GENOMIC IMPRINTING AND X CHROMOSOME INACTIVATION IN THE MOUSE

A Dissertation
Presented to the Faculty of the Graduate School
of Cornell University
In Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by
Xu Wang
August 2011

GENOMIC IMPRINTING AND X CHROMOSOME INACTIVATION IN THE MOUSE

Xu Wang, Ph. D.
Cornell University 2011

Genomic imprinting is a special form of epigenetic modification of the genome in which gene expression differs in an allele-specific manner depending on the parent-of-origin. The degree of imprinting is often tissue- and/or developmental stage-specific, and may be altered in some diseases including cancer. To date, 99 genes have been shown to undergo genomic imprinting in mouse, and 60 are imprinted in humans, with an overlapping set of 43 imprinted in both species. This list is far from complete, and obtaining an exhaustive identification of imprinted genes would expand our understanding of the regulation and evolution of the phenomenon. To search for novel imprinted genes, I applied custom SNP microarray and Illumina mRNA sequencing technologies to the transcriptomes of reciprocal F1 mouse brains and placentas. In brain, I identified 26 genes with parent-of-origin dependent differential allelic expression. Pyrosequencing verified 17 of them, including three novel imprinted genes. In placenta, I confirmed the imprinting status of 23 known imprinted genes, and found that 12 genes reported previously to be imprinted in other tissues are also imprinted in placenta. Through a well-replicated design using an orthogonal allelic-expression technology, I verified five novel imprinted genes that were not previously known to be imprinted in mouse. After repeated application to multiple tissues and

developmental stages this approach will yield a complete catalog of imprinted genes, shedding light on the mechanism and evolution of imprinted genes and diseases associated with genomic imprinting.

X-inactivation in female eutherian mammals has long been considered to occur at random in embryonic and postnatal tissues. After RNA-seq data revealed what appeared to be a chromosome-wide bias toward under-expression of paternal alleles in mouse tissue, I applied pyrosequencing to mouse brain cDNA samples from reciprocal cross F1 progeny of divergent strains, and found a small but consistent and highly statistically significant tendency to under-express the paternal X chromosome. Allelic bias in expression is also influenced by the sampling effect of X inactivation and by *cis*-acting regulatory variation, and for each gene we quantified the contributions of these effects in two different strain combinations while controlling for variability in *Xce* alleles.

BIOGRAPHICAL SKETCH

Xu Wang was born on August 28, 1981, in the city of Jinan, Shandong province in China. After he graduated from Shandong Experimental High School in 2000, he went to Shanghai and studied biological sciences in Department of Biological Sciences at Fudan University. He worked with Dr. Li (Felix) Jin and was involved in two projects: genetic linkage analysis and association study of human anthropometric traits, and characterization of genetic structure and migration patterns of Chinese and East Asia populations. After receiving his B.S. degree in 2004, he became a member of Bioinformatics Research Group of Chinese National Human Genome Center at Shanghai (CHGC), analyzing the SNP genotyping data of the phase I HapMap Project from the Illumina platform. In 2005, he went to graduate school in the field of Genetics and Development at Cornell University, majoring genetics and minoring biometry and computational biology, under the direction of Dr. Andy Clark. In grad school, he had three years of teaching experience as a teaching assistant for the Human Genomic course. With Dr. Andy Clark and many collaborators, he compared the human linkage map and the population recombination rate, and investigated gene expression, genomic imprinting, random and imprinted X inactivation, *cis*-eQTL effect, hybrid dysregulation and whole genome methylation patterns in various organisms including human, mouse, horse, donkey, mule, hinny, chicken, honeybee and nasonia, using next generation sequencing and allele-specific pyrosequencing technologies. Upon completion of his Ph.D. study in 2011, he will become a Postdoctoral Research Associate with Dr. Andy Clark.

To Liande Wang, Ruhuan Su and Xianfei Sun

ACKNOWLEDGMENTS

First of all I would like to thank my advisor Dr. Andy Clark for his advice and guidance for all my projects, for his trust and patience when I met difficulties in my experiments and for his support both financially and intellectually. He provided the opportunity for me to conduct genetic and genomic research with cutting edge technologies in his laboratory. I learned a lot from him and I cannot finish my Ph. D. study without him. I thank all my other thesis commit members, Drs. Adam Siepel, Dan Barbash and Jason Mezey, for their help, advice and suggestions, and their time to read and review the thesis manuscript. I also thank all the past and current members of the Clark Lab for being both great colleagues and friends over the six years, especially Jian Li, Jian Lu, Xiaoyun Wang, Li Chi and Amanda Manfredo, for their discussion and help with my experiments.

I wish to thank Dr. Peter Schweitzer and Tom Stelick from Cornell sequencing core facility, Dr. Wei Wang, Jie Zhao and Ling Zhang from Cornell Microarray core facility, Dr. Jenny Xiang, Diana Shao and Jake Kresovich from Cornell Weill medical school DNA sequencing core facility, and Dr. Qi Sun, Jarek Pillardy, Robert Bukowski and Lalit Ponnala from Cornell Bioinformatics Service Unit, for their help with microarray and next generation sequencing experiments and data analysis.

I would like to thank our collaborators and members of their laboratories for providing help, materials, advice and discussions indispensable to all my projects including the work presented in my thesis: Dr. Paul Soloway, Dr. Yoon Jung Park and Jim Putnam for help with the mouse crosses and dissections; Dr. Paula Cohen for collecting mouse oocytes; Dr. John Schimenti for providing 129-BALB hybrid mouse ES-cell lines; Dr.

Doug Antczak, Don Miller and Becky Harman from Baker Institute, for providing and dissecting the horse, donkey, mule and hinny samples and their great work for the equids genomic imprinting and X inactivation projects; Dr. Monique Ho from University of Rochester Medical Center and Drs. Mark Roberson, Li Han and Jianjun Xie for providing the human term placenta samples; Dr. Sue Lamont and Erin Sandford from Iowa State University for providing the chicken brain samples; Dr. Jack Werren, Dr. Dave Wheeler and Rachel Edwards from University of Rochester for the nasonia projects and Dr. Gene Robinson from University of Illinois at Urbana-Champaign for the honeybee projects.

Last but not least I'd like to thank my wife Xianfei Sun, and all my family and friends for their support and encouragement during my Ph. D. study.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

3’UTR – 3’ untranslated region

ANOVA – analysis of variance

CBSU – Cornell Computational Biology Service Unit

cDNA – complementary deoxyribonucleic acid

CEPH – Centre d'Etude du Polymorphisme Humain

cRNA – complementary ribonucleic acid

DMR – differentially methylated regions

DNA – deoxyribonucleic acid

E – embryonic day

eQTL – expression quantitative trait loci

ES – embryonic stem cell

FDR – False discovery rate

Gbp – giga-basepairs

gDNA – genomic DNA

kb – kilo-base

kbp – kilo-basepairs

LCL – lymphoblastoid cell line

Mbp – mega-basepairs

MGI – Mouse Genome Informatics

mRNA – messenger ribonucleic acid

MSCI – meiotic sex chromosome inactivation

NIEHS – National Institute of Environmental Health Sciences

P – post-natal day

PAR – pseudo-autosomal region

PCR – polymerase chain reaction

RIN – RNA integrity number

RNA – ribonucleic acid

RPKM – Reads Per Kilobase of exon model per Million mapped reads

snoRNA – small nucleolar ribonucleic acid

SNP – single nucleotide polymorphism

TU – transcription unit

UPD – uniparental disomy

UTR – untranslated region

WCTC – Wellcome Trust Centre for human genetics

Xic – X inactivation center

Xm – maternally inherited X chromosome

Xp – paternally inherited X chromosome

CHAPTER 1

Testing the imprinting status of candidate imprinted genes using custom SNP arrays in mouse neonatal brain

Abstract

Imprinted genes are a subset of genes that are expressed in a parent-of-origin dependent manner. To date 99 genes have been shown to undergo imprinting in mouse, and 60 genes are imprinted in humans, with only 43 of these genes being imprinted in both species (http://igc.otago.ac.nz/home.html). We know that these lists are incomplete. Genome-wide screens for imprinted genes in humans have suffered from insufficient sample size to get reciprocal transmission of the two parental alleles and lack of appropriate tissue for assay (clearest cases of imprinting occur in fetal brain and placenta). Our goal is to discover novel imprinted genes in mouse and test the imprinting status of their human orthologs. I chose two mouse strains among the 15 with genome sequence information and extracted total RNA from post-natal day 2 brains of both parents and the reciprocal F1s. 181 genes were selected as candidate genes from 600 putative mouse imprinted genes predicted by Leudi *et al.* 2005. Additional candidates were selected among orthologs of 63 potentially imprinted genes in humans (identified by uniparental expression in collaboration with Perlegen). Then I designed 25-mer oligo probes for a total of 396 exonic SNPs in the test genes and had a microarray prepared by Agilent. By hybridizing the Cy3 and Cy5 labeled cRNA made from the RNA samples to this microarray, the relative abundance of the two SNP alleles was quantified in the two reciprocal F1 strains. The data analysis identifies 10 candidate imprinted genes, all of which have lower fold-change compared to the known imprinted gene controls. I verified novel candidates in mouse

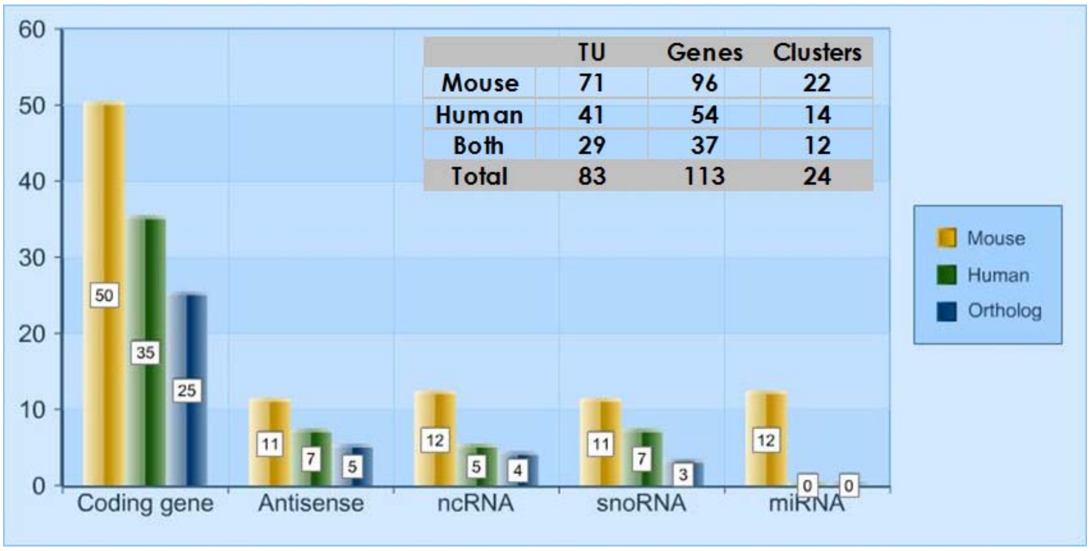
using pyrosequencing method, and found none of them were imprinted. Given that I observed expected results for the positive and negative control genes, I conclude that the computation prediction method has very little power and extremely high false positive rate for discovering novel imprinted genes.

Introduction

Genomic imprinting is a remarkable epigenetic phenomenon. At a small number of mammalian loci, only one of the two alleles is expressed, or one of the two alleles is predominantly expressed. The gist is that gene expression depends on the sex of the transmitting parent. These genes are called imprinted genes. Some inheritable molecular imprints, for example, differential methylation, or histone modification, are established on the paternal or maternal allele to repress the expression of one allele. The imprinting status of an imprinted gene is often tissue specific and/or developmental stage specific, and may be altered in human diseases including cancer.

In 2005, there are 96 imprinted genes in mouse genome, including non-coding transcript, snoRNA and miRNA, which belong to 83 transcriptional units defined by Marison *et al.* 2005 (Figure 1). 54 of them are maternally expressed and 42 of them show paternal specific expression. In the human genome, there are 53 known imprinted genes which belong to 41 transcriptional units. 19 of them are only maternally expressed and 34 of them show paternal specific expression. 37 genes (29 transcriptional units) are imprinted in both human and mouse. 1 of the 37 genes, *Zim2*, is reported to be oppositely imprinted in human and mouse. It is estimated that about 1% of the genes in human genome are imprinted, which is around 200 genes (Murphy and Jirtle 2003). So there are still 100 imprinted genes to be discovered.

Figure 1. Summary of known imprinted genes in human and mouse. Data from Morison et al. 2005. TU: transcription units. The term TU has been defined as a group of transcripts that contain a common core of genetic information having the same orientation, which does not necessarily correspond to protein-coding regions.



	TU	Genes	Clusters
Mouse	71	96	22
Human	41	54	14
Both	29	37	12
Total	83	113	24

Using a machine learning approach, Luedi *et al.* 2005 predicted 600 genes out of a total of 23,788 annotated autosomal genes in mouse to be imprinted (2.5%); 384 (64%) of these candidate imprinted genes were predicted to exhibit maternal expression. Here, to identify novel imprinted genes, I selected a subset of the predicted imprinted genes, designed a custom SNP microarray and hybridized to cRNAs from reciprocal F1 mouse brain samples of two different cross combinations. 10 candidate genes from the microarray results were chosen for allele-specific pyrosequencing verification and none of them are imprinted. The results suggest that the computation prediction method has very little power and suffers from extremely high false positive rate.

Materials and Methods

Mouse Strain and tissue selection

Because most of the imprinted genes are imprinted and expressed in brain and placenta, I choose the neonatal brain as the testing tissue. Two mouse strain combinations are selected to maximize the 600 predicted imprinted genes covered by our study. Four mouse strains (C57BL/6, C3H/HeJ, AKR/J, PWD/PhJ) were purchased from the Jackson Laboratory (www.jax.org). I performed two pairs of mouse reciprocal crosses (C57BL/6 x C3H/HeJ, C3H/HeJ x C57BL/6, AKR/J x PWD/PhJ, PWD/PhJ x AKR/J). Total RNA samples were extracted from the P2 F1 mouse whole brains using the Qiagen RNeasy Mini Kit. RNA concentrations and A_{260} nm/A_{280} nm ratios were checked with a NanoDrop ND-1000 Spectrophotometer. RNA integrity was checked using the Agilent 2100 Bioanalyzer. All of the samples have a RIN (RNA integrity number) of 10.

All procedures involving mice have been approved by the Institutional Animal Care and Use Committee at Cornell University (protocol number 2002-0075, approved for three years beginning 01/27/2006). Cornell University is accredited by AAALAC.

Selection of candidate and positive/negative control genes

I selected genes from the 600 predicted imprinted gene list with SNPs in the transcripts between the two strain combinations. The SNP information was downloaded from three mouse SNP databases. The first is the Mouse Genome

Informatics (MGI) SNP database from Jackson lab. It has 6,348,627 SNPs for a total of 87 commonly used mouse strains, including the four strains I selected. The second is Wellcome Trust Centre for human genetics (WCTC) SNP database, which has 13,370 genotyped SNPs for 480 strains uniformly distributed across the mouse genome. The third one is NIEHS & Perlegen SNP database, with 8,322,543 SNPs in 15 non-reference mouse strains (Frazer et al. 2007). A second set of candidate genes were selected from the results of a survey for novel imprinted genes in human (Pollard et al. 2008). In collaboration with Drs. Katie Pollard and Kelly Frazer, we did a genome-wide scan of expression level of paternal and maternal alleles by hybridizing both gDNA and cDNA to Perlegen resequencing arrays, which contain 7109 coding SNPs in 68 human cultured lymphoblastoid cell line samples from the CEPH panels. The mouse orthologs of human candidate imprinted genes from this study were also included in our array design. Because oligo dT primer is used in the cDNA synthesis step, all selected genes have at least one SNP within 800bp of 3'-UTR region. To make sure that there is enough expression level for selected genes in the tissues we used, I checked the Affymetrix Mouse Genome 430 v2.0 GeneChip array data from NCBI Gene Expression Omnibus published by Lindsley RC and Murphy KM (Nov. 2004, GSE1986).

Known paternally and maternally expressed imprinted genes in mouse were selected as positive control genes. Non-imprinted genes are needed to be chosen as negative control genes. There are more than 300 genes in mouse genome for which homozygous transgenic mutation is lethal. If such a gene is imprinted in mouse, it has monoallelic expression from one parent, therefore, the heterozygous mutation inherited from this parent in the inbred strain background will also be lethal. I examined the phenotype description of all 300 genes and selected those with clear description that there is no observable difference between heterozygotes and wild types. The gene selection is summarized in (Table 1).

Microarray probe design and experiment design

To quantify the allelic expression ratios for target genes, we need to hybridize the cRNA to probes that could distinguish the paternal and maternal alleles. The array is designed as a combination of an expression array and a SNP genotyping array. I used 25 nucleotides oligos as probes to the target sequence. The SNP position is located in the middle of the probe, or off by 1bp depending on the probe quality. Instead of using staggering probes, we decided to target multiple SNPs per gene. We will have higher confidence if more than one SNPs show different allelic expression signal. Because mismatch probes often have some signal, I included all 4 nucleotides (A, G, C and T) at the SNP position, with the three mismatched oligos as negative controls. The criteria for the probe design are as follows: the probe is 25 nucleotide long and it is within 3'end 800bp region; the Tm range of the probes is 62±6 °C; the maximum hairpin $\Delta G < -4$ kcal/mol, and the maximum self dimer $\Delta G < -7$ kcal/mol; the maximum homopolymer repeat nucleotide length ≤ 4. I then blasted the target sequence (201bp) and probe sequence against the mouse RefSeq database to make sure that they only have unique hit. Since we are using cRNAs for hybridization, the probe should be on the sense strand.

Table 1. Summary of gene and SNP selection for Agilent array design.

	Total #(genes)	Genes covered	SNP selected
600 genes list	600	274	773
Perlegen	56	31	73
Positive	96	23	81
Negative	85	48	114
Total	837	376	1041

For the microarray experiment design, I selected two strain combinations (AKR-PWD and B6-C3H) and did reciprocal crosses for these strains. In each strain combination, I labeled one paternal strain with Cy3 and the other with Cy5, and hybridized the cRNA from them to one array. I also labeled the reciprocal F1s with Cy3 and Cy5 respectively, and hybridized them to one array. Dye swaps were included to control the dye effect. In total, there are 8 arrays with 1,900 probes each on a single slide (Figure 2 and Figure 3).

Labeling and hybridization

cDNA samples were synthesized using MMLV RT in Agilent Low RNA Input Linear Amplification Kit PLUS, Two-Color (Cat. No. 5188-5340). Positive and negative control RNA mix from Agilent RNA Spike-In Kit (Cat. No. 5188-5279) was included in the starting total RNA samples. Then cRNAs were synthesized from the cDNA as template and labeled with Cy3 or Cy5. Labeled cRNAs were purified and quantified, before hybridized to the Agilent microarray. The hybridization and washing conditions were from the manufacture's protocols. The slide was then dried and scanned using GenePix 4000B scanner at 5μm resolution.

Microarray data analysis

The background subtraction of the probe intensities was done in GenePix software. Then I used R package `smida` for spatial and dye normalization. Custom R script was used for spike-in normalization. After normalization, overall the intensity of perfect match probes is higher than the mismatch probe. Probes with a normalized intensity greater than 30 were defined as informative probes. The SNPs with at least 6 (out of 8) informative perfect match probes are classified as informative SNPs. About 40% of all the SNPs are informative.

Pyrosequencing verifications

Pyrosequencing primers were designed for positive control genes and the 10 candidate imprinted genes. PCR products for Pyrosequencing were amplified using biotin labeled forward (or reverse) primer. The Pyrosequencing was done on a PSQ™ 96 MA Pyrosequencer (Biotage, AB) with the Pyro Gold Reagents (Biotage, AB). The relative level of the two parental alleles was quantified by the software for PSQ™ 96 MA Pyrosequencer (Version 2.02 RC 5.8, Biotage, AB) using the allele quantification method.

Figure 2. Custom microarray experimental design.

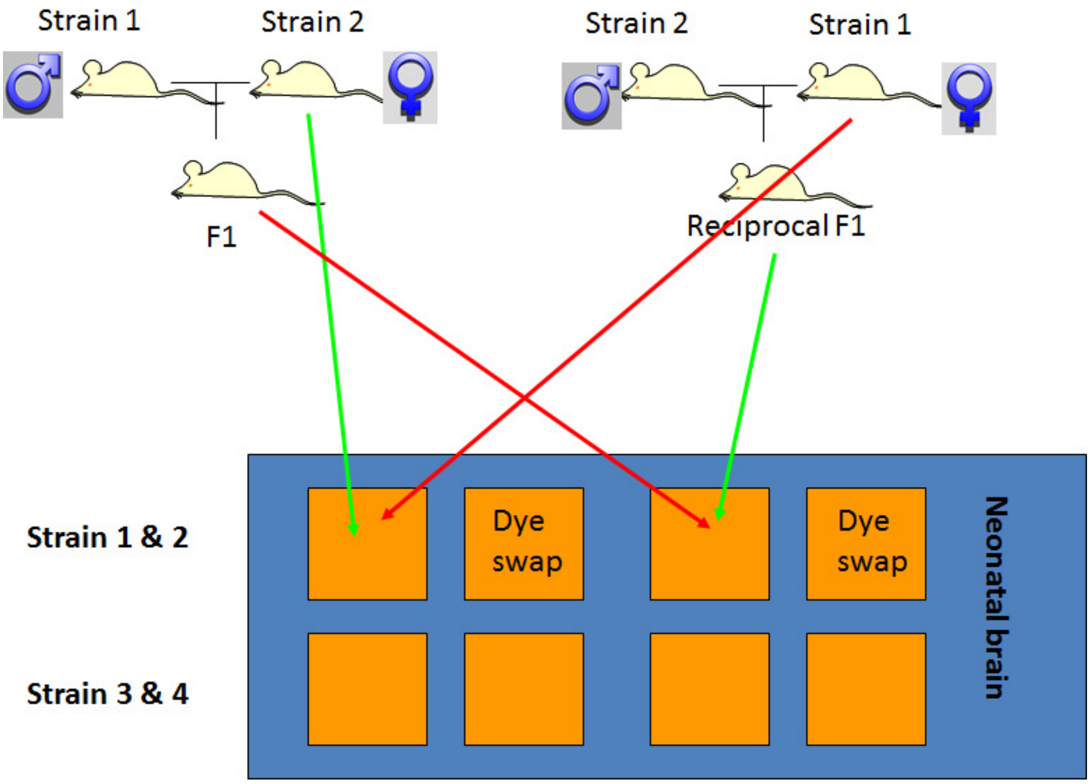
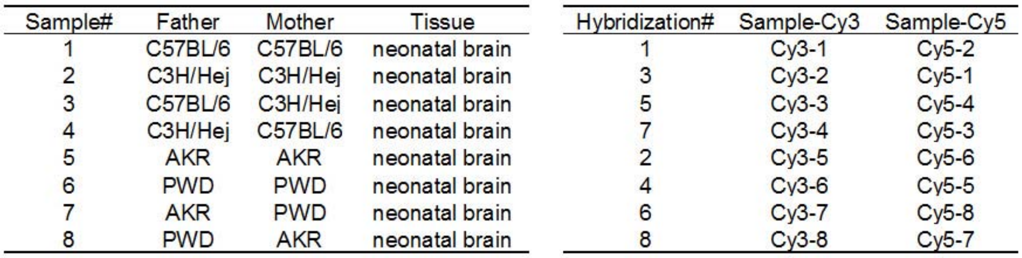
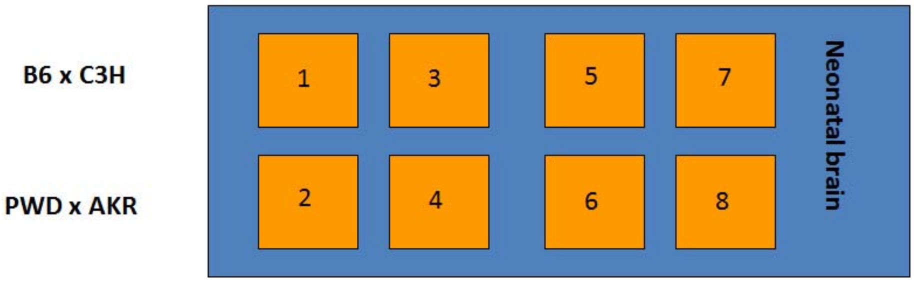


Figure 3. Sample information and microarray hybridization. Sample-Cy3 is the cRNA sample labeled with Cy3, and Sample-Cy5 indicates the cRNA samples labeled with Cy5.



Sample#	Father	Mother	Tissue
1	C57BL/6	C57BL/6	neonatal brain
2	C3H/Hej	C3H/Hej	neonatal brain
3	C57BL/6	C3H/Hej	neonatal brain
4	C3H/Hej	C57BL/6	neonatal brain
5	AKR	AKR	neonatal brain
6	PWD	PWD	neonatal brain
7	AKR	PWD	neonatal brain
8	PWD	AKR	neonatal brain

Hybridization#	Sample-Cy3	Sample-Cy5
1	Cy3-1	Cy5-2
3	Cy3-2	Cy5-1
5	Cy3-3	Cy5-4
7	Cy3-4	Cy5-3
2	Cy3-5	Cy5-6
4	Cy3-6	Cy5-5
6	Cy3-7	Cy5-8
8	Cy3-8	Cy5-7



Results

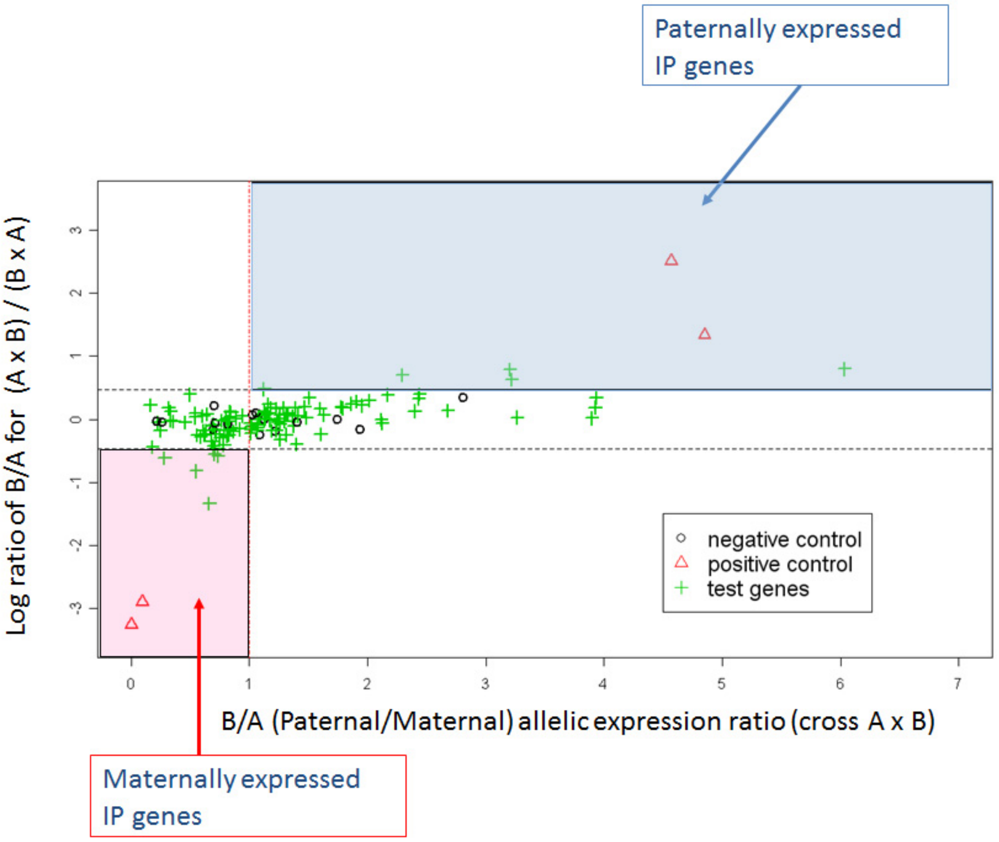
Identification of candidate imprinted genes from microarray data

After normalization, the relative fold change of the allele-specific probe intensity in the two reciprocal F1 crosses is plotted (Figure 4). From the plot, we observed that for the non-imprinted negative control genes, the fold differences are less than 1.5 if the reference and alternative allele probes are flipped (one is greater than 1.0 and one is less than 1.0). If the two probes show the same direction (both are greater than 1.0 or both are less than 1.0), the fold change is less than 2.2 (Table 2), which is consistent with the fact that they are not imprinted. Based on the results of negative control genes, I draw an arbitrary cutoff to identify candidate imprinted genes for verification.

Table 2. Fold change of negative control genes.

Gene	SNP_ID	reciprocal F1 ratio		Fold of difference
		Probe1	Probe2	
ENSMUSG00000025793	37238091	0.712	0.820	1.152
ENSMUSG00000025793	37238095	0.698	1.029	1.474
ENSMUSG00000028717	37697710	1.091	1.977	1.811
ENSMUSG00000026193	47460976	0.817	0.974	0.839
ENSMUSG00000029910	47963016	2.812	1.281	2.196
ENSMUSG00000024909	50578926	1.935	2.849	1.473
ENSMUSG00000019979	50888995	0.584	0.948	1.622
ENSMUSG00000019979	50888997	1.061	0.854	1.242
ENSMUSG00000000555	52584131	1.219	1.912	1.568
ENSMUSG00000041324	52620099	1.748	1.777	1.016
ENSMUSG00000055254	52641889	1.026	0.882	1.163
ENSMUSG00000054191	60643738	1.127	1.165	0.967
ENSMUSG00000004565	61076766	0.703	0.434	1.619

Figure 4. Plot of Agilent microarray results. Plotted on the *x*-axis is the paternal/maternal expression ratio in AxB cross (A is the mother). Plotted on y-axis is the $\log_2$ ratio of the ratio of B/A in the two reciprocal F1s.



Paternally expressed IP genes

Maternally expressed IP genes

Verification of positive control and candidate imprinted genes

Two positive control genes (*Meg3* and *Peg5*) showed clear imprinting pattern with fold change of 300-1800 (Table 3). *Peg5* (also known as *Nnat*) is a known imprinted genes expressed only from paternal allele in mouse neonatal brain (Kagitani et al. 1997; Kikyo et al. 1997). We correctly detected this gene as imprinted in our microarray. *Meg3* (also known as Gtl2) is a known maternally expressed imprinted genes but its imprinting status in neonatal brain is not clear (Miyoshi et al. 2000; Schmidt et al. 2000; Takada et al. 2000). I found that *Meg3* is imprinted in P2 neonatal brain. Pyrosequencing primers were designed for positive control genes, and both of them were confirmed to be imprinted (Figure 5 and Figure 6).

From the 150 informative target genes on the microarray, with the arbitrary cutoff, I identified 10 candidate imprinted genes (Table 4). The fold of change of these genes is not as high as the positive control genes (Table 5). Then I used allele-specific pyrosequencing to verify them. Five of the candidate genes (*B230120H23Rik*, *E030037J05Rik*, *Ednra*, *Prkar2b* and *Shrm*) showed roughly equal expression from both parental alleles. The others showed an expression QTL effect with higher expression level in the AKR (or C3H) strain.

We did not identify novel imprinted genes because we only tested about 150 genes and the bioinformatics support is relative weak. But the microarray approach worked perfectly and it can detect known imprinted genes. Also, pyrosequencing is an accurate way of quantifying allele-specific expression with 1-2% confidence interval of the ratio.

Table 3. Microarray results for the positive control genes *Peg5* and *Meg3*.

Gene	SNP_ID	Reciprocal F1 ratios		Fold of differenc	Exp allele
		Probe1	Probe2		
Nnat (*peg5*)	37546101	4.850	0.224	21.648	P
	37546102	4.566	0.014	323.705	P
	37546103	14.396	0.341	42.230	P
Gtl2 (*Meg3*)	61717777	0.004	7.186	1847.324	M
	61717778	0.094	74.708	792.774	M

Figure 5. Pyrosequencing verification for mouse *Peg5* in neonatal brain.

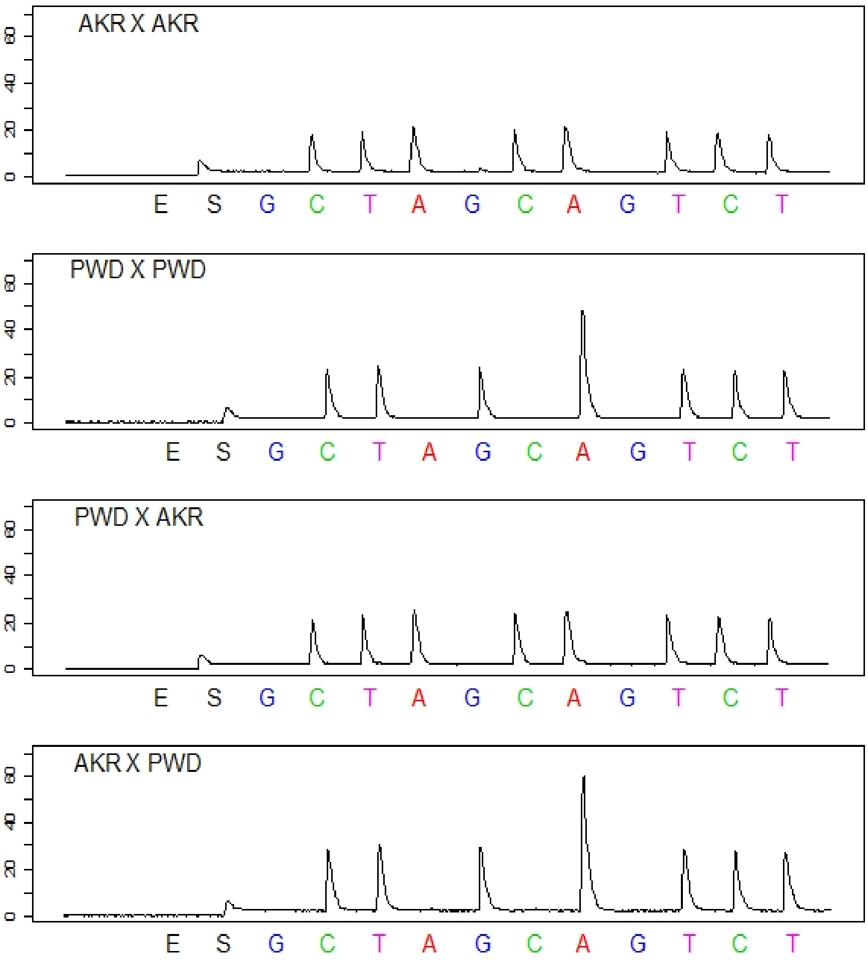


Figure 6. Pyrosequencing verification for mouse *Meg3* in neonatal brain.

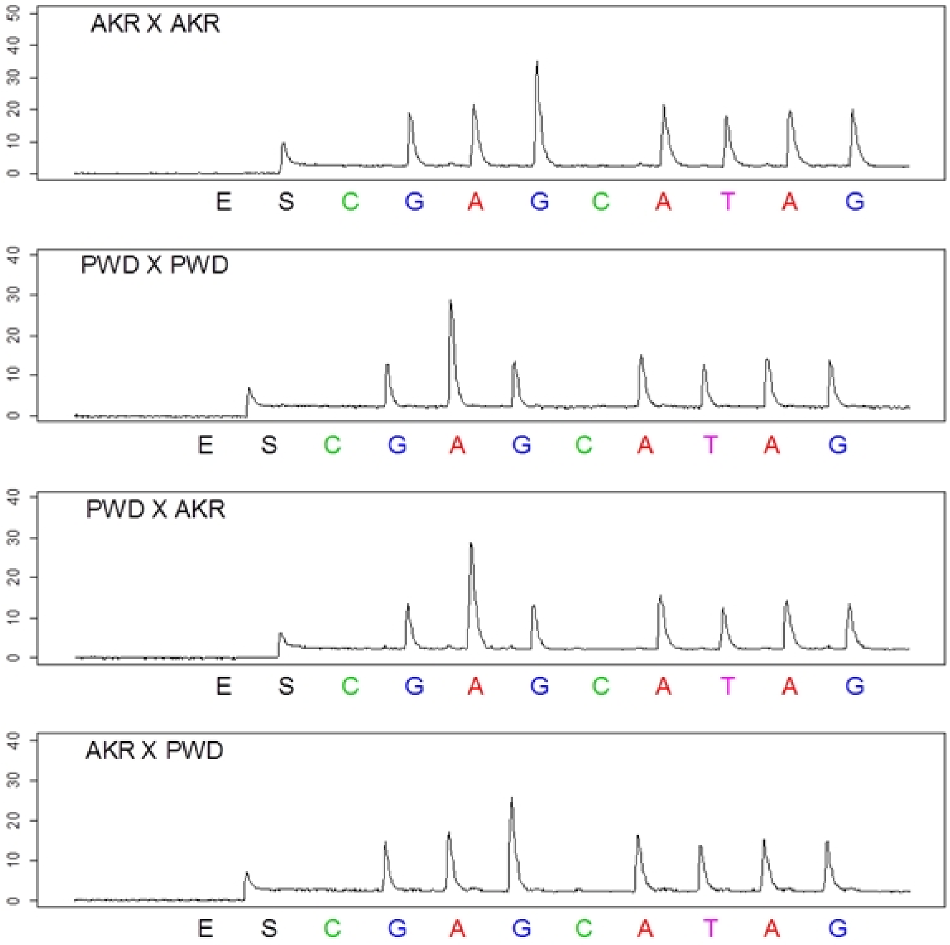


Table 4. List of candidate imprinted genes from custom SNP microarray study.

No	Gene name	Fold	IP status	Gene description	Gene Function
1	*Cuzd1*	3.828	M	CUB and zona pellucida-like domains 1	plays an essential role in trypsinogen activation
2	*H2-Aa*	4.278	P	histocompatibility 2, class II antigen A, alpha	major histocompatibility protein class II alpha chain
3	*Insl5*	6.656 4.615	M	insulin-like 5	is expressed in mouse brain and has a role in mobilization of calcium
4	*E030037J05Rik*	4.991	P	tensin 1	Homozygous knock out mice show disruption of cell-matrix junctions in
5	*B230120H23Rik*	3.604	M	sterile-alpha motif and leucine zipper containing kinase AZK	involved in MAPK signaling pathway
6	*Olfr544*	6.255	P	olfactory receptor 544	G-protein-coupled receptor
7	*Ednra*	3.001	M	endothelin receptor type A	is crucial for early neural crest cell patterning
8	*Bcl2l14*	6.15	P	Bcl2-like 14	apoptosis facilitator
9	*Prkar2b*	21.99	M	protein kinase, cAMP dependent regulatory, type II	Disruption in mice causes a lean phenotype, nocturnal hyperactivity,
10	*Shrm*	3.018	P	shroom family member 3	facilitates neural tube closure by regulating cell shape changes

Table 5. Fold of change of candidate imprinted genes in the microarray study.

Gene	SNP_ID	Reciprocal F1 ratios		Fold of difference	Strain combination
		Probe1	Probe2		
Cuzd1	50472329	0.733	2.806	3.828	AKR x PWD
H2-Aa	62694722	3.218	0.752	4.278	AKR x PWD
Insl5	38306343	0.552	3.675	6.656	AKR x PWD
Insl5	38306346	0.284	1.184	4.165	B6 x C3H
B230120H23Rik	38119858	0.702	2.531	3.604	AKR x PWD
E030037J05Rik	51276876	2.294	0.460	4.991	AKR x PWD
Olfr544	61044690	6.031	0.964	6.255	AKR x PWD
Ednra	60637726	0.716	2.147	3.001	AKR x PWD
Prkar2b	61590861	0.661	14.548	21.994	B6 x C3H
Shrm	46606289	1.124	0.372	3.018	B6 x C3H
Bcl2l14	60950825	3.202	0.521	6.150	B6 x C3H

Discussion

Custom microarray for detecting allele-specific expression

In this study, I designed a custom microarray to use for combined expression analysis and SNP genotyping. The imprinted status of the known imprinted genes and non-imprinted genes was correctly identified from the array results. Therefore, the allele-specific microarray method is valid for detecting allelic expression ratios. However, there are several limitations for the array method. Some mismatch probes have substantial signal intensity, making a large fraction of the probes uninformative. There are computational methods to predict the affinity of perfect match probes, but they cannot predict the affinity difference between the perfect match and mismatch probes. In addition, the microarray can only include known gene models, so it cannot query all transcripts and different splice variants in the transcriptome.

Computational prediction for novel imprinted genes

Here, to discover novel imprinted genes in mouse and test the imprinting status of their human orthologs, I chose two mouse strain pairs (AKR x PWD, B6 x C3H) and extracted total RNA from post-natal day 2 brain of both parents and the reciprocal F1s. 156 genes were selected as candidate genes from 600 putative mouse imprinted genes predicted by Luedi *et al.* (2005). Additional candidates were selected among orthologs of 63 potentially imprinted genes in humans (identified by uniparental expression in collaboration with Perlegen Sciences). Then I designed 25-mer oligo probes for a total of 396 exonic SNPs in the test genes and had a microarray printed by Agilent. By hybridizing the Cy3 and Cy5 labeled cRNAs made from the RNA samples to this microarray, we were able to tell the relative abundance of the two SNP alleles in the two reciprocal F1 strains. By applying an arbitrary cutoff, I found 10 candidate imprinted genes, but allele-specific pyrosequencing could verify none of them.

Computational prediction methods, including machine learning approaches, are very powerful for predicting functional elements and transcription factor binding sites in genomes. However, the prerequisite for successful prediction is sufficient and correct training data. The computational predictions for novel imprinted genes using DNA sequence related data suffer from the following limitations. First, there are very limited known imprinted clusters serving as training data. There are about 100 known imprinted genes in human and mouse, but they are not independent. Instead, they arrange in only 20 clusters. If you use 10 clusters as the training data and the other 10 as the verification set, the lack of data will result in an over-fitting problem in statistical learning, causing high false positive rate. Second, the tissue and developmental-stage specificity of genomic imprinting was ignored for prediction. Among the 100 known imprinted genes, for a specific tissue and developmental stage, only a subset of them is imprinted. The prediction method ignored this, and it does not know which tissues and stages are appropriate for verification of the candidates. Third, there are maternally and paternally expressed genes. Classifying the training set to these two directions further reduces the training data, causing more severe over-fitting. In most of the cases, the paternally and maternally expressed imprinted genes are present in the same imprinting cluster. Fourth, some imprinted genes show preferential maternal/paternal expression. The degree of imprinting for partial imprinted genes was ignored. Fifth, within a single imprinting cluster, there are multiple transcribed genes, but not all of them are imprinted. Some of them are imprinted but the others show biallelic expression. It is extremely difficult to distinguish such genes in the same cluster by DNA sequence related features. Sixth, because there are only 20 imprinting clusters, for the non-imprinted control gene selected, there are many different genes (all other genes in the genome) to choose from. In 2005, there were not many studies for a list of confirmed non-imprinted genes, therefore the selected non-imprinted genes could be imprinted or partially imprinted. To solve the problem in my previous point, non-imprinted genes in an imprinting cluster were not selected in the training set. The bias in selection of non-imprinted control genes will also cause over-fitting and high false positive rate. Seventh, the mechanism of genomic imprinting is not fully understood yet. The pattern discovered from the training data could only identity novel imprinted genes with the same mechanisms. Some potential imprinted genes with unknown mechanisms will be missed. Last but not least, in the machine learning approach, more than 7000 different features were used for predicting the imprinting status. Using such a large number of features will likely increase the possibility of over-fitting. Our microarray study did not confirm any of the prediction imprinted genes in mouse brain. Daelemans *et al.* (2010) performed a similar allele-specific microarray study targeting 932 genes in human term placenta (Daelemans et al. 2010). Their array design is enriched for the 124 predicted imprinted genes in human (Luedi et al. 2007), and none of them were confirmed either.

Will using epigenetic marks improve the computational prediction for novel imprinted genes?

Genomic imprinting is an epigenetic phenomenon, so using only DNA related features is not sufficient. I think including data from the epigenetic marks such as DNA methylation pattern and different histone modifications will increase the power, because you are adding more tissue-specific predictors. However, the epigenetics marks are dynamic during development, and are highly tissues-specific. Therefore, to get correct results, all epigenetics marks used in the computational prediction must be from the same tissue. Also, the predicted candidate genes should also be verified in this tissue. Including epigenetics marks from a variety of different cell lines at different developmental stages would not help much in terms of predicting power.

The path to a complete list of imprinted genes in the genome

The ultimate goal is to understand the evolutionary constraints imposed by the effective loss of diploidy engendered by genomic imprinting. Developing a more complete list of imprinted genes will help fully test ideas about genomic conflict. Bioinformatic prediction has its own limitations due to lack of training data and the over-fitting problem. We would like to develop a novel method that could query the imprinted status of all transcribed genes in the genome.

CHAPTER 2

Identification of novel imprinted genes in mouse neonatal brain from RNA-seq data[1]

Abstract

Imprinted genes display differential allelic expression in a manner that depends on the sex of the transmitting parent. The degree of imprinting is often tissue-specific and/or developmental stage-specific, and may be altered in some diseases including cancer. Here we applied Illumina sequencing of the transcriptomes of reciprocal F1 mouse neonatal brains and identified 26 genes with parent-of-origin dependent differential allelic expression. Allele-specific Pyrosequencing verified 17 of them, including three novel imprinted genes. The known and novel imprinted genes all are found in proximity to previously reported differentially methylated regions (DMRs). Ten genes known to be imprinted in placenta had sufficient expression levels to attain a read depth that provided statistical power to detect imprinting, and yet all were consistent with non-imprinting in our transcript count data for neonatal brain. Three closely linked and reciprocally imprinted gene pairs were also discovered, and their pattern of expression suggests transcriptional interference. Despite the coverage of more than 5000 genes, this scan only identified three novel imprinted RefSeq genes in neonatal brain, suggesting that this tissue is nearly exhaustively characterized. This approach has the potential to yield a complete catalog of imprinted genes after application to multiple tissues and developmental stages, shedding light on the mechanism, bioinformatic prediction, and evolution of imprinted genes and diseases associated with genomic imprinting.

[1] This study was published in 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.

Introduction

To date, 98 genes have been shown to undergo genomic imprinting in mouse, and 56 genes are imprinted in humans, with an overlapping set of 38 genes imprinted in both species (Morison et al. 2005). For neither species is the list of imprinted genes complete. Genome-wide bioinformatic predictions face the challenge of a high false positive rate, mostly because the training set of known imprinted genes is small, and we do not know all the signals driving tissue- and time-specificity of imprinting (Luedi et al. 2005; Luedi et al. 2007). Attempts at exhaustive scans for imprinted genes in humans have encountered several drawbacks, including the challenge of using the most appropriate tissue and developmental stage, a problem exacerbated by reliance on lymphoblastoid cell lines (LCLs) (Pollard et al. 2008). Many imprinted genes show tissue- and developmental stage-specific expression, and many are expressed and imprinted only in specific stages of brain development. Human studies also face the challenge of a low number of informative heterozygous SNPs, so that allele-specific assays are useful for only a small subset of individuals. Hence, pedigree information is needed to distinguish genomic imprinting from stochastic monoallelic expression (Lomvardas et al. 2006; Gimelbrant et al. 2007). These factors greatly amplify the effort and cost needed for a transcriptome-wide scan for imprinted genes in humans. By contrast, large-scale mouse studies have used uniparental disomy (Cattanach and Kirk 1985; Ferguson-Smith et al. 1991; Cattanach et al. 1992; Schulz et al. 2006; Ogata et al. 2008; Yamazawa et al. 2008) to detect parent-of-origin effects. While this approach has led to the discovery of many imprinted genes, and to the refinement of phenotypic analysis of the consequences of disruptions in imprinting, not all genomic regions are covered by uniparental disomies, and there is a risk that such aberrant genome configurations may distort expression patterns. Microarray-based approaches using allele-specific probes can only detect nearly "all-or-none" imprinting with confidence, because quantitative differences between maternal vs. paternal allelic expression have high error due to the cross hybridization of the perfect-match and mismatch probes (Bjornsson et al. 2008; Serre et al. 2008). In fact, genomic imprinting may occur as a continuum from complete uniparental expression to a slight but significant bias in the parental allele that is expressed, and a technology that could reliably detect quantitative differences in allele-specific expression at a transcriptome scale would greatly accelerate imprinting research.

Materials and Methods

Mouse Strains

Please see Chapter 1 for the mouse strains information and the method for total RNA extraction.

Illumina sequencing of the transcriptome

One Illumina Genome Analyzer run was performed for each reciprocal F1 of PWD and AKR mice at the Genome Center at Washington University. cDNA was synthesized using a modified version of the SMART Technology (ClonTech). To improve sequence coverage, we used a size selection procedure that removed cDNAs shorter than 1.3 kb in length. One Illumina Genome Analyzer run of each reciprocal F1 sample was run on the Illumina Genome Analyzer.

Synopsis. Mouse total RNA was converted to first strand cDNA using a modified-SMART protocol. The first strand cDNA was then PCR amplified and size fractionated in 6% polyethylene glycol (PEG)/0.55M sodium chloride (NaCl) to enrich for cDNA ≤1250bp.   SMART adapters were then excised from the cDNAs using *Mme*I and the adapters were removed from the reaction using 11% PEG/0.5M NaCl. The purified cDNA population then was fragmented and used as the source for a standard Illumina fragment library.

Modified-SMART. First strand cDNA was produced from mouse total RNA according to a modified version of the Clontech SMART protocol (E. Mardis, personal communication), using approximately 1 μg of total RNA and SuperScript II (Invitrogen).

Cycle optimization PCR and production PCR. The modified-SMART cDNA was used as the template in a PCR reaction to determine the number of cycles at which the reaction is no longer exponential. The cycle optimization reaction used 1 μl of the first strand cDNA reaction. Aliquots were removed at 2 cycle timepoints between 16 and 24 cycles. They were then run on a Flashgel (Lonza) for 5 min at 275 v, and the optimum cycle number was determined by observation.

The production PCR consisted of eight 100 μl reactions identical in composition to the cycle optimization reaction except that 2 μl of first-strand cDNA was used for each reaction and the empirically determined cycle optimum number was used for amplification of all eight reactions. The PCR products were purified and concentrated with two Qiaquick columns (Qiagen), according to the manufacturer's protocol, and eluted with 30 μl Buffer EB (Qiagen) per column.

Size fractionation. To fractionate cDNA ≤1250 bp, the amplified cDNA from the production PCR reactions was resuspended in a 300 μl reaction of 6% PEG-8000, 0.55 M NaCl and carboxylate paramagnetic beads. The mixture was vigorously vortexed and incubated for 10 min at room temperature. The reaction was placed on a magnetic particle collector (MPC, Invitrogen) for two min and the supernatant, containing the ≤ 1250 bp fraction, was transferred to a clean tube. This cDNA fraction was purified over a Qiaquick column according to the manufacturer's protocol, and eluted in 50 µl Buffer EB.

Adapter removal and cDNA purification. The 5' and 3' adapters added during cDNA synthesis, contain *Mme*I recognition sequences that are removed by digestion in a 100 µl reaction containing 1x NEB Buffer 4 (20 mM Tris-acetate, 50 mM potassium acetate, 10 mM magnesium acetate, 1 mM dithiothreitol, pH 7.9 @ 25°C), 10 µg of 10mg/ml BSA, 64 µM S-adenosylmethionine (New England Biolabs) and 12 units *Mme*I (New England Biolabs) for 30 min at 37°C. The digested cDNA was purified and concentrated with 1 Qiaquick column according to the manufacturer's protocol, and eluted with 30 µl Buffer EB.

A second round of PEG/NaCl fractionation further removes the adapter fragments released by *Mme*1 digestion. Here, the cDNAs purified by Qiaquick column were resuspended in a 300 µl reaction of 11% PEG-8000, 0.5M NaCl and carboxylate paramagnetic beads. The mixture was vigorously vortexed and incubated for 10 min at room temperature. The reaction was placed on an MPC for two min and the supernatant was then discarded. The paramagnetic beads were washed twice with 70% ethanol and air dried. The tube containing the paramagnetic beads was removed from the MPC and the beads were resuspended in 50 µl Buffer EB with vigorous vortexing. The reaction was placed on the MPC for two min and the supernatant was transferred to a clean tube. This fraction contains cDNA >150 bp and free of 5' & 3' adapters.

Nebulization/Covaris shearing and Illumina/Illumina library prep. Sample B17 (PWD/PhJ x AKR/J): The cDNA was sheared by nebulization (2 min at 50 PSI) and the sheared DNA was purified/concentrated with a single Qiaquick column according to the manufacturer's protocol. Sample B21 (AKR/J x PWD/PhJ): The cDNA was sheared with the Covaris S2 System in 75% glycerol with the following program: 10 cycles of 4 treatments, 60 sec each; Duty cycle = 20%; intensity = 10; 1000 cycles/burst. The cDNA was purified/concentrated by ethanol precipitation.

The sheared cDNAs were then prepared for Illumina sequencing according to the manufacturer's protocols. Libraries were prepared from a 150-200 bp size-selected fraction following adapter ligation and agarose gel separation. The libraries were sequenced using a single end read protocol with 32 bp of data collected per run on the Illumina Genome Analyzer. Data analysis and base calling were performed by the Illumina instrument software.

Illumina sequence data analysis

We obtained 33,519,739 reads (1072.63 Mbp total) from the PWD/PhJ x AKR/J cross (PWD x AKR for short) in seven lanes, and 35,510,887 reads (1136.35 Mbp total) in eight lanes for the reciprocal cross, AKR/J x PWD/PhJ (AKR x PWD for short).

Both runs have high sequence quality with 95% of the bases passing Q20 (Figure S1.1).

We used a local version of the NCBI BLAST program (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi) to align the 32-bp reads to the mouse RefSeq database (http://www.ncbi.nlm.nih.gov/RefSeq/). The parameters for the `blastn` program were optimized for short reads and our purpose. We did the BLAST jobs on 180 nodes of the CBSU clusters (http://cbsuapps.tc.cornell.edu/index.aspx) using the P-BLAST utility. 23.82% of the total reads in the PWR x AKR cross were aligned to the RefSeq database with 3.57 hits/read. 31.18% of the total reads in the AKR x PWD cross were aligned to the RefSeq database with 3.02 hits/read (Table 6). High quality SNP-containing reads were filtered out, with unique match to the RefSeq gene (or different transcripts of the same Entrez gene). Relative expression level of the two parental alleles was estimated by the relative counts of Illumina reads at the SNP positions in the Perlegen mouse SNP database. 59 out of the 98 known imprinted genes in mouse are in the mouse RefSeq database. We assembled them into ace files according to the BLAST alignment information. 20 novel SNPs were called in 12 known imprinted genes from the Illumina assembly (Table 7).

Table 6. Summary of BLASTN results.

Cross	Lane	Total reads	Matched reads	% of match	No match	Poly A/N*	Total hits	hits/read
PWD x AKR	s1	4,619,970	1,202,604	26.03%	3,409,387	7,979	4,369,063	3.63
	s2	4,295,871	1,200,111	24.73%	3,088,792	6,968	4,408,399	3.67
	s3	4,722,842	1,095,349	23.19%	3,615,132	12,361	3,852,524	3.52
	s4	4,853,113	1,126,465	23.21%	3,713,847	12,801	4,009,494	3.56
	s6	5,158,778	1,193,386	23.13%	3,953,819	11,573	4,249,988	3.56
	s7	5,053,146	1,173,701	23.23%	3,868,074	11,371	4,178,607	3.56
	s8	4,816,019	1,116,624	23.19%	3,688,339	11,056	3,919,609	3.51
AKR x PWD	s1	4,096,916	1,241,763	30.31%	2,777,901	77,252	3,749,816	3.02
	s2	4,339,623	1,322,613	30.48%	2,922,408	94,602	3,996,721	3.02
	s3	4,447,068	1,361,104	30.61%	2,990,873	95,091	4,126,160	3.03
	s4	4,397,822	1,348,417	30.66%	2,956,351	93,054	4,073,444	3.02
	s5	4,399,210	1,369,262	31.13%	2,932,067	97,881	4,150,208	3.03
	s6	4,509,790	1,417,377	31.43%	2,992,308	100,105	4,289,787	3.03
	s7	4,493,249	1,444,386	32.15%	2,946,392	102,471	4,345,896	3.01
	s8	4,827,209	1,576,487	32.66%	3,140,904	109,818	4,746,400	3.01

*: low complexity reads including polyA and polyT

Table 7. Known imprinted genes covered by assembly of Illumina reads.

RefSeq_ID	PWD x AKR		AKR x PWD		Gene_name	Chr	RefSeq_len
	AKR count	PWD count	AKR count	PWD count			
NM_011245	16	0	0	20	*Rasgrf1*	chr9	4243
NR_002864	168	0	6	74	*Peg13*	chr15	4745
XR_035484	1	339	193	1	*Gtl2*	chr12	1890
NM_001077507	181	214	101	96	*Gnas*	chr2	3733
NM_001033962	3	10	4	2	*Ube3a*	chr7	4910
NM_010514	52	43	20	27	*Igf2*	chr7	4038
NM_008672	9	12	61	76	*Nap1l4*	chr7	2283
NM_009876	0	8	13	0	*Cdkn1c*	chr7	1849
NM_010345	2	3	6	10	*Grb10*	chr11	5015
NM_021432	22	0	0	67	*Nap1l5*	chr6	1909
NM_181595	36	54	67	26	*Ppp1r9a*	chr6	9547
NR_001592	2	14	61	1	*H19*	chr7	2615

Estimation of the relative parental expression

To identify the SNP positions in the mouse RefSeq database, we used the SNP genotype and information in the Perlegen mouse SNP database (http://mouse.perlegen.com). Perlegen Sciences and NIEHS genotyped 8 million SNPs among 15 mouse strains with a genome coverage of 70%, including PWD and AKR. The SNP density is approximately 3 SNPs/kb and most of the genic regions are covered in the database. The genome coordinates of the reviewed and validated mouse RefSeq sequences (starting with NM and NR, see http://www.ncbi.nlm.nih.gov/RefSeq/key.html#status) were downloaded from the UCSC genome browser (www.genome.ucsc.edu, July 2007 assembly). The SNP positions in the RefSeq sequences were filtered by the RefSeq gene coordinates. To correct for gaps in the RefSeq-genomic sequence alignments, we also did text matches using 20 bp upstream and downstream the SNP positions. A total of 206,589 Perlegen SNPs were found in 18,797 RefSeq sequences (Table 8 andTable 9), with an average SNP density of 11 SNPs/RefSeq sequence (Figure 7). 4,127 SNPs with missing data in the Perlegen SNP database were called based on the Illumina sequence reads. The genotypes of all the high quality Perlegen SNPs (*q*-score $\geq$ 10, Mismatch $\leq$ 4 for alternative allele, Mismatch $\leq$ 3 for reference allele and match length $\geq$ 28) covered in the Illumina reads were summarized in the two reciprocal F1s. 175,687 (84.71%) of the 207,407 Perlegen RefSeq SNPs were not covered or not informative (less than 1 SNP count in both direction). In the 31,720 Illumina-covered Perlegen SNPs, 25,289 (83.21%) were confirmed by Illumina reads, and 4,127 (13.58%) Perlgen SNPs with missing data (N) in AKR and PWD strains were called based on the Illumina sequence information (Figure 8). The newly called SNPs were included in the data analysis. From the results, the genotype of the Illumina short-read sequence identified SNPs are consistent with the Perlegen SNP, indicating high sequence quality of our Illumina Genome Analyzer run. There are only 161 inconsistent SNPs, most of which are the complementary allele and could come from the antisense transcript of the RefSeq gene.

The expression level of the RefSeq transcripts were quantified by the number of perfectly matched reads in the Illumina sequence data. 15,491 RefSeq genes were covered by at least one perfect match read in each of the two reciprocal crosses (Figure 9).

In order to do the quality control and filter out the true SNP-containing reads, several criteria were considered. The Illumina sequence SNPs (Perlegen SNP that are present in our Illumina reads) were classified to six categories according to their consistency with the Perlegen SNP information (Figure 10). Classes 1-5 are the consistent SNPs. Class 1 includes SNPs that are polymorphic between AKR and PWD strains. These are the SNPs we want to use in our study to quantify the relative parental expression. Class 2 SNPs are also consistent but the SNP is not polymorphic between AKR and PWD strains. Classes 3-5 are SNPs that have missing data (N) in the Perlegen database. The rest of the Illumina SNPs are classified in class 0, which are the inconsistent SNPs. Most of the Illumina SNPs have a quality score 20 or above

(Figure 10). The distribution of the number of mismatches showed that the pattern class 1 SNPs are consistent with perfectly matched reference and alternative alleles, an attribute not seen in any other SNP classes (Figure 11). So the class 1 SNPs were used in the following analysis. Regarding the match length of the SNP-containing reads, more than 80% have a full length match (32 bp), and most of the reads have a match length of 25 or more. The `blastn` algorithm is a local alignment algorithm, so if there are SNPs in the first or last 2 bp of a read, the alignment will be truncated, although it is still considered a full length match (Figure 12). Two sets of filter criteria were used before the final SNP counts for each RefSeq gene were summarized (Table 10). Both Filter 1 and Filter 2 are conservative and the reads after the filtering all matched uniquely to the Entrez gene database (could be more than one RefSeqs due to alternative splicing).   Since there is no lane effect, the AKR and PWD counts in the two reciprocal crosses are summarized by RefSeq genes and by SNPs. 326 class 1 SNPs are not polymorphic in the Illumina sequence data due to the repetitive match of the SNP-containing sequence in the mouse genome, so we do not know where transcripts bearing these SNPs are coming from. Such SNPs are excluded from the final analysis.

Table 8. Categories of Perlegen SNP in the RefSeq sequences.

Categories	Coordinates filter	Text match	Status	# of RefSeq sequences	% Total
Conisistent	YES	YES	Consistent	172,104	83.31%
Corrected by text match *	YES	YES	Inconsistent, corrected by text match	15,679	7.59%
Coordinates information only **	YES	NO	Coordinates only	18,806	9.10%
Total				206,589	

*: There are discrepency between the RefSeq and genomic sequence alignment. The gaps in the alignment could be corrected by the text match.

**: Not all coordinate filtered SNPs have text match results, because of the exon-intron boundary problem.

Table 9. Consistency of Perlegen SNP alleles and the RefSeq alleles.

Categories	Example			Counts	% of Total
	RefSeq	P_reference	P_alternative		
RefSeq allele is conisistent with Perlegen reference allele	A	A	C	202,718	98.13%
RefSeq allele is conisistent with Perlegen alternative allele *	A	C	A	3,829	1.85%
RefSeq allele is NOT conisistent with Perlegen alleles **	A	G	T	42	0.02%
Total				206,589	

*: The Perlegen reference allele is from the reference genome sequence, which is C57BL/6. Not all the RefSeq sequences come from the C57BL/6 strain whose genome

**: The inconsistent SNPs could be due to Perlegen genotyping error, RefSeq sequencing error or RNA editing.

Figure 7. Distribution of number of Perlegen SNPs per RefSeq genes.

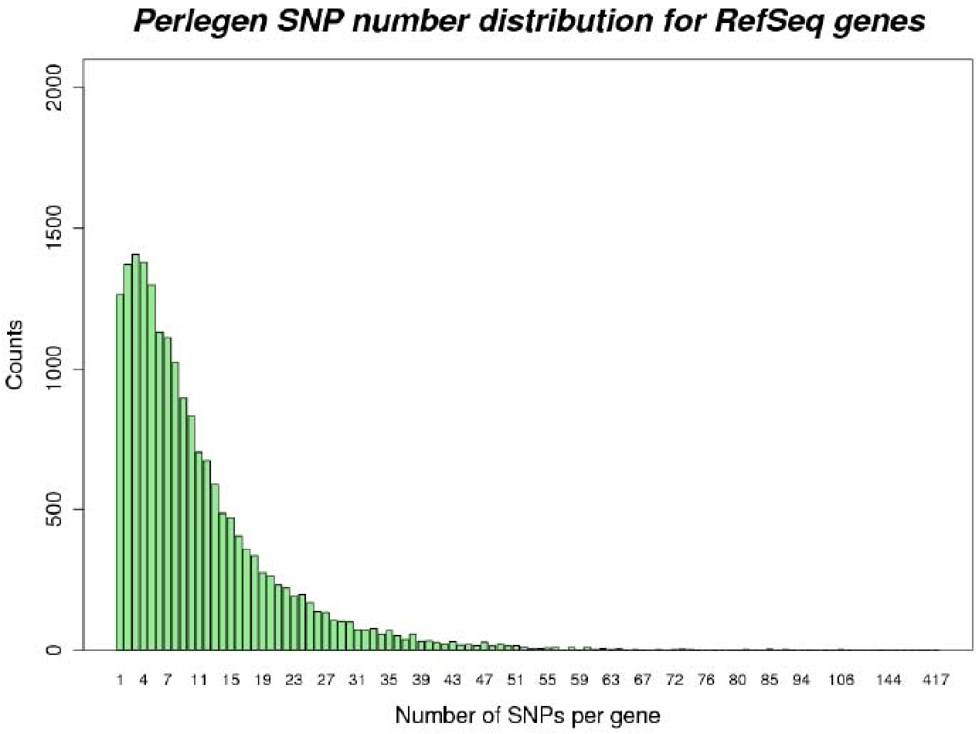
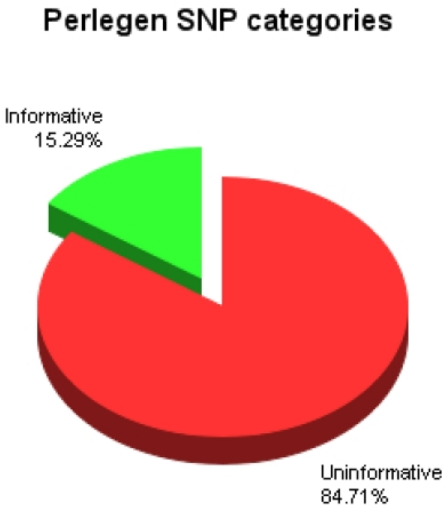


Figure 8. SNP calling for the Perlegen missing data.

(A). Summary of coverage of the Perlegen RefSeq SNPs. Informative: SNP counts >= 1 in both the two reciprocal crosses. **(B).** Summary of the Illumina Informative SNPs. **Confirmed:** Perlegen SNPs that present in Illumina and the genotypes agree with each other. **Called:** Perlegen SNPs with missing data in AKR and PWD that called based on the Illumina information.

A. Summary of coverage of the Perlegen RefSeq SNPs.



B. Summary of the Illumina Informative SNPs.

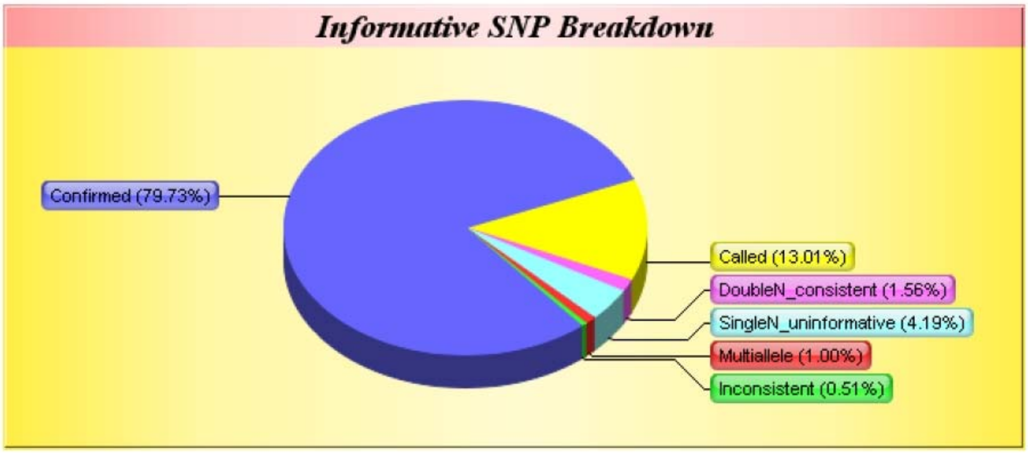


Figure 9. Coverage distribution in the Illumina data. Histogram of number of perfect match reads per gene for 15,491 RefSeq genes covered by at least one perfect match read in each of the two reciprocal crosses. (Only RefSeq genes with 10,000 or less perfect match reads are shown. There are 124 genes with number of perfect match reads > 10,000.)

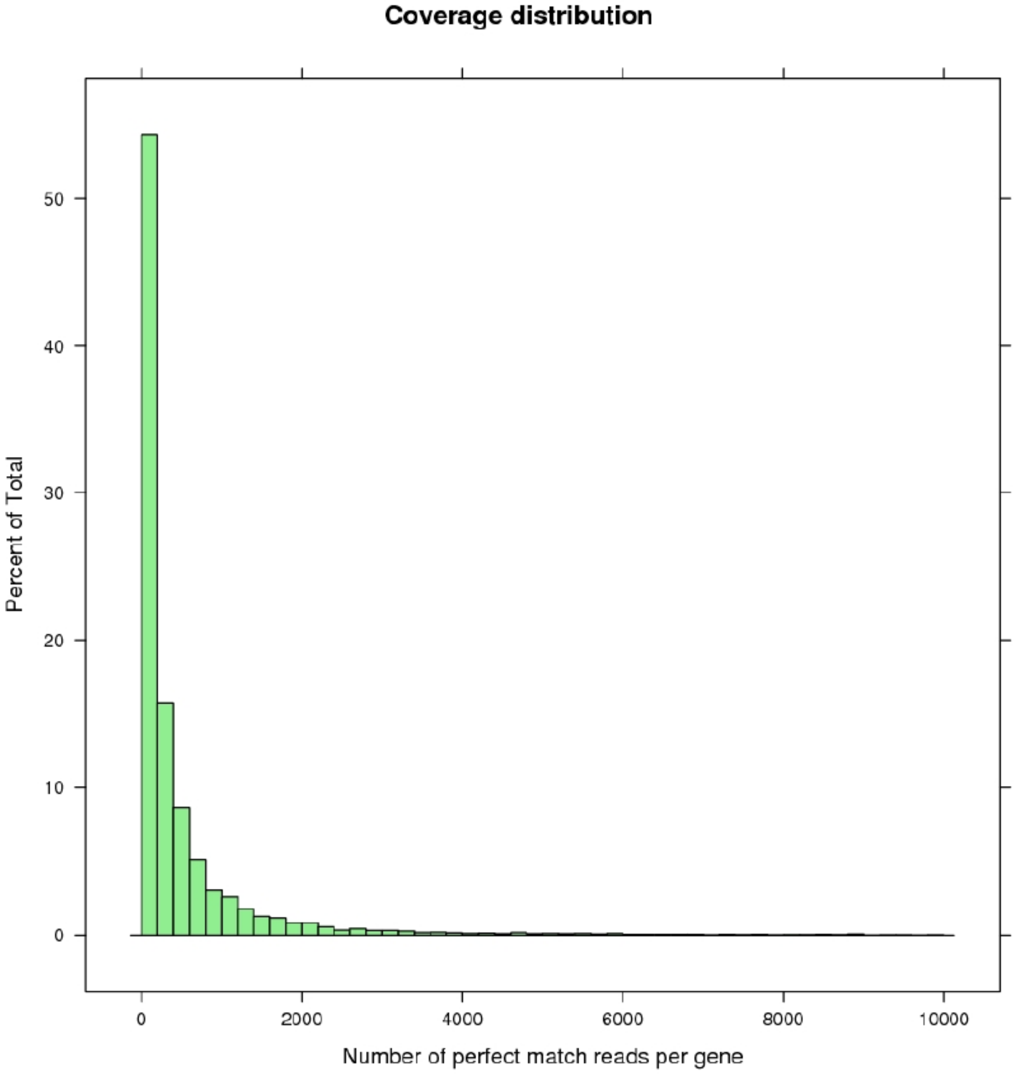


Table 10. Criteria for SNP filtering of the Illumina data.

Filter 1
• Type 1 SNP only.
• Q-score of the SNP position >= 10
• Mismatch score for reads containing reference allele: 3
• Mismatch score for reads containing alternative allele: 4
Filter 2
• Type 1 SNP only.
• Q-score of the SNP position >= 10
• SNP within 1.3kb to the 3'-end.
• Mismatch score for reads containing reference allele: 3
• Mismatch score for reads containing alternative allele: 4
• Match length >= 28.

Figure 10. Quality score distribution of Illumina SNPs by SNP class. Shown here is lane 1 for PWD x AKR cross.

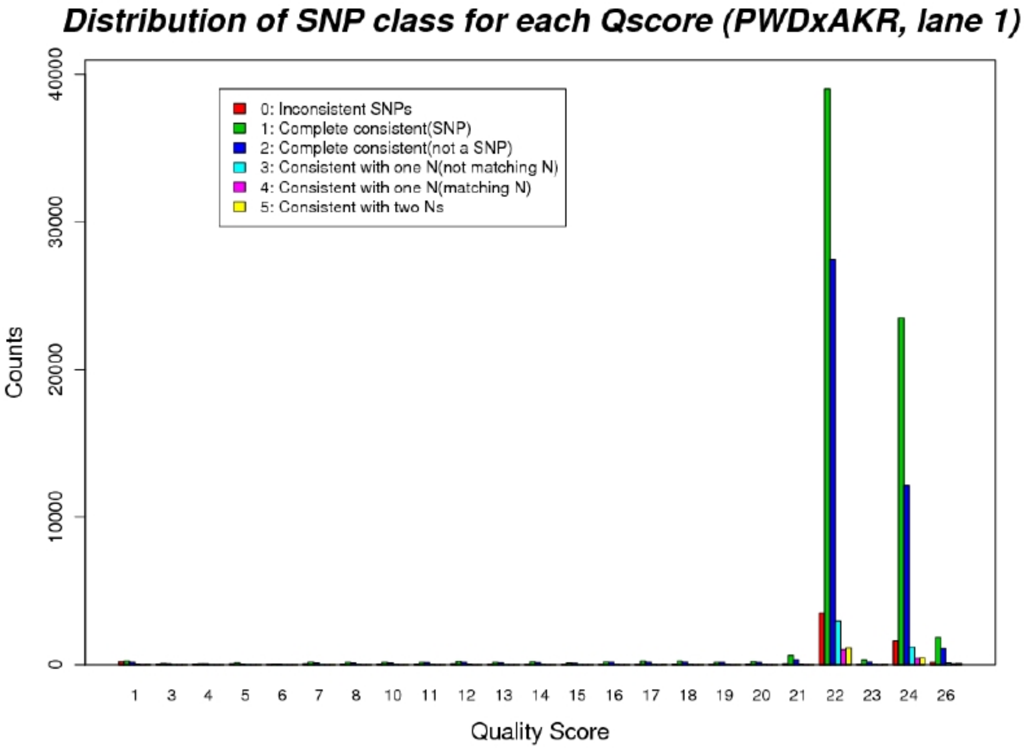


Figure 11. Mismatch score distribution of Illumina SNPs by SNP class. Shown here is lane 1 from PWD x AKR cross.

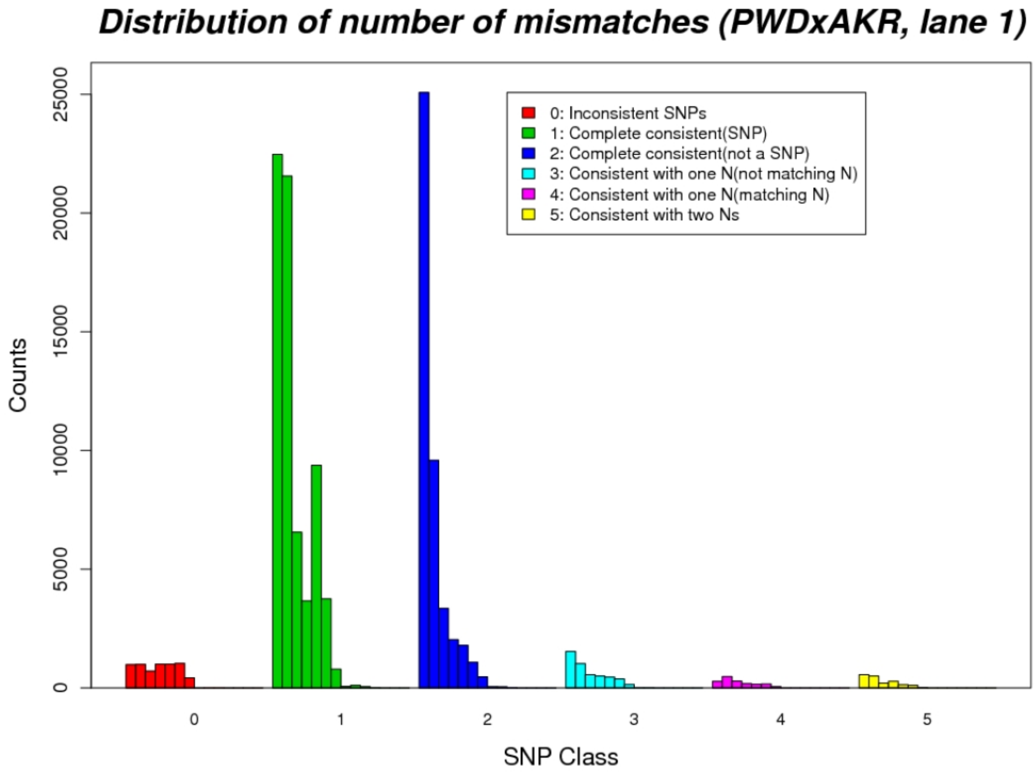
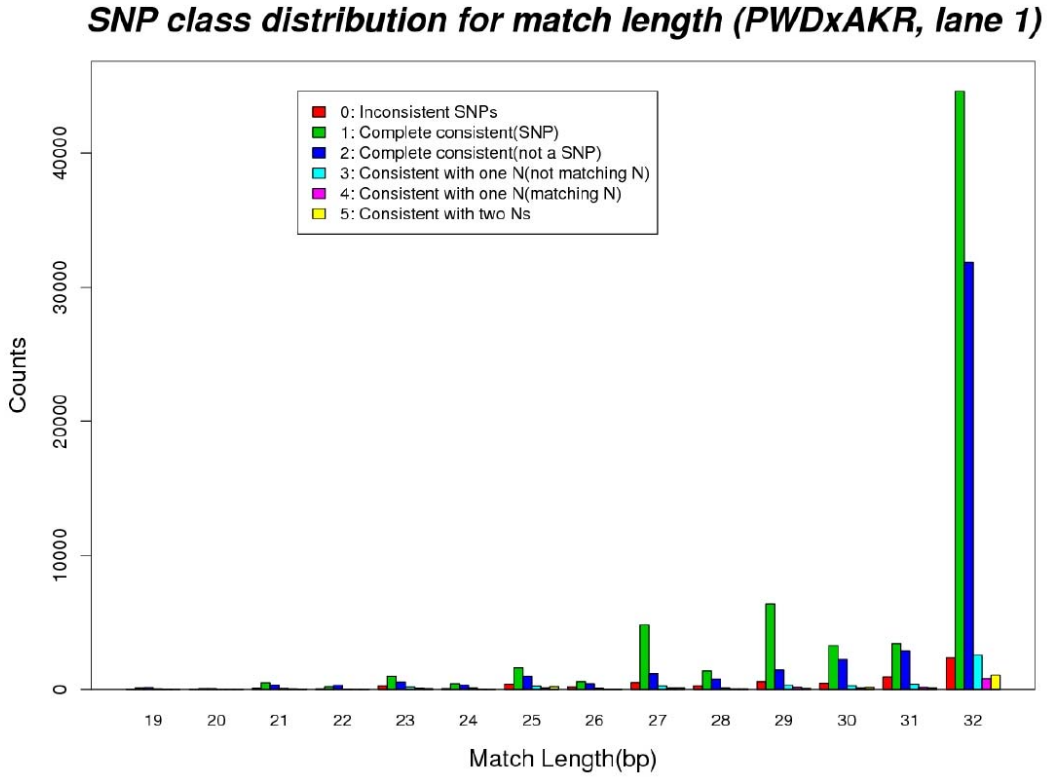


Figure 12. BLASTN match length distribution of Illumina SNPs by SNP class. Shown here is lane 1 from PWD x AKR cross.



Detecting genomic imprinting and Statistical analysis

We have the filtered AKR and PWD allele counts for the two reciprocal F1s. We define p_1 as the AKR allele proportion in the PWD x AKR cross and p_2 as the AKR allele proportion in the AKR x PWD cross (Table 11). If a gene has equal expression from the two parental alleles, p_1 and p_2 will be approximately 0.5. If a gene is an expression QTL (eQTL) with higher expression from the AKR-derived allele, p_1 will be approximately equal to p_2 and both p_1 and p_2 will be greater than 0.5. A paternally expressed imprinted gene will have the pattern of $p_1 > 0.5$ and $p_2 < 0.5$, whereas a maternally expressed imprinted gene will have the pattern of $p_1 < 0.5$ and $p_2 > 0.5$ (Table 12). The advantage of having the reciprocal crosses is that we could distinguish an eQTL from true genomic imprinting.

A formal statistical test is needed to test the significance. We did not use Fisher's exact test because it is a conservative test and results in substantial loss of power, especially when the total counts are small (Lehmann and Romano 2005). Rather, we used a modern statistical method, the Storer-Kim method for two independent binomials to test whether there is a significant difference between the two binomial parameters, p_1 and p_2 (Storer and Kim 1990). The *P*-values were calculated using Wilcox's code (Wilcox 2003) in R (version 2.60, www.r-project.org). The 95% confidence intervals for p_1 and p_2 were also obtained by the Wilson method (Wilson 1927) (R, the `binom` package). False discovery rate (*q*-value) was calculated by the `qvalue` package in R (Storey et al. 2004).

Table 11. Definition of p_1 and p_2.

Cross	AKR allele counts	PWD allele counts	
PWD x AKR	a	b	**p1** = a/(a+b)
AKR x PWD	c	d	**p2** = c/(c+d)

Table 12. Detecting genomic imprinting.

$p1 = p2 = 0.5$	Relatively equal expression from the two parental copies.
$p1 = p2 > 0.5$	eQTL with higher expression from the AKR strain
$p1 = p2 < 0.5$	eQTL with higher expression from the PWD strain
$p1 != p2, p1 > 0.5, p2 < 0.5$	Paternally expressed imprinted candidate gene
$p1 != p2, p1 < 0.5, p2 > 0.5$	Maternally expressed imprinted candidate gene

Sanger and Pyrosequencing verification

We designed Pyrosequencing PCR and sequencing primers for the candidate imprinted genes using Pyrosequencing Assay Design Software Version 1.0.6 (Biotage AB). To guarantee that there are no SNPs within the primers, SNP positions in the Perlegen database were labeled and excluded when designing the primers. PCR amplification for Pyrosequencing was done using TaqGold Enzyme (Applied Biosystems) with a 45 cycles of 3-step PCR (95°C for 45s, 46-58°C for 30s and 72°C for 10-20s) followed by a final extension of 10 min. The PCR products (80-300 bp) were purified by Exonuclease I and Shrimp Alkaline Phosphatase and sequenced bidirectionally using the original Pyro PCR primers on ABI 3730xl DNA analyzer (Applied Biosystems) with BigDye Terminator v3.1. The sequence chromatograms were analyzed with CodonCode Aligner version 2.0.4 (CodonCode Corporation Software for DNA Sequencing). PCR products for Pyrosequencing were amplified using the same condition with biotin labeled forward (or reverse) primer. The Pyrosequencing was done on a PSQ™ 96 MA Pyrosequencer (Biotage, AB) with the Pyro Gold Reagents (Biotage, AB). The relative level of the two parental alleles was quantified by the software for PSQ™ 96 MA Pyrosequencer (Version 2.02 RC 5.8, Biotage, AB) using the allele quantification method.

Results

Illumina sequencing results and SNP coverage

Short-read sequencing (*e.g*. Illumina sequencing) of transcripts provides many advantages in imprinting studies by providing a large number of sequence tags that allow simple counting of transcripts encoded by the two transmitted parental alleles. In this study, we performed quantitative assessments of genomic imprinting in transcripts from reciprocal cross progeny of the AKR/J and PWD/PhJ mouse strains. Total RNA was extracted from postnatal day 2 (P2) F1 female mouse whole brains. One run of Illumina sequencing was done for each F1 female brain cDNA sample. We obtained 1072.63 Mbp of sequence data from the PWD x AKR cross (listing female strain first) and 1136.35 Mbp from AKR x PWD in 32 bp reads with high quality (Figure 13). On average, 27.74% of the reads were aligned to the NCBI RefSeq mouse genome database. Sequence heterogeneity between alleles was great enough to produce poor performance by ELAND in mapping reads to the genome, so this mapping was performed with the NCBI BLAST program (Table 6). Altogether, 33,519,739 and 35,510,887 reads were aligned to the RefSeq database in the respective reciprocal crosses. The sequences covered 15,491 RefSeq genes with at least one perfectly matching Illumina read in each of the two reciprocal crosses. Within these genes, we identified 814,360 and 884,828 reads spanning Perlegen SNPs for the two respective reciprocal crosses (Frazer et al. 2007). After quality control filtering (Table 10), 320,804 and 327,451 high quality SNP-containing reads remained, allowing identification of parent-of-origin of each read (see Methods for more details). 5,533 RefSeq genes (5,076 unique Entrez genes) were covered in our study with a total SNP count of four or more in both reciprocal crosses (Table 13). From the mouse Brain EST Database, among the 5,500 cDNA clones of polyA-containing 3'-end EST sequences in P4 cerebellum, 3,500 are distinct species (Matoba et al. 2000). This contrasts with a recent SAGE study of P30 mouse brain, where the number of matched GenBank transcripts with copy number five or more per cell was 4,161 (Chrast et al. 2000), but those data lacked the allele-specific identification. Based on this information, we could query the imprinting status of nearly all currently known transcribed genes with detectable expression in mouse neonatal brain with an informative number of counts.

Detecting genomic imprinting

The relative expression level of the two parental alleles was quantified from the counts of the AKR and PWD SNP alleles in the Illumina read data (Figure 14). We define p_1 to be the percentage of counts from AKR allele in PWD x AKR cross, and p_2 as the percentage of counts from AKR allele in AKR x PWD cross (Table 11). We identify a gene as a paternally expressed candidate imprinted gene if p_1 is significantly different from p_2 and where p_1>0.5 and p_2<0.5 (and, for maternally expressed genes, p_1<0.5, and p_2>0.5) (Table 12). The Storer-Kim test for two independent binomials (Storer and Kim 1990; Wilcox 2003) was used to test the significance of the difference between the two binomial parameters, p_1 and $p_{2,}$ for each gene covered in the study (Storer and Kim 1990). *q*-values for each gene were calculated, and a false discovery rate cutoff of 0.05 was applied (Storey et al. 2004). Using these criteria, we identified 13 paternally and 13 maternally expressed candidate imprinted genes with p_1>0.65, p_2<0.35 (p_1<0.65, p_2>0.35 for maternal genes) and *q*-value < 0.05, respectively (Table 14).

Table 13. Summary of gene coverage and total SNP counts after filtering.

Filter	Gene Coverage***		PWD x AKR		AKR x PWD	
	RefSeq genes	Entrez gene	AKR*	PWD**	AKR*	PWD**
Filter 1	5,533	5,076	175,560 (54.73%)	145,244 (45.27%)	174,300 (53.23%)	153,151 (46.77%)
Filter 2	4,467	4,116	145,778 (54.47%)	121,853 (45.53%)	133,507 (52.44%)	121,096 (47.56%)

*: Total counts of AKR alleles

**: Total counts of PWD alleles

***: Genes covered by a total count of 4 or more in both cross

Figure 13. Quality score distribution for the Illumina sequencing reads. Shown here is lane 1 for PWD x AKR cross. The orange bar is the average quality score for the first 25bp. The red bar represents the average quality score for all positions.

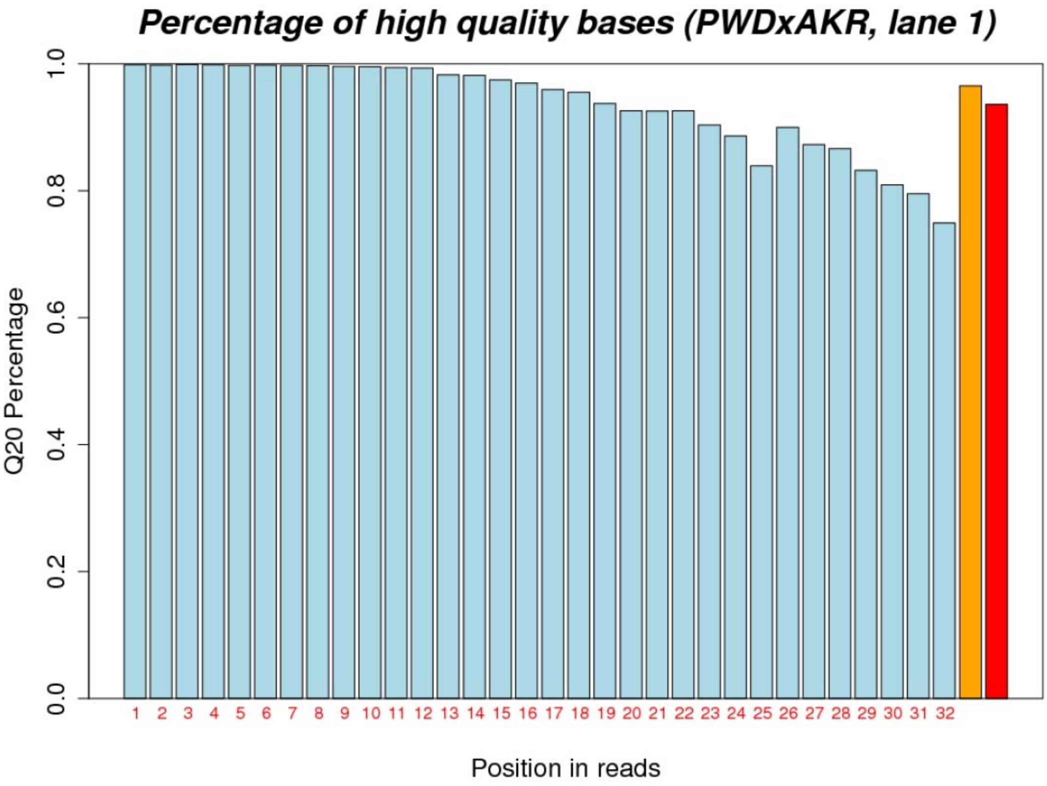


Table 14. Candidate imprinted genes identified by biased allelic counts among transcripts.

Known IP genes	PWD x AKR		AKR x PWD		q-value	AKR percentage		Known status†	Verified status	Sig_SNPs (q<0.1)¶	Pyrosequencing	
	AKR*	PWD*	AKR*	PWD*		p1	p2				p1	p2
Nnat[1]	1182	1	21	1853	0	99.90%	1.10%	IP	IP	4	100.00%	0.00%
Snrpn[2]	898	1	1	19	0	99.90%	5.00%	IP	IP	1	100.00%	0.00%
Snurf[2]	888	1	1	18	0	99.90%	5.30%	IP	IP	1	100.00%	0.00%
Peg13[3]	168	0	6	74	0	100.00%	7.50%	NR	IP	3	98.80%	3.00%
Nap1l5[3]	22	0	0	67	1.20E-19	100.00%	0.00%	NR	IP	1	100.00%	0.00%
Inpp5f_v2[4]	41	3	14	80	1.40E-17	93.20%	14.90%	IP	IP	2	91.90%	7.80%
Sgce[5]	9	0	0	54	2.00E-09	100.00%	0.00%	NR	IP	2	100.00%	1.50%
Rasgrf1[6]	16	0	0	20	7.50E-09	100.00%	0.00%	IP	IP	3	100.00%	0.00%
Impact[7]	15	6	8	83	1.20E-06	71.40%	8.80%	NR	IP	2	79.10%	19.80%
Zrsr1[8]	11	0	1	14	6.70E-05	100.00%	6.70%	IP	IP	0	97.50%	0.40%
Gtl2[9]	1	339	193	1	0	0.30%	99.50%	NR	IP	4	0.00%	100.00%
H19[10]	2	14	61	1	5.80E-10	12.50%	98.40%	NR	IP	3	9.40%	100.00%
Cdkn1c[11]	0	8	13	0	1.30E-04	0.00%	#######	NR	IP	1	3.60%	100.00%
Commd1[12]	12	33	22	7	2.60E-03	26.70%	75.90%	IP	IP	0	41.20%	72.50%

Novel IP genes	PWD x AKR		AKR x PWD		q-value	AKR percentage		Known status†	Verified status	Sig_SNPs (q<0.1)¶	Pyrosequencing	
	AKR*	PWD*	AKR*	PWD*		p1	p2				p1	p2
Inpp5f	359	19	89	1293	0	95.00%	6.40%	-	IP	7	83.20%	19.10%
2410042D21Rik	21	7	16	32	0.024	75.00%	33.30%	-	eQTL$	0	79.90%	83.60%
BC043301	8	0	3	9	0.042	100.00%	25.00%	-	eQTL	0	-	-
1810044A24Rik	7	20	25	5	1.10E-03	25.90%	83.30%	-	IP	1	20.60%	73.50%
Gyg	9	35	21	9	0.002	20.50%	70.00%	-	eQTL	1	40.90%	36.10%
Ppfia2	6	16	32	8	0.003	27.30%	80.00%	-	eQTL	0	-	-
Prim1	6	81	5	2	0.005	6.90%	71.40%	-	eQTL	1	-	-
Asns	24	60	27	14	0.005	28.60%	65.90%	-	eQTL	1	53.70%	56.30%
2010012O05Rik	6	17	41	16	0.01	26.10%	71.90%	-	eQTL	0	56.70%	57.60%
Rgs17	10	24	39	17	0.013	29.40%	69.60%	-	eQTL	0	54.50%	55.10%
Pdcl	5	13	61	23	0.018	27.80%	72.60%	-	eQTL	0	56.80%	58.90%
Blcap	6	13	15	2	0.025	31.60%	88.20%	-	IP	1	25.20%	73.70%

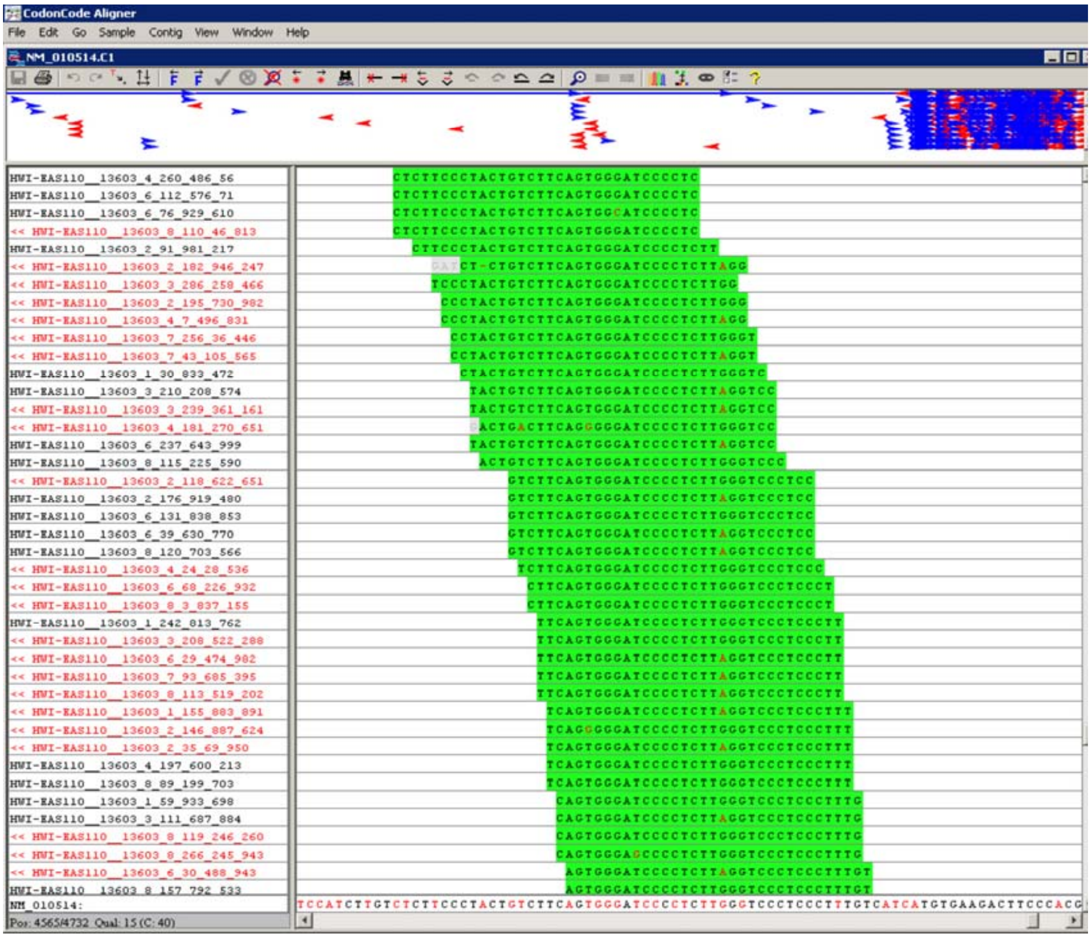
*: Counts of the AKR and PWD allele in the Illumina sequence data after filtering.

†: Reported imprinted status of the known imprinted genes in neonatal brain (IP: imprinted; NR: not reported).

¶: Number of significant SNPs with *q* -value ? 0.10 for each gene.

$: eQTL : Expression quantitative trait loci

Figure 14. Alignment of Illumina sequence reads for *Igf2* transcript. The top panel is the summary window or all 1,253 cDNA reads that mapped to the 4,038 bp Igf2 transcript (NM_010514). The blue arrows represent the sense reads and the red arrows represent antisense reads. From the figure, most of the reads were aligned to the 1 kb region near the 3'-end of the transcript. The bottom left panel gives the Illumina read names, and the bottom right gives the sequence alignment. Sense reads are printed in black font and the antisense reads are in red font. There are many overlapping 32-bp reads aligned uniquely to the transcript, with a quality score for each nucleotide. There is a SNP (A/G) located in the middle. By directly counting the number of reference and alternative nucleotides at the SNP, we were able to quantify the relative expression level of the two parental alleles.



```
HWI-EAS110__13603_4_260_486_56                CTCTTCCCTACTGTCTTCAGTGGGATCCCCTC
HWI-EAS110__13603_6_112_576_71                CTCTTCCCTACTGTCTTCAGTGGGATCCCCTC
HWI-EAS110__13603_6_76_929_610                CTCTTCCCTACTGTCTTCAGTGGCATCCCCTC
<< HWI-EAS110__13603_8_110_46_813             CTCTTCCCTACTGTCTTCAGTGGGATCCCCTC
HWI-EAS110__13603_2_91_981_217                  CTTCCCTACTGTCTTCAGTGGGATCCCCTCTT
<< HWI-EAS110__13603_2_182_946_247                GATCT-CTGTCTTCAGTGGGATCCCCTCTTAGG
<< HWI-EAS110__13603_3_286_258_466                TCCCTACTGTCTTCAGTGGGATCCCCTCTTGG
<< HWI-EAS110__13603_2_195_730_982                 CCCTACTGTCTTCAGTGGGATCCCCTCTTGGG
<< HWI-EAS110__13603_4_7_496_831                   CCCTACTGTCTTCAGTGGGATCCCCTCTTAGG
<< HWI-EAS110__13603_7_256_36_446                   CCTACTGTCTTCAGTGGGATCCCCTCTTGGGT
<< HWI-EAS110__13603_7_43_105_565                   CCTACTGTCTTCAGTGGGATCCCCTCTTAGGT
HWI-EAS110__13603_1_30_833_472                       CTACTGTCTTCAGTGGGATCCCCTCTTGGGTC
HWI-EAS110__13603_3_210_208_574                       TACTGTCTTCAGTGGGATCCCCTCTTAGGTCC
<< HWI-EAS110__13603_3_239_361_161                    TACTGTCTTCAGTGGGATCCCCTCTTAGGTCC
<< HWI-EAS110__13603_4_181_270_651                    GACTGACTTCAGGGGGATCCCCTCTTGGGTCC
HWI-EAS110__13603_6_237_643_999                       TACTGTCTTCAGTGGGATCCCCTCTTAGGTCC
HWI-EAS110__13603_8_115_225_590                        ACTGTCTTCAGTGGGATCCCCTCTTGGGTCCC
<< HWI-EAS110__13603_2_118_622_651                        GTCTTCAGTGGGATCCCCTCTTGGGTCCCTCC
HWI-EAS110__13603_2_176_919_480                           GTCTTCAGTGGGATCCCCTCTTAGGTCCCTCC
HWI-EAS110__13603_6_131_838_853                           GTCTTCAGTGGGATCCCCTCTTGGGTCCCTCC
HWI-EAS110__13603_6_39_630_770                            GTCTTCAGTGGGATCCCCTCTTAGGTCCCTCC
HWI-EAS110__13603_8_120_703_566                           GTCTTCAGTGGGATCCCCTCTTAGGTCCCTCC
<< HWI-EAS110__13603_4_24_28_536                           TCTTCAGTGGGATCCCCTCTTGGGTCCCTCCC
<< HWI-EAS110__13603_6_68_226_932                           CTTCAGTGGGATCCCCTCTTGGGTCCCTCCCT
<< HWI-EAS110__13603_8_3_837_155                            CTTCAGTGGGATCCCCTCTTGGGTCCCTCCCT
HWI-EAS110__13603_1_242_813_762                              TTCAGTGGGATCCCCTCTTGGGTCCCTCCCTT
<< HWI-EAS110__13603_3_208_522_288                           TTCAGTGGGATCCCCTCTTGGGTCCCTCCCTT
<< HWI-EAS110__13603_6_29_474_982                            TTCAGTGGGATCCCCTCTTAGGTCCCTCCCTT
<< HWI-EAS110__13603_7_93_685_395                            TTCAGTGGGATCCCCTCTTAGGTCCCTCCCTT
<< HWI-EAS110__13603_8_113_519_202                           TTCAGTGGGATCCCCTCTTAGGTCCCTCCCTT
<< HWI-EAS110__13603_1_155_883_891                            TCAGTGGGATCCCCTCTTAGGTCCCTCCCTTT
<< HWI-EAS110__13603_2_146_887_624                            TCAGGGGGATCCCCTCTTGGGTCCCTCCCTTT
<< HWI-EAS110__13603_2_35_69_950                              TCAGTGGGATCCCCTCTTAGGTCCCTCCCTTT
HWI-EAS110__13603_4_197_600_213                               TCAGTGGGATCCCCTCTTGGGTCCCTCCCTTT
HWI-EAS110__13603_8_89_199_703                                TCAGTGGGATCCCCTCTTGGGTCCCTCCCTTT
HWI-EAS110__13603_1_59_933_698                                 CAGTGGGATCCCCTCTTGGGTCCCTCCCTTTG
HWI-EAS110__13603_3_111_687_884                                CAGTGGGATCCCCTCTTAGGTCCCTCCCTTTG
<< HWI-EAS110__13603_8_119_246_260                             CAGTGGGATCCCCTCTTGGGTCCCTCCCTTTG
<< HWI-EAS110__13603_8_266_245_943                             CAGTGGGAGCCCCTCTTGGGTCCCTCCCTTTG
<< HWI-EAS110__13603_6_30_488_943                               AGTGGGATCCCCTCTTAGGTCCCTCCCTTTGT
HWI-EAS110__13603_8_157_792_533                                 AGTGGGATCCCCTCTTGGGTCCCTCCCTTTGT
NM_010514:                          TCCATCTTGTCTCTTCCCTACTGTCTTCAGTGGGATCCCCTCTTGGGTCCCTCCCTTTGTCATCATGTGAAGACTTCCCACG
Pos: 4565/4732  Qual: 15 (C: 40)
```

A total of 17 of the 26 candidate genes were verified to be imprinted by a combination of Sanger sequencing and Pyrosequencing. Of these, 14 are known imprinted genes. *Nnat* (*Peg5*), *Inpp5f_v2*, *Rasgrf1*, *Zrsr1* (*U2af1-rs1*), *Snrpn* and *Snurf* genes are known to be imprinted in mouse neonatal brain with paternal-only expression (Table 14) (Leff et al. 1992; Plass et al. 1996; Kagitani et al. 1997; Wang et al. 2004; Choi et al. 2005), and this was confirmed by both the Illumina sequence data and by Sanger sequencing and Pyrosequencing (Figure 15, Figure 16,Figure 17,Figure 18, Figure 19, Figure 20 and Figure 21). *Neuronatin* (*Nnat*), a gene on mouse chromosome 2, is known to be imprinted in mouse neonatal brain (Kagitani et al. 1997). In our data, *Nnat* showed 100% paternal monoallelic expression, with a *q*-value of zero (Table 14). Four SNPs within the last exon of the gene were covered by the Illumina reads. All of them showed 100% paternal expression as scored in 3,057 observed paternal allele-bearing reads in both reciprocal F1s (Figure 22A), a result verified by Sanger sequencing (Figure 22B) and by Pyrosequencing (Figure 22C).

Figure 15. SNP counts in the Illumina data for *Inpp5f* and *Inpp5f_v2*. Allele counts for Perlegen SNP NES16063345, NES16063347, NES16063348, NES16063351, NES16063353, NES16063354 and NES16063356. The blue bars (from left to right) stand for the counts from the paternal allele in PWD x AKR and AKR x PWD F1s respectively. The red bars represent the maternal allele counts.

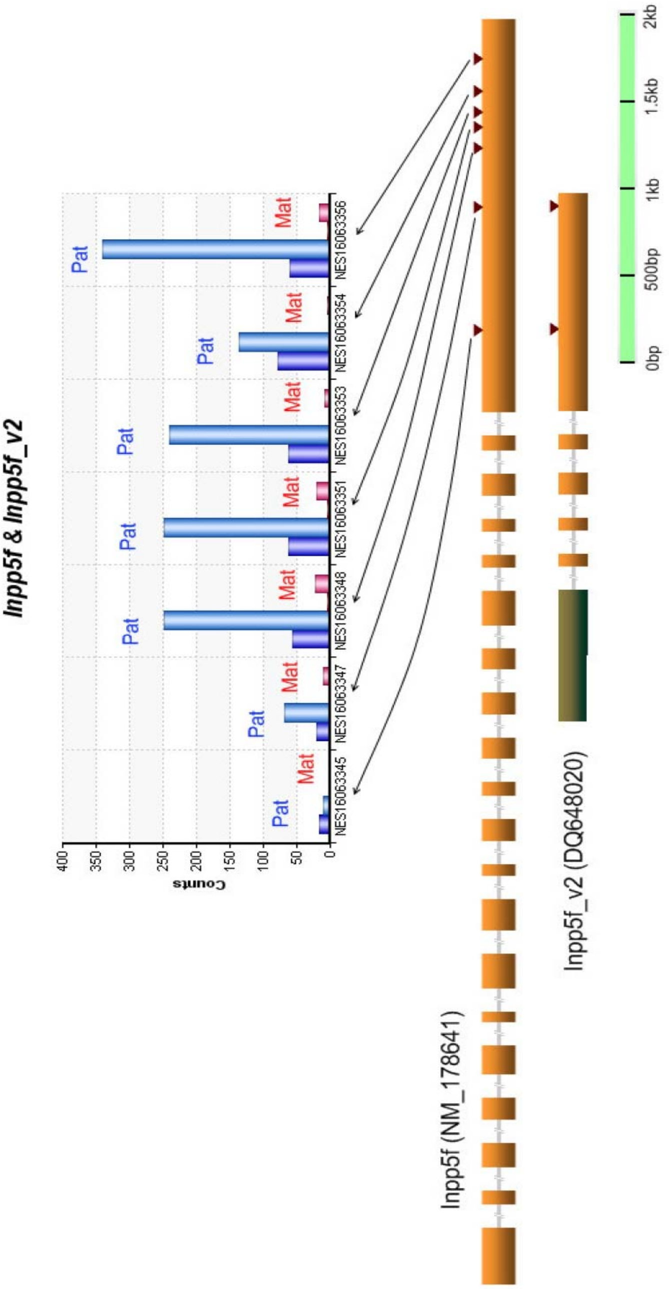


Figure 16. Sanger sequencing verification for *Inpp5f* at Perlegen SNP NES16063356. The target sequence is CTCTGA(C/T)AAGCA.

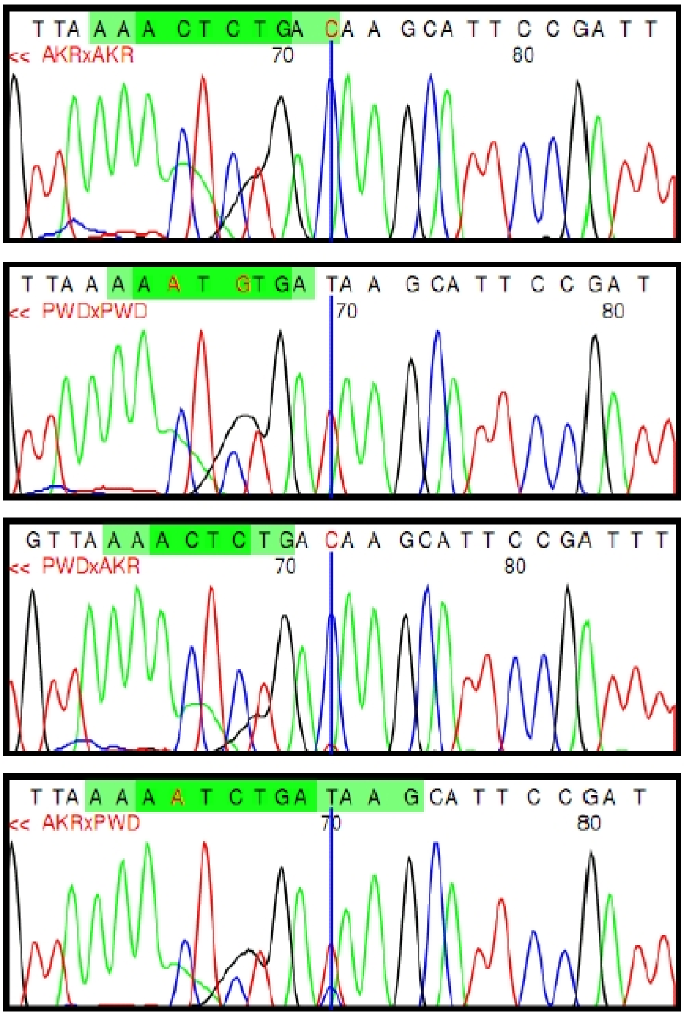


Figure 17. Pyrosequencing verification for *Inpp5f* at Perlegen SNP NES16063356. The target sequence is CTCTGA(C/T)AAGCA.

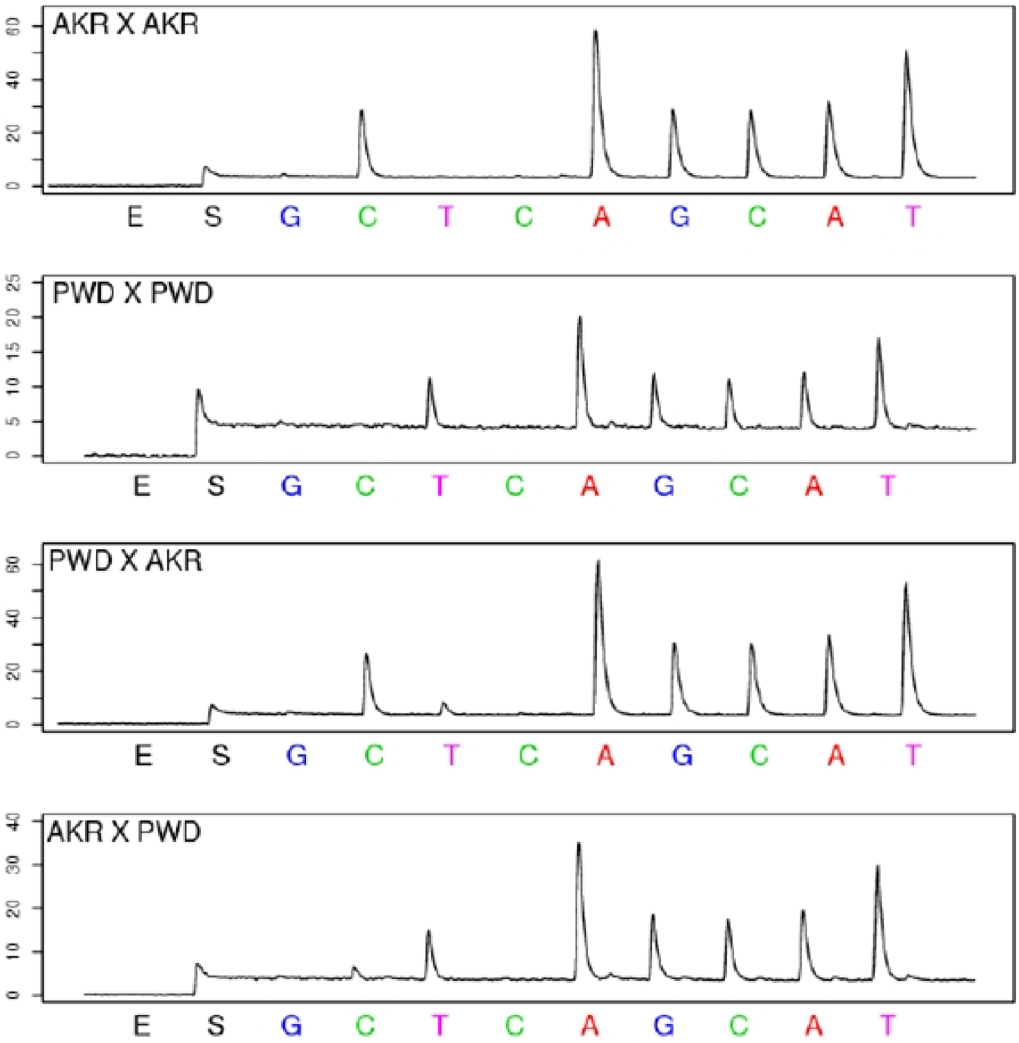


Figure 18. Pyrosequencing sequencing verification of *Inpp5f* for Perlegen SNP NES16063345. The target sequence is CGGTC(C/T)CAGTCT.

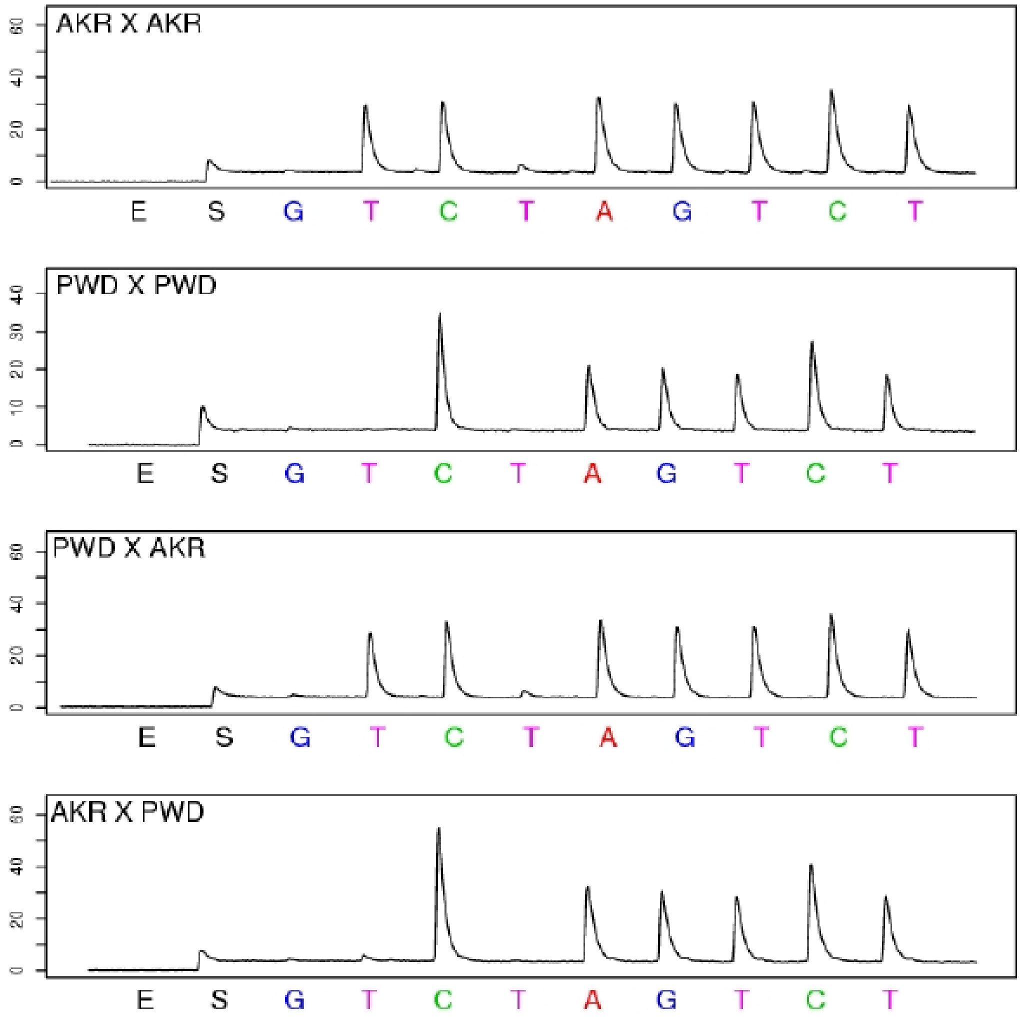


Figure 19. Pyrosequencing verification for *Rasgrf1*. The target sequence is T(C/T)ACGGGACAA.

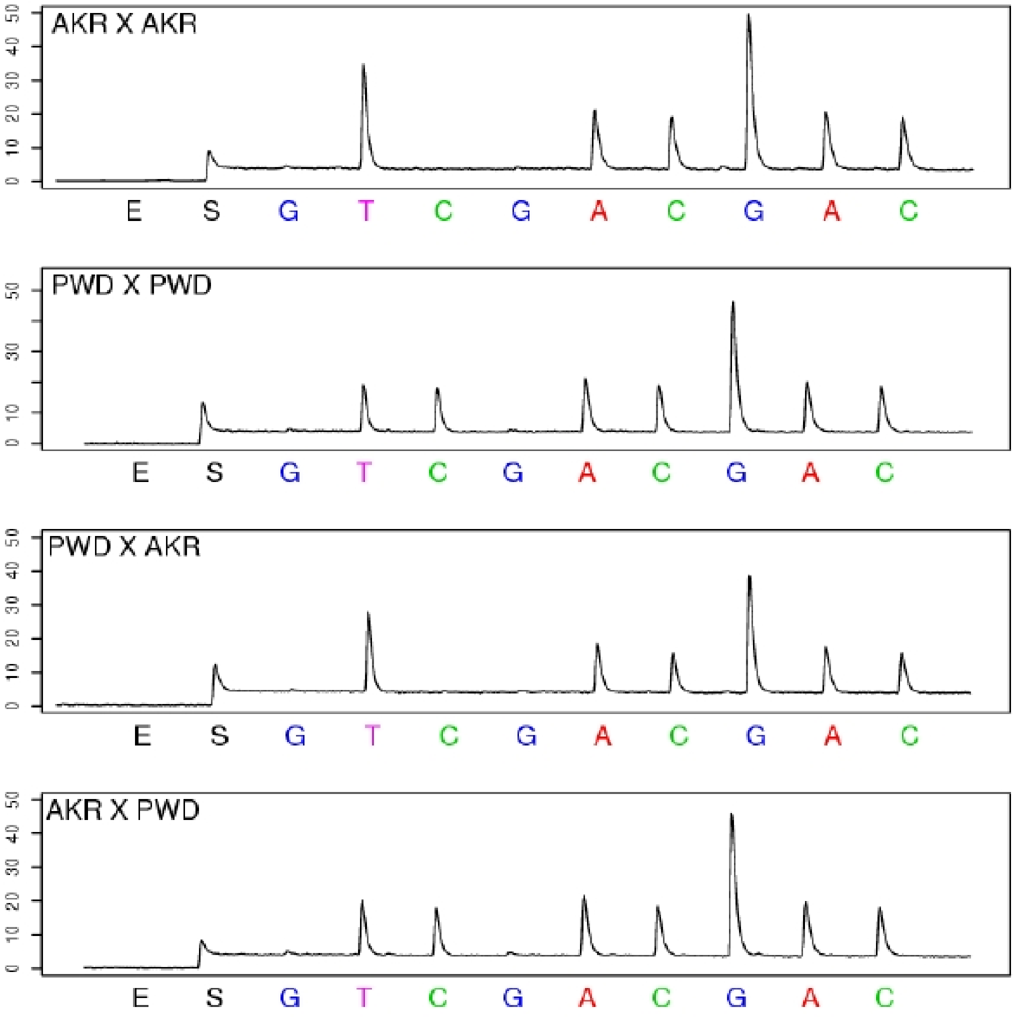


Figure 20. Sanger sequencing and pyrosequencing verification for *Zrsr1* at Perlegen SNP NES08366940. The target sequence is GGTAAAA(A/G)CTCAGA.

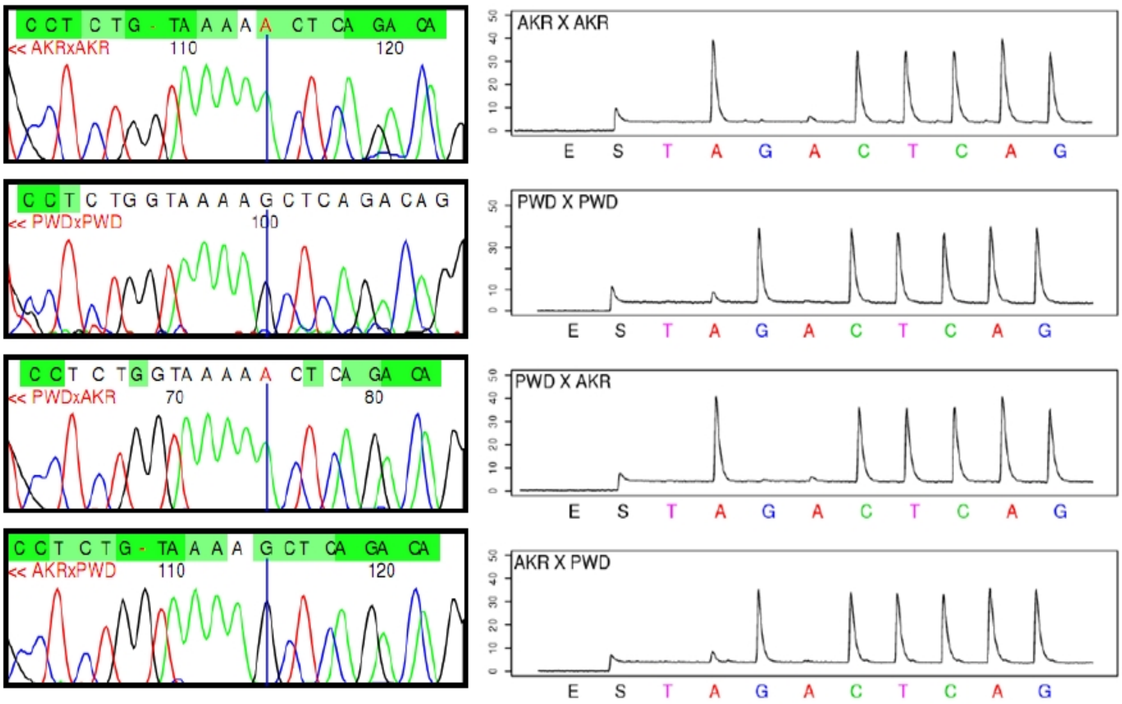


Figure 21. Pyrosequencing sequencing verification of *Snrpn-Snurf* at Perlegen SNP NES16116930. The target sequence is (G/T)GAAACCAAGTTCT.

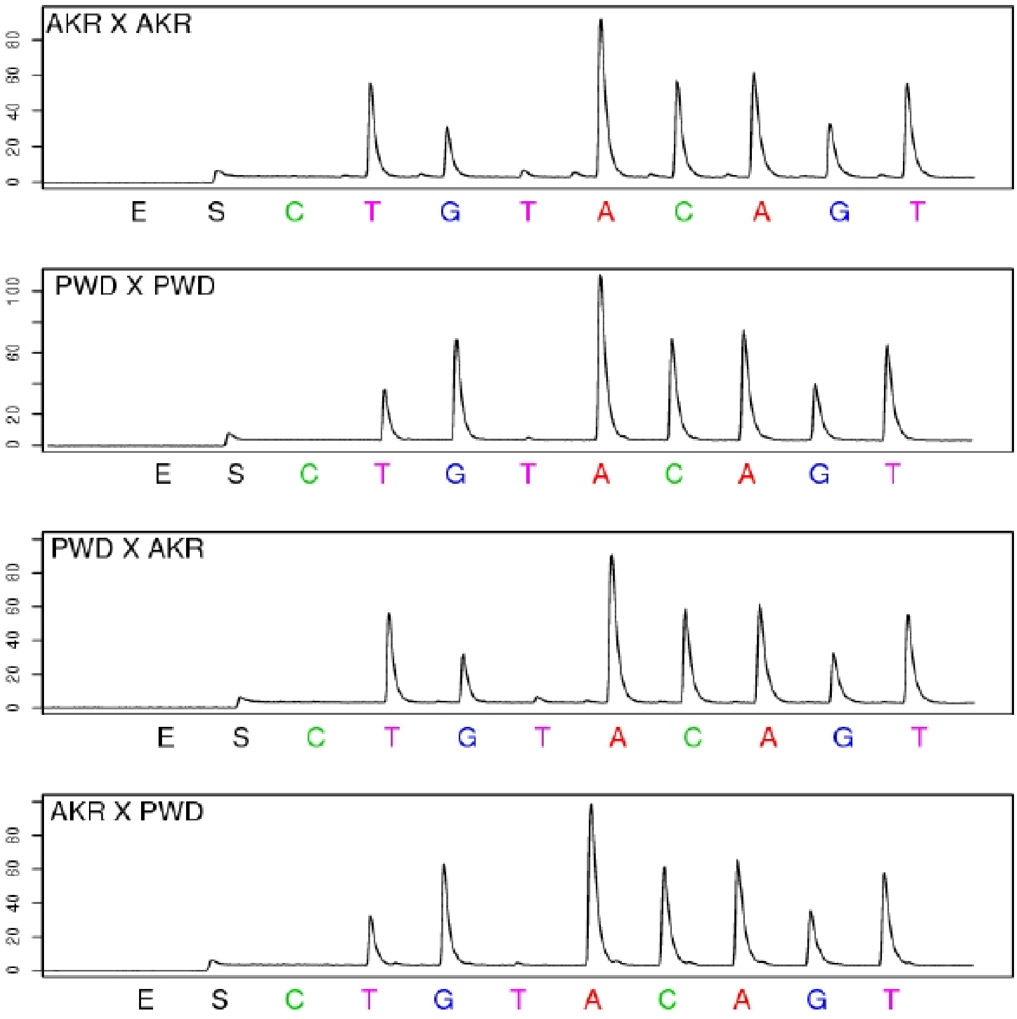
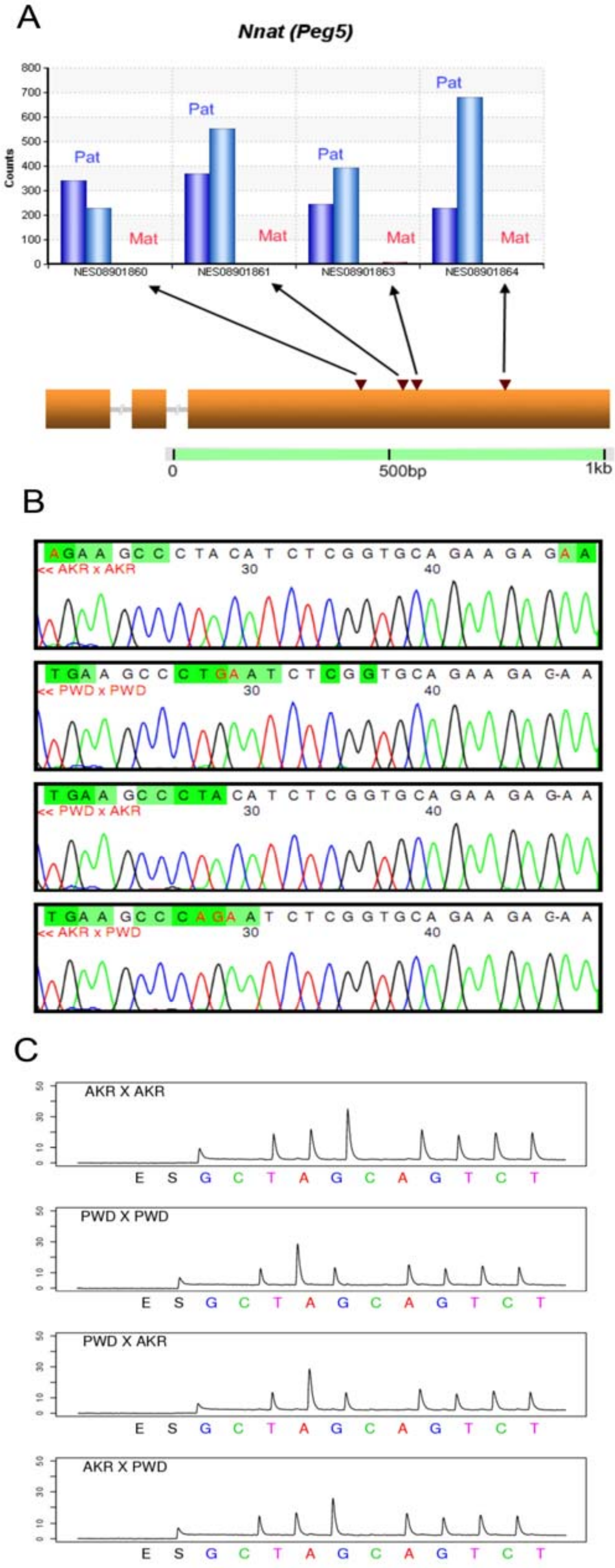


Figure 22. Verification for known imprinted gene *Nnat* (also known as *Peg5*).

(A). Allele counts for Perlegen SNP NES08901860, NES08901861, NES08901863 and NES08901864. The blue bars (from left to right) represent the Illumina read counts from the paternal allele in PWD x AKR and AKR x PWD F1s respectively (maternal genotype listed first). The red bars represent the maternal allele Illumina read counts.

(B). Sanger sequencing verification for Perlegen SNP NES08901861. We discovered an adjacent SNP position before NES08901861. The target sequence is GCCCT(AC/GA)ATCT.

(C). Pyrosequencing verification for Perlegen SNP NES08901861. The target sequence is GCCCT(AC/GA)ATCT.


A
Nnat (Peg5)
Counts
800
700
600
500
400
300
200
100
0
Pat
Mat
Pat
Mat
Pat
Mat
Pat
Mat
NES08901860
NES08901861
NES08901863
NES08901864
0
500bp
1kb
B
AGAA GCC CTA CA T CT CGGTGCA GA A GA GAA
<< AKR x AKR
30
40
TGAA GCC CTGAAT CT CGGTGCA GAA GA G-A A
<< PWD x PWD
30
40
TGAA GCC CTA CA T CT C GGTGCA GA A GA G-A A
<< PWD x AKR
30
40
TGAA GCC CAGAA T CT CGGTGCA GA A GA G-A A
<< AKR x PWD
30
40
C
AKR X AKR
50
30
10
0
E S G C T A G C A G T C T
PWD X PWD
50
30
10
0
E S G C T A G C A G T C T
PWD X AKR
50
30
10
0
E S G C T A G C A G T C T
AKR X PWD
50
30
10
0
E S G C T A G C A G T C T

The imprinting status of seven known imprinted genes have not been previously reported in neonatal brain, including the paternally expressed *Peg13*, *Sgce*, *Impact* and *Nap1l5* (Table 14; Figure 23, Figure 24, Figure 25 and Figure 26) (Piras et al. 2000; Smith et al. 2003) and the maternally expressed *Gtl2* (*Meg3*), *H19* and *Cdkn1c* (*P57*^{KIP2}) (Table 14; Figure 27 and Figure 28) (Hatada and Mukai 1995; Hagiwara et al. 1997; Hemberger et al. 1998; Schmidt et al. 2000). Our data support their being imprinted in P2 neonatal brain (Table 14). *Gtl2* (also known as *Meg3*) is a non-coding RNA gene on mouse chromosome 12, and it is reported to be imprinted in mouse placenta (Schmidt et al. 2000). Although the expression pattern of *Gtl2* has been determined in brain (Yevtodiyenko et al. 2004; McLaughlin et al. 2006), the imprinting status was not tested in neonatal brain. There is no Perlegen SNP covered in the Illumina data, but from the assembly of the Illumina reads, 4 novel SNPs were discovered and it is suggested that the *Gtl2* transcript (XR_035484) is expressed only from the maternal allele (Figure 29A). This is confirmed by Pyrosequencing (Figure 29B). Another splicing variant of *Gtl2*, NM_144513, was identified to be imprinted in our custom Agilent microarray survey of novel imprinted genes (See Chapter 1), with 1,847-fold difference in probe intensity in PWD x AKR cross and 793-fold in the reciprocal cross. A Perlegen SNP (NES17649478) in NM_144513 but not XR_035484 was verified by Pyrosequencing (Figure 29C). The analysis shows unambiguously that both splice variants are imprinted. Careful examination of the *in situ* images of fetal brain of uniparental disomic mice are consistent with our findings and suggest that there is maternal expression only (da Rocha et al. 2007).

Figure 23. Pyrosequencing sequencing verification for *Peg13*. The target sequence is TAG(C/T)TATAG.

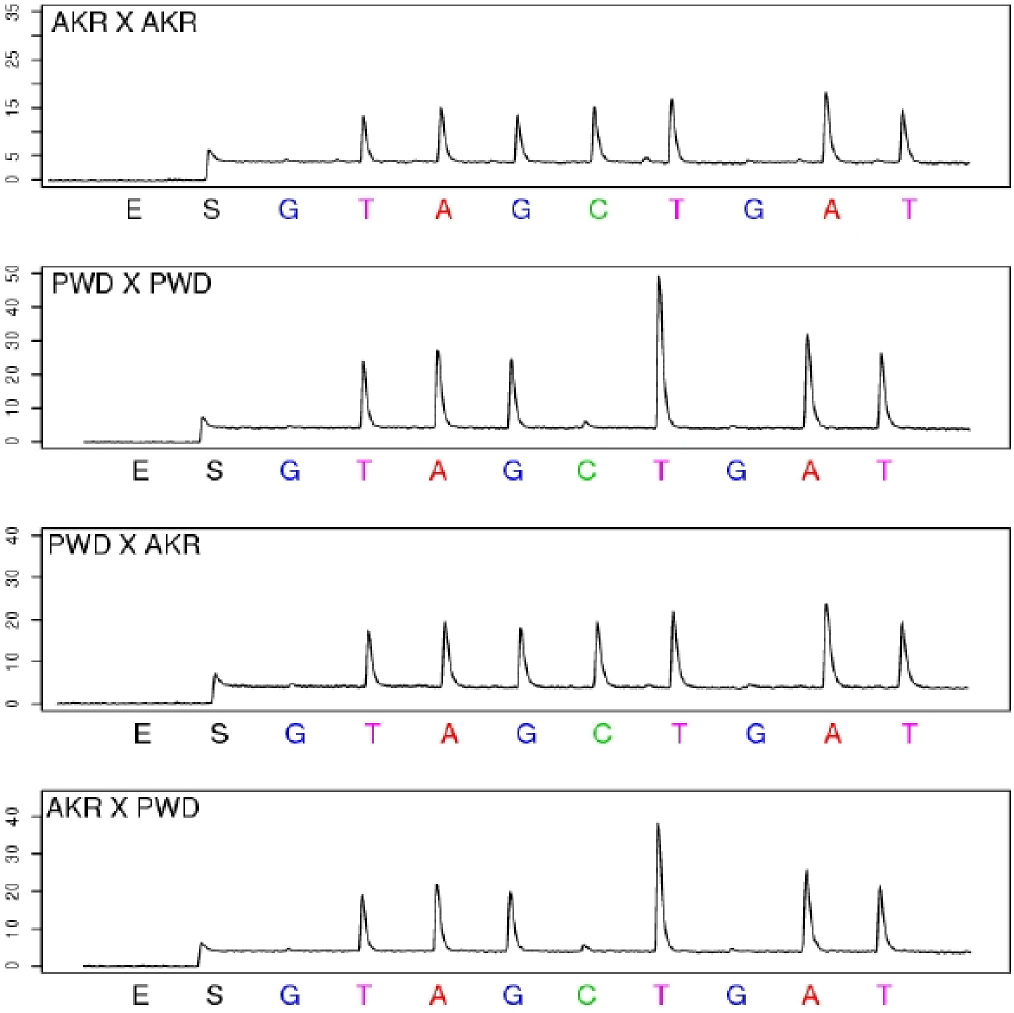


Figure 24. Sanger sequencing (left) and pyrosequencing verification (right) for known imprinted gene *Sgce* at Perlegen SNP NES10338539. The target sequence is ACC(G/A)TGACACA.

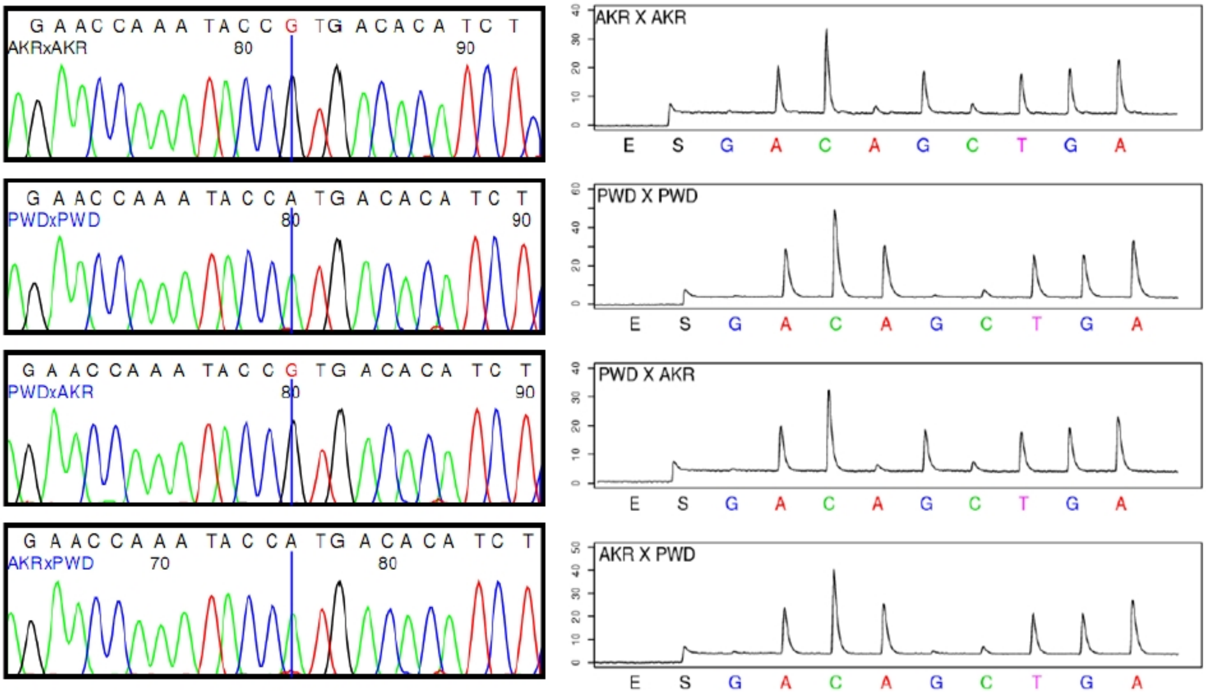


Figure 25. Pyrosequencing sequencing verification for known imprinted gene *Nap1l5*. The target sequence is AAT(A/G)CAAATATTTA.

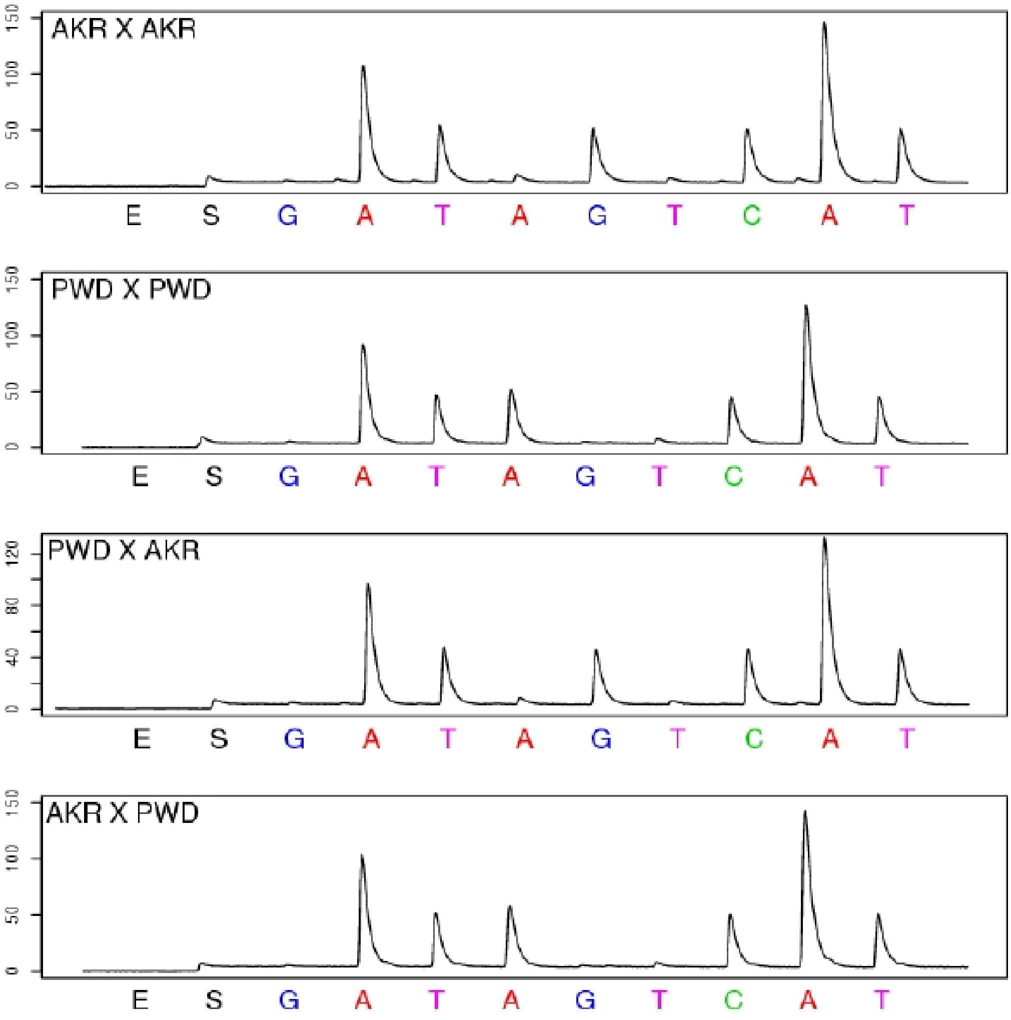


Figure 26. Pyrosequencing sequencing verification for known imprinted gene *Impact* at Perlegon SNP NES12698107. The target sequence is (G/A)TTCCTCAC.

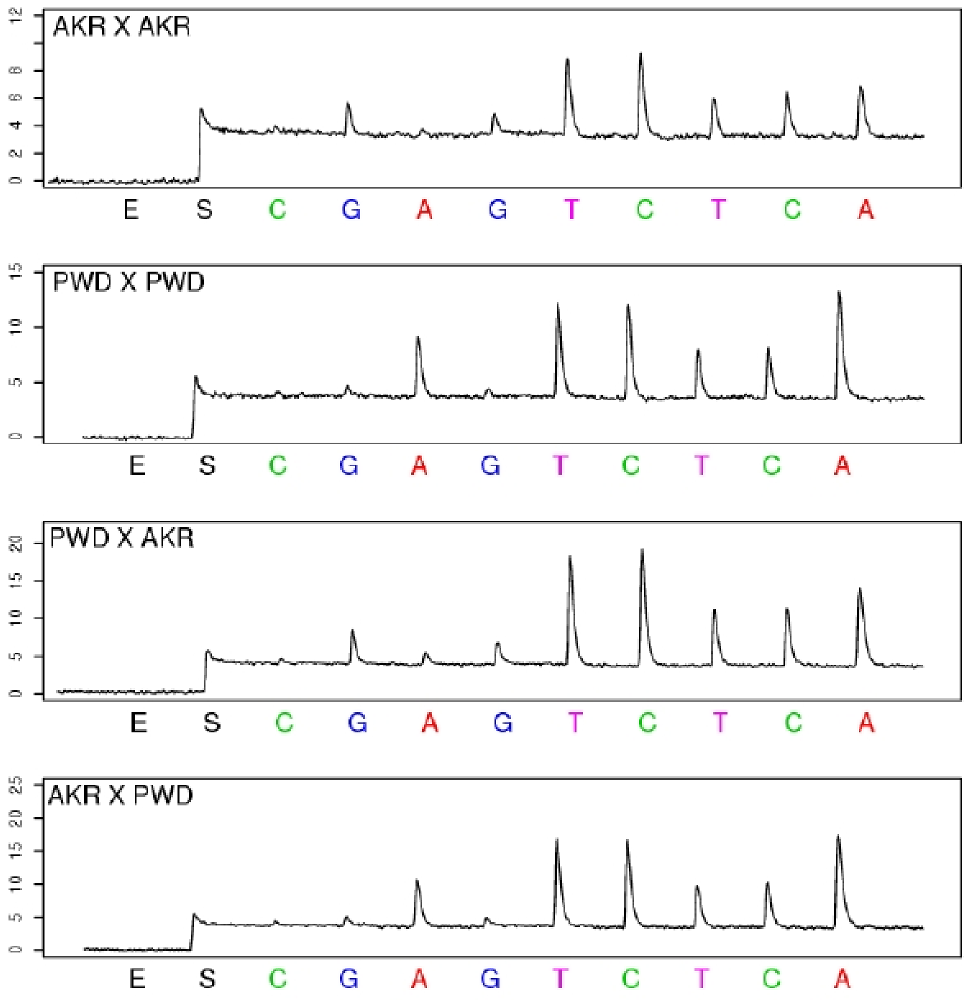


Figure 27. Pyrosequencing sequencing verification for *H19*. The target sequence is C(G/A)TCCATC.

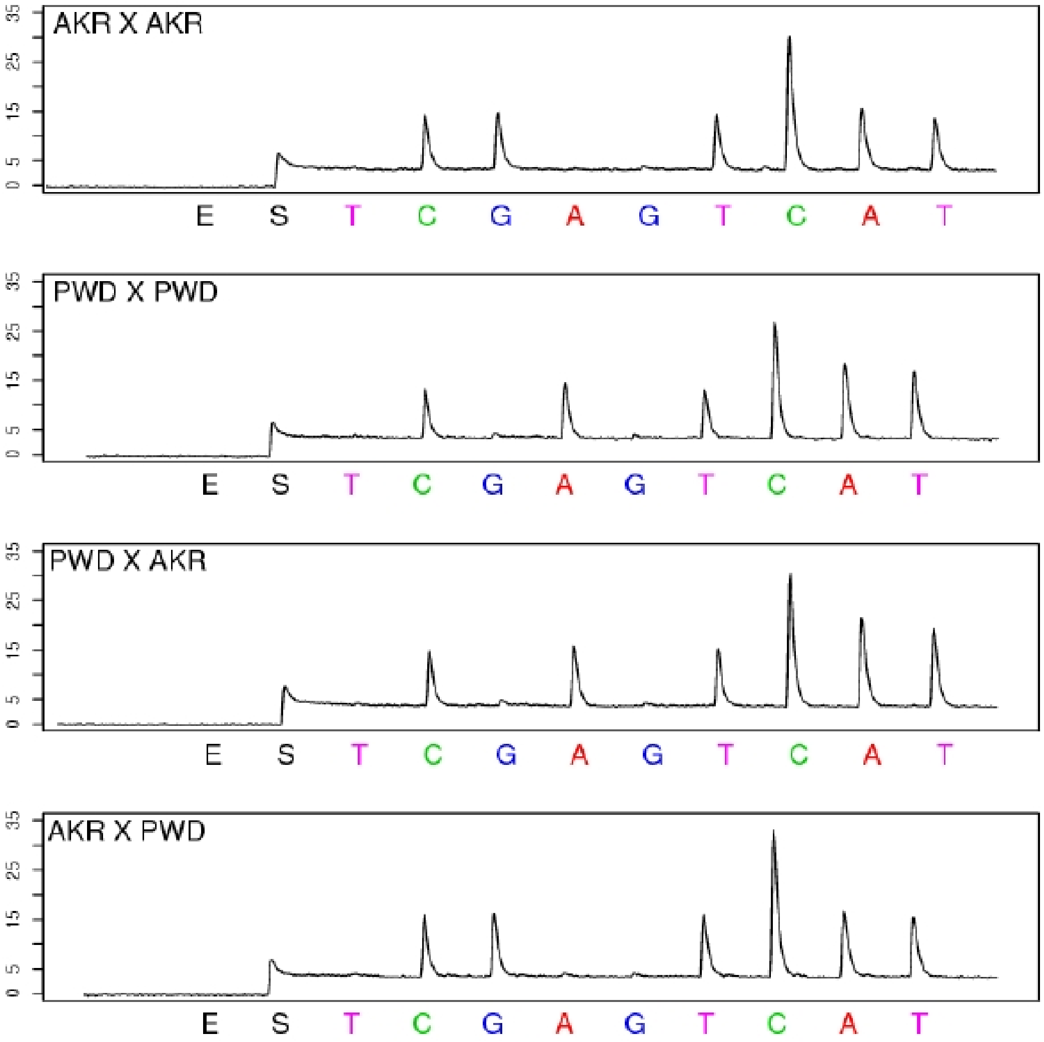


Figure 28. Pyrosequencing sequencing verification for known imprinted gene *Cdkn1c*. The target sequence is C(C/T)ACTTCAT.

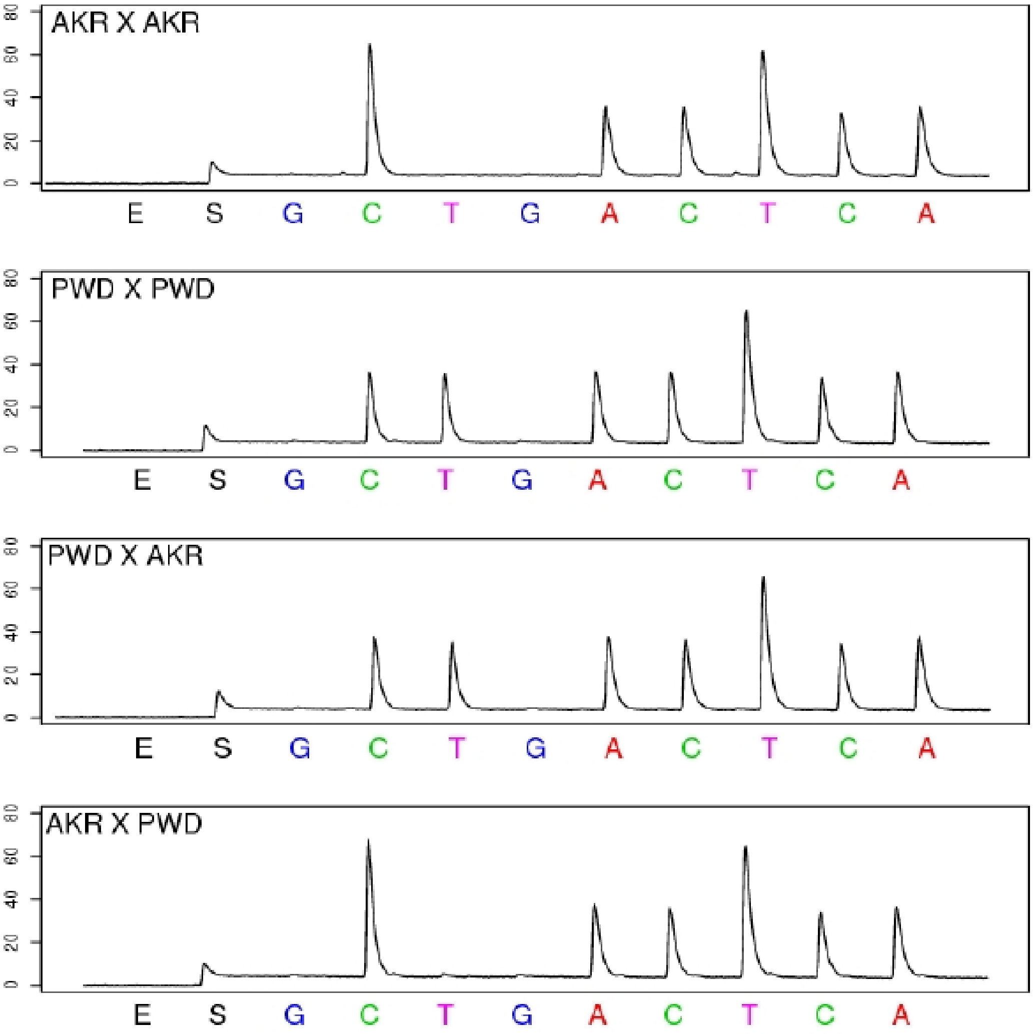
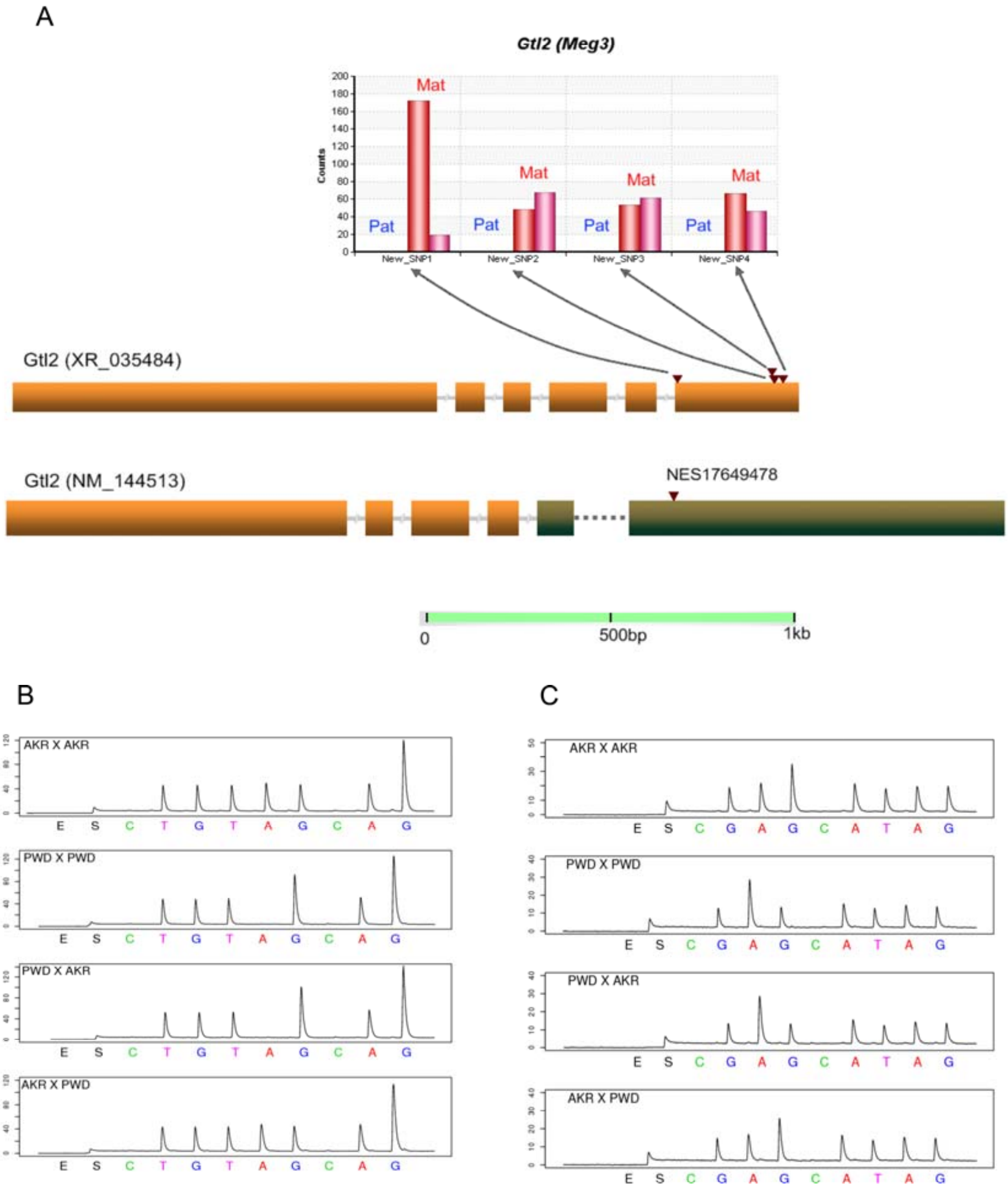


Figure 29. Verification for the known imprinted gene *Gtl2*.

(A). Allele counts for the 4 new SNPs discovered by assembling the Solexa reads. The blue bars (from left to right) stand for the counts from the paternal allele in PWD x AKR and AKR x PWD F1s respectively. The red bars represent the maternal allele counts. Four novel SNPs were discovered in one Gtl2 transcript (XR_035484), consistent with monoallelic expression from the maternal allele in both reciprocal crosses and confirmed by Pyrosequencing. Another splicing variant of Gtl2, NM_144513, previously was found by us to be imprinted using a custom Agilent allele-specific microarray (See Chapter 1), with an 1,847-fold difference in probe intensity in PWD x AKR cross and 793-fold in the reciprocal cross. A Perlegen SNP (NES17649478) in NM_144513 but not XR_035484 was verified by Pyrosequencing.   We conclude that both XR_035484 and NM_144513 are imprinted in the neonatal brain.

(B). Pyrosequencing verification for a novel SNP in *Gtl2*. The target sequence is TGT(A/G)GAGGGA.

(C). Pyrosequencing verification for Perlegen SNP NES17649478. The target sequence is GA(A/G)GATAG.


A
Gtl2 (Meg3)
Counts
200
180
160
140
120
100
80
60
40
20
0
Mat
Pat
New_SNP1
New_SNP2
New_SNP3
New_SNP4
Gtl2 (XR_035484)
Gtl2 (NM_144513)
NES17649478
0
500bp
1kb
B
AKR X AKR
PWD X PWD
PWD X AKR
AKR X PWD
E S C T G T A G C A G
C
AKR X AKR
PWD X PWD
PWD X AKR
AKR X PWD
E S C G A G C A T A G

Known and novel imprinted genes identified

We also discovered three novel imprinted genes by Illumina short-read sequencing, and verified by Sanger and Pyrosequencing. According to Choi et al. (Choi et al. 2005), *Inpp5f* is a splicing variant of the known imprinted gene *Inpp5f_v2*, sharing 4 exons and part of the last exon. There are seven SNPs covered in the sequence data for *Inpp5f*, with 2 of them shared by *Inpp5f_v2*. Since all seven SNPs show significant paternal-excess in expression, we conclude that *Inpp5f* is also imprinted in P2 neonatal brain. Formally, it is also possible that *Inpp5f* and *Inpp5f_v2* share the same 3' end. Two CpG islands near the gene region were reported before (Choi et al. 2005). CpG1 is not methylated and CpG2 is the DMR (Differentially Methylated Region) with only the paternal allele being methylated. Two previously reported non-imprinted genes, *1810044A24Rik* (Davies et al. 2004) and *Blcap* (Evans et al. 2005), are found to be predominantly maternally expressed novel imprinted genes in our sequence data (*q*-value 0.0011 and 0.025) and Pyrosequencing verified that they showed 80% expression from the maternal allele (Figure 30)(Figure 31). The imprinting status of *1810044A24Rik* was also verified by Pyrosequencing in reciprocal crosses of C57BL/6 and C3H/HeJ (Figure 30). The imprinting status for *Blcap* was not verified in C57BL/6 and C3H/HeJ due to lack of exonic SNPs. Two known imprinted genes, *Peg13* and *Nnat*, are located in the introns of *1810044A24Rik* and *Blcap*, respectively. The CpG island of *Peg13* is only methylated at the maternal allele (Smith et al. 2003). Five differentially methylated CpG sites within the gene region of *Nnat* were previously identified (Kikyo et al. 1997; Smith et al. 2003), so each of the three novel

Figure 30. Verification for known imprinted gene *1810044A24Rik*. **(A).** Sanger sequencing verification for Perlegen SNP NES12099717. The target sequence is TCCATA(T/C)GCCATA. **(B).** Pyrosequencing sequencing verification for Perlegen SNP NES12097854 in AKR and PWD reciprocal cross. The target sequence is (A/G)TTCAGGA. **(C).** Pyrosequencing sequencing verification for Perlegen SNP NES12098495 in C3H and B6 reciprocal cross. The target sequence is AG(C/T)TGCTT.

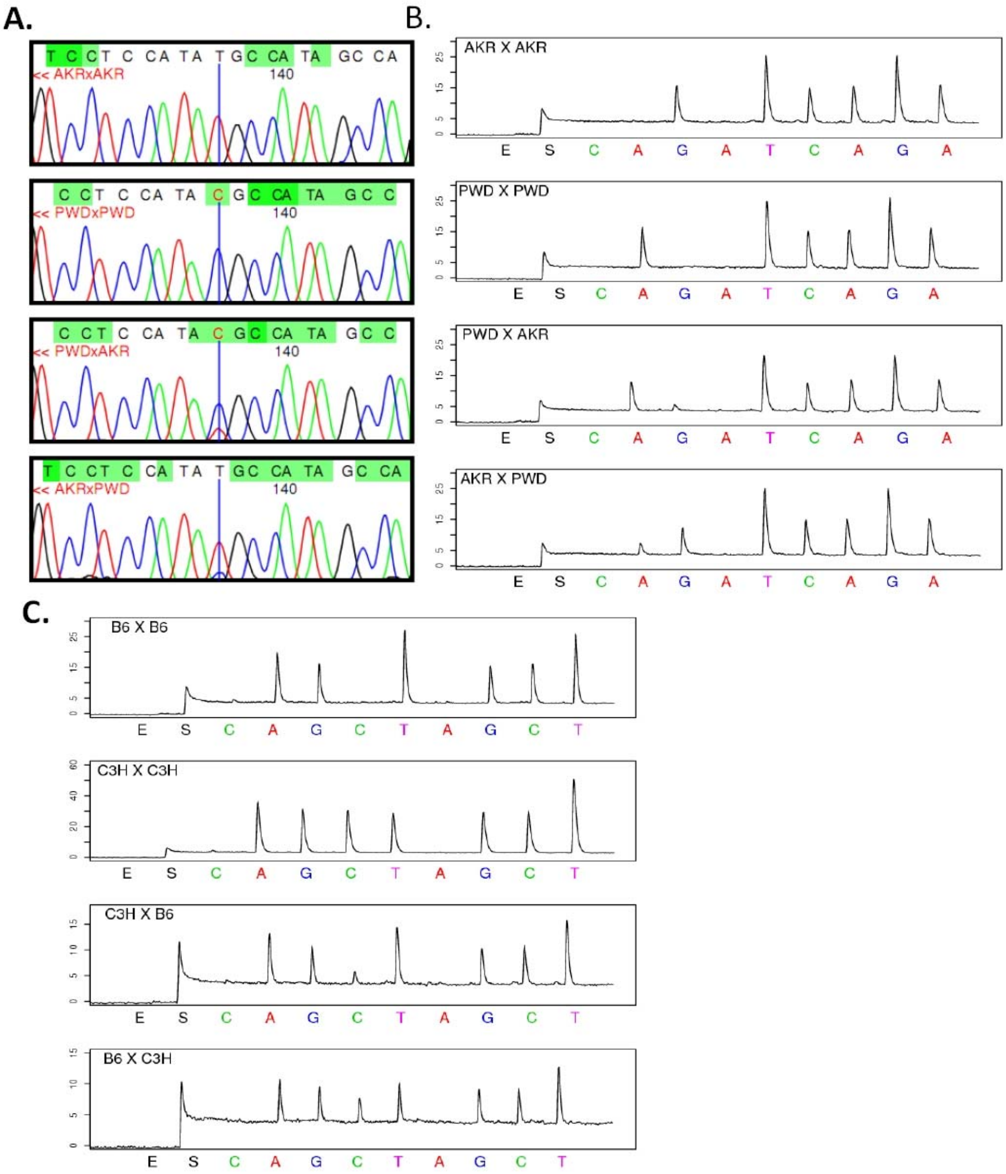
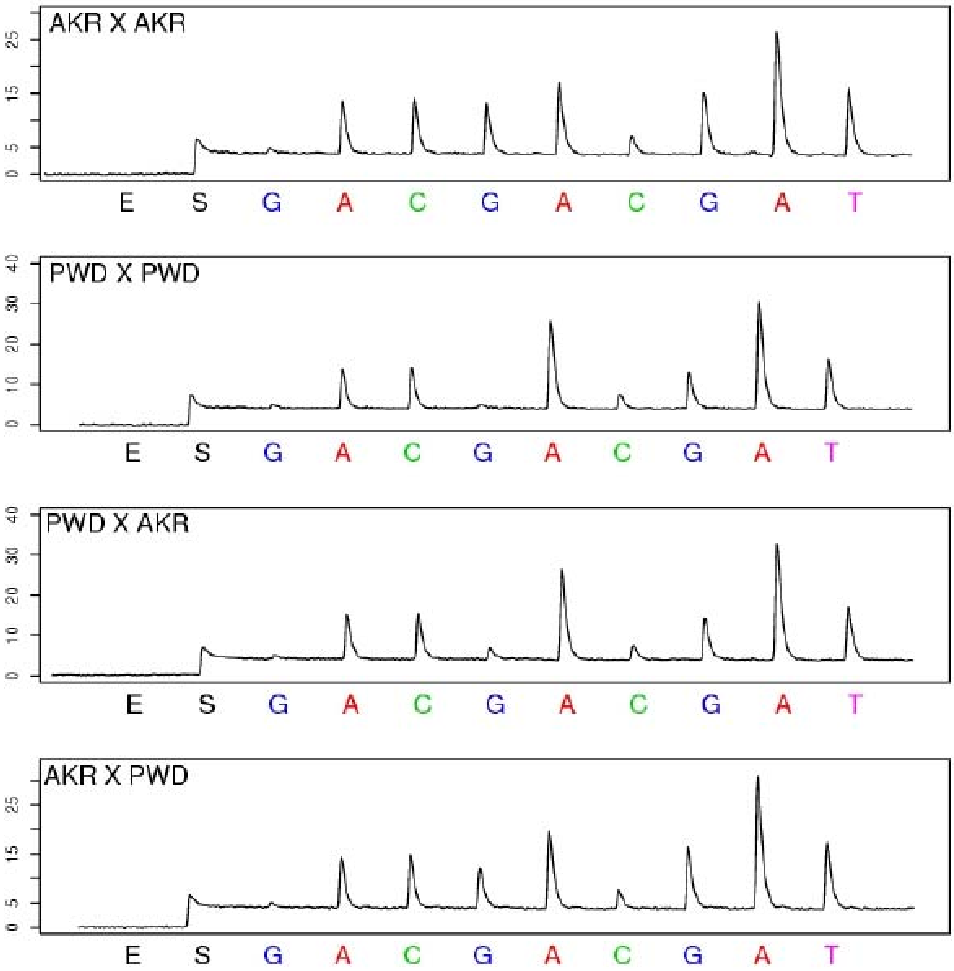


Figure 31. Pyrosequencing sequencing verification for novel imprinted gene *Blcap* at Perlegen SNP NES08901938. The target sequence is AC(A/G)AGAATA..



imprinted genes have DMRs near or within the gene regions. Nine genes attained marginal significance only after pooling across all SNPs, but showed no single SNP with a significantly skewed frequency. In all 9 cases, Pyrosequencing demonstrated unambiguously that they were not imprinted (Table 14).

Coverage of known imprinted genes in this study

Among the 98 known imprinted genes in mouse, 45 have both RefSeq ID and SNPs between AKR and PWD strains. 33 of the 45 known imprinted genes with SNPs were covered in our short-read sequence data. The remaining 12 genes were not covered by filtered high quality SNP-containing reads due to lack of detected expression in mouse neonatal brain (Table 15). 14 of 33 covered known imprinted genes are significant (Table 14). In the non-significant maternally expressed imprinted genes, *Ppp1r9a, Asb4*, *Calcr* and *Ube3a* have been reported as being imprinted in brain (Albrecht et al. 1997; Mizuno et al. 2002; Hoshiya et al. 2003; Ono et al. 2003), and they all have a marginally significant *P*-value. *Ube3a* imprinting was verified by Pyrosequencing. Genes that have too low a high-quality SNP-containing read count, such as *Gnas*, *Gatm*, *Tnfrsf23*, *Zim1*, *Dcn*, *Nap1l4*, *Osbpl5*, *Grb10* and *Slc22a2* have an imprinting status that remains inconclusive, but the data are not consistent with strong imprinting (Table 15). All known maternally expressed genes covered with adequate depth of sequence reads had a pattern of allelic bias consistent with their known imprinting status. *Gtl2*, *H19*, *Cdkn1c* and *Commd1* are significant in the Illumina data and they are verified to be imprinted in neonatal brain. *Ppp1r9a* has significant nominal *P*-value but is not significant after multiple test correction. However, the Illumina counts are consistent with preferential maternal expression (Table 16). *Asb4*, *Calcr*, *Ube3a* has marginal significant *P*-value due to the small number of SNP-containing reads covered in the data, suggesting that they might be imprinted in neonatal brain. We verified that *Ube3a* is imprinted in neonatal brain by the Pyrosequencing method, with the p_1 and p_2 ratios 0.392 and 0.755. The other genes covered in the data, *Gatm, Tnfrsf23, Zim1, Dcn, Nap1l4, Osbpl5,* and *Slc22a2* are not significant, which is consistent with the fact that they are known to be imprinted in placenta instead of neonatal brain (Table 16). *Gnas,* a known imprinted gene in the pituitary but not in the whole brain of mouse (Yu et al. 1998; Weinstein et al. 2000; Weinstein et al. 2001; Weinstein et al. 2004), is not statistically significant in the Illumina data. However, the Pyrosequencing verification showed 0.459/0.562 ratio of p_1 /p_2, suggesting that there is slightly higher expression from the allele inherited from mother. *Grb10* is imprinted in both placenta and brain (Blagitko et al. 2000; Mergenthaler et al. 2001; Hikichi et al. 2003) but does not show a significant difference between p_1 and p_2 in the Illumina data, despite adequate expression level to provide a test of adequate power. Subsequent Pyrosequencing verified the non-imprinting status in P2 neonatal brain (Table 16). In fact, *Grb10* is imprinted in mouse brain with paternal-only expression, but it shows maternal-only expression in other tissues (Hikichi et al. 2003). It could be possible that *Grb10* is imprinted in other stages of brain (for example, fetal brain) but not P2 brain in mouse, or it is possible that the imprinting status varies among strains, and the AKR x PWD F1 fail to imprint *Grb10*. For the paternally expressed known imprinted genes that are not statistically significant in our data, *Magel2* and *Peg3* are consistent with 100% paternal expression. *Rtl1* and *Copg2* may be maternally expressed, as suggested by the sequence count data, but there were too few reads to attain statistical significance. While *Copg2* is maternally expressed, and *Rtl1* is expressed from the paternally inherited allele, the microRNA-containing antisense transcript is expressed from the maternal allele (Seitz et al. 2003). *Igf2* and *Slc38a4* are consistent with non-imprinting and, consistent with the pattern of expression in human and mouse (DeChiara et al. 1991; Ohlsson et al. 1994; Jones et al. 2001; Charalambous et al. 2004),   *Igf2* is verified by Pyrosequencing to be biallelically expressed in brain (Figure 32 and Table 16).

Table 15. Summary of known imprinted genes covered in Illumina P2 brain reads.

Type	Count	Description
No_RefSeq	37	The gene is not in the RefSeq database
No_SNP	16	There is no SNP in the RefSeq transcript.
No_counts	12	There is SNP site within the transcripts, but there is no counts in the filtered Solexa data.
Covered	33	The gene is covered in the Solexa data.

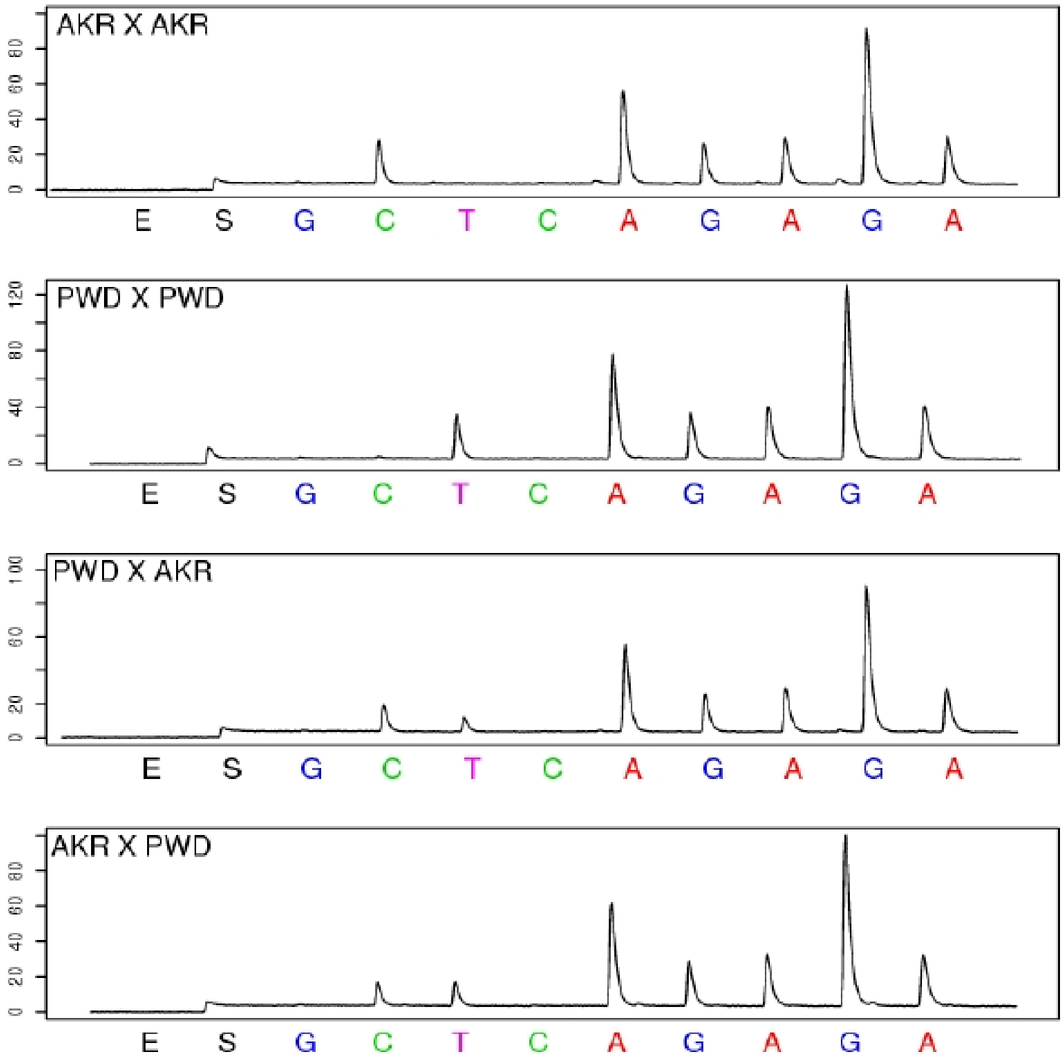
Table 16. Known imprinted genes covered in Illumina P2 brain data.

MGIsymbol	Exp. allele	PWD x AKR		AKR x PWD		Storer-Kim p-value	yrosquencin		Reported imprinting status in brain and placenta	# of perfectly matched reads
		AKR count	PWD count	AKR count	PWD count		p1	p2		
Gtl2 *	M	1	339	193	1	0	0.000	1.000	placenta	452
H19 *	M	2	14	61	1	1.68E-12	0.094	1.000	fetal brain	247
Cdkn1c *	M	0	8	13	0	9.46E-07	0.036	1.000	whole body of neonates	521
Commd1	M	12	33	22	7	2.52E-05	0.412	0.725	brain	574
Ppp1r9a *	M	36	54	67	26	1.18E-05			placenta, partially in neonatal brain	2117
Asb4	M	3	5	8	0	0.01292505			brain & placenta	158
Calcr	M	0	1	6	0	0.05666007			embryonic and adult brain	77
Ube3a *	M	3	10	4	2	0.08214854	0.392	0.755	brain	1660
Gnas *	M	181	214	101	96	0.2167333	0.459	0.562	other (embryos), Imprinted in pituitary	15998
Gatm	M	4	5	15	7	0.22626793			placenta	637
Tnfrsf23	M	2	2	2	0	0.37311385			placenta	17
Zim1	M	0	1	2	1	0.5			other tissue, biallelic in neonatal brain	19
Dcn	M	16	10	4	4	0.59255371			placenta	344
Nap1l4 *	M	9	12	181	214	0.7968901			placenta	1175
Osbpl5	M	3	2	5	4	0.96653274			placenta	157
Grb10 *	M	2	3	6	10	0.98294642	0.609	0.522	placenta and brain	508
Slc22a2	M	0	1	0	1	1			placenta	11
Nnat	P	1182	1	21	1853	0	1.000	0.000	neonatal brain	8561
Snurf	P	888	1	1	18	0	1.000	0.000	neonatal and adult brain	23679
Snrpn	P	889	1	1	19	0	1.000	0.000	neonatal and adult brain	23310
Peg13 *	P	168	0	6	74	0	0.988	0.030	adult brain	1088
Nap1l5 *	P	22	0	0	67	2.42E-22	1.000	0.000	adult brain	2329
Inpp5f_v2	P	38	3	14	80	4.29E-19	0.919	0.078	neonatal brain	5509
Sgce	P	9	0	0	54	6.01E-12			adult brain & placenta	313
Rasgrf1 *	P	16	0	0	20	2.56E-11	1.000	0.000	neonatal brain	93
Impact	P	15	6	8	83	6.24E-09	0.791	0.198	embryonic and adult brain	1953
Zrsr1	P	11	0	1	14	4.48E-07	0.975	0.004	embryonic, neonatal and adult brain	204
Magel2	P	2	0	0	5	0.01614973			adult brain	171
Rtl1	P	0	3	2	0	0.0576			brain and placenta	62
Peg3	P	2	0	0	2	0.125			embryonic, neonatal and adult brain	52
Igf2 *	P	52	43	20	27	0.17846458	0.641	0.499	placenta, biallelic in fetal brain	247
Copg2	P	0	1	2	0	0.22222222			embryonic, neonatal and adult brain	302
Slc38a4	P	2	2	12	6	0.58003101			brain and placenta	45

*: assembly information used.

■	known maternally expression imprinted genes that are significant in the Illumina data
■	known maternally expression imprinted genes that are not significant in the Illumina data
■	known paternally expression imprinted genes that are significant in the Illumina data
■	known paternally expression imprinted genes that are not significant in the Illumina data

Figure 32. Pyrosequencing sequencing verification for *Igf2*.The target sequence is (C/T)AAGAGGGGAT.



Closely-linked pairs of imprinted genes

Of the 10 sense-antisense pairs of known imprinted genes discovered to date (Morison et al. 2005), eight pairs are reciprocally imprinted (maternal expression for sense transcripts and paternal expression for antisense transcripts, or vice versa) (Barlow et al. 1991; Kay et al. 1994; Albrecht et al. 1997; Gould and Pfeifer 1998; Paulsen et al. 1998; Hu et al. 1999; Peters et al. 1999; Kim et al. 2000; Lee et al. 2000; Chamberlain and Brannan 2001; Kim et al. 2001; Sado et al. 2001; Fitzpatrick et al. 2002; Coombes et al. 2003; Seitz et al. 2003) (Table 17). The remaining two show only paternal expression (DeChiara et al. 1991; Moore et al. 1997; Jong et al. 1999). These cases of imprinting all were discovered and verified individually in different samples using different mouse strains (Table 17). In our Illumina sequence data, three reciprocally expressed closely linked sense-antisense (or sense-sense) pairs were covered adequately to perform statistical analysis (Table 18). Four of them are known imprinted genes (*Peg13*, *Nnat*, *Zrsr1*, *Commd1*) and two (*1810044A24Rik*, *Blcap*) are among our verified novel imprinted genes. *Peg13*, *Nnat* and *Zrsr1* are located in an intron of *1810044A24Rik*, *Blcap* and *Commd*, respectively. Interestingly, in the three pairs, *Peg13-1810044A24Rik, Nnat-Blcap* and *Zrsr1-Commd1*, the first gene is a paternally expressed imprinted gene with 100% monoallelic expression, whereas the second gene is maternally expressed partially imprinted gene (Figure 33). The pattern is consistent with the possibility that the monoallelic expression of the paternally expressed sense transcripts might reduce the relative expression of the paternal copy of the antisense transcript, resulting in predominantly maternal expression. Our hypothesis is that the paternally expressed imprinted gene is driving the apparent imprinting of the maternal gene, presumably through transcriptional interference. While this reciprocal imprinting has been noted in the literature (Sleutels et al. 2002;

Sleutels et al. 2003; Wang et al. 2004) , this is the first genome-wide study identifying multiple, well quantified cases in mouse neonatal brains.

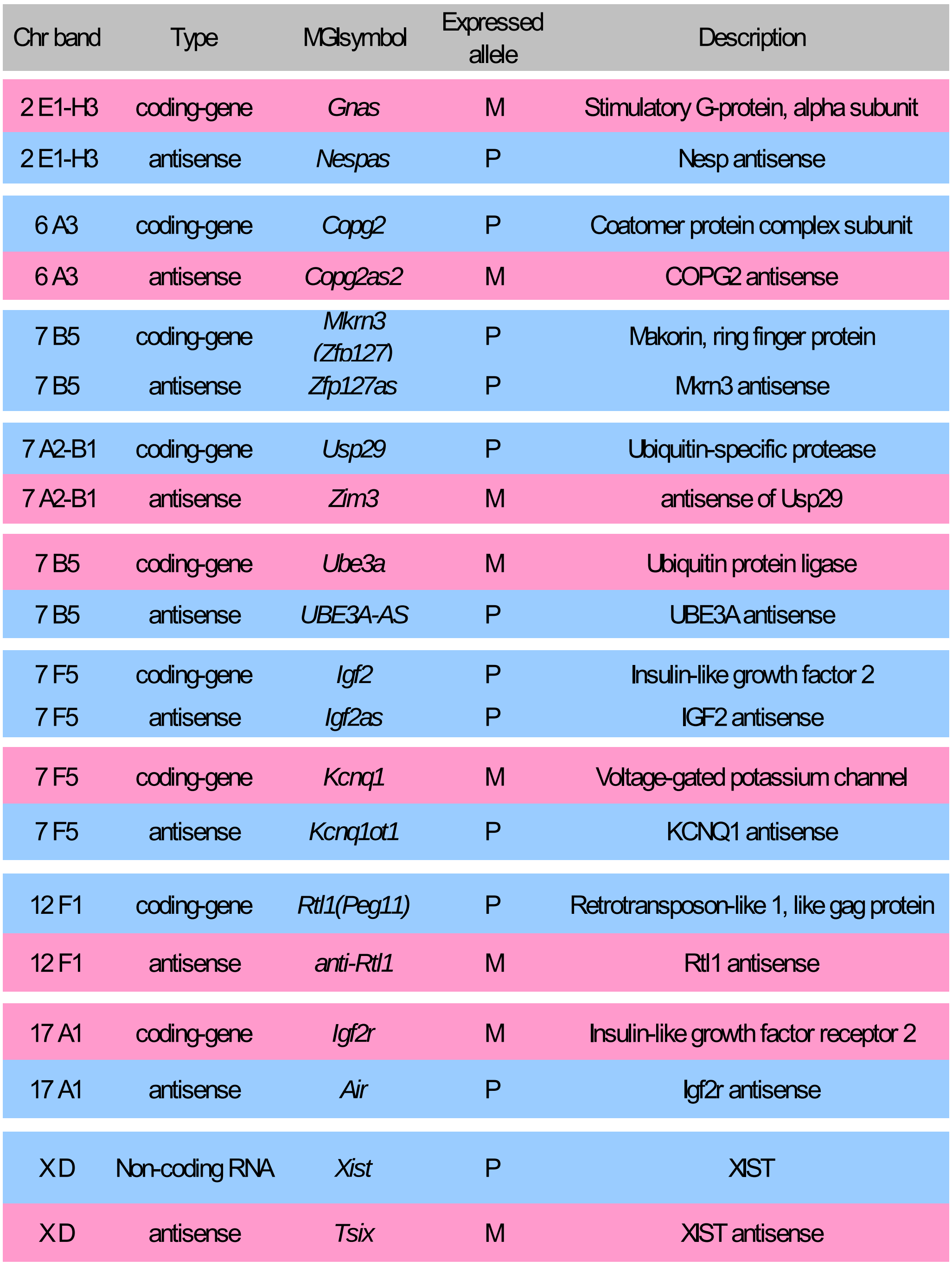
Transcriptome-wide pattern of imprinting status

To investigate the pattern of imprinting status for all the transcripts covered by our study, we plotted the 5,076 unique Entrez genes with counts of four or more in both reciprocal crosses across the mouse genome (Figure 34). We define imprinting status as the difference between the AKR percentages in the two reciprocal crosses, which is p_1-p_2 (Table 11). Most genes display a value of p_1-p_2 close to zero, indicating a lack of significant imprinting. The sense-antisense pairs and the imprinted genes in known imprinting clusters are clearly demonstrated in the genome-wide plots. There are 1,606 non-significant genes with a total count 25 or more in both reciprocal crosses, forming a good tissue-specific non-imprinted dataset for computational prediction and evolutionary analysis.

Paternal-brain and maternal placenta bias of imprinted genes

When paternally- and maternally-expressed imprinted genes covered in the sequence read data are compared, we discovered an excess of paternal expression (11 paternal and 6 maternal), and most of these (9 of 11) show strong monoallelic expression (90%-100%). Three of the maternally expressed genes are only partially imprinted in brain with 70%-80% expression from the maternal allele (Table 14). Overall there is a bias toward paternally expressed imprinted genes in brain, whereas of the 29 genes reported to be imprinted in placenta, only 8 are paternally expressed.

Table 17. Sense-antisense pairs in known imprinted genes.



Chr band	Type	MGIsymbol	Expressed allele	Description
2 E1-H3	coding-gene	*Gnas*	M	Stimulatory G-protein, alpha subunit
2 E1-H3	antisense	*Nespas*	P	Nesp antisense
6 A3	coding-gene	*Copg2*	P	Coatomer protein complex subunit
6 A3	antisense	*Copg2as2*	M	COPG2 antisense
7 B5	coding-gene	*Mkrn3 (Zfp127)*	P	Makorin, ring finger protein
7 B5	antisense	*Zfp127as*	P	Mkrn3 antisense
7 A2-B1	coding-gene	*Usp29*	P	Ubiquitin-specific protease
7 A2-B1	antisense	*Zim3*	M	antisense of Usp29
7 B5	coding-gene	*Ube3a*	M	Ubiquitin protein ligase
7 B5	antisense	*UBE3A-AS*	P	UBE3A antisense
7 F5	coding-gene	*Igf2*	P	Insulin-like growth factor 2
7 F5	antisense	*Igf2as*	P	IGF2 antisense
7 F5	coding-gene	*Kcnq1*	M	Voltage-gated potassium channel
7 F5	antisense	*Kcnq1ot1*	P	KCNQ1 antisense
12 F1	coding-gene	*Rtl1(Peg11)*	P	Retrotransposon-like 1, like gag protein
12 F1	antisense	*anti-Rtl1*	M	Rtl1 antisense
17 A1	coding-gene	*Igf2r*	M	Insulin-like growth factor receptor 2
17 A1	antisense	*Air*	P	Igf2r antisense
X D	Non-coding RNA	*Xist*	P	XIST
X D	antisense	*Tsix*	M	XIST antisense

Table 18. Closely linked and reciprocal imprinted genes covered in Illumina data.

Gene_name	Chr	PWD x AKR		AKR x PWD		q-value	AKR percentage		IP status
		AKR	PWD	AKR	PWD		p1	p2	
Peg13	chr15	168	0	6	74	0	1.000	0.075	100%
1810044A24Rik	chr15	7	20	25	5	1.08E-03	0.259	0.833	Partially
Nnat	chr2	1182	1	21	1853	0	0.999	0.011	100%
Blcap	chr2	6	13	15	2	0.0252	0.316	0.882	partially
Zrsr1	chr11	11	0	1	14	6.68E-05	1.000	0.067	100%
Commd1	chr11	12	33	22	7	2.55E-03	0.267	0.759	partially

Figure 33. Sense-antisense gene pairs covered by the Illumina sequence data. Gene structures of the three gene pairs showing nested structures. The blue shading represents the paternal allele and the pink shading indicates for the maternal allele. Boxes with dashed lines indicate no expression. The arrows represent the direction of transcription. The sum of the heights of the two parental exons for each gene is in proportion to the expression level, which is quantified by the total counts of the perfect-match Illumina reads. The relative heights of the exons for the paternal and maternal allele within the same gene represent the relative expression level of the two parental alleles. The short vertical lines under the exons indicate the SNP positions, and the total counts of the two reciprocal crosses for the maternal and paternal allele are labeled.

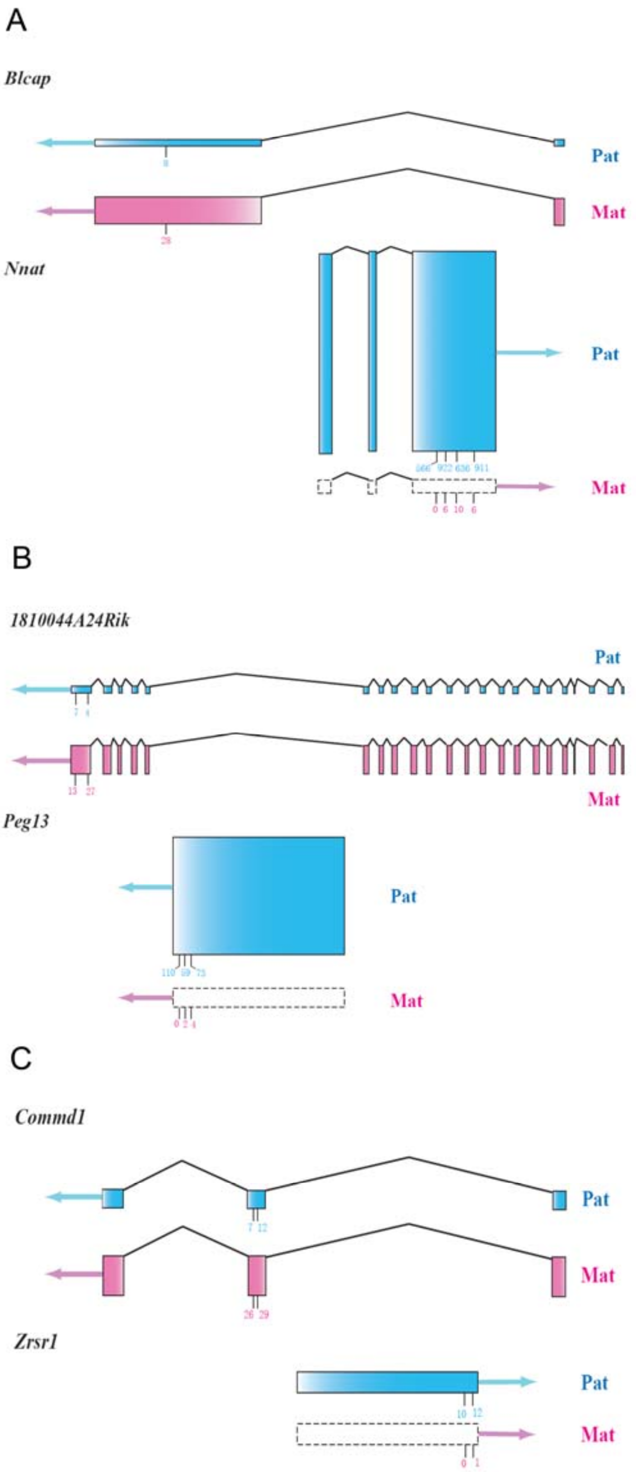
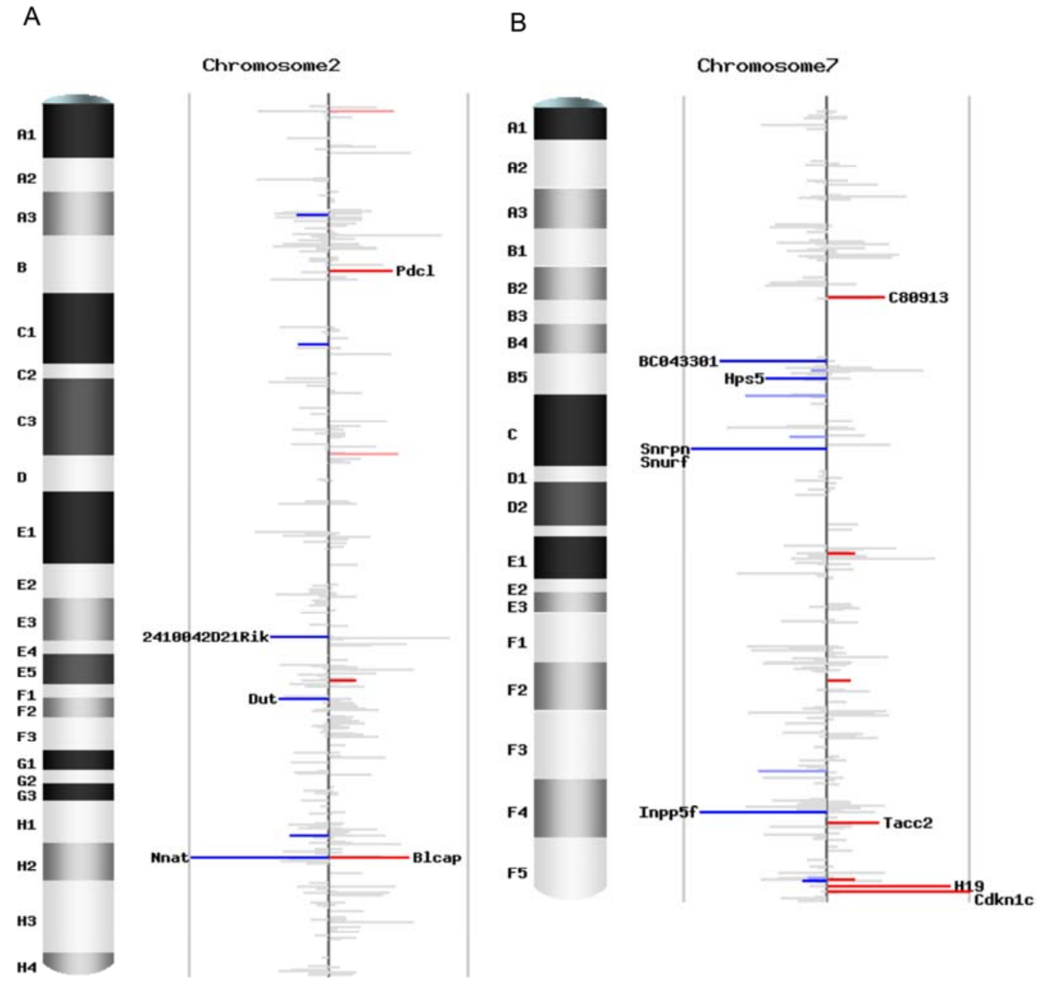

A
Blcap
Pat
8
Mat
28
Nnat
Pat
566
922
636
911
Mat
0
6
10
6
B
1810044A24Rik
Pat
7
4
Mat
13
27
Peg13
Pat
110
59
73
Mat
0
2
4
C
Commd1
Pat
7
12
Mat
26
29
Zrsr1
Pat
10
12
Mat
0
1

Figure 34. Chromosomal scans of imprinting status.

(A). Imprinting status for chromosome 2. **(B).** Imprinting status for chromosome 7. Each plot contains unique Entrez genes covered by SNP-containing Illumina reads with counts no less than 4 in both reciprocal crosses. The height of each bar is the difference of the AKR percentage in the two reciprocal crosses (p_1-p_2), representing the intensity of imprinting. The color stands for the direction of imprinting, blue for paternal expression and red for maternal expression. The intensity of the color represents the significance, grey for not significant (*q*-value ≥ 0.10), lighter blue and pink for marginally significant (0.05 ≤ *q*-value < 0.10), darker blue and red for significant (*q*-value < 0.05). The gene name is indicated if | p_1-p_2| ≥ 0.3.



Discussion

Quantifying allele-specific expression with accurate ratios by directly counting the SNPs

Genomic imprinting is not always an "all-or-none" effect with 100% from the paternal or maternal allele. Instead, the degree of imprinting falls on a continuum from complete uniparental expression to equal expression of the two parental alleles. Microarray hybridization can identify uniparental expression, but it cannot give reliable ratios of the two parental alleles, since there is no good means to quantify the affinity difference between perfect and mismatch probes. The method of direct Sanger sequencing of the cDNA is not quantitative and will miss those cases with quantitative differences between maternal vs. paternal expression. To solve these problems, we sequenced the entire transcriptomes of mouse reciprocal F1 neonatal brains by the Illumina/Illumina sequencing method, and obtained relative expression ratios of the two parental alleles by counting the allele-specific sequence reads at the SNP positions within the transcripts. The method is well validated by independent methods (Pyrosequencing and Sanger sequencing). We present discoveries of the imprinting status of many genes for the neonatal brain, including genes not known to be imprinted in any tissue. Scoring allele-specific expression by short read transcriptome sequencing will be widely used whenever allele-specific differential expression is of interest, including quantification of *cis*-acting regulatory SNP effects (Nagalakshmi et al. 2008).

The path to exhaustive profiling of tissue- and developmental stage-specific genomic imprinting

The discovery of imprinted genes in humans and mice remains sporadic due to the hit-or-miss way that these genes have been discovered. Different studies used different mouse strains, testing imprinting status in different tissues and developmental time points, and none of the studies published to date has employed a truly transcriptome-wide screen for imprinting. Our study shows a way to quantitatively assess in a highly uniform manner the imprinting status of the entire transcriptome for each tissue. The uniformity of the short-read sequencing approach has clear advantages, and paves the way toward a catalog of imprinting status of all transcribed genes in the mouse and human.

Imprinting of nested and closely-linked genes

Our short-read transcriptome sequencing approach identified three pairs of closely linked and reciprocally imprinted genes in which the paternally expressed genes showed 100% monoallelic expression whereas the maternally expressed genes are only partially imprinted in neonatal brains. These data are consistent with the scenario in which the paternally expressed gene is strongly imprinted, and by virtue of its imprinting, there is transcriptional interference, driving weaker expression of genes that are transcribed from the opposite strand (or are nested within the first transcript). This would impose an appearance of weak maternally expressed imprinting. The implications of the bias toward maternal expression in partially imprinted genes, paternal expression of strongly imprinted genes, and the apparent transcriptional interference of opposing strand transcripts all await further analysis to understand the mechanism regulating their imprinting as well as their functional and evolutionary implications.

How many imprinted genes are there in the genome?

It has been estimated that about 1% of the genes in the mammalian genome are imprinted. However, this estimate has a wide range, from around 100 genes (Luedi et al. 2007) to 600 genes (Luedi et al. 2005), to more than 2,000 genes (Nikaido et al. 2003). The variation is due to the ignorance of tissue-specificity of imprinting status and the inability to make inference about non-imprinted genes. Using our method, by counting the reads that correspond to the two parental alleles, we can specify the statistical confidence that a gene is not imprinted, as well as identifying those that are only partially imprinted. This enables determination of the statistical confidence that this list of imprinted genes is close to exhaustive in neonatal brains. In addition to the three novel imprinted genes we found in neonatal brain, we confirmed the imprinting status of 7 known imprinted genes and we also discovered the novel imprinting status in neonatal brain of 7 additional genes known to be imprinted in other tissues. With our coverage of more than 5,000 transcripts, we did not discover novel imprinting clusters, and only a small number of novel imprinted genes were found. Taken altogether, the data suggest that the list of genes that are imprinted in the neonatal brain is nearly complete, and the only remaining ones to be discovered are either expressed at very low levels, show a small parent-of-origin bias, or are imprinted in only a small portion of the brain.

CHAPTER 3

Identification of novel imprinted genes in mouse placenta from RNA-seq data[2]

Abstract

Many questions about the regulation, functional specialization, computational prediction, and evolution of genomic imprinting would be better addressed by having an exhaustive genome-wide catalog of genes that display parent-of-origin differential expression. As a first-pass scan for novel imprinted genes, we performed mRNA-seq experiments on E17.5 mouse placenta cDNA samples from reciprocal cross F1 progeny of AKR and PWD mouse strains, and quantified the allele-specific expression and the degree of parent-of-origin allelic imbalance. We confirmed the imprinting status of 23 known imprinted genes in the placenta, and found that 12 genes reported previously to be imprinted in other tissues are also imprinted in mouse placenta. Through a well-replicated design using an orthogonal allelic-expression technology, we verified five novel imprinted genes that were not previously known to be imprinted in mouse. Our data suggest that most of the strongly imprinted genes have already been identified, at least in the placenta, and that evidence supports perhaps 100 additional weakly imprinted genes. Despite previous appearance that the placenta tends to display an excess of maternally-expressed imprinted genes, with the addition of our validated set of placenta-imprinted genes, this maternal bias has disappeared.

[2] This work was published in Wang X, Soloway PD, Clark AG., 2011. A Survey for Novel Imprinted Genes in the Mouse Placenta by mRNA-seq. *Genetics*, [Epub ahead of print].

Introduction

Genomic imprinting occurs when the expression of the maternal and paternal copies of a gene differ in a parent-of-origin dependent manner (Reik and Walter 2001). Several of the mechanisms of genomic imprinting are shared by higher plants and therian mammals, involving differential DNA methylation, non-coding RNA and/or histone modifications (Delaval and Feil 2004; Pauler and Barlow 2006; Hudson et al. 2010), although imprinting almost certainly arose independently in these lineages. Imprinted genes are often expressed and imprinted in a tissue- and developmental stage-specific manner. Although known imprinted genes tend to be clustered in the genome, there has been an ascertainment bias in concentrating the search among nearby genes for new imprinted candidates, motivating a need for a more balanced genome-wide scan. The occurrence of medical disorders associated with defects in imprinting provides further motivation to produce an exhaustive identification of imprinted genes. Because one allele is virtually silenced, mutations transmitted from the expressing parent behave in a dominant fashion, as is seen in many human disorders associated with defects in imprinted genes (Jiang et al. 2004; Butler 2009).

To date, more than 100 imprinted genes have been discovered in the mouse (Morison et al. 2005), but the list is not exhaustive. Transcriptome-wide and genome-wide attempts to search for novel imprinted genes have exploited different approaches (Maeda and Hayashizaki 2006; Henckel and Arnaud 2010). Genome-wide bioinformatic predictions have successfully identified novel imprinted genes in human and mouse (Yang et al. 2003; Luedi et al. 2005; Luedi et al. 2007; Brideau et al.

2010), but the prediction power is low because the training set of known imprinted genes is small, and the genomic clustering of imprinted genes violates independence of the imprinting signals (Daelemans et al. 2010). Earlier experimental approaches such as expression microarrays on parthenogenetic and androgenetic embryos (Mizuno et al. 2002; Kuzmin et al. 2008; Sritanaudomchai et al. 2010), expression arrays on uniparental disomic (UPD) mice (Choi et al. 2001; Choi et al. 2005; Schulz et al. 2006), and allele-specific expression arrays on individuals with informative SNPs (Pollard et al. 2008; Brideau et al. 2010) have identified many novel imprinted genes on a larger scale than the single-gene approach. However, these methods require an abnormal configuration of the genome and can only cover a subset of genes included in the array design or the UPD region. DNA methylation-based methods have successfully identified a number of novel imprinted genes (Peters et al. 1999; Smith et al. 2003). This method first searches for differentially methylated regions (DMR), then examines the genes in close proximity to each novel DMR. Since not all imprinted genes have an associated DMR, even this approach will likely miss some novel imprinted genes. To overcome these problems and begin to identify imprinted genes transcriptome-wide in a variety of tissues, we (Wang et al. 2008) and other investigators (Babak et al. 2008; Gregg et al. 2010a; Gregg et al. 2010b) have carried out mRNA-seq studies to identify novel imprinted genes through differential allele-specific expression in reciprocal F1 plants and animals. Wang *et al*. and Babak *et al*. are the first studies using RNA-seq of mouse reciprocal crosses to search for novel imprinted genes. Wang *et al*. performed RNA-seq of mouse neonatal day 2 (P2) brains from reciprocal crosses of AKR and PWD strains. We discovered and confirmed 14 known and 3 novel imprinted genes in P2 brains. Babak *et al*. did transcriptome sequencing on embryonic day 9.5 (E9.5) embryos in CAST/EiJ and C57BL/6J reciprocal crosses and they found 14 imprinted genes which are all known in mouse. No novel imprinted genes emerged from this study. Recently, Gregg *et al*. published an RNA-seq study on embryonic and adult brains of CAST/EiJ and C57BL/6J reciprocal crosses. Whole E15 brain, adult cortex, and adult hypothalamus samples were sequenced and analyzed, and they claimed more than 1,300 genes showed differential parental allelic expression in the brain. It is clear that RNA-seq provides a powerful tool for scoring parent-of-origin differential expression, and that differences in targeted tissue, developmental stage, sequence quantity, methods of validation may contribute to differences across studies.

In the mouse, most of the known imprinted genes are expressed and imprinted in the brain and/or placenta (Morison et al. 2005). The placenta is a mammalian-specific organ, which has important nutritional transport and immune functions for fetal growth. The placenta has been a primary target organ in studies of genomic imprinting in terms of the number and importance of known imprinted genes (Wagschal and Feil 2006; Frost and Moore 2010), motivating this RNA-seq analysis of reciprocal F1 mice to discover novel imprinted genes. Since all three previous transcriptome-wide RNA-seq studies were focused on brain or embryonic tissue, our first-pass survey in mouse placenta will complement previous studies and provide information on a tissue of particularly focused interest to the imprinting community.

Materials and Methods

Mouse Strains and crosses

The four mouse strains (AKR/J, PWD/PhJ, C57BL/6 and CAST/EiJ) and the two reciprocal crosses with two strain combinations (PWD/PhJ x AKR/J, AKR/J x PWD/PhJ and C57BL/6 x CAST/EiJ, CAST/EiJ x C57BL/6) were described in (Wang et al. 2010). We dissected E17.5 placenta tissues from the F1 mice from the following crosses: AKR/J female x PWD/PhJ male cross (AKR x PWD for short), PWD/PhJ female x AKR/J male (PWD x AKR for short), C57BL/6 female x CAST/EiJ male (B6 x CAST for short), and CAST/EiJ female x C57BL/6 male (CAST x B6 for short). To minimize maternal contamination, we cut only the fetal side of the placenta tissue during dissection. E17.5 was chosen because it is late enough to be able to get enough tissue in these dissections, but early enough to make it fairly easy to avoid maternal contamination.

We extracted the total RNA samples from the placentas harboring the F1 progeny using Qiagen RNeasy Plus Mini Kit (Qiagen, CA). The RNA concentrations and A260 nm/A280 nm ratios were quantified with a NanoDrop ND-1000 Spectrophotometer (Thermo Scientific, DE). RNA integrity was tested using the Agilent 2100 Bioanalyzer (Agilent Technologies, Inc, CA). All of the samples had a RIN (RNA integrity number) in the range 9.6-10.0 (RIN_{max} = 10.0). AKR/J and PWD/PhJ genomic DNA were purchased from the Jackson Laboratory (www.jax.org).

Illumina mRNA sequencing of the F1 placenta transcriptome

Our initial mRNA-seq was performed on total RNA samples from one AKR female x PWD male and one PWD female x AKR male placenta using an Illumina Genome Analyzer (Illumina Inc., CA). The mRNA-seq libraries were made with 5 μg of starting total RNA samples using the mRNA-Seq 8-Sample Prep Kit (Illumina Inc., CA), following the Illumina protocol for mRNA sequencing sample preparation. Eight Illumina GA lanes were sequenced for the AKR x PWD library and seven lanes for the PWD x AKR library. Image analysis and base calling were performed by the Illumina instrument software. In total, our initial screen consisted of 66.0 million short reads (read length 44 bp) for the AKR x PWD cross, and 63.3 million reads for the PWD x AKR cross.

mRNA-seq Alignment and quantification of total and allele-specific expression

The reads were truncated to 40 bp and aligned to the mouse reference genome (NCBI B37) using BWA with a maximum of four mismatches (Li and Durbin 2009). On average, 55.2% of the reads were mapped uniquely to the reference genome. Alignment counts in the exon regions were summarized by custom scripts. To identify reads that mapped to the exon-intron junctions, we built a junction database by extracting all possible junction sequences, based on the gene and exon models from the Ensembl database (www.ensembl.org). 4.8% of the total reads were mapped to the exon-intron junctions. The exon and junction counts were normalized by the transcript length and the total number of mapped reads to compute RPKM (Mortazavi et al. 2008). We covered 12,532 unique Ensembl genes (41,110 Ensembl transcripts) with

RPKM ≥ 1. If we use a more stringent cut-off of RPKM ≥ 5, then 6,794 unique genes (20,026 transcripts) are retained.

To quantify the allele-specific expression in the two reciprocal crosses, at each identified SNP position we counted the reads with the reference allele as well as reads with the alternative allele (Wang et al. 2008). In addition to using the known sequence differences between the mouse strains used, we also performed *de novo* SNP calling from the uniquely aligned reads using SAMtools software (Li et al. 2009), followed by our own post-filtering scripts. To determine the transmission direction, we used the AKR/J allele information from the Sanger mouse genome project 2010-03 SNP release (http://www.sanger.ac.uk/resources/mouse/genomes/). Since there is imprinted X inactivation in the mouse placenta (Huynh and Lee 2005; Sado and Ferguson-Smith 2005), SNPs with an X-chromosome homolog will mistakenly suggest a maternally-expressed imprinted gene. To eliminate this bias, we BLATed all SNPs with sequence from 50 bp upstream and 50 bp downstream to the genome (http://genome.ucsc.edu/cgi-bin/hgBlat?command=start). We removed all SNPs with an X chromosome BLAT hit with matched with length 40 or more (equal to the read length).

In total, 43,510 high quality autosomal SNPs with 4 or more counts in each of the two reciprocal crosses were identified. 41,953 SNPs (96.4%) are within 1 kb upstream or downstream of Ensembl gene models. Manual annotation was performed for three known imprinted genes that are not in the Ensembl database (*Peg10*, *Rian*, *Mirg*). Because both the AKR and PWD alleles in the F1 transcriptome are mapped to the reference genome, which was assembled from the B6 strain, there will be genome mapping bias toward the AKR allele if we use the same cut-off for both alleles (AKR is closer to B6 than it is to the PWD strain in terms of genetic distance). To remove this mapping bias, we generated a pseudo-genome, by replacing the reference allele in the genome with the alternative allele. Then we redid the alignment with the same cut-off to the pseudo-genome. We used the averaged counts from the reference genes and pseudo-genome as the final SNP count summary.

Detection of significant parent-of-origin effects and identification of candidate imprinted genes

To select candidate imprinted genes for verification, we applied a formal statistical test to the 2 x 2 contingency table formed by the tally of reads of the two alleles in the two reciprocal crosses (Wang et al. 2008). In addition to using this statistical significance, we also filtered the results based on the magnitude of the allelic expression bias. We define p_1 as the AKR allele percentage in the AKR x PWD cross and p_2 as the AKR allele percentage in the PWD x AKR cross. For a 100% maternally expressed candidate imprinted gene, we expect $p_1 = 1$ and $p_2 = 0$. For partially imprinted genes with preferential maternal expression, we used a cut-off of $p_1 > 0.65$ and $p_2 < 0.35$ (and similarly, we selected the paternally expressed imprinted candidates with a cut-off of $p_1 < 0.35$ and $p_2 > 0.65$). This cut-off is somewhat arbitrary, but it was meant to avoid the inclusion of genes with a weak allelic imbalance that would otherwise be included in our imprinted candidate set only because they are so highly expressed that they become statistically significant. For graphical presentation and discussion, the metric $p_2 - p_1$ quantifies the parent-or-origin effect in the range from -1 to +1. If there is no parent-of-origin effect, then $p_2 - p_1 = 0$. If there is preferential expression of the paternal allele, then $p_2 - p_1 > 0$. If there is preferential expression of the maternal allele, then $p_2 - p_1 < 0$. In order to keep the false positive rate low, we only include candidate genes with two or more informative SNPs.

Verification of the candidate imprinted genes with allele-specific pyrosequencing

We selected three known imprinted genes (*Igf2*, *Peg10* and *Klf14*) and seven candidate imprinted genes (*Pde10a*, *Phf17*, *Gpsm2*, *Zfp64*, *Htra3*, *Phactr2* and *Trim23*) for allele-specific expression quantification using Pyrosequencing. To exclude the possibility of stochastic expression effects and sex-specific genomic imprinting, we verified these genes in placentas harboring 3 female and 1 male F1 progeny from each of AKR-PWD reciprocal crosses. To exclude strain-specific effects, 3 novel imprinted genes (*Pde10a*, *Phf17* and *Phactr2*) were verified in an additional 3 female and 1 male progeny bearing placentas from each of B6-CAST reciprocal crosses. The pyrosequencing assay design and sequencing protocol can be found in (Wang et al. 2010).

Results

mRNA-seq alignments, transcriptome coverage and SNP calling

This mRNA-seq study was performed on E17.5 placental tissues from reciprocal crosses of AKR and PWD mouse strains. We obtained 66 million 44-bp reads from placenta cDNA of a single AKR x PWD F1 individual and 63 million reads from the reciprocal PWD x AKR placental transcriptome. 60% of the reads could be uniquely mapped to the NCBI B37 mouse reference genome, with 55.2% of reads mapping to the exons and 4.8% mapping to the exon-intron junctions. The total expression levels were quantified by RPKM, which is a normalized per gene read counts (Mortazavi et al. 2008). In the RNA-seq data, there was coverage of 12,532 Ensembl unique genes (41,110 transcripts) with RPKM greater than 1, and 6,794 unique genes had an RPKM value greater than 5.

Informative SNP positions are needed to quantify the allele-specific expression. From *de novo* SNP calling based on the RNA-seq data, after quality filtering, we found 43,510 high-quality autosomal SNPs, 96.4% of which reside in known Ensembl gene models. To remove the genome mapping bias, we summarized the SNP counts by the average count when mapped to the reference genome and to a pseudo-genome of the alternate strain (See Materials and Methods).

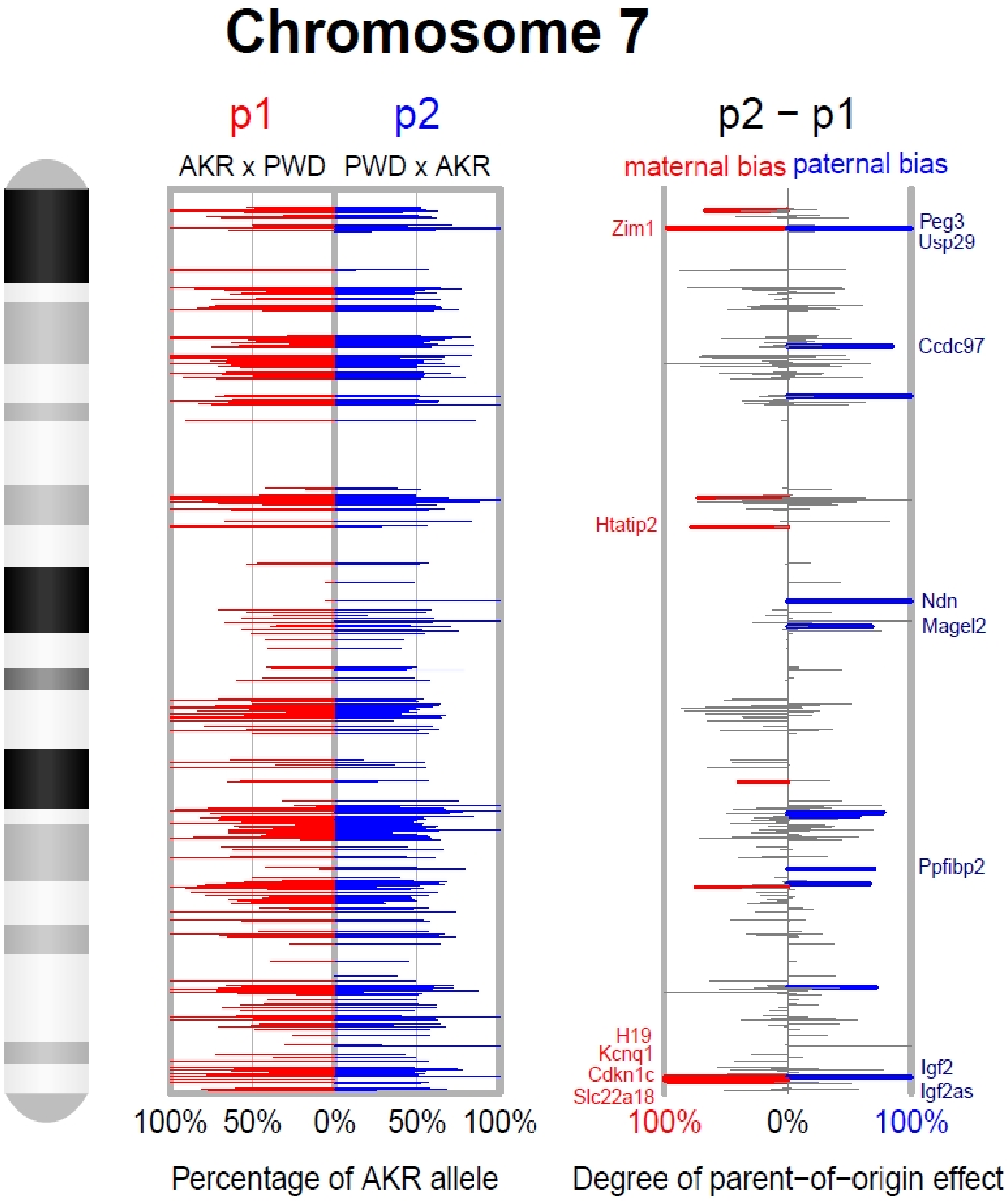
Detection of significant parent-of-origin effects

With the read counts at the informative SNP positions, we were able to determine the allele-specific expression ratio from the relative counts of the reference and alternative alleles (Wang et al. 2008). We define p_1 as the expression percentage from the AKR allele in placentas from the AKR female x PWD male cross and p_2 as the AKR allele percentage in the reciprocal cross. In regard to the direction of transmission, p_1 is the maternal allele percentage in AKR x PWD, and p_2 is the paternal percentage for PWD x AKR. The Storer-Kim test was used as a formal statistical test of the null hypothesis that $(p_2 - p_1) = 0$. Rejections of this null hypothesis identify novel imprinted candidate genes (See Materials and Methods). To further filter the data and reduce false positives, rather than relying only on the P-value of the Storer-Kim test, we also used an arbitrary cut-off of $p_1 > 0.65$ and $p_2 < 0.35$ for maternally expressed candidates, and $p_1 < 0.35$ and $p_2 > 0.65$ for paternally expressed ones. This allows for identification of partially imprinted genes when there are sufficiently many reads spanning the SNPs to make a confident call.

Out of the 5,557 unique genes covered with two or more informative SNPs, with the above criteria for significant parent-of-origin effect identification, we found 251 significant candidates with q-value < 0.01 and SNP coverage 4 or more in each of the two reciprocal crosses (criterion 1). Of these candidate genes, 120 have preferential maternal expression, and 131 have a paternally biased expression. If we use RPKM>1 and SNP coverage>10 in both reciprocal crosses as the criteria for inclusion, 216 significant candidates are left, with 115 paternal and 101 maternal candidates (criterion 2). If we use a more stringent cut-off of PRKM>3 and SNP coverage>20, only 113 candidates are retained, with 60 paternal and 53 maternal candidates

(criterion 3). To visualize the allelic expression ratio and the degree of parent-or-origin effect genome-wide, we made a plot for each autosome, and chromosome 7 is shown in (Figure 35) as an example. From these figures, we observed that most of the genes show nearly 50:50 allelic expression ratios, when we scan along the chromosomes.
A number of significant candidate imprinted genes emerged from the parent-of-origin effect plot.

Figure 35. Allele-specific expression ratio and the distribution along mouse chromosome 7 of parent-of-origin biased expression. **Left panel:** Allele-specific expression levels for genes on chromosome 7 for both AKR female x PWD male and PWD female x AKR male F1 fetal tissue from the placentas. The x-axis is the allelic expression ratio from the AKR allele (0% to 100%) in the two reciprocal crosses. The y-axis is the physical location along the chromosome. The red bar is drawn according to the AKR allelic expression ratio (p_1) in the AKR x PWD cross, and the blue bar is the AKR allelic expression ratio (p_2) in the PWD x AKR cross (throughout we adhere to the convention of listing crosses as female x male). **Right panel:** Degree of parent-of-origin effect on chromosome 7. Unique Ensembl genes covered with a high-quality SNP count of at least 4 in each of the two reciprocal crosses are included in the plot. The height of each bar is the degree of the parent-of-origin effect, which is computed as (p_2-p_1). The blue and red colors represent significant candidate imprinted genes with RPKM > 3 and *q*-value < 0.01. The gene name is displayed for the significant, known imprinted genes and candidate genes with $| p_2-p_1| > 0.4$.


Chromosome 7
p1
p2
AKR x PWD
PWD x AKR
p2 − p1
maternal bias
paternal bias
Zim1
Peg3
Usp29
Ccdc97
Htatip2
Ndn
Magel2
Ppfibp2
H19
Kcnq1
Cdkn1c
Slc22a18
Igf2
Igf2as
100%
50%
0%
50%
100%
Percentage of AKR allele
100%
0%
100%
Degree of parent−of−origin effect

Significant candidate imprinted genes that are previously known to be imprinted in mouse

Among the 251 candidate imprinted genes that we identified from criterion 1, 35 have been previously reported in the mouse literature to be imprinted. For each gene, the number of SNPs, total SNP counts, allelic expression ratios and the *q*-value are summarized in Table 1. We compared the expression direction of these genes in our RNA-seq data and the previously reported imprinting direction, and 35 out of 35 matched. 23 of the 35 genes were known to be imprinted in mouse placenta in various stages and crosses: *Igf2* (DeChiara et al. 1991; Hu et al. 1995), *Peg10* (Ono et al. 2003), *Sfmbt2* (Kuzmin et al. 2008), *Sgce* (Piras et al. 2000), *Plagl1* (Piras et al. 2000; Smith et al. 2002), *Slc38a4* (Mizuno et al. 2002; Smith et al. 2003), *Airn* (also known as *Igf2rAS*) (Wutz et al. 1997) , *Rtl1* (Seitz et al. 2003), *Mest* (Kaneko-Ishino et al. 1995), *Igf2as* (Moore et al. 1997) and *Dlk1*(Schmidt et al. 2000) are known to be paternally expressed in the mouse placenta; *H19* (Bartolomei et al. 1991), *Igf2r* (Barlow et al. 1991), *Cdkn1c* (Hatada and Mukai 1995), *Grb10* (Miyoshi et al. 1998), *Ppp1r9a* (Ono et al. 2003), *Klf14* (Parker-Katiraee et al. 2007), *Nesp* (Peters et al. 1999), *H13* (Wood et al. 2007), *Slc22a2* (Zwart et al. 2001), *Asb4* (Mizuno et al. 2002), *Slc22a18* (Dao et al. 1998) and *Kcnq1*(Gould and Pfeifer 1998) are preferentially expressed from the maternal allele. The remaining 12 genes are either known to be not imprinted in the placenta or the imprinting status in the placenta is not clear. In this study, we found that they are actually imprinted in E17.5 mouse placenta in AKR-PWD reciprocal crosses. *Peg3* (Kaneko-Ishino et al. 1995) is known to be imprinted in the human placenta (Hiby et al. 2001), however, the imprinting status in the mouse placenta had not been reported. *Ndn* (MacDonald and Wevrick 1997) and *Magel2* (Boccaccio et al. 1999) are both expressed in the mouse placenta, whereas the imprinting status was not clear. *Rian* (Hatada et al. 2001), *Zim1* (Kim et al. 1999), *Meg3* (Miyoshi et al. 2000), *Mirg* (Seitz et al. 2004), *Usp29* (Kim et al. 2000), *Impact* (Hagiwara et al. 1997), *Nnat* (Kagitani et al. 1997), *Zdbf2* (Kobayashi et al. 2009) and *Zrsr1* (Hatada et al. 1993) were not previously reported to be imprinted in the mouse placenta either. Therefore, we identified 12 candidate genes with novel mouse placenta imprinting status.

The *q*-value rank order is presented in (Table 19). We noticed that most of the known imprinted genes identified in our study have higher *q*-value rank relative to other genes, most of them are highly expressed in the placenta, and the imprinting status of most previously known imprinted genes is close to 100%. We conclude that most of the significant imprinted genes with highest degree of parent-of-origin bias have already been identified by the genomic imprinting community. The high concordance (35 out of 35) of known imprinted genes with the significance of our test of parent-of-origin effects on allelic expression ratios provides one measure of the confidence in the results despite the lack of replication at the RNA-seq stage.

Table 19. Imprinted genes identified in mouse placenta RNA-seq data that have been reported previously in the literature.

q-value ranking	Gene name	# of SNPs	AKRxPWD allele counts		PWDxAKR allele counts		p1	p2	qvalue	Expr. allele	Known expr. allele	Known in mouse placenta?
			AKR	PWD	AKR	PWD						
1	*Igf2*	3	2	2235	4418	0	0.09%	100.00%	0	P	P	YES
2	*Peg10*	13	9.5	3503	17293	3	0.27%	99.98%	0	P	P	YES
3	*Sfmbt2*	13	5	555	761.5	15	0.89%	98.07%	0	P	P	YES
4	*Peg3*	11	202	4944	6398	1	3.93%	99.98%	0	P	P	NO
5	*Rian*	31	1180	38	104.5	2080	96.88%	4.78%	0	M	M	NO
6	*H19*	2	2067	0	59.5	9319	100.00%	0.63%	0	M	M	YES
7	*Zim1*	10	1600	0	52.5	2365	100.00%	2.17%	0	M	M	NO
8	*Meg3*	26	866.5	0	52	1610	100.00%	3.13%	0	M	M	NO
9	*Igf2r*	11	482	0	57	1785	100.00%	3.10%	0	M	M	YES
10	*Cdkn1c*	1	468.5	0	26	671	100.00%	3.73%	5.3E-294	M	M	YES
11	*Sgce*	4	51	523	499	22.5	8.89%	95.69%	1.1E-213	P	P	YES
12	*Plagl1*	4	0	210	616.5	0	0.00%	100.00%	6.9E-200	P	P	YES
13	*Grb10*	10	846	148	972	1999	85.15%	32.72%	1.3E-189	M	M	YES
14	*Slc38a4*	2	18	250	439	0	6.73%	100.00%	2.3E-168	P	P	YES
16	*Mirg*	9	160	0	12	313.5	100.00%	3.69%	3.1E-111	M	M	NO
17	*Usp29*	1	0	94	128.5	0	0.00%	100.00%	8.08E-63	P	P	NO
18	*Impact*	2	6	89	293	16	6.32%	94.82%	2.71E-62	P	P	NO
20	*Airn*	6	0	75.5	158.5	0	0.00%	100.00%	3.22E-61	P	P	YES
21	*Rtl1*	5	12	145	154	13	7.64%	92.22%	4.37E-58	P	P	YES
22	*Ppp1r9a*	6	76	0	23.5	162.5	100.00%	12.63%	7.62E-43	M	M	YES
23	*Mest*	1	0	38.5	108.5	0	0.00%	100.00%	9.54E-34	P	P	YES
24	*Nnat*	2	0	37	147	5	0.00%	96.71%	4.63E-32	P	P	NO
28	*Zdbf2*	5	0	86	85.5	46	0.00%	65.02%	9.08E-25	P	P	NO
33	*Klf14*	3	54.5	6	6	55	90.08%	9.84%	1.43E-18	M	M	YES
34	*Nesp*	1	34.5	0	3	46	100.00%	6.12%	3.03E-18	M	M	YES
35	*H13*	2	152.5	74.5	118	282	67.18%	29.50%	5.36E-18	M	M	YES
38	*Zrsr1*	1	0	19	67.5	0	0.00%	100.00%	1.16E-17	P	P	NO
56	*Slc22a2*	2	22	0	2	40	100.00%	4.76%	1.51E-13	M	M	YES
80	*Asb4*	4	65	27	5	47	70.65%	9.62%	1.23E-11	M	M	YES
98	*Ndn*	1	0	28.5	13	0	0.00%	100.00%	1.22E-09	P	P	NO
121	*Igf2as*	1	0	41.5	7.5	0.5	0.00%	93.75%	2.99E-08	P	P	YES
167	*Slc22a18*	1	12	0	20	59.5	100.00%	25.16%	5.3E-06	M	M	YES
176	*Magel2*	1	1	17.5	16.5	3	5.41%	84.62%	1.01E-05	P	P	NO
180	*Kcnq1*	1	6	0	0	23	100.00%	0.00%	1.53E-05	M	M	YES
189	*Dlk1*	1	0	7	14	0	0.00%	100.00%	5.08E-05	P	P	YES

Identification and verification of novel imprinted genes in the mouse placenta

To confirm the novel imprinted candidates identified above, we need to quantify their allele-specific expression using an independent method. We performed pyrosequencing to quantify allele-specific expression in two reciprocal F1 placenta samples. Pyrosequencing is a highly quantitative method to profile the allelic expression ratio, with a measurement coefficient of variation of 2-5% (Marsh 2007). To exclude the possibility of random monoallelic expression for specific genes (Lomvardas et al. 2006; Gimelbrant et al. 2007), and potential sex-specific imprinting status (Gregg et al. 2010a), we verified the candidates in 4 AKR x PWD F1 individuals (3 females and 1 male) and 4 PWD x AKR F1 individuals (3 females and 1 male). The average allelic percentage is reported in (Table 20).

We selected a total of 10 candidate genes for verification, including three known imprinted genes as positive controls (*Igf2*, *Peg10* and *Klf14*). Among the top 20 candidates, only 2 are novel (*Pde10a* and *Phf17*), and we included both. Then we selected 5 additional novel candidates (*Gpsm2*, *Zfp64*, *Htra3*, *Trim23* and *Phactr2*) for verification (Table 21).

From the pyrosequencing results in Table 21, 8 of the 10 known and novel candidate genes we tested are verified to be imprinted; one candidate gene (*Trim23*) did not show good pyrosequencing signal due to low expression level; we observed biallelic expression for one candidate gene (*Gspm2*). Further examination of the *Gspm2* gene region reveals that the different SNPs are not consistent in RNA-seq data. Careful inspection of the RNA-seq read alignments suggests that the false-positive call may have been made because of poor read mapping, as the read depth is unusually variable around this gene. Therefore, we have an empirical false discovery rate of 1 out of 9 or 11% confirmed by our pyrosequencing verification results.

Igf2 and *Peg10* were correctly verified as paternally expressed imprinted genes, and *Klf14* as maternally expressed imprinting gene, which is consistent with the results in our RNA-seq data. Among the 7 novel candidates, 5 (*Pde10a*, *Phf17*, *Zfp64*, *Htra3* and *Phactr2*) were verified to be novel imprinted genes in the mouse placenta (Table 20), one test failed due to low expression, and one failed to validate.

Pde10a is the most significant novel candidate gene (*q*-value rank 15). It is located on chromosome 17, 3.6 Mbp away from the known imprinted gene, *Slc22a3*. It is a member of the phosphohydrolyase gene family, catalyzing the hydrolysis of the cAMP and cGMP to the respective nucleoside 5' monophosphate (Loughney et al. 1999; Soderling et al. 1999). Pyrosequencing primers were designed to target one of the 12 significant SNPs in this gene (Figure 36A). In the RNA-seq data, we observed expression primarily from the maternal allele in both AKR-PWD reciprocal crosses (Figure 36B). We verified it in four placentas from each of the two reciprocal crosses, and we found consistent preferential maternal expression (Figure 36C, D). To exclude the possibility of strain-specific imprinting, we also tested placenta tissue from B6-CAST reciprocal crosses, and we obtained the same results (Table 22). Thus, we conclude that *Pde10a* is a novel imprinted gene in the E17.5 placenta.

Table 20. Pyrosequencing verification for known/novel imprinted genes in mouse placenta.

q-value ranking	Gene name	p_1 RNA-seq	p_2 RNA-seq	Expr. allele in RNA-seq	p_1 pyro AKR x PWD	p_2 pyro PWD x AKR	Conclusion	Preferentially expressed allele	Notes
1	*Igf2*	0.09%	100.00%	P	6.10%	100.00%	Confirmed known	P	
2	*Peg10*	0.27%	99.98%	P	0.00%	100.00%	Confirmed known	P	
15	*Pde10a*	93.27%	8.40%	M	77.40%	14.83%	Novel	M	
19	*Phf17*	18.72%	71.39%	P	40.29%	75.26%	Novel	P	
25	*Gpsm2*	0.00%	96.04%	P	64.20%	63.75%	Not imprinted	Biallelic	inconsistent SNPs
26	*Zfp64*	18.44%	75.95%	P	28.41%	100.00%	Novel	P	
33	*Klf14*	90.08%	9.84%	M	91.15%	0.00%	Confirmed known	M	
40	*Htra3*	100.00%	4.00%	M	73.55%	43.69%	Novel	M	low expression level
47	*Trim23*	0.00%	100.00%	P	-	-	No signal	-	low expression level
145	*Phactr2*	66.79%	34.81%	M	62.32%	30.94%	Novel	M	

Table 21. RNA-seq read-counts summary for selected known/novel imprinted genes for verification in mouse placenta.

q-value ranking	Gene name	# of SNPs	AKR x PWD allele counts		PWD x AKR allele counts		p_1	p_2	*q*-value	Expr. allele	Expr. level AKR x PWD	Expr. level PWD x AKR
			AKR	PWD	AKR	PWD						
1	*Igf2*	3	2	2235	4418	0	0.09%	100.00%	0	P	1109.87	974.25
2	*Peg10*	13	9.5	3503	17293	3	0.27%	99.98%	0	P	1500.91	1031.52
15	*Pde10a*	12	347	25	44	480	93.27%	8.40%	1.24E-159	M	14.14	13.22
19	*Phf17*	7	64.5	280	584	234	18.72%	71.39%	1.80E-61	P	25.85	37.61
25	*Gpsm2*	2	0	59	48.5	2	0.00%	96.04%	7.30E-27	P	3.49	2.94
26	*Zfp64*	5	33	146	120	38	18.44%	75.95%	3.04E-25	P	11.72	6.78
33	*Klf14*	3	54.5	6	6	55	90.08%	9.84%	1.43E-18	M	4.37	3.27
40	*Htra3*	3	53	0	1	24	100.00%	4.00%	2.47E-17	M	0.72	0.52
47	*Trim23*	2	0	32	26	0	0.00%	100.00%	2.60E-15	P	1.68	1.96
145	*Phactr2*	6	87.5	43.5	59	111	66.79%	34.81%	5.04E-07	M	8.07	11.26

Table 22. Pyrosequencing verification results table in AKR-PWD and B6-CAST reciprocal crosses (percent of AKR/B6 allele is shown in the table).

SampleID	Sex	Mother	Father	*Igf2*	*Peg10*	*Klf14*_1*	*Klf14*_2	*Htra3*_1	*Htra3*_2	*Gpsm2*
AKR/PWD allele				C/T	G/A	C/T	C/T	A/G	A/G	G/A
AP1	Male	AKR	PWD	7.00%	0.00%	93.50%	97.10%	70.60%	72.70%	62.90%
AP2	Female	AKR	PWD	5.90%	0.00%	78.20%	90.80%	96.00%	95.80%	64.50%
AP3	Female	AKR	PWD	5.80%	0.00%	89.50%	98.90%	61.80%	65.80%	64.80%
AP4	Female	AKR	PWD	5.70%	0.00%	96.60%	84.60%	61.50%	64.20%	64.60%
PA1	Male	PWD	AKR	100.00%	100.00%	0.00%	0.00%	46.20%	47.20%	63.70%
PA2	Female	PWD	AKR	100.00%	100.00%	0.00%	0.00%	37.70%	42.30%	64.40%
PA3	Female	PWD	AKR	100.00%	100.00%	0.00%	0.00%	42.70%	43.30%	64.80%
PA4	Female	PWD	AKR	100.00%	100.00%	0.00%	0.00%	46.10%	44.00%	62.10%

SampleID	Sex	Mother	Father	*Zfp64*_1	*Zfp64*_2	*Phf17*_1	*Phf17*_2	*Pde10a*	*Phactr2*_1	*Phactr2*_2
AKR/PWD allele				G/A	G/A	C/T	C/T	T/C	C/T	C/T
AP1	Male	AKR	PWD	34.70%	35.10%	43.00%	40.30%	78.00%	57.30%	64.60%
AP2	Female	AKR	PWD	0.00%	0.00%	43.40%	42.20%	78.40%	NA	58.70%
AP3	Female	AKR	PWD	39.60%	37.60%	41.90%	40.50%	71.60%	63.20%	66.20%
AP4	Female	AKR	PWD	40.00%	40.30%	35.60%	35.40%	81.60%	NA	63.90%
PA1	Male	PWD	AKR	100.00%	100.00%	85.00%	75.50%	23.70%	29.60%	26.00%
PA2	Female	PWD	AKR	NA	NA	72.30%	71.60%	17.40%	NA	29.40%
PA3	Female	PWD	AKR	100.00%	100.00%	73.90%	71.10%	0.00%	33.60%	33.40%
PA4	Female	PWD	AKR	NA	NA	77.10%	75.60%	18.20%	32.90%	31.70%

SampleID	Sex	Mother	Father	*Phf17*_1	*Phf17*_2	*Pde10a*_1	*Pde10a*_2	*Phactr2*_1	*Phactr2*_2
B6/CAST allele				C/T	C/T	T/C	T/C	C/T	C/T
BC1	Male	C57BL/6	CAST	43.20%	41.50%	NA	23.20%	23.60%	20.30%
BC2	Female	C57BL/6	CAST	38.70%	37.80%	NA	23.40%	26.00%	31.60%
BC3	Female	C57BL/6	CAST	42.30%	40.30%	NA	0.00%	28.00%	27.10%
BC4	Female	C57BL/6	CAST	45.70%	44.20%	NA	22.20%	26.80%	30.40%
CB1	Male	CAST	C57BL/6	77.30%	75.40%	71.80%	65.50%	50.10%	52.00%
CB2	Female	CAST	C57BL/6	74.30%	74.40%	NA	62.60%	57.40%	52.70%
CB3	Female	CAST	C57BL/6	74.80%	74.30%	63.80%	68.30%	55.80%	51.80%
CB4	Female	CAST	C57BL/6	76.10%	76.70%	61.90%	55.50%	57.10%	52.30%

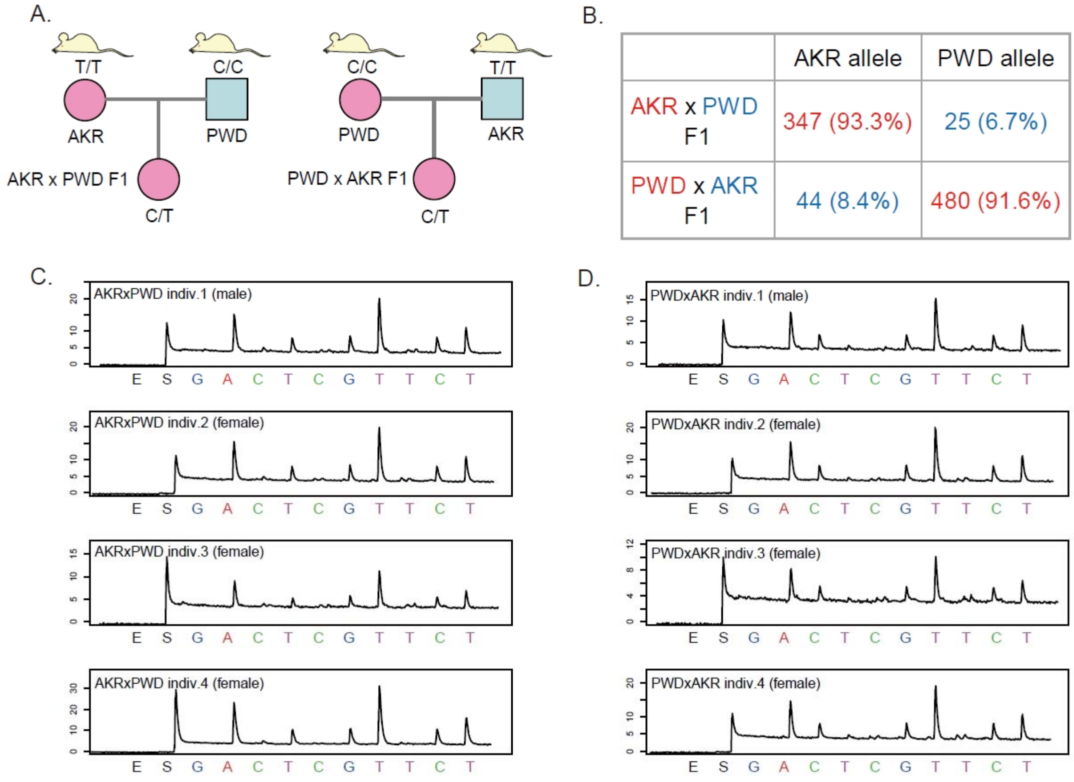
Figure 36. Verification of the novel candidate imprinted gene *Pde10a*, a preferentially maternally-expressed imprinted gene.

(A). The mouse crossing scheme used to generate the AKR-PWD reciprocal F1 placentas. One informative SNP within the gene is shown in the Figure with a T allele in AKR and a C allele in PWD.

(B). SNP allelic counts summary table for the *Pde10a* gene, showing preferential maternal expression.

(C). Pyrograms for verification in 4 individuals of AKR x PWD cross (1 male and 3 females). Target sequence analyzed: AA(C/T)GTTTTCTT.

(D). Pyrograms for verification in 4 individuals of PWD x AKR cross (1 male and 3 females).



	AKR allele	PWD allele
AKR x PWD F1	347 (93.3%)	25 (6.7%)
PWD x AKR F1	44 (8.4%)	480 (91.6%)

Phf17 is the second most significant novel candidate in the list. It is located on mouse chromosome 4 and it is not near any of the known imprinting cluster. *Phf17* (aka *Jade1*) is a component of the HBO1 complex which has a histone H4-specific acetyltransferase activity and performs most of the histone H4 acetylation *in vivo* (Foy et al. 2008). Imprinting of genes involved in histone modifications are particularly interesting, as they may provide a means for amplification of the imprinting signal, and for propagating the effect to other target genes. Pyrosequencing verifications confirmed preferential paternal expression in both AKR-PWD and B6-CAST crosses (Table 20 and Table 22).

Phactr2 is a phosphatase and actin regulator, and it is identified in our RNA-seq study as a maternally expressed imprinted candidate. This gene had not previously been known to be imprinted in mouse. We verified it in multiple individuals of both AKR-PWD and B6-CAST crosses, and it is confirmed to be preferentially expressed from the maternal allele (Table 20 and Table 22). In a recent Illumina ASE BeadArray survey of novel imprinted genes in human term placenta, human *PHACTR2* is found to be partially imprinted, with a maternal allelic bias (Daelemans et al. 2010). Therefore, the imprinting status of *Phactr2* is conserved between mouse and human. *Phactr2* is on mouse chromosome 10, 104 kbp downstream of a paternally expressed known imprinted gene, *Plagl1*. *Phactr2* is transcribed in the opposite direction to *Plagl1*, which could be another reciprocally-imprinted sense-antisense pair (Wang et al. 2008).

Among the 7 novel candidates tested, two other genes, *Zfp64* and *Htra3* have also been verified to be partially imprinted in the mouse placenta. *Zfp64* is on mouse chromosome 2, 6 Mbps from a known imprinted gene *Nesp*. *Zfp64* is a Krüppel family transcription factor that is under the control of *Runx2*, and participates in Notch signalling to regulate differentiation in mesenchymal cells (Sakamoto et al. 2008). *Htra3* is a serine protease whose activity is absolutely required for its activity in TGF-beta signalling inhibition (Tocharus et al. 2004). *Htra3* was initially discovered to have a strong 100%:0% allelic bias, but the verification results showed only partial imprinting (with a 75%:25% allelic bias). This could be due to the low expression level of *Htra3* in the mouse placenta (RPKM < 1).

Finally, considering the last two imprinting validation tests, the pyrosequencing signal from *Trim23* was too low to determine the allelic expression percentage. Therefore we could neither confirm nor exclude the imprinting status of *Trim23*. *Gpsm2* was shown in our pyrosequencing assay to be not imprinted in the mouse placenta (Table 20). Overall the empirical verification rate is quite high (8 out of 9 successful tests), compared to other recently published transcriptome-wide surveys.

Assessment of the degree of maternal contamination in our placenta samples

One caution about identifying novel imprinted genes in the mouse and human placenta is the potential for maternal contamination (Proudhon and Bourc'his 2010). The placenta is a complex organ that consists of many different tissue and cell types. For term and near-term placenta, the contact of maternal and fetal tissues at the interface is challenging to separate by dissection, resulting in the potential for maternal contamination (Proudhon and Bourc'his 2010). In some studies of novel imprinted genes in the placenta, the possibility for maternal contamination cannot be excluded.

Several approaches were used to minimize maternal contamination in our samples. The first was to take special precautions during the dissection. From every sample collected, tissue was only taken from the middle of the placenta and only from what was clearly the fetal side. Then we washed the tissue many times in PBS to remove maternal blood. Second, we quantified the degree of contamination and chose the samples for RNA-seq that had the least maternal contamination (based on allelic expression ratio of several known imprinted genes with 100% paternal expression in placenta). If there were maternal contamination, paternally expressed imprinted genes would display expression from the maternal allele, and the degree of leakage could be used as a criterion to select the best samples. Third, several uterus samples near the placenta were collected at the same time, which allowed us to check the uterus expression level of a gene to determine the potential for contamination. Fourth, with the transcriptome-wide allelic expression profile, maternal contamination would be reflected by an allelic bias throughout the genome. By quantifying read counts of maternal alleles transcriptome-wide, it has been possible to estimate the degree of maternal contamination, and use this estimate to normalize SNP counts in the candidate imprinted genes.

Before we quantify the maternal bias introduced by maternal tissue contamination, we need to understand what other factors could also contribute to the deviation from 50:50 expression ratio of the two parental alleles. First, there is the possibility of global eQTL effects. As we observed from the allelic expression from a single gene, not all genes show 50:50 ratios.   If the AKR allele is associated with a *cis*-regulatory element, it could have higher expression from AKR allele in both reciprocal crosses. If we sum the SNP counts over all genes, it should be close to 50:50.   Second, since we are aligning reads with both the AKR and PWD sequences to the B6 reference genome, there will be a mapping bias toward the AKR allele, because the mouse strain genealogy shows that the AKR strain is closer to the B6 strain. So it was important to quantify and remove the mapping bias before we could assess the degree of maternal contamination (See Materials and Methods). Finally, imprinted X inactivation takes place in the mouse placenta, which means that the X-linked genes in females will be primarily expressed from the maternal allele (Sado and Ferguson-Smith 2005). If a gene/SNP has X chromosome homology, the reads might actually be from the X chromosome, which would create a spurious maternal bias. Consequently, in this analysis the X chromosomal genes were not assessed for imprinting status.

To illustrate these confounding factors for the deviation from 50:50 allelic expression, we present an example in (Table 23). Under a null model, if there is not any global eQTL effect or maternal bias or mapping bias, the allelic expression ratio will be 50:50 in both AKR x PWD and PWD x AKR crosses. Suppose there is 5% mapping bias. We would then always observe 55% expression from the AKR allele in both reciprocal crosses. If there is 5% maternal contamination, we would detect 55%

expression of the AKR allele in the AKR x PWD cross, because AKR is the mother in this cross, but 45% expression of the AKR allele in the PWD x AKR cross because PWD is the mother (Table 23). To quantify the degree of maternal contamination, we compute $(p_{1\_overall} - p_{2\_overall})/2$ as an metric whose expectation is zero if there is no maternal contamination (where $p_{1\_overall}$ is the total AKR allelic expression percentage from the AKR x PWD cross summing over all genes in the transcriptome, and $p_{2\_overall}$ is the total AKR expression percentage from the PWD x AKR cross, again summing over the transcriptome). With this metric, eQTL effects will cancel out, leaving a bias for un-imprinted genes only if there is maternal contamination.

In our placenta data, the total AKR allelic percentages are 51.99% and 51.52% in the AKR x PWD and PWD x AKR crosses respectively, before correcting the alignment bias (Table 23). After the mapping bias correction, the percentages are 50.50% and 50.17%, indicating that there is roughly a 1.5% mapping bias (Table 23). The maternal contamination is estimated to be 0.15% (Table 23), a quite tolerably low figure. For genes with moderate and high expression levels in our placenta samples, the effect of maternal contamination was negligible.

Table 23. Quantification of global maternal contamination percentage.

		Expected (5% maternal/mapping bias)			Observed total allelic %		
Maternal contamination		NO	NO	YES	mapping bias correction		after removal of SNPs with X-homology
Strain mapping bias		NO	YES	NO	before	after	
AKR x PWD	AKR(mat) %	50%	55%	55%	51.99%	50.50%	50.49%
	PWD(pat)%	50%	45%	45%	48.01%	49.50%	49.51%
PWD x AKR	AKR(pat) %	50%	55%	45%	51.52%	50.17%	50.18%
	PWD(mat)%	50%	45%	55%	48.48%	49.83%	49.82%
Quantification of maternal contamination		0%	0%	5%	-	0.17%	0.15%

Maternally expressed placenta-only imprinted genes: artifacts of maternal contamination?

Because of the maternal contamination problem, the imprinting status has been questioned for 13 placenta-only known imprinted genes (Proudhon and Bourc'his 2010). All are known to be maternally expressed imprinted genes. Among these genes, *Gatm*, *Pon3*, *Th*, *Tspan32*, *Cd81*, *Tssc4*, *Tnfrsf23* and *Osbpl5* have sufficient SNP coverage in our data to determine the imprinting status with confidence (Table 24). The genes, *Tfpi2*, *Pon2* and *Dcn*, do not show significant parent-of-origin effect in our data, suggesting that they may not be imprinted, at least at stage E17.5 in the AKR-PWD strain combination (Table 24). *Ppp1r9a* is detected to be imprinted with preferential maternal expression. *Nap1l4* is discovered to be a maternally expressed imprinted gene in the placenta (Engemann et al. 2000). Others have suggested that there may be leaky expression from the paternal allele (Umlauf et al. 2004). When we examined this gene in detail, we found four SNPs in the gene region, two in the exons and two in the introns. One exonic SNP shows biallelic expression, and the other one shows preferentially maternal expression (Table 25). The parent-of-origin effect is not significant if we sum over the two SNPs (Table 24). There are also two SNPs covered by the Illumina reads in the intron, with preferential paternal expression (Table 25). This gene maybe imprinted and there might be antisense non-coding transcript in the intronic region, or there may be complications from alternative splice products. Further investigation is needed to determine the imprinting status of *Nap1l4*.

Table 24. Coverage of known placenta-only imprinted genes whose imprinting status has been questioned.

Gene name	Mouse chr	Expr. allele	AKR x PWD allele counts		PWD x AKR allele counts		p_1	p_2	q-value	Conclusion
			AKR	PWD	AKR	PWD				
Gatm	Prox2	M	-	-	-	-	-	-	-	Not enough SNP coverage
Tfpi2	Prox6	M	0	8	4	8	0.00%	33.33%	0.16215	Not imprinted*
Ppp1r9a	Prox6	M	76	0	23.5	163	100.00%	12.63%	7.60E-43	Imprinted
Pon3	Prox6	M	-	-	-	-	-		-	Not enough SNP coverage
Pon2	Prox6	M	35.5	38	55	106	48.30%	34.16%	0.07218	Not imprinted
Th	Dist7	M	-	-	-	-	-	-	-	Not enough SNP coverage
Tspan32	Dist7	M	-	-	-	-	-	-	-	Not enough SNP coverage
Cd81	Dist7	M	-	-	-	-	-	-	-	Not enough SNP coverage
Tssc4	Dist7	M	-	-	-	-	-	-	-	Not enough SNP coverage
Nap1l4	Dist7	M	7	14.5	29	22	32.56%	56.86%	0.16468	could be imprinted
Tnfrsf23	Dist7	M	-	-	-	-	-	-	-	Not enough SNP coverage
Osbpl5	Dist7	M	-	-	-	-	-	-	-	Not enough SNP coverage
Dcn	Dist10	M	130	29	284	123	81.76%	69.83%	0.01028	Not imprinted

Table 25. Allelic expression ratios for SNPs in *Nap1l4* gene region.

Mourse chr	SNP position	AKRxPWD allele counts		PWDxAKR allele counts		*p1*	*p2*	qvalue	Direction	SNP type	Conclusion
		AKR	PWD	AKR	PWD						
chr7	150700360	7	6	20	19	53.85%	51.28%	1	Biallelic	exonic	ENSMUSE00000498169
chr7	150701525	9.5	0	15	16	100.00%	48.39%	0.0096	P	intron	NA
chr7	150707592	32	8	0	5	80.00%	0.00%	0.0039	P	intron	NA
chr7	150734989	0	8.5	9	3	0.00%	75.00%	0.0048	M	exonic	ENSMUSE00000667974

Maternal contamination could not only create false positive calls for maternally expressed imprinted genes, but also may result in a paternally expressed imprinted gene to be a false negative. *Zdbf2* could be one of such example. *Zdbf2* is detected in our data to be imprinted with preferential paternal expression, but it has been previously reported to be biallelically expressed in the placenta (Kobayashi et al. 2009). However, this could also be due to a different imprinting status of the same gene in different developmental stages/mouse strain combinations.

Is there a bias toward more maternally expressed imprinted genes in the placenta?

Contrasting patterns of genomic imprinting in the brain and placenta raises a series of questions about the mechanism and evolution of the control of imprinting. Previously, in a literature review of the tissue specificity and maternal vs. paternal expression of imprinted genes (Morison et al. 2005), it was noted that there is a paternal-brain/maternal placenta bias (Wang et al. 2008; Proudhon and Bourc'his 2010). The genes imprinted in the brain but not the placenta tend to be paternally expressed, whereas the genes imprinted in the placenta but not the brain tend to be maternally expressed (Figure 37A and Table 26) (*P*-value = 0.0001322, Fisher’s exact test). Our previous study also provided some suggestive evidence that the brain-paternal bias might be real (Wang et al. 2008). Here, we would like to ask whether the maternal-placenta bias is also true, or whether there might be an artifact due to the potential maternal contamination or limited sampling. We covered 35 known imprinted genes and verified 5 additional novel imprinted genes in this study. If we break them down by the direction of imprinting, we do not see a bias toward more maternally expressed genes (Figure 37B) (*P*-value = 0.6821, one-sided exact binomial test). If we examine all 251 candidates and classify them by their expression bias, we still see roughly equal numbers of paternally and maternally expressed candidates (Figure 37C), and the degree of allelic bias is statistically homogeneous between the two sets of reciprocal offspring.

Table 26. Tissue of imprinted genes and the imprinting direction.

Class of selected genes	Count	Gene names
Genes imprinted in brain but NOT placenta	19	*Nnat (peg5), Copg2, Nap1l5, Peg3, Usp29, Zfp264, Peg12 (Frat3), Mkrn3 (Zfp127), Magel2, Ndn, Snurf-Snrpn, Inpp5f_v2, Rasgrf1, Impact, Zdbf2, Calcr, Ube3a, Commd1, Kcnk9*
Genes imprinted in placenta but NOT brain	16	*Sfmbt2, Dlk1 (Peg9), Nesp, Gatm, Ascl2 (Mash2), Phemx (Tssc6), Tssc4, Slc22a18, Phlda2 (Tdag51), Nap1l4, Tnfrsf23, Osbpl5, Dcn, Grb10 (Meg1), Slc22a2, Slc22a3*
Genes imprinted in both brain and placenta	9	*Sgce, Peg10, Igf2, Plagl1, Rtl1(Peg11), Slc38a4, Asb4, Klf14, H13*
Other*	13	*Gnasxl, Mest (Peg1), Ins2, Dio3, Gnas, Zim1, Zim2, Zim3, Cd81 (Impt1), Kcnq1, Cdkn1c (p57KIP2), Htr2a, Igf2r*
Total	**57**	

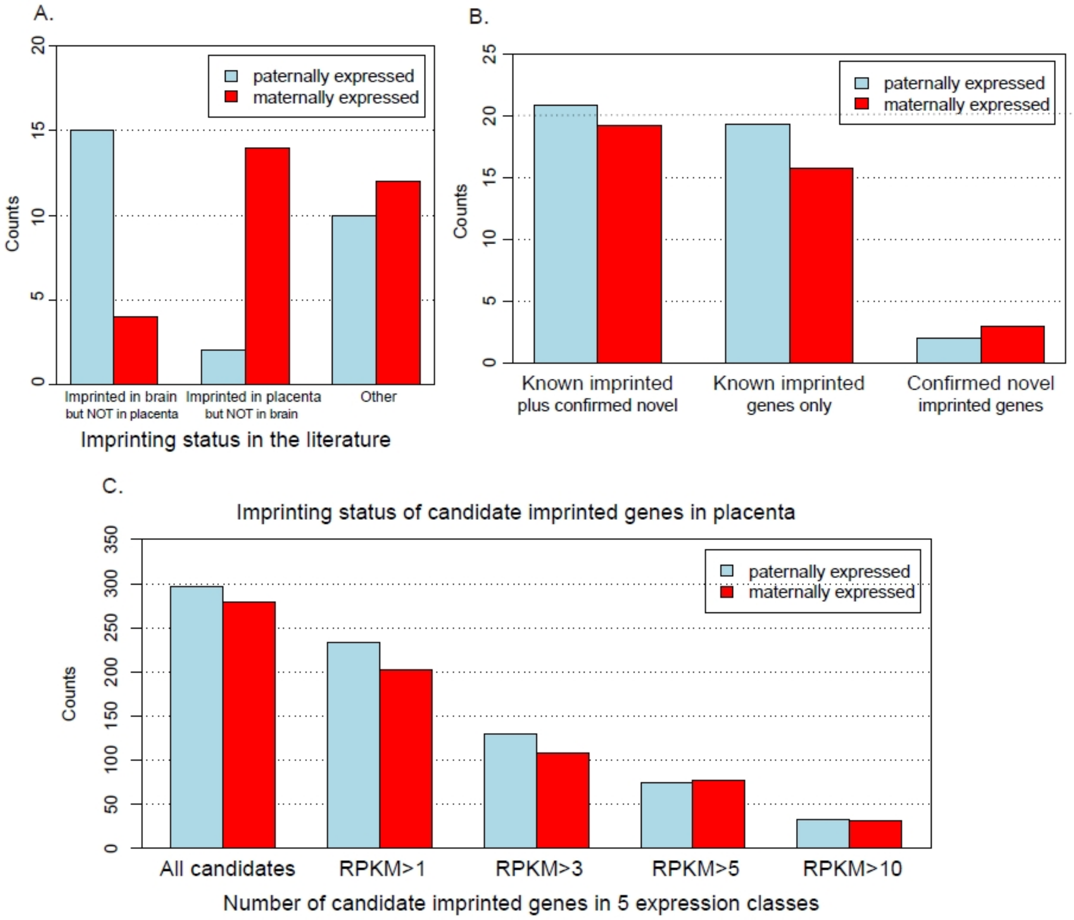
*This category is genes that are imprinted in other tissues (not brain and placenta).

Figure 37. Paternal vs. maternal novel and candidate imprinted genes.

(A). Paternal vs. maternal imprinting status in brain and placenta in the literature.

(B). The number of paternally and maternally expressed known and novel imprinted genes in our study.

(C). The number of paternally and maternally expressed candidate imprinted genes identified in our study.



Discussion

The number of imprinted genes in the mouse genome

Different studies present quite a wide range of estimates of the number of imprinted genes in the mouse genome, ranging from 100 genes (Luedi et al. 2007), to 600 genes (Luedi et al. 2005), to 1,300 genes (Gregg et al. 2010b), to 2,000 genes (Nikaido et al. 2003). There are several reasons for the broad range of these estimates. First, different studies used widely varying approaches, so they will have different false positive rates as well as different coverage and sensitivity. Second, different studies examined different tissues and developmental stages. In our study, we found 251 candidate imprinted genes in the E17.5 placenta falling in the set with a statistical false discovery rate of 0.01, but we also show empirically that the false discovery rate is more like 11%. Most of the top genes in the list are already known to be imprinted, indicating that the genomic imprinting community has done a commendable job of identification of the imprinted genes. Exhaustive enumeration of imprinted genes will require a large community-wide effort, including multiple replicates from multiple lines, with samples of many tissues and many developmental time points. If the results are to be interpreted with confidence based on RNA-seq data alone, a blocked and replicated design is essential (Auer and Doerge 2010).

Our intention here was to apply RNA-seq in a simple, unreplicated design to serve as a means of nominating candidates for subsequent validation. Among our candidate imprinted genes, we selected 10 for validation with biological replication and an independent assay for allele-specific expression. One pyrosequencing assay failed, but of the remaining nine, eight of the imprinting candidate genes were soundly confirmed. The candidates were chosen from a list with a theoretical false discovery rate of 0.01, whereas we observed that 1/9, or 11% of the candidates were false discoveries. The discrepancy between the *q*-value and the true verification rate could arise from several causes, most of which are expected to inflate the false positive rate of an unreplicated RNA-seq study. First, for lowly expressed genes, with only a few mRNA copies in the transcriptome, there is a chance during library construction that only one of the two alleles might be randomly ligated to the adapter and included in the final pool. After sequencing, the gene would resemble a monoallelically expressed gene, when in fact it is not. This is different from the random monoallelic expression that has been reported previously (Lomvardas et al. 2006; Gimelbrant et al. 2007), where single cells appear to fail to express both alleles. In applying quantitative RNA-seq for allele-specific expression, it is critical to assure that high library complexity is attained in order to avoid this allelic dropout caused by an insufficiently complex library. We might not get conclusive results for lowly expressed genes, so we need other independent methods to verify candidates with low expression levels. Second, sequencing bias and mis-alignments could also be a source of discordance. For the statistical test and subsequent inference of a *q*-value, several assumptions producing ideal experimental conditions are made: there is no sequencing bias, no mis-alignments, and the SNP-containing read counts are in proportion to the allelic expression ratio. However, in practice, these assumptions are easily violated. As a result, SNPs that truly have technical problems will be among the candidates that are found to be statistically significant by the Storer-Kim test, and these will be false positive calls. This is another reason why we need independent verification using an orthogonal technology like pyrosequencing. To account for these factors, we used more stringent filters. With our criterion 3 (defined as RPKM>3 and SNP coverage>20), only 113 significant candidates were left. Among the 113 genes, most of the known ones (23/35) and the confirmed novel ones (4/5) are preserved. Thus, by applying expression level and SNP coverage cut-offs, the degree of library complexity and SNP bias problems will be reduced, resulting in a lower false discovery rate. We will reach the theoretical FDR only if we completely remove these effects and meet all the ideal experimental conditions, and the most obvious way to improve the situation is by replication. But it is important to note that even with only a single replicate of RNA-seq runs from each cross, valid, verifiable novel imprinted genes were discovered.

Many pairs of known imprinting genes occur as overlapping sense-antisense pairs (Morison et al. 2005; Wang et al. 2008). With a double-strand cDNA RNA-seq library, the allelic expression from the sense vs. antisense transcripts cannot be distinguished, so SNPs that fall in regions where both strands are transcribed may produce false negative calls. By closely examining the SNPs within the candidates, we found some problematic genes with inconsistent SNPs or overlapping sense-antisense gene models. This could also contribute to the low verification rate. In the future, methods that allow preparation of strand-specific RNA-seq libraries should solve this problem (Levin et al. 2010).

Given the various limitations of RNA-seq studies, we conclude that an independent verification such as pyrosequencing or other allele-specific methods is necessary to confirm the imprinting status. It is also important to examine biological replicates, ideally from individuals from different strains to test the possibility of strain-specific effects. A much larger study, with a well-replicated and blocked design of multiple RNA-seq runs (Auer and Doerge 2010) would be needed to generate a definitive count of the number of imprinted genes. From our data, ~4.5% (251) of the 5,527 genes having sufficient data to perform the test exhibit significant imprinting in the placenta. Given the empirical FDR of 11% for this test, 224 genes are expected to be verified. However, the 11% false positive rate was seen among the subset of genes with the lowest q-values, and if all 251 genes were tested, it would likely be higher. On the other hand, the gene list of 251 was generated using strict selection criteria (RPKM >1, p_1 > 0.65 or p_1<0.35), and the un-measured false negative rate be inflated. Therefore, while the experiment produces an estimate of 224 imprinted genes, the uncertainty in false positive and false negative rates suggest that a range of 100-250 genes may be the most supportable. Because this study was restricted to the E17.5 placental tissue in AKR-PWD crosses, the true number of imprinted genes across all tissues and stages is likely to be larger.

Artifacts in novel imprinted gene identification

There are various sources of artifacts in the identification of imprinted genes (Proudhon and Bourc'his 2010). The first one is that there may be random monoallelic expression instead of genomic imprinting. We verified our candidates in multiple individuals to exclude this possibility. Second is that the allelic bias could be generated by an eQTL effect. In our study, we used reciprocal F1s, allowing us to distinguish parent-of-origin effects from the eQTL effects. Third, there may be a strain-specific PCR bias. Random primers were used in the Illumina library preparation, making PCR bias unlikely, and our confirmation method using pyrosequencing did not employ the same PCR primers. The fourth class of artifact is maternal contamination in the dissected placenta tissues. We took pains to avoid and to quantify the maternal contamination in our samples, and our quantitative analysis demonstrates that these efforts were successful (Table 23). Another artifact that might spuriously lead to allelic bias is homology to the X chromosome. Males inherit the X chromosome from the mother, so the X-linked genes in males will have 100% maternal expression. In female mouse embryos and placental tissues of fetal origin, there is imprinted X inactivation, resulting in preferential expression from the maternal allele. If an autosomal gene/SNP has X homology, there could be non-specific amplification during RT-PCR or mis-alignment for the RNA-seq. Either case would result in spurious identification of a maternally expressed imprinted gene. This could happen even with zero maternal contamination. Careful attention to this possibility during read-mapping should minimize its impact, although it is hard to exclude the possibility entirely.

Is there a paternal-brain and maternal-placenta bias?

Previous literature indicates that there is a maternal bias to allelic expression of imprinted genes in the placenta (Morison et al. 2005; Proudhon and Bourc'his 2010). This could be real, or it may be due to over-estimation of maternally expressed imprinted genes due to maternal contamination or under-estimation of the paternally expressed imprinted genes. From our results, we did not observe any bias toward maternally expressed imprinted genes in the placenta (Figure 37). We think this is simply because some paternally expressed genes are not known to be imprinted in placenta. In the 12 known imprinted genes identified in our data without prior reports of placenta imprinting, 8 of them are paternally expressed. In the list of novel candidate imprinted genes, we did not find any bias toward maternally expressed genes. This is also consistent with the minimal maternal contamination estimated in our study.

Conclusion

We have shown that even an unreplicated RNA-seq study can identify a highly informative set of genes showing parent-of-origin allelic expression differences that validated with a quite acceptable rate (89%).   This provides an excellent set of candidates for genes showing genomic imprinting, including 5 novel genes that we validated by pyrosequencing in multiple biological samples.   The finding that *Phf17* shows strong paternally expressed imprinting is especially intriguing, as this gene is part of a histone H4 transacetylase complex, and may specify a parent-of-origin differential histone acetylation.   It is not immediately clear why *Pde10a*, a cAMP and cGMP phosphodiseterase should be maternally expressed and imprinted in the placenta, but the allelic expression bias is well validated. A larger scale RNA-seq study with this reciprocal cross design, sequencing to greater coverage and using biological replication would also be highly informative, allowing assessment of splice isoform-specific imprinting, sex difference in imprinting, inter-strain variability, and more.

CHAPTER 4

Biased paternal X inactivation in mouse brain[3]

Abstract

X-inactivation in female eutherian mammals has long been considered to occur at random in embryonic and postnatal tissues. Methods for scoring allele-specific differential expression with a high degree of accuracy have recently motivated a quantitative reassessment of the randomness of X inactivation. After RNA-seq data revealed what appeared to be a chromosome-wide bias toward under-expression of paternal alleles in mouse tissue, we applied pyrosequencing to mouse brain cDNA samples from reciprocal cross F1 progeny of divergent strains and found a small but consistent and highly statistically significant excess tendency to under-express the paternal X chromosome. The bias toward paternal X inactivation is reminiscent of marsupials (and extraembryonic tissues in eutherians), suggesting that there may be retained an evolutionarily conserved epigenetic mark driving the bias. Allelic bias in expression is also influenced by the sampling effect of X inactivation and by cis-acting regulatory variation (eQTL), and for each gene we quantify the contributions of these effects in two different mouse strain combinations while controlling for variability in Xce alleles. In addition, we propose an efficient method to identify and confirm genes that escape X inactivation in normal mice by directly comparing the allele-specific expression ratio profile of multiple X-linked genes in multiple individuals.

[3] This work was published in Wang X, Soloway PD, Clark AG., 2010. Paternally biased X inactivation in mouse neonatal brain. *Genome Biol.*, **11(7):**R79.

Introduction

In placental mammals, dosage compensation is achieved during embryonic development by random inactivation of one of the two female X chromosomes (Straub and Becker 2007; Payer and Lee 2008). In male germline tissue, both sex chromosomes are inactivated through meiotic sex chromosome inactivation (MSCI). In the mouse placenta, the paternal X chromosome (Xp) is inactivated in extra-embryonic tissues. In female zygotes, at the two-cell stage, Xp is activated and X-linked genes are transcribed from both parental X chromosomes. In the mouse, starting from the eight-cell stage, the paternal X is inactivated through a process known as imprinted X inactivation (Huynh and Lee 2001; Huynh and Lee 2005; Heard and Disteche 2006). Subsequently the paternal X is reactivated and, in the mouse, random X inactivation occurs around the implantation stage (about day 6.5) in the embryonic tissue, with only one of the two X chromosomes remaining activated (Cheng and Disteche 2004), while the extraembryonic tissues retain imprinted X inactivation and express only the maternal X. This would seem to be a cumbersome way to accomplish dosage compensation, and an evolutionary perspective may shed light on the origins of the process. In humans, there remains some controversy surrounding the presence of imprinted X inactivation. There is some evidence of imprinted inactivation in pre-implantation embryos, but it has not been fully confirmed (Goto et al. 1997; van den Berg et al. 2009). Most placental mammals appear to perform dosage compensation in the same fashion as the mouse, whereas in marsupials X inactivation is not complete but instead preferentially silences the paternal allele in both embryonic and extraembryonic tissues (Cooper et al. 1990). In the egg-laying monotremes (platypus and echidna), both alleles of X-linked genes are transcribed, and some of the genes do not display dosage compensation while others show some degree of compensation by gene-specific transcriptional inhibition (Deakin et al. 2008). This is consistent with the fact that the platypus X chromosomes are not homologous to the human X, but instead have molecular sequence similarity to the chicken Z chromosome (Warren et al. 2008), and birds do not appear to effect dosage compensation by Z inactivation (Arnold et al. 2008).

In eutherian mammals, imprinted X inactivation is reported in extraembryonic tissues, and in embryonic tissue early in development prior to random X inactivation. Skewed X inactivation can affect the severity of human disorders such as PHACES (posterior fossa malformations, hemangiomas, arterial anomalies) (Levin and Kaler 2007), Rett Syndrome (Krepischi et al. 1998) and other diseases (Martinez et al. 2005; Talebizadeh et al. 2005; Shimizu et al. 2006). However, aside from extraembryonic tissues, it is widely thought that placental mammals inactivate one or the other X chromosome in a purely random fashion (except the loci that clearly influence choice such as the *Xce* alleles, and *Xist* polymorphisms). Two earlier studies found possible parental influence on the biased expression of the maternal allele, but their data are only from a single X-linked gene, and so it is not possible to distinguish between explanations involving single gene effects (such as imprinting) or those that would generate chromosome-wide patterns (such as X-inactivation) (Forrester and Ansell 1985; Fowlis et al. 1991). In this report, we quantify the relative paternal and maternal expression level of 33 X-linked genes from P2 neonatal brains of 18 female mice for each of the two reciprocal F1 progeny of AKR and PWD strains. These data reveal a significant and consistent elevated expression level from the maternal X, consistent with preferential paternal X inactivation in normal non-extraembryonic tissue. The same pattern of preferential paternal X inactivation was also seen in our examination of reciprocal F1 progeny of the B6 and CAST strains.

Not all X-linked genes are subject to X inactivation. In humans, Carrel and Willard (2005) reported that roughly 15% of the X-linked genes are expressed from both alleles. To date, in the mouse, four genes that escape X inactivation have been discovered outside the pseudo-autosomal region (PAR) (Adler et al. 1991; Agulnik et al. 1994; Disteche et al. 2002). Human studies have nearly completed a scan for genes that escape X inactivation by thorough testing of murine-human hybrid cell lines, as well as human fibroblast samples (Brown and Willard 1989; Brown et al. 1997; Carrel et al. 1999; Carrel and Willard 2005). Early mouse studies employed female mice carrying the T(X;16)16H (T16H) translocation (Agulnik et al. 1994; Greenfield et al. 1998), and recently Yang *et al.* (Yang et al. 2010) showed from RNA-seq of mouse hybrid cell lines that biallelic expression is found for 13 of the 393 X-linked genes examined. Here, we employ a novel method to detect X inactivation status using normal somatic tissue (P2 neonatal brains) from reciprocal mouse crosses, by comparing the allele-specific expression profiles among many X-linked genes and autosomal genes in multiple individuals. We confirm the status of two known mouse genes that escape X inactivation, and see a consistent pattern wherein one gene partially escapes X inactivation. We also test 13 orthologs of known genes that escape

X inactivation in humans and find that all are subject to X inactivation in mouse. The method presented here is a valuable complement to the current methods, and by applying it to all mouse and human X-linked genes, it will be possible to build an exhaustive catalog of mouse and human X inactivation escapers.

Materials and Methods

Mouse Strains and crosses

Four mouse strains (AKR/J, PWD/PhJ, C57BL/6 and CAST/EiJ) were purchased from the Jackson Laboratory (Brondum-Nielsen and Pedersen 2001). We performed reciprocal crosses with two strain combinations (PWD/PhJ x AKR/J, AKR/J x PWD/PhJ, C57BL/6 x CAST/EiJ, CAST/EiJ x C57BL/6). 18 female P2 F1 mice were generated from 5 litters from the PWD/PhJ x AKR/J cross (PWD x AKR for short). 18 female P2 F1 mice were generated from 4 litters from the AKR/J x PWD/PhJ cross (AKR x PWD for short). 11 female P2 F1 mice were generated from 3 litters from the C57BL/6 x CAST/EiJ cross (B6 x CAST for short). 11 female P2 F1 mice were generated from 4 litters from the CAST/EiJ x C57BL/6 cross (CAST x B6 for short). Total RNA samples were extracted from the P2 F1 mouse whole brains using the Qiagen RNeasy Lipid Tissue Mini Kit. RNA concentrations and A260nm/A280nm ratios were checked with a NanoDrop ND-1000 spectrophotometer. RNA integrity was checked using the Agilent 2100 Bioanalyzer. All of the samples have a RIN (RNA integrity number) in the range 9.8-10.0 (RIN_{max} = 10.0).

All procedures involving mice have been approved by the Institutional Animal Care and Use Committee at Cornell University (protocol number 2002-0075). Cornell University is accredited by AAALAC.

Illumina sequencing of the transcriptome and allele-specific expression analysis

Experimental procedures, statistical methods, and data from our original RNA-seq study are available (Wang et al. 2008) (see Chapter 2).

Quantification of allele-specific expression of 35 genes by pyrosequencing

33 X-linked genes (*Ctps2*, *Plxna3*, *Syn1*, *Phf6*, *Taf1*, *Utx*, *Syap1*, *Maoa*, *Zfx*, *Xist*, *Usp9x*, *Ddx3x*, *Ikbkg*, *Prkx*, *Eif2s3x*, *Nxt2*, *Gpm6b*, *Nudt11*, *Zbtb33*, *Sh3bgrl*, *Fundc1*, *Wdr13*,*Hcfc1*, *Rbmx*, *Uba1*, *L1cam*, *Ofd1*, *Crsp2*, *Cstf2*, *Ids*, *Jarid1c*, *Tsix* and *Xite*) and 8 autosomal genes (*Pex7*, NM_023057, *Prkar2b*, *Hibadh*, *Rgs17*, *Cab39l*, *Trpm6* and *Tmem109*) were selected for quantification of expression level from the two parental alleles using pyrosequencing in 18 female brain samples from each of AKR-PWD reciprocal crosses. 18 X-linked genes and the same 2 autosomal genes were examined in 11 female brain samples from each of B6-CAST reciprocal crosses. The X-linked gene selection criteria include genes having a detectable expression level in the Illumina sequence data, including known mouse genes that escape X inactivation (Disteche et al. 2002) and orthologs to human genes that escape X inactivation and are subject to X inactivation, respectively (Carrel and Willard 2005). Genes were selected to span the entire mouse X chromosome with a relatively even distribution. The eight autosomal genes were selected at random among the genes that have a detectable expression level in the Illumina sequence data. In addition, one male sibling of the tested females was included from each litter and was

used as a pyrosequencing control, since males should have 100% maternal allele expression, if there is no Y homolog of that gene.

Pyrosequencing PCR and sequencing primers were designed for the selected X-linked and autosomal control genes with the pyrosequencing Assay Design Software Version 1.0.6 (Biotage AB). To guarantee that there were no SNPs within the primers, SNP positions in the Perlegen SNP database (Frazer et al. 2007) were labeled and excluded when designing the primers. The detailed PCR amplification and allele-specific pyrosequencing protocol can be found in (Wang et al. 2008). Pyrosequencing was done twice for each gene in each sample, and the mean difference is 1.90%, with standard deviation 1.52%, indicating high reproducibility.

Statistical analysis

Cluster analysis of the X-linked genes. 33 X-linked genes and 2 autosomal genes were clustered using the Agglomerative Nesting Hierarchical Clustering method (Blashfield 1991), which is implemented in the `cluster` package (version 1.11.11) (Buettner et al. 2004) in R (version 2.62) (Bullard et al. 2010). Absolute Pearson Correlation distance was used as the dissimilarity measure.

Nested ANOVA methods. To determine whether there is significant maternal bias and/or sampling effect, a three-factor nested ANOVA model was implemented.

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{k(j)} + \varepsilon$$

In this model, y_{ijk} is the response variable of observed PWD maternal/paternal expression ratio for the i^{th} gene (*cis*-regulatory effect), k^{th} individual (sampling effect) in j^{th} cross (parent-of-origin effect). μ is the mean PWD expression ratio. α_i is the fixed effect for individual genes (i = 1,…,27). β_j is the parent-of-origin effect (j = 1, 2 for the PWD x AKR and AKR x PWD crosses). $\gamma_{k(j)}$ is the sampling effect nested within the parent-of-origin effect (k = 1,…, 18). The data were analyzed in SAS using the Proc Mixed procedure, with gene and mother as fixed factors, and individual as random factor nested in mother.

Estimation of number of brain-forming cells during X inactivation. For each of the PWD-AKR reciprocal crosses, we simulated the mean sampling variance 1,000 times for the number of brain-forming stem cells *N* ranging from 30 to 150, using the `rbinom()` function in R (version 2.62) (Bullard et al. 2010). The mean and 95% CI were estimated by interpolation.

Results

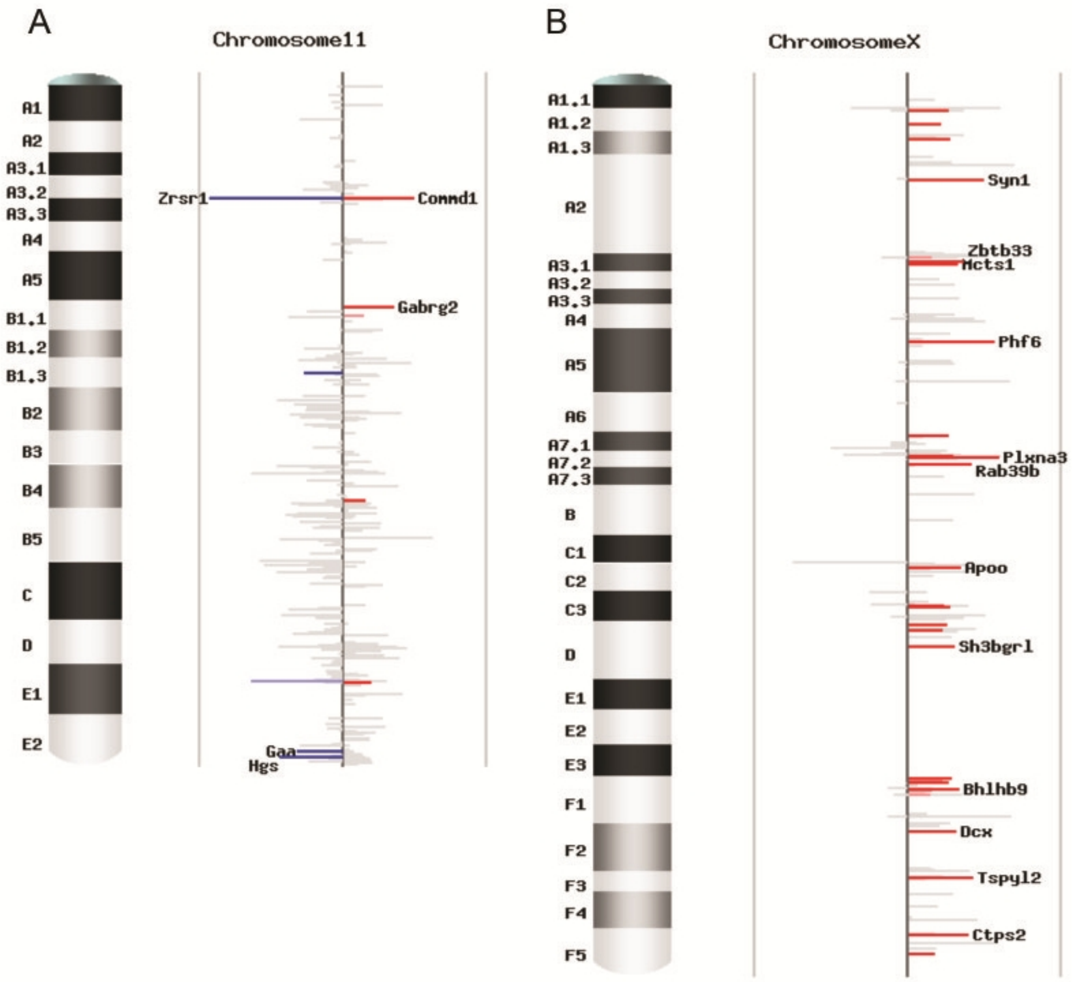
Maternal bias in transcriptome-wide differential allelic expression

In our previous effort to identify novel imprinted genes in mouse (Wang et al. 2008), we performed an “RNA-seq” study in which >69 million sequence reads were sampled from the transcriptomes of reciprocal F1 female P2 neonatal brains (AKR/J and PWD/PhJ strains) by Illumina short-read sequencing. Relative expression ratios of the two parental alleles were obtained by directly counting the allele-specific sequence reads at the SNP positions within the transcripts (Wang et al. 2008). 5,076 unique Entrez genes had a coverage of four or more sequence reads overlapping each SNP position in both reciprocal crosses across the mouse genome. The imprinting status was quantified as the difference between the AKR percentages in the F1 progeny derived from the two reciprocal crosses. For most genes this difference in expression was close to zero, indicating a lack of significant imprinting (Wang et al. 2008) . The known imprinted genes and novel imprinted gene candidates had an obvious and highly statistically significant bias in allelic expression. When we compared the pattern of skewed allelic expression of autosomes with the X chromosome, we noted that for every autosome, there was approximately the same number of preferentially paternally and maternally expressed genes. However, X chromosomal genes showed consistently elevated maternal expression, and there was not a single significant paternally over-expressed gene (Figure 38). Because we saw exclusively maternal over-expression in progeny of both reciprocal crosses of PWD and AKR strains, the results cannot be explained by differences in alleles at the X chromosome control element (*Xce)*, a locus that influences in an allele-specific manner the probability of X inactivation (Cattanach and Isaacson 1967)**.**

There are three possible explanations for the maternal bias in X-linked expression. First, the pattern might be driven by each X-linked gene having its own independent factors driving its imprinting. Second, since the RNA-seq data are from only two mice, we cannot exclude the possibility of a sampling effect caused by the small number of cells at the time of X inactivation. X inactivation initiates when the total number of cells committed to become brain is only 10 to 50 (Gartler and Riggs 1983). If X inactivation occurs as an independent Bernoulli trial for each cell, then the count of cells expressing maternal vs. paternal alleles would have a binomial variance. Such sampling effects will yield an X inactivation process that may still be truly random for all single cells, but in aggregate there may appear to be a bias due to the small cell sample size at the time of X inactivation. This phenomenon was seen in humans by an allele-specific methylation assay of the *AR* (androgen receptor) gene (X chromosome inactivation assay) (Amos-Landgraf et al. 2006). The third possibility is that there may be preferential inactivation of the paternal X chromosome, in violation of the standard notion of random X-inactivation, and that this bias may act on top of the sampling effect. In this study we applied pyrosequencing to multiple F1 progeny samples to determine whether the skewed allelic expression we saw in our mouse imprinting study was due to such a sampling effect.

Figure 38. Chromosomal scans of imprinting status for chromosome 11 and X.

Each plot contains unique Entrez genes covered by SNP-containing Illumina reads with counts no less than 4 in each reciprocal cross. The height of each bar is the difference of the AKR percentage in the two reciprocal crosses (p1-p2), representing the intensity of imprinting. The color indicates for the direction of expression bias, blue for paternal over-expression and red for maternal over-expression. The intensity of the color represents the significance, grey for not significant (*q*-value ≥ 0.10), lighter blue and pink for marginally significant (0.05 ≤ *q*-value < 0.10), darker blue and red for significant (*q*-value < 0.05). The gene name is indicated for the instances where $|\ p_1\text{-}p_2| \geq 0.3$. Data are from Chapter 2.


A
Chromosome11
A1
A2
A3.1
A3.2
A3.3
A4
A5
B1.1
B1.2
B1.3
B2
B3
B4
B5
C
D
E1
E2
Zrsr1
Commd1
Gabrg2
Gaa
Hgs
B
ChromosomeX
A1.1
A1.2
A1.3
A2
A3.1
A3.2
A3.3
A4
A5
A6
A7.1
A7.2
A7.3
B
C1
C2
C3
D
E1
E2
E3
F1
F2
F3
F4
F5
Syn1
Zbtb33
Mcts1
Phf6
Plxna3
Rab39b
Apoo
Sh3bgrl
Bhlhb9
Dcx
Tspyl2
Ctps2

The maternal bias is unlikely to be due to individual imprinted genes

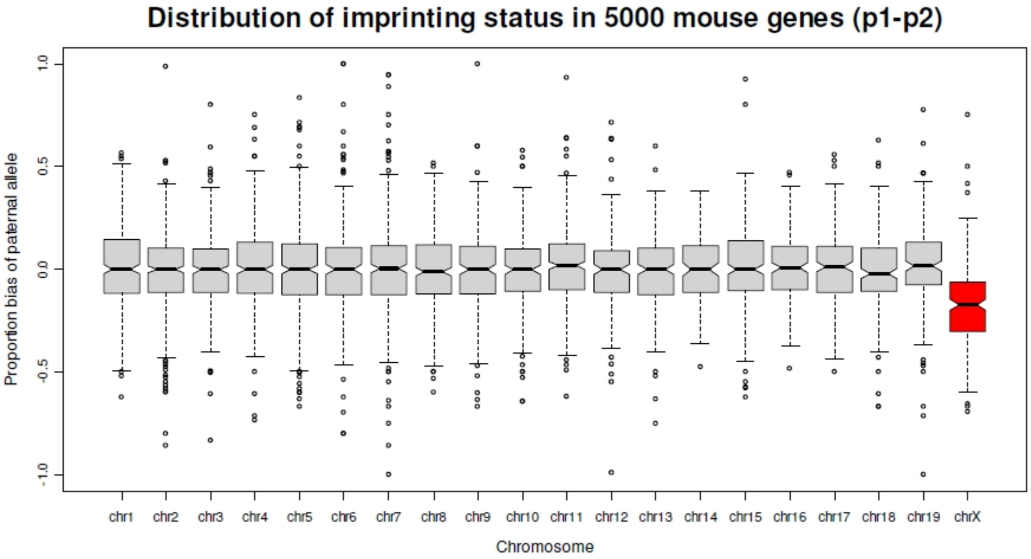
To determine whether the maternal bias is due to several X-linked imprinted genes or a chromosome-wide effect, we plotted the distribution of the difference in expression between reciprocal F1 progeny for the X chromosome from our RNA-seq data (Figure 39). The distributions of all autosomes are centered near zero (mean is 0.000975), whereas the distribution for the X chromosome is shifted to a mean of -0.176. Pairwise Kolmogorov-Smirnov tests revealed a significant difference between the X chromosome and autosomal allelic bias ($P < 10^{-12}$ for all chromosomes), but no significant heterogeneity among autosomes, indicating that the bias in X-linked allelic expression is a chromosome-wide effect (Table 27). Further verification in multiple individual mice confirmed that none of the 26 tested X-linked candidate imprinted genes are consistent with classical genomic imprinting. We observed variable allele-specific expression ratios in multiple individuals of the two reciprocal crosses. If the maternal bias that we observed were caused by independent imprinting of each gene, and if there is no prior reason to assume a bias toward maternal or paternal imprinting, then the chance that all 26 genes are maternally-expressed imprinted genes would be $(1/2)^{26}$, a vanishingly small number. We conclude that biased X inactivation is a much more parsimonious explanation than maternally-biased imprinting for the observed maternal bias in allelic expression of so many X-linked genes.

Table 27. Kolmogorov-Smirnov tests of p_1-p_2 distribution of different chromosome pairs.

	Chr 1	Chr 2	Chr 3	Chr 4	Chr 5	Chr 6	Chr 7	Chr 8	Chr 9	Chr 10	Chr 11	Chr 12	Chr 13	Chr 14	Chr 15	Chr 16	Chr 17	Chr 18	Chr 19	ChrX	All autosomes
Chr 1	-	0.2588	0.2137	0.9550	0.7966	0.6913	0.7087	0.5559	0.4206	0.3658	0.9495	0.0449	0.1937	0.4402	0.9977	0.9380	0.8126	0.0643	0.5611	0	0.2914
Chr 2	0.2296	-	0.9527	0.6511	0.5522	0.7462	0.4561	0.9523	0.9782	1.0000	0.1888	0.7016	0.9592	0.6598	0.3107	0.6342	0.1798	0.3211	0.1338	0	0.9180
Chr 3	0.1858	0.9300	-	0.4201	0.6395	0.9249	0.4190	0.7752	0.9422	0.9973	0.1588	0.6755	0.7653	0.6303	0.4345	0.8094	0.2571	0.5809	0.3073	0	0.8923
Chr 4	0.9242	0.6034	0.3738	-	0.8124	0.8193	0.9488	0.9666	0.8799	0.6294	0.6147	0.1118	0.4432	0.8103	0.7138	0.7491	0.5353	0.1593	0.2144	0	0.7635
Chr 5	0.7516	0.4968	0.5958	0.7644	-	0.8971	0.9418	0.8039	0.8216	0.7992	0.6489	0.2624	0.5387	0.6841	0.8025	0.5627	0.7821	0.1253	0.1244	0	0.7874
Chr 6	0.6358	0.7022	0.8856	0.7708	0.8520	-	0.6914	0.3950	0.9218	0.7791	0.3747	0.4931	0.7036	0.8578	0.8500	0.6459	0.3834	0.1349	0.4253	0	0.5896
Chr 7	0.6658	0.4168	0.3742	0.9236	0.9172	0.6424	-	0.5371	0.3716	0.6472	0.9062	0.2146	0.6205	0.7261	0.8744	0.7943	0.7841	0.0427	0.4909	0	0.4818
Chr 8	0.5108	0.9244	0.7374	0.9472	0.7550	0.3522	0.4832	-	0.9955	0.7565	0.2615	0.2878	0.8890	0.9711	0.6552	0.5921	0.3303	0.3276	0.0843	0	0.9286
Chr 9	0.3818	0.9618	0.9030	0.8338	0.7730	0.8822	0.3374	0.9868	-	0.9371	0.1287	0.4117	0.7346	0.7981	0.7073	0.6021	0.4116	0.7365	0.1688	5.11E-15	0.9451
Chr 10	0.3330	0.9996	0.9918	0.5796	0.7480	0.7186	0.6018	0.7042	0.8946	-	0.4170	0.5664	0.9071	0.8741	0.4593	0.6397	0.3876	0.4026	0.2074	0	0.9118
Chr 11	0.9182	0.1618	0.1382	0.5718	0.5988	0.3364	0.8656	0.2312	0.1076	0.3936	-	0.1246	0.3860	0.8249	0.7825	0.8325	0.8555	0.0062	0.5422	0	0.1094
Chr 12	0.0414	0.6508	0.6232	0.1026	0.2318	0.4454	0.1868	0.2562	0.3494	0.5178	0.1052	-	0.8447	0.2924	0.1374	0.4313	0.2847	0.2119	0.1959	5.55E-16	0.2772
Chr 13	0.1666	0.9332	0.7102	0.4016	0.4888	0.6458	0.5770	0.8476	0.6844	0.8598	0.3440	0.7858	-	0.8826	0.1991	0.5899	0.5278	0.2473	0.3489	1.11E-16	0.6609
Chr 14	0.3826	0.6120	0.5772	0.7500	0.6364	0.8054	0.6624	0.9512	0.7464	0.8286	0.7712	0.2538	0.8300	-	0.5270	0.8882	0.9336	0.3169	0.4979	7.79E-13	0.7673
Chr 15	0.9926	0.2802	0.3938	0.6610	0.7540	0.8078	0.8272	0.5868	0.6338	0.4150	0.7516	0.1202	0.1638	0.4688	-	0.8316	0.6601	0.2203	0.4923	0	0.4799
Chr 16	0.9072	0.5884	0.7568	0.6974	0.5074	0.5720	0.7442	0.5238	0.5374	0.5724	0.7794	0.3696	0.5216	0.8408	0.7712	-	0.9471	0.1021	0.6725	0	0.5873
Chr 17	0.7618	0.1532	0.2314	0.4706	0.7274	0.3440	0.7248	0.2976	0.3668	0.3456	0.7988	0.2442	0.4626	0.8940	0.5866	0.9142	-	0.0566	0.5698	5.55E-16	0.2752
Chr 18	0.0548	0.2854	0.5142	0.1326	0.1022	0.1132	0.0376	0.2942	0.6662	0.3560	0.0048	0.1770	0.2016	0.2654	0.1780	0.0804	0.0452	-	0.0353	2.22E-16	0.1147
Chr 19	0.5182	0.1132	0.2628	0.1906	0.1002	0.3722	0.4362	0.0664	0.1424	0.1816	0.4868	0.1670	0.3076	0.4366	0.4342	0.6094	0.5080	0.0270	-	0	0.0683
Chr X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-	0
All autosomes	0.2644	0.8924	0.8724	0.7312	0.7676	0.5666	0.4472	0.9074	0.9296	0.8950	0.1030	0.2470	0.6324	0.7388	0.4498	0.5586	0.2678	0.1118	0.0608	0.0000	-

🟨	K-S test p-value
🟩	K-S test p-value by bootstrap

Figure 39. Distribution of imprinting status of 5000 genes covered by the RNA-seq study in mouse brain. Boxplot of the imprinting status for autosomes and the X chromosome. The *y*-axis is proportion bias from the paternal allele (p_1-p_2). From the plot, for all autosomes, the mean is very close to zero. However, the mean for X chromosome is -0.17, which indicate a 17% maternal bias in allele-specific expression. The difference between X and autosome is extremely significant from non-parametric statistical test of distributions. So this is a chromosome-wide effect, rather than effect of single imprinted genes.



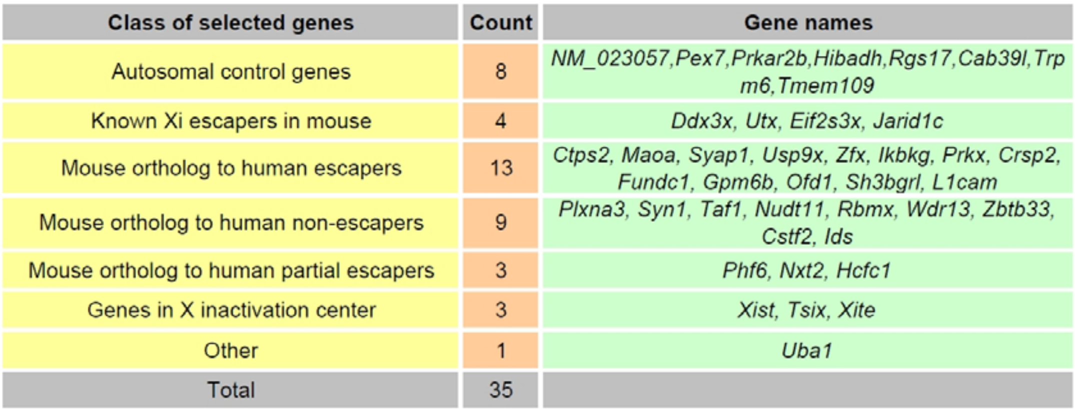
Sources of variability in allele-specific expression

To further elucidate the cause of maternal bias in expression of X-linked genes in (Figure 38), we employed pyrosequencing to quantify the parental expression ratios of 33 X-linked genes and 8 autosomal genes in 18 female P2 brains in each of the PWD and AKR reciprocal crosses (Marsh 2007). First, we selected genes that had a detectable level of expression in our Illumina RNA-seq data. We included the known mouse genes that escape X inactivation as well as mouse orthologs to human genes that escape X inactivation, genes with variable X inactivation status, and genes that are subject to normal X inactivation (Carrel and Willard 2005). We also randomly selected eight autosomal genes as controls (Table 28).

There are three possible sources of variability for the allele-specific expression ratio we quantified by pyrosequencing: a sampling effect, a *cis*-regulatory effect (also called an eQTL effect) and a parent-of-origin effect. We already explained the sampling and parent-of-origin effects as possible causes of the maternal expression bias. An eQTL effect occurs when there is a *cis*-regulatory polymorphism near the gene. In this case, if the PWD allele of the regulatory variant confers elevated expression, then for all progeny and in both reciprocal crosses, the effect of the PWD *cis*-acting effect will be to increase the PWD allele expression relative to the AKR allele. The eQTL effect may be different for each gene. Since the eQTL effect drives a bias in expression among progeny of both reciprocal crosses, it cannot cause the observed maternal bias.

We illustrate the possible patterns of differential allelic expression under the three different effects in (Figure 40). For autosomal genes and X-linked genes that are subject to X inactivation, because there is no sampling effect (no X inactivation), there will not be much variability (Figure 40A). The only source of allele-specific variability is the measurement error of the pyrosequencing assay. For the X-linked genes that are subject to X inactivation, because there are only a few brain-forming cells at the time of X inactivation, there is a sampling effect over the counts of cells expressing one X or the other, and the among-individual variance will be large (Figure 40B). The standard model for X-inactivation posits that the offspring from the two reciprocal crosses should have essentially the same mean and variance in their allele-specific expression ratios. Among a set of X-linked genes that display both a sampling effect and an eQTL effect, there will be differences in mean expression percentages from the PWD allele, but the means for the two reciprocal crosses are still expected to be the same (Figure 40C). Only if there is a parent-of-origin effect will the means of the PWD expression percentages be different between the two reciprocal crosses, and the bias will be in the same direction for every single gene that is subjected to X inactivation (Figure 40D).

Table 28. Gene selection for Pyrosequencing.



Class of selected genes	Count	Gene names
Autosomal control genes	8	*NM_023057,Pex7,Prkar2b,Hibadh,Rgs17,Cab39l,Trpm6,Tmem109*
Known Xi escapers in mouse	4	*Ddx3x, Utx, Eif2s3x, Jarid1c*
Mouse ortholog to human escapers	13	*Ctps2, Maoa, Syap1, Usp9x, Zfx, Ikbkg, Prkx, Crsp2, Fundc1, Gpm6b, Ofd1, Sh3bgrl, L1cam*
Mouse ortholog to human non-escapers	9	*Plxna3, Syn1, Taf1, Nudt11, Rbmx, Wdr13, Zbtb33, Cstf2, Ids*
Mouse ortholog to human partial escapers	3	*Phf6, Nxt2, Hcfc1*
Genes in X inactivation center	3	*Xist, Tsix, Xite*
Other	1	*Uba1*
Total	35	

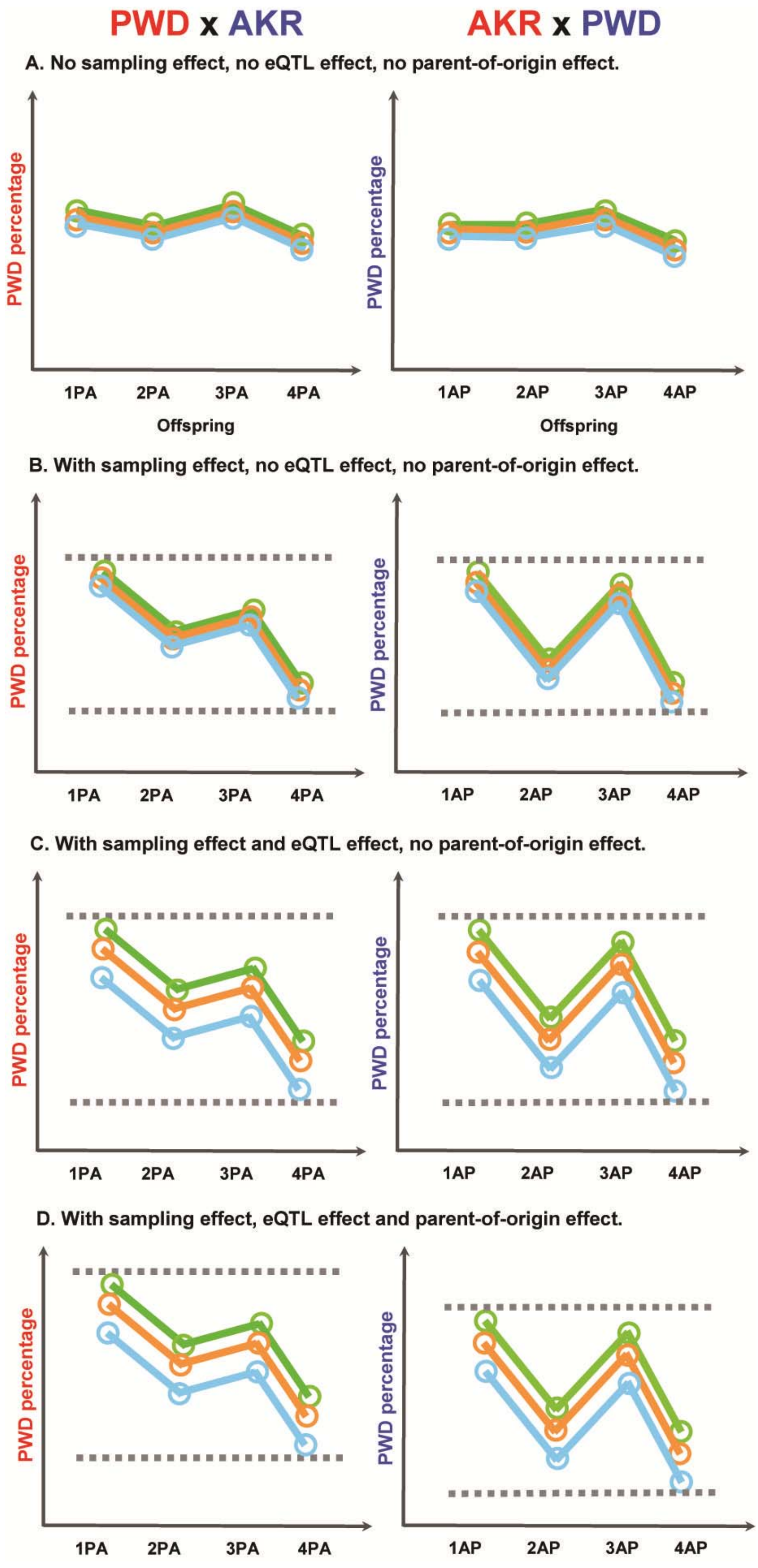
Figure 40. Three effects that cause the allele-specific expression variability. In these plots, the *y*-axis quantifies the proportion of expression from the PWD allele (PWD percentage). The *x*-axis provides an arbitrary index for different individuals from the reciprocal crosses. The left panels show offspring from the PWD x AKR cross, and the left panels show offspring from the AKR x PWD cross. Different colors represent different X-linked genes.

(A) A diagram to illustrate the allele-specific expression results when there is no sampling effect, no eQTL effect and no parent-of-origin effect. In this case, there is little variability of PWD allelic expression among individuals or among the two reciprocal crosses. The only source of variability is the pyrosequencing measurement error. This is the case for the autosomal genes and X-linked genes that escape X inactivation.

(B) A diagram to illustrate the sampling effect caused by random X inactivation. In this diagram, the X inactivation process itself is random, but the number of brain-forming cells is small during the time of X inactivation, resulting in sampling variation among individuals. Although individuals are expected to show a 1:1 expression ratio, if each cell randomly and independently inactivates one or the other X chromosome, then we expect to see a binomial distribution of counts of cells inactivating the maternal vs. paternal X. If the count of cells is small, the variance in expression ratios could be large, and a maternal bias observed in a small number of individuals might be explained by this sampling effect. The sampling effect of X inactivation also drives the observed co-variation of allelic bias in expression of all X-linked genes.

(C) A diagram to illustrate the eQTL effect. If there is a cis-regulatory polymorphism near the respective gene, it may drive differential allelic expression yielding allelic expression counts different from 1:1. The regulatory variant might drive higher expression from the PWD or the AKR allele, so the mean PWD expression percentage is not 50%. Such an effect would be allele-specific (or strain-specific), and would not explain differences in expression between reciprocal crosses or a maternal bias.

(D) A diagram of preferential paternal X inactivation. Here the X inactivation is NOT random and the paternal X is preferentially inactivated. In this case we will observe greater expression from the maternal allele. The bias is like that of a biased coin. For small numbers of tosses, not all samples will show a skewed ratio of heads to tails, but with a sufficiently large sample, the bias will appear as a shift in the mean. In this cartoon, a comparison of the two reciprocal crosses shows that the allele-specific expression profile is shifted.



A. No sampling effect, no eQTL effect, no parent-of-origin effect.

B. With sampling effect, no eQTL effect, no parent-of-origin effect.

C. With sampling effect and eQTL effect, no parent-of-origin effect.

D. With sampling effect, eQTL effect and parent-of-origin effect.

Combined effect of sampling and preferential paternal X inactivation

In our pyrosequencing experiment, the three sources of variation, namely sampling effects, eQTL effects, and parent-of-origin effects are superimposed, and all may contribute to the variability in allele-specific expression percentages. We will now show how statistical tests allow quantitative partitioning of these effects from the PWD percentages of these X-linked genes across the 36 individual female progeny:

(i). Sampling effect. We studied 26 genes that are subjected to X inactivation, shown in Figure 41. In Figure 41A, the X-linked genes vary in parallel with each other, indicating that from one mouse to another, the allele-specific expression ratio of these genes covary in a concerted fashion. If by chance in one mouse 70% of inactivated X chromosomes were paternal and 30% were maternal, this sampling effect would produce a consistent pattern of excess maternal expression in all the X-linked genes examined (or at least those that undergo normal X-inactivation). Among different individual mice, we expect to see such sampling variation due to the small number of brain-forming stem cells at the time of X inactivation early in development.

Figure 41. Allele-specific expression ratio of 37 genes in P2 brains of 18 female F1 progeny from each of the two reciprocal crosses between AKR and PWD strains.

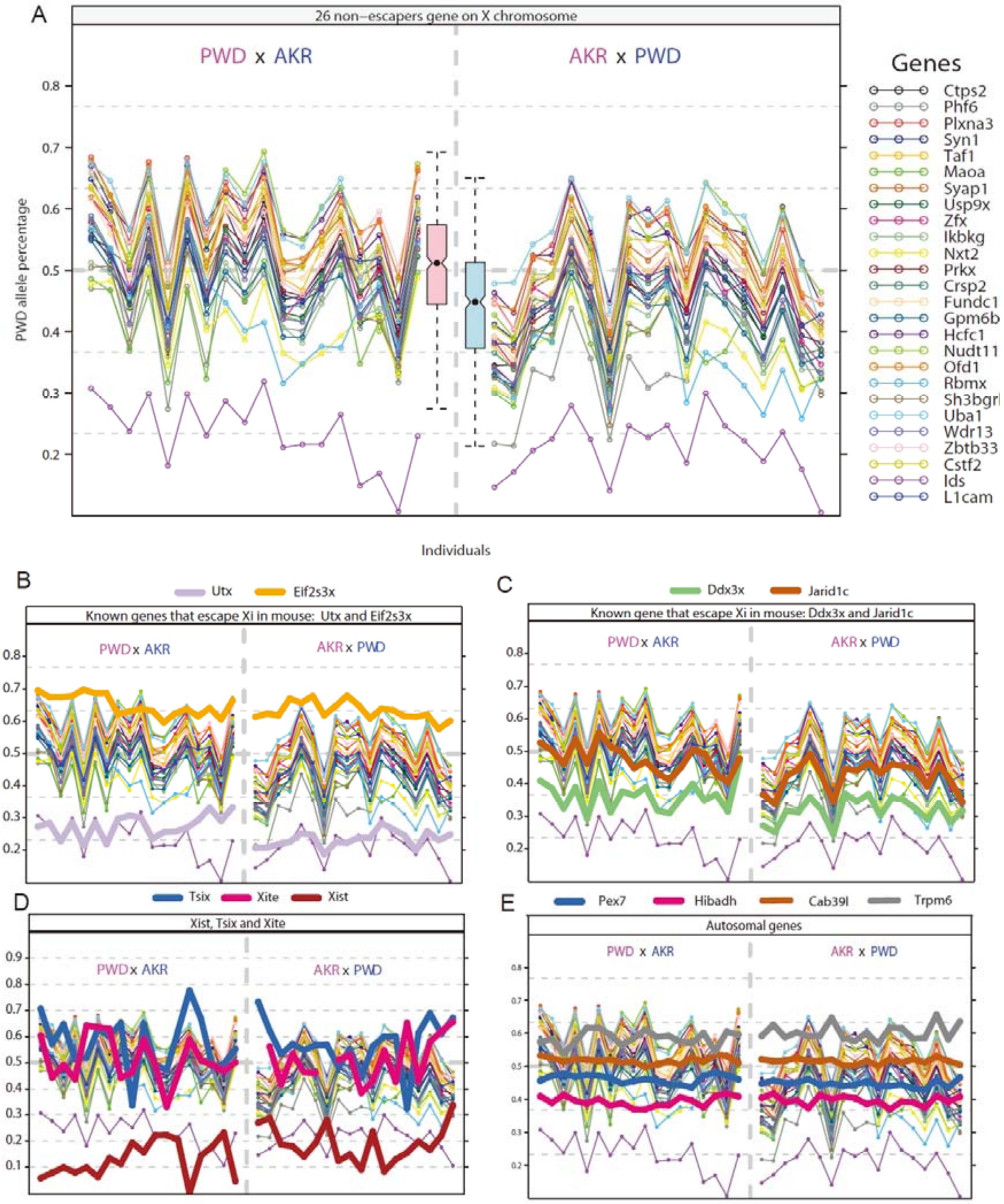
(A). Allele-specific expression profiling of 26 genes that are subject to X inactivation. The pink boxplot in the middle is the distribution of PWD expression percentage from the PWD x AKR cross for all X-linked genes that are subject to X inactivation. It is labeled pink because PWD is the maternal allele in this cross. The blue boxplot is the distribution of PWD expression percentage from the AKR x PWD cross. It is labeled blue because PWD is the paternal allele in this cross.

(B). Allele-specific expression profiling of known genes that escape X inactivation in mouse: *Utx* and *Eif2s3x*.

(C). Allele-specific expression profiling of known genes that escape X inactivation in mouse: *Ddx3x* and *Jarid1c*.

(D). Allele-specific expression profiling of four autosomal genes: *Cab39l* , Pex7, *Hibadh* and *Trpm6*.

(E). Allele-specific expression profiling of *Xist*, *Tsix* and *Xite* transcripts.


A
26 non−escapers gene on X chromosome
PWD x AKR
AKR x PWD
PWD allele percentage
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Individuals
Genes
Ctps2
Phf6
Plxna3
Syn1
Taf1
Maoa
Syap1
Usp9x
Zfx
Ikbkg
Nxt2
Prkx
Crsp2
Fundc1
Gpm6b
Hcfc1
Nudt11
Ofd1
Rbmx
Sh3bgrl
Uba1
Wdr13
Zbtb33
Cstf2
Ids
L1cam
B
Utx
Eif2s3x
Known genes that escape Xi in mouse: Utx and Eif2s3x
PWD x AKR
AKR x PWD
C
Ddx3x
Jarid1c
Known gene that escape Xi in mouse: Ddx3x and Jarid1c
PWD x AKR
AKR x PWD
D
Tsix
Xite
Xist
Xist, Tsix and Xite
PWD x AKR
AKR x PWD
0.9
0.1
E
Pex7
Hibadh
Cab39l
Trpm6
Autosomal genes
PWD x AKR
AKR x PWD

(ii). *cis*-regulatory effect. Within each individual, not all the genes have the same level of allele-specific expression from the PWD allele. This is because the two alleles differ in *cis*-regulatory activity, and the *cis*-regulatory differences are specific to each gene. If there is a strain-specific *cis*-regulatory SNP near the gene, it will produce an elevated relative expression from the allele coming from one strain, in the offspring of both reciprocal crosses

(iii). Preferential paternal X inactivation. In addition to the sampling and eQTL effects, we also observed a parent-of-origin effect of random X inactivation. The average PWD expression percentage for 26 genes that are subject to X inactivation in the PWD x AKR cross is 50.4%, whereas the average in the AKR x PWD cross is 44.0% (Figure 41A). This difference, while quantitatively modest, is highly statistically significant.

Statistical analysis of the three factors affecting X expression ratios

In order to quantify the three effects discussed above and to assess their statistical significance, a nested analysis of variance (ANOVA) model was implemented. We assume that each individual represents an independent sampling trial at the time of X inactivation. There are two fixed factors, "*cis*-regulatory" and "parent-of-origin", as well a random factor "sampling" nested within "parent-of-origin". The "*cis*-regulatory" factor refers to the consistent allelic bias as one might see if there were *cis*-acting (eQTL) factors that result in, for example, an over- or under-expression of the AKR allele relative to the PWD allele. Our data cover 27 genes that are subject to

X inactivation (26 genes in Figure 41A and *Ddx3x*), and because each gene may have a different magnitude of such *cis*-acting expression effects, the *cis*-regulatory factor has 27 levels. The "parent-of-origin" factor represents the differences seen in allelic bias between reciprocal crosses (PWD x AKR and AKR x PWD). The "sampling" factor is nested in the "parent-of-origin" factor, with 18 independent trials from each of the two reciprocal crosses. From the nested ANOVA results (Table 29), there is a significant "*cis*-regulatory" effect ($P < 0.001$), indicating that there is highly significant heterogeneity in allelic expression across these X-linked genes (Table 29 and Table 30; Figure 41A andFigure 42A). Some genes have higher average expression from the PWD allele, and some genes have higher average expression from the AKR allele (Figure 43). The "parent-of-origin" effect is also highly significant ($P = 0.0045$), suggesting preferential paternal X inactivation (Table 29; Figure 43). We saw the same trend of preferential inactivation of the paternal allele in the B6 and CAST strain combination (Figure 42A). The "sampling" effect nested in the parent-of-origin factor is significant as well ($P < 0.0001$), showing a substantial amount of variation of the sampling effect (Table 29; Figure 41A andFigure 42A). We also applied a non-parametric test by rank transformation (Conover and Iman 1981), all three effects remain highly significant, with $P < 0.0001$, $P = 0.0051$ and $P < 0.0001$ for the *cis*-regulatory, parent-of-origin and sampling effect respectively (Table 31). The effect size was estimated by variance component analysis. The sampling effect explains 30.9% of the total variance. The parent-of-origin effect explains 14.3% of the total variance, and the *cis*-regulatory effect explains 48.3% of the total variance (Table 32). We applied the method of Least-Squares Means to obtain an LS-mean for PWD mothers (in the PWD x AKR cross) of 0.4985 (SE = 0.01464, Table 30), and the LS-mean for AKR mothers (AKR x PWD cross) 0.4355 (SE = 0.01464). In B6-CAST reciprocal crosses, the estimate for CAST mothers (CAST x B6 cross) is 0.6706 (SE = 0.02403), and the estimate for B6 mothers (B6 x CAST cross) is 0.6160 (SE = 0. 02403). Since we found a similar degree of maternal bias (about 6%) in B6-CAST progeny as in PWD-AKR progeny, we analyzed the two datasets together. The *P*-value of the "parent-of-origin" effect for the pooled data is even smaller ($P < 0.0020$) (Table 33). We conclude that the maternal bias or the degree of preferential paternal X inactivation is about 6%.

Table 29. Analysis of variance table of the PWD-AKR data of X-linked genes subject to X inactivation. Type III sums of squares are reported.

Source	Sum of Squares	Mean Square	DF	F Value	Pr > F
Gene	10.96479	0.421723	26	524.83	<.0001
Mother	1.822754	1.822754	1	9.25	0.005
individual(mother)	6.700906	0.197085	34	245.27	<.0001
Residual	1.428698	0.000804	1778		

Table 30. Least-squares means (LS-means) of fixed effects gene and mother.

Effect	mother	gene	Estimate	Error	DF	t Value	Pr > |t|
mother	AKR		0.4355	0.01464	34	29.74	<.0001
mother	PWD		0.4985	0.01464	34	34.04	<.0001
gene		Crsp2	0.4749	0.01086	1778	43.73	<.0001
gene		Cstf2	0.5233	0.01086	1778	48.19	<.0001
gene		Ctps2	0.4655	0.01092	1778	42.65	<.0001
gene		Ddx3x	0.3398	0.01087	1778	31.27	<.0001
gene		Fundc1	0.5021	0.01086	1778	46.24	<.0001
gene		Gpm6b	0.4353	0.01087	1778	40.03	<.0001
gene		Hcfc1	0.5614	0.0109	1778	51.52	<.0001
gene		Ids	0.2228	0.01086	1778	20.52	<.0001
gene		Ikbkg	0.4247	0.01086	1778	39.12	<.0001
gene		L1cam	0.4735	0.01087	1778	43.58	<.0001
gene		Maoa	0.4027	0.01087	1778	37.06	<.0001
gene		Nudt11	0.5587	0.0109	1778	51.27	<.0001
gene		Nxt2	0.3968	0.01086	1778	36.54	<.0001
gene		Ofd1	0.5578	0.01086	1778	51.37	<.0001
gene		Phf6	0.3902	0.01087	1778	35.91	<.0001
gene		Plxna3	0.5537	0.01086	1778	50.99	<.0001
gene		Prkx	0.4756	0.01086	1778	43.8	<.0001
gene		Rbmx	0.4246	0.01101	1778	38.57	<.0001
gene		Sh3bgrl	0.4225	0.0117	1778	36.11	<.0001
gene		Syap1	0.517	0.01086	1778	47.61	<.0001
gene		Syn1	0.508	0.01087	1778	46.75	<.0001
gene		Taf1	0.5007	0.01086	1778	46.11	<.0001
gene		Uba1	0.5704	0.01086	1778	52.53	<.0001
gene		Usp9x	0.4475	0.01086	1778	41.21	<.0001
gene		Wdr13	0.4512	0.01086	1778	41.55	<.0001
gene		Zbtb33	0.5293	0.01086	1778	48.75	<.0001
gene		Zfx	0.4781	0.01104	1778	43.29	<.0001

Table 31. Nonparametric analysis of variance table of the PWD-AKR data of X-linked genes subject to X inactivation.

Source	Error Term	DF	F Value	Pr > F
gene	MS(Residual)	1778	300.87	<.0001
mother	0.9996 MS(individual(mother)) + 0.0004 MS(Residual)	34	8.96	0.0051
individual(mother)	MS(Residual)	1778	178.04	<.0001
Residual				

Table 32. Variance component analysis.

	REML		Type 1	
	Variance component estimate	% variance explained	Variance component estimate	% variance explained
gene	0.0059768	48.30%	0.0062052	49.24%
mother	0.00177	14.30%	0.001751	13.89%
individual(mother)	0.0038256	30.91%	0.0038431	30.49%
Error	0.000803	6.49%	0.0008035	6.38%

	ML		MIVQUE(0)	
	Variance component estimate	% variance explained	Variance component estimate	% variance explained
gene	0.0059138	51.51%	0.0063354	50.24%
mother	0.0009369	8.16%	0.0017555	13.92%
individual(mother)	0.0038256	33.32%	0.00387	30.69%
Error	0.0008035	7.00%	0.0006497	5.15%

Table 33. Analysis of variance table of the pooled data (PWD-AKR and B6-CAST crosses) of X-linked genes subject to X inactivation. Type III sums of squares are reported.

Source	Sum of Squares	Mean Square	DF	F Value	Pr > F
Gene	13.204705	0.356884	37	466.51	<.0001
Mother	2.153376	1.076688	2	6.98	0.0020
individual(mother)	9.265601	0.171585	54	224.29	<.0001
Residual	1.675383	0.000765	2190		

Figure 42. Allele-specific expression ratio of 20 genes in P2 brains of 11 female mice from each of the two reciprocal crosses between B6 and CAST strains.

(A). Allele-specific expression profiling of 11 genes that are subject to X inactivation.

(B). Allele-specific expression profiling of known mouse genes that escape X inactivation: *Utx* and *Eif2s3x*.

(C). Allele-specific expression profiling of known mouse genes that escape X inactivation: *Ddx3x* and *Jarid1c*.

(D). Allele-specific expression profiling of *Xist*, *Tsix* and *Xite* transcripts.

(E). Allele-specific expression profiling of two autosomal genes: NM_023057 and *Pex7*.

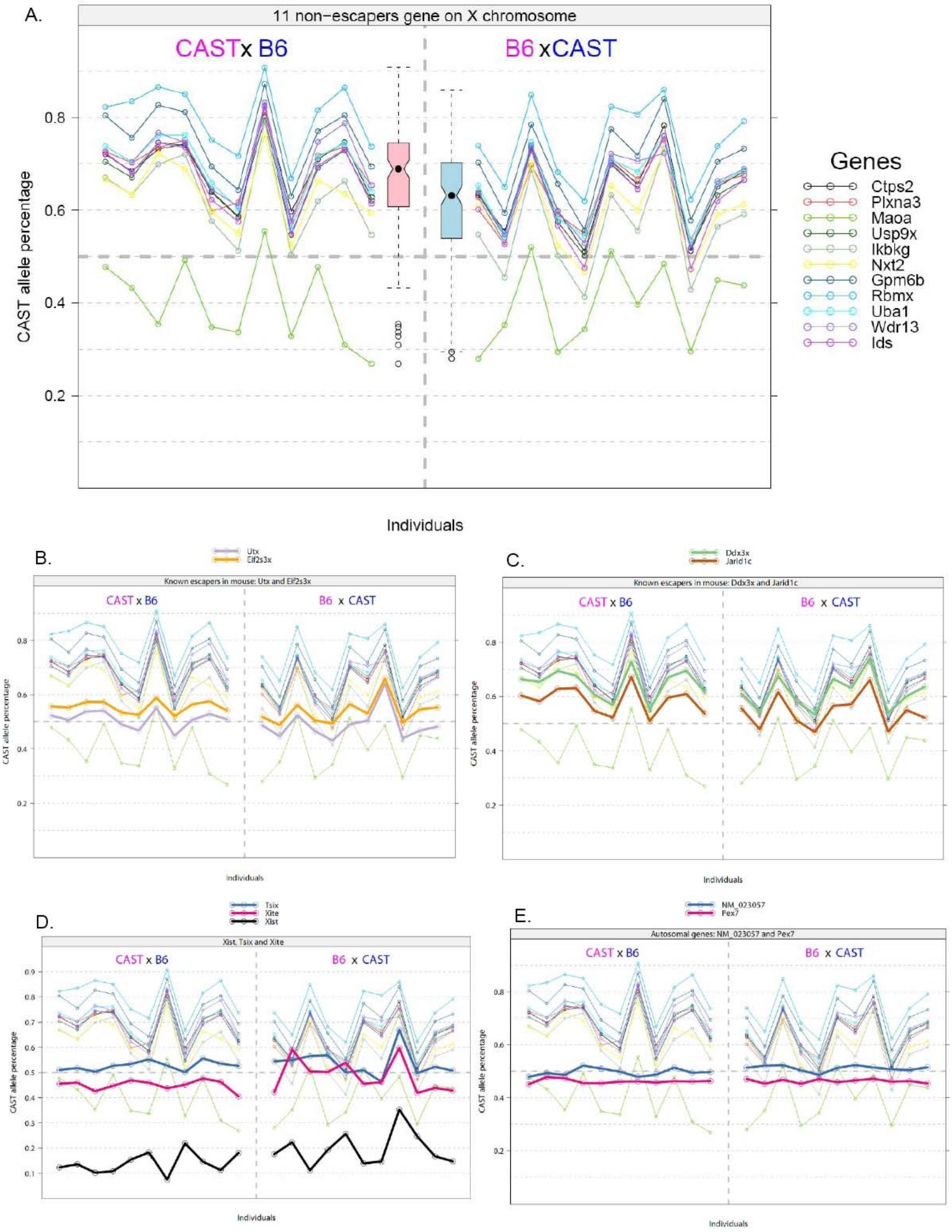
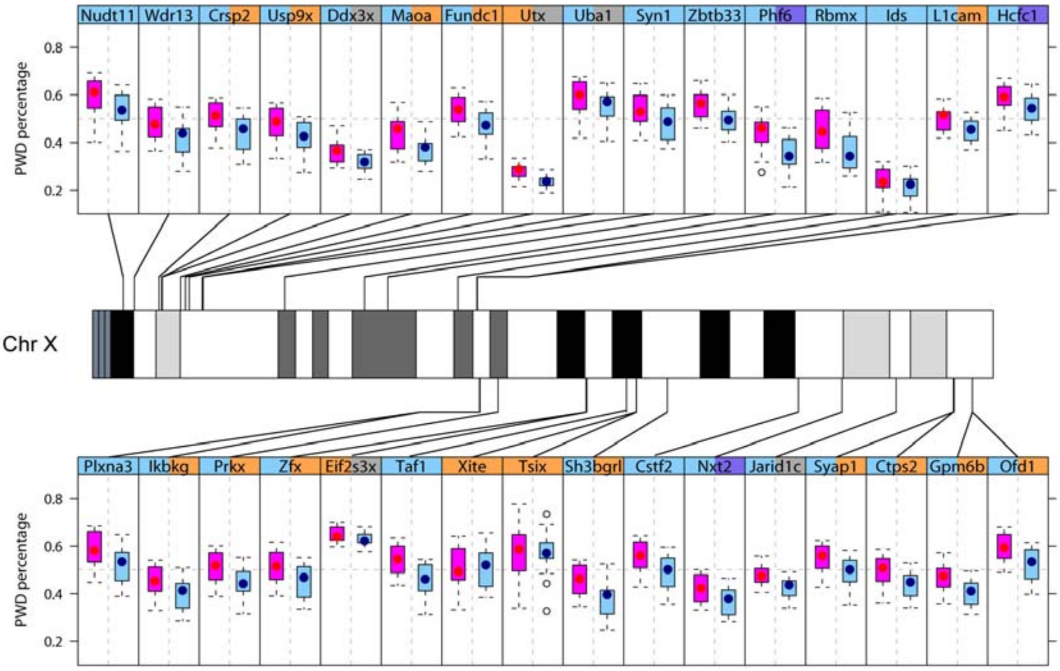

A.
11 non−escapers gene on X chromosome
CASTx B6
B6 xCAST
CAST allele percentage
0.2
0.4
0.6
0.8
Individuals
Genes
Ctps2
Plxna3
Maoa
Usp9x
Ikbkg
Nxt2
Gpm6b
Rbmx
Uba1
Wdr13
Ids
B.
Utx
Eif2s3x
Known escapers in mouse: Utx and Eif2s3x
CAST x B6
B6 x CAST
CAST allele percentage
Individuals
C.
Ddx3x
Jarid1c
Known escapers in mouse: Ddx3x and Jarid1c
CAST x B6
B6 x CAST
CAST allele percentage
Individuals
D.
Tsix
Xite
Xist
Xist, Tsix and Xite
CAST x B6
B6 x CAST
CAST allele percentage
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Individuals
E.
NM_023057
Pex7
Autosomal genes: NM_023057 and Pex7
CAST x B6
B6 x CAST
CAST allele percentage
Individuals

Figure 43. Distribution of the PWD allele expression percentage in F1 progeny of AKR and PWD reciprocal crosses. The mouse X chromosome map is diagrammed in the middle of the figure. Each panel is a boxplot of an X-linked gene with its chromosomal position labeled. The red box is the distribution of the PWD allele expression percentage in P2 brains of 18 F1mice from the PWD x AKR cross (mother listed first). The blue box is the distribution of the PWD allele expression percentage in P2 brains of 18 F1mice from the AKR x PWD cross. The gene name is listed at the top of the figure. The color of the left and right strip label depicts the known X-inactivation status in mouse and human, respectively (Orange: genes that escape X inactivation; Purple: genes that partially escape X inactivation; Blue: genes subject to X inactivation; Black: NA). Note that every gene that undergoes X inactivation shows a consistent bias toward excess inactivation of the paternal X (a sign test shows the bias to by highly significant, P < 1.5 x 10^{-8}).

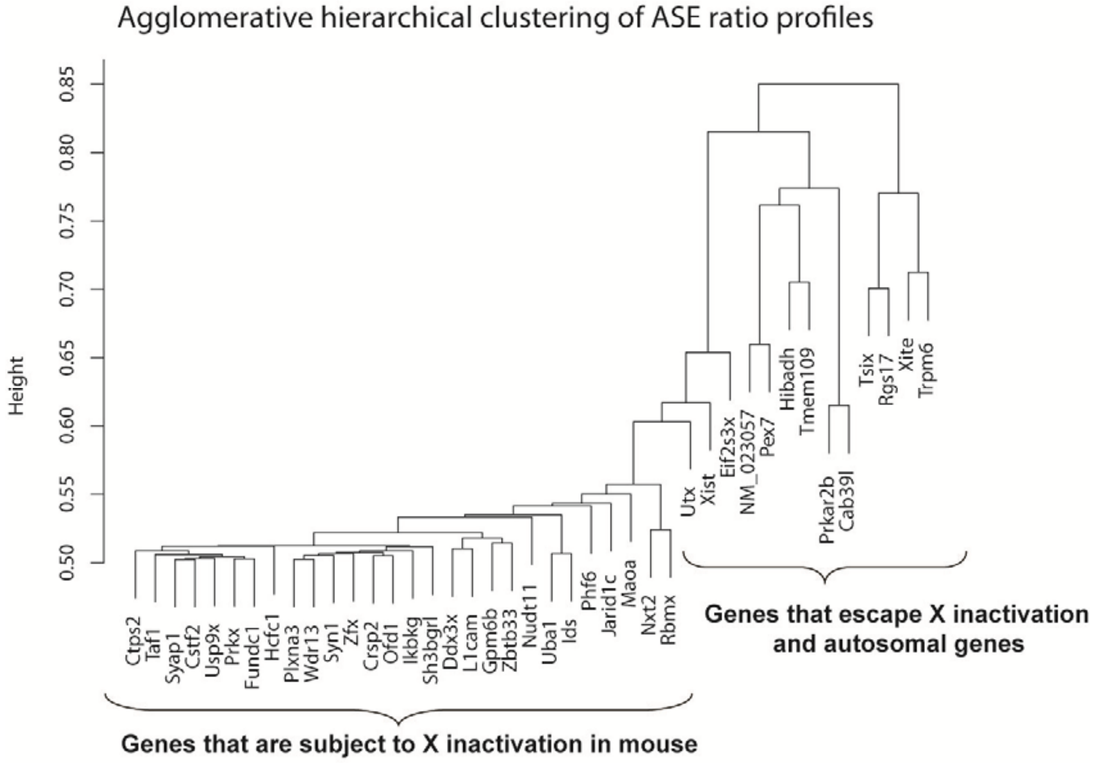


Identification of genes that escape X inactivation in normal mouse brains

One way to distinguish the genes that escape X inactivation from those that do not is to perform a cluster analysis based on the correlation in allelic bias across genes. We found a large and closely related cluster containing most of the X-linked genes (Figure 44), leaving the two known escapers (*Eif2s3x* and *Utx*) and the eight autosomal control genes (*NM_023057, Pex7*, *Prkar2b, Hibadh, Rgs17, Cab39l, Trpm6 and Tmem109*) outside the cluster. The genes within the cluster are the genes that are subject to X inactivation, because they are expected to vary in relative allelic expression in parallel with each other, as a consequence of the sampling variation in the brain-progenitor cells at the time of X inactivation during early development. The genes that escape X inactivation do not have this property of correlated allelic bias, and as expected they are clearly separated from the cluster. Similarly the autosomal control genes fall outside the cluster of genes that are X inactivated.

Unlike the X-linked genes that are subject to X inactivation, eight randomly chosen autosomal genes, *NM_023057* (on chromosome 2), *Pex7* (on chromosome 10), *Prkar2b* (on chromosome 12)*, Hibadh* (on chromosome 6)*, Rgs17* (on chromosome 10)*, Cab39l* (on chromosome 14)*, Trpm6* (on chromosome 19) and *Tmem109* (on chromosome 19), have much less among-individual variation in PWD expression percentage and did not show high correlation with the genes that are subject to X inactivation. This is exactly as expected: because the autosomal genes are biallelically expressed in the same way in all cells of all individuals, they should exhibit far less among-individual variation. To illustrate the profile for autosomal genes with an

Figure 44. Cluster analysis of the allele-specific expression ratios of X-linked genes in F1 progeny from AKR and PWD reciprocal crosses. Based only on the differential allelic expression, genes are clustered using a standard nested agglomerative hierarchical clustering (see text for details). The large cluster of genes to the left are all subject to normal X inactivation, while the genes that escape X inactivation fall on the deeper branches to the right.



eQTL effect, four of the eight autosomal genes tested are shown in Figure 41E. For all genes we observe no maternal bias (the mean is not significantly different between the PWD x AKR and AKR x PWD crosses). For *Cab39l* and *Pex7*, there is very little eQTL effect, so the PWD:AKR expression ratio is nearly 50%:50%. For *Trpm6*, there is a PWD dominant eQTL effect, and the PWD:AKR expression ratio is about 60%:40%. For *Hibadh*, there is an AWD dominant eQTL effect and the PWD:AKR expression ratio is about 40%:60%. Unlike the genes that are subject to X inactivation, the PWD:AKR expression ratios of the autosomal genes do not flip in the reciprocal crosses (Figure 41E). *NM_023057* and *Pex7* were also tested in the B6-CAST reciprocal crosses (Figure 42E).

For genes that escape X inactivation, since there is no sampling effect, we expect less among-individual variation in PWD expression ratios, just like the autosomal genes. Among the four known genes that escape X inactivation in mouse, allelic expression of *Eif2s3x* and *Utx* was much less variable among individual mice, and was not well correlated with the genes that do undergo X inactivation (Figure 41B and Figure 42B). This is consistent with their escaper status (Figure 43 and Figure 44). The other two previously reported genes in mouse, *Ddx3x* and *Jarid1c* (also known as *Smcx*), clustered with the genes that are subject to X inactivation. *Jarid1c* expression showed a weak correlation (Figure 41C and Figure 42C). This is consistent with the fact that *Jarid1c* only partially escapes X inactivation with approximately 30% expression from the inactivated X chromosome (Carrel et al. 1996; Li and Carrel 2008). The *Ddx3x* gene showed a perfect correlation with all the other X-inactivated genes, implying that

Ddx3x in fact displays normal X inactivation in neonatal mouse brain. The discrepancy could be due to tissue-specificity of X inactivation, or spurious expression effects resulting from the aberrant genomic configuration of the translocation mouse line used in other studies.

We also tested three genes in the *Xic* (X inactivation center), namely *Xist*, *Tsix* and *Xite*. We observed that *Tsix* and *Xite* are correlated with one another (Figure 41D and Figure 42D), which is consistent with the notion that *Xite* is regulating *Tsix* in *cis*. Note that the correlation is not perfect, because the low expression level of *Tsix* resulted in a weak pyrosequencing signal, and the expression level of *Xite* is even lower. However, we did detect expression of these two genes in the RNA-seq and pyrosequencing data based on the GenBank gene models. For *Xist*, we observed a large eQTL effect, with about 90% expression from the AKR allele in both AKR x PWD reciprocal crosses (Figure 41D and Figure 44), and about 80% expression from the B6 allele in both B6 x CAST reciprocal crosses (Figure 42D). The reason for this is the strength of the *Xce* (X controlling element) locus is different among mouse strains. *Xce* is mapped to a region near the *Xic* which contains the *Xite* gene, the promoter of *Tsix,* as well as the pairing region of the two X chromosomes (Simmler et al. 1993; Courtier et al. 1995; Chadwick et al. 2006; Valley and Willard 2006). Allelic differences in *Xce* in expression bias cluster into three groups with strength order $Xce^{a} < Xce^{b} < Xce^{c}$ (Plenge et al. 2000). In inter-strain F1 mice, the X chromosome with a stronger allele will have higher probability to be the active X chromosome (Plenge et al. 2000). Our observation of the allele-specific expression pattern of *Xist* in B6 and CAST crosses is consistent with the fact that the B6 *Xce* allele belongs to the *Xce*^{b} group and the CAST allele is an *Xce*^{c} allele (Plenge et al. 2000). So we expect a strong eQTL effect with higher expression of the B6 allele of *Xist*. From the AKR and PWD crosses, it is known that the strength of the AKR *Xce* allele is somewhere between *Xce*^{b} and *Xce*^{c}. Given our data, we conclude that the PWD *Xce* allele is stronger than that of AKR. The 90% allele-specific expression ratio seems to be unexpectedly high, but note that the bias in the final X inactivation ratio need not match the allele-specific expression of *Xist*. The *Xist* transcript is only expressed from the inactive X chromosome but the two *Xist* alleles may be expressed quantitatively at different levels, and the expression levels measured here are from heterogeneous pools of cells. It could be that the AKR allele expression level is higher in cells with inactive X from AKR strain than the PWD allele expression level in cells with inactive X from PWD strain, but the PWD expression level is sufficient to maintain the X inactivation status. Parent-of-origin influences of *Xce* on X chromosome biased allelic inactivation had been reported in heterozygous F2 mice (not significant in F1) in B6-CAST crosses (Chadwick and Willard 2005). Since the *Xce* is a strain-specific DNA sequence feature rather than an epigenetic mark, it is expected to be manifested as an eQTL effect. The parent-origin-effect of skewed random X inactivation that we observed cannot be explained as a canonical *Xce* effect.

We found the mouse orthologs of human genes that escape X inactivation (*Ctps2, Maoa, Syap1*, *Usp9x*, *Zfx*, *Ikbkg, Prkx, Crsp2, Fundc1, Gpm6b, Ofd1, Sh3bgrl, L1cam*) and the ones that partially escape X inactivation (*Phf6, Nxt2, Hcfc1*) (Carrel and

Willard 2005), are subject to X inactivation in mouse. The mouse orthologs of human genes subject to X inactivation (*Taf1*, *Syn1*, *Plxna3*, *Nudt11*, *Zbtb33*, *Wdr13*, *Rbmx*, *Uba1*, *Cstf2*, *Ids*) are also subject to X inactivation in mouse (Figure 41A, Figure 43 and Figure 44). This is consistent with the previous findings that human has more genes that escape X inactivation than mouse. We also confirmed 11 of the above genes in the B6 x CAST strain combination (Figure 42A). *Prkx*, a mouse X-inactivation escaper candidate gene whose X inactivation status is not determined (Disteche et al. 2002), is found to be a non-escaper in our data.

Sampling effect of X inactivation during early development in the mouse brain

We observed significant variation in allelic expression for the X-linked genes among 36 normal F1 individuals in the reciprocal crosses of AKR and PWD, as well as 22 F1 individuals in B6 and CAST reciprocal crosses. Because we do not see the same amount of variation for the autosomal control genes, we conclude that the variation in expression is due to a cellular sampling effect at the time of X inactivation (see also (Amos-Landgraf et al. 2006)). We found that the among-individual sampling effect (explaining 30.9% of the total allele-specific variance in the AKR-PWD cross) is larger than the parent-of-origin effect (explaining 14.3% of the total allele-specific expression variance).

The X inactivation process starts at an early stage (approximately at E6.5) when there are only a few brain-forming cells, and once X inactivation occurs in a cell, the X inactivation status is retained by the daughter cells. Here, we refer to the average number because the X inactivation does not initiate instantaneously but instead occurs over a short period of time. The average number of brain-forming cells at the time of X inactivation can be estimated from the among-individual sampling variance of relative gene expression levels (Amos-Landgraf et al. 2006). The larger the variation among individuals, the smaller number of cells there must have been during X inactivation. By simulating a random process of X-inactivation, and matching the observed and simulated variance, we estimated the average number of brain precursor cells during the time of X inactivation (Figure 45).

Parent-of-origin effect is chromosome-wide

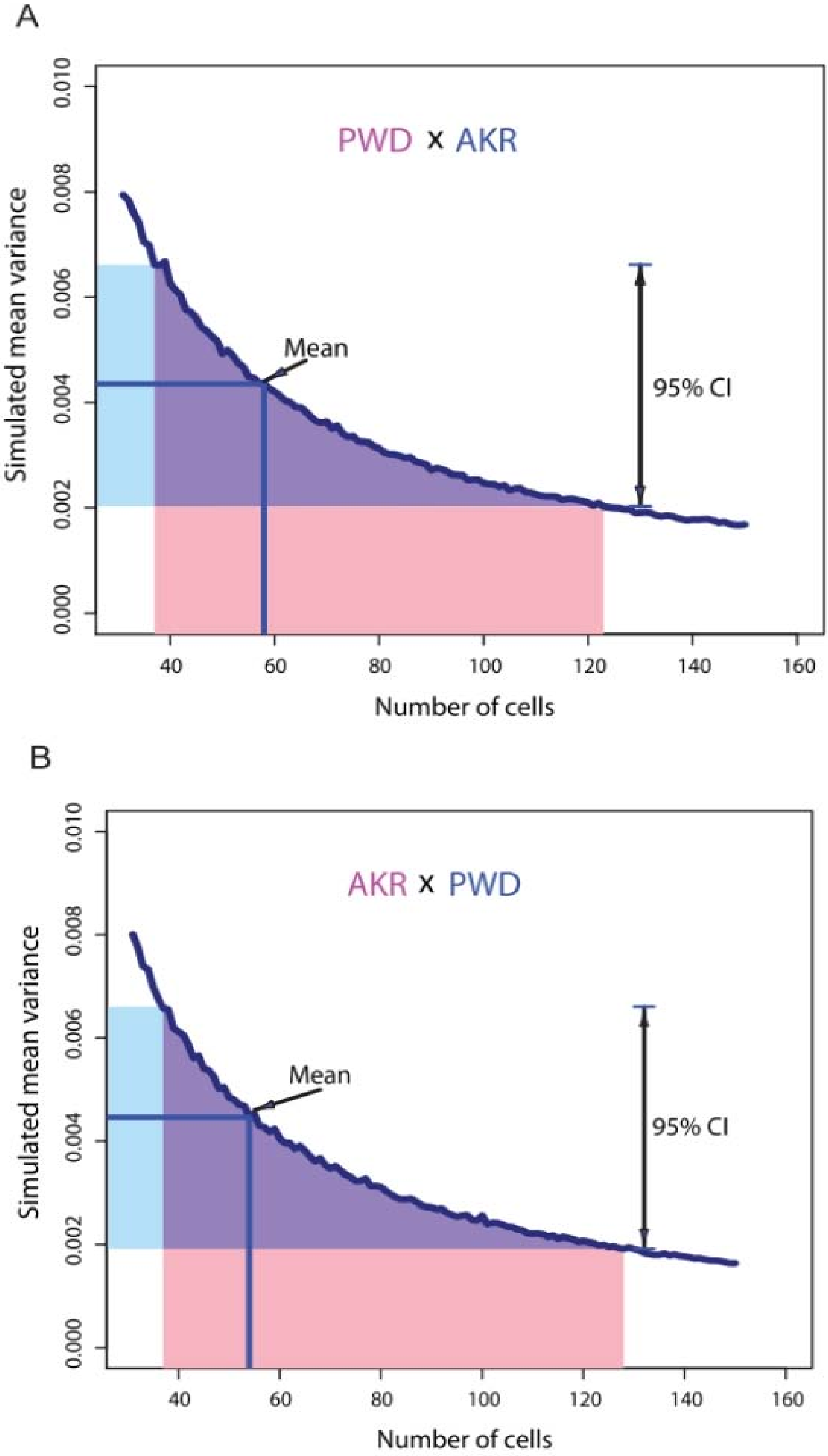
Analysis of the distribution of allele-specific expression of a set of X-linked genes allowed us to quantify the parent-of-origin effect for the X chromosome (Figure 43). We observed that the X-linked non-escaper genes in mouse showed a significant parent-of-origin effect, as well as larger sampling variation. In contrast, for the known escapers, we did not see a significant parent-of-origin effect and the sampling variance of gene expression is much smaller. The data from the 33 X-linked genes assayed are consistent with the parent-of-origin effect being chromosome-wide.

Figure 45. Estimation of the number of brain-forming cells at the time of X inactivation in mouse. Given the observed variance among individuals in relative expression levels, we calculated the maximum likelihood estimate for the number of cells present at the time of X inactivation (assuming X inactivation occurs at a single point in time and is irreversible). For the PWD x AKR cross, the average number of brain forming cells at the time of X inactivation is estimated to be 58, with 95% confidence interval from 37 to 123. For the AKR x PWD cross, the estimated number is 54, with 95% confidence interval from 37 to 128. The cell numbers estimated from the two reciprocal crosses are thus consistent with each other, and numerical simulations were also consistent with these results.

(A). Estimation of number of brain-forming cells at the time of X inactivation in F1 progeny of the PWD x AKR cross.

(B). Estimation of number of brain-forming cells at the time of X inactivation in F1 progeny of the AKR x PWD cross.

Binomial sampling was done with different sample sizes of brain-forming cells (x-axis) and for each sample the . The Y-axis is the simulated mean variance. The observed mean variance with 95% confidence interval is labeled.


A
PWD x AKR
Mean
95% CI
Simulated mean variance
0.000
0.002
0.004
0.006
0.008
0.010
40
60
80
100
120
140
160
Number of cells
B
AKR x PWD
Mean
95% CI
Simulated mean variance
0.000
0.002
0.004
0.006
0.008
0.010
40
60
80
100
120
140
160
Number of cells

Discussion

Is random X inactivation truly "random"?

Following the initial discovery that dosage compensation is accomplished in mammals by X inactivation (Lyon 1961), the process has been considered to occur through a random process in the embryonic tissues of eutherian mammals. This implies that each cell has an equal probability to inactivate either the paternal or the maternal copy of the X chromosome during random X inactivation (assuming equal influence of the two parental *Xce* alleles). Our data provide clear evidence that X inactivation can depart from a strictly random pattern, and in the mouse brain we find a small but significant and consistent preferential bias to inactivate the paternal X. The result is robust across multiple individual mice from two sets of reciprocal crosses. The average ratio of inactivated paternal and maternal X chromosomes is not 50:50. Instead, there is about 6% preferential paternal bias in X inactivation, a bias small enough that it is easy to see why it has been overlooked. At present it is not clear whether the bias is driven by incomplete erasure of the paternal X imprint (Lee et al. 1996; Jaenisch et al. 1998; Lee and Lu 1999), or whether the signal is totally erased and there follows a bias in the X-inactivation process itself. Formally there is also the possibility that the bias that we observe toward excess maternal expression could be due to preferential growth/proliferation of cells with the maternal active X, but the absence of any known mechanism that might drive this bias reduces its plausibility. The ultimate experimental answer will come from examination of differential X chromosome expression in appropriate tissues at the single cell level.

Further understanding the process of X inactivation

Two hypotheses may explain the preferential paternal X inactivation. First, the short time interval during the transition from imprinted X inactivation to random X inactivation in embryonic tissues may leave a residual imprint. During imprinted X inactivation, it is known that there might be a residual imprint on the maternal X chromosome that keeps it active, probably by repressing the *Xist* transcription in *cis* (Heard and Disteche 2006). If this is the case, then during reactivation of the paternal X chromosome, the short time interval may be insufficient to completely reset the *Xist/Tsix* status by erasure of its epigenetic marks. The other possibility is that erasure of *Xist* from the X chromosome could be complete after imprinted X inactivation, but that during the random X inactivation, by some unknown mechanism, the maternal X chromosome has a slightly higher chance to remain active. Additional experiments are needed to elucidate the mechanism of preferential paternal X inactivation in mouse.

Evolutionary considerations

Both marsupials and eutherian mammals achieve dosage compensation through X inactivation. For marsupials, the imprinted X inactivation status is retained in both the extraembryonic and embryonic tissues during development and later throughout adulthood (Cooper et al. 1990). Because the maternal expression of the X-linked genes is not 100%, the imprinted X inactivation is called incomplete or leaky X inactivation. Here, we found that the random X inactivation in eutherian mammals is not 50:50, but instead there is preferential paternal inactivation, suggesting the possibility that the imprinted X inactivation represents a remnant of the ancestral state. Classical evolutionary theory suggests that after the differentiation of the X and Y sex chromosomes, the Y chromosome degenerates, necessitating a means for adjusting dosage to resolve the X chromosome dosage imbalance (Vicoso and Charlesworth 2006; Straub and Becker 2007) . One possible mechanism for X inactivation is to always inactivate one of the parental X chromosomes. The inactivated X cannot be the maternal X because the only X chromosome that males possess is maternal. Paternal X-inactivation, as is found in marsupials, may represent the ancestral form of mammalian dosage compensation (Namekawa et al. 2007), although it is formally possible that the common ancestor of marsupials and eutherian mammals lacked dosage compensation, and that both lineages developed their own dosage compensation mechanisms independently.

Compared to random X inactivation, imprinted X inactivation runs a greater risk of error. If a recessive deleterious or lethal allele is transmitted from the mother, the fitness of the offspring will be severely reduced. For random (or nearly random) X inactivation, there are still half the cells expressing the normal allele. By expressing one of the two parental alleles in different cells, both dosage compensation and the problem of X hemizygosity are solved. As mentioned before, in marsupials the imprinted X inactivation is not complete, and we discovered that there is also preferential paternal X inactivation in mouse brain, but with much smaller degree of maternal bias than in marsupials. If the common ancestor of eutherian mammals and marsupials had some form of imprinted X inactivation, then the most parsimonious explanation would be that during evolution, there has been a trend from complete imprinted X inactivation in the ancestor of all mammals, to leaky imprinted X inactivation in marsupials, whereas the lineage leading to eutherian mammals developed random X inactivation with slight maternal bias.

Caveats for identifying X-linked imprinted genes outside extra-embryonic tissue

It is known that many imprinted genes are derived from retro-transposition events with the origin from the X chromosome, such as *Nap1l5, U2af1-rs1,* and *Inpp5f_v2*. Currently, there are four documented X-linked imprinted genes. *Xist* and *Tsix,* are imprinted in mouse, and they are imprinted in the extra-embryonic tissues (Kay et al. 1994; Sado et al. 2001). *Rhox5*, is imprinted at a preimplantation stage before the completion of X inactivation (Kobayashi et al. 2006). A candidate imprinted gene, *Xlr3b* was found by comparing the expression of 39, X$^{\text{maternal}}$O and 39, X$^{\text{paternal}}$O mice (Davies et al. 2005). The genes *Xlr3b, Xlr4b* and *Xlr4c* are were examined in normal female neonatal brain from reciprocal cross F1 progeny, and their imprinting status was variable. *Xlr3b* is clearly not imprinted in our data (not shown). In our previous RNA-seq study, we found four X-linked genes (*Syn1*, *Plxna3*, *Phf6* and *Ctps2*, Figure 38) with a significant parent-of-origin effect on expression. However, a subsequent study described in this paper showed that they are not imprinted, but the skewed expression ratio instead arose by a sampling effect of X inactivation. Further attempts to discover X-linked imprinted genes should use a larger sample size to distinguish and verify X-linked imprinted genes from the confounding of the preferential paternal X inactivation and the sampling effect.

Cataloging X inactivation escapers in mouse and human

To further understand the X inactivation process and the evolution of the X chromosome, it is essential to tabulate an exhaustive catalog of genes that escape X inactivation in both human and mouse. Unfortunately, there is no direct method to do this in a normal single cell. For an RNA gene that works in *cis*, such as *Xist*, it is possible to count the foci in single cells using a fluorescent staining approach (Lee 2000). However, for most of the X transcripts, the single cell method is too laborious to be applied at a genome-wide scale. Human-murine (Carrel and Willard 2005) fusion cell lines and human primary fibroblasts have been used with great success to discover human genes that escape X inactivation. In mice, the genes that escape X inactivation were found using T(X;16)16H (T16H) translocations. Currently, there is no published chromosome-wide survey of the X inactivation status of all X-linked genes in mice, although methods like ours and that of Yang *et al.* (Yang et al. 2010) could easily be extended to cover the entire X. Based on the known X inactivation escapers in mouse and human, 15% of X-linked genes in human escape X inactivation, whereas previous efforts found only several escapers in mouse (Brown and Greally 2003), and Yang *et al.* (Yang et al. 2010) estimate that 3.3% of X linked genes escape X inactivation in mouse cultured cells. In this paper, we found many orthologs of known human escapers to be non-escapers in mouse (all the non-escaper genes tested by both our method and Yang *et al.*'s were concordant with respect to escaper status), suggesting that mouse does have fewer escapers that does human. Although the method presented here is an indirect one, it opens the door to examine the X inactivation status for any polymorphic X-linked gene in normal mice in any tissue.

Conclusions

Analysis of allele-specific transcript abundance in tissues of F1 progeny from reciprocal crosses of mouse strains provides a remarkably informative way to dissect the sources of variation among individuals. A large part of the inter-individual variation in relative expression of the two X chromosomes is due to a sampling effect determined by the number of cells in the tissue at the time of X inactivation – fewer cells results in larger sampling variance. The promoters from the parental mouse strains may differ in their efficiency, producing a bias in expression that follows the allelic state in both reciprocal crosses. Such eQTL effects are widespread. The *Xce* effect also may lend a chromosome-wide bias to the choice of inactivated X. Escapers of X inactivation are readily identified by this method, and we confirm the relative paucity of X inactivation escapers in mouse compared to human. On top of all of these factors, this study establishes the existence of a significant parent-of-origin effect, showing that the paternal X chromosome has a roughly 6% greater tendency toward being inactivated in the mouse brain. This observation is consistent with an evolutionary model that posits paternal X inactivation as an ancestral state.

REFERENCE

Adler DA, Bressler SL, Chapman VM, Page DC, Disteche CM. 1991. Inactivation of the Zfx gene on the mouse X chromosome. *Proceedings of the National Academy of Sciences of the United States of America* **88**(11): 4592-4595.

Agulnik AI, Mitchell MJ, Mattei MG, Borsani G, Avner PA, Lerner JL, Bishop CE. 1994. A novel X gene with a widely transcribed Y-linked homologue escapes X-inactivation in mouse and human. *Human molecular genetics* **3**(6): 879-884.

Albrecht U, Sutcliffe JS, Cattanach BM, Beechey CV, Armstrong D, Eichele G, Beaudet AL. 1997. Imprinted expression of the murine Angelman syndrome gene, Ube3a, in hippocampal and Purkinje neurons. *Nature genetics* **17**(1): 75-78.

Amos-Landgraf JM, Cottle A, Plenge RM, Friez M, Schwartz CE, Longshore J, Willard HF. 2006. X chromosome-inactivation patterns of 1,005 phenotypically unaffected females. *American journal of human genetics* **79**(3): 493-499.

Arnold AP, Itoh Y, Melamed E. 2008. A Birds-Eye View of Sex Chromosome Dosage Compensation. *Annual review of genomics and human genetics*.

Auer PL, Doerge RW. 2010. Statistical design and analysis of RNA sequencing data. *Genetics* **185**(2): 405-416.

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. *Curr Biol* **18**(22): 1735-1741.

Barlow DP, Stoger R, Herrmann BG, Saito K, Schweifer N. 1991. The mouse insulin-like growth factor type-2 receptor is imprinted and closely linked to the Tme locus. *Nature* **349**(6304): 84-87.

Bartolomei MS, Zemel S, Tilghman SM. 1991. Parental imprinting of the mouse H19 gene. *Nature* **351**(6322): 153-155.

Bjornsson HT, Albert TJ, Ladd-Acosta CM, Green RD, Rongione MA, Middle CM, Irizarry RA, Broman KW, Feinberg AP. 2008. SNP-specific array-based allele-specific expression analysis. *Genome Res*.

Blagitko N, Mergenthaler S, Schulz U, Wollmann HA, Craigen W, Eggermann T, Ropers HH, Kalscheuer VM. 2000. Human GRB10 is imprinted and expressed from the paternal and maternal allele in a highly tissue- and isoform-specific fashion. *Human molecular genetics* **9**(11): 1587-1595.

Blashfield RK. 1991. Finding Groups in Data - an Introduction to Cluster-Analysis - Kaufman,L, Rousseeuw,Pj. *Journal of Classification* **8**(2): 277-279.

Boccaccio I, Glatt-Deeley H, Watrin F, Roeckel N, Lalande M, Muscatelli F. 1999. The human MAGEL2 gene and its mouse homologue are paternally expressed and mapped to the Prader-Willi region. *Human molecular genetics* **8**(13): 2497-2505.

Brideau CM, Eilertson KE, Hagarman JA, Bustamante CD, Soloway PD. 2010. Successful computational prediction of novel imprinted genes from epigenomic features. *Molecular and cellular biology* **30**(13): 3357-3370.

Brondum-Nielsen K, Pedersen ML. 2001. [Epigenetic modification of the genetic material. Genomic imprinting and its significance for disease in human beings]. *Ugeskr Laeger* **163**(23): 3218-3222.

Brown CJ, Carrel L, Willard HF. 1997. Expression of genes from the human active and inactive X chromosomes. *American journal of human genetics* **60**(6): 1333-1343.

Brown CJ, Greally JM. 2003. A stain upon the silence: genes escaping X inactivation. *Trends Genet* **19**(8): 432-438.

Brown CJ, Willard HF. 1989. Noninactivation of a selectable human X-linked gene that complements a murine temperature-sensitive cell cycle defect. *American journal of human genetics* **45**(4): 592-598.

Buettner VL, Longmate JA, Barish ME, Mann JR, Singer-Sam J. 2004. Analysis of imprinting in mice with uniparental duplication of proximal chromosomes 7 and 15 by use of a custom oligonucleotide microarray. *Mamm Genome* **15**(3): 199-209.

Bullard JH, Purdom E, Hansen KD, Dudoit S. 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC*

Bioinformatics **11**: 94.

Butler MG. 2009. Genomic imprinting disorders in humans: a mini-review. *J Assist Reprod Genet* **26**(9-10): 477-486.

Carrel L, Cottle AA, Goglin KC, Willard HF. 1999. A first-generation X-inactivation profile of the human X chromosome. *Proceedings of the National Academy of Sciences of the United States of America* **96**(25): 14440-14444.

Carrel L, Hunt PA, Willard HF. 1996. Tissue and lineage-specific variation in inactive X chromosome expression of the murine Smcx gene. *Human molecular genetics* **5**(9): 1361-1366.

Carrel L, Willard HF. 2005. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* **434**(7031): 400-404.

Cattanach BM, Barr JA, Evans EP, Burtenshaw M, Beechey CV, Leff SE, Brannan CI, Copeland NG, Jenkins NA, Jones J. 1992. A candidate mouse model for Prader-Willi syndrome which shows an absence of Snrpn expression. *Nature genetics* **2**(4): 270-274.

Cattanach BM, Isaacson JH. 1967. Controlling elements in the mouse X chromosome. *Genetics* **57**(2): 331-346.

Cattanach BM, Kirk M. 1985. Differential activity of maternally and paternally derived chromosome regions in mice. *Nature* **315**(6019): 496-498.

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.

Chadwick LH, Willard HF. 2005. Genetic and parent-of-origin influences on X chromosome choice in Xce heterozygous mice. *Mamm Genome* **16**(9): 691-699.

Chamberlain SJ, Brannan CI. 2001. The Prader-Willi syndrome imprinting center activates the paternally expressed murine Ube3a antisense transcript but represses paternal Ube3a. *Genomics* **73**(3): 316-322.

Charalambous M, Menheniott TR, Bennett WR, Kelly SM, Dell G, Dandolo L, Ward A. 2004. An enhancer element at the Igf2/H19 locus drives gene expression in both imprinted and non-imprinted tissues. *Developmental biology* **271**(2): 488-497.

Cheng MK, Disteche CM. 2004. Silence of the fathers: early X inactivation. *Bioessays* **26**(8): 821-824.

Choi JD, Underkoffler LA, Collins JN, Marchegiani SM, Terry NA, Beechey CV, Oakey RJ. 2001. Microarray expression profiling of tissues from mice with uniparental duplications of chromosomes 7 and 11 to identify imprinted genes. *Mamm Genome* **12**(10): 758-764.

Choi JD, Underkoffler LA, Wood AJ, Collins JN, Williams PT, Golden JA, Schuster EF, Jr., Loomes KM, Oakey RJ. 2005. A novel variant of Inpp5f is imprinted in brain, and its expression is correlated with differential methylation of an internal CpG island. *Molecular and cellular biology* **25**(13): 5514-5522.

Chrast R, Scott HS, Papasavvas MP, Rossier C, Antonarakis ES, Barras C, Davisson MT, Schmidt C, Estivill X, Dierssen M et al. 2000. The mouse brain transcriptome by SAGE: differences in gene expression between P30 brains of the partial trisomy 16 mouse model of Down syndrome (Ts65Dn) and normals. *Genome Res* **10**(12): 2006-2021.

Conover WJ, Iman RL. 1981. Rank Transformations as a Bridge between Parametric and Nonparametric Statistics. *American Statistician* **35**(3): 124-129.

Coombes C, Arnaud P, Gordon E, Dean W, Coar EA, Williamson CM, Feil R, Peters J, Kelsey G. 2003. Epigenetic properties and identification of an imprint mark in the Nesp-Gnasxl domain of the mouse Gnas imprinted locus. *Molecular and cellular biology* **23**(16): 5475-5488.

Cooper DW, Johnston PG, Vandeberg JL, Robinson ES. 1990. X-Chromosome Inactivation in Marsupials. *Australian Journal of Zoology* **37**(2-4): 411-417.

Courtier B, Heard E, Avner P. 1995. Xce haplotypes show modified methylation in a region of the active X chromosome lying 3' to Xist. *Proceedings of the National Academy of Sciences of the United States of America* **92**(8): 3531-3535.

da Rocha ST, Tevendale M, Knowles E, Takada S, Watkins M, Ferguson-Smith AC.

2007. Restricted co-expression of Dlk1 and the reciprocally imprinted non-coding RNA, Gtl2: implications for cis-acting control. *Developmental biology* **306**(2): 810-823.

Daelemans C, Ritchie ME, Smits G, Abu-Amero S, Sudbery IM, Forrest MS, Campino S, Clark TG, Stanier P, Kwiatkowski D et al. 2010. High-throughput analysis of candidate imprinted genes and allele-specific gene expression in the human term placenta. *BMC genetics* **11**: 25.

Dao D, Frank D, Qian N, O'Keefe D, Vosatka RJ, Walsh CP, Tycko B. 1998. IMPT1, an imprinted gene similar to polyspecific transporter and multi-drug resistance genes. *Human molecular genetics* **7**(4): 597-608.

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.

Davies W, Smith RJ, Kelsey G, Wilkinson LS. 2004. Expression patterns of the novel imprinted genes Nap1l5 and Peg13 and their non-imprinted host genes in the adult mouse brain. *Gene Expr Patterns* **4**(6): 741-747.

Deakin JE, Hore TA, Koina E, Marshall Graves JA. 2008. The status of dosage compensation in the multiple X chromosomes of the platypus. *PLoS Genet* **4**(7): e1000140.

DeChiara TM, Robertson EJ, Efstratiadis A. 1991. Parental imprinting of the mouse insulin-like growth factor II gene. *Cell* **64**(4): 849-859.

Delaval K, Feil R. 2004. Epigenetic regulation of mammalian genomic imprinting. *Current opinion in genetics & development* **14**(2): 188-195.

Disteche CM, Filippova GN, Tsuchiya KD. 2002. Escape from X inactivation. *Cytogenet Genome Res* **99**(1-4): 36-43.

Engemann S, Strodicke M, Paulsen M, Franck O, Reinhardt R, Lane N, Reik W, Walter J. 2000. Sequence and functional comparison in the Beckwith-Wiedemann region: implications for a novel imprinting centre and extended imprinting. *Human molecular genetics* **9**(18): 2691-2706.

Evans HK, Weidman JR, Cowley DO, Jirtle RL. 2005. Comparative phylogenetic analysis of blcap/nnat reveals eutherian-specific imprinted gene. *Molecular biology and evolution* **22**(8): 1740-1748.

Ferguson-Smith AC, Cattanach BM, Barton SC, Beechey CV, Surani MA. 1991. Embryological and molecular investigations of parental imprinting on mouse chromosome 7. *Nature* **351**(6328): 667-670.

Fitzpatrick GV, Soloway PD, Higgins MJ. 2002. Regional loss of imprinting and growth deficiency in mice with a targeted deletion of KvDMR1. *Nature genetics* **32**(3): 426-431.

Forrester LM, Ansell JD. 1985. Parental influences on X chromosome expression. *Genet Res* **45**(1): 95-100.

Fowlis DJ, Ansell JD, Micklem HS. 1991. Further evidence for the importance of parental source of the Xce allele in X chromosome inactivation. *Genet Res* **58**(1): 63-65.

Foy RL, Song IY, Chitalia VC, Cohen HT, Saksouk N, Cayrou C, Vaziri C, Cote J, Panchenko MV. 2008. Role of Jade-1 in the histone acetyltransferase (HAT) HBO1 complex. *J Biol Chem* **283**(43): 28817-28826.

Frazer KA, Eskin E, Kang HM, Bogue MA, Hinds DA, Beilharz EJ, Gupta RV, Montgomery J, Morenzoni MM, Nilsen GB et al. 2007. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature* **448**(7157): 1050-1053.

Frost JM, Moore GE. 2010. The importance of imprinting in the human placenta. *PLoS Genet* **6**: e1001015.

Gartler SM, Riggs AD. 1983. Mammalian X-chromosome inactivation. *Annu Rev Genet* **17**: 155-190.

Gimelbrant A, Hutchinson JN, Thompson BR, Chess A. 2007. Widespread monoallelic expression on human autosomes. *Science (New York, NY* **318**(5853): 1136-1140.

Goto T, Wright E, Monk M. 1997. Paternal X-chromosome inactivation in human trophoblastic cells. *Mol Hum Reprod* **3**(1): 77-80.

Gould TD, Pfeifer K. 1998. Imprinting of mouse Kvlqt1 is developmentally regulated. *Human molecular genetics* **7**(3): 483-487.

Greenfield A, Carrel L, Pennisi D, Philippe C, Quaderi N, Siggers P, Steiner K, Tam PP, Monaco AP, Willard HF et al. 1998. The UTX gene escapes X inactivation in mice and humans. *Human molecular genetics* **7**(4): 737-742.

Gregg C, Zhang J, Butler JE, Haig D, Dulac C. 2010a. Sex-specific parent-of-origin allelic expression in the mouse brain. *Science (New York, NY* **329**(5992): 682-685.

Gregg C, Zhang J, Weissbourd B, Luo S, Schroth GP, Haig D, Dulac C. 2010b. High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *Science (New York, NY* **329**(5992): 643-648.

Hagiwara Y, Hirai M, Nishiyama K, Kanazawa I, Ueda T, Sakaki Y, Ito T. 1997. Screening for imprinted genes by allelic message display: identification of a paternally expressed gene impact on mouse chromosome 18. *Proceedings of the National Academy of Sciences of the United States of America* **94**(17): 9249-9254.

Hatada I, Morita S, Obata Y, Sotomaru Y, Shimoda M, Kono T. 2001. Identification of a new imprinted gene, Rian, on mouse chromosome 12 by fluorescent differential display screening. *J Biochem* **130**(2): 187-190.

Hatada I, Mukai T. 1995. Genomic imprinting of p57KIP2, a cyclin-dependent kinase inhibitor, in mouse. *Nature genetics* **11**(2): 204-206.

Hatada I, Sugama T, Mukai T. 1993. A new imprinted gene cloned by a methylation-sensitive genome scanning method. *Nucleic acids research* **21**(24): 5577-5582.

Heard E, Disteche CM. 2006. Dosage compensation in mammals: fine-tuning the expression of the X chromosome. *Genes & development* **20**(14): 1848-1867.

Hemberger M, Redies C, Krause R, Oswald J, Walter J, Fundele RH. 1998. H19 and Igf2 are expressed and differentially imprinted in neuroectoderm-derived cells in the mouse brain. *Development genes and evolution* **208**(7): 393-402.

Henckel A, Arnaud P. 2010. Genome-wide identification of new imprinted genes. *Brief Funct Genomics* **9**(4): 304-314.

Hiby SE, Lough M, Keverne EB, Surani MA, Loke YW, King A. 2001. Paternal monoallelic expression of PEG3 in the human placenta. *Human molecular genetics* **10**(10): 1093-1100.

Hikichi T, Kohda T, Kaneko-Ishino T, Ishino F. 2003. Imprinting regulation of the murine Meg1/Grb10 and human GRB10 genes; roles of brain-specific promoters and mouse-specific CTCF-binding sites. *Nucleic acids research* **31**(5): 1398-1406.

Hoshiya H, Meguro M, Kashiwagi A, Okita C, Oshimura M. 2003. Calcr, a brain-specific imprinted mouse calcitonin receptor gene in the imprinted cluster of the proximal region of chromosome 6. *Journal of human genetics* **48**(4): 208-211.

Hu JF, Balaguru KA, Ivaturi RD, Oruganti H, Li T, Nguyen BT, Vu TH, Hoffman AR. 1999. Lack of reciprocal genomic imprinting of sense and antisense RNA of mouse insulin-like growth factor II receptor in the central nervous system. *Biochemical and biophysical research communications* **257**(2): 604-608.

Hu JF, Vu TH, Hoffman AR. 1995. Differential biallelic activation of three insulin-like growth factor II promoters in the mouse central nervous system. *Mol Endocrinol* **9**(5): 628-636.

Hudson QJ, Kulinski TM, Huetter SP, Barlow DP. 2010. Genomic imprinting mechanisms in embryonic and extraembryonic mouse tissues. *Heredity* **105**(1): 45-56.

Huynh KD, Lee JT. 2001. Imprinted X inactivation in eutherians: a model of gametic execution and zygotic relaxation. *Curr Opin Cell Biol* **13**(6): 690-697.

Huynh KD, Lee JT. 2005. X-chromosome inactivation: a hypothesis linking ontogeny and phylogeny. *Nat Rev Genet* **6**(5): 410-418.

Jaenisch R, Beard C, Lee J, Marahrens Y, Panning B. 1998. Mammalian X chromosome inactivation. *Novartis Foundation symposium* **214**: 200-209; discussion 209-213, 228-232.

Jiang YH, Bressler J, Beaudet AL. 2004. Epigenetics and human disease. *Annual review of genomics and human genetics* **5**: 479-510.

Jones BK, Levorse J, Tilghman SM. 2001. Deletion of a nuclease-sensitive region between the Igf2 and H19 genes leads to Igf2 misregulation and increased adiposity. *Human molecular genetics* **10**(8): 807-814.

Jong MT, Carey AH, Caldwell KA, Lau MH, Handel MA, Driscoll DJ, Stewart CL, Rinchik EM, Nicholls RD. 1999. Imprinting of a RING zinc-finger encoding gene in the mouse chromosome region homologous to the Prader-Willi syndrome genetic region. *Human molecular genetics* **8**(5): 795-803.

Kagitani F, Kuroiwa Y, Wakana S, Shiroishi T, Miyoshi N, Kobayashi S, Nishida M, Kohda T, Kaneko-Ishino T, Ishino F. 1997. Peg5/Neuronatin is an imprinted gene located on sub-distal chromosome 2 in the mouse. *Nucleic acids research* **25**(17): 3428-3432.

Kaneko-Ishino T, Kuroiwa Y, Miyoshi N, Kohda T, Suzuki R, Yokoyama M, Viville S, Barton SC, Ishino F, Surani MA. 1995. Peg1/Mest imprinted gene on chromosome 6 identified by cDNA subtraction hybridization. *Nature genetics* **11**(1): 52-59.

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.

Kikyo N, Williamson CM, John RM, Barton SC, Beechey CV, Ball ST, Cattanach BM, Surani MA, Peters J. 1997. Genetic and functional analysis of neuronatin in mice with maternal or paternal duplication of distal Chr 2. *Developmental biology* **190**(1): 66-77.

Kim J, Bergmann A, Wehri E, Lu X, Stubbs L. 2001. Imprinting and evolution of two Kruppel-type zinc-finger genes, ZIM3 and ZNF264, located in the PEG3/USP29 imprinted domain. *Genomics* **77**(1-2): 91-98.

Kim J, Lu X, Stubbs L. 1999. Zim1, a maternally expressed mouse Kruppel-type zinc-finger gene located in proximal chromosome 7. *Human molecular genetics* **8**(5): 847-854.

Kim J, Noskov VN, Lu X, Bergmann A, Ren X, Warth T, Richardson P, Kouprina N, Stubbs L. 2000. Discovery of a novel, paternally expressed ubiquitin-specific processing protease gene through comparative analysis of an imprinted region of mouse chromosome 7 and human chromosome 19q13.4. *Genome Res* **10**(8): 1138-1147.

Kobayashi H, Yamada K, Morita S, Hiura H, Fukuda A, Kagami M, Ogata T, Hata K, Sotomaru Y, Kono T. 2009. Identification of the mouse paternally expressed imprinted gene Zdbf2 on chromosome 1 and its imprinted human homolog ZDBF2 on chromosome 2. *Genomics* **93**(5): 461-472.

Kobayashi S, Isotani A, Mise N, Yamamoto M, Fujihara Y, Kaseda K, Nakanishi T, Ikawa M, Hamada H, Abe K et al. 2006. Comparison of gene expression in male and female mouse blastocysts revealed imprinting of the X-linked gene, Rhox5/Pem, at preimplantation stages. *Current Biology* **16**(2): 166-172.

Krepischi AC, Kok F, Otto PG. 1998. X chromosome-inactivation patterns in patients with Rett syndrome. *Human genetics* **102**(3): 319-321.

Kuzmin A, Han Z, Golding MC, Mann MR, Latham KE, Varmuza S. 2008. The PcG gene Sfmbt2 is paternally expressed in extraembryonic tissues. *Gene Expr Patterns* **8**(2): 107-116.

Lee JT. 2000. Disruption of imprinted X inactivation by parent-of-origin effects at Tsix. *Cell* **103**(1): 17-27.

Lee JT, Lu N. 1999. Targeted mutagenesis of Tsix leads to nonrandom X inactivation. *Cell* **99**(1): 47-57.

Lee JT, Strauss WM, Dausman JA, Jaenisch R. 1996. A 450 kb transgene displays properties of the mammalian X-inactivation center. *Cell* **86**(1): 83-94.

Lee YJ, Park CW, Hahn Y, Park J, Lee J, Yun JH, Hyun B, Chung JH. 2000. Mit1/Lb9 and Copg2, new members of mouse imprinted genes closely linked to Peg1/Mest(1). *FEBS letters* **472**(2-3): 230-234.

Leff SE, Brannan CI, Reed ML, Ozcelik T, Francke U, Copeland NG, Jenkins NA. 1992. Maternal imprinting of the mouse Snrpn gene and conserved linkage homology with the human Prader-Willi syndrome region. *Nature genetics* **2**(4): 259-264.

Lehmann EL, Romano JP. 2005. *Testing statistical hypotheses*. Springer, New York.

Levin JH, Kaler SG. 2007. Non-random maternal X-chromosome inactivation associated with PHACES. *Clin Genet* **72**(4): 345-350.

Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A. 2010. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* **7**(9): 709-715.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25**(14): 1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**(16): 2078-2079.

Li N, Carrel L. 2008. Escape from X chromosome inactivation is an intrinsic property of the Jarid1c locus. *Proceedings of the National Academy of Sciences of the United States of America* **105**(44): 17055-17060.

Lomvardas S, Barnea G, Pisapia DJ, Mendelsohn M, Kirkland J, Axel R. 2006. Interchromosomal interactions and olfactory receptor choice. *Cell* **126**(2): 403-413.

Luedi PP, Dietrich FS, Weidman JR, Bosko JM, Jirtle RL, Hartemink AJ. 2007. Computational and experimental identification of novel human imprinted genes. *Genome Res* **17**(12): 1723-1730.

Luedi PP, Hartemink AJ, Jirtle RL. 2005. Genome-wide prediction of imprinted murine genes. *Genome Res* **15**(6): 875-884.

Lyon MF. 1961. Gene action in the X-chromosome of the mouse (Mus musculus L.). *Nature* **190**: 372-373.

MacDonald HR, Wevrick R. 1997. The necdin gene is deleted in Prader-Willi syndrome and is imprinted in human and mouse. *Human molecular genetics* **6**(11): 1873-1878.

Maeda N, Hayashizaki Y. 2006. Genome-wide survey of imprinted genes. *Cytogenet Genome Res* **113**(1-4): 144-152.

Marsh S. 2007. *Pyrosequencing protocols*. Humana Press, Totowa, N.J.

Martinez R, Bonilla-Henao V, Jimenez A, Lucas M, Vega C, Ramos I, Sobrino F, Pintado E. 2005. Skewed X inactivation of the normal allele in fully mutated female carriers determines the levels of FMRP in blood and the fragile X phenotype. *Mol Diagn* **9**(3): 157-162.

Matoba R, Kato K, Saito S, Kurooka C, Maruyama C, Sakakibara Y, Matsubara K. 2000. Gene expression in mouse cerebellum during its development. *Gene* **241**(1): 125-131.

McLaughlin D, Vidaki M, Renieri E, Karagogeos D. 2006. Expression pattern of the maternally imprinted gene Gtl2 in the forebrain during embryonic development and adulthood. *Gene Expr Patterns* **6**(4): 394-399.

Mergenthaler S, Hitchins MP, Blagitko-Dorfs N, Monk D, Wollmann HA, Ranke MB, Ropers HH, Apostolidou S, Stanier P, Preece MA et al. 2001. Conflicting reports of imprinting status of human GRB10 in developing brain: how reliable are somatic cell hybrids for predicting allelic origin of expression? *American journal of human genetics* **68**(2): 543-545.

Miyoshi N, Kuroiwa Y, Kohda T, Shitara H, Yonekawa H, Kawabe T, Hasegawa H, Barton SC, Surani MA, Kaneko-Ishino T et al. 1998. Identification of the Meg1/Grb10 imprinted gene on mouse proximal chromosome 11, a candidate for the Silver-Russell syndrome gene. *Proceedings of the National Academy of Sciences of the United States of America* **95**(3): 1102-1107.

Miyoshi N, Wagatsuma H, Wakana S, Shiroishi T, Nomura M, Aisaka K, Kohda T, Surani MA, Kaneko-Ishino T, Ishino F. 2000. Identification of an imprinted gene, Meg3/Gtl2 and its human homologue MEG3, first mapped on mouse distal chromosome 12 and human chromosome 14q. *Genes Cells* **5**(3): 211-220.

Mizuno Y, Sotomaru Y, Katsuzawa Y, Kono T, Meguro M, Oshimura M, Kawai J, Tomaru Y, Kiyosawa H, Nikaido I et al. 2002. Asb4, Ata3, and Dcn are novel imprinted genes identified by high-throughput screening using RIKEN cDNA microarray. *Biochemical and biophysical research communications* **290**(5): 1499-1505.

Moore T, Constancia M, Zubair M, Bailleul B, Feil R, Sasaki H, Reik W. 1997.

Multiple imprinted sense and antisense transcripts, differential methylation and tandem repeats in a putative imprinting control region upstream of mouse Igf2. *Proceedings of the National Academy of Sciences of the United States of America* **94**(23): 12509-12514.

Morison IM, Ramsay JP, Spencer HG. 2005. A census of mammalian imprinting. *Trends Genet* **21**(8): 457-465.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**(7): 621-628.

Murphy SK, Jirtle RL. 2003. Imprinting evolution and the price of silence. *Bioessays* **25**(6): 577-588.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science (New York, NY* **320**(5881): 1344-1349.

Namekawa SH, VandeBerg JL, McCarrey JR, Lee JT. 2007. Sex chromosome silencing in the marsupial male germ line. *Proceedings of the National Academy of Sciences of the United States of America* **104**(23): 9730-9735.

Nikaido I, Saito C, Mizuno Y, Meguro M, Bono H, Kadomura M, Kono T, Morris GA, Lyons PA, Oshimura M et al. 2003. Discovery of imprinted transcripts in the mouse transcriptome using large-scale expression profiling. *Genome Res* **13**(6B): 1402-1409.

Ogata T, Kagami M, Ferguson-Smith AC. 2008. Molecular mechanisms regulating phenotypic outcome in paternal and maternal uniparental disomy for chromosome 14. *Epigenetics* **3**(4): 181-187.

Ohlsson R, Hedborg F, Holmgren L, Walsh C, Ekstrom TJ. 1994. Overlapping patterns of IGF2 and H19 expression during human development: biallelic IGF2 expression correlates with a lack of H19 expression. *Development (Cambridge, England)* **120**(2): 361-368.

Ono R, Shiura H, Aburatani H, Kohda T, Kaneko-Ishino T, Ishino F. 2003. Identification of a large novel imprinted gene cluster on mouse proximal chromosome 6. *Genome Res* **13**(7): 1696-1705.

Parker-Katiraee L, Carson AR, Yamada T, Arnaud P, Feil R, Abu-Amero SN, Moore GE, Kaneda M, Perry GH, Stone AC et al. 2007. Identification of the imprinted KLF14 transcription factor undergoing human-specific accelerated evolution. *PLoS Genet* **3**(5): e65.

Pauler FM, Barlow DP. 2006. Imprinting mechanisms--it only takes two. *Genes & development* **20**(10): 1203-1206.

Paulsen M, Davies KR, Bowden LM, Villar AJ, Franck O, Fuermann M, Dean WL, Moore TF, Rodrigues N, Davies KE et al. 1998. Syntenic organization of the mouse distal chromosome 7 imprinting cluster and the Beckwith-Wiedemann syndrome region in chromosome 11p15.5. *Human molecular genetics* **7**(7): 1149-1159.

Payer B, Lee JT. 2008. X chromosome dosage compensation: how mammals keep the balance. *Annu Rev Genet* **42**: 733-772.

Peters J, Wroe SF, Wells CA, Miller HJ, Bodle D, Beechey CV, Williamson CM, Kelsey G. 1999. A cluster of oppositely imprinted transcripts at the Gnas locus in the distal imprinting region of mouse chromosome 2. *Proceedings of the National Academy of Sciences of the United States of America* **96**(7): 3830-3835.

Piras G, El Kharroubi A, Kozlov S, Escalante-Alcalde D, Hernandez L, Copeland NG, Gilbert DJ, Jenkins NA, Stewart CL. 2000. Zac1 (Lot1), a potential tumor suppressor gene, and the gene for epsilon-sarcoglycan are maternally imprinted genes: identification by a subtractive screen of novel uniparental fibroblast lines. *Molecular and cellular biology* **20**(9): 3308-3315.

Plass C, Shibata H, Kalcheva I, Mullins L, Kotelevtseva N, Mullins J, Kato R, Sasaki H, Hirotsune S, Okazaki Y et al. 1996. Identification of Grf1 on mouse chromosome 9 as an imprinted gene by RLGS-M. *Nature genetics* **14**(1): 106-109.

Plenge RM, Percec I, Nadeau JH, Willard HF. 2000. Expression-based assay of an X-linked gene to examine effects of the X-controlling element (Xce) locus.

Mamm Genome **11**(5): 405-408.

Pollard KS, Serre D, Wang X, Tao H, Grundberg E, Hudson TJ, Clark AG, Frazer K. 2008. A genome-wide approach to identifying novel-imprinted genes. *Human genetics* **122**(6): 625-634.

Proudhon C, Bourc'his D. 2010. Identification and resolution of artifacts in the interpretation of imprinted gene expression. *Brief Funct Genomics*.

Reik W, Walter J. 2001. Genomic imprinting: parental influence on the genome. *Nat Rev Genet* **2**(1): 21-32.

Sado T, Ferguson-Smith AC. 2005. Imprinted X inactivation and reprogramming in the preimplantation mouse embryo. *Human molecular genetics* **14 Spec No 1**: R59-64.

Sado T, Wang Z, Sasaki H, Li E. 2001. Regulation of imprinted X-chromosome inactivation in mice by Tsix. *Development (Cambridge, England)* **128**(8): 1275-1286.

Sakamoto K, Tamamura Y, Katsube K, Yamaguchi A. 2008. Zfp64 participates in Notch signaling and regulates differentiation in mesenchymal cells. *J Cell Sci* **121**(Pt 10): 1613-1623.

Schmidt JV, Matteson PG, Jones BK, Guan XJ, Tilghman SM. 2000. The Dlk1 and Gtl2 genes are linked and reciprocally imprinted. *Genes & development* **14**(16): 1997-2002.

Schulz R, Menheniott TR, Woodfine K, Wood AJ, Choi JD, Oakey RJ. 2006. Chromosome-wide identification of novel imprinted genes using microarrays and uniparental disomies. *Nucleic acids research* **34**(12): e88.

Seitz H, Royo H, Bortolin ML, Lin SP, Ferguson-Smith AC, Cavaille J. 2004. A large imprinted microRNA gene cluster at the mouse Dlk1-Gtl2 domain. *Genome Res* **14**(9): 1741-1748.

Seitz H, Youngson N, Lin SP, Dalbert S, Paulsen M, Bachellerie JP, Ferguson-Smith AC, Cavaille J. 2003. Imprinted microRNA genes transcribed antisense to a reciprocally imprinted retrotransposon-like gene. *Nature genetics* **34**(3): 261-262.

Serre D, Gurd S, Ge B, Sladek R, Sinnett D, Harmsen E, Bibikova M, Chudin E, Barker DL, 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 Genetics* **4**(2): e1000006.

Shimizu Y, Nagata M, Usui J, Hirayama K, Yoh K, Yamagata K, Kobayashi M, Koyama A. 2006. Tissue-specific distribution of an alternatively spliced COL4A5 isoform and non-random X chromosome inactivation reflect phenotypic variation in heterozygous X-linked Alport syndrome. *Nephrol Dial Transplant* **21**(6): 1582-1587.

Simmler MC, Cattanach BM, Rasberry C, Rougeulle C, Avner P. 1993. Mapping the murine Xce locus with (CA)n repeats. *Mamm Genome* **4**(9): 523-530.

Sleutels F, Tjon G, Ludwig T, Barlow DP. 2003. Imprinted silencing of Slc22a2 and Slc22a3 does not need transcriptional overlap between Igf2r and Air. *The EMBO journal* **22**(14): 3696-3704.

Sleutels F, Zwart R, Barlow DP. 2002. The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature* **415**(6873): 810-813.

Smith RJ, Arnaud P, Konfortova G, Dean WL, Beechey CV, Kelsey G. 2002. The mouse Zac1 locus: basis for imprinting and comparison with human ZAC. *Gene* **292**(1-2): 101-112.

Smith RJ, Dean W, Konfortova G, Kelsey G. 2003. Identification of novel imprinted genes in a genome-wide screen for maternal methylation. *Genome Res* **13**(4): 558-569.

Sritanaudomchai H, Ma H, Clepper L, Gokhale S, Bogan R, Hennebold J, Wolf D, Mitalipov S. 2010. Discovery of a novel imprinted gene by transcriptional analysis of parthenogenetic embryonic stem cells. *Hum Reprod* **25**(8): 1927-1941.

Storer BE, Kim C. 1990. Exact Properties of Some Exact Test Statistics for Comparing 2 Binomial Proportions. *Journal of the American Statistical Association* **85**(409): 146-155.

Storey JD, Taylor JE, Siegmund D. 2004. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society Series B-Statistical Methodology* **66**: 187-205.

Straub T, Becker PB. 2007. Dosage compensation: the beginning and end of generalization. *Nat Rev Genet* **8**(1): 47-57.

Takada S, Tevendale M, Baker J, Georgiades P, Campbell E, Freeman T, Johnson MH, Paulsen M, Ferguson-Smith AC. 2000. Delta-like and gtl2 are reciprocally expressed, differentially methylated linked imprinted genes on mouse chromosome 12. *Curr Biol* **10**(18): 1135-1138.

Talebizadeh Z, Bittel DC, Veatch OJ, Kibiryeva N, Butler MG. 2005. Brief report: non-random X chromosome inactivation in females with autism. *Journal of autism and developmental disorders* **35**(5): 675-681.

Tocharus J, Tsuchiya A, Kajikawa M, Ueta Y, Oka C, Kawaichi M. 2004. Developmentally regulated expression of mouse HtrA3 and its role as an inhibitor of TGF-beta signaling. *Development, growth & differentiation* **46**(3): 257-274.

Umlauf D, Goto Y, Cao R, Cerqueira F, Wagschal A, Zhang Y, Feil R. 2004. Imprinting along the Kcnq1 domain on mouse chromosome 7 involves repressive histone methylation and recruitment of Polycomb group complexes. *Nature genetics* **36**(12): 1296-1300.

Valley CM, Willard HF. 2006. Genomic and epigenomic approaches to the study of X chromosome inactivation. *Current opinion in genetics & development* **16**(3): 240-245.

van den Berg IM, Laven JS, Stevens M, Jonkers I, Galjaard RJ, Gribnau J, van Doorninck JH. 2009. X chromosome inactivation is initiated in human preimplantation embryos. *American journal of human genetics* **84**(6): 771-779.

Vicoso B, Charlesworth B. 2006. Evolution on the X chromosome: unusual patterns and processes. *Nat Rev Genet* **7**(8): 645-653.

Wagschal A, Feil R. 2006. Genomic imprinting in the placenta. *Cytogenet Genome*

Res **113**(1-4): 90-98.

Wang X, Soloway PD, Clark AG. 2010. Paternally biased X inactivation in mouse neonatal brain. *Genome Biol* **11**(7): R79.

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.

Wang Y, Joh K, Masuko S, Yatsuki H, Soejima H, Nabetani A, Beechey CV, Okinami S, Mukai T. 2004. The mouse Murr1 gene is imprinted in the adult brain, presumably due to transcriptional interference by the antisense-oriented U2af1-rs1 gene. *Molecular and cellular biology* **24**(1): 270-279.

Warren WC Hillier LW Marshall Graves JA Birney E Ponting CP Grutzner F Belov K Miller W Clarke L Chinwalla AT et al. 2008. Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453**(7192): 175-183.

Weinstein LS, Liu J, Sakamoto A, Xie T, Chen M. 2004. Minireview: GNAS: normal and abnormal functions. *Endocrinology* **145**(12): 5459-5464.

Weinstein LS, Yu S, Ecelbarger CA. 2000. Variable imprinting of the heterotrimeric G protein G(s) alpha-subunit within different segments of the nephron. *American journal of physiology* **278**(4): F507-514.

Weinstein LS, Yu S, Warner DR, Liu J. 2001. Endocrine manifestations of stimulatory G protein alpha-subunit mutations and the role of genomic imprinting. *Endocrine reviews* **22**(5): 675-705.

Wilcox RR. 2003. *Applying contemporary statistical techniques*. Academic Press, Amsterdam ; Boston.

Wilson EB. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* **22**: 209-212.

Wood AJ, Roberts RG, Monk D, Moore GE, Schulz R, Oakey RJ. 2007. A screen for retrotransposed imprinted genes reveals an association between X chromosome homology and maternal germ-line methylation. *PLoS Genet* **3**(2): e20.

Wutz A, Smrzka OW, Schweifer N, Schellander K, Wagner EF, Barlow DP. 1997. Imprinted expression of the Igf2r gene depends on an intronic CpG island.

Nature **389**(6652): 745-749.

Yamazawa K, Kagami M, Ogawa M, Horikawa R, Ogata T. 2008. Placental hypoplasia in maternal uniparental disomy for chromosome 7. *American journal of medical genetics* **146A**(4): 514-516.

Yang F, Babak T, Shendure J, Disteche CM. 2010. Global survey of escape from X inactivation by RNA-sequencing in mouse. *Genome Res* **20**(5): 614-622.

Yang HH, Hu Y, Edmonson M, Buetow K, Lee MP. 2003. Computation method to identify differential allelic gene expression and novel imprinted genes. *Bioinformatics (Oxford, England)* **19**(8): 952-955.

Yevtodiyenko A, Steshina EY, Farner SC, Levorse JM, Schmidt JV. 2004. A 178-kb BAC transgene imprints the mouse Gtl2 gene and localizes tissue-specific regulatory elements. *Genomics* **84**(2): 277-287.

Yu S, Yu D, Lee E, Eckhaus M, Lee R, Corria Z, Accili D, Westphal H, Weinstein LS. 1998. Variable and tissue-specific hormone resistance in heterotrimeric Gs protein alpha-subunit (Gsalpha) knockout mice is due to tissue-specific imprinting of the gsalpha gene. *Proceedings of the National Academy of Sciences of the United States of America* **95**(15): 8715-8720.

Zwart R, Sleutels F, Wutz A, Schinkel AH, Barlow DP. 2001. Bidirectional action of the Igf2r imprint control element on upstream and downstream imprinted genes. *Genes & development* **15**(18): 2361-2366.