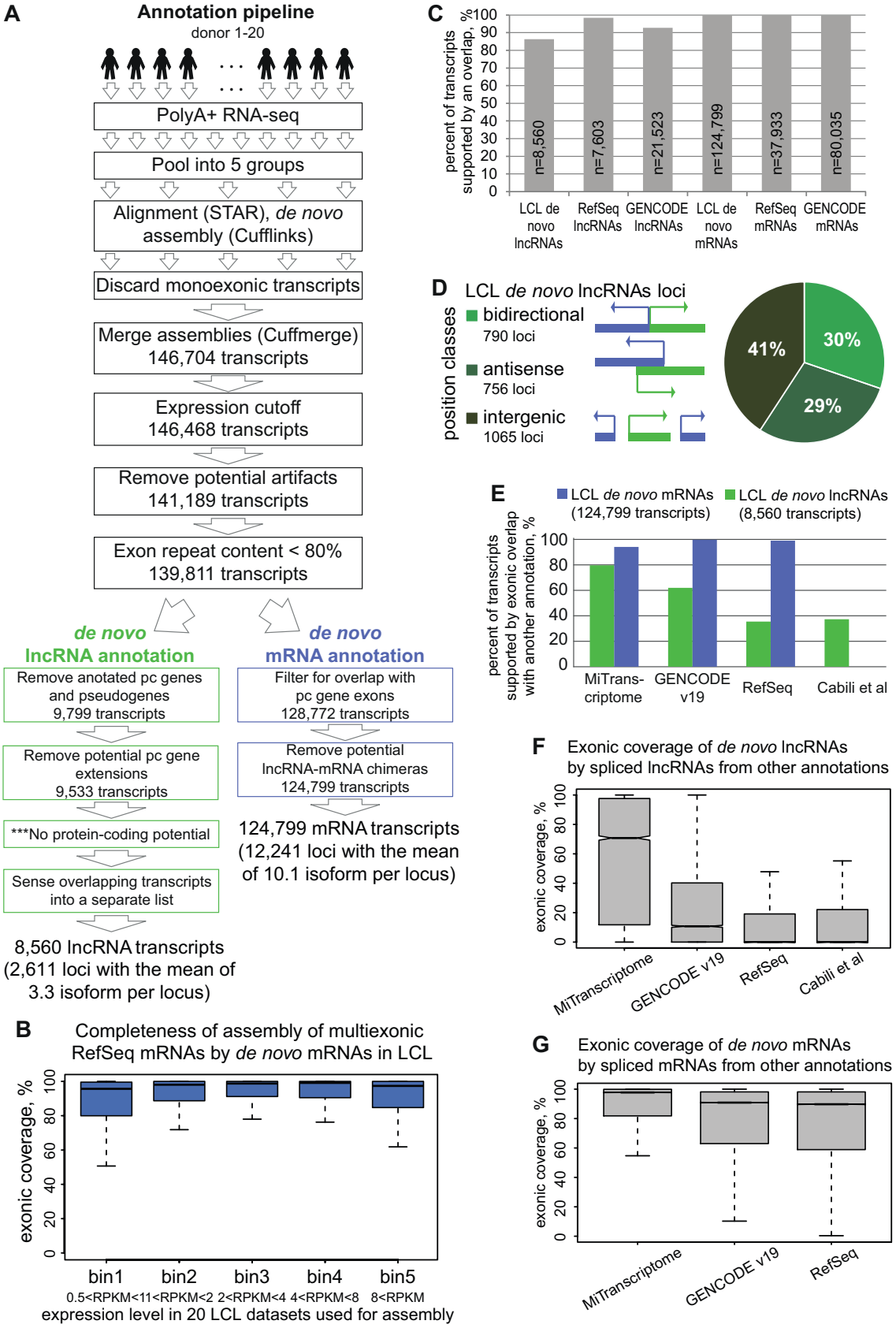


Figure S24. Defining the lncRNA transcriptome of LCL (Lymphoblastoid cell line)

A. Pipeline for *de novo* lncRNA and *de novo* mRNA identification in LCL with transcript numbers at

different stages.

B. Completeness of assembly. Exonic coverage of multi-exonic RefSeq mRNAs by *de novo* mRNA annotation in LCL. Genes are split into 5 bins according to their average expression in 20 donors LCL PolyA+ RNA-seq datasets used for the transcriptome assembly in order to account for the bias between expression level and assembly success. Median values from left to right: 95.7, 98.0, 98.8, 99.1, 97.4.

C. The majority of *de novo* lncRNAs in LCL are supported by an overlap with an EST. The bar plot shows the percentage of transcripts that have a sense exonic overlap with a spliced EST (human EST database downloaded from UCSC) for (from left to right): *de novo* lncRNAs, annotated in this study in LCL, lncRNAs annotated by RefSeq and GENCODE v19, *de novo* mRNAs, annotated in this study in LCL, mRNAs annotated by RefSeq and GENCODE v19. Numbers inside bars indicate total number of transcripts in each annotation.

D. Distribution of the *de novo* annotated lncRNA loci in LCL according to their position relative to protein-coding genes (*de novo*, GENCODE v19 and RefSeq mRNAs). The pie chart shows that 30% (790) of the lncRNA loci are bidirectional (light green), 29% (756) are antisense (medium green) and 41% (1,065) are intergenic (dark green) relative to protein-coding genes.

E. The majority of *de novo* lncRNAs in LCL are supported by an overlap with MiTranscriptome lncRNAs. The bar plot shows percentage of LCL *de novo* lncRNA (green) and mRNA (blue) transcripts supported by an exonic overlap with a multi-exonic lncRNA or mRNA respectively from MiTranscriptome, GENCODE v19, RefSeq and Cabili et al (only lncRNAs) annotations.

F. MiTranscriptome provides best exonic coverage for LCL *de novo* lncRNAs. The box plot shows percent exonic coverage of LCL *de novo* lncRNAs by multi-exonic lncRNAs from MiTranscriptome, GENCODE v19, RefSeq and Cabili et al annotations [2-5]. Number of LCL *de novo* lncRNA transcripts examined in each box – 8,560. Median values from left to right: 70.7, 10.7, 0, 0.Outliers are not displayed in the boxplot.

G. LCL *de novo* mRNAs are nearly fully exonically covered by MiTranscriptome and public annotations. The box plot shows percent exonic coverage of LCL *de novo* mRNAs by multi-exonic mRNAs from MiTranscriptome, GENCODE v19 and RefSeq annotations. Number of LCL *de novo* mRNA transcripts examined in each box – 124,799. Median values from left to right: 97.7, 90.9, 89.7. Outliers are not displayed in the boxplot.

Figure S25

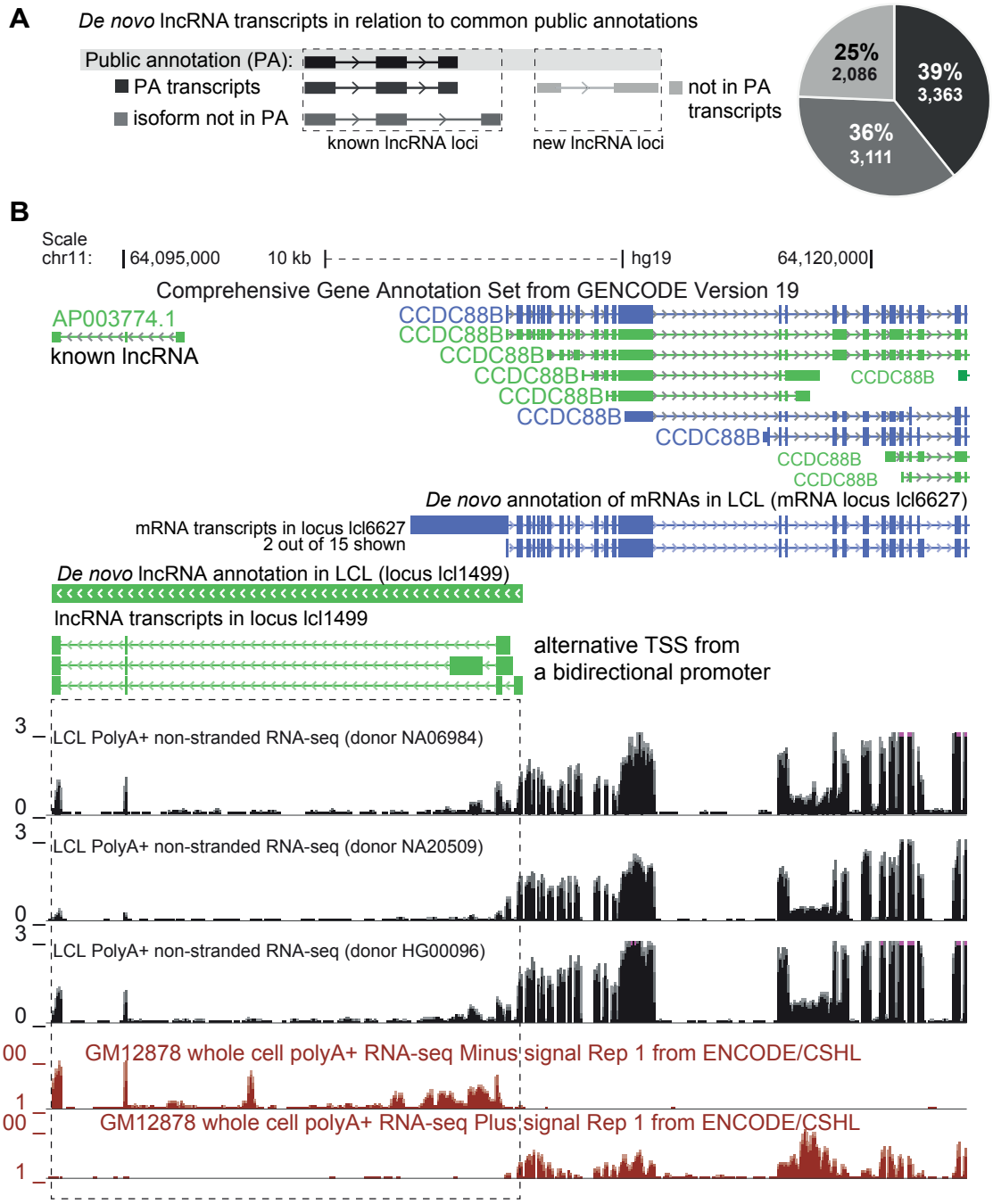


Figure S25. Identification of novel lncRNA transcripts and novel isoforms of known lncRNA loci in LCL cells

A. Distribution of 8,560 *de novo* LCL lncRNA transcripts annotated in the study according to their coverage by the 3 commonly used public annotations as described in Fig. 2A. Publicly annotated ("PA") transcripts constitute 39% of all LCL lncRNA transcripts (3,363 transcripts, black), "isoform not in PA" - 36% (3,111 transcripts, dark grey), "not in PA" – 25% (1,921 transcripts, light grey).

B. An example of a known lncRNA locus (AP003774.1 annotated by GENCODE v19) on chromosome 11 which was extended by our *de novo* annotation in LCL (locus lcl1499) with three new isoforms with an alternative TSS. From top to bottom: GENCODE v19 annotates a lncRNA AP003774.1 and a protein coding gene CCDC88B in this region; *de novo* mRNA annotation in LCL; *de novo* lncRNA annotation in LCL: lncRNA locus 1499 is formed by 3 novel lncRNA isoforms

bidirectional to the CCDC88B protein coding gene, the isoforms share two last exons with the GENCODE annotated lncRNA; Normalized non-strand-specific LCL PolyA+ RNA-seq signal: donor NA06984 - high expression of the lncRNA (average RPKM per transcript 5.8); donor NA20509 - lower expression (average RPKM per transcript 2.6); donor HG00096 - high expression (average RPKM per transcript 6.1); strand-specific PolyA+ RNA-seq signal for LCL sample from ENCODE (GM12878, RNA-sequencing track displayed from ENCODE RNA-seq public hub in UCSC browser) showing the expression of the extended lncRNA locus from reverse strand (Minus Signal) and expression of the protein coding gene from the forward strand (Plus Signal) (average RPKM of the three lncRNA transcripts - 2.9). Dashed box over RNA-seq signal outlines the area of lncRNA expression.

Figure S26

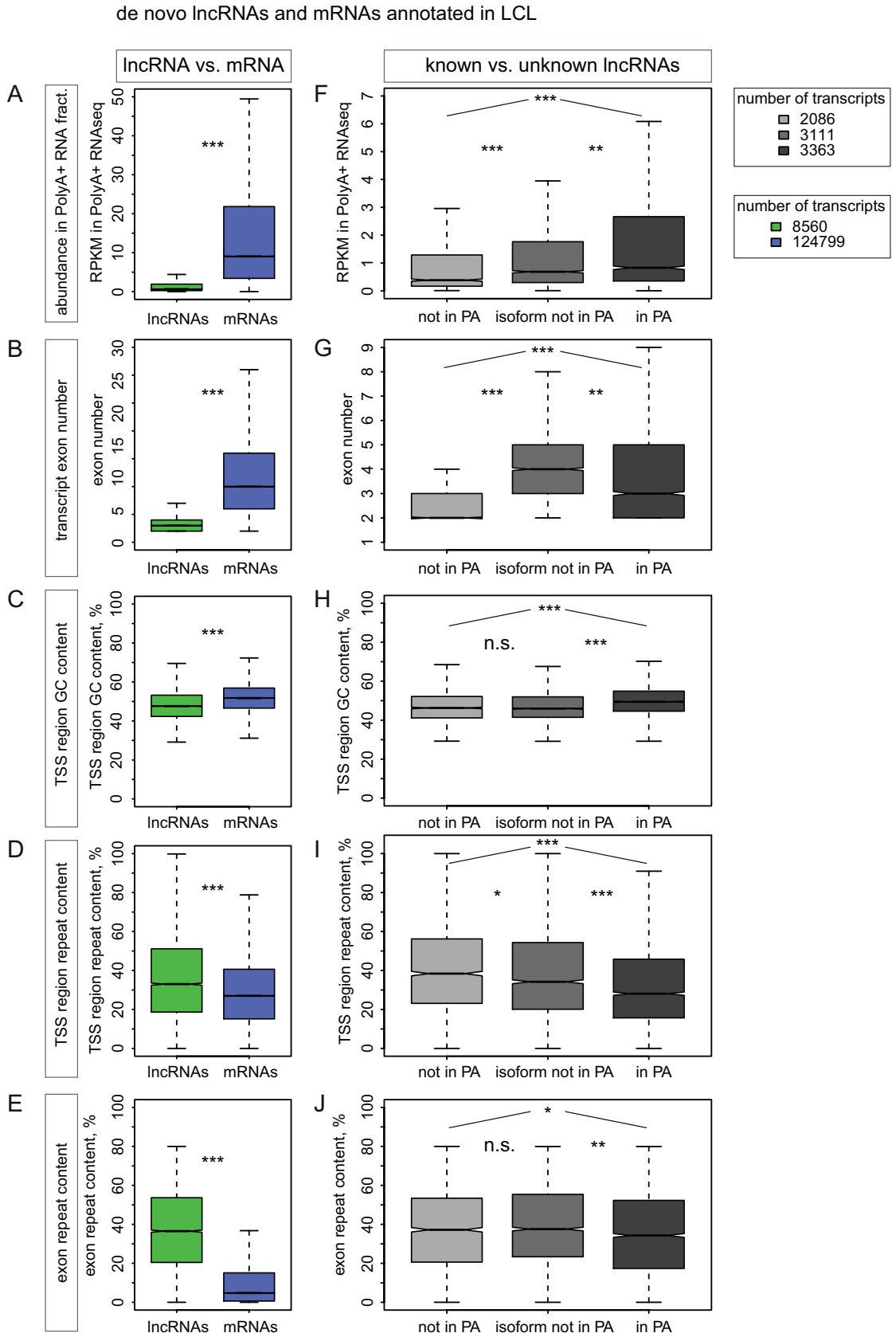


Figure S26. Analyzing features of LCL lncRNAs confirms that difficult to identify lncRNAs are more different from mRNAs than publicly annotated lncRNAs

A. The abundance in the PolyA enriched RNA fraction of LCL *de novo* lncRNA and mRNA transcripts calculated as average from all 462 available PolyA+ RNA-seq samples. Exon number (**B**), percent GC content of TSS region (TSS +/- 1.5 kb) (**C**), percent repeat coverage of TSS region (**D**) and exons (**E**) of *de novo* annotated lncRNAs and mRNA transcripts. Abundance in PolyA enriched RNA fraction (**F**), exon number (**G**), percent GC content of TSS region (**H**), percent repeat coverage of TSS region (**I**) and exons (**J**) of the three classes of *de novo* lncRNA as described in Figure S25A. Remarks: green - all *de novo* lncRNAs, blue – all *de novo* mRNAs, light gray – "not in PA" lncRNA transcripts, medium gray – "isoform not in PA" lncRNA transcripts, dark gray – "PA" lncRNA transcripts. Transcript numbers in each box are indicated on the right. *** - $p<10^{-10}$, ** - $p<10^{-5}$, * - $p<0.01$, n.s. – $p>0.01$. The box plots display the full population but p-values are calculated using Mann–Whitney U test with equalized sample size (Methods). Median values from left to right: **A**: 0.66, 9.05; **B**: 3, 10; **C**: 47.6, 51.7; **D**: 33.0, 27.0; **E**: 36.5, 4.8; **F**: 0.38, 0.68, 0.83; **G**: 2, 4, 3; **H**: 46.3, 45.9, 49.4; **I**: 38.4, 34.2, 28.1; **J**: 37.2, 37.6, 34.3. Outliers are not displayed in the boxplot.

Figure S27

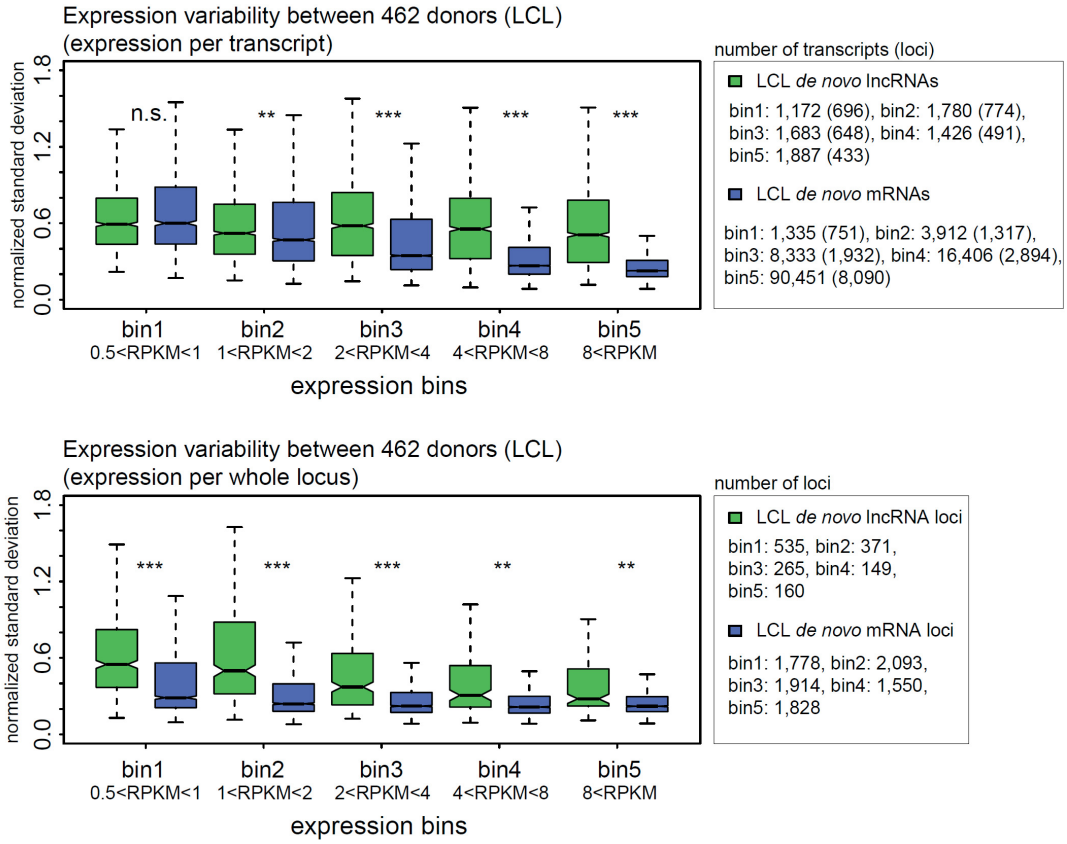


Figure S27. Higher LCL *de novo* lncRNA expression variability is not caused by lower lncRNA expression level.

The phenomenon of increased expression variability of lncRNAs compared to mRNAs is not biased to the absolute expression level of two types of transcripts in the LCL dataset. Normalized standard deviation of *de novo* LCL lncRNA (green) and mRNA (blue) transcripts **(top)** or loci **(bottom)** expression in LCL from 462 donors split into 5 expression bins according to their maximal expression level (RPKM) among 462 donors. Note, that in bin 1 of the upper box plot the difference between lncRNAs and mRNAs is not significant, in contrast to similar analysis performed in granulocytes (Figure S17A). This is likely caused by the increased number of donors (462 - LCL vs. 7 - granulocytes) and the way we split transcripts/loci into bins by their maximal expression among all donors. Thus, bin 1 most likely represents not expressed or marginally expressed transcripts, with an outlier reaching the maximal RPKM of 0.5 to 1, whose variability is strongly affected by detection bias. Such technical bias affects any transcript equally and thus lncRNA and mRNA expression variability is indistinguishable in bin 1. Note that variability of expression calculated over the whole locus (bottom plot) shows consistent lncRNA/mRNA difference in all bins. Remarks: Chr. X and Y were discarded from the analysis. *** - $p<10^{-10}$, ** - $p<10^{-10}$, * - $p<0.01$, n.s. – $p>0.01$. The box plots display the full population but p-values are calculated using Mann–Whitney U test with equalized sample size (Methods). Numbers on the right show number of transcripts/loci in each box of the boxplots. Number in brackets (top boxplot) indicates the number of loci the transcripts in each box initiate from. Median values from left to right: top plot: 0.59, 0.60, 0.51, 0.47, 0.58, 0.35, 0.55, 0.27, 0.51, 0.23; bottom plot: 0.55, 0.29, 0.50, 0.24, 0.37, 0.22, 0.31, 0.21, 0.28, 0.22.Outliers are not displayed in the boxplot.

Figure S28

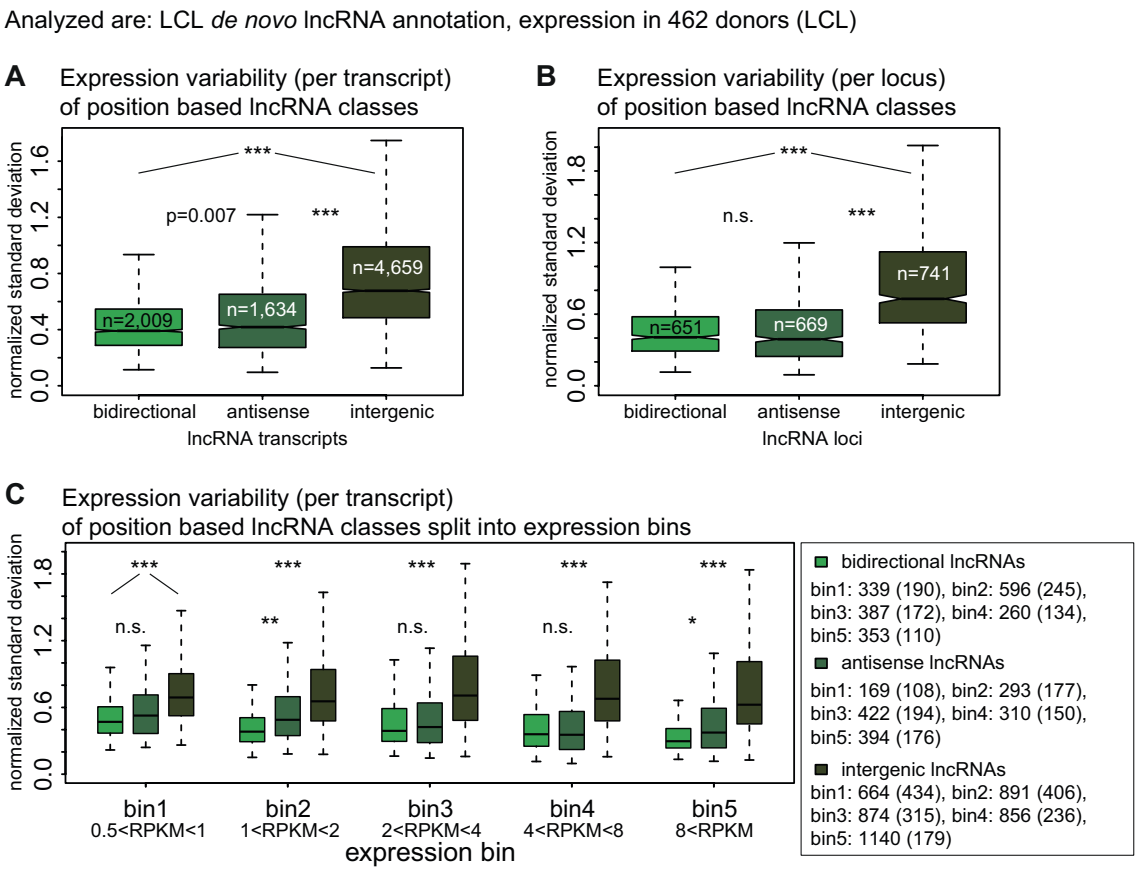


Figure S28. Bidirectional lncRNAs annotated in LCL show reduced variability

A. and **B.** Bidirectional *de novo* lncRNAs annotated in LCL show a decreased level of expression variability. The boxplot shows inter-individual variability of 3 classes of lncRNA transcripts (A)/ loci (B) split according to their position relative to protein coding genes (Figure S24D, Methods). Numbers inside the boxes on the right indicate number of transcripts/loci in each box.

C. Bidirectional LCL lncRNA transcripts are less variable than intergenic lncRNAs independently of expression bin. The boxplot shows inter-individual variability of 3 classes of lncRNAs. Transcripts were split into 5 bins according to their maximal expression among 462 donors. Numbers on the right show number of transcripts in each box of the boxplot. Numbers in brackets indicate the number of loci the transcripts in each box initiate from.

Remarks: Inter-individual variability is calculated as normalized standard deviation of transcripts expression in LCL samples from 462 donors. Remarks: Chr. X and Y were discarded from the analysis. *** - $p<10^{-10}$, ** - $p<10^{-5}$, * - $p<0.01$, n.s. – $p>0.01$. The box plots display the full population but p-values are calculated using Mann–Whitney U test with equalized sample size (Methods). Median values from left to right: **A**: 0.39, 0.42, 0.68; **B**: 0.41, 0.39, 0.73; **C**: 0.47, 0.53, 0.69; 0.38, 0.49, 0.65; 0.39, 0.42, 0.71; 0.36, 0.35, 0.68; 0.30, 0.37, 0.62. Outliers are not displayed in the boxplot. Antisense lncRNAs expression analysis in the LCL dataset (particularly that of expression over the whole locus (**B**)) could be biased to the expression of the overlapped mRNAs since the RNA-seq data was not strand-specific and thus expression variability of antisense lncRNAs is reduced accordant to the reduced mRNA expression variability (Fig. 5D).

Figure S29

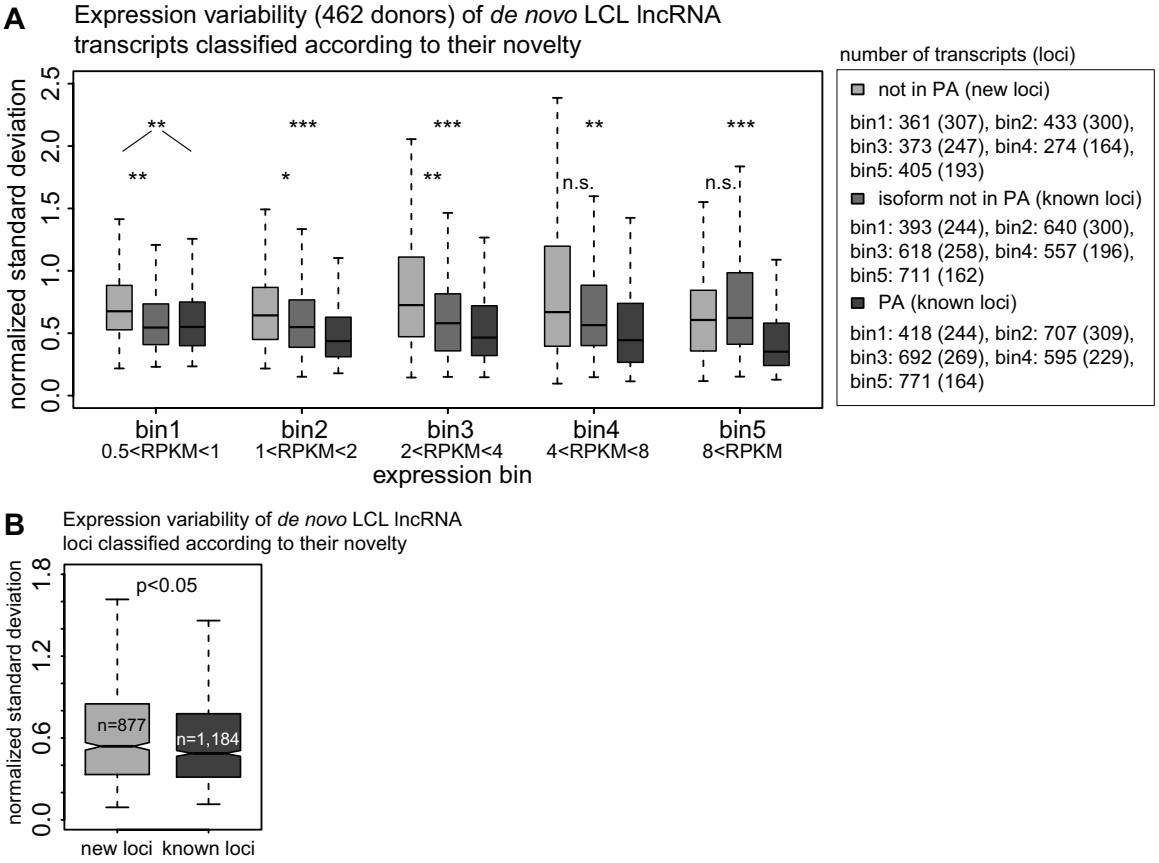


Figure S29. New lncRNAs are more variable than known lncRNAs from LCL de novo annotation: controls

A. Increased variability of "not in PA" and "isoform not in PA" lncRNA transcripts (Figure S25A in Additional File 1 and Fig. 5F) is persistent in all expression bins. The boxplot shows expression variability (normalized standard deviation) of "not in PA" LCL *de novo* lncRNA transcripts (light gray), "isoform not in PA" LCL *de novo* lncRNA transcripts (medium gray) and "PA" LCL *de novo* lncRNA transcripts (dark gray) between LCL samples from 462 donors. Transcripts were split into 5 bins according to their maximal expression of among 462 donors. Median values from left to right: 0.68, 0.55, 0.55; 0.64, 0.55, 0.44; 0.73, 0.58, 0.46; 0.67, 0.56, 0.44; 0.61, 0.62, 0.35. Numbers on the right show number of transcripts in each box of the boxplot. Numbers in brackets indicate the number of loci the transcripts in each box initiate from.

B. Variability of de novo LCL lncRNAs not present in public annotations is slightly increased over annotated lncRNAs when performing "per locus" analysis. The boxplot shows expression variability (normalized standard deviation) of "new" (light gray) and "known" (dark gray) lncRNA loci between LCL samples from 462 donors. Median values from left to right: 0.54, 0.49.

Remarks: chr X and Y were discarded from the analysis. *** - $p<10^{-10}$, ** - $p<10^{-5}$, * - $p<0.01$, n.s. – $p>0.01$. The box plots display the full population but p-values are calculated using Mann–Whitney U test with equalized sample size (Methods). Outliers are not displayed in the boxplot.

Figure S30

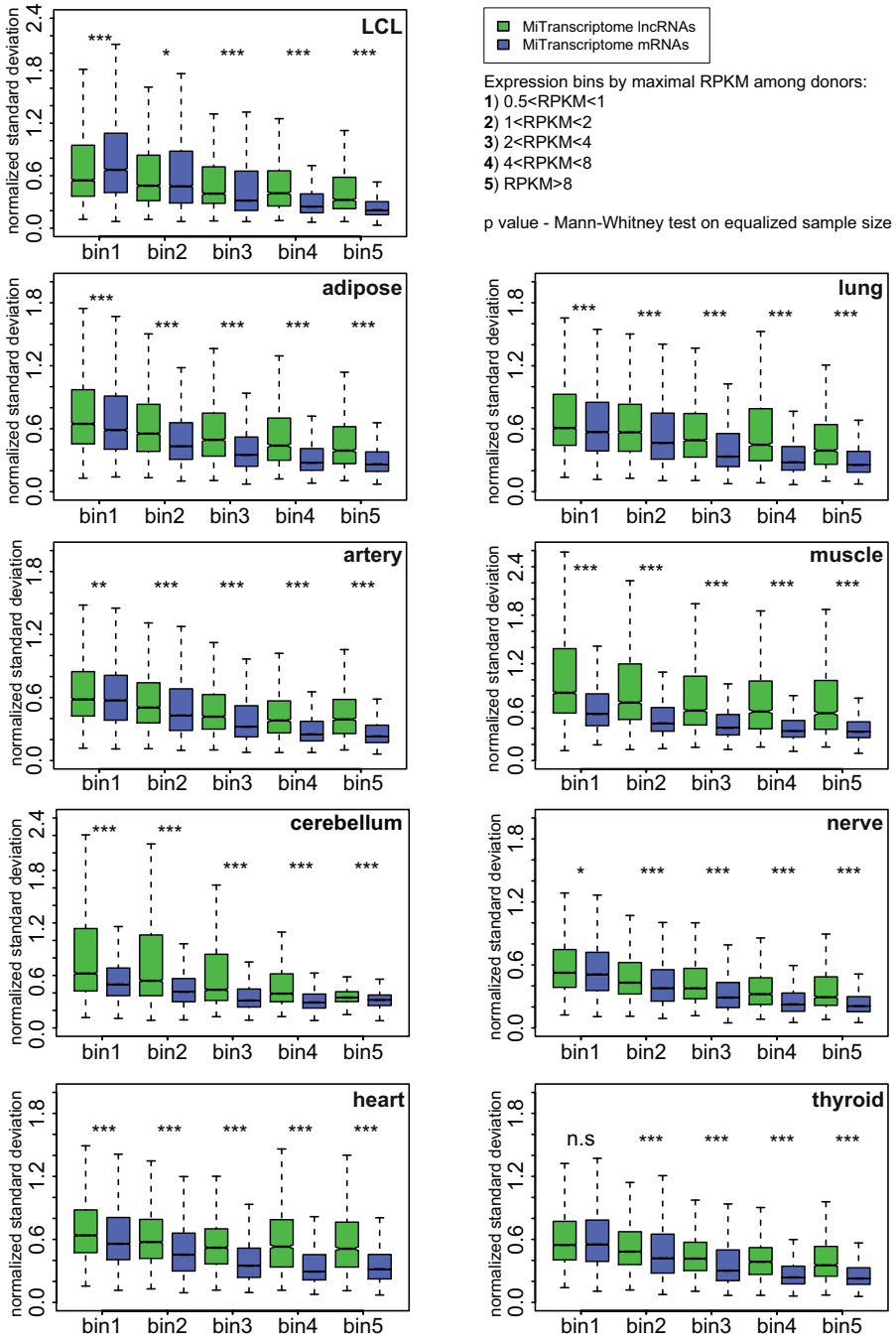
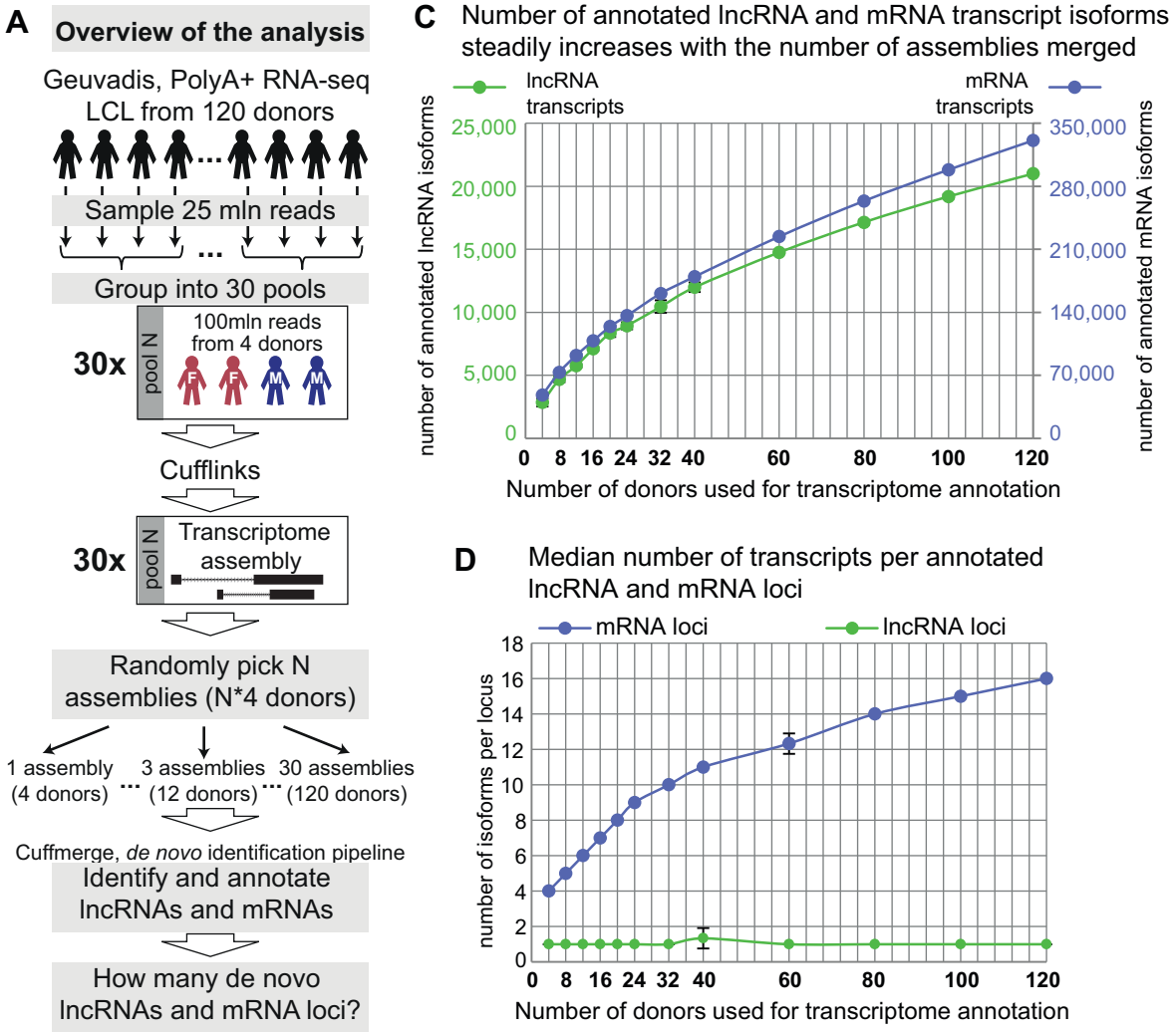


Figure S30. Confirmation of increased lncRNA expression variability in multiple human tissues using GTEx project data: expression level control

Binned normalized standard deviation of MiTranscriptome lncRNA (green) and mRNA (blue) transcripts expression between 20 donors in 9 tissues (as described for Fig. 6, Results, Methods). Transcripts were split into 5 expression bins according to their maximal expression level (RPKM) among 20 donors for each tissue. Chromosomes X and Y were discarded from the analysis. Remarks to the boxplots: *** - $p<10^{-16}$, ** - $p<10^{-10}$, * - $p<0.01$, n.s. – $p>0.01$. The box plots display the full population but p-values are calculated using Mann–Whitney U test with equalized sample size (Methods). Outliers are not displayed in the box plots. Tissue name is indicated for each box plot (top right). Median values from left to right: LCL: 0.54, 0.67, 0.49, 0.48, 0.40, 0.32, 0.40, 0.25, 0.32, 0.21 ; adipose: 0.54, 0.55, 0.48, 0.42, 0.42, 0.30, 0.39, 0.24, 0.35, 0.23 ; artery: 0.58, 0.57, 0.50, 0.43, 0.42, 0.32, 0.38, 0.25, 0.39, 0.23; cerebellum: 0.62, 0.50, 0.54, 0.41, 0.44, 0.31, 0.39, 0.29, 0.35, 0.32 ; heart: 0.64, 0.56, 0.57, 0.45, 0.52, 0.35, 0.53, 0.29, 0.51, 0.32; lung: 0.61, 0.57, 0.57, 0.46, 0.49, 0.33, 0.45, 0.28, 0.39, 0.26; muscle: 0.84, 0.58, 0.72, 0.46, 0.62, 0.41, 0.61, 0.37, 0.59, 0.36; nerve: 0.53, 0.51, 0.43, 0.38, 0.38, 0.29, 0.32, 0.22, 0.29, 0.21; thyroid: 0.55, 0.55, 0.48, 0.42, 0.42, 0.30, 0.39, 0.24, 0.35, 0.23.

Figure 31



B

number of replicates of random pick	number of assemblies	number of donors
3	1	4
3	2	8
3	3	12
3	4	16
3	5	20
3	6	24
3	8	32
3	10	40
3	15	60
3	20	80
3	25	100
1	30	120

Figure S31. *De novo* identification of lncRNA and mRNA loci in LCL using variable number of donors.

A. Algorithm for investigating the relation between the number of identified lncRNA loci and the number of donors analyzed. 120 donors (out of total 462 donors available from Geuvadis dataset [2]) were picked to be used in the study. Only unrelated samples with > 25 million reads were used. Each of the five population groups was represented by 12 females and 12 males (Additional File 11A). All

the 120 RNA-seq data sets were randomly down-sampled to give 25 million paired-end reads each. 120 donors were grouped into 30 pools (each pool contained 2 females and 2 males from the same population) of 100 (25x4) million reads each. Each pool was used to assemble LCL transcriptome using Cufflinks resulting in 30 transcriptome assemblies. We then used 1, 2, 3, 4, 5, 6, 8, 10, 15, 20, 25, 30 transcriptome assemblies to *de novo* annotate lncRNAs and mRNAs in LCL using the annotation pipeline established in the study and plot the number of annotated loci vs. number of donors as an output of the analysis (see Figure 7C and Additional File 11C).

B. For each number of assemblies (i.e. each data point) we performed random picking from the list of 30 assemblies (Additional File 11B).

C. Number of transcript isoforms increases similarly for lncRNA and mRNA loci with the increase of number of donors used for transcript annotation. The plot shows number of LCL *de novo* lncRNA (green) and mRNA (blue) transcripts annotated using different number of assemblies obtained from different number of donors. The Y-axis corresponding to lncRNAs (green) is placed on the left, the Y-axis corresponding to mRNAs (blue) is placed on the right. The range of values from 0 to 25,000 transcripts for lncRNAs and 14x fold more – from 0 to 350,000 for mRNAs. In spite of differing absolute numbers the dynamics of the increase is the same for lncRNAs and mRNAs. Maximum number of lncRNA transcripts - 20,992, maximum number of mRNA transcripts - 330,811 (data table Additional File 11C). Error bars that represent standard deviation in transcript number between three replicates of random assembly picking (**B**) are present for all data points but mostly not visible due to their low values.

D. Increasing donor numbers allows identification of an increasing number of isoforms per mRNA locus, whereas lncRNAs keep low median number of transcripts per locus while increasing the number of loci annotated in the genome. The plot shows the median transcript number in LCL *de novo* lncRNA (green) and mRNA (blue) loci annotated using different number of assemblies obtained from different number of donors. Error bars represent the standard deviation of loci number between 3 replicates of random picking for each number of assemblies used for identification (data table Additional File 11C). Error bars that represent standard deviation between three replicates are present for all data points but mostly not visible due to their low values.

E. Estimating how many unknown mRNA loci could be assembled and then discarded at the protein-coding capacity filtering step with the increasing number of donors.

Figure S32

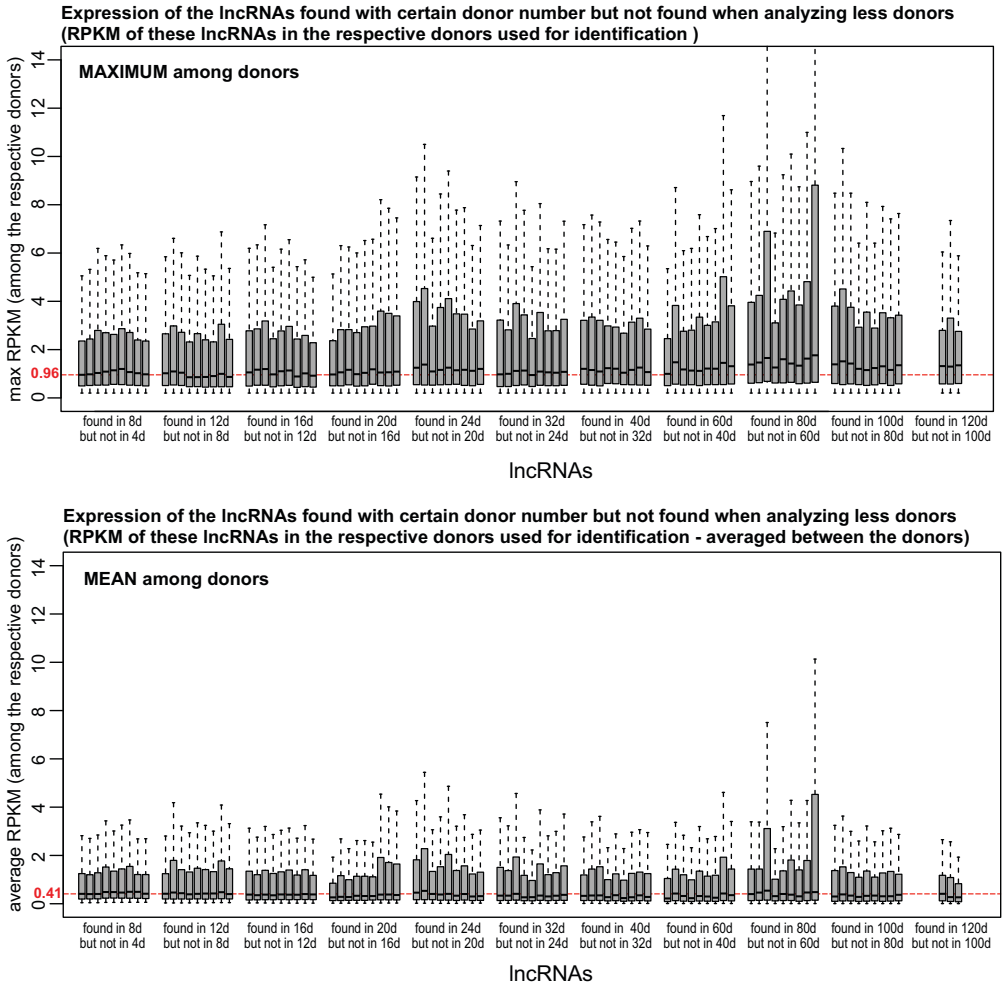


Figure S32. Increasing donor number does not tend to identify only marginally expressed lncRNAs

Overview: Expression level of the identified lncRNAs (Fig. 7C) is not decreasing with the increase of donor number and potential sensitivity of the pipeline. We asked if the lncRNA loci that we identified by adding more donors to the analysis were more lowly expressed than those identified using less donors. If so, this would indicate that the amount of sequencing data rather than number of individuals allows identification of new minimally expressed lncRNA loci. We plotted the expression level of lncRNA transcripts initiating from loci annotated using more donors, that could not be identified (defined as 50% sense overlap) using less donors. We found that expression of lncRNAs that require more and more donors to be identified does not anti-correlate with the donor number. Thus the identification of more lncRNAs in larger donor collections does not specifically identify lowly expressed transcripts.

The boxplot shows the maximal (**top**) and mean (**bottom**) (among the donors used for the annotation) expression level of *de novo* LCL lncRNA transcripts annotated using the indicated number of donors (indicated on the X axis). Only transcripts expressed from loci that have not been identified ("identified"= >50% sense overlap) using less donors (indicated on the X axis) are displayed. 9 boxes for each number of donors show the result for all the nine possible pairwise comparisons between three replicates of each donor-number annotations (e.g. "found in 8d but not in 4d": box1. lncRNA loci in 8-donor annotation replicate 1 not identified in 4-donor annotation replicate 1, box2. lncRNA loci in 8-donor annotation replicate 1 not identified in 4-donor annotation replicate 2, box3. lncRNA loci in 8-donor annotation replicate 1 not identified in 4-donor annotation replicate 3, box4. lncRNA loci in 8-donor annotation replicate 2 not identified in 4-donor annotation replicate 1, etc). 120-donor annotation only has 1 replicate, thus giving just three boxes. Outliers are not displayed in the boxplot. Horizontal dashed red line and red number indicate median level of the first box in the boxplot.

Figure S33

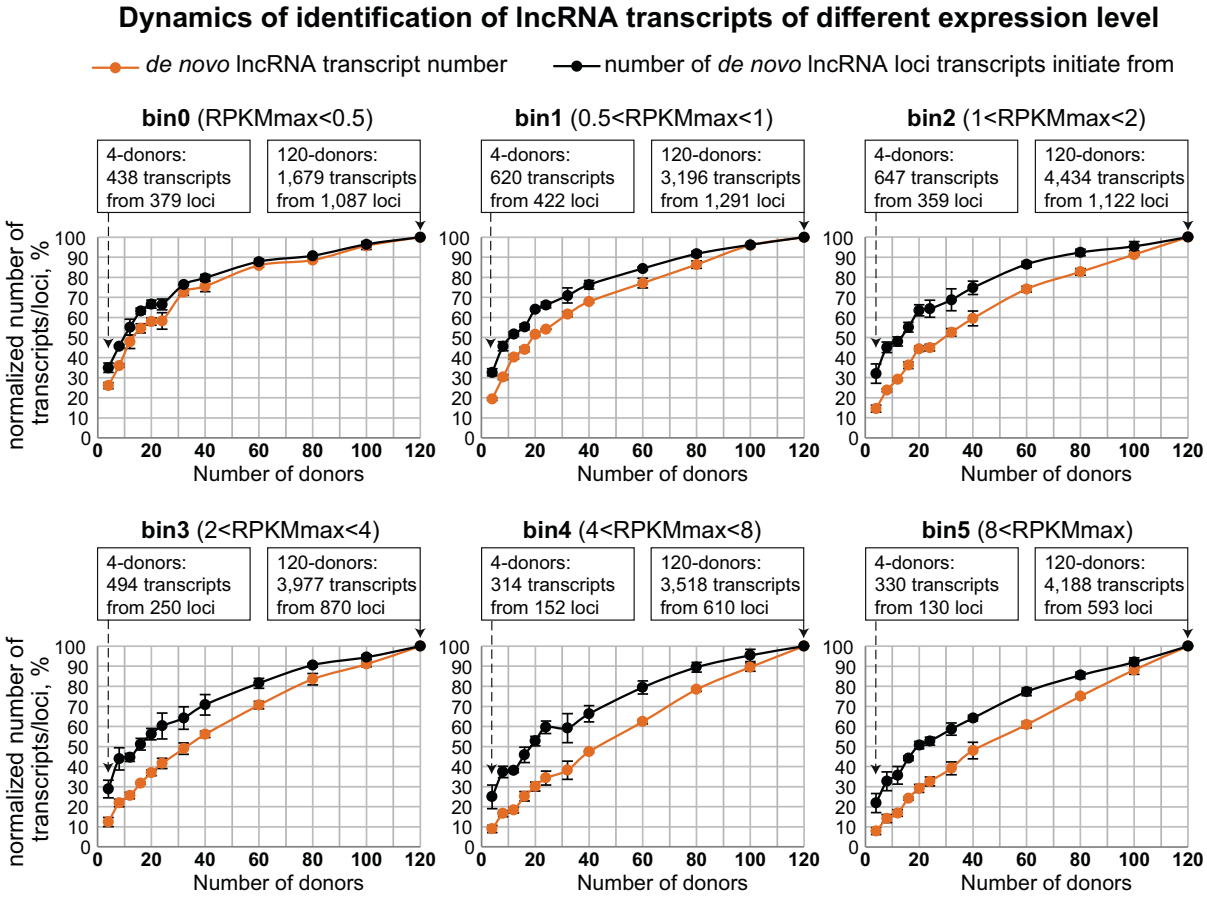


Figure S33. Increasing donor number identifies increased numbers lncRNAs in all expression bins.

Dynamics of identification upon donor number increase of transcripts split into 6 bins according to their maximal expression among donors used for their identification (Additional File 11E). Every plot shows number of transcripts (orange) and loci (black) these transcripts initiate from. Number of transcripts/loci is normalized to the number of transcripts/loci in 120-donor annotation. Absolute number of transcript/loci is given for 4-donor and 120-donor annotations (boxes above the plots). Error bars show standard deviation between 3 replicates for each donor number. Remarks: bin 0 has not been used in other figures and represent marginally expressed transcripts (RPKM<0.5).

Figure S34

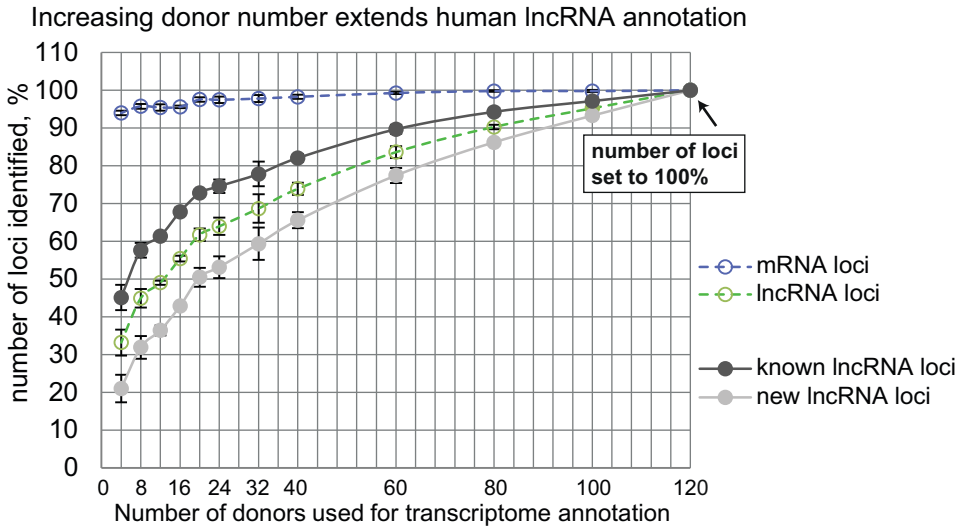


Figure S34. Number of new lncRNA increases more dramatically with donor number increase compared to known lncRNAs.

The dynamics of *de novo* identification of new (light gray) and known (dark gray) lncRNA loci in LCL using an increasing donor number (new – not covered by reference public annotations, known – covered by reference public annotations. As described for Fig. 2A). Dynamics for all lncRNA loci (dashed green line) and mRNA loci (dashed blue line) is indicated for comparison. The loci number is normalized to the total number of loci in the most comprehensive 120-donor annotation and set to 100% for each curve. Maximum number (100%) for new lncRNA loci: 2,063, maximum number (100%) for known lncRNA loci: 2,103. Error bars indicate standard deviation between 3 replicates of random picking for each number of assemblies used (Additional File 11C).

Figure S35

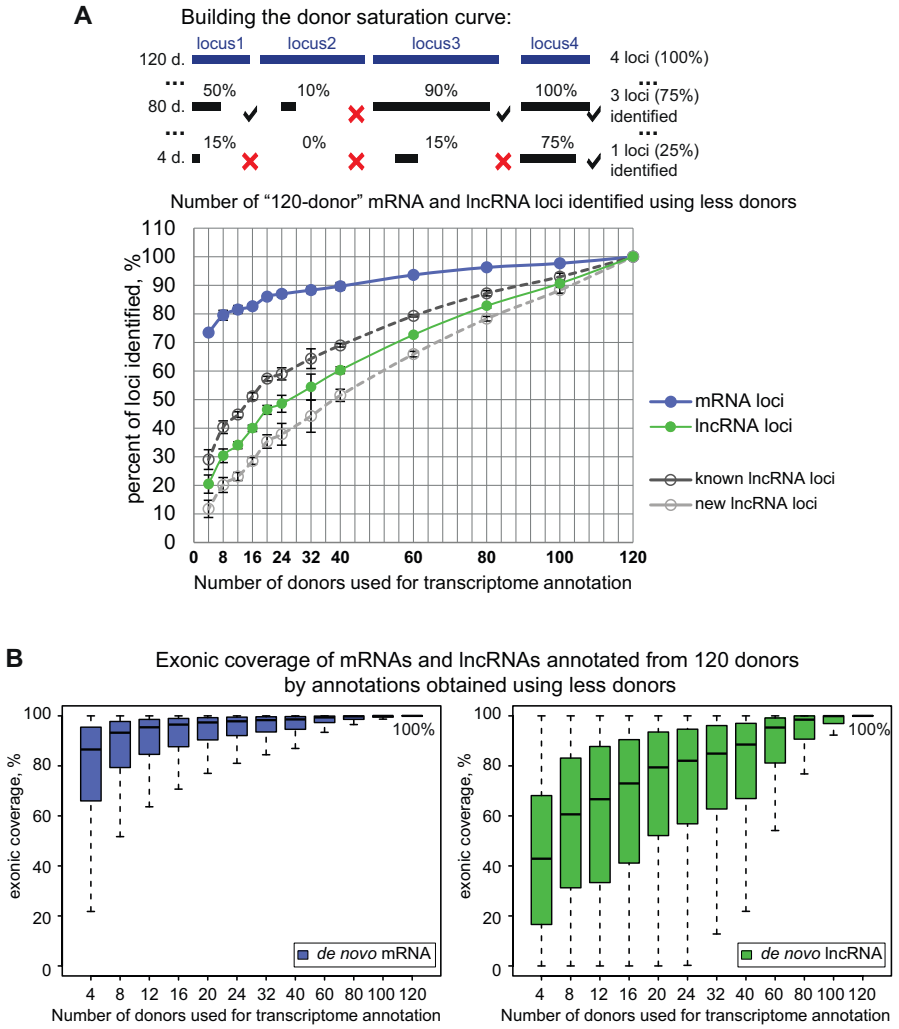


Figure S35. Donor saturation curve analysis of 120-donor lncRNA and mRNA identification using less donor in the identification pipeline

A. Donor saturation analysis. Top: definition of an "identified locus": lncRNA and mRNA annotation obtained using 120 donors (the most comprehensive annotation) is used as a reference for comparison and the number of loci in this annotation is set to 100%. When analyzing annotations obtained from fewer donors, a reference locus is called "identified" if it is covered by the down-sampled annotation to at least 50% of its length (black tick). In case the coverage is less the locus is not considered "identified" (red cross). Below: normalized number of 120-donor reference annotation loci "identified" when running the identification pipeline with fewer donors. Displayed are donor saturation curves for mRNA loci (blue), all lncRNA loci (green), only new lncRNA loci (unfilled light grey circles and dashed line) and only known lncRNA loci (unfilled dark grey circles and dashed line). Number loci in the 120-donor annotation was set to 100% for all the four displayed loci type. Error bars: representing standard deviation between 3 replicates are present for all data points but due to low values are mostly not visible (Additional File 11D).

B. Donor saturation curve of exon structure identification. The plot shows the percent exonic coverage (Supplemental Methods) matching the reference annotation generated from 120 donors (set to 100%), obtained using fewer donors. Left boxplot: *de novo* mRNA transcripts (blue), right boxplot: *de novo* lncRNA transcripts (green). Outliers are not displayed. Median exonic coverage values (from left to right): mRNAs – 86.6%, 93.3%, 95.4%, 96.5%, 97.4%, 97.8%, 98.3%, 98.6%, 99.4%, 99.8%, 100.0%, **100%**, lncRNAs – 42.9%, 60.6%, 66.7%, 73.0%, 79.4%, 82.0%, 84.9%, 88.5%, 95.3%, 98.5%, 99.9%, **100%**

SUPPLEMENTAL METHODS

1. Granulocyte isolation

Granulocytes and mononuclear cells (MNCs) were isolated from freshly collected blood using gradient density centrifugation (Ficoll-Plaque PREMIUM 1.078 g/ml, GE Healthcare Life Sciences). Briefly, 45 ml of fresh blood was centrifuged at 100g for 10 minutes at room temperature and the yellowish supernatant was discarded to remove the platelet rich plasma. The remainder was diluted approximately four fold with room temperature (RT) PBS (+2 mM EDTA) to 144 ml. 35 ml of diluted blood was carefully layered on top of 15 ml of Ficoll (equilibrated at RT) in a 50 ml Falcon tube and four such tubes were centrifuged at RT at 400g for 33 minutes with acceleration/brake at minimum. After centrifugation the upper layer was carefully removed and discarded, the MNC layer immediately on top of the Ficoll separation layer was collected into a new tube and washed in ice cold PBS (+2mM EDTA) by centrifugation at 300g for 10 minutes. The upper Ficoll layer was carefully removed and the underlying remaining layer containing the granulocyte population was depleted for erythrocytes using Cell Lysis Solution (Promega) in two 5 minute incubation steps followed by 5 minute 300g centrifugation at RT. Granulocytes were then washed in ice cold PBS (+2 mM EDTA) by centrifugation at 300g for 5 minutes. Both MNCs and granulocytes underwent one further ice-cold PBS (+2 mM EDTA) washing step (8 minutes at 200g) to remove residual platelets and to create a pellet for immediate RNA isolation

2. RNA isolation using TRI reagent

Pelleted cells were lysed in 1 ml of TRI reagent (Sigma-Aldrich T9424) per 10^7 cells by active pipetting / vortexing and incubated for 5 minutes at room temperature (RT) and the lysate frozen and stored at -80ºC for later RNA isolation. After thawing on ice, 0.1 ml BCP (Molecular Research Center, Inc.) was added per 1 ml of TRI reagent, followed by intensive shaking and 10 minute incubation at RT, and then centrifuged for 12 minutes at 12000g at 4ºC. The upper aqueous phase was transferred to a new tube with 0.5 ml isopropanol, vortexed and incubated for 10 minutes at RT to allow RNA precipitation. The RNA precipitate was pelleted by centrifugation at 12000g for 12 minutes at 4ºC, the supernatant was removed and the pellet washed with 1 ml of 70% ethanol (7500g, 5 minutes, 4ºC). After ethanol removal, the pellet was air-dried for few minutes and dissolved in RNA Storage Solution (RSS) (Ambion) and stored at 80ºC. DNaseI treatment was performed for 10 μg RNA per 50μl reaction, using the DNA-free kit (Ambion).

3. Reverse transcription

DNase I treated RNA was reversely transcribed into cDNA using RevertAid First Strand cDNA Kit (Fermentas). 0.6 μg of RNA per 20 μl was used and –RT control (lacking Reverse Transcriptase enzyme) was performed for each set of RT reactions.

4. Ribosomal RNA depletion

Total RNA was depleted for ribosomal RNA using the RiboZero rRNA removal kit Human/Mouse/Rat (Epicentre). 2-4 μg of DNase I treated RNA was used for each reaction. At the end of the protocol the volume of RiboZero treated RNA was adjusted to 180 μl and RNA was precipitated by adding 18 μl of 3M sodium acetate (Ambion), 2 μl of Glycogen (10mg/ml) and 600 μl of ice-cold 96% ethanol. Following overnight incubation at -20ºC the RNA precipitate was recovered by centrifugation (50 minutes, 16000g, 4ºC). The RNA pellet was washed as described above, dissolved in 1 μl nuclease-free water and diluted with 19.5 μl of Elute, Prime, Fragment Mix (TruSeq RNA Sample Prep Kit v2, Illumina). After incubation at 94ºC for 3 minutes to allow priming and fragmentation of RNA, 17 μl was transferred to DNA LoBind Tubes (Eppendorf) for immediate stranded library preparation.

5. Polyadenylated RNA enrichment

PolyA enriched RNA was prepared using the TruSeq RNA Sample Prep Kit v2 (Illumina) and 2-4 μg of DNase I treated RNA. At the end of the protocol the magnetic bead-bound RNA (Agencourt AMPure XP - PCR Purification, Beckman Coulter) was incubated with 19.5 μl of Elute, Prime, Fragment Mix (TruSeq RNA Sample Prep Kit v2, Illumina) at 94ºC for 3 minutes and the magnetic beads were pelleted using a magnetic stand. 17 μl of supernatant was transferred to a DNA LoBind Tube (Eppendorf) for immediate stranded or non-stranded library preparation.

6. Preparation of strand-specific RNA-seq libraries

First-strand cDNA synthesis: 8 μl of First Strand Master Mix (TruSeq RNA Sample Prep Kit v2, Illumina), supplemented with 1 μl SuperScript II per 9 μl First Strand Master Mix, was added to 17 μl of either PolyA+ or total RNA, followed by vortexing and incubation consequently at 25ºC for 10 minutes, 42ºC for 50 minutes, 70ºC for 15 minutes to perform reverse transcription reaction. First-strand cDNA was then cleaned using G-50 columns (Illustra ProbeQuant G-50 Micro Columns, GE

Healthcare). G-50 columns were preliminarily washed twice with 500 μl of 1 mM Tris-HCl pH 8.0 by centrifugation at 700g for 2 minutes at RT. First-strand cDNA was diluted to 30 μl by adding 5 μl of Elution Buffer (TruSeq RNA Sample Prep Kit v2, Illumina) and then added to the G-50 column and centrifuged at 700g for 2 min at RT. The eluate was then adjusted to 52 μl by adding nuclease-free water.

Second-strand cDNA synthesis: Second Strand master mix was prepared freshly as follows: 1 μl of 10 x Reverse Transcription Buffer (Invitrogen), 15 μl of 5 x Second Strand Syntheses Buffer (Invitrogen), 1 μl 50 mM $MgCl_2$, 1 μl of 100 mM DL-Dithiothreitol (DTT, Invitrogen), 2 μl of dUNTP mix (10 mM dATP, 10 mM dCTP, 10 mM dGTP, 10 mM dUTP) (Thermo Scientific), 0.5 μl of 10 U/μl E.coli DNA ligase (Invitrogen), 2 μl of 10 U/μl DNA Polymerase (Invitrogen) and 0.5 μl 2U/μl RNase H (Invitrogen). 23 μl of the master mix was added to the cleaned first-strand cDNA sample and incubated at 16ºC for 2 hours. The cDNA was then cleaned using magnetic beads (Agencourt AMPure XP - PCR Purification, Beckman Coulter). 135 μl of RT pre-vortexed beads was added to the reaction, the mixture was incubated for 15 minutes at RT to allow binding and put on the magnetic stand for 5 minutes to collect the beads. Supernatant was discarded and the bead pellet was washed twice with 200 μl freshly prepared 80% ethanol and air-dried for 15 minutes. The beads were then resuspended in 52.5 μl Resuspension Buffer (TruSeq RNA Sample Prep Kit v2, Illumina), incubated for 2 minutes at RT and put on magnetic stand for 2 minutes to collect the beads. 50 μl of the supernatant containing cleaned double stranded cDNA were transferred to a new DNA LoBind Tube.

End repair procedure: 40 μl of End Repair Mix (TruSeq RNA Sample Prep Kit v2, Illumina) was added to 50 μl of double stranded cDNA. The mixture was vortexed and incubated at 30ºC for 30 minutes. Afterwards the cDNA was cleaned by the procedure described above using 160 μl of magnetic beads. cDNA was eluted from beads using 20 μl of Resuspension Buffer.

3'end adenylation: After the end repair 3' ends of cDNA were adenylated by adding 12.5 μl of A-Tailing Mix (TruSeq RNA Sample Prep Kit v2, Illumina) and incubating the mix at 37ºC for 30 minutes and then at 70ºC for 5 minutes.

Adapter ligation: The adapters (barcodes) provided by TruSeq RNA Sample Prep Kit v2 were chosen according to the plan of how to pool the libraries to be sequenced on one lane. To ligate a desired adapter to cDNA 2.5 μl of Resuspension Buffer, 2.5 μl of Ligation Mix (TruSeq RNA Sample Prep Kit v2, Illumina) and 2.5 μl of the RNA Adapter Index (TruSeq RNA Sample Prep Kit v2, Illumina) were added to the reaction and incubated at 30ºC for 10 minutes. To stop the ligation we added 5 μl of Stop Ligation Buffer (TruSeq RNA Sample Prep Kit v2, Illumina) to the reaction. Afterwards the cDNA was cleaned and eluted twice by the procedure described above: first time with 42 μl of magnetic beads and 52.5 μl of Resuspension Buffer and second time with 50 μl of magnetic beads and 22.5 μl of Resuspension Buffer.

UDGase treatment: Removal of the second strand cDNA with incorporated dUTPs was performed by adding 2.3 μl of 10 x UNG Buffer (company) and 1 μl of 5 U/μl UDGase (company) to cleaned adapter-ligated cDNA and incubation at 37ºC for 30 minutes.

Library enrichment by PCR**:** To enrich for cDNA fragments PCR was performed by adding 5 μl of PCR Primer Cocktail (TruSeq RNA Sample Prep Kit v2, Illumina), 25 μl of PCR Master Mix, vortexing and running the PCR program: 98ºC 30 seconds, 8 cycles (98ºC 10 seconds, 60ºC 30 seconds, 72ºC 30 seconds), 72ºC 5 minutes). The PCR reaction was then cleaned and eluted twice by the procedure described above: first time with 42 μl of magnetic beads and 52.5 μl of Resuspension Buffer and second time with 53 μl of magnetic beads and 20 μl of Resuspension Buffer.

7. RNA-seq read alignment

Raw RNA-seq reads were aligned with STAR [6]. The default STAR options were optimized in order to give a stringent unbiased alignment. (i) the hg19 genome was built without pre-annotated splice junctions (SJ) to allow non-biased *de novo* splice junction identification: *STAR_2.3 --runMode genomeGenerate --genomeSAindexNbases 13 --genomeChrBinNbits 18 --genomeSAsparseD 2 --genomeDir [hg19genome_for_STAR] --genomeFastaFiles hg19.fa --runThreadN 8*. (ii) to be stringent in novel splice site discovery we only considered canonical SJs using *--outFilterIntronMotifs RemoveNoncanonical* (keeping only canonical SJ was also important for further Cufflinks transcriptome assembly from non-strand-specific data), required the minimum overhang length for SJs on both sides to be >= 16bp by *--outSJfilterOverhangMin 30 16 16 16* , required that a SJ is supported by at least 2 reads by *--outSJfilterCountTotalMin 4 2 2 2*, required the minimum overhang of a spliced read to be >=6 by *--alignSJoverhangMin 6*, required to only keep the spliced reads which passed the previous filtering requirements by *--outFilterType BySJout*. (iii) we required that intron size <=300kb *(--alignIntronMax 300000*) and the maximum gat between two read mate <=500kb *(--alignMatesGapMax 500000*). The alignment was performed using with the following command (all the not shown options were set to default), followed by sorting and indexing of the resulting BAM file: *STAR_2.3 --genomeDir [hg19genome_for_STAR] --readFilesIn READ1.fastq.gz READ2.fastq.gz --readFilesCommand zcat --outSJfilterOverhangMin 30 16 16 16 --outSJfilterCountTotalMin 4 2 2 2 --alignSJoverhangMin 6 --outFilterType BySJout --outSAMstrandField intronMotif --outFilterIntronMotifs RemoveNoncanonical --alignIntronMax 300000 --alignMatesGapMax 500000 --runThreadN 8 --outFileNamePrefix outprefix --outStd SAM --outSAMmode Full | samtools view -bS -> outprefix.bam*

#sort the bam file#: samtools sort outprefix.bam outprefix.sorted

#create indexed bam file#: samtools index outprefix.sorted.bam

Strand specificity of the aligned data was assessed using RSEQC package:

infer_experiment.py –r RefSeq_mRNAs.bed –i alignment.bam

Insert size of the libraries was assessed using RSEQC package:

inner_distance.py –r RefSeq_mRNAs.bed –i alignment.bam

8. Public gene annotations used in the study

We used a public annotation of lncRNAs provided by the GENCODE project (http://www.gencodegenes.org/releases/19.html). GENCODE v19 lncRNA annotation consists of 23,898 transcripts (21,523 multi-exonic transcripts). We also used lncRNA annotation provided by RefSeq (NR_* (>200nt cDNA length) annotation downloaded from the UCSC table browser on 3 June 2014) and lncRNA annotation published by Cabili et al [3]. RefSeq lncRNA annotation consists of 8,236 transcripts (7,603 multi-exonic transcripts). Cabili et al lncRNA annotation consists of 21,630 transcripts (21,595 multi-exonic transcripts). We used two public annotations of protein-coding genes: RefSeq (NM_* (mRNA) annotation downloaded from the UCSC table browser on 27 January 2014) and an annotation provided by the GENCODE project. RefSeq mRNA annotation consists of 39,562 transcripts (37,933 multi-exonic transcripts). GENCODE v19 mRNA annotation consists of 81,814 transcripts (80,035 multi-exonic transcripts). We used a public annotation of pseudogenes provided by the GENCODE project (GENCODE v19 – 17,572 pseudogenes). We used an annotation of repeat elements – RepeatMasker downloaded from the UCSC browser.

9. Calculation of GC content

GC content of selected regions was calculated using bedtools *nuc -fi hg19.fa -bed regions.bed*

10. RT- and qRT-PCR primer design

Primers for RT- and qRT-PCR were designed using Primer3 software (http://biotools.umassmed.edu/bioapps/primer3_www.cgi).

11. Annotating mRNAs and lncRNAs in primary granulocytes

Transcriptome assembly: 17 PolyA+ RNA-seq data sets (comprising ten different healthy donors) with total of 784 M mapped reads were used to create the granulocyte mRNA and lncRNA annotations used for further analysis (Additional File 2A, B). Although it is suggested that each sample's transcriptome should be assembled separately [7], PolyA+ datasets were pooled into 6 parts at the stage of alignment in order to increase the sensitivity of the transcriptome assembly (Additional File 2C). Pools were created so that they contained 24-37 M spliced reads each (85-102 M mapped reads for 4 x 100bp paired-end pools and ~ 220 M reads for 2 x 50 bp paired-end pools). Six alignment .BAM files were created as described in Methods and then used to perform six separate transcriptome assemblies using Cufflinks [7]. In order to prevent a bias towards identification of already annotated transcripts no reference gene annotation was provided to Cufflinks. In order to avoid Cufflinks pausing over problematic regions in the genome annotated pseudogenes (GENCODE v 19) were masked using -- *mask-file* option. Cufflinks was run for 6 .BAM files (6 pools' alignments) with the following options: *cufflinks --multi-read-correct --output-dir [output_dir] -F 0.01 -p 7 --library-type fr-unstranded (for stranded pools --library-type fr-firststrand) --mask-file pseudogenes.gtf PolyA_pool_N.sorted.bam*

Removal of mono-exonic transcripts: All single exon transcripts were removed from each of the 6 transcriptome assemblies using *gffread* tool (part of Cufflinks package): *gffread transcripts.gtf -T -U -o transcripts_multiexon.gtf*. The three main rationales for focusing on multi-exonic transcripts were the following. First, the majority of mono-exonic transcripts assembled by Cufflinks appear to be intronic signals and other artifacts. Second, mono-exonic transcripts assembled by Cufflinks are not continuous unlike spliced transcripts whose continuity is supported by the spliced reads spanning thousands of kilobases. Third, 13 out of 17 PolyA+ datasets we used for the assembly were not strand-specific. However, the presence of a splice site in a transcript allowed Cufflinks to infer the strand it was transcribed from. Previous publications extensively used non-strand-specific data for Cufflinks based annotation of multi-exonic lncRNAs in human and have shown that the error rate of inferring the strand from the canonical splice sites was negligible [3, 8]. Additionally, to be maximally stringent, we also removed potential strand specificity artifacts at later filtering steps.

Merging the annotations: The resulting six multi-exonic transcriptome annotations were merged using Cuffmerge with the following command: *cuffmerge -s hg19.fa --keep-tmp -p 8 --min-isoform-fraction 0 list_of_6_annotation_files.txt*. The resulting merged annotation contained 158,038 transcripts comprising 13,589 loci.

11.1. Filtering Steps

The merged transcriptome annotation was then filtered in order to create granulocyte mRNA and lncRNA annotations for further use in the study. It was necessary to *de novo* annotated mRNA genes as we noticed that comparison of *de novo* annotated lncRNAs to mRNAs annotated by RefSeq or GENCODE, was misleading due to the precision of the *de novo* annotation for granulocytes where only the isoforms actually expressed in granulocytes were annotated, and due to potential artifacts of *de novo* annotation missing from the curated, comprehensive and experimentally supported public mRNA annotations. Thus, in order to avoid potential technical biases, mRNA annotation was created *de novo* in granulocyte using the same pipeline that was used for lncRNAs. The following common filtering steps for both mRNAs and lncRNAs:

Expression cut off: We used 6 pools of RNA-seq data to increase the sensitivity of Cufflinks to lowly expressed spliced isoforms of lncRNAs. However, the increased sensitivity could potentially result in false positive transcripts. We checked if the assembled transcripts could be detected in at least one of the diverse granulocytes RNA-seq samples from 10 individuals used in the study. We used an expression level calculation method independent from Cufflinks (*RPKM_count.py* from RSeQC package) to analyze the expression of all the *de novo* assembled transcripts in all the available granulocyte RNA-seq datasets (17 PolyA+ RNA-seq datasets + 21 total RNA-seq datasets). If a transcript was not expressed (RPKM<=0.2) in any of the datasets, we called it an artifact and removed from the annotation. By this step 0.4% (631) of transcripts was removed resulting in residual 157,407 transcripts.

Filtering out transcripts potentially assigned to the wrong strand: Although Cufflinks infers the strand of the transcript from the direction of spliced junctions within this transcript, we wanted to be stringent and remove potential artifacts assigned to the wrong strand. For that we performed two steps of filtering. First, using a custom script, we checked if the *de novo* annotation contains transcripts that have a "mirror" transcript on the other strand (that is a transcript that has exons with > 30% reciprocal antisense overlap with exons of a transcript on another strand). Such transcripts arising from problems in the strand-specificity step could be potential artifacts and had to be removed in order to create a stringent set of transcripts. In each pair of such transcripts we then kept the one with the higher expression level. As an expression level estimate for each transcript we took the maximum RPKM among all the stranded RNA-seq datasets (four PolyA+ RNA-seq datasets and 21 total RNA-seq datasets). By this step 1.4% (2,273) of transcripts was removed from the annotation, resulting in residual 155,134 transcripts. 107 transcripts that fulfilled the criteria were not expressed in any of the stranded samples (RPKM<=0.2), and therefore could not be filtered out and were kept in the annotation and run through the next stage of filtering. Second, transcripts that had exons with >20% reciprocal antisense overlap with exons of an annotated mRNA (RefSeq or GENCODE v19) or lncRNA (GENCODE v19) expressed in any of the 6 pools were removed from the annotation. By this step 2.0% (3,142) of transcripts was removed from the annotation, resulting in residual 151,992 transcripts.

Size cut off: LncRNA transcripts are by definition longer than 200nt. Therefore we removed all the transcripts whose summary exon length was <200nt. By this step 0.02% (37) of transcripts was removed from the annotation, resulting in residual 151,955 transcripts.

Exon length cut off: To further remove potential artifacts from the annotation, we filtered out the transcripts with unusually long exons and with an unusually high exon/intron length ratio. To set the cutoff we checked the properties of the annotated mRNA and lncRNA genes and found that 99.4 % of GENCODE multi-exonic lncRNAs, 99.7% of GENCODE multi-exonic mRNAs and 99.7% RefSeq multi-exonic mRNAs do not have exons longer than 15kb and their exons constitute less than 90% of total gene length (Figure S1C). We removed all the transcripts not fulfilling these criteria. By this step

3.9% (5,939) of transcripts was removed from the annotation, resulting in residual 146,016 transcripts.

Repeat coverage cut off: lncRNAs are rich in repeat elements [9] and we allowed reads that mapped in several locations in the genome to be aligned (see *RNA sequencing read alignment* above). Therefore, we allowed some repeat elements to be mapped and potentially assembled as transcripts. It was necessary then to remove the transcripts assembled mainly from repeat regions and thus potentially being artifacts. Using the following command containing the custom script - *bed12ToBed6 -i annotation.bed | coverageBed -b stdin -a RepeatMaskUCSC.bed | perl repeat_coveage.pl > coverage_of_genes* - we performed a control check of annotated multi-exonic lncRNA and mRNA genes (lncRNAs by GENCODE v19, mRNAs by GENCODE v19 and RefSeq – see above Public gene annotations used in the study) and found that repeat coverage of exons of 95,5% of GENCODE multi-exonic lncRNAs, 99,87% of GENCODE multi-exonic mRNAs and 99,96% RefSeq multi-exonic mRNAs did not exceed 80% (Figure S1D). By that we set the cutoff to 80% for the filtering and, using the command given above, removed all the *de novo* transcripts whose exons were covered by repeats more than 80%. By this step 0.7% (1,029) of transcripts was removed from the annotation, resulting in residual 144,987 transcripts.

11.1.1. Filtering for mRNAs - Creating *de novo* mRNA annotation in primary granulocytes

Overlap with annotated protein-coding genes: As the protein-coding genome is very well annotated we defined *de novo* mRNAs as transcripts that overlapped exons of protein-coding genes annotated by RefSeq or GENCODE v19 in the sense orientation (we used intersectBed tool with the –split option). This filtering step resulted in 136,482 transcripts being called protein-coding based on their overlap with the annotation.

Filtering out transcripts spanning from mRNAs to annotated lncRNAs: Some *de novo* transcripts spanned over more than one gene, which can be caused by Cufflinks artificially joining spliced transcripts located close to each other. However it is also known that transcription from a gene can run through a downstream gene and use its splice sites to create a chimera transcript [10]. We aimed to remove such chimera transcripts. While it was possible to remove protein-coding transcripts that span more than one protein-coding gene, due to the poor annotation of lncRNAs, we could not exclude transcripts that comprise a chimera of two lncRNA genes merged together by Cufflinks. As our goal was to process *de novo* mRNA transcripts similarly to *de novo* lncRNA transcripts, we did not apply this filtering to *de novo* mRNAs or lncRNAs. However, we could exclude the case when an artifact chimeric transcript combines an mRNA with a lncRNA. D*e novo* lncRNAs were filtered not to share a sense exonic overlap with annotated protein-coding genes (see below) and we similarly removed *de novo* mRNA transcripts that spanned to a GENCODE v19 annotated lncRNA (note, that some annotated lncRNAs do have a sense exonic overlap with annotated mRNAs and we took care of such cases). By this step 3.7% (5,059) of *de novo* mRNA transcripts were removed from the annotation, resulting in residual 131,423 transcripts.

11.1.2. Filtering for lncRNAs - Creating *de novo* lncRNA annotation in granulocytes

Filtering out transcripts: protein-coding genes and pseudogenes: To form a preliminary *de novo* lncRNA annotation the transcripts that passed all the common filtering steps, but had any exonic sense overlap with a protein-coding gene (GENCODE or RefSeq) or a pseudogene (GENCODE) were removed. By this filtering step 94.6% (139,080) of all lncRNA transcripts were removed from the annotation, resulting in residual 7,862 transcripts.

Filtering out extensions of protein-coding genes: While creating the *de novo* mRNA set we also identified 3' or 5' extensions of annotated protein-coding genes. The previous lncRNA filtering step, which excluded all transcripts overlapping the exons of annotated protein-coding genes could leave in transcripts corresponding to these extensions. To exclude this possibility, we removed transcripts that had any exonic sense overlap with the *de novo* annotated mRNAs. By this filtering step 6.0% (476) of lncRNA transcripts were removed from the annotation, resulting in residual 7,386 lncRNA transcripts.

Removing transcripts that overlap protein-coding genes in the sense direction: Out of 7,386 lncRNA transcripts 317 overlapped annotated or *de novo* assembled mRNA genes in the sense direction. We removed these transcripts from the list of *de novo* lncRNAs to avoid confusing their expression with the expression of overlapped protein-coding gene during the expression variation analysis. After removing sense overlapping transcripts we obtained a list of 7,069 *de novo* lncRNA transcripts.

11.2. Combining de novo lncRNA and mRNA transcripts into genomic loci

Transcripts were initially grouped into loci by Cuffmerge. However, after the filtering and artifact removal, many transcripts were removed and the rest had to be slightly regrouped. For example, if a *de novo* transcript spanned both a protein-coding gene and a lncRNA gene, two separate loci would be first grouped into one by Cuffmerge. After the removal of the artifact spanning transcript, the transcripts corresponding to a protein-coding gene would form one locus, and transcripts corresponding to a lncRNA gene would form another. We redefined the locus definition to account for removal of some transcripts using a custom script. 131,423 *de novo* mRNA transcripts formed 10,029 genomic loci with a mean of 13.1 transcripts per locus (median – 10 transcripts per locus). 7,069 *de novo* lncRNA transcripts formed 1,691 genomic loci with a mean of 3.9 transcripts per locus (median – 1 transcript per locus).

11.3. Protein-coding potential calculation pipeline

We based our mRNA *de novo* annotation on filtering for transcripts exonically overlapping annotated protein-coding genes. On the other hand, we based the *de novo* annotation of lncRNAs filtering for transcripts that had no exonic overlap with annotated protein-coding genes. However, although we combined both RefSeq and GENCODE v19 public annotations for protein-coding genes, a possibility remained that within the lncRNA list there were unknown transcripts coding for proteins. To test the coding potential of transcripts remaining in the lncRNA list, we performed an estimation of protein-coding potential of each *de novo* annotated transcript. We used a combination of two previously developed tools: RNAcode [11] and Coding Potential Calculator or CPC [12]. We used a local version of CPC (cpc-0.0-r2) that was modified to work with HMMER 3.0 [13] instead of blastx using UniProtKB/Swiss-Prot database (Jan 2012) (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.fasta). cDNA of the transcript for HMMER input was created using the getfasta tool from the bedtools suite: *bedtools getfasta -bed [bed12 file] -fi mm10.fa -s -split -fo [cDNA FASTA file]*. Potential peptides originating from this transcript were determined using the transeq tool from EMBOSS6.5.7: *transeq -frame F [cDNA FASTA file] [translated protein sequence]*. The result of this program is a continuously translated protein sequence for each of the three forward frames. As this sequence contains many stop codons we reduced the runtime of HMMER by extracting those peptide sequences that are between two stop codons using a custom script. The first peptide sequence (before the first

stop) as well as the last peptide sequence (after the last stop) was retained if it was longer than 20 amino acids. All other sequences were retained if they contained a start codon (M) that was located more than 30 amino acids from the next stop. HMMER was used with the following command: *phmmer -E 0.0000000001 --cpu 3 –tblout [output file with alignments]*.

The exact criteria for calling a gene non-protein-coding were determined by analyzing a set of well-studied lncRNAs (H19, XIST, JPX, MALAT, NEAT1, TUSC7, ANRIL, MIAT, HULC, HOTTIP, HOTAIR and HOTAIRM1 - see Additional File 2D). We controlled for false positive results by then applying the pipeline to public annotations of coding and non-coding multi-exonic transcripts. With the chosen criteria 98.9% of RefSeq and 96.1% of GENCODE multi-exonic protein-coding transcripts were identified as protein-coding. Accordingly, 94.4% of multi-exonic GENCODE lncRNA transcripts were identified as non-protein-coding. The final criteria for calling a transcript non-coding were the following: CPC score of a transcript had to be less than 1.6. RNAcode score had to be less than 18 (in case genome alignments for 3 species were available for the transcript and RNAcode score could be calculated). We discarded all the de novo lncRNA transcripts that were identified by the pipeline as having protein-coding potential. Moreover, if more than 15% of transcripts in one locus were called protein-coding, we discarded the whole locus. Thus, we obtained the final de novo lncRNA annotation consisting of 6,249 lncRNA transcripts. We lastly fine-tuned the loci definition (for loci where some isoforms were removed by the protein-coding potential filtering) and obtained the final annotation of 1,591 lncRNA loci.

12. Calculating exonic coverage

Exonic coverage of one (“reference”) multiexonic annotation by another (“analyzed”) was calculated using a custom Perl script. For each transcript of the reference annotation we looked for transcripts of the analyzed annotation which would exonically overlap it in the sense orientation. From these transcripts the one that covered the highest percentage of the exonic length of the reference transcript was picked and the exonic coverage given by this transcript was used as an output of the analysis for a given reference transcript.

13. Creating granulocyte specificity estimation heat maps

Each heat map was created from a table listing all the transcripts/loci for each annotation and corresponding RPKMs (calculated by RPKM_count.py) in 36 (GRA_pap, GRA_tot and 34 public RNA-seq samples) samples. Prior to building the heat map, an expression cut off was applied filtering for transcripts expressed (RPKM>0.2) in at least one sample, RPKMs were normalized to the maximum RPKM among all the samples for each transcript (row) and the maximum was set to one. The data table was then sorted using a custom script to organize the columns such that the transcripts showing >70% expression level relative to the maximum would be place on the top and then the rest of the table would be sorted the same way for the next column. This procedure was done consequently for all the columns. We then picked the transcripts/loci that fulfill “granulocyte specificity” criteria and placed them in the upper part of the table, and the rest of the transcripts/loci, to the lower part of the table. Then we used *pheatmap* function in R without the clustering option for rows or columns, to create the final heat maps.

14. RT-PCR to test splicing efficiency calculation

To test bioinformatic calculation of splicing efficiency we picked efficiently spliced (defined as mean splicing efficiency per transcript>80%, maximal splicing efficiency per transcript>90%) and inefficiently spliced (defined as mean splicing efficiency per transcript<50%, maximal splicing efficiency per transcript<70%) *de novo* granulocyte transcripts, eight each (4 x lncRNAs, 4 x mRNAs). We preliminary filtered lncRNA and mRNA transcripts to contain at least one junction which could be tested using our assay, i.e. which would be short enough to allow amplification of both spliced and unspliced products by standard PCR (junction length<1500bp). We additionally filtered the transcripts for relatively high expression (RPKM>1) to facilitate RT-PCR amplification. We also did not pick transcripts that were antisense to another gene since RT-PCR is not strand-specific. List of picked lncRNA and mRNA transcripts and primers to amplify the short junction are given in Figure S13B. Size of the expected PCR product amplifying the unspliced junction is given in the right-most column of the table in Figure S13B. RT-PCR program: 95º 3min, (95º 30sec, 59º 30sec, 72º 1 min) for 35 cycles, 72º 7min. Only the 8x2 randomly picked transcripts (one junction each) were tested. Only one primer pair per junction was tested.

SUPPLEMENTAL REFERENCES

1. Smit AFA: **RepeatMasker**. *http://wwwrepeatmaskerorg* 1996-2005.
2. Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA, Gonzalez-Porta M, Kurbatova N, Griebel T, Ferreira PG *et al*: **Transcriptome and genome sequencing uncovers functional variation in humans**. *Nature* 2013, **501**(7468):506-511.
3. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL: **Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses**. *Genes & development* 2011, **25**(18):1915-1927.
4. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S *et al*: **GENCODE: the reference human genome annotation for The ENCODE Project**. *Genome research* 2012, **22**(9):1760-1774.
5. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM *et al*: **RefSeq: an update on mammalian reference sequences**. *Nucleic Acids Res* 2014, **42**(Database issue):D756-763.
6. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner**. *Bioinformatics* 2013, **29**(1):15-21.
7. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks**. *Nat Protoc* 2012, **7**(3):562-578.
8. Necsulea A, Kaessmann H: **Evolutionary dynamics of coding and non-coding transcriptomes**. *Nat Rev Genet* 2014, **15**(11):734-748.
9. Johnson R, Guigo R: **The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs**. *Rna* 2014, **20**(7):959-976.
10. Gingeras TR: **Implications of chimaeric non-co-linear transcripts**. *Nature* 2009, **461**(7261):206-211.
11. Washietl S, Findeiss S, Muller SA, Kalkhof S, von Bergen M, Hofacker IL, Stadler PF, Goldman N: **RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data**. *Rna* 2011, **17**(4):578-594.
12. Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, Gao G: **CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W345-349.
13. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M: **Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions**. *Nucleic Acids Res* 2013, **41**(12):e121.

Publication 2 - Additional File 2A-1

Human granulocyte samples sequenced in this study

Donor Number	ENA sample assession number	Sample name	Gender	Age	Date of collection	Gap between collections, weeks	PolyA+ RNAseq, library prep CC	PolyA+ RNAseq, library prep AK	Total RNAseq, library prep AK	CD19 (Bcell marker) MNC/GRA ratio, qRT-PCR*	CD19 (Bcell marker) RPKM** (RefSeq NM_001770)	CD28 (Tcell marker) RPKM** (RefSeq NM_006139)
1	ERS679730	**D1-1**	M	33	15.06.2012		nd	nd	ss^{100}	72.8	0.02	0.16
	ERS679731	**D1-2**		33	07.09.2012	12.0	$ns^{50}ns^{100}$	ss^{100}	ss^{100}	158.0	0.01	0.07
	ERS679732	**D1-3**		33	30.11.2012	12.0	nd	nd	ss^{100}	39.0	0.03	0.09
2	ERS679733	**D2-1**	F	62	19.06.2012		ns^{100}	nd	ss^{100}	214.9	0.00	0.05
	ERS679734	**D2-2**		62	11.09.2012	12.0	ns^{50}	ss^{100}	ss^{100}	200.7	0.02	0.07
	ERS679735	**D2-3**		62	29.11.2012	11.3	nd	nd	ss^{100}	113.5	0.03	0.14
3	ERS679736	**D3-1**	F	43	25.06.2012		nd	ss^{100}	ss^{100}	300.9	0.00	0.18
	ERS679737	**D3-2**		44	08.10.2012	15.0	ns^{100}	nd	ss^{100}	259.8	0.01	0.55
	ERS679738	**D3-3**		44	07.12.2012	8.6	ns^{50}	nd	ss^{100}	79.6	0.00	0.40
4	ERS679835	**D4-1**	M	46	03.07.2012		nd	nd	ss^{100}	55.3	0.02	0.10
	ERS679836	**D4-2**		47	28.09.2012	12.4	nd	nd	ss^{100}	479.6	0.02	0.13
	ERS679837	**D4-3**		47	06.12.2012	9.9	ns^{100}	nd	ss^{100}	171.1	0.02	0.21
5	ERS679838	**D5-1**	M	37	15.06.2012		nd	nd	ss^{100}	35.9	0.00	0.25
	ERS679839	**D5-2**		37	07.09.2012	12.0	ns^{100}	nd	ss^{100}	77.1	0.03	0.09
	ERS679840	**D5-3**		38	01.02.2013	21.0	nd	nd	ss^{100}	166.0	0.01	0.65
6	ERS679841	**D6-1**	M	31	25.06.2012		nd	ss^{100}	ss^{100}	59.4	0.02	0.10
	ERS679842	**D6-2**		31	21.09.2012	12.6	ns^{100}	nd	ss^{100}	84.0	0.01	0.48
	ERS679843	**D6-3**		31	07.12.2012	11.0	nd	nd	ss^{100}	111.7	0.05	0.25

Publication 2 - Additional File 2A-2

Human granulocyte samples sequenced in this study

Donor Number	ENA sample assession number	Sample name	Gender	Age	Date of collection	Gap between collections, weeks	PolyA+ RNAseq, library prep CC	PolyA+ RNAseq, library prep AK	Total RNAseq, library prep AK	CD19 (Bcell marker) MNC/GRA ratio, qRT-PCR*	CD19 (Bcell marker) RPKM** (RefSeq NM_001770)	CD28 (Tcell marker) RPKM** (RefSeq NM_006139)
7	ERS679844	**D7-1**	F	28	01.02.2013		nd	nd	ss^{100}	79.0	0.02	0.50
	ERS679845	**D7-2**		29	07.06.2013	18.0	nd	nd	ss^{100}	59.0	0.02	0.27
	ERS679846	**D7-3**		29	11.07.2013	4.9	ns^{50}	nd	ss^{100}	207.1	0.05	0.33
8	ERS684037	**D8**	M	42	19.06.2012	-	ns^{100}	nd	nd	nd	0.07	0.04
9	ERS684038	**D9**	F	43	06.07.2012	-	ns^{50}	nd	nd	55.9	0.09	0.55
10	ERS684039	**D10**	F	27	16.05.2013	-	ns^{50}	nd	nd	313.5	0.04	0.11

ns^{50}	non strand-specific 50bp paired-end RNA sequencing
ns^{100}	non strand-specific 100bp paired-end RNA sequencing
ss^{100}	strand-specific 100bp paired-end RNA sequencing
nd	not sequenced/not done
MNC	mononuclear cells fraction
GRA	granulocyte fraction
*	Ratio of the normalized (to TBP gene) qPCR signal of CD19 gene expression in Mononuclear cells (MNC) and granulocytes (GRA) isolated from the same blood sample
**	RPKM was calculated from RiboZero total RNA-seq samples ss100 for donors D1-D7, and for from available PolyA+ RNA-seq samples for donors D8-D10
	Sequenced samples

exon spanning qPCR primers

CD19_F	CCCACCAGGAGATTCTTCAA	Forward primer itself spans 2 exons and cannot be mapped using UCSC InSilico PCR tool
CD19_R	TGCTCGGGTTTCCATAAGAC	
TBP_F	ACAACAGCCTGCCACCTTAC	Housekeeping gene control for qRT-PCR
TBP_R	GCCATAAGGCATCATTGGAC	

Publication 2 Additional File 2B-1 List of human granulocyte RNA-seq datasets produced in the study

	RNA-seq sample name	RNA-seq sample name	read type	RNA fraction	Library prep AK/CC	Strand specific	number of input reads	uniquely mapped reads, %	multiple mapped reads, %	unmapped reads, %	uniquely mapped reads	total mapped reads	number of spliced reads, mln	Proper strand, inferexperiment.py "1+-,1-+,2++,2--", %	Strand leakage, inferexperiment.py "1++,1--,2+-,2-+", %
PolyA+ RNA-seq											756,609,485	784,337,705			
1	**Donor1_tp2_paRNA_100ns**	**D1-2_pa_100ns**	100bp PE	PolyA+	CC	No	38,433,109	94.6	2.9	2.5	36,342,348	37,449,221	14.2	ns	ns
2	**Donor1_tp2_paRNA_100ss**	**D1-2_pa_100ss**	100bp PE	PolyA+	AK	Yes	42,653,238	94.34	2.7	2.9	40,239,065	41,382,172	16.0	88.5	11.5
3	**Donor1_tp2_paRNA_50ns**	**D1-2_pa_50ns**	50bp PE	PolyA+	CC	No	65,588,108	92.25	3.7	3.8	60,505,030	62,905,554	7.4	ns	ns
4	**Donor2_tp1_paRNA_100ns**	**D2-1_pa_100ns**	100bp PE	PolyA+	CC	No	29,215,229	94.04	3.1	2.8	27,474,001	28,382,595	10.5	ns	ns
5	**Donor2_tp2_paRNA_100ss**	**D2-2_pa_100ss**	100bp PE	PolyA+	AK	Yes	45,285,047	94.14	2.8	3.0	42,631,343	43,903,853	16.9	88.7	11.3
6	**Donor2_tp2_paRNA_50ns**	**D2-2_pa_50ns**	50bp PE	PolyA+	CC	No	65,122,658	92.12	3.9	3.7	59,990,993	62,498,215	7.4	ns	ns
7	**Donor3_tp1_paRNA_100ss**	**D3-1_pa_100ss**	100bp PE	PolyA+	AK	Yes	41,452,428	94.22	2.8	3.0	39,056,478	40,200,565	15.6	88.4	11.6
8	**Donor3_tp2_paRNA_100ns**	**D3-2_pa_100ns**	100bp PE	PolyA+	CC	No	21,542,366	94.1	3.0	2.8	20,273,521	20,921,946	7.8	ns	ns
9	**Donor3_tp3_paRNA_50ns**	**D3-3_pa_50ns**	50bp PE	PolyA+	CC	No	72,563,423	92.22	3.9	3.7	66,917,989	69,711,680	8.4	ns	ns
10	**Donor4_tp3_paRNA_100ns**	**D4-3_pa_100ns**	100bp PE	PolyA+	CC	No	24,271,235	94.2	3.0	2.7	22,868,358	23,596,495	9.1	ns	ns
11	**Donor5_tp2_paRNA_100ns**	**D5-2_pa_100ns**	100bp PE	PolyA+	CC	No	27,670,669	94.4	2.8	2.8	26,110,043	26,882,055	10.2	ns	ns
12	**Donor6_tp1_paRNA_100ss**	**D6-1_pa_100ss**	100bp PE	PolyA+	AK	Yes	42,628,962	94.42	2.6	3.0	40,250,266	41,341,567	16.1	86.9	13.2
13	**Donor6_tp2_paRNA_100ns**	**D6-2_pa_100ns**	100bp PE	PolyA+	CC	No	29,981,079	94.44	2.7	2.8	28,314,131	29,123,620	11.2	ns	ns
14	**Donor7_tp3_paRNA_50ns**	**D7-3_pa_50ns**	50bp PE	PolyA+	CC	No	84,458,555	91.54	4.1	4.1	77,313,361	80,767,716	9.5	ns	ns
15	**Donor8_paRNA_100ns**	**D8_pa_100ns**	100bp PE	PolyA+	CC	No	27,139,813	94.29	3.0	2.7	25,590,130	26,390,754	10.1	ns	ns
16	**Donor9_paRNA_50ns**	**D9_pa_50ns**	50bp PE	PolyA+	CC	No	64,668,631	91.49	4.1	4.2	59,165,331	61,784,410	7.3	ns	ns
17	**Donor10_paRNA_50ns**	**D10_pa_50ns**	50bp PE	PolyA+	CC	No	90,932,644	91.9	3.9	3.9	83,567,100	87,095,286	9.9	ns	ns

ns non strand specific library

Raw granulocyte RNA-seq data were submitted to the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under study accession number PRJEB8740.

Raw granulocyte RNA-seq data were submitted to the GEO under study accession number GSE70390

Publication 2 Additional File 2B-2 List of human granulocyte RNA-seq datasets produced in the study

	RNA-seq sample name	RNA-seq sample name	read type	RNA fraction	Library prep AK/CC	Strand specific	number of input reads	uniquely mapped reads, %	multiple mapped reads, %	unmapped reads, %	uniquely mapped reads	total mapped reads	number of spliced reads, mln	Proper strand, inferexperiment.py "1+-,1-+,2++,2--", %	Strand leakage, inferexperiment.py "1++,1--,2+-,2-+", %
Total RNA-seq											1,111,729,702	1,213,423,977			
1	**Donor1_tp1_totalRNA_100ss**	**D1-1_tot_100ss**	100bp PE	total	AK	Yes	70,036,177	88.75	8.9	2.4	62,157,107	68,355,309	8.6	96.1	3.9
2	**Donor1_tp2_totalRNA_100ss**	**D1-2_tot_100ss**	100bp PE	total	AK	Yes	74,774,544	87.32	10.1	2.6	65,293,132	72,815,451	9.0	95.4	4.6
3	**Donor1_tp3_totalRNA_100ss**	**D1-3_tot_100ss**	100bp PE	total	AK	Yes	68,473,162	93.48	4.0	2.5	64,008,712	66,720,249	7.4	96.8	3.2
4	**Donor2_tp1_totalRNA_100ss**	**D2-1_tot_100ss**	100bp PE	total	AK	Yes	65,073,786	91.85	5.8	2.3	59,770,272	63,570,582	7.0	96.4	3.6
5	**Donor2_tp2_totalRNA_100ss**	**D2-2_tot_100ss**	100bp PE	total	AK	Yes	68,319,487	89.98	7.5	2.5	61,473,874	66,625,164	8.0	96.3	3.7
6	**Donor2_tp3_totalRNA_100ss**	**D2-3_tot_100ss**	100bp PE	total	AK	Yes	67,358,940	87.41	10.1	2.4	58,878,449	65,695,174	7.8	95.9	4.1
7	**Donor3_tp1_totalRNA_100ss**	**D3-1_tot_100ss**	100bp PE	total	AK	Yes	72,517,400	88.73	8.8	2.5	64,344,689	70,689,962	8.8	96.0	4.0
8	**Donor3_tp2_totalRNA_100ss**	**D3-2_tot_100ss**	100bp PE	total	AK	Yes	61,371,161	46.24	3.4	50.3	28,378,025	30,476,919	3.5	93.8	6.2
9	**Donor3_tp3_totalRNA_100ss**	**D3-3_tot_100ss**	100bp PE	total	AK	Yes	79,188,820	46.63	3.2	50.1	36,925,747	39,491,465	4.5	96.0	4.0
10	**Donor4_tp1_totalRNA_100ss**	**D4-1_tot_100ss**	100bp PE	total	AK	Yes	69,092,551	88.0	9.3	2.7	60,794,536	67,233,961	7.4	95.7	4.3
11	**Donor4_tp2_totalRNA_100ss**	**D4-2_tot_100ss**	100bp PE	total	AK	Yes	69,906,326	90.7	6.8	2.5	63,405,038	68,158,668	7.7	96.1	3.9
12	**Donor4_tp3_totalRNA_100ss**	**D4-3_tot_100ss**	100bp PE	total	AK	Yes	68,186,023	89.7	7.6	2.7	61,162,863	66,324,545	7.6	95.7	4.3
13	**Donor5_tp1_totalRNA_100ss**	**D5-1_tot_100ss**	100bp PE	total	AK	Yes	22,360,731	91.4	6.2	2.4	20,437,708	21,821,837	2.6	96.4	3.6
14	**Donor5_tp2_totalRNA_100ss**	**D5-2_tot_100ss**	100bp PE	total	AK	Yes	46,128,806	86.2	11.2	2.6	39,772,257	44,915,618	5.3	95.8	4.2
15	**Donor5_tp3_totalRNA_100ss**	**D5-3_tot_100ss**	100bp PE	total	AK	Yes	50,547,283	87.1	10.6	2.3	44,011,519	49,359,422	5.8	96.2	3.8
16	**Donor6_tp1_totalRNA_100ss**	**D6-1_tot_100ss**	100bp PE	total	AK	Yes	54,445,042	87.1	10.3	2.7	47,394,409	52,996,804	6.0	95.1	4.9
17	**Donor6_tp2_totalRNA_100ss**	**D6-2_tot_100ss**	100bp PE	total	AK	Yes	58,601,271	86.0	10.4	3.5	50,420,534	56,532,646	6.5	95.0	5.0
18	**Donor6_tp3_totalRNA_100ss**	**D6-3_tot_100ss**	100bp PE	total	AK	Yes	55,286,129	84.3	12.2	3.5	46,617,264	53,362,172	5.8	93.7	6.3
19	**Donor7_tp1_totalRNA_100ss**	**D7-1_tot_100ss**	100bp PE	total	AK	Yes	60,139,278	88.4	9.2	2.4	53,157,108	58,671,880	7.1	96.5	3.5
20	**Donor7_tp2_totalRNA_100ss**	**D7-2_tot_100ss**	100bp PE	total	AK	Yes	76,082,579	93.5	4.1	2.3	71,144,820	74,233,772	8.7	97.5	2.5
21	**Donor7_tp3_totalRNA_100ss**	**D7-3_tot_100ss**	100bp PE	total	AK	Yes	56,774,715	91.9	5.6	2.5	52,181,641	55,372,380	6.1	96.6	3.4

ns non strand specific library

Raw granulocyte RNA-seq data were submitted to the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under study accession number PRJEB8740.

Raw granulocyte RNA-seq data were submitted to the GEO under study accession number GSE70390

Publication 2 Additional File 2C - Pools used for human granulocyte transcriptome assembly											
Pool number	RNA-seq sample names pooled	read type	RNA fraction	Strand specific	number of input reads	uniquely mapped reads,%	multiple mapped reads, %	unmapped reads, %	uniquely mapped reads	total mapped reads	number of spliced reads, M
1	D6-3_pa_100ns,D3-1_pa_100ns,D8_pa_100ns,D4-2_pa_100ns	100bp PE	PolyA+	No	102,168,643	94.2	3.0	2.6	96,201,217	99,286,710	37,494,688
2	D1-2_pa_100ns, D2-2_pa_100ns,D5-2_pa_100ns	100bp PE	PolyA+	No	96,084,857	94.5	2.8	2.6	90,764,983	93,455,359	35,536,074
3	D10_pa_50ns,D9_pa_50ns,D3-2_pa_50ns	50bp PE	PolyA+	No	220,723,933	91.8	3.9	4.0	202,713,469	211,365,847	24,667,365
4	D2-2_pa_50ns,D4-3_pa_50ns,D7-3_pa_50ns	50bp PE	PolyA+	No	222,610,086	92.0	3.9	3.9	204,731,073	213,390,605	25,351,256
5	D5-1_pa_100ss, D2-2_pa_100ss	100bp PE	PolyA+	Yes	85,282,200	94.4	2.6	2.9	80,486,999	82,721,393	32,019,277
6	D3-2_pa_100ss, D4-1_pa_100ss	100bp PE	PolyA+	Yes	86,737,475	94.2	2.8	2.9	81,684,495	84,104,471	32,543,755
							sum		756,582,236	784,324,385	187,612,415

Publication 2 Additional File 2D - Well-known lncRNAs used to adjust RNAcode & CPC pipeline output

lncRNA gene	GENCODE ID of isoform	RNAcode output	CPC score	Outcome of protein coding potential estimation with set criteria
XIST	ENST00000429829.1 (main isoform)	10.17	-0.95	non-protein-coding
XIST	ENST00000602863.1	ns	-1.18	non-protein-coding
XIST	ENST00000416330.1	ns	-1.19	non-protein-coding
XIST	ENST00000602587.1	ns	-1.13	non-protein-coding
XIST	ENST00000602495.1	ns	-1.33	non-protein-coding
XIST	ENST00000445814.1	ns	-1.18	non-protein-coding
XIST	ENST00000434839.1	ns	-1.26	non-protein-coding
XIST	ENST00000433732.1	ns	-1.03	non-protein-coding
XIST	ENST00000417942.1	ns	-1.27	non-protein-coding
XIST	ENST00000421322.1	ns	-1.29	non-protein-coding
LOC100288798	ENST00000607353.1	ns	-0.48	non-protein-coding
MALAT1 (most isoforms single exon)	ENST00000508832.1 (multiexonic isoform)	ns	-1.11	non-protein-coding
NEAT1 (most isoforms single exon)	ENST00000499732.1 (multiexonic isoform)	ns	-0.81	non-protein-coding
HULC	ENST00000503668.1	ns	-0.75	non-protein-coding
TUSC7	ENST00000477805.1	3.25	-0.81	non-protein-coding
ANRIL	ENST00000580576.1	ns	-0.97	non-protein-coding
ANRIL	ENST00000428597.1	ns	-0.97	non-protein-coding
ANRIL	ENST00000421632.1	ns	-1.13	non-protein-coding
MIAT	ENST00000423278.1	4.89	-0.56	non-protein-coding
MIAT	ENST00000458302.1	ns	-0.98	non-protein-coding
MIAT	ENST00000456129.1	4.89	-0.41	non-protein-coding
MIAT	ENST00000455640.1	ns	-1.12	non-protein-coding
MIAT	ENST00000453023.1	ns	-1.16	non-protein-coding
MIAT	ENST00000452429.1	ns	-1.10	non-protein-coding
MIAT	ENST00000451141.1	ns	-0.99	non-protein-coding
MIAT	ENST00000450203.1	12.07	-1.11	non-protein-coding
MIAT	ENST00000449717.1	ns	-1.02	non-protein-coding
MIAT	ENST00000440347.1	12.07	-0.23	non-protein-coding
MIAT	ENST00000439738.1	ns	-1.07	non-protein-coding
MIAT	ENST00000436238.1	ns	-1.02	non-protein-coding
H19	ENST00000447298.1	ns	-0.77	non-protein-coding
H19	ENST00000446406.1	ns	-0.56	non-protein-coding
H19	ENST00000442037.1	ns	-0.75	non-protein-coding
H19	ENST00000439725.1	ns	-0.80	non-protein-coding
H19	ENST00000436715.1	ns	-0.75	non-protein-coding
H19	ENST00000431095.1	ns	-0.56	non-protein-coding
H19	ENST00000428066.1	ns	-0.36	non-protein-coding
H19	ENST00000422826.1	ns	-0.88	non-protein-coding
H19	ENST00000417089.1	ns	-0.63	non-protein-coding
H19	ENST00000414790.1	ns	-0.85	non-protein-coding
H19	ENST00000411861.1	ns	-0.78	non-protein-coding
H19	ENST00000411754.1	ns	-0.58	non-protein-coding

Publication 2 Additional File 2F

Validation of *de novo* granulocyte lncRNA transcripts not supported by public annotations by means of cloning and Sanger Sequencing

transcript number	Supplemental Fig.	targeted lncRNA transcript (GRA_denovo_LNC.)	type of lncRNA in relation to public annotations	locus (GRA_denovo_LNC.)	granulocyte cDNA sample used for PCR amplification	forward primer	reverse primer	genomic position of the product (In Silico PCR UCSC)	genomic span (In Silico PCR UCSC), bp	expected product length, bp
1	4A	**84.1**	not in PA	84	D7-3	TATGCCCACCTGTCCAAACC	ATTCCCTGTGGTGATGGCAG	chr5:57803491+57807383	3,893	276
2	4B	**117.4**	not in PA	117	D7-3	AGGATCCTCATGGCCTCTAGT	TGCCAATGACAAAAGCCGTG	chr1:182118924+182173727	54,804	821
3	4C	**152.4**	not in PA	152	D7-3	TAAACAAGCCACGGTGCAGA	AAAGCAGCGTTTCAGTGCAG	chr1:186584602+186595696	11,095	424
4	4D	**187.1**	not in PA	187	D3-1	CCAGCTAGCCCAGACTAGGA	ATGGCACCCAAATTCCCACC	chr6:30464776+30467666	2,891	294
5	4E	**294.12**	not in PA	294	D7-3	AGCAGATTCCAGAAGCAGCA	AGTGAGCCGGGGAGATTTTG	chr7:15276811+15390924	114,114	348
6	5A	**308.3**	not in PA	308	D3-1	GTCTGCCTGAAAGCCTGACT	ACTTTCTCTACTGGGCAGCC	chr7:30287566+30325701	38,136	541
7	5B	**397.1**	not in PA	397	D7-3	ATGCCCACCCAATTCACTGT	TGAGGTAGGGTGGGGTCAAA	chr8:37757836+37758599	764	256
8	5C	**421.3**	not in PA	421	D3-1	GACAGACCCTGAGGAACACG	ACTGTAGCCAGAGCAGACTT	chr8:117664317+117665856	1,540	452
9	5D	**639.3**	not in PA	639	D7-3	GTCAGGAGGTCATGTGGTGG	CAGCCAGGATCGTCAGGTTT	chr10:129855996+129860689	4,694	1,210
10	5E	**1060.1**	not in PA	1060	D3-1	AAACAGGAAGGCCCTCATCG	GCTGCAGTCAGGAAAGCTTG	chr1:85039788+85048776	8,989	671
11	6A	**772.3**	not in PA	772	D3-1	ACATGGGTCTGCAGCTACTG	TCTCAGAGAGTTGACACTGTGT	chr12:89821752+89849600	27,849	378
12	6A	**772.4**	not in PA	772	D3-1	ACATGGGTCTGCAGCTACTG	GGCATGGGCAGTTAGGTTCT	chr12:89821752+89866557	44,806	477
13	6A	**772.7**	not in PA	772	D3-1	ACATGGGTCTGCAGCTACTG	TGTCTCTCAGCAGTTGCAGA	chr12:89821752+89871097	49,346	411
14	6A	**772.13**	not in PA	772	D7-3	ACATGGGTCTGCAGCTACTG	AAAATGCCCGACAGCTCTCA	chr12:89821752+89845932	24,181	435
15	6B	**944.2**	not in PA	944	D3-1	ACACAAACTGCGATGGTTGC	AGCCTGAGCCAATTCCAGAC	chr15:91946403+91952107	5,705	587
16	6B	**944.3**	not in PA	944	D3-1	ACACAAACTGCGATGGTTGC	AGGATGGAGGGCTTAGCTGA	chr15:91946403+91955749	9,347	569
17	6B	**944.8**	not in PA	944	D7-3	ACACAAACTGCGATGGTTGC	AGGCTTTTCATCGGTGGAGT	chr15:91946403+92012935	66,533	624
18	6B	**944.1**	not in PA	944	D7-3	TACCGCTGCAGCTATGTCTG	AGGATGGAGGGCTTAGCTGA	chr15:91952073+91955749	3,677	204
19	7A	**1090.1**	not in PA	1090	D3-1	TTCGAGTGGAAACCCTCACG	AGAAAGCGGTTAGTGGCTGC	chr17:78121426+78122196	771	291
20	7B	**1036.1**	not in PA	1036	D3-1	GGGTCTCTCAGATGGAGGGT	CACTCTTGCCCTGTCCTGAG	chr17:38473655+38478648	4,994	1,104
21	7B	**1036.2**	not in PA	1036	D3-1	GGGTCTGAGGGAGGGATTCT	CACTCTTGCCCTGTCCTGAG	chr17:38474279+38478648	4,370	790
22	7C	**1110.3**	not in PA	1110	D3-1	AGAAGCAGGTGGTTTAGGGC	CAAGACCGTGGGAAAGGGAG	chr1:4791664+4796450	4,787	700
23	8A	**1374.11**	not in PA	1374	D3-1, D7-3	AGAACCGGGGCTGAATGAAG	TGCACTGCTTACACAGGAGG	chr20:641441-715207	73,767	1,118
24	8A	**1374.13**	not in PA	1374	D7-3	TGAATTTTGCAAGGCACCCG	CAGGCATTCTCTGGACACGT	chr2:185456502+185461400	4,899	407
25	8A	**1374.2**	not in PA	1374	D3-1	ACATTGCTCAGGATGCAGCT	TGGACCCCAGGTTAGGACTT	chr20:641358+668740	27,383	560
26	8B	**1547.1**	isoform not in PA	1547	D7-3	ACATGCCCCTGAAATGAGGA	GAATGTCTCTTGGACCCCCG	chr3:112304511+112315730	11,220	261
27	8C	**1576.4**	isoform not in PA	1576	D7-3	CTATGTTGTGCTGCATGGCG	CCACATTTTGCCGTTCCGAC	chr3:151877253+151923218	45,966	289

Publication 2 Additional File 2G-1

Validation of de novo granulocyte lncRNA transcripts not supported by public annotations by means of cloning and Sanger Sequencing: sequencing results

number of transcript	transcript (GRA_denovo_LNC.)	gene (GRA_denovo_LNC.)	code in the supplemental fig	Sanger seq result (primer T7) (not fully displayed)	Sanger seq result (primer SP6) (not fully displayed)	Sanger Seq clean from Vector(not fully displayed)	Sanger Seq clean from Vector (SP6) (not fully displayed)	SUBMITTED to genbank	genebank accession number
1	1036.1	**1036**	**1036.1_1**	GGCCGCCATGGCG	-	GGGTCTCTCAGATGGA	AGGGTGATTCAGATCCT	YES	KP992444
	1036.1	**1036**	**1036.1_2**	GCCATGGCGGCCG	GCTCTCCCATATGG	CACTCTTGCCCTGTC	GGGTCTCTCAGATGGA	YES	KP992445
	1036.1	**1036**	**1036.1_3**	GCCGCCATGGCGG	TATGGTCGACCTGC	GGGTCTCTCAGATGG	CACTCTTGCCCTGTCC	YES	KP992446
	1036.1	**1036**	**1036.1_4**	CGCCATGGCGGCC	-	GGGTCTCTCAGATGGA	AGGGTGATTCAGATCCT	YES	KP992447
	1036.1	**1036**	**1036.1_5**	GGCGGCCGCGGGA	-	CACTCTTGCCCTGTCCTGAGCCTCCAGGTTCA		YES	KP992448
	1036.1	**1036**	**1036.1_6**	CGCCATGGCGGCC	GCTCTCCCATATGG	CACTCTTGCCCTGTC	GGGTCTCTCAGATGGA	YES	KP992449
2	1036.2	**1036**	**1036.2_1**	ATGGCGGCCGCGG	-	GGGTCTGAGGGAGGGATTCTCCCGGATTCCC		YES	KP992450
	1036.2	**1036**	**1036.2_2**	ATGGCGGCCGCGG	-	TGGGTCTGAGGGAGGGATTCTCCCGGATTCC		YES	KP992451
	1036.2	**1036**	**1036.2_3**	GCCATGGCGGCCG	-	CACTCTTGCCCTGTCCTGAGCCTCCAGGTTCA		YES	KP992452
	1036.2	**1036**	**1036.2_4**	GCCATGGCGGCCG	-	CACTCTTGCCCTGTCCTGAGCCTCCAGGTTCA		YES	KP992453
	1036.2	**1036**	**1036.2_5**	GCCATGGCGGCCG	-	CACTCTTGCCCTGTCCTGAGCCTCCAGGTTCA		YES	KP992454
3	1060.1	**1060**	**1060.1_2**	TGGCGGCCGCGGG	-	GCTGCAGTCAGGAAAGCTTGGAGGCAGCAAG		YES	KP992442
4	1090.1	**1090**	**1090.1_1**	TGGCGGCCGCGGG	-	TTCGAGTGGAAACCCTCACGGCCGCCGGTGG		YES	KR021976
	1090.1	**1090**	**1090.1_2**	ATGGCGGCCGCGG	-	AGAAAGCGGTTAGTGGCTGCGCGTGGACACC		YES	KR021977
5	1110.3	**1110**	**1110.3_1**	GCCATGGCGGCCG	AGCTCTCCCATATG	CAAGACCGTGGGAAA	AGAAGCAGGTGGTTTA	YES	KR021978
	1110.3	**1110**	**1110.3_2**	CGCCATGGCGGCC	AGCTCTCCCATATG	CAAGACCGTGGGAAA	GGCCAGGAGGCCTCC	YES	KR021979
6	117.4	**117**	**117.4_1**	CCATGGCGGCCGC	-	TGCCAATGACAAAAGCCGTGTACTTTCTGAAT		YES	KR021980
	117.4	**117**	**117.4_2**	CCGCCATGGCGGC	-	AGGATCCTCATGGCCTCTAGTTTGGCCATTCA		YES	KR021981
	117.4	**117**	**117.4_3**	CCGCCATGGCGGC	-	TGCCAATGACAAAAGCCGTGTACTTTCTGAAT		YES	KR021982
	117.4	**117**	**117.4_4**	CGCCATGGCGGCC	-	TGCCAATGACAAAAGCCGTGTACTTTCTGAAT		YES	KR021983
	117.4	**117**	**117.4_6**	CGCCATGGCGGCC	-	AGGATCCTCATGGCCTCTAGTTTGGCCATTCA		YES	KR021980
7	1374.13	**1374**	**1374.13_1**	CCGCCATGGCGGC	-	CAGGTTCCATGTCCACTGGGGGCCGGGATCT		YES	KR021994
	1374.13	**1374**	**1374.13_2**	CGCCATGGCGGCC	-	TCCCTCAAAGCATGCAGGTTATGCTTCTTGAG		YES	KR021995
	1374.13	**1374**	**1374.13_3**	CCGCCATGGCGGC	-	CAGGTTCCATGTCCACTGGGGGCCGGGATCT		YES	KR021996
8	1374.11	**1374**	**1374.11_1**	GGCCGCCATGGCG	-	TGCACTGCTTACACAGGAGGAAACTGACAAAG		YES	KR021991
	1374.11	**1374**	**1374.11_2**	GCCATGGCGGCCG	-	AGAACCGGGGCTGAATGAAGGATGAATGAGG		YES	KR021992
	1374.11	**1374**	**1374.11_3**	CCGCCATGGCGGC	-	AGAACCGGGGCTGAATGAAGGATGAATGAGG		YES	KR021993
9	1374.2	**1374**	**1374.20_1**	GCCATGGCGGCCG	-	ACATTGCTCAGGATGCAGCTGCTTGTCCAGAG		YES	KR021997

Publication 2 Additional File 2G-2

Validation of de novo granulocyte lncRNA transcripts not supported by public annotations by means of cloning and Sanger Sequencing: sequencing results

number of transcript	transcript (GRA_denovo_LNC.)	gene (GRA_denovo_LNC.)	code in the supplemental fig	Sanger seq result (primer T7) (not fully displayed)	Sanger seq result (primer SP6) (not fully displayed)	Sanger Seq clean from Vector(not fully displayed)	Sanger Seq clean from Vector (SP6) (not fully displayed)	SUBMITTED to genbank	genebank accession number
9	1374.2	**1374**	**1374.20_2**	CCGCCATGGCGGC	-	ACATTGCTCAGGATG	-	YES	KR021998
10	152.4	**152**	**152.4_1**	CCGCCATGGCGGC	-	AAAGCAGCGTTTCAG	-	YES	KP992443
11	1547.1	**1547**	**1547.1_2**	ATGGCGGCCGCGG	-	GAATGTCTCTTGGAC	-	YES	KR021999
	1547.1	**1547**	**1547.1_3**	WNGCTCCGGCCGC	-	ACTGCCCCTGAAATG	-	YES	KR022000
	1547.1	**1547**	**1547.1_5**	WNGCTCCGGCCGC	-	ACATGCCCCTGAAATC	-	YES	KR022001
12	1576.4	**1576**	**1576.4_1**	SGNANGCTCCGGC	-	CTATGTTGTGCTGCAT	-	YES	KR024014
	1576.4	**1576**	**1576.4_2**	GCTCCGGGCCGCC	-	CCACATTTTGCCGTTC	-	YES	KR024015
13	84.1	**84**	**84.1_1**	NWNGCTCCGGCCG	-	TATGCCCACCTGTCCA	-	YES	KR024016
14	187.1	**187**	**187.1_1**	ATGGCGGCCGCGG	-	CCAGCTAGCCCAGAC	-	YES	KR024017
	187.1	**187**	**187.1_2**	GGCGGCCGCGGGA	-	CCAGCTAGCCCAGAC	-	YES	KR024018
15	294.12	**294**	**294.12_1**	ATGGCGGCCGCGG	-	AGCAGATTCCAGAAG	-	YES	KR021985
	294.12	**294**	**294.12_2**	GCCATGGCGGCCG	-	AGTGAGCCGGGGAGA	-	YES	KR021986
	294.12	**294**	**294.12_3**	TGGCGGCCGCGGG	-	AGTGAGCCGGGGAGA	-	YES	KR021987
	294.12	**294**	**294.12_4**	GCCATGGCGGCCG	-	AGTGAGCCGGGGAGA	-	YES	KR021988
	294.12	**294**	**294.12_5**	CGCCATGGCGGCC	-	AGTGAGCCGGGGAGA	-	YES	KR021989
	294.12	**294**	**294.12_6**	ATGGCGGCCGCGG	-	AGCAGATTCCAGAAG	-	YES	KR021990
16	308.3	**308**	**308.3_1**	GCCATGGCGGCCG	-	GTCTGCCTGAAAGCC	-	YES	KR024019
	308.3	**308**	**308.3_2**	GCCATGGCGGCCG	-	GTCTGCCTGAAAGCC	-	YES	KR024020
17	397.1	**397**	**397.1_1**	GCCATGGCGGCCG	-	TGAGGTAGGGTGGGC	-	YES	KR137563
	397.1	**397**	**397.1_2**	GCCATGGCGGCCG	-	TGAGGTAGGGTGGGC	-	YES	KR137564
	397.1	**397**	**397.1_3**	TGGCGGCCGCGGG	-	TGAGGTAGGGTGGGC	-	YES	KR137565
18	421.3	**421**	**421.3_1**	GGCCGCCATGGCG	AGCTCTCCCATATG	ACTGTAGCCAGAGCA	GACAGACCCTGAGGAA	YES	KR137566
	421.3	**421**	**421.3_2**	CCGCCATGGCGGC	-	GACAGACCCTGAGGA	-	YES	KR137567
19	639.3	**639**	**639.3_1**	GCCATGGCGGCCG	GAGCTCTCCCATAT	CAGCCAGGATCGTCA	GTCAGGAGGTCATGTG	YES	KR186191
	639.3	**639**	**639.3_3**	CCATGGCGGCCGC	-	CAGCCAGGATCGTCA	-	YES	KR186192
	639.3	**639**	**639.3_4**	CCGCCATGGCGGC	-	CAGCCAGGATCGTCA	-	YES	KR186193
	639.3	**639**	**639.3_6**	ATGGCGGCCGCGG	GCTCTCCCATATGG	CAGCCAGGATCGTCA	GTCAGGAGGTCATGTG	YES	KR186194

Publication 2 Additional File 2G-3

Validation of de novo granulocyte lncRNA transcripts not supported by public annotations by means of cloning and Sanger Sequencing: sequencing results

number of transcript	transcript (GRA_denovo_LNC.)	gene (GRA_denovo_LNC.)	code in the supplemental fig	Sanger seq result (primer T7) (not fully displayed)	Sanger seq result (primer SP6) (not fully displayed)	Sanger Seq clean from Vector(not fully displayed)	Sanger Seq clean from Vector (SP6) (not fully displayed)	SUBMITTED to genbank	genebank accession number
19	639.3	**639**	**639.3_7**	GCCATGGCGGCCG	-	GTCAGGAGGTCATGT	-	YES	KR186195
20	772.13	**772**	**772.13_1**	GCTCCGGCCGCCA	-	ACATGGGTCTGCAGC	-	YES	KR186196
	772.13	**772**	**772.13_3**	GCTCCGGCCGCCA	-	AAAATGCCCGACAGC	-	YES	KR186197
	772.13	**772**	**772.13_4**	GGCGGCCGCGGGA	GCTCTCCCATATGG	ACATGGGTCTGCAGC	AAAATGCCCGACAGCT	YES	KR186198
	772.13	**772**	**772.13_5**	GCCATGGCGGCCG	CCTGCAGGCGGCC	AAAATGCCCGACAGC	ACATGGGTCTGCAGCT	YES	KR186199
21	772.3	**772**	**772.3_1**	GCCATGGCGGCCG	AGCTCTCCCATATG	ACATGGGTCTGCAGC	TCTCAGAGAGTTGACA	YES	KR186200
22	772.4	**772**	**772.4_2**	GCCATGGCGGCCG	TCTCCCATATGGTC	ACATGGGTCTGCAGC	GGCATGGGCAGTTAGG	YES	KR186201
23	772.7	**772**	**772.7_1**	ATGGCGGCCGCGG	-	TGTCTCTCAGCAGTTC	-	YES	KR186202
24	944.1	**944**	**944.10_1**	NGCTCCGGCCGCC	-	ACCGCTGCAGCTATG	-	YES	KR137557
	944.1	**944**	**944.10_2**	GCTCCGGCCGCCA	-	TACCGCTGCAGCTATC	-	YES	KR137558
	944.1	**944**	**944.10_3**	NWNGCTCCGGCCG	-	NAAATTCGATTTACCG	-	YES	KR137559
25	944.2	**944**	**944.2_1**	ATGGCGGCCGCGG	-	AGCCTGAGCCAATTC	-	YES	KR137560
	944.2	**944**	**944.2_2**	GGCCGCCATGGCG	-	AGCCTGAGCCAATTC	-	YES	KR137561
	944.2	**944**	**944.2_3**	CCGCCATGGCGGC	-	AGCCTGAGCCAATTC	-	YES	KR137562
26	944.3	**944**	**944.3_1**	GCCATGGCGGCCG	-	ACACAAACTGCGATG	-	YES	KR137554
	944.3	**944**	**944.3_2**	CCATGGCGGCCGC	-	AGGATGGAGGGCTTA	-	YES	KR137555
	944.3	**944**	**944.3_3**	GGCCGCCATGGCG	-	AGGATGGAGGGCTTA	-	YES	KR137556
27	944.8	**944**	**944.8_1**	GCCATGGCGGCCG	-	AGGCTTTTCATCGGTC	-	YES	KR137551
	944.8	**944**	**944.8_2**	GCCATGGCGGCCG	-	AGGCTTTTCATCGGTC	-	YES	KR137552
	944.8	**944**	**944.8_3**	GCCATGGCGGCCG	-	AGGCTTTTCATCGGTC	-	YES	KR137553

Publication 2 Additional File 2H-1 - Overview of the publicly available RNA-seq datasets used in the study

Data set	type of RNA-seq	cell type	Short label (Figure 2A)	source	library	read length, bp	read type	Number of uniquely mapped reads, mln	download command (not fully displayed)
1	Total	Aortic adventitial fibroblasts	**AAF_tot**	ENCODE	strand spec	100	PE	296.7	wget http://hgdownload.cse.ucsc.edu/goldenP
2	Total	Aortic endothelial cells	**AEndC_tot**	ENCODE	strand spec	100	PE	326.9	wget http://hgdownload.cse.ucsc.edu/goldenP
3	PolyA+	B cells	**Bcell_pap**	ENCODE	strand spec	75	PE	140.4	wget http://hgdownload.cse.ucsc.edu/goldenP
4	Total	Dermal fibroblasts	**DF_tot**	ENCODE	strand spec	100	PE	337.8	wget http://hgdownload.cse.ucsc.edu/goldenP
5	PolyA+	Epidermal keratinocytes	**EK_pap**	ENCODE	strand spec	75	PE	241.9	wget http://hgdownload.cse.ucsc.edu/goldenP
6	Total	Epidermal melanocytes	**EM_tot**	ENCODE	strand spec	100	PE	260.2	wget http://hgdownload.cse.ucsc.edu/goldenP
7	Total	Follicle dermal papilla cells	**FDPC_tot**	ENCODE	strand spec	100	PE	255.9	wget http://hgdownload.cse.ucsc.edu/goldenP
8	PolyA+	HUVEC	**HUVEC_pap**	ENCODE	strand spec	75	PE	175.2	wget http://hgdownload.cse.ucsc.edu/goldenP
9	PolyA+	IMR90	**IMR90_pap**	ENCODE	strand spec	100	PE	165.0	wget http://hgdownload.cse.ucsc.edu/goldenP
10	Total	IMR90	**IMR90_tot**	ENCODE	strand spec	100	PE	217.4	wget http://hgdownload.cse.ucsc.edu/goldenP
11	PolyA+	K562	**K562_pap**	ENCODE	strand spec	75	PE	193.3	wget http://hgdownload.cse.ucsc.edu/goldenP
12	PolyA+	Lung fibroblasts	**LF_pap**	ENCODE	strand spec	75	PE	245.0	wget http://hgdownload.cse.ucsc.edu/goldenP
13	PolyA+	Mammary epithelial cells	**MEC_pap**	ENCODE	strand spec	75	PE	115.0	wget http://hgdownload.cse.ucsc.edu/goldenP
14	Total	Mammary epithelial cells	**MEC_tot**	ENCODE	strand spec	100	PE	135.6	wget http://hgdownload.cse.ucsc.edu/goldenP
15	Total	Undifferentiated mesenchymal stem cells from subcutaneous abdomen adipose tissue	**MSC_AT_tot**	ENCODE	strand spec	100	PE	256.1	wget http://hgdownload.cse.ucsc.edu/goldenP
16	Total	Undifferentiated mesenchymal stem cells from bone marrow	**MSC_BM_tot**	ENCODE	strand spec	100	PE	337.6	wget http://hgdownload.cse.ucsc.edu/goldenP
17	Total	Undifferentiated mesenchymal stem cells from umbical cord	**MSC_UC_tot**	ENCODE	strand spec	100	PE	221.0	wget http://hgdownload.cse.ucsc.edu/goldenP
18	Total	Mononuclear cells from peripheral blood	**MNC_PB_tot**	ENCODE	strand spec	100	PE	204.0	wget http://hgdownload.cse.ucsc.edu/goldenP
19	Total	Mononuclear cells from umbical cord blood	**MNC_UC_tot**	ENCODE	strand spec	100	PE	185.0	wget http://hgdownload.cse.ucsc.edu/goldenI
20	PolyA+	Monocytes	**Monocytes_pap**	ENCODE	strand spec	75	PE	170.8	wget http://hgdownload.cse.ucsc.edu/goldenP
21	Total	Undifferentiated osteoblasts	**Osteoblasts_tot**	ENCODE	strand spec	100	PE	370.9	wget http://hgdownload.cse.ucsc.edu/goldenP
22	Total	Placental epithelial cells	**PEC_tot**	ENCODE	strand spec	100	PE	285.7	wget http://hgdownload.cse.ucsc.edu/goldenP
23	Total	Saphenous vein endothelial cells	**SVEC_tot**	ENCODE	strand spec	100	PE	235.3	wget http://hgdownload.cse.ucsc.edu/goldenP
24	PolyA+	Skeletal muscle myoblasts	**SMM_pap**	ENCODE	strand spec	75	PE	219.9	wget http://hgdownload.cse.ucsc.edu/goldenP

Publication 2 Additional File 2H-1 - Overview of the publicly available RNA-seq datasets used in the study

Data set	type of RNA-seq	cell type	Short label (Figure 2A)	source	library	read length, bp	read type	Number of uniquely mapped reads, mln	download command (not fully displayed)
25	PolyA+	Skin fibroblasts	**SF_pap**	ENCODE	strand spec	75	PE	191.6	wget http://hgdownload.cse.ucsc.edu/goldenF
26	Total	Undifferentiated chondrocytes	**UnCh_tot**	ENCODE	strand spec	100	PE	364.8	wget http://hgdownload.cse.ucsc.edu/goldenF
27	PolyA+	Mixture of tissues*	**Mix_pap**	Illumina Human Body Map	strand spec	50	SE	575.1	wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR030
28	PolyA+	Lymphoblastoid cell line	**LCL_pap**	ENCODE	strand spec	75	PE	216.9	wget http://hgdownload.cse.ucsc.edu/goldenF
29	Total	Undifferentiated white preadipocytes	**UnWP_tot**	ENCODE	strand spec	100	PE	305.2	wget http://hgdownload.cse.ucsc.edu/goldenF
30	Total	Mobilized CD34+ cells	**CD34_tot**	ENCODE	strand spec	100	PE	204.8	wget http://hgdownload.cse.ucsc.edu/goldenI
31	PolyA+	Embryonic stem cells	**ESC_pap**	ENCODE	strand spec	75	PE	203.4	wget http://hgdownload.cse.ucsc.edu/goldenF
32	Total	Skeletal striated muscle cells	**SSMC_tot**	ENCODE	strand spec	100	PE	175.9	wget http://hgdownload.cse.ucsc.edu/goldenF
33	PolyA+	Fetal lung fibroblasts	**FLF_pap**	ENCODE	strand spec	75	PE	197.0	wget http://hgdownload.cse.ucsc.edu/goldenF
34	PolyA+	HeLa cells	**HeLa_pap**	ENCODE	strand spec	75	PE	211.0	wget http://hgdownload.cse.ucsc.edu/goldenF

* Mixture of tissues* Adrenal, Adipose, Brain, Breast, Colon, Heart, Kidney, Liver, Lung, Lymph Node, Ovary, Prostate, Skeletal Muscle, Testis, Thyroid, White Blood Cells

PE - paired end reads

SE - single end reads

Average number of uniquely mapped reads, mln 242.3

Publication 2 Additional File 2I - Pools used for human LCL transcriptome assembly (Geuvadis* raw RNAseq data used)

Pool number	population	gender	RNA-seq sample names pooled	Strand specific	number of input reads	uniquely mapped reads	number of spliced reads	multiple mapped reads	pools	number of input reads	uniquely mapped reads, %	uniquely mapped reads	multiple mapped reads	number of spliced reads
1	GBR	male	**HG00096**	No	27,256,165	24,773,114	7,962,899	1,581,025	1	116,797,342	90.1	105,280,495	6,900,266	35,064,475
1	GBR	female	**HG00097**	No	43,941,108	39,134,349	13,017,424	2,628,817						
1	GBR	female	**HG00100**	No	22,341,838	20,165,282	7,397,960	1,502,726						
1	GBR	male	**HG00101**	No	23,258,231	21,207,750	6,686,192	1,187,698						
2	FIN	female	**HG00171**	No	14,760,621	13,080,693	4,285,079	1,145,958	2	119,830,793	89.7	107,432,928	7,428,156	37,318,822
2	FIN	female	**HG00173**	No	17,350,282	14,511,190	4,670,129	1,139,365						
2	FIN	male	**HG00181**	No	24,721,834	22,836,439	8,032,229	1,351,730						
2	FIN	male	**HG00182**	No	26,648,092	24,114,994	7,657,925	1,552,848						
2	YRI	male	**NA18498**	No	36,349,964	32,889,612	12,673,460	2,238,255						
3	CEU	male	**NA06984**	No	22,136,382	19,162,772	5,647,074	1,315,979	3	109,172,552	89.4	97,571,729	7,388,696	32,497,314
3	CEU	female	**NA06985**	No	33,263,834	29,663,226	9,511,143	2,305,706						
3	CEU	male	**NA06986**	No	33,049,561	29,859,360	10,692,928	2,333,442						
3	CEU	female	**NA06989**	No	20,722,775	18,886,371	6,646,169	1,433,569						
4	YRI	female	**NA18499**	No	32,158,178	29,414,543	9,796,572	1,734,609	4	109,675,503	90.1	98,771,256	6,503,478	34,951,880
4	YRI	female	**NA18502**	No	56,870,036	50,467,250	18,507,407	3,499,107						
4	YRI	male	**NA18510**	No	20,647,289	18,889,463	6,647,901	1,269,762						
5	TSI	female	**NA20503**	No	25,288,232	22,851,777	7,599,493	1,509,228	5	125,339,286	90.2	113,035,596	7,496,387	37,932,930
5	TSI	female	**NA20504**	No	26,194,108	23,701,646	7,648,137	1,616,900						
5	TSI	male	**NA20509**	No	58,056,819	52,025,639	17,958,043	3,376,518						
5	TSI	male	**NA20510**	No	15,800,127	14,456,534	4,727,257	993,741						
									sum	580,815,476		522,092,004	35,716,983	177,765,421

RNA-seq type for all samples - PolyA+ RNA-seq

read type for all samples - 75bp paired end

* Lappalainen et al, Nature, 2013, "Transcriptome and genome sequencing uncovers functional variation in humans"

Publication 2 Additional File 2J-1 - Overview of the GTEx RNA-seq datasets used in the study

name of tissue in Figure S30	N_donor	**sex_s**	number of reads	Insert Size_l	LibraryLayout_s	**Run_s**	**SRA_Sample_s**	histological_type_s	is_tumor_s
Cells - EBV-transformed lymphocytes									
LCL	1	**female**	43,826,559	216	PAIRED	**SRR660295**	**SRS389735**	Blood	No
LCL	2	**female**	61,353,296	279	PAIRED	**SRR819975**	**SRS408820**	Blood	No
LCL	3	**male**	64,855,412	279	PAIRED	**SRR807925**	**SRS408125**	Blood	No
LCL	4	**male**	66,563,879	260	PAIRED	**SRR807849**	**SRS408122**	Blood	No
LCL	5	**male**	70,376,290	254	PAIRED	**SRR808968**	**SRS408198**	Blood	No
LCL	6	**male**	72,451,631	261	PAIRED	**SRR814794**	**SRS408476**	Blood	No
LCL	7	**female**	66,869,979	256	PAIRED	**SRR819096**	**SRS408763**	Blood	No
LCL	8	**male**	68,203,429	223	PAIRED	**SRR819430**	**SRS408787**	Blood	No
LCL	9	**male**	83,501,629	150	PAIRED	**SRR1095309**	**SRS525980**	Blood	No
LCL	10	**male**	92,961,693	150	PAIRED	**SRR1360359**	**SRS629948**	Blood	No
LCL	11	**female**	65,331,279	150	PAIRED	**SRR1399880**	**SRS637849**	Blood	No
LCL	12	**female**	80,810,369	350	PAIRED	**SRR1369264**	**SRS631488**	Blood	No
LCL	13	**male**	80,814,042	150	PAIRED	**SRR1415048**	**SRS639253**	Blood	No
LCL	14	**male**	77,314,489	150	PAIRED	**SRR1327951**	**SRS626624**	Blood	No
LCL	15	**female**	80,122,223	150	PAIRED	**SRR1419234**	**SRS639463**	Blood	No
LCL	16	**female**	72,032,056	150	PAIRED	**SRR1402391**	**SRS637980**	Blood	No
LCL	17	**female**	67,437,433	150	PAIRED	**SRR1469169**	**SRS648339**	Blood	No
LCL	18	**female**	70,380,569	150	PAIRED	**SRR1419086**	**SRS639456**	Blood	No
LCL	19	**male**	67,688,105	150	PAIRED	**SRR1381865**	**SRS634877**	Blood	No
LCL	20	**female**	67,720,605	150	PAIRED	**SRR1401885**	**SRS637957**	Blood	No
Adipose - Subcutaneous									
adipose	1	**female**	16,855,344	205	PAIRED	**SRR615778**	**SRS374916**	Adipose Tissue	No
adipose	2	**female**	40,345,893	231	PAIRED	**SRR663783**	**SRS389996**	Adipose Tissue	No
adipose	3	**female**	40,768,127	220	PAIRED	**SRR656059**	**SRS389039**	Adipose Tissue	No
adipose	4	**male**	59,854,637	249	PAIRED	**SRR658754**	**SRS389610**	Adipose Tissue	No
adipose	5	**female**	47,439,435	306	PAIRED	**SRR661267**	**SRS389858**	Adipose Tissue	No
adipose	6	**male**	16,312,304	188	PAIRED	**SRR660764**	**SRS389717**	Adipose Tissue	No
adipose	7	**male**	66,808,068	253	PAIRED	**SRR820448**	**SRS408850**	Adipose Tissue	No
adipose	8	**female**	63,858,469	238	PAIRED	**SRR821715**	**SRS408931**	Adipose Tissue	No
adipose	9	**female**	55,260,733	273	PAIRED	**SRR819820**	**SRS408811**	Adipose Tissue	No
adipose	10	**female**	67,982,959	268	PAIRED	**SRR813680**	**SRS408413**	Adipose Tissue	No
adipose	11	**female**	56,923,019	262	PAIRED	**SRR808019**	**SRS408129**	Adipose Tissue	No
adipose	12	**male**	58,124,266	231	PAIRED	**SRR813824**	**SRS408419**	Adipose Tissue	No
adipose	13	**male**	63,838,767	227	PAIRED	**SRR807543**	**SRS408109**	Adipose Tissue	No
adipose	14	**male**	55,115,316	229	PAIRED	**SRR817329**	**SRS408644**	Adipose Tissue	No
adipose	15	**female**	68,724,662	150	PAIRED	**SRR1500472**	**SRS650169**	Adipose Tissue	No
adipose	16	**male**	58,773,885	150	PAIRED	**SRR1420873**	**SRS639539**	Adipose Tissue	No
adipose	17	**male**	65,549,742	150	PAIRED	**SRR1444138**	**SRS644586**	Adipose Tissue	No
adipose	18	**male**	71,470,989	150	PAIRED	**SRR1318354**	**SRS625412**	Adipose Tissue	No
adipose	19	**female**	76,942,880	150	PAIRED	**SRR1430549**	**SRS643882**	Adipose Tissue	No
adipose	20	**male**	80,719,625	150	PAIRED	**SRR1419917**	**SRS639494**	Adipose Tissue	No
Artery - Tibial									
artery	1	**female**	61,003,998	240	PAIRED	**SRR612779**	**SRS374762**	Blood Vessel	No

Publication 2 Additional File 2J-2 - Overview of the GTEx RNA-seq datasets used in the study

name of tissue in Figure S30	N_donor	**sex_s**	number of reads	Insert Size_l	LibraryLayout_s	**Run_s**	**SRA_Sample_s**	histological_type_s	is_tumor_s
artery	2	**female**	55,512,319	214	PAIRED	**SRR615081**	**SRS374824**	Blood Vessel	No
artery	3	**male**	14,831,046	238	PAIRED	**SRR614984**	**SRS374828**	Blood Vessel	No
artery	4	**female**	28,142,221	229	PAIRED	**SRR615032**	**SRS374964**	Blood Vessel	No
artery	5	**male**	37,606,713	214	PAIRED	**SRR612604**	**SRS374700**	Blood Vessel	No
artery	6	**male**	37,910,528	224	PAIRED	**SRR615056**	**SRS375002**	Blood Vessel	No
artery	7	**male**	37,120,536	297	PAIRED	**SRR663185**	**SRS389982**	Blood Vessel	No
artery	8	**male**	42,438,208	301	PAIRED	**SRR809089**	**SRS408205**	Blood Vessel	No
artery	9	**male**	42,973,794	391	PAIRED	**SRR809755**	**SRS408247**	Blood Vessel	No
artery	10	**male**	53,143,118	262	PAIRED	**SRR657444**	**SRS389526**	Blood Vessel	No
artery	11	**female**	44,901,594	287	PAIRED	**SRR659839**	**SRS389655**	Blood Vessel	No
artery	12	**male**	47,078,655	319	PAIRED	**SRR659427**	**SRS389683**	Blood Vessel	No
artery	13	**female**	60,841,822	269	PAIRED	**SRR812625**	**SRS408379**	Blood Vessel	No
artery	14	**female**	65,479,260	249	PAIRED	**SRR812837**	**SRS408386**	Blood Vessel	No
artery	15	**male**	72,713,729	241	PAIRED	**SRR812366**	**SRS408370**	Blood Vessel	No
artery	16	**female**	76,453,677	150	PAIRED	**SRR1408727**	**SRS638324**	Blood Vessel	No
artery	17	**female**	69,205,722	150	PAIRED	**SRR1323252**	**SRS625690**	Blood Vessel	No
artery	18	**female**	65,489,844	150	PAIRED	**SRR1316078**	**SRS625206**	Blood Vessel	No
artery	19	**female**	71,348,616	150	PAIRED	**SRR1396804**	**SRS637662**	Blood Vessel	No
artery	20	**male**	66,488,451	150	PAIRED	**SRR1446436**	**SRS644695**	Blood Vessel	No
Brain - Cerebellum									
cerebellum	1	**female**	39,042,117	233	PAIRED	**SRR627429**	**SRS377766**	Brain	No
cerebellum	2	**male**	31,238,840	225	PAIRED	**SRR627462**	**SRS377775**	Brain	No
cerebellum	3	**female**	43,767,167	174	PAIRED	**SRR600876**	**SRS333582**	Brain	No
cerebellum	4	**female**	60,371,488	164	PAIRED	**SRR601098**	**SRS333391**	Brain	No
cerebellum	5	**male**	39,922,849	176	PAIRED	**SRR598396**	**SRS333341**	Brain	No
cerebellum	6	**female**	39,843,311	177	PAIRED	**SRR613747**	**SRS374906**	Brain	No
cerebellum	7	**female**	38,083,030	169	PAIRED	**SRR615249**	**SRS374911**	Brain	No
cerebellum	8	**male**	52,059,291	284	PAIRED	**SRR659412**	**SRS389682**	Brain	No
cerebellum	9	**male**	44,597,249	305	PAIRED	**SRR659331**	**SRS389645**	Brain	No
cerebellum	10	**female**	45,536,814	264	PAIRED	**SRR663453**	**SRS389966**	Brain	No
cerebellum	11	**male**	71,633,333	212	PAIRED	**SRR814563**	**SRS408462**	Brain	No
cerebellum	12	**female**	68,969,905	223	PAIRED	**SRR810957**	**SRS408309**	Brain	No
cerebellum	13	**female**	57,045,293	150	PAIRED	**SRR1467813**	**SRS648276**	Brain	No
cerebellum	14	**male**	76,662,861	150	PAIRED	**SRR1434250**	**SRS644056**	Brain	No
cerebellum	15	**male**	75,817,887	150	PAIRED	**SRR1478096**	**SRS648770**	Brain	No
cerebellum	16	**male**	71,889,594	150	PAIRED	**SRR1456514**	**SRS645909**	Brain	No
cerebellum	17	**male**	77,117,709	150	PAIRED	**SRR1466833**	**SRS648231**	Brain	No
cerebellum	18	**male**	70,536,195	150	PAIRED	**SRR1419675**	**SRS639483**	Brain	No
cerebellum	19	**female**	70,019,625	150	PAIRED	**SRR1362183**	**SRS630053**	Brain	No
cerebellum	20	**female**	85,401,884	150	PAIRED	**SRR1385036**	**SRS635380**	Brain	No
Heart - Left Ventricle									
heart	1	**male**	26,611,844	201	PAIRED	SRR604206	SRS333057	Heart	No
heart	2	**female**	27,992,839	229	PAIRED	SRR613510	SRS374867	Heart	No
heart	3	**female**	34,365,519	201	PAIRED	SRR612875	SRS374774	Heart	No

Publication 2 Additional File 2J-3 - Overview of the GTEx RNA-seq datasets used in the study

name of tissue in Figure S30	N_donor	**sex_s**	number of reads	Insert Size_l	LibraryLayout_s	**Run_s**	**SRA_Sample_s**	histological_type_s	is_tumor_s
heart	4	**female**	46,947,146	198	PAIRED	SRR612335	SRS374719	Heart	No
heart	5	**female**	30,854,322	259	PAIRED	SRR655792	SRS389056	Heart	No
heart	6	**male**	30,862,141	212	PAIRED	SRR659283	SRS389676	Heart	No
heart	7	**male**	17,614,685	286	PAIRED	SRR665514	SRS389845	Heart	No
heart	8	**female**	42,248,178	265	PAIRED	SRR661337	SRS389863	Heart	No
heart	9	**male**	37,185,448	303	PAIRED	SRR663404	SRS389834	Heart	No
heart	10	**female**	49,935,943	223	PAIRED	SRR810507	SRS408289	Heart	No
heart	11	**male**	46,852,982	150	PAIRED	SRR1394520	SRS637285	Heart	No
heart	12	**female**	47,730,023	150	PAIRED	SRR1434093	SRS644049	Heart	No
heart	13	**male**	51,292,672	150	PAIRED	SRR1321505	SRS625592	Heart	No
heart	14	**female**	49,953,488	150	PAIRED	SRR1370659	SRS631780	Heart	No
heart	15	**male**	18,511,881	150	PAIRED	SRR1454522	SRS645817	Heart	No
heart	16	**male**	54,807,313	150	PAIRED	SRR1400889	SRS637904	Heart	No
heart	17	**male**	45,528,414	150	PAIRED	SRR1352936	SRS629549	Heart	No
heart	18	**male**	46,821,194	150	PAIRED	SRR1310563	SRS623958	Heart	No
heart	19	**female**	26,379,643	150	PAIRED	SRR1347667	SRS629269	Heart	No
heart	20	**female**	48,532,060	150	PAIRED	SRR1401446	SRS637929	Heart	No
Lung									
lung	1	**female**	45,934,422	203	PAIRED	SRR600584	SRS333222	Lung	No
lung	2	**female**	35,393,248	188	PAIRED	SRR615539	SRS374980	Lung	No
lung	3	**female**	55,843,038	267	PAIRED	SRR655064	SRS389028	Lung	No
lung	4	**male**	45,278,740	203	PAIRED	SRR655625	SRS389042	Lung	No
lung	5	**male**	38,252,106	313	PAIRED	SRR821525	SRS408919	Lung	No
lung	6	**female**	20,293,190	330	PAIRED	SRR655601	SRS389012	Lung	No
lung	7	**female**	44,626,536	279	PAIRED	SRR663573	SRS389857	Lung	No
lung	8	**female**	50,001,934	280	PAIRED	SRR656886	SRS389310	Lung	No
lung	9	**male**	28,070,058	323	PAIRED	SRR661937	SRS389923	Lung	No
lung	10	**male**	54,900,438	297	PAIRED	SRR656874	SRS389309	Lung	No
lung	11	**male**	64,864,621	231	PAIRED	SRR817488	SRS408651	Lung	No
lung	12	**female**	70,340,992	223	PAIRED	SRR811866	SRS408349	Lung	No
lung	13	**female**	70,205,811	150	PAIRED	SRR1435066	SRS644096	Lung	No
lung	14	**female**	80,305,275	150	PAIRED	SRR1406176	SRS638184	Lung	No
lung	15	**male**	71,089,853	150	PAIRED	SRR1347795	SRS629275	Lung	No
lung	16	**male**	71,621,573	150	PAIRED	SRR1475524	SRS648648	Lung	No
lung	17	**male**	72,690,166	150	PAIRED	SRR1333659	SRS627204	Lung	No
lung	18	**female**	76,138,575	150	PAIRED	SRR1429660	SRS643842	Lung	No
lung	19	**male**	79,771,050	150	PAIRED	SRR1339029	SRS627793	Lung	No
lung	20	**male**	76,206,688	150	PAIRED	SRR1318118	SRS625401	Lung	No
Muscle - Skeletal									
muscle	1	**female**	22,571,004	265	PAIRED	SRR627439	SRS377779	Muscle	No
muscle	2	**male**	49,384,644	204	PAIRED	SRR598044	SRS333390	Muscle	No
muscle	3	**female**	45,435,025	182	PAIRED	SRR608264	SRS333125	Muscle	No
muscle	4	**male**	31,523,094	218	PAIRED	SRR607067	SRS333010	Muscle	No
muscle	5	**male**	44,376,301	228	PAIRED	SRR607117	SRS333513	Muscle	No

Publication 2 Additional File 2J-4 - Overview of the GTEx RNA-seq datasets used in the study

name of tissue in Figure S30	N_donor	**sex_s**	number of reads	Insert Size_l	LibraryLayout_s	**Run_s**	**SRA_Sample_s**	histological_type_s	is_tumor_s
muscle	6	**female**	24,482,180	226	PAIRED	SRR615862	SRS374999	Muscle	No
muscle	7	**male**	36,743,677	229	PAIRED	SRR613330	SRS374819	Muscle	No
muscle	8	**female**	42,520,553	328	PAIRED	SRR661155	SRS389809	Muscle	No
muscle	9	**male**	28,805,611	339	PAIRED	SRR813656	SRS408412	Muscle	No
muscle	10	**male**	42,947,973	323	PAIRED	SRR815044	SRS408495	Muscle	No
muscle	11	**male**	45,145,139	331	PAIRED	SRR658057	SRS389032	Muscle	No
muscle	12	**female**	44,906,664	300	PAIRED	SRR655887	SRS389091	Muscle	No
muscle	13	**female**	53,062,750	305	PAIRED	SRR818773	SRS408742	Muscle	No
muscle	14	**female**	54,677,479	304	PAIRED	SRR816226	SRS408580	Muscle	No
muscle	15	**female**	45,400,428	289	PAIRED	SRR810249	SRS408278	Muscle	No
muscle	16	**male**	53,426,767	277	PAIRED	SRR809348	SRS408230	Muscle	No
muscle	17	**female**	60,211,775	223	PAIRED	SRR809444	SRS408235	Muscle	No
muscle	18	**female**	76,711,863	150	PAIRED	SRR1403366	SRS638023	Muscle	No
muscle	19	**male**	65,062,674	150	PAIRED	SRR1342782	SRS628342	Muscle	No
muscle	20	**male**	66,861,094	150	PAIRED	SRR1380716	SRS634794	Muscle	No
Nerve - Tibial									
nerve	1	**male**	57,450,768	215	PAIRED	SRR614827	SRS374997	Nerve	No
nerve	2	**female**	50,989,133	209	PAIRED	SRR613987	SRS374947	Nerve	No
nerve	3	**female**	39,699,570	180	PAIRED	SRR612911	SRS374778	Nerve	No
nerve	4	**female**	33,573,685	197	PAIRED	SRR613591	SRS374880	Nerve	No
nerve	5	**female**	22,222,434	222	PAIRED	SRR615671	SRS374974	Nerve	No
nerve	6	**female**	22,506,268	208	PAIRED	SRR613867	SRS374892	Nerve	No
nerve	7	**male**	22,551,221	194	PAIRED	SRR613687	SRS374894	Nerve	No
nerve	8	**female**	43,069,891	233	PAIRED	SRR661421	SRS389871	Nerve	No
nerve	9	**male**	16,738,958	225	PAIRED	SRR662055	SRS389932	Nerve	No
nerve	10	**male**	40,869,770	261	PAIRED	SRR814052	SRS408432	Nerve	No
nerve	11	**male**	28,993,121	341	PAIRED	SRR659049	SRS389641	Nerve	No
nerve	12	**male**	78,113,473	215	PAIRED	SRR656770	SRS389305	Nerve	No
nerve	13	**male**	45,999,964	270	PAIRED	SRR662486	SRS389960	Nerve	No
nerve	14	**female**	45,687,119	258	PAIRED	SRR656083	SRS389068	Nerve	No
nerve	15	**female**	59,871,354	225	PAIRED	SRR818939	SRS408753	Nerve	No
nerve	16	**male**	59,988,648	279	PAIRED	SRR815422	SRS408511	Nerve	No
nerve	17	**female**	60,044,536	272	PAIRED	SRR815685	SRS408538	Nerve	No
nerve	18	**male**	67,807,778	216	PAIRED	SRR812080	SRS408358	Nerve	No
nerve	19	**female**	76,396,371	150	PAIRED	SRR1469214	SRS648341	Nerve	No
nerve	20	**male**	66,316,441	150	PAIRED	SRR1447019	SRS644722	Nerve	No
Thyroid									
thyroid	1	**female**	51,361,361	186	PAIRED	SRR604294	SRS333450	Thyroid	No
thyroid	2	**female**	64,194,854	179	PAIRED	SRR601359	SRS333389	Thyroid	No
thyroid	3	**male**	65,821,523	180	PAIRED	SRR598565	SRS333481	Thyroid	No
thyroid	4	**female**	59,131,414	182	PAIRED	SRR598100	SRS333268	Thyroid	No
thyroid	5	**male**	26,309,219	197	PAIRED	SRR614275	SRS374939	Thyroid	No
thyroid	6	**female**	31,771,033	187	PAIRED	SRR614743	SRS374952	Thyroid	No
thyroid	7	**male**	47,301,434	208	PAIRED	SRR657372	SRS389520	Thyroid	No

Publication 2 Additional File 2J-5 - Overview of the GTEx RNA-seq datasets used in the study

name of tissue in Figure S30	N_donor	**sex_s**	number of reads	Insert Size_l	LibraryLayout_s	**Run_s**	**SRA_Sample_s**	histological_type_s	is_tumor_s
thyroid	8	**male**	45,546,626	189	PAIRED	SRR663795	SRS389997	Thyroid	No
thyroid	9	**male**	42,595,504	307	PAIRED	SRR663392	SRS389833	Thyroid	No
thyroid	10	**male**	69,649,018	259	PAIRED	SRR658573	SRS389583	Thyroid	No
thyroid	11	**female**	55,257,414	238	PAIRED	SRR808658	SRS408175	Thyroid	No
thyroid	12	**male**	72,717,486	201	PAIRED	SRR821176	SRS408896	Thyroid	No
thyroid	13	**female**	77,729,936	150	PAIRED	SRR1323850	SRS625717	Thyroid	No
thyroid	14	**female**	72,814,234	150	PAIRED	SRR1499543	SRS650126	Thyroid	No
thyroid	15	**male**	63,147,359	150	PAIRED	SRR1498226	SRS650067	Thyroid	No
thyroid	16	**female**	63,488,871	150	PAIRED	SRR1318898	SRS625453	Thyroid	No
thyroid	17	**male**	63,926,058	150	PAIRED	SRR1312024	SRS624031	Thyroid	No
thyroid	18	**male**	64,752,806	150	PAIRED	SRR1313789	SRS624514	Thyroid	No
thyroid	19	**female**	69,941,031	150	PAIRED	SRR1334939	SRS627421	Thyroid	No
thyroid	20	**female**	79,585,098	150	PAIRED	SRR1330792	SRS627069	Thyroid	No

average read number:	52,093,468
maximal read number:	85,401,884
minimal read number:	14,831,046

Publication 2 Additional File 11A

List of 120 donors used in the donor saturation study with corresponding population and pools

N	Donor code	Gender	Population	Pool number	Pool name	N	Donor code	Gender	Population	Pool number	Pool name
1	HG00096	m	GBR	1	pool1_GBR	61	NA11840	f	CEU	16	pool4_CEU
2	HG00097	f	GBR	1	pool1_GBR	62	NA11881	m	CEU	16	pool4_CEU
3	HG00099	f	GBR	1	pool1_GBR	63	NA11893	m	CEU	16	pool4_CEU
4	HG00103	m	GBR	1	pool1_GBR	64	NA11894	f	CEU	16	pool4_CEU
5	HG00105	m	GBR	2	pool2_GBR	65	NA11920	f	CEU	17	pool5_CEU
6	HG00106	f	GBR	2	pool2_GBR	66	NA11993	f	CEU	17	pool5_CEU
7	HG00108	m	GBR	2	pool2_GBR	67	NA11994	m	CEU	17	pool5_CEU
8	HG00110	f	GBR	2	pool2_GBR	68	NA11995	f	CEU	18	pool6_CEU
9	HG00111	f	GBR	3	pool3_GBR	69	NA12005	m	CEU	17	pool5_CEU
10	HG00112	m	GBR	3	pool3_GBR	70	NA12006	f	CEU	18	pool6_CEU
11	HG00115	m	GBR	3	pool3_GBR	71	NA12043	m	CEU	18	pool6_CEU
12	HG00116	m	GBR	4	pool4_GBR	72	NA12144	m	CEU	18	pool6_CEU
13	HG00117	m	GBR	4	pool4_GBR	73	NA18486	m	YRI	19	pool1_YRI
14	HG00118	f	GBR	3	pool3_GBR	74	NA18489	f	YRI	19	pool1_YRI
15	HG00119	m	GBR	5	pool5_GBR	75	NA18498	m	YRI	19	pool1_YRI
16	HG00122	f	GBR	4	pool4_GBR	76	NA18499	f	YRI	19	pool1_YRI
17	HG00123	f	GBR	4	pool4_GBR	77	NA18502	f	YRI	20	pool2_YRI
18	HG00124	f	GBR	5	pool5_GBR	78	NA18520	f	YRI	20	pool2_YRI
19	HG00125	f	GBR	5	pool5_GBR	79	NA18858	f	YRI	21	pool3_YRI
20	HG00126	m	GBR	5	pool5_GBR	80	NA18867	f	YRI	21	pool3_YRI
21	HG00127	f	GBR	6	pool6_GBR	81	NA18868	m	YRI	20	pool2_YRI
22	HG00128	f	GBR	6	pool6_GBR	82	NA18873	f	YRI	22	pool4_YRI
23	HG00131	m	GBR	6	pool6_GBR	83	NA18909	f	YRI	22	pool4_YRI
24	HG00136	m	GBR	6	pool6_GBR	84	NA18912	f	YRI	23	pool5_YRI
25	HG00174	f	FIN	7	pool1_FIN	85	NA18916	f	YRI	23	pool5_YRI
26	HG00177	f	FIN	7	pool1_FIN	86	NA18917	m	YRI	20	pool2_YRI
27	HG00178	f	FIN	8	pool2_FIN	87	NA18923	m	YRI	21	pool3_YRI
28	HG00180	f	FIN	8	pool2_FIN	88	NA18934	m	YRI	21	pool3_YRI
29	HG00182	m	FIN	7	pool1_FIN	89	NA19092	m	YRI	22	pool4_YRI
30	HG00183	m	FIN	7	pool1_FIN	90	NA19093	f	YRI	24	pool6_YRI
31	HG00185	m	FIN	8	pool2_FIN	91	NA19095	f	YRI	24	pool6_YRI
32	HG00187	m	FIN	8	pool2_FIN	92	NA19098	m	YRI	22	pool4_YRI
33	HG00188	m	FIN	9	pool3_FIN	93	NA19107	m	YRI	23	pool5_YRI
34	HG00266	f	FIN	9	pool3_FIN	94	NA19117	m	YRI	23	pool5_YRI
35	HG00267	m	FIN	9	pool3_FIN	95	NA19130	m	YRI	24	pool6_YRI
36	HG00268	f	FIN	9	pool3_FIN	96	NA19141	m	YRI	24	pool6_YRI
37	HG00269	f	FIN	10	pool4_FIN	97	NA20504	f	TSI	25	pool1_TSI
38	HG00271	m	FIN	10	pool4_FIN	98	NA20505	f	TSI	25	pool1_TSI
39	HG00272	f	FIN	10	pool4_FIN	99	NA20507	f	TSI	26	pool2_TSI
40	HG00274	f	FIN	11	pool5_FIN	100	NA20508	f	TSI	26	pool2_TSI
41	HG00276	f	FIN	11	pool5_FIN	101	NA20509	m	TSI	25	pool1_TSI
42	HG00280	m	FIN	10	pool4_FIN	102	NA20513	m	TSI	25	pool1_TSI
43	HG00281	f	FIN	12	pool6_FIN	103	NA20514	f	TSI	27	pool3_TSI
44	HG00282	f	FIN	12	pool6_FIN	104	NA20515	m	TSI	26	pool2_TSI
45	HG00284	m	FIN	11	pool5_FIN	105	NA20516	m	TSI	26	pool2_TSI
46	HG00310	m	FIN	11	pool5_FIN	106	NA20518	m	TSI	27	pool3_TSI
47	HG00311	m	FIN	12	pool6_FIN	107	NA20519	m	TSI	27	pool3_TSI
48	HG00312	m	FIN	12	pool6_FIN	108	NA20524	m	TSI	28	pool4_TSI
49	NA06985	f	CEU	13	pool1_CEU	109	NA20525	m	TSI	28	pool4_TSI
50	NA06986	m	CEU	13	pool1_CEU	110	NA20527	m	TSI	29	pool5_TSI
51	NA06994	m	CEU	13	pool1_CEU	111	NA20528	m	TSI	29	pool5_TSI
52	NA07037	f	CEU	13	pool1_CEU	112	NA20529	f	TSI	27	pool3_TSI
53	NA07051	m	CEU	14	pool2_CEU	113	NA20530	f	TSI	28	pool4_TSI
54	NA07056	f	CEU	14	pool2_CEU	114	NA20531	f	TSI	28	pool4_TSI
55	NA07346	f	CEU	14	pool2_CEU	115	NA20534	m	TSI	30	pool6_TSI
56	NA07347	m	CEU	14	pool2_CEU	116	NA20539	m	TSI	30	pool6_TSI
57	NA07357	m	CEU	15	pool3_CEU	117	NA20541	f	TSI	29	pool5_TSI
58	NA10847	f	CEU	15	pool3_CEU	118	NA20582	f	TSI	29	pool5_TSI
59	NA11831	m	CEU	15	pool3_CEU	119	NA20589	f	TSI	30	pool6_TSI
60	NA11832	f	CEU	15	pool3_CEU	120	NA20756	f	TSI	30	pool6_TSI

Publication 2 Additional File11B-1 - List of randomly picked pools for each data point

	number of assemblies	number of donors	point on the plot	1	2	3	4	5	6	7	8	9	10
1	1	4	point1-1	pool3_CEU									
	1	4	point1-2	pool5_FIN									
	1	4	point1-3	pool5_TSI									
2	2	8	point2-1	pool6_TSI	pool2_YRI								
	2	8	point2-2	pool5_YRI	pool1_CEU								
	2	8	point2-3	pool5_CEU	pool3_TSI								
3	3	12	point3-1	pool4_CEU	pool5_YRI	pool5_FIN							
	3	12	point3-2	pool6_FIN	pool2_TSI	pool3_TSI							
4	3	12	point3-3	pool5_GBR	pool4_FIN	pool5_YRI							
4	4	16	point4-1	pool2_TSI	pool5_CEU	pool3_FIN	pool3_CEU						
	4	16	point4-2	pool3_CEU	pool2_FIN	pool1_GBR	pool5_GBR						
	4	16	point4-3	pool6_GBR	pool6_YRI	pool2_GBR	pool3_GBR						
5	5	20	point5-1	pool3_FIN	pool4_GBR	pool3_YRI	pool1_YRI	pool2_CEU					
	5	20	point5-2	pool1_GBR	pool3_CEU	pool4_CEU	pool3_YRI	pool5_TSI					
6	5	20	point5-3	pool1_FIN	pool3_FIN	pool1_TSI	pool6_TSI	pool5_YRI					
6	6	24	point6-1	pool6_YRI	pool6_FIN	pool1_FIN	pool1_TSI	pool5_FIN	pool4_FIN				
	6	24	point6-2	pool2_YRI	pool1_TSI	pool6_CEU	pool4_TSI	pool2_FIN	pool6_YRI				
	6	24	point6-3	pool4_FIN	pool2_CEU	pool1_CEU	pool2_TSI	pool4_YRI	pool3_TSI				
7	8	32	point7-1	pool4_GBR	pool6_TSI	pool6_GBR	pool1_TSI	pool5_YRI	pool5_FIN	pool2_GBR	pool3_FIN		
	8	32	point7-2	pool5_GBR	pool1_FIN	pool4_TSI	pool1_CEU	pool3_GBR	pool3_YRI	pool4_YRI	pool2_GBR		
	8	32	point7-3	pool2_TSI	pool5_YRI	pool6_TSI	pool4_FIN	pool4_GBR	pool5_FIN	pool3_CEU	pool4_CEU		
8	10	40	point8-1	pool1_GBR	pool6_CEU	pool6_FIN	pool3_CEU	pool1_FIN	pool6_GBR	pool3_GBR	pool4_GBR	pool4_CEU	pool2_GBR
	10	40	point8-2	pool2_CEU	pool3_YRI	pool2_TSI	pool1_YRI	pool5_FIN	pool1_CEU	pool3_TSI	pool6_TSI	pool5_CEU	pool3_CEU
	10	40	point8-3	pool4_TSI	pool2_CEU	pool6_TSI	pool5_CEU	pool5_YRI	pool3_GBR	pool2_GBR	pool2_FIN	pool2_TSI	pool6_FIN
9	15	60	point9-1	pool1_YRI	pool6_GBR	pool4_CEU	pool4_TSI	pool5_TSI	pool4_FIN	pool2_YRI	pool1_FIN	pool4_GBR	pool2_GBR
				pool6_FIN	pool5_GBR	pool2_CEU	pool5_FIN	pool4_YRI					
	15	60	point9-2	pool4_GBR	pool6_GBR	pool1_TSI	pool2_CEU	pool5_GBR	pool1_FIN	pool6_YRI	pool5_CEU	pool3_CEU	pool3_YRI
				pool3_FIN	pool5_FIN	pool6_TSI	pool2_YRI	pool1_GBR					
	15	60	point9-3	pool5_GBR	pool1_GBR	pool4_GBR	pool6_GBR	pool4_FIN	pool4_YRI	pool1_CEU	pool3_YRI	pool2_TSI	pool2_CEU
				pool1_YRI	pool3_TSI	pool3_FIN	pool5_CEU	pool6_YRI					
10	20	80	point10-1	pool5_YRI	pool5_TSI	pool6_GBR	pool2_CEU	pool4_GBR	pool5_GBR	pool6_CEU	pool1_CEU	pool4_YRI	pool1_GBR
				pool4_TSI	pool1_YRI	pool1_TSI	pool5_FIN	pool2_YRI	pool2_FIN	pool3_FIN	pool6_TSI	pool3_GBR	pool6_YRI
	20	80	point10-2	pool6_YRI	pool5_FIN	pool5_YRI	pool4_CEU	pool1_TSI	pool6_TSI	pool5_TSI	pool5_GBR	pool1_CEU	pool1_FIN
				pool4_YRI	pool2_GBR	pool3_GBR	pool3_CEU	pool6_GBR	pool2_CEU	pool4_TSI	pool6_CEU	pool2_YRI	pool4_FIN
	20	80	point10-3	pool2_CEU	pool3_FIN	pool5_GBR	pool6_FIN	pool2_GBR	pool3_TSI	pool2_FIN	pool4_YRI	pool1_CEU	pool4_FIN
				pool6_TSI	pool3_YRI	pool6_YRI	pool6_GBR	pool1_FIN	pool5_TSI	pool5_CEU	pool2_YRI	pool4_TSI	pool5_FIN

Publication 2 Additional File11B-2 - List of randomly picked pools for each data point

	number of assemblies	number of donors	point on the plot	1	2	3	4	5	6	7	8	9	10
11	25	100	point11-1	pool3_CEU	pool1_GBR	pool1_FIN	pool3_GBR	pool6_TSI	pool3_TSI	pool2_TSI	pool6_FIN	pool5_YRI	pool5_GBR
				pool4_TSI	pool5_CEU	pool4_YRI	pool4_CEU	pool6_CEU	pool4_FIN	pool2_GBR	pool2_FIN	pool3_YRI	pool6_YRI
				pool1_CEU	pool1_TSI	pool2_CEU	pool2_YRI	pool6_GBR					
	25	100	point11-2	pool6_CEU	pool3_TSI	pool4_TSI	pool2_CEU	pool5_GBR	pool4_FIN	pool3_FIN	pool5_TSI	pool5_YRI	pool6_YRI
				pool3_YRI	pool6_TSI	pool4_CEU	pool5_CEU	pool2_GBR	pool3_CEU	pool5_FIN	pool2_TSI	pool3_GBR	pool6_GBR
				pool2_FIN	pool1_FIN	pool4_GBR	pool1_CEU	pool6_FIN					
	25	100	point11-3	pool5_GBR	pool3_TSI	pool2_TSI	pool5_YRI	pool5_TSI	pool2_YRI	pool4_GBR	pool4_FIN	pool6_FIN	pool3_GBR
				pool2_FIN	pool3_YRI	pool6_CEU	pool2_GBR	pool5_FIN	pool5_CEU	pool3_CEU	pool1_GBR	pool6_GBR	pool1_FIN
				pool6_YRI	pool1_YRI	pool4_TSI	pool6_TSI	pool1_CEU					
12	30	120	point12-1	pool5_YRI	pool5_FIN	pool2_CEU	pool6_YRI	pool2_FIN	pool2_GBR	pool1_TSI	pool3_FIN	pool4_FIN	pool3_CEU
				pool5_CEU	pool6_GBR	pool4_YRI	pool3_GBR	pool4_GBR	pool3_YRI	pool1_CEU	pool1_FIN	pool2_YRI	pool4_CEU
				pool2_TSI	pool6_CEU	pool1_YRI	pool3_TSI	pool5_GBR	pool6_FIN	pool1_GBR	pool5_TSI	pool6_TSI	pool4_TSI

Publication 2 Additional File 11C

Number of de novo lncRNA and mRNA loci annotated using different number of transcriptome assemblies (donors)

–data for plotting Fig. 7B, Figure S32C,D,E and S34 in Additional File 1

					Fig.7B	Fig.S34	Fig.S34	Fig.S32C	Fig.S32D	Fig.7B	Fig.S32C	Fig.S32D	Fig.S32E
	number of assemblies	number of donors	number of replicate	point on the plot	number of lncRNA loci	number of new lncRNA loci	number of known lncRNA loci	number of lncRNA transcrips	median number of transcripts annotated in lncRNA loci	number of mRNA loci	number of mRNA transcrips	median number of transcripts annotated in lncRNA loci	number of loci discarded because of potential PC capacity
1	1	4	1	point1-1	1,520	513	1007	3,168	1	12,159	48,479	4	90
	1	4	2	point1-2	1,233	363	870	2,544	1	12,087	47,831	4	78
	1	4	3	point1-3	1,394	425	969	2,817	1	12,010	48,100	4	80
2	2	8	1	point2-1	1,818	636	1182	4,549	1	12,357	73,340	5	100
	2	8	2	point2-2	1,988	729	1259	4,929	1	12,366	74,638	5	107
	2	8	3	point2-3	1,807	611	1196	4,537	1	12,219	71,982	5	105
3	3	12	1	point3-1	2,069	779	1290	5,898	1	12,377	92,970	6	115
	3	12	2	point3-2	2,023	725	1298	5,624	1	12,162	90,261	6	111
	3	12	3	point3-3	2,036	750	1286	5,739	1	12,277	93,125	6	111
4	4	16	1	point4-1	2,320	877	1443	7,178	1	12,315	107,618	7	127
	4	16	2	point4-2	2,335	899	1436	7,182	1	12,325	107,904	7	127
	4	16	3	point4-3	2,275	876	1399	6,923	1	12,239	109,119	7	129
5	5	20	1	point5-1	2,624	1079	1545	8,353	1	12,588	125,091	8	138
	5	20	2	point5-2	2,597	1062	1535	8,626	1	12,590	124,328	8	132
	5	20	3	point5-3	2,494	983	1511	8,033	1	12,451	122,916	8	136
6	6	24	1	point6-1	2,591	1045	1546	8,674	1	12,412	136,070	9	124
	6	24	2	point6-2	2,633	1084	1549	8,893	1	12,593	136,253	9	141
	6	24	3	point6-3	2,773	1162	1611	9,223	1	12,599	136,229	9	142
7	8	32	1	point7-1	2,731	1145	1586	9,984	1	12,490	160,149	10	143
	8	32	2	point7-2	3,035	1320	1715	10,961	1	12,717	161,225	10	151
	8	32	3	point7-3	2,817	1208	1609	10,399	1	12,521	161,345	10	139
8	10	40	1	point8-1	3,041	1326	1715	11,931	2	12,567	178,957	11	150
	10	40	2	point8-2	3,155	1403	1752	12,387	1	12,710	179,695	11	157
	10	40	3	point8-3	3,041	1330	1711	11,625	1	12,629	179,933	11	170
9	15	60	1	point9-1	3,409	1551	1858	14,708	1	12,725	224,858	12	170
	15	60	2	point9-2	3,529	1630	1899	14,846	1	12,784	224,069	13	169
	15	60	3	point9-3	3,513	1612	1901	14,701	1	12,789	223,342	12	168
10	20	80	1	point10-1	3,735	1778	1957	17,024	1	12,803	262,758	14	176
	20	80	2	point10-2	3,782	1785	1997	17,345	1	12,832	264,468	14	186
	20	80	3	point10-3	3,769	1774	1995	17,040	1	12,859	263,438	14	182
11	25	100	1	point11-1	4,009	1939	2070	19,361	1	12,862	297,305	15	193
	25	100	2	point11-2	3,956	1919	2037	19,141	1	12,842	299,308	15	185
	25	100	3	point11-3	3,938	1916	2022	19,050	1	12,788	298,228	15	197
	30	120	1	point12-1	4,166	2063	2103	20,992	1	12,857	330,797	16	198

Publication 2 Additional File 11D

Number of de novo lncRNA and mRNA loci from "120 donors" annotation identified using less transcriptome assemblies (donors)

Data for plotting donor saturation curve - Figure S35A in Additional File 1.

	number of assemblies	number of donors	number of replicate	point on the plot	number of "120 donors annotation" lncRNA loci dentified using less donors	number of "120 donors annotation" mRNA loci dentified using less donors
1	1	4	1	point1-1	974	9,492
	1	4	2	point1-2	711	9,373
	1	4	3	point1-3	870	9,454
2	2	8	1	point2-1	1,232	10,293
	2	8	2	point2-2	1,375	10,408
	2	8	3	point2-3	1,182	9,979
3	3	12	1	point3-1	1,443	10,614
	3	12	2	point3-2	1,360	10,282
	3	12	3	point3-3	1,449	10,552
4	4	16	1	point4-1	1,646	10,596
	4	16	2	point4-2	1,717	10,754
	4	16	3	point4-3	1,630	10,540
5	5	20	1	point5-1	1,977	11,125
	5	20	2	point5-2	1,967	11,066
	5	20	3	point5-3	1,860	10,991
6	6	24	1	point6-1	1,900	11,029
	6	24	2	point6-2	2,020	11,289
	6	24	3	point6-3	2,146	11,238
7	8	32	1	point7-1	2,128	11,256
	8	32	2	point7-2	2,482	11,526
	8	32	3	point7-3	2,186	11,292
8	10	40	1	point8-1	2,475	11,363
	10	40	2	point8-2	2,571	11,687
	10	40	3	point8-3	2,491	11,548
9	15	60	1	point9-1	3,002	12,016
	15	60	2	point9-2	3,053	12,066
	15	60	3	point9-3	3,029	12,039
10	20	80	1	point10-1	3,427	12,380
	20	80	2	point10-2	3,485	12,367
	20	80	3	point10-3	3,437	12,390
11	25	100	1	point11-1	3,827	12,631
	25	100	2	point11-2	3,763	12,533
	25	100	3	point11-3	3,743	12,518
	30	120	1	point12-1	4,166	12,857

"**identified**" means covered by a lncRNA/mRNA locus in the other annotation on the same strand by at least 50% of length

Publication 2 Additional File 11E -1

Number of de novo lncRNAs from different expression bins identified from increasing number of donors

Data for plotting Figure S33 in Additional File 1.

					number of transcripts							number of loci						
	number of assemblies	number of donors	number of replicate		bin0 (RPKMmax<0.5)	bin1 (0.5<RPKMmax<1)	bin2 (1<RPKMmax<2)	bin3 (2<RPKMmax<4)	bin4 (4<RPKMmax<8)	bin5 (8<RPKMmax)			bin0 (RPKMmax<0.5)	bin1 (0.5<RPKMmax<1)	bin2 (1<RPKMmax<2)	bin3 (2<RPKMmax<4)	bin4 (4<RPKMmax<8)	bin5 (8<RPKMmax)
1	1	4	1	point1-1	435	619	725	597	384	408		point1-1	383	439	412	292	193	161
	1	4	2	point1-2	414	599	570	430	261	270		point1-2	353	396	304	217	125	107
	1	4	3	point1-3	464	641	647	456	298	311		point1-3	403	431	360	242	139	121
2	2	8	1	point2-1	598	965	1015	827	529	615		point2-1	499	587	487	343	210	194
	2	8	2	point2-2	618	1008	1056	954	634	659		point2-2	503	619	540	435	245	222
	2	8	3	point2-3	597	932	1087	828	591	502		point2-3	489	562	491	366	228	166
3	3	12	1	point3-1	742	1314	1319	1059	685	779		point3-1	550	685	565	402	237	241
	3	12	2	point3-2	820	1253	1303	944	665	639		point3-2	618	663	514	369	225	190
	3	12	3	point3-3	850	1299	1257	1041	588	704		point3-3	630	657	537	392	234	204
4	4	16	1	point4-1	890	1386	1611	1283	974	1034		point4-1	669	727	625	457	300	269
	4	16	2	point4-2	961	1389	1675	1255	875	1027		point4-2	692	721	641	462	284	262
	4	16	3	point4-3	901	1454	1533	1249	811	975		point4-3	699	693	589	416	254	255
5	5	20	1	point5-1	980	1637	1972	1527	1035	1202		point5-1	746	830	729	510	310	303
	5	20	2	point5-2	997	1679	2026	1467	1145	1312		point5-2	723	834	733	496	338	310
	5	20	3	point5-3	941	1635	1905	1411	992	1149		point5-3	703	817	677	462	317	289
6	6	24	1	point6-1	966	1733	2019	1595	1098	1263		point6-1	718	833	683	474	342	301
	6	24	2	point6-2	917	1735	1913	1595	1337	1396		point6-2	697	868	707	514	370	312
	6	24	3	point6-3	1052	1727	2050	1770	1185	1439		point6-3	755	863	775	585	378	324
7	8	32	1	point7-1	1189	1938	2280	1816	1238	1523		point7-1	826	871	718	514	322	329
	8	32	2	point7-2	1234	1960	2425	2028	1529	1785		point7-2	836	967	838	610	409	365
	8	32	3	point7-3	1233	2010	2285	1991	1264	1616		point7-3	830	908	759	551	352	350
8	10	40	1	point8-1	1318	2161	2678	2193	1666	1915		point8-1	882	982	801	577	418	371
	10	40	2	point8-2	1231	2190	2777	2301	1682	2206		point8-2	846	1014	876	664	419	388
	10	40	3	point8-3	1255	2157	2459	2192	1652	1910		point8-3	872	960	840	607	376	382

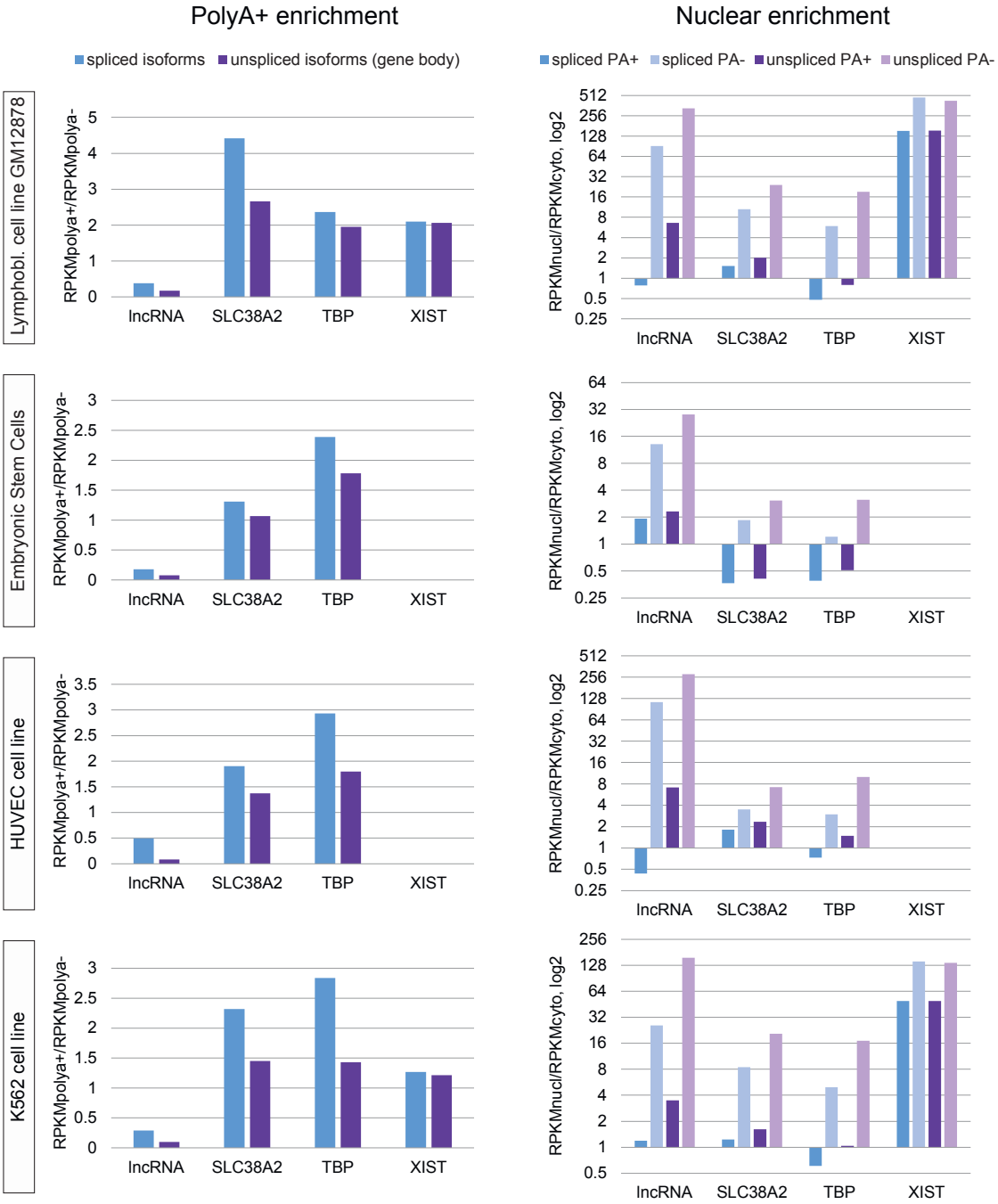
Publication 2 Additional File 11E -2

Number of de novo lncRNAs from different expression bins identified from increasing number of donors

Data for plotting Figure S33 in Additional File 1.

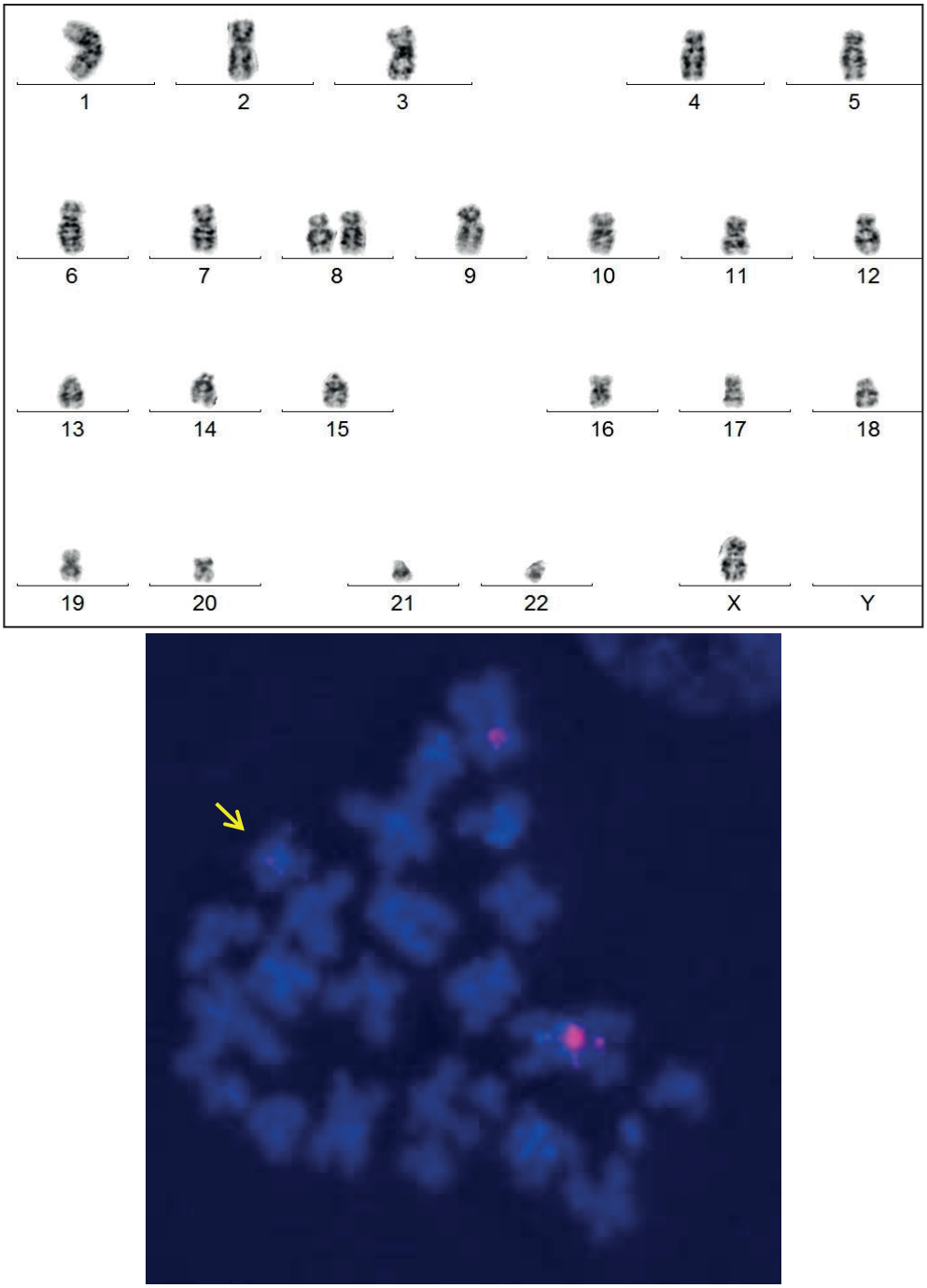
9	15	60	1	point9-1	1435	2550	3219	2859	2171	2474		point9-1	941	1094	952	695	462	445
	15	60	2	point9-2	1434	2448	3304	2843	2242	2575		point9-2	970	1098	977	699	495	466
	15	60	3	point9-3	1460	2400	3343	2724	2172	2602		point9-3	953	1077	982	734	497	464
10	20	80	1	point10-1	1477	2776	3581	3291	2789	3110		point10-1	986	1181	1017	781	533	503
	20	80	2	point10-2	1489	2807	3703	3446	2762	3138		point10-2	985	1201	1053	785	562	500
	20	80	3	point10-3	1496	2696	3717	3230	2721	3180		point10-3	989	1168	1039	797	542	518
11	25	100	1	point11-1	1577	3046	4051	3662	3230	3795		point11-1	1032	1240	1100	830	603	557
	25	100	2	point11-2	1604	3082	4062	3631	3125	3637		point11-2	1054	1254	1064	826	568	549
	25	100	3	point11-3	1640	3070	4036	3563	3095	3646		point11-3	1059	1231	1048	810	575	531
	30	120	1	point12-1	1679	3196	4434	3977	3518	4188		point12-1	1087	1291	1122	870	610	593

Supplemental Figure 1. Analysis of *LOC100288798* processing in additional cell lines.



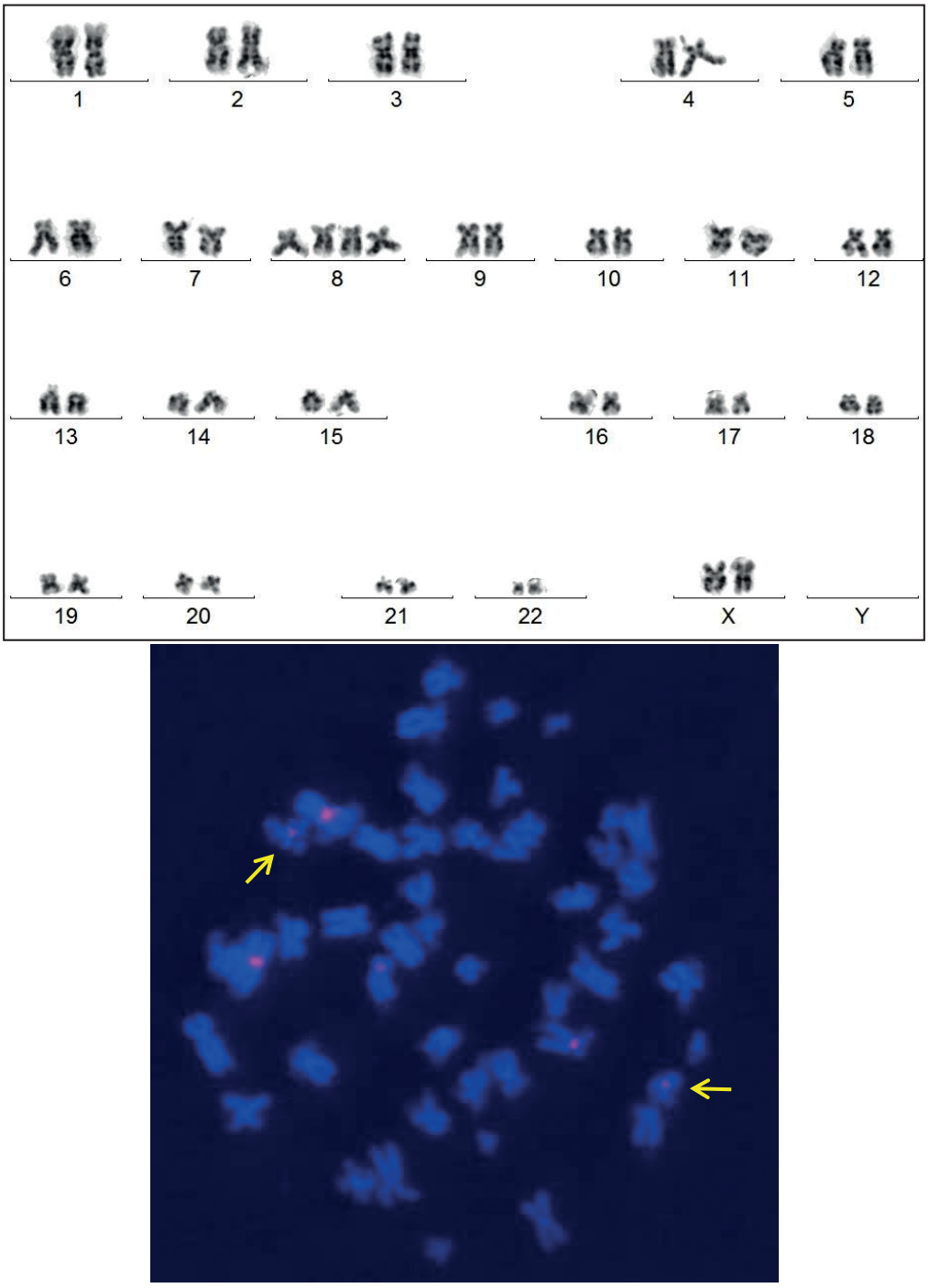
Analyzing ENCODE RNA-seq from different cell and RNA fractions confirms inefficient processing and distinct biology of spliced and unspliced isoforms of *LOC100288798* (marked as "lncRNA") in multiple cell types (Results, Fig. 1E,F): from top to bottom - Lymphoblastoid Cell line GM12878, Human Embryonic Stem Cells, HUVEC cell line and K562 cell line. Bar plots on the left show PolyA+ enrichment as described for Figure 1E and bar plots on the right show nuclear enrichment as described for Figure 1F for the four genes in the four cell lines. Embryonic stem cells and HUVEC do not express *XIST* lncRNA and thus there are no bars corresponding to *XIST* for these two cell lines

Supplemental Figure 2. Chromosome analysis of WT2 KBM7 cell line



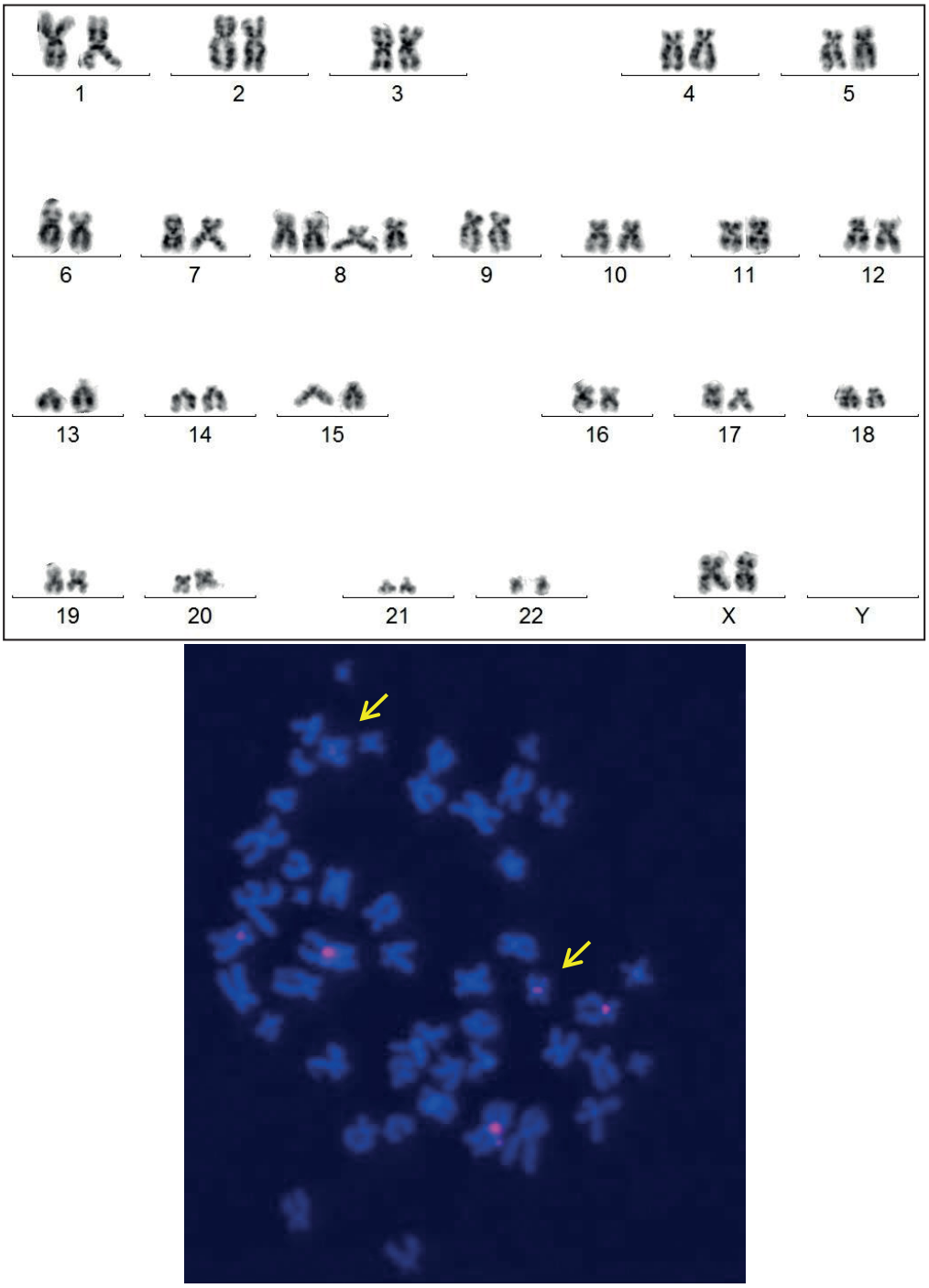
Top: Karyogram showing a haploid set of chromosomes (without the Y-chromosome), except for chromosome 8 which is disomic as reported before (details see text). Bottom: Because chromosomes 16 and 19 display a similar size in "G-bands produced with trypsin and Giemsa" (GTG) banding analysis we performed FISH analysis (Fluorescent In Situ Hybridisation) on metaphase chromosomes using a probe mix that label the centromere regions of chromosomes 1, 5 und 19. The result indicates the presence of both chromosomes 19 and 16. Yellow arrow indicates chromosome 19.

Supplemental Figure 3. Chromosome analysis of C1 KBM7 cell line



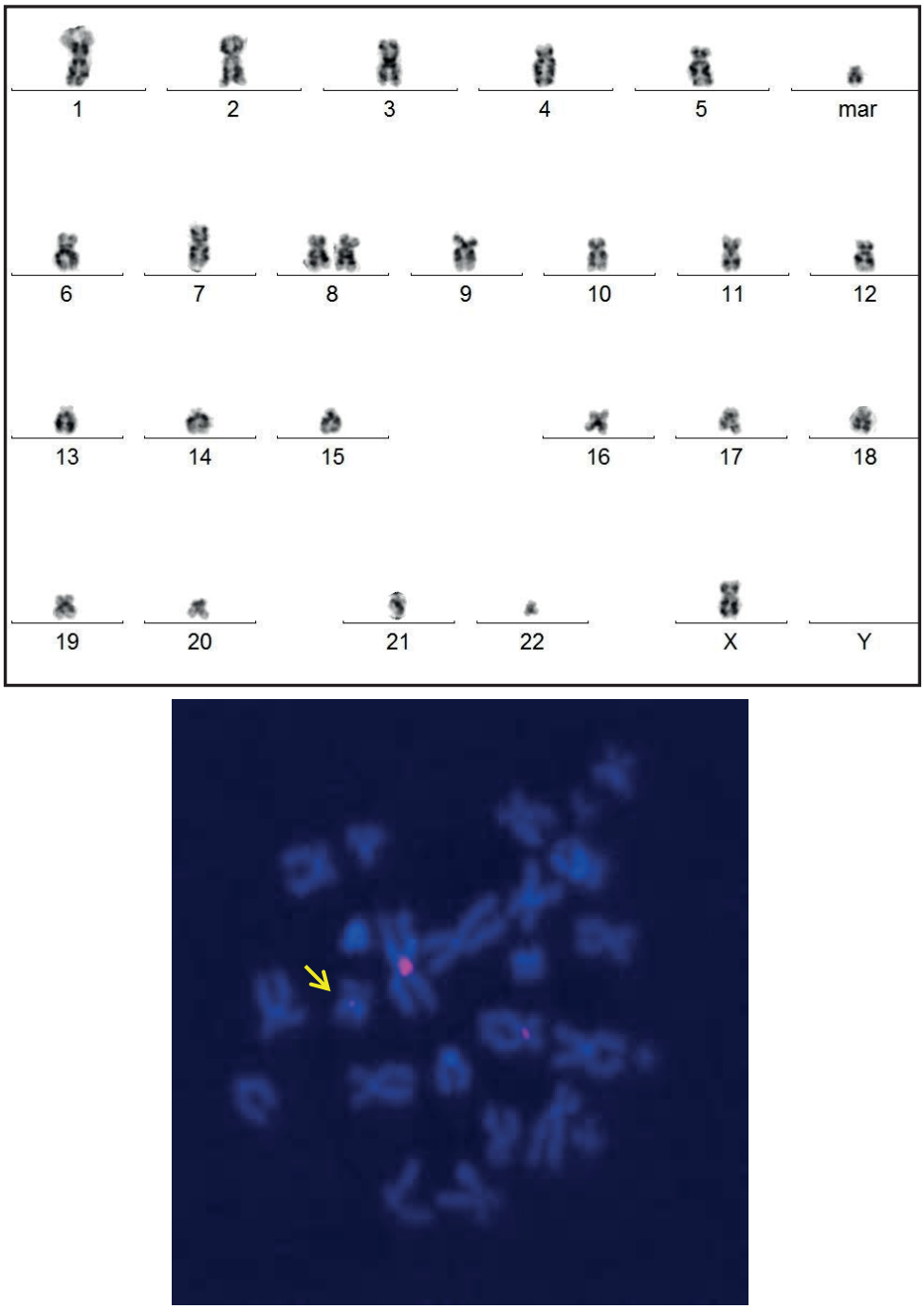
Top: Karyogram showing a diploid karyotype. Note that chromosome 8 is tetrasomic, as diploid KBM7 cells result from endoreduplication of haploid KBM7 cells, where chromosome 8 is disomic (details see text). Bottom: As chromosomes 16 and 19 could not be visually distinguished by GTG-banding analysis we performed FISH analysis of metaphase chromosomes with centromeric probes for chromosomes 1, 5 und 19. This analysis indicated the presence of both chromosomes 19 and 16. Yellow arrows indicate chromosomes 19.

Supplemental Figure 4. Chromosome analysis of 3kb2 KBM7 cell line



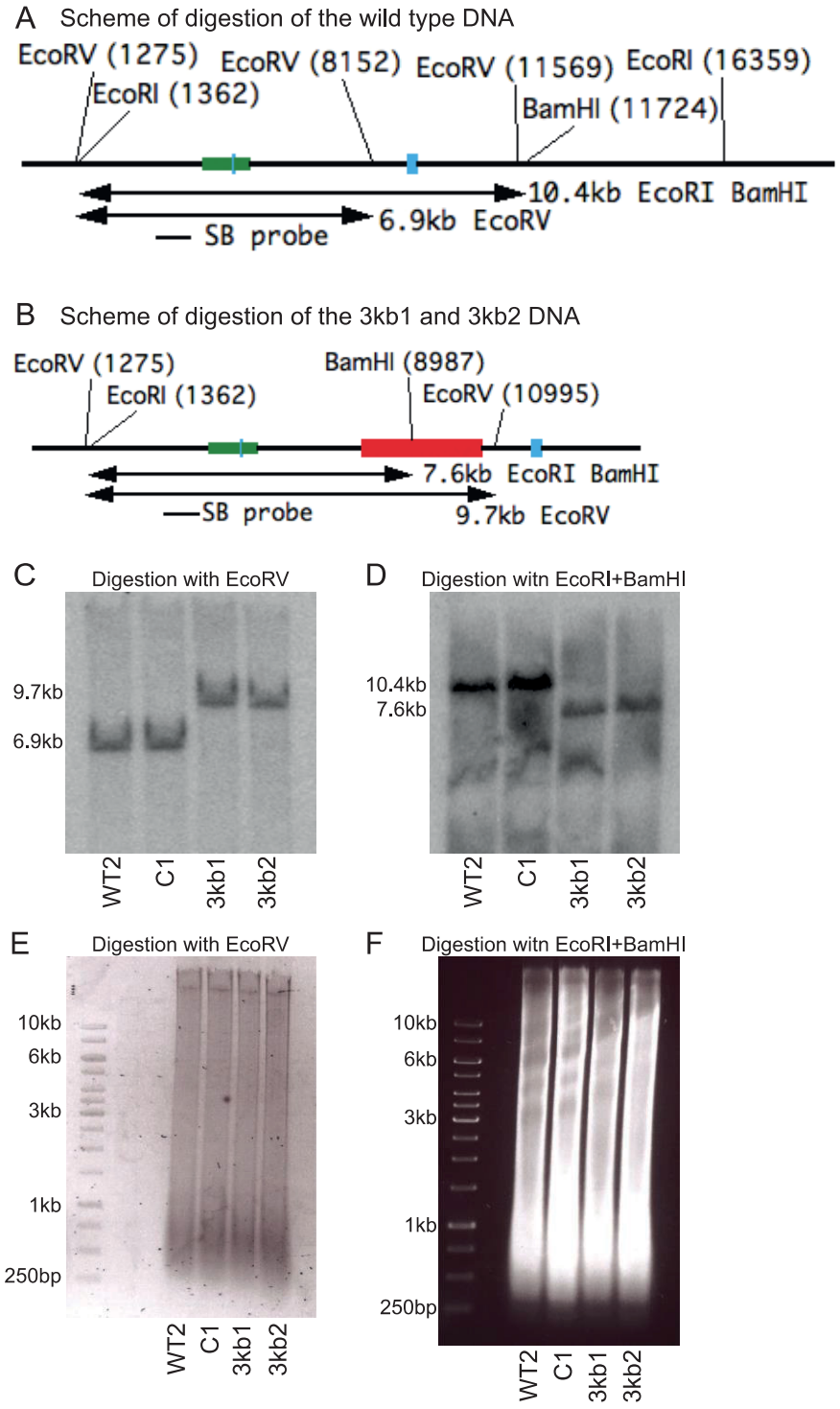
Top: Karyogram showing diploid chromosome set including all chromosomes, except Y. Note that chromosome 8 is tetrasomic, as diploid KBM7 cells result from endoreduplication of haploid KBM7 cells, where chromosome 8 is disomic (details see text). Bottom: As chromosomes 16 and 19 could not be visually distinguished by GTG-banding analysis we performed FISH analysis of metaphase chromosomes with centromeric probes for chromosomes 1, 5 und 19. This analysis indicated the presence of both chromosomes 19 and 16. Yellow arrows indicate chromosomes 19.

Supplemental Figure 5. Chromosome analysis of 100kb1 KBM7 cell line



Top: Karyogram showing a haploid set of chromosomes (without the Y-chromosome), except for chromosome 8 which is disomic as reported before (details see text). Bottom: Because chromosomes 16 and 19 display a similar size in GTG banding analysis we performed FISH analysis (Fluorescent In Situ Hybridisation) on metaphase chromosomes using a probe mix that label the centromere regions of chromosomes 1, 5 und 19. The result indicates the presence of both chromosomes 19 and 16. Yellow arrow indicates chromosome 19.

Supplemental Figure 6. Integrity of the locus remains upon 3kb1 and 3kb2 gene trap insertions

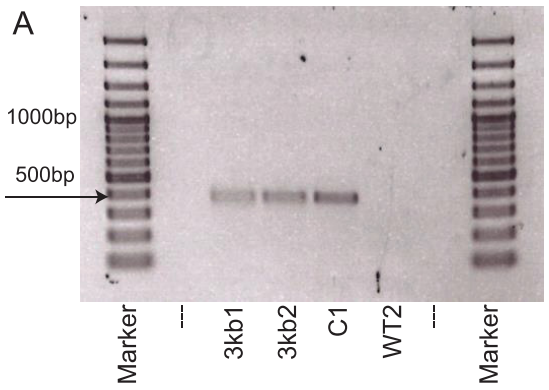


DNA-blot assay to validate the integrity of the genomic locus after the gene trap insertion in 3kb1 and 3kb2 KBM7 clones.

DNA-blot assay design: (**A**) wild type allele (displayed region - chr12:46,772,535-46,791,476), (**B**) 3kb gene trap insertion allele (displayed region - chr12:46,772,535-46,784,103). Genomic regions (h19) downloaded from the UCSC genome browser are shown with the positions (relative to the region start)

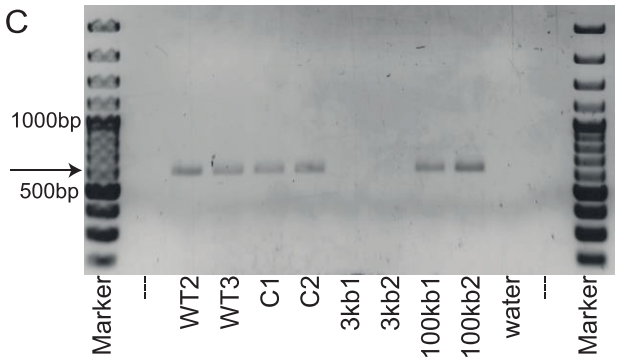
and names of the selected restriction enzyme cutting sites. Below the scheme of the region the sizes of restriction fragments relevant for the assay are shown together with the position of the 865bp Southern blot probe that was produced with primers SLC38ASBPF/R (forward SLC38ASBPF: CCTTTTCATTTGACCCTGGA, reverse SLC38ASBPR: ACTCAAAGGGGGTTGTTGTG). Green box: CpG island promoter of *SLC38A4-AS*, light blue boxes: exons 1 and 2 of *SLC38A4-AS*. Red box: gene trap cassette sequence (only present in **B** that describes the 3kb truncation allele).
(**C**) DNA blot of genomic DNA from WT2, C1, 3kb1 and 3kb2 cell lines digested with EcoRV enzyme, transferred to a membrane and hybridized with the probe indicated in **A** and **B**.
(**D**) DNA blot of genomic DNA from WT2, C1, 3kb1 and 3kb2 cell lines digested with EcoRI and BamHI enzyme, transferred to a membrane and hybridized with the probe indicated in **A** and **B**.
Note that in all cases the expected band sizes were obtained and that the small differences between neighboring bands result from unequal separation of genomic DNA in different wells of the agarose gel as shown by the respective agarose gel picture stained with ethidium bromide, obtained before transfer of the DNA to the membrane (**E** corresponding to **C**, and **F** corresponds to **D**).

Supplemental Figure 7. Splice acceptor sequence validation and wild type cell contamination test in 3kb1 and 3kb2 KBM7 cell lines



B

```
3kb1_trap_genetrap_F   AAGTTATGATCGCAGGCGCAATCTTCGCATTTCTTTTTTCCAGATGGTGAGCAAGGGCG
3kb2_trap_genetrap_F   AAGTTATGATCGCAGGCGCAATCTTCGCATTTCTTTTTTCCAGATGGTGAGCAAGGGCG
C1_trap_genetrap_F     AAGTTATGATCGCAGGCGCAATCTTCGCATTTCTTTTTTCCAGATGGTGAGCAAGGGCG
splice_acceptor              ------GATCGCAGGCGCAATCTTCGCATTTCTTTTTTC----------
                             **********************************
```

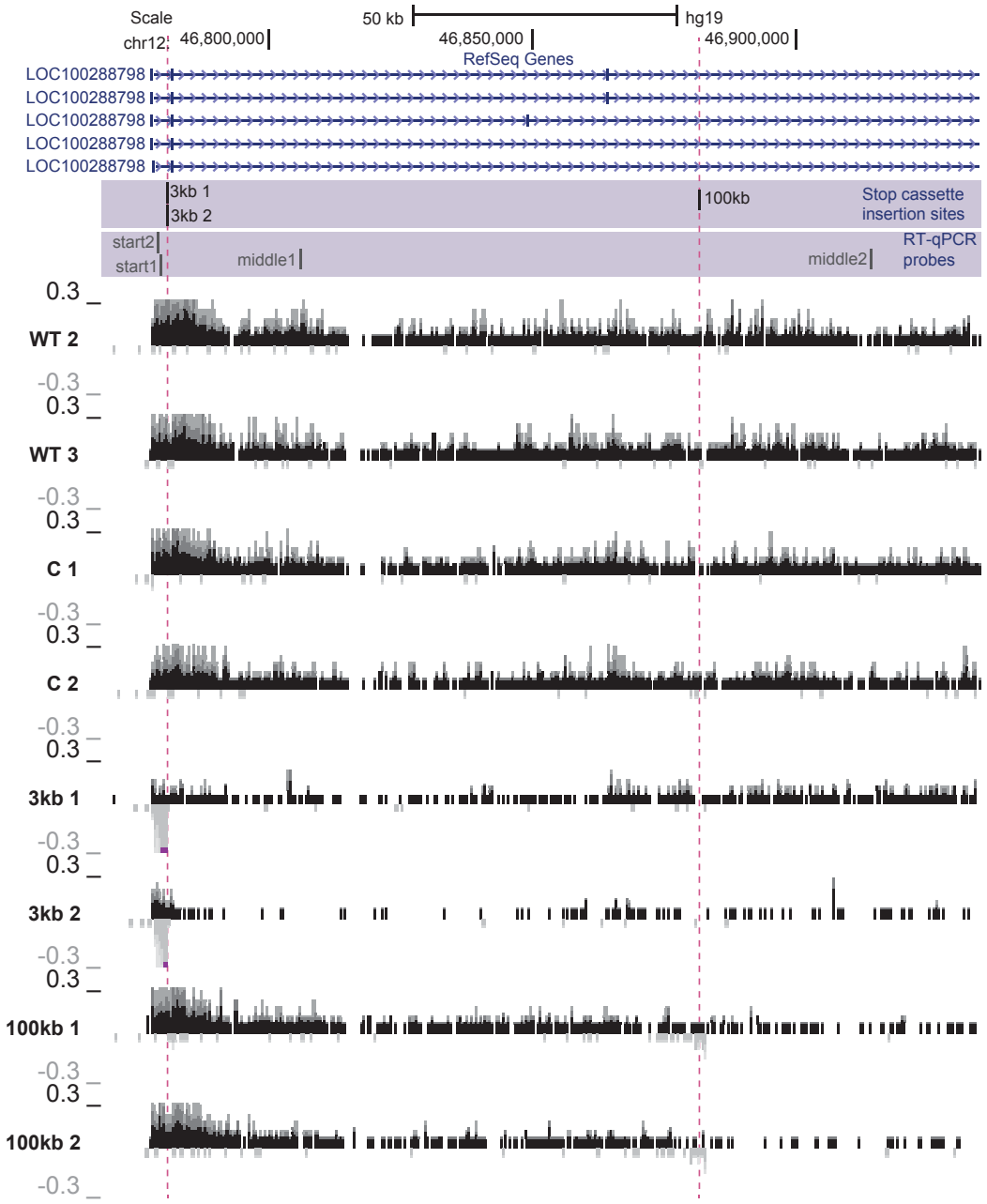


(**A**) PCR amplifying a 385bp long fragment of gene trap cassette containing the splice acceptor site (primers: forward – CCTACAGGTGGGGTCTTTCA, reverse - AAGTCGTGCTGCTTCATGTG) was performed on genomic DNA from 3kb1, 3kb2, C1 and WT2. The resulting fragment amplified by PCR was of the expected size (indicated by the horizontal arrow).

(**B**) Part of the sequence, obtained by Sanger sequencing of the PCR fragment shown in **A** with the indicated forward primer is displayed together with the splice acceptor sequence used in the gene trap cassette. Stars indicate matched bases.

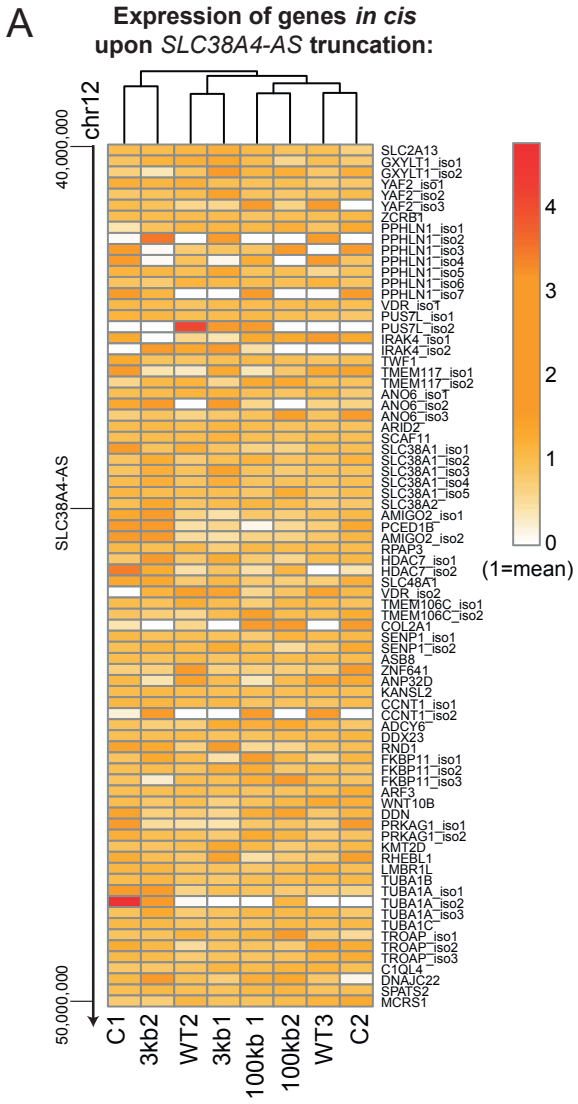
(**C**) PCR flanking the insertion site of the 3kb gene trap cassette insertion (primers: forward – TCAAAGTGTCTGCTGTTAGGTTG, reverse - TATTGCCTCCACAGCTCAAA) was performed on genomic DNA from WT2, WT3, C1, C2, 3kb1, 3kb2, 100kb1 and 100kb2 KBM7 cell lines. The gel picture shows that in all samples, except for 3kb1 and 3kb2, a PCR produced of expected fragment size was obtained: 608bp (indicated by the horizontal arrow). In 3kb1 and 3kb2 cell lines the size of the region targeted by the PCR primers is increased to 3,453bp by the insertion of the gene trap cassette and thus is too long to be amplified in a standard PCR reaction. The absence of signal in 3kb1 and 3kb2 samples indicates lack of detectable wild type cell contamination in both of these cell lines.

Supplemental Figure 8. Emergence of antisense transcription at the site of gene trap cassette insertion



Inspection of RNA-Seq signal of the eight clones reveals emergence of antisense transcription initiating at gene trap insertion sites. Top: chromosome coordinates, RefSeq annotation corresponding to first 150kb of *LOC100288798*, location of gene trap insertion sites, location of RT-qPCR probes. Bottom: RNA-seq signal, normalized to sample read number, pink dots indicate RNA-seq signal that exceeds the range presented inside the box. Name of the cell line is indicated on the left. Vertical dashed red lines indicate position of the 3kb and 100kb stop cassettes. Low density of RNA-seq signal piles indicate low expression and the smallest size corresponds to 1 read.

Supplemental Figure 9. Truncation of *SLC38A4-AS* lncRNA does not affect genes *in cis*



Heat map shows expression level (FPKM, Methods) of genes (name indicated on the right, "iso" stands for isoform when more than one isoform is displayed) in the 10Mbp region around *SLC38A4-AS* lncRNA transcription start site in the four truncation cell lines and the four control cell lines. Expression values are normalized to the mean FPKM among all 8 samples. Mean is set to 1. Only genes with mean FPKM > 1 are displayed: 47 genes (78 isoforms). Heat map color legend is displayed on the right. Heat map was built in R using *pheatmap* function with options *clustering_distance_cols= "canberra", clustering_distance_rows= "euclidean"*.

Publication 3 Supplemental Table 1A-1 - Overview of the publicly available RNA-seq datasets used in the study

Dataset #	label in Fig1B and 1C	RNA fraction	Sample	type	source	library type	read length, bp	read type	Number of uniquely mapped reads, mln	download command (not fully displayed)
1	Adult Brain(C,P)	PolyA+	Adult Brain	primary tissue	Cabili et al	non-strand-specific	75	PE	26.22	wget ftp://ftp-trace
2	Adult Colon(B,P)	PolyA+	Adult Colon	primary tissue	Human Bodymap project	non-strand-specific	50	PE	69.69	wget ftp://ftp-trace
3	Adult Heart(B,P)	PolyA+	Adult Heart	primary tissue	Human Bodymap project	non-strand-specific	50	PE	73.03	wget ftp://ftp-trace
4	Adult Kidney(B,P)	PolyA+	Adult Kidney	primary tissue	Human Bodymap project	non-strand-specific	50	PE	66.08	wget ftp://ftp-trace
5	Adult Liver(B,P)	PolyA+	Adult Liver	primary tissue	Human Bodymap project	non-strand-specific	50	PE	69.88	wget ftp://ftp-trace
6	Adult Liver(C,P)	PolyA+	Adult Liver	primary tissue	Cabili et al	non-strand-specific	75	PE	16.41	wget ftp://ftp-trace
7	Adult Lung(B,P)	PolyA+	Adult Lung	primary tissue	Human Bodymap project	non-strand-specific	50	PE	67.82	wget ftp://ftp-trace
8	Adult Skeletal Muscle (B,P)	PolyA+	Adult Sceletal Muscle	primary tissue	Human Bodymap project	non-strand-specific	50	PE	71.86	wget ftp://ftp-trace
9	Aortic Adventitial Fibroblasts(E,t)	Total	Aortic adventitial fibroblasts	primary cell type	ENCODE project	strand-specific	100	PE	296.7	wget http://hgdown
10	Aortic Endothelial Cells(E,t)	Total	Aortic endothelial cells	primary cell type	ENCODE project	strand-specific	100	PE	326.9	wget http://hgdown
11	B cells(E,P)	PolyA+	B cells	primary cell type	ENCODE project	strand-specific	75	PE	140.4	wget http://hgdown
12	Breast cancer MCF7(E,P)	PolyA+	MCF7 cell line	malignant cell line	ENCODE project	strand-specific	75	PE	228.9	wget http://hgdow
13	CD34 Mobilized(E,t)	Total	Mobilized CD34+ cells	primary cell type	ENCODE project	strand-specific	100	PE	204.8	wget http://hgdow
14	Dermal Fibroblasts(E,t)	Total	Dermal fibroblasts	primary cell type	ENCODE project	strand-specific	100	PE	337.8	wget http://hgdown
15	Embryonic Stem Cells(E,P)	PolyA+	Embryonic stem cells	primary cell type	ENCODE project	strand-specific	75	PE	203.4	wget http://hgdown
16	Epidermal Keratinocytes(E,P)	PolyA+	Epidermal keratinocytes (Nhek cell line)	normal cell line	ENCODE project	strand-specific	75	PE	241.9	wget http://hgdown
17	Epidermal Melanocytes(E,t)	Total	Epidermal melanocytes	primary cell type	ENCODE project	strand-specific	100	PE	260.2	wget http://hgdown
18	Fetal Lung Fibroblasts(E,P)	PolyA+	Fetal lung fibroblasts	primary cell type	ENCODE project	strand-specific	75	PE	197.0	wget http://hgdown
19	Follicle Dermal Papilla Cells(E,t)	Total	Follicle dermal papilla cells	primary cell type	ENCODE project	strand-specific	100	PE	255.9	wget http://hgdown

Publication 3 Supplemental Table 1A-2 - Overview of the publicly available RNA-seq datasets used in the study

Dataset #	label in Fig1B and 1C	RNA fraction	Sample	type	source	library type	read length, bp	read type	Number of uniquely mapped reads, mln	download command (not fully displayed)
20	HeLa(E,P)	PolyA+	HeLa cell line	malignant cell line	ENCODE project	strand-specific	75	PE	211.0	wget http://hgdowr
21	HUVEC(E,P)	PolyA+	HUVEC cell line	normal cell line	ENCODE project	strand-specific	75	PE	175.2	wget http://hgdowr
22	IMR90(E,P)	PolyA+	IMR90 cell line	normal cell line	ENCODE project	strand-specific	100	PE	165.0	wget http://hgdowr
23	IMR90(E,t)	Total	IMR90 cell line	normal cell line	ENCODE project	strand-specific	100	PE	217.4	wget http://hgdowr
24	K562(E,P)	PolyA+	K562 cell line	malignant cell line	ENCODE project	strand-specific	75	PE	193.3	wget http://hgdowr
25	Lung Fibroblasts(E,P)	PolyA+	Lung fibroblasts	primary cell type	ENCODE project	strand-specific	75	PE	245.0	wget http://hgdowr
26	Lymphoblastoid cells GM12878(E,P)	PolyA+	Lymphoblastoid cell line	primary cell type	ENCODE project	strand-specific	75	PE	216.9	wget http://hgdowr
27	Mammary Epithelial Cells(E,P)	PolyA+	Mammary epithelial cells	primary cell type	ENCODE project	strand-specific	75	PE	115.0	wget http://hgdowr
28	Mammary Epithelial Cells(E,t)	Total	Mammary epithelial cells	primary cell type	ENCODE project	strand-specific	100	PE	135.6	wget http://hgdowr
29	Mes.Stem Cells Adipose(E,t)	Total	Undifferentiated mesenchymal stem cells from subcutaneous abdomen adipose tissue	primary cell type	ENCODE project	strand-specific	100	PE	256.1	wget http://hgdowr
30	Mes.Stem Cells Bone Marrow(E,t)	Total	Undifferentiated mesenchymal stem cells from bone marrow	primary cell type	ENCODE project	strand-specific	100	PE	337.6	wget http://hgdowr
31	Mes.Stem Cells Umbical Cord(E,t)	Total	Undifferentiated mesenchymal stem cells from umbical cord	primary cell type	ENCODE project	strand-specific	100	PE	221.0	wget http://hgdowr
32	MNC Pheriph.Blood (B,P)	PolyA+	Mononuclear cells from Peripheral Blood	primary cell type	Human Bodymap project	non-strand-specific	50	PE	70.29	wget ftp://ftp-trace

Publication 3 Supplemental Table 1A-3 - Overview of the publicly available RNA-seq datasets used in the study

Dataset #	label in Fig1B and 1C	RNA fraction	Sample	type	source	library type	read length, bp	read type	Number of uniquely mapped reads, mln	download command (not fully displayed)
33	MNC Periph.Blood(E,t)	Total	Mononuclear cells from Peripheral Blood	primary cell type	ENCODE project	strand-specific	100	PE	204.0	wget http://hgdowr
34	MNC Umbical Cord(E,t)	Total	Mononuclear cells from umbical cord blood	primary cell type	ENCODE project	strand-specific	100	PE	185.0	wget http://hgdow
35	Monocytes(E,P)	PolyA+	Monocytes	primary cell type	ENCODE project	strand-specific	75	PE	170.8	wget http://hgdowr
36	Osteoblasts(E,t)	Total	Undifferentiated osteoblasts	primary cell type	ENCODE project	strand-specific	100	PE	370.9	wget http://hgdowr
37	Placenta(C,P)	PolyA+	Placenta	primary tissue	Cabili et al	non-strand-specific	75	PE	24.87	wget ftp://ftp-trace
38	Placental Epithelial Cells(E,t)	Total	Placental epithelial cells	primary cell type	ENCODE project	strand-specific	100	PE	285.7	wget http://hgdowr
39	Saphenous Vein Endothelial Cells(E,t)	Total	Saphenous vein endothelial cells	primary cell type	ENCODE project	strand-specific	100	PE	235.3	wget http://hgdowr
40	Skeletal Muscle Myoblasts(E,P)	PolyA+	Skeletal muscle myoblasts	primary cell type	ENCODE project	strand-specific	75	PE	219.9	wget http://hgdowr
41	Skeletal Striated Muscle Cells(E,t)	Total	Skeletal striated muscle cells	primary cell type	ENCODE project	strand-specific	100	PE	175.9	wget http://hgdowr
42	Skin Fibroblast(E,P)	PolyA+	Skin fibroblasts	primary cell type	ENCODE project	strand-specific	75	PE	191.6	wget http://hgdowr
43	Testes (C,P)	PolyA+	Testes	primary tissue	Human Bodymap project	non-strand-specific	50	PE	71.23	wget ftp://ftp-trace
44	Testes(B,P)	PolyA+	Testes	primary tissue	Cabili et al	non-strand-specific	75	PE	26.31	wget ftp://ftp-trace
45	Undifferentiated Chondrocytes(E,t)	Total	Undifferentiated chondrocytes	primary cell type	ENCODE project	strand-specific	100	PE	364.8	wget http://hgdowr
46	Undifferentiated White Preadipocytes(E,t)	Total	Undifferentiated white preadipocytes	primary cell type	ENCODE project	strand-specific	100	PE	305.2	wget http://hgdowr
						average read number			185.78	
						maximal read number			370.90	
						minimal read number			16.41	

Publication 3 Supplemental Table 1B

Overview of the ENCODE cell/RNA fractionation RNA-seq datasets used in the study

Data set #	Cell type	Cell fraction	RNA fraction	Number of uniquely mapped reads, mln	Raw RNA-seq fastq files names* (not fully displayed)
1	Lymphoblastoid cells GM12878	Whole cell	PolyA-	87.0	wgEncode wgEncodeCs
2	Lymphoblastoid cells GM12878	Whole cell	PolyA+	100.6	wgEncode wgEncodeCs
3	Lymphoblastoid cells GM12878	Cytosol	PolyA-	63.6	wgEncode wgEncodeCs
4	Lymphoblastoid cells GM12878	Cytosol	PolyA+	111.2	wgEncode wgEncodeCs
5	Lymphoblastoid cells GM12878	Nucleus	PolyA-	89.3	wgEncode wgEncodeCs
6	Lymphoblastoid cells GM12878	Nucleus	PolyA+	113.1	wgEncode wgEncodeCs
7	Embryonic Stem Cells	Whole cell	PolyA-	76.7	wgEncode wgEncodeCs
8	Embryonic Stem Cells	Whole cell	PolyA+	92.5	wgEncode wgEncodeCs
9	Embryonic Stem Cells	Cytosol	PolyA-	40.2	wgEncode wgEncodeCs
10	Embryonic Stem Cells	Cytosol	PolyA+	85.6	wgEncode wgEncodeCs
11	Embryonic Stem Cells	Nucleus	PolyA-	74.0	wgEncode wgEncodeCs
12	Embryonic Stem Cells	Nucleus	PolyA+	86.9	wgEncode wgEncodeCs
13	HeLa cell line	Whole cell	PolyA-	83.0	wgEncode wgEncodeCs
14	HeLa cell line	Whole cell	PolyA+	104.2	wgEncode wgEncodeCs
15	HeLa cell line	Cytosol	PolyA-	50.4	wgEncode wgEncodeCs
16	HeLa cell line	Cytosol	PolyA+	92.0	wgEncode wgEncodeCs
17	HeLa cell line	Nucleus	PolyA-	65.4	wgEncode wgEncodeCs
18	HeLa cell line	Nucleus	PolyA+	92.1	wgEncode wgEncodeCs
19	HUVEC cell line	Whole cell	PolyA-	63.5	wgEncode wgEncodeCs
20	HUVEC cell line	Whole cell	PolyA+	96.2	wgEncode wgEncodeCs
21	HUVEC cell line	Cytosol	PolyA-	58.0	wgEncode wgEncodeCs
22	HUVEC cell line	Cytosol	PolyA+	100.1	wgEncode wgEncodeCs
23	HUVEC cell line	Nucleus	PolyA-	89.0	wgEncode wgEncodeCs
24	HUVEC cell line	Nucleus	PolyA+	105.9	wgEncode wgEncodeCs
25	K562 cell line	Whole cell	PolyA-	85.5	wgEncode wgEncodeCs
26	K562 cell line	Whole cell	PolyA+	94.9	wgEncode wgEncodeCs
27	K562 cell line	Cytosol	PolyA-	55.4	wgEncode wgEncodeCs
28	K562 cell line	Cytosol	PolyA+	109.1	wgEncode wgEncodeCs
29	K562 cell line	Nucleus	PolyA-	92.4	wgEncode wgEncodeCs
30	K562 cell line	Nucleus	PolyA+	104.1	wgEncode wgEncodeCs
31	Epidermal keratinocytes (Nhek cell line)	Whole cell	PolyA-	78.3	wgEncode wgEncodeCs
32	Epidermal keratinocytes (Nhek cell line)	Whole cell	PolyA+	241.7	wgEncode wgEncodeCs
33	Epidermal keratinocytes (Nhek cell line)	Cytosol	PolyA-	38.8	wgEncode wgEncodeCs
34	Epidermal keratinocytes (Nhek cell line)	Cytosol	PolyA+	174.6	wgEncode wgEncodeCs
35	Epidermal keratinocytes (Nhek cell line)	Nucleus	PolyA-	134.3	wgEncode wgEncodeCs
36	Epidermal keratinocytes (Nhek cell line)	Nucleus	PolyA+	196.1	wgEncode wgEncodeCs
		average read number		95.16	
		maximal read number		241.74	
		minimal read number		38.78	

* Raw RNA-seq fastq files were downloaded from
http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCshlLongRnaSeq/

all samples:

library type	read length, bp	read type
strand-specific	75	PE

Publication 3 Supplemental Table 1D

Overview of KBM7 RNA-seq data obtained in the study

name of KBM7 cell line	Number of input reads	Uniquely mapped reads	% of input reads	Number of splices	% of input reads	Reads mapped to multiple loci	% of input reads
WT 2	44,532,797	34,171,152	76.7	1,193,957	2.7	9,454,116	21.2
WT 3	45,996,815	34,848,847	75.8	1,229,945	2.7	10,184,775	22.1
C 1	47,209,693	35,620,249	75.5	1,191,740	2.5	10,639,825	22.5
C 2	45,898,289	35,617,539	77.6	1,268,746	2.8	9,350,100	20.4
3kb 1	44,041,581	36,184,996	82.2	1,239,308	2.8	7,019,131	15.9
3kb 2	44,783,498	33,008,978	73.7	1,149,820	2.6	10,856,634	24.2
100 kb 1	46,095,996	35,545,970	77.1	1,221,678	2.7	9,631,867	20.9
100 kb 2	47,153,016	35,531,201	75.4	1,210,838	2.6	10,610,173	22.5

Publication 3 Supplemental Table 1F - Mock differential gene expression analyses

Comparing different combinations of mixed samples Mock comparisons

Control 1	1	2	**Control 2**	1	2	**Control 3**	1	2	**Control 4**	1	2
1	WT2	WT3	1	WT3	WT2	1	WT2	WT3	1	WT2	WT3
2	C1	C2	2	C1	C2	2	C2	C1	2	C1	C2
3	3kb1	3kb2	3	3kb1	3kb2	3	3kb1	3kb2	3	3kb2	3kb1
4	100kb1	100kb2	4	100kb1	100kb2	4	100kb1	100kb2	4	100kb1	100kb2
Control 5	1	2	**Control 6**	1	2	**Control 7**	1	2	**Control 8**	1	2
1	WT2	WT3	1	WT3	WT2	1	WT3	WT2	1	WT3	WT2
2	C1	C2	2	C2	C1	2	C1	C2	2	C1	C2
3	3kb1	3kb2	3	3kb1	3kb2	3	3kb2	3kb1	3	3kb1	3kb2
4	100kb2	100kb1	4	100kb1	100kb2	4	100kb1	100kb2	4	100kb2	100kb1
Control 9	1	2	**Control 10**	1	2	**Control 11**	1	2			
1	WT2	WT3	1	WT2	WT3	1	WT2	WT3			
2	C2	C1	2	C2	C1	2	C1	C2			
3	3kb2	3kb1	3	3kb1	3kb2	3	3kb2	3kb1			
4	100kb1	100kb2	4	100kb2	100kb1	4	100kb2	100kb1			

	Differentially expressed genes with 3-fold change										
n	control 1	control 2	control 3	control 4	control 5	control 6	control 7	control 8	control 9	control 10	control 11
1	none	NM_000584	NM_000397	NM_000397	none	NM_000397	NM_000397	none	none	NM_000397	NM_000397
2		NM_000700	NM_000655	NM_001080426		NM_001005463	NM_000655			NM_000655	NM_001005463
3		NM_001005463	NM_001080426	NM_001142966		NM_001080426	NM_001080426			NM_001080426	NM_001080426
4		NM_001142966	NM_001142966	NM_001676		NM_001142966	NM_001142966			NM_001142966	NM_001142966
5		NM_001769	NM_001144952	NM_002777		NM_001769	NM_001144952			NM_001144952	NM_014220
6		NM_002975	NM_001676	NM_002975		NM_014220	NM_001676			NM_001759	NM_018003
7		NM_004120	NM_001759	NM_018003		NM_018003	NM_001759			NM_002777	NM_032608
8		NM_014220	NM_001769	NM_032608		NM_032608	NM_002777			NM_002975	NM_144590
9		NM_018003	NM_002777	NM_144590		NM_144590	NM_002975			NM_005797	NM_144646
10		NM_020340	NM_002975	NM_144646		NM_144646	NM_005797			NM_014220	NM_145175
11		NM_144646	NM_005797	NM_145175		NM_145175	NM_014220			NM_018003	
12		NM_152342	NM_006203				NM_018003			NM_144590	
13			NM_018003				NM_144590			NM_144646	
14			NM_144590				NM_144646			NM_145175	
15			NM_144646				NM_145175				
16			NM_145175								
number of genes:	0	12	16	11	0	11	15	0	0	14	10

average number of differentially expressed (with 3-fold change) genes in mock comparisons 8.090909091