Figures and figure supplements

Mapping the mouse Allelome reveals tissue-specific regulation of allelic expression

Daniel Andergassen *et al*

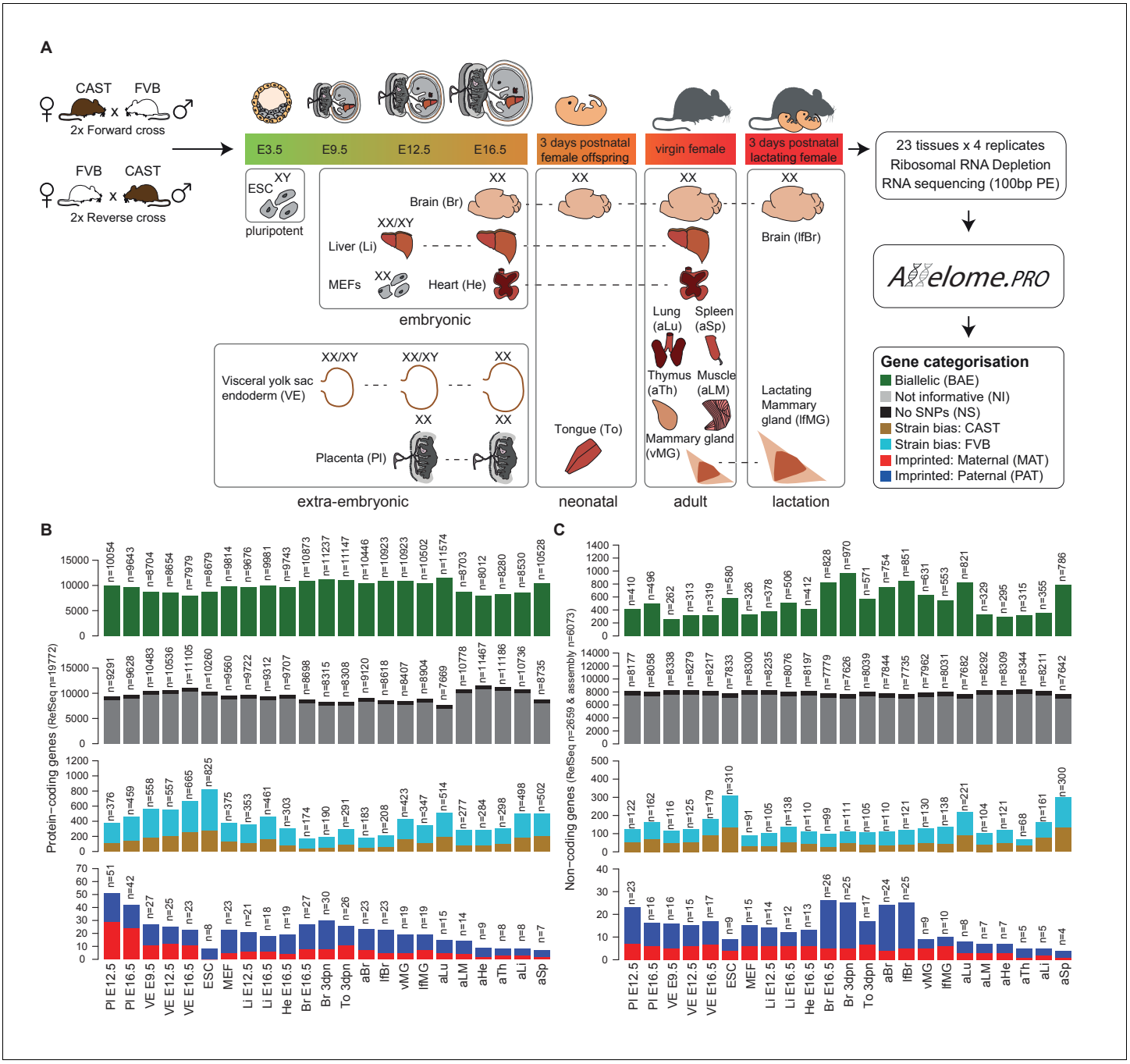

Figure 1. Defining the mouse Allelome. (**A**) Strategy for detecting allelic expression from RNA-seq data from 23 mouse tissues and developmental stages using Allelome.PRO. Every gene in the annotation is classified into one of seven different allelic expression categories indicated by different colors in the key and explained in the text. These colors are used in figures throughout the manuscript. The sex of the tissues is indicated by XX (female) and XY (male). Individuals were used except for indicated embryonic tissues where an entire litter was pooled (XX/XY). (**B**) Allelome.PRO classification of the allelic expression status of protein-coding genes in each tissue. (**C**) Allelome.PRO classification of non-coding genes. Tissues examined were placenta (Pl embryonic day (**E**) 12.5, E16.5), visceral yolk sac endoderm (VE E9.5, E12.5, E16.5), embryonic stem cells (ESC), mouse embryonic fibroblasts (MEF E12.5), embryonic liver (Li E12.5, E16.5), embryonic heart (He E16.5), embryonic and neonatal brain (Br E16.5, 3 days postnatal (dpn)), neonatal tongue (To 3dpn), adult brain (aBr), adult lactating female brain (lfBr), adult virgin mammary glands (vMG), adult lactating female mammary glands (lfMG), adult lung (aLu), adult leg muscle (aLM), adult heart (aHe), adult thymus (aTh), adult liver (aLi) and adult spleen (aSp). Embryo and placenta diagrams adapted from ***Hudson et al. (2011)***. Allelome.PRO settings: FDR 1%, allelic-ratio cutoff 0.7, minread 2.
DOI: 10.7554/eLife.25125.002

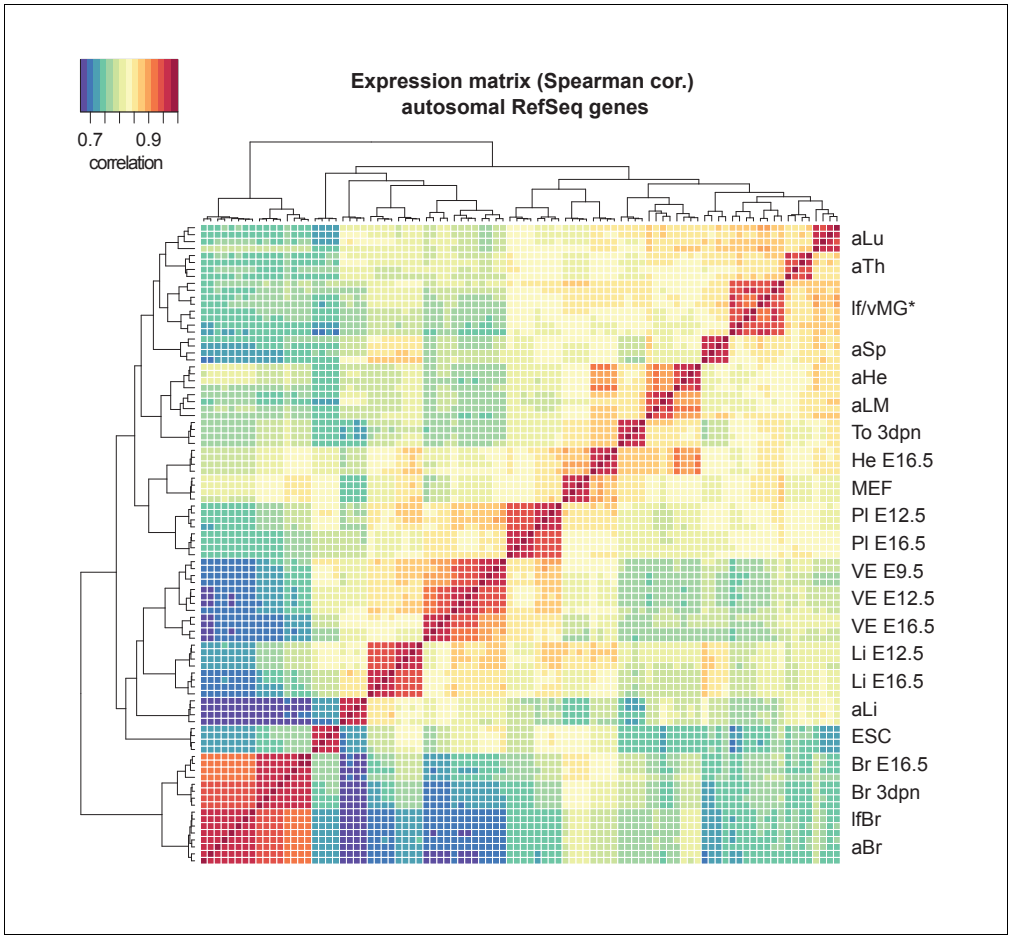

Figure 1—figure supplement 1. Clustering of tissues by their RNA-seq expression data matches the expected developmental relationships. Cluster analysis confirms identity and similarity of sequenced tissues and replicates. The Heatmap shows unsupervised clustering of a Spearman correlation matrix from log-transformed gene expression data across 92 samples (23 tissues x four replicates, Refseq genes). In an exception, clustering did not distinguish virgin and lactating mammary glands (*).

DOI: 10.7554/eLife.25125.003

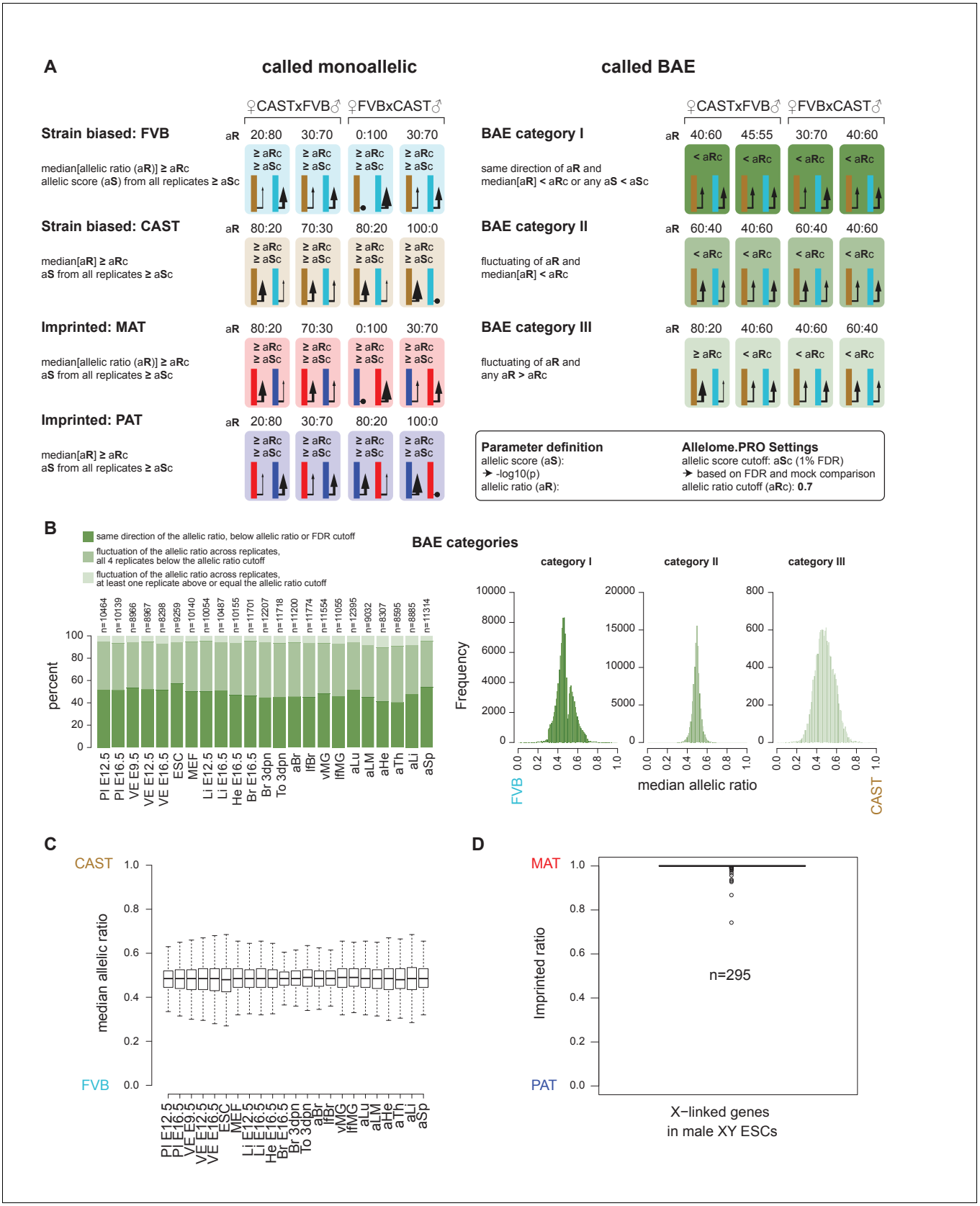

Figure 1—figure supplement 2. The Allelome.PRO pipeline output and quality controls. (**A**) Allelome.PRO strategy to classify monoallelic (strain biased or imprinted) and biallelic (BAE) expression (FDR 1%, allelic ratio cutoff 0.7, minread 2). The three different categories of biallelic expression are

Figure 1—figure supplement 2 continued on next page

Figure 1—figure supplement 2 continued

shown. For more details see the Materials and methods and the Allelome.PRO methods paper and manual (***Andergassen et al., 2015***). (**B**) Left: the percentage of each biallelic category in each tissue. Right: histograms showing the allelic ratio distribution for each of the three biallelic categories. (**C**) Allelic ratios of genes show a similar distribution between tissues (outliers not shown in boxplots). (**D**) X-linked genes align to the correct allele in XY males, indicating that detection of X-inactivation escapers is not due to an alignment artifact.
DOI: 10.7554/eLife.25125.004

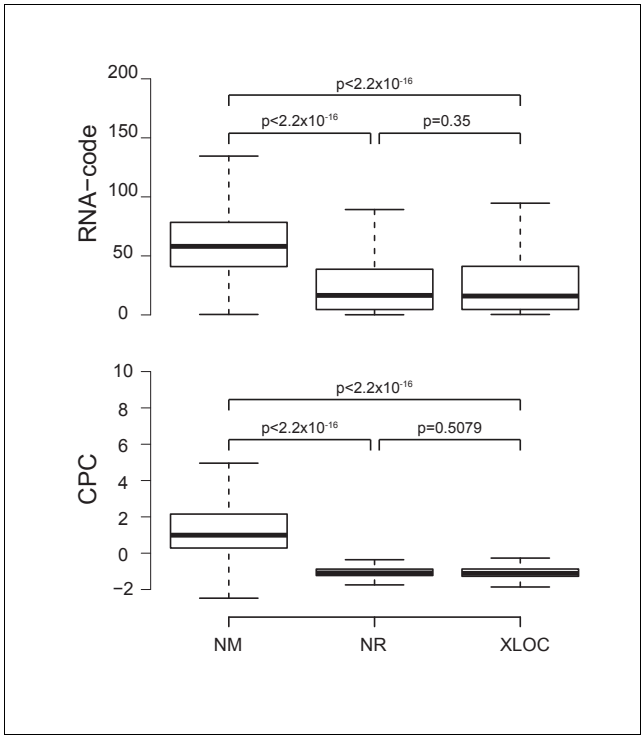

Figure 1—figure supplement 3. Novel non-coding RNAs show a similar non-coding potential to annotated Refseq non-coding RNAs. Novel annotated loci outside of RefSeq are non-protein coding. The boxplots show the estimated RNA-code (top) (***Washietl et al., 2011***) and Coding Potential Calculator (CPC) scores (bottom) (***Kong et al., 2007***) for mRNAs (RefSeq NM), non-coding RNAs (RefSeq NR) and loci annotated in this study (XLOC, outliers not shown). p-Values were calculated using a t-test. For more details see Materials and methods.
DOI: 10.7554/eLife.25125.005

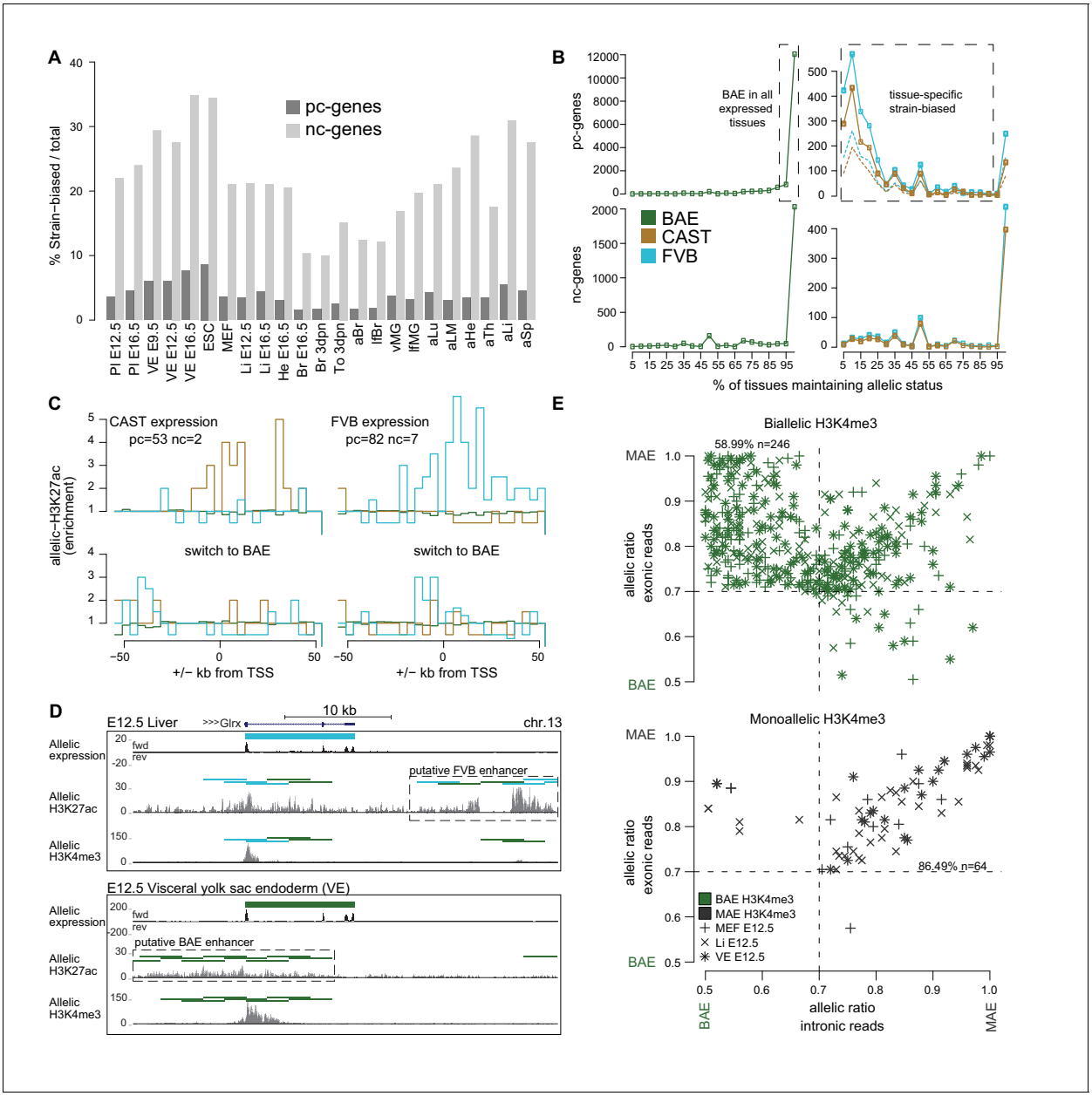

Figure 2. The Allelome reveals tissue-specific expression of strain-biased genes. (**A**) The percentage of strain-biased genes from total informative genes for each tissue for protein-coding (pc, black) and non-coding (nc, grey) genes. (**B**) The percentage of tissues where pc- and nc- genes maintained their biallic (left) or strain-biased (right), allelic expression status (calculated relative to number of tissues where a gene was informative). Allelome.PRO settings: FDR 1%, allelic-ratio cutoff 0.7, minread 2, dotted lines indicates the outcome with a 0.8 allelic ratio cutoff. (**C**) The enrichment of H3K27ac ±50 kb from the transcription start site (TSS) of genes that show strain-biased expression in either E12.5 VE or Li, and biallelic expression in the other tissue. Top: H3K27ac enrichment near strain-biased genes. The enrichment over random of allelic H3K27ac 4 kb windows was calculated. Bottom: The same analysis for the same set of genes where they show biallelic expression. Analysis detailed in Materials and methods. (**D**) An example of putative enhancer switching: *Glrx* switches from FVB strain-biased expression in liver to BAE expression in VE. This is associated with a switch in putative enhancers that matches the allelic expression status. Allelome.PRO settings for H3K27ac: FDR 1%, allelic-ratio cutoff 0.7, minread 1. (**E**) Strain-biased genes with biallelic H3K4me3 on their promoter are enriched for genes that show biallelic expression in their introns and strain-biased expression in their exons, indicating post-transcriptional differences in stability may explain strain-biased expression. Scatter plots comparing the allelic ratio of strain-biased genes in their exons and introns. Top: Strain-biased genes with biallelic H3K4me3 ChIP-seq enrichment on their promoter. Bottom: Strain-biased genes with supporting monoallelic enrichment of H3K4me3 on their promoter (key indicates H3K4me3 color code and tissue). Color code as in ***Figure 1***.

DOI: 10.7554/eLife.25125.006

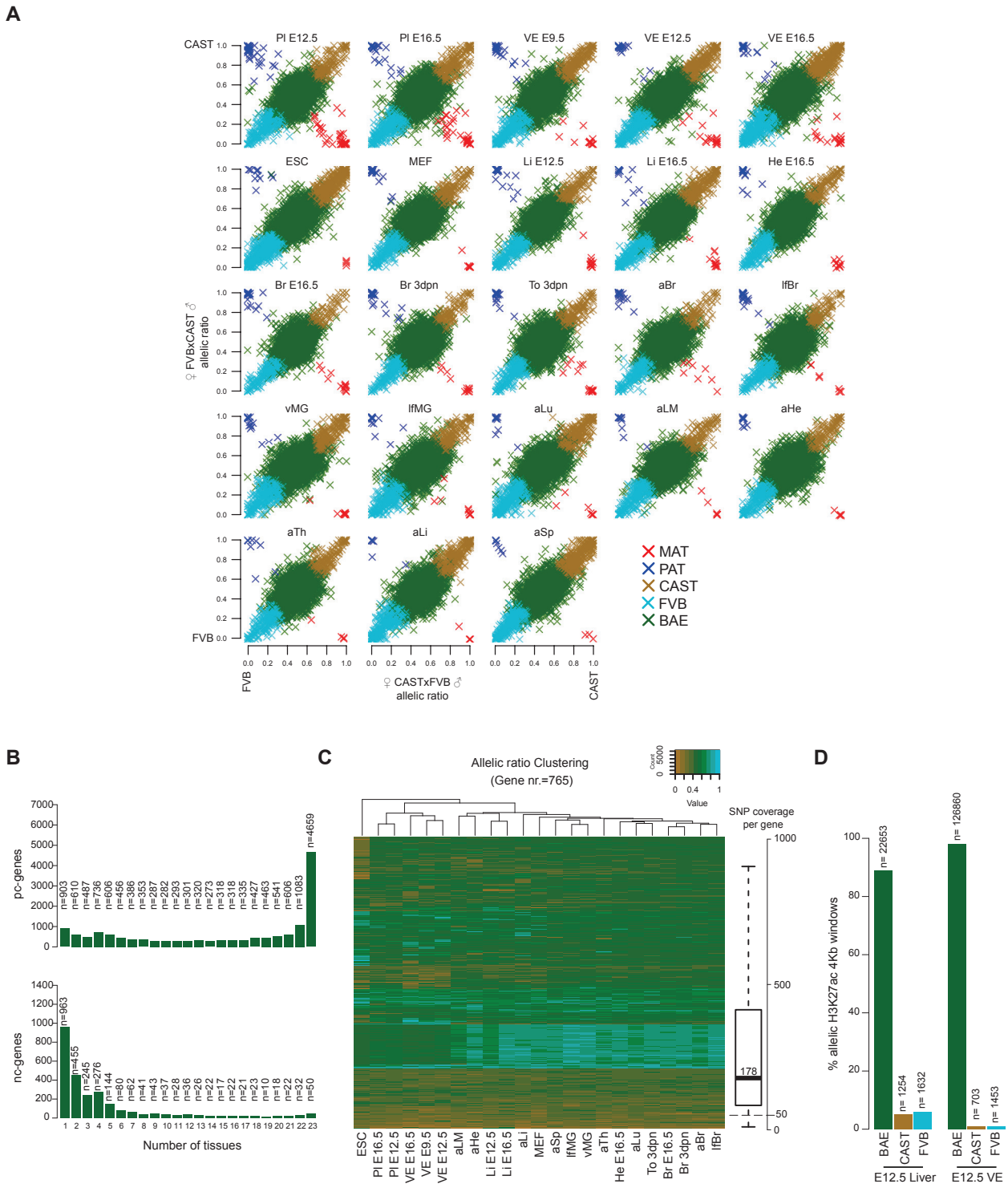

Figure 2—figure supplement 1. Characterization of the mouse Allelome reveals that protein-coding genes switch their allelic status among tissue and development. (**A**) Distribution of allelic ratios in mouse tissues. A scatterplot of the median allelic ratios of autosomes in the FVBxCAST (x2) versus

Figure 2—figure supplement 1 continued on next page

Figure 2—figure supplement 1 continued

CASTxFVB (x2) crosses determined by Allelome.PRO analysis of RNA-seq data from the 23 tissues examined (abbreviations defined in ***Figure 1***). Genes are color-coded according to their allelic bias classification as indicated in the key. (**B**) Most biallelic pc-genes (top) showed biallelic expression in the majority of tissues, while nc-genes (bottom) showed biallelic expression mainly in a single or few tissues. (**C**) Tissue clustering by allelic ratio reflects developmental relationships. Unsupervised clustered was performed using the median allelic ratios of the 765 strain-biased genes that were informative in all 23 tissues (Hierarchical Clustering Method (hClust) = 'complete' in R). Genes were color coded according to the allelic ratio using a gradient between 0 (CAST brown), 0.5 (BAE green) and 1 (FVB turquoise) (see key). The SNP coverage per gene is shown in the box plot on the right. (**D**) Biallelic enrichment of H3K27ac is frequent in the genome. Strain biased H3K27ac 4 kb enrichment windows occur at between 1–5% the frequency of biallelic enrichment windows.

DOI: 10.7554/eLife.25125.007

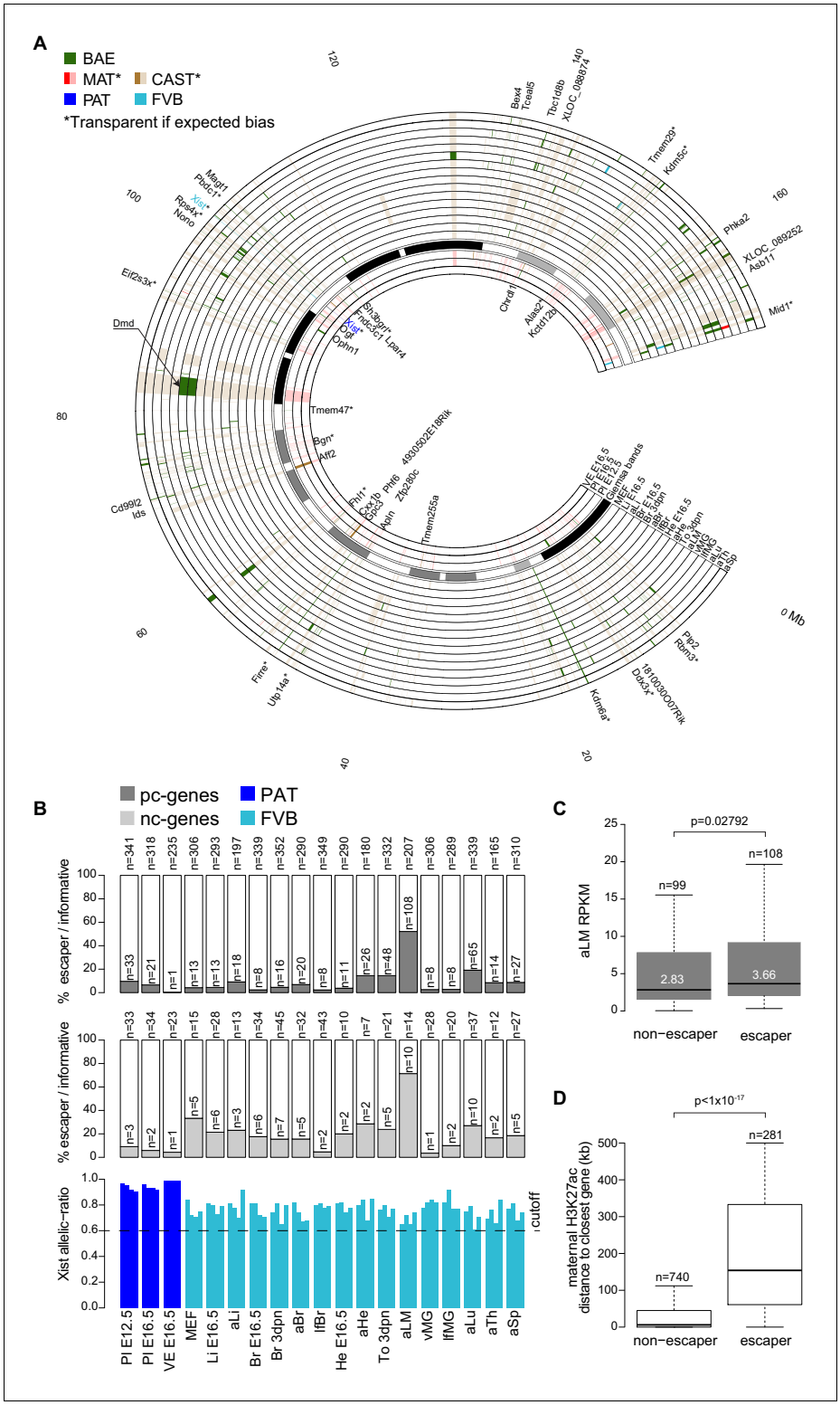

Figure 3. X chromosome inactivation (XCI) escapers appear to be highly tissue-specific. (**A**) Circos plot showing the mouse chromosome X Allelome for 19 female tissues (***Figure 1***). Outer layers: 16 embryonic, neonatal and adult tissues showing random XCI (FVB X preferentially inactivated in CAST/FVB F1 tissues (skewed XCI)). Middle layer: Giemsa banding (UCSC genome browser). Inside layers: three extra-embryonic tissues showing imprinted XCI (paternal X chromosome inactivated). The top 25/225 escapers from random XCI are indicated on the outside of the Circos plot, while the top 20/43 escapers from imprinted XCI are indicated on the inside (escapers ranked

Figure 3 continued on next page

Figure 3 continued

by number of tissues). An asterisk marks known escapers. *Dystrophin* (*Dmd*) escapes specifically in muscle (aLM and To 3dpn, indicated by arrow). Color code as in ***Figure 1*** except non-escapers are partially transparent (20% opacity, CAST in embryonic and adult tissues, MAT in extra-embryonic tissues). Allelome.PRO settings: FDR 1%, allelic-ratio cutoff 0.6, minread 2. (**B**) Top: the percentage of pc-genes (black) escaping XCI from all informative pc for each female tissue. The number of informative pc-genes is given above the barplot, while the number of pc escapers is given above each bar. Middle: the same analysis performed for nc-genes (grey). Bottom: the allelic ratio of *Xist* for each replicate in extra-embryonic tissues (*Xist* expressed paternally (blue)) and in non extra-embryonic tissues (*Xist* preferentially expressed from the FVB allele (turquoise)). (**C**) Leg muscle XCI escapers are expressed at a higher level than non-escapers. Protein coding X-chromosome escaper genes are expressed significantly higher than non-escapers on the X (t-test). Box plots indicate the expression levels of genes in the different categories (outliers not shown). (**D**) The distance of parental-specific H3K27ac window to the closest non-escaper (331) and escaper (36) gene in placenta E12.5. Maternal H3K27ac windows with a distance higher than 500 kb were not included in the analysis. A boxplot including median values is shown (outliers not shown). After correcting for sample size, a significant difference was observed between escapers and non-escapers (Fisher's exact test, $p<1\times10^{-17}$, details in Materials and methods).

DOI: 10.7554/eLife.25125.008

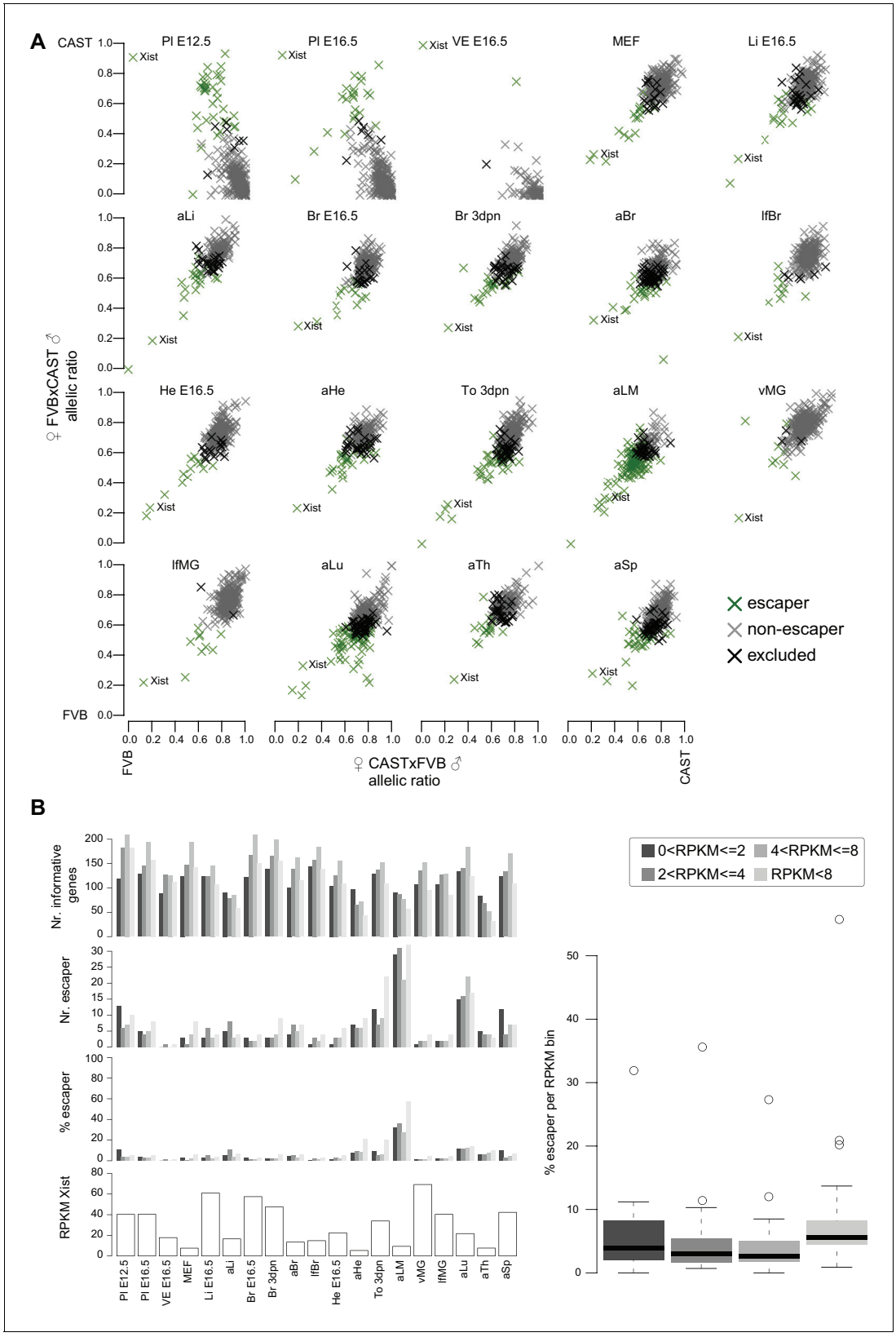

Figure 3—figure supplement 1. The distribution of X chromosome inactivation escapers across tissues. (**A**) X chromosome inactivation (XCI) escapers in 19 mouse tissues. A scatterplot of allelic ratios of X-linked genes in the FVBxCAST versus CASTxFVB crosses determined by Allelome.PRO analysis of RNA-seq data from the 19 XX tissues examined (abbreviations defined in ***Figure 1***). Escapers, non-escapers, and genes excluded from the analysis (see

Figure 3—figure supplement 1 continued on next page

Figure 3—figure supplement 1 continued

Materials and methods) are indicated in the key. *Xist* is indicated in each scatter plot. (**B**) XCI escapers are detected over a wide range of expression levels. First row: number of informative X-linked genes in each of 5 expression bins for each female tissue. Second row: the number of XCI escaper numbers for each expression bin for each tissue. Third row: percentage of informative X-linked genes that escape XCI in each expression bin for each tissue. Fourth row: *Xist* expression level does not correlate with the number XCI escapers. Shown are the *Xist* expression levels for each tissue. Right: Genes with low expression levels are not more likely to be classified as XCI escapers. Box plots display the percentage of informative genes that escape X-inactivation in each expression bin across the 19 female tissues assayed in this study.

DOI: 10.7554/eLife.25125.009

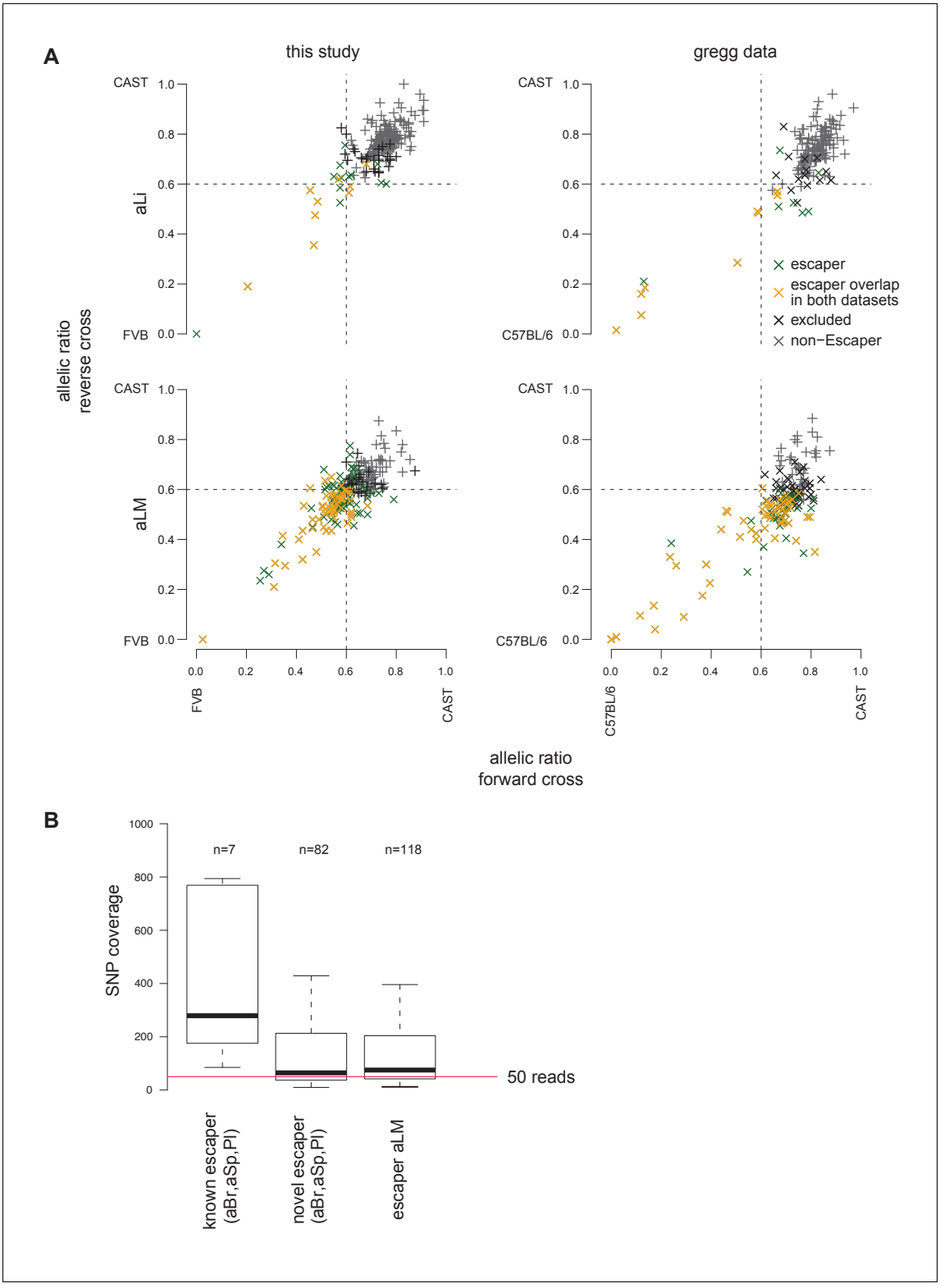

Figure 3—figure supplement 2. Validation of X chromosome inactivation escapers in an independent dataset. (**A**) Genes escaping X-inactivation confirmed in an independent dataset. Scatter plots of the median allelic ratio from the forward and reverse crosses (two replicates) of X-linked genes classified as escapers, non-escapers, or excluded from the analysis (as described in methods, color code in key). Left: adult liver (aLi) and adult leg muscle (aLM) from this study using FVBxCAST reciprocal crosses (FVBxCAST SNPs from Sanger). Right: re-analyzed adult liver and adult leg muscle data from the Gregg lab (***Bonthuis et al., 2015***) using C57BL/6xCAST reciprocal crosses (read number adjusted to match this study, C57BL/6xCAST

Figure 3—figure supplement 2 continued on next page

Figure 3—figure supplement 2 continued

SNPs from Sanger). Escapers detected in both datasets are indicated in yellow (see key). (**B**) Novel X-chromosome escapers show a lower SNP coverage than known escapers. Boxplots show the SNP coverage of known and novel escapers in adult brain (aBr), spleen (aSp) and placenta (Pl) (***Berletch et al., 2015***; ***Finn et al., 2014***). Outliers are not shown. Red line indicates a SNP coverage of 50.

DOI: 10.7554/eLife.25125.010

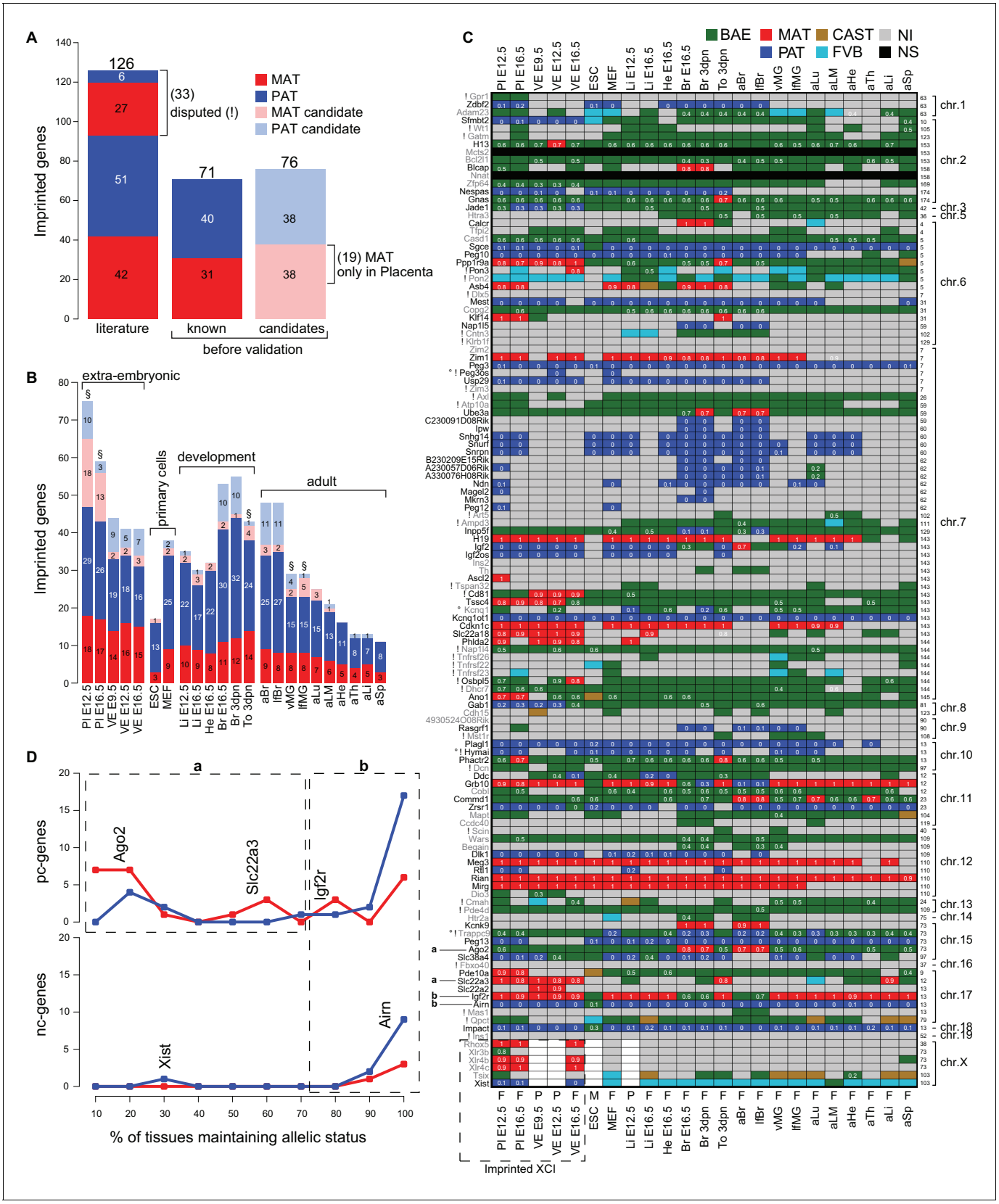

Figure 4. The Allelome reveals tissue-specific regulation of imprinted protein-coding genes. (**A**) The number of known and candidate imprinted genes detected in this study by RNA-seq before validation. Known genes were RefSeq genes listed by the Otago or Harwell imprinted databases on 24th

Figure 4 continued on next page

Figure 4 continued

Sept 2015 (***Glaser et al., 2006***; ***Williamson et al., 2013***). (**B**) The number of known and candidate novel imprinted genes found among different tissues and developmental stages. Tissues important for the energy transfer from the mother to the offspring are indicated (§). (**C**) The heatmap shows the allelic pattern for all 126 known imprinted genes among the different tissues. Gene names (left side) colored black are confirmed in this study, while gene names colored grey could not be confirmed. ! Indicates disputed genes. ° Probable RNA-seq strand bleed through. Examples given in D for variable ('a', in ≤70% expressing tissues) and consistent ('b' in >70% expressing tissues) imprinted expression are indicated. The chromosome number and the base pair coordinates (in Mb) for each gene are indicated on the right side. The imprinted allelic ratio for each gene and tissue (white) is given only if all four replicates show a bias in the same direction (1 = 100% expression from the maternal allele, 0 = 100% expression from the paternal allele). The sex for each tissue is indicated on the bottom of the heatmap: F (female, XX), M (male, XY), P (pooled XX/XY). Note: allelic analysis of the X chromosome can only be done for female tissues. (**D**) The percentage of tissues that maintain imprinted expression of protein-coding genes (top) and nc-genes (bottom) (calculated as the number of tissues showing imprinted expression, divided by the number of informative tissues for each gene). The analysis was done for the 69 known imprinted genes confirmed by RNA-seq in this study (*Peg3os* and *Hymai* were excluded due to probable RNA-seq bleed-through). Dotted boxes indicate genes that show variable ('a', in ≤70% expressing tissues) and consistent ('b' in >70% expressing tissues) imprinted expression. Examples are positioned according to the percentage of expressing tissues where they show imprinted expression. Color key as in ***Figure 1*** except novel maternal and paternal imprinted candidates are pale red and blue, respectively (Allelome.PRO settings: FDR 1%, allelic-ratio cutoff 0.7, minread 2).

DOI: 10.7554/eLife.25125.011

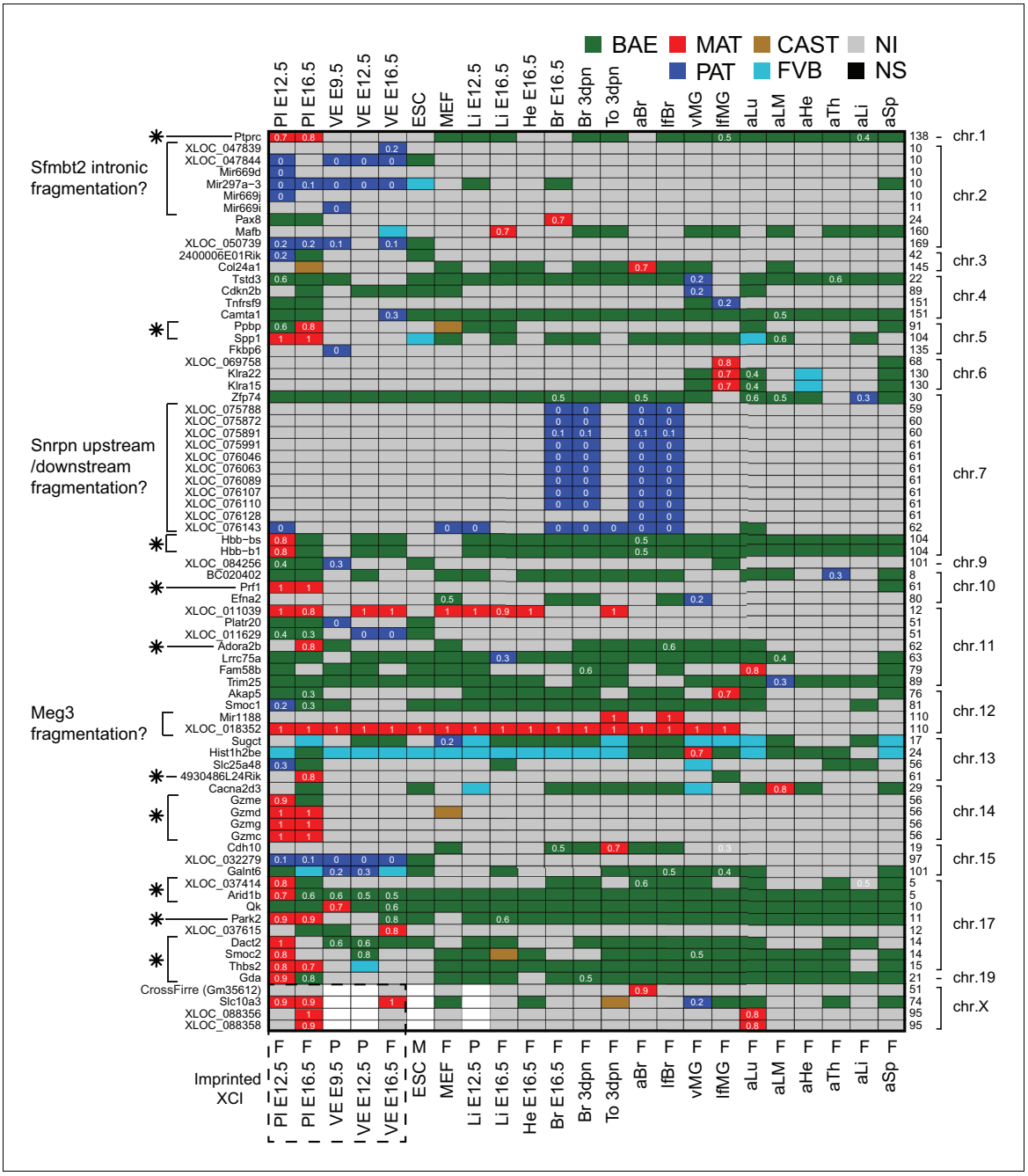

Figure 4—figure supplement 1. The allelic categorization of candidate novel imprinted genes across tissues. Novel imprinted genes identified in this study. The heatmap shows the allelic pattern among the different tissues for the 76 novel imprinted genes detected by RNA-seq. Bold gene names (left side)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

DOI: 10.7554/eLife.25125.012

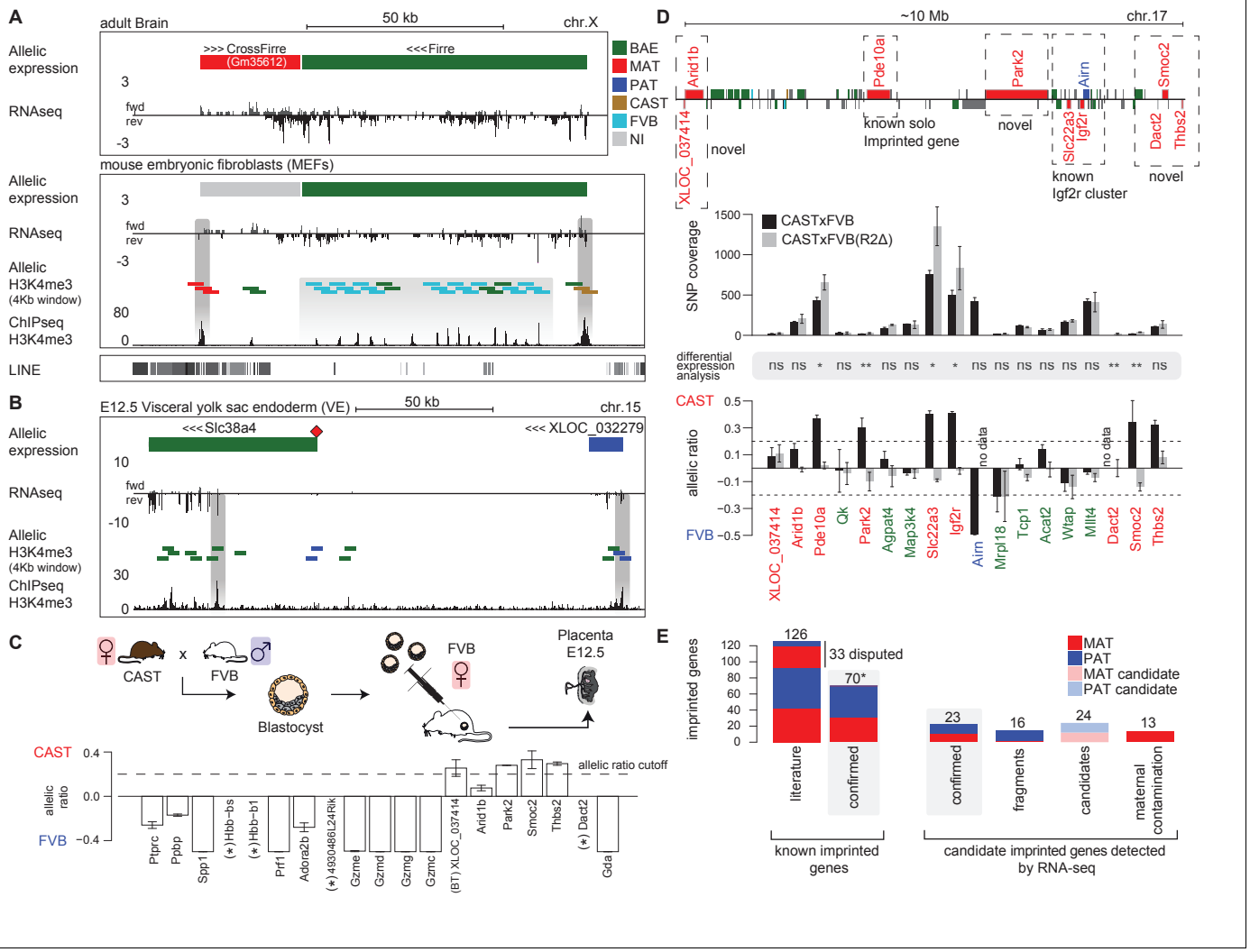

Figure 5. Novel validated imprinted genes belong to known clusters. (**A**) A novel X-linked imprinted nc-gene is transcribed anti-sense to *Firre* lncRNA. Maternal imprinted expression of *CrossFirre* (*Gm34612*) detected from RNA-seq in adult brain (top) was validated by maternal H3K4me3 enrichment in MEFs (middle). The gene body of *CrossFirre* is enriched for LINE repetitive elements (bottom). Highlighted in grey are H3K4me3 peaks over the *CrossFirre* promoter (MAT), the canonical *Firre* promoter (CAST), and multiple peaks in the *Firre* gene body (FVB) (UCSC genome browser screenshot). (**B**) *Slc38a4* forms a cluster with a novel imprinted lncRNA. The *Slc38a4* promoter is associated with maternal DNA methylation of the gametic differentially methylated region (gDMR, red square). In E12.5 VE, biallelic expression of *Slc38a4* is associated with biallelic H3K4me3 enrichment over an alternative TSS (highlighted by grey bar). Paternal expression of the novel upstream imprinted lncRNA XLOC_032279 was validated by paternal H3K4me3 enrichment (UCSC genome browser screenshot). (**C**) The allelic ratio in embryo-transferred placentas distinguishes maternal imprinted expression from maternal contamination. CAST (**F**) x FVB (**M**) blastocysts were transferred into pseudo-pregnant FVB females, and placentas collected at E12.5 and subject to ribosome RNA depleted RNA-seq. Genes showing maternal imprinted expression show a bias toward the maternal CAST allele. Genes expressed in maternal blood or decidua (maternal contamination) show a bias toward the FVB allele of the host mother. Novel imprinted genes that showed imprinted expression only in placenta or visceral yolk sac endoderm are displayed. Dotted line indicates 0.7 allelic ratio cutoff, (*) indicates genes with too low SNP coverage to be informative in this analysis. (**D**) The *Airn* promoter deletion (R2Δ) demonstrates that genes over a 10 Mb region are subject to imprinted silencing by *Airn*. First row: Known and novel imprinted genes detected by Allelome.PRO analysis of RNA-seq from E12.5 placenta in 10 Mb region surrounding the known *Igf2r* cluster (UCSC genome browser screenshot). Second row: The SNP coverage (reads over SNPs) of imprinted genes and selected biallelic controls between *Arid1b* and *Thbs2* on chromosome 17 for CASTxFVB and CASTxFVB(R2Δ) E12.5 placentas. Third row: Differential expression analysis calculated using Cuffdiff (version 2.2.1) ** adjusted p-value<0.01* adjusted p-value<0.05, ns non-significant. Fourth row: The allelic ratio (median and standard deviation) for the same genes calculated using the Allelome.PRO pipeline (0.5 = 100% maternal and −0.5 = 100% paternal expression). (**E**) A summary of known and candidate imprinted genes confirmed in this study. After validation, imprinted genes from the literature were confirmed or not (RefSeq genes listed by the Otago or Harwell imprinted databases on 24th Sept 2015 [***Glaser et al., 2006***;

Figure 5 continued on next page

Figure 5 continued

Williamson et al., 2013]). * 70 confirmed known genes include *Meg3*, *Rian* and *Mirg*, which our data indicates is a single imprinted transcript (***Figure 5—figure supplement 1B***). Novel imprinted candidates were classified as 'confirmed' (validated or supported), 'fragments' (supported nc-genes near known imprinted genes without evidence they are independent transcripts), 'candidates' (detected in one tissue by RNA-seq without further supporting evidence), and 'maternal contamination' (bias toward host mother allele in embryo transfer and/or expression in maternal decidua and blood). The 23 confirmed novel imprinted genes are shown in ***Table 1***. Classification of imprinted gene candidates following validation is further explained in the text.

DOI: 10.7554/eLife.25125.013

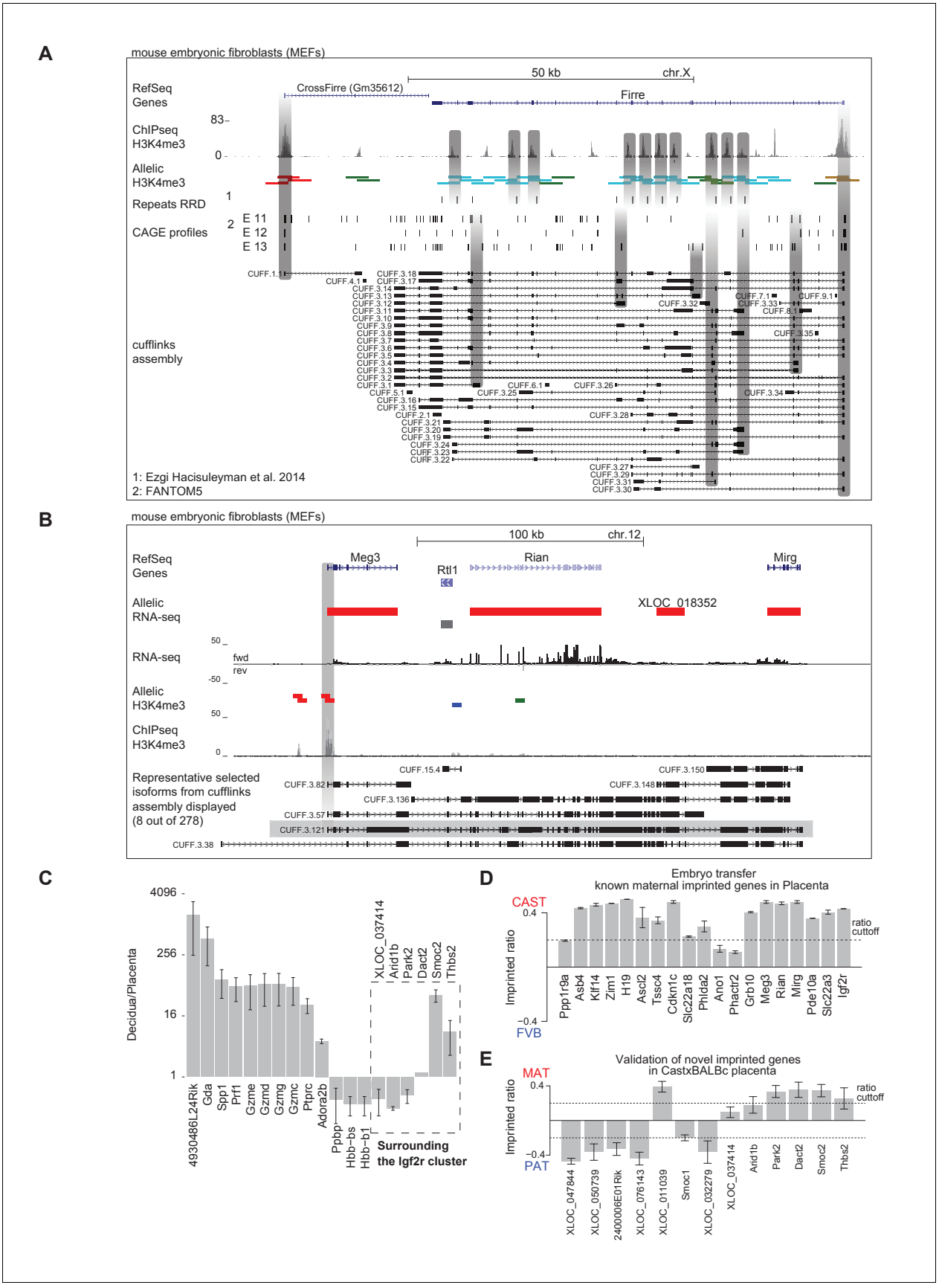
A
mouse embryonic fibroblasts (MEFs)
50 kb
chr.X
RefSeq Genes
CrossFirre (Gm35612)
Firre
ChIPseq H3K4me3
Allelic H3K4me3
Repeats RRD
CAGE profiles
E 11
E 12
E 13
cufflinks assembly
1: Ezgi Hacisuleyman et al. 2014
2: FANTOM5
B
mouse embryonic fibroblasts (MEFs)
100 kb
chr.12
RefSeq Genes
Meg3
Rtl1
Rian
Mirg
Allelic RNA-seq
XLOC_018352
RNA-seq
fwd
rev
Allelic H3K4me3
ChIPseq H3K4me3
Representative selected isoforms from cufflinks assembly displayed (8 out of 278)
C
Decidua/Placenta
4930486L24Rik
Gda
Spp1
Prf1
Gzme
Gzmd
Gzmg
Gzmc
Ptprc
Adora2b
Ppbp
Hbb-bs
Hbb-b1
XLOC_037414
Arid1b
Park2
Dact2
Smoc2
Thbs2
Surrounding the Igf2r cluster
D
Embryo transfer
known maternal imprinted genes in Placenta
CAST
FVB
Imprinted ratio
ratio cuttoff
Ppp1r9a
Asb4
Klf14
Zim1
H19
Ascl2
Tssc4
Cdkn1c
Slc22a18
Phlda2
Ano1
Phactr2
Grb10
Meg3
Rian
Mirg
Pde10a
Slc22a3
Igf2r
E
Validation of novel imprinted genes
in CastxBALBc placenta
MAT
PAT
XLOC_047844
XLOC_050739
2400006E01Rik
XLOC_076143
XLOC_011039
Smoc1
XLOC_032279
XLOC_037414
Arid1b
Park2
Dact2
Smoc2
Thbs2

Figure 5—figure supplement 1. Validation of novel imprinted genes. (**A**) The lncRNA *Firre* locus contains a full-length CAST biased transcript sense overlapped by multiple shorter FVB biased transcripts. The *Firre* locus in MEFs has a CAST biased H3K4me3 ChIP-seq peak at its canonical 5´end, and multiple internal FVB biased H3K4me3 peaks. 10/13 of the internal FVB H3K4me3 peaks overlap with the previously reported RRD repeats (***Hacisuleyman et al., 2014***). The *Firre* locus contains multiple cap analysis gene expression (CAGE) profiles from the FANTOM5 consortium, indicating transcription start sites (TSS). Cufflinks assembly of the MEF RNA-seq data finds transcripts with the canonical TSS plus alternative transcripts indicating six alternative TSS, all of which overlap FVB biased H3K4me3 peaks (see Materials and methods for details of alignment and Cufflinks assembly). (**B**) Known imprinted lncRNAs *Meg3*, *Rian*, and *Mirg* may form single maternally imprinted transcript in MEFs. Maternal enrichment of H3K4me3 is found over the *Meg3* TSS, but not the annotated *Rian* and *Mirg* TSS. Cufflinks assembly from RNA-seq data indicates that transcripts can extend over the entire *Meg3*/*Rian*/*Mirg* locus. (**C**) Maternal tissue in placental samples may result in false-positive maternal imprinted expression. The mean decidua/placenta RPKM ratio (3x E12.5 decidua and placenta) for extra-embryonic specific candidate maternal imprinted genes (19 genes) presented on a log(2) scale. The error bar represents the standard deviation between replicates. (**D**) Embryo transfer confirms maternal imprinted expression of known imprinted genes detected in placenta. Allelic ratio from E12.5 placentas from CAST (F) x FVB (M) embryos transferred into FVB host mother shown. (**E**) Novel placental imprinted genes are detected in BALBc x CAST reciprocal crosses. The imprinted ratio of novel placenta imprinted genes detected from ribosome RNA depleted RNA-seq from E12.5 BALBc/CAST placentas is shown. Dotted line indicates 0.7 allelic ratio cutoff.
DOI: 10.7554/eLife.25125.014

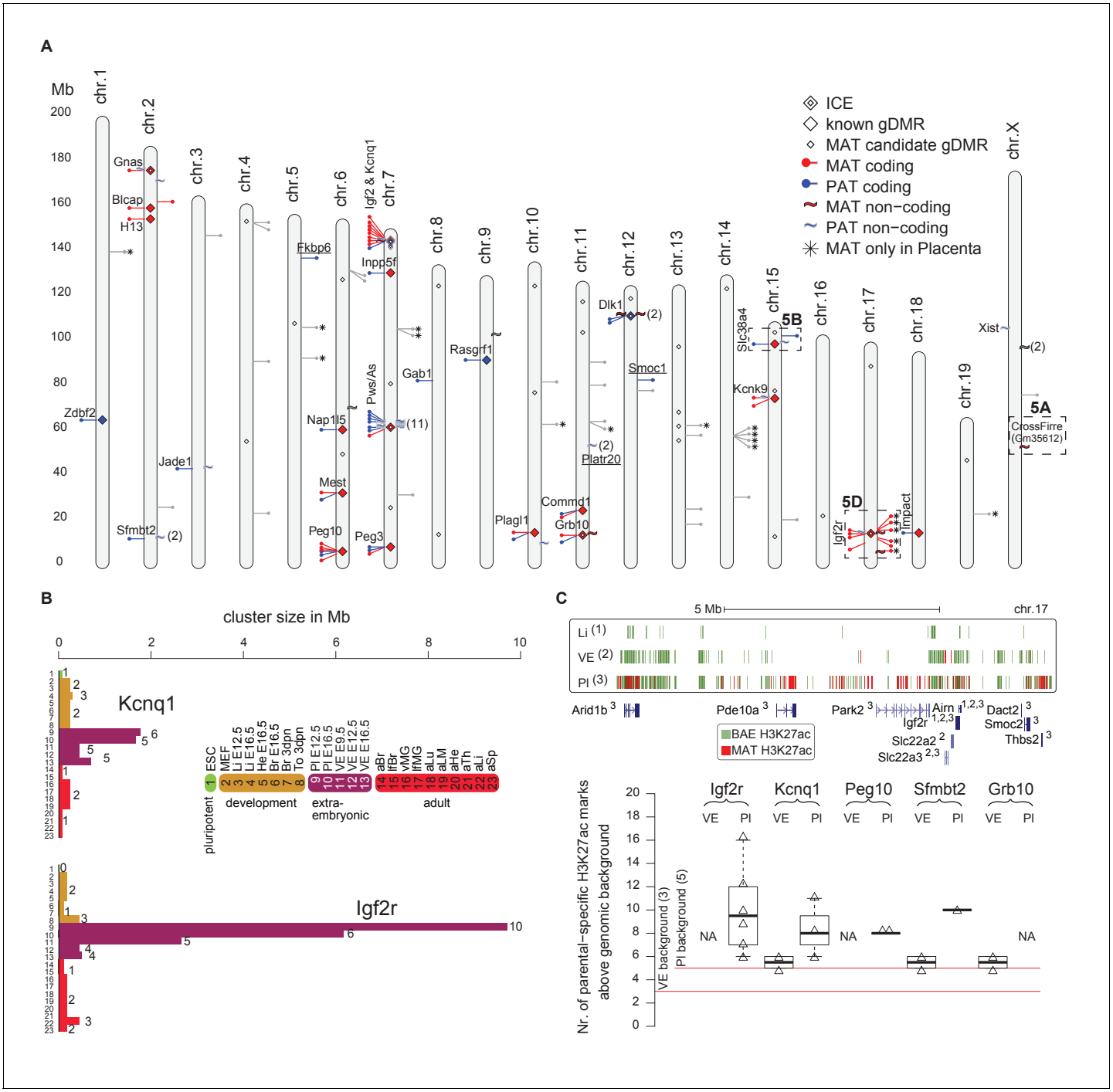

Figure 6. Tissue and developmental-specific expansion and contraction of imprinted clusters correlates with parental-specific histone modification. (**A**) A summary of imprinted genes detected in this study. Mouse chromosomes with the positions of known (left side of the chromosome) and novel supported or validated (right side of the chromosome) imprinted pc (–) and nc (~) genes. Candidate imprinted genes that are not supported or validated are indicated in grey. Imprint control elements (ICE), known and candidate gDMRs are indicated (***Proudhon et al., 2012***; ***Xie et al., 2012***). * Indicates maternally expressed genes restricted to placenta. The base pair coordinates (Mb) are indicated on the left side. Underlined are new imprinted clusters. Dashed boxes indicate the *Crossfirre* locus shown in ***Figure 5A***, the *Slc38a4* cluster shown in ***Figure 5B***, and the *Igf2r* cluster shown in ***Figure 5D***. Color code as in ***Figure 1***. For more details see Materials and methods and ***Supplementary file 1***, sheets G-J. (**B**) The *Igf2r* and *Kcnq1* cluster size during development and between tissues (tissue abbreviations as in ***Figure 1***). The number of imprinted genes for each developmental stage/tissue is indicated at the top of the bar. (**C**) Top: Allelic H3K27ac enrichment (4 kb sliding windows) over the expanded *Igf2r* cluster for E12.5 Liver, VE and placenta (UCSC genome browser screenshot). Numbers indicate tissue where a gene shows imprinted expression. Bottom: The number

Figure 6 continued on next page

Figure 6 continued

of parental-specific H3K27ac 4 kb sliding windows within non-overlapping 100 kb count windows for E12.5 VE and placenta (Pl). Counts over the background cutoff are shown (defined as the maximum count detected outside of imprinted regions for each tissue). For more details see Materials and methods. NA = not available (no parental-specific windows available for analysis).

DOI: 10.7554/eLife.25125.016

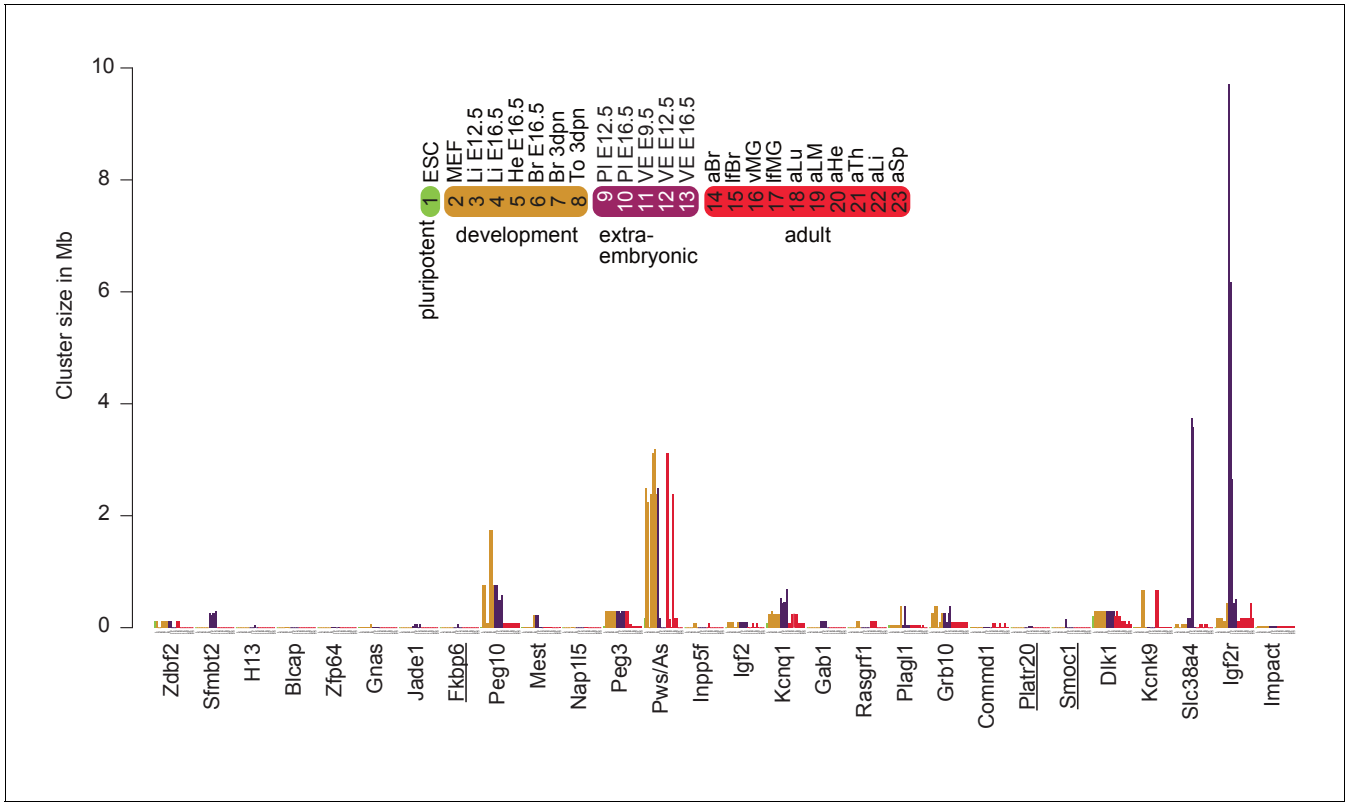

Figure 6—figure supplement 1. Expansion and contraction of imprinted clusters during development. The cluster size in Mb of confirmed and novel validated/supported imprinted regions was plotted. Underlined names indicate novel imprinted regions.
DOI: 10.7554/eLife.25125.017