universität wien

DISSERTATION

Titel der Dissertation

“Genome-wide investigation of RNA biology features of mouse long non-coding RNAs”

Verfasser

Dipl.-Ing.(FH) Philipp Günzl

angestrebter akademischer Grad

Doktor der Naturwissenschaften (Dr.rer.nat.)

Wien, 2015

Studienkennzahl lt. Studienblatt:	A 091 441
Dissertationsgebiet lt. Studienblatt:	Doktoratsstudium Genetik und Mikrobiologie
Betreuerin:	Prof. Denise P. Barlow, Ph.D.

The work presented in this thesis was performed under supervision and in the laboratory of Dr. Denise Barlow at CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences (Lazarettgasse 14, AKH BT 25.3, 1090 Vienna, Austria) during the period of April 2009 to September 2014.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Table 1: List of abbreviations

Abbreviation	Description
100PE	100bp paired-end RNA-seq
100SE	100bp single-end RNA-seq
50PE	50bp paired-end RNA-seq
ActD	Actinomycin D
bp	Basepair
BR	Biological replicate
CAGE	Cap-analysis gene expression
ceRNA	Competing endogenous RNA
ChIP-seq	Chromatin immunoprecipitation sequencing
CPC	Coding potential calculator (software)
CUT	Cryptic unstable transcript
Cyt	cytoplasmic
ENCODE	Encyclopedia of DNA elements
eRNA	Enhancer lncRNA
EST	Expressed sequence tag
EtOH	Ethanol
GRO-seq	Global run-on sequencing
kb	Kilobasepair
lncRNA	Long non-coding RNA
MEF	Mouse embryonic fibroblasts
mESC	Mouse embryonic stem cells
miRNA	Micro RNA
mRNA	Messenger RNA
ncRNA	Non-coding RNA
NELF	Negative elongation factor
NMD	Nonsense mediated decay
nt	Nucleotides
Nuc	nuclear
PCR	Polymerase chain reaction
PRC	Polycomb repressive complex
PROMPT	Promoter upstream transcript
qPCR	Quantitative real-time PCR
REF	Rat embryonic fibroblasts
rESC	Rat embryonic stem cells
RIP-seq	RNA immunoprecipitation sequencing
RNAi	RNA interference
RPKM	Reads per kilobase of transcript per million reads mapped
rRNA	Ribosomal RNA
RSeQC	RNA-seq Quality Control Package (software)
shRNA	Small hairpin RNA
snRNA	Small nuclear RNA
STAR	Spliced transcripts alignment to a reference (software)
TR	Technical replicate
UCSC	University of California, Santa Cruz
UTR	Untranslated region

1 ZUSAMMENFASSUNG

Die Weiterentwicklung der RNA Sequenzierung hat unser Verständnis von Säugetiergenomen grundsätzlich verändert. Große Teile unseres Genoms wurden lange Zeit nicht beachtet und als “junk DNA” bezeichnet, sie werden tatsächlich aber häufig in nichtkodierende RNA transkribiert. Tausende lange nicht kodierende (lnc)RNAs sind mittlerweile in den Genomen von Mensch und Maus annotiert und eine faszinierende Frage ist, ob all diese lncRNAs funktionell sind. Einige Dutzend lncRNAs sind bereits ausgiebig erforscht worden und es herrscht nun Einigkeit, dass viele davon die Expression von Genen steuern können und damit auf die Entwicklungsbiologie sowie die Entstehung von Krankheiten Einfluss nehmen. Ähnliche lncRNAs wurden in bestimmte Kategorien eingeteilt um mögliche Funktionen besser prognostizieren zu können. Die bekanntesten dieser Kategorien sind intergenische, bidirektionale, antisense und enhancer lncRNAs und jeder davon wurden bestimmte RNA Biologie Eigenschaften und Funktionen zugesagt. Es wurde bereits spekuliert, dass die RNA Biologie jeder lncRNA eine Voraussetzung oder Folge einer bestimmten Funktion sein könnte. Nachdem der Großteil der publizierten RNA Biologie Experimente jedoch mit unterschiedlichen lncRNA Annotationen, abweichenden Methoden und in verschiedenen Zelltypen gemacht wurden, können diese nur schwer miteinander verglichen werden. Daher habe ich Experimente durchgeführt um die drei RNA Biologie Merkmale RNA Stabilität, nukleärer Export und Spleißeffizienz gemeinsam in zwei Zelltypen genomweit untersucht. Ich habe embryonale Stammzellen und embryonale Fibroblasten der Maus als Modelle gewählt und nach RNA Sequenzierung und bioinformatischer Analyse herausgefunden, dass die RNA Biologie von lncRNAs zelltypspezifisch reguliert wird und sehr unterschiedlich zu mRNAs ist. Weiters zeige ich, dass die vier derzeitigen lncRNA Kategorien überraschenderweise sehr ähnliche Eigenschaften haben und weder durch ihre RNA Biologie noch durch ihre genomischen Transkripteigenschaften unterscheidbar sind. Ich habe daher alle lncRNAs nach ihrer RNA Biologie kategorisiert und sechs Gruppen definiert, jede mit einem einzigartigen RNA Biologie Muster. Ich habe dabei herausgefunden, dass etwa die Hälfte ähnliche Eigenschaften wie mRNAs hat und die andere Hälfte ineffizient gespleißt wird, instabil ist und kaum aus dem Nukleus exportiert wird. Zusätzlich habe ich alle Experimente in den entsprechenden Zelltypen der Ratte wiederholt und dabei entdeckt, dass die RNA Biologie Eigenschaften weitgehend zwischen Maus und Ratte konserviert sind. Wie die einzelnen RNA Biologie Eigenschaften von der Zelle reguliert werden ist großteils unbekannt, jedoch zeige ich, dass sie mit einigen genomischen Transkripteigenschaften signifikant korrelieren. Insgesamt habe ich in dieser Studie 76 RNA Sequenzierungsbibliotheken erstellt und 4,1 Milliarden Reads generiert die ich mit 2,3 Milliarden weiteren publizierten Reads analysiert habe. Diese tief sequenzierten und gut kontrollierten Datensätze werden für die wissenschaftliche Gemeinschaft außerordentlich hilfreich sein um lncRNA Forschung voran zu bringen.

1 ABSTRACT

Recent advances in RNA-sequencing changed our perception of mammalian genomes and established that major parts of the previously disregarded “junk DNA” are transcribed into non-coding RNA. Thousands of long non-coding (lnc)RNAs have since been annotated in human and mouse genomes and the most intriguing question is whether they are functional. A few dozen lncRNAs have been studied in detail and they are now increasingly recognized as important gene regulators in development and disease. Similarly to the field of proteomics, related lncRNAs were divided into classes to be able to extrapolate functions. The most prominent classes of intergenic, bidirectional, antisense and enhancer lncRNAs have been assigned certain RNA biology features and functions. It has been speculated that RNA biology might be a prerequisite for or consequence of function. However, the majority of RNA biology experiments was carried out using different lncRNA annotations, variable protocols and different cell types and therefore they are not comparable. Therefore, I conducted experiments to assay the three RNA biology features RNA stability, nuclear export and splicing efficiency of lncRNAs genome-wide using two cell types. I have chosen mouse embryonic cells as the primary model as they express a wealth of lncRNAs and selected embryonic fibroblasts as a second cell type. After RNA-sequencing, bioinformatic analyses and rigorous quality filtering, I find that lncRNA biology is cell type specifically regulated and markedly different from mRNAs. I show that the four current lncRNA classes have very similar features as they are neither distinguishable by their RNA biology features nor by their genomic transcript features. I therefore clustered lncRNAs based on their three RNA biology features and define six clusters, each having a unique RNA biology signature. I find that half of lncRNAs have mRNA-like RNA features whereas the other half is inefficiently spliced, rather unstable and less efficiently exported to the cytoplasm. Additionally, I repeated all experiments in the corresponding cell types of the rat and find that RNA biology features are largely conserved between mouse and rat. How each of the three RNA biology features is regulated by the cell remains unknown, however, they significantly correlate with certain genomic transcript features. All together, I prepared 76 RNA-seq libraries for this study and generated 4.1 billion reads which were analyzed alongside with 2.3 billion reads from published RNA-seq experiments. These deeply sequenced and well controlled datasets will be a valuable and comprehensive resource for the research community to investigate the expression states and RNA features of lncRNAs genome-wide.

2 INTRODUCTION

2.1 LncRNAs are pervasively transcribed in eukaryotic genomes

A fascinating question in biology is how organisms use and regulate their genomes to sustain their life cycle and respond to internal and external signals. While only 1.5% of the human genome code for proteins, the remaining 98.5% of intergenic space were often dismissed as "junk DNA" (Ohno, 1972; Zuckerkandl, 1992). The Human Genome Project nevertheless sequenced the whole genome and laid the foundation to genome-wide analysis of genes and their corresponding RNA transcripts. Mammalian genomes contain ~20,000 protein-coding genes that are dispersed with largely repetitive sequences. The conventional view that the intergenic space contains mainly nontranscribed sequences changed dramatically with the advent of genome-wide transcriptome studies such as cDNA sequencing, tiling arrays and massively parallel sequencing. The number of non-coding (nc)RNAs soon exceeded the number of mRNAs and it became evident that large parts of mammalian genomes are transcribed (Carninci et al., 2005; Djebali et al., 2012; ENCODE Project Consortium et al., 2007). The central dogma "DNA makes RNA and RNA makes protein" (Crick, 1970) has since been enriched by non-coding RNAs that do not need to encode proteins but can be functional by themselves (Wang and Chang, 2011). While it is now accepted that the genome is pervasively transcribed, it has remained an open question whether all ncRNAs are functional or whether many of them represent just spurious transcription that arises from accessible chromatin or in order to keep it accessible (Clark et al., 2011). The classification and nomenclature of ncRNAs is inconsistent, they can be grouped by size (e.g. small RNA, short interfering RNA, micro RNA, long ncRNA), by their interaction partners (e.g. piwi-interacting RNAs), by their cellular localization (e.g. small nucleolar RNA), by their position relative to mRNAs or genetic elements (e.g. bidirectional, antisense, intergenic and enhancer RNA) or by the way they function (e.g. cis-acting and trans-acting RNA). Non-coding RNAs are broadly classified into small (<200nt) and long (>200nt) ncRNAs, a rather arbitrary cutoff that is based on the limitations of current RNA isolation procedures. In this study, I only focused on the group of long non-coding (lnc)RNAs. The known functions of lncRNAs are manifold and beside the fact that many lncRNAs exhibit structural functions, they are now increasingly recognized as a crucial layer of gene regulatory networks (Guttman et al., 2011; Ulitsky and Bartel, 2013; Vance and Ponting, 2014). These regulatory lncRNAs can be grouped into *cis*-acting lncRNAs regulating genes on the same chromosome from where they are expressed and *trans*-acting lncRNAs regulating genes genome-wide by leaving the site of expression and associating with chromatin modifiers and other proteins (Rinn and Chang, 2012). LncRNAs evolved extremely rapidly between the closely related species mouse and rat and up to 61% of lncRNAs have been shown to be unique to the *Mus* genus (Kutter et al., 2012). In the last decade, the methods to annotate lncRNAs genome-wide have significantly improved and accelerated lncRNA research substantially.

2.2 Genome-wide methods to annotate lncRNAs

2.2.1 Transcript mapping by expressed sequence tags (ESTs)

The Sanger sequencing of cDNA libraries yielded enormous amounts of expressed sequence tags (ESTs) that have been instrumental to the discovery of genes in multiple organisms (Nagaraj et al., 2007). ESTs are with 200-800bp relatively short and represent parts of cDNAs that were reverse transcribed from RNA molecules of a particular cell type and organism. When the genome sequence of the organism was available, the ESTs were aligned and the gene structure assembled from many ESTs. Generation of ESTs and computational analysis have improved over the years and allowed more sensitive annotation of transcripts (Dias Neto et al., 2000). The possible applications of ESTs grew with their abundance in databases and soon included gene structure predictions, analysis of alternative splicing, investigation of tissue or disease-specific expression and the discovery and characterization of SNPs (Buetow et al., 1999; Kan et al., 2001; Modrek et al., 2001; Schmitt et al., 1999). In 2003, the FANTOM2 consortium published a mouse full-length cDNA encyclopedia containing 70,000 transcription units that were derived from 1.44 million 3'-end sequences and 0.54 million 5'-end sequences (Carninci et al., 2003). 55% to 65% of these ESTs could be attributed to known protein-coding genes by the BLAST algorithm, however, the majority of the remaining ESTs was attributed to lowly expressed non-coding RNA. The FANTOM3 consortium enhanced EST-based transcript mapping by incorporating cap-analysis gene expression (CAGE) tags to map proper 5'-ends (Carninci et al., 2005). This method yielded ~102,000 cDNAs of which ~34,000 lacked protein-coding sequence and were therefore defined as non-coding RNA, however, many of those were single-exon transcripts. They found many of those ncRNAs being conserved across species, although on average slightly less than 5' or 3' UTRs. In a first conclusion about ncRNA functions, they hypothesize that ncRNA transcription is either important for or a consequence of the underlying genomic location, and that the transcript itself might function by sequence-specific interactions with the DNA sequence from which it is derived (Carninci et al., 2005).

2.2.2 Tiling arrays detect transcribed regions

Tiling arrays are a subtype of microarrays that function by hybridizing chemically labeled RNA or DNA sequences with probes being fixed on glass slides. In contrast to traditional microarrays that are used to investigate known sequences, tiling arrays have been extensively exploited to characterize transcription in particular genomic regions, chromosomes and even genome-wide (Bertone et al., 2004; Cheng et al., 2005; Kampa et al., 2004). A tiling array study interrogating the transcriptomes of human chromosomes 21 and 22 in eleven cell lines found that more than 90% of transcribed nucleotides were located outside of annotated mRNAs (Kampa et al., 2004). They estimated that ten times more genomic DNA was transcribed into RNA as had been known and proposed a re-evaluation of the term

"gene" to incorporate the steadily increasing class of ncRNAs. The first genome-wide human tiling arrays detected ~10,600 previously unknown sequences, many of which were located in intergenic regions (Bertone et al., 2004). A more detailed view on the human transcriptome was provided by high-density tiling arrays (5-nucleotide-resolution) that were probed with polyadenylated and nonpolyadenylated RNA as well as nuclear and cytoplasmic RNA (Kampa et al., 2004). Their main conclusions were that the human transcriptome consists of heavily interlaced transcripts that form complex networks and that approximately 50% of all transcribed sequences are not polyadenylated. Despite unprecedented sensitivity, the main challenge of tiling arrays is definitely to confidently map proper transcript boundaries as well as exon models and to distinguish cross hybridizing sequences, e.g. from pseudogenes, from sequences originating from this position.

2.2.3 Massively parallel RNA-sequencing and transcriptome assembly

RNA-sequencing (RNA-seq) is a recently developed method that allows the detailed study of transcriptomes by using deep sequencing (also known as massively parallel sequencing or next-generation sequencing) of cDNA libraries. These libraries can be generated from RNA of virtually all organisms, RNA of particular cell types or dedicated RNA subsets. Three RNA-seq platforms (Illumina sequencing, Applied Biosystem SOLiD and Roche 454 Life Sciences) have been established, of which Illumina sequencing is by far the most widely used technology (Wang et al., 2009). For a typical Illumina RNA-seq experiment, the RNA sample of choice is fragmented and reverse transcribed into short cDNAs. After the addition of adaptors, the fragments are hybridized onto glass slides and bridge-amplified to give rise to clusters of ~1,000 clonal fragments. In a "sequence-by-synthesis" approach, each fragment is extended nucleotide-by-nucleotide from one side using fluorescently labeled nucleotides while high-throughput cameras capture the fluorescence of the clusters after each cycle (Pettersson et al., 2009). The sequence of each cluster is called by an algorithm that converts the sequence of fluorescent colors into a string of nucleotides. Each cluster gives one read and the number of sequencing cycles determines the length of the read. A single-end read is derived from only one end of the cDNA fragment while a paired-end read consists of two paired reads that come from both ends of the cDNA fragment. The cost of sequencing per base has steadily decreased over the last twenty years, with an even accelerated decrease after 2008 (Hayden, 2014). This price drop increased possible applications of RNA-seq and made it possible to use RNA-seq to compare the transcriptomes of thousands of healthy people and patients with a previously unprecedented depth. To increase the coverage of lowly abundant lncRNAs in RNA-seq experiments, tiling arrays were used to capture lncRNAs by hybridization and subsequent elution to enrich them ~380-fold (Mercer et al., 2012). The number of possible RNA-seq applications is ever increasing and includes now, among others, gene fusion detection, alternative splicing analysis, identification of disease relevant RNAs and single-cell transcriptomics (Ozsolak and Milos, 2011).

The generation of large RNA-seq datasets required the development of dedicated and sophisticated software solutions to process, quality-test and analyze data. Typically, RNA-seq data is aligned to a reference genome and expression levels quantified for a reference annotation. For each step, multiple algorithms are available and testing multiple parameters of numerous programs can be cumbersome. Over the years, few programs were more frequently used than others and became the standard for most applications. The alignment program Bowtie (Langmead et al., 2009) has frequently been used for short reads, however, with the increase of read length ultrafast programs such as TopHat (Trapnell et al., 2009) and STAR (Dobin et al., 2013) allowed *de novo* splice junction annotation. The assembly of aligned reads into transcripts is done by programs such as Scripture (Guttman et al., 2010) and more commonly Cufflinks (Trapnell et al., 2010, 2013), which can also directly estimate transcript abundances and calculate differential expression. Packages such as RSeQC help to analyze RNA-seq quality, strand specificity, GC bias, coverage uniformity and can also be used to estimate transcript abundances (Wang et al., 2012).

The first RNA-seq mouse transcriptome was generated from ~140 million 25bp reads of three mouse tissues (Mortazavi et al., 2008). While the overall coverage was low, they detected previously unknown mRNA promoters, exons and 3' UTRs as well as novel lncRNAs bearing miRNAs. In order to annotate lncRNAs, several pipelines were developed to filter out transcripts with protein-coding potential, however, most of them were based on assaying sequence features and reading frames (Kong et al., 2007) and protein-coding potential across multiple species (Washietl et al., 2011). The first genome-wide lncRNA annotation from RNA-seq data was generated from three mouse cell types (Guttman et al., 2010). Since then, RNA-seq significantly accelerated the progress of lncRNA research and within a few years lncRNA annotations of dozens of organisms were generated (Ulitsky and Bartel, 2013). RNA-seq experiments further helped to investigate different lncRNA classes, lncRNA biology and genome-wide transcriptome changes within development and disease as well as upon perturbations by drugs, genetic knock-outs and other treatments. Huge consortia such as ENCODE extensively used RNA-seq to show that 74.7% of the human genome is transcribed in any of the 15 investigated cell types (Djebali et al., 2012) and to evaluate lncRNA structure, evolution and expression (Derrien et al., 2012). The most common way to classify lncRNAs today is based on their position relative to genetic elements such as mRNAs and enhancers, however, the functions and RNA biology of these lncRNA classes might not be as homogenous as previously assumed.

2.3 LncRNAs are currently classified by position relative to mRNAs

2.3.1 Intergenic lncRNAs

The class of intergenic lncRNAs (also known as large intervening (linc)RNAs) has first been defined by two landmark papers that used H3K4me3 and H3K36me3 chromatin maps to

annotate ~1,600 intergenic lncRNAs in four mouse cell types (Guttman et al., 2009) and ~3,300 intergenic lncRNAs in six human cell types (Khalil et al., 2009). While this approach could not resolve exact exon structures and transcript boundaries, RNA-seq data in combination with advanced bioinformatic tools was later able to reconstruct complete transcriptomes with exact exon models *ab initio*. This powerful technique yielded thousands of novel and known intergenic lncRNAs across three mouse cell types (Guttman et al., 2010) and ~8,000 intergenic lncRNAs across 24 human tissues (Cabili et al., 2011). In these studies, the class of intergenic lncRNAs basically constitutes all assembled non-coding transcripts that are not in the vicinity of mRNAs (Figure 1). While this approach was useful to avoid complex loci with overlapping transcription and to simplify downstream bioinformatic analysis, it did not recapitulate the full spectrum of lncRNAs. Intergenic lncRNAs are therefore considered as a "catch-all class" that is "defined more by what they are not than by what they are" (Ulitsky and Bartel, 2013).

While dozens of lncRNAs classes with distinct functions, RNA biology features and expression patterns may exist, this intergenic subclass long served as a prototype for lncRNAs and insights from the study of intergenic lncRNAs were often projected to be true for all lncRNAs. Intergenic lncRNAs were largely annotated from fully processed polyA+ RNA, which led to the perception that they are frequently spliced (Moran et al., 2012; Ulitsky and Bartel, 2013). While intergenic lncRNAs were early declared to be evolutionary conserved and therefore considered as functional (Guttman et al., 2009; Khalil et al., 2009), it seems that sequence conservation has been overestimated for most of them (Ulitsky and Bartel, 2013). It is now accepted that exons of intergenic lncRNAs are more conserved than random intergenic regions but significantly less than mRNA exons (Derrien et al., 2012; Ulitsky and Bartel, 2013).

Several subsets of intergenic lncRNAs have been defined based on size and functions. For example, more than 2,000 very long intergenic (vlinc)RNAs with lengths from 50kb to 700kb were annotated in healthy and cancerous human cells, many of which are expressed from retroviral promoters (St Laurent et al., 2013). Cis-acting enhancer-like lincRNAs were found by RNAi knock-downs of seven transcripts from a set of ~3,000 manually curated human lincRNAs (Ørom et al., 2010). While it is certainly convenient to study the diverse class of intergenic lncRNAs as a whole, it seems that uncovering the full spectrum of functions and features will require a more differentiated approach and a more sophisticated way of analysis.

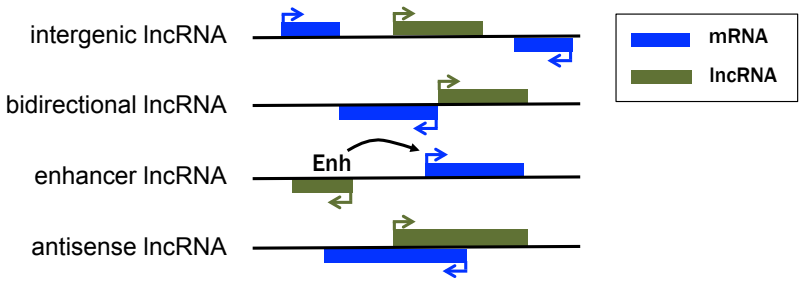


Figure 1: LncRNAs are currently classified by their position relative to mRNAs

Depicted is a schematic representation of the four currently used lncRNA subclasses. Intergenic lncRNAs are transcribed from the intergenic space and therefore do not overlap mRNAs. Bidirectional lncRNAs are expressed from a bidirectional promoter shared with an mRNA. Enhancer of mRNAs are often transcribed and give rise to enhancer lncRNAs. Antisense lncRNAs overlap mRNAs in antisense orientation. Details see text. Enh, enhancer.

2.3.2 Bidirectional lncRNAs

Bidirectional lncRNAs are transcribed from shared mRNA promoters in antisense orientation to the respective mRNA (Neil et al., 2009; Seila et al., 2008) (Figure 1). The transcript start sites of lncRNA and mRNA within this bidirectional CpG island promoter are thereby less than 1kb apart. Bidirectional promoters are abundant in mammalian genomes and have been suggested to provide a mechanism of endogenous gene regulation (Trinklein et al., 2004). Since the discovery that bidirectional promoters most frequently consist of one mRNA and one lncRNA (Katayama et al., 2005), it has been discussed whether bidirectional lncRNAs are involved in the regulation of their corresponding mRNA partner by keeping the chromatin accessible or whether they represent transcriptional noise being generated as a consequence of open chromatin (Brosius, 2005; Kowalczyk et al., 2012). In mouse ES cells it has been shown that divergently transcribed mRNA/lncRNA pairs are coordinately regulated through differentiation (Sigova et al., 2013). These lncRNAs have a median length of 2.7kb, are capped as well as polyadenylated and only half of their loci contained spliced transcripts. GRO-seq data indicated that transcription levels of paired mRNAs and lncRNAs are similar, however, the reduced stability of bidirectional lncRNAs led to ~10-fold lower steady-state levels. A detailed study of four bidirectional lncRNAs indicated that they are regulated by the positive transcription elongation factor (P-TEFb) and that they are rapidly degraded by the exosome (Flynn et al., 2011). It has also been shown that mRNAs that share their promoters with lncRNAs exhibit reduced noisy expression (Wang et al., 2011b). The explanation is that the bipromoter architecture enables transcription of the mRNA and the lncRNA to facilitate a constantly open chromatin and transcription factor binding. This allows more constant transcription compared to transcriptional bursts when previously inactive chromatin suddenly becomes active. In terms of gene regulatory functions, most studies detected a co-regulation of divergently transcribed mRNAs and lncRNAs, thereby establishing a gene-activating function for the majority of bidirectional lncRNAs (Uesaka et al., 2014). The transcription of bidirectional lncRNAs could be the cause for accessible chromatin or the consequence, in

both cases the transcript itself might not be functional. Bidirectional lncRNAs could also tether chromatin modifiers or RNAPII to regulate local chromatin and transcription *in cis*, thereby making the RNA molecule the functional entity. However, as an exception to proposed *cis*-acting mechanisms, the bidirectional lncRNA *Six3OS* has been found to modulate retinal development by concordantly regulating the neighboring mRNA *Six3 in trans* (Rapicavoli et al., 2011). In conclusion, bidirectional lncRNAs are mostly considered unstable and both the act of divergent lncRNA transcription and the bidirectional lncRNAs themselves could be functional.

2.3.3 Enhancer lncRNAs

Enhancer lncRNAs (also know as eRNAs) are generated by transcription of enhancer elements (Figure 1) and seem to be a hallmark of most active mammalian enhancers (Core et al., 2008; Seila et al., 2008). They are not to be confused with enhancer-like lncRNAs that form a group of intergenic lncRNAs and have been shown to activate the expression of neighboring mRNAs (Ørom et al., 2010). Enhancer lncRNAs were first described in mouse neuronal cells (Kim et al., 2010) and activated mouse macrophages (De Santa et al., 2010). In the former study, neuronal activity-regulated enhancers were defined by binding of the transcriptional co-activator p300/CBP and H3K4me1, which marks active chromatin regions such as enhancers and promoters. To remove uncharacterized promoters, genetic elements with H3K4me3 marks were removed. ChIP-seq data indicated RNAPII binding at ~3000 (25%) of these enhancer elements and RNA-seq data verified that ~2000 of them are actually bidirectionally transcribed, giving rise to transcripts with <2kb length. Transcription of these non-polyadenylated eRNAs was co-regulated with the expression of nearby target genes. The second study detected low-level intergenic transcription upstream of activated genes and found that they correspond to enhancer elements defined by p300 peaks, high H3K4me1 and low H3K4me3 (De Santa et al., 2010). These eRNAs are polyadenylated and were estimated to have a median length of ~500nt. Actinomycin D treatment indicated that they are very unstable and it was suggested that the exosome degrades them. Another study detected eRNAs in androgen-sensitive human prostate cancer cells at enhancers that regulate expression changes upon androgen stimulation (Wang et al., 2011a). It was soon speculated that these eRNAs are functional parts in the gene regulatory networks (Orom and Shiekhattar, 2011).

Enhancer lncRNAs were also detected by GRO-seq in human MCF7 cells after induction of estrogen signaling (Li et al., 2013b). Knock-downs of three eRNAs by miRNAs and locked nucleic acids (LNAs) reduced eRNA levels post-transcriptionally while nascent transcription levels were unchanged, which significantly impaired the activation of the neighboring target gene. The GAL4-Box-b tethering based reporter assay (Wang et al., 2011c) was used to independently verify that the RNA molecule is the functional entity in activating target gene expression rather than the process of transcription. In an attempt to elucidate the mechanism

of action, they show that eRNAs function primarily *in cis* by contributing to the generation and stabilization of Cohesin-mediated chromosomal loops between the estrogen-responsive enhancers and the promoter of the target mRNA genes (Li et al., 2013b). Enhancer lncRNAs have also been shown to activate gene expression by establishing chromatin accessibility of distinct regulatory regions (Mousavi et al., 2013) and to be required for p53-dependent enhancer activity (Melo et al., 2013). A recent study proposes that eRNAs could be involved in the transition of paused to elongating RNAPII complexes (Schaukowitch et al., 2014). This step is usually inhibited by the negative elongation factor (NELF) complex, which contributes to the genome-wide regulatory mechanism of RNAPII pausing. Enhancer RNAs were shown to directly bind the NELF complex and remove it from the RNAPII complex, thereby enabling its transition to productive RNA elongation. Knock-down of eRNAs did reduce target gene expression *in cis* but did not abolish chromosomal looping between the enhancer and the promoter.

While eRNAs are now accepted to mark active enhancers, the question is still whether all eRNAs have a function themselves or whether many of them represent transcriptional noise of open chromatin. This model proposes that chromosomal looping brings together the enhancer and the target promoter with high concentrations of RNAPII, thereby leading to random bidirectional transcription of the accessible enhancer. Open for debate is also the question whether the RNA product or the progression of RNAPII with its associated enzymes is important for enhancer function. Most probably, the class of eRNAs is not a homogenous functional class but there are numerous different functions of eRNAs. To reflect different properties, enhancer lncRNAs were subclassed into unidirectionally transcribed and polyadenylated eRNAs termed 1D-eRNAs and the bidirectionally non-polyadenylated transcribed 2D-eRNAs (Natoli and Andrau, 2012).

2.3.4 Antisense lncRNAs

Antisense lncRNAs overlap mRNAs in antisense orientation and have by definition a transcript complementarity to mRNAs (Figure 1). They are also called *cis*-natural antisense transcripts (*cis*-NATs) in order to distinguish them from exogenous antisense transcripts such as siRNAs and components of the *trans*-acting RNAi machinery. Antisense lncRNAs and mRNAs can overlap themselves by their 5'-ends (head to head), by their 3'-ends (tail to tail) or one can overlap the other completely. These sense/antisense pairs have very early sparked interest as they may provide a mechanism to regulate mRNA translation and mRNA stability (Mizuno et al., 1984). Among the first examples that antisense lncRNAs regulate the expression of their overlapped mRNAs was the imprinted lncRNA *Airn* (Sleutels et al., 2002). The first genome-wide analysis of antisense transcription from the FANTOM3 consortium identified thousands of sense/antisense pairs and expression profiling indicated that pairs can be concordantly or discordantly regulated (Katayama et al., 2005). Coexpression of mRNA and lncRNA would require that they use the same enhancer while differential expression can

be explained by separate regulation of mRNA and lncRNA or when one regulates the other. Another systematic study created sense/antisense libraries from HeLa cells, cloned and sequenced these self-hybridized sequences and found the lncRNA *p15AS* (also known as *ANRIL*) to overlap and epigenetically silence *p15* in leukemia (Yu et al., 2008). The early RNA-seq studies mostly dismissed antisense lncRNAs due to their often complex overlap with mRNAs and rather focused on intergenic lncRNAs that were easier to study (Cabili et al., 2011; Guttman et al., 2009). The RNA biology of mouse or human antisense lncRNAs as a group has, to my knowledge, never been thoroughly investigated. The mechanisms how antisense lncRNAs regulate their overlapped mRNAs are manifold, in the case of *Airn* it has been shown that *Airn* transcription through the overlapped *Igf2r* promoter represses *Igf2r* (Latos et al., 2012). The transcription factor PU.1 (encoded by the *SPI1* gene) has been shown to be regulated by antisense lncRNAs that modulate mRNA translation (Ebralidze et al., 2008). In a similar example, the nuclear localized antisense lncRNA *UCHL1-AS1* becomes exported upon Rapamycin treatment to promote the translation of *UCHL1* in dopaminergic neurons (Carrieri et al., 2012). *UCHL1-AS1* exert its function in the cytoplasm and is a perfect example of how cellular localization of lncRNAs is specifically regulated. The tumor suppressor gene *p21* has been shown to be repressed by its antisense lncRNA *p21-AS* by directing Ago1-dependent H3K27me3 to the *p21* promoter (Morris et al., 2008). In summary, antisense lncRNAs are a heterogeneous group that only share positional antisense overlap with mRNAs, however, their mode of gene regulation is very diverse.

2.4 LncRNAs are important gene regulators

2.4.1 LncRNAs are involved in genomic imprinting

The discovery that most imprinted regions express at least one lncRNA early raised questions whether they are implicated in the regulation of genomic imprinting (Koerner et al., 2009). The four imprinted lncRNAs *Xist*, *Airn*, *Kcnq1ot1* and *Nespas* have been shown to epigenetically repress mRNA genes *in cis* and were, together with *H19*, among the first confirmed functional lncRNAs. Their extensive study unraveled not only different silencing mechanisms but also provided details about lncRNA regulation and RNA biology.

The *Xist* lncRNA (18kb cDNA in mouse) in involved in the female process of dosage compensation which limits the expression of X-linked genes to only one of the two X chromosomes. In female ES cells and embryos prior to X inactivation, *Xist* is lowly expressed from both X chromosomes and rapidly degraded (Sheardown et al., 1997). Upon differentiation, the *Xist* lncRNA from the future inactive X chromosome is stabilized and spreads *in cis* across it by exploiting the three-dimensional structure of the X chromosome (Engreitz et al., 2013). *Xist* thereby recruits the polycomb repressive complex 2 (PRC2) to establish a transcriptionally silent nuclear compartment that is heavily enriched for the repressive histone modification H3K27me3 to silence the whole chromosome (Plath et al.,

2003). *Xist* is exclusively localized to the nucleus and is also well spliced, the latter of which could be a consequence of its evolution from a protein-coding gene that pseudogenized (Duret et al., 2006). The RNA half life of *Xist* is tightly regulated and lies between ~30min in XX ES cells and 5-7h in XX somatic cells with established X chromosome inactivation (Sheardown et al., 1997). The function of *Xist* therefore strongly correlates with its RNA stability. *Xist* is overlapped by the antisense lncRNA *Tsix*. In mice, but not in humans, *Tsix* induces repressive chromatin at the *Xist* promoter on the active X chromosome to prevent *Xist* expression and X chromosome inactivation as a means of generating paternal specific or imprinted X chromosome inactivation in some mouse extra-embryonic tissues (Sado et al., 2005).

The imprinted lncRNA *Airn* is expressed from the paternal chromosome and overlaps the *Igf2r* promoter in antisense orientation (Figure 2A). On the maternal chromosome, a repressive methylation imprint generated in the oocyte prevents *Airn* expression. *Airn* is an atypical RNA transcript as it is ~118kb long and mostly unspliced (Seidl et al., 2006). Only ~5% of nascent and ~35% of steady-state RNA levels are spliced (to ~1kb) and exported to the cytoplasm, while the unspliced isoform remains in the nucleus. Furthermore, unspliced *Airn* is unstable and has a half life of only ~90min, in contrast to the ~16h half life of the spliced products. *Airn* has been shown to repress paternal expression of *Igf2r* in all cell types and the two upstream genes *Slc22a2* in all and *Slc22a3* in some extra-embryonic tissues. The replacement of the *Airn* promoter with a constitutive promoter did not abolish silencing properties or alter the RNA biology of *Airn*, indicating that the functional regions are located within the *Airn* gene body (Stricker et al., 2008). Truncation experiments stopping *Airn* transcription after it had passed the *Igf2r* promoter also did not stop *Igf2r* repression, however, when *Airn* transcription was truncated before it reaches the *Igf2r* promoter, *Igf2r* was re-expressed from the paternal chromosome (Latos et al., 2012). In a final experiment, the *Airn* promoter was moved next to the *Igf2r* promoter and *Igf2r* was still silenced. This established that the RNA sequence is dispensable for *Igf2r* silencing and that the act of *Airn* transcription through the *Igf2r* promoter is sufficient to repress *Igf2r* (Latos et al., 2012). It has been hypothesized that the lack of *Airn* splicing, export and stabilization is explainable by the fact that its transcription is functional rather than the RNA molecule (Guenzl and Barlow, 2012; Pauler et al., 2007; Santoro and Pauler, 2013).

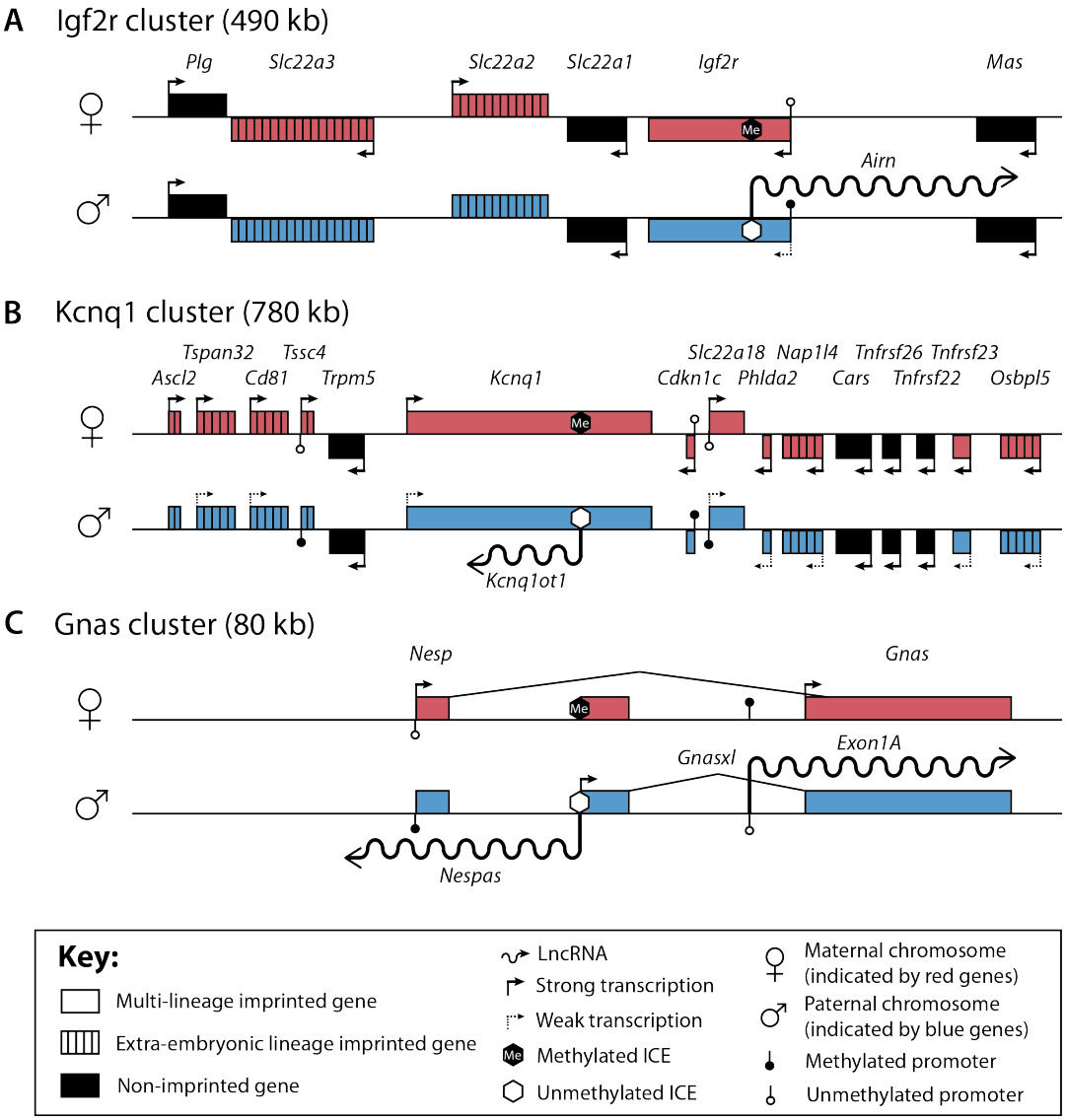


Figure 2: Imprinted lncRNAs repress neighboring mRNAs *in cis*

(**A**) The lncRNA *Airn* overlaps the promoter of the *Igf2r* mRNA in antisense orientation. Expression of *Airn* on the paternal chromosome leads to a repression of *Igf2r* expression *in cis*. *Slc22a2* and *Slc22a3* are not overlapped by Airn and are paternally repressed only in some extra-embryonic tissues. (**B**) The lncRNA *Kcnq1ot1* is expressed from the paternal chromosome from an intron of the mRNA gene *Kcnq1*. *Kcnq1ot1* silences *Kcnq1* and several neighboring genes *in cis*. (**C**) The lncRNA *Nespas* is expressed from the paternal chromosome antisense to *Nesp* and represses *Nesp* expression *in cis* by a yet unknown mechanism. Details see text. Clusters are not drawn to scale. Red boxes: imprinted expression from the maternal chromosome. Blue boxes: imprinted expression from the paternal chromosome. Genes that are bi-allelically expressed are depicted as black boxes. Open colored boxes: genes with multi-lineage imprinted expression. Colored boxed with black stripes: genes whose imprinted expression is restricted to the extra-embryonic lineage. Genes above the line are expressed from the (+) strand, genes below the line are expressed from the (-) strand. Arrows show transcriptional direction. Solid arrows indicate strong transcription whereas dashed arrows indicate weak transcription. LncRNAs are shown as wavy lines. See key for further details. ICE, imprint control element. Figure modified from Guenzl and Barlow, 2012.

The imprinted lncRNA *Kcnq1ot1* is expressed from the paternal chromosome and is located in antisense orientation within an intron of *Kcnq1* (Figure 2B). On the maternal chromosome, a repressive methylation imprint generated in the oocyte prevents *Kcnq1ot1* expression. Similar to *Airn*, *Kcnq1ot1* is mostly unspliced (~83kb), nuclear localized and has a relatively

short half life of <4h (Mohammad et al., 2010). It silences *in cis* the overlapped gene *Kcnq1* and the three upstream genes *Cdkn1c*, *Slc22a18* and *Phlda2* in multiple tissues and additional six genes in extra-embryonic tissues only (Mancini-Dinardo et al., 2006). In extra-embryonic tissues, *Kcnq1ot1* has been shown to interact with the histone methyltransferase G9a and the PRC2 complex in a lineage-specific manner to deploy repressive H3K9me3 and H3K27me3 chromatin modifications throughput the cluster (Pandey et al., 2008). However, as these repressive marks are not found at imprinted genes in the embryo (Mager et al., 2003; Wagschal et al., 2008), it could be that *Kcnq1ot1* also functions by transcriptional interference of *cis*-regulatory elements (Koerner et al., 2009). In fact, post-transcriptional depletion of *Kcnq1ot1* by RNAi did not affect imprinting maintenance in stem cells, further arguing that the RNA molecule might be dispensable for the regulation of imprinted genes (Golding et al., 2011).

The imprinted lncRNA *Nespas* is expressed from the paternal chromosome in antisense orientation to the mRNA gene *Nesp* (encoded by the *Gnas* gene) (Figure 2C). Similar to *Airn* and *Kcnq1ot1*, the maternal *Nespas* promoter carries a repressive methylation imprint established in the oocyte (Williamson et al., 2002). *Nespas* has been shown to repress *Nesp* in multiple tissues. The *Nespas* gene produces spliced isoforms (~2.2kb) as well as unspliced isoforms (~30kb), however, the ratio is unknown so far. It is functioning in the nucleus and its RNA stability is unknown. It is still unclear whether the act of *Nespas* transcription or the *Nespas* lncRNA itself is functional in repressing *Nesp* expression (Williamson et al., 2011).

The imprinted lncRNA *H19* is expressed from the maternal chromosome and is markedly different from the previously described imprinted lncRNAs. It is very mRNA-like as it is efficiently spliced to give rise to a ~2.2 kb lncRNA, exported to the cytoplasm and stable (Brannan et al., 1990). *H19* expression on the paternal chromosome is repressed by a DNA methylation imprint acquired in the sperm. *H19* is one of the most abundant and most conserved lncRNAs and it has been shown to be deregulated in numerous cancers. Furthermore, it acts as a growth repressor to limit the growth of the placenta before birth and has been suggested to have tumor suppressor activity (Hao et al., 1993). While the *H19* lncRNA is dispensable for imprinted expression of *Igf2*, it hosts *miR-675* which is rapidly released in response to stress to inhibit cell proliferation (Keniry et al., 2012).

Meg3 and *Rian* are two imprinted lncRNAs located in the *Dlk1* cluster and expressed from the maternal chromosome (Miyoshi et al., 2000). *Meg3* contains 12 exons and gives rise to several alternatively spliced isoforms that are cell type specific (Zhang et al., 2010). The ENCODE consortium found *Meg3* to be one of the most nuclear-enriched lncRNAs (Derrien et al., 2012). Its expression is implicated in growth suppression and activation of the p53 pathway and is therefore lost in most human tumor cell lines. In addition to these trans-acting functions, *Meg3* has been shown to control expression of the neighboring and oppositely imprinted *Dlk1* gene by recruiting PRC2 *in cis* (Zhao et al., 2010). *Rian* contains 20 exons and gives rise to alternatively spliced isoforms with a length of ~3.6kb. It is nuclear localized and a

CRISPR/Cas9 mediated deletion has shown that it negatively regulates *Dlk1* in brain where it is most highly expressed but not in heart or ovary (Han et al., 2014). *Rian* binds several chromatin modifiers such as PRC1, PRC2, JARID1B and JARID1C and is supposed to be implicated in gene regulatory networks controlling pluripotency and differentiation (Guttman et al., 2011; Zhao et al., 2010). Interestingly, it has been shown that *Meg3* and *Rian* are aberrantly silenced in many induced pluripotent stem (iPS) cell clones which failed to generate viable mice (Stadtfeld et al., 2010). Treatment of these cells with histone deacetylase inhibitors reactivated the locus and led to viable all-iPSC mice. This established that the expression state of a single locus can be used to identify fully reprogrammed ESC-like iPSC clones.

2.4.2 LncRNAs act *in trans* to regulate gene expression genome-wide

Many lncRNAs have been shown to regulate genes *in trans*. Strictly speaking, most likely every cytoplasmic RNA molecule exerts at least minor *trans*-effects by offering microRNA binding sites (Figure 3A). MicroRNA binding, whether specific or unspecific, reduces available microRNAs from the intracellular pool and could lead to expression changes of other microRNA targets. The resulting large-scale regulatory network forms the basis for the competing endogenous (ce)RNA hypothesis (Salmena et al., 2011). While many of these regulations may be unspecific, it has been shown that pseudogenes use this mechanism to specifically regulate their coding counterpart (Poliseno et al., 2010) and a lncRNA has also been shown to sponge specific microRNAs to regulate transcription factors involved in muscle differentiation (Cesana et al., 2011). However, many lncRNAs also exert *trans*-effects by recruiting chromatin modifiers and targeting them to distant loci to modify local chromatin environment and thereby gene expression (Figure 3B) (Khalil et al., 2009; Zhao et al., 2010). Genome-wide investigations of lncRNAs using RNA immunoprecipitation (RIP)-Chip (Khalil et al., 2009) and RIP-Seq (Zhao et al., 2010) found that hundreds of lncRNAs recruit PRC2, a chromatin modifying complex primarily involved in establishing repressive H3K27me3 marks, to repress multiple genes *in trans*. LncRNAs have also been shown to extensively bind active chromatin modifiers, such as WDR5, a component of the MLL histone methyltransferase complex driving H3K4me3 (Yang et al., 2014). While all of these studies suggest *trans*-effects, only one study excludes possible *cis*-effects by checking expression levels of at least 10 neighboring genes of respective lncRNAs in either direction (Khalil et al., 2009). Well-studied examples of *trans*-acting regulatory lncRNAs involved in fundamental cellular processes include among others *HOTAIR*, *PCAT-1*, *MALAT1*, *Braveheart*, *Fendrr*, *lincRNA-p21* and *Firre*.

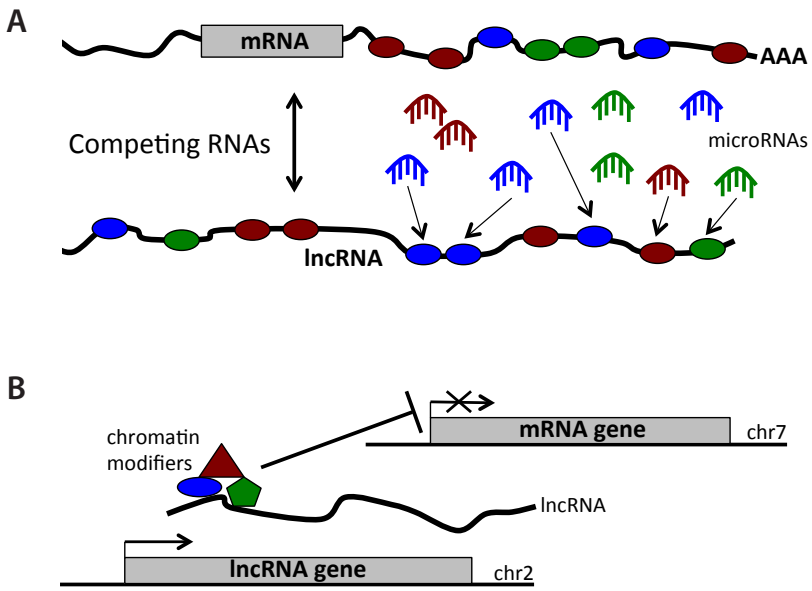


Figure 3: LncRNAs act *in trans* to regulate mRNAs

(**A**) A lncRNA and an mRNA share microRNA response elements (MREs, colored ovals) and therefore compete for the same microRNAs. If microRNAs preferentially bind the lncRNA, the mRNA will not be repressed any longer. (**B**) A lncRNA being expressed from chromosome 2 binds chromatin modifiers and targets them to an mRNA locus on chromosome 7 to regulate its expression *in trans*. Details see text.

Human *HOTAIR* is a 2.2kb spliced transcript and was among the first examples of lncRNAs that exert genome-wide regulatory effects in development and disease (Gupta et al., 2010; Rinn et al., 2007). *HOTAIR* is expressed from the *HOXC* cluster and represses via PRC2 targeting multiple *HOXD* genes spread across 40kb *in trans* (Rinn et al., 2007). During breast cancer progression, *HOTAIR* has been shown to be increasingly upregulated and a powerful predictor for metastasis and death (Gupta et al., 2010). Forced *HOTAIR* overexpression induced PCR2 occupancy and H3K27me3 at >800 genes, of which 39% showed altered gene expression levels. A handful of them were previously described to inhibit breast cancer progression and were downregulated upon *HOTAIR* overexpression. *HOTAIR* also served as a prototype for the hypothesis that lncRNAs can act as molecular scaffolds binding multiple chromatin modifiers to affect chromatin environment at target genes (Tsai et al., 2010). While the 3' domain binds the LSD1/CoREST/REST complex and the 5' domain of *HOTAIR* binds PRC2, this coupled recruitment of chromatin modifiers leads to coordinated H3K4me3 demethylation and gain of H3K27me3, respectively. However, unexpectedly, mouse *Hotair* is poorly conserved and a deletion of *Hotair* along with eight *Hoxc* genes did not lead to expression changes or altered chromatin patterns in *Hoxd* genes (Schorderet and Duboule, 2011). A targeted deletion of *Hotair* led to marginal de-repression of *Hoxd* genes and weak skeletal malformations (Li et al., 2013a). This discrepancy argues that lncRNAs might have different functions in mouse and humans or that a battery of genetic tests is required to confidently determine lncRNA functions (Bassett et al., 2014). The intergenic lncRNA *PCAT-1* was identified as one of 121 prostate cancer associated lncRNAs (PCATs) in a large-scale

RNA-seq study of 102 prostate tissues and cell lines (Prensner et al., 2011). Expression of *PCAT-1* is repressed by PRC2 but was upregulated in metastatic cancers and activated cell proliferation. The expression pattern of the 121 PCATs could distinguish benign, localized cancers and metastatic cancers and it was hypothesized that they could further be used to stratify patients by urine-based assays (Prensner et al., 2011).

MALAT1 is an ~8kb mostly unspliced nuclear lncRNA and belongs to the most highly abundant lncRNAs. Its half life ranges from ~9h to >12h and is conferred by the formation of a stabilizing triple helix (Brown et al., 2014; Tani et al., 2010). It is highly conserved across 33 mammalian species and has been first identified in lung adenocarcinomas (Ji et al., 2003). *MALAT1* is localized to nuclear speckles by two sequence elements (Miyagawa et al., 2012) and has been shown to be cleaved by RNase P at the 3' end to give rise to a small cytoplasmic tRNA-like RNA (Wilusz et al., 2008). Furthermore, *MALAT1* interacts with several splicing factors in nuclear speckles to regulate alternative splicing of pre-mRNAs (Tripathi et al., 2010). A systematic approach investigating *Malat1*-binding RNAs genome-wide confirmed binding of pre-mRNAs, however, it seems to be an indirect interaction mediated by proteins (Engreitz et al., 2014). *MALAT1* is frequently mutated, deleted or overexpressed in numerous human tumor entities, as reflected by its name "metastasis associated lung adenocarcinoma transcript 1" (Gutschner et al., 2013a). It has been shown that *MALAT1* binds PRC1 to modulate the three-dimensional localization of genes in the nucleus (Yang et al., 2011a) and it is indeed a potent regulator binding to hundreds of genomic loci, preferentially active genes (West et al., 2014). Knock-down of *MALAT1* in a human lung tumor cells using zinc finger-mediated RNA destabilization resulted in impaired migration and less tumor formation in mouse xenografts. In vivo knock-down of *MALAT1* using antisense oligonucleotides (ASOs) prevented metastasis formation after tumor implantation and was suggested to serve as a potential therapeutic approach (Gutschner et al., 2013b). However, despite numerous described functions for *MALAT1*, a knockout in human cells did not show an obvious phenotype and two *Malat1* knockout mouse models indicated that *Malat1* is dispensable for pre- and post-natal development (Eißmann et al., 2012; Zhang et al., 2012). This surprising results shows that genetic tests are required to confidently determine lncRNA functions (Bassett et al., 2014).

The intergenic lncRNA *Braveheart* is a spliced ~0.6kb RNA and has been shown to be required for cardiac lineage commitment in the mouse (Klattenhoff et al., 2013). It directly interacts with Suz12 of PRC2 to regulate a cardiovascular gene network *in trans*. A bidirectional lncRNA named *Fendrr* is implicated in the regulation of mouse heart and body wall development (Grote et al., 2013). It associates with the complexes PRC2 and TrxG/MLL to modulate *in trans* the expression of transcription factors that control cardiac mesoderm differentiation. Mutant *Fendrr* mice exhibit severe developmental defects and *Fendrr* loss has been shown to be perinatal lethal in mice due to multiple organ defects (Grote et al., 2013; Sauvageau et al., 2013). Together, *Braveheart* and *Fendrr* are among the first characterized

lncRNAs to have critical roles in lineage commitment and mammalian development by regulating multiple target genes *in trans* (Srivastava and Cordes Metzler, 2013).

LincRNA-p21 is a 3.1kb spliced and stable (half life >6h) RNA being transcribed ~15kb upstream of the cell cycle regulator *Cdkn1a* (also known as *p21*). It is induced by p53 upon DNA damage and has been shown to interact with the heterogenous nuclear ribonucleoprotein (hnRNP-K) through its 5' end to transcriptionally repress multiple genes *in trans* to facilitate apoptosis (Huarte et al., 2010). *LincRNA-p21* also associates with the RNA binding protein HuR to post-transcriptionally repress mRNA translation (Yoon et al., 2012). Additionally, a cis-regulatory function to activate expression of p21 has also been described for *lincRNA-p21* and will be discussed in the next chapter (Dimitrova et al., 2014).

The intergenic lncRNA *Firre* is an alternatively spliced and strictly nuclear RNA that forms expression foci and remains stable for >6h (Hacisuleyman et al., 2014). *Firre* escapes X chromosome inactivation and interacts with the nuclear matrix factor hnRNPU through a 156bp repeating sequence, which is also required for nuclear localization. *Firre* localizes to a ~5mb domain on the X chromosome *in cis* and to five other chromosomal loci *in trans*, all of which reside in spatial proximity to the *Firre* locus. Deletion of *Firre* resulted in a deregulation of >1000 genes, mouse ESC growth defects and increased Tgfβ signaling, the latter of which is consistent with previous findings that *Firre* strongly represses adipogenesis in mouse preadipocytes (Sun et al., 2013). This is the first known example of a lncRNA that modulates three-dimensional nuclear architecture.

2.4.3 LncRNAs act *in cis* to regulate neighboring gene expression

As already introduced in chapter 2.4.1, several well-studied imprinted lncRNAs have repressive *cis*-regulatory functions on neighboring genes, either by the transcript or by the mechanism of transcriptional interference. LncRNAs might also exert their *cis*-regulatory functions on neighboring genes by the act of transcription to keep chromatin accessible, however, they could also represent mere byproducts of open chromatin. Several lncRNA classes such as enhancer lncRNAs (Kim et al., 2010; Wang et al., 2011a) and upstream antisense RNAs (Core et al., 2008; Seila et al., 2008) correlate positively with expression of nearby mRNAs and are supposed to be the cause or consequence of open chromatin. The class of enhancer-like intergenic lncRNAs has been shown to increase expression of neighboring genes *in cis* by chromosome-looping and recruitment of the Mediator complex (Lai et al., 2013; Ørom et al., 2010). Well-studied examples of *cis*-acting regulatory lncRNAs involved in fundamental cellular processes include among others *Hottip*, *Mira*, *DBE-T*, *ANRIL*, *BACE1-AS*, *ZEB2-AS* and *lincRNA-p21*.

The human lncRNA HOTTIP is 3.8kb long, spliced and expressed from the 5' tip of the HOXA locus and activates the expression of several HOXA genes *in cis* (Wang et al., 2011c). It is mostly nuclear localized and single molecule RNA-fluorescence in situ hybridizations (RNA-

FISH) indicated low expression with ~0.3 estimated transcripts per cell. *HOTTIP* recruits the MLL1 complex by binding to WDR5 and targets it to the *HOXA* locus by chromosomal looping to facilitate H3K4me3 deposition and activation of *HOXA* genes. Interestingly, ectopic expression of *HOTTIP* did not recapitulate this effect, arguing for a need of *HOTTIP* transcription near its target genes (Wang et al., 2011c). This is one of the first examples of how a lncRNA exploits a tethering mechanism and chromosomal looping to bring activating chromatin modifiers to its target genes *in cis*. In mouse embryonic stem cells the intergenic lncRNA *Mira* exploits a similar mechanism to activate *Hoxa6* and *Hoxa7*. *Mira* is ~0.8kb long, unspliced and targets the Mll1 complex and thereby H3K4me3 to the two *Hoxa* genes by intrachromosomal looping (Bertani et al., 2011).

The human intergenic lncRNA *DBE-T* is a chromatin-associated nuclear RNA and has been the first described activatory lncRNA to be involved in a human genetic disease (Cabianca et al., 2012). Facioscapulohumeral muscular dystrophy (FSHD) is a common myopathy that is caused by deletions that reduce the copy number of a 3.3kb repeat. In healthy individuals, these repeats normally recruit PRC2 and its repressive mark H3K27me3, which also represses neighboring FSHD-causing genes. A reduction of repeat copy numbers below 11 units leads to decreased H3K27me3 levels and thereby expression of *DBE-T*, which recruits the TrxG protein Ash1L and its associated mark H3K36me2 to de-repress FSHD-causing genes *in cis* and induce the disease (Cabianca et al., 2012).

ANRIL (also known as *p15AS*) is an antisense lncRNA that spans over 126kb, is inefficiently spliced and nuclear localized. It has been first detected in familial cancers, is transcribed in antisense orientation to the *INK4b-ARF-INK4a* tumor suppressor gene cluster and completely overlaps *p15/INK4a* (Pasmant et al., 2007). While the three tumor suppressor genes are not expressed in normal cells, they become rapidly upregulated in senescent and oncogenic cells and are frequently deleted in various cancers. *ANRIL* has been shown to silence *p15/INK4a* in human leukemias through heterochromatin formation but not DNA methylation (Yu et al., 2008). It therefore interacts with CBX7 within PRC1 and SUZ12 within PCR2 to recruit H3K27me3 and repress the whole locus *in cis* (Kotake et al., 2011; Yap et al., 2010). A knock-down of *ANRIL* using shRNA or antisense oligonucleotides resulted in 4- to 8-fold upregulation of *p15/INK4a* expression and decreased proliferation. *Anril* knock-out in mice indicated a direct cis-repressor activity towards *p15/INK4a and p16/INK4b* (Visel et al., 2010). The importance of ANRIL is further underlined by genome-wide association studies that identified it as a susceptibility locus for coronary disease, intracranial aneurysm and type 2 diabetes (Pasmant et al., 2011).

BACE1-AS is a conserved and ~2kb long antisense lncRNA to *BACE1* in mouse and humans (Faghihi et al., 2008). *BACE1* is a crucial enzyme in the development of Alzheimer's disease as it cleaves the amyloid precursor protein to give rise to 42bp long amyloid-β peptides that accumulate in the brain and form plaques. As a consequence of cellular stress, *BACE1-AS* is upregulated, and forms an RNA duplex with its sense partner *BACE1*. This leads to *BACE1*

mRNA stabilization, increased BACE1 protein levels and enhanced amyloid-β production. A follow-up study could show that this antisense lncRNA-mediated stabilization of *BACE1* mRNA is due to a masking of a specific microRNA binding site (Faghihi et al., 2010). In vivo, levels of *BACE1-AS* were elevated in Alzheimer's disease patients as well as mice transgenic for the amyloid precursor protein, which is consistent with its proposed role as a driver for Alzheimer's disease. Another example of an antisense lncRNA regulating its sense partner is *Zeb2-as*, which forms a duplex with *Zeb2* mRNA and masks the 5' splice site of an intron (Beltran et al., 2008). The retained intron contains a ribosome entry site that facilitates translation and an increase of *Zeb2* protein levels. As a consequence, *Zeb2* transcriptionally represses *E-cadherin* and epithelial-mesenchymal transition (EMT) is triggered.

As already discussed in chapter 2.4.2, the intergenic *lincRNA-p21* post-transcriptionally represses translation of multiple mRNAs *in trans* but has also been shown to exhibit *cis*-regulatory effects on its neighboring gene *p21*. A deletion of the promoter and first exon of *lincRNA-p21* led to differential expression of 143 genes in unchallenged MEFs and 904 genes in DNA damaged MEFs (Dimitrova et al., 2014). However, the predominant target of *lincRNA-p21* seems to be *p21* as *lincRNA-p21* associates with hnRNP-K to act as a coactivator for p53-dependent *p21* transcription. The *trans*-deregulation of hundreds of genes could be secondary due to diminished *p21* levels rather than a direct consequence of *lincRNA-p21* depletion. These findings suggest that *lincRNA-p21* affects global gene expression by primarily regulating its neighboring gene *p21 in cis*.

2.5 RNA biology features of lncRNAs

2.5.1 RNA splicing

Maturation of RNA precursors consists of the four processes capping, splicing, 3' cleavage and polyadenylation, all of which happen simultaneously as the RNA is transcribed (Bentley, 2014). RNA splicing is the process in which the introns of nascent pre-mRNA are excised and the exons are joined to form the mature mRNA. It exists in all kingdoms of life although with major differences. While eukaryotes predominantly splice mRNAs and to a lesser extent non-coding RNAs, splicing in prokaryotes is much more rare (Rogozin et al., 2012). Three splicing pathways exist in eukaryotes with the spliceosomal pathway being the most important one followed by self-splicing and tRNA splicing. The spliceosome operates in the nucleus and consists of small nuclear (sn)RNAs and a range of associated proteins. It assembles new at each pre-mRNA and detects the 5' and 3' splice sites of the introns and a branch point sequence in the middle of the intron. Three canonical splice sites exist, of which GT-AG is most widely used (>99%) followed by GC-AG and AT-AC. Most RNAs are co-transcriptionally spliced, however, splicing is often completed post-transcriptionally (Bentley, 2014; Tilgner et al., 2012). While the majority of mRNAs is spliced, several single-exon mRNAs do not depend on splicing. Many lncRNAs such as *Airn*, *Kcnq1ot1* and *Malat1* have been shown to be

predominantly unspliced while others such as *H19* and *Braveheart* are efficiently spliced. It is currently unknown why lncRNAs are less efficiently spliced compared to mRNAs. As many lncRNAs are not processed properly (Cheng et al., 2005; Yang et al., 2011b), it could be that a general lack of efficient RNA processing also leads to a lower splicing efficiency. However, also polyadenylated lncRNAs can be inefficiently spliced (Seidl et al., 2006). One study found that co-transcriptional splicing is inefficient for lncRNAs (Tilgner et al., 2012), however, post-transcriptional RNA splicing of the steady state RNA population has to my knowledge never been thoroughly investigated genome-wide.

2.5.2 RNA export

The cellular localization of RNA molecules is crucial to their functions and therefore tightly regulated. Messenger RNAs need to be exported to the cytoplasm to become translated while lncRNAs have described functions in either the cytoplasm or the nucleus, or both. Cytoplasmic functions of lncRNAs include regulation of mRNA translation, direct regulation of mRNA stability and regulation of the miRNA pool as a competing endogenous (ce)RNA (Fatica and Bozzoni, 2014; Salmena et al., 2011). Nuclear lncRNAs predominantly have gene regulatory functions, however, structural functions are also described, e.g. for *Malat1* and *Neat1* in the organization of nuclear bodies such as paraspeckles (Bond and Fox, 2009; West et al., 2014) and for *Firre* in the modulation of nuclear architecture (Hacisuleyman et al., 2014). Nuclear gene regulatory functions of lncRNAs can be broadly classified into *cis*-acting and *trans*-acting functions. Three models are known how *cis*-acting lncRNAs regulate mRNAs on the same chromosome, all of which strictly happen in the nucleus: (i) lncRNAs bind chromatin modifiers while still being tethered to the nuclear transcription machinery and target them to neighboring mRNA loci (e.g. *HOTTIP* (Wang et al., 2011c)). (ii) lncRNAs induce the formation of repressive chromatin to regulate dosage compensation and genomic imprinting (e.g. *Xist* and *Kcnq1ot1* (Pandey et al., 2008; Terranova et al., 2008; Umlauf et al., 2004). (iii) lncRNAs also regulate genes by the act of transcription through genetic elements or promoters (e.g. Airn (Latos et al., 2012)).

The regulation of RNA export into the cytoplasm is complex and requires RNA to associate with proteins to form ribonucleoprotein complexes. Subsequently, these complexes are targeted to and translocate through nuclear pore complexes (NPC) that are contained in the nuclear envelope (Carmody and Wente, 2009). Proteins binding the polyA tail facilitate nuclear export and RNA binding proteins harbor nuclear retention signals that keep RNAs in the nucleus or induce their immediate transport back into the nucleus once they are shuttled to the cytoplasm. Each of the four RNA processing events capping, splicing, 3' cleavage and polyadenylation that give rise to mature RNAs trigger the recruitment of proteins that facilitate nuclear export. Inefficiently processed RNAs are not exported but will instead be targeted for degradation by the exosome (Carmody and Wente, 2009). It is still unknown whether or how lncRNAs are actively retained in the nucleus to fulfill their functions or whether many of them

lack mRNA properties such as RNA processing and recruitment of specific proteins and therefore exhibit decreased nuclear export efficiency.

2.5.3 RNA stability

The abundance of RNA molecules in cells is a prerequisite for function and is determined by two factors: the rate of RNA transcription and RNA stability. While the rate of transcription for each gene can be determined using global run-on sequencing (GRO-seq) (Core et al., 2008), the RNA stability can be examined using transcription inhibitors or RNA labeling followed by RNA-seq. In general, RNA degradation is a necessary cellular function to eliminate RNA molecules that are not useful to the cell anymore and this process has to be tightly controlled. The three major classes of RNA-degrading enzymes include endonucleases (cut RNA internally) as well as 5' exonucleases (hydrolyze RNA from the 5' end) and 3' exonucleases (hydrolyze RNA from the 3' end) (Houseley and Tollervey, 2009). Most of RNA degradation activities can be attributed to RNA processing in which excised introns and spacer fragments are recycled. RNA degradation also plays a role in surveillance pathways that monitor RNA quality and targets defective RNAs for nonsense mediated decay (NMD). The third role for RNA degradation involves the constitutive turnover of mRNAs and lncRNAs and is a key factor in the control of gene expression (Houseley and Tollervey, 2009). Polyadenylated transcripts such as mRNAs and many lncRNAs undergo progressive deadenylation and degradation by the exosome in the nucleus and the cytoplasm, a process that is specific for each different RNA species and determines their life spans.

The regulation of lncRNA stability is important to retain lncRNAs with housekeeping functions and eliminate lncRNAs that are unfavorable for the cell, e.g. because they are defective or they have potent regulatory functions that were only needed for a limited time span to react to a stimulus. The stability of lncRNAs has long been studied in order to evaluate possible functions. Single lncRNAs such as *Airn* (half life ~2h) (Seidl et al., 2006) and *Kcnq1ot1* (half life <4h) (Redrup et al., 2009) as well as lncRNA classes such as cryptic unstable transcripts (CUTs, half life 3-10min) (Wyers et al., 2005) in yeast and promoter upstream transcripts (PROMPTs, "highly unstable") (Preker et al., 2008) in humans have been shown to be rather unstable. A class of highly unstable cryptic antisense lncRNAs thought to be a byproduct of RNA polymerase II activity has recently been shown to be specifically degraded by the exonucleolytic RNA exosome (Core et al., 2008; Preker et al., 2008). The first genome-wide studies on lncRNA stability using microarrays (Clark et al., 2012) and RNA-seq (Tani et al., 2012) showed that many lncRNAs are unstable and that the diversity of lncRNA stability is increased compared to mRNAs.

2.6 Is RNA biology indicative for lncRNA function?

It seems that lncRNA functions can be split into the three major categories of (i) *trans*-acting lncRNAs that function by interacting with proteins, (ii) *cis*-acting lncRNAs that function by acting as tethers to recruit chromatin modifiers and lastly (iii) *cis*-acting lncRNAs that regulate neighboring genes by the act of transcription (Batista and Chang, 2013). A novel lncRNA can easily be grouped in any of the four well-studied classes of intergenic, bidirectional, enhancer and antisense lncRNAs (see chapter 2.3), however, these classes will give little information about possible functions and mechanisms of action. It was suggested that RNA biology could be an initial predictor for function and could provide a good start for further in-depth functional validation (Guenzl and Barlow, 2012; Kornienko et al., 2013). Intuitively, this makes sense, as lncRNAs that interact with proteins would need to be rather stable to reach certain abundance levels in order to fulfill their functions. Conversely, lncRNAs that are a byproduct of a functional transcriptional process do not necessarily need to be stabilized or otherwise processed. Due to a lack of functional data for the majority of lncRNAs, this hypothesis will await verification for some time. However, in the meantime one could look at the few known lncRNAs that have been assigned functions and mechanisms of actions or use conservation of RNA biology between closely related species as a proxy to estimate the importance of RNA biology for function. The drawback is that for many lncRNAs full information about RNA biology features is missing and if these features are known, they are not comparable as they had been assayed in different cell types, species and often with different techniques. A genome-wide dataset combining the three RNA biology features export, stability and splicing for the same lncRNA annotation in multiple cell types is clearly needed to unravel possible connections between RNA biology and function.

A combination of functional and RNA biology data is lacking for most lncRNAs. The unusual RNA biology of imprinted lncRNAs such as *Airn* and *Kcnq1ot1* (see chapter 2.4.1) early raised questions how they function despite the low stability and inefficient processing of their RNA transcripts (Pauler et al., 2007). The term “macro lncRNA” was coined to emphasize their RNA biology and distinguish them from mRNA-like lncRNAs (Guenzl and Barlow, 2012). It was suggested that macro lncRNAs could function by transcriptional overlap of mRNA promoters or enhancers, which would render the RNA molecule dispensable (Koerner et al., 2009; Pauler et al., 2012). Indeed, *Airn* represses *Igf2r* by transcriptional overlap (Latos et al., 2012). However, the RNA molecules of *Airn* as well as *Kcnq1ot1* have also been reported to interact with the H3K9 histone methyltransferase G9a to regulate distant neighboring genes in extra-embryonic tissues (Nagano et al., 2008; Pandey et al., 2008). This indicates that lncRNAs could employ two different mechanisms to regulate overlapped and distant genes in respective tissues. Other lncRNAs such as *Braveheart* and *Hotair* that interact with proteins are often spliced and stable and have an mRNA-like RNA biology. It has never been investigated whether “macro” lncRNAs are more widespread in the mouse genome than previously anticipated and whether they can be distinguished from mRNA-like lncRNAs.

If RNA biology indeed correlates with function, one would expect that when the function of a particular lncRNA is conserved also the RNA biology features are conserved. LncRNAs evolve rapidly, a third of all human lncRNAs is assumed to have arisen only in the primate lineage (Derrien et al., 2012). Between the closely related species mouse and rat up to 61% of lncRNAs have been shown to be unique to the *Mus* genus (Kutter et al., 2012). Those lncRNAs that are expressed in multiple species can therefore be assumed to have conserved (yet unknown) functions. Investigating the RNA biology of these conserved functional lncRNAs might reveal key insights into the connection between RNA biology and function. As an example, a lncRNA that is mRNA-like in mouse and rat has probably a similar function in both species and further argues that RNA biology is important for function. If, however, the majority of lncRNAs exhibit a different RNA biology in mouse and rat this could point towards the fact that RNA biology is irrelevant for function. In essence, knowledge of the RNA biology of lncRNAs together with secondary structures, protein-binding motifs and other features in multiple cell types and species would contribute to the functional characterization of lncRNAs (Mercer and Mattick, 2013). The interrogation of all these features genome-wide, rather than for individual lncRNAs, will provide a powerful platform to extrapolate functions for related lncRNAs, similar to large-scale proteomic studies, which might also be interesting for the pharmaceutical industry to develop lncRNA based therapeutics.

2.7 Aim of this study

The major aim of this study was to investigate the RNA biology (stability, export and splicing) of mouse lncRNAs genome-wide in order to gain more knowledge about possible connections between RNA biology, genomic transcript features and functional implications. I created genome-wide datasets for RNA stability, RNA export and RNA splicing in two cell types of the mouse and the rat and found that approximately half of all lncRNAs have an mRNA-like RNA biology while the other half is non-mRNA like. Furthermore, I show that current lncRNA classes are not distinguishable by RNA biology. As an alternative, I clustered lncRNAs by their RNA biology and show that the RNA biology of each cluster is evolutionary conserved in the rat. Additionally, this study reveals that certain genomic transcript features such as cDNA size and average exon size strongly correlate with RNA biology features. These datasets will be a valuable resource for the research community and will help to classify lncRNAs by their RNA biology features.

N.B.: Work from this thesis contributes to a manuscript that is currently prepared by myself and Florian Pauler (Guenzl et al., manuscript in preparation). Furthermore, I have published one review article for which I wrote the text and prepared figures (Guenzl and Barlow, 2012) and contributed text for a second review article (Kornienko et al., 2013). Both reviews are included in the appendix.

3 MATERIAL AND METHODS

3.1 Materials

3.1.1 Cell lines

Table 2: Cell lines

Cell Lines	Source
CCE mouse embryonic stem cells (mESC)	Provided by Erwin Wagner (CNIO, Madrid, Spain)
Primary mouse embryonic fibroblasts (MEF)	Self-established from FvB/N strain
Rat embryonic stem cells (rESC)	Obtained from the Rat Resource & Research Center
Primary rat embryonic fibroblasts (REF)	Self-established from Him:OFA strain

3.1.2 Cell culture reagents

Table 3: Cell culture reagents

Cell Culture Reagent	Company
Fetal calf serum	PAA
Gentamycin	Invitrogen
L-glutamine	Invitrogen
DMEM	Invitrogen
DMSO	Sigma
Trypsin 0,25% EDTA, red	Invitrogen
GS1-R Rat Pluripotent Stem Cell Culture Media	StemCells, Inc.

3.1.3 Chemicals

Table 4: Chemicals

Chemicals	Company
GoTaq DNA polymerase	Fermentas
5x GoTaq flexi buffer	Fermentas
Proteinase K	Applichem
Actinomycin D	Sigma
Agarose	Biozym
BCP	MRC
dNTP Mix 10mM	Fermentas
Ethanol 96%	Merck
Ethidiumbromide	Applichem
Glycogen	Roche
HCl	Merck
Isopropanol	Merck
MgCl2 25mM	Fermentas
RNA storage solution	Ambion
RNase Zap Spray	Ambion
RNase Zap Wipes	Ambion
Sodium acetate 3M	Ambion
TRI reagent	Sigma
Tris	Applichem
Betaine 5M	Sigma
AMPure XP beads	Beckman Coulter, Inc.

3.1.4 Kits

Table 5: Kits

Kit	Company
DNA-free Kit	Ambion
RevertAid First Strand cDNA Synthesis Kit	Fermentas
Mesa Green qPCR Mastermix Plus	Eurogentec
Ribo-Zero rRNA Removal Kit (H/M/R) Low Input	Epicentre
Ribo-Zero rRNA Removal Kit (H/M/R)	Epicentre
ScriptSeq RNA-Seq Library Preparation Kit	Epicentre
ScriptSeq v2 RNA-Seq Library Preparation Kit	Epicentre
ScriptSeq Index PCR Primers (Set 1)	Epicentre
TruSeq RNA Sample Prep Kit v2	Illumina
MinElute PCR Purification Kit	Qiagen
MinElute Gel Extraction Kit	Qiagen
RiboMinus Eukaryote Kit for RNA-seq	Life Technologies
Wizard SV Gel and PCR Clean-up System	Promega

3.1.5 Equipment

Table 6: Equipment

Equipment	Company
NanoDrop 1000 Spectrophotometer	Thermo Scientific
Bioanalyzer 2100	Agilent Technologies, Inc.
Experion Automated Electrophoresis System	Bio-Rad Laboratories GmbH
HiSeq2000	Illumina, Inc.
cBot	Illumina, Inc.
AbiPrism 7000 Sequence Detection System	Applied Biosystems
Thermal cycler PCT-200	MJ Research
Microcentrifuge 5415R	Eppendorf
Megafuge 1.0R	Heraeus
Avanti J-26 XP Centrifuge	Beckman Coulter
Qubit 1.0, 2.0	Invitrogen
RNA 6000 Nano Kit	Agilent Technologies, Inc.

3.1.6 PCR primers

Table 7: PCR primers

PCR assay	Primer name	Sequence (5'-3')	Reference
YMT2/B (mouse)	YMT2/B-F	CTGGAGCTCTACAGTGATGA	(Capel et al., 1999)
	YMT2/B-R	CAGTTACCAATCAACACATCAC	
Myog (mouse)	Myog-F	TTACGTCCATCGTGGACAGCAT	(Capel et al., 1999)
	Myog-R	TGGGCTGGGTGTTAGTCTTAT	
Sry (rat)	Sry-F	CATCGAAGGGTTAAAGTGCCA	(Buehr et al., 2008)
	Sry-R	ATAGTGTGTAGGTTGTTGTCC	
Il-2 (rat)	Il-2-F	CTAGGCCACAGAATTGAAAGATCT	(Buehr et al., 2008)
	Il-2-R	GTAGGTGGAAATTCTAGCATCATCC	

3.1.7 qPCR primers

Table 8: qPCR primers

PCR assay	Primer name	Sequence (5'-3')	Reference
5S rRNA (human)	5S-F	CTACGGCCATACCACCCT	(Zhang et al., 2011)
	5S-R	GGTATTCCCAGGCGGTCT	
5.8S rRNA (human)	5.8S-F	CTTAGCGGTGGATCACTCG	(Zhang et al., 2011)
	5.8S-R	AAGCGACGCTCAGACAGG	
18S rRNA (human)	18S-F	TCCTTTGGTCGCTCGCTCCT	(Zhang et al., 2011)
	18S-R	TCGCTCTGGTCCGTCTTGC	
28S rRNA (human)	28S-F	TTCGGGATAAGGATTGGCTCTA	(Zhang et al., 2011)
	28S-R	GGCTGTGGTTTCGCTGGAT	
Gapdh (mouse)	Gapdh-F	CATGGCCTTCCGTGTTCCTA	self-designed
	Gapdh-R	TGTCATCATACTTGGCAGGTTT	
Myc (mouse)	Myc-F	GAGCCCCTAGTGCTGCAT	self-designed
	Myc-R	CCACAGACACCACATCAATTTCTT	
Airn (mouse)	Airn-F	GACCAGTTCCGCCCGTTT	self-designed
	Airn-R	GCAAGACCACAAAATATTGAAAAGAC	
Kcnq1ot1 (mouse)	Kcnq1ot1-F	GCCCAAACCTTAGTCCTCCAT	self-designed
	Kcnq1ot1-R	GAAAGCACTCCTCCCCATTT	

3.2 Cells and Cell Culture

3.2.1 Ethics statement

According to the Austrian Laboratory Animal Act no animal experiments were performed in this study. The humane killing of laboratory animals is not defined as animal experimentation under the Austrian Laboratory Animal Act (Animal Experiments Act, Federal Law Gazette No. 501/1989). Therefore, approval of the study by an institutional ethics committee was not required. Mice were bred and housed at The Research Institute of Molecular Pathology, Dr. Bohr-Gasse 7, 1030 Vienna, Austria in strict accordance with national recommendations described in the "IMP/IMBA Common Institutional policy concerning the care and use of live animals" with the permission of the national authorities (Laboratory Animal Facility Permit MA58-0375/2007/4). Mouse and rat embryos and mouse tissues were obtained after humane killing of animals in a CO_2 chamber.

3.2.2 Mouse embryonic stem cells (mESC)

Undifferentiated CCE mESC are feeder independent and were grown in ES cell medium at 37°C in 5% CO_2 atmosphere on gelatinized dishes. Medium was replaced daily and cells were passaged every second or third day. ES cell medium contained HEPES-buffered DMEM medium supplemented with 15% FCS, 50µg/ml gentamicin, 2mM L-glutamin, 1x MEM (non-essential amino acids), 1mM sodium pyruvate, 0.1mM β-Mercaptoethanol and LIF.

3.2.3 Primary mouse embryonic fibroblasts (MEF)

Primary MEF cell lines were generated from E12.5 embryos from one FVB/N litter. 11 Embryos were dissected, the head and all organs removed and the remainings repeatedly passed through a 20G needle (dissection help from Quanah J. Hudson). Cells were cultured on 10cm dishes in MEF media (DMEM supplemented with 10% FCS, 2mM L-Glutamin and 50µg/ml Gentamycin) and split 1:3 every third day for 3 passages. The head was used for DNA isolation to sex-type all MEF cell lines. Two female MEF cell lines (#1 and #3) were used as biological replicates for all experiments.

3.2.4 Rat embryonic stem cells (rESC)

High and low-passage rESC (p37 & p23) were obtained from the Rat Resource & Research Center and expanded for three passages on irradiated MEF feeder layers in GS1-R Rat Pluripotent Stem Cell Culture Media (by Florian Pauler). I washed off rESC colonies by repeatedly pipetting media over the cells. High and low-passage rESC were used as biological replicates for all experiments. The mouse feeder contamination was assessed by PCR and Sanger sequencing.

3.2.5 Primary rat embryonic fibroblasts (REF)

Primary REF cell lines were generated from E13.5 embryos from one Him:OFA litter. 12 Embryos were dissected, the head and all organs removed and the remainings repeatedly passed through a 20G needle (dissection help from Quanah J. Hudson). Cells were cultured on 10cm dishes in MEF media and split 1:3 every third day for 3 passages. Because some cell lines were growing slowly, 30-50% conditioned media was used. The head was used for DNA isolation to sex-type all REF cell lines. Two female REF cell lines (#4 and a pool of #6, #10 and #12) were used as biological replicates for all experiments.

3.2.6 Sex typing of MEFs and REFs

Mouse and rat embryo heads were lysed in 1ml of a standard lysis buffer at 55°C. After addition of 300µl saturated NaCl solution and centrifugation (10min, 16,100xg, RT), the supernatant was added to 600µl Isopropanol. The precipitated DNA was pelleted by centrifugation (10min, 16,100xg, RT), the pellet washed once in 1ml 70% EtOH and resuspended in TE buffer. The concentration of the DNA samples was adjusted to be approximately 100ng/µl. For MEFs, PCR was carried out with primers specific for the Y-chromosomal locus YMT2/B and for the autosomal gene *Myogenin* as an internal control (Capel et al., 1999). For REFs, PCR was carried out with primers specific for the Y-chromosomal gene *Sry* and the autosomal gene *Interleukin-2* as an internal control (Buehr et al., 2008).

3.2.7 Quantification of mouse feeder contamination of rESC

DNA was isolated from TRIreagent samples after completion of RNA isolation according to the manufacturer's procedure. PCR was carried out with primer specific for a genomic region in the Myc gene bearing three SNPs between mouse FvB/N and rat Him:OFA. As a control, FvB/N mouse DNA and rESC DNA from a feeder-independent clone was used. PCR products were gel-purified and sent to Microsynth for Sanger sequencing. Nucleotide intensities were read out from the chromatograms and a ratio for each SNP calculated. For each sample, the three SNP ratios were averaged. After subtraction of background, the mouse feeder contamination of rESC was calculated to be in average 5.2% with a range from 2.0-13.0%. Two samples of the rESC RNA stability dataset with a contamination of 18% and 25% were not used for pooling of technical replicates.

3.2.8 FACS Sorting of B and T cells

B and T cells were harvested from mouse lymph nodes together with Martina Minnich (Group Meinrad Busslinger, IMP Vienna) and sorted by Fluorescence-activated cell sorting (FACS) with the help of Thomas Lendl in the Flow Cytometry Core Facility of IMP/IMBA. Approximately ten lymph nodes were each harvested from three female 6-week-old C57BL/6 mice into cold FACS buffer and mashed through a nylon mesh. After centrifugation (5min, 1500rpm, 4°C), the cells were washed with cold FACS buffer and counted in as CasyCounter. After another centrifugation (5min, 1500rpm, 4°C), cells were resuspended in 500µl cold FACS buffer and 1µl Fc block (1:500 of 2mg/ml solution) added for 10min to prevent unspecific antibody staining. Cells were divided into two fractions and stained using a set of antibodies specific for B cells or CD4+/CD8+ T cells (see Table 9 and Table 10). I prepared 2x antibody dilutions and incubated 250µl cell suspension with 250µl antibody dilution for 30min at 4°C in the dark. In the meantime, antibody controls were prepared the same way for each antibody used. After 30min, the cells were washed with FACS buffer, centrifuged (5min, 1500rpm, 4°C) and resuspended in 1ml FACS buffer. FACS sorting was done using a FACS Aria III sorter by first checking each dye against each other to see whether colors shine into other detectors. Then, for each of the two antibody stainings all three biological replicates were sorted into buffer, cells were pelleted and immediately resuspended in TRIreagent. As RNA concentrations were low after DNaseI treatment, I decided to pool RNA from biological replicates before RNA-seq.

Table 9: Antibodies for B cell staining

Conjugate	Antibody	Specificity	Conc.	Dilution	Clone	Company
FITC	B220	anti-mouse	0.5mg/ml	(1:2,500)	RA3-6B2	Biolegend
PE	IgD	anti-mouse	0.2mg/ml	(1:25,000)	11-26	eBiosciences
APC	CD19	anti-mouse	0.2mg/ml	(1:2,500)	1D3	eBiosciences

Table 10: Antibodies for CD4+ and CD8+ T cell staining

Conjugate	Antibody	Specificity	Conc.	Dilution	Clone	Company
FITC	CD8a	anti-mouse	0.5mg/ml	(1:10,000)	53-6.7	BD Biosciences
FITC	CD8b	anti-mouse	0.5mg/ml	(1:10,000)	53-5.8	BD Biosciences
PE	TCRb	anti-mouse	0.2mg/ml	(1:2,500)	H57-597	eBiosciences
PE-Cy5	CD4	anti-mouse	0.2mg/ml	(1:10,000)	GK1.5	Biolegend
APC	CD3e	anti-mouse	0.2mg/ml	(1:2,500)	145-2C11	eBiosciences

3.3 RNA localization: nuclear and cytoplasmic RNA extraction

The protocol for the extraction of nuclear and cytoplasmic RNA was adapted from Sambrook and Russel, Molecular Cloning, Third Edition. Briefly, cells were washed 3x with ice-cold PBS and scraped off into a glass Corex tube. After centrifugation (5min, 2,000xg, 4°C), cells were resuspended in ice-cold Lysis buffer and underlayed by an equal volume of Lysis buffer containing sucrose and NP-40. After centrifugation (20min, 10,000xg, 4°C) the heavier intact nuclei formed a pellet at the bottom of the tube while the cytoplasm stayed above the sucrose gradient. The cytoplasm phase was taken to another tube and an equal volume of 2x Proteinase K buffer added. The nuclei pellet was washed five times with ice-cold PBS, resuspended in Lysis buffer and an equal volume of 2x Proteinase K buffer was added. The nuclei were sheared by repeatedly passing them through a 19G needle. The cytoplasmic as well as the nuclear sample was incubated for 30min at 37°C to degrade proteins. A phenol/chloroform extraction removed contaminants and the RNA was precipitated by addition of 2.5V 96% EtOH and 0.1V 3M NaAc and an incubation for at least 1h at -20°C. The RNA was recovered by centrifugation (30min, 13,000xg, 4°C), the pellet washed once with 75% EtOH and the RNA dissolved in RSS. The efficiency of separation was quantified by qPCR for Gapdh and nuclear localized Air or Kcnq1ot1.

3.4 RNA stability: Actinomycin D treatment

To assess the stability of RNA transcripts, cells were treated with the transcription inhibitor Actinomycin D. Actinomycin D was dissolved in 96% EtOH to a stock concentration of 4µg/µl and added to cell culture media to a final concentration of 10µg/ml (1:400). Cells were treated for 0h, 1h or 4h with Actinomycin D or the vehicle control 96% EtOH in two technical replicates that were pooled after quality control by qPCR. The experiment was repeated within one week with cells of a different passage number or another cell line to have biological replicates that both were used for RNA-seq.

3.5 RNA isolation & DNase I treatment

3.5.1 RNA isolation

RNA was isolated using TRI Reagent with some modifications to the vendor's protocol. Briefly, cell culture media was removed and cells were lysed in TRI Reagent. After 5min of incubation, the cell homogenate was immediately used for RNA isolation used or stored at -20°C for later processing. Per 1.0ml of TRI Reagent 0.1ml BCP phase separation reagent was added. After 15s of shaking and incubation for 10min at RT, the sample was centrifuged (15min, 16,100xg, 4°C). The upper aqueous phase was transferred to a new 1.5ml tube, 0.5ml Isopropanol was added and the mixture was gently mixed by repeatedly inverting the tube. The RNA was allowed to precipitate for 10min at RT and pelleted by centrifugation (15min, 16,100xg, 4°C). After removal of the supernatant, the RNA pellet was washed with 1ml of 75% EtOH and again centrifuged (10min, 16,100xg, 4°C). The RNA pellet was allowed to air-dry for 5-10min and resuspended in appropriate amounts of RNA Storage Solution (Ambion). RNA was stored at -20°C and RNA concentration and purity was measured using a NanoDrop. RNA quality was assessed by running an RNA Nano Chip (Agilent Technologies, Inc.) on a 2100 Bioanalyzer and an RNA integrity number (RIN) >8 was required.

3.5.2 DNase I treatment

RNA was DNaseI treated using the DNA-free kit (Ambion) according to the vendor's protocol. Briefly, per reaction 10µg RNA was diluted with RNase-free water to 44µl and 5µl 10x buffer and 1µl DNaseI was added. The mixture was incubated for 30min at 37°C in a thermal heater. Then 5µl of the Inactivation Reagent were added, the sample 2x mixed during a 2min incubation step at RT and subsequently centrifuged (2min, 10,000xg, RT). The supernatant was carefully taken off and transferred to a new 1.5ml tube and precipitated by adding 2.5V 96% EtOH and 0.1V 3M NaAc. After an incubation at -20°C for at least 1h, the RNA was pelleted by centrifugation (30min, 16,100xg, 4°C) and washed with 1ml of 75% EtOH and again centrifuged (10min, 16,100xg, 4°C). The RNA pellet was allowed to air-dry for 5-10min and resuspended in appropriate amounts of RNA Storage Solution (Ambion).

3.6 Quantitative real-time PCR (qPCR)

3.6.1 Reverse transcription

1-2µg of total RNA were reverse transcribed using the RevertAid First Strand cDNA Kit according to the vendor's protocol. The RT-program was as follows: 10min 25°C, 60min 42°C and 10min 70°C.

3.6.2 Quantitative real-time PCR (qPCR)

Reverse transcription reactions were diluted 1:10 with embryo water and 5µl added to 20µl Mesa Green qPCR MasterMix Plus for SYBR Assays (Eurogentec) containing 100nM of primers. To control for DNA contamination, reverse transcription reactions without reverse transcriptase were prepared and assayed in parallel. A difference of >7 qPCR cycles between +RT and -RT reactions indicated no significant DNA contamination.

3.6.3 Primer design

qPCR primers were designed online using primer-blast on the NCBI website (http://www.ncbi.nlm.nih.gov/tools/primer-blast/) and the following parameters: PCR product size 70-150bp, primer melting temperature 58-60°C (59°C optimal), primer size 15-30 (20 optimal), GC content 30-80%, 3'GC clamp of 0 residues and amplicon maximal temperature 85°C.

3.7 Purification and fragmentation of RNA

3.7.1 Removal of ribosomal RNA (Ribo-Zero)

Ribosomal RNA was removed from total DNaseI treated RNA using the Ribo-Zero rRNA removal kit (Human/Mouse/Rat) (Epicentre) according to the vendor's protocol. Briefly, 1-4µg DNaseI treated RNA were diluted with RNase-free water to 26µl and 4µl 10x buffer and 10µl RNA Removal Solution was added. After an incubation step at 68°C for 10min and an incubation step at RT for 15min, the mixture was added to resuspended and previously washed microsphere beads supplemented with RNase inhibitor and incubated for 10min at RT with occasional vortexing. After a final incubation at 50°C for 10min, the mixture was applied onto Microsphere Removal Filter Units and centrifuged (1min, 10,000xg, RT). The flow-through contained the non-ribosomal RNA and after an addition of 80µl RNase-free water it was precipitated with 18µl 3M NaAc, 2µl Glycogen (10mg/ml) and 600µl of 96% EtOH and incubated at -20°C for at least 1h. The non-ribosomal RNA was pelleted by centrifugation (30min, 16,100xg, 4°C), washed with 1ml of 75% EtOH, briefly air-dried and resuspended in 1µl of RNase-free water. 19.5µl of Elute, Prime, Fragment Mix from the TruSeq RNA Sample Prep Kit v2 (Illumina) were added and the reaction incubated at 94°C for 8min. After the incubation, 17µl of total non-ribosomal RNA were transferred to a new 2ml low-bind tube and directly used for library preparation.

3.7.2 Enrichment of polyA+ RNA

PolyA+ RNA was enriched from total DNaseI treated RNA using the TruSeq RNA Sample Prep Kit v2 (Illumina) according to the vendor's protocol. Briefly, 1-4µg DNaseI treated RNA were

diluted in PCR tubes with RNase-free water to 50µl, 50µl of RNA Purification Beads were added and the mixture incubated for 5min at 65°C in a PCR machine. The reaction was transferred to a 2ml low-bind tube and put onto a magnetic stand for 5min. After removal of the supernatant containing most of the contaminants, the magnetic beads were washed with 200µl of Bead Washing Buffer. 50µl of Elution Buffer was added and the reaction incubated in a PCR tube for 2min at 80°C in the PCR machine to elute the enriched RNA from the magnetic beads. The RNA was transferred to a 2ml low-bind tube tube and 50µl of Bead Binding Buffer were added to allow specific rebinding of polyA+ RNA to the magnetic beads. After incubation for 5min at RT, the reaction was put onto a magnetic stand for 5min and the supernatant containing residual rRNA discarded and the beads washed with 200µl of Bead Washing Buffer. The polyA+ RNA was eluted by addition of 19.5µl of Elute, Prime, Fragment Mix and an incubation step at 94°C for 8min. After 5min on the magnetic stand, 17µl polyA+ RNA were transferred to a new 2ml low-bind tube and directly used for library preparation.

3.8 Strand-specific library preparation and RNA-seq

3.8.1 Epicentre's ScriptSeq (v1) RNA-Seq Library Preparation kit

3.8.1.1 RNA fragmentation

50-250ng of Ribo-Zero treated RNA were chemically fragmented for 5min at 85°C and the cDNA synthesis primer (random hexamers with tagging sequence) randomly annealed by putting the sample on ice.

3.8.1.2 cDNA synthesis

cDNA synthesis is initiated by the addition of the StarScript Reverse Transcriptase in the presence of dNTPs, RNase inhibitor and a StarScript buffer. The reactions were incubated for 5min at 25°C and for 20min at 45°C. 1µl Finishing Solution I was added to stop the reaction followed by a 3min incubation step at 95°C. The cDNA is now tagged at its 5' end.

3.8.1.3 3'-Terminal tagging of cDNA

The next step adds a tagging sequence to the 3' end of the cDNA, giving rise to di-tagged cDNA. Therefore, a master mix containing terminal-tagging oligos, Dithiothreitol and DNA polymerase was added and the mixture incubated for 15min at 37°C and for 3min at 95°C. 1µl Finishing Solution II was added to stop the reaction followed by a 10min incubation at 37°C and a 3min incubation step at 95°C.

3.8.1.4 Purification of the di-tagged cDNA

The di-tagged library was cleaned using the MinElute PCR Purification Kit (Qiagen) according to the manufacturer's procedure and eluted in 18µl EB buffer.

3.8.1.5 PCR amplification and addition of barcodes

The PCR amplification step generates the second strand of cDNA and incorporates the barcodes to be able to multiplex RNA-seq. PCR was carried out using 1.25U of the proof-reading FailSafe PCR Enzyme in presence of FailSafe PCR PreMix E, forward primer and the barcode-specific reverse primer. The PCR program was as follows: 95°C for 1min, 10-12 cycles of: 95°C for 30s, 55°C for 30s, 68°C for 3min and a final elongation step at 68°C for 7min, 10°C hold. Excess PCR primers were removed by incubating the reaction with 1µl Exonuclease I for 15min at 37°C.

3.8.1.6 Purification of the library

The libraries were purified using the MinElute PCR Purification Kit (Qiagen) according to the manufacturer's procedure or gel electrophoresis. For the latter, 50µl PCR product were supplemented with 10µl 6x loading dye and run for 2h at 80V in a 2% Agarose TAE gel containing 12µl SYBR Gold. Bands from 200-600bp were excised and libraries cleaned using the MinElute Gel Extraction Kit (Qiagen) according to the manufacturer's procedure.

3.8.2 Epicentre's ScriptSeq (v2) RNA-Seq Library Preparation kit

The ScriptSeq v2 RNA-Seq Library Preparation kit (Epicentre) was only used once to produce two RNA-seq libraries for a direct comparison against the dUTP/TruSeq protocol in the library preparation test (see Figure 6) and then discontinued. These libraries were prepared according to the manufacturer's procedure.

3.8.3 Illumina's TruSeq kit and its modifications

Strand-specific RNA-seq libraries from Ribo-Zero or polyA+ RNA were prepared using the TruSeq RNA Sample Prep Kit v2 (Illumina) according to the vendor's protocol with some modifications to preserve strand information. These modifications are published and include a filter step to remove unincorporated dNTPs after first-strand cDNA synthesis, the incorporation of dUTP instead of dTTP during second-strand cDNA synthesis and a final step to degrade the dUTP containing second-cDNA strand after adapter ligation using the enzyme Uracil-DNA Glycosylase (Sultan et al., 2012).

3.8.3.1 First-strand cDNA synthesis

8µl of First-Strand Master Mix (containing 1µl Superscript II per 9µl Master Mix) were added to 17µl of rRNA depleted or polyA+ RNA and shortly vortexed. The reverse transcription program was as follows: 10min 25°C, 50min 42°C, 15min 70°C, 4°C hold.

3.8.3.2 Clean-up of first-strand cDNA reaction

MicroSpin G-50 columns were centrifuged (1min, 700xg, RT) and washed twice with 500µl 1mM Tris-HCl pH 8.0. 5µl of Elution Buffer was added to the 25µl of first-strand cDNA reaction, the whole reaction applied onto the MicroSpin G-50 columns and centrifuged (1min, 700xg, RT). The volume of the eluate was measured and RNase-free water added to a total volume of 52µl.

3.8.3.3 Second-strand cDNA synthesis

23µl second-strand master mix were prepared per sample using the following reagents: 1µl RT buffer, 15µl 5x Second-Strand Synthesis Buffer, 1µl 50mM $MgCl_2$, 1µl 100mM DTT, 2µl dNTP mix (10mM each dATP, dCTP, dGTP, dUTP), 0.5µl E.coli DNA ligase (10U/µl), 2µl DNA Polymerase (10U/µl) and 0.5µl RNase H (2U/µl). After addition to the cleaned first-strand cDNA reaction, the reaction was incubated for 2h at 16°C. The reaction was cleaned by adding 135µl of magnetic AMPure XP beads and incubating the reaction for 15min at RT. After 5min on the magnetic stand, the supernatant was taken off and the beads washed twice with 200µl 80% EtOH. The beads were allowed to air-dry for 15min, resuspended in 52.5µl Resuspension Buffer and incubated for 2min at RT. After 2min on the magnetic stand, 50µl of the supernatant containing the purified cDNA were transferred to a new 2ml low-bind tube.

3.8.3.4 End repair

40µl End Repair Mix were added and the reaction incubated for 30min at 30°C. The reaction was cleaned as described above using 160µl of magnetic AMPure XP beads and eluting in 20µl Resuspension Buffer.

3.8.3.5 Adenylation of 3'ends

12.5µl A-Tailing Mix were added and the reaction incubated for 30min at 37°C and additionally for 5min at 70°C to remove adapter dimers and concatemers.

3.8.3.6 Adapter ligation

2.5µl Resuspension Buffer, 2.5µl Ligation Mix and 2.5µl of the desired RNA Adapter Index were added and the reaction incubated for 10min at 30°C. 5µl Stop Ligation Buffer were then added to stop the ligation. The reaction was cleaned twice as described above using 42µl of magnetic AMPure XP beads and 52.5µl Resuspension Buffer and for the second clean-up 50µl of magnetic AMPure XP beads and 22.5µl Resuspension Buffer.

3.8.3.7 Uracil-DNA glycosylase treatment

2.3µl of 10x UDGase buffer and 1µl of UDGase (5U/µl) were added to 20µl of cleaned adapter-ligated cDNA and incubated for 30min at 37°C.

3.8.3.8 Library enrichment by PCR

5µl of PCR Primer Cocktail and 25µl PCR Master Mix were added to 22.3µl of UDGase treated cDNA. The PCR program was as follows: 98°C for 30s, 8 cycles of: 98°C for 10s, 60°C for 30s, 72°C for 30s and a final elongation step at 72°C for 5min, 10°C hold.

3.8.4 Quantification, quality control and pooling of libraries

The RNA-seq libraries were quantified using the Qubit dsDNA Assay kit on a Qubit 1.0 or 2.0 Fluorometer. The quality and length distribution was assessed by running an Experion DNA 1K Analysis Chip on an Experion Automated Electrophoresis System. The molarity of each library was calculated by using the formula $nM = (conc * 10^6)/(size * 660)$, in which conc is the library concentration in ng/µl, size is the average size of the library (350bp for Ribo-Zero libraries, 400bp for polyA+ libraries) and 660 is the weight of a DNA basepair. 5µl of the library were diluted to 2nM using EB buffer containing 0.1% Tween-20. This 2nM dilution was quantified by Qubit and libraries were pooled with their actual measured molarities adjusted accordingly. Care was taken to only pool compatible barcodes as indicated in the Illumina manual. The final pooled library was quantified by Qubit and submitted to the Biomedical Sequencing Facility (BSF) at CeMM.

3.8.5 RNA-sequencing

RNA-seq was done by the biomedical sequencing facility (BSF) at CeMM. Briefly, 12-14pM of the library were loaded per lane of the flowcell and clonal clusters were generated from single molecules of cDNA using the cBot system. The flowcell was then loaded into the HiSeq 2000 and the clusters sequenced 50bp single-end (50SE) or 100bp paired-end (100PE). After the run, basepairs were called and barcodes demultiplexed. Raw RNA-seq data was provided as fastq files.

3.9 Basic analysis of RNA-seq data

3.9.1 Public data tracks used

For the analysis of RNA-seq data the following public data tracks were used (see Table 11).

Table 11: Public data tracks used in this study

Data track	Date of download from UCSC
RefSeq annotation (mm10)	08.03.2014
RefSeq annotation (rn5)	27.04.2014
Repeatmasker (mm10)	23.12.2013
Repeatmasker (rn5)	27.04.2014
mm10_chromInfo	Downloaded from UCSC
rn5_chromInfo	Downloaded from UCSC

3.9.2 Assessment of RNA-seq quality

The quality of RNA-seq data was assessed using the tool FastQC (available at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Data was checked for sequence quality per base, quality score per sequence, GC content per base, sequence duplication levels and overrepresented sequences. No dramatic differences were noticed.

3.9.3 Assessment of strand-specificity

The strand rule and strand-specificity of RNA-seq data was calculated using infer_experiment.py from the RSeQC package (Wang et al., 2012) with the following parameters:

```
infer_experiment.py -r transcript_annotation.bed12 -i
RNAseq_alignment.bam -s 1000000
```

3.9.4 Alignment of RNA-seq reads

RNA-seq reads were aligned to the mouse genome assembly mm10 or the rat genome assembly rn5 using the aligner STAR (version 2.3) (Dobin et al., 2013) with standard parameters and the following modifications: maximum intron size = 100kb (--alignIntronMax 100000), consider only canonical splice sites (--outFilterIntronMotifs RemoveNoncanonical). These options prevented the assembly of artifact transcripts that have either enormous introns >100kb thereby extending transcripts over multiple gene loci or were defined by non-canonical splice sites. The output of STAR is stored in a binary alignment map (BAM) file containing the aligned reads.

3.9.5 Preparation of data tracks for the UCSC genome browser

In order to visualize RNA-seq data on the UCSC genome browser (Fujita et al., 2011) the BAM files containing the aligned reads were first sorted and indexed using the functions sort and index from the SAMtools package (Li et al., 2009). Then, the BAM files were converted into two strand-specific wiggle (WIG) files using the script bam2wig.py from the RSeQC package (Wang et al., 2012). As WIG files are usually too large to directly load to UCSC they were further converted to BigWig (BW) files using the UCSC wigToBigWig tool. Bigwig files were loaded to a server and directly accessed by UCSC.

```
bam2wig.py -i RNAseq_alignment.bam -d strand-rule -o output -s
mm10_chromInfo.txt

wigToBigWig strandspecific_normalized.wig mm10_chromInfo.txt
strandspecific_normalized.bw -clip
```

3.9.6 RPKM calculation

The expression of transcripts from RNA-seq data was calculated using the script RPKM_count.py from the RSeQC package (Wang et al., 2012) using the following parameters:

```
RPKM_count.py -r transcript_annotation.bed12 -i
RNAseq_alignment.bam -d strand-rule -o output_prefix --skip-
multiple-hits --only-exonic
```

3.9.7 Analysis of RPKM saturation (RPKM error)

The accuracy of RPKM strongly depends on the sequencing depth, e.g. an RPKM of 10 can be inaccurate with 1 million reads whereas an RPKM of 0.5 can be very precise with 500 million reads. Instead of using only a strict RPKM cutoff and thereby ignoring different sequencing depths, I decided to use the method of read downsampling to judge whether a calculated RPKM is reliable or not. Therefore, RPKM saturation was calculated using the script RPKM_saturation.py from the RSeQC package (Wang et al., 2012) using the following parameters:

```
RPKM_saturation.py -r RefSeq_mm10.bed12 -d strand-rule -i
RNAseq_alignment.bam -o output_prefix
```

This script calculates 20 RPKM values for each of the transcripts of a given annotation using randomly downsampled reads starting from 5% to 100% of total reads (in 5% increments). An RPKM is considered stable (or saturated) when increased read depth does not alter the RPKM significantly, however, more than 100% of reads is not available. Hence, as a proxy, read numbers are reduced to see whether RPKM are stable with less reads. I calculated for each of the 20 RPKM values the RPKM error spread to the final RPKM being calculated with 100% of reads using the following formula, in which X is the percentage of downsampled reads (5%-100% of total reads).

$$RPKM\_Error(X) = \frac{|\, RPKM_{X\%} - RPKM_{100\%} \,|}{RPKM_{100\%}}$$

An RPKM is considered accurate when it is stable irrespective of whether 70%, 80%, 90% or 100% of reads are used. I calculated the final RPKM error for each transcript locus by averaging the RPKM error of five RPKM values (70%, 75%, 80%, 85% and 90%). In analyses that require stable RPKM (such as developmental regulation in chapter 4.2.7 and the RNA biology features in chapters 4.3 and 4.4) I filtered out transcripts that had an average RPKM error > 5%, in other words, I discarded transcripts whose RPKM error of five different RPKM calculated with 70-90% of reads was on average more than 5% different than the final RPKM.

3.9.8 Analysis of gene-body coverage

The gene-body coverage was analyzed using the script geneBody_coverage.py from the RSeQC package (Wang et al., 2012) using the following parameters:

```
geneBody_coverage.py -r RefSeq_mm10.bed12 -i
RNAseq_alignment.bam -o output_prefix
```

3.9.9 Analysis of splice junction coverage

The coverage of known RefSeq and novel splice junctions was calculated using the script inner_distance.py from the RSeQC package (Wang et al., 2012) using the following parameters:

```
junction_saturation.py -r RefSeq_mm10.bed12 -i
RNAseq_alignment.bam -o output_prefix
```

3.9.10 Analysis of the inner distance of paired-end RNA-seq reads

The inner distance of paired-end RNA-seq reads was analyzed using the script inner_distance.py from the RSeQC package (Wang et al., 2012) using the following parameters:

```
inner_distance.py -r RefSeq_mm10.bed12 -i RNAseq_alignment.bam -
o output_prefix
```

3.10 The lncRNA annotation

3.10.1 Public RNA-seq data used for lncRNA annotation

The following published polyA+ RNA-seq datasets (Merkin et al., 2012) were used to annotate mouse and rat lncRNAs and to calculate RPKMs (see Table 12). Mouse heart RNA-seq data from Merkin et al. was only available with short reads and was therefore only used for RPKM calculation to enable comparable analyses. Mouse heart RNA-seq data from the Sanger Institute (Keane et al., 2011) was available with 76PE and substituted the short reads from Merkin et al. for transcriptome assembly, however, this data is not stranded. Direction of assembled transcripts from unstranded data was inferred from directional splice junctions.

Table 12: Public RNA-seq data used to annotate lncRNAs and calculate RPKMs

Dataset	Accession Nr.	Read type	Nr of Reads	Strandedness
Mouse Brain	SRR594402	80x80 PE	118,824,353	0.9814
Mouse Colon	SRR594403	80x80 PE	87,447,334	0.9916
Mouse Heart (*)	SRR594395	36x36 PE	51,144,587	0.9930

Mouse Heart (*)	SRR594412	40SE		
Mouse Heart (**)	ERR032227-31, 38	76x76 PE	158,165,171	unstranded
Mouse Kidney	SRR594404	80x80 PE	118,885,190	0.9870
Mouse Liver	SRR594405	80x80 PE	134,045,721	0.9943
Mouse Lung	SRR594406	80x80 PE	62,362,901	0.9930
Mouse Sk. muscle	SRR594407	80x80 PE	117,171,737	0.9942
Mouse Spleen	SRR594408	80x80 PE	114,814,142	0.9762
Mouse Testis	SRR594409	80x80 PE	116,525,147	0.9954
Rat Brain	SRR594428	80x80 PE	96,368,839	0.9955
Rat Colon	SRR594429	80x80 PE	73,664,290	0.9912
Rat Heart	SRR594430	80x80 PE	67,008,998	0.9953
Rat Kidney	SRR594431	80x75 PE	116,656,722	0.9947
Rat Liver	SRR594432	80x80 PE	131,658,529	0.9966
Rat Lung	SRR594433	80x80 PE	84,087,214	0.9948
Rat Sk. muscle	SRR594434	80x80 PE	65,129,431	0.9914
Rat Spleen	SRR594435	80x75 PE	112,500,969	0.9833
Rat Testis	SRR594436	80x75 PE	114,820,645	0.9651

(*) Mouse heart RNA-seq data from Merkin et al. was used for RPKM calculation
(**) Mouse heart RNA-seq data from Keane et al. (Sanger Institute) was used for transcript assembly

3.10.2 Self-generated RNA-seq data used for lncRNA annotation

The following RNA-seq datasets were generated for transcriptome assembly and RPKM calculations (see Table 13).

Table 13: Self-generated RNA-seq data used to annotate lncRNAs and calculate RPKMs

Dataset	Read Type	Nr of Reads	Strand Specificity
mESC_pA	100PE	90,300,424	0.8687
MEF_BR1_pA	100PE	46,374,323	0.9091
MEF_BR2_pA	100PE	43,421,110	
rESC_pA	100PE	88,268,828	0.9331
REF_BR1_pA	100PE	42,939,061	0.9270
REF_BR2_pA	100PE	48,349,412	

3.10.3 Generation of the lncRNA annotation

The lncRNA annotation used in this study was generated by Florian Pauler. Shortly, transcripts were assembled from aligned RNA-seq reads from eleven tissues using the program cufflinks (version 2.1.1.) (Trapnell et al., 2013). The two parameters `--min-isoform-fraction` (default 0.10) and `--pre-mrna-fraction` (default 0.15) normally suppress lowly abundant splice isoforms and intra-intronic transcripts were set to 0 to increase lncRNA detection. To remove mRNAs, transcripts overlapping RefSeq mRNAs in sense (by 1bp) or antisense (by 20% of their cDNA) were filtered out. To remove remaining transcripts with potential protein-coding potential, a two step pipeline was employed: in step 1, the program RNAcode detects conserved protein patterns in multiple species alignments (Washietl et al., 2011) and in step 2 the program Coding Potential Calculator (CPC) analyzes six nucleotide sequence features (Kong et al., 2007).

3.10.4 Classification of lncRNAs

The grouping of lncRNAs into known subclasses was automated by Florian Pauler. Bidirectional lncRNAs were defined by starting from the same promoter (transcription start site +/- 1kb) as an mRNA and being transcribed in antisense direction. Overlapping lncRNAs were defined by a full, partial or intronic overlap of an mRNA in antisense direction. Enhancer RNAs were defined by an overlap of the lncRNA promoter (TSS +/- 1kb) within an H3K4me1 peak, an H3K27ac peak or a p300 peak in the tissue with highest expression. As public ChIP-seq data from the eleven tissues used is not available for rat, eRNAs could only be classified for the mouse annotation. Remaining transcripts were classified as intergenic lncRNAs.

3.10.5 Addition of RefSeq mRNAs to final lncRNA annotation

The final lncRNA annotation was complemented by the multi-exonic RefSeq mRNA annotation in order to have one combined annotation of mRNAs and lncRNAs for all downstream analyses. I refrained from using the assemblies to define mRNAs by myself as the non-coding pipeline is slow and can not handle tens of thousands putative mRNAs with an average of eleven exons, compared to three exons for lncRNAs. Alternatively, mRNAs could be defined from the assemblies by RefSeq mRNA overlap, however, this would also include unstable splice variants, fragments or fusion transcripts.

3.11 Analysis of RNA stability

3.11.1 Self-generated RNA-seq datasets

Table 14: Self-generated RNA-seq data used to calculate RNA stability

Dataset	Read Type	Nr. of Reads	Strand Specificity
mESC_0h_BR1	50SE	30,501,857	0.9620
mESC_0h_BR2	50SE	36,695,011	0.9437
mESC_1hActD_BR1	50SE	33,308,966	0.9630
mESC_1hActD_BR2	50SE	37,514,855	0.9517
mESC_1hEtOH_BR1	50SE	37,548,799	0.9674
mESC_1hEtOH_BR2	50SE	37,999,601	0.9534
mESC_4hActD_BR1	50SE	38,052,467	0.9554
mESC_4hActD_BR2	50SE	37,918,832	0.9317
mESC_4hEtOH_BR1	50SE	38,519,837	0.9542
mESC_4hEtOH_BR2	50SE	36,631,094	0.9567
MEF_0h_BR1	50SE	39,113,461	0.9471
MEF_0h_BR2	50SE	36,525,575	0.9574
MEF_1hActD_BR1	50SE	36,737,814	0.9539
MEF_1hActD_BR2	50SE	39,268,481	0.9491
MEF_1hEtOH_BR1	50SE	35,433,267	0.9393
MEF_1hEtOH_BR2	50SE	37,254,483	0.9570
MEF_4hActD_BR1	50SE	42,179,393	0.9503

MEF_4hActD_BR2	50SE	42,439,924	0.9449
MEF_4hEtOH_BR1	50SE	34,467,192	0.9720
MEF_4hEtOH_BR2	50SE	38,851,303	0.9593
rESC_0h_BR1	50SE	25,787,240	0.9428
rESC_0h_BR2	50SE	32,088,572	0.9382
rESC_1hActD_BR1	50SE	34,632,018	0.9344
rESC_1hActD_BR2	50SE	29,623,153	0.9330
rESC_1hEtOH_BR1	50SE	28,526,438	0.9288
rESC_1hEtOH_BR2	50SE	33,382,747	0.9350
rESC_4hActD_BR1	50SE	32,778,668	0.9329
rESC_4hActD_BR2	50SE	29,087,533	0.9302
rESC_4hEtOH_BR1	50SE	26,010,647	0.9401
rESC_4hEtOH_BR2	50SE	32,621,010	0.9402
REF_0h_BR1	50SE	34,352,856	0.9343
REF_0h_BR2	50SE	38,653,134	0.9490
REF_1hActD_BR1	50SE	35,734,464	0.9501
REF_1hActD_BR2	50SE	40,221,945	0.9502
REF_1hEtOH_BR1	50SE	39,700,369	0.9202
REF_1hEtOH_BR2	50SE	38,182,997	0.9442
REF_4hActD_BR1	50SE	43,322,613	0.9531
REF_4hActD_BR2	50SE	40,178,646	0.9431
REF_4hEtOH_BR1	50SE	36,226,698	0.9486
REF_4hEtOH_BR2	50SE	39,033,144	0.9421

RNA-seq reads of biological replicates were combined before alignment.

3.11.2 Normalization of RNA stability data

Treatment of cells with Actinomycin D leads to a stalling of RNA PolII and an inhibition of RNA synthesis. Short-lived RNA molecules are rapidly degraded while other RNA molecules can be stable for 24h and more. Analysis of RNA populations after Actinomycin D treatment to estimate RNA stability needs to be done carefully as it leads to a phenomenon that can dramatically skew data interpretation. As short-lived RNAs are degraded and removed from the RNA pool the total amount of RNA is reduced. The abundance of remaining RNAs is therefore relatively increased in the reduced pool compared to the larger pool before Actinomycin D treatment. I tried to correct this bias by normalizing the RPKMs of all transcripts to a basket of 10 stable housekeeping genes. I have chosen 15 commonly used housekeeping genes for mouse and rat and eliminated 5 for mouse (*Gapdh*, *Ppia*, *Ubc*, *Tubb4b*, *Uba52*) and rat (*Hprt*, *Rplp1*, *Rps14*, *Tubb4b*, *Ubc*) due to either too low expression or decrease upon ActD treatment. This left me with 10 housekeeping genes for both mouse (*Actb*, *B2m*, *Cnbp*, *Gusb*, *Rplp0*, *Rplp1*, *Rplp2*, *Rps14*, *Tmsb4x*, *Hprt*) and rat (*ActB*, B2m, *Cnbp*, *Gapdh*, *Gusb*, Ppia, *Rplp0*, *Rplp1*, *Tmsb4x*, *Uba52*) cell types.

3.11.3 Calculation of RNA stability

After normalization, RNA stability was calculated in two steps: Equation #1 shows the calculation of the intermediate RNA stability by normalizing the 4h Actinomycin D RPKM to the

0h untreated control RPKM. Equation #2 shows the calculation of the final RNA stability by normalizing the intermediate RNA stability to the 4h EtOH control treatment RPKM. Similarly, the RNA stability after 1h Actinomycin D treatment was calculated, however, the 1h Actinomycin D treatment was too short to yield a good enough resolution of RNA stability (see Figure 21C, 2nd lanes, compared to 4th lanes). As a result, for each transcript a specific RNA stability value was given representing percent of RNA that was left after 4h of Actinomycin D treatment relative to corresponding 0h untreated control and relative to EtOH control treatment.

Equation #1: $$RNA_{stability\ (intermediate)} = \frac{RPKM_{4hActD} * 100}{RPKM_{0h}}$$

Equation #2: $$RNA_{stability} = \frac{RNA_{stability\ (intermediate)} * 100}{RPKM_{4hEtOH}}$$

3.11.4 Applying RPKM and RPKM saturation cut offs

For analysis of RNA stability I applied two filter steps to remove transcript loci that are not stably detected. First, a cutoff required each loci to be expressed with an RPKM > 0.1 in the untreated control. Second, an RPKM saturation filter step removed loci whose RPKM stability (see chapter 3.11.3) was not stably detected. Therefore, I calculated RPKM saturation (see chapter 3.9.7) and calculated RNA stability as indicated in chapter 3.11.3 also for 20 samples with differing read numbers (5% to 100%). If the RPKM error of RNA stability of the 70-90% samples relative to the 100% sample was on average >5%, the whole locus was removed from further analysis, thereby ensuring that only stably detected RPKMs are further analyzed.

3.11.5 Quality control of RNA stability data

I checked the RNA stability data by visually inspecting stable und unstable RNAs in the UCSC genome browser (Fujita et al., 2011). Additionally, I plotted the RNA stability of expressed RefSeq mRNAs in heatmaps using the R function `pheatmap` to show the fraction of unstable mRNAs, the difference between 1h Actinomycin D and 4h Actinomycin D treatment and the largely unaffected EtOH control treatments.

3.12 Analysis of RNA export

3.12.1 Self-generated RNA-seq datasets

Table 15: Self-generated RNA-seq data used to calculate RNA export

Dataset	Read type	nr of reads	Strand specificity
mESC_cyt	50SE	121,078,769	0.9577
mESC_nuc	50SE	80,006,456	0.9606
MEF_cyt_BR1	50SE	39,753,442	0.9589
MEF_cyt_BR2	50SE	40,964,526	0.9587
MEF_nuc_BR1	50SE	35,191,578	0.9499
MEF_nuc_BR2	50SE	40,815,555	0.9430
REF_cyt	50SE	94,602,675	0.9121
REF_nuc	50SE	90,941,801	0.9647

cyt, cytoplasmic; nuc, nuclear; BR, biological replicate

3.12.2 Calculation of RNA export

RNA export was calculated for each transcript by the following equation and expressed as percent of total RNA detected in the cytoplasm:

$$RNA\_export = \frac{RPKM\_cyt * 100}{RPKM\_cyt + RPKM\_nuc}$$

3.12.3 Applying RPKM and RPKM saturation cut offs

For analysis of RNA export I applied two filter steps to remove transcript loci that are not stably detected. First, a cutoff required each loci to be expressed with an RPKM > 0.1 in either the cytoplasmic or the nuclear fraction. Second, an RPKM saturation filter step removed loci whose RPKM export (see chapter 3.12.2) was not stably detected. Therefore, I calculated RPKM saturation (see chapter 3.9.7) and calculated RNA export as indicated in chapter 3.12.2 also for 20 samples with differing read numbers (5% to 100%). If the RPKM error of RNA export of the 70-90% samples relative to the 100% sample was on average >5%, the whole locus was removed from further analysis, thereby ensuring that only stably detected RPKMs are further analyzed.

3.12.4 Quality control of RNA export data

I checked the RNA export data by visually inspecting nuclear und cytoplasmic RNAs in the UCSC genome browser (Fujita et al., 2011).

3.13 Analysis of RNA splicing

3.13.1 Self-generated RNA-seq datasets

Table 16: Self-generated RNA-seq data used to calculate RNA splicing

Dataset	Read Type	Nr of Reads	Strand Specificity
mESC_RZ (*)	100PE	134,089,235	0.9274
MEF_RZ_BR1	100PE	61,108,202	0.9597
MEF_RZ_BR2	100PE	56,901,263	0.9614
rESC_RZ_BR1	100PE	45,535,600	0.9476
rESC_RZ_BR2	100PE	57,584,616	0.9340
REF_RZ_BR1	100PE	37,869,122	0.9526
REF_RZ_BR2	100PE	51,653,271	0.9618
4x mESC_RZ (**)	100PE	125,802,515	0.9448
4x MEF_RZ (**)	100PE	213,227,892	0.9406

RZ, Ribo-Zero; BR, biological replicate
(*) pool from four libraries of hydrolysis experiment (see chapter 4.1.4)
(**) data from Daniel Andergassen & Quanah Hudson, libraries prepared by me

3.13.2 Calculation of RNA splicing

I created a splice junction annotation from the combined lncRNA and mRNA annotation in which for every single splice junction two 45bp blocks were present: one block 5bp away from the annotated junction on the exon side, and one block 5bp away from the annotated junction on the intron side (see Figure 22B,C). If the exon or intron was shorter than 50bp, the block size was reduced accordingly, but had to have a minimum length of 10bp. In total, 743,556 junctions were essayed and for each an exonic and an intronic RPKM were calculated. For RPKM calculation, I combined my mouse Ribo-Zero RNA-seq data with Ribo-Zero RNA-seq data of four biological replicates of undifferentiated primary E3.5 mESC and four biological replicates of primary E12.5 MEF, both isolated from Cast/FvB crosses (Daniel Andergassen & Quanah Hudson, manuscript in preparation). Increased mESC and MEF read numbers resulted in enhanced read coverage and thereby higher accuracy of splicing estimation. Percent RNA splicing was calculated for each junction by the following formula:

$$RNA_{splicing} = 100 - (\frac{RPKM_{intron} * 100}{RPKM_{exon}})$$

3.13.3 Applying RPKM and RPKM saturation cut offs

If the intron RPKM and the exon RPKM where equal and >1, RNA splicing was zero. If the intron RPKM was higher than the exon RPKM (i.e. RNA splicing had a negative value), RNA splicing was set to zero. If the intron RPKM or the exon RPKM was zero, the corresponding other RPKM had to be >1. Non-informative junctions (both RPKM are zero or >1) were

discarded from further analysis. Additionally, an RPKM saturation filter step removed junctions whose RNA splicing value (see chapter 3.13.2) was not stably detected. Therefore, I calculated RPKM saturation (see chapter 3.9.7) and calculated RNA splicing as indicated in chapter 3.13.2 also for 20 samples with differing read numbers (5% to 100%). If the RPKM error of RNA splicing of the 70-90% samples relative to the 100% sample was on average >10%, the junction was removed from further analysis, thereby ensuring that only stably detected junctions are further analyzed.

3.13.4 Averaging splicing values over transcripts and loci

Splicing values of remaining junctions were averaged for each transcript. Splicing values of transcripts were averaged for each locus, giving rise to a robust RNA splicing value for each locus.

3.14 Clustering of lncRNAs by RNA biology

LncRNAs and mRNAs were clustered according to their three RNA biology features (stability, export, splicing) using the popular clustering algorithm k-means (Hartigan and Wong, 1979). To make the clustering reproducible, a seed was set beforehand using the R function `set.seed(300)`. The clustering itself was carried out using the R function `kmeans()` with the parameters `centers=6`, `iter.na=500` and `nstart=10`. The number of clusters was empirically tested and six clusters seemed most suitable to represent the diverse RNA biology of lncRNAs and mRNAs. Fewer clusters led to reduced resolution by combining fundamentally different RNAs and more clusters resulted in the artificial splitting up of single clusters giving rise to very similar clusters.

3.15 Conservation of RNA biology

In order to assay the conservation of RNA biology, I first had to define mouse and rat RNAs that are annotated in syntenic regions. Therefore, the rat annotation (genome build rn5) was lifted over to the mouse genome build mm10 using the UCSC tool liftover with the following commands:

```
liftOver -minMatch=0.3 -minBlocks=0.3 -fudgeThick
rat_annotation.bed rn5ToMm10.over.chain.gz
rat_annotation.mm10.bed12 rat_annotation.unmapped.bed12
```

Using these parameters, I find mouse homologs for 99.35% of rat mRNAs and for 84.47% of rat lncRNAs. Next, I tested whether mm10 RNAs and rn5->mm10 RNAs overlap each other (meaning that they are annotated in the same syntenic region) and isolated those that overlap each other by >30% using intersectBed:

```
intersectBed -wao -s -f 0.3 -a rat_annotation.mm10.bed6 -b
mouse_lncRNAs.bed6 > intersect_out.bed
```

I find that 1,964 mouse lncRNA loci (21.50%) and 14,543 mouse mRNA loci (97.30%) have annotated RNAs in syntenic rat regions. I compared the RNA biology features of these conserved RNAs between mouse and rat.

3.16 RNA-seq data used for developmental regulation of lncRNAs

The following self-generated RNA-seq data derived from ScriptSeq v1 libraries were used to calculate developmental regulation of lncRNAs:

Table 17: Self-generated RNA-seq data used to calculate developmental regulation

Dataset	Read Type	Nr of Reads	Strand Specificity
Adult Heart	51PE	65,341,797	0.9702
Fetal Heart BR1	51PE	71,857,811	0.9814
Fetal Heart BR2	51PE	60,916,788	0.9815
Adult Liver	51PE	64,915,207	0.9801
Fetal Liver BR1	51PE	48,766,733	0.9644
Fetal Liver BR2	51PE	67,562,979	0.9774
Adult Spleen BR1	51PE	64,703,965	0.9537
Adult Spleen BR2	51PE	63,940,635	0.9536
Adult Spleen BR3	51PE	69,318,167	0.9527
B cells	51SE	260,046,325	0.9387
CD4+ T cells	51SE	251,421,559	0.9379
CD8+ T cells	51SE	261,429,795	0.9408

3.17 Analysis of ChIP-seq data to annotate enhancer lncRNAs

3.17.1 Public ChIP-seq datasets used

The following published H3K4me1, H3K4me3, H3K27ac and p300 ChIP-seq datasets of 18 mouse tissues (Shen et al., 2012) were used to map enhancers in order to annotate enhancer RNAs (eRNAs).

Table 18: Public ChIP-seq datasets used in this study

Histone modification	Tissue	Read Type	Aligned Reads	Nr of Peaks
Input	Bone Marrow	36SE	13,835,996	-
Input	Cerebellum (Brain)	36SE	16,967,638	-
Input	Cortex (Brain)	36SE	18,750,599	-
Input	Embryonic Brain	36SE	28,182,051	-
Input	Embryonic Heart	36SE	12,693,848	-
Input	Embryonic Limb	36SE	16,757,607	-
Input	Embryonic Liver	36SE	18,581,229	-

Input	Heart	36SE	15,871,588	-
Input	Intestine (Colon)	36SE	29,229,438	-
Input	Kidney	36SE	16,658,461	-
Input	Liver	36SE	17,576,197	-
Input	Lung	36SE	10,005,668	-
Input	Olfactory Bulb	36SE	31,125,180	-
Input	Placenta	36SE	17,893,882	-
Input	MEF	36SE	19,259,220	-
Input	mESC	36SE	15,730,741	-
Input	Spleen	36SE	11,700,231	-
Input	Testis	36SE	7,466,083	-
H3K4me1	Bone Marrow	36SE	26,363,871	29,506
H3K4me1	Cerebellum (Brain)	36SE	20,955,568	75,052
H3K4me1	Cortex (Brain)	36SE	35,152,810	41,683
H3K4me1	Embryonic Brain	36SE	45,269,644	79,665
H3K4me1	Embryonic Heart	36SE	25,547,373	64,861
H3K4me1	Embryonic Limb	36SE	16,872,321	75,763
H3K4me1	Embryonic Liver	36SE	27,775,925	56,342
H3K4me1	Heart	36SE	33,876,588	58,551
H3K4me1	Intestine (Colon)	36SE	18,953,493	66,681
H3K4me1	Kidney	36SE	18,295,701	57,889
H3K4me1	Liver	36SE	16,585,927	83,388
H3K4me1	Lung	36SE	18,717,952	45,542
H3K4me1	Olfactory Bulb	36SE	13,524,927	65,125
H3K4me1	Placenta	36SE	10,220,560	36,987
H3K4me1	MEF	36SE	20,180,338	96,225
H3K4me1	mESC	36SE	18,682,778	66,060
H3K4me1	Spleen	36SE	9,133,039	58,696
H3K4me1	Testis	36SE	21,972,808	31,808
H3K4me3	Bone Marrow	36SE	22,926,986	19,859
H3K4me3	Cerebellum (Brain)	36SE	22,262,180	22,725
H3K4me3	Cortex (Brain)	36SE	16,142,643	24,110
H3K4me3	Embryonic Brain	36SE	37,583,123	21,896
H3K4me3	Embryonic Heart	36SE	22,900,122	21,782
H3K4me3	Embryonic Limb	36SE	44,675,164	21,251
H3K4me3	Embryonic Liver	36SE	24,509,946	19,995
H3K4me3	Heart	36SE	14,965,378	23,945
H3K4me3	Intestine (Colon)	36SE	38,627,598	27,214
H3K4me3	Kidney	36SE	18,873,578	21,404
H3K4me3	Liver	36SE	27,648,883	20,889
H3K4me3	Lung	36SE	15,790,217	21,308
H3K4me3	Olfactory Bulb	36SE	14,577,172	18,190
H3K4me3	Placenta	36SE	23,954,349	30,306
H3K4me3	MEF	36SE	19,950,704	21,698
H3K4me3	mESC	36SE	22,131,262	15,622
H3K4me3	Spleen	36SE	9,964,841	21,281
H3K4me3	Testis	36SE	34,721,899	35,529
H3K27ac	Bone Marrow	36SE	23,409,037	30,006
H3K27ac	Cerebellum (Brain)	36SE	15,554,055	31,526
H3K27ac	Cortex (Brain)	36SE	12,623,031	36,895
H3K27ac	Embryonic Brain	36SE	15,991,870	36,916
H3K27ac	Embryonic Heart	36SE	17,327,725	40,788
H3K27ac	Embryonic Limb	36SE	14,784,293	35,470
H3K27ac	Embryonic Liver	36SE	17,735,247	34,600

H3K27ac	Heart	36SE	18,117,256	43,404
H3K27ac	Intestine (Colon)	36SE	15,329,666	43,939
H3K27ac	Kidney	36SE	16,965,342	46,004
H3K27ac	Liver	36SE	19,967,447	41,975
H3K27ac	Lung	36SE	7,877,134	44,446
H3K27ac	Olfactory Bulb	36SE	13,522,059	42,954
H3K27ac	Placenta	36SE	12,733,402	31,630
H3K27ac	MEF	36SE	15,135,913	40,390
H3K27ac	mESC	36SE	21,171,342	44,055
H3K27ac	Spleen	36SE	15,342,277	28,916
H3K27ac	Testis	36SE	13,630,097	23,907
p300	Heart	36SE	16,909,148	45,320
p300	Kidney	36SE	6,079,062	13,112
p300	Liver	36SE	9,223,591	7,268
p300	Lung	36SE	10,595,179	22,368

3.17.2 Alignment

ChIP-seq data was aligned to the mouse genome mm10 using bowtie (version 2.0.6) (Langmead et al., 2009) and the following commands:

```
bowtie2 -x bowtie2_index --very-sensitive-local -U
reads.fastq.gz --no-unal -p4 -k1
```

The numbers of aligned reads for each tissue and histone modification are listed in Table 18.

3.17.3 Peak calling

ChIP-seq peaks were called by the program MACS (version 1.4.) (Zhang et al., 2008) using the following commands:

```
macs14 --treatment ChIP.sorted.bam --control Input.sorted.bam --
format BAM -g mm --name ChIP_tissue
```

The numbers of detected peaks for each tissue and histone modification are listed in Table 18. For each histone modifications a combined track was prepared by concatenating respective peaks from all tissues. Peaks were merged when the distance of peaks was smaller than 1000bp. An enhancer track was prepared by concatenating H3K4me1, H3K27ac and p300 peaks from all tissues. Enhancers had to be at least 1kb away from H3K4me3 peaks. The final enhancer track listed 255,281 enhancers being defined by H3K4me1, H3K27ac or p300 binding and being devoid of H3K4me3.

3.18 Statistical analyses

Statistical analysis and plots were done in R statistical environment (r-project.org), if no package name is specified analyses were carried out using the base package. Boxplots depict

the median (bold line), the 25th and 75th percentile (lower and upper box end) and the 1.5-fold interquartile range (IQR) (whiskers). Outliers lying outside the 1.5-fold IQR are plotted individually. Notches around the median illustrate a rough estimate of the significance of the difference between the medians. If the notches of two boxes do not overlap, there is strong evidence (95% confidence interval) that the medians differ significantly (Graphical Methods for Data Analysis, John M. Chambers, 1983). The notches are calculated by the 1.58-fold IQR divided by the square root of the numbers of observation, thereby being sensitive to the number of observations. Correlation coefficients were calculated by Pearson's analysis, unless otherwise noted.

4 RESULTS

4.1 Optimization of RNA-seq workflow

4.1.1 The Ribo-Zero kit efficiently removes abundant rRNA species

Ribosomal RNA species account for 95-98% of a cellular transcriptome and to improve sequencing coverage of mRNAs and the less abundant lncRNAs they have to be removed from the RNA preparation prior to RNA-seq. The RiboMinus Eukaryote Kit for RNA-Seq (Life Technologies) was the first commercially available kit to remove rRNA but when we analyzed RNA-seq data derived from RiboMinus treated RNA we found that 46-73% of the reads still mapped to rRNA species (Huang et al., 2011). At the end of 2010, the Epicentre company launched the Ribo-Zero rRNA Removal Kit (Human/Mouse/Rat) as an alternative and I tested its efficiency by qPCR. Total DNaseI treated HeLa RNA was subjected to Ribo-Zero rRNA removal according to the protocol. Samples that were rRNA depleted and untreated control samples were reverse transcribed into cDNA and rRNA abundance measured by qPCR. The four rRNA species 5S, 5.8S, 18S and 28S were efficiently removed in the Ribo-Zero treated HeLa samples as rRNA levels were reduced by >99% relative to the untreated controls (Figure 4). I used this kit for human, mouse and rat RNA samples and found consistent performance. RNA-seq results of many libraries I prepared for internal and external collaborations indicated that after rRNA removal using the standard Ribo-Zero kit approximately 4-5% of RNA-seq reads remained that aligned to rRNA. The low input version Ribo-Zero rRNA Removal Kit (Human/Mouse/Rat) Low Input was more effective with only 0.5-1% of RNA-seq reads aligning to rRNA species after rRNA removal (data not shown).

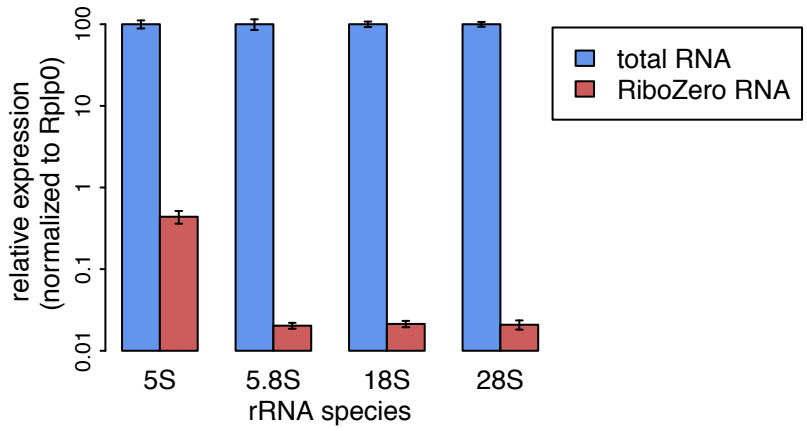


Figure 4: The Ribo-Zero kit efficiently removes abundant rRNA species

2µg of HeLa RNA were subjected to rRNA removal using the Ribo-Zero kit. Levels of 5S, 5.8S, 18S and 28S rRNA species were measured by qPCR in cDNA from Ribo-Zero treated samples and untreated control samples. Expression levels were normalized to the mRNA of the ribosomal protein *Rplp0*. Expression levels in total RNA were set to 100 and Ribo-Zero treated samples relative to it. Data is log10 transformed and depicted as mean, error bars represent variation of three technical replicates.

4.1.2 The ScriptSeq kit produces strand-specific libraries with 500ng input

The first RNA-seq libraries in the Barlow lab were made from double-stranded cDNA, a method that does not preserve strand specificity. As many non-coding RNAs overlap protein-coding genes in antisense orientation (e.g. Airn overlaps Igf2r), I strove to conduct strand-specific RNA-seq in order to be able to distinguish overlapping transcripts. The company Epicentre launched the ScriptSeq RNA-seq Library Preparation Kit (referred to as ScriptSeq v1 kit within this thesis) that allowed to prepare strand-specific RNA-seq libraries in half a day. I tested two parameters of the kit: the amount of RNA input (500ng vs. 1000ng) before Ribo-Zero treatment and the final library purification method (gel vs. column). The decision to test total RNA input instead of testing different amounts of Ribo-Zero treated RNA was made because Ribo-Zero treatment typically removes ~90-95% of an RNA sample, which makes accurate RNA quantification after Ribo-Zero treatment difficult. I used 500ng and two samples of 1000ng of mouse E14.5 fetal head total RNA that were Ribo-Zero treated and ScriptSeq libraries prepared with them. One of the 1000ng reactions was cleaned using a column while the second 1000ng reaction plus the 500ng reaction were cleaned using a gel. The libraries were sequenced (conditions: 50bp, single-end) and 65 million reads were taken from each dataset for a comparable analysis. In terms of gene expression, the samples with two different input amounts are well correlated (r=0.993) (Figure 5A, left), as are the samples with the two different library purification methods (r=0.979) (Figure 5A, right). The overall gene-body coverage was higher in the gel-cleaned samples, however, unexpectedly for total RNA, the gene-body coverage of all three ScriptSeq v1 libraries showed a strong 5'end bias (Figure 5B). The strand-specificity was similarly good for all samples, ranging from 94.51% to 95.58% of all reads mapping to the correct strand (data not shown), however, the number of unmapped reads was considerably higher in the column-cleaned samples compared to the gel-cleaned sample (8.78% vs. 2.93%) (Figure 5C). Taken together, these results established that RNA input can be as low as 500ng of total RNA and that library purification by gel is the preferred method. I then used the ScriptSeq v1 kit to prepare RNA-seq libraries for this study (see chapter 4.2.7 and 4.2.8) and also internal (Huang et al., 2011; Koerner et al., 2012) as well as external collaborations (Bürckstümmer et al., 2013).

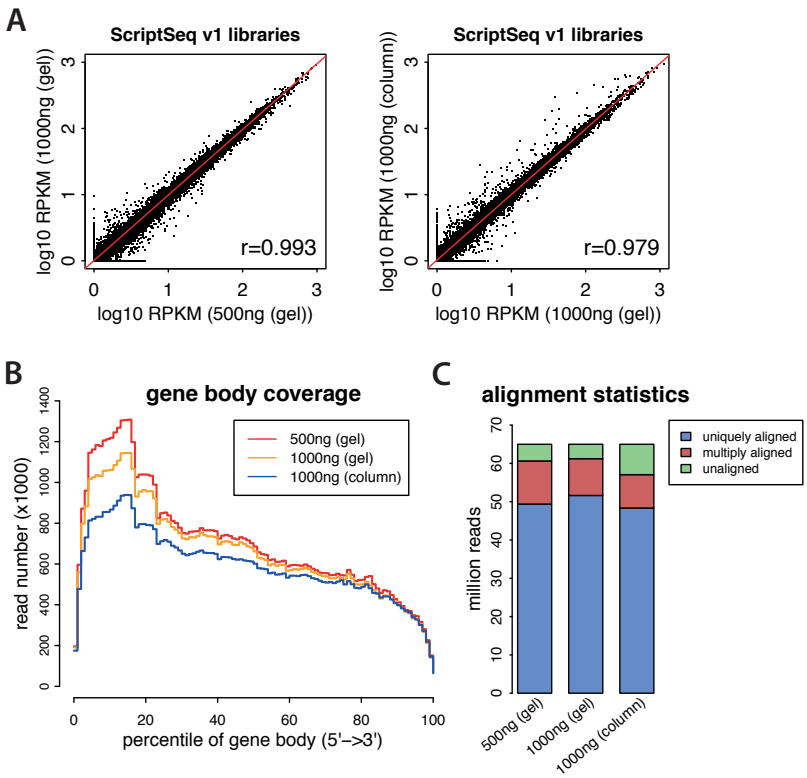


Figure 5: The ScriptSeq v1 kit produces strand-specific libraries with 500ng input

65 million reads (50bp, single-end) of mouse E.14.5 fetal head RNA-seq were used from each ScriptSeq v1 dataset to compare the influence of RNA input and library purification method. (**A**) RPKM of RefSeq transcripts (n=33,092) are plotted for two libraries with different input amounts (left) and two libraries with different library purification method (right). RPKM are log10 transformed. The Pearson correlation coefficient is indicated in the bottom right corners. (**B**) Gene-body coverage of the RefSeq annotation (n=33,092) was calculated for all three ScriptSeq v1 RNA-seq datasets. The gene-body of all transcripts was divided into percentiles to normalize for different transcript lengths. (**C**) Alignment statistics depicting the number of unmapped, multiply mapped and uniquely mapped reads of three ScriptSeq v1 libraries.

4.1.3 The dUTP/TruSeq protocol produces superior libraries for transcript assembly

In all the above-mentioned projects and collaborations, ScriptSeq v1 RNA-seq data was used to calculate gene expression. However, when I first started to assemble transcriptomes, I found that RNA-seq coverage from ScriptSeq v1 libraries across genes was variable and the read distribution indicated strong regional biases (Figure 6A, peaks in red tracks). In this study, I used RNA-seq data to estimate splicing efficiency of lncRNAs (see Figure 9 for an overview) and these biases invalidated splicing analysis and hindered transcriptome assemblies. Therefore, I tested two alternative methods to prepare strand-specific libraries. The company Epicentre launched an improved version of its original ScriptSeq v1 kit (ScriptSeq™ v2 RNA-Seq Library Preparation Kit, referred to as ScriptSeq v2 within this thesis) with a streamlined protocol, lower input (0.5-50ng RNA) and improved hexamer design to facilitate more even transcript coverage. The TruSeq™ RNA Sample Prep Kit (Illumina) was already established at CeMM but did not produce strand-specific libraries, however, a modified protocol adding two steps to the standard Illumina protocol rendered the

libraries strand-specific (Sultan et al., 2012). Specifically, these two steps are replacement of dTTP by dUTP during second-strand synthesis and a digestion by Uracil-DNA Glycosylase (New England Biolabs) at the end of library preparation to remove the dUTP containing second strand before PCR. I prepared libraries from the same adult spleen RNA sample using the ScriptSeq v2 kit (test of 8 vs. 12 PCR cycles) and the modified dUTP/TruSeq protocol (test of 0.5 vs. 4µg input) and sequenced the four libraries (conditions: 50bp, single-end). The Pearson correlation of RefSeq RPKM within the two dUTP/TruSeq or the two ScriptSeq v2 libraries is very good (r=0.982 and r=0.994, respectively) (Figure 6B). The correlation between a ScriptSeq v2 library (8 cycles) and a dUTP/TruSeq library (4µg) is less pronounced (r=0.667), indicating differences between the two library preparations (Figure 6C, left). The correlation between a ScriptSeq v1 library and a dUTP/TruSeq library (4µg) (r=0.482) is even lower and shows that a direct comparison of gene expression between these libraries could be flawed (Figure 6C, left). The gene-body coverage of the dUTP/TruSeq libraries is much higher and more evenly distributed compared to the ScriptSeq v2 libraries (Figure 6C) and also known splice junctions are better detected in the dUTP/TruSeq libraries (Figure 6D). When RNA-seq data of the four libraries is displayed on the genome browser UCSC (Figure 6E, green tracks for ScriptSeq v2 libraries, blue tracks for dUTP/TruSeq libraries) it becomes evident that the dUTP/TruSeq protocol produces superior libraries. The inefficiently spliced lncRNA Malat1 is evenly covered in the dUTP/TruSeq tracks (blue tracks) whereas the ScriptSeq v2 tracks show spikes of reads (green tracks), similar to ScriptSeq v1 data (red tracks). Six of the peaks in the ScriptSeq v1 and v2 libraries go beyond the cutoff as high as 25,000 (Figure 6E, indicated by arrows). Based on the data presented in Figure 6, I concluded that the modified dUTP/TruSeq protocol is better suited for transcriptome assembly, however, for gene expression purposes RNA-seq data generated with the ScriptSeq v1 and v2 kits is still useful.

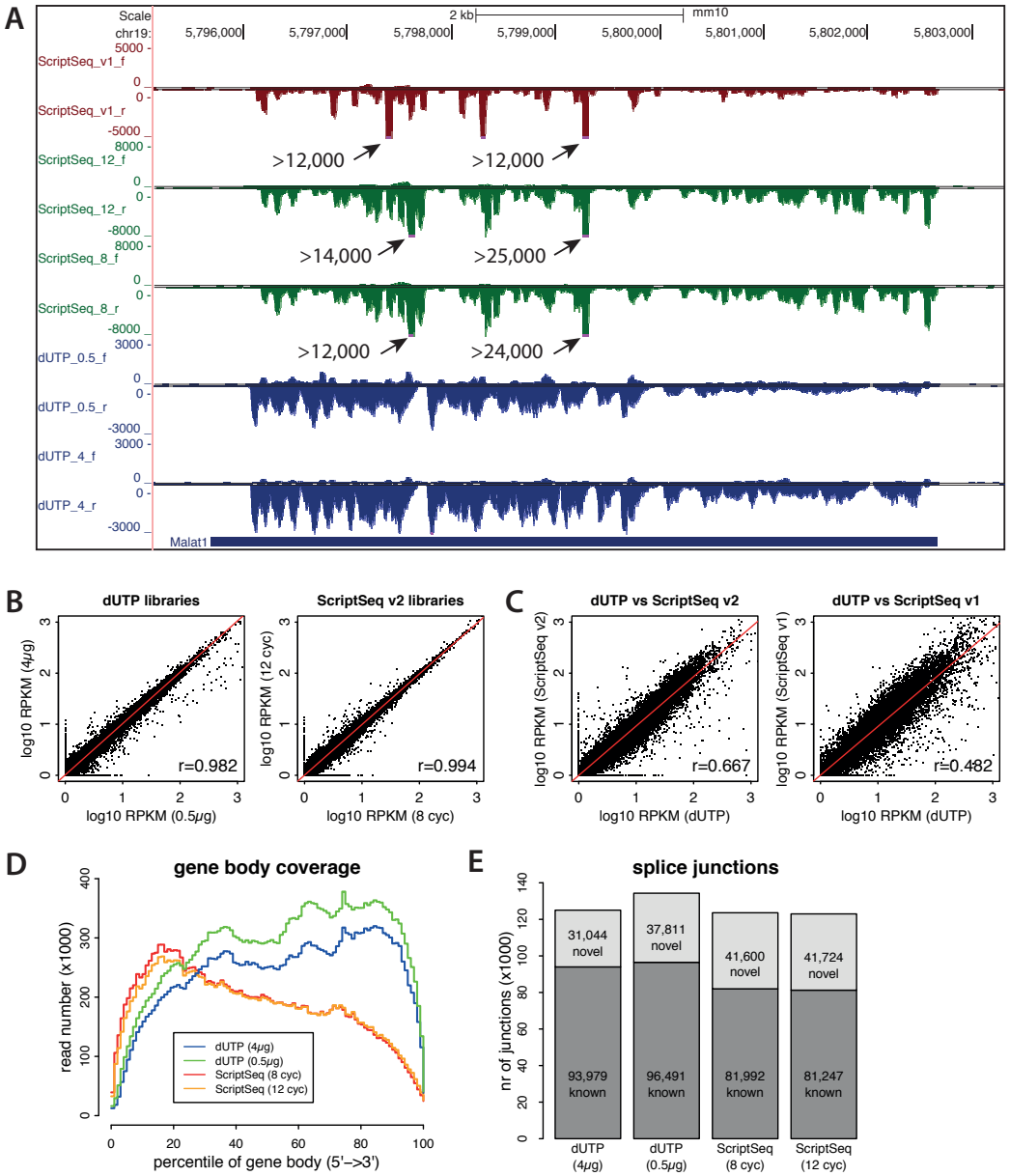


Figure 6: The dUTP/TruSeq protocol produces superior libraries for transcript assembly

25 million reads (50bp, single-end) of mouse adult spleen RNA-seq were used from each library test dataset to compare the ScriptSeq v2 kit (test of 8 vs 12 PCR cycles) and dUTP/TruSeq libraries (test of 0.5µg vs. 4µg input). (**A**) RNA-seq data of five libraries for Malat1 is displayed in the UCSC genome browser. Chromosome and genomic coordinates are displayed on top. Two tracks are displayed for every sample, the upper one corresponding to the forward strand the lower one corresponding to the reverse strand. The RefSeq annotation of Malat1 is displayed at the bottom. Arrows indicate regions of unusually high read coverage in the ScriptSeq v1 and v2 tracks. (**B**) RPKM of RefSeq transcripts (n=33,092) are plotted for two dUTP/TruSeq (left) and two ScriptSeq v2 libraries (right). The Pearson correlation coefficient is indicated in the bottom right corner. (**C**) RPKM of RefSeq transcripts (n=33,092) are plotted for ScriptSeq v2 (8 cycles) vs. dUTP/TruSeq (4µg) libraries (left) and ScriptSeq v1 vs dUTP/TruSeq libraries (right, visceral endoderm RNA-seq data taken from Kulinski et al., submitted) (**D**) Gene-body coverage of RNA-seq reads was calculated for all four library test datasets for the complete RefSeq annotation (n=33,092) (**E**) Depicted are the numbers of known (RefSeq) and novel splice junctions detected in the four RNA-seq datasets.

4.1.4 RNA hydrolysis time has only marginal influence on transcript assembly

While the ScriptSeq kits use full-length RNA as input, the dUTP/TruSeq protocol requires RNA hydrolysis prior to cDNA synthesis. The TruSeq protocol suggests 8min of hydrolysis, which should yield RNA fragments with a median length of 155bp. In case such a library is sequenced 100bp paired-end, 45bp in the middle of the fragment would be covered by both reads of the read pair, i.e. sequenced twice. To reduce this overlap, I tested three different hydrolysis times and determined the degree of read overlap and its influence on the number of splice junctions and on transcript assembly (Figure 7). I used DNaseI and Ribo-Zero treated RNA from undifferentiated CCE ES cells as input and prepared four libraries with hydrolysis times of 1min, 3min (2 replicates) and 8min, all of which were sequenced 100bp paired-end. In order to have a comparable analysis, I randomly sampled 19 million reads from each dataset before analysis. As can be seen in Figure 7A, the overlap of paired reads is reduced with shorter hydrolysis times. A negative inner distance indicates read overlap (in bp), while a positive inner distance indicates the distance (in bp) between two paired 100bp reads. The median overlap is reduced from 53bp (8min) to 33/38bp (3min replicates) and to 18bp (1min hydrolysis). I hypothesized that a greater overlap will lead to a reduced detection of splice junctions and as a consequence to poor transcript assembly. However, the number of splice junctions is only marginally affected by hydrolysis time (Figure 7B), as there is no clear trend visible between the three timepoints. In line with this finding, the transcript assembly is also very comparable, the median number of exons per assembled transcript comes in all samples very close to the RefSeq annotation (Figure 7C), a well annotated set of reference transcripts serving as the gold standard. One problem that arose was that it seemed that the longer the fragments are the lower their probability is to get sequenced. The four libraries were carefully quantified and pooled in equimolar ratios, however, the number of RNA-seq reads decreased approximately from 46 million (for 8min hydrolysis sample) to 37/31 million (for 3min hydrolysis replicates) and to 19 million (for 1min hydrolysis sample) reads. In order to overcome problems with decreased read numbers, I decided to use 3min of hydrolysis for all future RNA-seq experiments as it reduced the read overlap by ~20bp but still yielded enough RNA-seq reads.

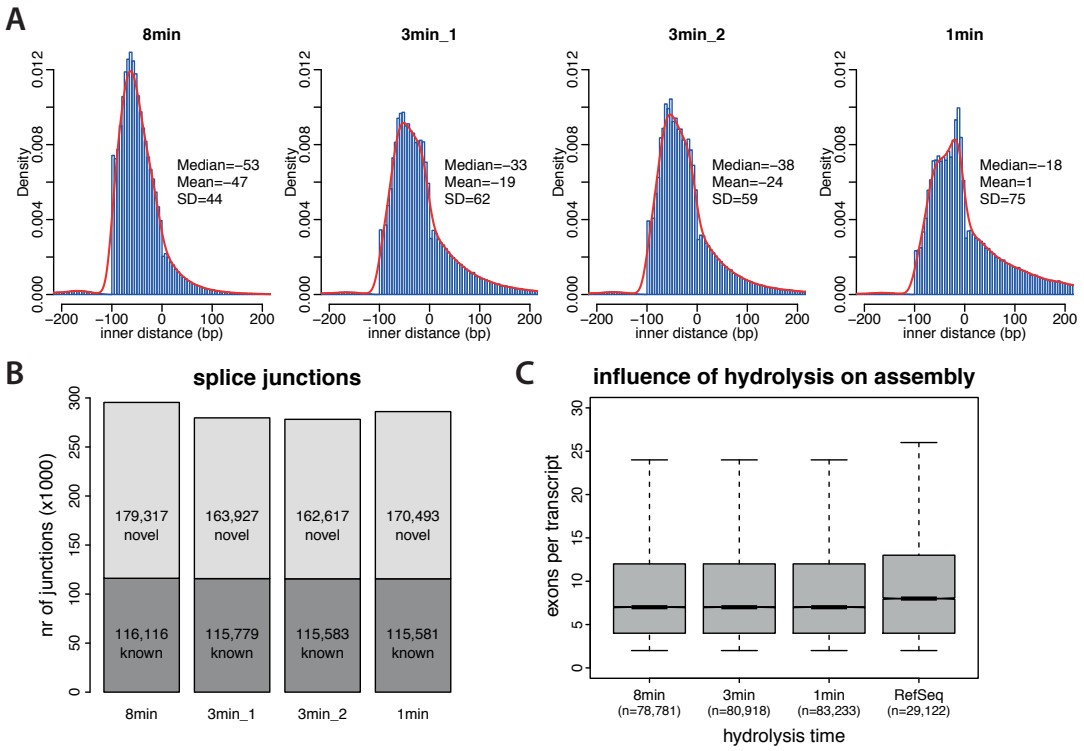


Figure 7: The RNA hydrolysis time has only marginal influence on transcript assembly

19 million reads (100bp, paired-end) of mouse undifferentiated CCE ES cell RNA-seq were used from each dataset with hydrolysis times of 8min, 3min (2 replicates) and 1min. (**A**) Analysis of the inner distance of sequenced cDNA fragments between the two read pairs in the four datasets. A negative inner distance indicates overlap of the two read pairs (in bp), a positive inner distance refers to the number of base pairs between the ends of the paired reads. (**B**) Depicted are the numbers of known (RefSeq) and novel splice junctions detected in four hydrolysis datasets. (**C**) Boxplots are showing the distribution of exon numbers of assembled transcripts of the four RNA-seq datasets. Exon numbers are derived from the cufflinks transcript files. The RefSeq mRNA annotation (n=29,122) serves as a reference. Single-exon transcripts were removed from assembled transcripts as well as the RefSeq annotation prior to the analysis.

4.1.5 100bp paired-end RNA-seq allows assembly of full-length transcripts

Now that the library preparation kit was established and the hydrolysis time optimized, the next question was which RNA-seq read type is most cost-effective to assemble transcripts. For this analysis I used 19 million 100bp paired-end reads from the 3min hydrolysis (replicate 1) dataset from above. Bioinformatically, I either removed the second read of the read pairs to give rise to a 100bp single-end dataset or trimmed the reads to give rise to a 50bp paired-end dataset. I then analyzed the number of detected splice junctions and assembled the transcriptome with cufflinks to determine the influence of the read type on transcript assembly (Figure 8). The transcriptome assembly relies on splice junctions and the more splice junctions are detected, the better lowly expressed and long transcripts are assembled. The number of detected known RefSeq splice junctions was comparable between all three datasets, however, the number of novel splice junctions increased by ~2.8 fold (50bp paired-end compared to 100bp single-end) and ~4 fold (50bp paired-end compared to 100bp paired-end) (Figure 8A). While the first increase is due to an increase in read length (at same sequencing depth), the second increase is a function of both increased read length and

increased coverage by read pairs. As a consequence of increased splice junction detection, the transcript assembly is also most effective with 100bp paired-end reads (Figure 8B). The numbers of exons per assembled transcript show almost the same distribution as the numbers of exons per transcript in the RefSeq annotation serving as the gold standard, that is why I decided to use 100bp paired-end RNA-seq for assembling transcripts throughout this study.

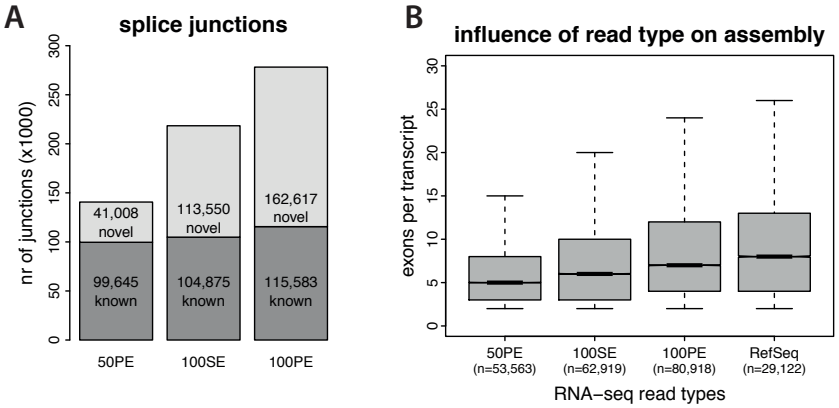


Figure 8: 100bp paired-end RNA-seq allows assembly of full-length transcripts

19 million reads (100bp paired-end) of mouse CCE ES cell RNA-seq were used to generate three datasets with the read types 50bp paired-end (50PE), 100bp single-end (100SE) and 100bp paired-end (100PE). (**A**) Depicted are the numbers of known (RefSeq) and novel splice junctions detected in each of these three RNA-seq datasets. (**B**) Boxplots are showing the distribution of exon numbers of cufflinks assembled transcripts of the three RNA-seq datasets. Exon numbers are derived from the cufflinks transcript files. The RefSeq mRNA annotation (n=29,122) serves as a reference. Single-exon transcripts were removed from assembled transcripts as well as the RefSeq annotation prior to the analysis.

4.2 Characterization of mouse and rat lncRNAs

4.2.1 Overview of annotation and analysis pipeline

The mouse and rat lncRNA annotations used in this study were generated by Florian Pauler (see chapter 3.10.1 for details). He assembled transcriptomes from my self-generated RNA-seq data from undifferentiated ES cells and primary embryonic fibroblasts (100bp paired-end, polyA, stranded) and published RNA-seq data from nine adult tissues (80bp paired-end, polyA, stranded) (Merkin et al., 2012) and applied several filter steps to obtain the final lncRNA annotations (Figure 9, left panel). The pipeline was set up for mouse RNA-seq data and then, with minor modifications (see methods section), also applied to rat RNA-seq data. RNA-seq data was aligned using STAR (Dobin et al., 2013) and transcriptomes were assembled using Cufflinks (Trapnell et al., 2013). After several filter steps to remove potentially coding transcripts and transcript artifacts, the assembled lncRNAs were grouped into loci representing transcriptional units. The lncRNA annotation was complemented with RefSeq mRNAs and then used in the second step to investigate RNA stability, RNA localization and RNA splicing (Figure 9, right panel). For each of these three RNA biology features, self-generated RNA-seq datasets of mouse and rat ES cells and embryonic fibroblasts were analyzed.

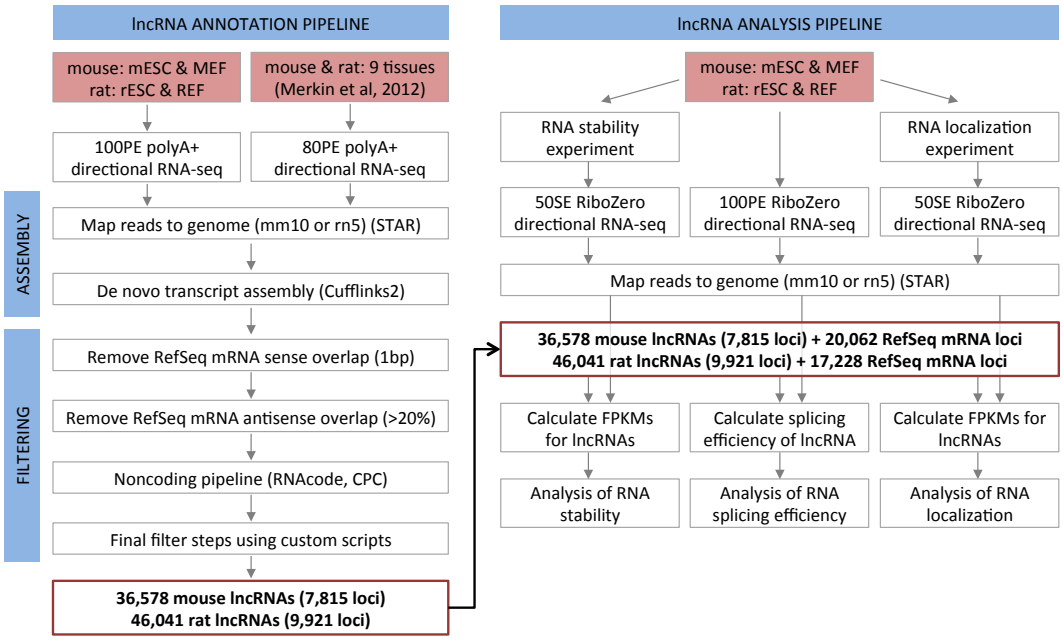


Figure 9: Overview of annotation and analysis pipeline

For this study, self-generated and public RNA-seq datasets of eleven cell types were used by Florian Pauler to generate novel and comprehensive lncRNA annotations for mouse and rat (left panel, see chapter 3.10.1 for details). The mouse and rat lncRNA annotations were complemented with RefSeq mRNAs and in the second step used to analyze RNA stability, RNA localization and RNA splicing (right panel). For each of these three RNA biology features self-generated RNA-seq datasets of ES cells and embryonic fibroblasts of mouse and rat were used. Abbreviations: mESC, mouse ES cells; MEF, mouse embryonic fibroblasts; rESC, rat ES cells; REF, rat embryonic fibroblasts; 100PE, 100bp paired-end RNA-seq; 50SE, 50bp single-end RNA-seq; mm10, mouse reference genome; rn5, rat reference genome; CPC, Coding Potential Calculator.

4.2.2 Statistics of read numbers, assembly and filtering

The mouse and rat lncRNA annotations generated by Florian Pauler are each derived from two self-generated and nine published RNA-seq datasets. Figure 10A shows the numbers of uniquely aligned, multiply aligned and unaligned reads of all mouse and rat datasets used for transcriptome assembly. All tissues were deeply sequenced, ranging from ~60 million reads for lung to ~160 million reads for heart (in mouse) and from ~60 million reads for skeletal muscle to ~130 million reads for liver (in rat). Throughout all tissues, ~80-90% of all reads were uniquely aligned, ~5-10% were multiply aligned and ~3-6% could not be aligned. The cufflinks transcriptome assembly generated between 100,000 and 250,000 multi-exonic transcripts per tissue (Figure 10B). In heart and skeletal muscle the fewest transcripts could be assembled in both species, while in testis, ES cells and embryonic fibroblasts the most transcripts were assembled. Transcripts overlapping RefSeq mRNAs on the sense strand (by 1bp of cDNA) or on the antisense strand (>20% of cDNA) were systematically filtered out (Figure 10C), as these transcripts have a high likelihood to be coding or contain coding sequence. This step ultimately removed ~90-95% of all assembled transcripts for eight of the eleven tissue. In the three remaining tissues testis, ES cells and embryonic fibroblasts of

mouse and rat ~35-60 million transcripts passed the sense mRNA filter, which is ~5 times higher than for the other tissues. While testis indeed expresses a wealth of putative lncRNAs, the number of transcripts in ES cells and embryonic fibroblasts decreased to similar levels as in the other tissues after applying the antisense mRNA filter. The reason for this phenomenon is that RNA-seq data of ES cells and embryonic fibroblasts is not as strand-specific as the remaining tissues, hence, highly expressed mRNAs are also assembled on the opposite strand and are only removed by the antisense mRNA filter. The filtered putative non-coding transcripts from all tissues were merged to give rise to ~93,000 mouse transcripts and ~119,000 rat transcripts (Figure 10D). A two-step pipeline including RNAcode (Washietl et al., 2011) and Coding Potential Calculator (Kong et al., 2007) stringently filtered out transcripts with estimated coding potential, reducing the number of lncRNAs to ~60,000 in mouse and ~88,000 in rat. A final filter step removed transcripts that were assembled antisense to highly expressed lncRNAs (due to incomplete strandedness, as for mRNAs) and transcripts that were derived from unmapped or random chromosome pieces. In total, this pipeline identified 36,578 mouse lncRNA transcripts in 7,815 loci (on average 4.68 lncRNAs per locus) and 46,041 rat lncRNA transcripts in 9,921 loci (on average 4.64 lncRNAs per locus) (Figure 10D).

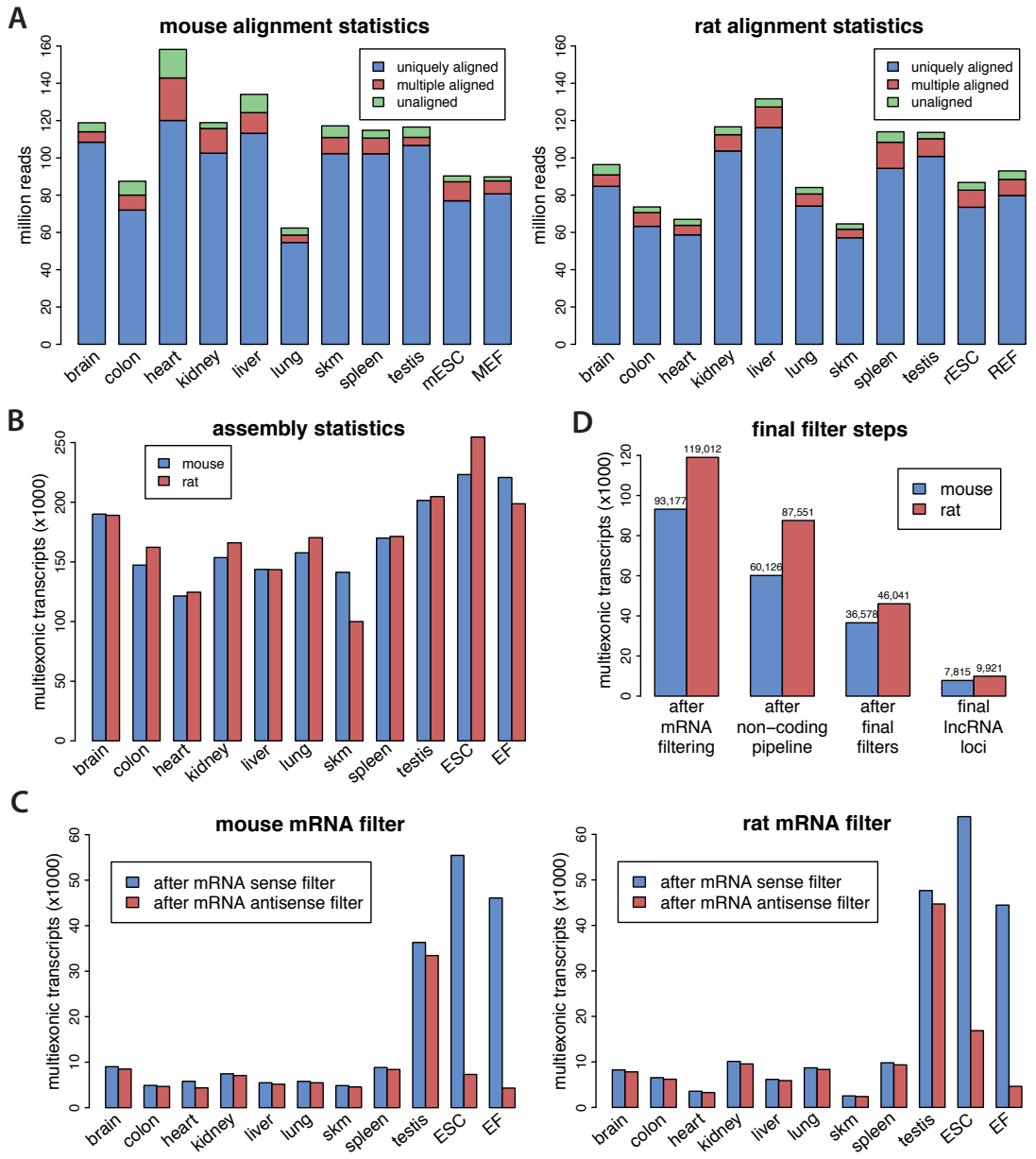


Figure 10: Statistics of read numbers, assembly and filtering

(**A**) Number of total RNA-seq reads per tissue for mouse (left) and rat (right). Total reads were split into uniquely aligned, multiply aligned and unaligned reads after STAR alignment. (**B**) Number of assembled multi-exonic transcripts per tissue for mouse (blue) and rat (red). (**C**) Number of assembled multi-exonic transcripts after removal of transcripts overlapping RefSeq mRNA in sense (by 1bp, blue) and after removal of transcripts overlapping RefSeq mRNA in antisense (by >20% of their cDNA, red) for mouse (left) and rat (right). (**D**) Number of assembled multi-exonic transcripts for mouse (blue) and rat (red) after merging of non-mRNA-overlapping transcripts from eleven tissues, after filtering of transcripts with predicted coding potential by the non-coding pipeline and after applying final filters (see method section for details). The last bars represent the final number of mouse and rat lncRNA loci after grouping of lncRNAs into loci. Abbreviations: skm, skeletal muscle; ESC, ES cells; EF, embryonic fibroblasts.

4.2.3 Intergenic lncRNAs are the most abundant class of lncRNAs

The lncRNA annotations of mouse and rat were complemented by the addition of multi-exonic RefSeq mRNAs, Figure 11A shows the number of lncRNA and mRNA loci in the mouse (left) and the rat genome (right). In rat, more lncRNA loci were annotated, however, with a reduced

number of mRNA loci the total number of gene loci is still lower than in mouse (27,877 mouse loci vs. 25,157 rat loci). The mouse lncRNAs were classified into the four major classes intergenic, antisense, enhancer and bidirectional lncRNAs (see chapter 3.10.4 for details). For rat, it was not possible to classify enhancer RNAs as no public ChIP-seq data was available for the eleven tissues used to map rat lncRNAs. In mouse, intergenic lncRNAs form the most abundant subclass representing almost 58% of all lncRNAs, whereas the remaining classes are equally distributed with each approximately 14% (Figure 11B, left). In rat, due to the lack of eRNAs, approximately 78% of all lncRNA belong to intergenic lncRNAs, while overlapping lncRNAs have a comparable share as in mouse and bidirectional lncRNAs are halved (Figure 11B, right).

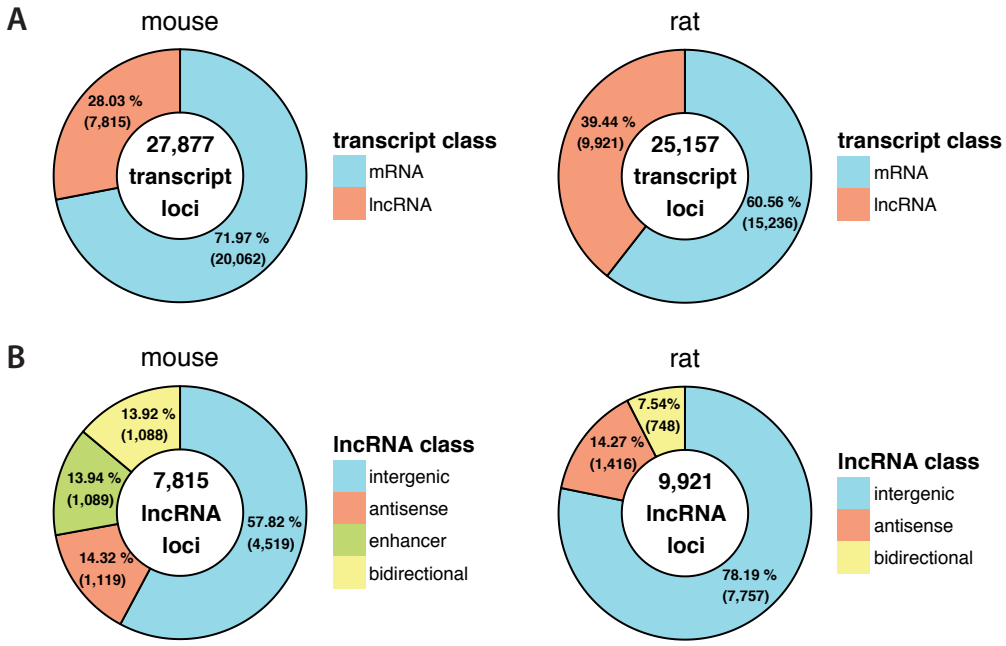


Figure 11: Intergenic lncRNAs are the most abundant class of lncRNAs

(**A**) Ring plots showing the number and percentages of assembled lncRNA loci and added mRNA loci for mouse (left) and rat (right). (**B**) Ring plots showing the number and percentages of lncRNA subclasses for mouse (left) and rat (right). Rat enhancer lncRNAs could not be defined due to a lack of public ChIP-seq data for the tissues that were used to map lncRNAs.

4.2.4 LncRNAs have unusual genomic transcript features

I calculated six genomic transcript features and compared lncRNAs to mRNAs as well as lncRNA subclasses among themselves. In general, these features can be grouped in two broader characteristics: transcript length (defined by the features exon count, cDNA size and locus size) and transcript composition (defined by average exon size, exon/intron ratio and exonic repeat coverage). I find that mouse lncRNAs are very different from mRNAs in terms of their genomic transcript features. They are significantly shorter than mRNAs, as indicated by three of the six features: the median number of exons per lncRNA is substantially lower than per RefSeq mRNA (3 vs 8 exons per transcript), as is the median cDNA size (1.47kb vs

2.58kb) and the median locus size (8.78kb vs 20.57kb) (Figure 12A). Mouse lncRNAs also have an unusual transcript composition as indicated by larger average exon sizes (254bp vs. 144bp), a higher exon/intron ratio (23.09% vs. 12.42%) and a dramatically increased exonic repeat coverage (26.55% vs. 1.67%) compared to mRNAs (Figure 12A). The results are very similar for the rat (Figure 12B). Taken together, lncRNAs are markedly different from mRNAs in terms of genomic transcript features, which may reflect their different evolution, conservation and function.

Within the four lncRNA subclasses, the differences are not as pronounced (Figure 12A, B). The exon count, cDNA size and locus size is very similar throughout all lncRNA subclasses in mouse and rat. In terms of average exon size and exon/intron ratio, bidirectional lncRNAs have slightly longer exons and a higher exon/intron ratio in both species. The exonic repeat coverage is similar in all lncRNA subclasses, only antisense lncRNAs tend to have less repeats in mouse and rat. This is most probably explainable by the fact that they overlap mRNAs and coding regions, which exhibit reduced repeat content. The four mouse lncRNA classes also exhibit similar steady-state RNA levels with a maximum difference of ~2-fold between intergenic and bidirectional lncRNAs (Figure 12C). In summary, this establishes that neither genomic transcript features nor abundance levels distinguish any of the currently used mouse lncRNA classes.

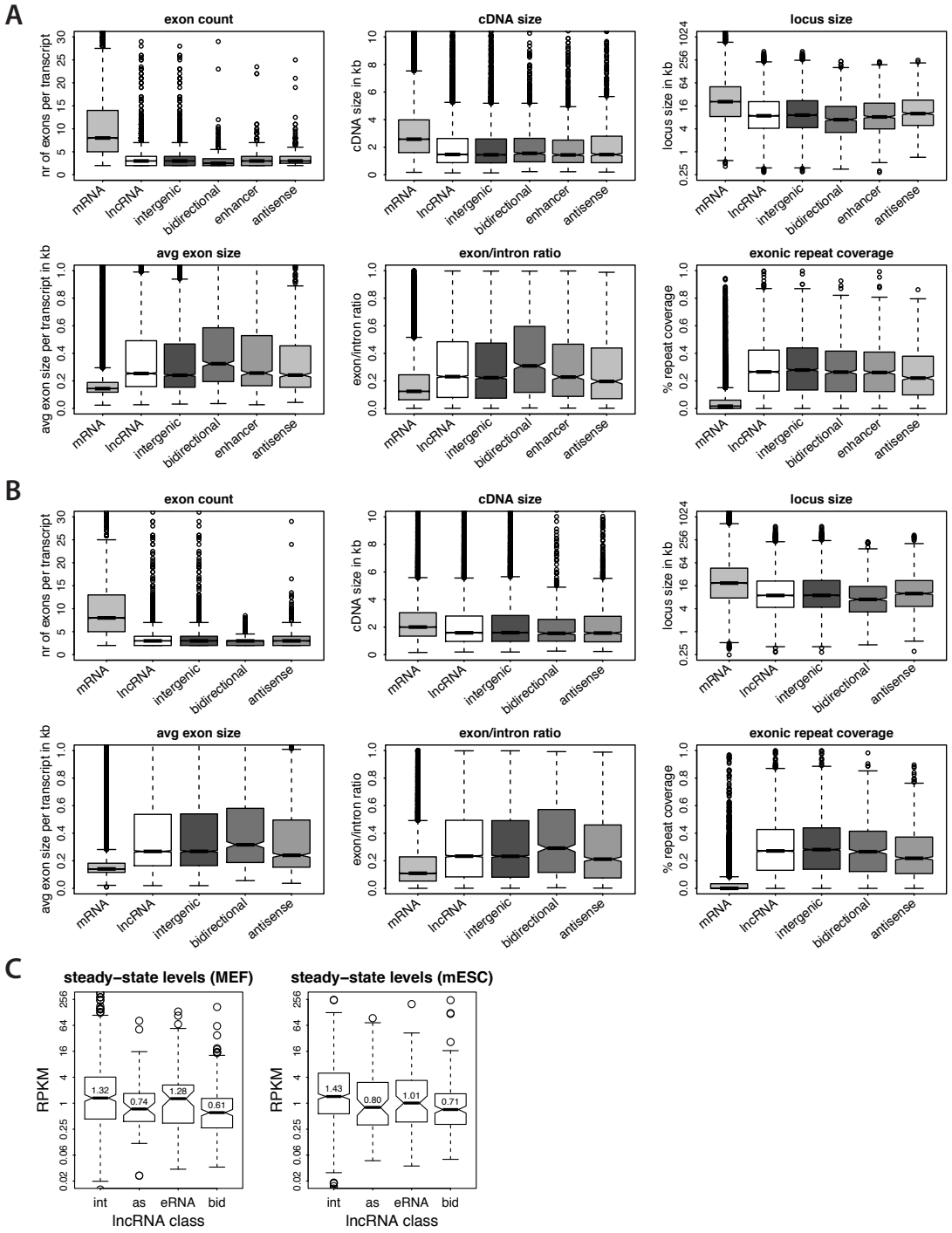


Figure 12: LncRNAs subclasses have very similar genomic transcript features

(**A**), (**B**) Six genomic transcript features were calculated for mouse and rat transcripts and averaged (by median) for each locus. Features of lncRNAs (white boxes) and mRNAs (left grey boxes) were plotted separately for (**A**) mouse and (**B**) rat. Additionally, lncRNAs were further divided into the subclasses intergenic, bidirectional, enhancer and antisense lncRNAs (shaded grey boxes). The exonic repeat coverage was calculated using the UCSC repeatmasker track, all other features were directly calculated from the annotation bed file. Statistical significance for lncRNA vs. mRNA comparisons was calculated using the R function `wilcox.test()` and the p-value was found in all six cases to be < 2.2e-16. Numbers of loci for each box can be found in Figure 11. **C**) RPKMs calculated from polyA+ RNA-seq data are plotted in log2 scale for four mouse lncRNA classes in MEF and mESC. Int, intergenic lncRNA; as, antisense lncRNA; eRNA, enhancer lncRNA; bid, bidirectional lncRNA.

4.2.5 LncRNAs are lowly expressed

While single lncRNAs can be very highly expressed, it is known that lncRNAs are on average considerably lower expressed than mRNAs (Cabili et al., 2011). The density plots in Figure 13 depict the distribution of log10 transformed RPKMs for mouse (Figure 13A) and rat (Figure 13B) lncRNAs (left) and mRNAs (right). As can be seen, the majority of expressed lncRNAs tend to have RPKMs between 0.01 and 1 while most of the expressed mRNAs have an RPKM of ~10. LncRNAs are therefore on average approximately 10-1000 fold lower expressed than mRNAs. In terms of lncRNA expression, testis is an outstanding tissue as it expresses a wealth of lncRNAs with an RPKM of ~1.23 (in mouse) and ~1.52 (in rat) (peaks of black lines) and thereby considerably higher than any other tissue (Cabili et al., 2011; Necsulea et al., 2014).

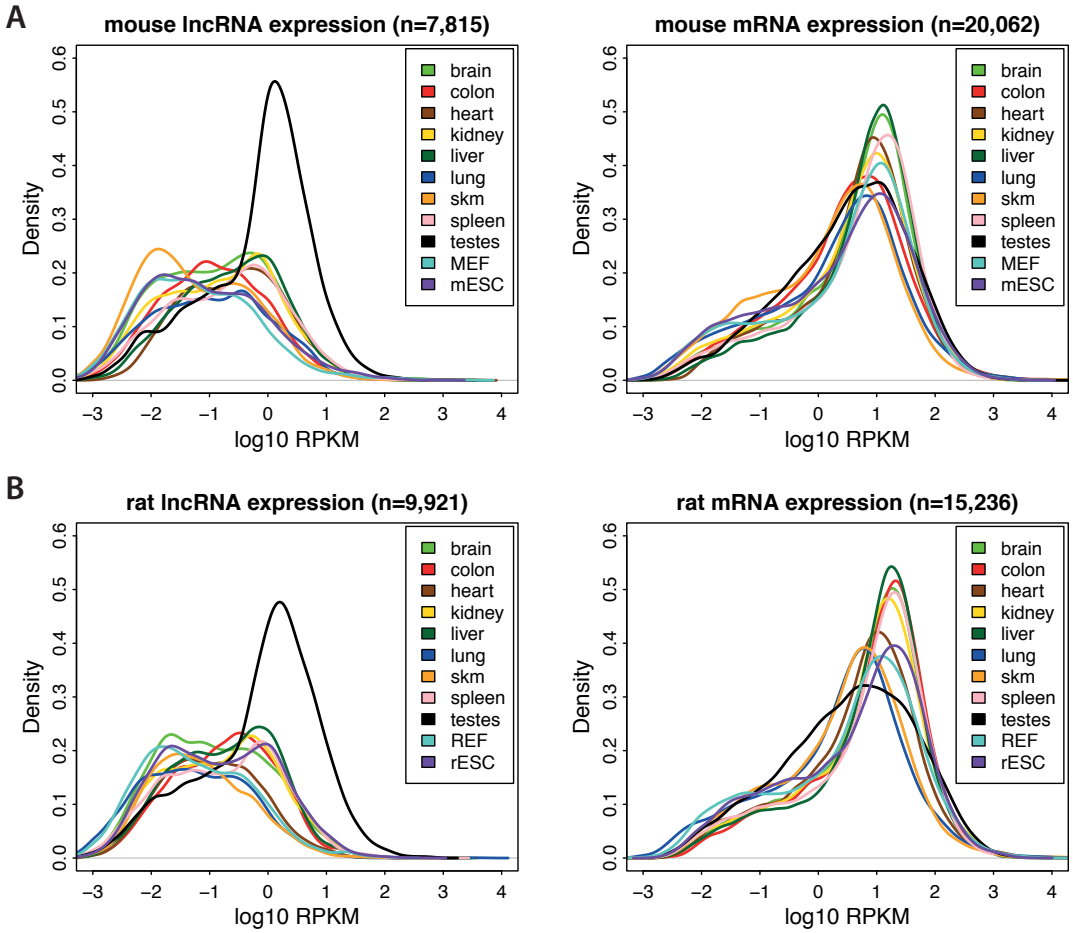


Figure 13: LncRNAs are lowly expressed

RPKMs were calculated for the mouse and rat annotation for eleven tissues. LncRNAs (left) and mRNAs (right) were plotted separately for mouse (**A**) and rat (**B**). RPKMs were log10 transformed and RPKM densities were calculated using the R function `density()`.

4.2.6 LncRNAs are tissue-specifically expressed

Although some structural and essential lncRNAs are known to be ubiquitously expressed, many lncRNAs are actually expressed only in one tissue (Cabili et al., 2011). In order to estimate the degree of tissue-specific expression of lncRNAs in my catalog, I calculated RPKMs from the eleven tissues and asked how much of the RPKM sum of all tissues is coming from each single tissue. Figure 14A depicts these fractional densities of 5,973 expressed lncRNAs and 18,177 expressed mRNAs of the mouse. Fractional densities were sorted in decreasing order and densities >0.5 (representing 50% of the RPKM sum coming from only one tissue) were considered as tissue specific expression. It becomes evident that 74.77% of all lncRNAs are expressed primarily from only one tissue, in contrast to 35.99% of all mRNAs. The remaining transcripts are expressed in more than one tissue with many of them being ubiquitously expressed in all tissues. Also in rat, 78.15% of lncRNAs are expressed primarily from one tissue in contrast to 32.98% of mRNAs (Figure 14B). In an attempt to further quantify that result, I sorted the RPKMs of each locus from the highest expressing tissue to the lowest expressing tissue. Figure 14C (left panel) shows that mouse mRNAs reach median expression levels of ~50% and ~30% in the second and third highly expressing tissue relative to the highest expressing tissue, respectively, while mouse lncRNAs reach only ~10% and ~3%. This trend of tissue-restricted expression of lncRNAs is also evident in rat (Figure 14C, right panel).

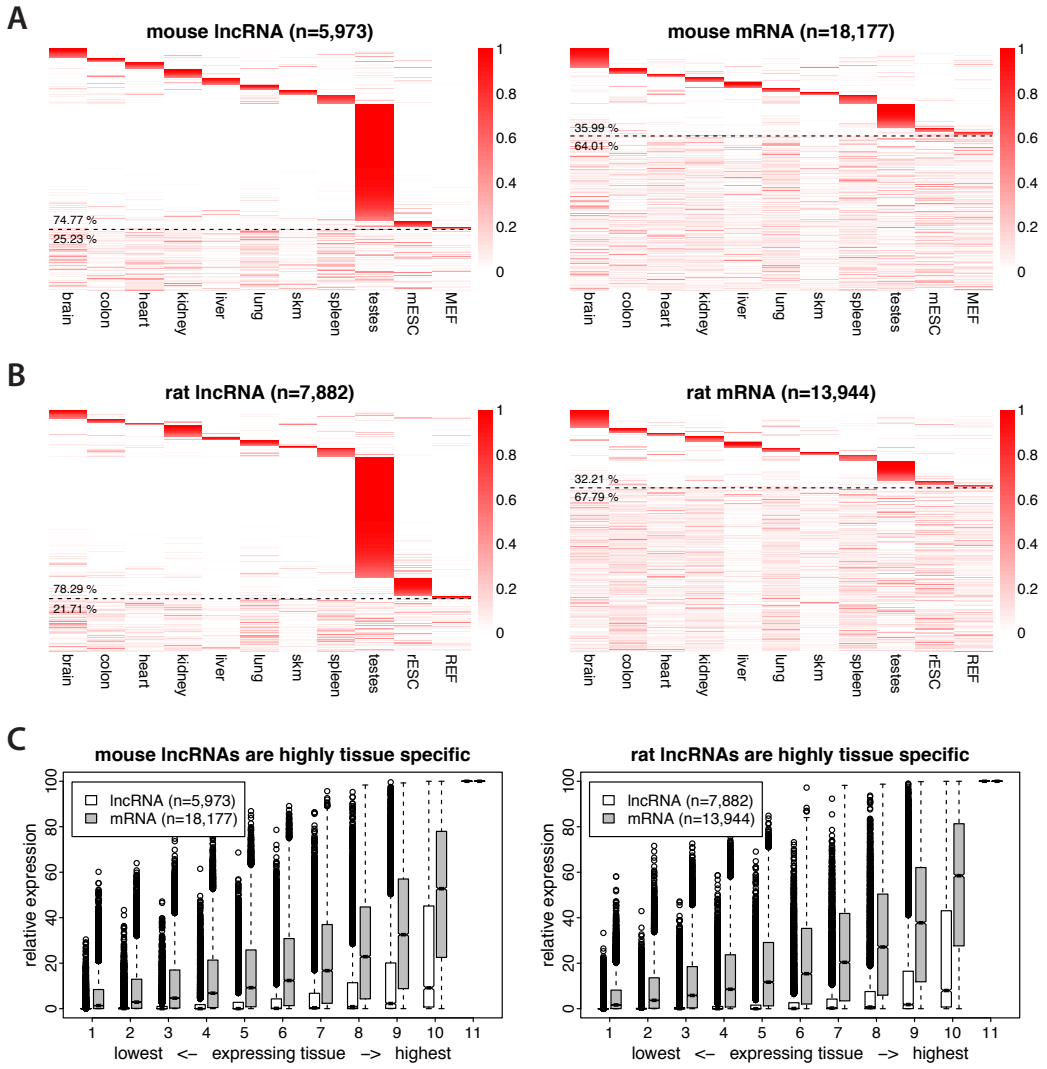


Figure 14: LncRNAs are more tissue-specifically expressed than mRNAs

RPKMs were calculated for the mouse and rat annotation for 11 tissues. An expression cut-off was applied to keep only loci that were stably detected with an RPKM>1 in at least one of the eleven tissues. (**A**,**B**) Fractional densities of mouse and rat lncRNAs (left) and mRNAs (right) were calculated by determining the fraction of the RPKM from each single tissue relative to the RPKM sum of all tissues. Loci with a fractional density of >0.5 (corresponding to >50% of the total expression of all tissues is coming from a single tissue) were considered as tissue-specifically expressed (red blocks). (**C**) RPKMs of the 11 tissues were sorted for each locus from lowest to highest expressing tissue (set to 100). The boxplots depict the RPKMs of lncRNAs and mRNAs in all tissues relative to the highest expressing tissue for mouse (left) and rat (right).

4.2.7 LncRNAs are developmentally regulated in liver and heart

Many lncRNAs have been shown to play important roles in mouse development and to be differentially expressed between different developmental stages (Grote et al., 2013; Herriges et al., 2014; Klattenhoff et al., 2013). In order to quantify the extent of developmental regulation of my lncRNA annotation, I investigated heart and liver development and sequenced one adult sample (6 weeks) and two fetal samples (E14.5) of each tissue (RNA-seq data provided by TM Kulinski and QJ Hudson, manuscript in revision). In Figure 15A I plotted fold-changes of lncRNAs and mRNAs between the adult and the two fetal tissues

(adult vs. fetal #1 and adult vs, fetal #2). As can be seen for heart development, 23.62% of expressed lncRNAs are >2-fold upregulated in the adult heart (11.30% of mRNAs) relative to the fetal heart whereas 41.24% are downregulated (29.5% of mRNAs). In liver this trend is also evident as 25.56% of lncRNAs are upregulated (15.39% of mRNAs) and 58.18% are downregulated (52.2% of mRNAs) during liver development. The reason why lncRNAs are more developmentally regulated than mRNAs in liver and heart development remains enigmatic, however, it fits well to the high degree of tissue-specific expression of lncRNAs (see Figure 14). As lncRNAs are considerably lower expressed than mRNAs (see Figure 13), care has to be taken that low expression does not lead to more variable RPKM and thereby to an overestimation of differential regulation. Therefore, I used two cut-offs to only include reliably detected transcripts. First, I demanded each transcript to be detected in at least one of the three (one adult, two fetal) samples with an RPKM > 1. Second, I calculated RPKM errors and kept only transcripts that have a RPKM error < 5% in at least one of the three samples (see chapter 3.9.7). This cut-off ensures that only those transcripts are further analyzed that are stably detected at the current sequencing depth and even if the sample is deeper sequenced does not change its final RPKM by more than 5% (see method section for details). Figure 15B shows the distribution of RPKM errors in the up- and down regulated lncRNAs and mRNAs in heart and liver. It becomes evident that lncRNAs are in fact less stably detected than mRNAs, however, their RPKM errors are mostly below 10% and therefore can not be the sole reason for increased developmental deregulation defined by a 2-fold change. As expected, genes higher expressed in the adult sample tend to have low RPKM errors in the adult sample but higher errors in the fetal samples (where they are lower expressed) whereas genes higher expressed in the fetal samples genes show the opposite pattern. The relative sequencing errors are similar between heart samples (Figure 15B, left) and liver samples (Figure 15B, right). Additionally, as a control, I compared the two biological replicates of the fetal samples of liver and heart and found that few transcripts are differentially expressed between the two replicas (grey lines, Figure 15A), indicating that the developmental regulation of lncRNAs can not be attributed to only increased variability of lowly expressed lncRNAs.

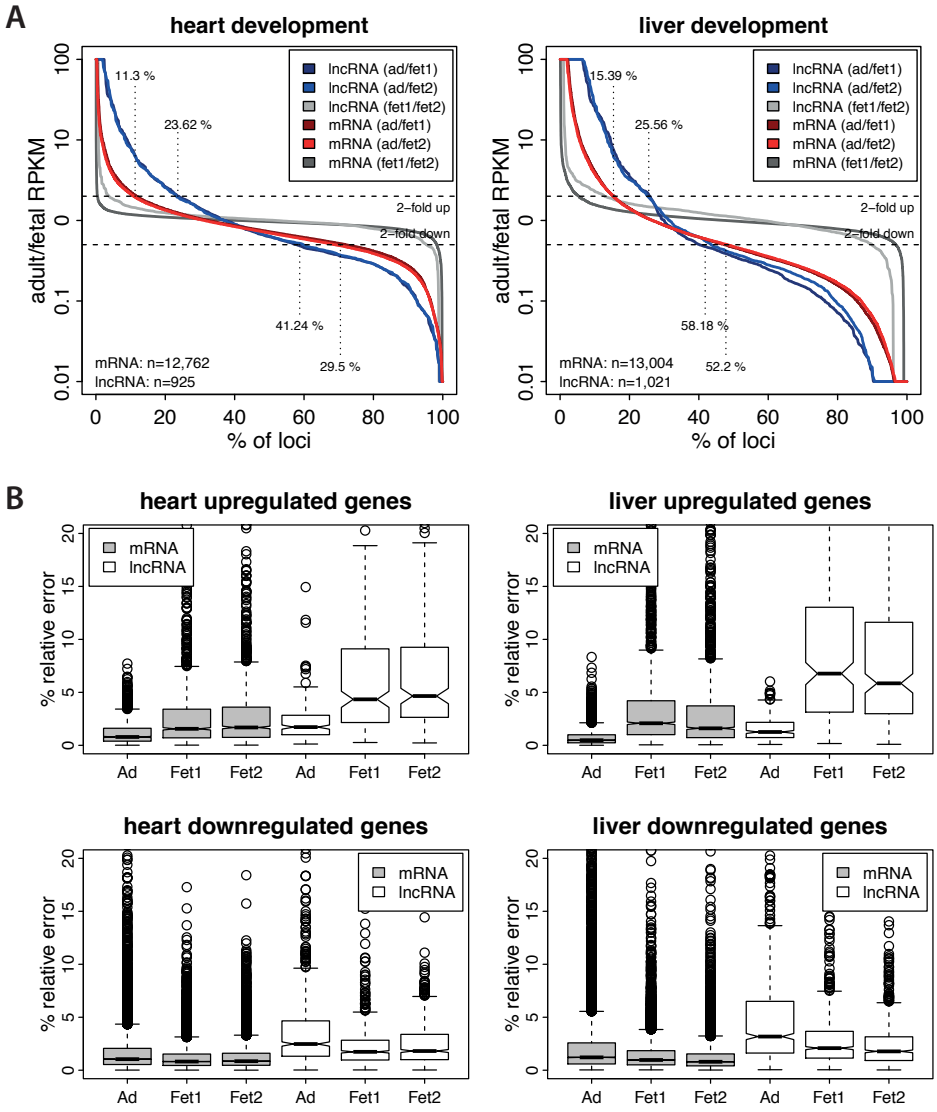


Figure 15: LncRNAs are developmentally regulated in liver and heart

(**A**) RPKMs were calculated for one adult (6 weeks) and two fetal samples (E14.5) of mouse heart (left) and mouse liver (right). Stably detected loci were chosen by RPKM > 1 and RPKM error < 5% cut-offs in at least one of the three samples. Fold-changes of adult over fetal samples (adult vs. fetal #1 and adult vs fetal #2) are plotted for lncRNAs (blue lines) and mRNAs (red lines). The comparisons between the two fetal samples (grey lines) serve as a control to rule out increased lncRNA variability as a major source for differential regulation. Numbers indicate the average percentages of lncRNAs and mRNAs that are up- or downregulated in the adult over the two fetal samples. (**B**) Distributions of RPKM errors are plotted for up- and down-regulated mRNAs and lncRNAs in heart (left figures) and liver (right figures). As an example, transcripts >2-fold upregulated in the adult over the fetal sample are obviously higher expressed in the adult than the fetal sample, hence, their relative sequencing error is low in the adult sample and higher in the fetal sample. Abbreviations: fet, fetal; ad, adult.

4.2.8 LncRNAs are differentially expressed in B, CD4+ and CD8+ T cells

In order to investigate tissue specific expression of lncRNAs also in pure cell types, I FACS-sorted B and CD4+ as well as CD8+ T cells and compared their transcriptomes to their host tissue spleen. Figure 16A depicts heatmaps of fractional densities (similar to Figure 14A) of 1,394 lncRNAs (left) and 12,826 mRNAs (right) and as can be seen a considerable fraction of lncRNAs appears to be B and T cell specific. This is remarkable as B and T cells were not

explicitly included in the range of tissues used for transcript assembly. The annotation pipeline therefore seems to be sensitive enough to annotate transcripts from rare cell types that infiltrated other organs. Mature B cells constitute a large portion of the spleen, whereas T cells are much more rare there. This physiologic situation is reflected in the heatmaps, as B cell specific lncRNAs also show expression in spleen (light red block) whereas T cell specific lncRNAs show much less expression in spleen (white block). Interesting is also the observation that B cell lncRNAs are mostly not found in T cells, and vice versa, whereas lncRNAs from T cells are in most cases found in CD4+ as well as CD8+ T cells. In order to estimate robustness of lncRNA expression I sequenced RNA from three biological replicates of mouse spleen and calculated fold change of lncRNAs and mRNAs between all three possible pairwise combinations (Figure 16B). After averaging the number of differentially expressed (> 2-fold) transcripts between the three pairwise comparisons, only less than 1% of lncRNAs and mRNAs are differentially expressed between three biological replicates of spleen. This shows that most lncRNAs are robustly detected by our sequencing pipeline and if differential expression is detected it can be attributed to the biological context rather than spurious variance of sequencing data. This finding further validates differential expression data from Figure 15A and Figure 16A. I further plotted fractional densities of expressed lncRNAs and mRNAs (as in Figure 15) for all ScriptSeq datasets (Figure 16C). It can be seen that most of the tissue specific lncRNAs are primarily expressed from the adult tissue of heart, liver and spleen from which they were assembled. Interestingly, while many lncRNAs primarily expressed in adult heart are also lowly expressed in fetal heart, and vice versa, adult liver and fetal liver share much less lncRNAs. In summary, this data shows that many lncRNAs are developmentally and tissue specifically expressed, which argues that they might have cell type specific functions such as the regulation of transcription factors or key signaling pathways.

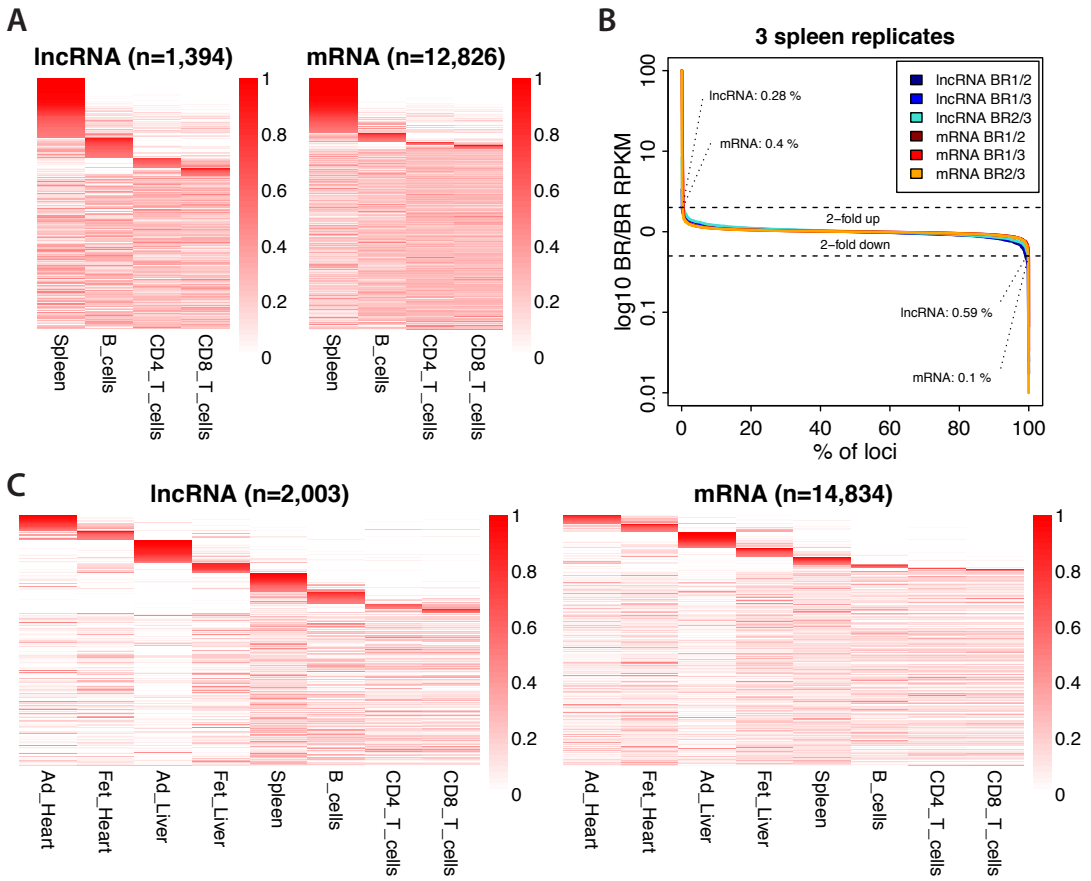


Figure 16: LncRNAs are differentially expressed in B, CD4+ and CD8+ T cells

RPKM of mouse lncRNAs and mRNAs were calculated for the heart and liver adult and fetal samples, for FACS sorted B cells, CD4+ and CD8+ T cells and for adult spleen. These RNA-seq datasets are all derived from ScriptSeq libraries and are therefore only compared among themselves. (**A**) An expression cut-off was applied to keep only loci that were stably detected with an RPKM>1 in at least one of the four indicated tissues. Fractional densities of lncRNAs (left) and mRNAs (right) were calculated by determining the fraction of the RPKM that comes from each single tissue relative to the RPKM sum of all tissues. Loci with a fractional density of >0.5 (corresponding to >50% of the total expression of all tissues is coming from a single tissue) were considered as tissue-specifically expressed (red blocks). (**B**) Stably detected loci were chosen by RPKM > 1 and RPKM error < 5% cut-offs in at least one of the three spleen replicates. Fold-changes of pairwise comparisons are plotted for lncRNAs (blue shades) and mRNAs (red shades). Numbers indicate the average percentages of lncRNAs and mRNAs that are up- or downregulated in the three pairwise comparisons. (**C**) as in (A) but with eight indicated tissues. Abbreviations: BR, biological replicate; fet, fetal; ad, adult.

4.3 Investigation of lncRNA biology (export, stability, splicing)

4.3.1 Cellular fractionation efficiently separates nuclear and cytoplasmic RNA

In order to investigate the cellular localization of lncRNAs, I incubated cells in a mild lysis buffer to disrupt cell membranes while keeping the nucleus intact (Figure 17A). After sucrose gradient centrifugation, I isolated RNA from the cytoplasmic fraction above the sucrose and from the intact heavier nuclei that formed a pellet below the sucrose. A qPCR analysis confirmed efficient separation of the nuclear and cytoplasmic RNA fractions. The left panel in Figure 17B shows the enrichment of the nuclear localized lncRNA *Airn* (Seidl et al., 2006) in

the nuclear fraction and its reduction in the cytoplasmic fraction of MEF cells compared to the *Gapdh* mRNA in two biological replicates. As *Airn* is not expressed in mouse ES cells, I decided to use the nuclear localized lncRNA *Kcnq1ot1* (Redrup et al., 2009) as a marker for ES cells, leading to similar results (Figure 17B, right panel). I prepared RNA-seq libraries from nuclear and cytoplasmic RNA fractions and sequenced them 50bp single-end. RNA-seq data of mESC and MEF loaded onto the UCSC genome browser shows that fourteen clustered histone mRNAs are efficiently exported and therefore strongly enriched in the cytoplasmic fraction (top panel, blue peaks) compared to the nuclear fraction (Figure 17C, top panel) whereas the *Kcnq1ot1* lncRNA is retained in the nucleus (bottom panel, green peaks) and almost absent in the cytoplasmic fraction (Figure 17C, bottom panel). This further confirms the efficient separation of cytoplasmic and nuclear RNA fractions in both cell types.

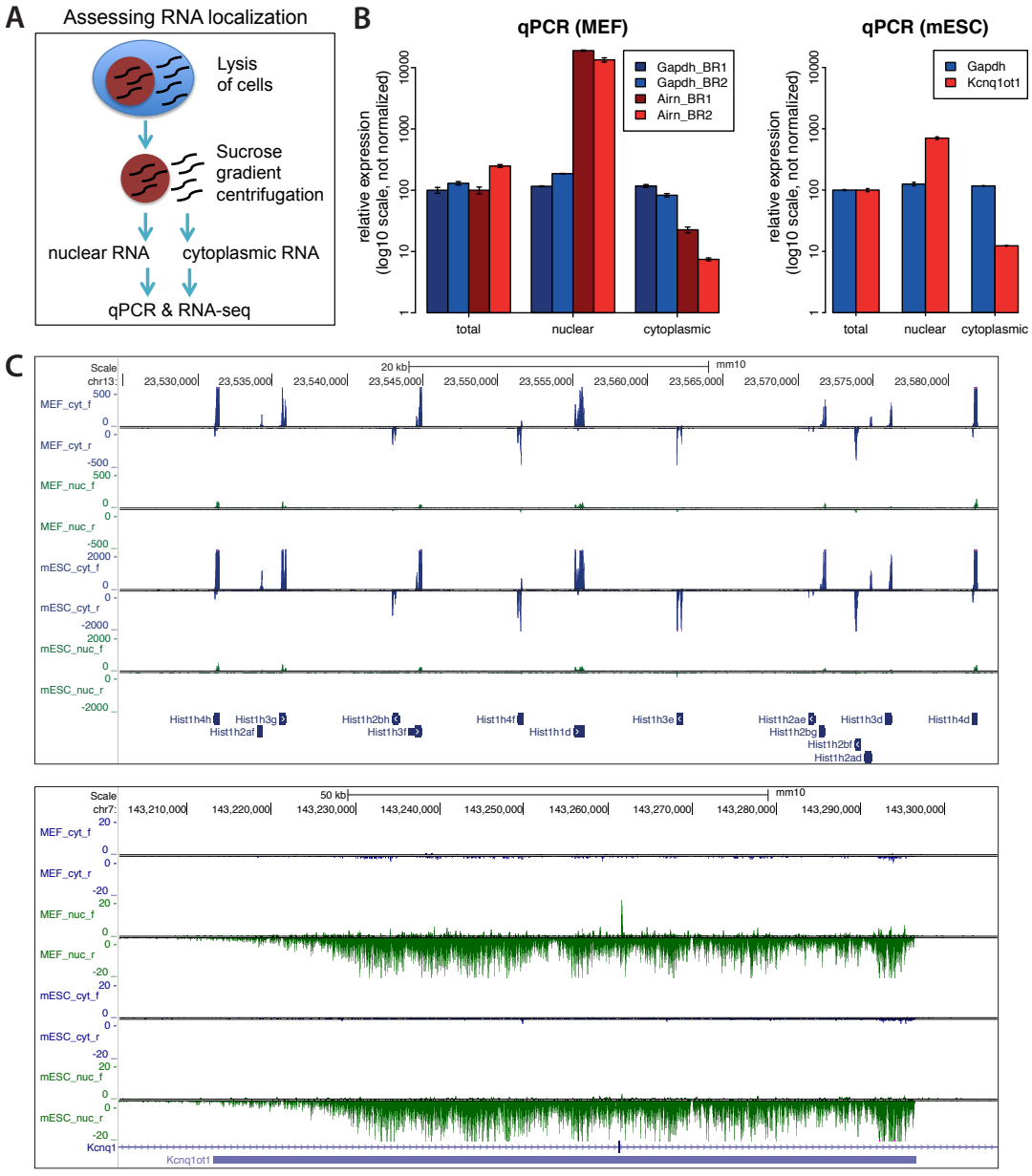


Figure 17: Cellular fractionation efficiently separates nuclear and cytoplasmic RNA

(**A**) Experimental outline to investigate RNA localization by subcellular RNA fractionation and RNA-seq. (**B**) Levels of *Air* and *Gapdh* (left) and *Kcnq1ot1* and *Gapdh* (right) were measured by qPCR analysis in the nuclear and cytoplasmic fractions of MEF (left) and mESC (right). Expression values were not normalized to a housekeeping gene. Expression levels in total RNA were set to 100 (in MEF replicates only biological replicate 1) and subcellular fractions relative to it. Data is log10 transformed and depicted as mean, error bars represent variation of three technical replicates. (**C**) RNA-seq data of cytoplasmic (blue tracks) and nuclear (green tracks) RNA fractions of MEF and mESC are displayed in the UCSC genome browser. Chromosome and genomic coordinates are displayed on top. Two tracks are displayed for every sample, the upper one corresponding to the forward strand and the lower one corresponding to the reverse strand. The RefSeq annotation showing a cluster of fourteen histone mRNAs (top) and the locus of the lncRNA *Kcnq1ot1* (bottom) is displayed beneath the tracks. The histone mRNAs are predominantly found in the cytoplasmic fractions (top, blue peaks) and nuclear retained *Kcnq1ot1* lncRNA predominantly found in the nuclear fractions (bottom, green peaks) in MEF and mESC. Abbreviations: BR, biological replicate; nuc, nuclear RNA fraction; cyt, cytoplasmic RNA fraction.

4.3.2 LncRNAs are less exported than mRNAs

I calculated the RNA export of each lncRNA locus and RefSeq mRNA from RNA-seq data by summing up each transcript's nuclear and cytoplasmic RPKM and determine which percentage of the RPKM sum comes from the cytoplasmic fraction (see methods section 3.12 for details). RNA export of a transcript is therefore defined by the percentage it is exported to the cytoplasm and ranges between 0% (exclusively detected in the nucleus), 50% (detected in the nuclear and the cytoplasmic fraction with the same RPKM) and 100% (exclusively detected in the cytoplasm). As mRNAs are transcribed in the nucleus and translocate to the cytoplasm to be translated, they are found in both fractions. The subcellular localization of lncRNAs is not as defined as their functions do not depend on translation in the cytoplasm, hence, many of them evade nuclear export and are therefore found predominantly in the nucleus. Figure 18A summarizes the RNA export data for mRNAs and lncRNAs in mESC and MEF (left) as well as REF (right). The median export of mRNAs seems to be very constant in all three cell types (52.26%, 52.60% and 53.07% for mESC, MEF and REF, respectively), arguing that the RNA fractionation experiment was equally efficient in all three experiments. The median export of lncRNAs is, however, statistically significantly lower in all three cell types compared to mRNAs as indicated by a Wilcoxon Rank-Sum test (p-values for all three comparisons < 2.2e-16). Also, RNA export of lncRNAs is more variable between the three cell types. MEF lncRNAs seem to be less exported than mESC lncRNAs (median of 16.68% compared to 28.41%), while REF lncRNAs are more efficiently exported (21.55%) compared to MEF lncRNAs. As lncRNAs are generally lower expressed than mRNAs, I wanted to make sure that RNA export of lncRNAs is as accurately determined as RNA export of mRNAs. I therefore applied not only an RPKM cutoff but also filtered out transcripts when their RPKM error was higher than 5% (see methods section 3.9.7 and 3.12 for details). In Figure 18B the distribution of RPKM errors for mouse (left) and rat (right) are plotted. As mRNAs are usually well expressed their median RPKM error is in all three celltypes very low (between 0.43% and 0.49%). In contrast, most lncRNAs are lowly expressed and so the medians of their RPKM errors are higher (between 1.13% and 1.20%), although still very much acceptable. These errors indicate that the RNA-seq libraries were sequenced deep enough to have robust

RPKM values even of lowly expressed lncRNAs. I also investigated how robust the RNA export calculation is between biological replicates, how similar RNA export is between the two mouse cell types mESC and MEF and whether RNA export is conserved between mouse and rat embryonic fibroblasts (Figure 18C). The RNA fractionation experiments seem to be quite reproducible as the correlations of mRNA and lncRNA export are good (r=0.77 for both) between two experiments that were carried out with cells of a different passage number and a week apart (Figure 18C, left panel). RNA export between two different cell types is more variable for mRNAs (r=0.61) and lncRNAs (r=0.51) (Figure 18C, middle panel). Interestingly, also mRNA export seems to be markedly different between mESC and MEF, although their median export is very similar (see Figure 18A). I also checked whether export of mRNAs and lncRNAs is conserved between mouse and rat. I therefore investigated mRNAs and lncRNAs whose genomic loci overlap >30% between mouse and the synthenic rat region. Indeed, the conservation of mRNA export between mouse and rat embryonic fibroblasts is quite good (r=0.72) and comparable to the variation seen for biological replicates, however, lncRNA export seems to be more variable between the two species (r=0.52) (Figure 18C, right panel).

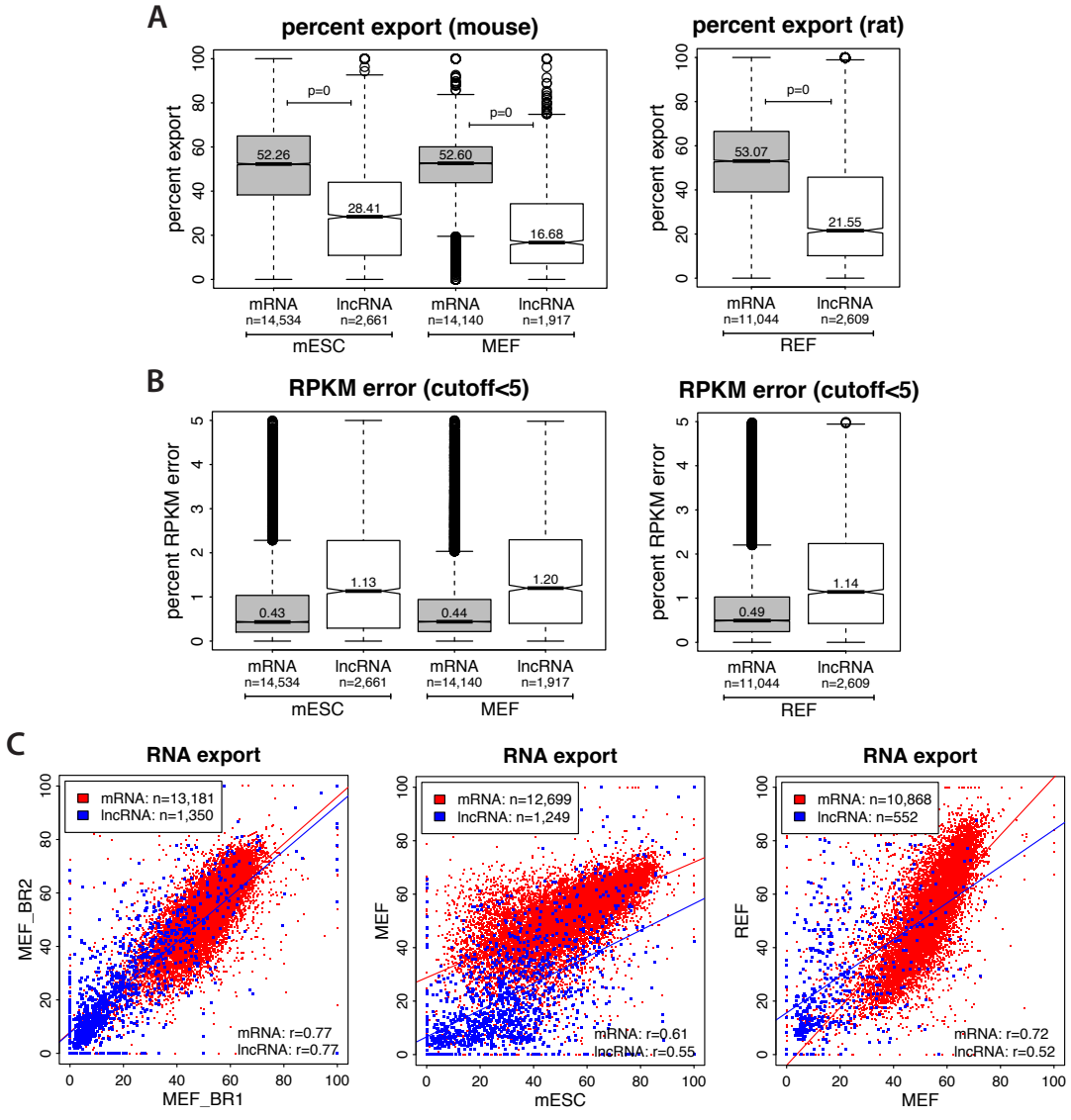


Figure 18: LncRNAs are less exported than mRNAs

(**A**) Boxplots are showing the distribution of RNA export of mRNAs (grey) and lncRNAs (white) for ES cells and embryonic fibroblasts of the mouse (left) and rat embryonic fibroblasts (right) as calculated in chapter 3.12. Numbers in the boxes indicate medians. (**B**) Boxplots are showing the distribution of RPKM errors of mRNAs (grey) and lncRNAs (white) for ES cells and embryonic fibroblasts of the mouse (left) and rat embryonic fibroblasts (right) as calculated in chapter 3.12.3. Numbers in the boxes indicate medians. (**C**) Scatter plots are showing the correlation between two MEF biological replicates (left panel), the correlation between mESC and MEF (middle panel) and the correlation between mouse and rat embryonic fibroblasts (right panel). Pearson's correlation coefficients were calculated using the R function `cor()` and are displayed in the bottom right corners. Numbers of mRNAs and lncRNAs are displayed in the top left corners. Trend lines were calculated using the R function `lm()`.

4.3.3 Actinomycin D treatment efficiently inhibits RNA synthesis

In order to investigate the stability of lncRNAs, I treated mouse and rat ES cells and embryonic fibroblasts with the transcriptional inhibitor Actinomycin D (ActD) or the vehicle control EtOH for 1h and 4h (Figure 19A). Each of the four cell types was assayed in two biological replicates, and each of those in two technical replicates. To confirm the effectiveness of the ActD treatment, I determined levels of the unstable mRNA *Myc* (Dani et al., 1984) in each of these samples prior to RNA-seq. *Myc* levels (Figure 19B, left panel) are decreased to ~50% in the mouse and rat ES cells and to ~20-30% in the embryonic fibroblasts after 1h ActD treatment and to ~5% in all samples after 4h of treatment. As a control, the levels of the stable *Gapdh* mRNA remained largely unaffected by ActD (Figure 19B, right panel). This indicated that the experiment worked in all biological and technical replicates, hence, RNA from corresponding technical replicates was pooled before RNA-seq library preparation. After RNA-seq, the RNA abundances of RefSeq mRNAs were assayed first in order to appreciate the genome-wide effects of ActD treatment. As can be seen in Figure 19C, in all four cell types 1h ActD treatment had only minor effects on mRNAs as only few mRNAs show reduced abundance, however, after 4h a considerable number of mRNAs is substantially reduced. The mRNA abundances in the EtOH control treatments seem to be largely unaffected and resemble the untreated control. RNA-seq data of mESC and MEF loaded onto the UCSC genome browser further confirms the efficiency of ActD treatment. Fourteen clustered histone mRNAs can all be considered unstable as after 1h ActD treatment their abundance is reduced and after 4h ActD treatment the majority of histone mRNAs is degraded, while in the 1h and 4h EtOH control treatments their abundance is unaffacted (Figure 20A+B, left). In contrast, the *Hprt* mRNA can be considered a stable mRNA as its abundance is constant in all ActD and EtOH treatments (Figure 20A+B, right).

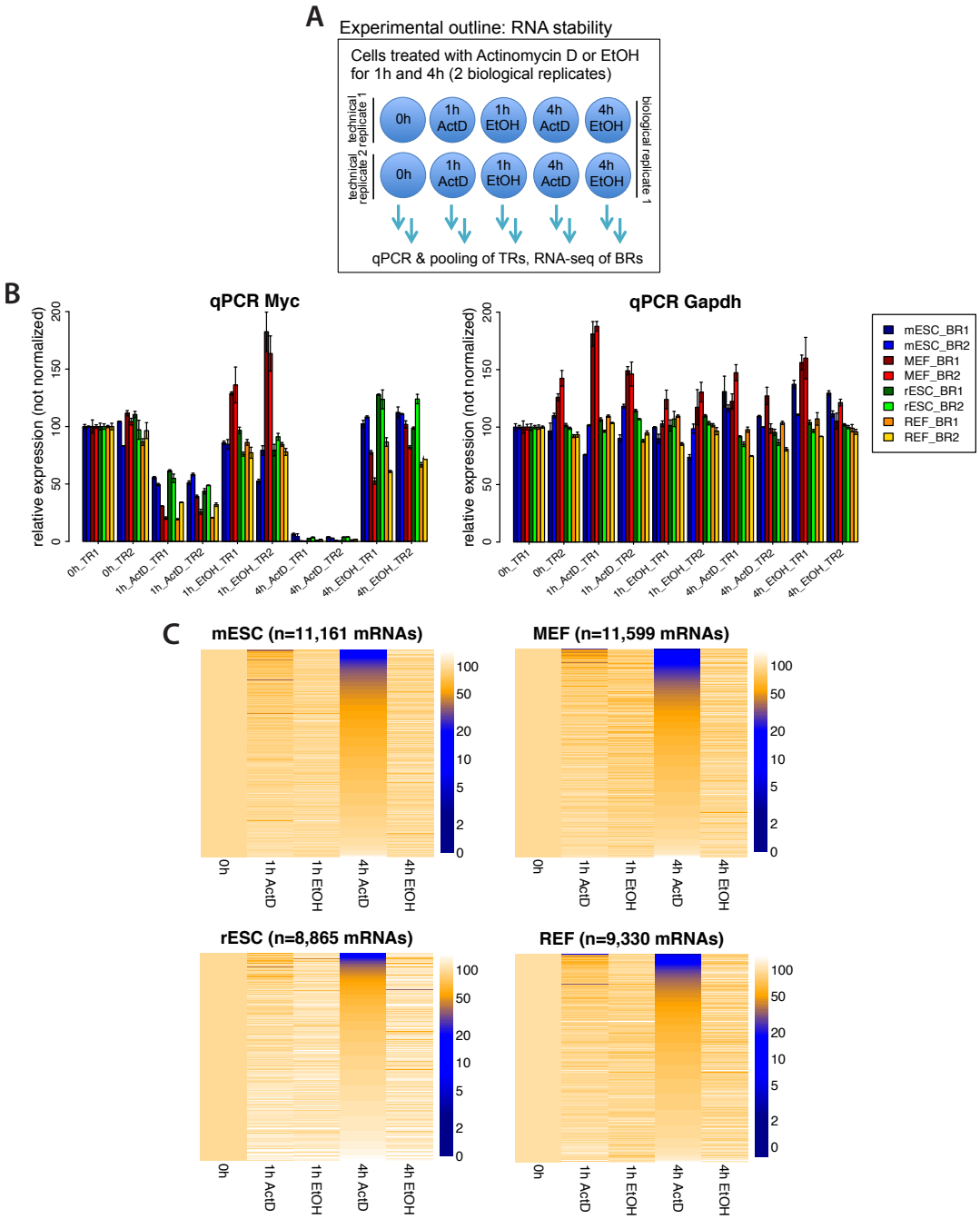


Figure 19: Actinomycin D treatment efficiently inhibits RNA synthesis

(**A**) Experimental outline to investigate RNA stability in mouse and rat ES cells and embryonic fibroblasts. Each cell type was assayed in biological replicates (BR), each of which was assayed in technical replicates (TR). Technical replicates were assayed by qPCR and pooled before RNA-seq. (**B**) qPCR analysis of *Myc* (left) and *Gapdh* (right) in biological replicates of four celltypes after ActD or EtOH treatment for 1h or 4h relative to the untreated control. Data is depicted as mean, error bars represent variation of three technical replicates. (**C**) Heatmaps depict the log10 transformed RPKM of RefSeq mRNAs in four celltypes after ActD or EtOH treatment for 1h or 4h relative to the untreated control (set to 100). In each panel column #4 (4h ActD treatment) was sorted in increasing order to illustrate the distribution of unstable and stable mRNAs. Abbreviations: BR, biological replicate; TR, technical replicate.

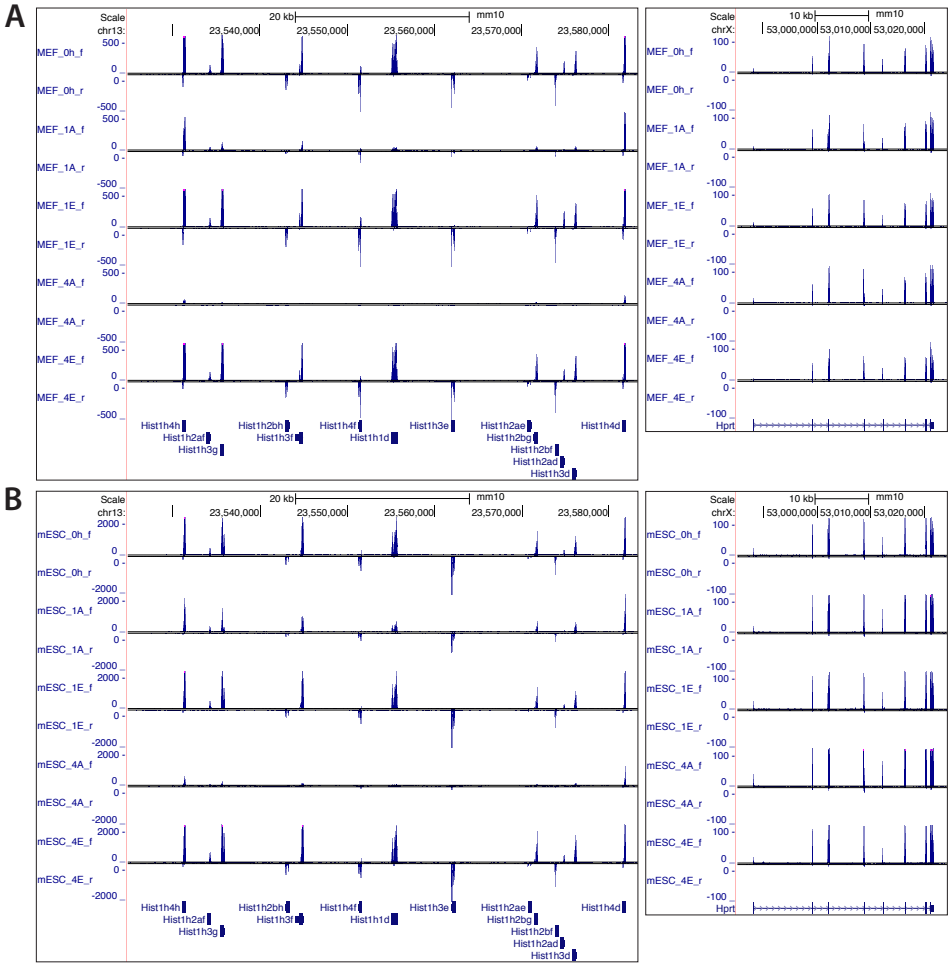


Figure 20: UCSC snapshots of RNA stability RNA-seq data

RNA-seq data of mESC (top panels) and MEF cells (bottom panels) treated with ActD or EtOH for 1h or 4h are displayed in the UCSC genome browser. Chromosome and genomic coordinates are displayed on top. Two tracks are displayed for every sample, the upper one corresponding to the forward strand (indicated by f) the lower one corresponding to the reverse strand (indicated by r). In the left panels, the genomic region of a cluster of fourteen histone mRNAs being expressed from the forward or reverse strand is depicted. Histone mRNA levels are reduced after 1h ActD treatment and nearly disappeared after 4h ActD treatment. These histone mRNAs can therefore be considered as unstable mRNAs. In both EtOH control treatments the RNA levels seem unaffected and very similar to the untreated control. In the right panels, the genomic region of the *Hprt* gene being expressed from the forward strand is shown. *Hprt* mRNA shows comparable abundances in all five samples in mESC and MEF, indicating that *Hprt* mRNA is a stable mRNA at least for 4h. 0h, untreated control; 1A, 1h ActD treatment; 1E, 1h EtOH treatment; 4A, 4h ActD treatment; 4E, 4h EtOH treatment.

4.3.4 LncRNAs are less stable than mRNAs

Reads of corresponding biological replicates were pooled before alignment to increase read numbers and facilitate accurate RPKMs of lowly expressed lncRNAs. RPKMs were normalized to a basket of ten housekeeping genes (see methods section 3.11.2). I defined RNA stability as percent of RNA that is left after 4h of ActD treatment relative to untreated control and relative to 4h EtOH control. RNA stability ranges from 0% (extremely unstable, no RNA left after 4h ActD treatment) to 100% (stable RNA, same levels as in untreated control) or even higher (e.g. 120%, when RNA is increased upon ActD treatment). Figure 21A

summarizes the RNA stability data of mRNAs and lncRNAs in mouse (left) and rat (right) ES cells and embryonic fibroblasts. As can be seen, lncRNAs are on average significantly less stable than mRNAs in all four celltypes, as indicated by Wilcoxon Rank-Sum tests. While mRNAs have a median stability of ~69-73% in mouse and ~78-81% in rat, lncRNAs have a median stability of 44.41% and 58.57% in mESC and MEF and 52.98% and 68.19% in rESC and REF. It seems that mRNAs have a similar stability in both species, whereas lncRNAs in mESC are less stable than MEF lncRNAs. This trend is also observed in rat, as rESC lncRNAs are less stable than REF lncRNAs. In order to accurately quantify RNA stability, I applied not only an RPKM cutoff but also selected transcripts based on the RPKM error (see methods section 3.9.7 for details). I discarded all loci with an error >10% and plotted the distribution of RPKM errors in Figure 21B for mouse (left) and rat (right). Overall, as mRNAs are higher expressed their median RPKM error is in all four celltypes between 1.70% and 2.01%. In contrast, lncRNAs are usually lowly expressed and so the medians of their RPKM errors are between 3.40% and 4.29%. These errors indicate that the RNA-seq libraries were sequenced deep enough to have robust RPKM values even of lowly expressed lncRNAs.

I also investigated the correlation of RNA stability between two different cell types of the same species (Figure 21C) and between the same cell types in two different species (Figure 21D). Overall, the correlations of mRNAs are good while the correlations of lncRNAs are considerably weaker. Specifically, the Pearson correlation coefficients of mRNAs between mESC and MEF (r=0.76) (Figure 21C, left) and between rESC and REF (r=0.67) (Figure 21C, right) show that RNA stability of mRNAs is well correlated, albeit considerably scattered. LncRNAs are, however, less well correlated between mESC and MEF (r=0.60) and especially between rESC and REF (r=0.29). I also checked whether stability of mRNAs and lncRNAs is conserved between mouse and rat. I therefore investigated mRNAs and lncRNAs whose genomic loci overlap >30% between mouse and the synthenic rat region. The RNA stability of mRNAs seems to be quite conserved in both cell types (r=0.75 for MEF vs. REF, r=0.74 for mESC vs. rESC). However, lncRNAs are again more diverse between the two species (r=0.40 for MEF vs. REF, r=0.31 for mESC vs. rESC). The increased diversity of RNA stability of lncRNAs is probably due to two facts: first, RNA stability of lncRNAs is generally more variable between the two mouse and the two rat cell types (as already seen in Figure 21A), and second, lncRNAs are lower expressed and as a consequence RPKM errors are higher than for mRNAs. Unfortunately, I could not lower the RPKM error cut-off to 5% (as I did for RNA export) because RNA stability libraries were not as deeply sequenced and a lower cut-off would have removed most of the lncRNAs. Additionally, increased diversity of lncRNA stability could point towards distinct functional roles of lncRNAs in specific cells or species.

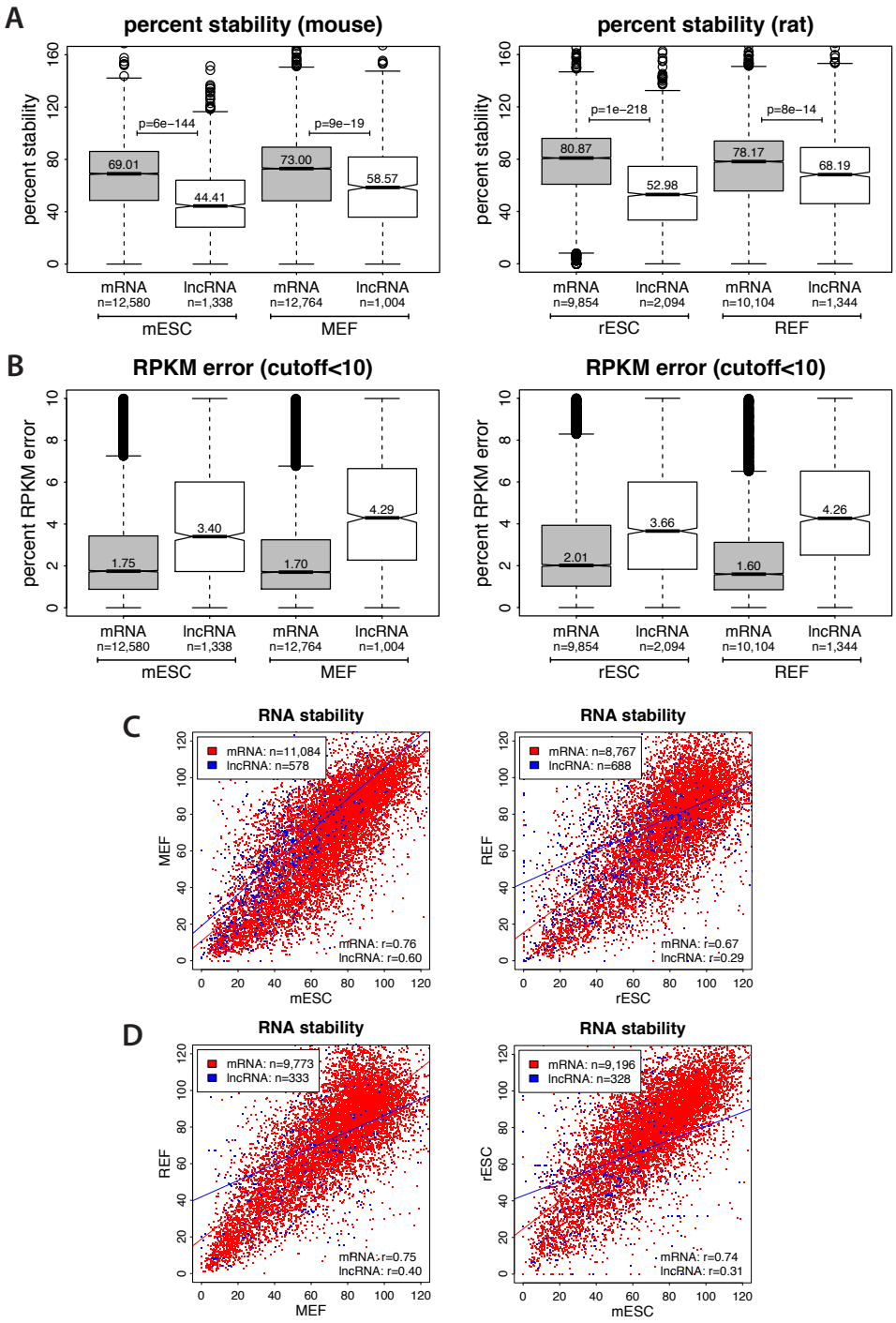


Figure 21: LncRNAs are less stable than mRNAs

(**A**) Boxplots are showing the distribution of RNA stability as calculated in chapter 3.11 for mRNAs (grey) and lncRNAs (white) for ES cells and embryonic fibroblasts of the mouse (left panel) and rat (right panel). Only values between 0 and 160 are plotted, some outliers may not be displayed. (**B**) Boxplots are showing the distribution of relative RPKM errors as calculated in chapter 3.11.4 for mRNAs (grey) and lncRNAs (white) for ES cells and embryonic fibroblasts of the mouse and rat. (**C,D**) Scatter plots are showing the correlations of RNA stability between two different cell types of the same species (mESC vs. MEF, left, and rESC vs. REF, right) and between the same cell types in two different species (MEF vs. REF, left, and mESC vs. rESC, right). Pearson's correlation coefficients were calculated using the R function `cor()` and are displayed in the bottom right corners. Numbers of mRNAs and lncRNAs are displayed in the top left corners. Trend lines were calculated using the R function `lm()`. Only loci with values between 0 and 120 are plotted.

4.3.5 Bioinformatic pipeline to investigate RNA splicing

Splicing efficiency of lncRNAs and mRNAs was calculated from Ribo-Zero RNA-seq data from mESC, MEF, rESC and REF. Figure 22A shows a schematic of the strategy that I followed to estimate RNA splicing efficiency, hoewever, this is only one out of four strategies that I tried (see discussion in chapter 5.4.3). For this approach, I calculated for each junction an RPKM for a 45bp exonic region and an RPKM for a 45bp intronic region, both 5bp away from the junction. From the intron RPKM and the exon RPKM I calculated the splicing efficiency for each splice junction (see chapter 3.13 for details). This approach had three main advantages: First, calculating the exonic and intronic RPKMs near the junction abolishes length biases and reduces the probability that intronic repeats or intronic transcripts skew RPKM calculation. Second, the RPKM of a 45bp region is more robust than calculating read pileups in a 10bp region or counting reads in a 5bp region. Third, the exonic and intronic regions used for RPKM calculation did not start right at the junction, but 5bp away from it, thereby allowing some ambiguity for inexact splice junction annotation. The fact that splice junctions are not perfectly mapped in lncRNAs as well as RefSeq mRNA can be seen in the coverage plots in Figure 22B (for mouse) and Figure 22C (for rat). Coverage plots for lncRNAs (top) and mRNAs (bottom) were split to display left splice junctions (exon left, intron right) and right splice junctions (intron left, exon right) separately. A sharp coverage change can be seen between exons and intron in all eight plots, however, this change seems to occur in two steps: directly at the splice junction the first step and a few base pairs away a second step, indicating that either splice junctions are not accurately annotated or that spliced reads are not perfectly mapped. While the former might be true for self-annotated lncRNAs, it certainly does not hold true for RefSeq mRNAs whose splice junctions have been accurately mapped by a variety of techniques (Pruitt et al., 2014). More likely is the explanation that reads that contain a splice junction within the last five basepairs are mapped as unspliced reads over the splice junction because the aligner allows up to five mismatches and thereby just ignores the splice junction. Another interesting conclusion can already be drawn from these coverage plots: the difference between the exonic and the intronic coverage is much higher in mRNAs than in lncRNAs, arguing that lncRNAs are overall less spliced than mRNAs in mouse and rat.

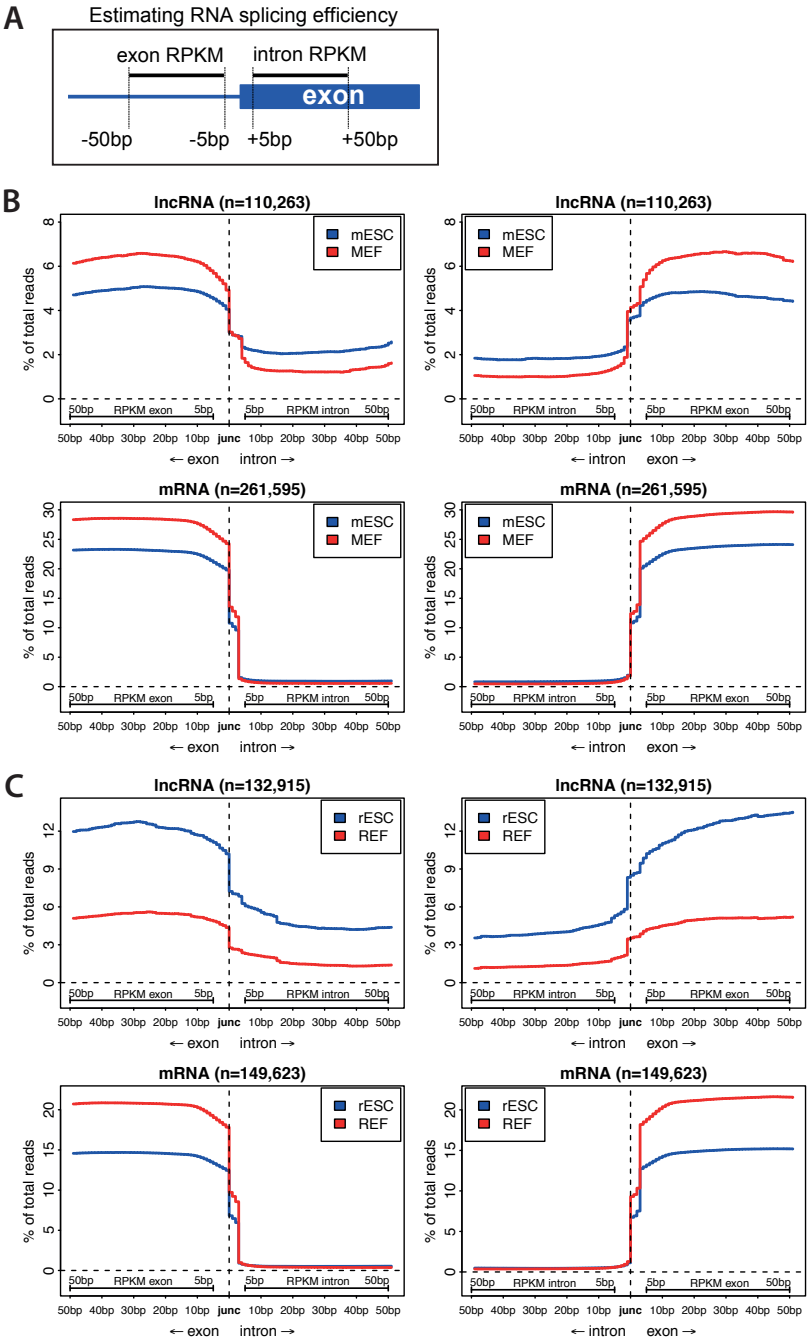


Figure 22: Bioinformatic pipeline to investigate RNA splicing

(**A**) Overview of the approach that I used to calculate RNA splicing. (**B,C**) Coverage plots of mouse (B) and rat (C) lncRNAs (top) and RefSeq mRNAs (bottom) displaying a 100bp region around left splice junctions (exon left, intron right) or right splice junctions (exon right, intron left). Blue (ES cells) and red (embryonic fibroblasts) lines indicate percent of total reads mapping to the regions of splice junctions +/- 50bp. At the bottom of each graph, two lines indicate the region that was used to calculate the exon and intron RPKM. These regions were 45bp long and started 5bp away from the junction. Abbreviations: junc, splice junction.

4.3.6 LncRNAs are less spliced than mRNAs

In order to investigate RNA splicing I generated 100bp paired-end Ribo-Zero RNA-seq data from mESC, MEF, rESC and REF and added additional 100bp paired-end Ribo-Zero RNA-

seq data from mESC and MEF generated for another project in the Barlow Lab (Daniel Andergassen & Quanah J. Hudson, manuscript in preparation). The total numbers of reads are 259 million for mESC, 331 million for MEF, 103 million for rESC and 89 million for REF (see Table 16 for details about read numbers). I calculated the splicing efficiency for each splice junction using the approach indicated in Figure 23A. Splicing efficiencies of junctions were averaged to give a splicing efficiency per transcript. Lastly, the splicing efficiencies of all transcripts in a locus were averaged to ultimately yield a splicing value for each lncRNA and mRNA locus. I find that mRNAs are as expected efficiently spliced while lncRNAs show signs of inefficient splicing in both cell types of the mouse and rat (Figure 23A). The median splicing efficiency of mRNAs in all four cell types resembles each other and is between ~97% and ~99%. The median splicing efficiency of lncRNAs is ~74% in mouse ES cells and ~85% in mouse embryonic fibroblasts (left panel). In rat (right panel), the median splicing efficiencies are ~87% in ES cells and ~90% in embryonic fibroblasts. In order to accurately estimate RNA splicing of lowly expressed lncRNAs, I applied not only an RPKM cutoff but also filtered out splice junctions when the RPKM error of their splicing efficiency was higher than 10% (see methods section 3.9.7 and 3.12 for details). This cut-off removed junctions with inaccurately calculated splicing efficiencies and the remaining ones were averaged per transcript and per locus to yield the final RPKM errors. In Figure 23B I plotted the distribution of RPKM errors for mouse (left) and rat (right). As can be seen, the errors of mRNAs and lncRNAs in the two mouse cell types are in the same range as in the two rat cell types. This is surprising as the number of RNA-seq reads for the mouse cells were ~2-3 times higher than for the rat cells. RPKM errors for well-expressed mRNAs are lower than for lowly expressed lncRNAs in all four cell types, however, the lncRNA medians do not exceed 2% and the mRNA medians are well below 0.5% indicating that RNA splicing data is accurate for mouse and rat.

I also investigated the correlation of RNA splicing between two different cell types of the same species (Figure 23C) and between the same cell types in two different species (Figure 23D). The Pearson correlation coefficients of mRNA splicing between mESC and MEF (r=0.80) (Figure 23C, left) and between rESC and REF (r=0.77) (Figure 23C, right) indicate that mRNA splicing is well correlated in the two mouse cell types and in the two rat cell types. Correlation of lncRNA splicing is lower but still quite good (r=0.72 for mESC vs. MEF, r=0.65 for rESC vs. REF). I also analyzed whether RNA splicing is conserved between mouse and rat. I therefore investigated mRNAs and lncRNAs whose genomic loci overlap >30% between mouse and the synthenic rat region. The conservation of RNA splicing is, however, for mRNAs poor (r=0.35 for MEF vs. REF, r=0.27 for mESC vs. rESC) and for lncRNAs (r=0.45 for MEF vs. REF, r=0.42 for mESC vs. rESC) only marginally better. The low correlation of mRNA stability is probably explainable by the fact that although nearly all mRNAs are efficiently spliced in mouse and rat, small differences in the deeply sequenced mouse datasets and the less deeply sequenced rat datasets account for these dramatically low correlation coefficients.

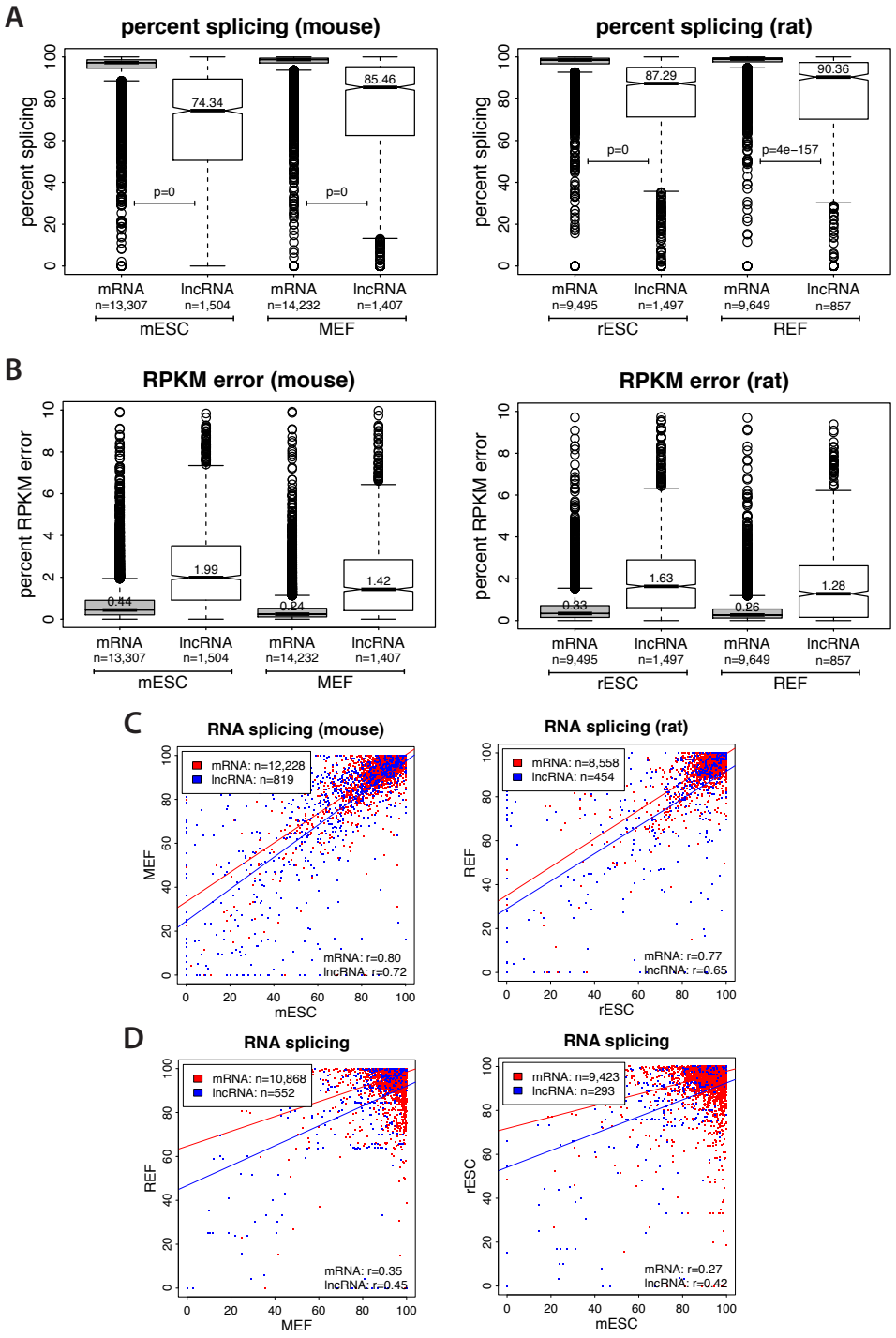


Figure 23: LncRNAs are less spliced than mRNAs

(**A**) Boxplots are showing the distribution of RNA splicing efficiency as calculated in chapter 3.13 for mRNAs (grey) and lncRNAs (white) for ES cells and embryonic fibroblasts of the mouse and rat. (**B**) Boxplots are showing the distribution of relative RPKM errors as calculated in chapter 3.13.3 for mRNAs (grey) and lncRNAs (white) for ES cells and embryonic fibroblasts of the mouse and rat. (**C,D**) Scatter plots are showing the correlations of RNA splicing between two different cell types of the same species (mESC vs. MEF, left, and rESC vs. REF, right) and between the same cell types in two different species (MEF vs. REF, left, and mESC vs. rESC, right). Pearson's correlation coefficients were calculated using the R function `cor()` and are displayed in the bottom right corners. Numbers of mRNAs and lncRNAs are displayed in the top left corners. Trend lines were calculated using the R function `lm()`.

4.3.7 Current lncRNA classes are not distinguished by RNA biology

One of the main questions I wanted to answer with this study is whether lncRNA subclasses can be distinguished by RNA biology. Figure 24A shows that the density curves of the four lncRNA subclasses heavily overlap each other for each of the three investigated RNA biology features (stability, export, splicing). The peak heights of the curves might be variable or slightly shifted to the left or right, but overall, no subclass can be differentiated by any of the three RNA biology features. When the individual lncRNAs of each subclass are plotted separately (Figure 24B) and each RNA biology feature is compared against each other, there is also no separation of any lncRNA subclass visible. LncRNAs of all four subclasses are equally distributed in all three comparisons and no separated cloud of a subclass becomes apparent. This indicates that the overall RNA biology of four major lncRNA subclasses is very similar and that it is unlikely that RNA biology plays an important role for the diverse functions of each lncRNA subclass.

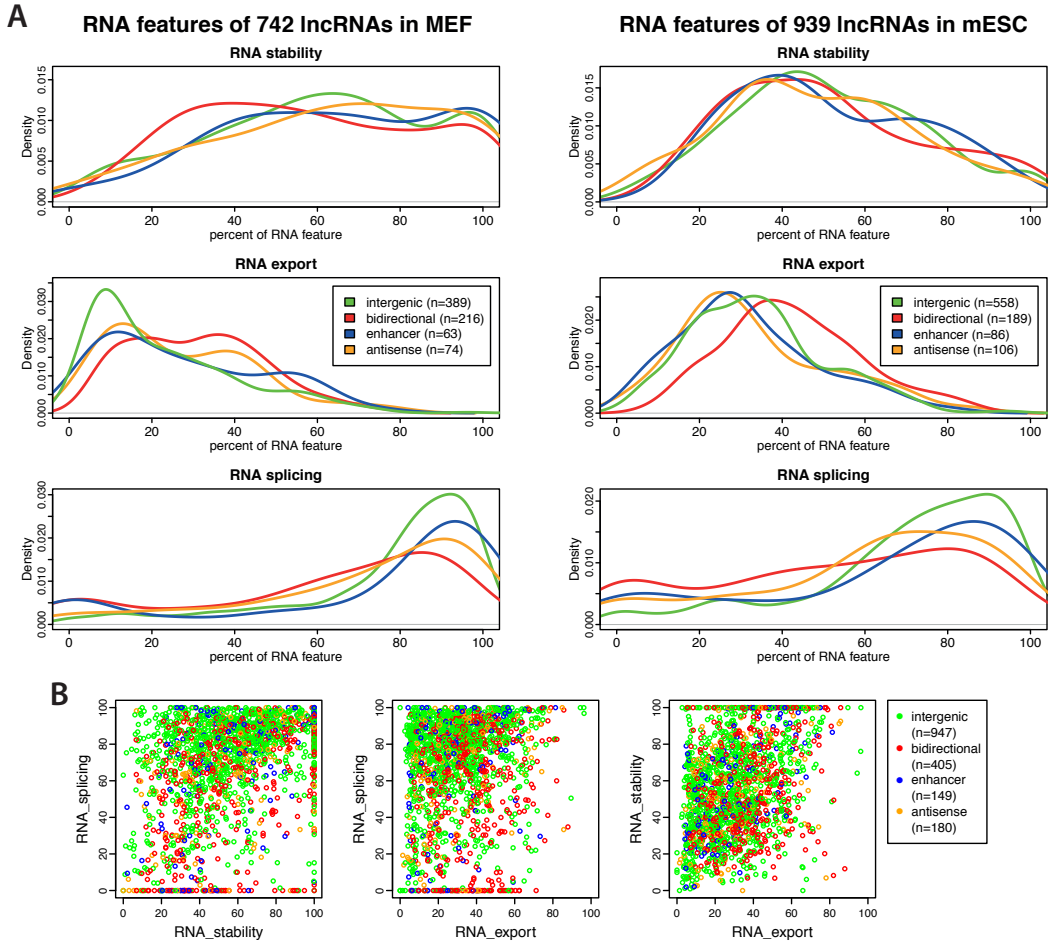


Figure 24: Current lncRNA classes are not distinguished by RNA biology

(**A**) Density plots are showing the distribution of the three RNA biology features stability, export and splicing for four lncRNA subclasses in mESC (left panel) and MEF (right panel). (**B**) Scatter plots are comparing the RNA biology features stability vs. splicing (left), export vs. splicing (middle) and export vs. stability (right) for 1,681 lncRNAs classified into four lncRNA subclasses. LncRNAs of mESC and MEF were combined in this plot, numbers of lncRNA loci per subclass are indicated on the right side.

4.4 Clustering of lncRNAs by RNA biology

4.4.1 Clustering of lncRNAs and mRNAs by RNA biology

I intersected the RNA export, RNA stability and RNA splicing datasets and isolated all lncRNAs and mRNAs that were retained in each of these three analyses. All together 1,681 lncRNAs (939 in mESC, 742 in MEF; 420 of them in both cell types) and 24,510 mRNAs (12,006 in mESC, 12,504 in MEF; 10,743 of them in both cell types) were kept for further analyses (Figure 25A). In order to compare the RNA biology of lncRNAs and mRNAs by clustering, I had to reduce the number of mRNAs. I randomly selected 300 mRNAs from the list of mRNAs being expressed in both cell types using the R function `sample()` and analyzed whether the distribution of each RNA biology feature of those 300 mRNA is similar compared to the total amount of mRNAs. Figure 25B shows that the overall distribution of RNA export, RNA stability and RNA splicing of the 300 mRNAs in mESC and MEF is representative for all mRNAs being expressed in mESC and MEF. I combined the 1,681 lncRNAs and 600 mRNAs (300 in mESC, 300 in MEF) and clustered them by their three RNA biology features using the popular k-means algorithm. The number of clusters has to be determined empirically beforehand and after testing three to seven clusters, I found that six clusters are optimal to recapitulate the diverse RNA biology of lncRNAs and mRNAs. Figure 25C shows the results of the k-means clustering in three scatterplots, each comparing two of the three RNA biology features. In the left plot, RNA splicing and RNA stability nicely separate cluster 5 (lowly spliced and stable) from cluster 6 (also lowly spliced but less stable). The remaining four clusters are more efficiently spliced with cluster 1 and 3 being highly spliced and cluster 2 and 4 being intermediate spliced. RNA stability clearly separates cluster 1 from cluster 3 and cluster 2 from cluster 4. While cluster 1 and 2 are both rather stable, cluster 3 and cluster 4 are unstable. In the middle plot, RNA splicing and RNA export do not separate the clusters as well, however, it becomes apparent that the clusters 1, 3 and 5 are more exported than the clusters 2, 4 and 6, respectively. In the right plot, RNA stability and RNA export separate clusters 1 to 4 with cluster 5 and 6 lying beneath them. Cluster 1 and 2 are both rather stable but cluster 1 is inefficiently exported while cluster 2 is efficiently exported. Cluster 3 and 4 are rather unstable and again are separated by their RNA export, with cluster 3 being more exported than cluster 4. Less visible, cluster 5 is stable and cluster 6 is unstable, with cluster 5 being more efficiently exported than cluster 6. Additionally, I marked the position of well-studied lncRNAs in MEF within these clusters. In a 3D plot, the clusters of Figure 25C become better visible (Figure 25D). Each clustered mRNA and lncRNA can be imagined to be located in a three-dimensional cube with the dimensions x, y and z. Its position inside this cube is determined by the three RNA biology features that define the x, y and z position. Looking at the cube from the front, from the side or from the top, each position will give a different appearance of the clusters (as shown in Figure 25C).

I analyzed the distribution of the three RNA biology features in each of the six clusters in more detail (Figure 25E). In terms of RNA splicing, clusters 1 to 3 are well spliced with

clusters 4 to 6 being gradually less efficiently spliced. RNA stability is gradually reduced from cluster 1 to cluster 4, with cluster 5 being rather stable and cluster 6 being unstable. RNA export defined clusters 1, 3 and 5 as being efficiently exported and cluster 2, 4 and 6 being at least partially nuclear retained. This further confirms that RNA biology is different for each of these clusters. Next, I investigated the distribution of mRNAs and lncRNA subclasses in each cluster. Figure 25F shows that mRNAs are predominantly found in cluster 1 to 3, only 12 out of 600 mRNAs are found in cluster 4 and 1 in cluster 5. Each of the four lncRNA subclasses is fairly equally present in all six clusters, reinforcing the notion that lncRNA subclasses can not be distinguished by RNA biology (see Figure 24). Figure 25G shows the distribution of mRNAs and lncRNA subclasses in each of the six clusters. Approximately half of all lncRNAs cluster together with mRNAs (in black) in clusters 1 to 3, indicating that they have an mRNA-like RNA biology. Accordingly, the other half has a non-mRNA-like RNA biology.

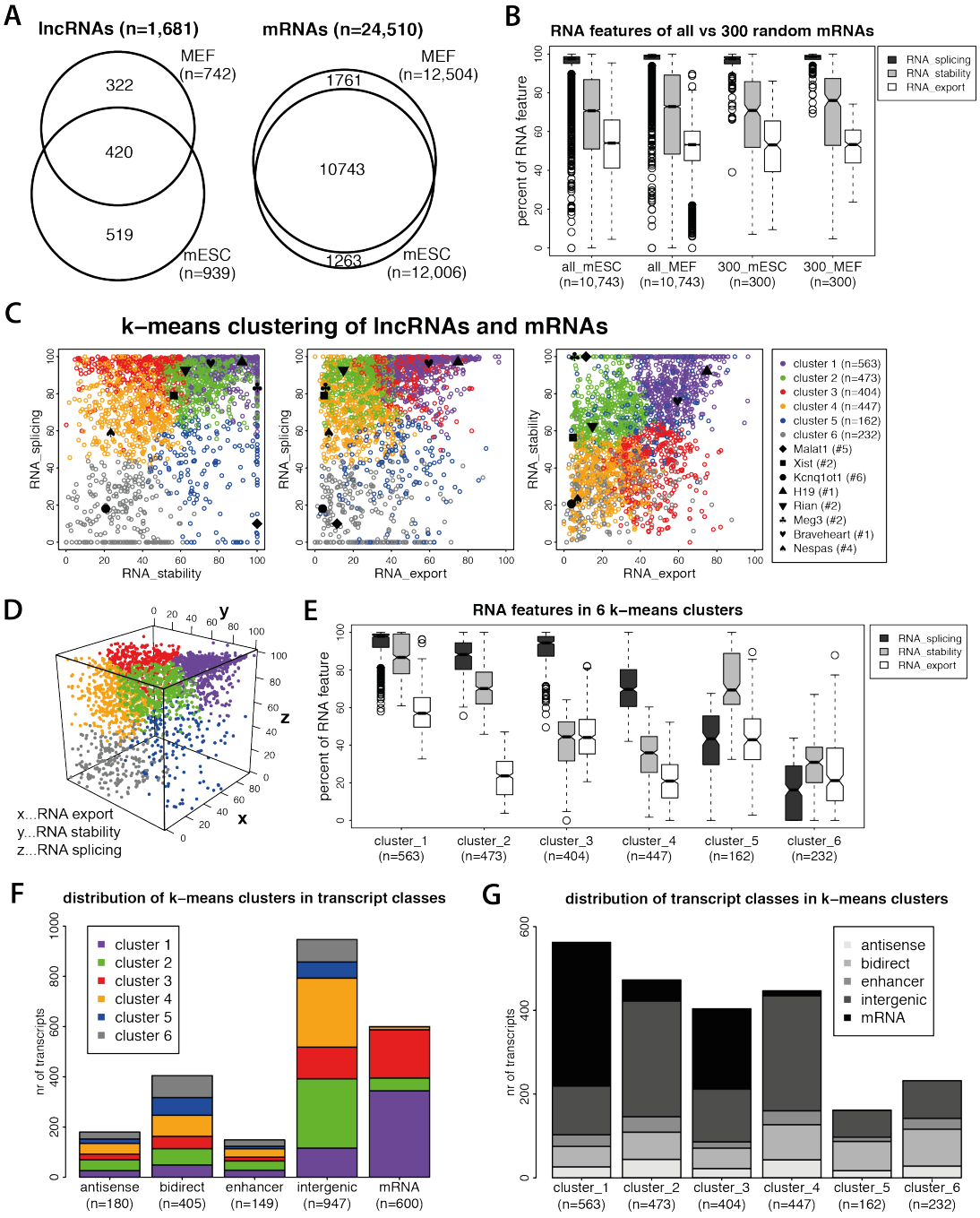


Figure 25: Clustering of lncRNAs and mRNAs by RNA biology

(**A**) Venn diagrams showing the numbers of lncRNA (left) and mRNA loci (right) for which all three RNA biology features (export, stability, splicing) are available in MEF and mESC. (**B**) For 10,743 mRNA loci all three RNA biology features are available in both mESC and MEF. The boxplot displays the distribution of RNA splicing, RNA stability and RNA export for these mRNAs in mESC and in MEF and for 300 randomly picked mRNAs. (**C**) Scatterplots showing the results of the k-means clustering of 1,681 lncRNA datapoints and 600 mRNA datapoints based on three RNA biology features. Each cluster is depicted in a different color. The left plot compares RNA splicing with RNA stability, the middle plot RNA splicing and RNA export and the right plot RNA stability and RNA export. The positions of eight well-studied lncRNAs are marked by black symbols, in the legend the number of their clusters in indicated. For these lncRNAs only the RNA biology features in MEF are shown. (**D**) Same data as in (C), but depicted in a three-dimensional plot. (**E**) Boxplot displaying the distribution of the three RNA biology features splicing, stability and export for each of the six k-means clusters. (**F**) Barplot depicting the cluster distribution in each of four lncRNA classes and mRNAs. The color scheme is identical to (C) and (D). (**G**) Barplot depicting the transcript class distribution in each of the six k-means clusters. Clusters 1 to 3 contain 97.83% of investigated mRNAs and 50.74% of lncRNAs, which are therefore mRNA-like.

4.4.2 Half of lncRNAs are in the same cluster in MEF and mESC

After clustering lncRNAs by their RNA biology in MEF and mESC, I tested to which extent lncRNAs that are expressed in both cell types are actually falling into the same cluster. I investigated the cluster affiliation of 420 lncRNAs and 300 mRNAs (see Figure 25A) for which RNA biology data is available from MEF and mESC and that were used for clustering in Figure 25C. Figure 26A shows that 51.43% of lncRNAs are in the same cluster in both cell types (left panel, blue box), compared to 68.33% of mRNAs (right panel, blue box). When I looked more closely from which cluster lncRNAs switch into which cluster, I found that some specific patterns emerged. The largest group of switching lncRNAs (12.62%, grey box) is in MEF in cluster 2 and in mESC in cluster 4. The second largest group (8.09%, yellow box) switches from cluster 2 to cluster 3 and the third largest group (5.48%, green box) from cluster 4 in MEF to cluster 6 in mESC. The remaining 22.38% (red box) contain lncRNAs that switch in any of the other 27 possible combinations between the six clusters in MEF and mESC. As mRNAs are predominantly found in only three clusters, their two main groups switch from cluster 1 in MEF to cluster 3 in mESC (9.67%, grey box) or from cluster 3 in MEF to cluster 1 in mESC (also 9.67%, green box). The remaining 12.34% (red box) contain mRNAs that switch in other combinations between the clusters in MEF and mESC.

The fact that lncRNAs and mRNAs switch preferably from a few clusters to a few other clusters in very specific combinations argues that it is not a random switching due to inaccurate mapping of RNA biology features but rather due to specific biological differences between MEF and mESC. I investigated whether the RNA biology features of lncRNAs (Figure 26B) and mRNAs (Figure 26C) switching from one cluster in MEF to another cluster in mESC are actually due to specific changes in RNA biology features. And indeed, when I plot the RNA biology features of 53 lncRNAs switching from cluster 2 in MEF to cluster 4 in mESC, I find that the main difference is the reduction of RNA stability from ~70% in MEF to ~40% in mESC (Figure 26B, left panel, indicated by arrow). When I compare this change to the RNA biology of the whole cluster 2 and 4, it becomes evident that while RNA splicing and RNA export is very similar, the main difference between these two clusters is also RNA stability (indicated by arrow). This establishes that the RNA stability of these 53 lncRNAs is apparently differentially regulated between MEF and mESC and thereby explains why they switch from cluster 2 to cluster 4. The second largest group of 34 lncRNAs switch from cluster 2 in MEF to cluster 3 in mESC. Again, I compared the RNA biology features of these 34 lncRNA in both cell types with the two respective clusters (Figure 26B, right panel). Similarily to the previous example, these 34 lncRNAs have a reduced RNA stability (from ~70% in MEF to ~45% in mESC, indicated by arrow) but also an increased RNA export (from ~20% in MEF to ~40% in mESC, indicated by arrow). The RNA biology features of cluster 2 and 4 show overall the same difference in RNA stability (indicated by parallel arrows), which explains why the 34 lncRNAs switch between these two clusters. The two main groups of switching mRNAs are also explainable by differences in their RNA stability (Figure 26C). 29 mRNAs switch from cluster 1 in MEF to cluster 3 in mESC due to a reduction of RNA stability from ~75% in MEF

to ~50% in mESC (left panel, indicated by arrow) and the same number of mRNAs switches from cluster 3 in MEF to cluster 1 in mESC due to an increase of RNA stability from ~55% in MEF to ~75% in mESC (right panel, indicated by arrow). The main difference between the clusters 1 and 3 is a decreased RNA stability, while RNA splicing and RNA export is in the same range. This establishes that also RNA stability of mRNAs is differentially regulated between MEF and mESC, which is the main driver for mRNAs to switch clusters. In summary, it seems that RNA splicing is the most constant RNA biology feature between MEF and mESC, followed by RNA export and lastly RNA stability, which seems to be main driver for cluster switches.

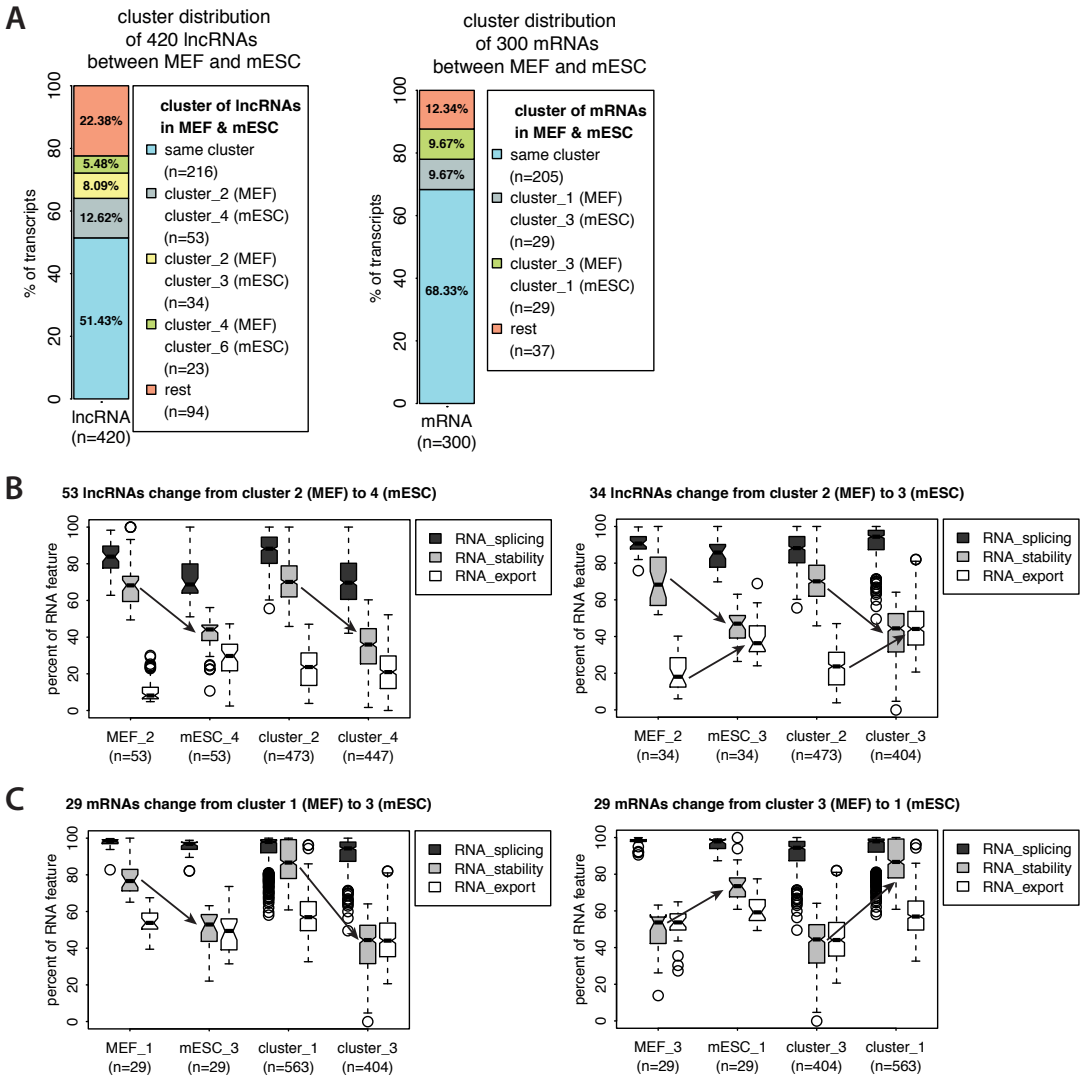


Figure 26: Half of lncRNAs are in the same cluster in MEF and mESC

(**A**) The cluster affiliation of 420 lncRNAs (left panel) and 300 mRNAs (right panel) being expressed in both MEF and mESC was analyzed to determine the fractions of lncRNAs and mRNAs that are constant or switching between the six clusters. Blue boxes indicate the fractions of lncRNAs and mRNAs that are in the same cluster in MEF and mESC. The remaining boxes indicate fractions that are in different clusters in MEF and mESC. In the left panel, three main groups of cluster switching lncRNAs are indicated. In the right panel, two main groups of cluster switching mRNAs are indicated. Red boxes indicate the fractions of lncRNAs and mRNAs that are in different clusters in MEF and mESC but are not included in the main groups of cluster switching lncRNAs or mRNAs. Percentages for each group are indicated inside the boxes, numbers for each group can be found in the legend. (**B**) RNA biology

features are plotted for lncRNAs that switch from cluster 2 in MEF to cluster 4 in mESC (n=53, left panel) and from cluster 2 in MEF to cluster 3 in mESC (n=34, right panel). As a comparison, RNA biology features for all transcripts in the clusters 2 and 4 (left panel) and the clusters 2 and 3 (right panel) are plotted. Arrows indicate the main difference of RNA biology between the cluster switching transcripts in the two cell types and the respective clusters. (**C**) RNA biology features are plotted for mRNAs that switch from cluster 1 in MEF to cluster 3 in mESC (n=29, left panel) and from cluster 3 in MEF to cluster 1 in mESC (n=29, right panel). As a comparison, RNA biology features for all transcripts in the clusters 1 and 3 (left panel) and the clusters 3 and 1 (right panel) are plotted. Arrows indicate the main difference of RNA biology between the cluster switching transcripts in the two cell types and the respective clusters.

4.4.3 Mouse RNA biology clusters are conserved in rat

In order to answer the question if RNA biology is conserved between mouse and rat, I analyzed whether the RNA biology features of each k-means cluster are comparable between mouse and rat. I define lncRNAs and mRNAs as conserved when their genomic loci overlap >30% between mouse and the synthenic rat region. Due to the fact that I have RNA export data only for REF, I refrained from clustering rat transcripts according to their RNA biology. Instead, I plotted the RNA biology features of conserved mouse and rat transcripts according to their mouse cluster (Figure 27). Figure 27A shows RNA splicing for conserved lncRNAs (left boxplot) and mRNAs (right boxplot). Grey boxes contain conserved mouse lncRNAs and white boxes contain the corresponding rat lncRNAs. As can be seen, mouse lncRNAs are efficiently spliced in cluster 1 to 3 and are gradually less spliced in cluster 4 to 6. Rat lncRNAs follow this trend, albeit with higher splicing efficiencies (as already seen in Figure 23A). In terms of mRNAs, mouse as well as rat mRNAs are all efficiently spliced. Figure 27B displays RNA stability and mouse lncRNAs have a gradually reduced RNA stability from cluster 1 to 6, with the only exception of cluster 5 that exhibits greater RNA stability. Again, rat lncRNAs tend to follow this pattern although their overall RNA stability is higher than for mouse lncRNAs (as already seen in Figure 21A). In terms of mRNAs, RNA stability decreases gradually from cluster 1 to 3 in mouse as well as rat. Figure 27C shows RNA export and mouse lncRNAs are well exported in clusters 1, 3 and 5 while they are less efficiently exported in clusters 2, 4 and 6. Rat lncRNAs again follow this “zig-zag” pattern very closely. Mouse mRNAs are well exported in cluster 1 and 3 and less exported in cluster 2 and rat lncRNAs show the same trend.

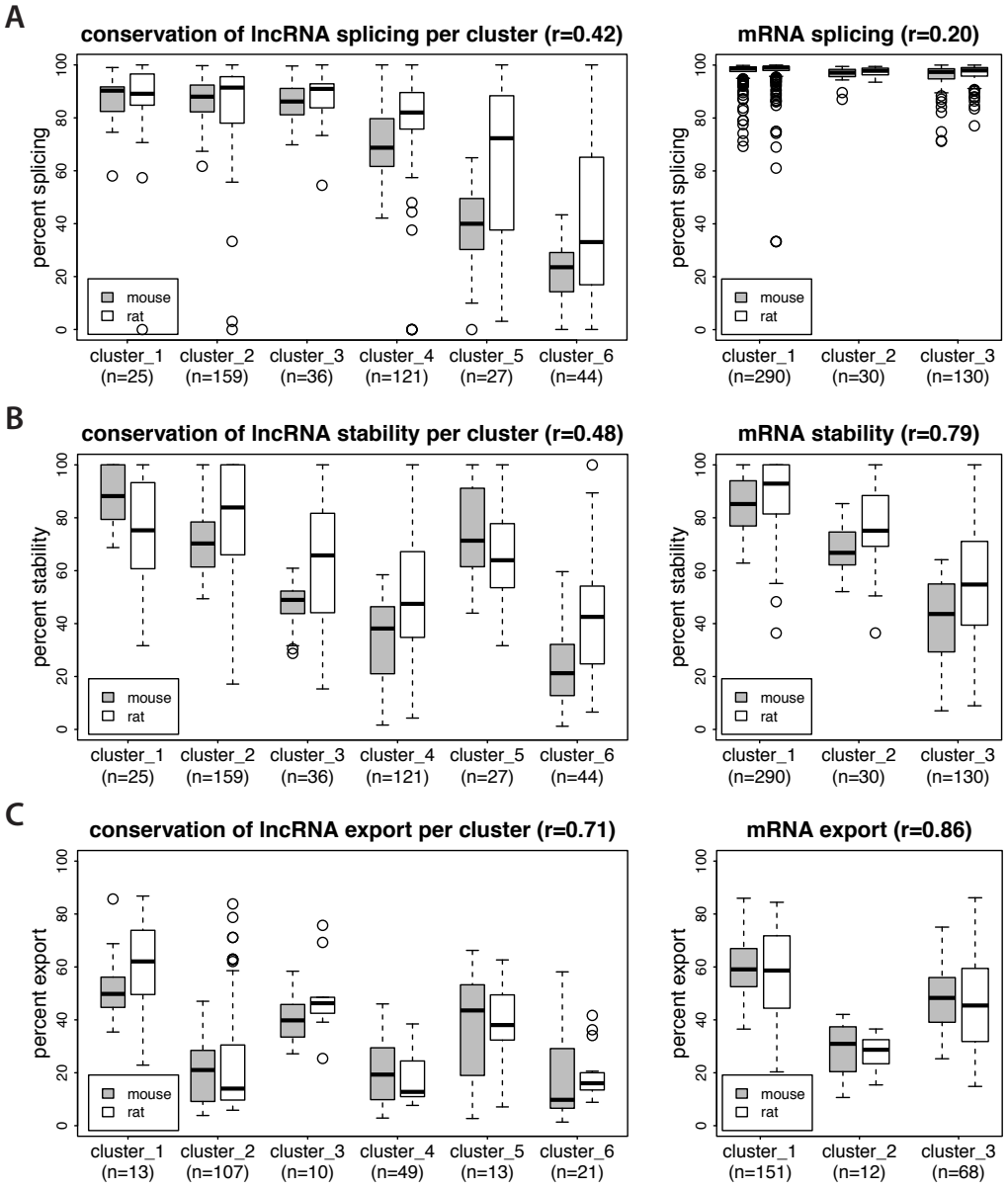


Figure 27: Mouse RNA biology clusters are conserved in the rat

For this analysis, I considered only lncRNAs and mRNAs that are conserved between mouse and rat. I defined conservation by a >30% overlap of their synthenic genomic loci in mouse and rat. Boxplots depict (**A**) RNA splicing (**B**) RNA stability and (**C**) RNA export of conserved lncRNAs (left boxplots) and mRNAs (right boxplots). LncRNAs are grouped according to the cluster the mouse lncRNAs are in (see Figure 25C). For mRNAs, only cluster 1 to 3 are shown as the remaining clusters do not contain >5 mRNAs. The Pearson correlation coefficients are indicated after the plot title.

4.4.4 Genomic transcript features differ in six RNA biology clusters

In Figure 12 I concluded that lncRNAs have significantly different genomic transcript features than mRNAs, however, four well-studied lncRNA subclasses do not differ much. After clustering of lncRNAs and mRNAs by RNA biology, I was interested to see whether any of the genomic transcript features is different between lncRNAs of these six RNA biology clusters. When the first three features exon count, cDNA size and locus size are compared it becomes evident that cluster 2, 4 an 6 have longer cDNAs, however, only cluster 2 and 4 also have a

higher exon count and locus size (Figure 28A). This indicates that lncRNAs in cluster 6 have the longest cDNA (median of 3kb, 25% are longer than 5kb), yet, they mostly have only 2 or 3 exons and their locus size is not larger than average. When the average exon length is compared between the six clusters, it becomes evident that lncRNAs in cluster 6 have in fact the longest exons, which might also explain why they have the lowest values of RNA stability, RNA splicing and RNA export. Accordingly, the exon/intron ratio is also highest for cluster 6. In contrast, the exonic repeat coverage does not differ significantly between lncRNAs in any of the six clusters (Figure 28A). I also investigated steady-state RNA levels of each RNA biology cluster in MEF and mESC (Figure 28B). In both cell types, cluster 1 has the highest median RPKM (1.81 in MEF, 3.80 in mESC) and cluster 6 the lowest median RPKM (0.35 in MEF, 0.45 in mESC). This hints towards the fact that lncRNAs in cluster 6 are either lower expressed than in cluster 1 or that lncRNAs in cluster 6 are less abundant because of their low RNA stability compared to lncRNAs in cluster 1. Presumably, both factors play a role as lncRNAs in cluster 3 and 4 are also rather unstable but still have ~2-fold higher RPKMs than cluster 6.

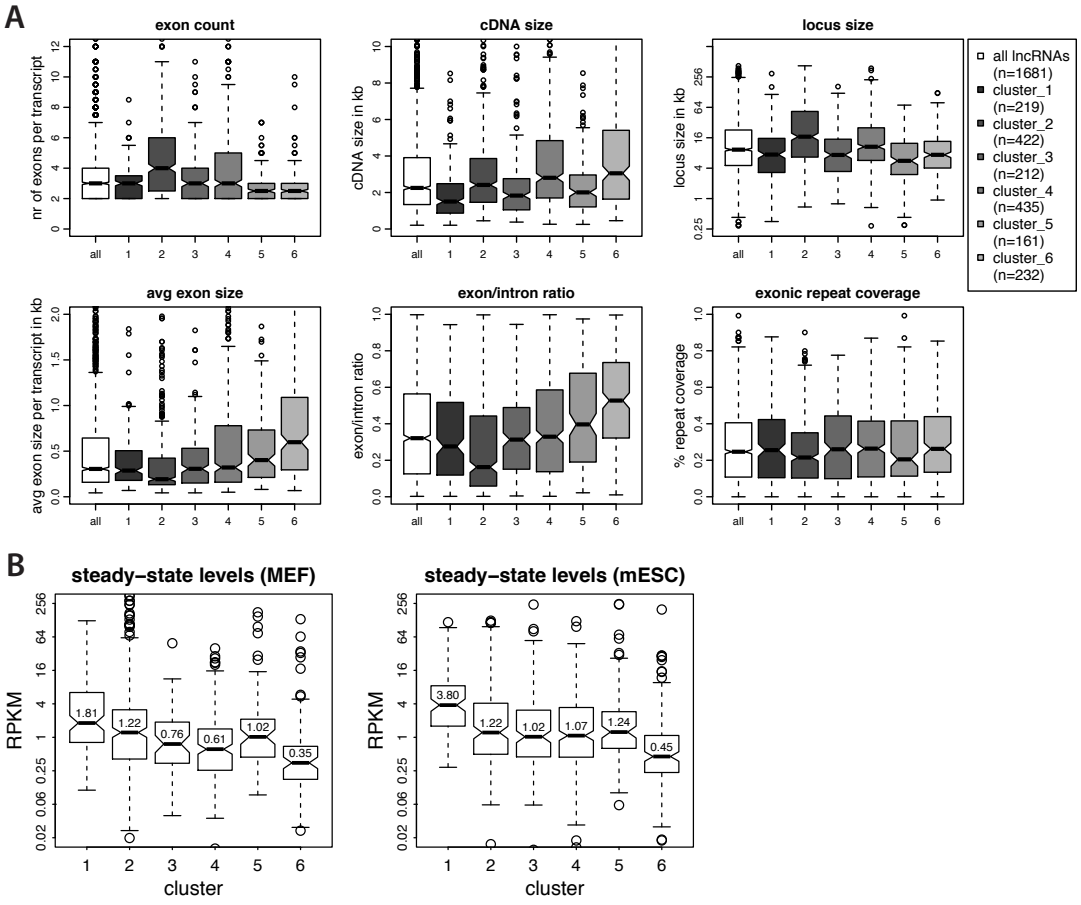


Figure 28: Genomic transcript features for six RNA biology clusters

(**A**) Six genomic transcript features were calculated for mouse lncRNAs clustered in Figure 25C and averaged (by median) for each locus. Boxplots depict features for all lncRNAs (white boxes) and lncRNAs of respective clusters (shaded grey boxes). The exonic repeat coverage was calculated using the UCSC repeatmasker track, all other features were directly calculated from the annotation bed file. (**B**) Boxplots depict RPKMs for MEF (left) and mESC (right) lncRNAs according to their cluster affiliation (Figure 25C). Numbers of medians are plotted inside each box.

4.4.5 RNA stability and RNA export correlate with RNA abundance

I analyzed whether certain RNA biology features of lncRNAs and mRNAs are correlated with RNA abundance (steady-state levels). As can be seen in Figure 29, RNA stability is indeed positively correlated with RNA abundance of mRNAs (MEF: r=0.44, mESC: r=0.45) and lncRNAs (MEF: r=0.28, mESC: r=0.32) in both cell types (left panel). Lowly abundant mRNAs have a median RNA stability of <60% compared to >90% for highly abundant mRNAs. LncRNAs increase their median RNA stability from ~50% to ~70% in MEF and from ~40% to ~60% in mESC as their abundance increases. Also RNA export is positively correlated with abundance of mRNAs (MEF: r=0.40, mESC: r=0.46) and mESC but not MEF lncRNAs (MEF: r=0.16, mESC: r=0.31) in both cell types (middle panel). The RNA export of mRNAs increases from ~50% to ~60% in MEF and from ~50% to even 70% in mESC as their abundance increases. LncRNAs follow the trend and slightly increase their export with increasing abundance in mESC, however, in MEF this trend is not observed as pronounced. RNA splicing (right panel) of mRNAs seems to be unaffected by their steady-state levels, lncRNAs show a trend towards more efficient splicing with higher abundance, although the significance is low and the notches of the boxplots overlap each other (see chapter 3.18).

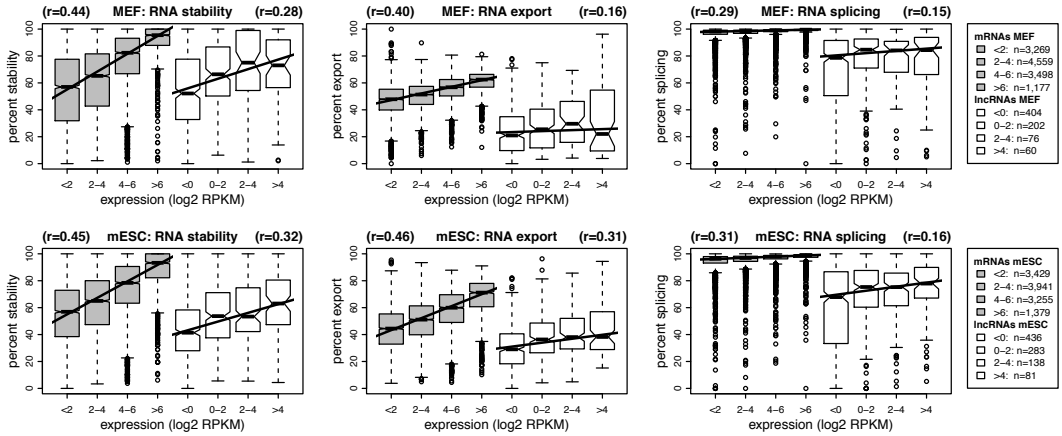


Figure 29: RNA stability and RNA export correlate with expression strength

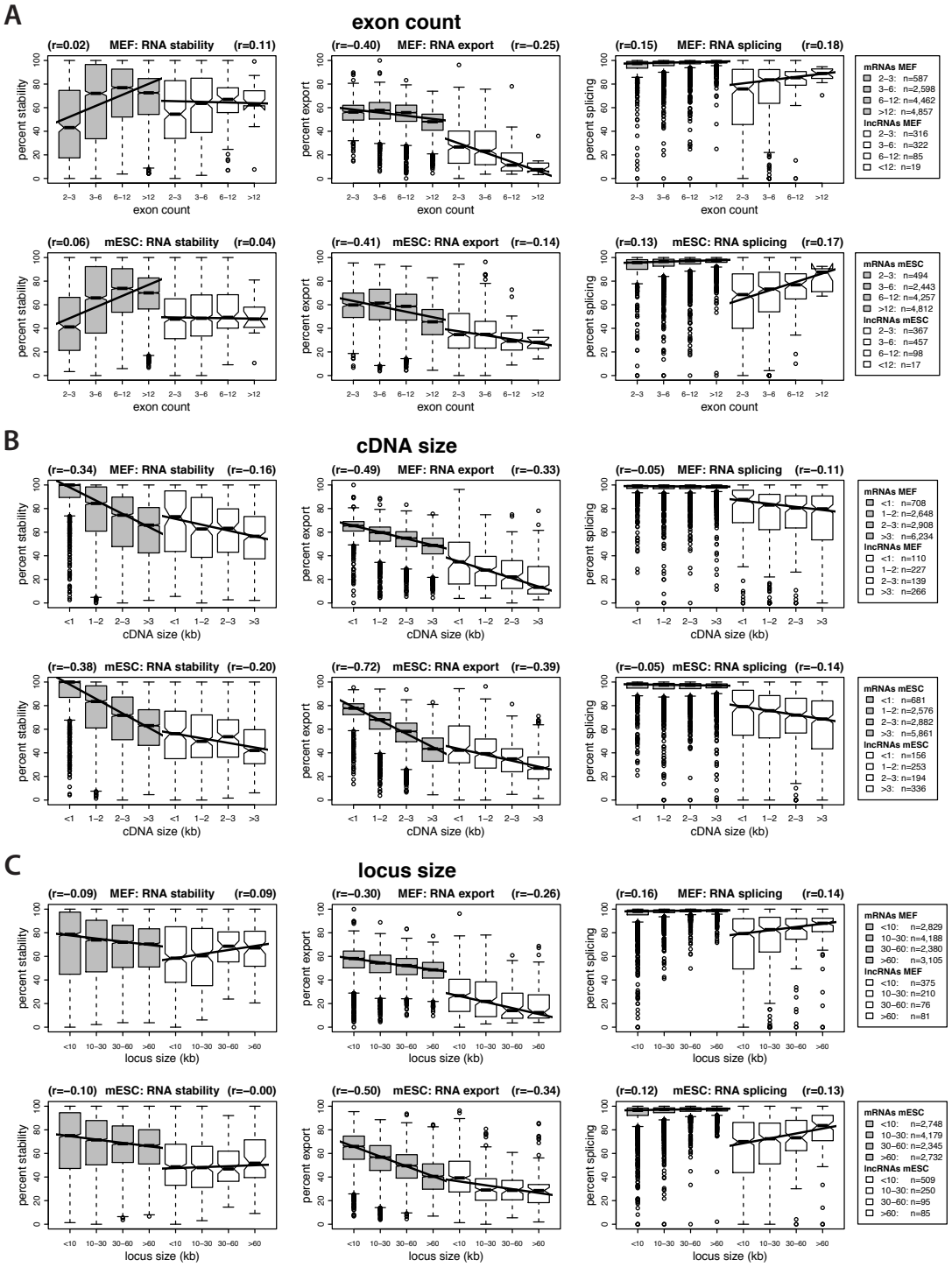
RPKM were calculated for mRNAs and lncRNAs using polyA+ RNA-seq data of MEF and mESC. Transcripts were binned by their log2 transformed steady-state levels (abundance measured by RPKM). For each bin, RNA stability (left panel), RNA export (middle panel) and RNA splicing (right panel) is depicted in boxplots. In contrast to the six genomic transcripts features in Figure 30, it was not possible to maintain the same bin definitions for mRNAs and lncRNAs as their expression strength is too different (see Figure 13). Statistical significance was judged by the overlap of the notches of the boxplots (see chapter 3.18) and the slope of the trend lines that were calculated from the respective medians by the R function `lm()`. Spearman's correlation coefficients are plotted above each plot on the left side (for mRNAs) and on the right side (for lncRNAs).

4.4.6 RNA biology correlates with certain genomic transcript features

Finally, I was interested whether RNA biology is correlated with genomic transcript features such as exon count, cDNA size, locus size, average exon size, exon/intron ratio or repeat coverage. For each of the six genomic transcript features I divided lncRNAs and mRNAs into four bins according to their genomic transcript features and plotted RNA biology features

separately for each bin. Figure 30 depicts RNA stability (left plots), RNA export (middle plots) and RNA splicing (right plots) for MEF (top panel) and mESC (bottom panel) for four bins of each of the six genomic transcript features. Trendlines were calculated from the medians of the four respective bins while Spearman correlation coefficients were calculated from unbinned data. In Figure 30A the correlation between the exon count and the RNA biology features stability, export and splicing in MEF and mESC is depicted. RNA stability of mRNAs increases with higher exon counts and levels off as they reach six exons, however, as the majority of mRNAs has more than six exons this is statistically not significant (MEF: r=0.02, mESC: r=0.06). For lncRNAs, this trend is also visible in MEF (r=0.11) but not mESC (r=0.04). RNA export is negatively correlated with the exon count of mRNAs (MEF: r=-0.40, mESC: r=-0.41) and to a lesser extent with the exon count of of lncRNAs (MEF: r=-0.25, mESC: r=-0.14). RNA splicing of mRNAs is not correlated with the exon count, however, lncRNAs show a clear trend in MEF (r=0.18) and mESC (r=0.17) that the more exons they have, the more they are spliced. Figure 30B depicts the negative correlation between cDNA size and RNA biology. While mRNAs in MEF and mESC show a dramatic reduction of RNA stability with longer cDNA lengths (MEF: r=-0.34, mESC: r=-0.38), lncRNAs show a less pronounced effect (MEF: r=-0.16, mESC: r=-0.20). In terms of RNA export, both mRNAs (MEF: r=-0.49, mESC: r=-0.72) and lncRNAs (MEF: r=-0.33, mESC: r=-0.39) are significantly less exported the longer the cDNA gets. RNA splicing is again unaffected in mRNAs (MEF: r=-0.05, mESC: r=-0.05) however, lncRNAs are by trend less efficiently spliced the longer their cDNA becomes (MEF: r=-0.11, mESC: r=-0.14). In Figure 30C the relationship between the locus size and RNA biology is shown. In contrast to cDNA length (Figure 30B), the locus length has only a marginal negative impact on mRNA and lncRNA stability in MEF and mESC. The RNA export efficiency of mRNAs (MEF: r=-0.30, mESC: r=-0.50) and lncRNAs (MEF: r=-0.26, mESC: r=-0.34) is negatively correlated with increasing locus sizes. RNA splicing of mRNAs is not affected, however, lncRNAs are by trend more efficiently spliced in MEF (r=0.14) and mESC (r=0.13) with increasing locus sizes. Figure 30D shows the correlation between the average exon size of mRNAs or lncRNAs and RNA biology. RNA stability of mRNAs is significantly decreased in MEF (r=-0.25) and mESC (r=-0.28) when their exons increase their average length. This trend is also observed for lncRNAs, with higher effects in MEF (r=-0.20) than mESC (r=-0.13). RNA export of mRNAs and lncRNAs seems not to be affected by the average exon length. RNA splicing is negatively correlated with longer exons in lncRNAs (MEF: r=-0.24, mESC: r=-0.21), but not mRNAs. Figure 30E depicts the relationship between the exon/intron ration and RNA biology. In both cell types, RNA stability of mRNAs (MEF: r=-0.11, mESC: r=-0.13) and lncRNAs (MEF: r=-0.22, mESC: r=-0.14) is marginally negatively affected by higher exon/intron ratios. RNA export shows a trend to be marginally increased with higher exon/intron ratios, however, the correlation coefficients are low. RNA splicing of mRNAs is not affected by the exon/intron ratio, however, splicing of lncRNAs is considerably reduced with higher exon/intron ratios in MEF (r=-0.24) and mESC (r=-0.25). In Figure 30F the relationship between exonic repeat coverage and RNA biology is

portrayed. It seems that among the six genomic transcript features, repeat coverage has the least effects on RNA stability, RNA export and RNA splicing. The only noticeable effect is that mRNAs with >40% exonic repeat content (n=197 in MEF, n=214 in mESC) are unstable and less exported than all other mRNAs. A considerable fraction of lncRNAs contains repeats in their exons, however, their RNA biology seems unaffected by varying repeat content.



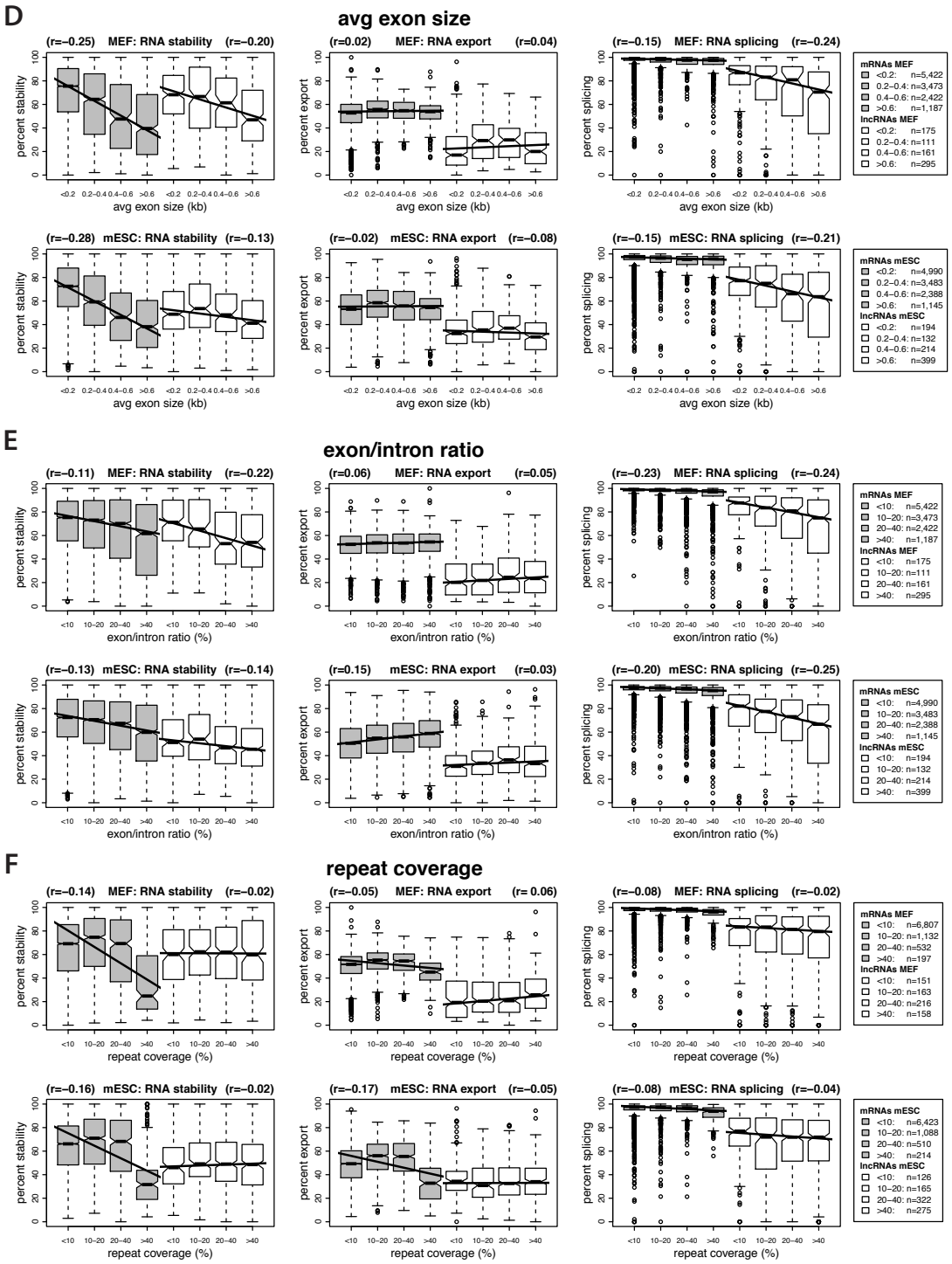


Figure 30: RNA biology of lncRNAs binned by their genomic transcript features

Six genomic transcript features were calculated for mouse lncRNAs and mRNAs and averaged for each locus. For each feature, lncRNAs and mRNAs were divided into four bins by their value of the respective feature. Boxplots depict features for mRNAs (grey boxes) and lncRNAs (white boxes). Depicted is RNA stability (left plots), RNA export (middle plots) and RNA splicing (right plots) for MEF (top panel) and mESC (bottom panel) for four bins of the six genomic transcript features (**A**) exon count, (**B**) cDNA size, (**C**) locus size, (**D**) average exon size, (**E**) exon/intron ratio and (**F**) exonic repeat coverage. The exonic repeat coverage was calculated using the UCSC repeatmasker track, all other features were calculated from the annotation bed file. Statistical significance was judged by the overlap of the notches of the boxplots (see chapter 3.18) and the slope of the trend lines that were calculated from the respective medians by the R function `lm()`. Spearman's correlation coefficients were calculated from unbinned data and are plotted above each plot on the left (for mRNAs) and right side (for lncRNAs).

5 DISCUSSION

5.1 Summary of results

In this study, I investigated the RNA biology of mouse and rat lncRNAs genome-wide. I therefore established and optimized an RNA-seq pipeline, applied it to 76 RNA samples and then used a comprehensive lncRNA annotation for the mouse and the rat to analyze lncRNA biology. I demonstrate, in agreement with published data (Cabili et al., 2011; Ulitsky and Bartel, 2013), that lncRNAs are lowly abundant, tissue-specifically expressed and developmentally regulated. Furthermore, I analyzed genomic transcript features such as cDNA length, exon/intron ratio and repeat coverage and confirm that lncRNAs differ significantly from mRNAs (Ulitsky and Bartel, 2013). In order to investigate the RNA biology of lncRNAs, I conducted experiments to assay RNA export, RNA stability and RNA splicing in ES cells and embryonic fibroblasts of the mouse and rat. After RNA-sequencing, bioinformatic analyses and rigorous quality filtering, I find that lncRNAs are significantly less exported, less stable and less spliced than mRNAs. I divided lncRNAs into the currently used classes of intergenic, antisense, enhancer and bidirectional lncRNAs and find that they are not distinguishable by any of the three RNA biology features. Therefore, I clustered lncRNAs and mRNAs according to their RNA biology and defined six clusters, each having a unique RNA biology signature. I show that the RNA biology of these six clusters is largely conserved between mouse and rat and that RNA biology significantly correlates with certain genomic transcript features. These finding indicate that RNA biology can be a guide to understanding lncRNA function. The compiled RNA-seq datasets that I generated for this thesis will provide a valuable research tool for the research community to further unravel properties and functions of lncRNAs.

5.2 Towards an efficient RNA-seq pipeline to annotate lncRNAs

5.2.1 ScriptSeq v1 and v2 kits produce biased RNA-seq libraries

When this study was started in 2011, the ScriptSeq v1 kit was the first commercially available kit that allowed the preparation of stranded RNA-seq libraries. With a novel chemistry, this kit was based on tagged random hexamers that prime the fragmented RNA and are extended to give rise to first-strand cDNA tagged at the 5'end. In a second step, terminal-tagging oligos that are 3' end blocked bind and tag the 3' end of the first-strand cDNA. This way the directionality is preserved when RNA is transcribed into cDNA. The fast and easy protocol was an advantage when multiple libraries had to be prepared, however, as shown in Figure 6A the RNA-seq coverage was uneven and the data therefore not optimal for transcriptome assembly and splicing analysis. Most probably, the problem lies in the design of the kit as it relies on tagged hexamers that apparently are not binding randomly to RNA or do not bind randomly due to their attached tag, thereby creating uneven transcript coverage. These

hexamers biases have been noted before in multiple RNA-seq datasets, thereby adversely affecting the uniformity of read coverage along expressed transcripts (Hansen et al., 2010). Calculation of RPKM from ScriptSeq data is still possible as the coverage biases occur in all transcripts and scale with their expression levels (Figure 6C). Several programs (e.g. cufflinks and the R package RNASeqBias) correct for the hexamers bias, however, when RPKMs of the same transcript are compared between tissues the biases cancel themselves out. After the bias problems of the ScriptSeq v1 kit emerged, I tested the improved ScriptSeq v2 kit. This kit was advertised to offer improved hexamer design to facilitate more even transcript coverage. However, the transcript coverage still showed strong regional biases and in direct comparison to the dUTP/TruSeq protocol its overall transcript coverage was poor (Figure 6). It seemed that the sequencing community did not use the ScriptSeq kits extensively, I know only very few studies that published ScriptSeq RNA-seq data. A PubMed search (Oct. 30, 2014) confirms this observation as the term “scriptseq” is only found once, compared to “truseq” that was found 27 times. Accordingly, in the GEO database the term “scriptseq” is present in 24 datasets, compared to 650 appearances of “truseq”. This shows that the ScriptSeq kits lack acceptance in the community although the protocol is easy to follow and fast. In terms of quality, they may be sufficient for simple gene expression purposes, however, for more complex applications such as transcriptome assembly and splicing analysis they might need improvement.

5.2.2 The dUTP/TruSeq protocol generates superior RNA-seq libraries

I established the dUTP/TruSeq protocol as an alternative to the ScriptSeq kits (Figure 4-3). This protocol combines the unstranded TruSeq kit with a few modifications to facilitate strandedness (Sultan et al., 2012). The effort to prepare dUTP/TruSeq libraries is significantly higher compared to ScriptSeq libraries, however, the quality difference was so convincing that I switched to the dUTP/TruSeq protocol. The transcript coverage is much more even and considerably higher than observed in ScriptSeq data (Figure 6). After I finished most of the RNA-seq libraries for this study, Illumina launched a stranded version of the original TruSeq kit, thereby incorporating the dUTP modifications into the kit. The TruSeq kits seem to be well accepted in the sequencing community and are widely used for DNA-seq and ChIP-seq experiments. I reduced RNA hydrolysis time to increase the size of the RNA fragments (Figure 7) and found that 100bp paired-end RNA seq data is most useful to assembly transcripts (Figure 8). Many of the studies that assembled transcriptomes used the TruSeq kit (Alvarez-Dominguez et al., 2013), although not all of them were strand-specific (Cabili et al., 2011; Keane et al., 2011). For transcriptome assembly, most studies also used paired-end RNA-seq data preferentially with a read length of at least 76bp. Longer reads are mapped more accurately and span more splice junctions and thereby assemble transcripts better. Taken together, the choice of the optimal RNA-seq library preparation method and the optimization of the dUTP/TruSeq protocol improved RNA-seq data quality considerably.

5.3 A comprehensive lncRNA annotation for the mouse and rat

5.3.1 Choosing the ideal lncRNA annotation for this study

In order to investigate the RNA biology of lncRNAs in an unbiased way, I needed a comprehensive lncRNA annotation that included all kinds of transcripts: from highly abundant to lowly abundant, from well spliced to inefficiently spliced and from stable to unstable. When this study started in 2011, the first mouse lncRNA annotations were already published (Guttman et al., 2009, 2010), however, they were obtained from only three mouse tissues and did not contain many of the lncRNAs that I was interested in. While numerous studies followed and annotated mostly well spliced intergenic lncRNAs, they often missed inefficiently spliced imprinted lncRNAs such as *Airn* and *Kcnq1ot1* (Huang et al., 2011). In this study, I therefore used a lncRNA annotation generated by Florian Pauler (see chapter 3.10.1 for details). This annotation was derived from eleven tissues and was probably the most comprehensive lncRNA annotation at the time this project started. Before this annotation became available, I tried in an alternative approach to annotate inefficiently spliced lncRNAs from total Ribo-Zero RNA-seq data (data not shown). In this approach, a pipeline detected putative inefficiently spliced lncRNAs as continuously transcribed regions and was indeed successful for some well expressed lncRNAs, however, it failed for lowly expressed lncRNAs as the annotations were heavily fragmented and the 5'- and 3'-ends were not correctly annotated. I tested if sequencing nuclear RNA would increase the coverage of precursors and nuclear lncRNAs but this did not eliminate the inherent problems of fragmentation (data not shown). The transcriptome assembly algorithms did not work well with Ribo-Zero data as they poorly assembled mRNAs and often skipped introns of lowly spliced lncRNAs and generated a wealth of splice variants. Also, the coverage of splice junctions was low as the majority of the reads in total RNA-seq data (74% in Ameur et al, 2011) maps outside known mRNA exons.

Florian Pauler then developed an algorithm to map lncRNAs using polyA+ RNA-seq data to get continuous exon models and proper 5'- and 3'-ends and subsequently use Ribo-Zero RNA-seq data to estimate the splicing efficiency for each exon model. This approach had the drawback that I could not analyze completely unspliced lncRNAs, however, the annotation pipeline detected even the rarest splice junctions by changing two parameters of the cufflinks assembly algorithm (see chapter 3.10.3). This strategy enabled even the detection of previously unknown splice variants for the *bona fide* unspliced lncRNA *Kcnq1ot1*. A diploma student in the Barlow laboratory verified 104/116 (89.66%) splice junctions of 60 lncRNAs by PCR, indicating that most splice junctions were annotated correctly (Christoph Dotter, Master's thesis, University of Vienna, 2014). The mouse lncRNA annotation contains 36,578 lncRNA transcripts in 7,815 lncRNA loci (Figure 11A). Florian Pauler found that these lncRNAs exhibit higher levels of chromatin marks indicating open chromatin (H3K4me3, H3K4me1 or H3K27ac) than random regions and that many of them have not been previously detected in other lncRNA annotations (Florian Pauler, manuscript in preparation). Only in

2014, two studies generated similar comprehensive lncRNA annotations: Necsulea et al. annotated ~9,000 lncRNA transcripts from mostly non-strand-specific datasets derived from seven mouse tissues (Necsulea et al., 2014) and Werber et al. annotated 1,403 lncRNA loci in six embryonic tissues (Werber et al., 2014). However, to my knowledge, no study annotated mouse lncRNAs from strand-specific RNA-seq data from eleven tissues.

The rat lncRNA annotation was also generated by Florian Pauler (see chapter 3.10.1) for the purpose to analyze the conservation of RNA biology features of mouse lncRNAs. Rat (*Rattus norvegicus*) and mouse (*Mus musculus*) diverged approximately 13-19 million years ago (Kutter et al., 2012) and therefore offer a perfect model to study the evolution of lncRNA biology in closely related species. Published rat lncRNA annotations are scarce, one study assembled rat lncRNAs from liver (Kutter et al., 2012) and another one more recently from six tissues (Wang et al., 2014a). The comprehensive rat lncRNA annotation used in this study contains 46,041 lncRNA transcripts in 9,921 lncRNA loci (Figure 11B).

5.3.2 LncRNAs are tissue specifically expressed and developmentally regulated

The comprehensive lncRNA annotation assembled from eleven mouse and rat tissues used in this study allowed the detailed study of lncRNA expression variation. In humans, a previous study assembled lncRNAs and investigated their expression levels across 19 cell types and found that ~78% of lncRNAs compared to only ~19% of mRNAs are tissue specifically expressed, irrespective of their low expression levels (Cabili et al., 2011). In mouse, a study investigated lncRNA expression across 30 primary cell types, however, their lncRNA catalog was only assembled from erythroblasts (Ter119+ cells) and their progenitors (Ter119- cells) (Alvarez-Dominguez et al., 2013). In this study, instead, I investigated mouse and rat lncRNA expression levels in the eleven tissues that have been used to assemble them (see chapter 3.10 for details), thereby ensuring that the lncRNA assembly includes many tissue types. I find that ~75% of mouse lncRNAs and ~78% of rat lncRNAs are tissue specifically expressed, compared to 36% of mouse mRNAs and 32% of rat mRNAs (Figure 14). Numerous reviews infer functionality from this extraordinary tissue specific expression of lncRNAs, however, as chromatin states and the accessibility of underlying cryptic promoters may also have tissue-specific distributions, lncRNA transcription may be a consequence of open chromatin or gene expression (Ulitsky and Bartel, 2013). From the presented data it can be conferred that only a minority of lncRNAs has potential housekeeping functions, as most of them are exclusively expressed in only one tissue. Whether all tissue-specific lncRNAs have a specific function in the respective tissue remains to be investigated.

I further investigated the developmental regulation of mouse lncRNAs in heart and liver and found that ~65% of lncRNAs (compared to ~41% of mRNAs) are differentially expressed in fetal and adult heart and ~84% of lncRNAs (compared to ~68% of mRNAs) in fetal and adult liver. While I conclude that the low expression of lncRNAs does not explain this difference

(see chapter 4.2.6), it raises the question whether these hundreds of lncRNAs have indeed functional roles in the development of the two organs. Numerous lncRNAs have been shown to be differentially regulated during developmental processes as diverse as retinal development (Blackshaw et al., 2004), ES cell differentiation (Guttman et al., 2011), erythrocyte maturation (Alvarez-Dominguez et al., 2013), cardiac development (Wamstad et al., 2012), muscle development (Cesana et al., 2011) in mouse and embryonic development in zebrafish (Pauli et al., 2012; Ulitsky et al., 2011). The identification of hundreds of developmentally regulated lncRNAs in RNA-seq studies is simple whereas functional validation of single candidates is cumbersome and requires sophisticated genetic tests (Bassett et al., 2014). Few developmentally regulated lncRNAs interact with cell-type specific transcription factors (Alvarez-Dominguez et al., 2013) or regulate microRNAs that in turn regulate transcription factors (Cesana et al., 2011). Indeed, it has been shown that lncRNAs are preferentially expressed in large gene deserts flanking transcription factor genes, most of which are implicated in embryonic development (Guttman et al., 2009; Pauli et al., 2012; Ulitsky and Bartel, 2013). It is unclear whether they are all implicated in regulating transcription factor expression *in cis* by remodeling the local chromatin landscape or whether many of them are co-regulated with neighboring transcription factor genes, as has been observed for *Six3* and *Six3os* (Rapicavoli et al., 2011). Future work is needed to estimate the integration of lncRNAs into developmental regulatory networks and define how exactly lncRNAs contribute to development and differentiation.

5.3.3 Currently used lncRNA subclasses are not distinguishable by genomic transcript features

The mouse lncRNA annotation was classified into the four widely used lncRNA subclasses intergenic, antisense, bidirectional and enhancer lncRNAs (see chapter 3.10.4 and Figure 11 for details). In mouse and rat, intergenic lncRNAs form by far the largest subclass, followed by antisense, bidirectional and enhancer lncRNAs with roughly equal shares (Figure 11). I calculated six genomic transcript features for each mouse and rat lncRNA locus: exon size, cDNA size, locus size, average exon length, exon/intron ratio and percent repeat content. I interrogated genomic transcript features of the lncRNA subclasses to answer the question whether any of these features is able to distinguish the classes. The hypothesis is, that the more discriminating features are found, the more significance these classes have and the more likely they are to employ different functions. When I compared the genomic transcript features of the four classes I find no apparent differences (Figure 12A). The only dissimilarity I noted is that bidirectional lncRNAs tend to have smaller loci but longer exons and therefore they have a higher exon/intron ratio. But beside that, lncRNAs of all four subclasses are overall very similar and can not be distinguished by any genomic transcript feature. While intergenic lncRNAs represent a “catch-all” class and therefore include a huge variety of lncRNAs (Ulitsky and Bartel, 2013), enhancer lncRNAs were described to be short and unspliced. However, it is surprising to find that enhancer lncRNAs are not different from

intergenic lncRNAs in their genomic transcript features. The median steady-state RNA levels of the four mouse lncRNA classes are also similar in mESC and MEF, with the maximum difference of ~2-fold between intergenic and bidirectional lncRNAs (Figure 12C). These results indicate that current lncRNA classes are, beside their “geographic” position in the genome, arbitrarily classified, which could hinder large-scale functional studies and dampen lncRNA research progression. A more sophisticated approach is needed to classify lncRNAs by RNA biology, structural aspects, protein-binding motifs and other features to enable extrapolation of functions from similar lncRNAs.

5.4 Most lncRNAs have an unusual RNA biology

5.4.1 LncRNAs are inefficiently exported to the cytoplasm

I have analyzed the RNA export of lncRNAs and mRNAs genome-wide to investigate differences within lncRNAs and between lncRNAs and mRNAs (Figure 18). The finding that lncRNAs are significantly less exported to the cytoplasm than mRNAs is not surprising, as mRNAs need to be exported to become translated in the cytoplasm while lncRNAs can function in both the cytoplasm and the nucleus. The ENCODE consortium assayed the cellular localization of human lncRNAs and mRNAs by sequencing nuclear and cytoplasmic RNA fractions and calculating nuclear/cytoplasmic RPKM expression ratios (Derrien et al., 2012). They find that lncRNAs are more enriched in the nucleus than mRNAs in five out of six human cell lines. The median RNA export of human lncRNAs in Derrien et al. is ~40% and the median export of mRNAs ~60%, which is well comparable with the results I obtained for mouse lncRNAs (Figure 18). Derrien et al. used a statistical test that revealed that 17% of lncRNAs are enriched in the nucleus and 4% in the cytoplasm. Interestingly, in Normal Human Epidermal Keratinocytes (NHEK) there is no difference in the RNA export between lncRNAs and mRNAs, indicating that cell type specific differences in lncRNA export exist. The Pearson correlation of RNA export of 98 lncRNAs being expressed in all six cell lines was between 0.5 and 0.9 for all pairwise comparisons, which is similar to the correlation of 1,249 lncRNAs between mESC and primary MEF in this study (r=0.55, Figure 18). Derrien et al. find *MEG3* and *XIST* to be the most nuclear enriched lncRNAs, both of which I also find almost exclusively in the nucleus (Figure 25C). Another study claims that the majority of human lncRNAs is essentially exported to the cytoplasm and bound by ribosomes (van Heesch et al., 2014). However, this study only generated 18 million 40bp reads for the nuclear fraction and therefore may have missed many nuclear lncRNAs. This study also applies a very high expression cutoff of 2,500 reads per transcript to call transcripts expressed, so only a few highly expressed lncRNAs (n=152) were examined. Furthermore, it was pointed out that “the most common misperception of lncRNAs is that they are predominantly localized in the nucleus” (Ulitsky and Bartel, 2013). It was argued that a relative nuclear enrichment of lncRNAs compared to mRNAs does not mean that the absolute number of lncRNAs is also higher in the nucleus. The reason is that polyadenylated RNAs are primarily localized in the

cytoplasm and are therefore not equally distributed between the nucleus and the cytoplasm. While ENOCDE used polyA+ RNA-seq data to calculate nuclear and cytoplasmic RPKM of their six cell lines, I use total non-ribosomal (Ribo-Zero) RNA-seq data to calculate RNA export. Although this may not abolish all biases, at least it reduces the bias seen with polyA+ RNA-seq data. A bias that is certainly introduced in my RNA export data is that I (as well as ENCODE) directly compare nuclear and cytoplasmic RPKM. As RPKM does not measure absolute abundances but rather is a relative measure of abundance in the respective sample, I may introduce a bias due to the unequal RNA content and RNA composition of the nucleus and the cytoplasm. In order to correct for this bias I would need to correct by the RPKMs of a set of transcripts that has perfectly equal absolute levels of RNA in the nucleus and the cytoplasm but this is not available. Another bias I may introduce is that I calculate RNA export from the exon model and disregard the intronic sequence of inefficiently spliced transcripts, which often contributes significantly to a unspliced transcript's length. This bias might lead to an overestimation of the export efficiency of lncRNAs, as many of them are inefficiently spliced. Whether nuclear lncRNAs evade nuclear export because they lack an export signal or miss mRNA-like processing such as splicing, capping and polyadenylation remains enigmatic. However, most lncRNAs are capped, polyadenylated and at least to some extent spliced (Moran et al., 2012). A recent study found that the nuclear retention of the human lncRNA *BORG* is due to an RNA motif (Zhang et al., 2014). Mutation of this pentamer motif leads to nuclear export of *BORG* and insertion of the pentamer into a cytoplasmic lncRNA results in subsequent nuclear localization. As this motif is present in many human lncRNAs and correlates with nuclear localization, it may serve as a general nuclear localization signal for lncRNAs.

5.4.2 LncRNAs exhibit low RNA stability

I have assayed the RNA stability of lncRNAs and mRNAs genome-wide to investigate differences within lncRNAs and between lncRNAs and mRNAs (Figure 21). For this experiment I had to consider several issues: (i) selecting the best method to assay RNA stability, (ii) choosing the right time point to measure RNA half life in order to achieve a good separation of stable and unstable RNAs without getting adverse effects of Actinomycin D on cell physiology, (iii) incorporating untreated and vehicle control treatments and technical as well as biological replicates and (iv) establishing an analysis pipeline that normalizes all RPKMs to a basket of housekeeping genes and takes into account biological replicates, control treatments and RPKM errors.

RNA stability can be measured by treating cells with a transcription inhibitor or by chemical labeling followed by pulse-chase analysis. The two most widely used transcription inhibitors are the anti-cancer drug Actinomycin D and the toxin α-Amanitin (Bensaude, 2011). Both drugs inhibit RNA synthesis as Actinomycin D intercalates into DNA and thereby inhibits RNA polymerase I/II elongation whereas α-Amanitin directly degrades RNA polymerase II. I have

decided to prefer Actinomycin D over α-Amanitin for the reason that it permeates cells faster and therefore is more appropriate for short term (1-4h) RNA stability studies. Actinomycin D can be dissolved in EtOH or DMSO and I have used EtOH as a vehicle because it was shown to have less influence on cells than DMSO (Wolfgang Allhoff, Diploma thesis, University of Vienna, 2012). Other studies use chemical labeling to investigate RNA stability. Therefore, newly synthesized RNA is labeled by uridine analogs such as 5'-bromo-uridine (BrU) (Tani et al., 2012) or 4-thiouridine (4sU) (Rabani et al., 2011), followed by a washout of the labeling agent and analysis of chemically labeled RNA degradation in a time course. These inhibitor-free techniques have less influence on cell physiology than transcription inhibitors, however, inherent limitations are the incomplete removal of the intracellular nucleotide pool leading to low sensitivity and the reuse of uridine analogs generated by RNA decay after the washout, which leads to complications during data analysis (Ross, 1995).

Choosing the optimal time points for Actinomycin D treatments is a balance between a long treatment on the one hand, leading to a good separation of stable and unstable RNAs, and a short treatment on the other hand, preventing negative effects of the inhibitor on cell physiology. Whereas some studies treat cells with Actinomycin D in a time course for up to 32h (Clark et al., 2012), I decided to focus on two time points only. The rational was to treat cells for 1h with Actinomycin D to remove unstable introns and for 4h to degrade unstable RNAs while keeping stable ones. These short time points therefore should allow a categorization into stable and unstable RNAs rather than determining exact RNA half-lives, for which more time points would have been necessary. In the course of bioinformatic analysis of 1h and 4h Actinomycin D samples, I found that 1h treatment did not have an effect on many lncRNAs and was therefore not suitable to classify stable and unstable lncRNAs. However, when I analyzed the 4h Actinomycin D samples I found that stable and unstable lncRNAs were well separated and that ~20% of mRNAs are unstable and fall below 30% of remaining abundance (Figure 19C). The RNA stability of individual lncRNA examples will be discussed in chapter 5.5.2. Longer Actinomycin D treatments would have led to a complete degradation of unstable RNAs and therefore to a reduced resolution of RNA stability. Additionally, I wanted to avoid adverse effects on cell physiology such as altered RNA processing and cell death.

When I first treated cells with Actinomycin D to investigate the RNA stability of a single human lncRNA for another project, I found that qPCR analysis results were very variable between replicates and across time points (unpublished data). For the genome-wide analysis of RNA stability, I therefore decided to conduct a well-controlled experiment and include technical as well as biological replicates. The minimal RNA stability sample set consisted of five samples: the untreated control, 1h and 4h Actinomycin D treatments and 1h and 4h EtOH control treatments. I decided to conduct this experiment in technical replicates, meaning that this sample set was simultaneously produced twice using cells of the same passage. Additionally, I repeated the whole experiment within a week with cells of a different passage to get biological replicates. All together, I produced 4x5 samples for each of the four cell types and

analyzed levels of the control mRNAs *Myc* (unstable) and *Gapdh* (stable) in all of them by qPCR. As *Myc* degradation indicated that the treatments were effective in all replicates (Figure 19B), I pooled the technical replicates before RNA-seq. This left me with 2x5 samples for each cell type, for all of which I prepared RNA-seq libraries. The mRNA stability was very similar between biological replicates (r>0.9 for all comparisons), therefore I pooled reads of biological replicates before alignment to increase read coverage and reduce RPKM errors.

The first step in the analysis pipeline after read alignment and RPKM calculation is the normalization of RPKM to a basket of ten stable housekeeping genes. This step is necessary to account for the effect that the RPKM is relatively increased in the Actinomycin D treated samples because the RNA pool decreases (Clark et al., 2012; Sharova et al., 2009). Next, I calculated RNA stability by normalizing the 4h Actinomycin D RPKM to the untreated control RPKM and the 4h EtOH RPKM. The EtOH control treatment controls for effects the EtOH has on gene expression and cell physiology. I also observed that some RNAs are increased upon Actinomycin D treatment and have an RNA stability of >100%. This might be due to specific effects of Actinomycin D that are known for some transcripts (Cassé et al., 1999), due to transcription triggered by a p53 response (Ljungman et al., 1999) or due to artifacts created by three-fold normalizations and RPKM variation. As there is no way to discriminate between these three possibilities, I put their RNA stability to 100% and considered them as stable. I also may introduce a bias, similar to the RNA export datasets, as I calculate RNA stability from the exon model and disregard the intronic sequence of inefficiently spliced transcripts, which can contribute significantly to a transcript's length. This bias leads most probably to an overestimation of the RNA stability of lncRNAs, as many of them are inefficiently spliced.

A large-scale RNA stability study using Actinomycin D and microarrays determined the average half life in mouse Neuro-2a cells to be 4.8h for lncRNAs and 7.7h for mRNAs (Clark et al., 2012). These RNA half-lives can not be directly compared to the percent RNA stability values that I calculated, however, after transforming them using the decay law formula (http://www.calculator.net/half-life-calculator.html) they correspond to an RNA stability of ~56.12% for lncRNAs and ~69.76% for mRNAs. This fits very well to the RNA stability I calculated for lncRNAs (mESC 44.41%, MEF 58.57%) and mRNAs (mESC 69.01%, MEF 73.00%) (Figure 21A). A second study using BrU-labeling of HeLa cells followed by RNA-seq (BRIC-seq) determined the average half life for lncRNAs to be ~7.0h and for mRNAs to be ~6.9h (Tani et al., 2012). Although a class of short-lived lncRNAs (half life <4h) was defined in this study, it is unexpected that the average half life of lncRNAs is higher than for mRNAs. However, Tani et al. only analyzed ~5 million 36p reads in total and state themselves that they are biased towards highly expressed lncRNAs, which I show to be more stable (Figure 29). Overall, the comparison of genome-wide RNA stability data is difficult, as can be seen in multiple genome-wide mRNA datasets (Tani and Akimitsu, 2012). Three different studies using Actinomycin D and microarrays for mouse ES cells, Neuro-2a cells and NIH3T3 fibroblasts find median mRNA half-lives from 4.9h to 7.1h, indicating that either cell-type specific differences exist or the method varies significantly from lab to lab.

The most interesting question about RNA stability is definitely why are so many lncRNAs unstable and how are they degraded? Are they actively targeted and degraded? Or are they degraded by nonsense-mediated decay (NMD) because they lack (m)RNA processing and thereby stabilization? It has been suggested that RNA modifications could be implicated in the regulation of RNA stability (Pan, 2013). Indeed, N6-methyl-adenosine is the most prevalent RNA modification and is selectively bound by the human YTH domain family 2 (YTHDF2) protein to regulate degradation of mRNAs and a few lncRNAs (Wang et al., 2014b). If lncRNAs had the same stability as mRNAs they would constitute an even larger portion of a cell's RNA mass, which could negatively influence several cellular functions. DNA replication could be hindered due to the prevalent formation of RNA-DNA hybrids, RNA-binding proteins could be sequestered by abundant lncRNAs and be refrained from their actual targets, expression of mRNAs could be skewed by potent regulatory lncRNAs and finally, accumulation of random and potentially highly expressed RNA fragments could reduce the overall efficiency and fitness of a cell (Houseley and Tollervey, 2009). The variability of overall lncRNA stability seen in different cell types (Figure 21) may be explainable by the fact that RNA stabilizing and destabilizing proteins, miRNAs and components of the exosome and NMD machinery show tissue-specific expression and activities.

5.4.3 LncRNAs are inefficiently spliced

Lastly, I also analyzed the RNA splicing efficiency of lncRNAs and mRNAs genome-wide to investigate differences within lncRNAs and between lncRNAs and mRNAs (Figure 23). RNA splicing efficiency of lncRNAs is probably the least understood RNA biology feature of the three I investigated. The main reason is that most studies use polyA+ RNA-seq data to annotate and analyze lncRNAs and therefore bias themselves towards the fully processed and spliced isoforms. In order to completely understand the complex transcription and processing of RNAs, it was argued that RNA-seq of total non-ribosomal RNA gives unique insights as also nascent and immature transcripts can be detected (Ameur et al., 2011). It is known that many lncRNAs also have an unspliced isoform with different RNA biology features and maybe even different function. In the case of *Airn*, the minor spliced isoform is rather stable and exported to the cytoplasm whereas the major unspliced isoform is retained in the nucleus and unstable (Seidl et al., 2006). The unspliced *Airn* isoform is a byproduct of transcription, which has been shown to be the functional process in repressing *Igf2r* expression (Latos et al., 2012). The minor spliced *Airn* isoforms exported to the cytoplasm lack any silencing function. I analyzed splicing efficiencies of steady-state RNAs for each lncRNA and mRNA locus using total non-ribosomal RNA-seq data. A 100% splicing efficiency therefore indicates that no unspliced isoform exists, whereas a 50% splicing efficiency hints towards equal steady-state levels of a spliced and an unspliced isoform. Accordingly, the lower the splicing efficiency gets, the higher the abundance of the unspliced isoform becomes relative to the spliced isoform.

In chapter 4.3.5 I mentioned that I tried four different strategies to calculate RNA splicing from Ribo-Zero RNA-seq data. Figure 31 shows a schematic of these strategies of which only the fourth approach was satisfying in terms of accuracy and number of lncRNAs that passed the analysis. For approach #1, Florian Pauler set up a pipeline to define splicing efficiency by calculating a ratio of the exonic and intronic read pileups (+/- 10bp from the junction) and averaging all junction ratios per transcript and locus (Figure 31A, #1). This pipeline was, however, not satisfying as we detected major differences within junctions of the same transcript and within biological replicates during the quality check of well-studied control lncRNAs. For approach #2, Florian Pauler set up a pipeline that, for each splice junction, counts the number of reads that span this splice junction and the number of reads that splice away at this splice junction (Figure 31A, #2). A splicing ratio was calculated from these numbers for each junction and these ratios again averaged over transcripts and loci. This pipeline gave good splicing estimations for well-studied control lncRNAs, however, the number of lncRNAs that passed this analysis was too low because we demanded at least ten reads per splice junction to be accurate. For approach #3, I decided to calculate RPKM for each exon and intron of each transcript, average all exon RPKM and all intron RPKM and calculate an exon/intron ratio (Figure 31A, #3). This strategy also yielded accurate splicing data for well-studied control lncRNAs, however, I anticipated problems with the different length and variable repeat content of exons and introns in mRNAs and lncRNAs, which would create biases that may have invalidated the splicing analysis of many transcripts. For approach #4, I combined the advantages of the above-mentioned approaches and calculated for each junction an RPKM for a 45bp exonic region and an RPKM for a 45bp intronic region, both 5bp away from the junction (Figure 31A, #4), as indicated in chapter 4.3.5. The three main advantages of this approach are that length biases (as in approach #3) are avoided, that the RPKM of a 45bp region is more robust than calculating read pileups in a 10bp region or counting reads in a 5bp region (as in approach #1 and #2) and lastly that some ambiguity for inexact splice junction annotation was allowed (+/- 5bp).

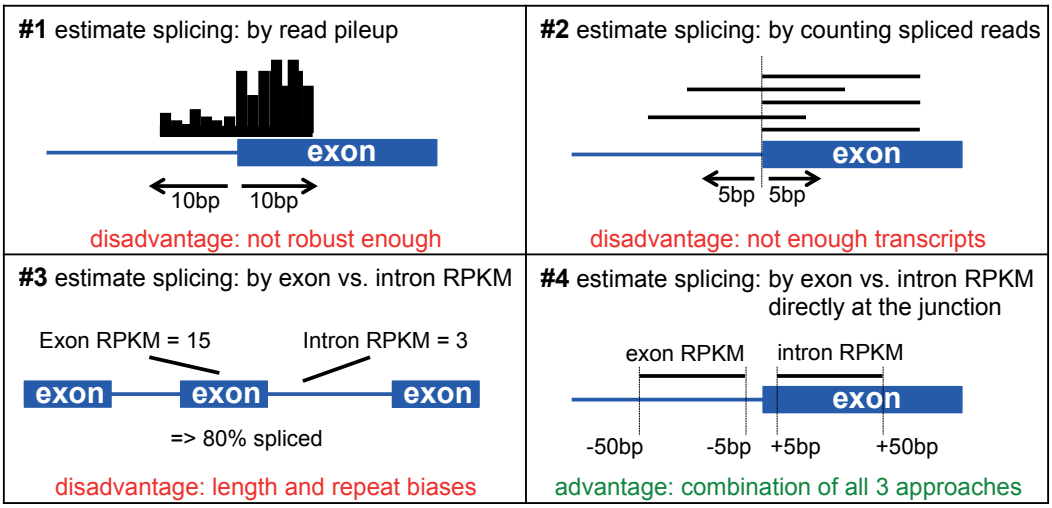


Figure 31: Four strategies to calculate RNA splicing efficiency
Overview of approaches #1 to #4 that I tested to calculate RNA splicing efficiencies, of which only approach #4 yielded accurate and sufficient data (see text for details).

I find that lncRNAs are as a group inefficiently spliced in mESC and MEF compared to mRNAs (Figure 23). In an attempt to benchmark my results, I first analyzed the splicing efficiencies of mRNAs, which are known to be efficiently spliced. As expected, >75% of mRNAs are >95% spliced in both cell types, arguing that my pipeline detects well spliced RNAs indeed as efficiently spliced. Additionally, the spliced lncRNA *H19* is also ~99% spliced in my pipeline. The analysis whether the pipeline also correctly detects inefficiently spliced RNAs is more difficult, as no homogenous set of lowly spliced RNAs is known. I analyzed whether known inefficiently spliced lncRNAs such as *Malat1*, *Neat1*, *Airn* and *Kcnq1ot1* are indeed classified as such (Hutchinson et al., 2007; Redrup et al., 2009; Seidl et al., 2006). And as expected, all of the four lncRNAs have splicing values of ~10% to ~40% and can therefore be considered as inefficiently spliced (Figure 25). This shows that RNAs are correctly defined as efficiently or inefficiently spliced. The question that arises is why lncRNAs are incompletely spliced? Is it because they do not need to be spliced in order to be functional? Is it their local chromatin environment that hinders efficient splicing or rather the lncRNA transcript that is unable to attract certain splicing cofactors?

RNA precursors are in general co-transcriptionally and post-transcriptionally spliced, a process that is very efficient for mRNAs and less efficient for lncRNAs (Figure 23). Two studies investigated co-transcriptional splicing of nascent transcripts, however, genome-wide studies investigating splicing of steady-state RNAs are, to my knowledge, lacking so far. The first study investigated co-transcriptional splicing by RNA-seq of chromatin-associated RNAs as well as polyA+ and polyA- fractions of the nucleus and cytoplasm (Tilgner et al., 2012). They find that cytoplasmic RNAs, no matter if polyA+ and polyA-, are almost completely spliced (median splicing = 100%), indicating that RNAs have been spliced before nuclear export. Chromatin-associated RNAs that are still in the process of being transcribed are, in contrast, not completely spliced (median splicing = 75.0%), indicating that splicing has started before transcription is completed and therefore occurs simultaneously with transcription. When chromatin-associated RNAs are split into lncRNAs and mRNAs, Tilgner et al. find that lncRNAs peak at ~15% co-transcriptional splicing and mRNAs at ~65%, arguing that co-transcriptional splicing is significantly more efficient for mRNAs. Whereas splicing is completed for most mRNAs post-transcriptionally, many lncRNAs evade splicing post-transcriptional completely and unspliced isoforms remain (Tilgner et al., 2012). It has been speculated that co-transcriptional or post-transcriptional splicing could be affecting RNA binding proteins that in turn regulate RNA stability and RNA localization (Bentley, 2014). Co-transcriptional splicing may offer a direct opportunity for the chromatin environment to influence splicing, a hypothesis that could explain why lncRNAs as a group partially evade splicing. A slight positive correlation has been observed between expression status and co-

transcriptional splicing levels, which could also explain why lowly expressed lncRNAs as a group are more often inefficiently spliced than mRNAs (Tilgner et al., 2012). A second study investigated co-transcriptional splicing in human adult and fetal brain and liver samples and found that mRNAs as well as lncRNAs are co-transcriptionally spliced. They find clear trends that brain samples are less well spliced than liver samples and that in both tissues fetal samples are less well spliced than adult samples (Ameur et al., 2011). This argues for tissue-specific and developmental regulation of RNA splicing and could explain the differences I see for splicing in ES cells and embryonic fibroblasts in mouse and rat (Figure 23A). Their experiments also confirm that intronic RNA-seq reads indeed stem from unprocessed transcripts rather than from already excised introns, a result that supports my interpretation that a splicing value of e.g. 50% means that the spliced and the unspliced isoform are equally abundant, rather than that the excised introns of this RNA are more stable and that is why I call it inefficiently spliced.

In conclusion, evidence that can explain why most lncRNAs are inefficiently spliced is lacking. Rather than a different chromatin environment or an altered binding of splicing cofactors, it could also be that mRNA splicing has been selected for during evolution and as lncRNAs are little conserved their splicing efficiency either eroded or did never evolve in the first place. Highly spliced lncRNAs could then have been evolved from former mRNAs that just lost their coding potential. For *Xist*, a spliced lncRNA, it has indeed been shown that it evolved from a protein-coding gene that pseudogenized (Duret et al., 2006).

5.4.4 Currently used lncRNA subclasses exhibit strikingly similar RNA biology features

Genome-wide lncRNA mapping has led to the identification of four major lncRNA subclasses in the mouse, each being based on the “geographic” position relative to mRNAs or genetic elements (see chapter 2.3). The large group of intergenic lncRNAs is crudely defined by being distant to mRNA genes and is considered to be a “catch-all” class (Ulitsky and Bartel, 2013). Antisense lncRNAs overlapping mRNAs on the antisense strand and bidirectional lncRNAs being expressed from bidirectional mRNA promoters form separate subclasses, whose lncRNAs are, in contrast to intergenic lncRNAs, located near mRNA genes. The class of enhancer lncRNAs is formed by transcripts that arise from expressed enhancers, most of which reside in the intergenic space. Functions have only been determined for single members of these classes, however, they were projected to represent the function of the whole class (Guttman et al., 2011; Kim et al., 2010; Ørom et al., 2010). The RNA biology of these subclasses has never been thoroughly investigated and compared among themselves, however, certain RNA biology features were attributed to some of these classes (see chapter 2.3). I find that, in fact, none of the four lncRNAs subclasses can be distinguished by any of the three RNA biology features (Figure 24). This is insofar surprising, as it invalidates at least one out of two hypotheses. One hypothesis is that lncRNA biology is indicative for function

(Guenzl and Barlow, 2012; Pelechano and Steinmetz, 2013). According to this theory, each lncRNA has a certain RNA biology that determines or is a result of its mechanism of action. The other hypothesis is that each of the four subclasses combines transcripts with similar functions. Now, if both hypotheses were true, this would mean that lncRNA subclasses are distinguishable from each other due to their RNA biology features. However, the data indicate that the RNA biology features of the subclasses are very similar and that they can not be distinguished by any RNA biology feature (Figure 24). This points to the interpretation that either the RNA biology does not play a role for different functions or that the lncRNAs in the four classes have similar functions. However, at least the latter seems unlikely due to the described functional diversity of lncRNAs (Mercer and Mattick, 2013). In summary, these analyses pose the following question: Are current lncRNA classes arbitrarily chosen just by their position relative to mRNAs and no functional differences exist between these classes? Maybe there is a better way to find classes of lncRNAs that have similar functions: a classification of lncRNAs by their RNA biology features.

5.5 Clustering of lncRNAs by their RNA biology features

5.5.1 Clustering of lncRNAs by their RNA biology

The scientific community defined several classes of lncRNAs by grouping together transcripts that share certain features with the ultimate goal to extrapolate functions determined for single candidates to all members of the respective class. Each of the four lncRNA classes was annotated in separate studies and therefore features that were used to justify separate classes are hard to compare. Incorporating the results from this study, it seems that the only feature that is different for these lncRNA classes is their “geographic” position in the genome while the RNA biology features and genomic transcript features are very similar (Figure 12 & Figure 24). In search for a more sophisticated way of classifying lncRNAs, I classified mouse lncRNAs based on their three RNA biology features export, stability and splicing. I refrained from using hard cut-offs to split lncRNAs into arbitrary classes, instead, I used the popular k-means clustering algorithm to cluster lncRNAs with similar RNA biology features (Figure 25C, D, method described in chapter 3.14). Statistically, clusters might be defined by data points densely located around the centers and some loose data points in the cluster periphery, a distribution I do not really see in the RNA biology clusters (Figure 25C, D). However, I am confident that using the k-means clustering algorithm offers a very useful way of classifying lncRNAs by their RNA biology. As a control, I also applied the Partitioning Around Medoids (PAM) clustering algorithm (Kaufman and Rousseeuw, 1987) and obtained very similar results, only very few lncRNAs at the periphery were located in other clusters and the position of the cluster centers were very close to those in the k-means clustering. The question that automatically arises when dealing with k-means clustering is how the number of clusters is chosen beforehand. In an empirical approach, I clustered lncRNAs using four to eight clusters and chose six clusters as the ideal number based on the best representation of RNA biology

diversity among lncRNAs (data not shown). Four or five clusters led to a fusion of transcripts with widely different RNA biology in the same clusters and seven or eight clusters led to fragmented clusters with diminished differences in their RNA biology features.

The distribution of clusters overall is as expected, the majority of mRNAs and ~50% of lncRNAs are efficiently spliced and form the first three clusters. These clusters mainly differ in their RNA stability, which is gradually decreasing from cluster 1 to cluster 3 while the RNA export is similar for cluster 1 and 3 and reduced in cluster 2. LncRNAs in cluster 1 to 3 can be considered as mRNA-like as their RNA biology is very similar to mRNAs. The remaining ~50% of lncRNAs are less efficiently spliced and form the second three clusters, their RNA biology is clearly distinct from mRNAs which makes them non-mRNA-like. Cluster 1 and cluster 6 form the two extremes: while lncRNAs in cluster 1 are stable, fully spliced and exported, lncRNAs in cluster 6 are inefficiently spliced, unstable and inefficiently exported. LncRNAs that are supposed to function by their transcript (e.g. *Xist*, *H19*, *Rian*, *Meg3* and *Braveheart*) cluster together with mRNAs while the two imprinted lncRNAs *Kcnq1ot1* and *Nespas* that are inefficiently spliced and hypothesized to function by the act of transcription are non-mRNA-like and located in clusters 6 and 4, respectively. Unfortunately, many interesting lncRNAs introduced in chapter 2.4 are not expressed in mESC or MEF or are not confidently detected in the RNA stability or RNA export datasets. In line with the finding that current positionally defined lncRNAs classes are not distinguished by their RNA biology, I also find that none of them is enriched in any of the six RNA biology clusters (Figure 25G).

5.5.2 The RNA biology of well studied lncRNA examples

In this study, I generated a catalog of three RNA biology features for each lncRNA and mRNA in mESC and MEF. In order to judge whether this catalog is accurate, I compared published data of eight well studied lncRNAs with my data. *Malat1* is known to be an unspliced lncRNA that is stable for ~10h and nuclear retained (Hutchinson et al., 2007; Tani et al., 2010; Tripathi et al., 2010). Indeed, in my data *Malat1* belongs to cluster 5 and is only ~10% spliced, 100% stable after 4h and only ~10% exported. *Xist* is reported to be efficiently spliced and nuclear localized (Pontier and Gribnau, 2011). Its half life was determined to be between 3h and 6h, depending on the investigated cell type (Sun et al., 2006). In my data, *Xist* is indeed well spliced, nuclear retained and ~50% stable (corresponds to ~4h half life) and therefore belongs to cluster 2. The imprinted lncRNA *Kcnq1ot1* has been reported to be unspliced, nuclear localized and rather unstable with a half life of only ~2h in neuroblastoma cells (Clark et al., 2012; Redrup et al., 2009). The clustering puts *Kcnq1ot1* into cluster 6, which contains lncRNAs that are inefficiently spliced, lowly exported and unstable. *H19* has long been the prototype of an mRNA-like lncRNA, as it is well spliced, exported and rather stable (half life in C2C12 myoblastic cells >4h) (Berteaux et al., 2005; Keniry et al., 2012; Milligan et al., 2000). I find that *H19* indeed clusters with mRNAs in cluster 1 as it is ~99% spliced, very stable and efficiently exported. *Rian* and *Meg3* are both expressed upstream from the imprinted *Dlk1*

gene and are alternatively spliced and predominantly localized in the nucleus (Hatada et al., 2001). The RNA biology is indeed very similar for the two lncRNAs, they are both efficiently spliced, nuclear localized and therefore in cluster 2. Not too much is known about the RNA biology of the *trans*-acting *Braveheart* lncRNA, except that it is spliced and that ~33% reside in the nucleus (Klattenhoff et al., 2013). I find *Braveheart* in cluster 1 as it is well spliced, ~60% exported and quite stable in MEF. The last example is *Nespas*, an imprinted lncRNA that expresses an unspliced transcript as well as alternatively spliced isoforms (Williamson et al., 2002). It is functioning in the nucleus and was suggested to be rather unstable (Ball et al., 2001; Yang and Kuroda, 2007). In MEF, *Nespas* locates to cluster 4 and is as expected nuclear localized, unstable and shows ~60% splicing efficiency. The overall RNA biology of well-studied lncRNAs is therefore well reflected in my dataset, however, biological differences between investigated cell types may occur. This catalog of three RNA biology features in two cell types of the mouse and the rat will be a valuable tool for the lncRNA research community to get information about their favorite lncRNA candidate and to compare it to well studied related lncRNAs in order to extrapolate possible functions and mechanisms of action.

5.5.3 RNA biology of lncRNA clusters is evolutionary conserved

If the hypothesis were true that RNA biology is a prerequisite or a consequence for the particular function of a lncRNA, one would expect that lncRNAs have a similar RNA biology throughout various cell types and across species. The significance of RNA biology can indeed be judged by the observations that the majority of lncRNAs falls into the same clusters in the two cell types mESC and MEF and that the RNA biology is largely conserved between the mouse and the rat. Figure 26 shows that 51.43% of lncRNAs and 68.33% of mRNAs have a very similar RNA biology in mESC and in MEF, as they fall into the same RNA biology clusters. Two reasons might explain why lncRNAs are more likely to switch their RNA biology cluster between two cell types. First, 97.8% of mRNAs fall into clusters 1 to 3 and therefore they also mostly switch between these three clusters, while lncRNAs switch between all six clusters due to their great diversity of RNA biology. Second, the RNA biology of lncRNAs is more diverse between mESC and MEF compared to mRNAs, as indicated in Figure 18A, Figure 21A and Figure 23A. I further investigated the top two fractions of lncRNAs and mRNAs that switched from one particular cluster to another one (Figure 26) and found that a certain pattern emerged. It seems that RNA stability is the main driver for lncRNAs and mRNAs to switch their cluster, indicating that cell type specific mechanisms might regulate RNA stability and thereby lncRNA abundance (Rabani et al., 2011) while the other two RNA biology features largely remain stable.

While lncRNA conservation has recently been described to depend on the four dimensions sequence, structure, function, and expression from syntenic loci (Diederichs, 2014), the conservation of RNA biology as a prerequisite for function has been underestimated so far. In Figure 27 I show that the RNA biology of syntenically transcribed lncRNAs in each cluster is

largely conserved between the mouse and the rat. While RNA export shows the best conservation (r=0.71 for lncRNAs, r=0.86 for mRNAs), the conservation of RNA stability is less pronounced (r=0.48 for lncRNAs, r=0.79 for mRNAs). The conservation of RNA splicing, however, is more difficult to interpret. The overall correlation is low (r=0.42 for lncRNAs, r=0.20 for mRNAs) but this is most likely attributable to the fact that median lncRNA splicing efficiencies vary between mouse and rat (Figure 23A). This becomes evident especially in the clusters 4 to 6, which are more efficiently spliced in rat than in mouse (Figure 27A). Albeit, the overall trend is very similar in both species, cluster 1 to 3 are well spliced and cluster 4 to 6 exhibit gradually decreased splicing efficiencies. It is interesting to see that a high splicing efficiency seems to be conserved between mouse and rat: a lncRNA or mRNA that is efficiently spliced in either of the two species does not show signs of inefficient splicing in the other species. This argues that the capability of a transcript to be spliced is not lost between mouse and rat. The very low correlation of mRNA splicing efficiencies (r=0.20) seems to be an artifact arising from the concentration of data points at the top, as mRNAs are very well spliced in both species. Taken together, from the data presented here it can be concluded that the RNA biology clusters are similar in mESC and MEF and seem to be conserved between mouse and rat. These findings underline the hypothesis that RNA biology is an important prerequisite for lncRNA function (Rabani et al., 2011).

5.5.4 RNA biology clusters exhibit variable genomic transcript features

While I have shown that current lncRNA subclasses are not distinguishable by their genomic transcript features and their abundance levels (Figure 12, discussed in chapter 5.3.3), I find that the six RNA biology clusters indeed exhibit differences in their genomic transcript features and their abundance levels (Figure 28). Especially lncRNAs in cluster 6 tend to have the longest cDNA sizes, the longest average exon sizes and the highest exon/intron ratio. Their abundance is also ~2x to ~8x lower than the abundance of lncRNAs in all other RNA biology clusters, which could be a reason why many inefficiently spliced lncRNAs have not been detected so far. Two reasons could explain why genomic transcript features are different for some of the six clusters. First, these features could be implicated in lncRNA function and the clustering could group together functionally similar lncRNAs. Second, the three RNA biology features export, stability and splicing and the genomic transcript features could, at least to some extent, be correlated. In the latter case, one or more of the genomic transcript features could predispose how efficiently a lncRNA is spliced or exported to the cytoplasm. RNA stability of lncRNAs might also to some extent be “coded” in their genomic transcript features, however, as discussed in the previous chapter RNA stability could provide a cell type specific mechanism to regulate abundance levels of lncRNAs .

5.5.5 RNA abundance levels correlate with RNA stability and RNA export

In order to further examine why some lncRNAs are mRNA-like and others are not, I investigated whether RNA abundance is correlated with RNA biology. Therefore, I split lncRNAs and mRNAs into four RPKM bins and plotted RNA stability, RNA export and RNA splicing for each bin (Figure 29). As discussed in chapter 4.4.5, RNA stability and RNA export of lncRNAs and mRNAs correlate well with RNA abundance. Highly abundant mRNAs or lncRNAs (e.g. with housekeeping functions) could indeed be more stable as they are key to a cell's life cycle. Alternatively, another explanation might be that stable mRNAs are simply more abundant because they are have longer half lives and reduced degradation rates compared to unstable lncRNAs that are degraded and exhibit decreased abundance levels. This correlation has been noted before for mRNAs, however, it was not found for lncRNAs in the same study which was based on expression arrays rather than RNA-seq (Clark et al., 2012). The explanation for the positive correlation of steady-state levels and export efficiency might be that RNA is transcribed in the nucleus and exported to the cytoplasm. Hence, highly expressed and thereby highly abundant mRNAs might accumulate in the cytoplasm rather than in the nucleus, thereby leading to the observed shift in RNA export efficiency. As lncRNAs are less exported, they also exhibit a less pronounced correlation between RNA abundance and RNA export. The ENCODE project assayed RNA export in six human cell lines, however, they did not comment on a correlation between export and RNA abundance levels (Derrien et al., 2012). In conclusion, the presented analyses also establish that the reduced RNA stability, RNA export and RNA splicing efficiency of lncRNAs compared to mRNAs are not just due to ~100x reduced expression levels of lncRNAs (see Figure 13), as corresponding RPKM bins (e.g. with a log2 RPKM between 2 and 4) also indicate RNA biology differences between mRNAs and lncRNAs.

5.5.6 Is RNA biology influenced by genomic transcript features?

In order to find patterns explaining why lncRNAs exhibit this extraordinary diversity of RNA biology features, I examined the relationship between RNA biology and genomic transcript features and I find that many of them often correlate with each other. Most notably, lncRNA stability is negatively correlated with cDNA size, average exon size and exon/intron ratio (Figure 30). LncRNA export is negatively correlated with exon count, cDNA size and locus size. Finally, lncRNA splicing is positively affected by exon count and negatively influenced by average exon size and exon/intron ratio. Most of these patterns are also evident in mRNAs, which could hint towards general principles regulating RNA molecules. While I discussed in chapter 5.4 how RNA stability, RNA export and RNA splicing could be specifically regulated by the cell and why lncRNAs are so different from mRNAs, it can be concluded from this data that the genomic transcript features could also have an important influence on RNA stability, RNA export and RNA splicing. The six genomic transcript features are encoded in the genome and could form a prerequisite for RNA biology, e.g. that a long lncRNA is much more

likely to be exported and to have cytoplasmic functions than a short one, or that a lncRNA with large exons is likely to be lowly spliced and like other lowly spliced lncRNAs in cluster 6 also unstable and not exported (see Figure 25E).

For mRNAs, some of these correlations have been noted before. Significant negative correlations between mRNA stability and cDNA size have been previously found in humans and Escherichia coli, but not in Saccharomyces cerevisiae or Bacillus subtilis (Feng and Niu, 2007). In this review, it was suggested that longer RNA molecules are more likely to suffer from endonucleolytic attacks by RNA endonucleases and mechanical damage. The negative correlation between RNA export and cDNA size has also previously been detected in humans (Solnestam et al., 2012). While the molecular mechanisms explaining this finding are still lacking, it was mentioned that the variable export efficiency of short and long mRNAs did not lead to changes in corresponding protein levels.

5.6 LncRNAs are promising drug targets to modulate gene expression

The emergence of lncRNAs as key regulators of mammalian genomes sparked great interest to pharmacologically target them to fight developmental defects and diseases such as cancer. Many of the currently available pharmacologically used molecules repress expression of genes or inhibit receptors, undesired gene products and fusion proteins. However, in many situations it would be desirable to activate expression of genes such as tumor suppressor genes, transcription factors, growth factors and deficient genes in genetic diseases (Wahlestedt, 2013). While it has proven difficult to pharmacologically increase gene expression, it might be easier to repress the repressor and thereby indirectly upregulate the desired gene. As many lncRNAs have been shown to directly and specifically repress single mRNA genes, a strategy to target them would lead to a de-repression of the mRNA genes. In the case of antisense lncRNAs (e.g. *BACE1-AS* and *Zeb2-as*) that form RNA duplexes with their sense partners, short antisense oligonucleotides (ASOs) could be used to block the formation of duplexes and the subsequent regulation of the mRNA. ASOs have successfully been used *in vivo* to specifically upregulate expression of *BDNF* by blocking its interaction with the repressing antisense lncRNA *BDNF-AS* (Modarresi et al., 2012). Two ASO drugs have already been approved by the Federal Drug Administration (FDA), one targets a key cytomegalovirus mRNA to prevent retinitis (Roehr, 1998) and the other one reduces apolipoprotein B mRNA levels to lower cholesterol levels in familial hypercholesterolemia patients (Athyros et al., 2008). This proves that antisense oligonucleotides could be used to target gene regulatory lncRNAs with the ultimate goal to influence mRNA expression. As an example, the Alzheimer's disease driving gene *BACE1* could be indirectly targeted by blocking its interaction with the stabilizing lncRNA *BACE1-AS*. Other lncRNAs such as the FSHD causing *BDE-T* and overexpressed *HOTAIR* in cancer might also represent promising therapeutic targets. However, of uttermost importance is the detailed understanding of the

mechanism of action and the RNA biology of lncRNAs. LncRNAs function by interacting with other RNA molecules, with DNA or with proteins and targeting lncRNAs therefore might need stratified approaches. Unstable or lowly expressed lncRNAs might be difficult to target and nuclear localized lncRNAs will require different drug delivery strategies than cytoplasmic lncRNAs. *Cis*-regulatory lncRNAs might have only one target while *trans*-acting lncRNAs influence hundreds of genes and could cause potent side effects. It will also be essential to know whether the RNA molecule is functional or the act of transcription, the latter of which will not be affected by targeting the RNA molecule. Targeting of lncRNAs that arise from open chromatin or keep chromatin open by constant transcription will likely have no effect on *cis*-regulated genes. In essence, a successful pharmacological targeting of lncRNAs depends on the complete understanding of the molecular mechanism and the underlying RNA biology. And of course, toxicity, off-target effects and delivery of lncRNA-based drugs have to be thoroughly investigated and closely monitored (Wahlestedt, 2013). Several companies such as RaNA Therapeutics, OPKO-CURNA, Moderna Therapeutics and Dicerna Pharmaceuticals already work on RNA-targeted medicines that selectively activate protein expression, targeted upregulation of gene expression, messenger RNA therapeutics and silencing of undruggable disease targets, respectively.

Furthermore, lncRNAs could also be used as diagnostic biomarkers and prognostic factors in cancer and other diseases. *HOTAIR* and *MALAT1* are overexpressed in numerous tumor entities and could potentially be detected in saliva, urine or blood (Reis and Verjovski-Almeida, 2012; Vitiello et al., 2014). The prostate cancer antigen 3 (PCA3) lncRNA can be detected in the urine and has already been approved by the US Food and Drug Administration as a biomarker for prostate cancer as it is more sensitive and more specific than the widely used PSA blood test (Sartori and Chan, 2014). *HOTAIR* is a strong biomarker candidate in metastatic oral cancer as it is found in higher concentrations in saliva of these patients (Vitiello et al., 2014). As a prognostic biomarker, *HOTAIR* has been associated with decreased survival in numerous cancers and higher levels were found in the blood of patients suffering from *HOTAIR* overexpressing colorectal cancer (Svoboda et al., 2014). Other examples of potential prognostic biomarkers currently being investigated include *ANRIL*, *MALAT1*, *GAS5* and *Sox2ot* (Vitiello et al., 2014).

5.7 Significance of datasets presented in this study

Within the scope of this thesis, I produced numerous RNA-seq datasets to study the RNA biology features of lncRNAs. While the majority of published transcriptome annotations and RNA biology experiments are difficult to compare among themselves as they were carried out using variable protocols and different cell types, I decided to do all experiments in the same cell types to be able to draw informative conclusions. I have chosen mouse ES cells as the primary model as they express a wealth of lncRNAs and selected embryonic fibroblasts as a

second cell type to investigate how lncRNA biology varies between cell types. Additionally, I repeated all experiments in the corresponding cell types of the rat to be able to assay conservation and evolution of RNA biology features. In summary, for ES cells and embryonic fibroblasts of the mouse and the rat I sequenced polyA-enriched RNA for transcriptome assembly (Table 13) and total non-ribosomal RNA (Table 16) for the analysis of splicing efficiency. For the investigation of RNA stability, I sequenced RNA from the same cell types after treatment with the transcription inhibitor Actinomycin D treated or the vehicle control EtOH (Table 14). I further isolated nuclear and cytoplasmic RNA fractions of these cells (except rat ES cells) and sequenced them (Table 15). For the analysis of developmental regulation of lncRNAs, I sequenced two fetal and one adult sample of the liver and the heart, FACS sorted B cells as well as CD4+ and CD8+ T cells and last but not least three replicates of spleen (Table 17). All together, I prepared 76 RNA-seq libraries for this study and generated 4.1 billion reads which were analyzed alongside with 2.3 billion reads from published RNA-seq experiments. These deeply sequenced and well controlled datasets will be a valuable and comprehensive resource for the research community to investigate the expression states and RNA features of lncRNAs genome-wide.

6 REFERENCES

Alvarez-Dominguez, J.R., Hu, W., Yuan, B., Shi, J., Park, S.S., Gromatzky, A.A., van Oudenaarden, A., and Lodish, H.F. (2013). Global discovery of erythroid long non-coding RNAs reveals novel regulators of red cell maturation. Blood.

Ameur, A., Zaghlool, A., Halvardson, J., Wetterbom, A., Gyllensten, U., Cavelier, L., and Feuk, L. (2011). Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. Nat. Struct. Mol. Biol. *18*, 1435–1440.

Athyros, V.G., Kakafika, A.I., Tziomalos, K., Karagiannis, A., and Mikhailidis, D.P. (2008). Antisense technology for the prevention or the treatment of cardiovascular disease: the next blockbuster? Expert Opin. Investig. Drugs *17*, 969–972.

Ball, S.T., Williamson, C.M., Hayes, C., Hacker, T., and Peters, J. (2001). The spatial and temporal expression pattern of Nesp and its antisense Nespas, in mid-gestation mouse embryos. Mech. Dev. *100*, 79–81.

Bassett, A.R., Akhtar, A., Barlow, D.P., Bird, A.P., Brockdorff, N., Duboule, D., Ephrussi, A., Ferguson-Smith, A.C., Gingeras, T.R., Haerty, W., et al. (2014). Considerations when investigating lncRNA function in vivo. eLife *3*, e03058.

Batista, P.J., and Chang, H.Y. (2013). Long noncoding RNAs: cellular address codes in development and disease. Cell *152*, 1298–1307.

Beltran, M., Puig, I., Peña, C., García, J.M., Alvarez, A.B., Peña, R., Bonilla, F., and de Herreros, A.G. (2008). A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition. Genes Dev. *22*, 756–769.

Bensaude, O. (2011). Inhibiting eukaryotic transcription: Which compound to choose? How to evaluate its activity? Transcription *2*, 103–108.

Bentley, D.L. (2014). Coupling mRNA processing with transcription in time and space. Nat. Rev. Genet. *15*, 163–175.

Bertani, S., Sauer, S., Bolotin, E., and Sauer, F. (2011). The noncoding RNA Mistral activates Hoxa6 and Hoxa7 expression and stem cell differentiation by recruiting MLL1 to chromatin. Mol. Cell *43*, 1040–1046.

Berteaux, N., Lottin, S., Monté, D., Pinte, S., Quatannens, B., Coll, J., Hondermarck, H., Curgy, J.-J., Dugimont, T., and Adriaenssens, E. (2005). H19 mRNA-like noncoding RNA promotes breast cancer cell proliferation through positive control by E2F1. J. Biol. Chem. *280*, 29625–29636.

Bertone, P., Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Rinn, J.L., Tongprasit, W., Samanta, M., Weissman, S., et al. (2004). Global identification of human transcribed sequences with genome tiling arrays. Science *306*, 2242–2246.

Blackshaw, S., Harpavat, S., Trimarchi, J., Cai, L., Huang, H., Kuo, W.P., Weber, G., Lee, K., Fraioli, R.E., Cho, S.-H., et al. (2004). Genomic analysis of mouse retinal development. PLoS Biol. *2*, E247.

Bond, C.S., and Fox, A.H. (2009). Paraspeckles: nuclear bodies built on long noncoding RNA. J. Cell Biol. *186*, 637–644.

Brannan, C.I., Dees, E.C., Ingram, R.S., and Tilghman, S.M. (1990). The product of the H19 gene may function as an RNA. Mol. Cell. Biol. *10*, 28–36.

Brosius, J. (2005). Waste not, want not--transcript excess in multicellular eukaryotes. Trends Genet. TIG *21*, 287–288.

Brown, J.A., Bulkley, D., Wang, J., Valenstein, M.L., Yario, T.A., Steitz, T.A., and Steitz, J.A. (2014). Structural insights into the stabilization of MALAT1 noncoding RNA by a bipartite triple helix. Nat. Struct. Mol. Biol. *21*, 633–640.

Buehr, M., Meek, S., Blair, K., Yang, J., Ure, J., Silva, J., McLay, R., Hall, J., Ying, Q.-L., and Smith, A. (2008). Capture of authentic embryonic stem cells from rat blastocysts. Cell *135*, 1287–1298.

Buetow, K.H., Edmonson, M.N., and Cassidy, A.B. (1999). Reliable identification of large numbers of candidate SNPs from public EST data. Nat. Genet. *21*, 323–325.

Bürckstümmer, T., Banning, C., Hainzl, P., Schobesberger, R., Kerzendorfer, C., Pauler, F.M., Chen, D., Them, N., Schischlik, F., Rebsamen, M., et al. (2013). A reversible gene trap collection empowers haploid genetics in human cells. Nat. Methods *10*, 965–971.

Cabianca, D.S., Casa, V., Bodega, B., Xynos, A., Ginelli, E., Tanaka, Y., and 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, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. *25*, 1915–1927.

Capel, B., Albrecht, K.H., Washburn, L.L., and Eicher, E.M. (1999). Migration of mesonephric cells into the mammalian gonad depends on Sry. Mech. Dev. *84*, 127–131.

Carmody, S.R., and Wente, S.R. (2009). mRNA nuclear export at a glance. J. Cell Sci. *122*, 1933–1937.

Carninci, P., Waki, K., Shiraki, T., Konno, H., Shibata, K., Itoh, M., Aizawa, K., Arakawa, T., Ishii, Y., Sasaki, D., et al. (2003). Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. Genome Res. *13*, 1273–1289.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., et al. (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., et al. (2012). Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. Nature *491*, 454–457.

Cassé, C., Giannoni, F., Nguyen, V.T., Dubois, M.F., and Bensaude, O. (1999). The transcriptional inhibitors, actinomycin D and alpha-amanitin, activate the HIV-1 promoter and favor phosphorylation of the RNA polymerase II C-terminal domain. J. Biol. Chem. *274*, 16097–16106.

Cesana, M., Cacchiarelli, D., Legnini, I., Santini, T., Sthandier, O., Chinappi, M., Tramontano, A., and Bozzoni, I. (2011). A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. Cell *147*, 358–369.

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., et al. (2005). Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science *308*, 1149–1154.

Clark, M.B., Amaral, P.P., Schlesinger, F.J., Dinger, M.E., Taft, R.J., Rinn, J.L., Ponting, C.P., Stadler, P.F., Morris, K.V., Morillon, A., et al. (2011). The reality of pervasive transcription. PLoS Biol. *9*, e1000625; discussion e1001102.

Clark, M.B., Johnston, R.L., Inostroza-Ponta, M., Fox, A.H., Fortini, E., Moscato, P., Dinger, M.E., and Mattick, J.S. (2012). Genome-wide analysis of long noncoding RNA stability. Genome Res. *22*, 885–898.

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845–1848.

Crick, F. (1970). Central dogma of molecular biology. Nature *227*, 561–563.

Dani, C., Blanchard, J.M., Piechaczyk, M., El Sabouty, S., Marty, L., and Jeanteur, P. (1984). Extreme instability of myc mRNA in normal and transformed human cells. Proc. Natl. Acad. Sci. U. S. A. *81*, 7046–7050.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. *22*, 1775–1789.

Dias Neto, E., Correa, R.G., Verjovski-Almeida, S., Briones, M.R., Nagai, M.A., da Silva, W., Zago, M.A., Bordin, S., Costa, F.F., Goldman, G.H., et al. (2000). Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. Proc. Natl. Acad. Sci. U. S. A. *97*, 3491–3496.

Diederichs, S. (2014). The four dimensions of noncoding RNA conservation. Trends Genet. TIG *30*, 121–123.

Dimitrova, N., Zamudio, J.R., Jong, R.M., Soukup, D., Resnick, R., Sarma, K., Ward, A.J., Raj, A., Lee, J.T., Sharp, P.A., et al. (2014). LincRNA-p21 activates p21 in cis to promote Polycomb target gene expression and to enforce the G1/S checkpoint. Mol. Cell *54*, 777–790.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. (2012). Landscape of transcription in human cells. Nature *489*, 101–108.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinforma. Oxf. Engl. *29*, 15–21.

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

Ebralidze, A.K., Guibal, F.C., Steidl, U., Zhang, P., Lee, S., Bartholdy, B., Jorda, M.A., Petkova, V., Rosenbauer, F., Huang, G., et al. (2008). PU.1 expression is modulated by the balance of functional sense and antisense RNAs regulated by a shared cis-regulatory element. Genes Dev. *22*, 2085–2092.

Eißmann, M., Gutschner, T., Hämmerle, M., Günther, S., Caudron-Herger, M., Groß, M., Schirmacher, P., Rippe, K., Braun, T., Zörnig, M., et al. (2012). Loss of the abundant nuclear non-coding RNA MALAT1 is compatible with life and development. RNA Biol. *9*, 1076–1087.

ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799–816.

Engreitz, J.M., Pandya-Jones, A., McDonel, P., Shishkin, A., Sirokman, K., Surka, C., Kadri, S., Xing, J., Goren, A., Lander, E.S., et al. (2013). The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. Science *341*, 1237973.

Engreitz, J.M., Sirokman, K., McDonel, P., Shishkin, A.A., Surka, C., Russell, P., Grossman, S.R., Chow, A.Y., Guttman, M., and Lander, E.S. (2014). RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent Pre-mRNAs and chromatin sites. Cell *159*, 188–199.

Faghihi, M.A., Modarresi, F., Khalil, A.M., Wood, D.E., Sahagan, B.G., Morgan, T.E., Finch, C.E., St Laurent, G., Kenny, P.J., and 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.

Faghihi, M.A., Zhang, M., Huang, J., Modarresi, F., Van der Brug, M.P., Nalls, M.A., Cookson, M.R., St-Laurent, G., and Wahlestedt, C. (2010). Evidence for natural antisense transcript-mediated inhibition of microRNA function. Genome Biol. *11*, R56.

Fatica, A., and Bozzoni, I. (2014). Long non-coding RNAs: new players in cell differentiation and development. Nat. Rev. Genet. *15*, 7–21.

Feng, L., and Niu, D.-K. (2007). Relationship between mRNA stability and length: an old question with a new twist. Biochem. Genet. *45*, 131–137.

Flynn, R.A., Almada, A.E., Zamudio, J.R., and Sharp, P.A. (2011). Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome. Proc. Natl. Acad. Sci. U. S. A. *108*, 10460–10465.

Fujita, P.A., Rhead, B., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Cline, M.S., Goldman, M., Barber, G.P., Clawson, H., Coelho, A., et al. (2011). The UCSC Genome Browser database: update 2011. Nucleic Acids Res. *39*, D876–D882.

Golding, M.C., Magri, L.S., Zhang, L., Lalone, S.A., Higgins, M.J., and Mann, M.R.W. (2011). Depletion of Kcnq1ot1 non-coding RNA does not affect imprinting maintenance in stem cells. Dev. Camb. Engl. *138*, 3667–3678.

Grote, P., Wittler, L., Hendrix, D., Koch, F., Währisch, S., Beisaw, A., Macura, K., Bläss, G., Kellis, M., Werber, M., et al. (2013). The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. Dev. Cell *24*, 206–214.

Guenzl, P.M., and Barlow, D.P. (2012). Macro lncRNAs: a new layer of cis-regulatory information in the mammalian genome. RNA Biol. *9*, 731–741.

Gupta, R.A., Shah, N., Wang, K.C., Kim, J., Horlings, H.M., Wong, D.J., Tsai, M.-C., Hung, T., Argani, P., Rinn, J.L., et al. (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature *464*, 1071–1076.

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

Gutschner, T., Hämmerle, M., Eissmann, M., Hsu, J., Kim, Y., Hung, G., Revenko, A., Arun, G., Stentrup, M., Gross, M., et al. (2013b). The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. Cancer Res. *73*, 1180–1189.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223–227.

Guttman, M., Garber, M., Levin, J.Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M.J., Gnirke, A., Nusbaum, C., et al. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat. Biotechnol. *28*, 503–510.

Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., Bruhn, L., et al. (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature *477*, 295–300.

Hacisuleyman, E., Goff, L.A., Trapnell, C., Williams, A., Henao-Mejia, J., Sun, L., McClanahan, P., Hendrickson, D.G., Sauvageau, M., Kelley, D.R., et al. (2014). Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. Nat. Struct. Mol. Biol. *21*, 198–206.

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

Hansen, K.D., Brenner, S.E., and Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res. *38*, e131.

Hao, Y., Crenshaw, T., Moulton, T., Newcomb, E., and Tycko, B. (1993). Tumour-suppressor activity of H19 RNA. Nature *365*, 764–767.

Hatada, I., Morita, S., Obata, Y., Sotomaru, Y., Shimoda, M., and Kono, T. (2001). Identification of a new imprinted gene, Rian, on mouse chromosome 12 by fluorescent differential display screening. J. Biochem. (Tokyo) *130*, 187–190.

Hayden, E.C. (2014). Technology: The $1,000 genome. Nature *507*, 294–295.

Van Heesch, S., van Iterson, M., Jacobi, J., Boymans, S., Essers, P.B., de Bruijn, E., Hao, W., MacInnes, A.W., Cuppen, E., and Simonis, M. (2014). Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes. Genome Biol. *15*, R6.

Herriges, M.J., Swarr, D.T., Morley, M.P., Rathi, K.S., Peng, T., Stewart, K.M., and Morrisey, E.E. (2014). Long noncoding RNAs are spatially correlated with transcription factors and regulate lung development. Genes Dev. *28*, 1363–1379.

Houseley, J., and Tollervey, D. (2009). The many pathways of RNA degradation. Cell *136*, 763–776.

Huang, R., Jaritz, M., Guenzl, P., Vlatkovic, I., Sommer, A., Tamir, I.M., Marks, H., Klampfl, T., Kralovics, R., Stunnenberg, H.G., et al. (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, M.J., Kenzelmann-Broz, D., Khalil, A.M., Zuk, O., Amit, I., Rabani, M., et al. (2010). A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell *142*, 409–419.

Hutchinson, J.N., Ensminger, A.W., Clemson, C.M., Lynch, C.R., Lawrence, J.B., and Chess, A. (2007). A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. BMC Genomics *8*, 39.

Ji, P., Diederichs, S., Wang, W., Böing, S., Metzger, R., Schneider, P.M., Tidow, N., Brandt, B., Buerger, H., Bulk, E., et al. (2003). MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. Oncogene *22*, 8031–8041.

Kampa, D., Cheng, J., Kapranov, P., Yamanaka, M., Brubaker, S., Cawley, S., Drenkow, J., Piccolboni, A., Bekiranov, S., Helt, G., et al. (2004). Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. Genome Res. *14*, 331–342.

Kan, Z., Rouchka, E.C., Gish, W.R., and States, D.J. (2001). Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. Genome Res. *11*, 889–900.

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C.C., Suzuki, M., Kawai, J., et al. (2005). Antisense transcription in the mammalian transcriptome. Science *309*, 1564–1566.

Keane, T.M., Goodstadt, L., Danecek, P., White, M.A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M., et al. (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. Nature *477*, 289–294.

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

Khalil, A.M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B.E., van Oudenaarden, A., et al. (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc. Natl. Acad. Sci. U. S. A. *106*, 11667–11672.

Kim, T.-K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., et al. (2010). Widespread transcription at neuronal activity-regulated enhancers. Nature *465*, 182–187.

Klattenhoff, C.A., Scheuermann, J.C., Surface, L.E., Bradley, R.K., Fields, P.A., Steinhauser, M.L., Ding, H., Butty, V.L., Torrey, L., Haas, S., et al. (2013). Braveheart, a long noncoding RNA required for cardiovascular lineage commitment. Cell *152*, 570–583.

Koerner, M.V., Pauler, F.M., Huang, R., and Barlow, D.P. (2009). The function of non-coding RNAs in genomic imprinting. Dev. Camb. Engl. *136*, 1771–1783.

Koerner, M.V., Pauler, F.M., Hudson, Q.J., Santoro, F., Sawicka, A., Guenzl, P.M., Stricker, S.H., Schichl, Y.M., Latos, P.A., Klement, R.M., et al. (2012). A downstream CpG island controls transcript initiation and elongation and the methylation state of the imprinted Airn macro ncRNA promoter. PLoS Genet. *8*, e1002540.

Kong, L., Zhang, Y., Ye, Z.-Q., Liu, X.-Q., Zhao, S.-Q., Wei, L., and Gao, G. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res. *35*, W345–W349.

Kornienko, A.E., Guenzl, P.M., Barlow, D.P., and Pauler, F.M. (2013). Gene regulation by the act of long non-coding RNA transcription. BMC Biol. *11*, 59.

Kotake, Y., Nakagawa, T., Kitagawa, K., Suzuki, S., Liu, N., Kitagawa, M., and Xiong, Y. (2011). Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15(INK4B) tumor suppressor gene. Oncogene *30*, 1956–1962.

Kowalczyk, M.S., Higgs, D.R., and Gingeras, T.R. (2012). Molecular biology: RNA discrimination. Nature *482*, 310–311.

Kutter, C., Watt, S., Stefflova, K., Wilson, M.D., Goncalves, A., Ponting, C.P., Odom, D.T., and Marques, A.C. (2012). Rapid turnover of long noncoding RNAs and the evolution of gene expression. PLoS Genet. *8*, e1002841.

Lai, F., Orom, U.A., Cesaroni, M., Beringer, M., Taatjes, D.J., Blobel, G.A., and Shiekhattar, R. (2013). Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. Nature *494*, 497–501.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Latos, P.A., Pauler, F.M., Koerner, M.V., Şenergin, H.B., Hudson, Q.J., Stocsits, R.R., Allhoff, W., Stricker, S.H., Klement, R.M., Warczok, K.E., et al. (2012). Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. Science *338*, 1469–1472.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinforma. Oxf. Engl. *25*, 2078–2079.

Li, L., Liu, B., Wapinski, O.L., Tsai, M.-C., Qu, K., Zhang, J., Carlson, J.C., Lin, M., Fang, F., Gupta, R.A., et al. (2013a). Targeted disruption of Hotair leads to homeotic transformation and gene derepression. Cell Rep. *5*, 3–12.

Li, W., Notani, D., Ma, Q., Tanasa, B., Nunez, E., Chen, A.Y., Merkurjev, D., Zhang, J., Ohgi, K., Song, X., et al. (2013b). Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature *498*, 516–520.

Ljungman, M., Zhang, F., Chen, F., Rainbow, A.J., and McKay, B.C. (1999). Inhibition of RNA polymerase II as a trigger for the p53 response. Oncogene *18*, 583–592.

Mager, J., Montgomery, N.D., de Villena, F.P.-M., and Magnuson, T. (2003). Genome imprinting regulated by the mouse Polycomb group protein Eed. Nat. Genet. *33*, 502–507.

Mancini-Dinardo, D., Steele, S.J.S., Levorse, J.M., Ingram, R.S., and Tilghman, S.M. (2006). Elongation of the Kcnq1ot1 transcript is required for genomic imprinting of neighboring genes. Genes Dev. *20*, 1268–1282.

Melo, C.A., Drost, J., Wijchers, P.J., van de Werken, H., de Wit, E., Oude Vrielink, J.A.F., Elkon, R., Melo, S.A., Léveillé, N., Kalluri, R., et al. (2013). eRNAs are required for p53-dependent enhancer activity and gene transcription. Mol. Cell *49*, 524–535.

Mercer, T.R., and Mattick, J.S. (2013). Structure and function of long noncoding RNAs in epigenetic regulation. Nat. Struct. Mol. Biol. *20*, 300–307.

Mercer, T.R., Gerhardt, D.J., Dinger, M.E., Crawford, J., Trapnell, C., Jeddeloh, J.A., Mattick, J.S., and Rinn, J.L. (2012). Targeted RNA sequencing reveals the deep complexity of the human transcriptome. Nat. Biotechnol. *30*, 99–104.

Merkin, J., Russell, C., Chen, P., and Burge, C.B. (2012). Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. Science *338*, 1593–1599.

Milligan, L., Antoine, E., Bisbal, C., Weber, M., Brunel, C., Forné, T., and Cathala, G. (2000). H19 gene expression is up-regulated exclusively by stabilization of the RNA during muscle cell differentiation. Oncogene *19*, 5810–5816.

Miyagawa, R., Tano, K., Mizuno, R., Nakamura, Y., Ijiri, K., Rakwal, R., Shibato, J., Masuo, Y., Mayeda, A., Hirose, T., et al. (2012). Identification of cis- and trans-acting factors involved in the localization of MALAT-1 noncoding RNA to nuclear speckles. RNA N. Y. N *18*, 738–751.

Miyoshi, N., Wagatsuma, H., Wakana, S., Shiroishi, T., Nomura, M., Aisaka, K., Kohda, T., Surani, M.A., Kaneko-Ishino, T., and Ishino, F. (2000). Identification of an imprinted gene, Meg3/Gtl2 and its human homologue MEG3, first mapped on mouse distal chromosome 12 and human chromosome 14q. Genes Cells Devoted Mol. Cell. Mech. *5*, 211–220.

Mizuno, T., Chou, M.Y., and Inouye, M. (1984). A unique mechanism regulating gene expression: translational inhibition by a complementary RNA transcript (micRNA). Proc. Natl. Acad. Sci. U. S. A. *81*, 1966–1970.

Modarresi, F., Faghihi, M.A., Lopez-Toledano, M.A., Fatemi, R.P., Magistri, M., Brothers, S.P., van der Brug, M.P., and Wahlestedt, C. (2012). Inhibition of natural antisense transcripts in vivo results in gene-specific transcriptional upregulation. Nat. Biotechnol. *30*, 453–459.

Modrek, B., Resch, A., Grasso, C., and Lee, C. (2001). Genome-wide detection of alternative splicing in expressed sequences of human genes. Nucleic Acids Res. *29*, 2850–2859.

Mohammad, F., Mondal, T., Guseva, N., Pandey, G.K., and Kanduri, C. (2010). Kcnq1ot1 noncoding RNA mediates transcriptional gene silencing by interacting with Dnmt1. Dev. Camb. Engl. *137*, 2493–2499.

Moran, V.A., Perera, R.J., and Khalil, A.M. (2012). Emerging functional and mechanistic paradigms of mammalian long non-coding RNAs. Nucleic Acids Res. *40*, 6391–6400.

Morris, K.V., Santoso, S., Turner, A.-M., Pastori, C., and Hawkins, P.G. (2008). Bidirectional transcription directs both transcriptional gene activation and suppression in human cells. PLoS Genet. *4*, e1000258.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods *5*, 621–628.

Mousavi, K., Zare, H., Dell'orso, S., Grontved, L., Gutierrez-Cruz, G., Derfoul, A., Hager, G.L., and Sartorelli, V. (2013). eRNAs promote transcription by establishing chromatin accessibility at defined genomic loci. Mol. Cell *51*, 606–617.

Nagano, T., Mitchell, J.A., Sanz, L.A., Pauler, F.M., Ferguson-Smith, A.C., Feil, R., and Fraser, P. (2008). The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. Science *322*, 1717–1720.

Nagaraj, S.H., Gasser, R.B., and Ranganathan, S. (2007). A hitchhiker's guide to expressed sequence tag (EST) analysis. Brief. Bioinform. *8*, 6–21.

Natoli, G., and Andrau, J.-C. (2012). Noncoding transcription at enhancers: general principles and functional models. Annu. Rev. Genet. *46*, 1–19.

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

Neil, H., Malabat, C., d' Aubenton-Carafa, Y., Xu, Z., Steinmetz, L.M., and Jacquier, A. (2009). Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. Nature *457*, 1038–1042.

Ohno, S. (1972). So much "junk" DNA in our genome. Brookhaven Symp. Biol. *23*, 366–370.

Orom, U.A., and Shiekhattar, R. (2011). Noncoding RNAs and enhancers: complications of a long-distance relationship. Trends Genet. TIG *27*, 433–439.

Ørom, U.A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., Huang, Q., et al. (2010). Long noncoding RNAs with enhancer-like function in human cells. Cell *143*, 46–58.

Ozsolak, F., and Milos, P.M. (2011). RNA sequencing: advances, challenges and opportunities. Nat. Rev. Genet. *12*, 87–98.

Pan, T. (2013). N6-methyl-adenosine modification in messenger and long non-coding RNA. Trends Biochem. Sci. *38*, 204–209.

Pandey, R.R., Mondal, T., Mohammad, F., Enroth, S., Redrup, L., Komorowski, J., Nagano, T., Mancini-Dinardo, D., and Kanduri, C. (2008). Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. Mol. Cell *32*, 232–246.

Pasmant, E., Laurendeau, I., Héron, D., Vidaud, M., Vidaud, D., and Bièche, 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., and Bièche, I. (2011). ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol. *25*, 444–448.

Pauler, F.M., Koerner, M.V., and Barlow, D.P. (2007). Silencing by imprinted noncoding RNAs: is transcription the answer? Trends Genet. TIG *23*, 284–292.

Pauler, F.M., Barlow, D.P., and Hudson, Q.J. (2012). Mechanisms of long range silencing by imprinted macro non-coding RNAs. Curr. Opin. Genet. Dev. *22*, 283–289.

Pauli, A., Valen, E., Lin, M.F., Garber, M., Vastenhouw, N.L., Levin, J.Z., Fan, L., Sandelin, A., Rinn, J.L., Regev, A., et al. (2012). Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res. *22*, 577–591.

Pelechano, V., and Steinmetz, L.M. (2013). Gene regulation by antisense transcription. Nat. Rev. Genet. *14*, 880–893.

Pettersson, E., Lundeberg, J., and Ahmadian, A. (2009). Generations of sequencing technologies. Genomics *93*, 105–111.

Plath, K., Fang, J., Mlynarczyk-Evans, S.K., Cao, R., Worringer, K.A., Wang, H., de la Cruz, C.C., Otte, A.P., Panning, B., and Zhang, Y. (2003). Role of histone H3 lysine 27 methylation in X inactivation. Science *300*, 131–135.

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

Pontier, D.B., and Gribnau, J. (2011). Xist regulation and function explored. Hum. Genet. *130*, 223–236.

Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M.S., Mapendano, C.K., Schierup, M.H., and Jensen, T.H. (2008). RNA exosome depletion reveals transcription upstream of active human promoters. Science *322*, 1851–1854.

Prensner, J.R., Iyer, M.K., Balbin, O.A., Dhanasekaran, S.M., Cao, Q., Brenner, J.C., Laxman, B., Asangani, I.A., Grasso, C.S., Kominsky, H.D., et al. (2011). Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nat. Biotechnol. *29*, 742–749.

Pruitt, K.D., Brown, G.R., Hiatt, S.M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C.M., Hart, J., Landrum, M.J., McGarvey, K.M., et al. (2014). RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. *42*, D756–D763.

Rabani, M., Levin, J.Z., Fan, L., Adiconis, X., Raychowdhury, R., Garber, M., Gnirke, A., Nusbaum, C., Hacohen, N., Friedman, N., et al. (2011). Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. Nat. Biotechnol. *29*, 436–442.

Rapicavoli, N.A., Poth, E.M., Zhu, H., and Blackshaw, S. (2011). The long noncoding RNA Six3OS acts in trans to regulate retinal development by modulating Six3 activity. Neural Develop. *6*, 32.

Redrup, L., Branco, M.R., Perdeaux, E.R., Krueger, C., Lewis, A., Santos, F., Nagano, T., Cobb, B.S., Fraser, P., and Reik, W. (2009). The long noncoding RNA Kcnq1ot1 organises a lineage-specific nuclear domain for epigenetic gene silencing. Dev. Camb. Engl. *136*, 525–530.

Reis, E.M., and Verjovski-Almeida, S. (2012). Perspectives of Long Non-Coding RNAs in Cancer Diagnostics. Front. Genet. *3*, 32.

Rinn, J.L., and Chang, H.Y. (2012). Genome regulation by long noncoding RNAs. Annu. Rev. Biochem. *81*, 145–166.

Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E., et al. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell *129*, 1311–1323.

Roehr, B. (1998). Fomivirsen approved for CMV retinitis. J. Int. Assoc. Physicians AIDS Care *4*, 14–16.

Rogozin, I.B., Carmel, L., Csuros, M., and Koonin, E.V. (2012). Origin and evolution of spliceosomal introns. Biol. Direct *7*, 11.

Ross, J. (1995). mRNA stability in mammalian cells. Microbiol. Rev. *59*, 423–450.

Sado, T., Hoki, Y., and Sasaki, H. (2005). Tsix silences Xist through modification of chromatin structure. Dev. Cell *9*, 159–165.

Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P.P. (2011). A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? Cell *146*, 353–358.

De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B.K., Muller, H., Ragoussis, J., Wei, C.-L., and Natoli, G. (2010). A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS Biol. *8*, e1000384.

Santoro, F., and Pauler, F.M. (2013). Silencing by the imprinted Airn macro lncRNA: transcription is the answer. Cell Cycle Georget. Tex *12*, 711–712.

Sartori, D.A., and Chan, D.W. (2014). Biomarkers in prostate cancer: what's new? Curr. Opin. Oncol. *26*, 259–264.

Sauvageau, M., Goff, L.A., Lodato, S., Bonev, B., Groff, A.F., Gerhardinger, C., Sanchez-Gomez, D.B., Hacisuleyman, E., Li, E., Spence, M., et al. (2013). Multiple knockout mouse models reveal lincRNAs are required for life and brain development. eLife *2*, e01749.

Schaukowitch, K., Joo, J.-Y., Liu, X., Watts, J.K., Martinez, C., and Kim, T.-K. (2014). Enhancer RNA Facilitates NELF Release from Immediate Early Genes. Mol. Cell *56*, 29–42.

Schmitt, A.O., Specht, T., Beckmann, G., Dahl, E., Pilarsky, C.P., Hinzmann, B., and Rosenthal, A. (1999). Exhaustive mining of EST libraries for genes differentially expressed in normal and tumour tissues. Nucleic Acids Res. *27*, 4251–4260.

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

Seidl, C.I.M., Stricker, S.H., and Barlow, D.P. (2006). The imprinted Air ncRNA is an atypical RNAPII transcript that evades splicing and escapes nuclear export. EMBO J. *25*, 3565–3575.

Seila, A.C., Calabrese, J.M., Levine, S.S., Yeo, G.W., Rahl, P.B., Flynn, R.A., Young, R.A., and Sharp, P.A. (2008). Divergent transcription from active promoters. Science *322*, 1849–1851.

Sharova, L.V., Sharov, A.A., Nedorezov, T., Piao, Y., Shaik, N., and Ko, M.S.H. (2009). Database for mRNA half-life of 19 977 genes obtained by DNA microarray analysis of pluripotent and differentiating mouse embryonic stem cells. DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes *16*, 45–58.

Sheardown, S.A., Duthie, S.M., Johnston, C.M., Newall, A.E., Formstone, E.J., Arkell, R.M., Nesterova, T.B., Alghisi, G.C., Rastan, S., and Brockdorff, N. (1997). Stabilization of Xist RNA mediates initiation of X chromosome inactivation. Cell *91*, 99–107.

Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V.V., et al. (2012). A map of the cis-regulatory sequences in the mouse genome. Nature *488*, 116–120.

Sigova, A.A., Mullen, A.C., Molinie, B., Gupta, S., Orlando, D.A., Guenther, M.G., Almada, A.E., Lin, C., Sharp, P.A., Giallourakis, C.C., et al. (2013). Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. Proc. Natl. Acad. Sci. U. S. A. *110*, 2876–2881.

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

Solnestam, B.W., Stranneheim, H., Hällman, J., Käller, M., Lundberg, E., Lundeberg, J., and Akan, P. (2012). Comparison of total and cytoplasmic mRNA reveals global regulation by nuclear retention and miRNAs. BMC Genomics *13*, 574.

Srivastava, D., and Cordes Metzler, K.R. (2013). Fending for a Braveheart. EMBO J. *32*, 1211–1213.

Stadtfeld, M., Apostolou, E., Akutsu, H., Fukuda, A., Follett, P., Natesan, S., Kono, T., Shioda, T., and Hochedlinger, K. (2010). Aberrant silencing of imprinted genes on chromosome 12qF1 in mouse induced pluripotent stem cells. Nature *465*, 175–181.

St Laurent, G., Shtokalo, D., Dong, B., Tackett, M.R., Fan, X., Lazorthes, S., Nicolas, E., Sang, N., Triche, T.J., McCaffrey, T.A., et al. (2013). VlincRNAs controlled by retroviral elements are a hallmark of pluripotency and cancer. Genome Biol. *14*, R73.

Stricker, S.H., Steenpass, L., Pauler, F.M., Santoro, F., Latos, P.A., Huang, R., Koerner, M.V., Sloane, M.A., Warczok, K.E., and Barlow, D.P. (2008). Silencing and transcriptional properties of the imprinted Airn ncRNA are independent of the endogenous promoter. EMBO J. *27*, 3116–3128.

Sultan, M., Dökel, S., Amstislavskiy, V., Wuttig, D., Sültmann, H., Lehrach, H., and Yaspo, M.-L. (2012). A simple strand-specific RNA-Seq library preparation protocol combining the Illumina TruSeq RNA and the dUTP methods. Biochem. Biophys. Res. Commun. *422*, 643–646.

Sun, B.K., Deaton, A.M., and Lee, J.T. (2006). A transient heterochromatic state in Xist preempts X inactivation choice without RNA stabilization. Mol. Cell *21*, 617–628.

Sun, L., Goff, L.A., Trapnell, C., Alexander, R., Lo, K.A., Hacisuleyman, E., Sauvageau, M., Tazon-Vega, B., Kelley, D.R., Hendrickson, D.G., et al. (2013). Long noncoding RNAs regulate adipogenesis. Proc. Natl. Acad. Sci. U. S. A. *110*, 3387–3392.

Svoboda, M., Slyskova, J., Schneiderova, M., Makovicky, P., Bielik, L., Levy, M., Lipska, L., Hemmelova, B., Kala, Z., Protivankova, M., et al. (2014). HOTAIR long non-coding RNA is a

negative prognostic factor not only in primary tumors, but also in the blood of colorectal cancer patients. Carcinogenesis *35*, 1510–1515.

Tani, H., and Akimitsu, N. (2012). Genome-wide technology for determining RNA stability in mammalian cells: historical perspective and recent advantages based on modified nucleotide labeling. RNA Biol. *9*, 1233–1238.

Tani, H., Nakamura, Y., Ijiri, K., and Akimitsu, N. (2010). Stability of MALAT-1, a nuclear long non-coding RNA in mammalian cells, varies in various cancer cells. Drug Discov. Ther. *4*, 235–239.

Tani, H., Mizutani, R., Salam, K.A., Tano, K., Ijiri, K., Wakamatsu, A., Isogai, T., Suzuki, Y., and Akimitsu, N. (2012). Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. Genome Res. *22*, 947–956.

Terranova, R., Yokobayashi, S., Stadler, M.B., Otte, A.P., van Lohuizen, M., Orkin, S.H., and Peters, A.H.F.M. (2008). Polycomb group proteins Ezh2 and Rnf2 direct genomic contraction and imprinted repression in early mouse embryos. Dev. Cell *15*, 668–679.

Tilgner, H., Knowles, D.G., Johnson, R., Davis, C.A., Chakrabortty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T.R., and Guigó, R. (2012). Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. Genome Res. *22*, 1616–1625.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinforma. Oxf. Engl. *25*, 1105–1111.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515.

Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat. Biotechnol. *31*, 46–53.

Trinklein, N.D., Aldred, S.F., Hartman, S.J., Schroeder, D.I., Otillar, R.P., and Myers, R.M. (2004). An abundance of bidirectional promoters in the human genome. Genome Res. *14*, 62–66.

Tripathi, V., Ellis, J.D., Shen, Z., Song, D.Y., Pan, Q., Watt, A.T., Freier, S.M., Bennett, C.F., Sharma, A., Bubulya, P.A., et al. (2010). The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. Mol. Cell *39*, 925–938.

Tsai, M.-C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J.K., Lan, F., Shi, Y., Segal, E., and Chang, H.Y. (2010). Long noncoding RNA as modular scaffold of histone modification complexes. Science *329*, 689–693.

Uesaka, M., Nishimura, O., Go, Y., Nakashima, K., Agata, K., and Imamura, T. (2014). Bidirectional promoters are the major source of gene activation-associated non-coding RNAs in mammals. BMC Genomics *15*, 35.

Ulitsky, I., and Bartel, D.P. (2013). lincRNAs: genomics, evolution, and mechanisms. Cell *154*, 26–46.

Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H., and Bartel, D.P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell *147*, 1537–1550.

Umlauf, D., Goto, Y., Cao, R., Cerqueira, F., Wagschal, A., Zhang, Y., and Feil, R. (2004). Imprinting along the Kcnq1 domain on mouse chromosome 7 involves repressive histone methylation and recruitment of Polycomb group complexes. Nat. Genet. *36*, 1296–1300.

Vance, K.W., and Ponting, C.P. (2014). Transcriptional regulatory functions of nuclear long noncoding RNAs. Trends Genet. TIG *30*, 348–355.

Visel, A., Zhu, Y., May, D., Afzal, V., Gong, E., Attanasio, C., Blow, M.J., Cohen, J.C., Rubin, E.M., and Pennacchio, L.A. (2010). Targeted deletion of the 9p21 non-coding coronary artery disease risk interval in mice. Nature *464*, 409–412.

Vitiello, M., Tuccoli, A., and Poliseno, L. (2014). Long non-coding RNAs in cancer: implications for personalized therapy. Cell. Oncol. Dordr.

Wagschal, A., Sutherland, H.G., Woodfine, K., Henckel, A., Chebli, K., Schulz, R., Oakey, R.J., Bickmore, W.A., and Feil, R. (2008). G9a histone methyltransferase contributes to imprinting in the mouse placenta. Mol. Cell. Biol. *28*, 1104–1113.

Wahlestedt, C. (2013). Targeting long non-coding RNA to therapeutically upregulate gene expression. Nat. Rev. Drug Discov. *12*, 433–446.

Wamstad, J.A., Alexander, J.M., Truty, R.M., Shrikumar, A., Li, F., Eilertson, K.E., Ding, H., Wylie, J.N., Pico, A.R., Capra, J.A., et al. (2012). Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. Cell *151*, 206–220.

Wang, K.C., and Chang, H.Y. (2011). Molecular mechanisms of long noncoding RNAs. Mol. Cell *43*, 904–914.

Wang, D., Garcia-Bassets, I., Benner, C., Li, W., Su, X., Zhou, Y., Qiu, J., Liu, W., Kaikkonen, M.U., Ohgi, K.A., et al. (2011a). Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. Nature *474*, 390–394.

Wang, F., Li, L., Xu, H., Liu, Y., Yang, C., Cowley, A.W., Wang, N., Liu, P., and Liang, M. (2014a). Characteristics of Long Non-coding RNAs in the Brown Norway Rat and Alterations in the Dahl Salt-Sensitive Rat. Sci. Rep. *4*, 7146.

Wang, G.-Z., Lercher, M.J., and Hurst, L.D. (2011b). Transcriptional coupling of neighboring genes and gene expression noise: evidence that gene orientation and noncoding transcripts are modulators of noise. Genome Biol. Evol. *3*, 320–331.

Wang, K.C., Yang, Y.W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., Lajoie, B.R., Protacio, A., Flynn, R.A., Gupta, R.A., et al. (2011c). A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature *472*, 120–124.

Wang, L., Wang, S., and Li, W. (2012). RSeQC: quality control of RNA-seq experiments. Bioinforma. Oxf. Engl. *28*, 2184–2185.

Wang, X., Lu, Z., Gomez, A., Hon, G.C., Yue, Y., Han, D., Fu, Y., Parisien, M., Dai, Q., Jia, G., et al. (2014b). N6-methyladenosine-dependent regulation of messenger RNA stability. Nature *505*, 117–120.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. *10*, 57–63.

Washietl, S., Findeiss, S., Müller, S.A., Kalkhof, S., von Bergen, M., Hofacker, I.L., Stadler, P.F., and Goldman, N. (2011). RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. RNA N. Y. N *17*, 578–594.

Werber, M., Wittler, L., Timmermann, B., Grote, P., and Herrmann, B.G. (2014). The tissue-specific transcriptomic landscape of the mid-gestational mouse embryo. Dev. Camb. Engl. *141*, 2325–2330.

West, J.A., Davis, C.P., Sunwoo, H., Simon, M.D., Sadreyev, R.I., Wang, P.I., Tolstorukov, M.Y., and Kingston, R.E. (2014). The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites. Mol. Cell *55*, 791–802.

Williamson, C.M., Skinner, J.A., Kelsey, G., and Peters, J. (2002). Alternative non-coding splice variants of Nespas, an imprinted gene antisense to Nesp in the Gnas imprinting cluster. Mamm. Genome Off. J. Int. Mamm. Genome Soc. *13*, 74–79.

Williamson, C.M., Ball, S.T., Dawson, C., Mehta, S., Beechey, C.V., Fray, M., Teboul, L., Dear, T.N., Kelsey, G., and Peters, J. (2011). Uncoupling antisense-mediated silencing and DNA methylation in the imprinted Gnas cluster. PLoS Genet. *7*, e1001347.

Wilusz, J.E., Freier, S.M., and Spector, D.L. (2008). 3' end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA. Cell *135*, 919–932.

Wyers, F., Rougemaille, M., Badis, G., Rousselle, J.-C., Dufour, M.-E., Boulay, J., Régnault, B., Devaux, F., Namane, A., Séraphin, B., et al. (2005). Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. Cell *121*, 725–737.

Yang, P.K., and Kuroda, M.I. (2007). Noncoding RNAs and intranuclear positioning in monoallelic gene expression. Cell *128*, 777–786.

Yang, L., Lin, C., Liu, W., Zhang, J., Ohgi, K.A., Grinstein, J.D., Dorrestein, P.C., and Rosenfeld, M.G. (2011a). ncRNA- and Pc2 methylation-dependent gene relocation between nuclear structures mediates gene activation programs. Cell *147*, 773–788.

Yang, L., Duff, M.O., Graveley, B.R., Carmichael, G.G., and Chen, L.-L. (2011b). Genomewide characterization of non-polyadenylated RNAs. Genome Biol. *12*, R16.

Yang, Y.W., Flynn, R.A., Chen, Y., Qu, K., Wan, B., Wang, K.C., Lei, M., and Chang, H.Y. (2014). Essential role of lncRNA binding for WDR5 maintenance of active chromatin and embryonic stem cell pluripotency. eLife *3*, e02046.

Yap, K.L., Li, S., Muñoz-Cabello, A.M., Raguz, S., Zeng, L., Mujtaba, S., Gil, J., Walsh, M.J., and Zhou, M.-M. (2010). Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. Mol. Cell *38*, 662–674.

Yoon, J.-H., Abdelmohsen, K., Srikantan, S., Yang, X., Martindale, J.L., De, S., Huarte, M., Zhan, M., Becker, K.G., and Gorospe, M. (2012). LincRNA-p21 suppresses target mRNA translation. Mol. Cell *47*, 648–655.

Yu, W., Gius, D., Onyango, P., Muldoon-Jacobs, K., Karp, J., Feinberg, A.P., and Cui, H. (2008). Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. Nature *451*, 202–206.

Zhang, B., Arun, G., Mao, Y.S., Lazar, Z., Hung, G., Bhattacharjee, G., Xiao, X., Booth, C.J., Wu, J., Zhang, C., et al. (2012). The lncRNA Malat1 is dispensable for mouse development but its transcription plays a cis-regulatory role in the adult. Cell Rep. *2*, 111–123.

Zhang, B., Gunawardane, L., Niazi, F., Jahanbani, F., Chen, X., and Valadkhan, S. (2014). A novel RNA motif mediates the strict nuclear localization of a long noncoding RNA. Mol. Cell. Biol. *34*, 2318–2329.

Zhang, X., Rice, K., Wang, Y., Chen, W., Zhong, Y., Nakayama, Y., Zhou, Y., and Klibanski, A. (2010). Maternally expressed gene 3 (MEG3) noncoding ribonucleic acid: isoform structure, expression, and functions. Endocrinology *151*, 939–947.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

Zhang, Z.-W., Cheng, J., Xu, F., Yuan, M., Du, J.-B., Shang, J., Wang, Y., Du, L., Li, Z.-L., and Yuan, S. (2011). Mammal cells double their total RNAs against diabetes, ischemia reperfusion and malaria-induced oxidative stress. Mol. Med. Camb. Mass *17*, 533–541.

Zhao, J., Ohsumi, T.K., Kung, J.T., Ogawa, Y., Grau, D.J., Sarma, K., Song, J.J., Kingston, R.E., Borowsky, M., and Lee, J.T. (2010). Genome-wide identification of polycomb-associated RNAs by RIP-seq. Mol. Cell *40*, 939–953.

Zuckerkandl, E. (1992). Revisiting junk DNA. J. Mol. Evol. *34*, 259–271.

7 CURRICULUM VITAE

Personal Information

Name:	Philipp Günzl
Date of birth:	03.02.1985
Place of birth:	Vienna, Austria
Nationality:	Austrian

Education

10/2004 - 10/2008	Diploma degree in Biotechnology University of Applied Sciences, Vienna, Austria Thesis: “The protective role of the PI3K/PTEN pathway in innate immune responses” (Supervisor: Dr. Gernot Schabbauer) Medical University of Vienna, Vienna, Austria
10/2003 - 09/2004	Civilian Service (NÖ Landeskinderheim “Schwedenstift”)
09/1995 - 06/2003	Secondary school Perchtoldsdorf, Austria
09/1991 - 06/1995	Elementary school Perchtoldsdorf, Austria

Employment

04/2009 - 10/2014	Doctoral studies as part of the Vienna Biocenter and DK-RNA Biology international PhD program University of Vienna, Vienna, Austria Thesis: “Genome-wide investigation of RNA biology features of mouse long non-coding RNAs” (Supervisor: Prof. Denise Barlow) CeMM Center for Molecular Medicine GmBH, Vienna, Austria

Manuscripts in preparation

Guenzl PM, Kornienko AE, Hudson QJ, Penz T, Schoenegger A, Minnich M, Bock C & Pauler FM. **RNA biology defines evolutionary conserved lncRNA subclasses**. (manuscript in preparation)

Dotter CP, Guenzl PM, Huang R, Andergassen D, Mayer D, Tamir I, Sommer A, Thorvaldsen JL, Hudson QJ, Steenpass L, Barlow DP, Horsthemke B, Bartolomei MS, Pauler FM. **Sensitive map of mouse lncRNAs identifies extended transcription units regulated by genomic imprinting**. (manuscript in preparation)

Kornienko AE, Dotter CP, Guenzl PM, Gisslinger H, Gisslinger B, Cleary C, Kralovics R, *Pauler FM, * Barlow DP. **Human granulocyte transcriptome annotation and analysis reveals increased expression variability for lncRNAs compared to mRNAs in healthy individuals**. (manuscript in preparation)

Andergassen D, Dotter CP, Kulinski TM, Guenzl PM, Bammer P, Barlow DP, Pauler FM, Hudson QJ. **Allelome.PRO, a pipeline to define allele-specific genome features**. (manuscript in preparation)

Kulinski TM, Casari RT, Guenzl PM, Wenzel D, Andergassen D, Hladik A, Datlinger P, Farlik M, Theussl HC, Penninger JM, Knapp S, Bock C, Barlow DP and Hudson QJ. **Imprinted expression in cystic embryoid bodies shows an embryonic and not an extra-embryonic pattern**. (in revision)

Publications

Bürckstümmer 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, Stöckl 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. **A reversible gene trap collection empowers haploid genetics in human cells**. Nat Methods. 2013 Oct;10(10):965-71.

Guenzl PM, Raim R, Kral J, Brunner J, Sahin E, Schabbauer G. **Insulin hypersensitivity induced by hepatic PTEN gene ablation protects from murine endotoxemia**. PLoS One. 2013 Jun 25;8(6):e67013.

Kornienko AE, Guenzl PM, Barlow DP, Pauler FM. **Gene regulation by the act of long non-coding RNA transcription**. BMC Biol. 2013 May 30;11:59.

Guenzl PM, Barlow DP. **Macro lncRNAs: a new layer of cis-regulatory information in the mammalian genome**. RNA Biol. 2012 Jun;9(6):731-41.

Koerner MV, Pauler FM, Hudson QJ, Santoro F, Sawicka A, Guenzl PM, Stricker SH, Schichl YM, Latos PA, Klement RM, Warczok KE, Wojciechowski J, Seiser C, Kralovics R, Barlow DP. **A downstream CpG island controls transcript initiation and elongation and the methylation state of the imprinted Airn macro ncRNA promoter**. PLoS Genet. 2012;8(3):e1002540.

Huang R, Jaritz M, Guenzl P, Vlatkovic I, Sommer A, Tamir IM, Marks H, Klampfl T, Kralovics R, Stunnenberg HG, Barlow DP, Pauler FM. **An RNA-Seq strategy to detect the complete coding and non-coding transcriptome including full-length imprinted macro ncRNAs**. PLoS One. 2011;6(11):e27288.

Günzl P, Bauer K, Hainzl E, Matt U, Dillinger B, Mahr B, Knapp S, Binder BR, Schabbauer G. **Anti-inflammatory properties of the PI3K pathway are mediated by IL-10/DUSP regulation**. J Leukoc Biol. 2010 Dec;88(6):1259-69.

Schabbauer G, Matt U, Günzl P, Warszawska J, Furtner T, Hainzl E, Elbau I, Mesteri I, Doninger B, Binder BR, Knapp S. **Myeloid PTEN promotes inflammation but impairs bactericidal activities during murine pneumococcal pneumonia**. J Immunol. 2010 Jul 1;185(1):468-76.

Günzl P, Schabbauer G. **Recent advances in the genetic analysis of PTEN and PI3K innate immune properties**. Immunobiology. 2008;213(9-10):759-65.

Stipends and awards

DOC fellowship of the Austrian Academy of Sciences

01/2011 - 06/2013 (2.5 years)

Selection into the international Vienna Biocenter PhD program

01/2009

Conferences

Keystone Symposia's 2012 Meeting on Non-Coding RNAs

Snowbird, UT, USA

March 31 - April 5, 2012

(poster presentation)

3rd EMBO Meeting
Vienna, Austria
September 10 - September 13, 2011
(poster presentation)

Epigenetic Regulation in Cell Fate & Disease
Vienna, Austria
March 17 - March 19, 2010
(poster presentation)

1st EMBO Meeting
Amsterdam, The Netherlands
August 29 - September 1, 2009
(poster presentation)

European Human Genetics Conference
Vienna, Austria
May 23 - May 26, 2009

Invited Talks

Stipend Weekend of the Austrian Academy of Sciences
Vienna, Austria
Feb 24, 2012

Medical University of Graz
Graz, Austria
May 17, 2010

8 ACKNOWLEDGEMENTS

This thesis would not have been possible without the help of many people. First of all, I am sincerely grateful to my supervisor Denise Barlow for guiding me through my PhD and for sharing her knowledge and passion for science. I am further very thankful to Florian Pauler with whom I worked closely together on this and numerous other projects. Florian implemented most bioinformatic pipelines in the lab and gave crucial input to successfully finish this project. He further wrote literally hundreds of small scripts that made bioinformatic analysis so much easier for me.

I want to thank all past and present members of the Barlow lab for establishing such an inspiring and fun work atmosphere. Thanks to Irena Vlatkovic, for teaching me several RNA techniques at the very beginning; Alexandra Kornienko, for countless important and non-important discussions and bringing great Russian delicacies such as the saltiest cheese ever that nobody liked except me; Daniel Andergassen, for help with R scripting and broadening my musical interest; Christoph Dotter, for bioinformatic assistance and teaching me how to best apply the command awk; Federica Santoro, for spreading pleasant Italian spirit and her delicious cup cakes; Quanah Hudson, for help with mouse dissections; Martha Körner, for scientific advice and a great trip to Edinburgh and finally Tomek Kulinski, Philipp Bammer, Daniela Mayer, Wolfgang Allhoff, Ru Huang, Stephan Hütter, Rita Casari and Markus Muckenhuber for numerous scientific and entertaining discussions, daily coffee breaks and excellent cakes. Special thanks go to our marvelous technical assistants Justyna Konecka, Ruth Klement and Kasia Warczok for perfectly organizing the lab.

I am especially grateful to the Austrian Academy of Sciences for awarding me a DOC fellowship for 2.5 years and the RNA doctoral program headed by Prof. Andrea Barta for additional funding. I want to thank Giulio Superti-Furga and CeMM for providing a great international and professional atmosphere and for assembling a scientifically excellent bunch of people, I am confident CeMM will accelerate genome research in Austria.

I further want to thank my diploma supervisor Dr. Gernot Schabbauer for introducing me to the world of science and for a very successful period of time, which certainly helped me to pass the international VBC PhD selection to start my PhD.

I am also very grateful to my parents Hannelore and Michael who supported me throughout my studies, gave me crucial advice and always encouraged me to follow my passion. A big thank you goes to all of my soccer-playing, basketball-playing, concert-listening, life-discussing and partying friends for providing emotional balance and often needed distraction.

Last but not least, my loving thanks go to my enchanting partner Alexandra Hebar. Since we first met at the beginning of our studies, it was a pleasure to have you on this journey. I really enjoy your continuous support, your curiosity and your way of challenging me to encourage me to give my best. Thank you for loving me!

9 APPENDIX

Review #1:

Guenzl PM, Barlow DP. **Macro lncRNAs: a new layer of cis-regulatory information in the mammalian genome**. RNA Biol. 2012 Jun;9(6):731-41.

Review #2:

Kornienko AE, Guenzl PM, Barlow DP, Pauler FM. **Gene regulation by the act of long non-coding RNA transcription**. BMC Biol. 2013 May 30;11:59.