universität wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation /Title of the Doctoral Thesis

Mapping the mouse Allelome reveals tissue-specific regulation

verfasst von / submitted by

Mag. Daniel Andergassen

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doctor of Philosophy (PhD)

Wien, 2016 / Vienna 2016

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 794 685 490
Dissertationsgebiet lt. Studienblatt / field of study as it appears on the student record sheet:	Molekulare Biologie
Betreut von / Supervisor:	Denise P. Barlow, PhD Quanah J. Hudson, PhD

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

The biggest thank goes to my main supervisor Denise Barlow, especially for accepting me as a diploma student and for supporting me in my choice to return to Vienna to restart my PhD in her lab after 10 month as a PhD student at ETH Zürich. Since at that time Denise was within 1.5 years of retirement she had to convince the CeMM faculty that I could manage to finish my PhD within only 2 years and consequently agreed to continue to support me even in her retirement. At this point I have to thank to the CeMM faculty, especially to the two directors Giulio Superti-Furga and Anita Ender, who allowed me to return at CeMM to finish my PhD. In addition I thank Denise for constant support, knowledge and attention to the details. It is a great honor for me to call myself "the last PhD student of Denise Barlow" a person that heavily influenced the field of epigenetics.

The same amount of thanks goes to the two postdocs Quanah Hudson and Florian Pauler from the lab, who supervised me throughout my time as a diploma and PhD student. Especially Quanah, who first of all paid my salary with his grant money and supervised me through bench work, mouse dissections and scientific writing. I thank Florian for teaching me his bioinformatics expertise, which was a key component of both projects, and for challenging me from the first day on in the lab. In particular I thank Quanah and Florian for their special sort of clever and witty humor and for trying to answer all my questions independent of the quality.

A special thanks goes to Christoph Dotter for investing a lot of time to improve Allelome.PRO and for being a wonderful team player. Without his help and intellectual input it would have been impossible to finish the two projects and consequently my PhD in this short time. I thank Alexandra Kornienko for wonderful discussions, helping me to improve my scientific writing and for being a wonderful working colleague and friend. I would like to thank Philipp who took his time to teach me how to prepare the famous "Günzl Golden Hand" RNA-seq libraries and for the nice discussions. In addition I would like to thank to Markus Muckenhuber and Philipp Bammer, Tomasz Kulinski for the intellectual input during the coffee breaks.

Vor allem möchte ich meiner Freundin Stefanie danken, die mich in den letzten eineinhalb Jahren sehr unterstützt hat. Ein großes Dankeschön geht auch an meine Eltern und meine Schwester, die es mir ermöglicht haben meine Zukunft so zu gestalten wie ich sie mir vorstelle.

DECLARATION

This PhD thesis is written in a cumulative format and consists of the following two publications where the author of the thesis is the first author:

Publication 1 - Accepted research article

"**Allelome.PRO, a pipeline to define allele-specific genomic features from high-throughput sequencing data**"

Published in Nucleic Acids research on July 21, 2015 with Open Access

Article webpage: http://nar.oxfordjournals.org/content/early/2015/07/21/nar.gkv727.full

Impact factor: 9.112

Publication 2 - Submitted research article (in revision)

"**Mapping the mouse Allelome reveals tissue-specific regulation**"

Submitted to Genome Research on June 2, 2016

Impact factor: 14.63

ABBREVIATIONS

4C-seq	Circular chromosome conformation capture sequencing
Airn	Antisense *Igf2r* non-coding RNA
bp	Base pairs
ChIP-seq	Chromatin immunoprecipitation sequencing
Chr	Chromosome
Dmd	Dystrophin
DMR	Differentially methylated region
DNA	Deoxyribonucleic acid
DNMT1	DNA methyltransferase 1
EHMT2	Euchromatic histone-lysine N-methyltransferase 2
eQTL	Expression quantitative trait loci
ESCs	Embryonic stem cells
EXEL	Extra-embryonic-linage
FDR	False discovery rate
gDMR	Germline differentially methylated region
GWAS	Genome-wide association study
H2AK119ub1	Monoubiquitylation of H2A lysine 119
H3K27ac	Acetylation of H3 lysine 27
H3K27me3	Trimethylation of H3 lysine 27
H3K4me3	Trimethylation of H3 lysine 4
H3K9me3	Trimethylation of H3 lysine 9
HDAC3	Histone deacetylase 3
ICE	Imprinted control element
ICM	Inner cell mass
Igf2r	Insulin-like growth factor 2 receptor
kb	Kilobase
lncRNA	Long non-protein-coding RNA
Mb	Megabase
MEFs	Mouse embryonic fibroblasts
ML	Multi-linage
mRNA	Messenger RNA
Pde10a	Phosphodiesterase 10A
PGCs	Primordial germ cells
PN	Pronucleus
PRC1/PRC2	Polycomb repressive complex 1/2
PYS	Parietal yolk sac
RAP	RNA antisense purification
RefSeq	NCBI Reference Sequence Database
RMAE	Random monoallelic expression
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
Slc22a2/3	Solute carrier family 22 member 2/3
SNP	Single-nucleotide polymorphism
TF	Transcription factor
VE	Visceral yolk sac endoderm
VYS	Visceral yolk sac
Xce	X-chromosome controlling element
XCI	X-chromosome inactivation
Xic	X-chromosome inactivation center
Xist	X-inactive specific transcript

ZUSAMMENFASSUNG

In Säugetieren können genetische und epigenetische Unterschiede zwischen den elterlichen Allelen zu allele-spezifischer Genexpression (AGE) führen. RNA-seq wurde seit der Entwicklung von Hochdurchsatzsequenzierung verwendet um AGE in Mensch- und Mausgeweben zu detektieren. Da man die DNA-Sequenz der Eltern benötigt, erweist sich ein genomweiter Nachweis von allele-spezifischer Genexpression im Menschen als schwierig. Ein geeigneteres Modell um AGE nachzuweisen stellen Kreuzungen von genetisch unterschiedlichen Stämmen von Labormäuse dar, da die genetischen Unterschiede zwischen vielen Stämmen bereits bekannt sind. Obwohl AGE Detektion mittels RNA-seq eine anerkannte Methode ist, gab es zu Beginn dieser Studie keine bioinformatische Software, die präzise und mit einer niedrigen Fehlerquote, AGE detektiert. Deshalb entwickelten wir im ersten Teil dieser Studie die benutzerfreundliche Software Allelome.PRO, welche mit hoher Präzision allele-spezifische Genexpression oder Histonmodifikationen erfasst. Zusätzlich klassifiziert die Software jedes Gen in einem Zelltyp in eine der folgenden Kategorien: biallelisch, Mausstamm-spezifisch, geprägt oder nicht-informativ. Dadurch wird das gesamte allele-spezifische Bild aller aktiven Gene – „Allelome“ genannt - abgebildet. Im nächsten Schritt verwendeten wir Allelome.PRO um AGE für Protein und nicht-Protein-kodierenden (nk) Gene in 23 Geweben, in jeweils verschiedenen Entwicklungsstadien der Maus, zu identifizieren um das Maus Allelome zu erhalten. Diese Entwicklungsstadien beinhalten pluripotente, embryonale, extra-embryonale, neugeborene und adulte Gewebe. Diese an der Maus neuartige Analyse führte zu folgenden drei Erkenntnissen. Erstens wurden viele Gene mit gewebespezifischer Mausstamm-spezifischer Expression identifiziert und dass dieses Expressionsmuster von naheliegenden Mausstamm-spezifischen Expressions-Aktivatoren reguliert wird. Diese Aktivatoren werden von Histonen mit der H3K27ac Modifikation markiert. Zweitens wurden in 19 weiblichen Geweben eine unerwartet große Zahl an Genen, die dem Prozess der X-Chromosom-Inaktivierung entkommen, genannt „Escaper“, gefunden. Im Gegensatz zu den meisten Studien, die von niedrigen Escaper-Prozenten in der Maus berichten, fanden wir einen dem Menschen ähnlichen Prozentsatz von 15 Prozent. Bemerkenswert ist der hohe Escaper-Anteil von 50 Prozent im adulten Beinmuskel. Drittens konnten wir zeigen das geprägte Gengruppen viel größer sind als bisher angenommen und die Größe dramatisch zwischen Geweben und Entwicklungsstadien variiert. Insbesondere zeigen wir anhand

von genetischen Mausmodellen, dass sich die geprägte *Igf2r* Gengruppe in der Plazenta über zehn Megabasen erstreckt und somit die größte, geprägte Region in der Maus repräsentiert. Zusammenfassend zeigen unsere Ergebnisse, dass AGE welche durch genetische Unterschiede zwischen den Allelen oder durch epigenetische Prozesse, wie X-Chromosom-Inaktivierung oder genomischer Prägung entstehen, überraschend oft gewebespezifisch ist.

ABSTRACT

In mammals, genetic or epigenetic differences between the parental alleles can results in allele-specific expression. Since the development of high-throughput sequencing, RNA-seq has been used to detect allele-specific expression in human and mouse tissues. Genome-wide detection of allelic expression in human is difficult since it requires genotyping of the parents to distinguish the alleles. In contrast the inbred mouse model is a powerful system to map allele-specific expression since the genetic differences between different laboratory strains are known. Although RNA-seq is a powerful tool, a bioinformatics pipeline with increased sensitivity and low levels of false positive calls was lacking. Here we developed Allelome.PRO, a user-friendly, fully automated bioinformatics pipeline, which robustly detects allele-specific expression or chromatin modification from high-throughput sequencing data. The pipeline automatically characterizes the allelic profile from all genes in one cell type into biallelic, strain-biased, imprinted or non-informative and thus provides the full allelic expression picture, "the Allelome". Next we used the pipeline to generate the most comprehensive survey of allelic expression for protein and non-coding genes yet known, by conducting RNA-seq on 23 mouse tissues throughout the development, including pluripotent, embryonic, extra-embryonic, neonatal and adult tissues, to map for the first time "the mouse Allelome". The mouse Allelome reveals that tissue-specific strain-biased expression is correlated with nearby strain-biased H3K27ac enrichment, implying regulation by tissue-specific allelic enhancers. Next we mapped X-chromosome inactivation (XCI) escaper genes in 19 female tissues. In contrast to most previous reports in mouse that reported lower numbers, we found an average of 15% escapers per tissue similar to human, with the notable exception of adult leg muscle that showed 50% escaper genes. In addition we show that imprinted clusters are much larger than previously known, and change their size dramatically among tissues and during development. In particular we genetically demonstrate here that the *Igf2r* cluster extends over 10Mb in placenta, representing the largest imprinted region in mouse. In summary we find that allelic expression arising from genetic differences between the alleles or from epigenetic processes such as XCI and genomic imprinting, is surprisingly highly tissue-specific.

INTRODUCTION

1. Allele-specific expression in mammals

Mammalian cells have two sets of chromosomes, one inherited from the mother and one from the father. Such a diploid system contains every gene locus twice on autosomes, while genes on the sex chromosomes in males or on the mitochondrial DNA have only one parental gene copy. Given that both gene copies on the parental alleles are regulated independently, genes can be expressed from both alleles (biallelic) or only from one allele (monoallelic). Genes showing biallelic expression represent the largest group, while only a small proportion is considered to show monoallelic expression. Monoallelic expression can occur randomly (RMAE) or be specific to one allele due to epigenetic or genetic differences between the alleles, as in the case of imprinted and strain-biased expression respectively (Barlow and Bartolomei, 2014; Reinius and Sandberg, 2015).

Random X-chromosome inactivation (XCI) in female placental mammals is one of the most prominent examples of RMAE. XCI results in inactivation of most of the genes on one X-chromosome in females and is important for dosage compensation between the sexes. This process is controlled by the long non-protein-coding RNA (lncRNA) *Xist* (reviewed in (Wutz, 2011)). Furthermore RMAE was shown to be important for B or T lymphocytes development by allowing the expression of immunoglobulin or T-cell receptor respectively only from a single allele, a process called allelic exclusion (Brady et al., 2010). The first genome-wide study of RMAE in mouse and human used single nucleotide polymorphism (SNP) sensitive microarrays on clonal cell population of lymphoblastoid cells and found that 10% of genes show mitotically stable RMAE (Gimelbrant et al., 2007; Zwemer et al., 2012). A recent study detected 12-24% genes showing RMAE from F1 mouse crosses in early blastomeres by using single cell RNA sequencing (RNA-seq) (Deng et al., 2014). A functional role for the high number of genes showing RMAE might be to generate cellular diversity (Chess, 2012; Reinius and Sandberg, 2015). Important to keep in mind here is that RMAE in single cells appears to be biallelic gene expression when the whole cell population is assayed. In contrast to RMAE, strain-biased expression results from genetic differences between the alleles, and can thus be detected over a whole cell population. Such allelic difference can arise for example from SNPs in regulatory regions such as promoter or enhancers, or influence posttranscriptional processes. (Lappalainen et al., 2013; Leung et al., 2015). Genomic imprinting is an

epigenetic process *i.e.,* not dependent on genetic differences, that leads to parental-specific gene expression that can be detected over a whole cell population. Imprinted expression is regulated by differentially methylated regions (gDMR), that are established in either the male or female germline and then maintained in somatic cells on the same parental allele throughout life. Such gDMRs have been shown in 7 cases so far to act as the imprinted control element (ICE), which regulates imprinted expression of small gene clusters *in cis* (Barlow, 2011). For several imprinted clusters the unmethylated ICE has been shown to act as a promoter for a imprinted lncRNA that silences the entire gene cluster (Barlow and Bartolomei, 2014).

Following the identification of lncRNAs involved in regulating X-inactivation and imprinted expression and a limited number of other lncRNAs, the advent of next-generation sequencing identified several thousand novel lncRNAs in mouse and human. Recently a study in human analyzed thousands of tumor, tissue and cell line samples and identified approximately 60,000 lncRNAs, which is more than twice the number of protein-coding genes (Iyer et al., 2015). Compared to protein-coding genes, lncRNAs are highly tissue-specific, lowly expressed and show significantly more inter-individual expression variability between people (Cabili et al., 2011; Kornienko et al., 2016). Although the function of most lncRNAs identified by RNA-seq is unknown, and it cannot be excluded that some may be transcriptional noise, the list of functional lncRNAs continues to grow (Ulitsky and Bartel, 2013). To date a clear lncRNA classification system is lacking, however lncRNAs can be separated into three general groups: non-functional lncRNAs (transcriptional noise or no function assigned yet), *cis*-acting lncRNAs that silence overlapping protein-coding genes by transcription interference (function mediated by transcription) and lncRNAs that regulate gene expression *in cis* or *trans* (function mediated by lncRNA product) (reviewed in (Kornienko et al., 2013; Quinn and Chang, 2015; Ulitsky and Bartel, 2013)). The function of the high number of novel lncRNAs clearly requires further investigation and a determination of their allelic expression status could provide essential information on their possible function.

1.1 High-throughput based approach to identify allele-specific expression

Since the advent of RNA-seq, genome-wide mapping of allele-specific expression has been performed in many human and mouse tissues (Baran et al., 2015; Lagarrigue et al., 2013; Pickrell et al., 2010). Mapping allelic expression requires SNPs for allele-specific assignment of RNA-seq reads. This approach was used previously to identify

imprinted and strain-biased expression, as well as genes that escape XCI in females and other genes showing RMAE (Babak et al., 2015; Berletch et al., 2015; Deng et al., 2014; Keane et al., 2011). Allelic expression can be reliably detected genome-wide in an outbred population, such as in humans or in F1 hybrids derived from inbred mouse strains. However, distinguishing strain-biased from parental-specific expression in an outbred population requires the genotype of the parents, information that is not always available for human data.

Assaying F1 tissues from reciprocal crosses between inbred mouse strains is a powerful approach to map allele-specific expression, especially when the number of SNPs between the two strains is maximized by choosing strains that are genetically distant from each other (Figure 1). In the forward cross, strain 1 is the mother and strain 2 the father, while in the reverse cross, strain 2 is the mother and strain 1 the father. Allelic expression can then be quantified by sequencing the RNA of the F1 offspring (forward and reverse cross) and by assigning the reads based on their SNPs to the corresponding allele. This method allows the detection of genes that show biallelic expression (equal genotype distribution), strain-biased expression (biased genotype distribution for one strain) and imprinted expression (biased genotype distribution for one parent).

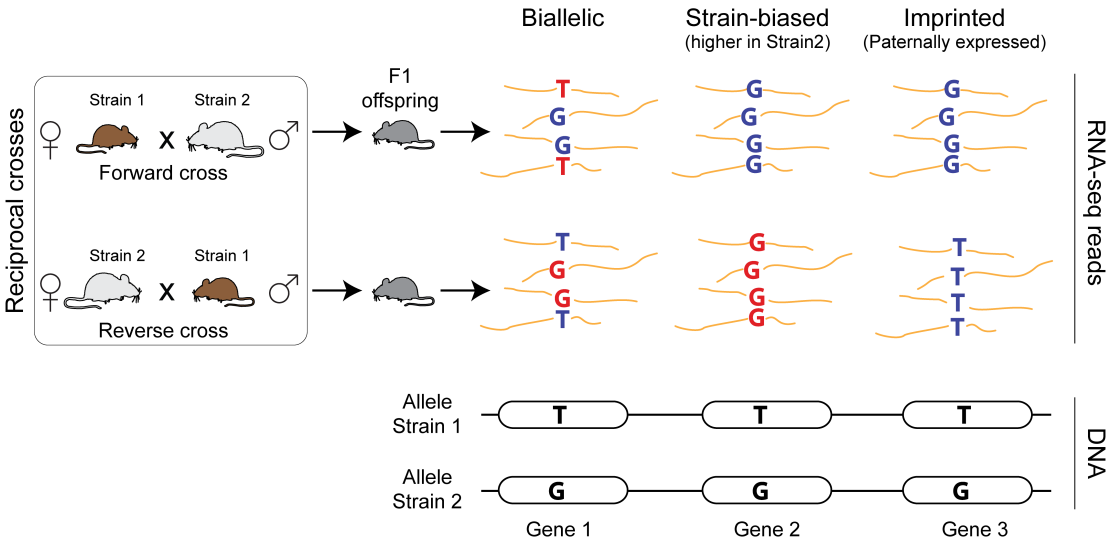


Figure 1. Genome-wide strategy to detect allele-specific expression using RNA-seq. First, reciprocal mouse crosses of two genetically distinct inbred mouse strains (Strain 1 and Strain 2) are conducted. In the forward cross, strain 1 is the mother and strain 2 the father, while in the reverse cross, strain 2 is the mother and strain 1 the father. Second, RNA-seq is performed on isolated F1 tissues from both crosses. Third, RNA-seq reads overlapping SNPs are assigned to the corresponding allele. The allelic read distribution within a gene locus indicates whether a gene shows biallelic expression (equal genotype distribution), strain-biased expression (biased genotype distribution for one strain) or imprinted expression (biased genotype distribution for one parent).

Several studies used F1 mouse hybrids and high-throughput sequencing to identify imprinted expression among different tissues and development (Babak, 2012; Babak et al., 2008; Babak et al., 2015; Lagarrigue et al., 2013; Proudhon and Bourc'his, 2010; Tran et al., 2014; Wang et al., 2008; Xie et al., 2012). Together, all these studies cover the majority of mouse organs and they found only a few hundred imprinted genes, most of which were either already listed in imprinted databases or were extensions of known imprinted genes (Glaser et al., 2006; Williamson CM, 2013). In contrast, one study reported more than thousand novel imprinted genes in different developmental stages of the mouse brain (Gregg et al., 2010b). However, reanalysis of this data demonstrated that the majority of the reported imprinted genes were false-positive to due technical and biological variation (DeVeale et al., 2012). Although RNA-seq is a powerful tool to quantify allelic expression from F1 crosses, it is now appreciated that the following aspects are important in order to reduce the number of false positive calls (reviewed in (Wang and Clark, 2014)): (i) Improvement of the library complexity by using more input RNA, (ii) the use of biological and technical replicates, (iii) independent validation of the candidates, (iv) empirical calculation of the false discovery rate (FDR) cut-off by mock comparisons using the real data, (v) correction of alignment bias and (vi) the introduction of an allelic ratio cut-off. Prior to this thesis, a fully automated, robust and user-friendly bioinformatics pipeline, which defines the allelic expression status for each gene, such as strain-biased, imprinted and biallelic expression was lacking.

2. Genetic causes of allelic expression

Genome wide association studies (GWAS), in the past decade led to the identification of a comprehensive catalogue of genetic loci or variants in human that affect traits, which influence morphology, physiology, behavior or diseases risk (Albert and Kruglyak, 2015). In particular a group of variants, called expression quantitative loci (eQTLs), alter gene expression levels of one or more genes and have been extensively mapped in yeast, plants, rodents and humans (Brem et al., 2002; Schadt et al., 2003). Such eQTLs are determined by analyzing genetically distinct population that can be outbred population or experimental crosses between different genetic backgrounds. For the analysis, each sample has to be first genotyped to detect genomic difference such as SNPs between the alleles, followed by measuring the expression level for each gene in a given cell type. Further statistical tests are than performed to link the sequence variant to gene expression. eQTLs can be categorized by the proximity to

the influenced gene (distant or nearby) or if they affect gene expression *in cis* or *in trans* (Brem et al., 2002; Rockman and Kruglyak, 2006) (Figure 2).

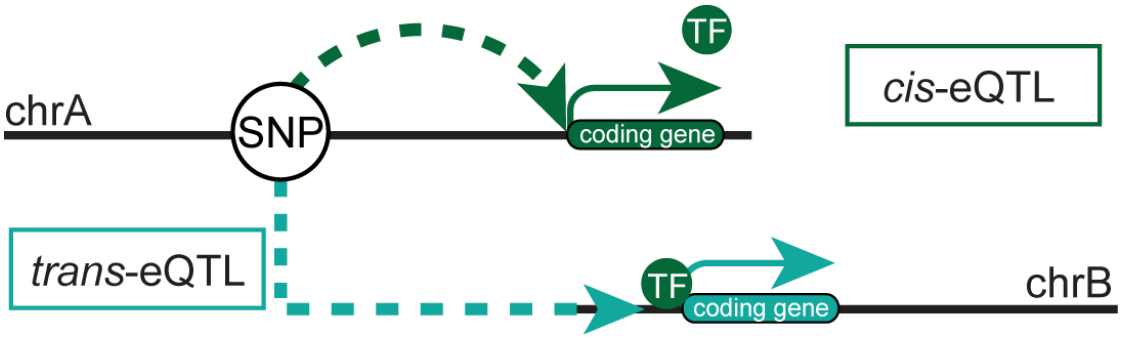


Figure 2. Illustration of how an expression quantitative trait loci (eQTL) may influence gene expression *in cis* or *in trans*. eQTLs that act *in cis* affect gene expression on the same chromosome, one example being SNPs within regulatory regions that disrupt target genes. eQTLs that act *in trans* affect gene expression on another chromosome, for example by disrupting the activity or targets of a transcription factor (TF).

Cis-acting eQTLs affect gene expression on the same chromosome, which leads to allele-specific or strain-biased expression. In contrast, *trans*-acting eQTLs, influence expression of transcription factors or other genes *in cis* that consequently may alter gene expression changes of their targets *in trans*.

2.1 Strain-biased expression in hybrids of genetically distinct mouse strains

Strain-biased gene expression is the result of *cis*-acting eQTLs and has been mapped genome-wide in human and mouse for some tissues using RNA-seq (Keane et al., 2011; Lappalainen et al., 2013; Proudhon and Bourc'his, 2010). F1 hybrids between different inbred mouse strains are a powerful model to study strain-biased expression, given that all the SNPs and deletions between the 17 commonly used laboratory mice have been mapped and can be used to assign RNA-seq reads to the corresponding allele (Keane et al., 2011). Another advantage of the inbred mouse model is the ability to generate biological replicates with the same genetic background, which increases the statistical power in calling allelic expression. High tissue-specific variation for strain-biased expression that might be explained by SNPs causing allelic differences within regulatory regions (Keane et al., 2011) was demonstrated by analyzing different organs from F1 hybrids.

2.2 Regulation of allele-specific expression

In mammals, genetic differences between the alleles may causes differential gene expression by influencing transcription factor binding on promoters or regulatory regions such as enhancers (Leung et al., 2015). In this study allele-specific enhancers

were mapped in human, using chromatin immunoprecipitation sequencing (ChIP-seq) for H3K27ac and this mark correlated with allele-specific expression. A high correlation was observed between enhancers biased towards one allele and nearby allelic expression. Genetic differences between the alleles might also cause changes in the nuclear localization or organization that might lead to allele-specific expression (Amendola and van Steensel, 2014). In addition to genetic variation influencing allele-specific expression at the transcriptional level, genetic difference can also have an affect on post-transcriptional gene regulation, leading to allele-specific splicing, influencing RNA stability or degradation due to allele-specific miRNA binding sides (Gilad et al., 2008; Lappalainen et al., 2013; Li et al., 2012; Majewski and Pastinen, 2011).

3. X-chromosome inactivation (XCI)

Female mammals have two X-chromosomes, whereas males only have one X. In order to compensate for this difference in gene dosage between the sexes, one female X-chromosome is inactivated (Lyon, 1961). Studies in mouse and human defined the X-chromosome inactivation center (Xic), which includes the lncRNA *Xist*, that is expressed exclusively from the inactive chromosome (Borsani et al., 1991; Brockdorff et al., 1991; Brown et al., 1991; Rastan, 1983). The process of XCI includes counting the number X-chromosomes relative to autosomes, choosing and initiating silencing of one entire X-chromosome per diploid set of autosomes, and then maintaining silencing on the inactive chromosome (Augui et al., 2011). The mechanism how each cell chooses one of the two X-chromosomes for silencing is still unclear. Deletion experiments in mouse embryonic stem cells (ESCs) demonstrated that the lncRNA *Xist* is responsible for the initiation of silencing *in cis* by coating the entire chromosome (Marahrens et al., 1997; Penny et al., 1996), but not required for maintaining stable repression (Brown and Willard, 1994; Csankovszki et al., 1999; Wutz and Jaenisch, 2000). Epigenetic processes such as DNA methylation and chromatin modification maintain the stable inactive state of the X-chromosome (Wutz, 2011).

3.1 XCI during mouse development

The first wave of XCI is parental-specific and occurs between the two and four cell stage in female mouse embryos when the paternal X-chromosome is inactivated (Figure 3) (Okamoto et al., 2004). This imprinted XCI is preserved in extra-

embryonic linages, such as the trophectoderm that differentiates to placenta cells types, or the primitive endoderm that originates from the inner cell mass and gives rise to the visceral and parietal yolk sac (VYS and PYS) endoderm (Takagi and Sasaki, 1975; West et al., 1977). Reactivation of the inactive X-chromosome takes place in the inner cell mass (ICM) of the blastocyst from E3.5, cells that will later form the embryo proper (Mak et al., 2004; Okamoto et al., 2004). Both X-chromosomes remain active in the early post-implantation embryo, until a second wave of XCI starting at E5.5, randomly inactivates one of the two X-chromosomes (Mak et al., 2004). The randomly repressed X-chromosome is thereafter transmitted clonally through mitosis (Krietsch et al., 1982; Ohhata and Wutz, 2013). Reactivation of the inactive X-chromosome occurs a second time in the primordial germ cells (PGCs) of E12.5 embryos in order to have two active X-chromosome before the initiation of the oogenesis (Sugimoto and Abe, 2007).

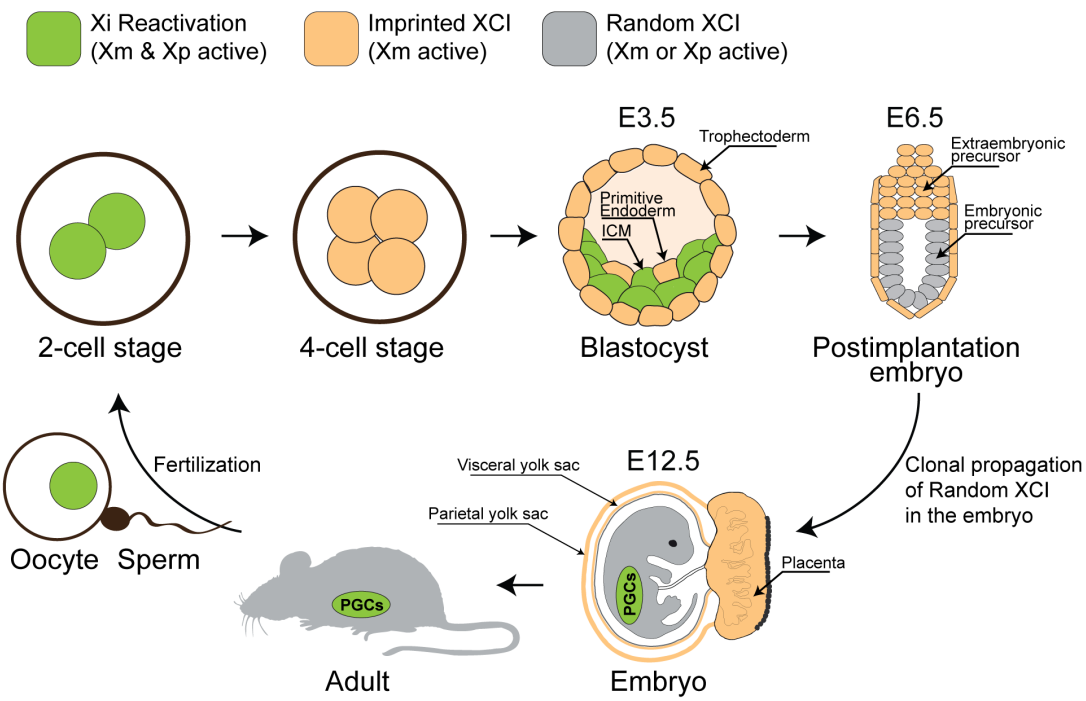


Figure 3. The X-chromosome inactivation (XCI) cycle during female mouse development. The first wave of XCI occurs between the two and four cell stage in female mouse embryos when the paternal X-chromosome is inactivated. While the paternal X-chromosome remains inactivated in extra-embryonic linages, reactivation occurs in the inner cell mass (ICM) of the blastocyst (E3.5), which later forms the embryo. The second wave of XCI occurs randomly in the early post-implantation embryo (E5.5). The inactivated X-chromosome is subsequently transmitted clonally through mitosis. Reactivation of the inactivated X-chromosome occurs a second time in the primordial germ cells (PGCs) of E12.5 embryos.

3.2 Regulation of XCI

The mouse was extensively used as a model to understand the process of random XCI, given that embryonic stem cells (ESCs) with two active X-chromosomes can be derived from the blastocyst and used to study random XCI during *in vitro* differentiation. In female ESCs the lncRNA *Xist* is lowly expressed from both alleles and upon differentiation is upregulated from the future silent X-chromosome in order to initiate silencing (Marahrens et al., 1997; Penny et al., 1996). Another key player in the process of XCI is the lncRNA *Tsix*, which is expressed antisense to *Xist* from the future active X-chromosome and helps maintain silencing of *Xist* (Lee and Lu, 1999). *Xist* is spliced and polyadenylated but remains in the nucleus to coat the entire X-chromosome *in cis*. Recently it was shown that *Xist* reaches distant regions by using the three-dimensional chromosome structure, and then starts spreading from these sides (Engreitz et al., 2013). This process of coating might lead to the formation of a repressive compartment by excluding transcription initiation factors such as RNA polymerase II or the splicing machinery (Chaumeil et al., 2006). The last step of the XCI initiation process is the deposition of repressive marks, such as histone H2A-lysine 119-monoubiquitylation (H2AK119ub1) and H3-lysine 27-trimethylation (H3K27me3) on regulatory regions by the Polycomb Repressive Complex (PRC1 and 2) (Calabrese et al., 2012; Plath et al., 2003). The maintenance of the silent X-chromosome is independent of *Xist* expression and requires DNA methylation on promoters catalyzed by the DNA methyltransferase 1 (DNMT1) (Wutz and Jaenisch, 2000).

The spliced form of the lncRNA *Xist* is 17kb in length and poorly conserved, with the exception of some repetitive regions that have been shown to be essential for silencing. Deletion of the most conserved repeat, the A-repeat within exon 1, showed that *Xist* is still capable of coating the entire X-chromosome, but the genes remain active (Wutz et al., 2002). Additional repetitive regions in exon 1 (F and B repeat) have been recently reported to be essential to recruit Jarid2, a member of PRC2 (da Rocha et al., 2014). Recently a robust method called RNA antisense purification (RAP) combined with quantitative mass spectrometry was developed that allows the identification of direct binding partners to the lncRNA *Xist* (Engreitz et al., 2015; Engreitz et al., 2013; McHugh et al., 2015). This method enabled the identification of 10 proteins, one of them SAF-A (also known as HNRNPU) was already shown to link *Xist* with chromatin (Hasegawa et al., 2010), and SHARP (also know as SPEN) a

protein known to interact with Histone deacetylase 3 (HDAC3). These results suggest that *Xist* can coat the X-chromosomes by binding to SAF-A, which is linked to the chromatin and directly binds SHARP in order to recruit HDAC3. Finally HDAC3 removes histone acetylation from the chromatin, which leads to the repression of active genes and the recruitment of PRC2 (Engreitz et al., 2015). Similar result have been reported by others using screening approaches of haploid ESCs or by using biochemical strategies (Chu et al., 2015; Moindrot et al., 2015; Monfort et al., 2015).

3.3 Skewed XCI between genetically different mouse strains

During female embryonic development one copy of the X-chromosome is randomly silenced in the epiblast (Mak et al., 2004). Genetic variation between the X-chromosome inactivation center (Xic) and skewed XCI ratios have been linked in female F1 hybrids of genetically distinct inbred mouse strains (Cattanach and Isaacson, 1965, 1967). The responsible *cis*-acting region that influences the XCI ratio was mapped and termed the X-chromosome controlling element (Xce) (Cattanach, 1975). The Xce includes the Xic containing the lncRNA *Xist*, which is responsible for the initiation of XCI (Borsani et al., 1991; Brockdorff et al., 1991; Brown et al., 1991; Rastan, 1983). Based on the XCI skewing ratio observed in different F1 hybrids of inbred mouse strains four different strengths of Xce have been categorized (ordered by strength: $Xce^{a} < Xce^{b} < Xce^{c} < Xce^{d}$) (Calaway et al., 2013). In females the X-chromosome with the stronger Xce has a lower chance to be silenced, and thus be active in more cells compared to the X-chromosome with the weaker Xce. The weakest resistance to XCI was observed for the species *Mus musculus domesticus*, which contains, depending on the strain, the Xce^{a} or Xce^{b} allele. For example the 129 strain contains the Xce^{a} allele and the C57BL/6 strain the Xce^{b} allele. The strongest allele was observed for the *Mus musculus castaneus* (Xce^{c}) and *Mus spretus* (Xce^{d}) (Calaway et al., 2013). Therefore the XCI skewing ratio is dependent on the combination of Xce in F1 hybrids, for example, the combination of the Xce^{a} and Xce^{b} alleles results in a 40/60 skewing ratio, and the combination of the Xce^{b} and Xce^{c} alleles results in a 30/70 skewing ratio. Note F1 hybrids containing the same allele, for example, twice Xce^{a} or twice Xce^{c} show no bias in the XCI ratio (Krietsch et al., 1986).

3.4 XCI escaper genes

After XCI occurs, almost all the X-linked genes on the repressed chromosome are silent, with the exception of some escaper genes that are expressed from both alleles (Berletch et al., 2011). In human 15% of escaper genes have been reported based on gene expression profiles from fused human and mouse cells that always inactivate the human X-chromosome (Brown et al., 1997; Carrel and Willard, 2005). In contrast to humans only 3% of genes have been reported to escape XCI in mouse (Yang et al., 2010). In this study RNA-seq was used to identify mouse escaper genes by calculating allele-specific expression from kidney cells derived from F1 crosses (C57BL/6 x *Mus spretus*), and by additionally selected for complete skewing (C57BL/6 X-chromosome 100% silent). Extra-embryonic tissues such as E17.5 placenta and trophoblast stem cells have been used to identify genes that escape imprinted XCI *in vivo* (paternal X-chromosome 100% silent) (Calabrese et al., 2012; Finn et al., 2014). However genes that escape random XCI in somatic tissues can be detected as genes that deviate from the expected XCI skewing ratio, or genes that do not show the complete skewing expected when on allele of the *Xist* genes is deleted (X-chromosome with deletion of *Xist* 100% active) (Deng et al., 2013). Recent studies used this approach to identify escaper genes in several adult organs such as brain, spleen and ovary (Berletch et al., 2015), or during mouse embryonic stem cell differentiation (Marks et al., 2015), and found that some XCI escaper genes can be detected in multiple tissues while others escape in a tissue-specific manner (Berletch et al., 2015). Another difference between mouse and human is the distribution of escaper genes over the X-chromosome. In mouse escaper genes are distributed randomly over the entire chromosome, whereas in human the majority of escapers are located within one large domain (Carrel and Willard, 2005; Yang et al., 2010). These findings suggest that in human large chromatin domains regulate escaper genes, whereas in mouse each individual escaper has the potential to escape XCI, independent of the location on the X-chromosome (Berletch et al., 2011). Given that escaper genes are highly tissue-specific a more comprehensive set of tissues has to be analyzed in order to understand the organization and the escape mechanism in mouse and human.

4. Parental-specific expression

Mammals have two copies of each chromosome, with the exception of sex chromosomes in males. One set of chromosomes is inherited from the mother and one from the father. In such a diploid system each gene is present twice and commonly expressed from both alleles. The epigenetic process of genomic imprinting leads to parental-specific gene expression in a diploid cell, even though both gene copies are present. Parental-specific expression was demonstrated only for a few hundred genes of approximately 25,000 genes in human and mouse (Lander et al., 2001; Mouse Genome Sequencing et al., 2002). Genomic imprinting is a consequence of parental inheritance and thus independent of sex. Evidence for the biological importance of genomic imprinting in mammals was provided by pronuclear injection experiments, which demonstrated that both parental copies of the genome are needed in order to develop a functional embryo (McGrath and Solter, 1984; Surani et al., 1984). Further genetic deletion experiments in mouse demonstrated that specific regions in the genome behave differently depending on their parental origin (Cattanach and Kirk, 1985). In 1991 the first three imprinted genes were discovered in the mouse: the first *Igf2r* (insulin-like growth factor 2 receptor), the second the *Igf2* growth factor itself and the third the lncRNA *H19* (Barlow et al., 1991; Bartolomei et al., 1991; DeChiara et al., 1991). Since that time approximately 150 mouse imprinted genes, and about half of this number in human, have been reported in different tissues and development stages and listed in imprinted databases such as Harwell and Otago (Glaser et al., 2006; Williamson CM, 2013). From the 150 mouse imprinted genes 126 are reported in both databases and are annotated in the NCBI Reference Sequence Database (RefSeq), a comprehensive, well-annotated and non-redundant gene annotation (Pruitt et al., 2014). From the remaining imprinted genes, 33 are disputed from the literature because of non-reproducible data, or doubt cast by the identified maternal expression being restricted to the placenta, which could be explained by maternal blood or decidua contamination (Glaser et al., 2006; Okae et al., 2012; Proudhon and Bourc'his, 2010).

4.1 Parental-specific DNA methylation regulates imprinted expression

Parental-specific gene expression is controlled by gametic differential methylated regions or gDMRs at imprint control regions, where differences in the DNA methylation pattern between oocyte and sperm are established. Notably, the majority of gDMRs are found on the maternal chromosome (17 maternal and 4 paternal)

(Hanna and Kelsey, 2014; Proudhon et al., 2012; Xie et al., 2012). Interestingly it is the unmethylated ICE that is associated with gene silencing *in cis* of entire gene clusters (Barlow, 2011; Bartolomei and Ferguson-Smith, 2011). In mouse, methylation of the ICE in the germline is set during the maturation of the oocyte and up to E18.5 in sperm, by the *de novo* DNA methylation complex DNMT3A and DNMT3L (Figure 4) (Bourc'his et al., 2001; Hanna and Kelsey, 2014; Kaneda et al., 2004; Morgan et al., 2005). Immediately after fertilization, before the pronuclei fuse together, the maternal pronucleus is passively demethylated while the paternal is actively demethylated. During this process the imprints on the ICE are not erased, but stably maintained during mitosis by the DNA methyltransferase DNMT1 (Li et al., 1993; Morgan et al., 2005; Ooi et al., 2009). In somatic cells the imprints are maintained during the entire lifespan, while in primordial germ cells (PGCs) the imprints are erased at E12.5 by an unknown mechanism in order to reset them for the next generation during gametogenesis (Lee et al., 2002).

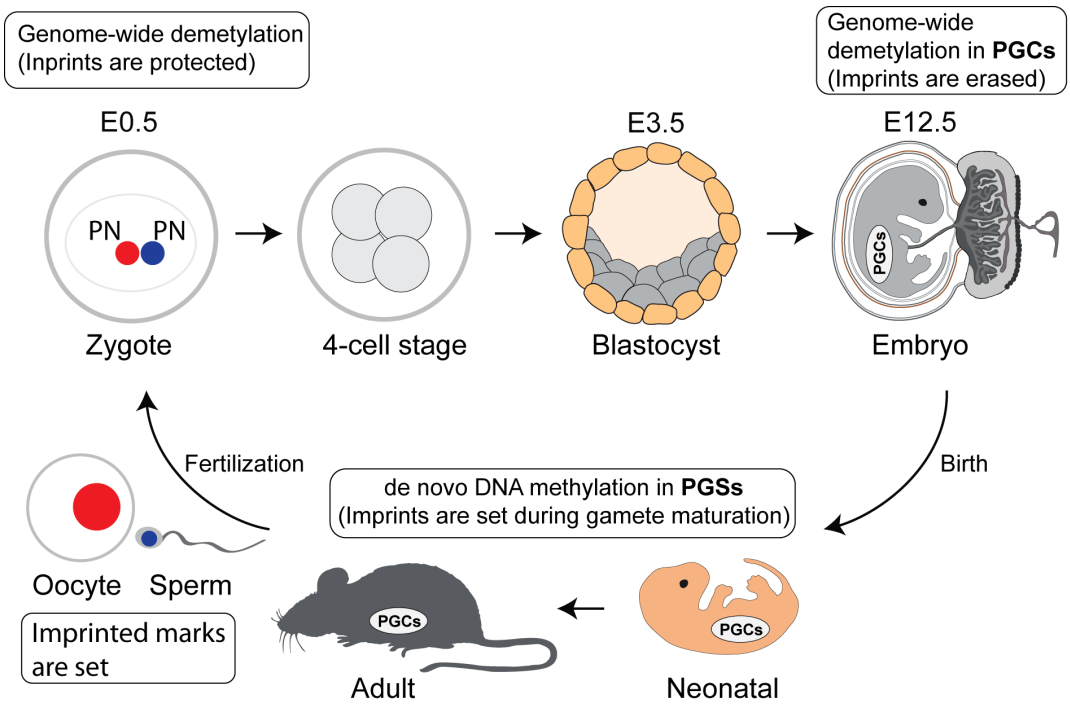


Figure 4. The methylation cycle of the imprint control element (ICE) during mouse development. In mouse, after fertilization before the pronuclei (PN) fuse together, the maternal pronucleus (red) is passively demethylated while the paternal pronucleus (blue) is actively demethylated. During this process the imprints are not erased. In somatic cells the imprints are maintained during the entire lifespan, while in primordial germ cells (PGCs) the imprints are erased at E12.5 in order to reset them for the next generation during gametogenesis. Methylation of the ICE in the germline is then set by the *de novo* DNA methylation complex DNMT3A and DNMT3L during postnatal maturation of the oocyte, and in the male gonad until E18.5 prior to meiosis.

4.2 Imprinted genes are mainly organized in clusters

From the 150 imprinted genes reported to day more than 80% are organized in imprinted gene clusters (Barlow, 2011; Bartolomei and Ferguson-Smith, 2011; Ferguson-Smith, 2011). Imprinted genes that lie outside of known imprinted clusters are defined as "solo" imprinted genes. Imprinted gene clusters or imprinted solo genes are under the control of gDMRs, 17 maternal, which are set during oogenesis and 4 paternal, acquired during spermatogenesis (Proudhon et al., 2012; Xie et al., 2012) (Figure 5). Such gDMRs are presumed to be the imprint control element (ICE) of the entire cluster or directly control the promoter of solo imprinted genes. To date the ICE was defined for seven clusters (*Igf2r-Airn*, *Kcnq1*, *Pws/As*, *Gnas*, *Igf2-H19*, *Dlk1, Grb10*) by genetic deletion of the gDMR, which resulted in biallelic expression in the affected imprinted gene cluster (Bressler et al., 2001; Fitzpatrick et al., 2002; Lin et al., 2003; Shiura et al., 2009; Thorvaldsen et al., 1998; Williamson et al., 2006; Wutz et al., 1997)

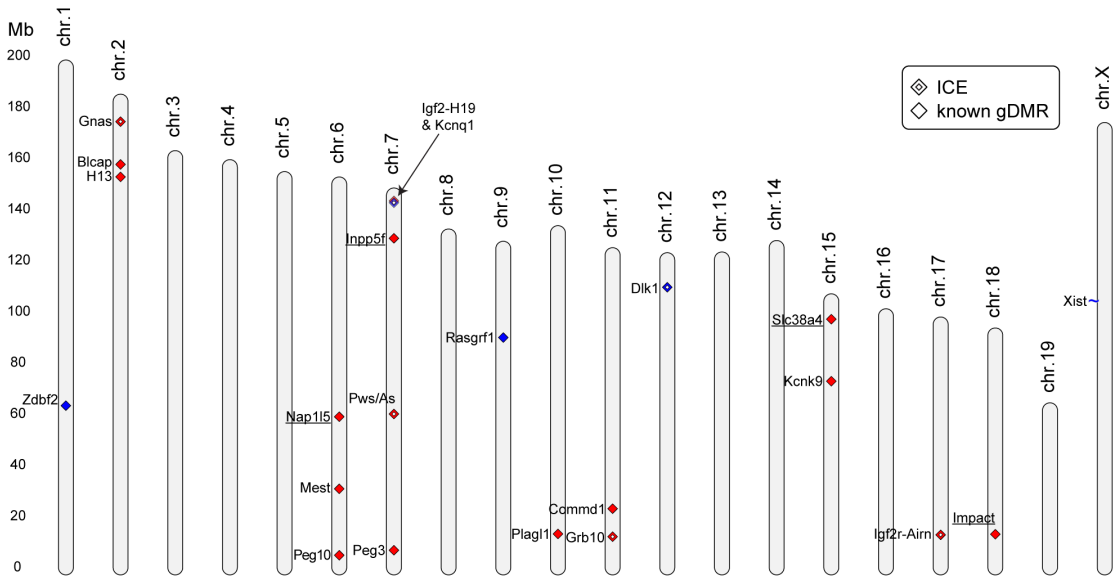


Figure 5. Mouse chromosomes indicating known imprint control element (ICE) and germline differently methylated regions (gDMRs). Genetic deletion experiments demonstrated seven gDMRs to be the imprint control element of imprinted gene clusters. Solo imprinted genes are underlined; gDMRs in red show maternal methylation and in blue paternal methylation. The base pairs coordinates (Mb) are shown on the left side.

These genetically defined clusters vary between 80 – 3700kb in their genomic size and contain between 3 and 12 imprinted protein-coding genes and at least one imprinted lncRNA (exception: *Grb10* cluster) (Barlow, 2011; Guenzl and Barlow, 2012). Another feature of these clusters is that protein-coding genes are expressed from one parental allele, whereas the imprinted lncRNA is expressed from the

opposite allele. Interestingly deletion of the ICE restores biallelic expression only if inherited from the parental allele that expresses the imprinted lncRNA. This indicated that the lncRNA is responsible for silencing the protein-coding genes and was later confirmed by truncating the lncRNAs *Airn, Kcnq1ot1* and *Nespas*, in the *Igf2r*, *Kcnq1* and *Gnas* cluster (Mancini-Dinardo et al., 2006; Sleutels et al., 2002; Williamson CM, 2013). The promoter of these three lncRNAs sits in the ICE region and is therefore directly regulated by parental-specific DNA methylation. In contrast, the promoter of lncRNA *H19* in the *Igf2-H19* cluster is not within the gDMR and not involved in imprinted silencing (Arney, 2003). In the *Igf2* cluster, CTCF binds to the unmethylated gDMR on the maternal allele and blocking access of *Igf2* and *Ins2* to distant enhancers. The same enhancers then exclusively promote *H19* expression on the maternal allele. On the paternal allele CTCF cannot bind to the methylated gDMR and thus not block enhancer access resulting in paternal expression of *Igf2* and *Ins2*. The methylated gDMR might silence *H19* through DNA methylation, by an unknown mechanism (Bell and Felsenfeld, 2000). Although there are different models how ICEs regulate imprinted expression of large gene clusters, two of which are detailed above, methylation on the ICE is associated with protein-coding gene expression and lncRNA repression.

4.3 Imprinted expression is highly tissue-specific

Several decades of work have characterized imprinted expression in different tissues and developmental stages of the mouse, work that has been annotated by imprinted databases that list 150 imprinted genes (Glaser et al., 2006; Williamson CM, 2013). From those that have been tested in multiple tissues approximately one third shows tissue-specific imprinted expression, restricted to extra-embryonic tissues such as the placenta and visceral yolk sac (VYS), or to specific brain regions (Prickett and Oakey, 2012). A major problem in the identification of tissue-specific imprinted expression in a whole organ is the masking of parental-specific expression in one cell type from biallelic expression in a neighboring tissue (Kulinski et al., 2013). However the surrounding tissue can also lead to the identification of falsely imprinted genes. This was observed for the placenta, an organ surrounded by maternal tissues such as decidua and blood, which can lead to the mistaken identification of maternally imprinted genes due to maternal contamination (Okae et al., 2012; Proudhon and Bourc'his, 2010).

Since the advent of RNA-seq, many groups used F1 hybrids to identify genome-wide imprinted expression in different tissue and developmental stages including: E9.5 embryo (Babak et al., 2008), embryonic and adult brain (Gregg et al., 2010b; Wang et al., 2008), E17.5 placenta (Wang et al., 2011), mouse embryonic fibroblasts (MEFs) (Tran et al., 2014) and trophoblast stem cells (Calabrese et al., 2015). Recently a study mapped imprinted expression in 33 mouse tissues and development stages, including 26 novel assayed tissues, combined with the 7 published datasets (listed above) and found the highest number of imprinted genes in embryonic, extra-embryonic and brain tissues, which is in agreement with the proposed role for genomic imprinting during development and in maternal behavior (Babak et al., 2015). In addition this study observed that the majority of genes show imprinted expression during early development and either maintain it in adult or lose it entirely, whether this was due to loss of expression or switching from imprinted to biallelic expression was not investigated. Although many organs have been tested for imprinted expression, reviews suggest a role for genomic imprinting during neonatal suckling, maternal care and thermogenesis (Peters, 2014; Stringer et al., 2014). In order to test this suggestion, tissues involved in neonatal suckling such as mammary glands and neonatal tongue or brown adipose might be interesting tissues to investigate in the future.

4.3.1 The *Igf2r* cluster shows tissue-specific regulation

The imprinted *Igf2r* cluster on chromosome 17 in mouse is a well-studied powerful model to understand tissue-specific regulation of imprinted expression (Figure 6). The cluster contains the four maternally expressed protein-coding genes *Igf2r* (insulin-like growth factor 2 receptor), *Slc22a2* (solute carrier family 22 member 2), *Slc22a3* (solute carrier family 22 member 3), *Pde10a* (phosphodiesterase 10A) and the paternally expressed lncRNA *Airn* (antisense *Igf2r* non-coding RNA) (Andergassen, 2012; Barlow et al., 1991; Zwart et al., 2001).

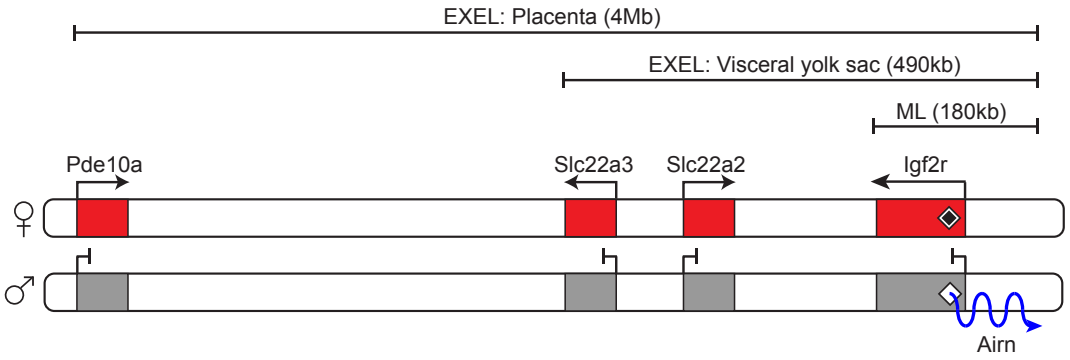


Figure 6. The *Igf2r* cluster in mouse shows tissue-specific regulation of imprinted expression. *Igf2r* and the lncRNA *Airn* show multi-linage (ML) imprinted expression in nearly every cell type, while *Slc22a2*, *Slc22a3* and *Pde10a* show extra-embryonic-linage (EXEL) specific expression. Genes in red show maternal expression while genes in blue are expressed paternally and grey indicates not expressed.

Truncation of the lncRNA *Airn* to 5% of its length restores biallelic expression of the 4 maternally expressed protein-coding genes, which demonstrates that the lncRNA silences the entire cluster *in cis* (Andergassen, 2012; Santoro et al., 2013; Sleutels et al., 2002; Zwart et al., 2001). The promoter of *Airn* sits within the 3.7kb gDMR in intron 2 of *Igf2r*, and gains methylation on the maternal allele during oogenesis, resulting in paternal expression of *Airn* (Lyle et al., 2000; Stoger et al., 1993). Deletion of the gDMR on the paternal allele shows the same phenotype as truncating the lncRNA *Airn*, namely restoring biallelic expression of the entire cluster and thus demonstrated it to be the ICE (Sleutels et al., 2002; Wutz et al., 1997; Wutz et al., 2001). *Airn* and *Igf2r* show imprinted expression in almost every tissues, with the exception of embryonic stem cells, testis and post-mitotic neurons, and thus show multi-linage (ML) imprinted expression (Lerchner and Barlow, 1997; Szabo and Mann, 1995a, b; Yamasaki et al., 2005). In VYS an extra-embryonic-lineage (EXEL), the cluster size expands to 490kb including *Airn*, *Igf2r*, *Slc22a2* and *Slc22a3* (Hudson et al., 2011). Genome-wide mapping of imprinted expression in E17.5 placenta, identified the maternally expressed gene *Pde10a*, approximately 4Mb distant from the *Igf2r* cluster that was classified as a solo imprinted gene (Wang et al., 2011). During my diploma work, I could reproduce maternal expression of *Pde10a* in E12.5 placenta and convincingly show regulation by *Airn*, by genetically demonstrating reactivation of *Pde10a* from the silent allele in the absence of the lncRNA (Andergassen, 2012). This result expands the *Igf2r* cluster size to 4Mb in placenta including the genes *Airn*, *Igf2r*, *Slc22a3* and *Pde10a* (Andergassen, 2012). In summary the *Igf2r* cluster shows tissue-specific regulation of imprinted expression, resulting in expansion and contraction of the cluster size.

4.4 Regulation of Imprinted clusters by repressive lncRNAs

Recently a study in human reported 60,000 lncRNAs by analyzing thousands of tumors, tissues and cell line samples (Iyer et al., 2015). LncRNAs are by definition longer than 200 base pairs and non-protein-coding. The largest proportion of lncRNAs is fully spliced, located in intergenic regions and has been suggested to regulate gene expression *in trans* (Guttman et al., 2011; Quinn and Chang, 2015). In contrast, imprinted lncRNAs previously called 'macro' lncRNAs are much longer (10-100kb), inefficiently spliced, unstable, mainly nuclear localized and regulate gene expression *in cis* (Guenzl and Barlow, 2012; Koerner et al., 2009). The majority of imprinted lncRNAs are transcribed from the unmethylated ICE and act *in cis* to prevent upregulation of all genes that belong to the imprinted cluster. This was demonstrated by truncating the lncRNAs *Airn, Kcnq1ot1* and *Nespas*, in the *Igf2r*, *Kcnq1* and *Gnas* cluster to less than 5% of their length. This resulted in a loss of imprinted silencing for all genes in the cluster, both genes overlapped by the lncRNA and distant non-overlapped genes (Mancini-Dinardo et al., 2006; Sleutels et al., 2002; Williamson CM, 2013). The *Igf2r* cluster provides a powerful model to investigate the mechanism of short and long range silencing, given that the lncRNA *Airn* silences the overlapping gene *Igf2r* in almost every tissue and the distant non-overlapping genes *Slc22a2*, *Slc22a3* and *Pde10a* only in extra-embryonic tissues (Andergassen, 2012; Barlow et al., 1991; Hudson et al., 2011; Zwart et al., 2001). Experiments that truncated the imprinted lncRNA *Airn* to different lengths demonstrated that transcription over the *Igf2r* promoter alone, is sufficient for silencing (Latos et al., 2012) (Figure 7A). This finding demonstrates that the sequence of *Airn* is not important to silence overlapping genes and represents the first example of transcription interference in mammals, a well-established mechanism in other species such as bacteria, yeast and flies (Kornienko et al., 2013; Mazo et al., 2007). However other experiments indicated that the sequence of *Airn* might be necessary in order to silence the non-overlapping distant gene *Slc22a3* in placenta. According to this model *Airn* localizes to the *Slc22a3* promoter and recruits the H3K9dimethylase EHMT2 (Euchromatic histone-lysine N-methyltransferase 2), leading to the deposition of H3K9me3 methylation marks and transcriptional repression (Figure 7B) (Nagano et al., 2008). However, this result might also be explained by the enhancer interference model (reviewed in (Pauler et al., 2012)). Under this model the essential enhancers required for placental expression of the distant genes *Slc22a3* and *Pde10a* are located

in the gene body of *Airn* and form active loops to its promoters. Transcription of *Airn* on the paternal allele might than interfere with the binding of transcription factors to enhancers and thus block the formation of an active loop (Figure 7C). This model has parallels to a new model proposing how the *cis*-acting lncRNA *Xist* targets distant genes. According to this model the lncRNA *Xist* reaches distant regions by exploding the three-dimensional chromosomes structure, followed by deposition of H3K27me3 by PRC2 (Figure 7D) (Engreitz et al., 2013).

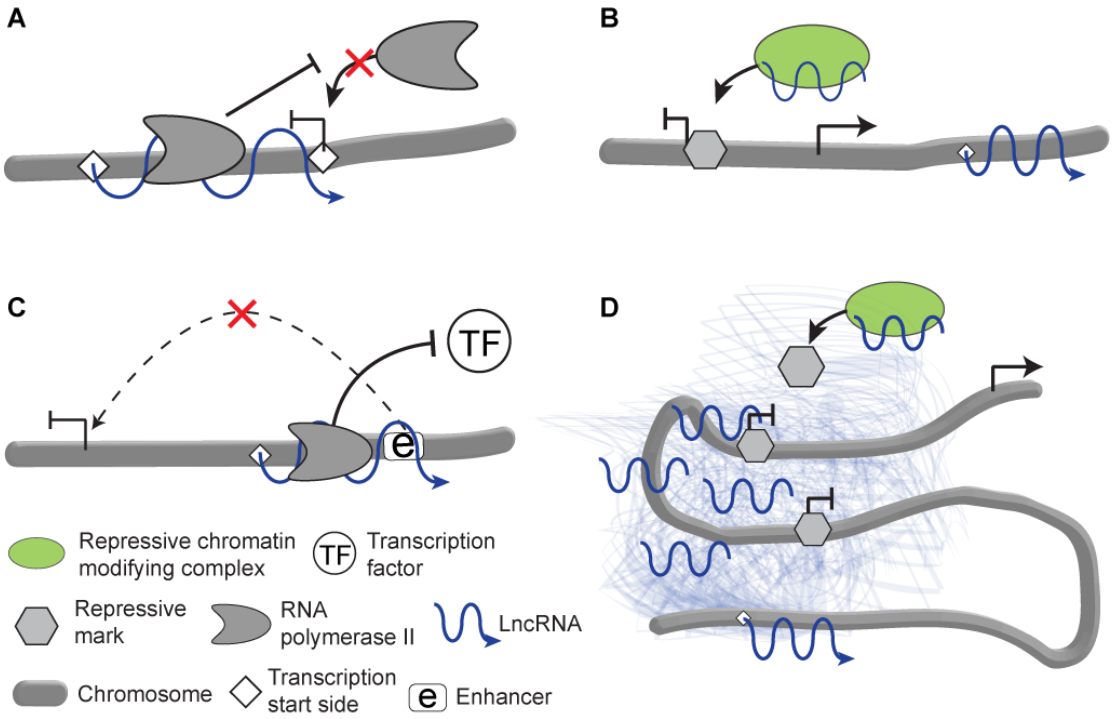


Figure 7. Silencing mechanism of *cis*-acting lncRNAs. (**A**) LncRNA silencing of overlapping genes by transcription interference. LncRNA silencing of distant target genes might use one of the proposed mechanisms: (**B**) LncRNA mediated targeting of repressive chromatin modifiers to its targets (Nagano et al., 2008) (**C**) Disrupting the enhancers activity of target genes that are located in the lncRNA gene body by transcription interference (Pauler et al., 2012) (**D**) LncRNA locus interacts with its targets and recruits repressive chromatin modifiers (Engreitz et al., 2013).

5. Aim of the PhD project

The first goal of my PhD project was to generate a user-friendly, fully automated bioinformatics pipeline, which robustly detects allele-specific expression or chromatin modification from high-throughput sequencing data, generated from tissues from F1 crosses from inbred mouse strains. This method should be able to characterize allele-specific expression from all genes in one cell type into biallelic, strain-biased, or imprinted.

The second goal of my PhD project was to apply the pipeline on a range of mouse tissues including pluripotent, embryonic, extra-embryonic, neonatal and adult tissues, to generate a comprehensive picture of allelic expression during development, "the mouse Allelome". The generated dataset will provide the most comprehensive survey of allele-specific expression, including all the genes showing biallelic or strain-biased expression, all the imprinted genes and all the X-chromosome escaper genes. The complete picture of all the imprinted genes in the mouse will help us to understand the genomic organization of imprinted genes and the identification of the key tissues or developmental stage in which imprinted expression might have an important role. The identification of all the X-chromosome inactivation escaper genes in a comprehensive set of female tissues might reveal the mechanism how genes escape the silencing of a whole chromosome. Finally the identification of strain-biased expression might be useful to identify the genes that cause the phenotype between different mouse strains in a tissue-specific manner. In summary this project will help us to understand how allele-specific expression is regulated during the mouse development.

RESULTS

1. Publication 1: "Allelome.PRO, a pipeline to define allele-specific genomic features from high-throughput sequencing data"

Authors:

Daniel Andergassen†, Christoph P. Dotter†, Tomasz M. Kulinski, Philipp M. Guenzl, Philipp C. Bammer, Denise P. Barlow, Florian M. Pauler* and Quanah J. Hudson*

† These authors contribute equally to the paper as first authors
* Corresponding authors

Published in Nucleic Acids research, 2015, Vol. 43, No. 21, doi: 10.1093/nar/gkv727

Received March 09, 2015; Revised June 09, 2015; Accepted July 06, 2015; first published online July 21, 2015

Open Access Article

Impact factor: 9.112

Article webpage: http://nar.oxfordjournals.org/content/early/2015/07/21/nar.gkv727.full

1.1 Contributions

D.A., C.P.D., T.M.K., Q.J.H., F.M.P. and D.P.B. planed the study and provided intellectual input. D.A., C.P.D., Q.J.H., F.M.P. and D.P.B. wrote the manuscript. D.A. and C.P.D. programmed the Allelome.PRO software using the programming languages Bash, Perl, R and prepared the figures for the manuscript. D.A. isolated and cultured mouse embryonic fibroblasts and isolated RNA. Q.J.H. performed the ChIP for the chromatin mark H3K4me3. P.M.G. and P.C.B generated the RNA-seq and ChIP-seq libraries. All authors read and approved this manuscript.

Nucleic Acids Research Advance Access published August 3, 2015

doi: 10.1093/nar/gkv727

Allelome.PRO, a pipeline to define allele-specific genomic features from high-throughput sequencing data

Daniel Andergassen†, Christoph P. Dotter†, Tomasz M. Kulinski, Philipp M. Guenzl, Philipp C. Bammer, Denise P. Barlow, Florian M. Pauler* and Quanah J. Hudson*

CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Lazarettgasse 14, AKH BT 25.3,1090 Vienna, Austria

Received March 09, 2015; Revised June 09, 2015; Accepted July 06, 2015

ABSTRACT

Detecting allelic biases from high-throughput sequencing data requires an approach that maximises sensitivity while minimizing false positives. Here, we present Allelome.PRO, an automated user-friendly bioinformatics pipeline, which uses high-throughput sequencing data from reciprocal crosses of two genetically distinct mouse strains to detect allele-specific expression and chromatin modifications. Allelome.PRO extends approaches used in previous studies that exclusively analyzed imprinted expression to give a complete picture of the 'allelome' by automatically categorising the allelic expression of all genes in a given cell type into imprinted, strain-biased, biallelic or non-informative. Allelome.PRO offers increased sensitivity to analyze lowly expressed transcripts, together with a robust false discovery rate empirically calculated from variation in the sequencing data. We used RNA-seq data from mouse embryonic fibroblasts from F1 reciprocal crosses to determine a biologically relevant allelic ratio cutoff, and define for the first time an entire allelome. Furthermore, we show that Allelome.PRO detects differential enrichment of H3K4me3 over promoters from ChIP-seq data validating the RNA-seq results. This approach can be easily extended to analyze histone marks of active enhancers, or transcription factor binding sites and therefore provides a powerful tool to identify candidate *cis* regulatory elements genome wide.

INTRODUCTION

Mammalian cells are diploid and thus contain two copies of every gene locus, one inherited from the male, and one from the female parent. Mitochondrial genes, plus genes on the sex chromosomes in males, are the only exception to this rule. Since each diploid gene locus has the possibility to be expressed independently from either parental chromosome, different allelic states of expression can arise. The majority of mouse genes are considered to show equal or 'biallelic' expression from both parental alleles based on the absence of parental-specific phenotypes in the majority of genes analyzed by gene knockout (1). Genes that deviate from biallelic expression by showing preferential expression of one of the two parental alleles are described as showing 'monoallelic' expression. To date, only a small subset of mammalian genes is known to show monoallelic expression. When either parental allele can show preferential expression, this is known as random monoallelic expression (RMAE). However, when one parental allele consistently and heritably shows preferential expression, this is known as parental-specific or imprinted monoallelic expression (IMAE).

Random monoallelic expression has been shown to affect clustered gene families, such as the allelic exclusion of the B- and T-cell receptor genes that allows clonal lymphocytes to express a single receptor with a unique specificity (2), the 'singular' expression of the clustered olfactory receptor genes that allows neurons to discriminate olfactory signals (3), and more recently, the stochastic monoallelic expression of the cadherin-related PCDH neuronal receptor clusters that may act in neuronal self recognition (4). All X-chromosome linked genes in female placental mammals also show random monoallelic expression, due to RMAE of a single locus containing the *Xist* long non-coding (lnc)

*To whom correspondence should be addressed. Tel: +43 1 40160 70 030; Fax: +43 1 40160 970 000 Email: QHudson@cemm.oeaw.ac.at
Correspondence may also be addressed to Florian M. Pauler. Tel: +43 1 40160 70 030; Fax: +43 1 40160 970 000; Email: FPauler@cemm.oeaw.ac.at
†These authors contributed equally to the paper as first authors.
Present addresses:
Christoph P. Dotter, Institute of Science and Technology Austria, 3400 Klosterneuburg, Austria.
Tomasz M. Kulinski, Institute of Biochemistry and Biophysics, PAS, 02-106 Warsaw, Poland.
Philipp Bammer, Friedrich Miescher Institute for Biomedical Research, 4058 Basel, Switzerland.

RNA, which controls X-chromosome inactivation (5). The X-chromosome can also display imprinted paternal-specific inactivation in some rodent extra-embryonic tissues, due to preferential paternal expression of the *Xist* lncRNA (6). In all these cases, RMAE can occur in inbred mouse strains, and thus can be initiated from genetically identical parental alleles, indicating an epigenetic mechanism.

In contrast to the clustered gene families mentioned above that use RMAE to generate specificity in clonal cells, up to 10% of solo autosomal genes were reported to show RMAE in isolated cell lines that could be stably propagated (7,8). Similarly, an estimated 12–24% of expressed genes showed monoallelic gene expression in single cells of F1 mouse pre-implantation embryos, indicating that this could be a widespread phenomenon that may play a role in generating diversity in individual cells (9). In cases of true RMAE an important point to bear in mind is that, although a gene may show monoallelic expression that can be detected by single cell assays, at the population level the gene will appear biallelic if the allele expressed is random in each cell.

In addition to IMAE and RMAE, a third category of monoallelic expression that may occur in outbred individuals is non-random monoallelic expression or strain bias. Such strain bias may occur due to genetic differences between the alleles that affect expression of certain genes. For example, expression differences could arise from nucleotide polymorphisms influencing the interaction of promoters and enhancers with transcription factors and thereby affecting transcription rates. Such polymorphisms could also act at a post-transcriptional level by influencing miRNA binding and RNA stability, or allele-specific processing, such as alternative splicing or alternative UTR generation (10–12). The *Xist* lncRNA that controls X-chromosome inactivation in female cells can also show a strain bias due to genetic variation at the X-inactivation center (*Xic*) locus that influences the likelihood of *Xist* being expressed from that chromosome (13). *Mus musculus castaneus* (CAST/EiJ) mice are known to possess a stronger *Xic* allele than *Mus musculus domesticus*, thus in the FVB/N x CAST/EiJ reciprocal crosses used in our study, the FVB/N X-chromosome will be preferentially inactivated (13,14).

Imprinted monoallelic expression primarily affects small clusters of unrelated genes (15). Currently 96 of the 123 known imprinted genes either lie in genetically characterized imprinted clusters, or, due to their close proximity are likely to lie in clusters (Supplementary Table S2). Thus most imprinted genes are clustered. A novel feature of several gene clusters showing IMAE in contrast to those showing RMAE, is their association with a long non-coding (lnc) RNA (16–18), that in four cases has been shown to induce imprinted gene silencing (reviewed in (15)). While some solo genes clearly show imprinted expression, the imprinted status of many has been challenged (19–21). Thus, the number of solo imprinted genes is not yet known. The defining characteristic of an imprinted gene is preferential expression from one parental chromosome. However, the exact ratio of parental-specific expression that constitutes imprinted expression has not yet been defined. The total number of known imprinted genes is also relatively low, only ∼0.5% of protein-coding genes and approximately equal numbers of maternally-expressed and paternally-expressed imprinted genes are known. This total number of imprinted genes was obtained from examination of a limited set of tissues such as embryo, placenta and fetal brain that are predicted to use imprinted gene expression to regulate pre- and post-natal growth of the mammalian embryo (22–24). However, it has only recently been appreciated that imprinted expression shows considerable tissue-specificity (25) and also developmental regulation (15). Given that only a limited number of tissues and developmental stages have been assayed so far, and even fewer studies of different mammalian taxonomic strains conducted, it is not known if the total number of imprinted genes has been underestimated. This possible underestimation of the total number of imprinted genes has implications for understanding the biological function of imprinted gene expression in mammals.

In recent years many studies have used high-throughput RNA sequencing (RNA-seq) of tissues from reciprocal crosses between genetically distinct inbred mouse strains to identify imprinted expression (26–30). These studies based on a few tissue types only found a small number of novel imprinted genes compared to those listed in publically available databases (www.otago.ac.nz/IGC). In contrast, one study reported parental-specific expression of 1300 transcripts in embryonic and adult mouse brain (31). However, a subsequent study indicated that the vast majority of these transcripts were false positives, and emphasized the need for careful controls including the use of biological replicates, the need to empirically determine the false positive rate, and the need for independent validation of the imprinted status of the gene (32). With these three requirements in mind, we developed Allelome Profiler (Allelome.PRO), an automated and user-friendly bioinformatic pipeline based on a previously described method (32,33), but modified to improve the robustness and sensitivity of imprinted expression detection, and also to detect strain bias gene expression as well as biallelically expressed and silent genes. Critically, in addition to a false discovery rate cutoff based on a statistical score, we introduced an allelic ratio cut-off for both parental and strain bias that removes loci showing a minor allelic bias with high sequencing coverage, thus enabling the allelic status of all genes to be categorised. This cut-off was determined from the expression patterns of known imprinted genes and from X-linked genes on X-chromosomes showing skewed X-inactivation. We use primary mouse embryo fibroblasts (MEFs) and we define different allelic states of expression as imprinted, strain-biased, biallelic, non-informative (due to low or no expression) or having no single nucleotide polymorphisms (SNPs). We also show that Allelome.PRO can detect allelic differences in high-throughput chromatin immunoprecipitation sequencing (ChIP-seq) data and demonstrate that H3K4me3, a promoter mark associated with active transcription, can be used as an independent validation of the RNA-seq allelome. Together this approach allows a high-resolution analysis of the entire allelome of any cell type and has the potential to expand our understanding of genetic and epigenetic mechanisms underlying IMAE, RMAE and the phenotypic differences between strains.

MATERIALS AND METHODS

Generation of mouse embryonic fibroblasts (MEFs)

CAST/EiJ (CAST) mice were purchased from the Jackson Laboratory (www.jax.org) and FVB/NJ (FVB) from Charles River to generate reciprocal crosses. After reciprocal mating (CASTxFVB and FVBxCAST), mouse embryonic fibroblasts (MEFs) were derived from E12.5 embryos after removing the head, viscera and urogenital system. The remaining carcass was homogenised to a single cell suspension using Trypsin/EDTA (Gibco) and plated on 6 cm dishes. Female MEFs from passage 2 of a confluent 10 cm plate were used for the RNA analysis, whereas MEFs from passage 5 of three confluent T175 cm^2 flasks were used for ChIP. The sex of the embryos was determined by PCR combining a Y-chromosome specific assay and an autosomal assay (34).

RNA and ChIP-seq sample preparation

Total RNA and DNA were extracted using TRI-reagent (Sigma–Aldrich T9424) according to the manufacturers protocol. Total RNA was DNaseI treated using the DNA-Free™ kit (Ambion). Ribosomal RNA was depleted from total DNaseI-treated RNA using the RiboZero rRNA removal kit (Human/Mouse/Rat) (Epicentre). Strand-specific RNA-seq libraries were prepared employing the TruSeq RNA Sample Prep Kit v2 (Illumina) modified as described for strand-specific sequencing (35). Native ChIP for H3K4me3 (antibody: cat. 07-473, lot 2019729, Millipore) was conducted as described (36). ChIP-seq libraries were prepared using the TruSeq ChIP Sample Prep Kit (Illumina). 100 bp paired end sequencing for RNA-seq and 50 bp single end sequencing for ChIP-seq were performed by the Biomedical Sequencing Facility (BSF) in Vienna using the Illumina HiSeq 2000 platform.

Alignment of sequencing data

Raw RNA sequencing data was aligned using STAR (version 2.3.1z12) (37), GSNAP (version 2014.07.04) (38) and TopHat (version 2.0.12, bowtie 2.2.3) (39) to allow a comparison between the three aligners. Reads mapping to multiple locations were excluded using specific parameters (STAR: –outFilterMultimapNmax 1), combining only output files that contained uniquely aligned reads (GSNAP), or by removing secondary alignments identified by the SAM flag (TopHat). Additional parameters for the STAR alignment were a maximum intron size of 100 000 bp and outfiltering of non-canonical splice junctions. SNP-tolerant alignment was enabled for GSNAP by providing information about the SNP variants between the two crosses. TopHat was run using a RefSeq based transcriptome index and parameters chosen to exclude novel junctions as well as novel insertions and deletions. As sequencing was done in a strand specific manner, the aligned reads were subsequently separated according to strand using a custom Perl script. STAR alignment of ChIPseq data was conducted with different parameters to disable spliced reads, i.e. a maximum intron size of 1, and prevent soft clipping by enforcing end-to-end alignment (–alignEndsType EndToEnd). All aligned BAM files were sorted afterwards using SAMtools (version 0.1.19).

Preparation of annotation files

The NCBI RNA reference sequences collection (RefSeq) annotation was downloaded from the UCSC genome browser on 2 July 2014. Transcripts <100 bp were removed and the remaining transcripts were separated by transcriptional orientation and used for strand specific analysis of RNA-seq by Allelome.PRO. An annotation of ±2 kb windows around the transcription start site (TSS) of RefSeq annotations was used to analyze ChIP-seq by Allelome.PRO. Sliding window annotations for the whole genome were created using makewindows from the BEDtools suite (version 2.20.1).

The Single Nucleotide Polymorphism (SNP) annotation file was created from a VCF file containing SNP variant data of 18 mouse strains, downloaded from the Sanger institute (40). SNP information for the strains CAST/EiJ and FVB/NJ were extracted and converted to the required browser extensible data (BED) format using a custom script that is available together with Allelome.PRO (details see manual, Supplementary Material). Note that the name field in this file contains SNP information. Only homozygous high quality SNPs were used and SNPs overlapping annotated pseudogenes were removed.

Reference list of imprinted genes

We defined the list of known imprinted genes by first merging the lists provided by the Harwell and Otago databases (http://www.mousebook.org/imprinting-gene-list, http://igc.otago.ac.nz, (41–43)). Genes that were not annotated by the RefSeq or UCSC database were then removed. We defined multiple imprinted isoforms from the same gene, and groups of closely linked lncRNAs with the same reported imprinted expression pattern to be a single imprinted gene. These cases are indicated in Supplementary Table S2 where we define 123 known imprinted genes. The Allelome.PRO results for RNA-seq and ChIP-seq for these genes are also included in this table, together with information from the literature including where imprinted expression is reported to occur, and if the imprinted status of the gene is disputed.

Saturation curves

Saturation curves were created by Allelome.PRO runs on random sampled subsets of aligned reads. Random sampling was performed using the Picard toolset (version 1.111) for sampling rates of 5, 10, 15, 20, 25, 30, 35, 40, 50, 60, 70, 80 and 90% of total reads. Three technical replicates were produced using different random seeds (1, 2 and 3). Reads were separated according to the transcribed strand after sampling to allow strand-specific analysis by Allelome.PRO. Basic statistical analysis of the resulting data was performed using R (44).

Simulation of sequencing errors

Aligned RNA-seq reads from the region surrounding the *Igf2r* imprinted cluster (GRCm38/mm10 chr17:12350000–

13000000) were extracted from the BAM files of the two forward and two reverse FVB/CAST MEF crosses and converted to FASTQ format. With a custom made Perl script we randomly generated errors for each base in a read at different frequencies (1%, 5%, 10% or 15%), and repeated this three times. Then we re-aligned the FASTQ files and ran Allelome.PRO.

Determining experimental error from *in silico* mixing of CAST/EiJ and FVB/N reads

100bp paired end RNA-seq data from FVB/N adult heart and CAST/EiJ adult heart was aligned to the reference genome using STAR. For CAST 85.3% reads were uniquely aligned, while for FVB 86.7% reads were uniquely aligned. To create *in silico* the two forward and two reverse crosses needed to input into Allelome.PRO, we took aliquots of reads from each strain and then combined them. Based on the alignment rate we calculated the number of input reads needed to uniquely align 3 million reads for each strain, and took sequentially four aliquots of this amount of reads from the FASTQ file (the FASTQ file lists the reads as they come off the Illumina machine, and therefore the order should be random), and then combined FVB and CAST aliquots to create the four technical replicates. We aligned the four technical replicates using STAR, assigned two replicates as forward and two as reverse crosses, and then used Allelome.PRO to calculate allelic expression. By combining equal numbers of CAST and FVB reads we expect most genes to have an allelic ratio around 0.5, but strain bias genes will show unequal ratios. However, we do not expect to find any imprinted genes. Therefore, we defined a false discovery rate (FDR) for imprinted expression as the percentage of informative genes (biallelic, strain bias, imprinted) called imprinted. We determined the FDR with no allelic ratio cutoff (plotted as 0.5) and at allelic ratio cutoffs of 0.6, 0.7, 0.8, 0.9 and 1.0, and at minread settings of 1, 2 and 3 (the minread parameter of Allelome.PRO defines the minimum number of reads over a SNP required before it is included in the analysis).

RESULTS

Allelome.PRO requirements

The Allelome Profiler (Allelome.PRO) pipeline uses custom Perl, shell and R scripts to analyze allelic specific features in massive parallel DNA sequencing data (see manual, Supplementary Material). Allelome.PRO is designed for Linux based operating systems and uses efficient software suites to optimize both the runtime and memory footprint, with SAMtools and BEDtools being the only dependencies (45,46). The Allelome.PRO pipeline depends on data obtained from genetically distinct individuals or pooled samples from two strains and requires three files to be provided by the user in order to start the fully automated analysis (Figure 1A). First, a file defining single nucleotide polymorphisms (SNPs) between the two strains is required in browser extensible data (BED4) format (note the special requirements for the name field detailed in the methods). Second, an annotation file defining the genomic regions to be analyzed must be provided in BED6 format (i.e. a BED file

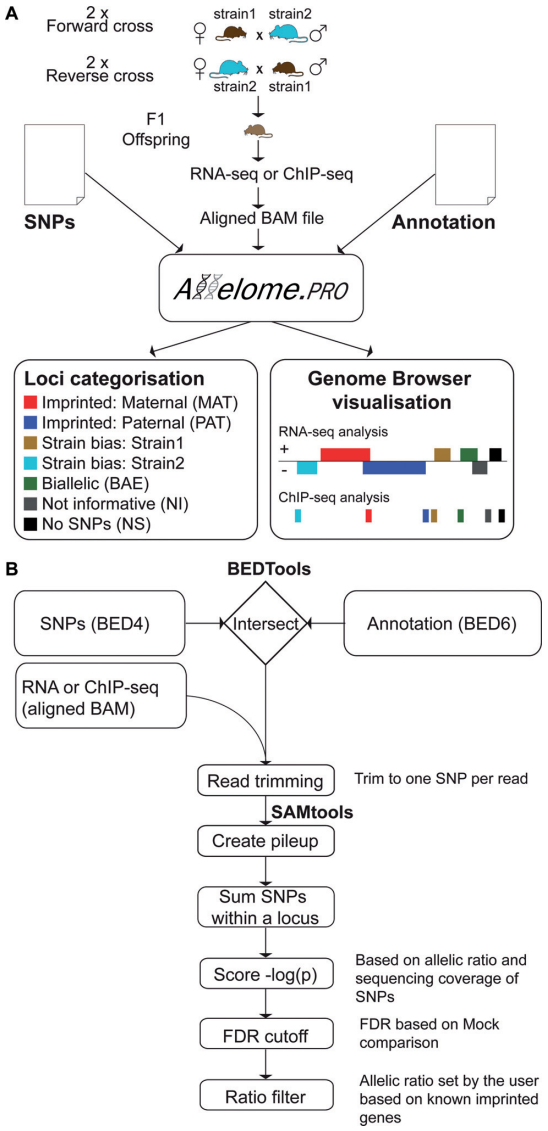


Figure 1. Allelome.PRO workflow to detect allele-specific genome features using RNA-seq and ChIP-seq data. (**A**) Allelome.PRO requires three types of input files: A SNP file (BED6), an annotation file (BED6) and 4 aligned BAM files from F1 reciprocal crosses (2 each of forward and reverse cross). The output categorizes the candidates in the annotation file into the following seven categories: Imprinted: MAT, maternally expressed (red) and PAT, paternally expressed (blue); Strain-biased: Strain1 expressed (brown) and Strain2 expressed (turquoise); BAE, biallelic expression (green); NI, non-informative (e.g. due to low coverage) (gray); NS, no SNP located inside the locus (black). Allelome.PRO further provides a result file (BED6) that can be uploaded to the UCSC genome browser for visual inspection. (**B**) The Allelome.PRO algorithm starts by using BEDtools to intersect the SNP file with the annotation file. The resulting intersection of SNPs located within the annotated candidates is then used to filter out aligned reads that do not overlap any of these SNPs. In the next step, a 1:1 relationship between aligned reads and SNPs is established by trimming the reads so that each read overlaps just one SNP. Subsequently a pileup file of reads at the SNP positions is created using SAMtools. Read counts for the two alleles are summed up over all SNPs within a locus. A binomial distribution is used to assess the significance of the observed allelic biases, and the resulting allelic score is defined as the negative logarithm of the derived *P* value ($-\log_{10}(P)$). An allelic score cutoff based on a user-set false discovery rate (FDR) is then empirically calculated using mock comparisons. The final allelic ratio cutoff filters out remaining candidates with an allelic ratio below a user-set limit (see the manual for details, Supplementary material).

with six fields as defined on http://genome.ucsc.edu, (47)). Overlapping regions with identical names (fourth field in the BED file) are merged into single loci. Finally, aligned sequencing data must be provided as an aligned compressed binary version of the Sequence Alignment Map (BAM) file (45). Allelome.PRO is capable of analysing any DNA sequencing data, however we have tested and optimized it for the massive parallel sequencing of cDNA ends (RNA-Seq) and of chromatin immunoprecipitation (ChIP-Seq). In order to apply statistical testing for allele specific enrichment Allelome.PRO requires four biological samples for RNA-Seq or ChIP-Seq. These are two replicate samples from the F1 offspring of a forward cross between strain1 (mother) and strain2 (father), and two replicates from a reverse cross, where the strains of the mother and father are reversed (Figure 1A). Multiple efficient software solutions are available to map short sequences from massive parallel sequencing to reference genomes, typically called aligners, and all of these report alignments as BAM files. In order to allow maximum flexibility Allelome.PRO is not dependent on a specific aligner, but rather requires one BAM file per biological replicate. The output of Allelome.PRO provides a categorization for each locus in the annotation file (Figure 1A). Furthermore, Allelome.PRO provides a BED file that allows a visual display of the data and can be viewed on any genome browser, such as the UCSC genome browser (http://genome.ucsc.edu, (47)).

Allelome.PRO pipeline

The Allelome.PRO pipeline operates using a number of discrete sequential steps (Figure 1B). First reads from the aligned BAM files that overlap SNPs within the loci provided in the annotation file are extracted using filters that require BEDtools (46). This step limits analysis to reads informative for allelic analysis reducing the number of reads that need to be processed in subsequent steps and thereby improving the efficiency of the pipeline. The extent to which runtime is reduced depends on the number of SNPs, the proportion of the genome covered by the annotation file, and the genomic distribution of the sequencing data. Next, reads overlapping multiple SNPs are trimmed using a custom script, so that each read covers a single SNP and is counted only once, a necessary step for statistical analysis. SAMtools (45) is then used to generate a pileup file of reads at the SNP positions. Pileup files are used to calculate the total number of reads aligning to each allele. These numbers are summed up for all covered SNPs within each annotated locus separately for each of the four biological samples. A binomial test, implemented in R (44), is then used to assess the significance of deviation of the observed allelic biases from the expected 1:1 distribution for biallelic expression for each of the four samples. An allelic score is then calculated for each sample by negative logarithm transformation of the P value ($-\log_{10}(P)$) (29). Two scores are then calculated for each loci by comparing the four samples with each other, a parental bias and strain bias summary score. Loci are then assigned into allelic categories based on whether the allelic score is over the empirically derived false discovery rate (FDR) cutoff and a user defined allelic ratio. Calculation of the summary scores, the FDR and definition of the allelic ratio cutoff are described in detail below. The Allelome.PRO program can be downloaded at the following link: https://sourceforge.net/projects/allelomepro/.

Validation of Allelome.PRO using RNA-seq and ChIP-seq of F1 MEFs

To validate the Allelome.PRO pipeline and define allelic expression and H3K4me3 enrichment in a pure cell type, we performed RNA-seq and ChIP-seq on female F1 MEFs derived from reciprocal crosses between the inbred mouse strains CAST/EiJ (CAST) and FVB/NJ (FVB). We performed two biological replicates from the forward and reverse cross to match the Allelome.PRO requirements. Sequencing reads were aligned to the GRCm38/mm10 genome using the STAR aligner (version 2.3.1z12) (37). For RNA-seq we performed strand-specific ribosomal depleted 100 bp paired-end RNA sequencing (see 'Materials and Methods' section). Ribosomal depletion of total RNA was chosen rather than polyA enrichment to allow analysis of intron located SNPs. On average we obtained 106.6 (±3.3) million total reads per biological replicate, 72% (±4%) of which were uniquely aligned. For ChIP-seq of H3K4me3 we applied 50 bp single-end sequencing and obtained 48.6 (±2.2) million total reads per biological replicate and 93% (±1%) uniquely aligned reads. For the Allelome.PRO run we downloaded SNP variant data from the Sanger institute (40) and extracted 20.4 million high quality SNPs between CAST and FVB (see 'Materials and Methods' section). We then used this data to validate and optimize the Allelome.PRO pipeline as described in the following sections.

Calculation of the allelic score and false discovery rate

Two allelic scores, a parental bias score and a strain bias score, were calculated for each annotated region (RefSeq gene for RNA-seq, RefSeq gene TSS ± 2 kb for H3K4me3 ChIP-seq) in each F1 sample from two forward and two reverse FVB (F) and CAST (C) crosses (CF1, CF2, FC1, FC2). Previous approaches using a similar experimental design and statistical method calculated an allelic score for imprinted expression (imprinted score) from RNA-seq data for the 4 possible reciprocal comparisons (32,33). By calculating scores for the individual samples we were able to include SNPs covered in single samples that would be excluded in the reciprocal comparison approach, thereby increasing the power of the analysis. The parental bias score was calculated using summed maternal and paternal reads over SNPs per loci (MAT >0, PAT <0), and the strain bias score using summed strain 1 (CAST) and strain 2 (FVB) reads over SNPs per loci (CAST >0, FVB <0). To distinguish different categories of allelic enrichment we made two comparisons between the scores of the four samples, a parental bias and strain bias comparison as illustrated in the reciprocal tables (Figure 2A). Two summary scores were then calculated for each locus, a parental or imprinted score (i.score) and a strain bias score (s.score). If a consistent positive or negative bias was seen in all 4 samples, the lowest value was taken as the summary score, otherwise if the direction of the bias was not consistent the summary score

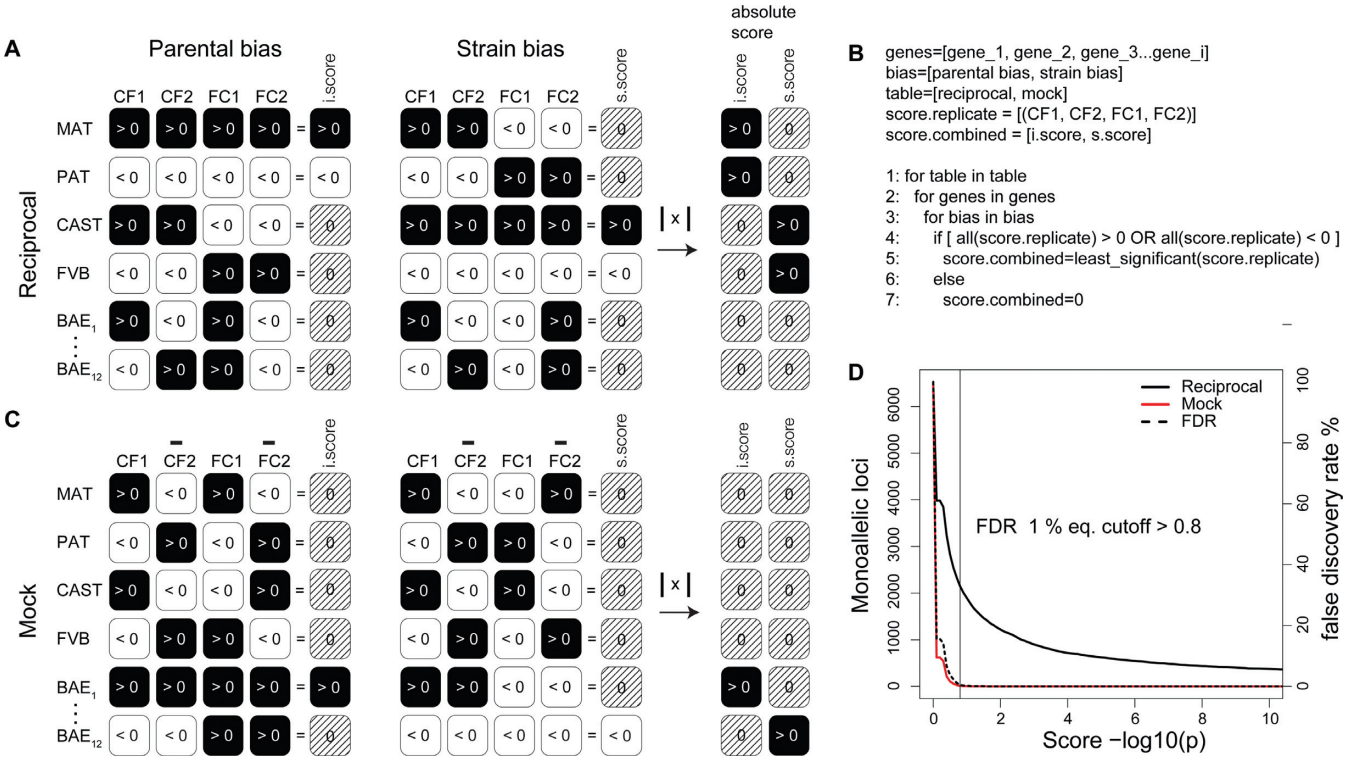


Figure 2. False discovery rate (FDR) allelic score cutoff based on mock analysis. (**A**) Two allelic scores were calculated for each annotated loci for each of the four samples, a parental bias score (MAT >0, black; PAT <0, white) and a strain bias score (CAST >0, black; FVB <0, white). The allelic score is defined as the negative logarithm of the binomial distribution of reads coming from one allele versus both alleles ($-\log_{10}(P)$). Reciprocal analysis was conducted to categorize allelic enrichment for each loci by comparing the parental bias scores (left) and strain bias scores (right) between the four samples. The allelic score patterns in the four samples for each allelic enrichment category are displayed: parental biased (MAT, PAT), strain-biased (CAST, FVB) and biallelic genes (BAE, only 2 of 12 possible biallelic combinations are displayed). A summary imprinted score (i.score) and strain-biased score (s.score) is calculated by comparing the four samples. If the bias is in the same direction for all four samples then the minimum score is taken, while if direction of bias is inconsistent for any of the four samples then the score is set to 0 (striped pattern). Each loci can have either an i.score value (imprinted) or an s.score value (strain biased), while the other score equals zero, or both the i.score and s.score equal 0 (biallelic). The absolute value of the i.score and s.score are calculated and then used for calculating the false discovery rate (FDR) in (D). (**B**) Pseudocode illustrating how the final allelic score (i.score or s.score) is derived from the allelic scores of the four biological replicates. (**C**) Mock analysis of parental bias and strain bias allelic scores to calculate i.scores and s.scores are conducted as for the reciprocal analysis in (A), except the scores of one sample from each cross are inverted. This results in the removal of parental bias and strain bias genes, which no longer have a consistent direction of bias and therefore have a score of 0. In contrast, 4 from 12 possible biallelic score combinations now have parental scores or strain bias scores in the same direction, resulting in a summary i.score or s.score value different from 0. These score values should be low compared to true allelic biases as they showed random deviations from a 0.5 ratio representing the technical and biological variation in the data. The absolute values of these mock scores are then compared to the values calculated in the reciprocal analysis to calculate the FDR in (D). (**D**) The false discovery rate (FDR) was estimated as the number of detected candidates with allelic biases (parental and strain bias) in the mock analysis, divided by the number of detected candidates with allelic biases in the reciprocal analysis. In this example RefSeq genes on the forward strand were analyzed in E12.5 mouse embryonic fibroblasts (MEFs) RNA-seq data using an FDR of 1%.

was set to 0. Using this approach, each locus had a value for only one score (either the i.score or the s.score), while the other score equalled 0, or both scores were 0. That is, loci showed parental-specific enrichment (i.score >0 or <0, s.score = 0), strain-specific enrichment (i.score = 0, s.score >0 or <0) or non-enrichment (biallelic or non-informative, i.score = 0, s.score = 0) (Figure 2B the logic for the allelic score calculation). Finally, the absolute value (>0) of the i.score and s.score was calculated, a step necessary for the calculation of the false discovery rate (FDR) as described in the following section (Figure 2A and C).

There are 16 possible combinations of positive (>0) and negative (<0) allelic scores for the four samples, two of which show allelic biases in the same direction for parental bias, and two for strain bias (Figure 2A). It is expected by chance that 4 in 16 biallelic loci will also show an allelic bias in the same direction for all four samples, for either the parental or strain bias comparisons, leading to an i.score or s.score >0 or <0, although the allelic score should be lower than for true imprinted or strain bias loci as the allelic ratio should be close to 0.5. Therefore, we sought to reduce the number of false positives by setting a FDR based on the level of random allelic enrichment in our data empirically determined by mock analysis (Figure 2C). This approach was based on mock analysis of the four samples as reported earlier (32), but with several modifications. Previously, mock comparisons between samples of the same genotype were used to determine the FDR, as no difference in allelic expression is expected (32,33). Thus two mock comparisons are possible with four reciprocal comparisons. In contrast, we calculate a score for each sample rather than for the comparisons, enabling four scores to be used for the mock analysis and allowing the reciprocal and mock analysis to be performed in the same way. To perform mock anal-

ysis we negated the allelic scores in one biological replicate of each cross (CF2, FC2), and then performed the analysis in an identical manner as for the reciprocal analysis generating an i.score and s.score for each locus in the annotation (Figure 2C). Using this approach loci previously showing parental or strain bias (Figure 2A) now had an i.score and s.score of 0, while some biallelic loci (expected 4 in 12) now had an i.score or s.score >0 or <0 (Figure 2C). This mock analysis gave an estimate of the technical and biological variation in our data and was used to calculate the FDR. To calculate a single FDR for monoallelic enrichment we first pooled the absolute value of the i.score and s.score for all loci for both the reciprocal and mock analysis (Figure 2A and C). This differed from previous approaches that compared only parental allele bias to calculate the FDR (32,33), and increased the robustness of the FDR due to the larger number of strain bias loci compared to parental bias loci (~20-fold higher in this study). The FDR (%) was calculated as the number of loci exceeding the score cutoff in the mock analysis, divided by the number of regions exceeding the score cutoff in the reciprocal analysis, multiplied by 100. For each run Allelome.PRO provides a plot showing the number of monoallelic loci at different score cutoffs in the reciprocal and the mock analysis as well as the corresponding FDR. A vertical line indicates the score cutoff at the user defined FDR, which in this study was 1% (Figure 2D).

Empirical determination of an allelic enrichment cutoff

Defining allelic enrichment by an allelic score FDR cutoff alone can lead to artefacts, as biallelic genes can overcome the FDR cutoff if by chance all four samples share the same direction of bias and SNP coverage is high enough. Following this, we observed that if we analyzed our data with the FDR cutoff as the only filtering criteria, some highly expressed genes with small deviations from a biallelic ratio could produce scores over the FDR cutoff. Loci showing a minor allelic ratio bias are often not validated by independent methods (32), and even if validated there is no evidence that such minor biases are biologically meaningful. Therefore, we empirically determined an allelic ratio cutoff from our data based on known imprinted and strain bias genes, and used this to filter loci over the FDR cutoff to further reduce false positives.

In order to determine a biologically relevant allelic ratio cutoff we first plotted the distribution of allelic ratios for all 65 genes classified as imprinted by the FDR cutoff in analysis of our RNA-seq data (Figure 3A). Of these 43 have been reported to show imprinted expression previously (29,42,43). Notably, a biphasic distribution was obtained with most known imprinted genes showing allelic ratios >0.85, and most of the novel imprinted genes identified by our RNA-seq showing much lower allelic ratios. Six known imprinted genes (*H13, Gnas, Inpp5f, Phactr2, Cobl, Trappc9*) were also clustered in this low ratio group. However, these are all genes with reported tissue-specific imprinted expression in a tissue other than MEFs (27,48–52). No novel imprinted gene was identified in the RefSeq annotation with this allelic ratio cut-off.

To determine an appropriate FDR cutoff for genes with a strain biased expression pattern, we made use of

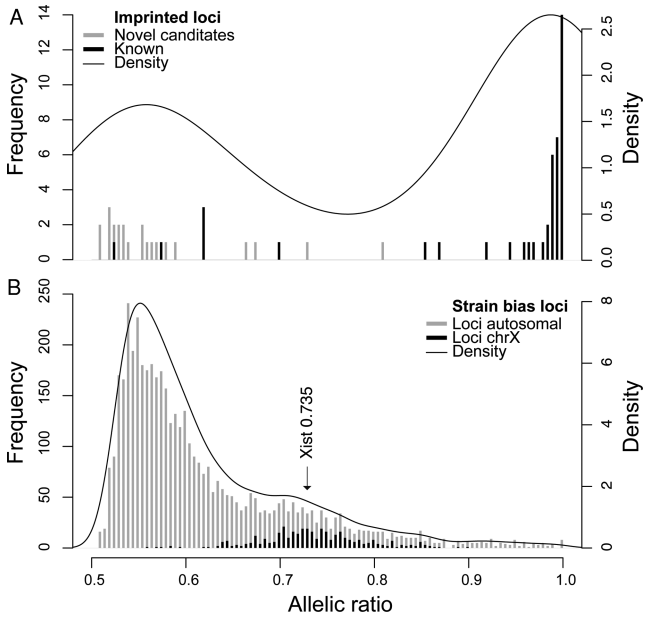


Figure 3. Setting the allelic ratio. (**A**) The allelic ratio distribution for the 65 parental bias genes with an i.score higher than the FDR cutoff in RNA-seq data from MEFs. Plotted are both a histogram in grey and a density curve for this distribution. The black bars overlapping the histogram indicate known imprinted genes. (**B**) The allelic ratio distribution for all strain-biased loci with an allelic score higher than the FDR cutoff in RNA-seq data from MEFs. Plotted are both a histogram in gray and a density curve for this distribution. The black bars overlapping the histogram indicate strain-biased loci on the X chromosome. Densities were estimated using a Gaussian kernel function calculated in R.

known skewed X-inactivation in our MEFs. This is a well-documented effect in female cells from crosses between *M. musculus domesticus* (FVB) and *M. musculus castaneus* (CAST) that results in the predominant inactivation of the FVB derived X-chromosome (13,14). The allelic ratio distribution of genes on autosomes showed a prominent peak between 0.5 and 0.6, close to a biallelic expression ratio (Figure 3B). However, we also noted a shoulder peak around 0.7, which followed the distribution of the X-linked genes (Figure 3B, black bars). Following this, genes on the X-chromosome showed a mean allelic ratio of 0.735 in our analysis. The majority of X-linked genes showing a significant strain bias over a ratio of 0.7 (>85%). Therefore, in order to have a single allelic cutoff for strain biased and parental biased genes and to distinguish allelic bias from biallelic expression, we used an allelic ratio >0.7 cutoff together with a 1% FDR cutoff for further analyses. At an allelic ratio cutoff of 0.7, two novel candidate imprinted genes were detected in addition to the 37 known imprinted genes detected with an allelic ratio cutoff of 0.85. These genes were not detected in a previous study of MEFs (Table 1) (53), and were not validated by differential H3K4me3 analysis (Table 2), indicating that they were false positives. This indicates that a lower parental bias allelic ratio cutoff is acceptable when novel candidates are subject to independent validation.

Table 1. RefSeq genes showing imprinted expression in MEFs detected by Allelome.PRO

Chr.	Start	End	Str.	Candidate	Cov. SNPs	Allelic ratio	RPKM	Cat.	Tran et al.
chr1	63200357	63314575	+	Zdbf2	49	0.03	2.39	PAT	confirmed
chr2	174281236	174295436	–	Nespas	19	0.08	3.61	PAT	confirmed
chr6	4674349	4747204	–	Sgce	64	0.01	0.61	PAT	confirmed
chr6	4747305	4760516	+	Peg10	24	0	59.85	PAT	confirmed
chr6	5383385	5433021	+	Asb4	43	0.86	6.39	MAT	confirmed
chr6	30733505	30748466	+	Mest	34	0	1.04	PAT	confirmed
chr6	58905232	58907126	–	Nap1l5	1	0	6.71	PAT	not inform.
chr7	6675442	6696432	–	Zim1	64	0.99	10.45	MAT	not inform.
chr7	6705959	6730419	–	Peg3	108	0	218.19	PAT	confirmed
chr7	6706759	6707624	+	Peg3os	3	0	8.43	PAT°	not inform.
chr7	6730582	6967219	+	Usp29	73	0.01	1.11	PAT	not inform.
chr7	59619157	59678878	–	Ipw	12	0	0.34	PAT	not inform.
chr7	59969576	59974431	–	D7Ertd715e	9	0	8.17	PAT	confirmed
chr7	59982500	60005156	–	Snurf	27	0	1.71	PAT	confirmed
chr7	59982501	60140219	–	Snrpn	32	0.01	0.08	PAT	confirmed
chr7	61705849	61927574	–	A230057D06Rik	75	0.01	0.27	PAT	not inform.
chr7	61943900	61982303	–	A330076H08Rik	26	0	0.56	PAT	not inform.
chr7	62348276	62349927	+	Ndn	1	0.01	156.97	PAT	confirmed
chr7	62461870	62464510	–	Peg12	6	0.01	27.94	PAT	confirmed
chr7	142575531	142578146	–	H19	8	1	5666.39	MAT	confirmed
chr7	142650767	142661035	–	Igf2	15	0	207.89	PAT	confirmed
chr7	142659692	142670356	+	Igf2os	14	0.01	8.13	PAT	confirmed
chr7	143107253	143427042	+	Kcnq1	27	0.06	0.03	PAT°	not inform.
chr7	143213110	143296547	–	Kcnq1ot1	358	0.01	2.45	PAT	confirmed
chr7	143458338	143461050	–	Cdkn1c	2	0.99	93.04	MAT	confirmed
chr10	13090787	13131695	+	Plagl1	183	0.01	18.43	PAT	confirmed
chr11	11930498	12037420	–	Grb10	437	0.99	328.94	MAT	confirmed
chr11	22972004	22976496	+	Zrsr1	7	0	7.25	PAT	not inform.
chr12	85686719	85709087	+	Batf	5	0.19	1.17	PAT	not inform.
chr12	109452822	109463336	+	Dlk1	18	0.13	0.47	PAT	confirmed
chr12	109540995	109571729	+	Meg3	89	1	4.05	MAT	confirmed
chr12	109603944	109661711	+	Rian	278	1	236.84	MAT	confirmed
chr12	109734980	109749457	+	Mirg	91	1	11.8	MAT	confirmed
chr14	31260374	31323896	–	Dnah1	11	0.73	0.17	MAT	not inform.
chr15	72589619	73061204	–	Trappc9	313	0.3	0.44	PAT°	not inform.
chr15	72805599	72810324	–	Peg13	27	0.01	9.54	PAT	confirmed
chr15	96994822	97055956	–	Slc38a4	179	0.04	178.84	PAT	confirmed
chr17	12682405	12769706	–	Igf2r	257	0.98	336.9	MAT	confirmed
chr17	12741310	12859884	+	Airn	251	0.04	1.11	PAT	confirmed
chr18	12972251	12992948	+	Impact	118	0.02	224.23	PAT	confirmed

Columns show the GRCm38/mm10 location (Chromosome, Start, End, Strand) of each candidate, as well as the allelic ratio (maternal reads over total reads) and the Allelome.PRO classification indicating whether the gene shows paternal or maternal expression. The last column shows the results published in a previous MEF study, which used JF1/M x TgOG2 reciprocal crosses compared to FVB/N × CAST/EiJ crosses used in this study (53). The false positive call for *Kcnq1* (°) and *Peg3os* (°) was the result of bleed-through in strand specific sequencing from the antisense overlapping *Kcnq1ot1* and *Peg3* respectively. *Trappc9* (°) was called because of a sense overlap with *Peg13*.

Table 2. H3K4me3 ChIPseq data results confirm the RNA-seq allelome

Chr.	Start	End	Candidate	ChIP-seq		RNA-seq	
				Allelic Ratio	Result	Allelic Ratio	Result
chr1	63198357	63275268	Zdbf2*	0.06	PAT	0.03	PAT
chr2	152685147	152689147	Mcts2	0.03	PAT	-	NS
chr2	174293436	174297436	Nespas	0.00	PAT	0.08	PAT
chr6	4745204	4749204	Sgce	0.01	PAT	0.01	PAT
chr6	4745305	4749305	Peg10	0.00	PAT	0.00	PAT
chr6	5381385	5385385	Asb4	0.51	CAST	0.86	MAT
chr6	30731505	30740052	Mest*	0.01	PAT	0.00	PAT
chr6	58905126	58909126	Nap1l5	0.02	PAT	0.00	PAT
chr7	6694432	6698432	Zim1	0.94	MAT	0.99	MAT
chr7	6728419	6732419	Peg3	0.00	PAT	0.00	PAT
chr7	6704759	6708759	Peg3os	-	NI	0	PAT°
chr7	6728582	6732582	Usp29	0.00	PAT	0.01	PAT
chr7	59676878	59680878	Ipw	-	NS	0	PAT
chr7	59972431	59976431	D7Ertd715e	-	NI	0	PAT
chr7	60003156	60007156	Snurf	0.00	PAT	0.00	PAT
chr7	60003156	60142219	Snrpn*	0.02	PAT	0.01	PAT
chr7	61925574	61929574	A230057D06Rik	-	NI	0.01	PAT
chr7	61980303	61984303	A330076H08Rik	0.02	PAT	0.00	PAT
chr7	62346276	62350276	Ndn	0.04	PAT	0.01	PAT
chr7	62374978	62378978	Magel2	0.02	PAT	-	NI
chr7	62418139	62422139	Mkrn3	0.02	PAT	-	NI
chr7	62462510	62466510	Peg12	0.02	PAT	0.01	PAT
chr7	142576146	142580146	H19	0.99	MAT	1.00	MAT
chr7	142655481	142663035	Igf2*	0.01	PAT	0.00	PAT
chr7	142657692	142661692	Igf2os	0.02	PAT	0.01	PAT
chr7	143105253	143109253	Kcnq1	-	NI	0.06	PAT°
chr7	143294547	143298547	Kcnq1ot1	0.01	PAT	0.01	PAT
chr7	143459050	143463050	Cdkn1c	-	NI	0.99	MAT
chr10	13088787	13092787	Plagl1	0.00	PAT	0.01	PAT
chr11	12025971	12039420	Grb10*	0.98	MAT	0.99	MAT
chr11	22970004	22974004	Zrsr1	0.01	PAT	0.00	PAT
chr12	85684719	85688719	Batf	0.46	NI	0.19	PAT
chr12	109450822	109455454	Dlk1*	0.11	PAT	0.13	PAT
chr12	109538995	109547397	Meg3*	1.00	MAT	1.00	MAT
chr12	109601944	109605944	Rian	-	NI	1	MAT
chr12	109732980	109736980	Mirg	-	NI	1	MAT
chr14	31321896	31325896	Dnah1	0.46	BAE	0.73	MAT
chr15	73053812	73063204	Trappc9*	0.51	BAE	0.3	PAT°
chr15	72808324	72812324	Peg13	0.00	PAT	0.01	PAT
chr15	97053956	97057956	Slc38a4	0.01	PAT	0.04	PAT
chr17	12767706	12771706	Igf2r	1.00	MAT	0.98	MAT
chr17	12739310	12743310	Airn	0.01	PAT	0.04	PAT
chr18	12970251	12974251	Impact	0.01	PAT	0.02	PAT

Column details and symbol (°) are as in Table 1. The chromosome start/end indicates the region containing all transcription start sites of the target gene that were evaluated. Asterisks (*) indicate genes for which multiple windows were evaluated. Gray font indicates discordance between the allelic ratio for H3K4me3 and RNA-seq (see text for details).

Genome-wide allele-specific expression in MEFs

Previous studies analyzed RNA-seq, ChIP-seq or DNA methylation-seq to detect either parental or strain specific allelic enrichment, but none defined the allelic enrichment status, or allelome, of all annotated loci in a given cell type or tissue. In order to do this, we defined biallelic loci as those loci not identified as showing a parental or strain bias, but with enough SNP coverage to theoretically overcome the allelic score FDR at an allelic ratio cutoff of 0.7. Those loci with a lower SNP coverage (non-expressed or very lowly expressed genes) were defined as non-informative, while the final category included loci with no SNPs. In summary, by introducing an allelic ratio cutoff in combination with an allelic score cutoff, Allelome.PRO was able to categorize all loci in an annotation into 7 categories: maternal bias (MAT), paternal bias (PAT), strain 1 bias (CAST), strain 2 bias (FVB), biallelic (BAE), non-informative (NI) and no SNP (NS).

Next, we investigated how many RNA-seq sequencing reads were necessary to categorize allelic data, and which alignment program produced the best results with RNA-seq data. The saturation curves for the STAR aligner showed saturation for the number of imprinted genes (red and blue) already at low sampling rates (10–20%, or 10–20 million reads per replicate) (Figure 4A, left) (37). In contrast, the numbers of strain-biased (brown and turquoise) and biallelic genes (green) continued to increase with increasing number of reads, although the slope decreased with higher sample rates indicating the data was near saturation. With increased sequencing depth the number of biallelically expressed genes increased in parallel with a decrease in the number of non-informative genes. Saturation curves were also produced from the data aligned with GSNAP and TopHat (Figure 4A, middle and right, respectively) (38,39). The saturation curves were broadly similar for all three aligners, with the exception of strain-biased genes. Both STAR and GSNAP detected more CAST than FVB strain bias genes (494 CAST versus 391 FVB (STAR) and 583 CAST versus 240 FVB (GSNAP)), but in contrast TopHat detected more FVB than CAST strain bias genes (338 CAST versus 1009 FVB). This is likely due to an alignment bias, as the FVB strain is more closely related to the C57BL/6J reference strain than CAST, and therefore TopHat may have difficultly aligning some CAST reads leading to false positive FVB strain bias genes and the failure to detect some CAST bias genes. We observed that STAR aligned more reads over SNP positions than GSNAP or TopHat, which, combined with the much shorter runtime, convinced us to use STAR for further analysis. Still, one of our main concerns in choosing STAR over GSNAP was that STAR does not offer an option for SNP-sensitive analysis like GSNAP. However, when we correlated allelic ratios determined by GSNAP and by STAR they showed very high correlation ($R^2 = 0.99$), indicating that alignment biases did not affect the overall results.

The results for the Allelome.PRO run with STAR aligned data showed 40 imprinted genes, 10 of which were maternally expressed while 30 showed paternal expression. Furthermore, we detected 494 CAST bias (299 on chromosome X) and 391 FVB bias genes (three on chromosome X), confirming the X-inactivation between these strains. Of the remaining genes, 12140 were classified as showing biallelic expression, 8930 as non-informative and 1208 could not be assessed as they contained no SNP (Figure 4B). The allelic ratios of the detected imprinted genes are displayed in Figure 4C with detailed information including genomic location, number of covered SNPs, and allelic ratio given in Table 1. Additionally, details for all informative SNPs over detected imprinted genes is given in Supplementary Table S1, and the full table including all informative genes is available from the Gene Expression Omnibus (GEO, accession number GSE69168). Only 31 of the 123 known imprinted genes were detected as imprinted (Supplementary Table S2). In most cases, this was likely due to tissue-specific imprinted expression where the genes were called non-informative (not expressed) or biallelic, although six genes could not be assessed for imprinted expression due to strain bias, and seven genes were not assessed because they were not included in the RefSeq annotation that we used. Our results showed a high level of agreement with a previous RNA-seq study conducted in MEFs from a JF1/M $\times$ TgOG2 reciprocal crosses, with 27 of 32 reported imprinted genes detected (Table 1) (53). Of the five genes that we did not detect, one had no SNP in our cross (*Nnat*), two were not part of the RefSeq annotation that we used (*AK050713* and *Rtl1as*), and we excluded one from our annotation due to its small size (*AF357425*). However, using Allelome.PRO with sliding window annotations (2, 4, 6 and 8 kb) we could confirm imprinted expression of *AK050713*, *Rtl1as* and *AF357425* (data not shown). The fifth candidate, *Blcap*, was categorized as biallelically expressed in our data, which was in agreement with reports that this gene only exhibits imprinted expression in brain (54). We detected 13 imprinted gene candidates by RNA-seq in our study that were not detected in a previous study of MEFs (53). Two were probable false positives due to overlap with other imprinted genes: *Kcnq1* due to anti-sense overlap with *Kcnq1ot1* and the incomplete strand-specificity of our sequencing technique, and *Trappc9* due to sense overlap with *Peg13*. *Kcnq1* and *Trappc9* were lowly expressed compared to their overlapping genes, and visual inspection of the genome browser revealed that these long genes showed an increased signal in the overlap region with the shorter *Kcnq1ot1* and *Peg13*, further indicating that this signal from this overlapping region was responsible for the probable false positive call. We called *Dlk1* as paternally imprinted, while the previous study classified it as paternally biased and did not include it in their final list (53). The remaining 10 imprinted gene candidates that we detected were characterized in the previous study as either lacking SNPs, being non-expressed or low expressed, or no data was presented (53). This indicates the increased sensitivity of our method due to the large number of SNPs used, our ability to detect SNPs in introns due to the use of total RNA-seq, and the Allelome.PRO approach of summing all covered SNPs within a gene, all of which together enabled us to detect imprinted expression of lowly expressed genes. Similarly, using all SNPs covered in at least one replicate, and summing up all SNPs within a gene, enabled us to also call lowly expressed long non-coding (lnc) RNAs with high confidence. This was illustrated by the example of *Igf2r* and *Airn* (Figure 4D). Coverage across the

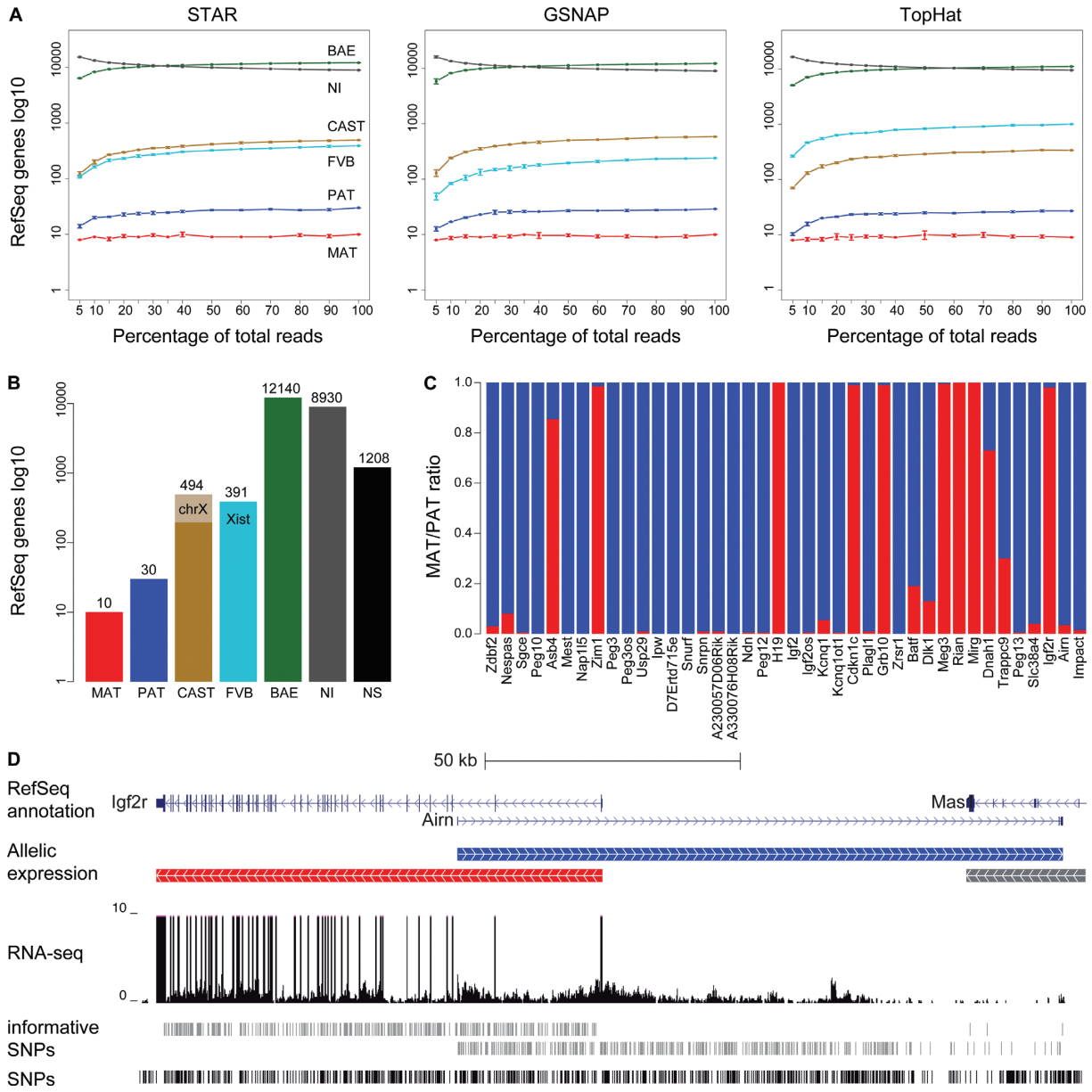


Figure 4. Allelome defined in MEFs using RNA-seq. (**A**) Saturation curves showing the Allelome.PRO results for different samplings of the total RNA-seq reads from MEFs for three different aligners, STAR, GSNAP and TopHat (from left to right). Reads were sampled from total uniquely aligned reads in three technical replicates (STAR: 77.47 ± 1.89, GSAP: 76.89 ± 1.94, TopHat: 94.47 ± 2.02 millions of reads per replicate). The six curves in each plot represent the categories listed in Figure 1 except the 'No SNP' category, which is omitted. The curves for imprinted genes (blue, red) show saturation at low sample rates and little differences between the three aligners. The curves for strain-biased genes (brown, turquoise) show an increase of strain-biased genes with increasing read number, although the slope decreases with higher sample rates it does not plateau. The three aligners detect different numbers of strain bias genes, with STAR and GSNAP detecting more CAST than FVB biased genes, while TopHat detects more FVB than CAST strain bias genes. All aligners show an increase in the number of biallelic genes detected, and a decrease in the number of non-informative genes with increasing number of sequencing reads. (**B**) Categorization of RefSeq genes as produced by Allelome.PRO for strand-specific RNA-seq data of MEFs. Genes were categorized into seven categories, as listed in Figure 1 with numbers given above. The pale brown bar shows the amount of CAST strain-biased genes on chromosome X. *Xist* shows a strain bias towards the FVB allele as indicated on the turquoise bar. (**C**) The imprinted genes from (B) in more detail. The ratio between maternal and paternal allele is illustrated as red and blue bars. Genes were sorted by chromosome number and genomic location and gene names are given on the x-axis. This Figure is also part of the Allelome.PRO output. (**D**) Ribosomal depletion followed by 100bp paired end deep sequencing allows the detection of SNPs within the introns of protein coding genes and lowly expressed long non-coding (lnc) RNAs. UCSC genome browser screenshot showing data on the protein-coding gene *Igf2r* and the long non-coding (lnc) RNA *Airn* in MEFs. The tracks depict (from top to bottom): RefSeq genes, Allelome.PRO allelic expression categorization, RNA-seq aligned reads, informative SNPs on the forward and reverse strand in grey, and total SNPs in black.

Airn gene body was much lower than coverage of *Igf2r* exons, but due to the large number of available SNPs (grey bottom track, informative SNPs) typical for a non-coding gene, the *Airn* lncRNA was still confidently called as showing paternal expression.

In summary, using Allelome.PRO to analyze RNA-seq data we confirmed previously reported imprinted genes in MEFs and detected additional genes, most of which were previously reported to show imprinted expression in another tissue. We also detected strain biased genes, including X-linked genes confirming a known X-chromosome inactivation bias, as well as classifying biallelic expressed and non-informative genes, thus defining for the first time the entire allelome of a tissue.

Validation of allele specific expression by H3K4me3 ChIP-seq

Previously most RNA-seq studies investigating imprinted expression validated their results using methods that assay allelic expression using a single SNP in cDNA, for example, by pyrosequencing (32), Sanger sequencing (29), or Sequenom assays (31,53). Here, we validated our RNA-seq results using differential enrichment of the active H3K4me3 mark over promoters detected by Allelome.PRO using multiple SNPs from ChIP-seq data. Differential enrichment or H3K4me3, H3K27ac or H3K36me3 was used before as a proxy for imprinted expression of some imprinted genes, but not as a general validation of imprinted or allelic expression (30,55). Here we used 4kb windows surrounding the TSS of RefSeq genes as an annotation file for Allelome.PRO to analyze ChIP-seq data for H3K4me3 marks in MEFs. If a gene had multiple isoforms with different start sites, SNPs from all sites were combined, as each gene was treated as a single locus in this analysis. Using this approach our results were broadly similar to the RNA-seq results. We found 31 parental specific promoter marks, 5 maternal and 26 paternal (Figure 5A and B). The details of the informative SNPs for these genes are shown in Supplementary Table S1. 382 CAST specific promoter marks (272 on chromosome X) and 183 FVB specific promoter marks (1 on chromosome X) were found, confirming the X-inactivation bias seen by RNA-seq (Figure 5A). 13061 promoter regions were classified as showing biallelic marks and 8654 regions were non-informative, while 892 regions were not assessed because they contained no SNP (Figure 5A). A table including SNP details for all informative promoters is available from GEO (accession number GSE69168).

A high level of agreement was seen between imprinted genes detected in MEFs by RNA-seq and H3K4me3 ChIP-seq using Allelome.PRO. In total 43 genes were detected as showing parental specific expression and/or parental specific histone marks (Table 2). A comparison between the RNA-seq and ChIP-seq results showed that 28 out of 40 genes called as showing imprinted expression by RNA-seq also had differential enrichment of H3K4me3 over their promoter. Five of 12 genes not confirmed by ChIP-seq were found to show imprinted expression in MEFs by RNA-seq in an independent study, indicating that they do show imprinted expression in this tissue (Tables 1 and 2) (53). One showed a CAST bias in H3K4me3 enrichment (*Asb4*), while 4 others had non-informative H3K4me3 data (*D7Ertd715e*, *Cdkn1c*, *Rian* and *Mirg*). Seven of 12 genes not confirmed by ChIP-seq were also not found in a previous study of MEFs (Tables 1 and 2) (53). Two of these genes were probable artefacts caused by sense and anti-sense transcriptional overlap by other imprinted genes as mentioned previously (*Kcnq1* and *Trappc9*), demonstrating the value of differential H3K4me3 enrichment assays to resolve such issues. *Peg3os* also shows an antisense overlap with the highly expressed paternal imprinted gene Peg3, and was not confirmed by ChIP-seq, so it cannot be excluded that it is also not a false positive. Two other genes were the only novel candidate imprinted genes that we detected by the RNA-seq analysis (*Batf* and *Dnah1*). As mentioned previously, these genes were the only 2 of 40 imprinted expression candidates detected by RNA-seq that were not detected with the higher allelic ratio cutoff of 0.85, and this together with the lack of validation by H3K4me3 differential enrichment indicates that they are false positives. The remaining two genes not detected in the previous study of MEFs had high allelic ratios in our RNA-seq, but had non-informative ChIP-seq results (*A230057D06Rik*) or no SNP in the assayed region (*Ipw*). The three genes detected by ChIP-seq, but not by RNA-seq in our study were all known imprinted genes with high allelic ratios, and either had no SNP in the gene body (*Mcts2*), or had a non-informative RNA-seq result (*Magel2* and *Mkrn3*). In the previous study of MEFs, *Mcts2*, a 795bp single exon gene, was also reported to have no SNP, while *Magel2* and *Mkrn3* were described as not being expressed (53). We found *Magel2* to be lowly expressed, resulting in a non-informative result by RNA-seq. However, in our data *Mkrn3* was highly expressed, but three of four SNPs were excluded due to overlap with pseudogenes leading to the non-informative RNA-seq call. In summary, 33 out of 43 RefSeq genes that we detected as showing imprinted expression were found either by RNA-seq and ChIP-seq, or by RNA-seq and a previous study of MEFs, or in all three datasets, making these high confidence imprinted genes in MEFs. Additionally, six genes were found to be imprinted in either RNA-seq or ChIP-seq data, but not in the complementary dataset due to lack of SNPs (two genes) or a non-informative result (3 genes). The remaining five genes were excluded as probable false positives. Thus, 38 RefSeq genes were identified as showing imprinted expression in MEFs.

To examine the results generated by Allelome.PRO in detail, we used the well-characterized *Igf2r* imprinted gene cluster (Figure 5C). The RNA-seq and ChIP-seq results confirmed that *Igf2r* was only expressed from the maternal allele, whereas the macro lncRNA *Airn* was only expressed from the paternal allele. The extra-embryonic-lineage specific imprinted genes *Slc22a2* and *Slc22a3* are not expressed in MEFs and were therefore classified as non-informative in both analyses, as were *Mas1*, *Mrgprh*, and *Pnldc1*. *Tcp1* and *Mrpl18* showed biallelic expression and biallelic H3K4me3 marks. Overall, Allelome.PRO analysis for RNA-seq data showed the expected pattern for the imprinted *Igf2r* cluster that was confirmed by the Allelome.PRO H3K4me3 ChIP-seq analysis. As mentioned above, one example where the ChIP-seq and RNA-seq analyses disagreed was the maternally expressed gene *Asb4* (Figure 5D). The RefSeq annotation only contained the long isoform of *Asb4*, which

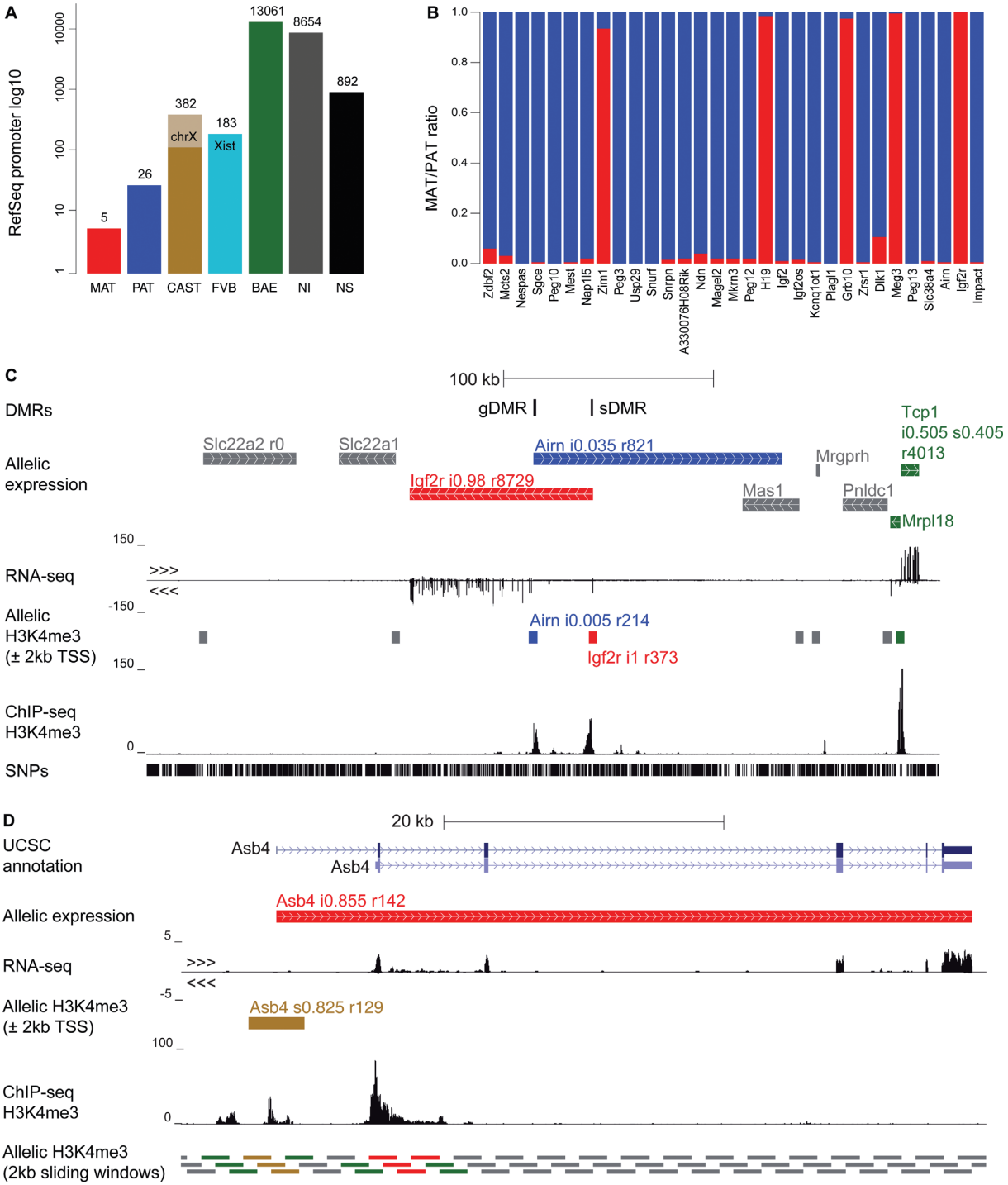


Figure 5. Allelome validated using ChIP-seq H3K4me3. (**A**) Categorization of RefSeq promoter regions (±2 kb windows over RefSeq gene transcription start sites) as produced by the Allelome.PRO for H3K4me3 ChIP sequencing data of MEFs. The CAST strain bias on chromosome X is seen as well as the categorization of *Xist* as showing a FVB strain bias. (**B**) The maternal/paternal ratio is shown as red/blue bars for the imprinted genes from (A). Promoter windows are named after their respective genes and sorted as in Figure 4C. (**C**) Allelic expression in the *Igf2r* cluster is validated by differential H3K4me3 enrichment in MEFs. UCSC genome browser screenshot showing the *Igf2r* imprinted gene cluster and adjacent genes together with the Allelome.PRO output for RNA-seq and ChIP-seq. The tracks depict (from top to bottom): gametic and somatic differentially DNA methylated regions (gDMRs and sDMRs), Allelome.PRO allelic expression categorization, strand-specific RNA-seq, Allelome.PRO H3K4me3 allelic enrichment categorization, H3K4me3 ChIP-seq, and total SNPs in black. (**D**) Sliding windows detect differential H3K4me3 peaks outside annotated transcription start sites. Allelome.PRO detected maternal expression of *Asb4* from RNA-seq data using the RefSeq annotation, but a CAST strain bias was detected by Allelome.PRO analysis of H3K4me3 ChIP-seq data using the RefSeq promoter annotation. However, analysing the ChIP-seq data instead with a 2 kb sliding window annotation revealed maternal H3K4me3 enrichment over the promoter of a shorter isoform of Asb4 that was annotated by UCSC and appeared from RNA-seq data to be the predominant isoform in MEFs. UCSC genome browser screenshot showing (from top to bottom) the UCSC gene annotation, Allelome.PRO allelic expression categorization from RefSeq, strand specific RNA-seq, Allelome.PRO H3K4me3 allelic enrichment categorization using RefSeq annotation, H3K4me3 ChIP-seq, and Allelome,PRO H3K4me3 allelic enrichment categorization using a 2 kb sliding window annotation.

showed a strain specific H3K4me3 peak at its transcription start site. However, Allelome.PRO run using a sliding window annotation for the same ChIP-seq data revealed the presence of a maternal peak at the start site of an UCSC-annotated shorter isoform of *Asb4*, confirming maternal expression in the RNA-seq results. Overall, ChIP-seq results showed agreement with RNA-seq results in 18180 (84%) of 21649 cases where at least one SNP was present in both analyses (Figure 6). Strain-biased expression showed the lowest validation rate with only 257 (29%) of 879 cases validated by ChIP-seq. Biallelic expression was confirmed for 10 922 (90%) of 12 058 candidates and 6973 genes were non-informative in both analyses.

DISCUSSION

Since the development of high-throughput sequencing technologies a number of studies have sought to detect imprinted expression from RNA-seq data from various tissues using a variety of experimental and analysis approaches (26,27,29,31,32,56,57). Additionally, detection of differential allelic expression has potential for the mapping of *cis*-regulatory elements, so-called *cis* expression quantitative loci or *cis*-eQTLs (40,58–61). Furthermore, differential allelic expression analysis has been employed as a tool for studying alternative mRNA processing (12). We developed Allelome.PRO as a user-friendly and efficient tool to capture the genome wide state of differential allelic features, thus providing a single tool to aid in these different applications.

Allelome.PRO provides a robust and sensitive tool to detect allelic enrichment

Detection of imprinted expression requires biological replicates and an empirical method to set the FDR from the data in order to take account of the biological and technical variation in the system and minimize the chance of false positives (32). On the other hand the experimental setup and analysis pipeline has to be sensitive enough to detect all imprinted genes in a given tissue. Additionally, previous analytical pipelines require a high level of bioinformatic expertise to implement. In contrast, Allelome.PRO is an efficient package that can function with minimal computer resources (tested on iMac 5.1, 3Gb RAM, Dual-core 2.16 GHz processor), and that based on a limited number of user-set parameters will automatically process the aligned sequencing data provided to set the FDR, categorize allelic enrichment of all loci in an annotation file, and output the analyzed data both as a table, as summary graphs and as a BED file that can be uploaded and viewed on a genome browser (detailed in the manual, Supplementary material).

Allelome.PRO was developed based on the approach taken by Babak and colleagues to detect imprinted expression from RNA-seq data (32,33). Following this we used tissue from F1 offspring from two reciprocal crosses (four samples), combining allelic counts of multiple SNPs within candidate loci, and calculated allelic scores based on the binomial distribution. In contrast to previous approaches (32,33), who calculated the allelic score based on the four possible reciprocal comparisons between the samples, we calculated the allelic score for each of the four samples separately and detected allelic bias as loci where the direction of the bias was the same in all four samples. This allowed us to include all SNPs covered in at least one sample increasing the sensitivity of our method. Additionally, this approach of calculating allelic scores for each sample could be adapted to include more than four samples to increase statistical confidence in situations where reciprocal crosses from inbred strains are not available but SNPs are well-characterized, as is the case in humans. Outbred species where SNPs have not been characterized could also be examined if a SNP calling program such as SAMtools or GATK is first used to call SNPs *de novo* (45,62). To maximize sensitivity to detect allelic expression, we performed RNA-seq using rRNA depleted total RNA, which provided increased coverage of SNP-rich intronic regions. In order to count all reads covering a SNP we trimmed reads covering multiple SNPs so that only a single SNP was counted, rather than excluding SNPs by their distance to other SNPs as done previously (32,33). All of these steps together helped increase the sensitivity of our approach.

In order to empirically calculate the FDR based on the data, previous approaches used the two possible mock comparisons between F1 samples of the same genotype, a comparison that should lack allelic differences (32,33). In contrast, we did mock analysis by inverting the scores of two biological replicates and comparing the four samples, exactly as for calculating the normal score, thereby using the variation in biallelic genes to calculate the FDR. Additionally, in contrast to previous approaches that calculated the FDR by comparing the imprinted score for reciprocal and mock comparisons alone (32,33), we included both imprinted and strain bias scores in our reciprocal to mock comparisons to calculate a single allelic FDR cutoff. Basing an FDR cutoff on imprinted genes alone does not allow a low cutoff to be set due to the limited number of imprinted genes. For example, with the 40 imprinted genes detected by RNA-seq in this study our FDR cutoff of 1% would not be reached until 0 genes are detected in the mock comparisons, making an effective FDR cutoff of 0%. Therefore, by including the several hundred strain bias genes (885 in our study) in the FDR calculation we are able to set a lower FDR cutoff than in previous studies (32,33), increasing the robustness of our pipeline.

A key innovation in the Allelome.PRO pipeline compared the approach taken by Babak *et al.* (32,33), is the introduction of an allelic ratio cutoff, in addition to the allelic score FDR cutoff, to further reduce false positives. One of the caveats of using the binomial distribution is that even small deviations from a 0.5 ratio could result in a score over the FDR if the amount of reads is high enough. As small differences in the ratio are likely due to chance, and even if true, are unlikely to cause a phenotype, we defined an allelic ratio cutoff to separate true allelic biases from stochastic variations. The introduction of an allelic ratio threshold was also proposed by Wang and Clark (63), who suggested a 0.65 ratio based on their experience that imprinted candidates below this ratio could rarely be validated. Other allelic ratios thresholds used in previous studies range from 0.6 (64), to 0.8 (53). Based on the allelic ratio distribution of known imprinted and strain-biased genes in our RNA-

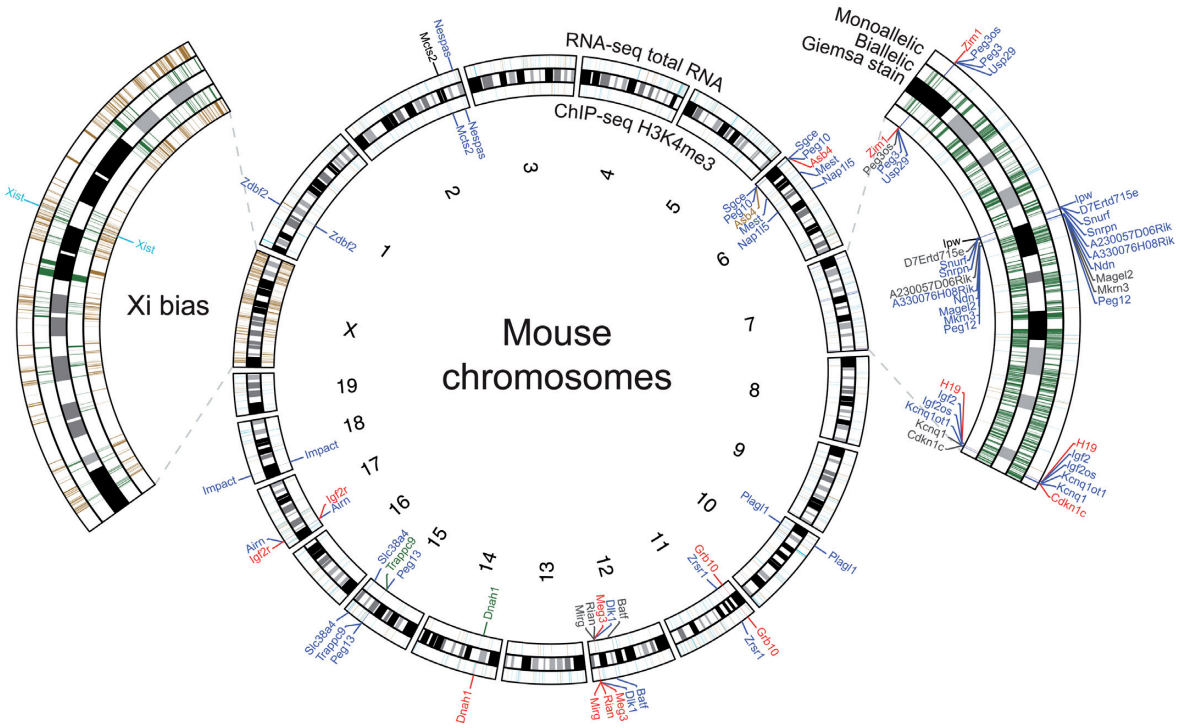


Figure 6. Allelome defined in MEFs. Allelome.PRO results are shown for all chromosomes on a circular representation. The tracks of the main plot in the middle are (from outside to inside): Mono-allelically expressed genes (i.e. strain-biased and imprinted), Giemsa chromosome staining (47) and genes showing allele specific H3K4me3 peaks in their promoter regions based on ChIP-seq. The two enlarged chromosomes were selected to demonstrate both the large amount of imprinted genes on chromosome 7 as well as the strain-biased X inactivation. In addition to the two outmost tracks showing mono-allelic genes also biallelic genes are shown in green for these two chromosomes (inner tracks).

seq data from MEFs, we chose 0.7 as an allelic ratio cutoff. Above this allelic ratio cutoff, 95% of imprinted gene candidates were known imprinted genes, and 85% of X-linked genes that showed a strain bias over the FDR cutoff were included. The introduction of an allelic ratio cutoff allows imprinted and strain bias loci to be distinguished from biallelic loci. In order to classify all annotated loci in an annotation file we defined non-informative genes as those with less SNP coverage than theoretically necessary to overcome the allelic ratio cutoff. This enabled Allelome.PRO to classify all loci as either parental biased (imprinted), strain biased, biallelic, non-informative or lacking SNPs. Therefore, in contrast to previous approaches that sought to detect imprinted expression from RNA-seq data, we provide a tool to categorize the entire allelic expression of all annotated loci in a given tissue, allowing other allelic expression types to be identified and investigated. Moreover, Allelome.PRO is flexible in that it will function with any annotation and sequencing data provided, as demonstrated in this study where both RNA-seq and H3K4me3 ChIP-seq were analyzed and showed a high correlation with each other.

To further test the robustness of the Allelome.PRO pipeline we simulated different rates of sequencing error in our MEF RNA-seq data in the region surrounding the *Igf2r* imprinted gene cluster, and assessed the effect on the Allelome.PRO results (Supplementary Figure S1). We found that *Airn* and *Igf2r* were correctly called imprinted even at error rates of 10 and 15% when the number of aligned reads was greatly reduced. The biallelic gene *Mrpl18* also remained biallelic at a 10% error rate before becoming non-informative at a 15% error rate. However, at these higher error rates the biallelic gene *Tcp1* gene became FVB strain biased, perhaps because FVB has less SNPs with the reference genome compared to CAST, and therefore more FVB reads may align, indicating that strain biased calls may be affected by high rates of sequencing errors (Supplementary Figure S1). To test the impact of experimental error in our method on the Allelome.PRO results we mixed *in silico* reads from FVB and CAST adult heart to create four pools that mimicked two forward and reverse crosses required for Allelome.PRO (detailed in methods). The allelic ratios showed the expected Gaussian distribution centered around 0.5, with deviations from the mean likely due to strain biased genes (Supplementary Figure S2A). As no imprinted genes are expected in such a mixing experiment, we defined the FDR as the percentage of informative genes called imprinted. The FDR was low (0.15%) and could be further decreased by increasing the allelic ratio cutoff or minread parameter (Supplementary Figure S2B), although increasing the minread parameter also decreased the number of informative genes, thus reducing sensitivity (Supplementary Figure S2C). At the 0.7 allelic ratio cutoff and minread 1 settings used in this manuscript, the FDR was reduced to 0.01%. In summary, analysis of the effects of sequencing errors and general experimental errors show the robustness of the Allelome.PRO pipeline in defining allelic expression.

Allelome.PRO defines the MEF Allelome using RNA-seq and ChIP-seq

We detected allele specific expression in MEFs using Allelome.PRO to analyze RNA-seq data, and then validated the results using Allelome.PRO analysis of H3K4me3 ChIP-seq data and by comparison to a previous study of imprinted expression in MEFs (53). We detected 40 genes showing imprinted expression from RNA-seq data using the RefSeq annotation, and 31 genes from ChIP-seq data using 2 kb ± RefSeq TSS, giving 43 genes that were detected by one or both methods. Twenty eight of the 40 genes that were detected by RNA-seq were validated by differential H3K4me3 enrichment over their promoters (70%), and a further five by detection in a previous study of MEFs (53). Another six genes were detected only by RNA-seq or ChIP-seq, but all were known imprinted genes and had high allelic ratios (0.03 or less), making it likely that they also show imprinted expression in MEFs. In addition to these 38 genes, we were able to detect three of five additional imprinted genes reported by a previous study of MEFs (53) using a sliding window annotation to assay our RNA-seq data. We did not initially detect these genes because they are not in the RefSeq annotation or were excluded because of small size. This indicates that we may detect a limited number of additional imprinted genes if we use other annotations in addition to RefSeq. More imprinted genes were detected by RNA-seq than ChIP-seq indicating it was more sensitive, although 5 of 40 genes appeared to be false positives. In contrast, there was no indication that any of the 31 genes detected by ChIP-seq were falsely called.

Some cases where ChIP-seq did not confirm RNA-seq may be due to incomplete annotation by RefSeq. This can arise if neighbouring imprinted lncRNAs are in fact continuous transcripts. If this is not annotated then multiple genes would be called by RNA-seq, and only 1 by ChIP-seq, due to the single promoter. For example, the neighbouring Riken transcripts A330076H08Rik and A230057D06Rik showed this pattern, with both being called imprinted by RNA-seq and being expressed at a similar low level, but only A330076H08Rik was also detected by ChIP-seq. Genes can have alterative start sites that may not be annotated by RefSeq, which would affect our ChIP-seq analysis based on windows around the RefSeq TSS. This was demonstrated by the example of *Asb4*, which was detected as imprinted in the RNA-seq analysis, but not validated by ChIP-seq analysis using a RefSeq promoter annotation. However, using a sliding window annotation we could validate imprinted expression of *Asb4*, finding differential H3K4me3 enrichment at an alternative promoter annotated in the UCSC gene track (47). Therefore, the validation rate of RNA-seq results by H3K4me3 ChIP-seq could be increased by using a more extensive annotation than RefSeq, such as UCSC genes, using peak calling programs, or by using the unbiased sliding windows approach.

Of the five genes considered false positive, three overlapped known imprinted genes: *Kcnq1* and *Peg3os* were called because of incomplete strand-specificity leading to bleed-through from the antisense overlapping *Kcnq1ot1* and *Peg3* respectively, while *Trappc9* was called because of sense overlap with *Peg3*. All 3 of these genes were lowly expressed compared with the overlapping imprinted genes (RPKM <5% the overlapping gene, Table 1), and were not confirmed by H3K4me3 ChIP-seq analysis. We generated strand-specificity using a method based on dUTP incorporation into second strand cDNA synthesis and subsequent uracil-N-glycosylase degradation (35), where bleed-through may occur due to incomplete degradation of the second strand or spurious second strand synthesis by reverse transcriptase (65). The remaining two genes (*Batf* and *Dnah*) were novel imprinted candidates, and were considered false positives because they had lower allelic ratios than the other imprinted genes detected by RNA-seq, and they were not validated by ChIP-seq or by being previously detected in other studies. These two genes were relatively lowly expressed and could be excluded if we adjusted the minread parameter in the Allelome.PRO pipeline. By increasing this parameter to only include SNPs covered by at least three reads (instead of 1 read) we observed that *Batf* and *Dnah* were then called non-informative. Additionally, *Kcnq1* was also then categorized as non-informative, indicating that increasing this parameter also made the pipeline more resistant to false calls due to bleed-through from the opposite strand. Therefore, we suggest that increasing the minreads parameter should be used to decrease the number of false-positives due to low coverage where no additional validation method, such as ChIP-seq, is available.

Allelome.PRO categorizes the allelic enrichment status of all loci in an annotation in a given tissue, enabling other categories in addition to imprinted genes to be investigated. In MEFs by analysis of RNA-seq data we found 885 genes showing strain bias expression, 34% were CAST biased X-linked genes due to a known bias in X-inactivation (14). The detection of 583 autosomal strain bias genes was a similar number to other studies that employed RNA-seq to investigate eQTLs in mouse adipose tissue (66) and adult liver (28). The number of strain bias genes detected in our study was over 20-fold higher than the number of imprinted genes, illustrating the importance of reciprocal crosses to detect true imprinted expression, and identifying genes that may explain the differences in the phenotype between the CAST and FVB strains. However, only 29% of strain biased genes were validated by the H3K4me3 ChIP-seq analysis, in contrast to the high validation rate of imprinted genes mentioned above, and a 90% validation rate for biallelic genes. Around 46% of genes categorized as strain-biased in RNA-seq showed biallelic H3K4me3 marks. Therefore, it is possible that in some cases the strain-biased levels of these transcripts arose not from allele specific transcription, but rather from allele specific post-transcriptional processing, for example, alternative splicing or alternative UTR generation, or due to strain-biased effects on miRNA-binding and RNA stability (10–12). Besides defining a set of genes that can be used as controls for studies of imprinted and strain biased genes, the identification of biallelically expressed genes can also be of interest in itself. For example, in our study we identified 120 biallelic genes on the X-chromosome despite the bias in X-inactivation. Of these genes 48 were validated by ChIP-seq as showing biallelic expression, including five of nine known X-inactivation escaper genes in our annotation (67). If the allelic cutoff for RNA-seq and ChIP-seq was reduced to 0.6 then seven

biallelic genes were detected including four known escapers, making the remaining three genes strong novel X-inactivation escaper candidates.

In summary, using Allelome.PRO we were able to define the entire allelic expression status of all RefSeq genes from RNA-seq data. Validation of this allelome by differential H3K4me3 enrichment detected from ChIP-seq data created a high confidence set for each category of allelic expression. We also demonstrated that a high confidence allelome could be generated from RNA-seq data alone by changing the user-set minreads parameter in Allelome.PRO, resulting in lowly expressed genes from all categories being classified as non-informative.

Applications of the Allelome.PRO pipeline

Most imprinted genes show tissue-specific imprinted expression, the pattern of which has only been relatively comprehensively characterized for a small number (25,68). Allelome.PRO in conjunction with RNA-seq, and validation by H3K4me3 ChIP-seq, provides a robust and sensitive method to assay a wide range of tissues and developmental time points, thus providing a complete picture of tissue-specific imprinted expression. In addition to known imprinted genes, novel tissue-specific imprinted genes may be uncovered in tissues that have not been thoroughly examined for imprinted expression previously. Such an approach would also classify strain biased genes into those that are found in multiple tissues, and those that show tissue-specificity and are therefore candidates to explain strain difference phenotypes in a particular organ or tissue.

Expression quantitative trait loci (eQTL) are defined as genomic loci that regulate gene expression and can be identified by combining whole genome association studies (GWAS) with differential expression analysis (10). Differentially expressed genes can be identified by differential expression analysis between two genotypes, or by allelic expression analysis from RNA-seq data (28,66). Mapping of *cis*-regulatory regions that may explain differences in expression then requires several generations of breeding from inbred strains in order to generate haplotypes that can then be subject to linkage analysis (10). In this study we demonstrated that Allelome.PRO could detect differential enrichment of H3K4me3 over promoters, indicating that it could also be used to detect differential enrichment in the genome of other histone modifications or chromatin binding proteins from ChIP-seq data. Allelic enrichment of enhancer marks, such as H3K27ac or H3K4me1, could identify eQTLs or enhancers that may regulate nearby strain bias genes detected by RNA-seq. This approach has the advantage over conventional eQTL analysis in that analysis is focused on enhancers rather than on all genetic variation between strains. Additionally, analysis can be conducted on the F1 generation, avoiding the extra breeding required for linkage analysis.

In summary, Allelome.PRO is a novel user-friendly pipeline to investigate allele specific features in high-throughput data using any compatible annotation and SNP file. In this study we showed the use of Allelome.PRO on expression and histone mark data, but allele specific differences of other features like DNA methylation or transcription factor binding could be investigated as well. Furthermore this pipeline is not limited to just one organism but instead it could be used in reciprocal crosses of strains from any given organism as long as a database of SNPs is available to distinguish the two alleles. By integrating analysis of different genomic features, such as expression and histone modifications, Allelome.PRO could be used as part of a toolset to investigate allele specific gene regulation.

ACCESSION NUMBERS

RNA-seq and ChIP-seq data are deposited in the Gene Expression Omnibus (GEO) with the accession number GSE69168. Analyzed data can be viewed on the UCSC genome browser at the following link: https://opendata.cemm.at/barlowlab/. The Allelome.PRO program can be downloaded from the following link: https://sourceforge.net/projects/allelomepro/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

We thank Florian Breitwieser for advice during the early stages of this project. High-throughput sequencing was conducted by the Biomedical Sequencing Facility (BSF) at CeMM in Vienna.

FUNDING

Austrian Science Fund [FWF P25185-B22, FWF F4302-B09, FWFW1207-B09]. Funding for open access charge: Austrian Science Fund.
Conflict of interest statement. None declared.

REFERENCES

1. White,J.K., Gerdin,A.K., Karp,N.A., Ryder,E., Buljan,M., Bussell,J.N., Salisbury,J., Clare,S., Ingham,N.J., Podrini,C. *et al.* (2013) Genome-wide generation and systematic phenotyping of knockout mice reveals new roles for many genes. *Cell*, **154**, 452–464.
2. Mostoslavsky,R., Alt,F.W. and Rajewsky,K. (2004) The lingering enigma of the allelic exclusion mechanism. *Cell*, **118**, 539–544.
3. Rodriguez,I. (2013) Singular expression of olfactory receptor genes. *Cell*, **155**, 274–277.
4. Chen,W.V. and Maniatis,T. (2013) Clustered protocadherins. *Development*, **140**, 3297–3302.
5. Gendrel,A.V. and Heard,E. (2014) Noncoding RNAs and Epigenetic Mechanisms During X-Chromosome Inactivation. *Annu. Rev. Cell Dev. Biol.*, **30**, 561–580.
6. Dupont,C. and Gribnau,J. (2013) Different flavors of X-chromosome inactivation in mammals. *Curr. Opin. Cell Biol.*, **25**, 314–321.
7. Gimelbrant,A., Hutchinson,J.N., Thompson,B.R. and Chess,A. (2007) Widespread monoallelic expression on human autosomes. *Science*, **318**, 1136–1140.
8. Zwemer,L.M., Zak,A., Thompson,B.R., Kirby,A., Daly,M.J., Chess,A. and Gimelbrant,A.A. (2012) Autosomal monoallelic expression in the mouse. *Genome Biol.*, **13**, R10.
9. Deng,Q., Ramskold,D., Reinius,B. and Sandberg,R. (2014) Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, **343**, 193–196.
10. Gilad,Y., Rifkin,S.A. and Pritchard,J.K. (2008) Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet.*, **24**, 408–415.

11. Majewski,J. and Pastinen,T. (2011) The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet.*, **27**, 72–79.
12. Li,G., Bahn,J.H., Lee,J.H., Peng,G., Chen,Z., Nelson,S.F. and Xiao,X. (2012) Identification of allele-specific alternative mRNA processing via transcriptome sequencing. *Nucleic Acids Res.*, **40**, e104.
13. Chadwick,L.H., Pertz,L.M., Broman,K.W., Bartolomei,M.S. and Willard,H.F. (2006) Genetic control of X chromosome inactivation in mice: definition of the Xce candidate interval. *Genetics*, **173**, 2103–2110.
14. Calaway,J.D., Lenarcic,A.B., Didion,J.P., Wang,J.R., Searle,J.B., McMillan,L., Valdar,W. and Pardo-Manuel de Villena,F. (2013) Genetic architecture of skewed X inactivation in the laboratory mouse. *PLoS Genet.*, **9**, e1003853.
15. Barlow,D.P. and Bartolomei,M.S. (2014) Genomic imprinting in mammals. *Cold Spring Harbor Perspect. Biol.*, **6**, a018382.
16. Koerner,M.V., Pauler,F.M., Huang,R. and Barlow,D.P. (2009) The function of non-coding RNAs in genomic imprinting. *Development*, **136**, 1771–1783.
17. Bonasio,R. and Shiekhattar,R. (2014) Regulation of transcription by long noncoding RNAs. *Annu. Rev. Genet.*, **48**, 433–455.
18. 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.
19. Hudson,Q.J., Seidl,C.I., Kulinski,T.M., Huang,R., Warczok,K.E., Bittner,R., Bartolomei,M.S. and Barlow,D.P. (2011) Extra-embryonic-specific imprinted expression is restricted to defined lineages in the post-implantation embryo. *Dev. Biol.*, **353**, 420–431.
20. Okae,H., Hiura,H., Nishida,Y., Funayama,R., Tanaka,S., Chiba,H., Yaegashi,N., Nakayama,K., Sasaki,H. and Arima,T. (2012) Re-investigation and RNA sequencing-based identification of genes with placenta-specific imprinted expression. *Hum. Mol. Genet.*, **21**, 548–558.
21. Proudhon,C. and Bourc'his,D. (2010) Identification and resolution of artifacts in the interpretation of imprinted gene expression. *Brief. Funct. Genomics*, **9**, 374–384.
22. Varmuza,S. and Miri,K. (2015) What does genetics tell us about imprinting and the placenta connection? *Cell Mol. Life Sci.*, **72**, 51–72.
23. Stringer,J.M., Pask,A.J., Shaw,G. and Renfree,M.B. (2014) Post-natal imprinting: evidence from marsupials. *Heredity*, **113**, 145–155.
24. Patten,M.M., Ross,L., Curley,J.P., Queller,D.C., Bonduriansky,R. and Wolf,J.B. (2014) The evolution of genomic imprinting: theories, predictions and empirical tests. *Heredity*, **113**, 119–128.
25. Prickett,A.R. and Oakey,R.J. (2012) A survey of tissue-specific genomic imprinting in mammals. *Mol. Genet. Genomics*, **287**, 621–630.
26. Wang,X., Sun,Q., McGrath,S.D., Mardis,E.R., Soloway,P.D. and Clark,A.G. (2008) Transcriptome-wide identification of novel imprinted genes in neonatal mouse brain. *PLoS One*, **3**, e3839.
27. Wang,X., Soloway,P.D. and Clark,A.G. (2011) A survey for novel imprinted genes in the mouse placenta by mRNA-seq. *Genetics*, **189**, 109–122.
28. Lagarrigue,S., Martin,L., Hormozdiari,F., Roux,P.F., Pan,C., van Nas,A., Demeure,O., Cantor,R., Ghazalpour,A., Eskin,E. *et al.* (2013) Analysis of allele-specific expression in mouse liver by RNA-Seq: a comparison with Cis-eQTL identified using genetic linkage. *Genetics*, **195**, 1157–1166.
29. Babak,T., Deveale,B., Armour,C., Raymond,C., Cleary,M.A., van der Kooy,D., Johnson,J.M. and Lim,L.P. (2008) Global survey of genomic imprinting by transcriptome sequencing. *Curr. Biol.: CB*, **18**, 1735–1741.
30. Xie,W., Barr,C.L., Kim,A., Yue,F., Lee,A.Y., Eubanks,J., Dempster,E.L. and Ren,B. (2012) Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell*, **148**, 816–831.
31. Gregg,C., Zhang,J., Weissbourd,B., Luo,S., Schroth,G.P., Haig,D. and Dulac,C. (2010) High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *Science*, **329**, 643–648.
32. DeVeale,B., van der Kooy,D. and Babak,T. (2012) Critical evaluation of imprinted gene expression by RNA-Seq: a new perspective. *PLoS Genet.*, **8**, e1002600.
33. Babak,T. (2012) Identification of imprinted loci by transcriptome sequencing. *Methods Mol. Biol.*, **925**, 79–88.
34. 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.
35. Sultan,M., Dokel,S., Amstislavskiy,V., Wuttig,D., Sultmann,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.
36. Regha,K., Sloane,M.A., Huang,R., Pauler,F.M., Warczok,K.E., Melikant,B., Radolf,M., Martens,J.H., Schotta,G., Jenuwein,T. *et al.* (2007) Active and repressive chromatin are interspersed without spreading in an imprinted gene cluster in the mammalian genome. *Mol. Cell*, **27**, 353–366.
37. 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. *Bioinformatics*, **29**, 15–21.
38. Wu,T.D. and Nacu,S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**, 873–881.
39. Kim,D., Pertea,G., Trapnell,C., Pimentel,H., Kelley,R. and Salzberg,S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**, R36.
40. 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.
41. Morison,I.M., Paton,C.J. and Cleverley,S.D. (2001) The imprinted gene and parent-of-origin effect database. *Nucleic Acids Res.*, **29**, 275–276.
42. Morison,I.M., Ramsay,J.P. and Spencer,H.G. (2005) A census of mammalian imprinting. *Trends Genet.*, **21**, 457–465.
43. Williamson,C.M., Blake,A., Thomas,S., Beechey,C.V., Hancock,J., Cattanach,B.M. and Peters,J. (2013) World Wide Web Site - Mouse Imprinting Data and References.
44. R Core Team. (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.
45. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and Genome Project Data Processing, S. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
46. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
47. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
48. Smith,R.J., Dean,W., Konfortova,G. and Kelsey,G. (2003) Identification of novel imprinted genes in a genome-wide screen for maternal methylation. *Genome Res.*, **13**, 558–569.
49. Choi,J.D., Underkoffler,L.A., Wood,A.J., Collins,J.N., Williams,P.T., Golden,J.A., Schuster,E.F. Jr, Loomes,K.M. and Oakey,R.J. (2005) A novel variant of Inpp5f is imprinted in brain, and its expression is correlated with differential methylation of an internal CpG island. *Mol. Cell. Biol.*, **25**, 5514–5522.
50. Wood,A.J., Roberts,R.G., Monk,D., Moore,G.E., Schulz,R. and Oakey,R.J. (2007) A screen for retrotransposed imprinted genes reveals an association between X chromosome homology and maternal germ-line methylation. *PLoS Genet.*, **3**, e20.
51. Peters,J. and Williamson,C.M. (2008) Control of imprinting at the Gnas cluster. *Adv. Exp. Med. Biol.*, **626**, 16–26.
52. Shiura,H., Nakamura,K., Hikichi,T., Hino,T., Oda,K., Suzuki-Migishima,R., Kohda,T., Kaneko-ishino,T. and Ishino,F. (2009) Paternal deletion of Meg1/Grb10 DMR causes maternalization of the Meg1/Grb10 cluster in mouse proximal Chromosome 11 leading to severe pre- and postnatal growth retardation. *Hum. Mol. Genet.*, **18**, 1424–1438.
53. Tran,D.A., Bai,A.Y., Singh,P., Wu,X. and Szabo,P.E. (2014) Characterization of the imprinting signature of mouse embryo fibroblasts by RNA deep sequencing. *Nucleic Acids Res.*, **42**, 1772–1783.
54. Schulz,R., McCole,R.B., Woodfine,K., Wood,A.J., Chahal,M., Monk,D., Moore,G.E. and Oakey,R.J. (2009) Transcript- and tissue-specific imprinting of a tumour suppressor gene. *Hum. Mol. Genet.*, **18**, 118–127.

55. Mikkelsen,T.S., Ku,M., Jaffe,D.B., Issac,B., Lieberman,E., Giannoukos,G., Alvarez,P., Brockman,W., Kim,T.K., Koche,R.P. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.
56. Babak,T., DeVeale,B., Tsang,E.K., Zhou,Y., Li,X., Smith,K.S., Kukurba,K.R., Zhang,R., Li,J.B., van der Kooy,D. *et al.* (2015) Genetic conflict reflected in tissue-specific maps of genomic imprinting in human and mouse. *Nat. Genet.*, **47**, 544–549.
57. Crowley,J.J., Zhabotynsky,V., Sun,W., Huang,S., Pakatci,I.K., Kim,Y., Wang,J.R., Morgan,A.P., Calaway,J.D., Aylor,D.L. *et al.* (2015) Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. *Nat. Genet.*, **47**, 353–360.
58. Serre,D., Gurd,S., Ge,B., Sladek,R., Sinnett,D., Harmsen,E., Bibikova,M., Chudin,E., Barker,D.L., Dickinson,T. *et al.* (2008) Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic cis-acting mechanisms regulating gene expression. *PLoS Genet.*, **4**, e1000006.
59. Xu,X., Wang,H., Zhu,M., Sun,Y., Tao,Y., He,Q., Wang,J., Chen,L. and Saffen,D. (2011) Next-generation DNA sequencing-based assay for measuring allelic expression imbalance (AEI) of candidate neuropsychiatric disorder genes in human brain. *BMC Genomics*, **12**, 518.
60. Lee,R.D., Song,M.Y. and Lee,J.K. (2013) Large-scale profiling and identification of potential regulatory mechanisms for allelic gene expression in colorectal cancer cells. *Gene*, **512**, 16–22.
61. Gee,F., Clubbs,C.F., Raine,E.V., Reynard,L.N. and Loughlin,J. (2014) Allelic expression analysis of the osteoarthritis susceptibility locus that maps to chromosome 3p21 reveals cis-acting eQTLs at GNL3 and SPCS1. *BMC Med. Genet.*, **15**, 53.
62. McKenna,A., Hanna,M., Banks,E., Sivachenko,A., Cibulskis,K., Kernytsky,A., Garimella,K., Altshuler,D., Gabriel,S., Daly,M. *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.
63. Wang,X. and Clark,A.G. (2014) Using next-generation RNA sequencing to identify imprinted genes. *Heredity*, **113**, 156–166.
64. Smith,R.M., Webb,A., Papp,A.C., Newman,L.C., Handelman,S.K., Suhy,A., Mascarenhas,R., Oberdick,J. and Sadee,W. (2013) Whole transcriptome RNA-Seq allelic expression in human brain. *BMC Genomics*, **14**, 571.
65. Parkhomchuk,D., Borodina,T., Amstislavskiy,V., Banaru,M., Hallen,L., Krobitsch,S., Lehrach,H. and Soldatov,A. (2009) Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res*, **37**, e123.
66. Hasin-Brumshtein,Y., Hormozdiari,F., Martin,L., van Nas,A., Eskin,E., Lusis,A.J. and Drake,T.A. (2014) Allele-specific expression and eQTL analysis in mouse adipose tissue. *BMC Genomics*, **15**, 471.
67. Yang,F., Babak,T., Shendure,J. and Disteche,C.M. (2010) Global survey of escape from X inactivation by RNA-sequencing in mouse. *Genome Res.*, **20**, 614–622.
68. Kulinski,T.M., Barlow,D.P. and Hudson,Q.J. (2013) Imprinted silencing is extended over broad chromosomal domains in mouse extra-embryonic lineages. *Curr. Opin. Cell Biol.*, **25**, 297–304.

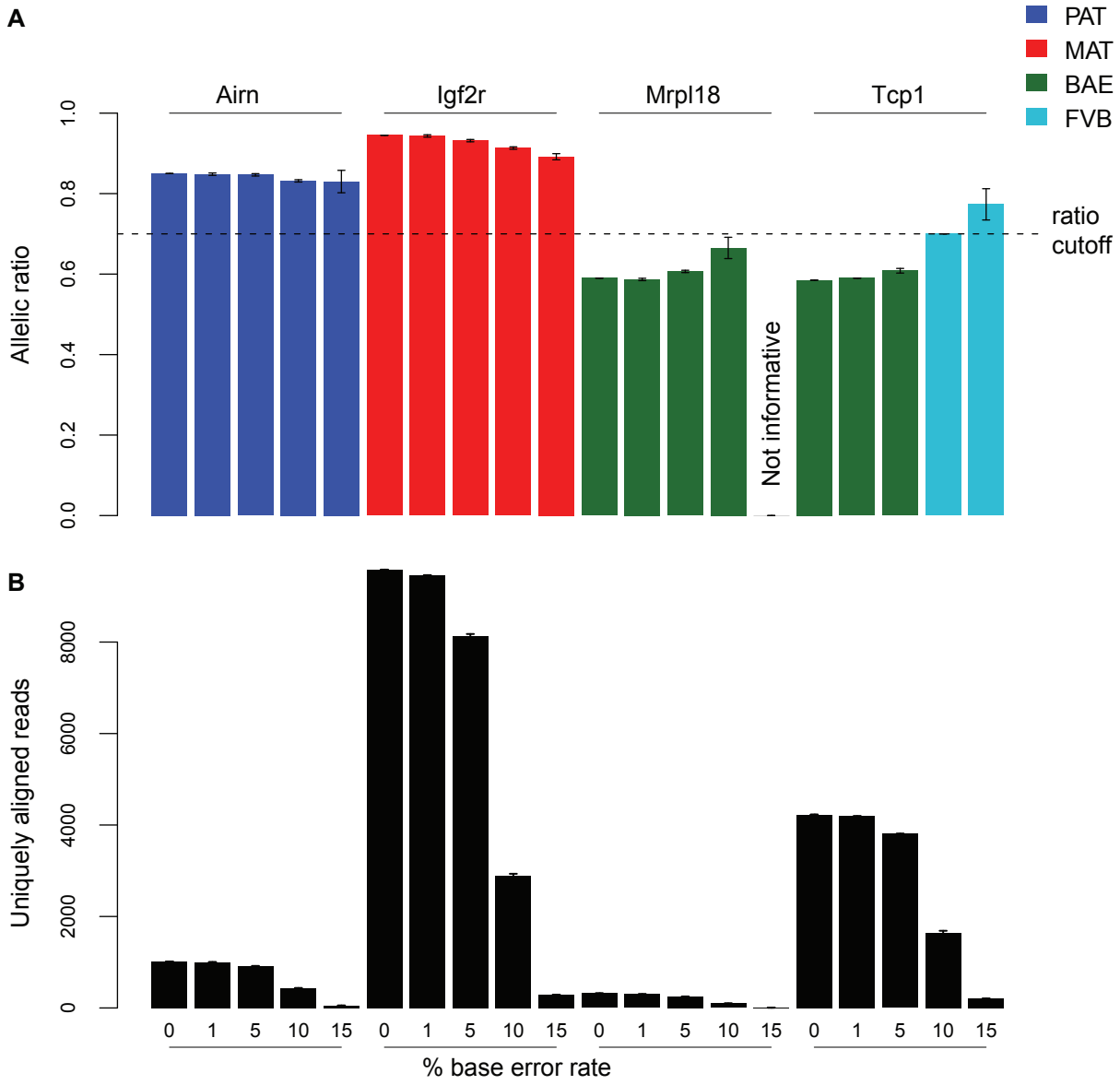


Figure S1. Allelic ratios generated by Allelome.PRO are robust over a wide range of sequencing error rates

(A) Allelic ratios of genes in the Igf2r region at different sequencing error rates. The imprinted ratio is plotted for *Airn* and *Igf2r*, and the strain biased ratio for *Mrpl18* and *Tcp1*. PAT paternal, MAT maternal, BAE biallelic, FVB strain bias.

(B) The number of uniquely aligned reads for each gene at different sequencing error rates.

Error bars show the standard deviation from generating random sequencing errors 3 times.

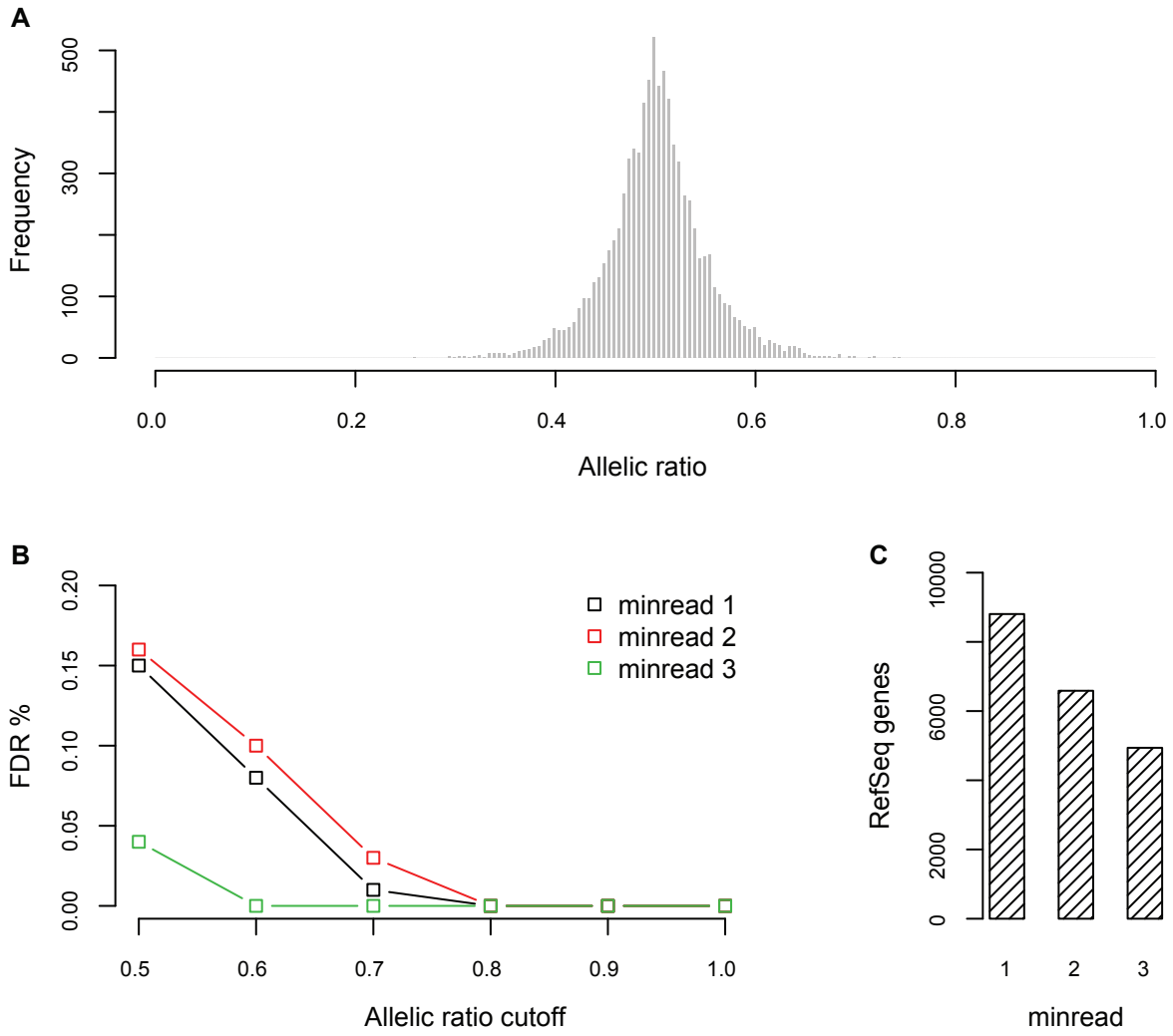


Figure S2. Determination of experimental error by in silico mixing of RNA-seq data from FVB/N and CAST/EiJ adult heart.

(A) Allelic ratio distribution of informative RefSeq genes generated by Allelome.PRO from a 50:50 mix of uniquely aligned reads from FVB and CAST.

(B) Determination of imprinted expression false discovery rate (FDR), defined as the percentage of informative genes called imprinted by Allelome.PRO at different allelic ratio cutoffs and minread parameter settings. The FDR is reduced when the minread parameter or the allelic ratio cutoff is increased.

(C) Increasing the minread parameter decreases the number of informative genes. The number of informative RefSeq genes at different minread settings is shown for an allelic ratio cutoff of 0.7.

AlleleSeq.PRO

Defining allele-specific expression in high throughput sequencing data

MANUAL

Daniel Andergassen & Christoph Dotter

Contents

1 Introduction

Allelome.PRO was developed in the group of Denise Barlow at the CeMM Research Center of Molecular Medicine of the Austrian Academy of Sciences (Vienna, Austria) as a fully automated user-friendly bioinformatics pipeline which uses high throughput sequencing data of four tissue samples from reciprocal crosses from genetically distinct mouse strains to detect allele-specific features. These features include allele-specific expression and allele-specific histone marks as demonstrated in the original publication:

Andergassen D, Dotter CP, Kulinski TM, Guenzl PM, Barlow DP, Pauler FM and Hudson QJ. Allelome.PRO, a pipeline to define allele-specific genomic features from high-throughput sequencing data. (Manuscript submitted 2015) [1]

When using this tool for a publication please cite the original publication.

2 Installation

2.1 Hardware requirements

During the runs of Allelome.PRO for the original publication we ran Allelome.PRO for data of one strand from strand specific RNA sequencing data using the RefSeq annotation. This translated into on average 41.8 million reads per sample for an annotation of around 15,000 genes per run, with a SNP file containing around 20 million SNPs. For these runs we observed that Allelome.PRO allocated a maximum of 3 GB of memory. The pipeline was designed to run inside a computer cluster environment with sufficient memory but runs also on a computer with less resources (tested on an iMac 5.1, 3Gb RAM, Dual-core 2.16Ghz processor).

2.2 Software dependencies

The pipeline was designed for Linux-like operating systems and was tested on Linux and Mac OS X.

Software required by the core pipeline

The pipeline requires the following programs/toolsets:

- bedtools ($\geq$ version 2.20.1) [2]
- SAMtools ($\geq$ version 0.1.19) [3]
- R ($\geq$ version 3.0.2) [4] + plyr package (will be automatically installed if internet access is possible)
- Perl ($\geq$ version 5.20.0)

All required software has to be located within the paths provided by the `PATH` environment variable. For instructions on how to set the PATH variable for your system please refer to one of the following pages:

- Instructions for setting the PATH variable in Linux/UNIX based systems
- Instructions for setting the PATH variable in Mac OS X.

Suggested additional software

For the alignment of RNA sequencing data as well as ChIP sequencing data we suggest the use of the STAR aligner (version $\geq$ 2.3.1) [5]. This is based on a comparison of three different aligners as described in the original publication.

2.3 Allelome.PRO content

The program archive is available at https://sourceforge.net/projects/allelomepro/. The archive contains the main pipeline shell script `allelome_pro.sh` as well as a folder `scripts` which contains scripts used by the main pipeline. In addition to that a helper script to create the SNP bed file (see 3.1.3) is included. A summary of all deployed files is given in table 1.

Script	Description
<base>/	
allelome_pro.sh	The main script that is **called by the user**.
<base>/scripts/	
bamtrim.sh bamtrim.pl	Trim reads covering multiple SNPs so they just cover one. This is done to prevent multiple counting of reads. (for details please refer to the publication).
pileup_filter.pl	Handles spliced reads/indels in the pileup file.
read_count.pl	Sums up number of reads for each variant at SNP positions.
score.R	Statistical scoring and categorisation of the candidates.
bed_creator.sh	Creates color-coded output bed files.
<base>/helperscripts/	
createSNPbedfile.sh	Prepares the SNP input file (see chapter 3.1.3). usage: createSNPbedfile.sh <vcf_file> <snp_file>
separate_BAM_strand.pl	Divides reads from an aligned BAM file into two files usage: separate_BAM_strand.pl containing reads from the forward and reverse strand, respectively (see chapter 3.3.1).

Table 1: Allelome.PRO content. This table lists the scripts that are part of the Allelome.PRO core pipeline and shortly describes their purpose. All scripts besides the main script are located in the scripts folder.

3 Usage

The usage of Allelome.PRO requires three steps:

1. Prepare the required input files.
2. Set up the configuration file for the needs of your analysis
3. Run the pipeline

3.1 Input file requirements

The pipeline requires three types of input files: An annotation file, four BAM files containing the aligned sequencing data and a file containing information about the SNPs between the crosses used. The exact format requirements along with examples for each will be described in the next three sections.

3.1.1 The annotation file

The annotation file is in BED6 format (see UCSC format description for more details), meaning it has six columns containing the information listed in table 2. In the simplest case it contains one line per candidate that should be categorised. It is also possible to include multiple lines per candidate (e.g. multiple PCR products for one gene, multiple isoforms of the same gene in an annotation such as RefSeq) which will be combined to one during the analysis. This is made possible by the fact that the pipeline groups lines in the annotation if they have the same name and are located on the same chromosome and strand, summing up SNPs covered by at least one of the grouped entries. Users should keep this grouping feature in mind when they curate their annotation to avoid errors. If, for example, the user wants to score different isoforms independently the names in the annotation will need to be different from each other (e.g. start with consecutive numbers).

Column	Information
1	Chromosome (written as e.g. chr1, chrX)
2	Start Position
3	End Position
4	Name (e.g. gene name)
5	Score (not used here)
6	Strand (e.g. +,- or . for not defined)

Table 2: The BED6 format. This format is used for both the annotation file and the SNP file.

To give some examples, here are the annotations used by the original publication:

- The "RefSeq Genes" annotation [6] (obtained via the UCSC table browser [7]).
- Sliding window annotations over the whole genome.
- A custom annotation of RefSeq Gene promoter regions.

3.1.2 The aligned BAM files

The pipeline requires four aligned BAM files derived from samples of two reciprocal crosses with two samples being from one cross, while the other two are from the other cross. This was tested for samples from RNA sequencing and ChIP sequencing experiments on an Illumina® sequencing system. The BAM files are ideally sorted by

leftmost coordinates (as done by samtools sort), but we also implemented an option to sort the BAM files before processing (see section 3.2).

3.1.3 The SNP file

The SNP file is also in BED6 format (see table 2) with the additional requirement that the name consists of only two letters indicating the two SNP variants present in the two strains of the crosses. The order of these two letters is important. The first letter indicates the variant in strain 1, while the second one indicates the variant in strain 2. **The way the SNP file is created therefore defines which strain will be "strain 1" and which one will be "strain 2" during the course of the analysis.** This should be kept in mind as some parameters in the configuration file need to be set according to this definition. SNP positions have to be based on the same reference genome that the BAM files were aligned to (e.g. mm10). One source for SNP data is the FTP site of the Sanger institute. The downloaded compressed VCF file (e.g. `mgp.v3.snps.rsIDdbSNPv137.vcf.gz` as used in the publication) can then be extracted and further processed using the included helper script `createSNPbedfile.sh`. The script takes two parameters, the first one being the VCF file, the second one being the desired SNP bed file name. Once started it lists all the strains for which the VCF file contains information and lets the user select the strains he wants to use. Afterwards it extracts the variant information for each SNP with different variants in the two crosses. Only high confidence SNPs homozygous in both strains are considered.

Once all input files are prepared, proceed with setting up the configuration file.

3.2 The configuration file

The configuration file is written in the parameter=value format (note: no spaces allowed). The available parameters, along with a short description are listed in table 3. A template configuration file containing documentation is included in the distribution.

General Parameters	
`pipe_location`	Complete path to the folder where the main pipeline script is located.
`ratio`	Minimum allelic ratio above which allele-specific expression is called biologically relevant. Default value: 0.7 - represents a 70:30 ratio between the two alleles.
`fdr_param`	The false-discovery rate set as how many candidates were called allele-specific in the mock analysis compared to the results. Default value: 1 - represents 1%, meaning that for each 100 genes categorised as allele-specific in the results, one call is allowed in the mock comparison
`minreads`	Minimum number of reads that must cover a SNP position for the SNP to be included in the analysis. Default value: 1 - include all covered SNPs.

Experiment-specific Parameters	
`outputdir`	Path where the job directory containing all output files will be created. The job directory name will contain the date, sample file names, annotation file name, and the FDR and minreads parameter.
`annotation`	Annotation file containing the candidates to analyse (see 3.1.1).
`main_title`	Title of the analysis, used for plot captions and file names. Note: No spaces or special characters are allowed here.
`y_axis`	Y axis for the result plot, describes the type of candidates. Examples: RefSeq Genes, windows.
`sorted`	Specify whether the BAM files are sorted or not. Default value: 1 - the files are sorted; 0 - unsorted.

Cross-specific Parameters	
`snp_file`	The SNP file used in this analysis (see 3.1.3).
`strains`	Labels for the two strains, separated by a semicolon (;). Example: CAST;FVB

`for_c1`
`for_c2`
`rev_c1`
`rev_c2`
	The four BAM files containing the aligned sequencing data. for_c1 and for_c2 are the samples of the forward cross, meaning the cross where the mother is of strain 1 and the father is of strain 2. rev_c1 and rev_c2 are the samples of the reverse cross in the opposite direction. Strains are defined via SNP file (see 3.1.3).

Table 3: Parameters in the configuration file.

3.3 Run

Syntax for starting the pipeline:
`<pipeline-dir>/allelome_pro.sh -c <path-to-configfile>`

3.3.1 Strand-specific analysis as performed in Andergassen and Dotter et al

It is well accepted that genes show complex spatial organization resulting in transcription from the same genomic region albeit from different DNA strands. These overlapping transcription units can show profound differences in their allelic expression pattern. To resolve this complexity RNA-Seq methodologies have been developed that retain the information of the strand that a particular RNA was transcribed from. To keep the core pipeline of Allelome.PRO as simple as possible we have not implemented an automatic "strand specific" analysis yet. We describe a workflow in the original publication that is based on the separate analysis of RNA-Seq reads originating from the forward and from the reverse strand and provide the necessary script to follow this workflow in this package. This script, `separate_BAM_strand.pl` can be found in the helperscripts folder of this package. The syntax for the script is as follows: `separate_BAM_strand.pl <bam_file> <strand_rule> <output_folder>` and it creates two files named `<bam_file>.fwd.bam` and `<bam_file>.rev.bam`. The strand rule indicates how the reads should be divided. For more documentation on the choice of strand rule please refer to the documentation in the header of the script. The pipeline can then be started for each strand separately. In addition to the required separation of reads it is also advisable to split the used annotation into forward and reverse strand as well and using the matching annotation for each of the two pipeline runs. Afterwards the results can be combined by concatenating the respective result tables. An option to generate combined graphical output is not implemented yet.

3.4 Output

The result directory contains the files and folders listed in table 4

Name	Description
<main>/	
<main_title>_IG.txt	Contains all loci categorised as imprinted after both FDR and ratio filtering.
<main_title>_SG.txt	Contains all loci categorised as strain biased after both FDR and ratio filtering.
<main_title>_locus_full.txt	Information about the categorisation of all loci in the annotation.
<main_title>_SNP_full.txt	Information about the categorisation of all SNPs in the annotation.
<main_title>.pdf	Graphical output of the allelome data.
info.txt	Additional information about the run
BED_files/	
<main_title>_locus.bed	BED6-file containing all loci, color-coded according to their categorisation
<main_title>_SNP.bed	BED6-file containing all SNPs, color-coded according to their categorisation
debug	Folder containing all files created during the run.

Table 4: Result files and folders. All resulting files and folders created by the pipeline. <main> represents the main output folder created by the pipeline, while <main-title> represents the title specified in the configuration file.

3.4.1 Result tables

The four result tables can be separated into two groups. First, <main_title>_IG.txt and <main-title>_SG.txt contain only the respective subset of the annotated loci categorised as showing imprinted and strain biased expression, respectively. The columns are listed in table 5. Colums 8 and 9 are different between the two files, with <main_title>_IG.txt containing I_score and I_ratio while <main_title>_SG.txt contains S_score and S_ratio.

Columns 1-6 of the files <main_title>_locus_full.txt and _SNP_full.txt are the same as listed in table 5. The seventh column, total_reads_min, shows the minimum number of reads covering SNPs in this locus across the four replicates. Column 8 is the same RPSM_min column as before, columns 9 and 10 contain the imprinting score and ratio, columns 11 and 12 contain the strain score and ratio and column 13 contains the tag.

Nr.	Column	Description
1	**chr**	
2	**start**	
3	**end**	Locus information from the annotation file.
4	**name**	
5	**strand**	
6	**cov_SNP_min**	Minimum number of SNPs covered in a single biological replicate.
7	**RPSM_min**	Minimum RPSM (**R**eads **P**er **S**NP per **M**illion total SNP covering reads) value across the four replicates. This gives an estimate of the expression level of the locus.
8	**I_score** or **S_score**	The "imprinting score", calculated as allelic score between maternal and paternal allele or the "strain score", calculated as allelic score between the alleles of strains 1 and 2. These scores are equal to the minimum score across the four replicates.
9	**I_ratio** or	Average maternal:total ratio across the four replicates.
	S_ratio	Average strain1:total ratio across the four replicates.
10	**tag**	Result of the categorisation.

Table 5: Result table columns. Listed are the columns of the files `<main-title>_IG.txt` and `<main-title>_SG.txt`, together with a short description of the information contained in them.

3.4.2 Graphical output

The pdf file produced by the pipeline, `<main_title>.pdf`, contains five different plots.

1. A barplot displaying the overall results of the categorisation.
2. A stacked barplot visualising the allelic ratios of the candidates defined as showing imprinted expression.
3. A graph illustrating the calculation of the score cutoff for the loci based on the false discovery rate (for more information please refer to the paper).
4. The same graph for the SNPs.
5. A graph showing the distribution of allelic ratios among the candidates with a score significant enough to pass the FDR cutoff. This graph can be used to determine whether the set ratio cutoff was a good choice and should aid in setting an allelic ratio cutoff. The upper graph shows the imprinted genes, while the bottom one shows strain biased genes.

Since X inactivation in females represents a special case of monoallelic expression, candidates from chromosome X are handled in a special way in these plots.

Chromosome X candidates in plot 1:
Candidates on chromosome X which were categorised as either imprinted or strain biased are displayed as lighter bars stacked atop the autosomal candidates. The numbers above the stacked bars represent the total numbers of imprinted and strain biased candidates from all chromosomes.

Chromosome X candidates in plots 2-4:
In these plots candidates from chromosome X are omitted. This is because imprinted genes on chromosome X are most likely the result of parental specific X inactivation (e.g. in extra-embryonic mouse tissues) and not themselves regulated in an imprinted fashion. This is also the reason why imprinted and strain biased candidates from chromosome X are not included in the calculations for the FDR cutoff.

Chromosome X candidates in plot 5:
Here the chromosome X candidates are not displayed in the plot for the imprinted genes (top) but indicated in the plot for the strain biased genes (bottom). This behaviour was chosen because the amount of imprinted chromosome X genes in a situation of parental specific X inactivation is very high compared to the number of autosomal imprinted candidates. This would distort the ratio distribution.

3.4.3 Result bed files

The two BED9 files `<main_title>_locus.bed` and `<main_title>_SNP.bed` include the data of the full result tables `<main_title>_locus_full.txt` and `_SNP_full.txt`, respectively. These files were created to include the categorisation of the candidates via a color code in the color column of the bed file. Furthermore information about relevant ratios and read counts are encoded in the name column. This is done by appending one or more of the following suffixes to the name of the locus/SNP:

- `_i<ratio>`: `<ratio>` gives the ratio of reads supporting the maternal variant over total reads (shown for imprinted and biallelic loci/SNPs).
- `_s<ratio>`: `<ratio>` gives the strain1 reads/total reads ratio (shown for strain biased and biallelic loci/SNPs).
- `_r<number>`: `<number>` gives the minimum number of SNP covering reads across the four replicates (shown for all loci with SNPs/all SNPs).

- <var1>|<rc_v1_fwd>|<rc_v1_rev>||<var2>|<rc_v2_fwd>|<rc_v2_rev>: Variant info (SNP file only) providing information about the two SNP variants (<var1>/<var2>) and the read counts in the forward (<rc_v1_fwd>, <rc_v2_fwd>) and the reverse cross (<rc_v1_rev>, <rc_v2_rev>) for variant 1 and 2, respectively. Read counts were summed up across the two replicates for each cross to enhance readability.

These BED files can then be used to visualise the results using a genome browser of choice (e.g. the UCSC genome browser) by uploading them as custom tracks.

Bibliography

[1] Andergassen D, Dotter CP, Kulinski TM, Guenzl PM, Barlow DP, Pauler FM, et al. Allelome.PRO, a pipeline to define allele-specific genomic features from high-throughput sequencing data. submitted. 2015;.

[2] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar;26(6):841–842.

[3] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug;25(16):2078–2079.

[4] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2013.

[5] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan;29(1):15–21.

[6] Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, et al. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. 2014 Jan;42(Database issue):D756–D763.

[7] Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, et al. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 2004 Jan;32(Database issue):D493–D496.

2. Publication 2: "Mapping the mouse Allelome reveals tissue-specific regulation"

Authors:

Daniel Andergassen, Christoph P. Dotter, Daniel Wenzel, Verena Sigl, Philipp C. Bammer, Markus Muckenhuber, Daniela Mayer, Tomasz M. Kulinski, Hans-Christian Theussl, Josef M. Penninger, Christoph Bock, Denise P. Barlow*, Florian M. Pauler* and Quanah J. Hudson*

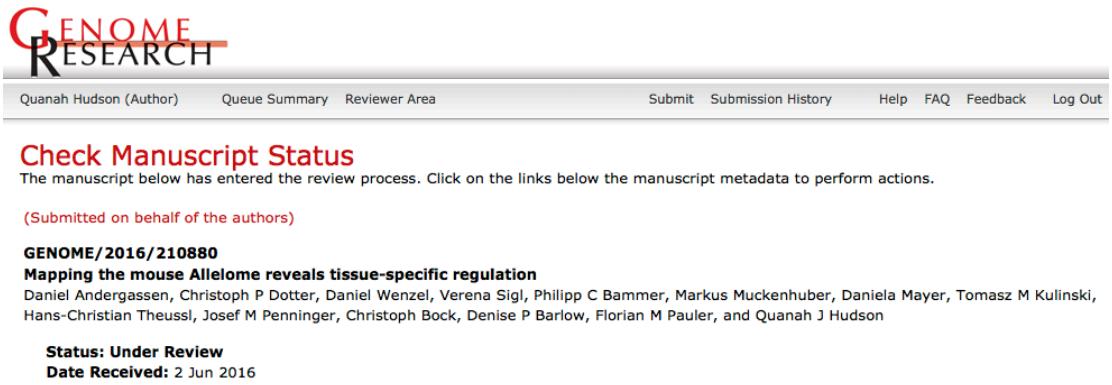
* Corresponding authors

Submitted to Genome Research

Submitted June 02, 2016

Impact factor: 14,63

Current Status: In revision


GENOME RESEARCH

Quanah Hudson (Author) Queue Summary Reviewer Area Submit Submission History Help FAQ Feedback Log Out

Check Manuscript Status

The manuscript below has entered the review process. Click on the links below the manuscript metadata to perform actions.

(Submitted on behalf of the authors)

GENOME/2016/210880

Mapping the mouse Allelome reveals tissue-specific regulation

Daniel Andergassen, Christoph P Dotter, Daniel Wenzel, Verena Sigl, Philipp C Bammer, Markus Muckenhuber, Daniela Mayer, Tomasz M Kulinski, Hans-Christian Theussl, Josef M Penninger, Christoph Bock, Denise P Barlow, Florian M Pauler, and Quanah J Hudson

Status: Under Review
Date Received: 2 Jun 2016

2.1 Contributions

D.A., Q.J.H., F.M.P. and D.P.B. planed the study, provided intellectual input, and wrote the manuscript. D.A., Q.J.H., dissected the majority of mouse tissues. D.A. generated the RNA-seq and ChIP-seq libraries, performed the analysis and prepared the figures for the manuscript. Q.J.H. and P.C.B. prepared the ChIP for the active chromatin marks. C.P.D. generated scripts that automatized the analysis of all the samples and isolated the RNA of the adult tissues. D.W. and H-C.T. isolated embryonic stem cells from mouse blastocysts, which where than tested by D.M. for genetic aberration and cultured by P.C.B in 2i. V.S. showed me how to microdissect mammary glands from adult mice. J.M.P provided supervision for D.W. and V.S.. T.M.K. and M.M helped with the isolation of visceral yolk sac endoderm. C.B. supported me during the revision. All authors read and approved this manuscript.

TITLE PAGE:

Mapping the mouse Allelome reveals tissue-specific regulation

Daniel Andergassen[1], Christoph P. Dotter[1,4], Daniel Wenzel[2], Verena Sigl[2], Philipp C. Bammer[1,5], Markus Muckenhuber[1,6], Daniela Mayer[1,5], Tomasz M. Kulinski[1,7], Hans-Christian Theussl[3], Josef M. Penninger[2], Christoph Bock[1], Denise P. Barlow[1]*, Florian M. Pauler[1,4]* and Quanah J. Hudson[1,8]*

[1]CeMM, Research Center for Molecular Medicine of the Austrian Academy of Sciences, Lazarettgasse 14, AKH BT 25.3,1090 Vienna, Austria

[2]IMBA, Institute of Molecular Biotechnology of the Austrian Academy of Sciences, Dr. Bohr Gasse 3, 1030 Vienna, Austria

[3]IMP/IMBA Transgenic Service, Institute of Molecular Pathology (IMP), Dr. Bohr Gasse 7, 1030 Vienna, Austria

*To whom correspondence should be addressed

quanah.hudson@univie.ac.at, florian.pauler@ist.ac.at, dbarlow@cemm.oeaw.ac.at

Present Address:

[4]ISTA, Institute of Science and Technology Austria, Am Campus 1, 3400 Klosterneuburg, Austria

[5]FMI, Friedrich Miescher Institute for Biomedical Research, 4058 Basel, Switzerland

[6]IMB, Institute of Molecular Biology, Ackermannweg 4, 55128 Mainz, Germany

[7]IBB, Institute of Biochemistry and Biophysics Polish Academy of Sciences, Pawinskiego 5a, 02-106 Warszawa, Poland

[8]IMBA, Institute of Molecular Biotechnology of the Austrian Academy of Sciences, Dr. Bohr Gasse 3, 1030 Vienna, Austria

ABSTRACT

Allele-specific expression is important in development and disease, but a complete picture is lacking. To address this, here we map the mouse Allelome by characterizing allelic expression in 23 tissues during development, including 19 female tissues enabling genes escaping X chromosome inactivation (XCI) to also be detected. By using ribosomal RNA depleted total RNA-seq we were able to map allelic expression of lncRNAs that may be lowly expressed or non-polyadenylated. Analysis of this resource in combination with ChIP-seq and genetic models reveals previously unappreciated aspects of regulation of allelic expression. We demonstrate that allelic expression arising from genetic differences between the alleles, or from the epigenetic processes XCI and genomic imprinting, is surprisingly highly tissue-specific. We find that tissue-specific enhancers marked by H3K27ac can explain tissue-specific allelic expression caused by genetic differences. We detect a higher rate of escape from XCI than most previous studies in mouse, with a mean level of ~15% genes escaping being more similar to reports in human, while leg muscle shows an unexpectedly high rate of ~50% escapers. By surveying an extensive range of tissues during development, and due to the robustness and sensitivity of our approach, we are able to provide a high confidence list of imprinted genes in mouse. This confirms that most imprinted genes (>90%) are clustered, and that cluster size varies dynamically during development, and can be substantially larger than previously thought with genetic the *Igf2r* cluster extending over 10Mb in placenta, making it the largest autosomal *cis*-regulated region.

INTRODUCTION

Allele-specific expression can occur in different contexts during mammalian development and affect a wide-range of processes. Random monoallelic expression at the single cell level has been reported to be relatively common, and plays an important role in the maturation of the lymphoid cell lineage where allelic exclusion of T and B cell receptors is required (Reinius and Sandberg 2015). At the tissue level such cases appear biallelic, but genetic and epigenetic differences between the alleles can lead to consistent biases in populations of cells or the whole organism.

Genetic differences between the alleles of mammalian genes frequently cause allele-specific expression differences in human and mouse (Lappalainen et al. 2013; Crowley et al. 2015). The sequence of the two alleles can vary at single nucleotide polymorphisms (SNPs) that can alter gene expression by modulating transcription factor binding to gene promoters or distal and proximal activating regions called enhancers (Leung et al. 2015). Active enhancers are marked by the H3K27ac histone modification (Creyghton et al. 2010), and can activate one or more promoters by direct interaction (Shlyueva et al. 2014). Allelic expression can also be caused by epigenetic differences between the alleles, notably in the developmentally important processes of X chromosome inactivation (XCI) and genomic imprinting. In both cases a long non-coding (lnc) RNA causes the initiation of silencing, of the whole chromosome in XCI, or a cluster of imprinted genes in the case of imprinted lncRNAs.

In female mammals the *Xist* lncRNA is expressed from one of the two X chromosomes leading to widespread epigenetic silencing of X-linked genes apart from a subset that escape XCI, reported to be 3% in mouse and 15% in human (Berletch et al. 2011), although other reports indicate that the number of escapers in mouse may be higher at around 13% (Calabrese et al. 2012). *Xist* uses the three-dimensional structure of the X-chromosome to gain access to distant parts of the chromosome from which it spreads to eventually coat the whole inactive X and cause XCI (Engreitz et al. 2013). Current evidence indicates that *Xist* initiates silencing by interacting with SPEN that then recruits HDAC3 to cause hypoacetylation of the X chromosome (Chu et al. 2015; McHugh et al. 2015; Monfort et al. 2015). A series of factors are then recruited that establish the repressive chromatin state required to maintaining silencing, including the Polycomb Repressive Complexes 1 and 2 (PRC1 and 2), DNMT1, SAF-A and ASH2L (Wutz 2011).

Imprinted genes are mostly clustered with allele-specific silencing regulated by a distant differential DNA methylated imprint control element (ICE). In the most common mechanism the unmethylated ICE acts as a promoter for a lncRNA that silences a cluster of genes, as shown for *Airn* and *Kcnq1ot1* in the *Igf2r* and *Kcnq1* clusters (Sleutels et al. 2002; Mancini-Dinardo et al. 2006). Both *Airn* and *Kcnq1ot1* have been associated with the histone modifying enzymes EHMT2 and PRC2, and are

thought to guide deposition of H3K9me2 and H3K27me3 to silence distant genes in these clusters (Nagano et al. 2008; Terranova et al. 2008). However, *Airn* directly silences the overlapped *Igf2r* by transcriptional interference, a process not requiring these enzymes (Mager et al. 2003; Nagano et al. 2008; Latos et al. 2012). It has been hypothesized that disruption of enhancer activity may be the first step in initiating silencing of imprinted genes distant from the lncRNA locus (Pauler et al. 2012).

Extensive studies on the influence of SNPs on allelic expression and disease association have been performed in human adult tissues or cell culture models (Leung et al. 2015). RNA-seq on mouse tissues from F1 crosses have been used to detect expression quantitative trait loci (eQTLs) (Keane et al. 2011), escape from XCI (Berletch et al. 2015) and imprinted expression (Babak et al. 2008; Wang et al. 2008; Wang et al. 2011; DeVeale et al. 2012; Okae et al. 2012; Babak et al. 2015), but studies of total allelic expression have been lacking. We have pioneered an approach to classify allelic expression of all genes in a tissue from RNA-seq data (Andergassen et al. 2015), and apply this here to map the allelic expression states of protein-coding (pc) and non-coding (nc) genes in 23 different mouse tissues and developmental stages to define the mouse Allelome. This revealed that biases in allelic expression of pc-genes are highly tissue-specific, while nc-genes tended to show a consistent bias when expressed. Following this, in the 19 females tissues we confirmed reports that XCI escapers can be tissue-specific (Berletch et al. 2015), and found an unusually high proportion of escapers in leg muscle (>50%). By assembling a high confidence list of validated or supported imprinted genes, we found that an even larger proportion than previously thought belong to clusters (>90%), that these clusters can be much larger than previously reported, and that they expand and contract during development, reaching their maximum in extra-embryonic tissues. In particular we found that the *Igf2r* cluster expanded to 10Mb in placenta, representing the largest *cis* co-regulated region outside of the X chromosome. For all types of allelic expression that we investigated we found an association with nearby allele-specific H3K27ac enrichment, indicating that allele-specific expression may be mediated through genetic differences in enhancers or by epigenetic repression established on enhancers during XCI and imprinted silencing.

RESULTS

The mouse gene expression Allelome shows tissue-specific variation

To investigate how allelic expression varies between tissues and during development, we first generated a near complete picture of allelic expression, or the mouse Allelome. We chose a range of 23 pluripotent, embryonic, extra-embryonic, neonatal and adult tissues, including a developmental series for selected tissues (Fig. 1A, Table S1A-B). We placed an emphasis on tissues where imprinted expression has been suggested to play an important role, such as in the energy transfer between the mother and embryo and in neonatal and maternal behavior (Peters 2014; Stringer et al. 2014), which includes tissues like brain and placenta reported to show the most imprinted expression (Babak et al. 2015). Therefore, our samples include a developmental series of brain and the extra-embryonic placenta and visceral yolk sac endoderm (VE) tissues, as well as the neonatal tongue and virgin and lactating mammary gland and brain from the lactating female.

For each tissue we collected four F1 samples from two reciprocal crosses between FVB/NJ (FVB) and CAST/EiJ (CAST) mice. To enable analysis of X chromosome allelic expression we collected single female (XX) organs, except for embryonic day (E) 12.5 liver, E9.5 VE and E12.5 VE where tissues from a litter were pooled (mix of XX/XY), and embryonic stem (ES) cells, which were derived from male (XY) blastocysts. Unsupervised clustering of all total RNA sequencing (RNA-seq) samples confirmed the quality of the dataset by showed that replicates of the same tissue clustered together, closest to the same organ at different developmental stages as expected (Supplemental Fig. S1A). We analyzed this data for biases in allelic expression using Allelome.PRO with an allelic ratio cutoff of 0.7 (Andergassen et al. 2015), a custom annotation and SNPs from the Sanger database (Keane et al. 2011). We previously validated the Allelome.PRO strategy in F1 crosses of inbred mouse strains (Andergassen et al. 2015), and the approach is described in more detail in this paper and accompanying manual, as well as in the Supplemental Methods and Fig. S2A. To generate a comprehensive annotation that covered all transcripts present in our dataset we combined the RefSeq mouse annotation for pc- and nc-genes (Pruitt et al. 2014), with nc-loci not in RefSeq detected by reference based assembly from our data, as detailed in the Methods. Analysis with RNAcode and CPC indicated that the coding potential of our nc-loci was significantly less than for pc-genes, but not distinguishable from RefSeq nc-genes (Kong et al. 2007; Washietl et al. 2011) (Supplemental Fig. S1B). In summary, our combined annotation had a total of 20743 pc-gene and 9068 nc-gene loci (including 2659 RefSeq).

Using this approach we classified allelic expression of pc- and nc-genes in the above 23 tissues as showing biallelic expression (BAE), not informative due to no or low expression (NI), not informative due to no SNPs (NS), strain-biased for CAST or for FVB, imprinted maternal expression (MAT) or

paternal expression (PAT). The total number of BAE pc-genes showed limited variation between tissues and developmental stages, varying 1.4 fold between 7,979-11,574 genes from the 19,772 annotated autosomal pc-genes, or 40-59% of the total (Fig. 1B, first row). The number of non-informative pc-genes showed a reciprocal pattern for each tissue, varying between 7669-11467 genes (39-58.0% of the total), while 723 genes (3.7%) could not be assessed due to a lack of SNPs (Fig. 1B, second row). Low tissue-specific variation in the number of BAE pc-genes is partly explained by genes that showed biallelic expression in multiple tissues, with 31% of biallelic genes showing biallelic expression in all 23 tissues (Supplemental Fig. S2B). In contrast, the number of pc-genes showing strain-biased and imprinted expression varied greatly between tissues. Genes showing strain-biased expression varied 4.7 fold among the different tissues from between 174-825 genes, or 0.9-4.2% of the total (Fig. 1B, third row). Overall strain bias towards the FVB allele was 1.9 fold higher than strain bias towards the CAST allele, which may reflect an alignment bias due to FVB having a shorter genetic distance to the C57BL/6 reference genome. Genes showing imprinted expression showed the most tissue-specific variation, varying 7.2 fold between the different tissues from 7-51 genes, or 0.035-0.258% of the total (Fig. 1B, fourth row).

The proportion of nc-genes classified BAE per tissue was much lower than for pc-genes, and showed greater tissue-specific variation, varying 3.7 fold from 262-970 genes or 3.0-11.1% of the total (Fig. 1C, first row). High variation is likely due to the known tight tissue-specific expression of lncRNAs that make up most of the nc-genes (Necsulea et al. 2014), as was further indicated by the high proportion of nc-genes that were non-informative in each tissue (87.3-95.5% of the total, Fig. 1C, second row). Reflecting this, in contrast to pc-genes only a small minority of nc-genes were biallelically expressed in all 23 tissues (50 of 2,673, 1.8%) and the majority showed biallelic expression in one tissue only (963 of 2,673, 36.9%) (Supplemental Fig. S2B). As for pc-genes, a low proportion of nc-genes could not be assessed due to a lack of SNPs (700 genes or 8.0% of the total). There was a similar high degree of variation in the number of nc-genes showing strain bias (4.5 fold, 68-310 transcripts, 0.77-3.5% of the total) or imprinted expression (6.5 fold, 4-26 transcripts, 0.045-0.29% of the total) as was seen for pc-genes (Fig. 1C, third and fourth rows).

In summary, mapping the Allelome revealed tissue-specific variation in the number of strain-biased and imprinted genes for both pc- and nc-genes, while the number of BAE genes was similar between tissues for pc-, but not nc-genes. Interestingly, the number of pc- and nc-genes in each allelic expression category appeared to co-vary between tissues, with the total number of pc- and nc-informative genes showing a high correlation ($r^2 = 0.77$, $p<10^{-4}$, Pearson).

Andergassen D et. al
Fig. 1 Defining the mouse Allelome

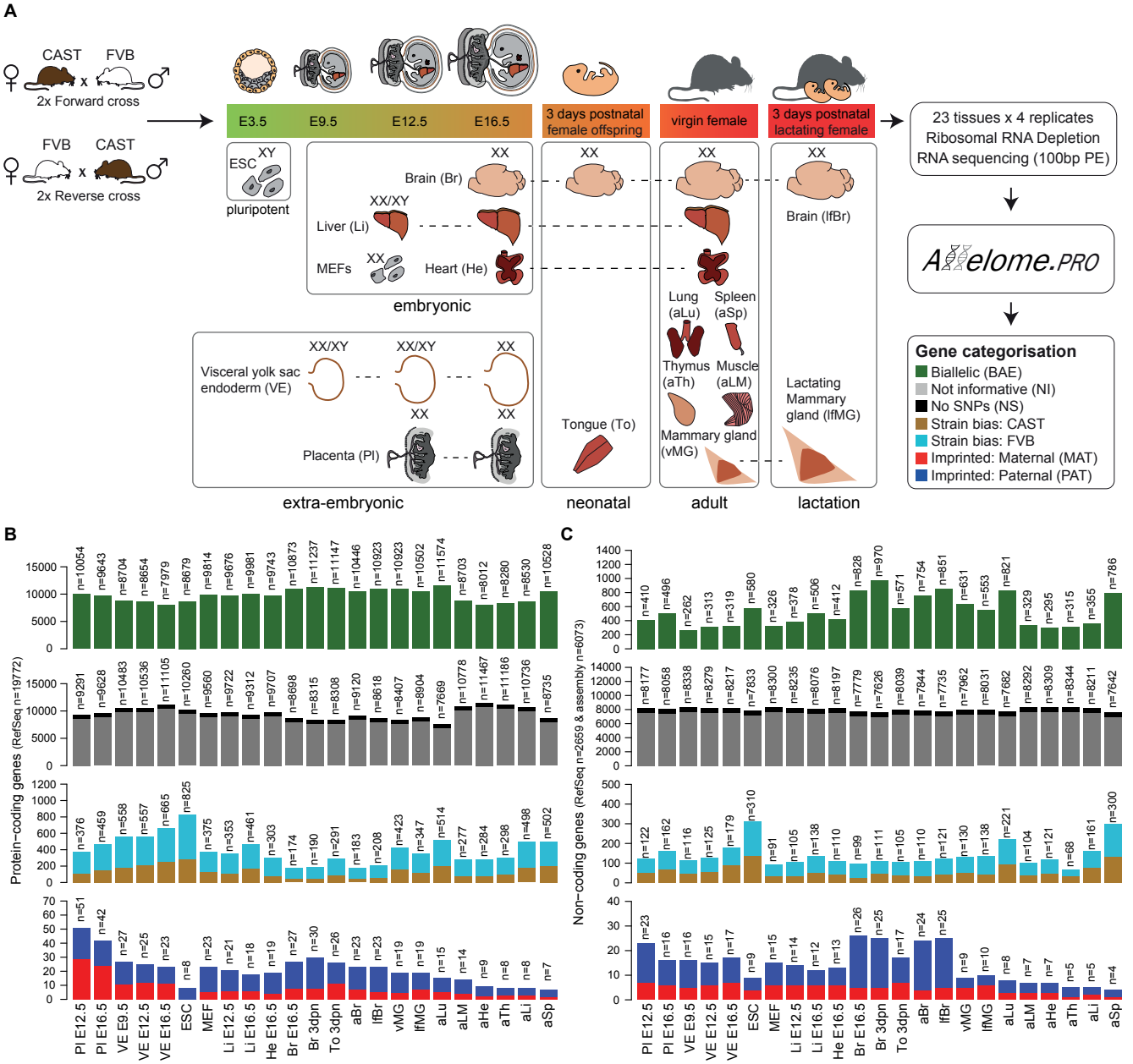


Figure 1. Defining the mouse Allelome.

(A) Strategy for detecting allelic expression from RNA-seq data from 23 mouse tissues and developmental stages using Allelome.PRO. The sex of the tissues is indicated by XX (female) and XY (male). Individuals were used except for indicated embryonic tissues where an entire litter was pooled (XX/XY).

(B) Allelome.PRO classification of the allelic expression status of protein-coding genes in each tissue.

(C) Allelome.PRO classification of non-coding genes.

Tissues examined were placenta (Pl embryonic day (E) 12.5, E16.5), visceral yolk sac endoderm (VE E9.5, E12.5, E16.5), embryonic stem cells (ESC), mouse embryonic fibroblasts (MEF E12.5), embryonic liver (Li E12.5, E16.5), embryonic heart (He E16.5), embryonic and neonatal brain (Br E16.5, 3 days postnatal (dpn)), neonatal tongue (To 3dpn), adult brain (aBr), adult lactating female brain (lfBr), adult virgin mammary glands (vMG), adult lactating female mammary glands (lfMG), adult lung (aLu), adult leg muscle (aLM), adult heart (aHe), adult thymus (aTh), adult liver (aLi) and adult spleen (aSp). Embryo and placenta diagrams adapted from (Hudson et al. 2011). Allelome.PRO settings: FDR 1%, allelic-ratio cutoff 0.7, minread 2.

Tissue-specific strain-biased expression correlates with strain-biased enhancer marks

The variation in the absolute number of strain-biased pc- and nc-genes was also reflected in the proportion of strain-biased genes relative to the number of informative genes per tissue (Fig. 2A). The proportion of pc-genes showing strain-biased expression (1.6% (embryonic brain) - 8.7% (ESC)) was generally lower than for nc-genes (10.0% (neonatal brain) - 34.8% (E16.5 VE)). This may reflect the known high evolution rate of lncRNAs that may lead to strain specific lncRNAs (Necsulea et al. 2014), and is in line with recent findings that lncRNAs vary significantly more than pc-genes between people (Kornienko et al. 2016). However, we found that the number of pc- and nc- strain-biased genes detected per tissue was correlated (r^2=0.71, $p<10^{-3}$, Pearson), indicating that some may be co-regulated.

We next investigated if the allelic status of genes is constant between tissues. We found that most biallelic genes remained biallelic wherever they are expressed for both pc- and nc-genes (Fig. 2B left), although the majority of nc-genes were expressed only in one tissue whereas most pc- genes are expressed in multiple tissues (Supplemental Fig. S2B). Most strain-biased nc-genes did not change their allelic status between tissues, whereas pc-genes could be categorized into two groups based on whether they maintained their allelic status between tissues or not (Fig. 2B right). The first group (134 CAST and 249 FVB) maintained strain-biased expression in 95-100% of the tissues, whereas the second group containing the majority of strain-biased genes (433 CAST and 569 FVB) maintained their allelic status in only 10% of the tissues.

We next sought to determine if tissue-specific strain-biased expression may be explained by the activity of strain-biased enhancers. To investigate this we used Allelome.PRO to detect allelic enrichment of H3K27ac chromatin immunoprecipitation and sequencing (ChIP-seq) data from FVB x CAST reciprocal crosses for fetal liver and VE, and compared this to the RNA-seq analysis for these tissues. We chose genes that switched from strain-biased in one tissue to biallelic in the other, and then examined H3K27ac enrichment ± 50kb from the transcription start site (TSS) (detailed in Methods). For both CAST and FVB strain-biased genes we found strain-biased H3K27ac enrichment both upstream and downstream of the TSS matching the strain-biased expression, while no enrichment was seen when the same genes were biallelic in the other tissue (Fig. 2C, Table S1C). This enrichment was explained by 39/144 (27%) strain-biased to BAE switchers (Table S1C), such as *Glrx* where a change from FVB biased expression in liver to BAE in VE was correlated with a putative switch in enhancer usage matching the allelic expression status (Fig. 2D). The other strain-biased switchers may be explained by enhancers outside of the ± 50kb TSS window or by tissue-specific post-transcriptional degradation, for example, due to tissue-specific expression of miRNAs. Altogether these results indicate that tissue-specific strain-biased expression can occur due to a switch

Andergassen D et. al
Fig. 2 The Allelome reveals tissue-specific expression of strain-biased genes

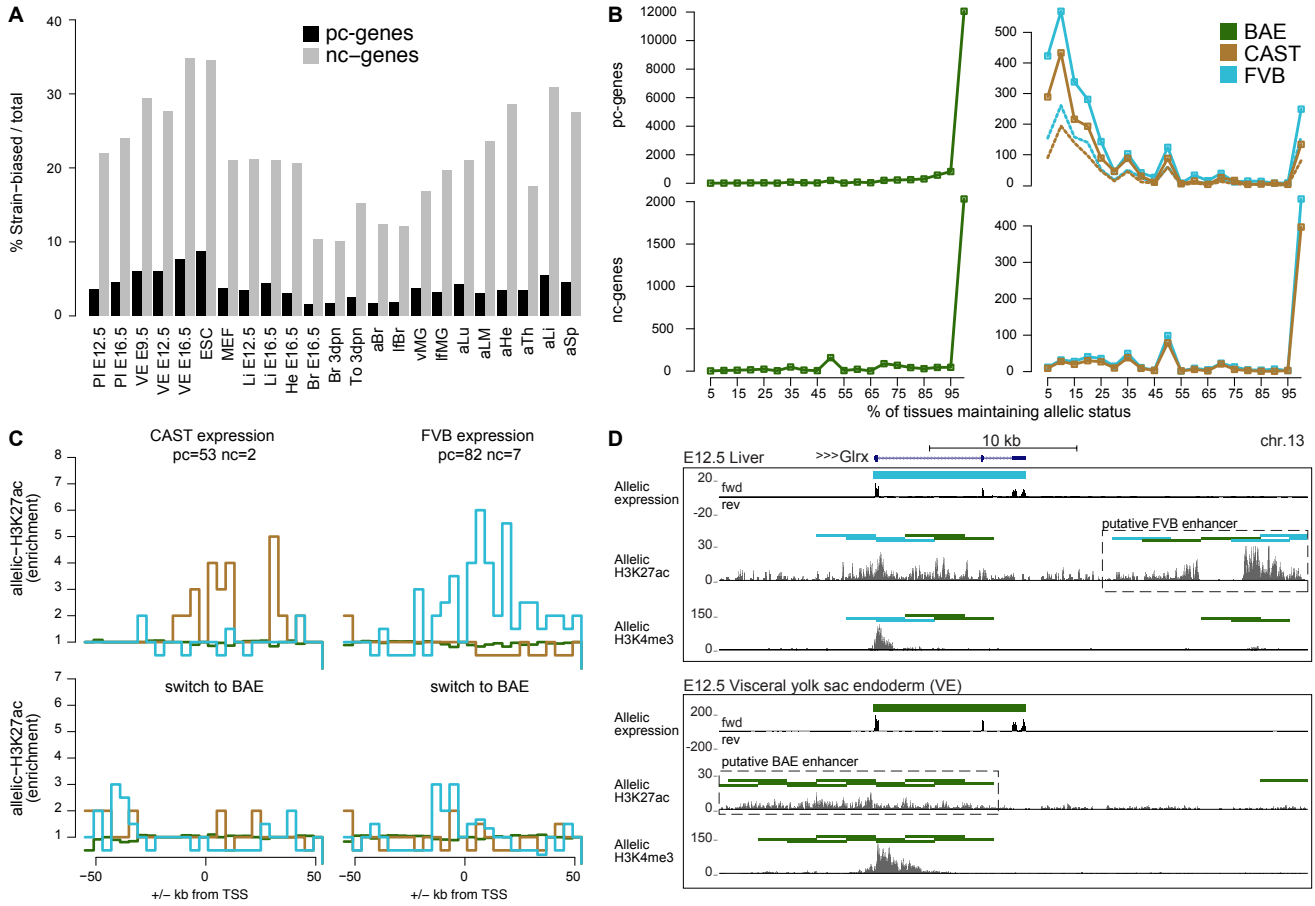


Figure 2. The Allelome reveals tissue-specific expression of strain-biased genes

(A) The percentage of strain-biased genes from total informative genes for each tissue for protein-coding (pc, black) and non-coding (nc, grey) genes.

(B) The percentage of tissues where pc- and nc- genes maintained their biallic (left) or strain-biased (right), allelic expression status (calculated relative to number of tissues where a gene was informative). Allelome.PRO settings: FDR 1%, allelic-ratio cutoff 0.7, minread 2, dotted lines indicates the outcome with a 0.8 allelic ratio cutoff.

(C) The enrichment of H3K27ac ±50kb from the transcription start site (TSS) of genes that show strain-biased expression in either E12.5 VE or Li, and biallelic expression in the other tissue. Top: H3K27ac enrichment near strain-biased genes. The enrichment over random of allelic H3K27ac 4Kb windows was calculated. Bottom: The same analysis for the same set of genes where they show biallelic expression. Analysis detailed in Supplemental Experimental Procedures.

(D) An example of putative enhancer switching: Glrx switches from FVB strain-biased expression in liver to BAE expression in VE. This is associated with a switch in putative enhancers that matches the allelic expression status.

in enhancer usage between tissues, from an enhancer that shows strain-biased activity to one that shows biallelic activity.

Escape from X-inactivation is tissue-specific and correlates with increased distance from monoallelic enhancers

We used 19 female tissues (Fig. 1A) to define the Allelome for the X-chromosome, 16 epiblast-derived embryonic, neonatal and adult tissues showing random X chromosome inactivation (XCI), and 3 extra-embryonic tissues showing imprinted XCI (Fig. 3A). In inbred mouse strains, both X-chromosomes in epiblast derived tissues have an equal chance to express the *Xist* lncRNA gene leading to random inactivation, however in CAST/FVB F1 mice the FVB allele is preferentially inactivated due to a bias in *Xist* allele expression (Chadwick et al. 2006; Calaway et al. 2013). In extra-embryonic tissues *Xist* is expressed only from the paternal allele, leading to inactivation of the paternal allele (Kay et al. 1994). Therefore, in this F1 cross XCI escapers can be detected as genes that do not show the expected CAST bias (epiblast-derived tissues) or MAT bias (extra-embryonic tissues) in expression, but rather show BAE or an unexpected bias (Note: to conservatively call BAE escapers a lower allelic ratio cutoff of 0.6 was used for this analysis, see Methods for details). Using this approach to define the mouse Escapome we detected 250 candidate escaper genes (225 random XCI and 43 imprinted XCI escapers, 18 escape in both) from 1044 X chromosome genes (792 pc- and 252 nc-genes), with a further 178 genes (14.6%) unable to be assessed due to a lack of SNPs (Fig. 3A, Table S1D, E). These included 31 out of 55 previously reported XCI escaper genes (Finn et al. 2014; Wu et al. 2014; Berletch et al. 2015). The high number of escapers detected could be due to the sensitivity of our method, although escapers were detected across a range of expression levels (Supplemental Fig. S3A). Expression of X chromosome genes varied between tissues from 165-352 pc-genes (14.4-28.8% of total) and 7-45 nc-genes (2.7-17.8% of total) (Fig. 3B top, middle). The number of genes escaping XCI varied considerably between tissues from 1-108 pc-genes (0.4-52.1% of informative genes), and from 1-10 for nc-genes including *Xist* (0.4-71.4%). Genes that escaped in a high proportion of tissues, or ubiquitous escapers, included a high number of known escapers such as *Kdm6a* and the nc-gene *Firre* (Fig. 3A), but the majority of escaper candidates showed tissue-specific escaping, for example *Dystrophin* (*Dmd*) in muscle (tongue and leg muscle, Fig. 3A). Ubiquitous and tissue-specific escaping was recently reported using a similar approach to define XCI escapers in brain, spleen, and ovary (Berletch et al. 2015). In our study we found 123 genes escaping in more than one tissue and 127 escaping in a single tissue, with *Xist* the only gene that escaped in all tissues.

We detected the highest number of escapers in leg muscle, with 118 of 221 informative pc- and nc-genes escaping (53.3%), while more than 50 escapers were also detected for tongue and lung (Fig. 3B). These leg muscle escapers showed a significant increase in expression compared to non-escapers, with a median 1.7 fold increase close to the expected doubling of expression (Fig. 3C). Interestingly

Andergassen D et. al

Fig. 3 X chromosome inactivation (XCI) escapers appear to be highly tissue-specific

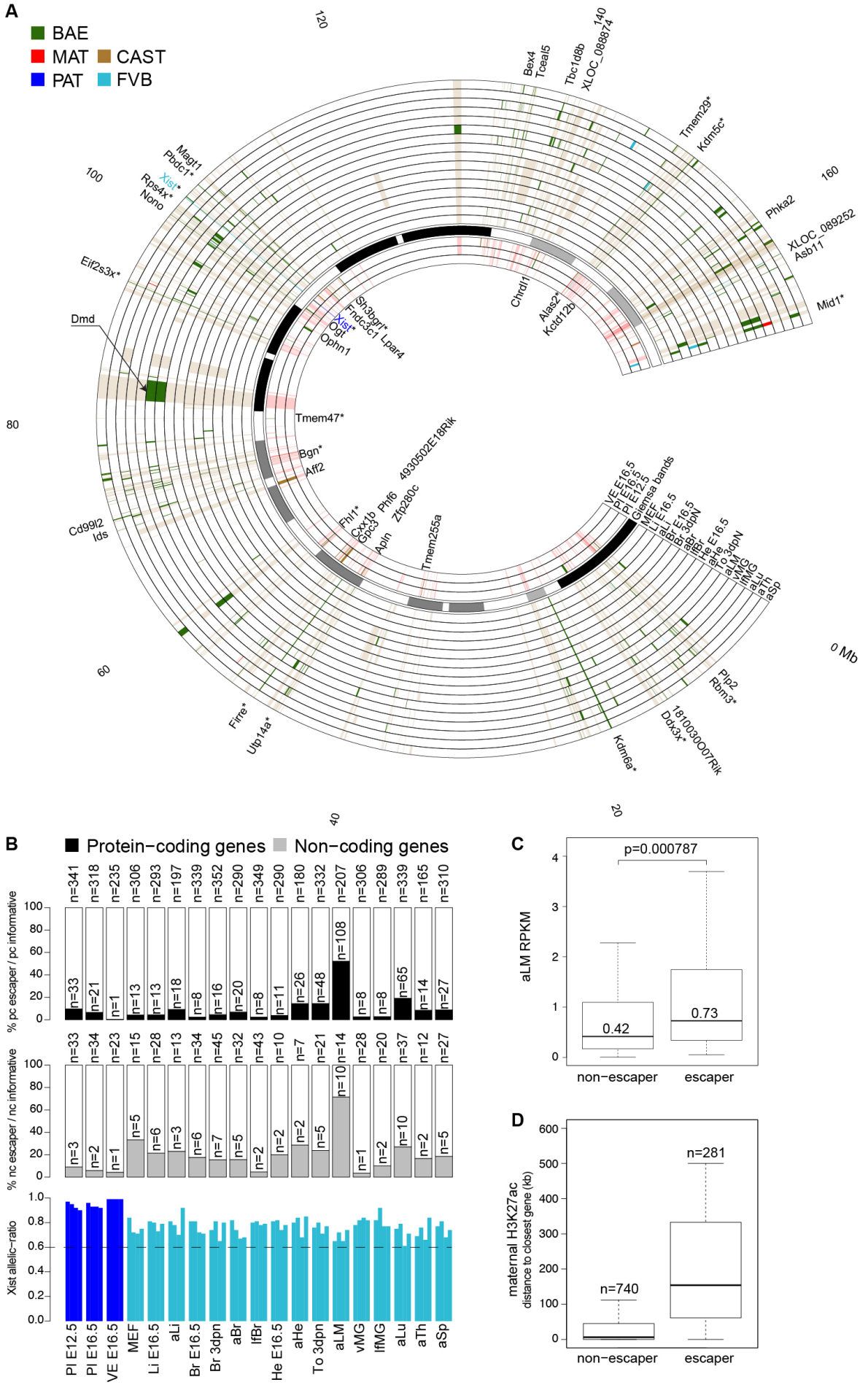


Figure 3. X chromosome inactivation (XCI) escapers appear to be highly tissue-specific

(A) Circos plot showing the mouse chromosome X Allelome for 19 female tissues (Fig. 1). Outer layers: 16 embryonic, neonatal and adult tissues showing random XCI (FVB X preferentially inactivated in CAST/FVB F1 tissues (skewed XCI)). Middle layer: Giemsa banding (UCSC genome browser). Inside layers: 3 extra-embryonic tissues showing imprinted XCI (paternal X chromosome inactivated). The top 25/225 escapers from random XCI are indicated on the outside of the Circos plot, while the top 20/43 escapers from imprinted XCI are indicated on the inside (escapers ranked by number of tissues). An asterisk marks known escapers. Dystrophin (Dmd) escapes specifically in muscle (aLM and To 3dpn, indicated by arrow). Color code as in Fig. 1. Allelome.PRO settings: FDR 1%, allelic-ratio cutoff 0.6, minread 2.

(B) Top: the percentage of pc-genes (black) escaping XCI from all informative pc for each female tissue. The number of informative pc-genes is given above the barplot, while the number of pc escapers is given above each bar.
Middle: the same analysis performed for nc-genes (grey).
Bottom: the allelic ratio of Xist for each replicate in extra-embryonic tissues (Xist expressed paternally (blue)) and in non extra-embryonic tissues (Xist preferentially expressed from the FVB allele (turquoise)).

(C) The expression level of escaper and non-escaper genes (pc and nc-genes) on the X chromosome for adult leg muscle. A boxplot including median values is shown (outliers not shown).

(D) The distance of parental-specific H3K27ac window to the closest non-escaper (331) and escaper (36) gene in placenta E12.5. Maternal H3K27ac windows with a distance higher than 500Kb were not included in the analysis. A boxplot including median values is shown (outliers not shown). After correcting for sample size, a significant difference was observed between escapers and non-escapers (Fisher's exact test, $p<1x10-20$, details in Supplemental Methods).

33 of 53 escapers in the tongue (62%), another muscular tissue, overlapped with escapers found in leg muscle. To investigate if the number of escapers could be explained by different degrees of XCI in different tissues, we examined the *Xist* allelic-ratio (Fig. 3B, bottom) and expression levels (Supplemental Fig. S3B). However, this indicated that all tissues showed the expected imprinted or skewed XCI, and there was no significant correlation between the number of escapers and the *Xist* allelic ratio or *Xist* expression levels.

To determine if differences in allelic H3K27ac enrichment may explain escaper status, we compared the distance to H3K27ac maternal enrichment 4kb windows in E12.5 placenta, a tissue that shows imprinted XCI of the paternal allele (Fig. 3D). We found that escapers tended to be further away from the nearest maternal H3K27ac window than non-escapers ($p<10^{-20}$, analysis described in the Methods). This is in agreement with previous reports that *Xist* causes silencing by targeting deposition of repressive chromatin to regulatory elements that then remain marked by H3K27ac on the active allele (Calabrese et al. 2012). Given that only *Xist* escapes in all tissues, together our data indicate that tissue-specific escape from XCI may be due to these elements being targeted by *Xist* in a tissue-specific manner.

Imprinted expression shows tissue-specific regulation

Tissue-specific imprinted expression indicates tissue-specific regulation and that there may be a tissue-specific function for imprinted expression (Prickett and Oakey 2012; Babak et al. 2015). Therefore, in order to gain a comprehensive picture of tissue-specific imprinted expression, we used Allelome.PRO to map the mouse Imprintome in our 23 tissues and developmental stages. We previously showed that our total RNA-seq approach combined with Allelome.PRO analysis robustly and sensitively detects imprinted expression in MEFs (Andergassen et al. 2015). The Harwell and Otago imprinting databases annotate a total of 126 RefSeq genes, 33 of which are disputed in the literature (downloaded 24th Sept 2015, (Glaser et al. 2006; Williamson CM 2013). In our analysis of RNA-seq data we found 71 of these known genes including 5 disputed genes (*Pon3, Peg3os, Cd81, Osbpl5,* and *Hymai*). Three of these genes disputed in placenta (*Pon3*, *Cd81* and *Osbpl5*) (Proudhon and Bourc'his 2010; Okae et al. 2012), we have previously confirmed to show imprinted expression in VE (Hudson et al. 2011; Kulinski et al. 2015). *Peg3os* and *Hymai* may be false positives due to overlap with known imprinted genes that show the same imprinted expression status, with *Peg3os* overlapped in anti-sense by *Peg3* (incomplete strand-specificity of RNA-seq) and *Hymai* overlapped in sense by *Plagl1*. The remaining 27 known genes that were not confirmed by RNA-seq fell into five categories: no SNP in the gene body (*Mcts2* (later confirmed by differential H3K4me3 promoter enrichment), *Nnat*), a imprinted bias detected below the 0.7 allelic ratio cutoff (*Adam23, Bcl2l1, Zfp64, Casd1, Copg2, Kcnq1, Cobl, Wars, Begain, Dio3, Htr2a*), only detected as biallelically expressed (*Mapt, Ccdc40*), non-informative in all tissues examined (low or no expression, *Htra3,*

Andergassen D et. al

Fig. 4 The Allelome reveals tissue-specific regulation of imprinted protein-coding genes

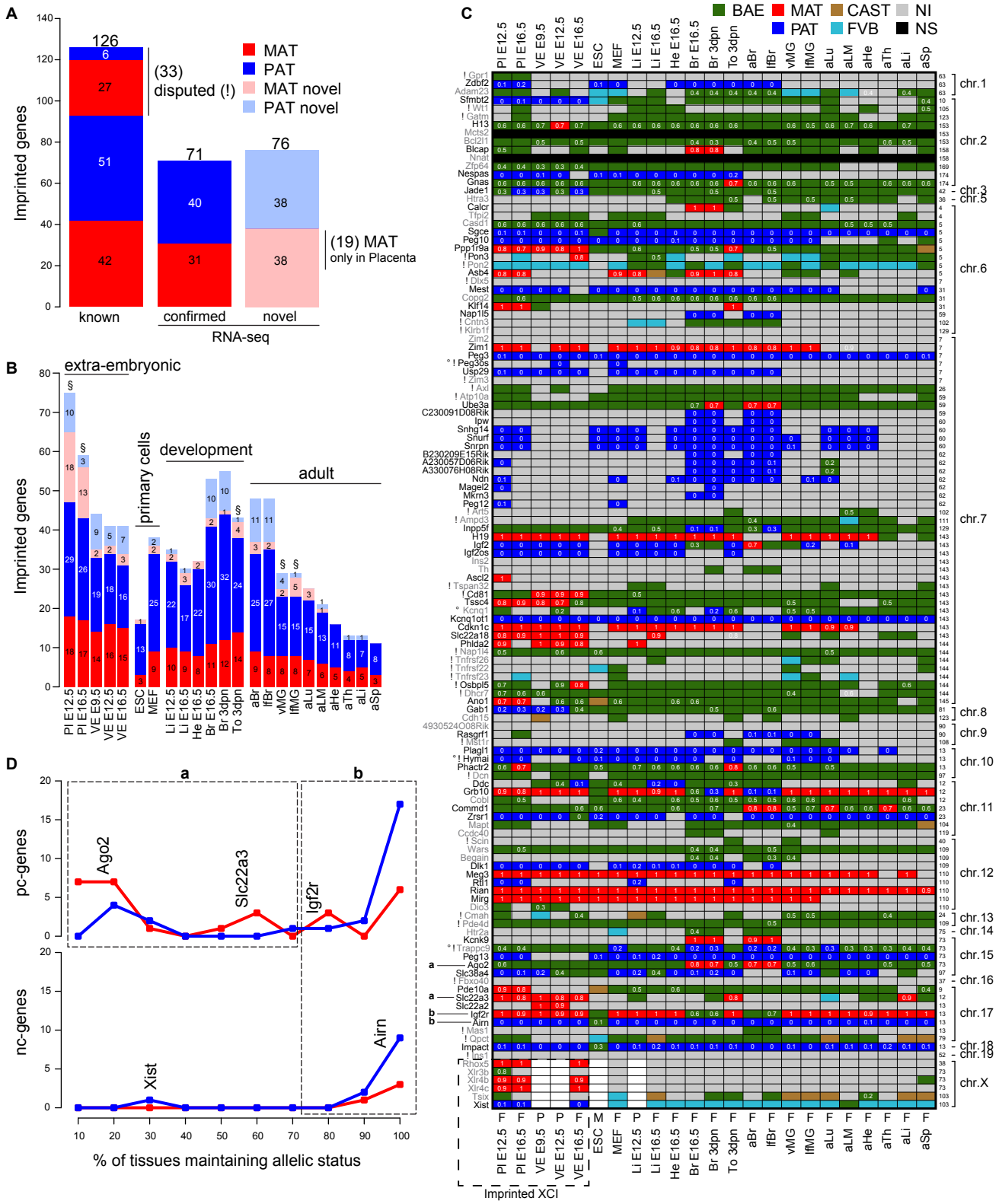


Figure 4. The Allelome reveals tissue-specific regulation of imprinted protein-coding genes

(A) The number of known, confirmed and novel imprinted genes detected in this study by RNA-seq. Known genes were RefSeq genes listed by the Otago or Harwell imprinted databases on 24th Sept 2015 (Glaser et al. 2006; Williamson CM 2013).

(B) The number of known and novel imprinted genes found among different tissues and developmental stages. Tissues important for the energy transfer from the mother to the offspring are indicated (§).

(C) The heatmap shows the allelic pattern for all 126 known imprinted genes among the different tissues. Gene names (left side) colored black are confirmed in this study, while gene names colored grey could not be confirmed. ! Indicates disputed genes. ° Probable RNA-seq strand bleed through. Examples given in D for variable ('a', in ≤70% expressing tissues) and consistent ('b' in >70% expressing tissues) imprinted expression are indicated. The chromosome number and the base pair coordinates (in Mb) for each gene are indicated on the right side. The imprinted allelic ratio for each gene and tissue (white) is given only if all 4 replicates show a bias in the same direction (1=100% expression from the maternal allele, 0=100% expression from the paternal allele). The sex for each tissue is indicated on the bottom of the heatmap: F (female, XX), M (male, XY), P (pooled XX/XY). Note: allelic analysis of the X chromosome can only be done for female tissues.
(D) The percentage of tissues that maintain imprinted expression of protein-coding genes (top) and nc-genes (bottom) (calculated as the number of tissues showing imprinted expression, divided by the number of informative tissues for each gene). The analysis was done for the 69 known imprinted genes confirmed by RNA-seq in this study (Peg3os and Hymai excluded due to probable RNA-seq bleed-through). Dotted boxes indicate genes that show variable ('a', in ≤70% expressing tissues) and consistent ('b' in >70% expressing tissues) imprinted expression. Examples are positioned according to the percentage of expressing tissues where they show imprinted expression.

Color key as in Fig. 1 except novel maternal and paternal imprinted candidates are pale red and blue respectively (Allelome.PRO settings: FDR 1%, allelic-ratio cutoff 0.7, minread 2).

Tfpi2, Zim2, Ins2, Th, 4930524O08Rik, Rhox5, Xlr3b, Xlr4b, Xlr4c), or the tissue reported to display imprinted expression was not assayed in this study (*Cdh15, Tsix*). In addition to known imprinted genes, this study identified 76 candidates, of which 38 were maternally expressed and 38 paternally expressed (Fig. 4A and S4A).

We examined our data for the distribution of imprinted expression between tissues and developmental stages (Fig. 4B). We found the highest number of known imprinted genes in placenta, brain and in neonatal tongue. In general extra-embryonic tissues and tissues collected during post-implantation development (embryonic and neonatal) expressed higher numbers of imprinted genes than pluripotent and adult tissues (with the exception of brain). Interestingly the number of imprinted genes detected tended to decrease within the same tissue during development (see placenta, brain, liver, heart). Tissues important for the energy transfer from the mother to the offspring, the placenta, neonatal tongue and mammary glands, showed a relatively high number of imprinted genes. However, we found a similar pattern in imprinted expression between brain and mammary glands collected from virgin and lactating females, indicating no obvious role for imprinted expression during lactation.

Tissue-specific imprinted expression could be directly explained by gene expression patterns, with a gene showing imprinted expression wherever it is expressed, or the allelic status of imprinted genes could switch between tissues. To investigate this we examined the imprinted status across tissues of all known imprinted genes confirmed by RNA-seq in our study (Fig. 4C, D). This analysis showed that imprinted pc-genes can be categorized into two groups based on the consistency of their allelic status in tissues where they are expressed. The first group ('a') showed variable imprinted expression (in ≤70% of expressing tissues, e.g. *Ago2* and *Slc22a3*), while a second group ('b') showed consistent imprinted expression (in >70% of expressing tissues, e.g. *Igf2r*). Imprinted nc-genes showed consistent imprinted expression where they were expressed (group 'b', e.g. *Airn*), with the exception of *Xist* whose imprinted expression was restricted to extra-embryonic tissues. In summary, we found a large proportion of known imprinted pc-genes (47%, group 'a') changed allelic status between tissues indicating tissue-specific regulation of imprinted expression, while in contrast imprinted nc-genes showed imprinted expression wherever they were expressed. This is in agreement with a recent study in human that found that most imprinted genes were tissue-specific and showed biallelic expression in another tissue (Baran et al. 2015).

Novel validated imprinted genes belong to known clusters

This study identified a relatively high number of 76 novel imprinted gene candidates that require further validation (Fig. 4A, B and S4A). We have previously shown that differential enrichment of H3K4me3 over promoters as detected by Allelome.PRO analysis of ChIP-seq data from F1 crosses can validate imprinted expression (Andergassen et al. 2015). Here we used 4kb sliding windows for

unbiased detection of differential H3K4me3 enrichment for selected tissues to validate novel imprinted genes (Table S1A, F-H). Using this approach we were able to validate the X-linked lncRNA *Gm35612* as a MAT imprinted gene in embryonic and adult tissues transcribed anti-sense to *Firre,* a lncRNA involved in regulating nuclear architecture (Hacisuleyman et al. 2014) (Fig. 5A). The previously reported MAT X-linked imprinted genes in brain *Xlr3b, Xlr4b,* and *Xlr4c* were classified non-informative in our data due to low SNP coverage (Davies et al. 2005; Raefski and O'Neill 2005). *Gm35612* was only detected as imprinted from RNA-seq data in adult brain where it was relatively highly expressed, while it was non-informative in all other tissues, likely due to difficulty in aligning reads in its repetitive gene body. However, the promoter of *Gm35612* contains a non-repetitive region, which showed maternal H3K4me3 enrichment in MEFs supporting imprinted expression of *Gm35612*. As the only imprinted gene on the X chromosome outside of extra-embryonic tissues that we validated, further investigation of this region for a connection to imprinted XCI may be warranted. Interestingly, although analysis of RNA-seq data from MEFs and other tissues indicated that *Firre* was biallelically expressed, as expected for a known XCI escaper, in MEFs we found a CAST biased H3K4me3 enrichment on its promoter, in line with the XCI bias in silencing the FVB allele, while the *Firre* gene body contained multiple H3K4m3 peaks enriched for a FVB bias (Fig. 5A). This indicates that the *Firre* locus may contain overlapping CAST and FVB biased transcripts, but is classified as biallelic when these transcripts are grouped together in the RefSeq annotation.

In addition to confirming the allelic expression status of imprinted genes, H3K4me3 enrichment can indicate the start site of independent transcripts, distinguishing novel imprinted gene candidates from 5´or 3´extensions of known imprinted genes (Supplemental Fig. S4A). In placenta and VE we found a novel paternally expressed candidate XLOC_032279 upstream of the known "solo" imprinted gene in placenta *Slc38a4* that was supported by paternal enrichment of H3K4me3 over its promoter (Fig. 5B). In E12.5 VE *Slc38a4* showed biallelic expression due to expression from an alternative downstream promoter, but paternal enrichment over the canonical promoter that is a maternally methylated gametic differentially methylated region (gDMR, red diamond Fig. 5B) that has not yet been validated as an ICE. This indicated that paternal expression of *Slc38a4* may be masked by a higher level of expression from the biallelic isoform, which was supported by a paternal bias in expression below the allelic ratio cutoff (Fig. 4C). Furthermore, paternal expression in E9.5 VE and biallelic expression without a bias in E16.5 VE indicated that *Slc38a4* switches from an imprinted to BAE isoform during VE development, while XLOC_032279 maintained imprinted expression at all stages (Fig. 4C and S4A). These results indicate that XLOC_032279 is an independent imprinted gene that may belong to an imprinted cluster together with *Slc38a4* that was previously thought to be a solo imprinted gene.

Maternal imprinted expression in placenta requires validation to distinguish it from expression in maternal decidua, blood and blood vessels that 'contaminate' the placenta. To get an indication if

Andergassen D et. al
Fig. 5 Novel validated imprinted genes belong to known clusters

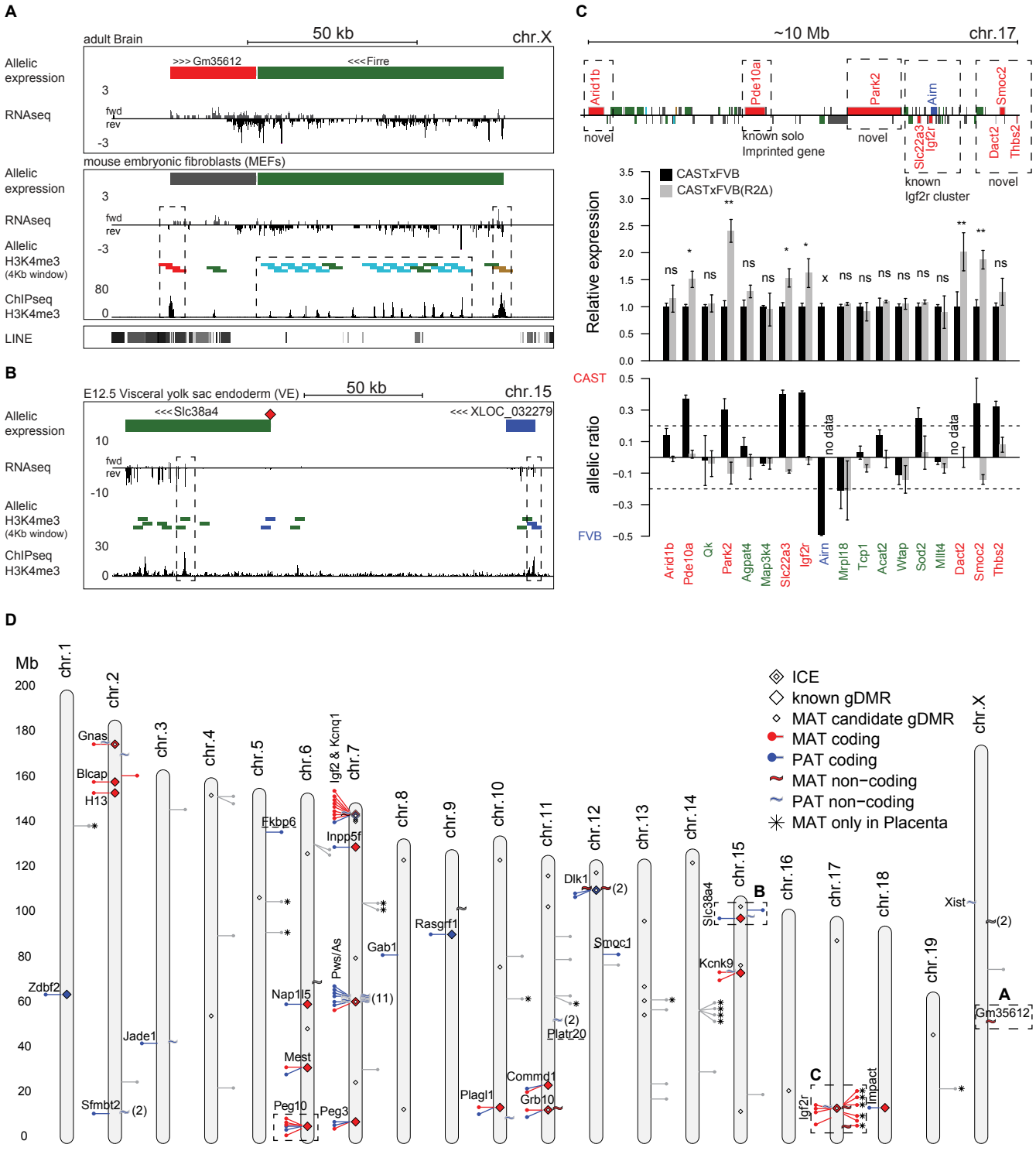


Figure 5. Novel validated imprinted genes belong to known clusters

(A) A novel X-linked imprinted nc-gene is transcribed anti-sense to Firre lncRNA. Maternal imprinted expression of Gm34612 detected from RNA-seq in adult brain (top) was validated by maternal H3K4me3 enrichment in MEFs (middle). The gene body of Gm34612 is enriched for LINE repetitive elements (bottom). (UCSC genome browser screenshot).

(B) Slc38a4 forms a cluster with a novel imprinted lncRNA. The Slc38a4 promoter is associated with maternal DNA methylation of the gametic differentially methylated region (gDMR, red square). In E12.5 VE, biallelic expression of Slc38a4 is associated with biallelic H3K4me3 enrichment over an alternative TSS. Paternal expression of the novel upstream imprinted lncRNA XLOC_032279 was validated by paternal H3K4me3 enrichment (UCSC genome browser screenshot).

(C) The Igf2r cluster is extended over 10Mb in placenta.
Top: Known and novel imprinted genes detected by Allelome.PRO analysis of RNA-seq from E12.5 placenta in 10Mb region surrounding the known Igf2r cluster (UCSC genome browser screenshot).
Middle: The relative expression (mean RPKM mutant/wildtype) between CASTxFVB and CASTxFVB(R2Δ) for imprinted genes and selected biallelic controls between Arid1b and Thbs2. * adjusted p-value <0.05, ns non-significant.
Bottom: The allelic ratio (median and standard deviation) for the same genes calculated using the Allelome.PRO pipeline (0.5 = 100% maternal and -0.5 = 100% paternal expression).

(D) A summary of imprinted genes detected in this study. Mouse chromosomes with the positions of known (left side of the chromosome) and novel supported or validated (right side of the chromosome) imprinted pc (–) and nc (~) genes. Candidate imprinted genes that are not supported or validated are indicated in grey. Imprint control elements (ICE), known and candidate gDMRs are indicated (Proudhon et al. 2012; Xie et al. 2012). * Indicates maternally expressed genes restricted to placenta. The base pair coordinates (Mb) are indicated on the left side. Color code as in Fig. 1. For more details see Supplemental Methods and Table S1F-H.

contaminating decidua material could explain apparent maternal imprinted expression of the 19 maternal candidates detected only in placenta (Supplemental Fig. S4A), we compared expression of these genes in placenta with expression in separated decidua as detected by RNA-seq, an approach that has been previously taken using RT-qPCR (Okae et al. 2012) (Supplemental Fig. S4B). We found that 11/19 candidates (58%) had a decidua/placenta ratio >5 indicating they could result from maternal contamination. Furthermore, 3 candidates with a low decidua/placenta ratio are genes expressed specifically in blood (*Ppbp*, *Hbb-bs*, *Hbb-b1*), indicating that they too could result from expression in maternal tissue. Interestingly, we noticed that 5 of the novel placenta candidates plus the known solo gene *Pde10a*, that shows imprinted expression only in placenta, were in close proximity to the known *Igf2r* cluster on chromosome 17. Therefore, we took advantage of the existing R2Δ mouse model that has a deletion of the *Igf2r* imprint control element (ICE) and *Airn* promoter (Wutz et al. 2001), to genetically test if these genes are part of the *Igf2r* imprinted cluster (Fig. 5C). We compared expression in CAST/FVB with CAST/R2D E12.5 placentas and found either a reduction in allelic ratio from maternal biased to biallelic and/or a significant increase in expression for all candidates near the *Igf2r* cluster (*Arid1b*, *Pde10a*, *Park2*, *Dact2*, *Smoc2*, *Thbs2*). This demonstrates that imprinted silencing of these genes is controlled by the *Airn* lncRNA, extending the size of the *Igf2r* imprinted cluster from a previously known size of 450kb to around 10Mb (7.7Mb upstream and 1.9Mb downstream of the ICE), and making it the largest *cis*-regulated autosomal region.

Following validation we classified the 76 novel imprinted candidates into 4 categories: (1) not confirmed, (2) candidate (3) supported candidate and (4) validated candidate, while novel lncRNAs upstream or downstream of a known imprinted lncRNA were considered fragments unless supported by H3K4me3 enrichment on their promoter and classified into the same 4 categories (Table S1F-H). Placental candidates were considered not confirmed if they had a decidua/placenta expression ratio >5 and no other evidence to support their imprinted status. 'Candidates' were detected by RNA-seq in one tissue or developmental stage only with no other supporting evidence. 'Supported candidates' were those that were found in multiple tissues or developmental stages (multiple tissues for placenta candidates), or genes that were located within 7Mb of a gDMR or known imprinted region found in our study (based on the maximum distance in the *Igf2r* cluster, extra evidence was required for placenta candidates). Finally, 'validated candidates' were supported by parental-specific H3K4me3 enrichment on their promoter or genetic validation (*Igf2r* cluster). This led to a total of 23 supported or validated candidates that were included with the 70 confirmed imprinted genes in subsequent analysis (Fig. 5D, Table S1H). Five of these novel imprinted genes (*Qk, Park2, Dact2*, *Fkbp6* and *Platr20*) were recently confirmed by others (Babak et al. 2015; Calabrese et al. 2015; Strogantsev et al. 2015). Although it is conceivable that some of the unsupported candidates that we found may be later confirmed, or that tissues we did not examine may have novel imprinted genes, given the extensive nature of our study, we do not expect that many more imprinted genes will be discovered.

Together 90.3% of the known and novel imprinted genes occurred in clusters, compared to 83.9% of the non-disputed imprinted genes prior to this study.

Tissue and developmental-specific expansion and contraction of imprinted clusters

By combining the detected imprinted expression in a comprehensive survey of mouse development we confirmed that imprinted genes are regulated in clusters. Tissue-specific imprinted expression indicated differences in regulation of imprinted expression, so we compared the size of each imprinted region in each tissue to determine if cluster size changed during development (Fig. 6A and S5A, Table S1I). Generally, we found that cluster size was at a minimum in pluripotent cells, then expanded in post-implantation and extra-embryonic tissues, before retracting to a minimal size in adult tissues (Supplemental Fig. S5A). Exceptions to this were the *Pws/As* and *Kcnk9* clusters where the cluster size in adult brain was equivalent to the maximum that was also found in embryonic and neonatal brain. Interestingly we observed that 19/28 imprinted regions (68%) showed the maximum size in extra-embryonic tissues (Table S1I). In particular we noticed that the *Kcnq1* and *Igf2r* imprinted clusters, where imprinted silencing is known to be controlled by an lncRNA, showed a dramatic cluster expansion in extra-embryonic tissues, particularly in placenta (Fig. 6A).

To investigate how this massive expansion may be regulated we examined allelic H3K27ac enrichment around imprinted clusters in embryonic liver, and the extra-embryonic VE and placenta (Fig. 6B). In the *Igf2r* cluster we found that maternal enrichment of H3K27ac correlated with cluster size, with no enrichment detected for embryonic liver, maternal enrichment windows detected within 2Mb of the ICE in VE, and up to 7Mb away in placenta (Fig. 6B, upper panel). Genome wide we found only a background level of parental-specific H3K27ac enrichment in embryonic liver, whereas VE and placenta showed a significant enrichment of parental-specific H3K27ac in imprinted regions (see Methods for details). Quantifying this for VE and placenta we found that the extent of H3K27ac expansion correlated with cluster size for these tissues (Fig. 6B, lower panel).

Altogether we found that all types of allele-specific expression that we examined were highly tissue-specific. Specifically we could also distinguish a clear developmental pattern in the numbers of genes showing imprinted expression, with imprinted clusters expanding during development, particularly in extra-embryonic tissues, and then contracting in the adult. For all types of allele-specific expression we found an association with nearby allele-specific H3K27ac enrichment, indicating that allele-specific expression due to both genetic and epigenetic causes may be mediated through enhancers.

Andergassen D et. al
Fig. 6 Tissue and developmental-specific expansion and contraction of imprinted clusters correlates with parental-specific histone modification

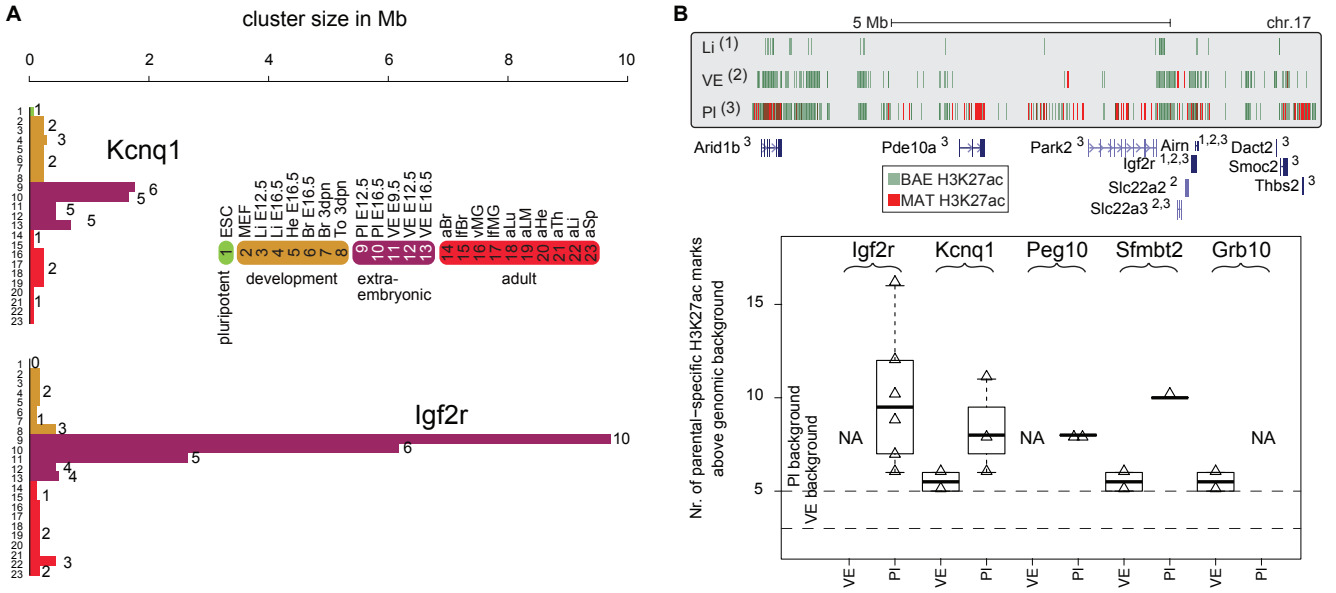


Figure 6. Tissue and developmental-specific expansion and contraction of imprinted clusters correlates with parental-specific histone modification

(A) The Igf2r and Kcnq1 cluster size during development and between tissues (tissue abbreviations as in Fig. 1). The number of imprinted genes for each developmental stage/tissue is indicated at the top of the bar.

(B) Top: Allelic H3K27ac enrichment (4kb sliding windows) over the expanded Igf2r cluster for E12.5 Liver, VE and placenta (UCSC genome browser screenshot). Numbers indicate tissue where a gene shows imprinted expression.
Bottom: The number of parental-specific H3K27ac 4kb sliding windows within non-overlapping 100kb count windows for E12.5 VE and placenta (Pl). Counts over the background cutoff are shown (defined as the maximum count detected outside of imprinted regions for each tissue). For more details see Supplemental Methods.

DISCUSSION

Biases in allelic expression in mammals due to genetic or epigenetic causes can have significant phenotypic consequences, but a comprehensive profile of this has been lacking. Here using our approach that classifies the allelic expression status of all genes in a tissue, we profiled allelic expression in 23 mouse tissues and developmental stages from RNA-seq data. This revealed that strain-biased expression, the extent of XCI and imprinted expression were highly tissue-specific. In particular, we show that imprinted gene cluster size varies between tissues and during development, and that they are at their maximum size in extra-embryonic tissues. Interestingly, we found that allelic expression was associated with differential enrichment of H3K27ac in adjacent regions.

Genetic polymorphisms can lead to expression biases in humans, but the outbred nature of the human population makes it difficult to assess the effect the same polymorphism has on allelic expression in different tissues. By using replicates of F1 tissues from crosses of inbred mouse strains we were able to assess the allelic expression of strain-biased genes in different tissues with the same genetic background. It could be that strain-biased expression is simply a reflection of tissue-specific expression, and that a bias is observed wherever a gene is expressed. However, we found that more often strain-biased genes showed a switch in allelic status between tissues, indicating that strain-biased expression most likely results from genetic differences in tissue-specific enhancers that control tissue-specific expression (Leung et al. 2015).

Xist expression leading to XCI, and the parental-allele specific DNA methylated ICEs that control imprinted expression, are present in almost every cell type during development, so it might be expected that if a gene subject to epigenetic silencing by these processes is expressed then it would always be silenced on one allele. However, this is not the case. We found that so-called ubiquitous XCI escapers that escape in many tissues were in the minority, with most XCI escapers escaping silencing only in 1 or 2 tissues. Similarly, a high proportion of imprinted genes showed tissue-specific imprinted expression where they switched to BAE in another tissue. These results showed that biases in allelic expression are generally tissue-specific, whether they arise from genetic or epigenetic causes, indicating that tissue-specific features are responsible for switches in allelic status.

Related to the variation tissue-specific imprinted expression, we found that the size of imprinted clusters varied during development, showing a minimum size in pluripotent ESC and a maximum in extra-embryonic tissues. These results have interesting parallels with XCI, with ESC showing 2 active alleles prior to the onset of random XCI, while extra-embryonic tissues are the only post-implantation tissues to show imprinted XCI (Wutz 2011). Specifically we showed that the *Igf2r* imprinted cluster is much larger than previously thought extending over 10Mb in placenta, or around 10% of mouse chromosome 17. The scale of the region effected by imprinted silenced by *Airn* is reminiscent of XCI

by *Xist*. Early in XCI *Xist* recruits PRC1 and PRC2 (Wutz 2011), repressive histone modifying complexes that have also been associated with *Airn* (Terranova et al. 2008), further indicating that they may act by a similar mechanism to cause silencing of distant genes.

The H3K27ac histone modification marks open chromatin and has been associated with active enhancers (Creyghton et al. 2010). Following this we found that a switch from strain-biased to BAE between tissues may be explained by tissue-specific enhancer usage associated with the corresponding H3K27ac enrichment. We also found an association between allelic H3K27ac enrichment and genes subject to XCI and imprinted silencing. *Xist* is reported to target H3K27me3 deposition to regions that remain marked by H3K27ac on the active allele (Calabrese et al. 2012). Consistent with this we found that the distance to allele-specific enrichment of H3K27ac was greater for XCI escapers than for genes subject to XCI. Enhancers explain tissue-specific expression, so it follows that tissue-specific silencing seen for XCI and imprinted silencing may be explained by actions on tissue-specific enhancers. Following this we found that the size of an imprinted cluster in a particular tissue correlated with size of the region showing parental-allele specific H3K27ac enrichment. Together these results indicate that all types of allele-specific expression that we observed may be mediated by allele-specific actions on enhancers.

METHODS

Mouse strains

Mice were bred and housed according to Austrian regulations under Laboratory Animal Facility Permit MA58-0375/2007/4. FVB/NJ (FVB) mice were purchased from Charles River and CAST/EiJ (CAST) from the Jackson Laboratory. The FVB.129P2-Airn<R2D> (R2Δ) mouse has a deletion that includes the *Airn* promoter and the imprint control element (ICE) of the *Igf2r* imprinted cluster (Wutz et al. 2001). F1 tissues were collected in replicates and frozen and stored at -80°C until further processed. For further details see Supplemental Methods and Table S1A.

RNA and ChIP-seq

RNA was extracted from TRI-reagent using standard protocols (Sigma-Aldrich T9424). DNaseI treated (DNA-Free™ Ambion) total RNA (1-3 μg) was depleted for Ribosomal RNA using the RiboZero rRNA removal kit (Human/Mouse/Rat, Epicentre) or enriched for polyA containing mRNA (Illumina). Strand-specific RNA-seq libraries were generated using the TruSeq RNA Sample Prep Kit v2 (Illumina) modified as previously described (Sultan et al. 2012). Native ChIP for H3K4me3, H3K27ac, and H3K27me3 was performed as previously described (Regha et al. 2007). The TruSeq ChIP Sample Prep Kit (Illumina) was then used to prepare ChIP-seq libraries. RNA-seq and ChIP-seq was then performed on a Illumina HiSeq. For further details see Table S1A.

Allelic RNA and ChIP-seq analysis

Allele-specific expression and histone modification enrichment was detected from RNA-seq and ChIP-seq data as previously described using the Allelome.PRO program (Andergassen et al. 2015). Further details on bioinformatics and statistical analysis is provided in the Methods.

DATA ACCESS

All sequencing data was deposited at the NCBI GEO data repository under accession numbers GSE75957 and GSE69168. Data can be viewed on the UCSC genome browser at …

[GEO reviewer access: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=qdybomkudtalzcx&acc=GSE75957]

[UCSC hubs: https://opendata.cemm.at/barlowlab/revision/]

ACKNOWLEDGEMENTS

This work was partly supported by the Austrian Science Fund (FWF P25185-B22, FWF F4302-B09, FWF W1207-B09). High-throughput sequencing was conducted by the Biomedical Sequencing

Facility (BSF) at CeMM in Vienna. Rita Casari and Philipp Guenzl helped with tissue collection. We thank Anton Wutz for critical reading of the manuscript.

REFERENCES

Andergassen D, Dotter CP, Kulinski TM, Guenzl PM, Bammer PC, Barlow DP, Pauler FM, Hudson QJ. 2015. Allelome.PRO, a pipeline to define allele-specific genomic features from high-throughput sequencing data. *Nucleic Acids Res* **43**(21): e146.

Babak T, Deveale B, Armour C, Raymond C, Cleary MA, van der Kooy D, Johnson JM, Lim LP. 2008. Global survey of genomic imprinting by transcriptome sequencing. *Current biology : CB* **18**(22): 1735-1741.

Babak T, DeVeale B, Tsang EK, Zhou Y, Li X, Smith KS, Kukurba KR, Zhang R, Li JB, van der Kooy D et al. 2015. Genetic conflict reflected in tissue-specific maps of genomic imprinting in human and mouse. *Nature genetics* **47**(5): 544-549.

Baran Y, Subramaniam M, Biton A, Tukiainen T, Tsang EK, Rivas MA, Pirinen M, Gutierrez-Arcelus M, Smith KS, Kukurba KR et al. 2015. The landscape of genomic imprinting across diverse adult human tissues. *Genome Res* **25**(7): 927-936.

Berletch JB, Ma W, Yang F, Shendure J, Noble WS, Disteche CM, Deng X. 2015. Escape from X inactivation varies in mouse tissues. *PLoS genetics* **11**(3): e1005079.

Berletch JB, Yang F, Xu J, Carrel L, Disteche CM. 2011. Genes that escape from X inactivation. *Human genetics* **130**(2): 237-245.

Calabrese JM, Starmer J, Schertzer MD, Yee D, Magnuson T. 2015. A survey of imprinted gene expression in mouse trophoblast stem cells. *G3* **5**(5): 751-759.

Calabrese JM, Sun W, Song L, Mugford JW, Williams L, Yee D, Starmer J, Mieczkowski P, Crawford GE, Magnuson T. 2012. Site-specific silencing of regulatory elements as a mechanism of X inactivation. *Cell* **151**(5): 951-963.

Calaway JD, Lenarcic AB, Didion JP, Wang JR, Searle JB, McMillan L, Valdar W, Pardo-Manuel de Villena F. 2013. Genetic architecture of skewed X inactivation in the laboratory mouse. *PLoS genetics* **9**(10): e1003853.

Chadwick LH, Pertz LM, Broman KW, Bartolomei MS, Willard HF. 2006. Genetic control of X chromosome inactivation in mice: definition of the Xce candidate interval. *Genetics* **173**(4): 2103-2110.

Chu C, Zhang QC, da Rocha ST, Flynn RA, Bharadwaj M, Calabrese JM, Magnuson T, Heard E, Chang HY. 2015. Systematic discovery of Xist RNA binding proteins. *Cell* **161**(2): 404-416.

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA et al. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America* **107**(50): 21931-21936.

Crowley JJ, Zhabotynsky V, Sun W, Huang S, Pakatci IK, Kim Y, Wang JR, Morgan AP, Calaway JD, Aylor DL et al. 2015. Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. *Nature genetics* **47**(4): 353-360.

Davies W, Isles A, Smith R, Karunadasa D, Burrmann D, Humby T, Ojarikre O, Biggin C, Skuse D, Burgoyne P et al. 2005. Xlr3b is a new imprinted candidate for X-linked parent-of-origin effects on cognitive function in mice. *Nature genetics* **37**(6): 625-629.

DeVeale B, van der Kooy D, Babak T. 2012. Critical evaluation of imprinted gene expression by RNA-Seq: a new perspective. *PLoS genetics* **8**(3): e1002600.

Engreitz JM, Pandya-Jones A, McDonel P, Shishkin A, Sirokman K, Surka C, Kadri S, Xing J, Goren A, Lander ES et al. 2013. The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science* **341**(6147): 1237973.

Finn EH, Smith CL, Rodriguez J, Sidow A, Baker JC. 2014. Maternal bias and escape from X chromosome imprinting in the midgestation mouse placenta. *Dev Biol* **390**(1): 80-92.

Glaser RL, Ramsay JP, Morison IM. 2006. The imprinted gene and parent-of-origin effect database now includes parental origin of de novo mutations. *Nucleic Acids Res* **34**(Database issue): D29-31.

Hacisuleyman E, Goff LA, Trapnell C, Williams A, Henao-Mejia J, Sun L, McClanahan P, Hendrickson DG, Sauvageau M, Kelley DR et al. 2014. Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. *Nature structural & molecular biology* **21**(2): 198-206.

Hudson QJ, Seidl CI, Kulinski TM, Huang R, Warczok KE, Bittner R, Bartolomei MS, Barlow DP. 2011. Extra-embryonic-specific imprinted expression is restricted to defined lineages in the post-implantation embryo. *Dev Biol* **353**(2): 420-431.

Kay GF, Barton SC, Surani MA, Rastan S. 1994. Imprinting and X chromosome counting mechanisms determine Xist expression in early mouse development. *Cell* **77**(5): 639-650.

Keane TM, Goodstadt L, Danecek P, White MA, 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**(7364): 289-294.

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

Kornienko AE, Dotter CP, Guenzl PM, Gisslinger H, Gisslinger B, Cleary C, Kralovics R, Pauler FM, Barlow DP. 2016. Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans. *Genome Biol* **17**(1): 14.

Kulinski TM, Casari MR, Guenzl PM, Wenzel D, Andergassen D, Hladik A, Datlinger P, Farlik M, Theussl HC, Penninger JM et al. 2015. Imprinted expression in cystic embryoid bodies shows an embryonic and not an extra-embryonic pattern. *Dev Biol* **402**(2): 291-305.

Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA, Gonzalez-Porta M, Kurbatova N, Griebel T, Ferreira PG et al. 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**(7468): 506-511.

Latos PA, Pauler FM, Koerner MV, Senergin HB, Hudson QJ, Stocsits RR, Allhoff W, Stricker SH, Klement RM, Warczok KE et al. 2012. Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. *Science* **338**(6113): 1469-1472.

Leung D, Jung I, Rajagopal N, Schmitt A, Selvaraj S, Lee AY, Yen CA, Lin S, Lin Y, Qiu Y et al. 2015. Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature* **518**(7539): 350-354.

Mager J, Montgomery ND, de Villena FP, Magnuson T. 2003. Genome imprinting regulated by the mouse Polycomb group protein Eed. *Nature genetics* **33**(4): 502-507.

Mancini-Dinardo D, Steele SJ, Levorse JM, Ingram RS, Tilghman SM. 2006. Elongation of the Kcnq1ot1 transcript is required for genomic imprinting of neighboring genes. *Genes & development* **20**(10): 1268-1282.

McHugh CA, Chen CK, Chow A, Surka CF, Tran C, McDonel P, Pandya-Jones A, Blanco M, Burghard C, Moradian A et al. 2015. The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature* **521**(7551): 232-236.

Monfort A, Di Minin G, Postlmayr A, Freimann R, Arieti F, Thore S, Wutz A. 2015. Identification of Spen as a Crucial Factor for Xist Function through Forward Genetic Screening in Haploid Embryonic Stem Cells. *Cell reports* **12**(4): 554-561.

Nagano T, Mitchell JA, Sanz LA, Pauler FM, Ferguson-Smith AC, Feil R, Fraser P. 2008. The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. *Science* **322**(5908): 1717-1720.

Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grutzner F, Kaessmann H. 2014. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**(7485): 635-640.

Okae H, Hiura H, Nishida Y, Funayama R, Tanaka S, Chiba H, Yaegashi N, Nakayama K, Sasaki H, Arima T. 2012. Re-investigation and RNA sequencing-based identification of genes with placenta-specific imprinted expression. *Human molecular genetics* **21**(3): 548-558.

Pauler FM, Barlow DP, Hudson QJ. 2012. Mechanisms of long range silencing by imprinted macro non-coding RNAs. *Curr Opin Genet Dev* **22**(3): 283-289.

Peters J. 2014. The role of genomic imprinting in biology and disease: an expanding view. *Nature reviews Genetics* **15**(8): 517-530.

Prickett AR, Oakey RJ. 2012. A survey of tissue-specific genomic imprinting in mammals. *Mol Genet Genomics* **287**(8): 621-630.

Proudhon C, Bourc'his D. 2010. Identification and resolution of artifacts in the interpretation of imprinted gene expression. *Brief Funct Genomics* **9**(5-6): 374-384.

Proudhon C, Duffie R, Ajjan S, Cowley M, Iranzo J, Carbajosa G, Saadeh H, Holland ML, Oakey RJ, Rakyan VK et al. 2012. Protection against de novo methylation is instrumental in maintaining parent-of-origin methylation inherited from the gametes. *Molecular cell* **47**(6): 909-920.

Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM et al. 2014. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res* **42**(Database issue): D756-763.

Raefski AS, O'Neill MJ. 2005. Identification of a cluster of X-linked imprinted genes in mice. *Nature genetics* **37**(6): 620-624.

Regha K, Sloane MA, Huang R, Pauler FM, Warczok KE, Melikant B, Radolf M, Martens JH, Schotta G, Jenuwein T et al. 2007. Active and repressive chromatin are interspersed without spreading in an imprinted gene cluster in the mammalian genome. *Molecular cell* **27**(3): 353-366.

Reinius B, Sandberg R. 2015. Random monoallelic expression of autosomal genes: stochastic transcription and allele-level regulation. *Nature reviews Genetics* **16**(11): 653-664.

Shlyueva D, Stampfel G, Stark A. 2014. Transcriptional enhancers: from properties to genome-wide predictions. *Nature reviews Genetics* **15**(4): 272-286.

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

Stringer JM, Pask AJ, Shaw G, Renfree MB. 2014. Post-natal imprinting: evidence from marsupials. *Heredity* **113**(2): 145-155.

Strogantsev R, Krueger F, Yamazawa K, Shi H, Gould P, Goldman-Roberts M, McEwen K, Sun B, Pedersen R, Ferguson-Smith AC. 2015. Allele-specific binding of ZFP57 in the epigenetic regulation of imprinted and non-imprinted monoallelic expression. *Genome Biol* **16**: 112.

Sultan M, Dokel S, Amstislavskiy V, Wuttig D, Sultmann H, Lehrach H, Yaspo ML. 2012. A simple strand-specific RNA-Seq library preparation protocol combining the Illumina TruSeq RNA and the dUTP methods. *Biochemical and biophysical research communications* **422**(4): 643-646.

Terranova R, Yokobayashi S, Stadler MB, Otte AP, van Lohuizen M, Orkin SH, Peters AH. 2008. Polycomb group proteins Ezh2 and Rnf2 direct genomic contraction and imprinted repression in early mouse embryos. *Dev Cell* **15**(5): 668-679.

Wang X, Soloway PD, Clark AG. 2011. A survey for novel imprinted genes in the mouse placenta by mRNA-seq. *Genetics* **189**(1): 109-122.

Wang X, Sun Q, McGrath SD, Mardis ER, Soloway PD, Clark AG. 2008. Transcriptome-wide identification of novel imprinted genes in neonatal mouse brain. *PLoS One* **3**(12): e3839.

Washietl S, Findeiss S, Muller SA, Kalkhof S, von Bergen M, Hofacker IL, Stadler PF, Goldman N. 2011. RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. *Rna* **17**(4): 578-594.

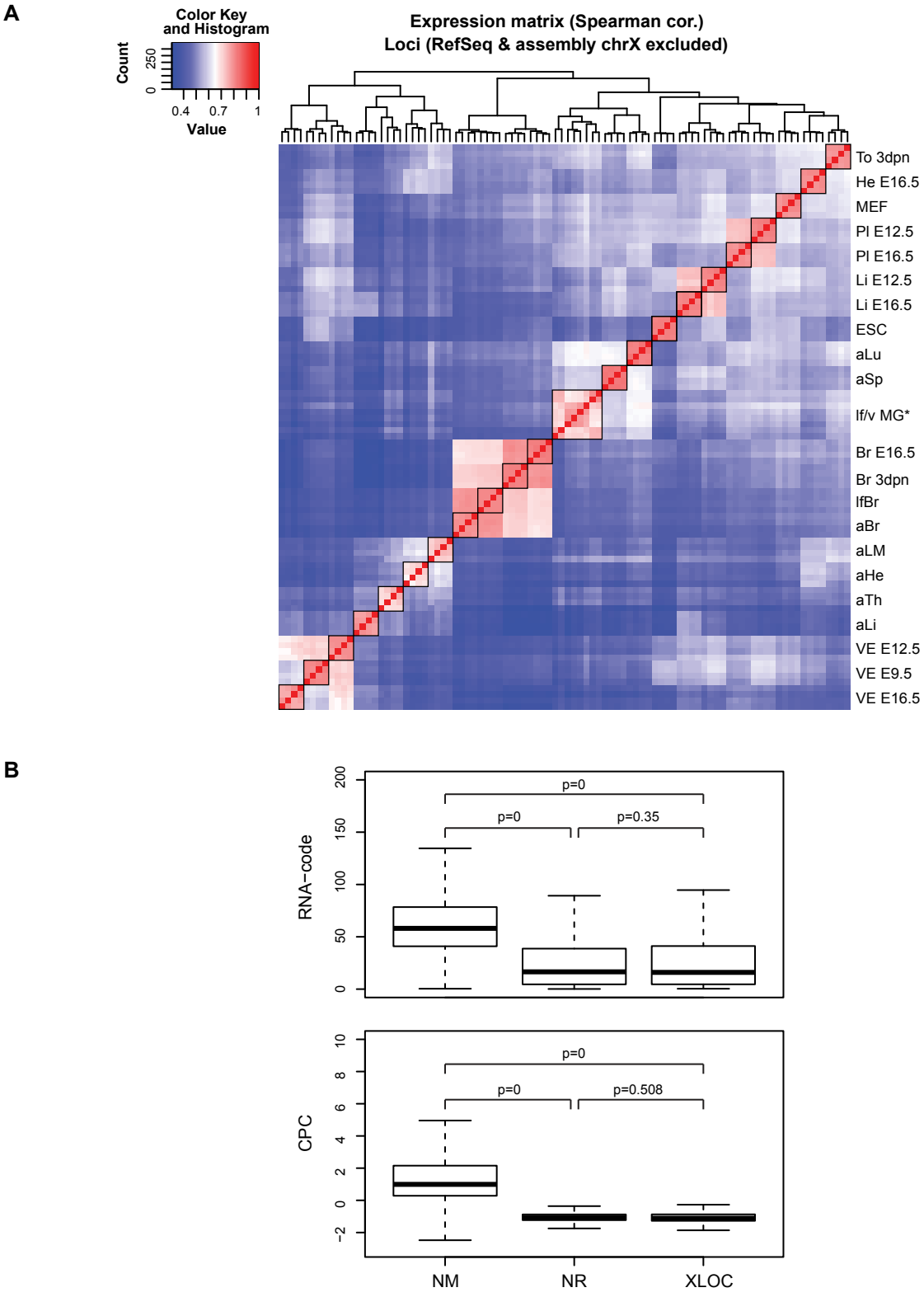
Williamson CM BA, Thomas S, Beechey CV, Hancock J, Cattanach BM, and Peters J. 2013. World Wide Web Site - Mouse Imprinting Data and References.

Wu H, Luo J, Yu H, Rattner A, Mo A, Wang Y, Smallwood PM, Erlanger B, Wheelan SJ, Nathans J. 2014. Cellular resolution maps of X chromosome inactivation: implications for neural development, function, and disease. *Neuron* **81**(1): 103-119.

Wutz A. 2011. Gene silencing in X-chromosome inactivation: advances in understanding facultative heterochromatin formation. *Nature reviews Genetics* **12**(8): 542-553.

Wutz A, Theussl HC, Dausman J, Jaenisch R, Barlow DP, Wagner EF. 2001. Non-imprinted Igf2r expression decreases growth and rescues the Tme mutation in mice. *Development* **128**(10): 1881-1887.

Xie W, Barr CL, Kim A, Yue F, Lee AY, Eubanks J, Dempster EL, Ren B. 2012. Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell* **148**(4): 816-831.



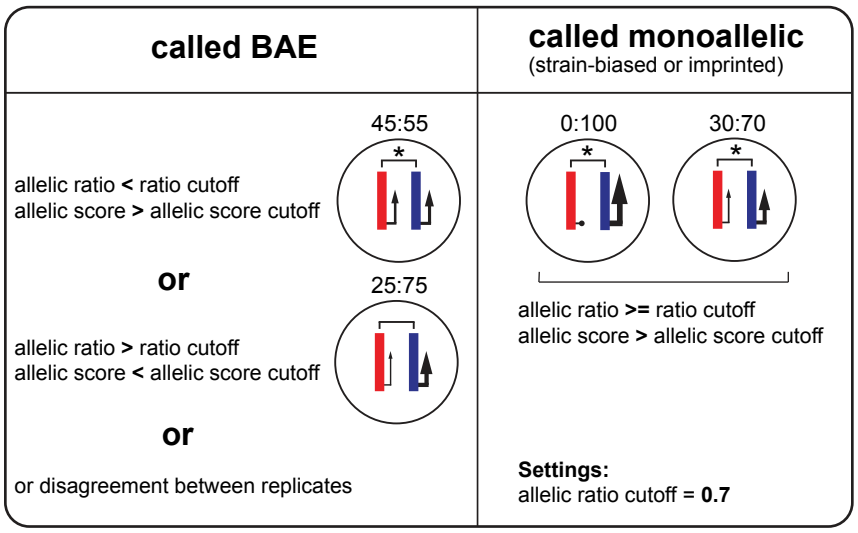
Supplemental Figure 1

(A) Cluster analysis confirms identity and similarity of sequenced tissues and replicates. The Heatmap shows unsupervised clustering of a Spearman correlation matrix from log-transformed gene expression data across 92 samples (23 tissues x 4 replicates). In an exception, clustering did not distinguish virgin and lactating mammary glands (*).

(B) Novel annotated loci outside of RefSeq are non-protein coding. The boxplots show the estimated RNA-code (top) (Washietl et al. 2011) and Coding Potential Calculator (CPC) scores (bottom) (Kong et al. 2007) for mRNAs (RefSeq NM), non-coding RNAs (RefSeq NR) and loci annotated in this study (XLOC). P-values were calculated using a t-test. For more details see Supplemental Methods.

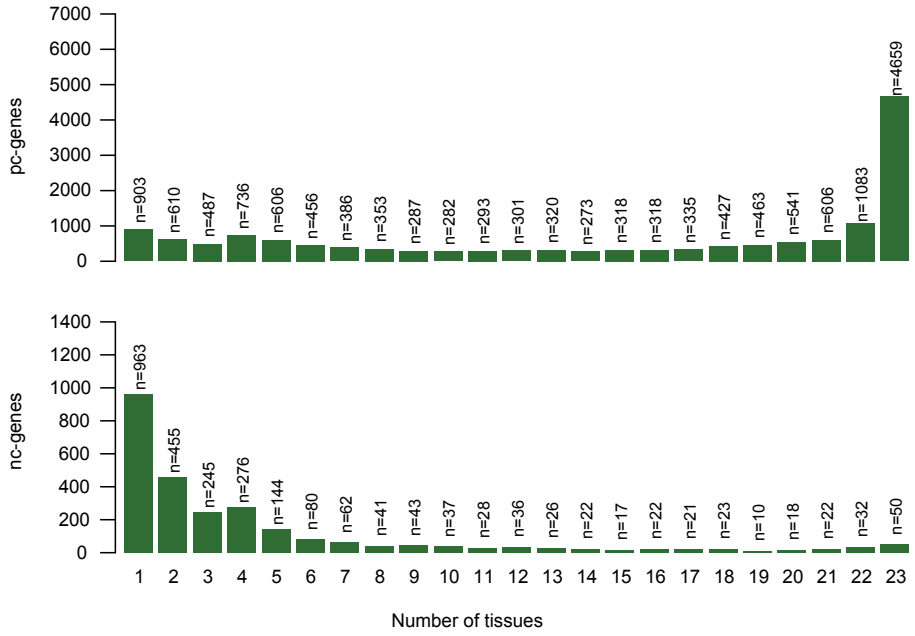
Andergassen D et. al
Supp. Fig. 2

A



Cutoffs based on the median allelic ratio and minimum allelic score from all replicates

B

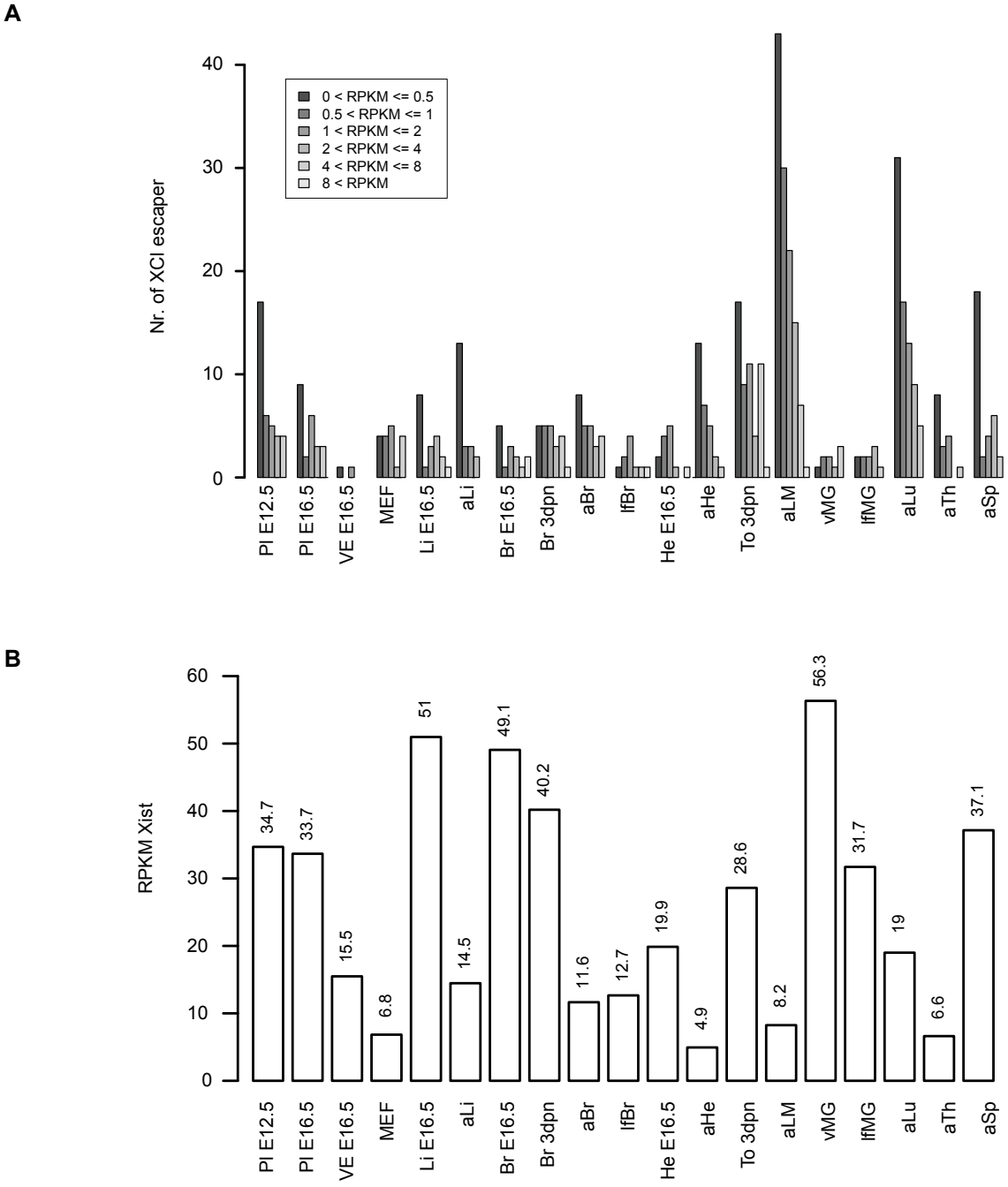


Supplemental Figure 2

(A) Allelome.PRO strategy to classify biallelic (BAE) and monoallelic (imprinted or strain biased) expression (FDR cutoff 1%, allelic ratio cutoff 0.7, minread 2, * passes FDR). For more details see the Supplemental Methods and the Allelome.PRO methods paper and manual (Andergassen et al. 2015).

(B) Most biallelic pc-genes (top) showed biallelic expression in the majority of tissues, while nc-genes (bottom) showed biallelic expression mainly in a single or few tissues.

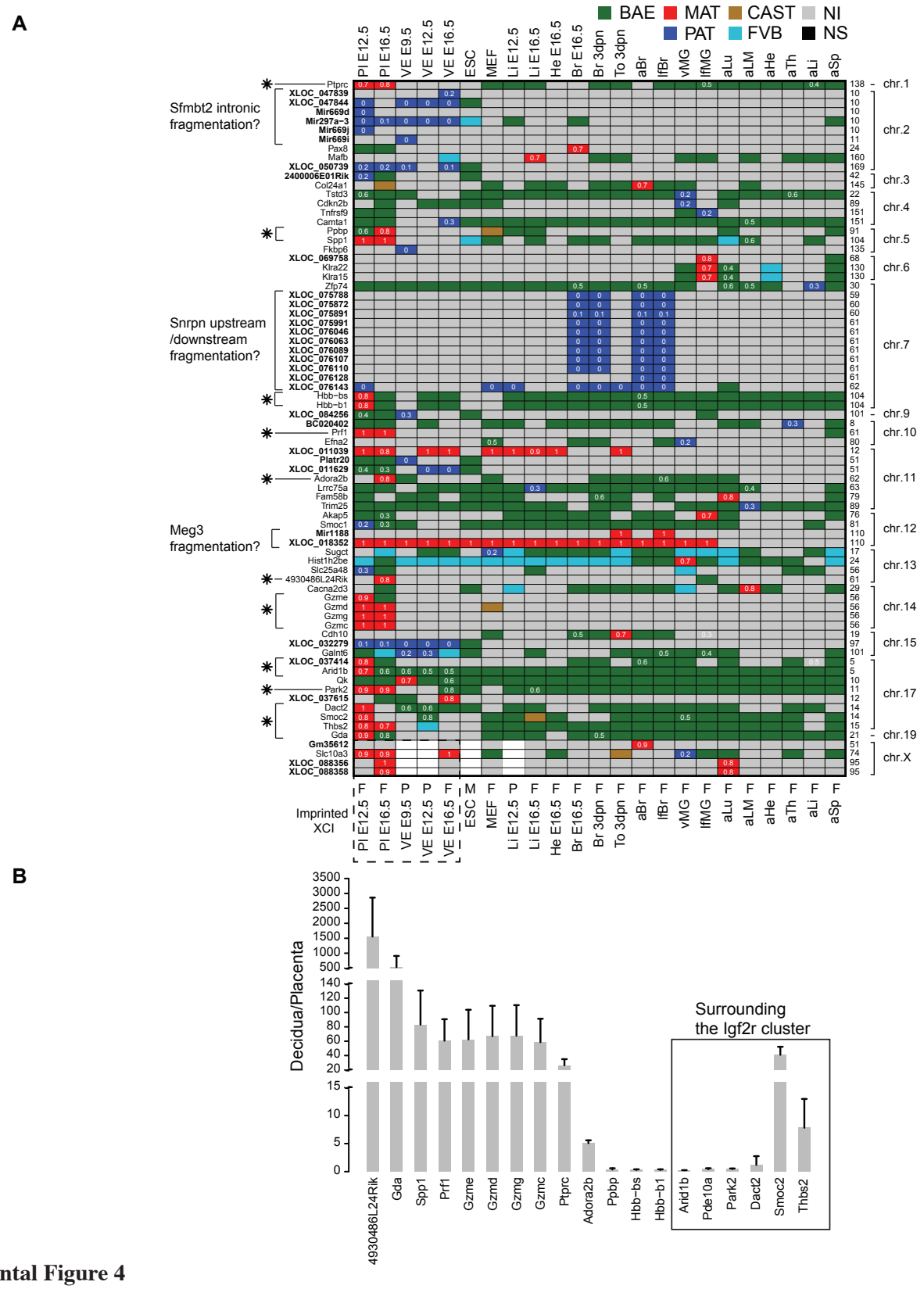
Andergassen D et. al
Supp. Fig. 3



Supplemental Figure 3

(A) XCI escapers show a wide range of expression levels. XCI escaper numbers for each female tissue and developmental stage are plotted for the indicated expression bins.

(B) Expression level of the lncRNA Xist among the different female tissues.

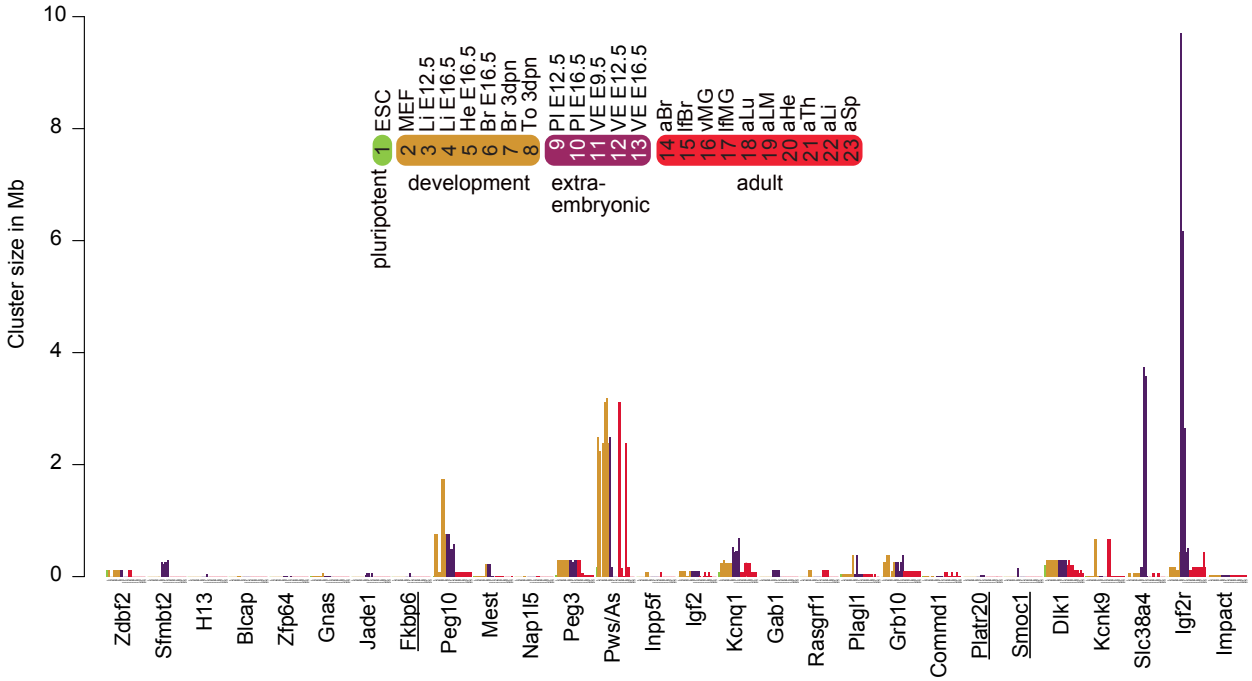


Supplemental Figure 4

(A) Novel imprinted genes identified in this study. The heatmap shows the allelic pattern among the different tissues for the 76 novel imprinted genes detected by RNA-seq. Bold gene names (left side) indicate non-coding genes. (*) Indicates maternally expressed genes restricted to placenta. The chromosome number and the base pair coordinates (in Mb) are indicated on the right side. The imprinted allelic ratio for each gene and tissue (white) is given only if the allelic direction agrees for all 4 replicates (1=100% expression from the maternal allele, 0=100% expression from the paternal allele). The sex for each tissue is indicated on the bottom of the heatmap: F (female), M (male), P (pooled). Note allelic analysis of the X chromosome can only be done in females.

(B) Maternal tissue in placental samples may result in false positive maternal imprinted expression. The mean decidua/placenta RPKM ratio (3x E12.5 decidua and placenta) for placental-only candidate maternal imprinted genes (18 RefSeq genes and the known solo imprinted gene Pde10a). The error bar represents the standard deviation between replicates.

Andergassen D et. al
Supp. Fig. 5



Supplemental Figure 5

Expansion and contraction of imprinted clusters during development. The cluster size in Mb of confirmed and novel validated/ supported imprinted regions was plotted. Underlined names indicate novel imprinted regions.

Supplemental Methods

Tissue isolation for RNA-seq

To determine the mouse expression Allelome from RNA-seq data we collected samples from 23 F1 mouse tissues and developmental stages (2x CASTxFVB and 2x FVBxCAST, maternal allele always on the left) representing pluripotent (1), embryonic (5), extra-embryonic (5), neonatal (2), adult (8) and lactating female (2) tissues (Fig. 1A, Table S1A). We collected 19 tissues samples from individual females (XX), while for embryonic day (E) 12.5 liver, and E9.5 and E12.5 visceral yolk sac endoderm (VE) we pooled males and females from 1 litter (XX/XY). Embryonic stem cells (ESCs) were derived from male (XY) clones. Sex was confirmed by PCR for individual extra-embryonic, embryonic and neonatal tissues samples (Capel et al. 1999).

Tissues were dissected and frozen immediately in liquid nitrogen, with the exception of ESCs, MEFs, VE and mammary glands that were processed differently before freezing and storing at -80°C. For ESCs and MEFs cells were centrifuged, washed in PBS and then frozen. Visceral yolk sac and mammary glands were processed as previously described to isolate VE and mammary epithelial cells (Joshi et al. 2010; Hudson et al. 2011). ESCs were derived following an established protocol (Bryja et al. 2006; Kulinski et al. 2015), and adapted to 2i media without feeders (ESGRO-2i Medium, Millipore) (Ying and Smith 2003; Ying et al. 2008). Note that the RNA-seq data from MEFs was described in an earlier study (Andergassen et al. 2015).

To determine genes subject to imprinted silencing by *Airn* long non-coding (lnc) RNA and exclude maternal contamination, we crossed CASTxR2Δ and collected placentas from 3 wild type (WT) and 3 mutant embryos from embryonic day (E) 12.5, as well as 3x CAST decidua from WT embryos.

Tissue isolation for ChIP-seq

For ChIP-seq experiments we collected material from FVBxCAST reciprocal crosses. Individuals were collected for adult liver, placenta, mammary glands and neonatal tongue, while samples were pooled for embryonic liver and VE (for details see Table S1A). Note that the ChIP-seq data from MEFs was described in an earlier study (Andergassen et al. 2015).

Preparation of the input files

The RNA and ChIP-seq data was aligned with STAR (Dobin et al. 2013) as previously described (Andergassen et al. 2015). In order to make a fair comparison between tissues we

equalized the number of uniquely aligned reads used for Allelome.PRO analysis of total RNA-seq from the 23 tissues, and H3K27ac ChIP-seq from E12.5 embryonic liver, VE and placenta. For each tissue we took the number of reads from the start of the unaligned FASTQ file required to obtain approximately 15 million uniquely aligned pairs (30 million reads) per sample for RNA-seq data, and 20 million uniquely aligned single end reads for H3K27ac ChIP-seq data. We then realigned these reads and performed all subsequent analysis on this data (Table S1B). For the analysis of the CAST/FVB placenta, CAST/R2Δ placenta and CAST decidua polyA RNA-seq samples we used approximately 18 million uniquely aligned single end reads per sample. For H3K4me3 and H3K27me3 ChIP-seq data all available reads were analyzed.

RNA-seq annotation file

To define allelic expression using the Allelome.PRO pipeline, we downloaded the RefSeq annotation from the UCSC genome browser (GRCm38/mm10) on July 15th 2015 and removed transcripts with a gene body length less than 100bp. To annotate the regions not covered by RefSeq we combined the reads from the 4 samples for each tissue using SAMtools (version 1.2) and used Cufflinks (version 2.2.1) to perform a reference based assembly. Next we used Cuffmerge to merge the assemblies from all tissuea together with the RefSeq annotation (-g RefSeq.gtf). Then we discarded transcripts overlapping RefSeq in sense orientation and single exon transcripts.

To predict whether the novel annotated transcripts are protein-coding or non-coding we used the Coding Potential Calculator (CPC) based on sequence features (Kong et al. 2007), modified as described previously (Kornienko et al. 2016), and RNA-code based on evolutionary signature (Washietl et al. 2011). We used the two pipelines for each transcript in the annotation (n=171389) and assigned the smallest CPC and RNA-code score to each locus. A t-test was performed between mRNAs (RefSeq NM), non-codings RNA (RefSeq NR) and genes annotated in this study (XLOC) for the CPC and RNA-code score. As expected, we observed a highly significant difference between RefSeq mRNAs (NM) and RefSeq non-coding RNAs (NR) for both the CPC and the RNA-code score (p=0). The same significant difference was observed between mRNAs (NM) and genes annotated in this study (XLOC) (p=0). However we observed no significant difference between RefSeq non-coding RNAs (NR) and genes annotated in this study (XLOC) showing that the bulk of novel annotated genes is non-coding. (CPC p=0.508, RNAcode p=0.35, Supplemental Fig. S2B).

The final annotation consists of 23521 RefSeq genes (20743 protein-coding (NM) and 2778 NR non-coding) and 6290 assembled non-coding genes (XLOC) outside of the RefSeq annotation.

ChIP-seq annotation files

Sliding windows were used to define allelic ChIP (4kb sliding windows for H3K27ac and H3K4me3 (2kb intervals) and 20kb sliding window for H3K27me3 (10kb intervals)).

SNP annotation files

The SNP annotation file containing 20,601,830 high confidence SNPs between the CAST/EiJ and FVB/NJ strains was extracted from the Sanger database as described previously (Keane et al. 2011; Andergassen et al. 2015). For RNA-seq, but not ChIP-seq, SNPs overlapping retroposed genes including pseudogenes (RetroGenes V6 from UCSC genome browser) were removed leaving 20,453,039 SNPs. For the CAST x FVB.129P2-Airn-R2D (R2Δ) cross we used only CAST/FVB SNPs where the FVB allele was shared with all 3 sequenced 129 strains (16,988,479 SNPs).

Allelome.PRO analysis of RNA and ChIP-seq data

Allele-specific expression and histone modification enrichment was detected from RNA-seq and ChIP-seq data as previously described using the Allelome.PRO program (Andergassen et al. 2015). Briefly, for each tissue a gene was classified as showing an allelic bias if the minimum allelic score (strain or imprinted score) of all biological replicates passed the FDR cutoff (defined by mock comparisons), and the median allelic ratio was above or equal to the allelic ratio cutoff. Informative genes that did not fulfill these criteria were classified as biallelic. A gene was classified as informative in a given tissue if a minimum SNP coverage was reached for all replicates. This was defined as the minimum SNP coverage required to pass the FDR cutoff assuming that the allelic ratio would be equal to the allelic ratio cutoff. For allelic analysis with an allelic ratio cutoff of 0.7 (all analysis except the Escapome analysis), the minimum SNP coverage required for a gene to be called informative varied between tissues from 11 to 13 reads, with a mean of 12 reads. For the Escapome analysis with an allelic ratio cutoff of 0.6, the minimum SNP coverage required for a gene to be called informative varied between tissues from 13 to 48 reads, with a mean of 27 reads.

For RNA-seq informative genes had a mean of 49 informative SNPs with a minread parameter of 2.

The Allelome.PRO settings used in this study:

Allelome RNA-seq: FDR 1%, allelic ratio cutoff 0.7, minread 2
Escapome RNA-seq: FDR 1%, allelic ratio cutoff 0.6, minread 2
CASTxR2Δ RNA-seq: FDR 1%, minread 1 (no allelic ratio cutoff)
Imprinted gene validation, H3K4me3 ChIP-seq: FDR 1%, allelic ratio cutoff 0.7, minread 2
H3K27ac ChIP-seq enrichment: FDR 1%, allelic ratio cutoff 0.7, minread 1

H3K27me3 ChIP-seq enrichment: FDR 1%, allelic ratio cutoff 0.7, minread 1
(Note: minread = minimum number of reads that must cover a SNP for it to be included in the analysis).

Calculating enrichment of H3K27ac near strain-biased genes that switch allelic status

Informative H3K27ac 4kb windows were extracted from the Allelome.PRO output for E12.5 liver and VE. Windows mapping to the X chromosome, and windows overlapping +/- 2kb of the transcription start side (TSS) of all RefSeq isoforms and non-coding loci were removed using BEDtools (version 2.20.1)(Quinlan and Hall 2010). The remaining H3K27ac windows were assigned to genes in our annotation if they were within +/- 50kb of the TSS (genes without SNPs were excluded, a window could be associated with more than one gene). For informative windows assigned to genes, we calculated the distance to the TSS (upstream (-) or downstream (+) taken from the middle of the window). We then shuffled the allelic status of the H3K27ac windows 100x to generate a random dataset that we subsequently used to calculate enrichment over random.

We selected a subset of genes for further analysis that showed strain-biased expression for CAST or FVB in liver or VE that then switched to BAE in the other tissue. In addition we called H3K4me3 peaks using MACS and performed an inner join (multiIntersectBed) from the 4 replicates for each tissue (Zhang et al. 2008). Next we removed strain-biased switchers where H3K4me3 peaks did not overlap the promoter (+/- 2kb of the annotated TSS) in both tissues. For these CAST (53 pc and 2 nc-genes) and FVB (82 pc and 2 nc-genes) switchers we then calculated H3K27ac allelic enrichment over random for each category (BAE, CAST, FVB) +/- 50kb from the TSS, when showing strain-biased or biallelic expression. The number of windows detected for each category were counted in 4kb bins over +/-50kb from the TSS, and enrichment over random calculated for each bin by dividing this number by the mean count for this category from the 100x shuffled allelic tags for the same genes. The H3K27ac enrichment was then plotted for BAE, CAST and FVB for each expression status (CAST, BAE (switching from CAST), FVB and BAE (switching from FVB) (Fig. 2C). To test for significant enrichment we performed a t-test comparing the 25 bins from the real data to the mean of the random data from the 25 bins (Table S1C).

Detecting X chromosome inactivation escapers

We detected X chromosome escapers as genes that deviated from the expected maternal bias in imprinted X chromosome inactivation (XCI) in extra-embryonic tissues, or the expected CAST bias due to skewed XCI in CASTxFVB crosses in other tissues. To increase the stringency in defining XCI escapers, we used a lower allelic ratio cutoff of 0.6 for the

Allelome.PRO analysis, compared to all other analysis in this study (Note: genes below the allelic ratio cutoff are classified biallelic, and therefore escapers). In addition, we excluded genes where all 4 replicates showed the expected MAT or CAST bias and a median allelic ratio above the cutoff, but were classified biallelic by Allelome.PRO because 1 or more replicates were below the FDR cutoff due to low expression. This approach enabled us to avoid setting an arbitrary RPKM cutoff for escapers.

Validation of the adult Leg Muscle XCI escapers

Next we tested if the 118 escapers in the leg muscle displayed the expected doubling of expression compared to the 103 non-escapers (Fig. 3C). We observed significant increase in the expression level of escapers compared to non-escapers (p=0.000787, Wilcoxon test), with a 1.73 fold median expression change (RPKM median: escapers 0.42, non-escapers 0.73).

Distance to H3K27ac maternal windows for XCI escapers and non-escapers in placenta

Maternal H3K27ac 4kb windows mapping to the X chromosome were extracted from the Allelome.PRO output for placenta E12.5. Windows overlapping +/- 2kb of the TSS of all RefSeq isoforms and non-coding loci were removed using BEDtools (version 2.20.1)(Quinlan and Hall 2010). For each maternal window (744 windows) the distance to the nearest escaper (36) and non-escaper (331) genes in E12.5 placenta was calculated using the Bedtools parameter closest "first". Maternal H3K27ac windows with a distance higher than 500kb were excluded from the analysis. Distances to the nearest maternal H3K27ac window were then plotted as a boxplot for both escapers and non-escapers (Fig. 3D).

To determine if the greater distance to the nearest maternal H3K27ac window observed for escapers was significant, we applied a statistical approach to correct for sample size. We compared the distance to the nearest maternal H3K27ac window for the 36 escapers with 36 non-escapers chosen randomly from the 331 non-escaper genes, and calculated a p-value (t-test). This was repeated a total of 10x, and then the p-values were combined using Fishers's exact test (sumlog method from the metap package in R). This indicated that H3K27ac maternal windows were significantly more distant from escapers than non-escapers (p= 1.318789e-21).

Reference: Michael Dewey (2014). metap: Meta-analysis of significance values. R package version 0.6. http://CRAN.R-project.org/package=metap

Published list of known imprinted genes

The list of 126 known imprinted genes was constructed by merging the Harwell and Otago imprinted databases and removing genes not annotated in RefSeq

(http://www.mousebook.org/imprinting-gene-list, www.otago.ac.nz/IGC (Glaser et al. 2006; Williamson CM 2013).

Detection of genes subject to imprinted silencing by *Airn*

To determine if novel imprinted genes detected in placenta near the *Igf2r* cluster belonged to the cluster, we examined whether imprinted silencing of these genes was regulated by *Airn*. Paternal deletion of the imprint control element (ICE) and *Airn* promoter (R2Δ) results in loss of imprinted expression for all genes in the *Igf2r* cluster (Wutz et al. 2001). Therefore, we compared the expression and allelic ratio calculated by Allelome.PRO of the novel candidates between RNA-seq for 3x CAST/FVB and 3x CAST/R2Δ E12.5 placentas (Fig. 5C, Table S1A). Differential gene expression was calculated using Cuffdiff (version 2.2.1) to compare the CASTxFVB and CASTxR2Δ samples and the q-value (corrected p-value) plotted (* 0.05 ≥ q-value > 0.01, ** 0.01 ≥ q-value > 0.001,*** 0.001 ≥ q-value). The allelic ratio for each replicate and genotype (CASTxFVB and CASTxR2Δ) was calculated using the Allelome.PRO pipeline, and then the mean and standard deviation was plotted. For the Allelome.PRO analysis we used a SNP annotation filtered for CAST/FVB SNPs where the FVB allele was shared with all 3 sequenced 129 strains (16,988,479 SNPs). This was necessary as the R2Δ allele was made on a 129 background, and therefore the region near the *Igf2r* may still be of 129 origin.

Validation of novel imprinted candidates

Following validation we classified the novel imprinted candidates into four categories (Table S1F-H):

I. **Candidate**: Detected in one tissue by RNA-seq (Note: maternal placenta candidates excluded from this category as they required further evidence)

II. **Supported Candidate**: Detected in multiple tissues or developmental stages (placenta candidates in different tissues), or located near a known imprinted region (<7Mb, distance defined in this study based on the distance of *Arid1b* to the ICE in the *Igf2r* cluster).

III. **Validated Candidate**: Imprinted expression confirmed by parental-specific H3K4me3 ChIP-seq enrichment on the promoter (Table S1F), or by showing a loss of imprinted expression by deleting the ICE as demonstrated for the *Igf2r* cluster.

IV. **Maternal Contamination (candidate excluded)**: Genes showing maternal expression restricted to placenta with a Decidua/Placenta expression ratio >5, and not supported by any other validation method were defined as maternal contamination. Additionally, blood-specific genes detected as maternally expressed in placenta were also classified as maternal contamination and excluded.

Note: lncRNA candidates were defined as fragments if they were in proximity to a known imprinted lncRNA that showed the same allelic status, and were not supported as an independent transcript by H3K4me3 enrichment on their promoter. Fragments were classified into the same 4 validation categories as independent transcripts.

Parental-specific H3K27ac enrichment within imprinted regions

Autosomal parental-specific H3K27ac 4kb sliding windows were extracted from the Allelome.PRO output for placenta and VE. Windows overlapping +/- 2kb of the TSS of all RefSeq isoforms and non-coding loci were removed using BEDtools (version 2.20.1) (Quinlan and Hall 2010). Next we counted the overlap of parental-specific H3K27ac windows with 100kb non-overlapping count windows using the count function of BEDtools (intersect –c). For each tissue we generated a BED file including the imprinted region based on confirmed known imprinted genes, and supported and validated candidates (Table S1I). For each tissue, the imprinted region BED file was joined with the associated BED file containing the parental-specific window counts using BEDtools (intersect). Next we filtered for 100kb count windows that contained at least one 4kb parental-specific window count. In VE, 14 were overlapping imprinted regions while 13 were outside. We observed significant enrichment of H3K27ac in imprinted regions (t-test, p= 5.466139-07). In placenta, 65 count windows overlapped imprinted regions while 3737 were outside. To correct for sample size we randomly took 65 count windows outside imprinted regions and compared count number to windows in imprinted regions. We repeated this a total of 100x and combined the p-values using Fisher's exact test (sumlog method from the metap package in R), showing a significant enrichment of H3K27ac in imprinted regions in placenta (p=0).

The maximum count outside of imprinted regions was then used as cutoff to define the background (VE=3 and Pl=5). The remaining 100kb count windows within imprinted regions were then plotted in R (Fig. 6B).

REFERENCES

Andergassen D, Dotter CP, Kulinski TM, Guenzl PM, Bammer PC, Barlow DP, Pauler FM, Hudson QJ. 2015. Allelome.PRO, a pipeline to define allele-specific genomic features from high-throughput sequencing data. *Nucleic Acids Res* **43**(21): e146.

Bryja V, Bonilla S, Arenas E. 2006. Derivation of mouse embryonic stem cells. *Nature protocols* **1**(4): 2082-2087.

Capel B, Albrecht KH, Washburn LL, Eicher EM. 1999. Migration of mesonephric cells into the mammalian gonad depends on Sry. *Mechanisms of development* **84**(1-2): 127-131.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**(1): 15-21.

Glaser RL, Ramsay JP, Morison IM. 2006. The imprinted gene and parent-of-origin effect database now includes parental origin of de novo mutations. *Nucleic Acids Res* **34**(Database issue): D29-31.

Hudson QJ, Seidl CI, Kulinski TM, Huang R, Warczok KE, Bittner R, Bartolomei MS, Barlow DP. 2011. Extra-embryonic-specific imprinted expression is restricted to defined lineages in the post-implantation embryo. *Dev Biol* **353**(2): 420-431.

Joshi PA, Jackson HW, Beristain AG, Di Grappa MA, Mote PA, Clarke CL, Stingl J, Waterhouse PD, Khokha R. 2010. Progesterone induces adult mammary stem cell expansion. *Nature* **465**(7299): 803-807.

Keane TM, Goodstadt L, Danecek P, White MA, 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**(7364): 289-294.

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

Kornienko AE, Dotter CP, Guenzl PM, Gisslinger H, Gisslinger B, Cleary C, Kralovics R, Pauler FM, Barlow DP. 2016. Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans. *Genome Biol* **17**(1): 14.

Kulinski TM, Casari MR, Guenzl PM, Wenzel D, Andergassen D, Hladik A, Datlinger P, Farlik M, Theussl HC, Penninger JM et al. 2015. Imprinted expression in cystic embryoid bodies shows an embryonic and not an extra-embryonic pattern. *Dev Biol* **402**(2): 291-305.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**(6): 841-842.

Washietl S, Findeiss S, Muller SA, Kalkhof S, von Bergen M, Hofacker IL, Stadler PF, Goldman N. 2011. RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. *Rna* **17**(4): 578-594.

Williamson CM BA, Thomas S, Beechey CV, Hancock J, Cattanach BM, and Peters J. 2013. World Wide Web Site - Mouse Imprinting Data and References.

Wutz A, Theussl HC, Dausman J, Jaenisch R, Barlow DP, Wagner EF. 2001. Non-imprinted Igf2r expression decreases growth and rescues the Tme mutation in mice. *Development* **128**(10): 1881-1887.

Ying QL, Smith AG. 2003. Defined conditions for neural commitment and differentiation. *Methods in enzymology* **365**: 327-341.

Ying QL, Wray J, Nichols J, Batlle-Morera L, Doble B, Woodgett J, Cohen P, Smith A. 2008. The ground state of embryonic stem cell self-renewal. *Nature* **453**(7194): 519-523.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**(9): R137.

DISCUSSION

1. Generating a bioinformatics pipeline to detect allele-specific genome features.

Following the development of RNA-seq, genome-wide mapping of allele-specific expression has been performed in human and mouse tissues (Babak et al., 2015; Baran et al., 2015; Crowley et al., 2015; Keane et al., 2011; Pickrell et al., 2010). Mapping of allelic expression requires single nucleotide polymorphisms or SNPs to assign RNA-seq reads to the corresponding allele. Genome-wide detection of allelic expression in human is difficult since the genotype of the SNP cannot be assigned to the parental allele without also genotyping both parents. In contrast the mouse model is a powerful system to map allele-specific expression since inbred strains are available, and the SNPs between the strains are annotated (Keane et al., 2011). Several groups used this system to map imprinted expression from reciprocal mouse crosses in different tissues and development stages and found only a limited number of novel imprinted genes (Babak, 2012; Babak et al., 2008; Babak et al., 2015; Lagarrigue et al., 2013; Proudhon and Bourc'his, 2010; Tran et al., 2014; Wang et al., 2008; Xie et al., 2012). In contrast, one study using a similar method reported more than thousand novel imprinted genes in the mouse brain (Gregg et al., 2010b) many of which were reported to differ between male and female brains (Gregg et al., 2010a). However reanalysis of the data demonstrated that the majority of novel identified imprinted genes were false positives to due technical and biological variation that was not considered in the original study (DeVeale et al., 2012). Although RNA-seq is a powerful tool to quantify parental-specific expression, the following aspects are essential in order to have less false positive calls (Wang and Clark, 2014): Improvement of the library complexity by using more input RNA, the use of biological and technical replicates, independent validation of the candidates, false discovery rate (FDR) cut-off and introduction of an allelic ratio cut-off.

With these requirements in mind in my first publication we developed Allelome.PRO, a user-friendly, fully automated bioinformatics pipeline to robustly detect allele-specific expression or chromatin modifications, from high-throughput sequencing data. The pipeline automatically characterizes all genes into biallelic, strain-biased, imprinted or non-informative and thus provides the full allelic expression picture, called 'the Allelome'. We demonstrated the high sensitivity of the Allelome.PRO pipeline by mapping for the first time the expression Allelome in MEFs from RNA-

seq data. In addition we used these datasets to define a biological relevant allelic ratio cut-off based on known imprinted genes and skewed X-chromosome inactivation. We also compared the expression Allelome to allele-specific H3K4me3 enrichment over promoters from ChIP-seq data and demonstrated that allelic enrichment is a robust method to independently validate allele-specific expression.

1.1 Allelome.PRO a robust user-friendly pipeline to define the Allelome

Mapping allelic expression requires a robust approach to reduce the number of false positives, but sensitive enough to detect all the genes that show an allelic bias. Previously, a user-friendly pipeline to characterize allele-specific expression that did not require high levels of bioinformatics skills was lacking. In contrast, to map allelic expression using Allelome.PRO, the user only has to provide a SNP and gene annotation and four aligned sequencing files, generated from reciprocal crosses. The pipeline runs on a standard desktop computer and only needs a few parameters to be entered including the FDR and allelic ratio cut-off in order to characterize the genes in the annotation file into biallelic, strain-biased, imprinted or not informative due to low expression or lacking SNPs in the gene body. In order to increase the detection sensitivity we generated ribosomal RNA depleted libraries that allow the detection of SNPs in introns and in non-polyadenylated lncRNAs. To increase the robustness of the pipeline, the Allelome.PRO algorithm automatically calculates for each gene and replicate, an allelic score and uses this data to generate mock comparisons in order to empirically set the false discovery rate cut-off. This approach was used in earlier studies to robustly call imprinted expression (DeVeale et al., 2012). The disadvantage of the allelic score alone is that a highly expressed gene with a small allelic ratio bias can lead to a score above the cut-off. Therefore I used the MEF allelic expression data to define a single biological relevant allelic ratio cut-off for imprinted and strain-biased expression based on known imprinted genes and skewed XCI. Consequently I investigated the allelic ratio for all the genes that show parental-specific expression in autosomes and strain specific expression on the X-chromosome, based one the allelic score alone. I detected 65 genes showing a parental bias, including 43 known imprinted genes with an allelic ratio higher than 0.85 and 22 novel candidates representing much lower ratios. The majority of expressed genes on the X-chromosome were skewed towards the *Mus musculus castaneus* strain with a mean allelic ratio of 0.735, a known effect between *Mus musculus castaneus* and *Mus musculus domesticus* (Calaway et al., 2013). The allelic ratio analysis of imprinted

and strain-biased expression revealed that a single allelic ratio cut-off of 0.7 robustly detects 95% known imprinted genes and 85% of strain-biased genes showing the expected skewing pattern. In addition, the allelic ratio cut-off allows the separation of genes showing an allelic bias from biallelic and non-informative genes. A biallelic gene is defined to have enough SNP coverage to be above the FDR cut-off assuming an allelic ratio equal to the allelic ratio cut-off. Furthermore, we demonstrated that Allelome.PRO is a useful tool to identify genes that escape XCI in females, by assuming that escapers are not showing the expected skewing ratio and rather show biallelic expression.

Next we demonstrated that allele-specific enrichment of H3K4me3 over promoters generally confirms the allelic expression status (84%). We obtained the highest validation rate for biallelic (90%) and imprinted genes (70%). However for strain-biased expression, the allelic promoter mark validated only 29%. This result could indicate that only one third of strain-biased genes arise from allelic transcription, and thus be controlled by allelic differences in *cis*-acting regulatory regions such as promoter and enhancers. The remaining strain-biased genes may be explained by allelic difference influencing post-transcriptional regulation, such as alternative splicing, miRNA binding or RNA stability (Gilad et al., 2008; Lappalainen et al., 2013; Li et al., 2012; Majewski and Pastinen, 2011) or by alignment biases towards one mouse strain.

In addition, Allelome.PRO can be used for other histone modification such as H3K27ac, a mark for active enhancers and promoters, indicating that any high-throughput sequencing data from F1 tissues could be examined. In this case, the allelic information of such enhancer marks would make it easier to identify the nearby target genes assuming allelic enrichment matches allelic expression. Moreover Allelome.PRO could also be used for other organism as long the two strains are inbred and the SNPs are available.

2. The mouse Allelome reveals tissue-specific regulation

For my second publication I used Allelome.PRO to map allele-specific expression for protein and non-coding genes in 23 mouse tissues during development including pluripotent, embryonic, extra-embryonic, neonatal and adult, to define for the first time the mouse Allelome. In order to analyze allele-specific expression on the X-chromosome, which enables the identification of XCI escaper genes, we mainly

collected female tissues. The generated dataset provides the most comprehensive survey of allelic expression to date, including biallelic, strain-biased and imprinted expression and all the genes that escape XCI. I found that strain-biased expression for protein-coding genes is highly tissue-specific and correlates with nearby allele-specific H3K27ac modification of presumptive enhancer regions. In addition, I identified high tissue-specific variation in XCI escaper genes with the highest number of 50% escaper genes in adult leg muscle, compared to an average of 10% over all tissues. The complete picture of imprinted expression in mouse identified a substantial number of novel imprinted genes but revealed that most of the novel validated imprinted genes belong to known clusters. In addition, I found that imprinted clusters are much larger than previously thought and change their size between tissue and development. In particular, I found that the *Igf2r* cluster expands to 10Mb in E12.5 placenta representing the largest imprinted region in mouse.

2.1 Tissue-specific strain-biased expression is regulated by enhancer switching

In order to generate a comprehensive allelic expression profile for protein and non-coding genes during mouse development I combined the robust RefSeq annotation with the loci obtained from transcriptome assembly to run Allelome.PRO. This is important since the RefSeq annotation mainly contains protein-coding genes and is closer to the FVB/NJ (FVB) laboratory mouse strain than the CAST/EiJ (CAST) strain, given that for the reference sequence the laboratory mouse strain C57Bl6 was used. The mouse Allelome for protein-coding genes revealed high tissue-specific regulation for strain-biased and imprinted expression, compared to genes showing biallelic expression. The low tissue-specific expression of biallelic genes might be explained by housekeeper genes that are essential for cell viability, and this suggestion is supported by the finding that 31% are expressed in all assayed tissues. In contrast for non-coding genes, I found high tissue-specific variation for both biallelic and monoallelic expression, which can be explained by known highly tissue-specific expression for many non-coding genes (Cabili et al., 2011; Guttman et al., 2009). Tissue-specific biases in allelic expression can have two explanations. First, gene expression could be absent in some tissues - so called 'tissue-specific expression'. Second, the allelic status can switch from one informative state to another, for example from biallelic expression to strain-biased or imprinted expression or *vice versa*. To further investigate this I calculated the percentage of tissues that maintain the allelic status for protein-coding genes. I found that the

majority of biallelically-expressed protein-coding genes remain biallelic wherever they are expressed. In contrast, strain-biased and imprinted protein-coding genes include a high gene number that switch from monoallelic to biallelic expression. This finding reveals tissue-specific regulation for strain-biased and imprinted expression that might be explained by tissue-specific enhancers that act to override the previous allelic state. Recently, a study mapped allelic expression and H3K27ac enrichment in human tissues to demonstrate that allele-specific genetic variants in enhancer regions cause allele-specific transcription factor binding resulting in allele-specific expression (Leung et al., 2015). However, genome-wide detection of allelic features in human is difficult since genotyping of the parents is required to know the SNPs between the alleles. Compared to human, the inbred mouse model has advantages for investigating tissue-specific regulation of allelic features, since the SNPs between the different strains are known and biological replicates of the same genetic background can be generated increasing the statistical power of the analysis. Using this system we first performed H3K27ac ChIP-seq from F1 hybrids of E12.5 liver and visceral yolk sac endoderm (VE) to calculate allele-specific enrichment using Allelome.PRO. Next I selected all the genes that switch from biallelic to strain-biased expression between the two tissues and investigated allelic enrichment ± 50kb from the transcription start side. Overall I found strain-biased H3K27ac enrichment matching strain-biased expression, while the same subset of genes showing biallelic expression in the other tissues showed no enrichment. This enrichment explained 27% of genes switching their allelic status, while the others might have an enhancer outside of the examined ± 50kb window, or might be explained by tissue-specific post-transcriptional regulation. Our findings are in agreement with the reports in human that allelic differences in regulatory regions cause allelic expression (Leung et al., 2015).

2.2 XCI escaper are highly tissue-specific

The process of XCI in female mammals is important for dosage compensation between the sexes and thus responsible to silence all the X-linked genes from one X-chromosome. However some genes have been reported to escape this process in mouse (3%) and human (15%) and are thus expressed from both X-chromosomes (Berletch et al., 2015; Berletch et al., 2011). Such escapers can be identified in RNA-seq data as genes on the X-chromosome that do not show the expected XCI skewing or imprinted pattern and rather show biallelic expression. A recent study analyzed a limited number of tissues for escaper genes in mouse and found that XCI escape

occurs in a tissue-specific manner (Berletch et al., 2015). Here we provide the largest survey of XCI escaper genes by analyzing 19 female mouse tissues and developmental stages and we report 250 escaper candidate genes, including 31 previously reported. Approximately half of the detected candidate genes escape in more than one tissue, while the others escape only in single tissue, confirming the reports of tissue-specific escaping. Overall I found that the average percentage of X-chromosome escaping in the mouse is similar as reported for human (Carrel and Willard, 2005). However to our surprise I found that in mouse adult leg muscle 50% of genes escape XCI. I validated this result by demonstrating that the high number of escapers show the expected two-fold increase in gene expression levels compared to non-escapers. Interestingly an overlap of 62% of the escaper genes was observed between the muscle tissues tongue and leg muscle, including the gene Dystrophin (*Dmd*) an important protein for muscle function. Mutation in *Dmd* causes the disease Duchenne muscular dystrophy and mainly affects males as they only have one copy on their single X-chromosome. In contrast females have two copies and thus rarely affected by the disorder. However, escaping of *Dmd* from X-chromosome inactivation might be an additional advantage for females, since skewed X-inactivation could lead to monoallelic expression of the mutant gene in the muscle syncytium, a cell with multiple nuclei. Overall the high number of escapers might be explained by the sensitivity of our approach, using total RNA-seq to detect intronic SNPs, an extended annotation and an unprecedented number of investigated tissues. These results are the foundation of future studies to understand how tissue-specific escaping is regulated.

2.3 Novel imprinted genes expand know imprinted clusters

In the early days of genomic imprinting novel imprinted genes were detected in mouse using a diverse set of approaches including testing candidates mapping close to known imprinted loci or known genes, or from observing parental-specific effects of single gene knockouts (reviewed in (Barlow and Bartolomei, 2014)). The development of RNA-seq now allows a genome-wide approach and can be applied to both protein-coding and non-coding genes. Several groups have now mapped imprinted expression in mouse F1 hybrids in different tissues (Babak, 2012; Babak et al., 2008; Babak et al., 2015; Lagarrigue et al., 2013; Proudhon and Bourc'his, 2010; Tran et al., 2014; Wang et al., 2008; Xie et al., 2012). This data plus the previous decades of work in the imprinting field lead to the discovery of approximately 150 imprinted genes, in different tissues and developmental stages, listed in the current

Harwell and OTAGO imprinted databases (Glaser et al., 2006; Williamson CM, 2013). However from this catalogue only 126 are annotated in RefSeq, the most validated gene annotation for the mouse genome. From this subset 33 genes are disputed mainly because of weak or not reproducible data or due to maternal placenta contamination (Glaser et al., 2006; Okae et al., 2012; Proudhon and Bourc'his, 2010). Recently a study mapped imprinted expression in 33 male mouse tissues using polyA+ RNA-seq and found a high number of known imprinted genes early in development, in extra-embryonic tissues and in the brain (Babak et al., 2015). In contrast we used a combination of total RNA-seq and a custom annotation based on our data in order to identify lowly expressed imprinted lncRNAs, female tissues to investigate imprinted expression on the X-chromosome and a developmental series of the same tissues to map imprinted expression during development. In addition, I selected relevant tissues for postnatal feeding, such as virgin and lactating mammary glands or the neonatal tongue, organs for which genomic imprinting was suggested to be important (Peters, 2014; Stringer et al., 2014). In our RNA-seq dataset I detect 71 known imprinted genes including 5 that have been previously disputed and 76 novel imprinted candidates. I could not confirm the remaining 27 'known' imprinted genes in our study for a variety of reasons: lack of SNPs in the gene body (2 genes), low expression just below our robust ratio cut-off (11 genes), not expressed or biallelic expressed (12 genes), or the tissues in which imprinted expression was reported was not analyzed in our study (2 genes). In agreement with the literature I find the highest number from the 126 known imprinted genes in extra-embryonic (42%) tissues such as the placenta and the visceral yolk sac endoderm (VE) and during embryonic and neonatal development (36.5%). An exception of this rule is the brain that shows a high number of known imprinted genes throughout mouse development (36.5%). In addition, I also detect a decrease in the imprinted gene number within the same tissues during development indicating an important role for imprinted expression early in development. Similar as I have observed for strain-biased protein-coding genes, imprinted protein-coding genes switch to a different informative allelic state in a different tissue, revealing tissue-specific regulation of imprinted expression. However, imprinted lncRNA do not switch their allelic status and thus show imprinted expression wherever they are transcribed. This might be explained by most imprinted lncRNAs being directly regulated by a gDMR. The relatively high number of known imprinted genes in the placenta (37.5%), neonatal tongue (30%) and mammary glands (18.2%) supports a role for imprinted expression in the energy

transfer from the mother to the offspring. Notably I observed no significant change in imprinted expression profile between mammary glands and maternal brains dissected from virgin and breast-feeding females, suggesting that imprinted expression might not be relevant during lactation.

Although Allelome.PRO robustly identifies imprinted expression, different factors can lead to false positive calls. From the 76 novel imprinted candidates identified in this study 19 show maternal expression restricted to placenta, which might result from maternal blood or maternal decidua contamination of the placenta. In addition, I detect 19 novel imprinted lncRNAs located adjacent to known imprinted lncRNAs with the same parental bias, indicating they may be an extension rather than novel independent transcripts. However, I also detect several novel imprinted candidates located close (*i.e.,* within 7Mb) to known imprinted genes indicating they might be part of the same cluster. In order to validate the novel imprinted candidates, I used Allelome.PRO to map allele-specific H3K4me3 enrichment on promoters using a sliding window annotation. In addition, this method allows the identification of the transcription start side and thus distinguishes independent imprinted transcripts from an extension of a nearby imprinted gene. This approach validated a single maternally expressed lncRNA (*Gm35612*) on the X-chromosome, with the exception of *Xist* that shows imprinted expression restricted to extra-embryonic tissues. The novel imprinted lncRNA *Gm35612* is expressed antisense to the lncRNA *Firre* that is involved in the nuclear organization (Hacisuleyman et al., 2014). Since *Gm35612* is the only imprinted locus on the X-chromosome in addition to the *Xist* lncRNA, there might be a link to imprinted X-chromosome inactivation in extra-embryonic tissues, which needs to be further investigated. In addition, the validation of the paternally expressed non-coding gene *XLOC_032279* on chromosome 15 in several extra-embryonic tissues located in the proximity to the known solo imprinted gene *Slc38a4*, indicates that both are part of a novel imprinted gene cluster. In order to get an indication whether the placenta specific candidates are explained by maternal decidua contamination, I sequenced RNA prepared from isolated decidua and calculated the decidua/placenta expression ratio for each candidate. I observed that approximately 60% have high ratios, indicating that most of them might be false positive to due maternal contamination. Remarkably, I observed that 5 candidates (*Arid1b, Park2, Dact2, Smoc2, Thbs2*) with an overall low decidua/placenta ratio are in proximity to the *Igf2r* cluster, indicating that these genes might be also regulated by the lncRNA

Airn and consequently part of the cluster. Therefore I examined regulation by *Airn* using a cross between CAST and a mouse lacking the *Airn* promoter (FVB.129P2-Airn-R2D). This cross enables the expression levels and allelic ratio changes of the candidates to be compared in the presence and absence of the *Airn* lncRNA. Our results show either a significant up regulation in expression or a switch from maternal to biallelic ratios in in the absence of *Airn*, demonstrating that the 5 genes are part of the imprinted cluster. The novel imprinted genes expand the *Igf2r* cluster size from 4Mb to 10Mb, representing the largest imprinted region in mouse. In summary, I confirm 70 known imprinted genes and validate 23 from the 76 candidates detect by RNA-seq. This generates a robust set of 93 imprinted genes. In addition I speculate that 90.3% of the confirmed or validated imprinted genes are organized in clusters, by assigning novel imprinted genes to known imprinted regions, using the *Igf2r* cluster size as a benchmark. Although we find a high number of genes showing tissue and developmental specific imprinted expression, we do not expect to find many more imprinted genes since we and others have covered the majority of tissues during mouse development.

2.4 Imprinted clusters expand and contract during tissue and development

In order to investigate the tissue and developmental dynamic of imprinted clusters, I used our robust set of 93 imprinted genes and examined the size of all clusters during tissue and development. Overall I detect the smallest cluster size in embryonic stem cell that represent a pluripotent cell type. The cluster size expands in embryonic and neo-natal development and contracts again to a minimum in adults. The largest expansion in extra-embryonic tissues, in particular in the placenta was observed for the *Kcnq1* and *Igf2r* cluster, both regulated by *cis*-acting imprinted lncRNAs. In order to investigate how a ubiquitously expressed imprinted lncRNA can dramatically affect the size of imprinted clusters in a tissue-specific manner I examined allele-specific H3K27ac in the *Igf2r* cluster using Allelome.PRO. I choose three different tissues, representing different sizes of the cluster: the embryonic liver representing the smallest cluster size, VE showing a minor extension and the placenta with a massive extension. Surprisingly I detect parental-specific maternal enrichment of H3K27ac correlating with the size of the cluster. This enrichment was significant for the *Igf2r* imprinted region in extra-embryonic tissues that show a cluster extension, while in embryonic liver, I detect only background H3K27ac levels. Interestingly the maternal H3K27ac enrichment over the *Igf2r* cluster in placenta shows a similar profile as

detected over the entire X-chromosome in placenta, a tissue showing imprinted XCI. Since previous studies in trophoblast stem cells demonstrated that *Xist* might target regulatory regions by the deposition of H3K27me3 on the paternal allele, leading to maternal H3K27ac enrichment (Calabrese et al., 2012), parental-specific H3K27me3 in the *Igf2r* cluster might be interesting to investigate in the future.

3. Future outlook

The development of Allelome.PRO, an easy to use and fully automated bioinformatics pipeline, which detects allele-specific genomic features from high-throughput sequencing data allowed us to map the mouse Allelome. The Allelome is defined as the allelic expression status of all genes in a comprehensive set of tissues during the mouse development and includes all the biallelic and strain-biased genes, all the imprinted genes and all the genes that escape XCI. In addition I demonstrated genetically that the ubiquitously expressed imprinted lncRNA *Airn* dramatically affects the size of imprinted clusters in a tissue-specific manner, with the largest cluster size in placenta including 10% of chromosome 17. This huge region regulated by *Airn* shows similarities to XCI and might show similarities in their silencing mechanisms.

The results in this Thesis could allow several lines of further investigation that would increase our understanding of how lncRNAs regulate allelic expression. For example, the demonstration that the *cis*-acting lncRNA *Xist* reaches distant regions by exploding the three-dimensional chromosome structure that exists on both alleles, before the deposition of H3K27me3 by PRC2 (Engreitz et al., 2013) could also apply to the expanded *Igf2r* cluster in placenta. In order to test this model for the *Igf2r* cluster I would perform Circular chromosome conformation capture sequencing analysis (4C-seq) with the *Airn* gene body as bait in F1 tissues that represent different cluster sizes, allowing the identification of allele-specific interactions using Allelome.PRO. Next I would investigate if these interactions correlate with the maternal enhancer windows. To further strengthen this analysis I would repeat the experiment using the mouse model with the truncated non-functional version of *Airn* to ask whether the interactions are reestablished on both alleles in the absence of *Airn*.

Another project that might be interesting to investigate is whether some of the thousand lncRNAs outside of imprinted region act similar to imprinted lncRNAs that silence entire gene clusters *in cis*. The targets of imprinted lncRNAs are relatively

easy to detect since they are expressed from the opposite parental allele. I hypothesize that other non-imprinted lncRNAs might function similarly, however since those are not controlled by gDMRs and thus show biallelic expression, all the target genes are silenced on both alleles during a short developmental time window and are therefore difficult to detect. For the mouse Allelome project we derived F1 ESCs from several female and male blastocysts. The idea of this project is to generate 96 heterozygote lncRNA knockouts in these ESCs by introducing promoter deletion with CRISPR/Cas9 in order to mimic allele-specific expression. Next I would differentiate these clones and calculate allele-specific expression. For a *cis*-acting repressive lncRNA I would expect that the distant targets show the opposite allelic expression pattern to the lncRNA. In order to get 96 strong lncRNA candidates I would filter for those that are upregulated during differentiation, mainly localized in the nucleus and conserved between different mouse strains. This experiment would finally provide evidence if lncRNAs outside of imprinted regions regulate entire gene clusters *in cis*.

Finally it might be interesting to characterize the imprinted maternally expressed lncRNA *Gm35612* that I discovered by mapping imprinted expression in 19 female tissues on the X-chromosome. Since this lncRNA is the only imprinted gene on the X-chromosome outside of extra-embryonic tissues that I could validate, it may be interesting to further investigate this region for a connection to imprinted XCI. For this project I would first investigate if this promoter is the missing gDMR on the X-chromosome by analyzing public available whole-genome bisulfite sequencing data from sperm and oocyte. If I have enough evidence I would like to then generate a mouse with a promoter deletion and cross it with a CAST mouse. Next I would sequence the placentas of the F1 offspring and run it thought the Allelome.PRO pipeline to investigate if imprinted XCI is lost.

Together the results in this thesis form a solid foundation to investigate further the molecular causes of allele-specific expression and its potential to influence development and disease.

REFERENCES

1. Albert, F.W., and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. Nature reviews Genetics *16*, 197-212.

2. Amendola, M., and van Steensel, B. (2014). Mechanisms and dynamics of nuclear lamina-genome interactions. Current opinion in cell biology *28*, 61-68.

3. Andergassen, D. (2012). Extension of imprinted silencing in the Igf2r cluster in extra-embryonic tissues (Universität Wien).

4. Arney, K.L. (2003). H19 and Igf2--enhancing the confusion? Trends Genet *19*, 17-23.

5. Augui, S., Nora, E.P., and Heard, E. (2011). Regulation of X-chromosome inactivation by the X-inactivation centre. Nature reviews Genetics *12*, 429-442.

6. Babak, T. (2012). Identification of imprinted loci by transcriptome sequencing. Methods in molecular biology *925*, 79-88.

7. Babak, T., Deveale, B., Armour, C., Raymond, C., Cleary, M.A., van der Kooy, D., Johnson, J.M., and Lim, L.P. (2008). Global survey of genomic imprinting by transcriptome sequencing. Current biology : CB *18*, 1735-1741.

8. Babak, T., DeVeale, B., Tsang, E.K., Zhou, Y., Li, X., Smith, K.S., Kukurba, K.R., Zhang, R., Li, J.B., van der Kooy, D.*, et al.* (2015). Genetic conflict reflected in tissue-specific maps of genomic imprinting in human and mouse. Nature genetics *47*, 544-549.

9. Baran, Y., Subramaniam, M., Biton, A., Tukiainen, T., Tsang, E.K., Rivas, M.A., Pirinen, M., Gutierrez-Arcelus, M., Smith, K.S., Kukurba, K.R.*, et al.* (2015). The landscape of genomic imprinting across diverse adult human tissues. Genome Res *25*, 927-936.

10. Barlow, D.P. (2011). Genomic imprinting: a Mammalian epigenetic discovery model. Annu Rev Genet *45*, 379-403.

11. Barlow, D.P., and Bartolomei, M.S. (2014). Genomic imprinting in mammals. Cold Spring Harbor perspectives in biology *6*.

12. Barlow, D.P., Stoger, R., Herrmann, B.G., Saito, K., and Schweifer, N. (1991). The mouse insulin-like growth factor type-2 receptor is imprinted and closely linked to the Tme locus. Nature *349*, 84-87.

13. Bartolomei, M.S., and Ferguson-Smith, A.C. (2011). Mammalian genomic imprinting. Cold Spring Harbor perspectives in biology *3*.

14. Bartolomei, M.S., Zemel, S., and Tilghman, S.M. (1991). Parental imprinting of the mouse H19 gene. Nature *351*, 153-155.

15. Bell, A.C., and Felsenfeld, G. (2000). Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. Nature *405*, 482-485.

16. Berletch, J.B., Ma, W., Yang, F., Shendure, J., Noble, W.S., Disteche, C.M., and Deng, X. (2015). Escape from X inactivation varies in mouse tissues. PLoS genetics *11*, e1005079.

17. Berletch, J.B., Yang, F., Xu, J., Carrel, L., and Disteche, C.M. (2011). Genes that escape from X inactivation. Human genetics *130*, 237-245.

18. Borsani, G., Tonlorenzi, R., Simmler, M.C., Dandolo, L., Arnaud, D., Capra, V., Grompe, M., Pizzuti, A., Muzny, D., Lawrence, C.*, et al.* (1991). Characterization of a murine gene expressed from the inactive X chromosome. Nature *351*, 325-329.

19. Bourc'his, D., Xu, G.L., Lin, C.S., Bollman, B., and Bestor, T.H. (2001). Dnmt3L and the establishment of maternal genomic imprints. Science *294*, 2536-2539.

20. Brady, B.L., Steinel, N.C., and Bassing, C.H. (2010). Antigen receptor allelic exclusion: an update and reappraisal. Journal of immunology *185*, 3801-3808.

21. Brem, R.B., Yvert, G., Clinton, R., and Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. Science *296*, 752-755.

22. Bressler, J., Tsai, T.F., Wu, M.Y., Tsai, S.F., Ramirez, M.A., Armstrong, D., and Beaudet, A.L. (2001). The SNRPN promoter is not required for genomic imprinting of the Prader-Willi/Angelman domain in mice. Nature genetics *28*, 232-240.

23. Brockdorff, N., Ashworth, A., Kay, G.F., Cooper, P., Smith, S., McCabe, V.M., Norris, D.P., Penny, G.D., Patel, D., and Rastan, S. (1991). Conservation of position and exclusive expression of mouse Xist from the inactive X chromosome. Nature *351*, 329-331.

24. Brown, C.J., Ballabio, A., Rupert, J.L., Lafreniere, R.G., Grompe, M., Tonlorenzi, R., and Willard, H.F. (1991). A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. Nature *349*, 38-44.

25. Brown, C.J., Carrel, L., and Willard, H.F. (1997). Expression of genes from the human active and inactive X chromosomes. American journal of human genetics *60*, 1333-1343.

26. Brown, C.J., and Willard, H.F. (1994). The human X-inactivation centre is not required for maintenance of X-chromosome inactivation. Nature *368*, 154-156.

27. 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 & development *25*, 1915-1927.

28. Calabrese, J.M., Starmer, J., Schertzer, M.D., Yee, D., and Magnuson, T. (2015). A survey of imprinted gene expression in mouse trophoblast stem cells. G3 *5*, 751-759.

29. Calabrese, J.M., Sun, W., Song, L., Mugford, J.W., Williams, L., Yee, D., Starmer, J., Mieczkowski, P., Crawford, G.E., and Magnuson, T. (2012). Site-specific silencing of regulatory elements as a mechanism of X inactivation. Cell *151*, 951-963.

30. Calaway, J.D., Lenarcic, A.B., Didion, J.P., Wang, J.R., Searle, J.B., McMillan, L., Valdar, W., and Pardo-Manuel de Villena, F. (2013). Genetic architecture of skewed X inactivation in the laboratory mouse. PLoS genetics *9*, e1003853.

31. Carrel, L., and Willard, H.F. (2005). X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature *434*, 400-404.

32. Cattanach, B.M. (1975). Control of chromosome inactivation. Annu Rev Genet *9*, 1-18.

33. Cattanach, B.M., and Isaacson, J.H. (1965). Genetic control over the inactivation of autosomal genes attached to the X-chromosome. Zeitschrift fur Vererbungslehre *96*, 313-323.

34. Cattanach, B.M., and Isaacson, J.H. (1967). Controlling elements in the mouse X chromosome. Genetics *57*, 331-346.

35. Cattanach, B.M., and Kirk, M. (1985). Differential activity of maternally and paternally derived chromosome regions in mice. Nature *315*, 496-498.

36. Chaumeil, J., Le Baccon, P., Wutz, A., and Heard, E. (2006). A novel role for Xist RNA in the formation of a repressive nuclear compartment into which genes are recruited when silenced. Genes & development *20*, 2223-2237.

37. Chess, A. (2012). Mechanisms and consequences of widespread random monoallelic expression. Nature reviews Genetics *13*, 421-428.

38. Chu, C., Zhang, Q.C., da Rocha, S.T., Flynn, R.A., Bharadwaj, M., Calabrese, J.M., Magnuson, T., Heard, E., and Chang, H.Y. (2015). Systematic discovery of Xist RNA binding proteins. Cell *161*, 404-416.

39. Crowley, J.J., Zhabotynsky, V., Sun, W., Huang, S., Pakatci, I.K., Kim, Y., Wang, J.R., Morgan, A.P., Calaway, J.D., Aylor, D.L*., et al.* (2015). Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. Nature genetics *47*, 353-360.

40. Csankovszki, G., Panning, B., Bates, B., Pehrson, J.R., and Jaenisch, R. (1999). Conditional deletion of Xist disrupts histone macroH2A localization but not maintenance of X inactivation. Nature genetics *22*, 323-324.

41. da Rocha, S.T., Boeva, V., Escamilla-Del-Arenal, M., Ancelin, K., Granier, C., Matias, N.R., Sanulli, S., Chow, J., Schulz, E., Picard, C*., et al.* (2014). Jarid2 Is Implicated in the Initial Xist-Induced Targeting of PRC2 to the Inactive X Chromosome. Molecular cell *53*, 301-316.

42. DeChiara, T.M., Robertson, E.J., and Efstratiadis, A. (1991). Parental imprinting of the mouse insulin-like growth factor II gene. Cell *64*, 849-859.

43. Deng, Q., Ramskold, D., Reinius, B., and Sandberg, R. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science *343*, 193-196.

44. Deng, X., Berletch, J.B., Ma, W., Nguyen, D.K., Hiatt, J.B., Noble, W.S., Shendure, J., and Disteche, C.M. (2013). Mammalian X upregulation is associated with enhanced transcription initiation, RNA half-life, and MOF-mediated H4K16 acetylation. Dev Cell *25*, 55-68.

45. DeVeale, B., van der Kooy, D., and Babak, T. (2012). Critical evaluation of imprinted gene expression by RNA-Seq: a new perspective. PLoS genetics *8*, e1002600.

46. Engreitz, J., Lander, E.S., and Guttman, M. (2015). RNA antisense purification (RAP) for mapping RNA interactions with chromatin. Methods in molecular biology *1262*, 183-197.

47. 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.

48. Ferguson-Smith, A.C. (2011). Genomic imprinting: the emergence of an epigenetic paradigm. Nature reviews Genetics *12*, 565-575.

49. Finn, E.H., Smith, C.L., Rodriguez, J., Sidow, A., and Baker, J.C. (2014). Maternal bias and escape from X chromosome imprinting in the midgestation mouse placenta. Dev Biol *390*, 80-92.

50. Fitzpatrick, G.V., Soloway, P.D., and Higgins, M.J. (2002). Regional loss of imprinting and growth deficiency in mice with a targeted deletion of KvDMR1. Nature genetics *32*, 426-431.

51. Gilad, Y., Rifkin, S.A., and Pritchard, J.K. (2008). Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet *24*, 408-415.

52. Gimelbrant, A., Hutchinson, J.N., Thompson, B.R., and Chess, A. (2007). Widespread monoallelic expression on human autosomes. Science *318*, 1136-1140.

53. Glaser, R.L., Ramsay, J.P., and Morison, I.M. (2006). The imprinted gene and parent-of-origin effect database now includes parental origin of de novo mutations. Nucleic Acids Res *34*, D29-31.

54. Gregg, C., Zhang, J., Butler, J.E., Haig, D., and Dulac, C. (2010a). Sex-specific parent-of-origin allelic expression in the mouse brain. Science *329*, 682-685.

55. Gregg, C., Zhang, J., Weissbourd, B., Luo, S., Schroth, G.P., Haig, D., and Dulac, C. (2010b). High-resolution analysis of parent-of-origin allelic expression in the mouse brain. Science *329*, 643-648.

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

57. 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.

58. 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.

59. 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. Nature structural & molecular biology *21*, 198-206.

60. Hanna, C.W., and Kelsey, G. (2014). The specification of imprints in mammals. Heredity *113*, 176-183.

61. Hasegawa, Y., Brockdorff, N., Kawano, S., Tsutui, K., Tsutui, K., and Nakagawa, S. (2010). The matrix protein hnRNP U is required for chromosomal localization of Xist RNA. Dev Cell *19*, 469-476.

62. Hudson, Q.J., Seidl, C.I., Kulinski, T.M., Huang, R., Warczok, K.E., Bittner, R., Bartolomei, M.S., and Barlow, D.P. (2011). Extra-embryonic-specific imprinted expression is restricted to defined lineages in the post-implantation embryo. Dev Biol *353*, 420-431.

63. Iyer, M.K., Niknafs, Y.S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T.R., Prensner, J.R., Evans, J.R., Zhao, S., *et al.* (2015). The landscape of long noncoding RNAs in the human transcriptome. Nature genetics *47*, 199-208.

64. Kaneda, M., Okano, M., Hata, K., Sado, T., Tsujimoto, N., Li, E., and Sasaki, H. (2004). Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. Nature *429*, 900-903.

65. 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.

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

67. Kornienko, A.E., Dotter, C.P., Guenzl, P.M., Gisslinger, H., Gisslinger, B., Cleary, C., Kralovics, R., Pauler, F.M., and Barlow, D.P. (2016). Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans. Genome Biol *17*, 14.

68. 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.

69. Krietsch, W.K., Fehlau, M., Renner, P., Bucher, T., and Fundele, R. (1986). Expression of X-linked phosphoglycerate kinase in early mouse embryos

homozygous at the Xce locus. Differentiation; research in biological diversity *31*, 50-54.

70. Krietsch, W.K., Fundele, R., Kuntz, G.W., Fehlau, M., Burki, K., and Illmensee, K. (1982). The expression of X-linked phosphoglycerate kinase in the early mouse embryo. Differentiation; research in biological diversity *23*, 141-144.

71. Kulinski, T.M., Barlow, D.P., and Hudson, Q.J. (2013). Imprinted silencing is extended over broad chromosomal domains in mouse extra-embryonic lineages. Current opinion in cell biology *25*, 297-304.

72. Lagarrigue, S., Martin, L., Hormozdiari, F., Roux, P.F., Pan, C., van Nas, A., Demeure, O., Cantor, R., Ghazalpour, A., Eskin, E.*, et al.* (2013). Analysis of allele-specific expression in mouse liver by RNA-Seq: a comparison with Cis-eQTL identified using genetic linkage. Genetics *195*, 1157-1166.

73. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W.*, et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

74. Lappalainen, T., Sammeth, M., Friedlander, M.R., t Hoen, P.A., Monlong, J., Rivas, M.A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G.*, et al.* (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506-511.

75. Latos, P.A., Pauler, F.M., Koerner, M.V., Senergin, 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.

76. Lee, J., Inoue, K., Ono, R., Ogonuki, N., Kohda, T., Kaneko-Ishino, T., Ogura, A., and Ishino, F. (2002). Erasing genomic imprinting memory in mouse clone embryos produced from day 11.5 primordial germ cells. Development *129*, 1807-1817.

77. Lee, J.T., and Lu, N. (1999). Targeted mutagenesis of Tsix leads to nonrandom X inactivation. Cell *99*, 47-57.

78. Lerchner, W., and Barlow, D.P. (1997). Paternal repression of the imprinted mouse Igf2r locus occurs during implantation and is stable in all tissues of the post-implantation mouse embryo. Mechanisms of development *61*, 141-149.

79. Leung, D., Jung, I., Rajagopal, N., Schmitt, A., Selvaraj, S., Lee, A.Y., Yen, C.A., Lin, S., Lin, Y., Qiu, Y.*, et al.* (2015). Integrative analysis of haplotype-resolved epigenomes across human tissues. Nature *518*, 350-354.

80. Li, E., Beard, C., and Jaenisch, R. (1993). Role for DNA methylation in genomic imprinting. Nature *366*, 362-365.

81. Li, G., Bahn, J.H., Lee, J.H., Peng, G., Chen, Z., Nelson, S.F., and Xiao, X. (2012). Identification of allele-specific alternative mRNA processing via transcriptome sequencing. Nucleic Acids Res *40*, e104.

82. Lin, S.P., Youngson, N., Takada, S., Seitz, H., Reik, W., Paulsen, M., Cavaille, J., and Ferguson-Smith, A.C. (2003). Asymmetric regulation of imprinting on the maternal and paternal chromosomes at the Dlk1-Gtl2 imprinted cluster on mouse chromosome 12. Nature genetics *35*, 97-102.

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

84. Lyon, M.F. (1961). Gene action in the X-chromosome of the mouse (Mus musculus L.). Nature *190*, 372-373.

85. Majewski, J., and Pastinen, T. (2011). The study of eQTL variations by RNA-seq: from SNPs to phenotypes. Trends Genet *27*, 72-79.

86. Mak, W., Nesterova, T.B., de Napoles, M., Appanah, R., Yamanaka, S., Otte, A.P., and Brockdorff, N. (2004). Reactivation of the paternal X chromosome in early mouse embryos. Science *303*, 666-669.

87. Mancini-Dinardo, D., Steele, S.J., 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 & development *20*, 1268-1282.

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

89. Marks, H., Kerstens, H.H., Barakat, T.S., Splinter, E., Dirks, R.A., van Mierlo, G., Joshi, O., Wang, S.Y., Babak, T., Albers, C.A.*, et al.* (2015). Dynamics of gene silencing during X inactivation using allele-specific RNA-seq. Genome Biol *16*, 149.

90. Mazo, A., Hodgson, J.W., Petruk, S., Sedkov, Y., and Brock, H.W. (2007). Transcriptional interference: an unexpected layer of complexity in gene regulation. Journal of cell science *120*, 2755-2761.

91. McGrath, J., and Solter, D. (1984). Completion of mouse embryogenesis requires both the maternal and paternal genomes. Cell *37*, 179-183.

92. McHugh, C.A., Chen, C.K., Chow, A., Surka, C.F., Tran, C., McDonel, P., Pandya-Jones, A., Blanco, M., Burghard, C., Moradian, A.*, et al.* (2015). The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. Nature *521*, 232-236.

93. Moindrot, B., Cerase, A., Coker, H., Masui, O., Grijzenhout, A., Pintacuda, G., Schermelleh, L., Nesterova, T.B., and Brockdorff, N. (2015). A Pooled shRNA Screen Identifies Rbm15, Spen, and Wtap as Factors Required for Xist RNA-Mediated Silencing. Cell reports *12*, 562-572.

94. Monfort, A., Di Minin, G., Postlmayr, A., Freimann, R., Arieti, F., Thore, S., and Wutz, A. (2015). Identification of Spen as a Crucial Factor for Xist Function

through Forward Genetic Screening in Haploid Embryonic Stem Cells. Cell reports *12*, 554-561.

95. Morgan, H.D., Santos, F., Green, K., Dean, W., and Reik, W. (2005). Epigenetic reprogramming in mammals. Human molecular genetics *14 Spec No 1*, R47-58.

96. Mouse Genome Sequencing, C., Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M.*, et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520-562.

97. 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.

98. Ohhata, T., and Wutz, A. (2013). Reactivation of the inactive X chromosome in development and reprogramming. Cell Mol Life Sci *70*, 2443-2461.

99. Okae, H., Hiura, H., Nishida, Y., Funayama, R., Tanaka, S., Chiba, H., Yaegashi, N., Nakayama, K., Sasaki, H., and Arima, T. (2012). Re-investigation and RNA sequencing-based identification of genes with placenta-specific imprinted expression. Human molecular genetics *21*, 548-558.

100. Okamoto, I., Otte, A.P., Allis, C.D., Reinberg, D., and Heard, E. (2004). Epigenetic dynamics of imprinted X inactivation during early mouse development. Science *303*, 644-649.

101. Ooi, S.K., O'Donnell, A.H., and Bestor, T.H. (2009). Mammalian cytosine methylation at a glance. Journal of cell science *122*, 2787-2791.

102. 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.

103. Penny, G.D., Kay, G.F., Sheardown, S.A., Rastan, S., and Brockdorff, N. (1996). Requirement for Xist in X chromosome inactivation. Nature *379*, 131-137.

104. Peters, J. (2014). The role of genomic imprinting in biology and disease: an expanding view. Nature reviews Genetics *15*, 517-530.

105. Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., Veyrieras, J.B., Stephens, M., Gilad, Y., and Pritchard, J.K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature *464*, 768-772.

106. 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.

107. Prickett, A.R., and Oakey, R.J. (2012). A survey of tissue-specific genomic imprinting in mammals. Mol Genet Genomics *287*, 621-630.

108. Proudhon, C., and Bourc'his, D. (2010). Identification and resolution of artifacts in the interpretation of imprinted gene expression. Brief Funct Genomics *9*, 374-384.

109. Proudhon, C., Duffie, R., Ajjan, S., Cowley, M., Iranzo, J., Carbajosa, G., Saadeh, H., Holland, M.L., Oakey, R.J., Rakyan, V.K.*, et al.* (2012). Protection against de novo methylation is instrumental in maintaining parent-of-origin methylation inherited from the gametes. Molecular cell *47*, 909-920.

110. 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*, D756-763.

111. Quinn, J.J., and Chang, H.Y. (2015). Unique features of long non-coding RNA biogenesis and function. Nature reviews Genetics *17*, 47-62.

112. Rastan, S. (1983). Non-random X-chromosome inactivation in mouse X-autosome translocation embryos--location of the inactivation centre. Journal of embryology and experimental morphology *78*, 1-22.

113. Reinius, B., and Sandberg, R. (2015). Random monoallelic expression of autosomal genes: stochastic transcription and allele-level regulation. Nature reviews Genetics *16*, 653-664.

114. Rockman, M.V., and Kruglyak, L. (2006). Genetics of global gene expression. Nature reviews Genetics *7*, 862-872.

115. Santoro, F., Mayer, D., Klement, R.M., Warczok, K.E., Stukalov, A., Barlow, D.P., and Pauler, F.M. (2013). Imprinted Igf2r silencing depends on continuous Airn lncRNA expression and is not restricted to a developmental window. Development *140*, 1184-1195.

116. Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G.*, et al.* (2003). Genetics of gene expression surveyed in maize, mouse and man. Nature *422*, 297-302.

117. Shiura, H., Nakamura, K., Hikichi, T., Hino, T., Oda, K., Suzuki-Migishima, R., Kohda, T., Kaneko-ishino, T., and Ishino, F. (2009). Paternal deletion of Meg1/Grb10 DMR causes maternalization of the Meg1/Grb10 cluster in mouse proximal Chromosome 11 leading to severe pre- and postnatal growth retardation. Human molecular genetics *18*, 1424-1438.

118. 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.

119. Stoger, R., Kubicka, P., Liu, C.G., Kafri, T., Razin, A., Cedar, H., and Barlow, D.P. (1993). Maternal-specific methylation of the imprinted mouse Igf2r locus identifies the expressed locus as carrying the imprinting signal. Cell *73*, 61-71.

120. Stringer, J.M., Pask, A.J., Shaw, G., and Renfree, M.B. (2014). Post-natal imprinting: evidence from marsupials. Heredity *113*, 145-155.

121. Sugimoto, M., and Abe, K. (2007). X chromosome reactivation initiates in nascent primordial germ cells in mice. PLoS genetics *3*, e116.

122. Surani, M.A., Barton, S.C., and Norris, M.L. (1984). Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis. Nature *308*, 548-550.

123. Szabo, P.E., and Mann, J.R. (1995a). Allele-specific expression and total expression levels of imprinted genes during early mouse development: implications for imprinting mechanisms. Genes & development *9*, 3097-3108.

124. Szabo, P.E., and Mann, J.R. (1995b). Biallelic expression of imprinted genes in the mouse germ line: implications for erasure, establishment, and mechanisms of genomic imprinting. Genes & development *9*, 1857-1868.

125. Takagi, N., and Sasaki, M. (1975). Preferential inactivation of the paternally derived X chromosome in the extraembryonic membranes of the mouse. Nature *256*, 640-642.

126. Thorvaldsen, J.L., Duran, K.L., and Bartolomei, M.S. (1998). Deletion of the H19 differentially methylated domain results in loss of imprinted expression of H19 and Igf2. Genes & development *12*, 3693-3702.

127. Tran, D.A., Bai, A.Y., Singh, P., Wu, X., and Szabo, P.E. (2014). Characterization of the imprinting signature of mouse embryo fibroblasts by RNA deep sequencing. Nucleic Acids Res *42*, 1772-1783.

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

129. Wang, X., and Clark, A.G. (2014). Using next-generation RNA sequencing to identify imprinted genes. Heredity *113*, 156-166.

130. Wang, X., Soloway, P.D., and Clark, A.G. (2011). A survey for novel imprinted genes in the mouse placenta by mRNA-seq. Genetics *189*, 109-122.

131. Wang, X., Sun, Q., McGrath, S.D., Mardis, E.R., Soloway, P.D., and Clark, A.G. (2008). Transcriptome-wide identification of novel imprinted genes in neonatal mouse brain. PLoS One *3*, e3839.

132. West, J.D., Frels, W.I., Chapman, V.M., and Papaioannou, V.E. (1977). Preferential expression of the maternally derived X chromosome in the mouse yolk sac. Cell *12*, 873-882.

133. Williamson CM, B.A., Thomas S, Beechey CV, Hancock J, Cattanach BM, and Peters J (2013). World Wide Web Site - Mouse Imprinting Data and References.

134. Williamson, C.M., Turner, M.D., Ball, S.T., Nottingham, W.T., Glenister, P., Fray, M., Tymowska-Lalanne, Z., Plagge, A., Powles-Glover, N., Kelsey, G.*, et al.* (2006). Identification of an imprinting control region affecting the expression of all transcripts in the Gnas cluster. Nature genetics *38*, 350-355.

135. Wutz, A. (2011). Gene silencing in X-chromosome inactivation: advances in understanding facultative heterochromatin formation. Nature reviews Genetics *12*, 542-553.

136. Wutz, A., and Jaenisch, R. (2000). A shift from reversible to irreversible X inactivation is triggered during ES cell differentiation. Molecular cell *5*, 695-705.

137. Wutz, A., Rasmussen, T.P., and Jaenisch, R. (2002). Chromosomal silencing and localization are mediated by different domains of Xist RNA. Nature genetics *30*, 167-174.

138. Wutz, A., Smrzka, O.W., Schweifer, N., Schellander, K., Wagner, E.F., and Barlow, D.P. (1997). Imprinted expression of the Igf2r gene depends on an intronic CpG island. Nature *389*, 745-749.

139. Wutz, A., Theussl, H.C., Dausman, J., Jaenisch, R., Barlow, D.P., and Wagner, E.F. (2001). Non-imprinted Igf2r expression decreases growth and rescues the Tme mutation in mice. Development *128*, 1881-1887.

140. Xie, W., Barr, C.L., Kim, A., Yue, F., Lee, A.Y., Eubanks, J., Dempster, E.L., and Ren, B. (2012). Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. Cell *148*, 816-831.

141. Yamasaki, Y., Kayashima, T., Soejima, H., Kinoshita, A., Yoshiura, K., Matsumoto, N., Ohta, T., Urano, T., Masuzaki, H., Ishimaru, T.*, et al.* (2005). Neuron-specific relaxation of Igf2r imprinting is associated with neuron-specific histone modifications and lack of its antisense transcript Air. Human molecular genetics *14*, 2511-2520.

142. Yang, F., Babak, T., Shendure, J., and Disteche, C.M. (2010). Global survey of escape from X inactivation by RNA-sequencing in mouse. Genome Res *20*, 614-622.

143. Zwart, R., Sleutels, F., Wutz, A., Schinkel, A.H., and Barlow, D.P. (2001). Bidirectional action of the Igf2r imprint control element on upstream and downstream imprinted genes. Genes & development *15*, 2361-2366.

144. Zwemer, L.M., Zak, A., Thompson, B.R., Kirby, A., Daly, M.J., Chess, A., and Gimelbrant, A.A. (2012). Autosomal monoallelic expression in the mouse. Genome Biol *13*, R10.