Supplementary Materials for

The Landscape of Genomic imprinting across diverse adult human tissues

Yael Baran[1], Meena Subramaniam[2], Anne Biton[2], Taru Tukiainen[3,4], Emily K Tsang[5,6], Manuel A Rivas[7], Matti Pirinen[8], Maria Gutierrez-Arcelus[9], Kevin S Smith[5], Kim R Kukurba[5,10], Rui Zhang[10], Celeste Eng[2], Dara G Torgerson[2], Cydney Urbanek[11], the GTEx Consortium, Jin Billy Li[10], Jose R. Rodriguez-Santana[12], Esteban G. Burchard[2,13], Max A. Seibold[11,14,15], Daniel G MacArthur[3,4,16], Stephen B Montgomery[5], Noah A Zaitlen[2]†*, Tuuli Lappalainen[17,18]†*

Correspondence to: tlappalainen@nygenome.org , noah.zaitlen@ucsf.org

This PDF file includes:

Materials and Methods
Figs. S1-S24
Captions for tables S1-S7
Captions for data S1-S5

Other Supplementary Materials for this manuscript includes the following:

Tables S1-4, S6-7 as an Exel file with multiple tabs
Table S5 as a zipped txt file
Data files S1-S5

Materials and Methods

1. Trio validation data

1.1 Genetics of Inherited Muscle Disease Cohort

Imprinting was assessed in skeletal muscle from six individuals for whom both individual and parental genotypes were available, hence allowing for the parental origin of the alleles to be determined. More specifically, the data set consists of five trios and one child-mother pair that were exome sequenced as part of a larger cohort (dbGaP accession

phs000655.v1.p1) the six probands each being affected with a neuromuscular disorder of currently unknown cause.

Exome capture was performed with Agilent SureSelect Human All Exon Kit v2 and the exome DNA was sequenced using Illumina HiSeq 2000 sequencer. Sequencing reads were aligned to the human reference genome (hg19) before calling single-nucleotide variants (SNVs) and small indels using the GATK version 3.0. A modified version of the Ensembl Variant Effect Predictor was used for variant annotation. The genotypes of the probands of the five trios and one parent-child pair were phased using the PhaseByTransmission tool of the GATK toolkit.

Non-strand specific RNA sequencing was performed for the poly-A selected mRNA isolated from the skeletal muscle biopsies using Illumina Tru Seq RNA Sample Preparation protocol with 76 bp paired-end sequencing reads. The sequencing was done on an Illumina HiSeq 2000 with five of the individuals sequenced to coverage of 50M and one to 500M of mapped paired-end reads. RNA-seq reads were aligned using Tophat version v1.4.1 with the UCSC human genome release version hg19 as the reference.

All exome and RNA sequencing was performed at the Broad Institute of Harvard and MIT following the same protocols used in the GTEx project, and the same ASE pipeline was used for this a data as for the GTEx data.

1.2 Genes-environments & Admixture in Latino Americans (GALA II)

Imprinting was assessed in nasal epithelium and whole blood from 10 trios from the GALA II cohort (Torgerson et al. 2011). More specifically, the data set consists of ten trios of Puerto Rican/Latino origin that were exome-sequenced. The ten probands were then RNA-sequenced in both tissues. Latino asthmatic probands and parents are Puerto Rican islanders recruited as part of the ongoing Genes Environments & Admixture in Latino Americans (GALA II) study described elsewhere (Borrell et al. 2013; Kumar et al. 2013; Nishimura et al. 2013). The nasal airway epithelium RNA-seq data used was previously published in a separate analysis of asthma differential expression (Poole et al. 2014). Probands had no history of smoking or recent nasal steroid use (within 4 weeks of recruitment). Methods for nasal airway brushing collection are described elsewhere (Poole et al. 2014). The study was approved by local institutional review boards, and written assent/consent was received from all subjects and their parents.

Exome capture was performed with Nimblegen SeqCap EZ Human Exome Library v3.0 and the exome DNA was sequenced using Illumina HiSeq 2500 sequencer to an average read depth of 39x. Sequencing reads were aligned to the human reference genome (hg19) before calling single-nucleotide variants (SNVs) and small indels using the GATK version 3.3-0. The genotypes of the probands of the ten trios were phased using the parent-offspring trio phasing of Beagle, and parental origin of each allele was assigned by matching phased parental haplotypes to offspring haplotypes within a 2000bp window above and below the SNP. If an exact match could not be found between the parental and offspring haplotypes, and the parental origin could not be inferred solely from the

genotypes at the SNP, then the parental origin of the allele was considered to be ambiguous.

Both nasal airway and blood RNA-seq libraries from the 10 probands were constructed and barcoded with the Illumina Tru Seq RNA Sample Preparation version 2 protocol (Illumina, San Diego, Calif). Barcoded RNA-seq libraries were run on flow cells of an Illumina HiSeq 2000 according to standard protocols using 2x100 paired end sequencing. RNA-seq reads were aligned using Tophat version v2.0.9 with the UCSC human genome release version hg19 as the reference. The same ASE pipeline was used for this data as for the GTEx data.

2. mmPCR-seq validation data

Microfluidic multiplex PCR sequencing (mmPCR-seq) (Zhang et al. 2014) was used to validate allelic ratios measured by RNA-Seq in 89 sites in 24 genes: ATP10A, COPG2, CPA4, ERLIN2, GRB10, H19, IGF2, KCNQ1, NAA60, NLRP2, NTM, PEG3, PHLDA2, PLAGL1, PPP1R9A, RB1, RBP5, SLC22A18, SNHG14, SNRPN, SNURF, UBE3A, WRB, ZNF331. The analysis was done for 121 GTEx samples from 9 individuals. The details of the experiment are in (Rivas et al. 2015). Briefly, PCR primers were designed to amplify the loci surrounding each site, and cDNA obtained from the RNA samples was amplified in multiplex PCR reactions using the Fluidigm Access Array. The pooled mmPCR libraries were sequenced on a MiSeq yielding 75 bp paired-end reads, and the data were aligned with STAR (Dobin et al. 2013) – alignment with TopHat yielded very similar results. Allelic counts were retrieved using an identical pipeline as for the GTEx RNA-seq data. The low number of individuals in the validation experiment did not allow us to use the statistical models developed for population-level RNA-seq data, but the allelic ratios obtained from mmPCR-seq data showed that the allelic counts themselves are reliable.

3. Long-read RNA-seq validation data

Standard RNA-seq data with relatively short reads can suffer from alignment errors and difficulty of determining the structure and annotation of the sequenced transcripts (Cho et al. 2014; Li et al. 2014). To this end, we analyzed long read strand-specific RNA-seq data (2 x 250 bp) from 34 GTEx samples from 5 individuals.

3.1. Library preparation, sequencing, and data processing

RNA sequencing was performed using a strand specific protocol with poly-A selection of mRNA. Strand specific RNA sequencing was performed at the Broad Institute using a large-scale, automated variant of the Illumina Tru Seq™ RNA Sample Preparation protocol. Briefly, 200 ng of total RNA was used from each sample as the starting material. This method uses oligo dT beads to select poly-A mRNA from the total RNA sample. The selected RNA is then heat fragmented and randomly primed before cDNA

synthesis from the RNA template. The resultant cDNA then goes through Illumina library preparation (end repair, base 'A' addition, adapter ligation, and enrichment) using Broad designed indexed adapters for multiplexing of samples, with 400 bp fragment size. After enrichment, the samples are qPCR quantified and equimolar pooled before proceeding to Illumina sequencing which was done on the Illumina HiSeq 2000 to a target depth of 100M reads. The entire process occurs in a 96-well format and all samples were electronically tracked through the process in real-time including reagent lot numbers, specific automation used, time stamps for each process step, and automatic registration.

RNA-seq data were aligned with Tophat version v1.4.1 to the UCSC human genome release version hg19. Gencode version 12 was used as a transcriptome model for the alignment as well as all gene and isoform quantifications. Unaligned reads were merged back in to create a final bam. Allele-specific expression was analyzed as for the other data sets.

3.2. Long read data analysis and results

Allele-specific expression estimates from the 2 x 250 bp and the standard 2 x 75 bp data are fully concordant (rho = 0.99 based on 707 sites with >=20 reads), showing that alignment error does not affect to our original estimates of monoallelic expression. Furthermore, using these data, we manually assessed each of the initial 21 novel or provisional genes to verify that (1) the transcript structure in the data corresponded to the gene annotation, (2) SNPs in the ASE analysis were in regions that correspond to the annotated transcripts, and (3) monoallelic SNPs did not overlap with known imprinted genes and showed no signs of switching between monoallelic/biallelic expression along the gene (Supplementary Figure 7).

The long read RNA-seq data showed that 4 out of 21 novel/provisional genes (LA16c-306E5.2, RP11-701H24.3, AL132709.5, RP11-395B7.2) were inconsistent with the gene annotations, showing either ambiguous (although often likely imprinted) transcription, or in one case imprinting derived from heterozygous SNPs that overlapped regions of a different known imprinted gene. These genes were removed from downstream analysis. All the other novel/provisional genes were relatively consistent with the Gencode annotation, although future work is needed to elucidate full transcript structure and gene annotation in the imprinted regions. See also section 9 for discussion of patterns observed in specific genes.

4. RNA-seq allele counts

For all the heterozygous sites per individual identified from genetic data, we calculated the number of REF and ALT alleles in RNA-sequencing data using the same pipeline for all the data sets – this has also been used in the original papers of each of the studies. We used only uniquely mapped reads, and required base quality >10. We excluded heterozygous sites with potential mapping errors: 50bp mapability <1 in the UCSC mapability track, and >5% bias in simulated RNA-sequencing data (Panousis et al. under review).

5. Methylation analysis

The Gencord data set includes methylation data from the Illumina 450K array from 107 fibroblast samples, 111 LCL samples, and 66 T-cell samples. In our analysis, we used normalized β–values; further details of the experiment and data processing are available in the original publication.

6. Statistical Method

We first describe the proposed model and the filtering steps we take. We then describe the classification based on the statistics output by our model. We conclude this section with a simulation study to examine edge properties of our approach.

6.1 Statistical Model

The input to our model is the genotypes (typed and imputed) of each individual, and the counts of RNA-seq alleles overlapping each SNP in each individual. We use the following notation for the count data:
n_{ij} - number of reads mapped to SNP *j* in individual *i*
r_{ij} - number of reads with the *ref* allele mapped to SNP *j* in individual *i*
h_{ij} - number of reads mapped to SNP j in individual *i* and phased to haplotype 1 (of unknown parental origin and with arbitrary haplotype numbering; these counts were generated by phasing the genotypes and combining alleles of the same haplotype)

We perform the analysis for each gene and for each tissue separately. Tissue indices are therefore discarded from the notation.

We use the following error probabilities:
p_g - genotyping error rate: set to 0.001 for non-imputed SNPs, and to 0.05 for imputed SNPs
p_s - sequencing error rate: set to 0.001
p_p - phasing error rate (we assume phasing errors in different SNPs along the gene are independent): set to 0.2

We say that a SNP is *informative* for a given individual in a given tissue if the individual is heterozygous and the SNP is covered by >=8 RNA-seq reads. Although our model accounts for genotyping and phasing error, SNPs covered by a small number of reads are uninformative. We therefore arbitrarily chose a threshold 8 of reads and show via simulation that our method is robust to false positives at this depth (see Section 6.5 below). Informativeness is tissue-specific since a heterozygous SNP may be covered and therefore informative for a given individual in one tissue but not in another. We denote by *site* any combination of (individual, informative SNP). Sites are tissue-specific since informative SNPs are tissue-specific.

A complete list of all symbols is provided in the Supplementary Text.

6.2 Filtering Steps

We first apply a series of filtering steps to address several of the technical and functional confounders described above:

a. Filtering of RNA-seq reads according to mapping and base quality to reduce the effect of mapping and sequencing errors, and filtering of SNPs with unreliable mapping to further remove SNPs where allelic mapping error is likely.

b. To filter out SNPs with high genotyping error rates, SNPs with a Hardy Weinberg p-value smaller than 10^{-3} are discarded from the analysis. Figure S1b depicts allele counts for SNPs in a gene for which more than half of the SNPs were removed due to deviation from HWE. Since multiple SNPs in this gene show only reference counts, this gene could potentially be handled also by the "flip test" that we describe in (d) below. In the Geuvadis dataset 2.7% of the SNPs failed the HWE filter.

c. To reduce the effects of NMD, for each gene we discarded individuals carrying a heterozygous premature stop SNP in that gene. Furthermore, NMD causing variants result in monoallelic expression only on heterozygous state, and these variants are typically rare ((Rivas et al. 2015), although see also (Andres et al. 2010)), and it is therefore very unlikely that NMD could cause a confounded imprinting signal with monoallelic expression in the vast majority of individuals. (Note that analogously, an eQTL can cause monoallelic expression only in individuals heterozygous for the eQTL, i.e. <=50% of individuals under HWE).

d. We apply a tissue-specific "flip test" to verify that the pattern of monoallelic expression is consistent with imprinting. We assume that with imprinting, the identity of a monoallelically expressed allele, either *ref* or *alt*, is independent of parent of origin, and therefore has an equal probability of being either of them. Genotyping error, RNA-seq sequencing error, and allelic bias in RNA-seq mapping are unlikely to flip randomly between the alleles and will therefore fail this test. eQTLs would cause random flipping of monoallelic expression only when the regulatory variant and the coding variants analyzed for monoallelic expression are not in LD; otherwise the SNPs in LD with an eQTL will also have consistently higher expression on one of the two alleles. We apply the flip test as follows:

1) For each SNP, we identify all individuals whose expression patterns appear monoallelic, and classify each such monoallelic site as *ref* or *alt* according to the over-expressed allele. Specifically, monoallelic sites are those for which the likelihood of the count data under the imprinted model is higher than the likelihoods under either the balanced or imbalanced models (these models are described next in section "Generative Model").
2) For each SNP, we compute a p-value for the null hypothesis that the fraction of *ref* sites out of all monoallelic sites was drawn from a binomial distribution with p=0.5. We remove all SNPs with p-value < 0.001.

We observe that this filter removes many genotyping and mapping errors. Figure S1c depicts an example of a gene that shows signs of imprinting before the flip test is applied but not afterwards. In the Geuvadis dataset 0.65% of the SNPs failed the "flip test".

e. For genes with a small number of individuals the flip test will not be well powered. We therefore test for imprinting only genes for which at least one SNP with both *ref* and *alt* monoallelic expression patterns have been observed. This will filter out genes in which monoallelicity either does not exist or is allele-specific and thus possibly driven by genotyping or mapping errors as discussed above. In the Geuvadis dataset, from the set of all genes showing some monoallelic expression, only 2215 qualified by this criterion. We therefore include statistical results of all genes in Supplementary Table S5 such that future studies with additional samples can leverage our analysis to determine additional imprinted genes.

6.3 Generative Model

We model individual *i*'s status in gene *g* and tissue *t* as being classified into one of three allelic expression classes:
(a) BAL (balanced) - The gene is expressed biallelically and evenly from both gene copies.
(b) IMB (imbalanced) - The gene exhibits allelic imbalance, *i.e.* one gene copy has a moderately higher expression level than the other. Such imbalance may result, for example, from an eQTL, in which case the expression level is sequence-dependent.
(c) IMP (imprinted) - The gene exhibits imprinting, *i.e.* one gene copy has a considerably higher expression level than the other, potentially depending on the parental origin. We assume that in this scenario one of the copies is nearly completely silenced.

Each allelic expression class is characterized by a Beta distribution (see Figure S3). For the balanced class, the distribution describes the relative expression of the reference allele level (*i.e.* the fraction of ref counts out of total counts) of SNPs residing in balanced genes. For the imbalanced and imprinted classes, the distribution describes the relative expression level of alleles residing in the over-expressed gene copy in imbalanced and imprinted genes, respectively. Given the allelic expression class and the phase information, the expression levels of multiple SNPs in the same gene are independently drawn from the relevant beta distribution. Independent sampling of relative expression levels along the gene is done as to account for isoform-specific silencing, splicing QTLs, and other biological effects that may cause inconsistency in allelic expression patterns in proximal sites, as well as for over-dispersion due to technical artifacts. Finally, given the allelic ratio and the total count data in a given site, the reference allele counts are drawn from the corresponding Binomial distribution.

A gene-specific and tissue-specific multinomial distribution determines the probability that any given individual belongs to each of the three expression classes. The multinomial distribution is characterized by the following parameters that sum to 1:
θ_g^{bal} - fraction of individuals belonging to the balanced class in gene *g*
θ_g^{imb} - fraction of individuals belonging to the imbalanced class in gene *g*

θ_g^{imp} - fraction of individuals belonging to the imprinted class in gene *g*
This model therefore allows for both imprinted and non-imprinted individuals in the same gene and tissue.

6.4 Model Computations

We first compute the likelihood that the count data observed for SNP *j* in individual *i* results from a genotyping error:

$$p_{ij}^{gerr} = p_g \cdot \left[\frac{1}{2} f_{bin}(r_{ij}|\, p_s, n_{ij}) + \frac{1}{2} f_{bin}(r_{ij}|\, 1 - p_s, n_{ij})\right]$$

where $f_{bin}(x|\, p, n)$ is the binomial probability density function with parameters (p, n), and p_g is set differently for typed and imputed SNPs, as explained above in *input data*.

We then compute the likelihood of the count data for the entire gene *g* given that individual *i* belongs to the balanced class:

$$L_{ig}^{bal} = \prod_{snp\ j \in g} \left[p_{ij}^{gerr} + (1 - p_g) \int_{x=0}^{1} f_{beta}^{bal}(x) \cdot f_{bin}(r_{ij}|x, n_{ij}) dx\right]$$

where $f_{beta}^{bal}(x)$ is the beta probability density function with parameters specific to the balanced class; this is the distribution from which the relative expression level of the reference alleles in balanced genes are sampled.

Similarly, the likelihood of the count data for gene *g* given that individual *i* belongs to the imprinted class, and that gene copy 1 is expressed (and gene copy 2 is silenced) is:

$$L_{ig}^{imp1} = \prod_{snp\ j \in g} \left[p_{ij}^{gerr} + (1 - p_g) \int_{x=0}^{1} f_{beta}^{imp}(x) \left[\left(1 - p_p\right) \cdot f_{bin}(h_{ij}|x, n_{ij}) + p_p \cdot f_{bin}(n_{ij} - h_{ij}|x, n_{ij})\right] dx\right]$$

and the likelihood of the same data given that gene copy 2 is the expressed one is:

$$L_{ig}^{imp2} = \prod_{snp\ j \in g} \left[p_{ij}^{gerr} + (1 - p_g) \int_{x=0}^{1} f_{beta}^{imp}(x) \left[\left(1 - p_p\right) \cdot f_{bin}\left(n_{ij} - h_{ij} | x, n_{ij}\right) + p_p \cdot f_{bin}(h_{ij} | x, n_{ij}) \right] dx \right]$$

where $f_{beta}^{imp}(x)$ is the beta probability density function with parameters specific to the imprinted class. Note that when a phasing error occurs in a given site, that site would seem to be expressed from the wrong gene copy. Therefore, given that a genotyping error did not occur, we draw the counts from the correct distribution with probability $\left(1 - p_p\right)$ and from the wrong distribution with probability p_p.

Since the numbering of the two gene copies is arbitrary, the likelihood of the count data for gene *g* given that individual *i* belongs to the imprinted class follows as:

$$L_{ig}^{imp} = \frac{1}{2} L_{ig}^{imp1} + \frac{1}{2} L_{ig}^{imp2}$$

The computations for the imbalanced class are identical to those for the imprinted class, but replacing f_{beta}^{imp} with f_{beta}^{imb}, which is the distribution from which the relative expression level of the over-expressed alleles in imbalanced genes are drawn.

Estimation of the Beta distribution parameters is described in Section "*Estimation of Beta parameters*" below.

6.5 Per-Gene Statistics

For each gene in each tissue we compute a set of statistics to summarize different aspects of expression and imprinting across individuals.

The overall likelihood of the RNA-seq data observed for gene *g* in a given tissue over all individuals is a function of the parameters (θ_g^{bal}, θ_g^{imb}, θ_g^{imp}):

$$L_g = \prod_{i \in individuals} \left[\theta_g^{bal} \cdot L_{ig}^{bal} + \theta_g^{imb} \cdot L_{ig}^{imb} + \theta_g^{imp} \cdot L_{ig}^{imp} \right]$$

Optimizing over $(\theta_g^{bal}, \theta_g^{imb}, \theta_g^{imp})$ yields $(\hat{\theta}_g^{bal}, \hat{\theta}_g^{imb}, \hat{\theta}_g^{imp})$, the maximum likelihood estimates for these parameters:

$$\left(\hat{\theta}_g^{bal}, \hat{\theta}_g^{imb}, \hat{\theta}_g^{imp}\right) = \underset{(\theta_g^{bal}, \theta_g^{imb}, \theta_g^{imp})}{\operatorname{argmax}} L_g$$

We compute the following likelihood statistics:

$$IMPGLR_g = \frac{\max\limits_{(\theta_g^{bal}, \theta_g^{imb}, \theta_g^{imp})} L_g}{\max\limits_{(\theta_g^{bal}, \theta_g^{imb}, \theta_g^{imp}=0)} L_g}$$

$IMPGLR_g$ is a generalized likelihood ratio statistic that quantifies the evidence for the existence of imprinting in gene g. Specifically, it is the ratio of the maximum data likelihood when allowing for imprinting to the maximum data likelihood under the assumption that imprinting does not exist in gene g.

$$IMPLR_g = \frac{\max\limits_{(\theta_g^{bal}=0, \theta_g^{imb}=0, \theta_g^{imp}=1)} L_g}{\max\limits_{(\theta_g^{bal}, \theta_g^{imb}, \theta_g^{imp}=0)} L_g}$$

$IMPLR_g$ is a likelihood ratio statistic that compares the hypotheses of gene g being imprinted and not imprinted under the assumption that all individuals share the same imprinting status. Specifically, it is the ratio of the data likelihood when all individuals are imprinted to the maximum data likelihood when all individuals are allowed to be either balanced or imbalanced, but not imprinted, in gene g.

$$HETLR_g = \frac{\max\limits_{(\theta_g^{bal}, \theta_g^{imb}, \theta_g^{imp})} L_g}{\max\left(\max\limits_{(\theta_g^{bal}, \theta_g^{imb}, \theta_g^{imp}=0)} L_g, \max\limits_{(\theta_g^{bal}=0, \theta_g^{imb}=0, \theta_g^{imp}=1)} L_g\right)}$$

$HETLR_g$ is a likelihood ratio statistic that quantifies the evidence for the existence of between-individuals heterogeneity in the imprinting status of gene g. Specifically, it is the ratio of the maximum data likelihood to the maximum data likelihood under the assumption that all individuals are either imprinted or not imprinted.

In the above likelihood statistics, maximum likelihood estimates for $\left(\theta_g^{bal}, \theta_g^{imb}, \theta_g^{imp}\right)$ are computed using the Expectation Maximization algorithm.

We compute the probabilities of gene g being balanced, imbalanced or imputed under the strict assumption that all individuals are either balanced, imbalanced or imputed:

$$z_g^{bal} = \frac{P^{bal} \prod_i L_{ig}^{bal}}{P^{bal} \prod_i L_{ig}^{bal} + P^{imb} \prod_i L_{ig}^{imb} + P^{imp} \prod_i L_{ig}^{imp}}$$

$$z_g^{imb} = \frac{P^{imb}\prod_i L_{ig}^{imb}}{P^{bal}\prod_i L_{ig}^{bal} + P^{imb}\prod_i L_{ig}^{imb} + P^{imp}\prod_i L_{ig}^{imp}}$$

$$z_g^{imp} \frac{P^{imp}\prod_i L_{ig}^{imp}}{P^{bal}\prod_i L_{ig}^{bal} + P^{imb}\prod_i L_{ig}^{imb} + P^{imp}\prod_i L_{ig}^{imp}}$$

In the equations above $P^{bal}, P^{imb}, P^{imp}$ are the prior probabilities of a given gene being balanced, imbalanced and imprinted. We set these probabilities to 0.8, 0.19 and 0.01, respectively. These are conservative estimates, which are meant to produce a gross scaling of the probabilities of the three classes.

τ_g - mean fraction of higher-frequency allele (either ref or alt) out of total counts, computed over all informative sites in gene *g*
φ_g – mean fraction of reference counts out of total counts, computed over all informative sites in gene *g*

6.6 Per-gene, per-individual Statistics

In addition to the statistics summarizing information over all individuals, we compute a series of statistics for each individual in a given gene and tissue. We denote by $P^{bal}, P^{imb}, P^{imp}$ the prior probabilities of a given gene being balanced, imbalanced and imprinted in any given individual. We set these probabilities to 0.8, 0.19 and 0.01, respectively, identically to the per-gene prior probabilities as described in "*Per-Gene Statistics*" above.
We compute the conditional probabilities of gene *g* in individual *i* belonging to each of the three classes as follows:

$$z_{ig}^{bal} = \frac{P^{bal}L_{ig}^{bal}}{P^{bal}L_{ig}^{bal} + P^{imb}L_{ig}^{imb} + P^{imp}L_{ig}^{imp}}$$

$$z_{ig}^{imb} = \frac{P^{imb}L_{ig}^{imb}}{P^{bal}L_{ig}^{bal} + P^{imb}L_{ig}^{imb} + P^{imp}L_{ig}^{imp}}$$

$$z_{ig}^{imp} = \frac{P^{imp}L_{ig}^{imp}}{P^{bal}L_{ig}^{bal} + P^{imb}L_{ig}^{imb} + P^{imp}L_{ig}^{imp}}$$

We count as balanced/imbalanced/imprinted any individual whose $z_{ig}^{bal}/z_{ig}^{imb}/z_{ig}^{imp}$ probability exceeds 0.7, respectively. We compute these statistics twice: First considering all individuals with at least one informative SNP in gene *g*, and second considering only individuals with at least two informative SNPs. We denote these sets of statistics as (bal1,imb1,imp1) and (bal2,imb2,imp2), respectively. Compared with the first set, the second one includes fewer but more confident counts. Finally, the statistics total1, total2 give the total number of individuals with at least 1,2 informative SNPs in gene *g*,

respectively. Note that not all individuals will be classified into a category as there must be substantial evidence to exceed the threshold, and therefore bal1+imb1+imp1 is often smaller than total1 (and same for total2).

6.7 Estimation of Beta parameters

For the parameter estimation procedure only we assume a simplistic mixture model in which the expression pattern of every gene can be classified into one of three different classes (BAL, IMB, IMP). Each class is characterized by a Beta distribution, from which the fraction of the over-expressed allele counts are drawn for the relevant sites. The parameters of the three classes are denoted as $(\alpha_{bal}, \beta_{bal})$, $(\alpha_{imb}, \beta_{imb})$ and $(\alpha_{imp}, \beta_{imp})$, respectively. The fraction of genes belonging to each class are denoted as q^{bal}, q^{imb} and q^{imp}, respectively.

Given the gene's class and the number of counts observed in the relevant sites, the reference counts are assumed to be drawn independently, regardless of the individual of origin, and discarding phase information. The likelihood of the count data as a function of the mixture model parameters is computed as

$$L(\boldsymbol{R}, \boldsymbol{N}|q, \alpha, \beta) = \prod_{\substack{g\in \\ genes}} \sum_{\substack{class\in \\ \{bal,imb,imp\}}} q^{classs} \prod_{\substack{i\in \\ individuals}} \prod_{snp\, j \in g} \left[p_{ij}^{gerr} + \left(1 - p_g\right) p_{ij}^{class}\right]$$

In the equation above R are the ref counts and N are the total counts, provided per individual and per site, and p_{ij}^{gerr} is computed as explained in section *"Model Computations"* above. p_{ij}^{class} is computed per site and per class as follows:

$$p_{ij}^{bal} = P\left(r_{ij}\middle|\alpha_{bal}, \beta_{bal}, n_{ij}\right) = \int_{x=0}^{1} f_{beta}^{bal}(x) \cdot f_{bin}\left(r_{ij}\middle|x, n_{ij}\right)$$

$$p_{ij}^{imb} = P\left(r_{ij}\middle|\alpha_{imb}, \beta_{imb}, n_{ij}\right) = \frac{1}{2}\int_{x=0}^{1} f_{beta}^{imb}(x) \cdot f_{bin}\left(r_{ij}\middle|1-x, n_{ij}\right) + \frac{1}{2}\int_{x=0}^{1} f_{beta}^{imb}(x) \cdot f_{bin}\left(r_{ij}\middle|x, n_{ij}\right)$$

$$p_{ij}^{imp} = P\left(r_{ij}\middle|\alpha_{imp}, \beta_{imp}, n_{ij}\right) = \frac{1}{2}\int_{x=0}^{1} f_{beta}^{imp}(x) \cdot f_{bin}\left(r_{ij}\middle|1-x, n_{ij}\right) + \frac{1}{2}\int_{x=0}^{1} f_{beta}^{imp}(x) \cdot f_{bin}\left(r_{ij}\middle|x, n_{ij}\right)$$

We use the Stochastic Expectation Maximization (SEM) algorithm (Celeux and Diebolt 1985) to estimate q, α and β. SEM is a modification of the EM algorithm in which the hidden variables are simulated according to their posterior probabilities, instead of being replaced by their expectations. SEM has been shown to produce stable estimates in similar problems.

We declare the following hidden indicator variables:

$$z_{ig}^{gerr} = \begin{cases} 1 & \text{if a genotyping error occurred in snp } j \text{ in individual } i \\ 0 & \text{otherwise} \end{cases}$$

$$z_{g}^{bal} = \begin{cases} 1 & \text{if gene } g \text{ belongs to the balanced class} \\ 0 & \text{otherwise} \end{cases}$$

$$z_{g}^{imb} = \begin{cases} 1 & \text{if gene } g \text{ belongs to the imbalanced class} \\ 0 & \text{otherwise} \end{cases}$$

$$z_{g}^{imp} = \begin{cases} 1 & \text{if gene } g \text{ belongs to the imprinted class} \\ 0 & \text{otherwise} \end{cases}$$

Initialization: We set $q^{(0)} = (q^{bal(0)}, q^{imb(0)}, q^{imp(0)})$ to $(1/3, 1/3, 1/3)$. In order to initialize the Beta parameters, we compute $\text{abs}(\frac{r_{ij}}{n_{ij}} - 0.5)$ for all sites in the dataset, and use the first decile to set $(\alpha^{bal(0)}, \beta^{bal(0)})$, the ninth decile to set $(\alpha^{imb(0)}, \beta^{imb(0)})$, and the tenth decile to set $(\alpha^{imp(0)}, \beta^{imp(0)})$. The motivation here is to initialize the distributions to reflect balanced, moderately imbalanced and highly imbalanced expression patterns, respectively. Given the sites in the relevant decile, the initialization is performed by optimizing each (α, β) pair using an interior-point procedure.

Iteration: In the ith iteration of the algorithm we perform the two following stages:

1. Drawing from expectation:
   a. Draw the hidden variables $z_g^{bal}, z_g^{imb}, z_g^{imp}$ for every gene g from their posterior multinomial distributions given q^{i-1} and $(\alpha, \beta)^{(i-1)} = (\alpha^{bal(i-1)}, \beta^{bal(i-1)}, \alpha^{imb(i-1)}, \beta^{imb(i-1)}, \alpha^{imp(i-1)}, \beta^{imp(i-1)})$.
   b. Draw the hidden variables z_{ij}^{gerr} for every individual i and for every SNP j, given the gene classifications drawn in 1(a).

2. Maximization: Set q^i and $(\alpha, \beta)^i$ by optimizing over the resulting complete data likelihood. The Beta parameters are obtained using an interior point procedure.

We ran the above procedure on the Geuvadis dataset. Following twenty iterations we observed that the absolute value of change in all parameters was smaller than 0.01, and used these estimates as the model's parameters. The parameter values are $(\alpha_{bal}, \beta_{bal}) = (45.5, 44.5)$, $(\alpha_{imb}, \beta_{imb}) = (6.3, 5.7)$, $(\alpha_{imp}, \beta_{imp}) = (0.64, 0.15)$.

6.8 Classification

The complete results of our approach applied to each data set are given in Supplementary Tables 3-5. Using these data we classified a set of imprinted genes in each tissue. Given the lack of parental inheritance information we took a conservative approach, requiring substantial evidence from multiple statistics and filters in order to classify a gene as imprinted. However, there were many genes that had moderate evidence of imprinting and these results can be combined with future studies in order to determine their validity. Below we describe the thresholds for classification in each of the four categories Imprinted, Biallelic, consistent with Imprinted, and consistent with Biallelic. Five genes classified as imprinted were removed based on validation results from the family or long-read RNA-seq analyses (see sections 1 and 3).

Imprinted

For identification of novel genes we required at least five individuals and two SNPs. For at least one SNP both alleles had to occur as the monoallelically expressed allele in at least one individual. We did not allow for novel heterogeneously imprinted genes (**hetlr** < 1.0), but removed this requirement if the gene was previously identified as imprinted in humans. If all of these filters were passed a gene was classified as imprinted if $\theta_g^{imp} > 0.7$, $\theta_g^{bal} < 0.05$, and **impglr** > 31.52 to account for the multiple hypotheses via a conservative Bonferoni correction for all genes/tissues examined (0.05/N=2532972 gene-tissue pairs examined). For the smaller set of previously identified genes we required $\theta_g^{imp} > 0.5$ and **impglr** > 13.7 (0.05/N=233 known gene-tissue pairs examined). The putatively imprinted genes defined according to these criteria are listed in Table S4. However, due to lack of comprehensive family data needed to confirm a parent of origin effect and formally define a false discovery rate obtained with these thresholds, and because the **glr** statistics are not guaranteed to be chi-square distributed due to the boundary condition of $\theta_g^{imp} = 0$ under the null, in the final analysis described in the main text, focusing on biological variation in imprinting, we used a more stringent threshold. Specifically, for novel genes, we required **impglr** > 50 if $\theta_g^{imp} = 1.00$ and **impglr** > 100 otherwise. For previously identified genes we required **impglr** > 40.

Biallelic

We next examined each of these imprinted genes in all tissues to determine, which tissues showed strong evidence of biallelic expression. We classified a gene as biallelic if it was not imprinted, and at least half of the individuals were classified as balanced by our method. Recall that individuals are only classified as biallelic if $z_{ig}^{bal} > 0.7$ and that not all individuals are classified as any of balanced, imbalanced or imprinted. That is, the method will give an unknown classification for ambiguous individuals with a small number of reads.

For genes with a small number of individuals, SNPs, or reads, there was not enough information to classify genes *de novo* as either imprinted or biallelic according to the metrics above. However, we did examine genes with lower thresholds conditional on the existence of a classified imprinted gene in at least one tissue. We created two additional categories “consistent with Imprinted” and “consistent with Biallelic”:

consistent with Imprinted

These are the set of genes in which the statistical model suggested imprinting, but there was not enough evidence for classification. We required that θ_g^{imp} accounted for 90% of the population of individuals. We used the same statistic, θ_g^{imp}, as in the classification procedure, but did not include a cutoff for the **impglr** test statistic allowing for individuals with low read counts. For example, a gene whose data consists of a single individual with read counts 20 *ref* and 0 *alt* at a single SNP would be classified into this category.

consistent with Biallelic

These are the set of genes in which the statistical model suggested biallelic expression, but there was not enough evidence for classification. We required that the combined set of biallelic and imbalanced individuals, as determined by the z_{ig}^{bal} and z_{ig}^{imb} statistics, accounted for at least 50% of the population of classified individuals.

7. Symbol and naming reference

Names and symbols used in this section and Tables S3-S5:

impglr – generalized LR statistic, quantifies the gain in likelihood obtained by allowing for imprinting in this gene (full description in supp methods)

implr - LR statistic, compares the assumption that all individual are imprinted to the assumption that none are (full description in supp methods)

hetlr – generalized LR statistic, quantifies the gain in likelihood obtained by allowing for heterogeneity in imprinting status across individuals (full description in supp methods)

$\hat{\theta}_g^{bal}, \hat{\theta}_g^{imb}, \hat{\theta}_g^{imp}$ – estimates for fraction of balanced, imbalanced and imprinted individuals

$z_g^{bal}, z_g^{imb}, z_g^{imp}$ - probabilities of the gene being balanced, imbalanced and imprinted in all individuals

bal2 – number of individuals with at least two informative SNPs who are classified as balanced

imb2 – number of individuals with at least two informative SNPs who are classified as imbalanced

imp2 – number of individuals with at least two informative SNPs who are classified as imprinted

total2 – number of individuals with at least two informative SNPs

bal1 – number of individuals with at least one informative SNP who are classified as balanced

imb1 – number of individuals with at least one informative SNP who are classified as imbalanced

imp1 – number of individuals with at least one informative SNP who are classified as imprinted

total1 – number of individuals with at least one informative SNP

both – a SNP monoallelic for both alleles was observed (yes/no)
known – a known imprinted gene (yes/no)
RME – a known random monoallelic expression gene (yes/no)
eQTLt – tissues in which known strong eQTLs exist
eQTLv – linear regression coefficient from allelic ratio ~ het|hom eQTL genotype

indn – number of individuals with at least one informative SNP
snpn– number of SNPs which are informative in at least one individual
indn2 – number of individuals with at least two informative SNP
snpn3 – number of SNPs which are informative in at least three individuals
τ_g – mean fraction of higher-frequency allele (either ref or alt) out of total counts, computed over all sites in gene g
φ_g – mean the fraction of reference counts out of total counts, computed over all sites in gene g

8. Known Imprinted Genes in Human and Mouse

The set of known imprinted genes in human and mouse were collected from the Otago database (Morison et al. 2001) by searching under category for status "imprinted" (Table S6). Provisional genes and genes with conflicting evidence were not considered amongst the known set of genes. We also added all genes from the Otago pdf 1101Summary-table.pdf, and human non-provisional "imprinted" genes from geneimprint.org. In the event that these data sources are not up to date with the latest literature, some of our novel genes may in fact be previously known.

We compared all identified imprinted genes with imprinting status in the mouse as retrieved from the Otago database and the results are shown in Figure S9. The results indicate that while there exists a high overlap between human and mouse imprinted genes, there are also uniquely human/mouse imprinted genes. However, not all tissues/developmental stages have been examined for imprinting.

9. Additional notes on specific genes

In this section, we briefly discuss the imprinting patterns of individual genes, based on additional inspection, especially of the novel imprinted genes discovered in this study, and other noteworthy observations.

- *IGF2*: We verified with visual inspection of both the long- and short-read RNA-seq data that the *IGF2* locus does not show signs of ambiguous transcription that could explain the different expressed allele in brain and muscle. Specifically, RNA-seq reads in the IGF2 region do not overlap the nearby H19 gene.
- *INPP5F*: The 3' exons of *INPP5F* overlap *INPP5F_V2*, which is annotated as a transcript of INPP5F according to both Gencode and RefSeq even though it is a previously known imprinted retrogene inside the non-imprinted *INPP5F* (Monk et al. 2011). Because of this, and all of our ASE data from *INPP5F* being from the region

overlapping *INPP5F_V2*, we treated *INPP5F_V2* as a separate gene (except in gene expression level analysis), observing signs of imprinting in several tissues. However, the long read data indicated expression both from *INPP5F* and *INPP5F_V2*, and fully distinguishing these transcripts is not possible in our data. Thus, lack of imprinting or a leaky signal (Supplementary Figure 1a) could be due to expression from the non-imprinted *INPP5F*.

- *SYCE1, THEGL,* and *MAGI2* have not been classified as imprinted in previous studies, and in our analysis there is strong support for imprinting only in the Geuvadis or Gencord cell line data sets. For the sake of consistency, these genes are included in this study, but we note that their imprinting status in primary tissues remains unclear.
- *SNHG14* is in the middle of a well-known imprinted region, but the annotation of the transcripts in this region remains unclear. We observe monoallelic transcription that is consistent with the Gencode annotation of *SNHG14*, but future work is needed to describe the full gene annotation of this region.
- NLRP2 is expressed in a monoallelic manner, but the trio data indicates non-parental origin of monoallelic expression (Supplementary Table 7).
- LPAR6 is a putative novel imprinted gene that is located in an intron of a previously identified RB1 gene. In our data, LPAR6 shows a stronger signal of imprinting than RB1 – although the signal is particularly strong in LCLs rather than primary tissue samples.

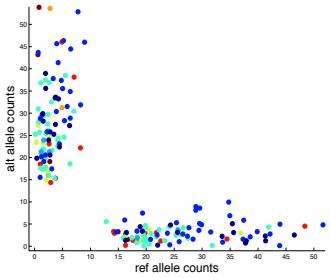


Fig. S1.

An example of partial imprinting in the *INPP5F_V2* in Gencord fibroblasts where the silenced allele is expressed albeit in a lower level. The plot depicts shows the alt vs ref allele counts for all SNPs and all individuals. Different colors indicate different SNPs.

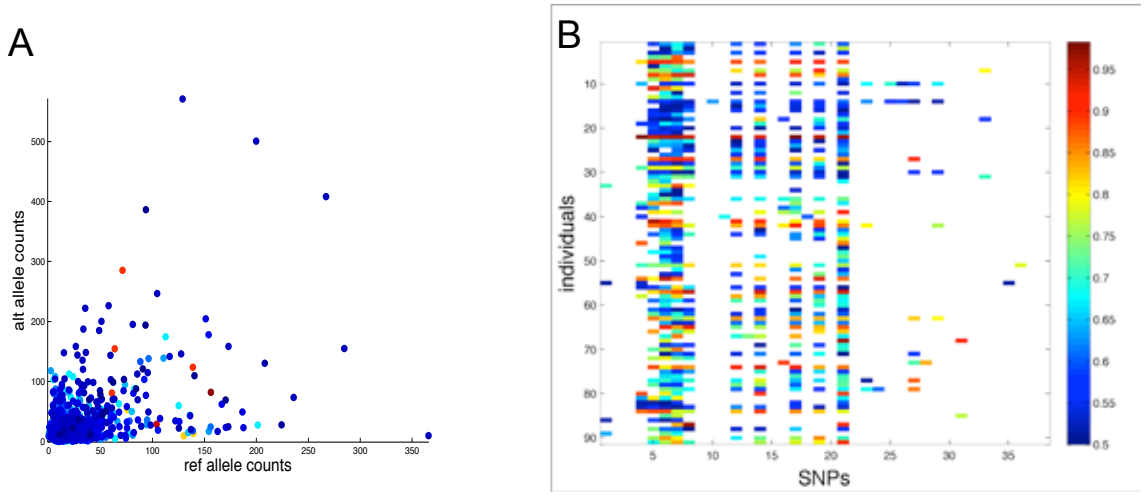


Fig. S2.

Random monoallelic expression (RME) manifested in the in Geuvadis LCL dataset. The gene P2RX5 was found to exhibit RME (Gimelbrant et al. 2007). A) shows the alt vs ref counts (as explained in Fig. S1). Figure B gives, for every heterozygous site with coverage >= 30, the relative frequency of the over-expressed allele (a number between 0.5 and 1). The monoallelic sites tend to appear in specific individuals, in whom the clonality level is presumably high.

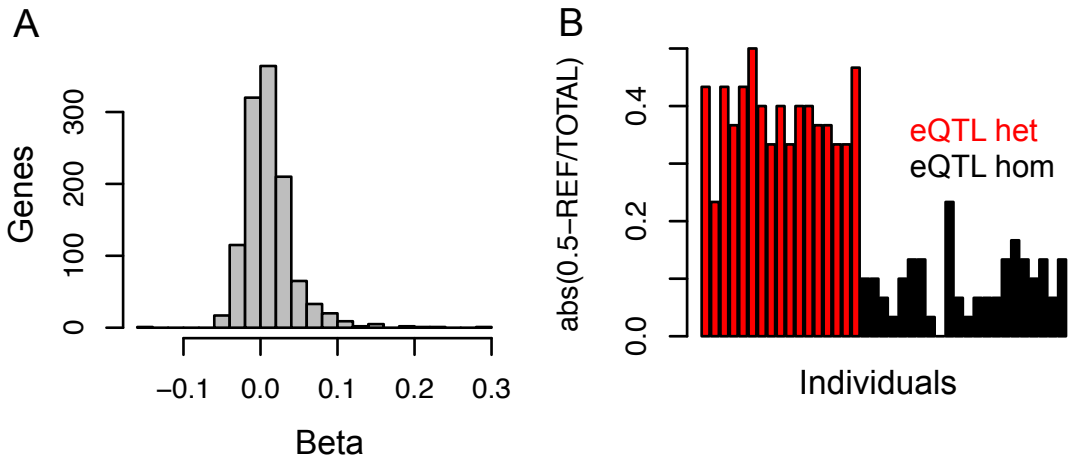


Fig. S3.

eQTL effects on monoallelic expression. Regulatory variants (eQTLs) are expected to lead to allelic imbalance in individuals that are heterozygous for eQTLs (max 50% of total). We investigated this in eQTL genes (identified by GTEx Consortium) by selecting one ASE site per gene per individual *i*, calculating allelic imbalance of that site as a_i = (|0.5-REF/TOTAL|), and modeling it as $a_i = \beta g_i + e$ where g_i is an indicator variable for whether the individual is heterozygous for the eQTL variant. A) shows the distribution of β from that model, showing that for most eQTLs, the effect of eQTLs on ASE is small. Other tissues show similar distributions. B) shows the pattern of allelic imbalance in the gene with the highest beta in thyroid, ENSG00000144115, with eQTL heterozygotes (red) having nearly monoallelic expression compared to homozygotes (black). Note that this is the worst case example rather than a typical eQTL effect on ASE. Altogether, an eQTL signal could be confused for a signal of imprinting only if the eQTL is extremely strong, has a high minor allele frequency, and none of the >=50% of individuals homozygous for the eQTL (thus lacking eQTL-induced ASE) have any heterozygous sites for ASE analysis. We did not find these patterns in our imprinted genes.

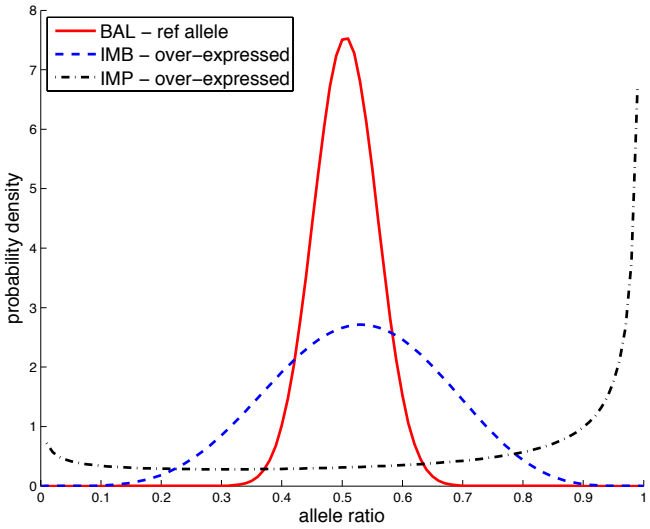


Fig. S4.

Three Beta distributions characterize the three expression classes of the model. Balanced class: The relative frequency of the reference allele is sampled from a Beta distribution concentrated around even expression, with a slight bias towards the reference allele. Imprinted class: The relative frequency of the allele residing on the over-expressed gene copy is sampled from a distribution strongly shifted towards 1. Imbalanced class: The relative frequency of the allele residing on the over-expressed gene copy is sampled from a high-variance Beta distribution shifted towards 1. Note that according to this distribution, it is common for an allele residing on the over-expressed gene copy to have an allele ratio less than 0.5. This situation often occurs in practice due to different biological and technical effects that cause inconsistency in imbalance levels along the gene.

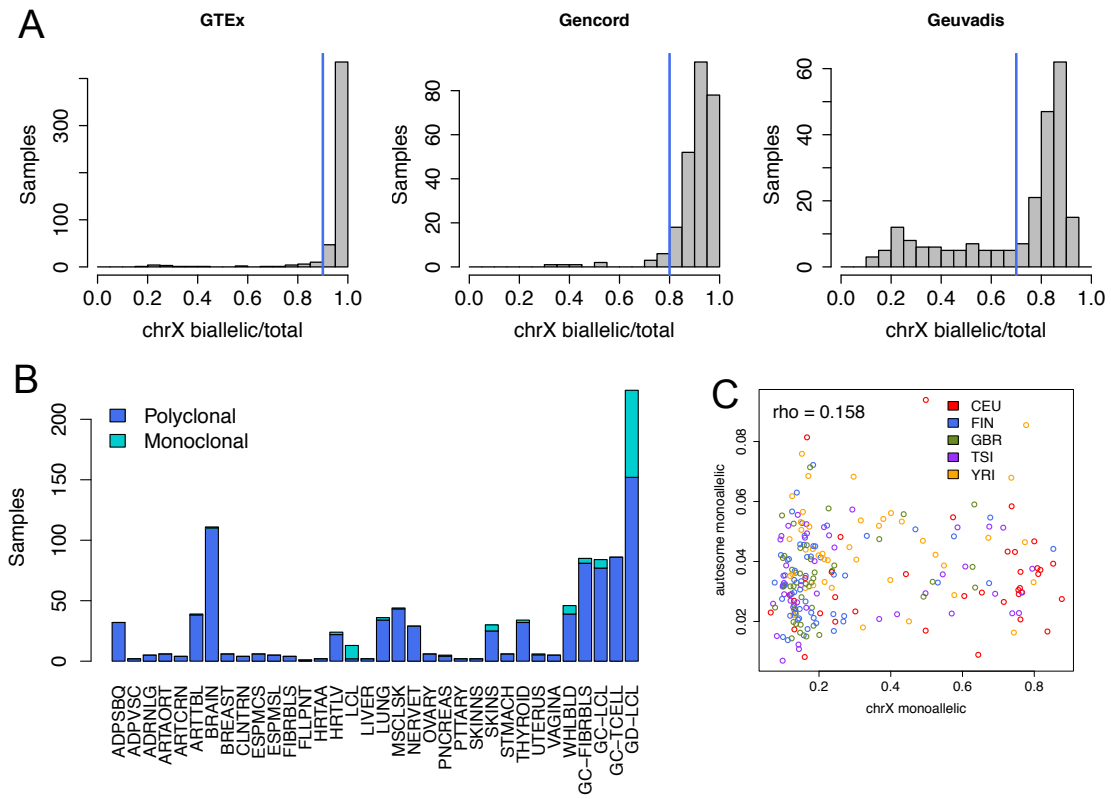


Fig. S5.

Clonality of samples as assessed from X-chromosome data. A) shows histograms of the proportion of biallelic sites in the X-chromosome per sample, in the three data sets used in this study. Different genotyping platforms and RNA-seq coverage make the exact proportions differ between the data sets; thus we estimated a separate threshold for each dataset (show in blue) for a sample to be considered potentially monoclonal. (B) shows this breakdown for each tissue. The LCLs of both GTEx and Geuvadis data sets have a substantial degree of clonality. (C) shows the very slight correlation of the degree of clonality from chrX and overall proportion of ASE in Geuvadis data. This shows that genome-wide ASE is not driven by epigenetic effects that are inherited in clonal cell lines. Furthermore, the differences between populations show the much higher degree of clonality in the CEU and YRI cell lines that are much older than the other cell lines (Coriell; personal communication).

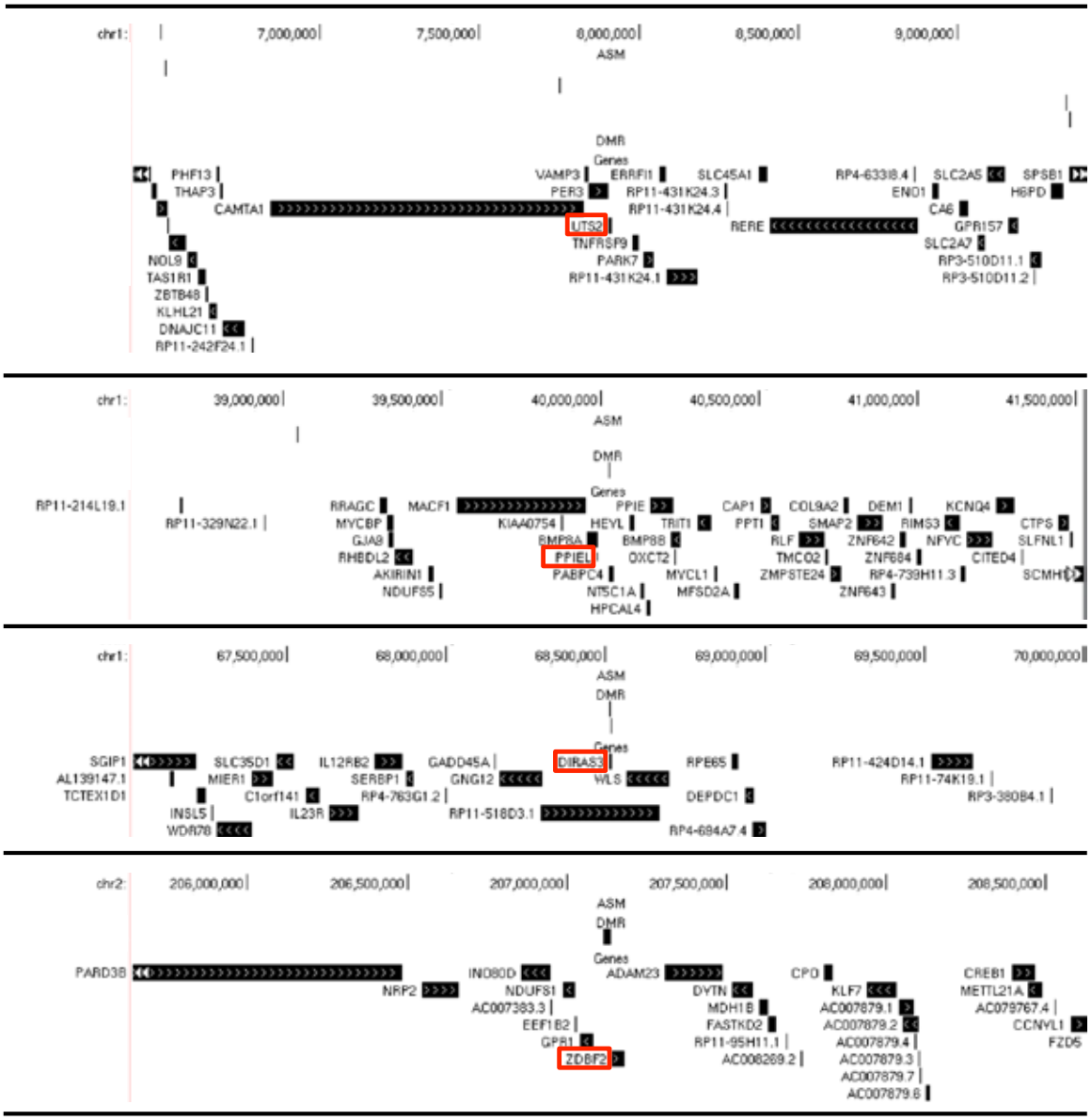
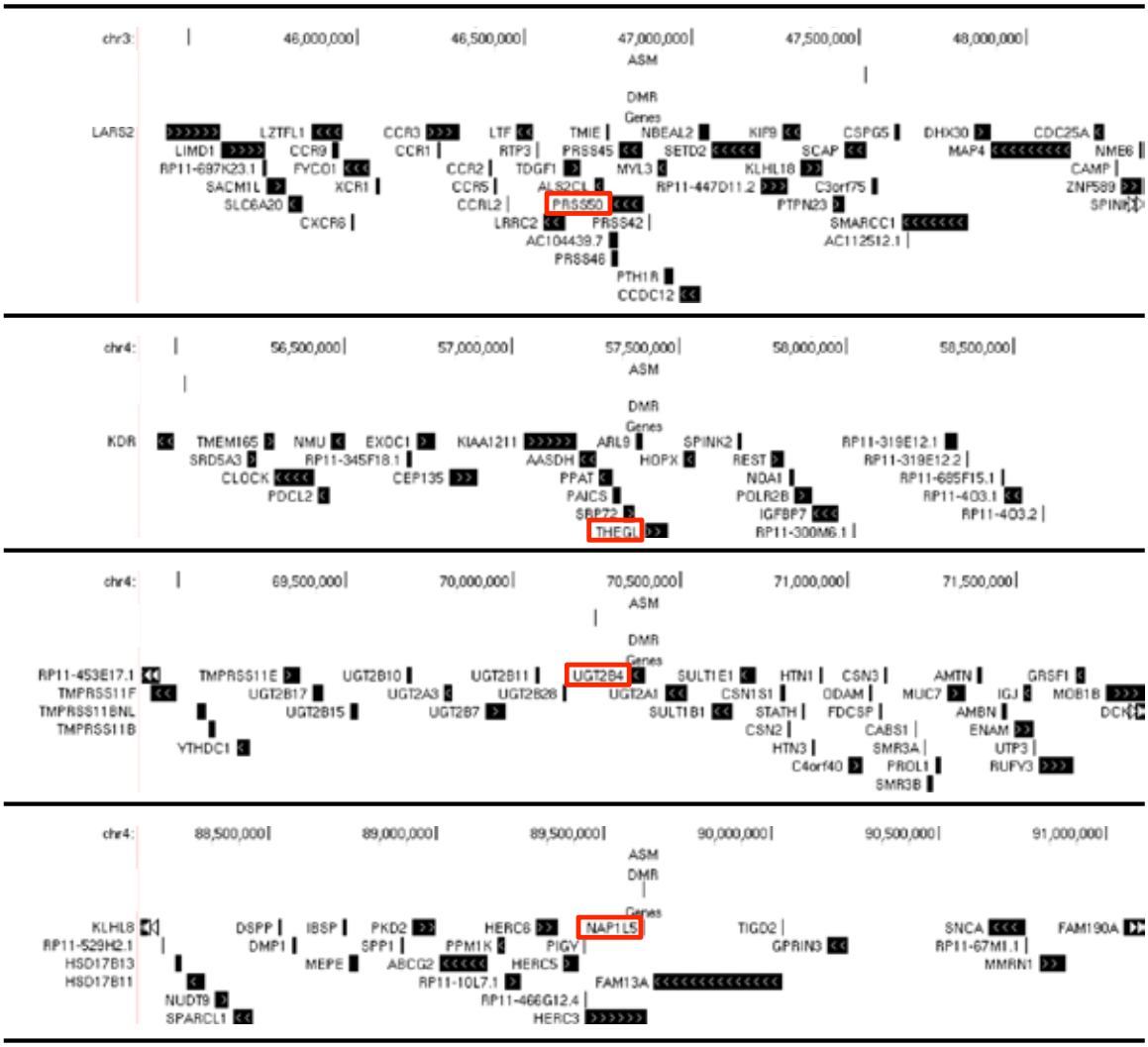
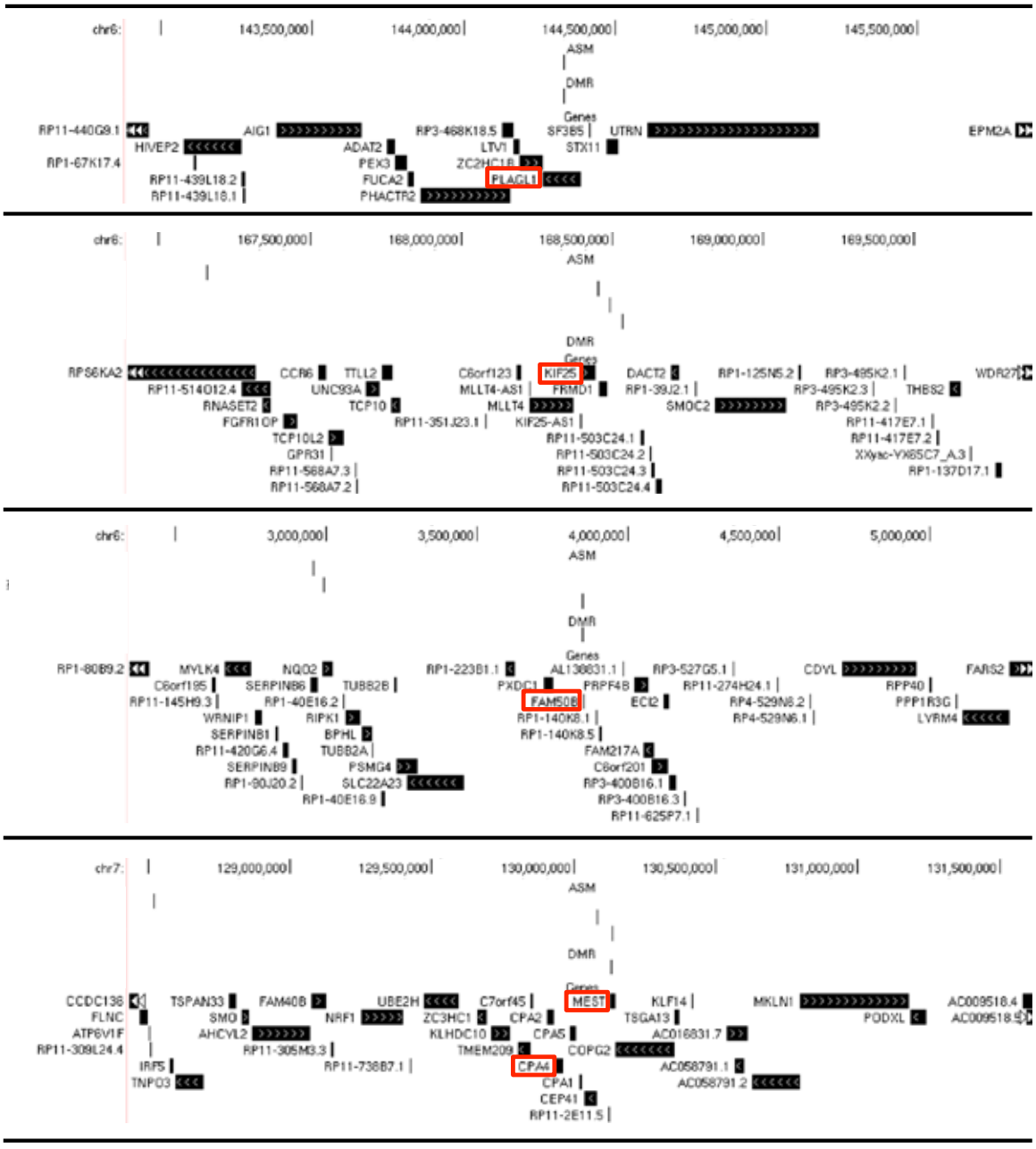
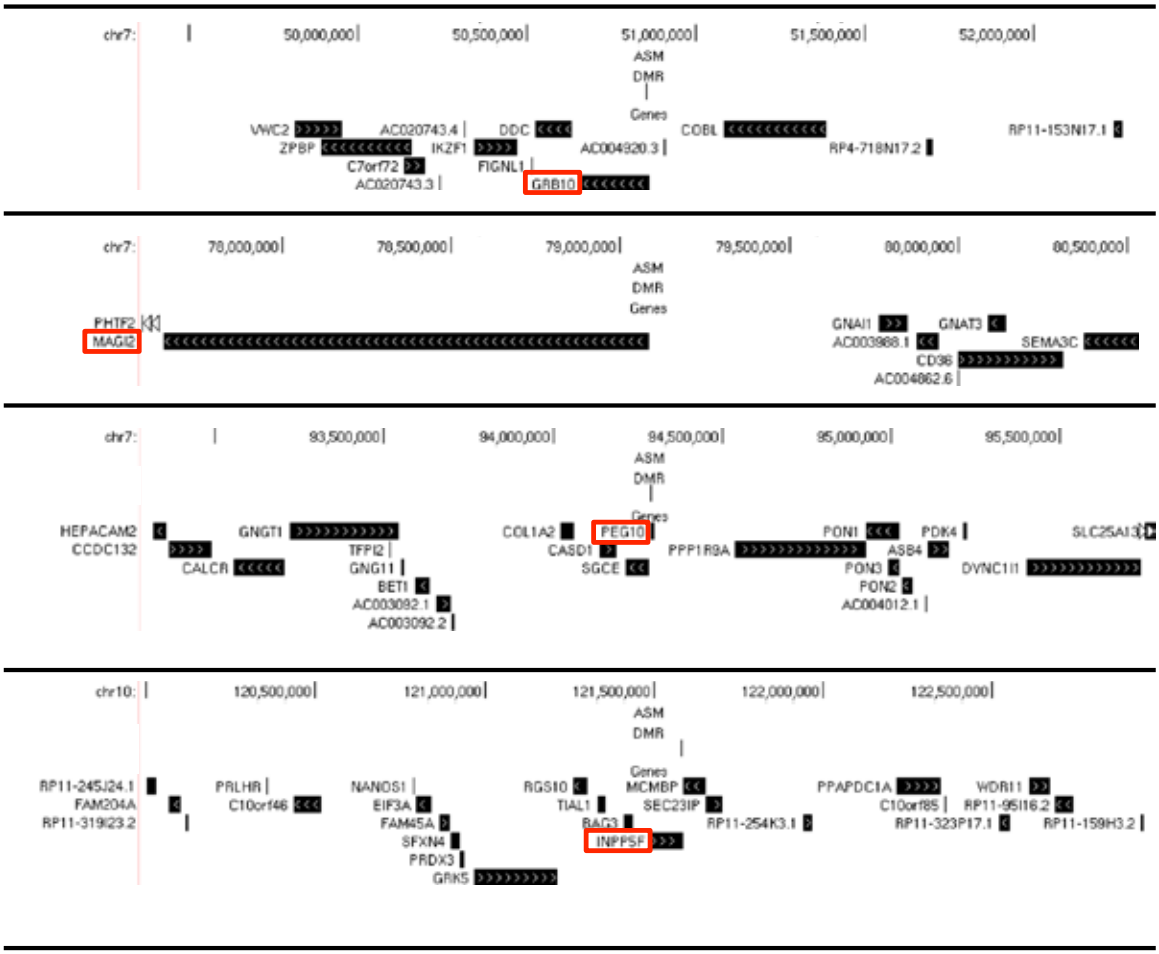


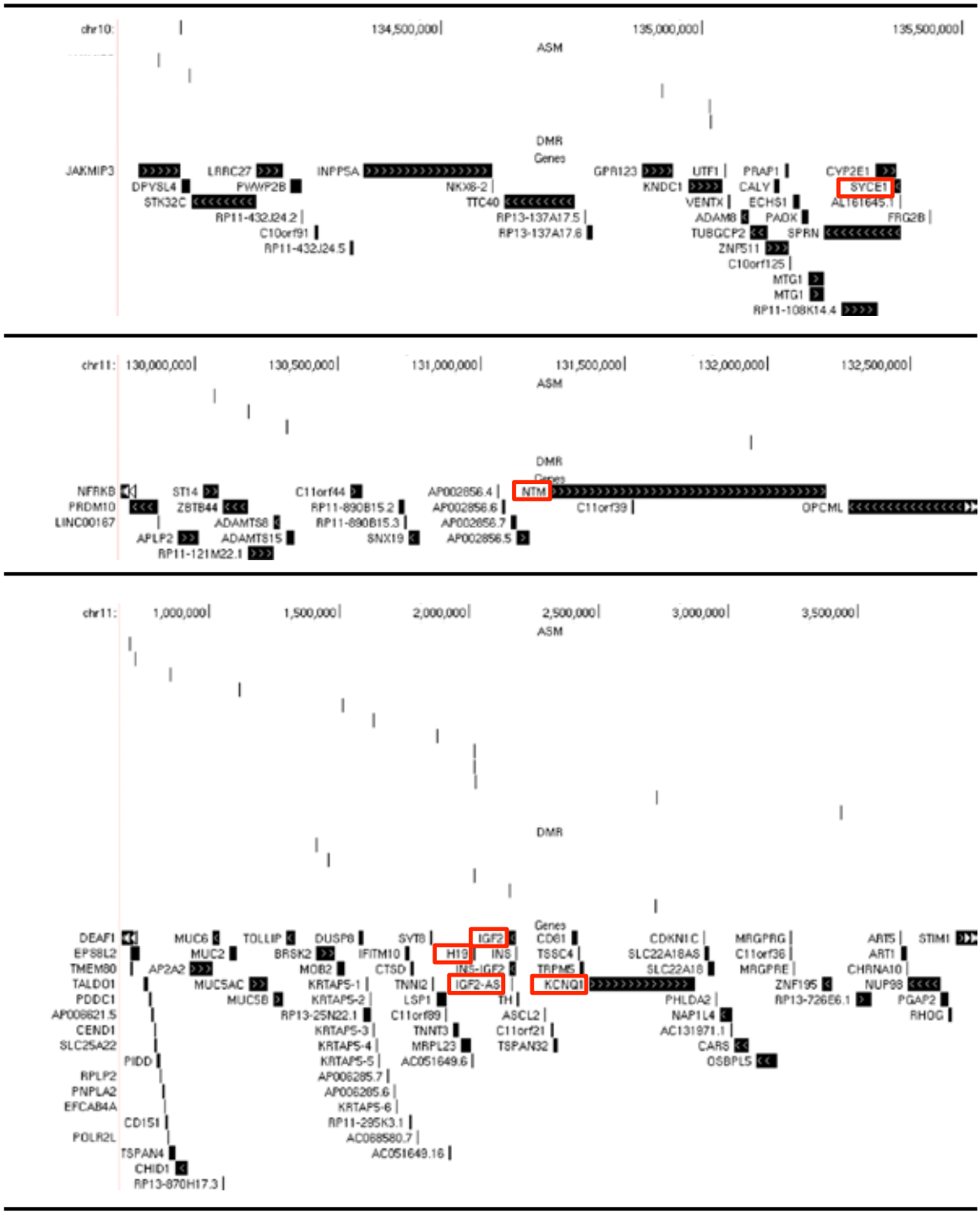
Fig. S6.

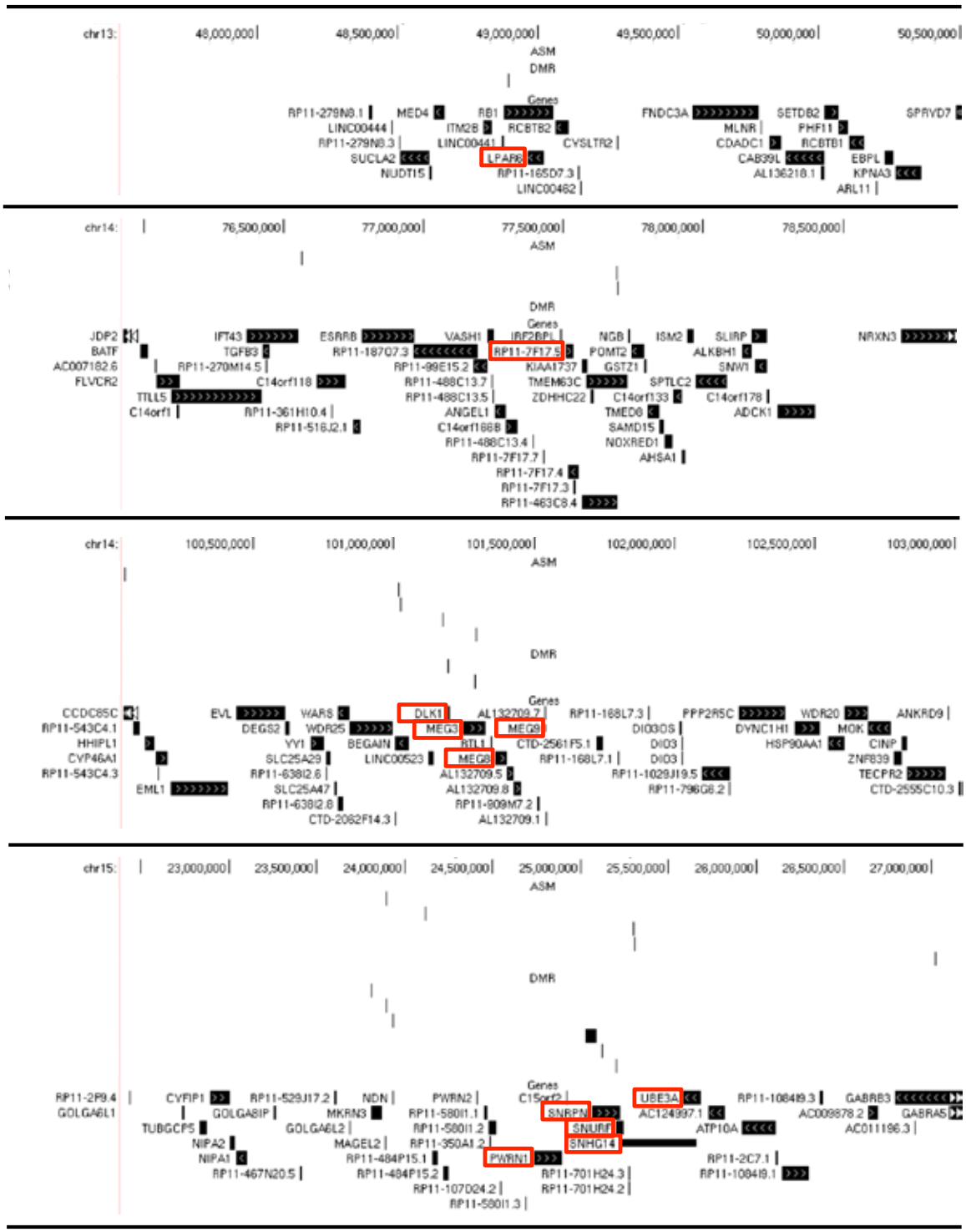
Genomic regions of all imprinted loci, with the 47 genes detected as imprinted marked in red. DMR denotes differentially methylated regions from (Court et al. 2014) and ASM denotes regions with allele-specific methylation according to (Kuleshov et al. 2014). Note that absence of data of DMR or ASM regions does not necessarily imply their full absence due to potential tissue-specific effects. (Continues on the next 7 pages)

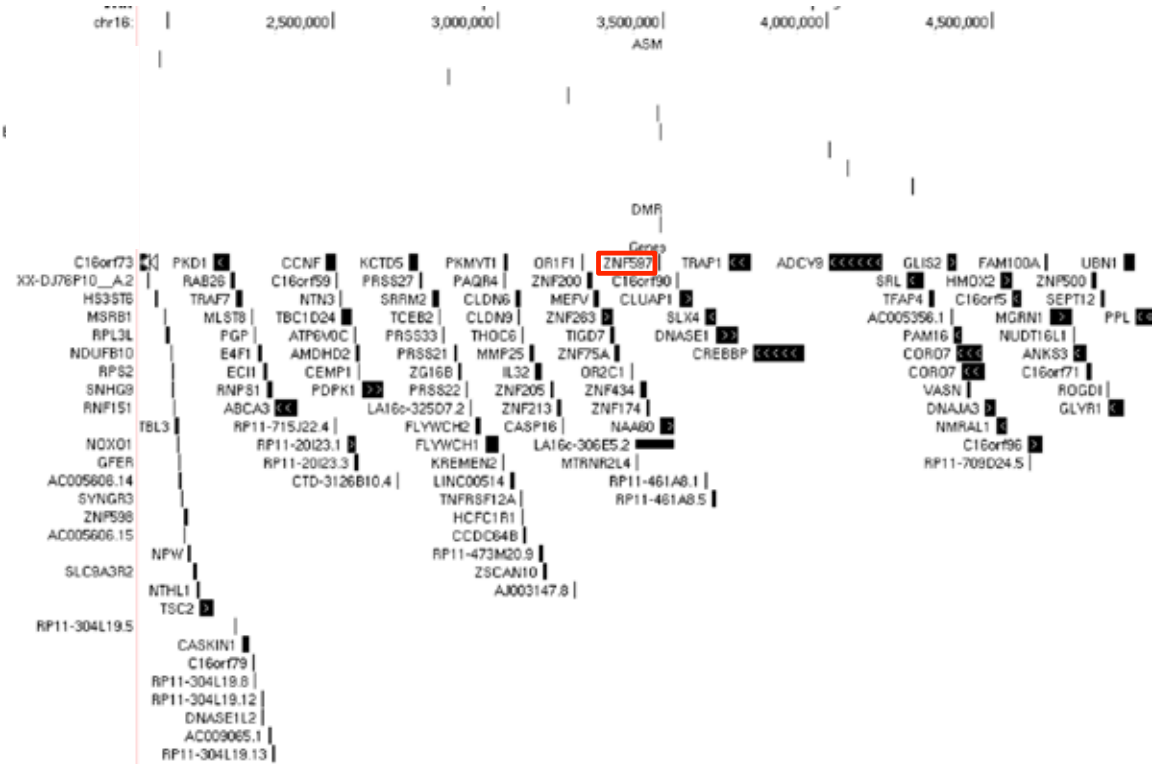

chr3:
46,000,000
46,500,000
47,000,000
47,500,000
48,000,000
ASM
DMR
Genes
LARS2
LZTFL1
CCR3
LTF
TMIE
NBEAL2
KIF9
CSPG5
DHX30
CDC25A
LIMD1
CCR9
CCR1
RTP3
PRSS45
SETD2
SCAP
MAP4
NME6
RP11-697K23.1
FYCO1
CCR2
TDGF1
MYL3
KLHL18
CAMP
SACM1L
XCR1
CCR5
ALS2CL
RP11-447D11.2
C3orf75
ZNF589
SLC6A20
CCRL2
PRSS50
PTPN23
SPINK
CXCR6
LRRC2
PRSS42
SMARCC1
AC104439.7
AC112512.1
PRSS46
PTH1R
CCDC12
chr4:
56,500,000
57,000,000
57,500,000
58,000,000
58,500,000
ASM
DMR
Genes
KDR
TMEM165
NMU
EXOC1
KIAA1211
ARL9
SPINK2
RP11-319E12.1
SRD5A3
RP11-345F18.1
AASDH
HOPX
REST
RP11-319E12.2
CLOCK
CEP135
PPAT
NOA1
RP11-685F15.1
PDCL2
PAICS
POLR2B
RP11-403.1
SRP72
IGFBP7
RP11-403.2
THEGL
RP11-300M6.1
chr4:
69,500,000
70,000,000
70,500,000
71,000,000
71,500,000
ASM
DMR
Genes
RP11-453E17.1
TMPRSS11E
UGT2B10
UGT2B11
UGT2B4
SULT1E1
HTN1
CSN3
AMTN
GRSF1
TMPRSS11F
UGT2B17
UGT2A3
UGT2B28
UGT2A1
CSN1S1
ODAM
MUC7
IGJ
MOB1B
TMPRSS11BNL
UGT2B15
UGT2B7
SULT1B1
STATH
FDCSP
AMBN
DCK
TMPRSS11B
CSN2
CABS1
ENAM
YTHDC1
HTN3
SMR3A
UTP3
C4orf40
PROL1
RUFY3
SMR3B
chr4:
88,500,000
89,000,000
89,500,000
90,000,000
90,500,000
91,000,000
ASM
DMR
Genes
KLHL8
DSPP
IBSP
PKD2
HERC6
NAP1L5
TIGD2
SNCA
FAM190A
RP11-529H2.1
DMP1
SPP1
PPM1K
PIGV
GPRIN3
RP11-67M1.1
HSD17B13
MEPE
ABCG2
HERC5
MMRN1
HSD17B11
RP11-10L7.1
FAM13A
NUDT9
RP11-466G12.4
SPARCL1
HERC3

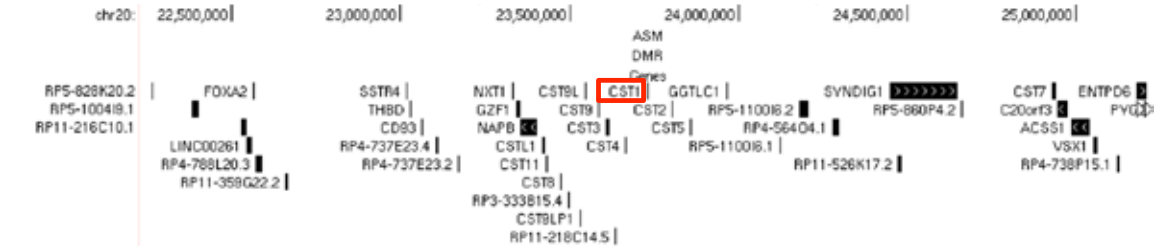

chr6:
143,500,000
144,000,000
144,500,000
145,000,000
145,500,000
ASM
DMR
Genes
RP11-440G9.1
AIG1
RP3-468K18.5
SF3B5
UTRN
EPM2A
HIVEP2
ADAT2
LTV1
STX11
RP1-67K17.4
PEX3
ZC2HC1B
RP11-439L18.2
FUCA2
PLAGL1
RP11-439L18.1
PHACTR2
chr6:
167,500,000
168,000,000
168,500,000
169,000,000
169,500,000
ASM
DMR
Genes
RPS6KA2
CCR6
TTLL2
C6orf123
KIF25
DACT2
RP1-125N5.2
RP3-495K2.1
WDR27
RP11-514O12.4
UNC93A
MLLT4-AS1
FRMD1
RP1-39J2.1
RP3-495K2.3
THBS2
RNASET2
TCP10
MLLT4
SMOC2
RP3-495K2.2
FGFR1OP
RP11-351J23.1
KIF25-AS1
RP11-417E7.1
TCP10L2
RP11-503C24.1
RP11-417E7.2
GPR31
RP11-503C24.2
XXyac-YX65C7_A.3
RP11-568A7.3
RP11-503C24.3
RP1-137D17.1
RP11-568A7.2
RP11-503C24.4
chr6:
3,000,000
3,500,000
4,000,000
4,500,000
5,000,000
ASM
DMR
Genes
RP1-80B9.2
MYLK4
NQO2
RP1-223B1.1
AL136831.1
RP3-527G5.1
CDYL
FARS2
C6orf195
SERPINB6
TUBB2B
PXDC1
PRPF4B
RP11-274H24.1
RPP40
RP11-145H9.3
RP1-40E16.2
FAM50B
ECI2
RP4-529N6.2
PPP1R3G
WRNIP1
RIPK1
RP1-140K8.1
RP4-529N6.1
LYRM4
SERPINB1
BPHL
RP1-140K8.5
RP11-420G6.4
TUBB2A
FAM217A
SERPINB9
PSMG4
C6orf201
RP1-80J20.2
SLC22A23
RP3-400B16.1
RP1-40E16.9
RP3-400B16.3
RP11-625P7.1
chr7:
129,000,000
129,500,000
130,000,000
130,500,000
131,000,000
131,500,000
ASM
DMR
Genes
CCDC136
TSPAN33
FAM40B
UBE2H
C7orf45
MEST
KLF14
MKLN1
AC009518.4
FLNC
SMO
NRF1
ZC3HC1
CPA2
TSGA13
PODXL
AC009518.5
ATP6V1F
AHCYL2
KLHDC10
CPA5
AC016831.7
RP11-309L24.4
RP11-305M3.3
TMEM209
COPG2
IRF5
RP11-736B7.1
CPA4
AC058791.1
TNPO3
CPA1
AC058791.2
CEP41
RP11-2E11.5

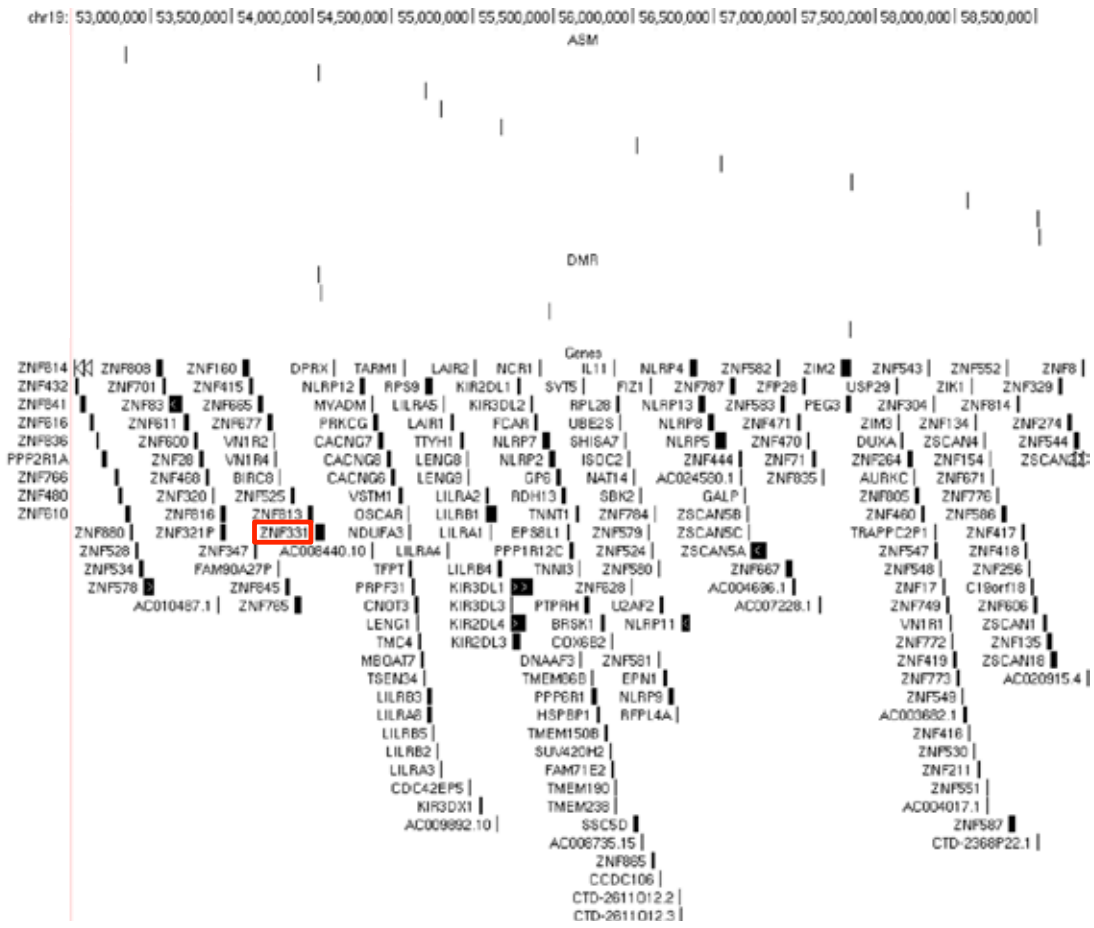

chr7:
50,000,000
50,500,000
51,000,000
51,500,000
52,000,000
ASM
DMR
Genes
VWC2
AC020743.4
DDC
COBL
RP11-153N17.1
ZPBP
IKZF1
AC004920.3
RP4-718N17.2
C7orf72
FIGNL1
AC020743.3
GRB10
chr7:
78,000,000
78,500,000
79,000,000
79,500,000
80,000,000
80,500,000
ASM
DMR
Genes
PHTF2
MAGI2
GNAI1
GNAT3
AC003988.1
SEMA3C
CD36
AC004862.6
chr7:
93,500,000
94,000,000
94,500,000
95,000,000
95,500,000
ASM
DMR
Genes
HEPACAM2
CCDC132
GNGT1
TFPI2
CALCR
GNG11
BET1
AC003092.1
AC003092.2
COL1A2
PEG10
CASD1
SGCE
PPP1R9A
PON1
PDK4
SLC25A13
ASB4
PON3
PON2
AC004012.1
DYNC1I1
chr10:
120,500,000
121,000,000
121,500,000
122,000,000
122,500,000
ASM
DMR
Genes
RP11-245J24.1
FAM204A
RP11-319I23.2
PRLHR
C10orf46
NANOS1
EIF3A
FAM45A
SFXN4
PRDX3
GRK5
RGS10
TIAL1
RAG3
INPPSF
MCMBP
SEC23IP
RP11-254K3.1
PPAPDC1A
C10orf85
RP11-323P17.1
WDR11
RP11-95I16.2
RP11-159H3.2

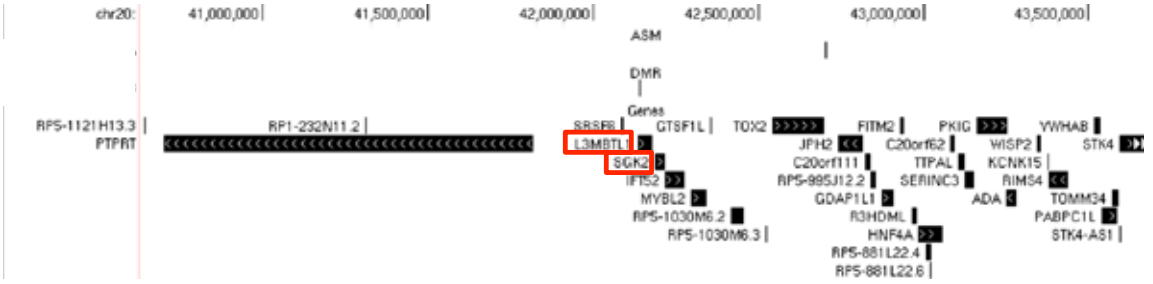

chr10:
134,500,000
135,000,000
135,500,000
ASM
DMR
Genes
JAKMIP3
LRRC27
INPPSA
GPR123
UTF1
PRAP1
CYP2E1
DPYSL4
PWWP2B
NKX6-2
KNDC1
CALY
SYCE1
STK32C
TTC40
VENTX
ECHS1
AL161645.1
RP11-432J24.2
RP13-137A17.5
ADAM8
PADK
FRG2B
C10orf91
RP13-137A17.6
TUBGCP2
SPRN
RP11-432J24.5
ZNF511
C10orf125
MTG1
MTG1
RP11-108K14.4
chr11:
130,000,000
130,500,000
131,000,000
131,500,000
132,000,000
132,500,000
ASM
DMR
Genes
NFRKB
ST14
C11orf44
AP002856.4
NTM
PRDM10
ZBTB44
RP11-890B15.2
AP002856.6
C11orf39
OPCML
LINC00167
ADAMTS8
RP11-890B15.3
AP002856.7
APLP2
ADAMTS15
SNX19
AP002856.5
RP11-121M22.1
chr11:
1,000,000
1,500,000
2,000,000
2,500,000
3,000,000
3,500,000
ASM
DMR
Genes
DEAF1
MUC6
TOLLIP
DUSP8
SYT8
IGF2
CD81
CDKN1C
MRGPRG
ART5
STIM1
EPS8L2
MUC2
BRSK2
IFITM10
H19
INS
TSSC4
SLC22A18AS
C11orf36
ART1
TMEM80
AP2A2
MOB2
CTSD
INS-IGF2
TRPM5
SLC22A18
MRGPRE
CHRNA10
TALDO1
MUC5AC
KRTAP5-1
TNNI2
IGF2-AS
KCNQ1
ZNF195
NUP98
PDDC1
MUC5B
KRTAP5-2
LSP1
TH
PHLDA2
RP13-726E6.1
PGAP2
AP006621.5
RP13-25N22.1
C11orf89
ASCL2
NAP1L4
RHOG
CEND1
KRTAP5-3
TNNT3
C11orf21
AC131971.1
SLC25A22
KRTAP5-4
MRPL23
TSPAN32
CARS
PIDD
KRTAP5-5
AC051649.6
OSBPL5
RPLP2
AP006285.7
PNPLA2
AP006285.6
EFCAB4A
KRTAP5-6
RP11-295K3.1
CD151
AC068580.7
POLR2L
AC051649.16
TSPAN4
CHID1
RP13-870H17.3


chr13:
48,000,000
48,500,000
49,000,000
49,500,000
50,000,000
50,500,000
ASM
DMR
Genes
RP11-279N8.1
MED4
RB1
FNDC3A
SETDB2
SPRYD7
LINC00444
ITM2B
RCBTB2
MLNR
PHF11
RP11-279N8.3
LINC00441
CYSLTR2
CDADC1
RCBTB1
SUCLA2
LPAR6
CAB39L
EBPL
NUDT15
RP11-165D7.3
AL136218.1
KPNA3
LINC00462
ARL11
chr14:
76,500,000
77,000,000
77,500,000
78,000,000
78,500,000
ASM
DMR
Genes
JDP2
IFT43
ESRRB
VASH1
IRF2BPL
NGB
ISM2
SLIRP
NRXN3
BATF
TGFB3
RP11-187O7.3
RP11-7F17.5
POMT2
ALKBH1
AC007182.6
RP11-270M14.5
RP11-99E15.2
KIAA1737
GSTZ1
SNW1
FLVCR2
C14orf118
RP11-488C13.7
TMEM63C
SPTLC2
TTLL5
RP11-488C13.5
ZDHHC22
C14orf133
C14orf178
C14orf1
RP11-361H10.4
ANGEL1
TMED6
ADCK1
RP11-516J2.1
C14orf166B
SAMD15
RP11-488C13.4
NOXRED1
RP11-7F17.7
AHSA1
RP11-7F17.4
RP11-7F17.3
RP11-463C8.4
chr14:
100,500,000
101,000,000
101,500,000
102,000,000
102,500,000
103,000,000
ASM
DMR
Genes
CCDC85C
EVL
WARS
DLK1
AL132709.7
RP11-168L7.3
PPP2R5C
WDR20
ANKRD9
RP11-543C4.1
DEGS2
WDR25
MEG3
MEG9
DIO3OS
DVNC1H1
MOK
HHIPL1
YY1
BEGAIN
RTL1
CTD-2561F5.1
DIO3
HSP90AA1
CINP
CYP46A1
SLC25A29
LINC00523
MEG8
RP11-168L7.1
DIO3
ZNF839
RP11-543C4.3
RP11-638I2.6
AL132709.5
RP11-1029J19.5
TECPR2
EML1
SLC25A47
AL132709.8
RP11-796G8.2
CTD-2555C10.3
RP11-638I2.8
RP11-909M7.2
CTD-2062F14.3
AL132709.1
chr15:
23,000,000
23,500,000
24,000,000
24,500,000
25,000,000
25,500,000
26,000,000
26,500,000
27,000,000
ASM
DMR
Genes
RP11-2P9.4
CYFIP1
RP11-529J17.2
NDN
PWRN2
C15orf2
UBE3A
RP11-108M9.3
GABRB3
GOLGA6L1
GOLGA8IP
MKRN3
RP11-580I1.1
SNRPN
AC124997.1
AC009878.2
GABRA5
TUBGCP5
GOLGA6L2
RP11-580I1.2
SNURF
ATP10A
AC011196.3
NIPA2
MAGEL2
RP11-350A1.2
SNHG14
NIPA1
RP11-484P15.1
PWRN1
RP11-2C7.1
RP11-467N20.5
RP11-484P15.2
RP11-701H24.3
RP11-108M9.1
RP11-107D24.2
RP11-701H24.2
RP11-580I1.3


chr16:
2,500,000
3,000,000
3,500,000
4,000,000
4,500,000
ASM
DMR
Genes
C16orf73
XX-DJ76P10__A.2
HS3ST6
MSRB1
RPL3L
NDUFB10
RPS2
SNHG9
RNF151
NOXO1
GFER
AC005606.14
SYNGR3
ZNF598
AC005606.15
SLC9A3R2
RP11-304L19.5
PKD1
RAB26
TRAF7
MLST8
PGP
E4F1
ECI1
RNPS1
ABCA3
TBL3
CCNF
C16orf59
NTN3
TBC1D24
ATP6V0C
AMDHD2
CEMP1
PDPK1
RP11-715J22.4
RP11-20I23.1
RP11-20I23.3
CTD-3126B10.4
KCTD5
PRSS27
SRRM2
TCEB2
PRSS33
PRSS21
ZG16B
PRSS22
LA16c-325D7.2
FLYWCH2
FLYWCH1
KREMEN2
LINC00514
TNFRSF12A
HCFC1R1
CCDC64B
RP11-473M20.9
ZSCAN10
AJ003147.8
PKMYT1
PAQR4
CLDN6
CLDN9
THOC6
PRSS23
MMP25
IL32
ZNF205
ZNF213
CASP16
LA16c-306E5.2
MTRNR2L4
OR1F1
ZNF200
MEFV
ZNF263
TIGD7
ZNF75A
OR2C1
ZNF434
ZNF174
NAA60
ZNF597
C16orf90
CLUAP1
SLX4
DNASE1
CREBBP
TRAP1
RP11-461A8.1
RP11-461A8.5
ADCY9
GLIS2
FAM100A
UBN1
SRL
HMOX2
ZNF500
TFAP4
C16orf5
SEPT12
AC005356.1
MGRN1
PPL
PAM16
NUDT16L1
CORO7
ANKS3
CORO7
C16orf71
VASN
ROGDI
DNAJA3
GLYR1
NMRAL1
C16orf96
RP11-709D24.5
NPW
NTHL1
TSC2
CASKIN1
C16orf79
RP11-304L19.8
RP11-304L19.12
DNASE1L2
AC009065.1
RP11-304L19.13


chr20:
22,500,000
23,000,000
23,500,000
24,000,000
24,500,000
25,000,000
ASM
DMR
Genes
RP5-828K20.2
RP5-1004I9.1
RP11-216C10.1
FOXA2
LINC00261
RP4-788L20.3
RP11-359G22.2
SSTR4
THBD
CD93
RP4-737E23.4
RP4-737E23.2
NXT1
GZF1
NAPB
CSTL1
CST11
CST8
RP3-333B15.4
CST8LP1
RP11-218C14.5
CST9L
CST9
CST3
CST4
CST1
CST2
CST5
GGTLC1
RP5-1100I6.2
RP4-564O4.1
RP5-1100I6.1
SYNDIG1
RP5-880P4.2
RP11-526K17.2
CST7
C20orf3
ACSS1
VSX1
RP4-738P15.1
ENTPD6
PYGB


chr19: 53,000,000 53,500,000 54,000,000 54,500,000 55,000,000 55,500,000 56,000,000 56,500,000 57,000,000 57,500,000 58,000,000 58,500,000
ASM
DMR
Genes
ZNF814 ZNF432 ZNF841 ZNF616 ZNF836 PPP2R1A ZNF766 ZNF480 ZNF610
ZNF808 ZNF701 ZNF83 ZNF611 ZNF600 ZNF28 ZNF468 ZNF320 ZNF616 ZNF880 ZNF321P ZNF528 ZNF534 ZNF578 AC010487.1
ZNF160 ZNF415 ZNF685 ZNF677 VN1R2 VN1R4 BIRC8 ZNF525 ZNF813 ZNF331 ZNF347 FAM90A27P ZNF845 ZNF765
DPRX MVADM PRKCG CACNG7 CACNG8 CACNG6 VSTM1 OSCAR NDUFA3 AC008440.10 TFPT PRPF31 CNOT3 LENG1 TMC4 MBOAT7 TSEN34 LILRB3 LILRA6 LILRB5 LILRB2 LILRA3 CDC42EP5 KIR3DX1 AC009892.10
TARM1 NLRP12 RPS9 LAIR1 LAIR2 LENG8 LENG9 LILRA2 LILRB1 LILRA1 LILRA4 LILRB4 KIR3DL1 KIR3DL3 KIR2DL4 KIR2DL3
NCR1 KIR2DL1 KIR3DL2 FCAR NLRP7 NLRP2 GP6 RDH13 TNNT1 EPS8L1 PPP1R12C TNNI3 ZNF628 PTPRH BRSK1 COX6B2 DNAAF3 TMEM86B PPP6R1 HSPBP1 TMEM150B SUV420H2 FAM71E2 TMEM190 TMEM238 SSC5D AC008735.15 ZNF885 CCDC106 CTD-2611O12.2 CTD-2611O12.3
SYT5 NLRP7 SHISA7 ISOC2 NAT14 SBK2 ZNF784 ZNF579 EPN1 NLRP9 RFPL4A ZNF581 ZNF524 ZNF580 U2AF2
IL11 FIZ1 RPL28 UBE2S NLRP8 NLRP5 ZNF444 AC024560.1 GALP ZSCAN5B ZSCAN5C ZSCAN5A ZNF667 AC004686.1 AC007228.1
NLRP4 ZNF787 NLRP13 ZNF471 ZNF470 ZNF71 ZNF835
ZNF582 ZFP28 ZNF583 ZNF304 ZNF814 ZIM2 PEG3
ZIM3 DUXA AURKC ZNF805 TRAPPC2P1 ZNF547 ZNF548 ZNF17 ZNF749 VN1R1 ZNF772 ZNF419 ZNF773 ZNF549 AC003682.1 ZNF416 ZNF530 ZNF211 ZNF551 AC004017.1 ZNF587 CTD-2368P22.1
ZNF543 USP29 ZNF134 ZSCAN4 ZNF264 ZNF460 ZNF805 ZNF586 ZNF417 ZNF418 ZNF256 C19orf18 ZNF606 ZSCAN1 ZNF135 ZSCAN16 AC020915.4
ZNF552 ZIK1 ZNF154 ZNF671 ZNF776
ZNF8 ZNF329 ZNF274 ZNF544 ZSCAN


chr20: 41,000,000 41,500,000 42,000,000 42,500,000 43,000,000 43,500,000
ASM
DMR
Genes
RP5-1121H13.3 PTPRT
RP1-232N11.2
SRSF6 L3MBTL1 SGK2 IFT52 MYBL2 RP5-1030M6.2 RP5-1030M6.3
GTSF1L
TOX2
JPH2 C20orf111 RP5-995J12.2 GDAP1L1 R3HDML HNF4A RP5-881L22.4 RP5-881L22.6
FITM2 C20orf62 TTPAL SERINC3 ADA
PKIG WISP2 KCNK15 RIMS4 TOMM34 PABPC1L STK4-AS1
YWHAB STK4

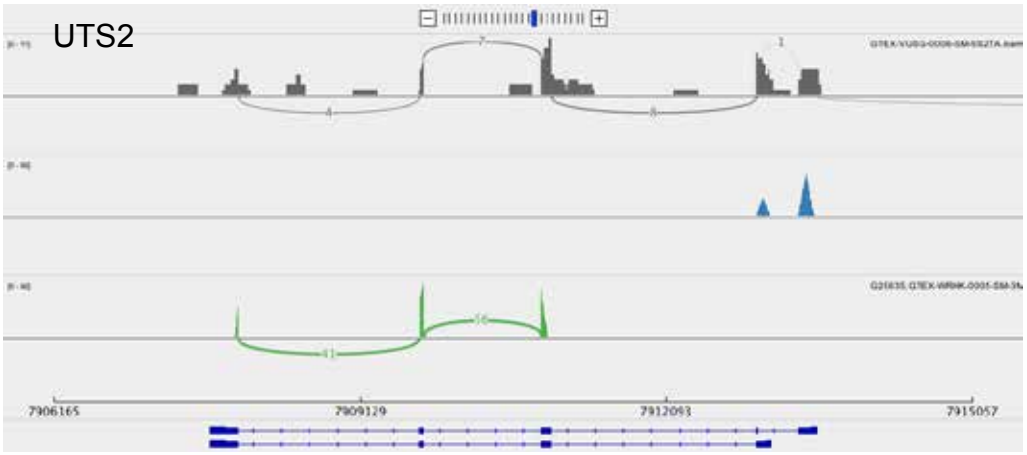
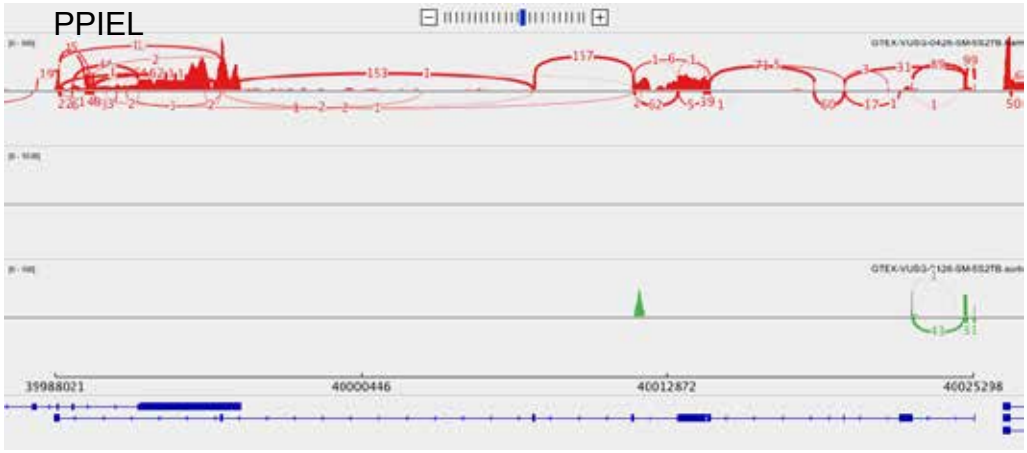
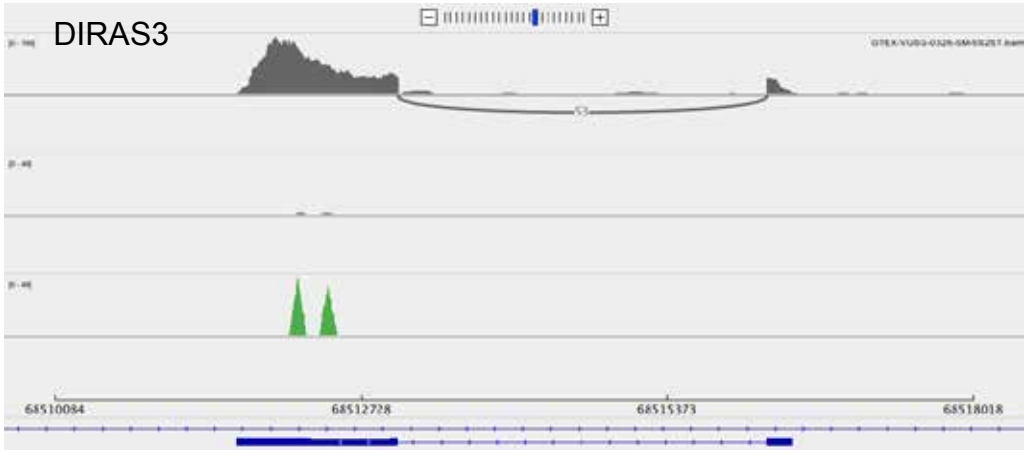


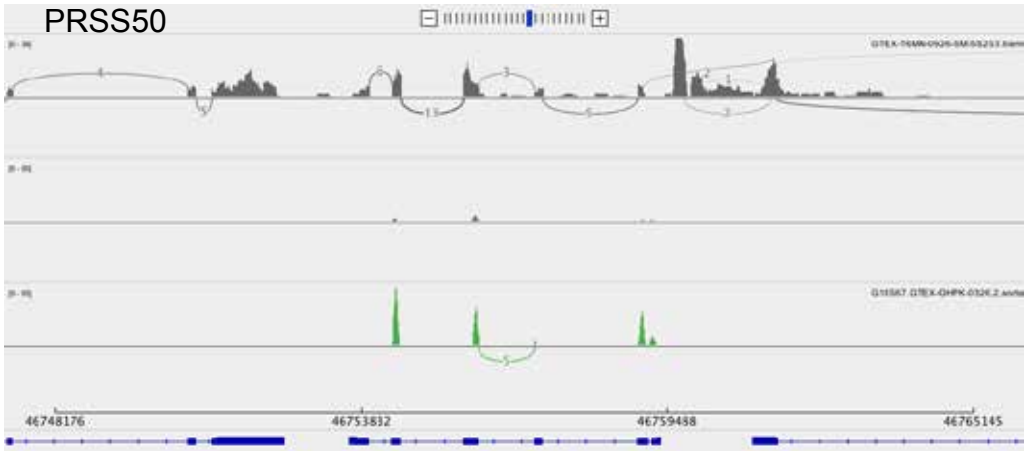
Fig. S7. Transcript structure in the novel or provisional imprinted genes, shown as a sashimi plot for one sample for each gene, based on GTEx data. The top row shows the coverage and splicing (loops with numbers) based on all long-read 2x250 bp RNA-seq reads in the region. The middle row shows reads that carry the reference allele of heterozygous sites, and the bottom shows reads that carry the alternative allele of heterozygous sites. The allelic read data is with 2x250 bp reads of the same sample as the full coverage track, when there was a long-read sample with a heterozygous SNP (these are marked with a red top row). Otherwise allelic data is from 2 x 75 bp reads using different samples from the same tissue for the full coverage track and the allelic read tracks (these are marked with a grey top row). The gene annotation in the bottom is Refseq, which is incorporated in IGV; while the rest of our analysis is based on Gencode v12, these are typically very similar, and it is known that neither is perfect.

The plots are shown for UTS2, PPIEL, DIRAS3, PRSS50, UGT2B4, KIF25, SYCE1, NTM, LPAR6, RP11-7F17.7, MEG9, SNHG14, ZNF331, NLRP2, CST1, and finally as an example of an excluded gene, RP11-701H24.3. The NTM gene is too big to be shown in its entirety, and for some genes, two plots are shown to capture the entire transcripts. Refseq gene annotation is lacking for RP11-7F17.7 and SNHG14, but the data are consistent with the Gencode annotation used in the rest of the analysis. MAGI2 and THEGL are not shown because they are not imprinted in GTEx data, but their transcript structure appears consistent with the annotation. This was not the case for RP11-701H24.3, LA16c-306E5.2, LA16c-306E5.2, and RP11-395B7.2, which were excluded from downstream analysis. The transcript annotations for the last three are too large to be shown as a plot.

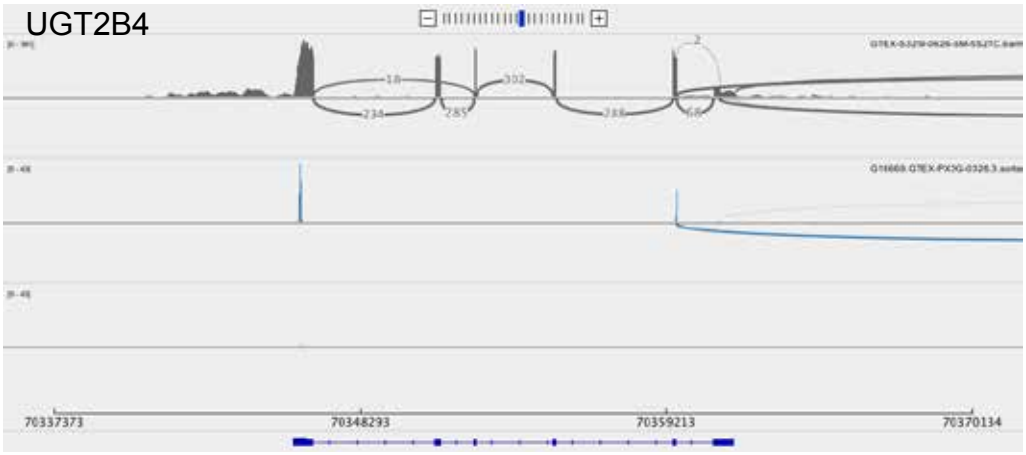
(Continues on the next 6 pages)


PPIEL
39988021
40000446
40012872
40025298

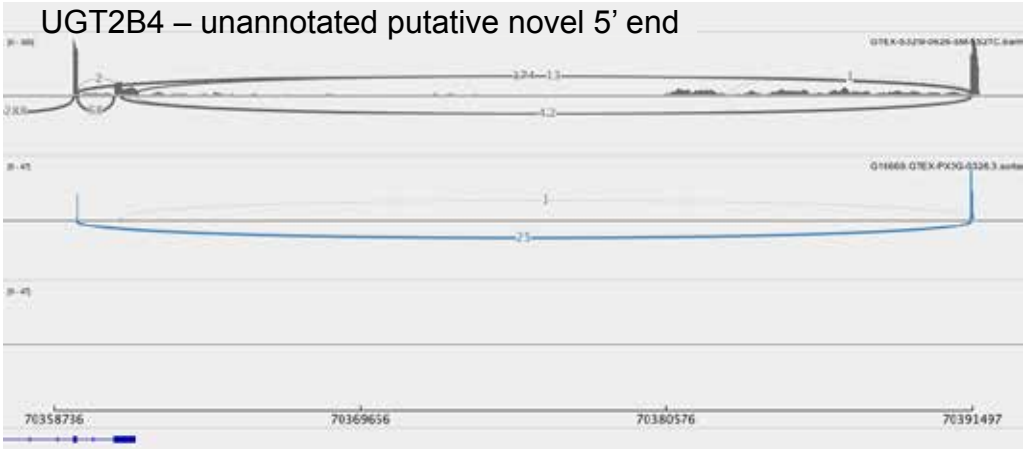

DIRAS3
68510084
68512728
68515373
68518018


PRSS50
46748176
46753832
46759488
46765145

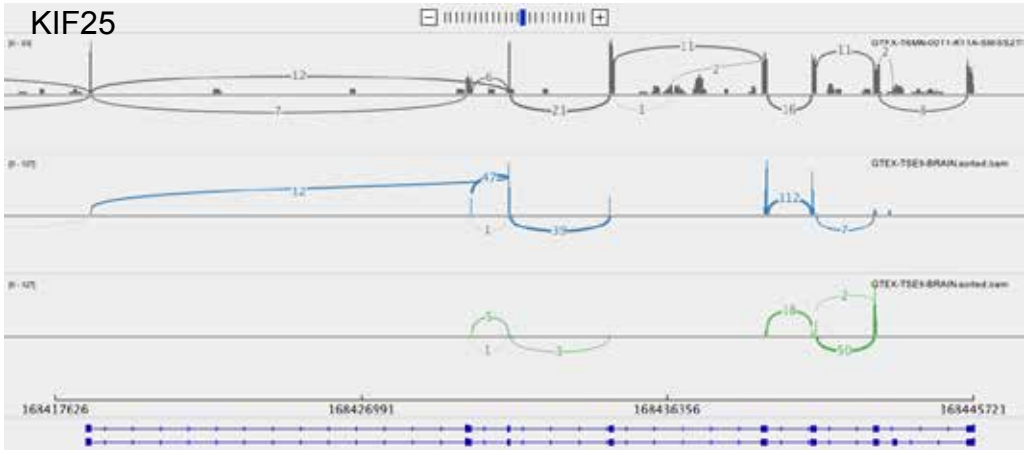
UGT2B4



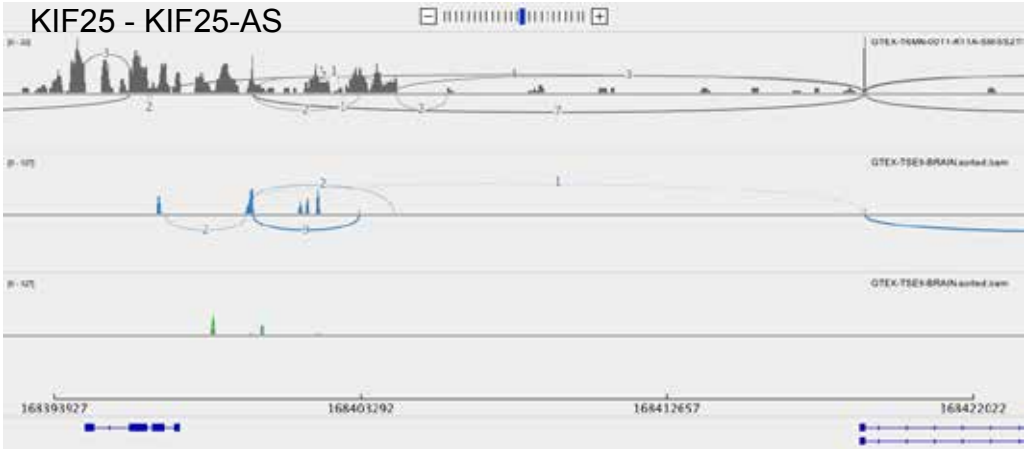
UGT2B4 – unannotated putative novel 5’ end



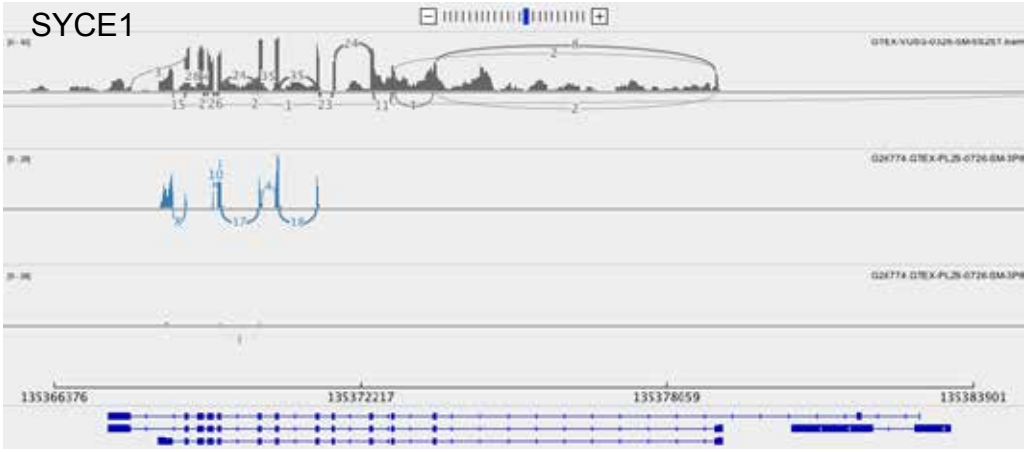
KIF25



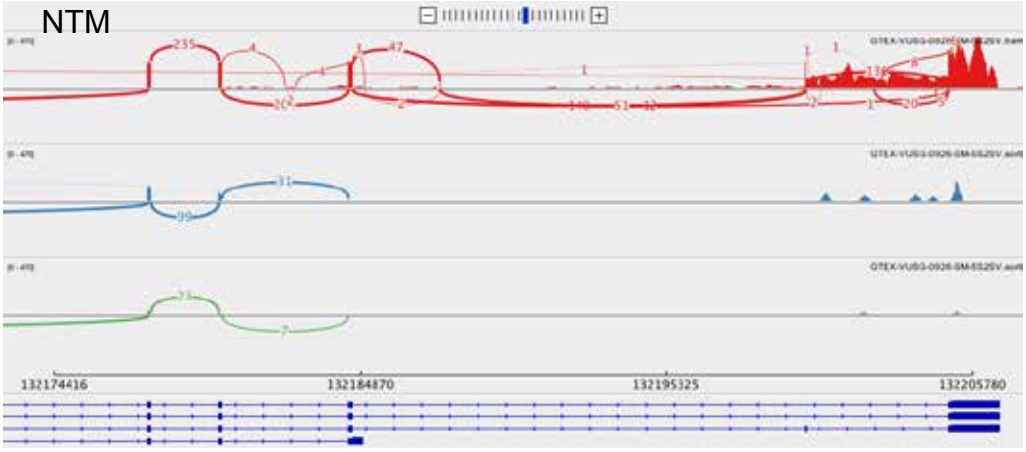
KIF25 - KIF25-AS

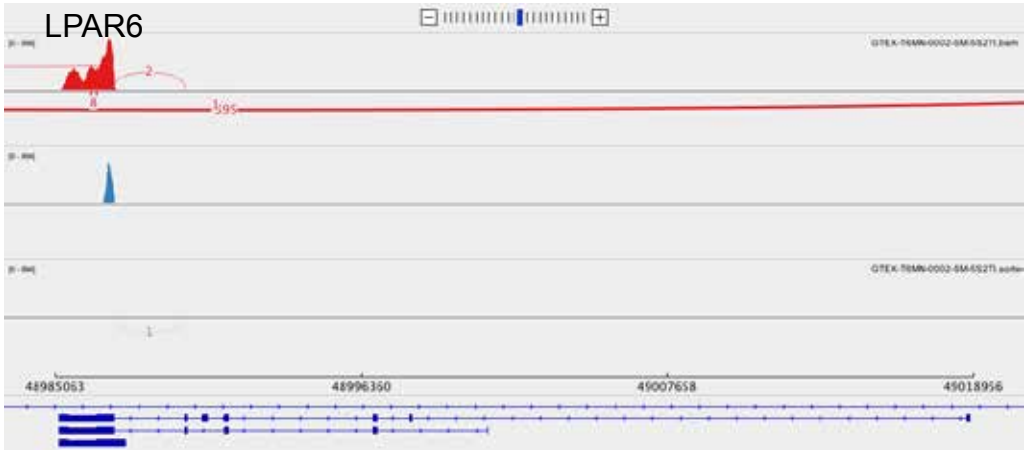


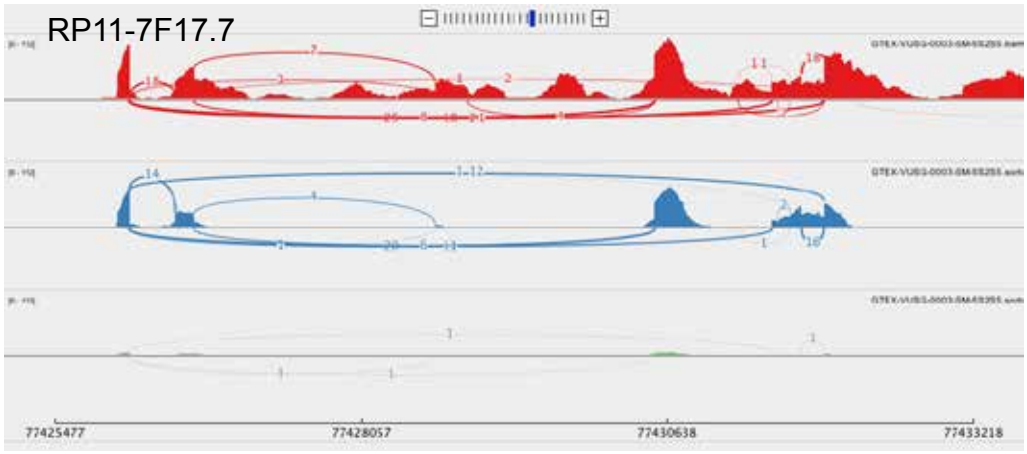
SYCE1

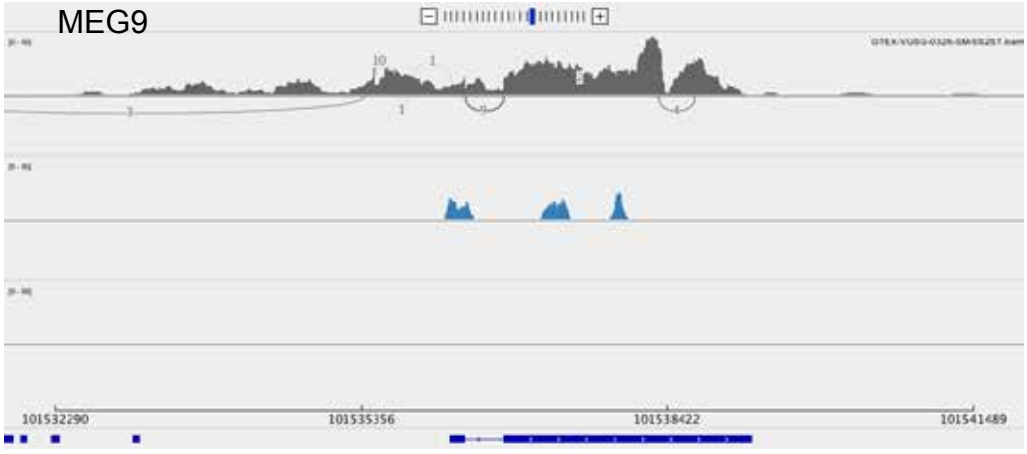


NTM

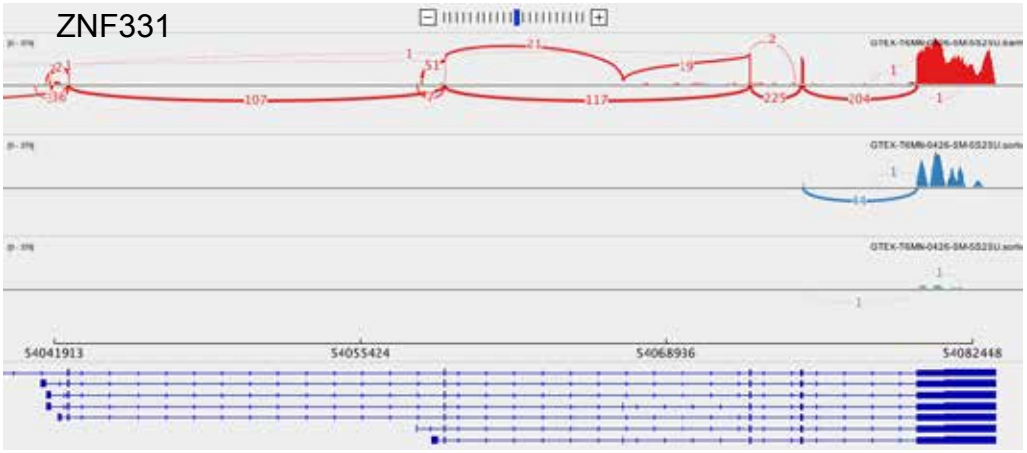


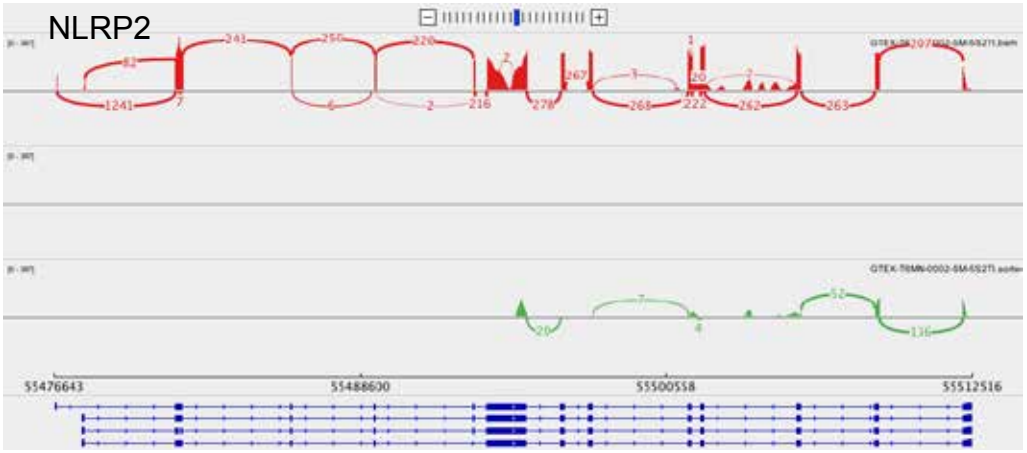

LPAR6
48985063
48996360
49007658
49018956


RP11-7F17.7
77425477
77428057
77430638
77433218

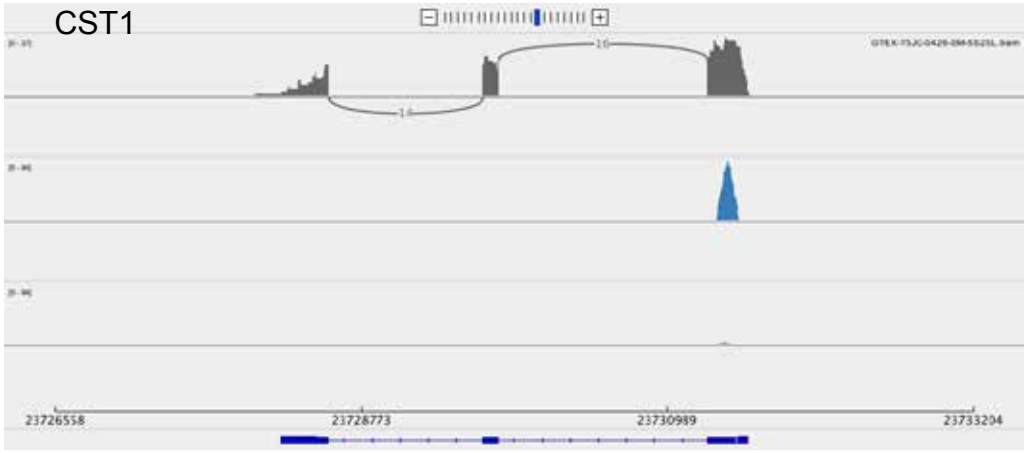

MEG9
101532290
101535356
101538422
101541489


SNHG14
25226809
25239445
25252081
25264717

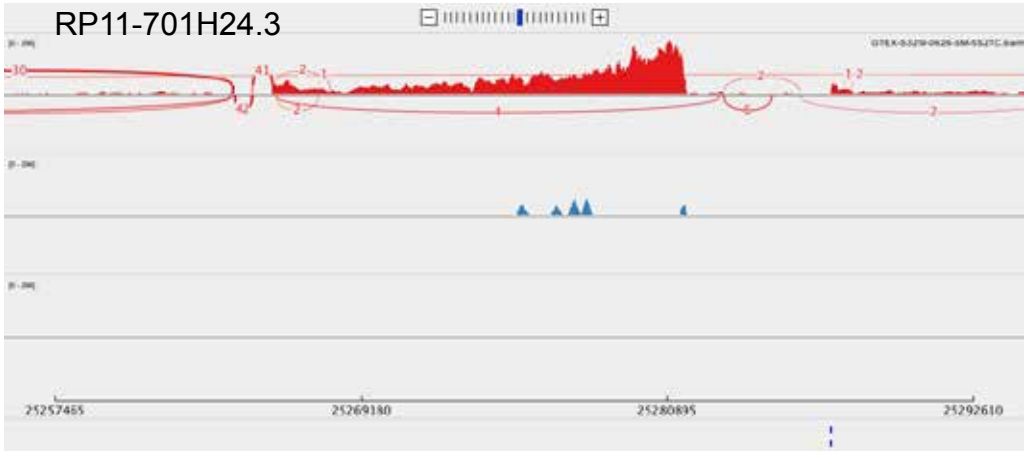

ZNF331
54041913
54055424
54068936
54082448


NLRP2
55476643
55488630
55500558
55512516

CST1



RP11-701H24.3



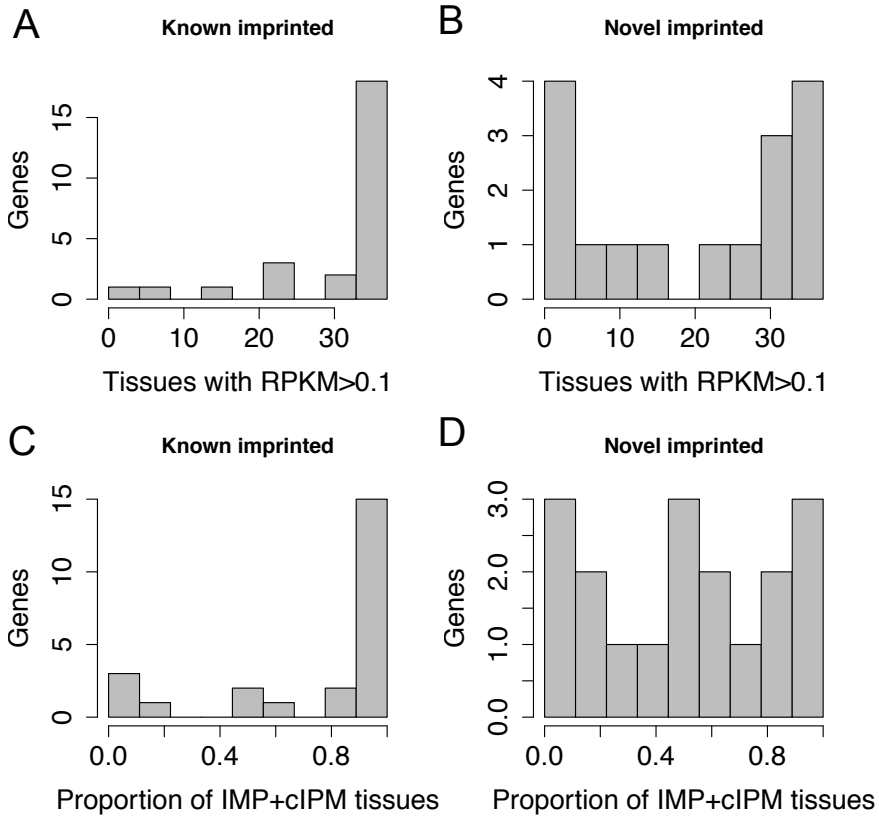


Fig. S8.

Known versus novel imprinted genes. Of the 42 imprinted genes identified in this study, 28 were included in previous catalogs ("known") and 12 were novel. A) and (b) show the number of tissues where the genes are expressed at >0.1 RPKM for known and novel genes, respectively. C) and (d) show the proportion of tissues where the genes are imprinted ((IMP + cIMP) / (IMP + cIMP + BI + cBI)) for known and novel genes, respectively.

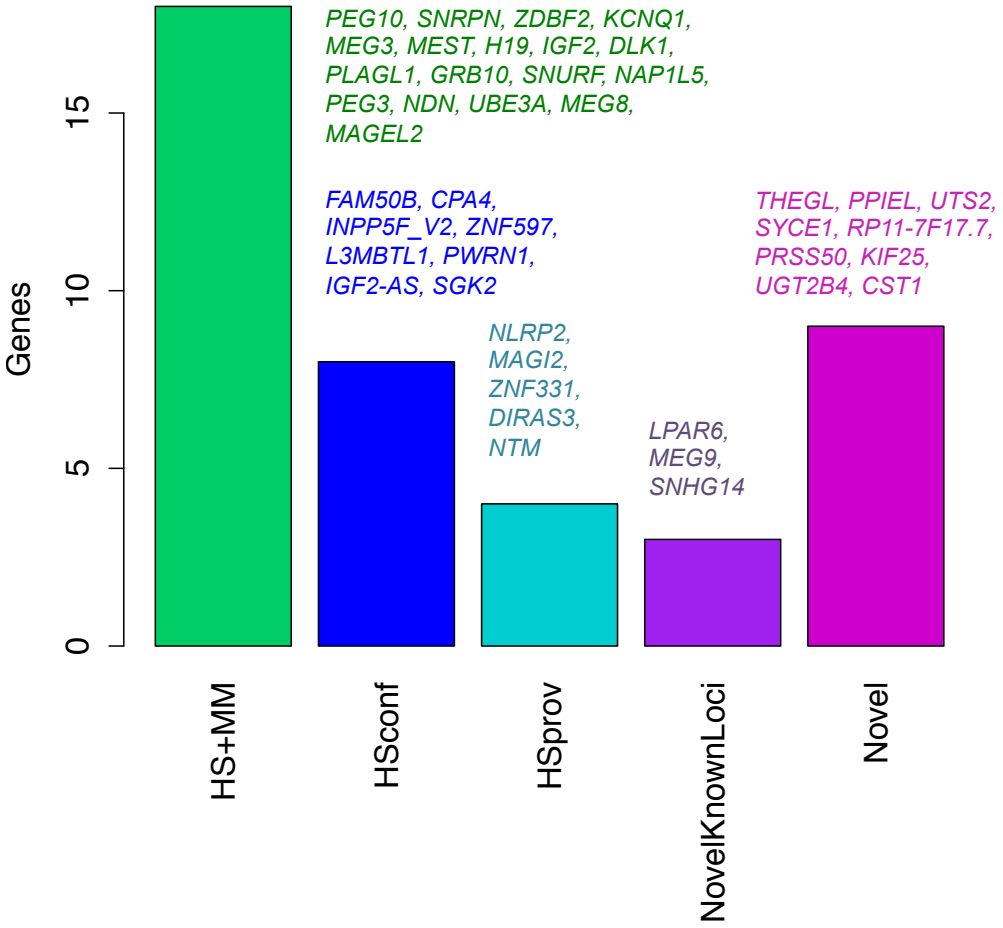


Fig. S9.

Detected imprinted genes classified by previously known imprinting in both humans and the mouse (green), only in humans (blue for confidently identified, cyan for provisional), novel genes that are inside known imprinted loci (red) and fully novel genes (green).

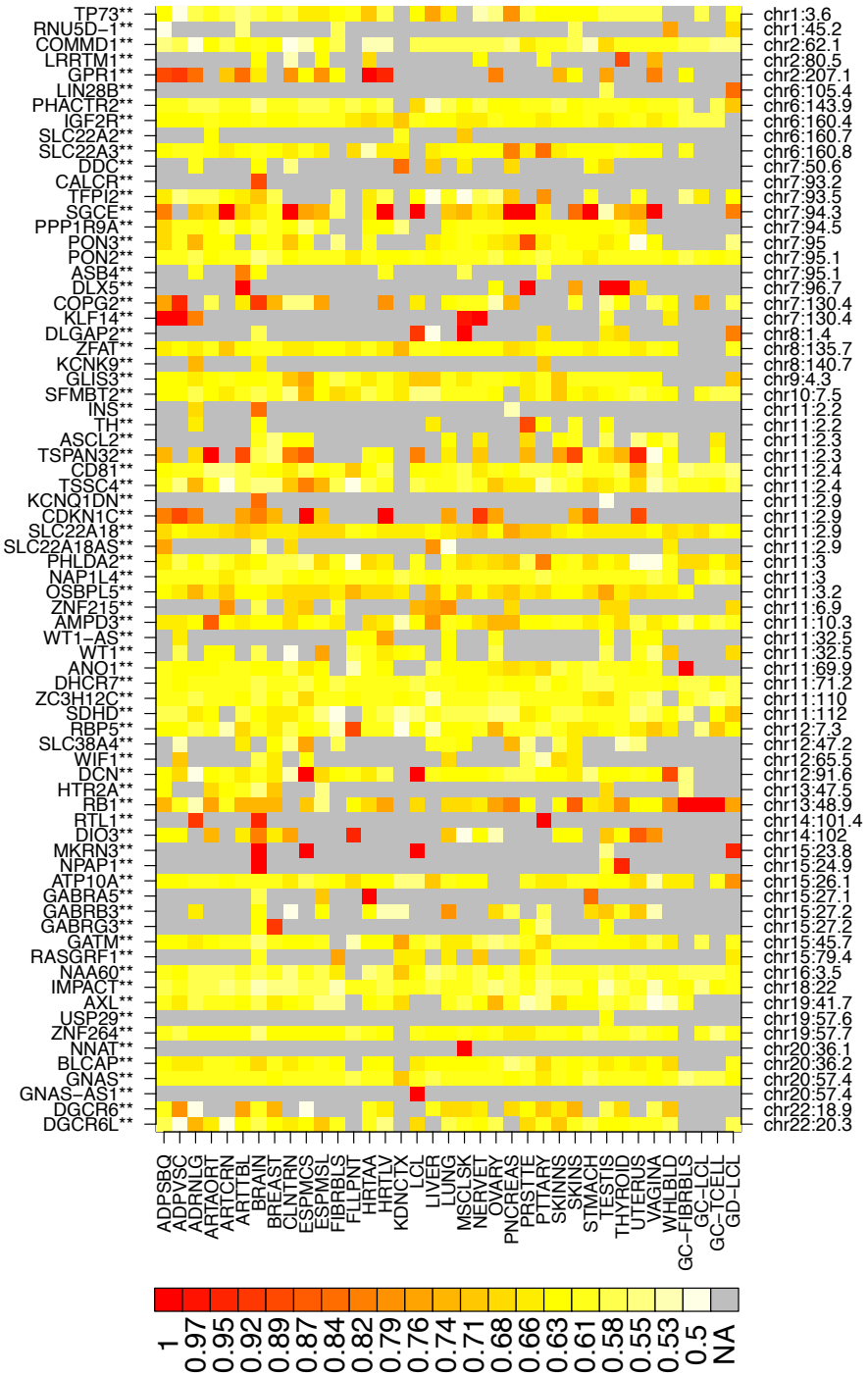


Fig. S10.

Levels of imprinting (τ) detected in all previously characterized human imprinted genes where we had data, excluding those found imprinted and shown in Fig. 2.

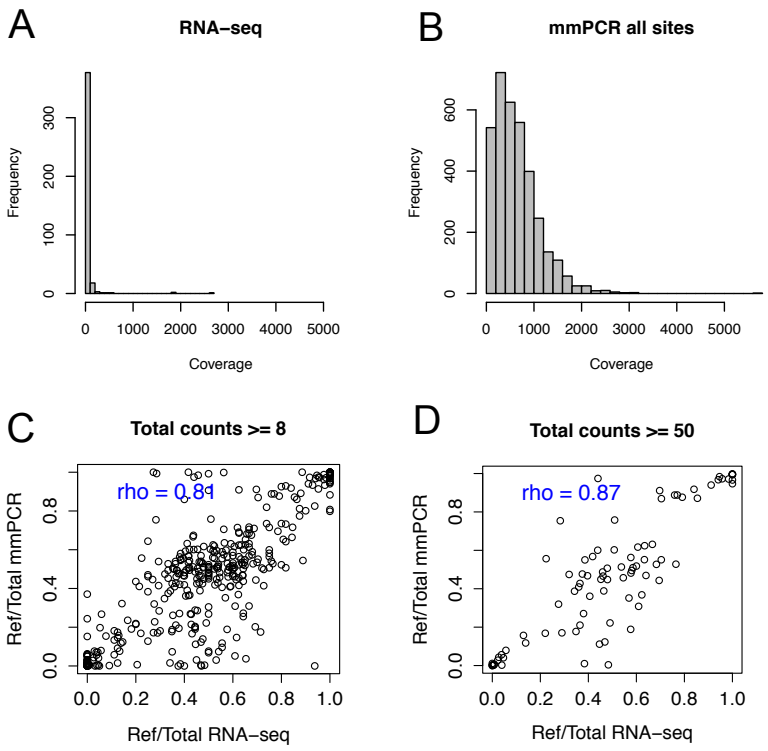


Fig. S11.

Validation of allelic ratios by mmPCR. Coverage per validated site in the original GTEx RNA-seq data (A) and in mmPCR data (B) shows a dramatically higher coverage in mmPCR data. The correlation of allelic ratios in the two data sets is high, as shown by (C) for sites with >=8 reads (the minimum in our analysis) and in (D) for sites with >=50 reads. Importantly, while the correlation is not perfect, the patterns do not suggest any systematic bias affecting allelic ratios in RNA-seq.

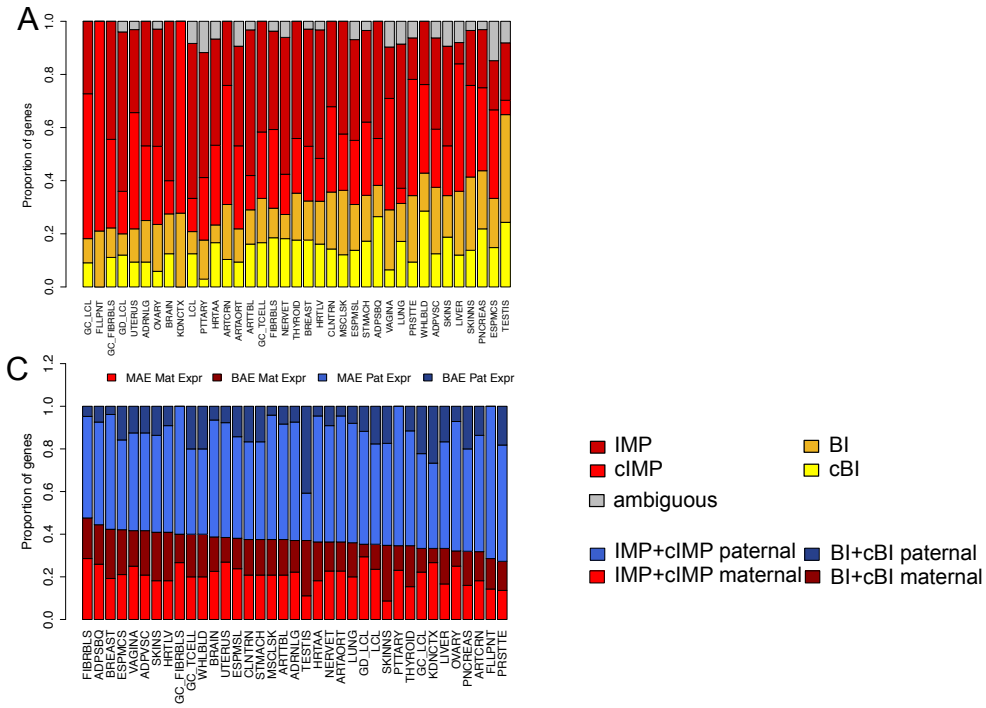


Fig. S12.

A) Imprinted and biallelic genes per tissue in all the 42 genes classified as imprinted in this study. C) Shows the imprinting status of genes with previously known information on the expressed parent. The larger amount of paternally expressed genes is a previously known fact, but interestingly, biallelic expression is more common in paternally imprinted genes (see also Fig. 3a). Furthermore, testis is an outlier in prevalent biallelic expression in paternally expressed genes.

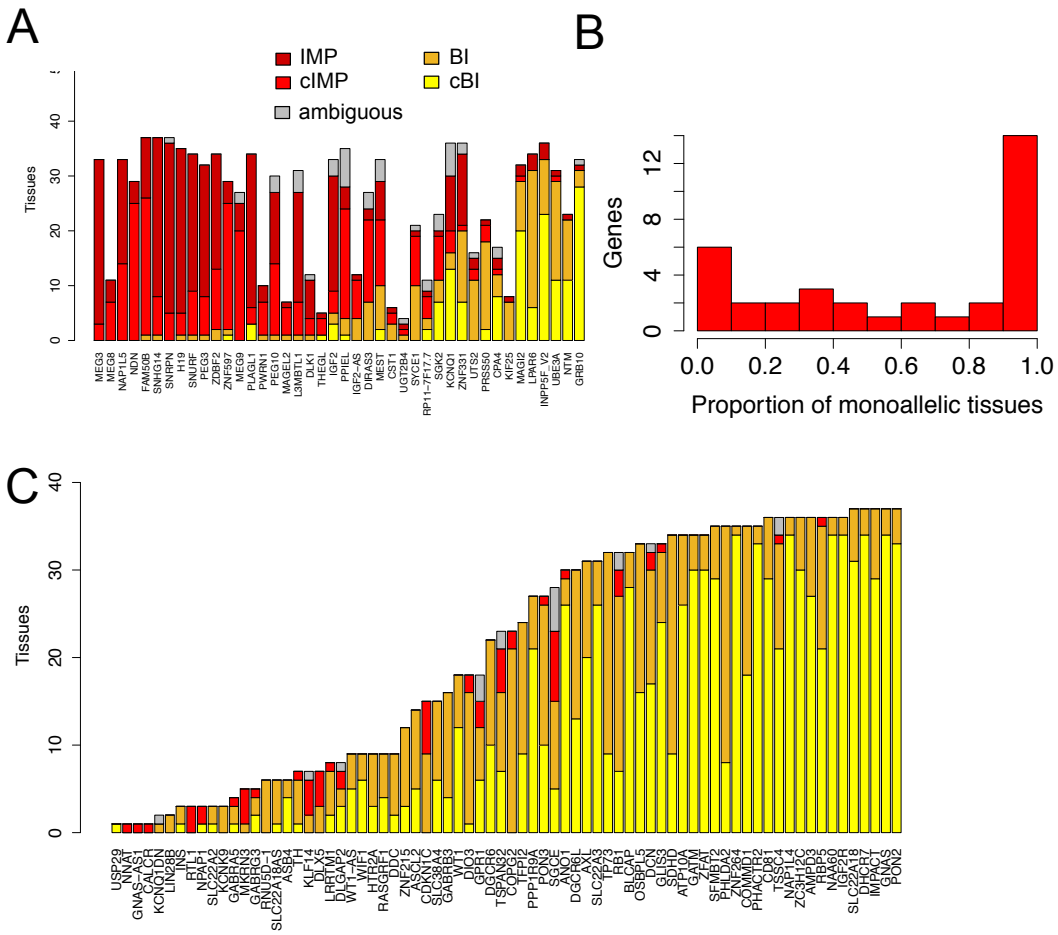


Fig. S13.

Imprinting status across tissues of each of the genes classified as imprinted in this study (A), and previously known genes that were not classified as imprinted and included in (a) (C). B) summarizes the data in (A) by showing the distribution of (cIMP+IMP)/(cIMP+IMP+cBI+BI) tissues for each gene.

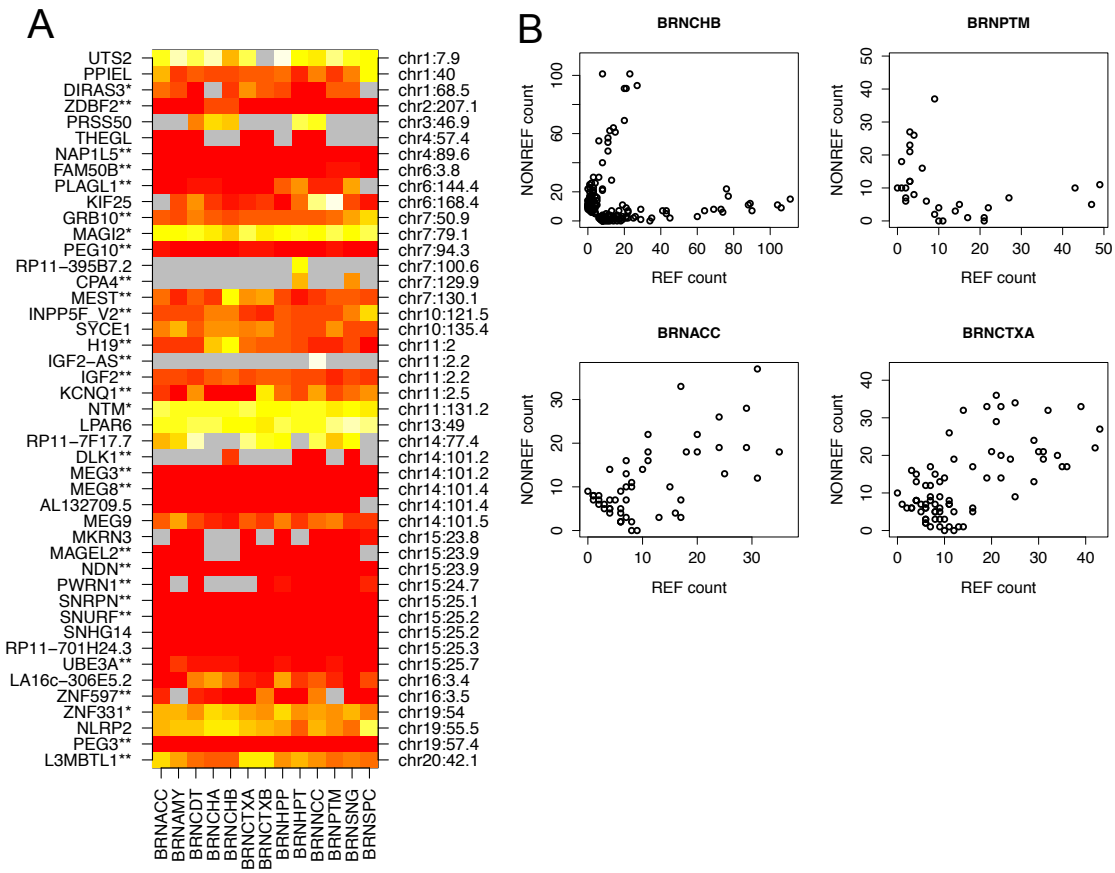


Fig. S14.

Imprinting across 13 sub-regions of the brain. A) shows a heatmap of τ (analogously to Fig. 2) for the brain sub-regions, demonstrating a generally consistent pattern across the different regions. Closer inspection revealed solid signs of tissue heterogeneity only in L3MBTL1, with four example sub-tissues shows in (B) – the top row tissues appear to have imprinted genes, whereas the bottom contains biallelic genes.

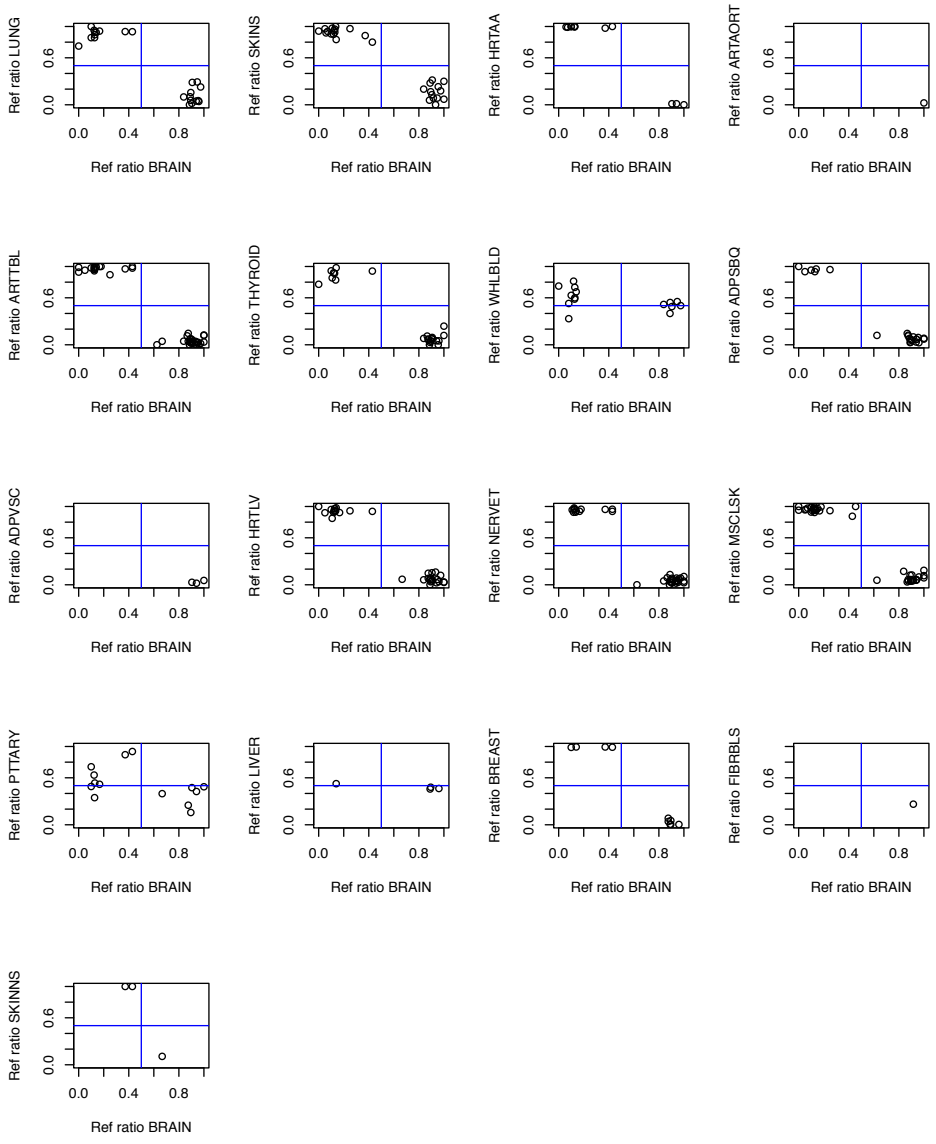


Fig. S15.

Allelic ratio of IGF2 of each of the tissues compared to allelic ratio in the brain. Each dot represent a SNP in one individual. Most tissues show a clear reversal of the direction of imprinting compared to the brain. See also section 9.

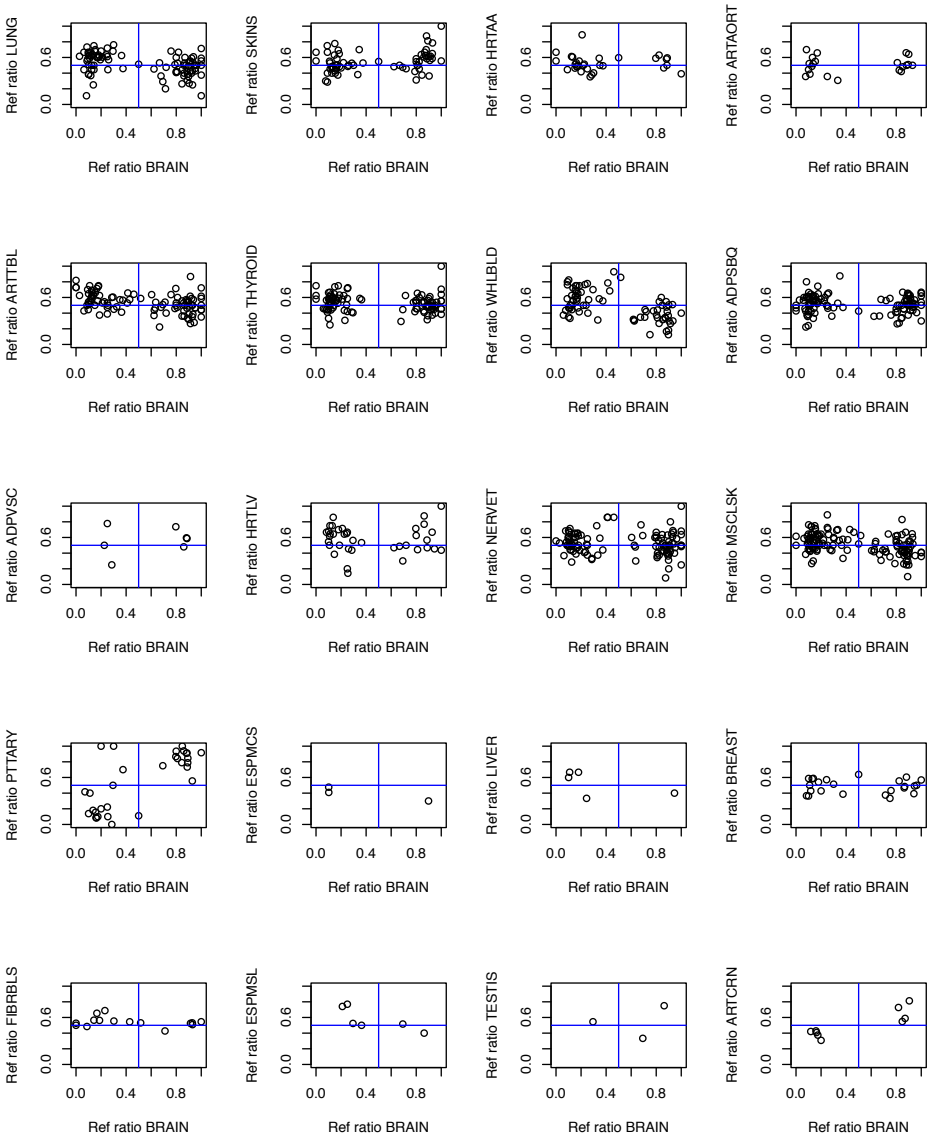


Fig. S16.

Allelic ratio of GRB10 of each of the tissues compared to allelic ratio in the brain. Each dot represent a SNP in one individual. Most tissues do not show strong imprinting, but for example blood shows an allelic bias to the opposite direction to that in the brain, and pituitary glad shows imprinting consistent with the brain.

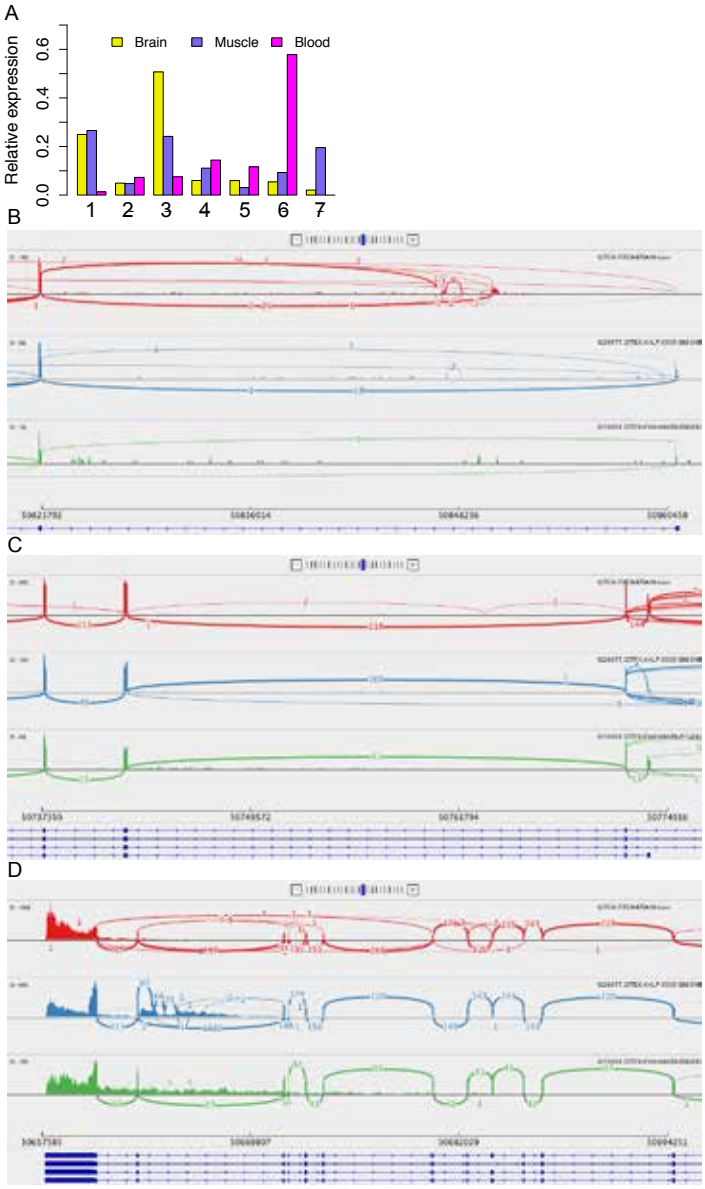


Fig. S17.

Transcript structure in GRB10 for brain, whole blood (WHLBLD) and muscle (MSCLSK), of which brain shows a strong signal of imprinting (Fig. S16) while blood and muscle show a biased allelic expression in the opposite direction. (A) shows relative expression levels of transcripts 1-7 corresponding to ENST00000406641, ENST0000047396, ENST00000439599, ENST00000398791, ENST00000461886, ENST00000398810, respectively. The sashimi plots in a-c show RNA-seq coverage and splicing in brain (red), blood (blue) and muscle (green) with clear differences especially in the 5' end, suggesting different promoters. However, analysis of allelic expression of the SNPs located in different parts of the transcripts yielded inconclusive results of putative transcript-specific imprinting.

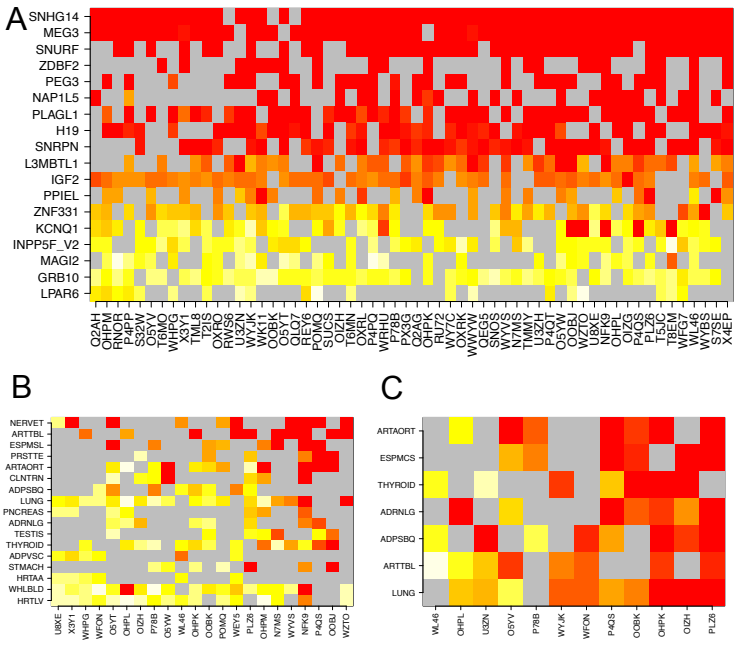


Fig. S18.

Variable imprinting levels between individuals in GTEx data. In order to avoid potential caveats of phasing errors, we calculated τ as an average of all the SNPs of an individual for a given gene and tissue. A) Shows a summary per gene, taking an average of the tissues, and B) shows an example from KCNQ1, and C) from PPIEL. There is no correlation between τ and the number of SNPs per sample, suggesting that individual errors e.g. in genotyping are unlikely to drive the patterns.

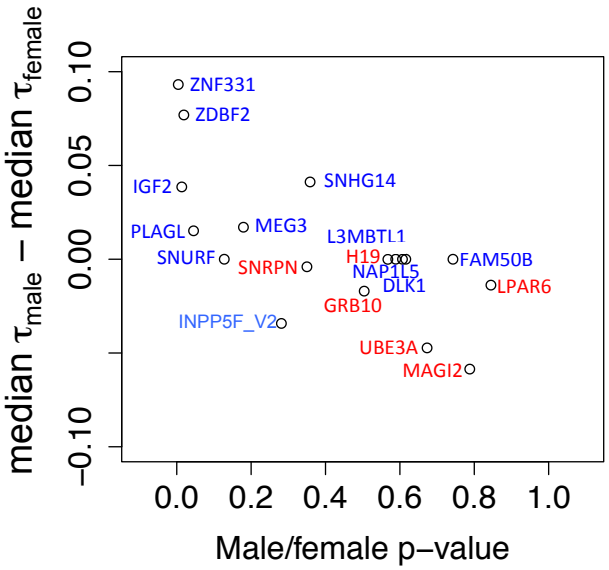


Fig. S19.

Gender difference in imprinting in the muscle for genes classified as imprinted in this tissue. For each gene, we compared the τ values of males and females: the y-axis shows the median τ per gene for males minus median τ per gene for females, and the Mann-Whitney p-value of the the comparison is on the x-axis.

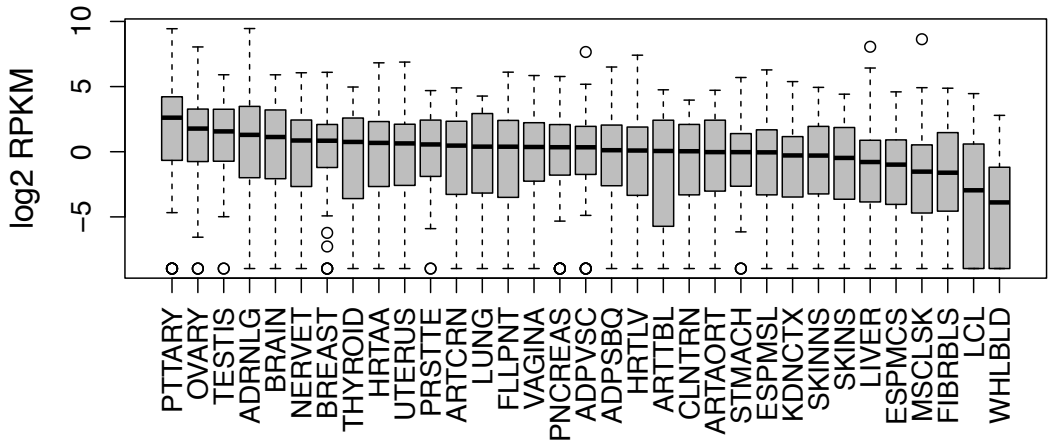


Fig. S20.

Distribution of expression levels per tissue of the 42 genes detected as imprinted in this study. All the 42 genes were included in each tissue, regardless of their imprinting status.

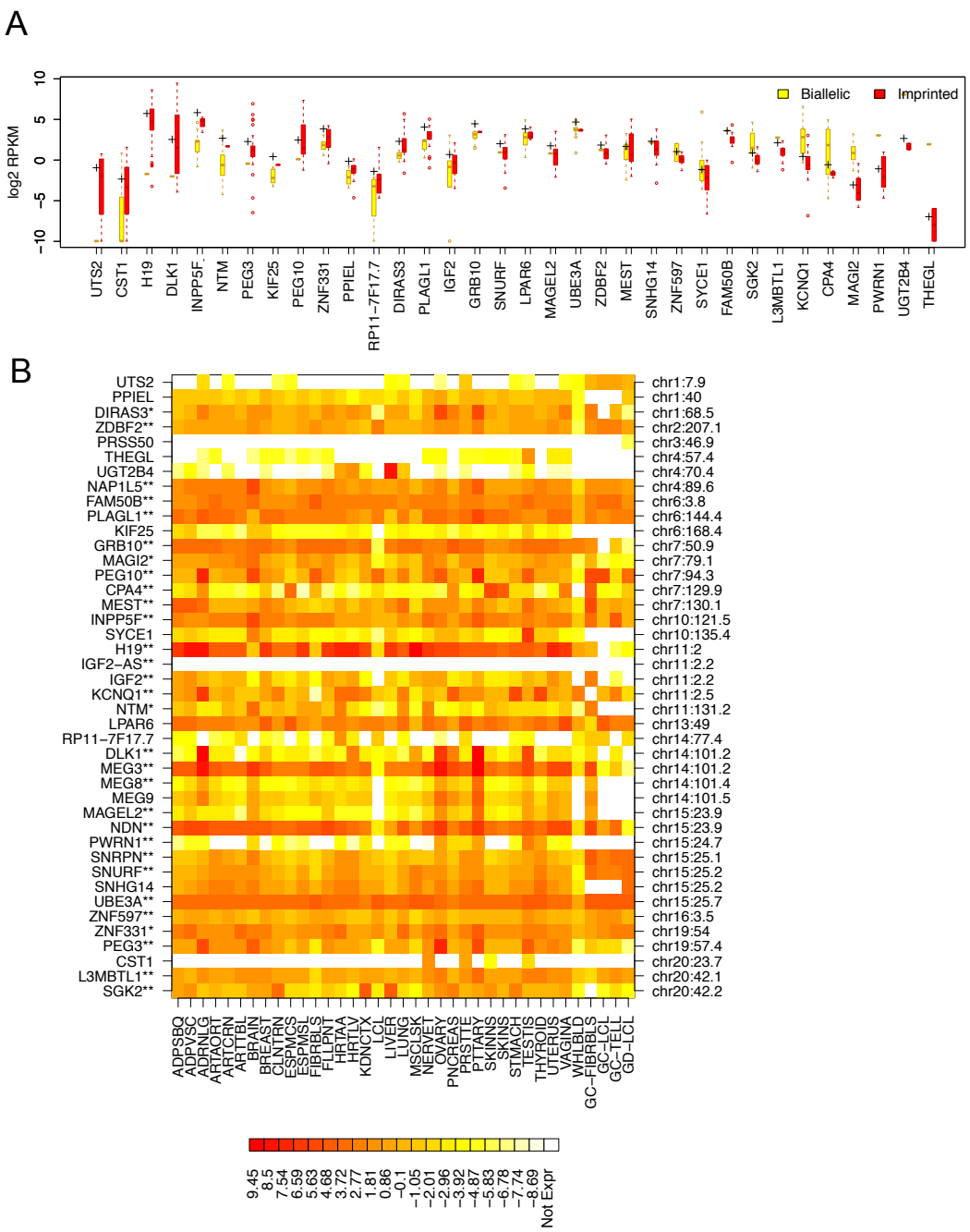


Fig. S21.

A) shows expression levels in tissues with imprinted and biallelic genes for the 42 genes detected as imprinted at least one tissue in this study, measured as median log2 RPKM. The cross denotes the expression level of 2 x imprinted median, which can be considered the expected value for tissues with biallelic genes. B) shows the distribution of expression levels as log2 RPKM for each biallelic or imprinted gene in all tissues.

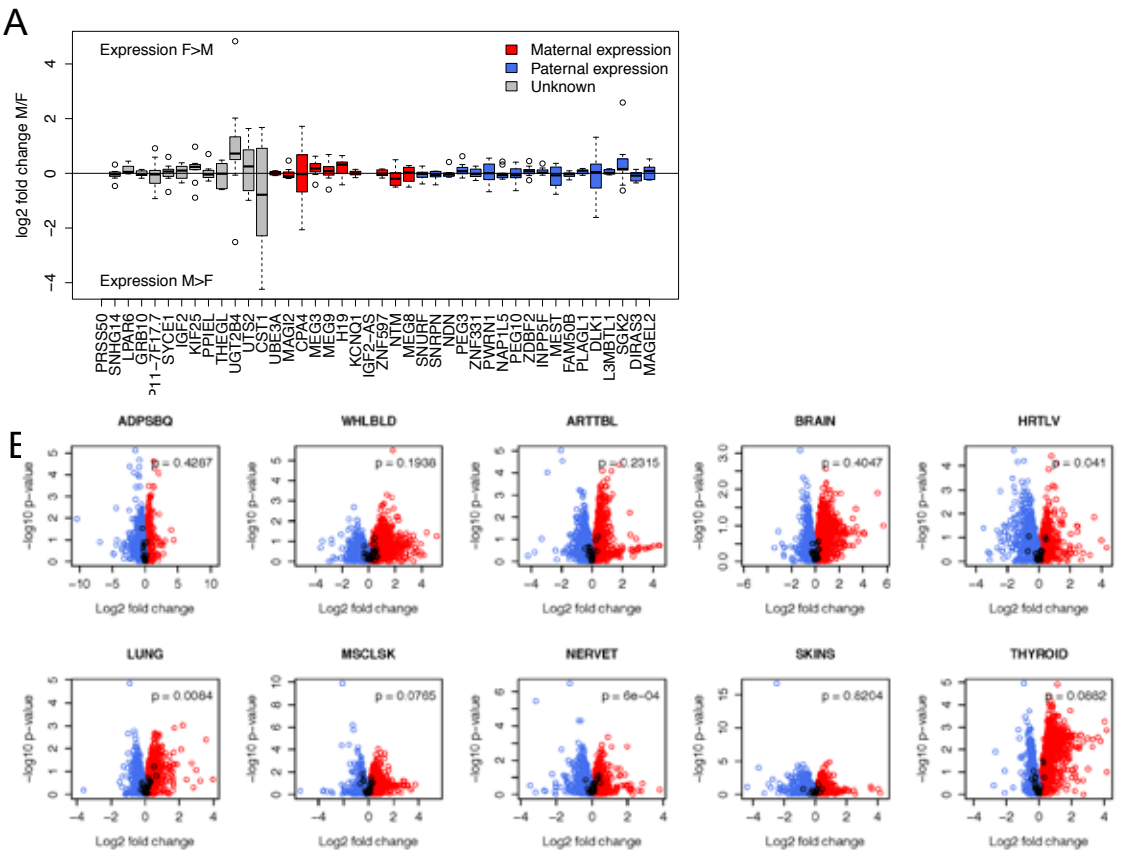


Fig. S22.

Gender differences in gene expression. A) shows log2 fold changes in expression levels measures as RPKM from all the tissues per gene. Values above 0 indicate higher expression in females. Maternally versus paternally expressed genes do not show significant differences in the proportion of genes with higher expression in females or males ($p = 0.27$ based on permutation of maternal/paternal labels. We also performed this analysis separately for each tissue (B), but did not find individual genes with significant expression differences. However, in muscle and nerve, imprinted genes show more differential expression between males and females than other genes. Imprinted genes are marked in black, and the nominal p-values are from a Mann-Whitney test of comparing absolute log2 fold change distribution of imprinted genes to all other genes.

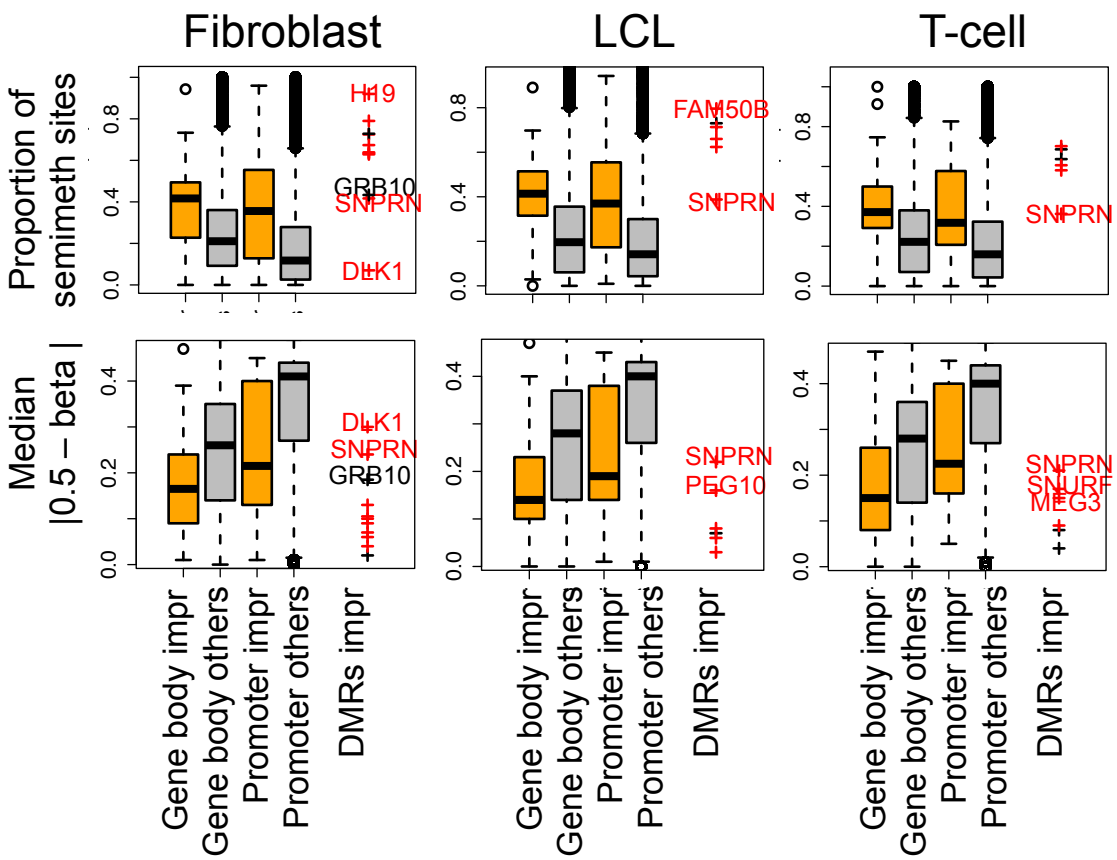


Fig. S23.

Methylation analysis. For each of the three cell types, we show two methylation statistics that summarize data across individuals and methylation sites – with the aim to detect methylation of only one allele as expected in imprinted loci. Other statistics showed similar trends (data not shown). The top row shows the proportion of semimethylated sites (sites with methylation beta value x-y) over the total, pooling all sites and individuals. The bottom row shows the median deviation from semimethylation (0.5) over all sites and individuals. These statistics are compared between imprinted genes in each of the three cell types and all other genes, both for gene body and promoters (boxplots). The gene names show the same statistics calculated from differentially methylated regions (DMRs; (Court et al. 2014)) for genes with a known DMR, with the gene names in red indicating an imprinted status in this study for the corresponding cell type. The DMRs are for SNRPN, MEG3, ZNF331, KCNQ1, SNURF, FAM50B, PPIEL, DLK1, PEG10, IGF2, DIRAS3, PLAGL1, MEST, H19, and GRB10.

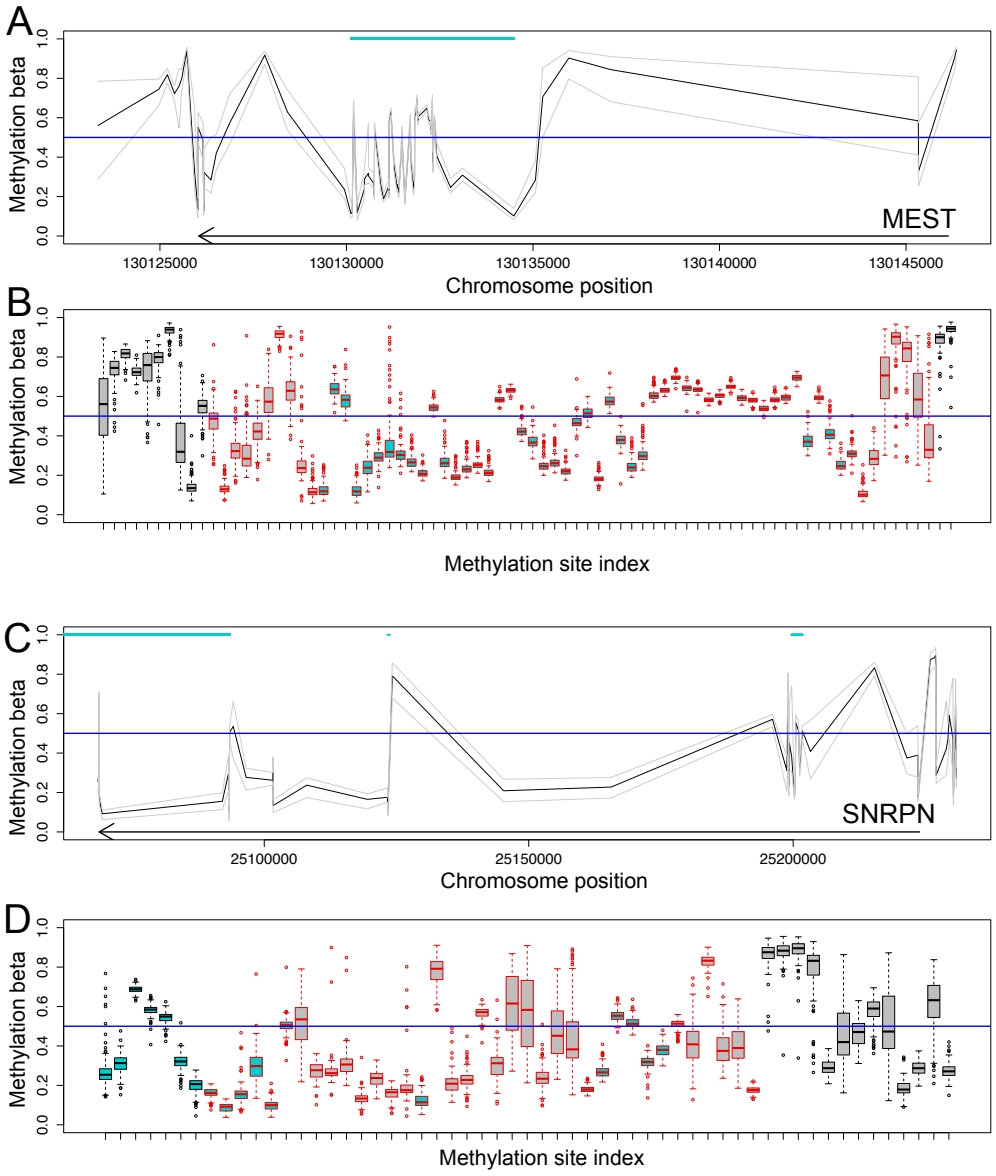


Fig. S24.

Examples of methylation landscapes in imprinted loci in *MEST* (a,b) and *SNRPN* (c,d) in Gencord fibroblast data. The lineplots in (a,c) show the methylation level (beta) as a median across individuals (black line) and the grey lines denote 10^{th} and 90^{th} quantiles. The arrow shows the gene region, and the cyan bars denotes differentially methylated regions. The boxplots in (b,d) are the same data, showing the full population distribution of each site and without scaling of the x-axis according to chromosomal position. The red boxes are for the gene region, and cyan color denotes differentially methylated regions.

Table S1.
Samples of the primary data sets of this study. The clonality proportion is estimated from monoallelic expression in the X chromosome in female samples; see Fig. S5. The validation data and samples are described in the text and the Supplementary Text.

Table S2.
Sources and characteristic of monoallelic expression

Table S3.
Summary statistics of the 42 identified as imprinted. Definitions of the column names are in Supplementary Text section 7.

Table S4.
Summary statistics of putatively imprinted genes. Definitions of the column names are in Supplementary Text section 7.

Table S5.
Summary statistics of all analyzed genes. Definitions of the column names are in Supplementary Text section 7.

Table S6.
List of known and putatively imprinted genes in human. HS and MM denote human and mouse, respectively, and the notes indicate data source or status other than the Otago database (Morison et al. 2001). The classification in our data is summarized across tissues with a hierarchy of (consistent with biallelic) < biallelic < (consistent with imprinting) < imprinted.

Table S7.
Results of the analysis of family validation data.

Data S1.
Scatterplots of read counts for the imprinted genes for each tissue, with each dot representing read counts per SNP per individual.

Data S2.
Scatterplots of read counts for the imprinted genes for each tissue, with each dot representing read counts per haplotype per individual, thus combining data from multiple SNPs (including imputed variants) when an individual has several heterozygous sites per gene.

Data S3.
Scatterplots of read counts for the known imprinted genes (as defined in 8) for each tissue, with each dot representing read counts per SNP per individual.

Data S4.
Scatterplots of read counts for the known imprinted genes (as defined in 8) for each tissue, with each dot representing read counts per haplotype per individual, thus combining data from multiple SNPs (including imputed variants) when an individual has several heterozygous sites per gene.

Data S5.
The software implementing all methods described in this paper.

References

Andres AM, Dennis MY, Kretzschmar WW, Cannons JL, Lee-Lin SQ, Hurle B, Program NCS, Schwartzberg PL, Williamson SH, Bustamante CD et al. 2010. Balancing selection maintains a form of ERAP2 that undergoes nonsense-mediated decay and affects antigen presentation. *PLoS Genet* **6**(10): e1001157.

Borrell LN, Nguyen EA, Roth LA, Oh SS, Tcheurekdjian H, Sen S, Davis A, Farber HJ, Avila PC, Brigino-Buenaventura E et al. 2013. Childhood obesity and asthma control in the GALA II and SAGE II studies. *American journal of respiratory and critical care medicine* **187**(7): 697-702.

Celeux G, Diebolt J. 1985. The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational statistics quaterly* **2**(1): 73-82.

Cho H, Davis J, Li X, Smith KS, Battle A, Montgomery SB. 2014. High-resolution transcriptome analysis with long-read RNA sequencing. *PLoS One* **9**(9): e108095.

Court F, Tayama C, Romanelli V, Martin-Trujillo A, Iglesias-Platas I, Okamura K, Sugahara N, Simon C, Moore H, Harness JV et al. 2014. Genome-wide parent-of-origin DNA methylation analysis reveals the intricacies of human imprinting and suggests a germline methylation-independent mechanism of establishment. *Genome Res* **24**(4): 554-569.

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

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

Jeon Y, Sarma K, Lee JT. 2012. New and Xisting regulatory mechanisms of X chromosome inactivation. *Current opinion in genetics & development* **22**(2): 62-71.

Kuleshov V, Xie D, Chen R, Pushkarev D, Ma Z, Blauwkamp T, Kertesz M, Snyder M. 2014. Whole-genome haplotyping using long reads and statistical methods. *Nat Biotechnol* **32**(3): 261-266.

Kumar R, Nguyen EA, Roth LA, Oh SS, Gignoux CR, Huntsman S, Eng C, Moreno-Estrada A, Sandoval K, Penaloza-Espinosa RI et al. 2013. Factors associated with degree of atopy in Latino children in a nationwide pediatric sample: the Genes-environments and Admixture in Latino Asthmatics (GALA II) study. *The Journal of allergy and clinical immunology* **132**(4): 896-905 e891.

Li S, Tighe SW, Nicolet CM, Grove D, Levy S, Farmerie W, Viale A, Wright C, Schweitzer PA, Gao Y et al. 2014. Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nat Biotechnol* **32**(9): 915-925.

Monk D, Arnaud P, Frost JM, Wood AJ, Cowley M, Martin-Trujillo A, Guillaumet-Adkins A, Iglesias Platas I, Camprubi C, Bourc'his D et al. 2011. Human imprinted retrogenes exhibit non-canonical imprint chromatin signatures and reside in non-imprinted host genes. *Nucleic Acids Res* **39**(11): 4577-4586.

Morison IM, Paton CJ, Cleverley SD. 2001. The imprinted gene and parent-of-origin effect database. *Nucleic Acids Res* **29**(1): 275-276.

Nishimura KK, Galanter JM, Roth LA, Oh SS, Thakur N, Nguyen EA, Thyne S, Farber HJ, Serebrisky D, Kumar R et al. 2013. Early-life air pollution and asthma risk in minority children. The GALA II and SAGE II studies. *American journal of respiratory and critical care medicine* **188**(3): 309-318.

Poole A, Urbanek C, Eng C, Schageman J, Jacobson S, O'Connor BP, Galanter JM, Gignoux CR, Roth LA, Kumar R et al. 2014. Dissecting childhood asthma with nasal transcriptomics distinguishes subphenotypes of disease. *The Journal of allergy and clinical immunology* **133**(3): 670-678 e612.

Rivas MA, Pirinen M, Conrad DF, Lek M, Tsang EK, Karczewski KJ, Maller JB, Kukurba KR, DeLuca D, Fromer M et al. 2015. Impact of predicted protein-truncating genetic variants on the human transcriptome. *Submitted*.

Torgerson DG, Ampleford EJ, Chiu GY, Gauderman WJ, Gignoux CR, Graves PE, Himes BE, Levin AM, Mathias RA, Hancock DB et al. 2011. Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. *Nat Genet* **43**(9): 887-892.

Zhang R, Li X, Ramaswami G, Smith KS, Turecki G, Montgomery SB, Li JB. 2014. Quantifying RNA allelic ratios by microfluidic multiplex PCR and sequencing. *Nat Methods* **11**(1): 51-54.