Supplementary Information for Analysis of allelic expression patterns in clonal somatic cells by single-cell RNA-seq

Björn Reinius[1,2,4], Jeff E. Mold[1,4], Daniel Ramsköld[1,2], Qiaolin Deng[2], Per Johnsson[2], Jakob Michaëlsson[3], Jonas Frisén[1] and Rickard Sandberg[1,2,*]

[1]Department of Cell and Molecular Biology, Karolinska Institutet, 171 77 Stockholm, Sweden. [2]Ludwig Institute for Cancer Research, 171 77 Stockholm, Sweden. [3]Center for Infectious Medicine, Department of Medicine, Karolinska Institutet, Karolinska University Hospital Huddinge, 141 86 Stockholm, Sweden.

[4]these authors contributed equally
*correspondence to: Rickard Sandberg (Rickard.Sandberg@ki.se)

Contents

Supplementary Figure Legends

Supplementary Figure 1 | Cell-type-specific gene expression in somatic mouse cells.
(**a**) Clustering by principal component analysis (PCA) using all genes. (**b**) Number of detected Ensembl genes (RPKM>1) per cell, shown as boxplots. (**c**) Expression (RPKM) of three neuronal marker genes in single cells, shown as boxplots. (**d**) Expression (RPKM) of three hepatocyte marker genes in single cells, shown as boxplots. (**e**) Expression (RPKM) of three fibroblast marker genes in single cells, shown as boxplots. (**f**) PCA on liver cells using hepatocyte marker genes, showing two outlier cells with low marker expression, which were consequently considered not to be hepatocytes.

Supplementary Figure 2 | Gene expression correlations between single-cell RNA-seq libraries from mouse cells.
Violin plots showing the distribution of Spearman correlations for all pair-wise comparisons of cells or split-cell fractions, with median values represented as circles and the 25th to 75th percentile marked as black bars.

Supplementary Figure 3 | Thresholds for allelic calls and SNP filtering.
We called allelic gene expression (as "biallelic", "C57/Reference monoallelic", "CAST/Alternative monoallelic", or "not detected") based on SNP-containing reads. For monoallelic calls we required at least 3 SNP-containing reads derived from an allele for it to be considered expressed and 98% or more of all SNP-containing reads being derived from the expressed allele. The figures investigate the accuracy of the criteria for making allelic calls based on X-chromosome (chrX) genes in male cells (that only carry a maternal X-chromosome, Xm). We therefore expect only monoallelic calls for the Xm allele, and treat biallelic calls and monoallelic calls to the wrong allele (paternal X-chromosome, Xp) as false calls.
(**a**) Histogram showing the fraction false (Xp-mapping) reads for all chrX genes (≥1 read over the whole male dataset for inclusion) on the x-axis and number of genes on the y-axis. (**b**) Boxplots showing percent X-chromosome genes with falsely called (Xp) allelic expression per cell on the y-axis at different read thresholds (1, 2, 3, 4, or 5 reads) over SNPs when to call an allele expressed. The percent false allele calls was calculated: $100x(X_{biallelic} + Xp_{monoallelic}) / (X_{biallelic} + Xp_{monoallelic} + Xm_{monoalleic})$ in which X denotes the number of X-linked genes with a particular allelic call. Very few false calls were made when requiring 3 or more SNP-spanning reads. (**c**) Filtering of reference SNPs detected by exome sequencing and expressed in the human T-cells, to remove SNPs with a strong genetic or imprinted allelic expression bias. The fraction reference (Ref) reads over the whole data set of T-cells is shown on the y-axis for each SNP, ordered on the x-axis. SNPs with fraction SNP reads >0.3 and <0.7 were included in the analyses of allelic expression in human T-cells, resulting in 1,010 heterozygous autosomal SNPs (806 unique genes) expressed in the donor's T-cells.

Supplementary Figure 4 | Probability of allelic calls in single cells.
(**a**) Fraction of genes with different allelic calls in fibroblasts (y-axis) as a function of gene expression level (x-axis). Genes were binning by their expression (RPKM) in single fibroblast cells, and the fraction of each allelic call in each expression-level bin was calculated. High-expressed genes were generally biallelic, while low-expressed genes more often had monoallelical expression. The red line, which shows the fraction of genes obtaining any allele-informative call, approached 1. Even at high expression levels a small fraction of genes in cells will by chance not have reads covering otherwise informative SNPs, which can occur when SNPs are located within alternative exons or UTRs that are not expressed in all cells. Importantly, in the analyses of our study we calculated the fractions of genes with monoallelic and biallelic expression against the fractions of genes with an allele-informative call within the cell. (**b**) Density plot (Kernel density estimation) of genes with allelic expression bias (fraction C57 reads) on the x-axis, shown for autosomal genes expressed above increasing RPKM thresholds (colored lines). Data from mouse fibroblasts. (**c**) Scatter plot of allelic expression bias (fraction C57 reads) as observed for individual genes in individual cells (each dot represents a gene in a single cell), with gene expression level (RPKM) on the x-axis. The scatter plot shows the data of autosomal genes from 30 randomly selected individual fibroblast cells.

Supplementary Figure 5 | Coherence in allelic information among separate SNPs within the same gene.
To test the coherence in allelic detection for separate SNPs located within the same genes we sampled two random SNPs (separation > 43bp to avoid SNPs detected within the same sequence read) for each multi-SNP-containing autosomal gene expressed in fibroblasts. The SNP sampling was performed

independently for each gene and cell (using the RNA-seq libraries from non-clonal fibroblasts), so that diverse SNP pairs for each gene were sampled over the cells. The figures show density plots of the fraction C57-derived reads for such sampled SNP pairs (SNP1 and SNP2), and demonstrate the reproducibility in the allelic detection over separate SNPs within the same cell and gene. (**a-e**) The subpanels show the results when requiring an increasing number of SNP-informative reads spanning the sampled SNPs. Raising the minimum-reads threshold pushes the analysis to include an increasing fraction high-expressed genes, which tend to be have biallelic expression, and thus intermediate (biallelic) states become more pronounced at higher thresholds. The SNP-to-SNP reproducibility was high already at the minimum of 3 reads (rho=0.67, $P<2.2x10^{-16}$). rho: Spearman's rank correlation coefficient. *P*-value: rank-based measure of the association between paired SNPs using asymptotic *t* approximation.

Supplementary Figure 6 | Overview of the aRME analysis of single-cell data.
Schematic overview of the workflow in the analyses of allelic expression, included to accompany the descriptions in **Online Methods** and in **Supplementary Note 2**.

Supplementary Figure 7 | Allelic expression of the X chromosome and two autosomes in clonal mouse fibroblasts.
Allelic calls for genes expressed above mean RPKM 1 or 20 in clonally related cells, shown for chromosome X, 1, and 12. Rows represent cells, grouped by their clonal identity, as indicated by the color code to the left: cl1 (female CAST x C57, light blue), cl2 (female CAST x C57, orange), cl3 (female CAST x C57, brick red) cl4 (female CAST x C57, purple), cl5 (female CAST x C57, cyan), cl6 (male C57 x CAST, beige), cl7 (female CAST x C57, dark blue). Columns represent genes, ordered by their position on the chromosome, and the allelic call for each gene and cell is indicated by color: biallelic (green), C57 monoallelic (blue), CAST monoallelic (red), SNP not detected or not available (gray). The purple lines to the right of chr1 and chr12 exemplify cells with chromosomal abnormalities, and such cells were excluded from the allelic analyses of the affected chromosomes. Such abnormalities were either the complete loss of one parental chromosome (as in chr1: fibroblast.23) or the apparent breakage of one parental X-chromosome copy (as in chr12: fibroblast.36, with the breakpoint indicated by an orange line). (**a-c**) Chromosome X (chrX), included for demonstrating the accuracy in allelic detection by single cell RNA-seq. Female cells (cl1-5 and cl7) express one parental X-chromosome in a clonal manner due to X-chromosome inactivation (XCI), and male cells (cl6) only express the maternal X chromosome, as expected. *Xist* ("master regulator of XCI") is expressed from the silenced X-chromosome in female cells, and is not expressed in male cells. Genes that escape XCI show biallelic expression or expression from the silenced X-chromosome copy, as exemplified by *Xist* and the *5530601H04Rik/2610029G23Rik* locus which are highlighted in panel (c) along with downstream genes included for comparison (Xm: maternal chrX, Xa: active chrX). (**d-f**) Chromosome 1, included for demonstrating the presence of few, and low-expressed, clonal aRME genes in a landscape dominated by dynamic aRME. All genes classified as undergoing clonal aRME are named along the chromosome, and two of them are further highlighted in panel (f) along with two downstream genes included for comparison. Note that no clonal aRME gene on chr1 was expressed above 20 RPKM. (**g-i**) Chromosome 12, included for demonstrating that the assay could correctly identify imprinted genes expressed in fibroblasts, according to their parental-specific monoallelic expression in single cells. Genes in the *Meg3/Rian/Mirg* imprinted locus are highlighted in (i) along with two downstream genes.

Supplementary Figure 8 | Properties of genes that escape X-chromosome inactivation.
(**a**) The fraction C57-derived reads in individual female and male fibroblasts, shown separately for chromosome X (chrX) and autosomes. This demonstrates random X-chromosome inactivation (XCI) in female cells (in the chrX panel) as well as the lack of any strong mapping-bias towards the C57 or CAST genomes (in the autosomes panel). The *P*-value (two-sided Wilcoxon test) indicates that although female cells have one inactive X chromosome there is still significant transcription occurring from this "inactive" X copy, as expected due to genes escaping XCI. (**b**) The percent of female fibroblasts having either the C57 chrX or the CAST chrX active, among randomly picked (non-clonal) fibroblasts. The CAST chrX constituted the active X copy (Xa) in a significantly higher proportion of cells than the C57 chrX. *P*-value according to a binominal test. (**c**) The percent of allele-informative reads derived from C57 and CAST shown for chrX and autosomes in bulk RNA extracted from a female C57 x CAST fibroblast culture, and sequenced using strand-specific mRNA Truseq. This population analysis of allelic gene expression validates the CAST chrX dominance observed using single-cell transcriptomics presented in panel (b). (**d**) *Xist* expression levels in individual female and

male fibroblasts. *Xist* was detected in female cells, in line with the occurrence of XCI. *P*-value according to a two-sided Wilcoxon test. (**e**) The mean percent of expressed genes (RPKM>1) on chrX with expression from the inactive X chromosome (Xi) in individual female fibroblasts. The mean percent of genes with allelic calls from the (non-existing) paternal X chromosome (Xp) in male cells is shown as an estimate of technical false positive detections. The error bar represents +/- SEM. (**f**) The percent of cells with Xi expression in female fibroblasts, i.e. "XCI escape events", shown for allele-informative genes along the X chromosome. Male data for Xp detection (blue dots) are included as a background of technical false positive detections for each gene. Genes with reproducible escape events over several cells are named and colored according to their Bonferroni-adjusted *P*-values, calculated for each gene using Fisher's exact test on the proportion of female cells with escape compared to the gene-specific background of technically false positive detection (proportion of male cells with Xp detection). Data for all assayed genes are presented in **Supplementary Table 2**. Cytoband A1.1 and D contained a significant enrichment (EASE Score < 0.05) of escapee genes given their gene densities. "Xic": X-inactivation centre. (**g**) Left: Boxplots of the fraction Xi-derived reads for *Xist* in all *Xist*–expressing female fibroblasts. Right: Boxplots of the fraction Xi-derived reads for XCI escapee genes, plotted specifically for cells that had simultaneous (biallelic) presence of RNA from both the Xa and the Xi alleles. The colors of the boxes indicate whether the genes had a tendency of higher expression from the Xa or Xi allele. *P*-value calculated with two-sided Wilcoxon tests versus equal expression. (**h**) We identified one gene, *Fundc1*, that escaped XCI in only 9% (6/69) of non-clonal female cells but in 100% (4/4) of cells in one of the clones (cl3). Given *Fundc1*'s escape frequency in non-clonal cells, the probability of independent escape events occurring in 4/4 clonal cells is $P=5.7\text{x}10^{-5}$ (binominal distribution, $P=0.027$ after adjustment for multiple testing considering all genes and clones). Given the above, the probability of *Fundc1* independently never escaping in any cell of the other female clones (0/78) is $8.3\text{x}10^{-4}$. (**i**) We tested whether female and male cells had different degrees of dynamic aRME, and found that their levels of aRME were on par ($P>0.05$, two-sided Wilcoxon test).

Supplementary Figure 9 | Visualization of the allelic detection of X-chromosome genes and imprinted genes.
(**a**) Sequence read coverage distributions over SNP-containing exons of an X-inactivated gene (*Pdk3*) and an XCI escapee gene (*2610029G23Rik*), shown for a bulk-population stranded mRNA TruSeq library ("Bulk") and for individual fibroblast cells. The gray graphs show read coverage and the pink/purple horizontal lines below each read-coverage plot represent individual reads aligned to the plus or minus strand. The color codes at the bottom indicate the four bases, and a colored vertical bar is shown across the reads and read-coverage plots when a non-reference (CAST-specific) base was present in the library. (**b**) Sequence read coverage distributions over SNP-containing exons of two previously known imprinted genes (*Sgce* and *Impact*), as in (a). (**c**) Barplots showing the number of reads derived from the maternal and paternal allele of *Sgce* and *Impact*, in the stranded mRNA TruSeq bulk-population sequencing. As expected, *Sgce* was exclusively paternally expressed. However, for *Impact* we observed a "leaky" expression from the presumed silenced maternal allele. This was not due to incorrectly annotated C57/CAST SNPs, since biallelic expression was observed also within SNPs, as demonstrated in (b). The strand specificity of the bulk mRNA TruSeq library in (b) further demonstrates that the maternal reads did not derive from antisense transcription on the maternal allele. (**d**) Barplots showing the number of cells with maternal monoallelic, biallelic and paternal monoallelic expression of *Sgce* and *Impact*, in single cell RNA-seq experiments, confirming leaky expression from the maternal allele for *Impact*.

Supplementary Figure 10 | Bias in allelic expression due to the alleles' genetic background.
(**a**) Genes, ordered and colored according to their chromosomal localization, with –log10 *P*-values for strain-specific expression (Fischer's exact test) on the y-axis. The dotted lines mark $P=0.05$. 25% of the genes (n=2540) had significant tendency ($P<0.05$, and ~1.25% explained by false positives) to preferentially transcribe from either the CAST or the C57 allele, and to variable degree in strength of the effect. (**b**) Gene scatter with fraction of cells with consistent maternal or paternal monoallelic (x-axis) plotted against the –log10 *P*-values (Fischer's exact test). (**c**) Boxplots showing that the degree dynamic aRME was on par in cells of the two reciprocal crosses (P-value<0.05, two-sided Wilcoxon test).

Supplementary Figure 11 | Visualization of the allelic detection of clonal aRME genes.
Sequence read coverage distributions over SNP-containing exons for four clonal aRME genes (*1700012B15Rik*, *Alkbh3*, *Fbxw17* and *Cck*), expressed at different levels (mean RPKM) in fibroblast cells; shown for three clone 6 (cl6) and three clone 7 (cl7) cells. The coverage in a bulk-population

stranded mRNA TruSeq library ("Bulk") is shown in the upper lanes. The gray graphs show read coverage, and the pink/purple horizontal lines below each read-coverage plot represent individual reads aligned to the plus or minus strand. The color codes at the bottom indicate the four bases, and a colored vertical bar is shown across the reads and read-coverage plots when a non-reference (CAST-specific) base is present in the library.

Supplementary Figure 12 | Fluorescence-activated cell sorting (FACS) of T-cells isolated from human blood following vaccination with YFV-17D.
(**a**) Gating strategy to identify and sort the HLA-B7 and HLA-A2 restricted YFV-specific $CD8^+$ T-cells. Single lymphocytes were detected by sequential gating on FSC-area versus FSC-height, FSC-area versus SSC-area, and FSC-area versus FSC-width, followed by gating on live, $CD14^-CD19^-CD3^+$ cells, and subsequently on $CD4^-$ T-cells. (**b**) Representative FACS scatter plots depicting the sort gate and frequency of HLA-B7 restricted and HLA-A2 restricted YFV-specific $CD8^+$ T-cells. During the early (acute) response to the vaccine, T-cell responses to the HLA-A2 restricted epitope were greater than that observed for the HLA-B7 restricted epitope. In the late phase (memory) of the immune response to the vaccine, the frequency of HLA-B7 restricted YFV-specific $CD8^+$ T-cell responses were reduced approximately 10-fold, typical of the contraction in T-cell responses after resolution of a viral infection, as indicated in the schematics in **Fig. 3a**.

Supplementary Figure 13 | Monoallelic expression in human T-cells at different expression level thresholds.
(**a**) Boxplots, showing percent autosomal genes with monoallelic expression in individual human *in vivo*-expanded T-cells (n=545) at different RPKM thresholds. (**b**) Boxplots, showing percent autosomal genes with monoallelic expression in individual human *ex vivo*-expanded T-cells (n=347) at different RPKM thresholds.

Supplementary Figure 14 | Percent monoallelic expression in acute and memory T-cells and the number of expressed genes per cell.
(**a**) Number of genes expressed (RPKM>1) in single T-cells collected from human blood during the acute phase (day 15, HLA-B7 cells) or memory phase (day 136, HLA-B7 cells), and in all T-cells. (**b**) Boxplots showing that the difference in degree monoallelic genes between *in vivo* T-cells from the acute phase (day 15, HLA-B7 cells) and memory phase (day 136, HLA-B7 cells) remained after re-sampling cells so that the two groups had the same distribution of number of mapped reads per cell. Thus, the observed difference in degree aRME between the acute and memory phase was not caused by differences in sequence depth. *P*-values: two-sided Wilcoxon test. (**c**) Histograms showing distributions of the number of unique mapped reads per cell, as described in (b).

Supplementary Figure 15 | Identification of genes with clonal aRME or allelic imbalance in human T-cells.
Genes, ordered and colored according to their chromosomal localization, with –log10 *P*-values for clone-consistent allele-specific monoallelic expression (Fischer's exact test) towards the reference ("Ref") or the alternative ("Alt") allele on the y-axis, shown for 9 *ex vivo* T-cell clones (clone A-I). The dotted lines mark $P=0.05$. The number of cells and eligible autosomal genes are indicated for each clone. All genes with P-value<0.05 are named (both strictly monoallelic genes and genes with allelic imbalance). Volcano scatter plots for each clone are shown in **Supplementary Fig. 16**.

Supplementary Figure 16 | Volcano scatter plots of clonal aRME genes in human T-cells.
Volcano gene scatter plots with fraction of cells with consistent monoallelic expression for the reference allele ("Ref") or the alternative allele ("Alt") on the x-axes, plotted against the –log10 *P*-values (Fischer's exact test) on the y-axes. All genes with $P<0.05$ are named (both strictly monoallelic genes and genes with allelic imbalance), and genes with $P<0.05$ and consistent monoallelic expression in ≥90% of cells were classified as clonal aRME genes. Clonal aRME genes with additional observed clonal aRME or allelic imbalance in a second clone are highlighted in yellow. The "E-values" to the right of each plot designate the expected number of false positive genes above the indicated *P*-value thresholds and clones, as determined by random sampling of same number of cells from all sequenced T-cells (mean over 300 permutations).

Supplementary Figure 17 | cDNA yield from large and small fibroblasts.
(**a**) cDNA yield per cell after 18 cycles of Smart-seq2 cDNA amplification (n large cells=22, n small cells=46). *P*-value according to a two-sided Wilcoxon test. (**b**) cDNA yield per splitted cell lysate after

18 cycles of Smart-seq2 cDNA amplification (n fractions from large cells=14, n fractions from small cells=18). The diameters of the dissociated cells were 25-35µm for large cells and 10-20µm for small cells. *P*-value according to a two-sided Wilcoxon test. (**c**) Sequencing depth (number of unique mapped reads per sequenced library) for large and small fibroblast cells or split-cell fractions. *P*-value according to a one-sided Wilcoxon test, testing for lower sequencing depth in small cells or in split fractions from small cells as compared to those of large cells.

Supplementary Figure 18 | Expression of fibroblast marker genes in large and small cells.
Boxplots showing the expression (RPKM) of three fibroblast marker genes in large fibroblasts (diameter of dissociated cell: 25-35μm, n=22 cells) and small fibroblasts (diameter of dissociated cell: 10-20μm, n=46 cells).

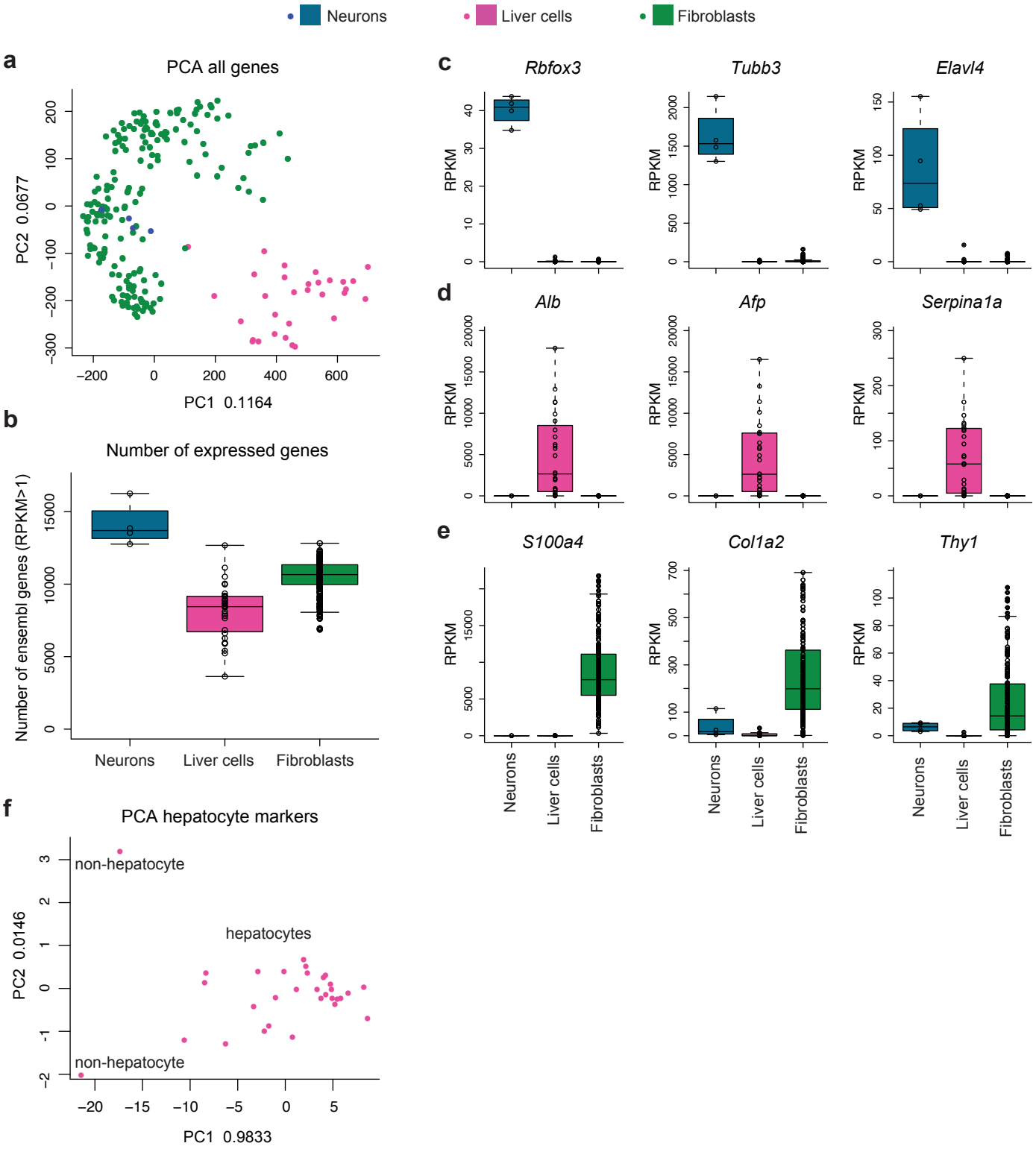
Supplementary Figure 19 | Split-cell analysis of large and small fibroblasts.
The figures show inferred monoallelic expression and allelic dropouts from split-cell experiments on large (diameter of dissociated cell: 25-35μm, n=7 pairs of split-cell fractions) and small fibroblasts (diameter of dissociated cell: 10-20μm, n=9 pairs of split-cell fractions). (**a-b**) Analytically computed percent monoallelic expression for genes, binned by gene expression (RPKM). Shown are means and standard deviations computed from the 7 and 9 individual split-cell experiments performed on large and small fibroblasts, respectively. (**c-d**) The percent allelic loss (i.e. the probability of losing all RNA molecules from an allele) for genes, binned by gene expression (RPKM). Shown are means and standard deviations computed from the 7 and 9 individual split-cell experiments performed on large and small fibroblasts, respectively. These experiment support a roughly 40-50% sensitivity in small and large fibroblasts respectively, since allelic losses at single RNA copy levels stabilize around 60 and 50% respectively. (**e**) Percent aRME in large and small fibroblast, when applying a stricter threshold (RPKM ≥28) on small fibroblasts and then assessing aRME expression of this exact same gene set in small and large fibroblasts.

Supplementary Figure 20 | The degree dynamic aRME correlates with the soma and nuclear size.
To further validate the relationship between cellular size and degree dynamic aRME, we dissociated a fibroblast culture and picked single cells of variable size under a fluorescence microscope. Before picking and lysing the cells in Smart-seq2 lysis buffer, we recorded the size (area) of the cellular soma and nucleus (visualized by Hoechst). (**a**) Examples of microscopy images for picked cells of various sizes. (**b**) Histograms of cellular soma and nuclear sizes. (**c-d**) Boxplots of the percent aRME in cells binned by the cells' soma or nuclei sizes. (**e**) Boxplots of cDNA yield per cell, for Smart-seq2 libraries, binned by the cells' soma sizes. rho: Spearman's rank correlation coefficient. *P*-value: rank-based measure of the association between degree aRME and soma size using asymptotic *t* approximation.

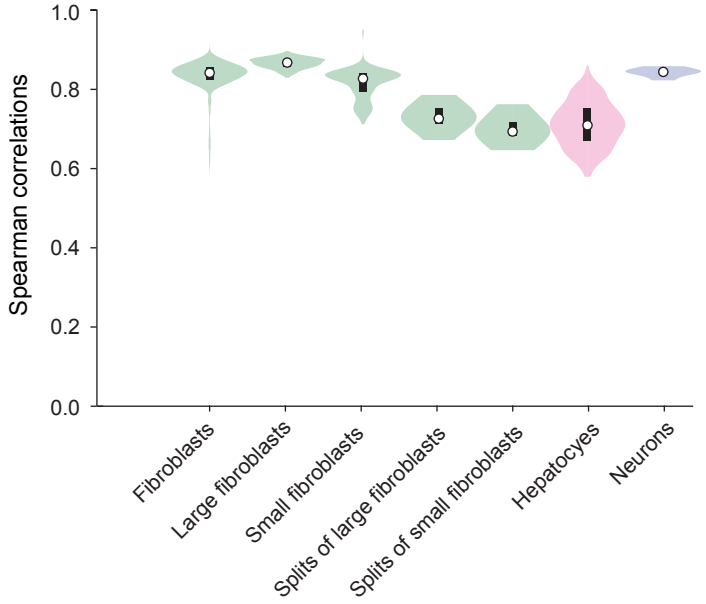
Supplementary Figure 21 | RNA content and monoallelic expression in mouse primary and *in vivo* cells.
Scatter plot (each ring represents a cell), showing the relationship between cellular RNA content and percent genes with monoallelic expression in the cell. The black curve shows a logistic curve fitted to the data points. The Spearman correlation is 0.77 ($P=4x10^{-40}$). This plot used the same set of genes for all cells, including genes expressed above an average of 20 RPKM across all cells.

Supplementary Figure 22 | Cell-cycle classification of mouse fibroblasts.
(**a**) A schematic depicting the different cell-cycle phases. (**b**) Identification of 100 cell-cycle genes presenting the greatest variability. A principal component analysis (PCA) plot using these most-variable cell-cycle genes identified three populations of cells, as shown in **Fig. 4d**, corresponding to G0, G1 and S/G2/M phases. (**c**) Upper row: The PCA plot in **Fig. 4d** with cells colored according to their expression levels of a subset of the cell-cycle marker genes (*Gas1, Top2a, Ccnd1, Ccne2* and *Ccna2*). Lower row: Boxplots showing the expression level of these genes in single cells of the different cell-cycle phases. *P*-values according to two-sided Wilcoxon tests. (**d**) Boxplots of the number of detected genes (RPKM>1) in fibroblasts from the G0, G1 or S/G2/M phases. *P*-values according to two-sided Wilcoxon tests. "ns": not significant (*P*-value>0.05). (**e**) Sequencing depth (number of unique mapped reads per sequenced library) for G0, G1 or S/G2/M cells. *P*-values according to a one-sided Wilcoxon test, testing for lower sequencing depth in G0 cells. (**f**) Boxplots of the degree aRME in large and small cells of the G0 cell-cycle phase. Together with the data in **Fig. 4a-b** this demonstrates that cellular size and cell-cycle phase are two separate cellular characteristics that both have impact on the level of dynamic aRME in the cell.

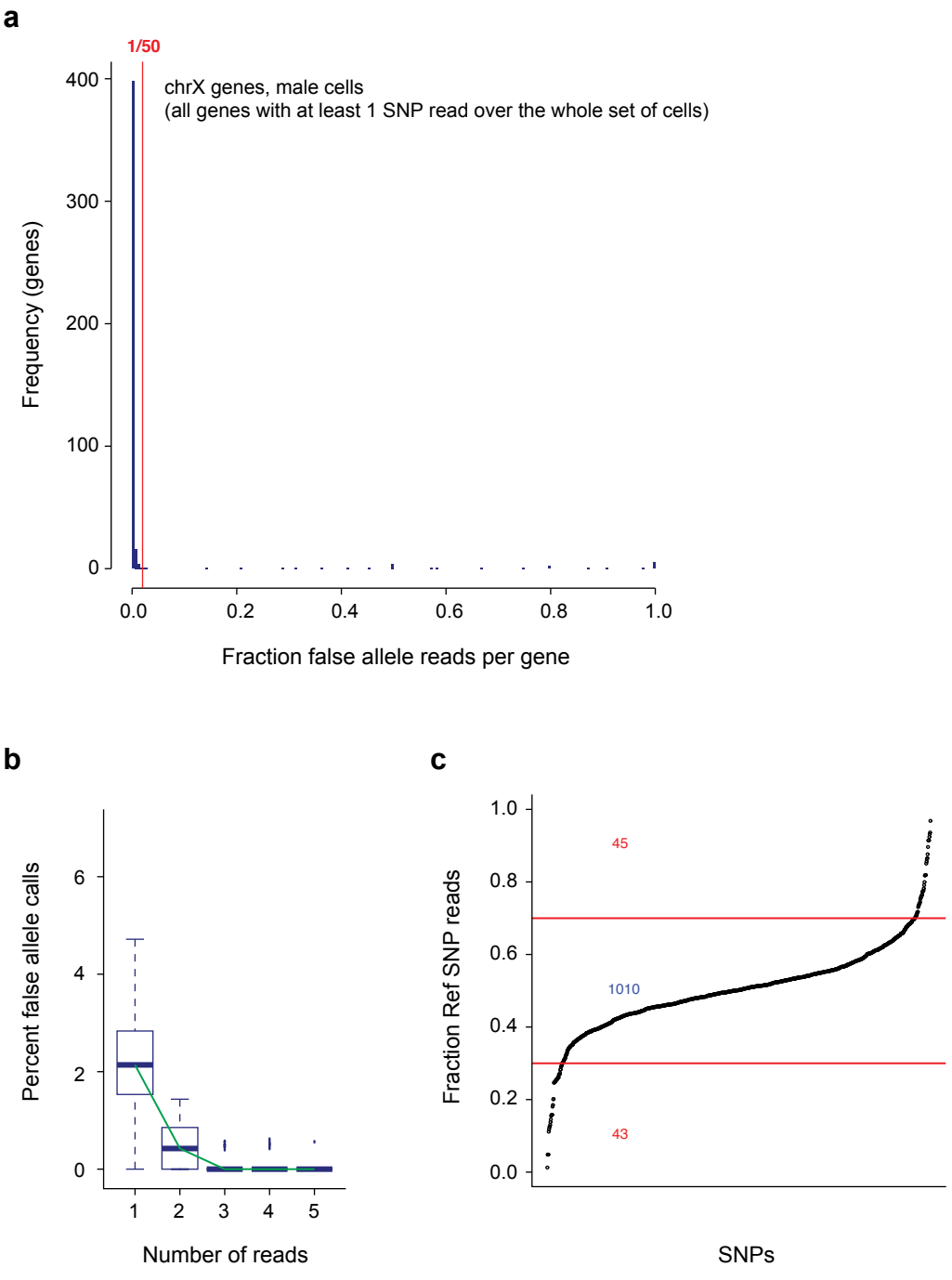
Supplementary Figure 1



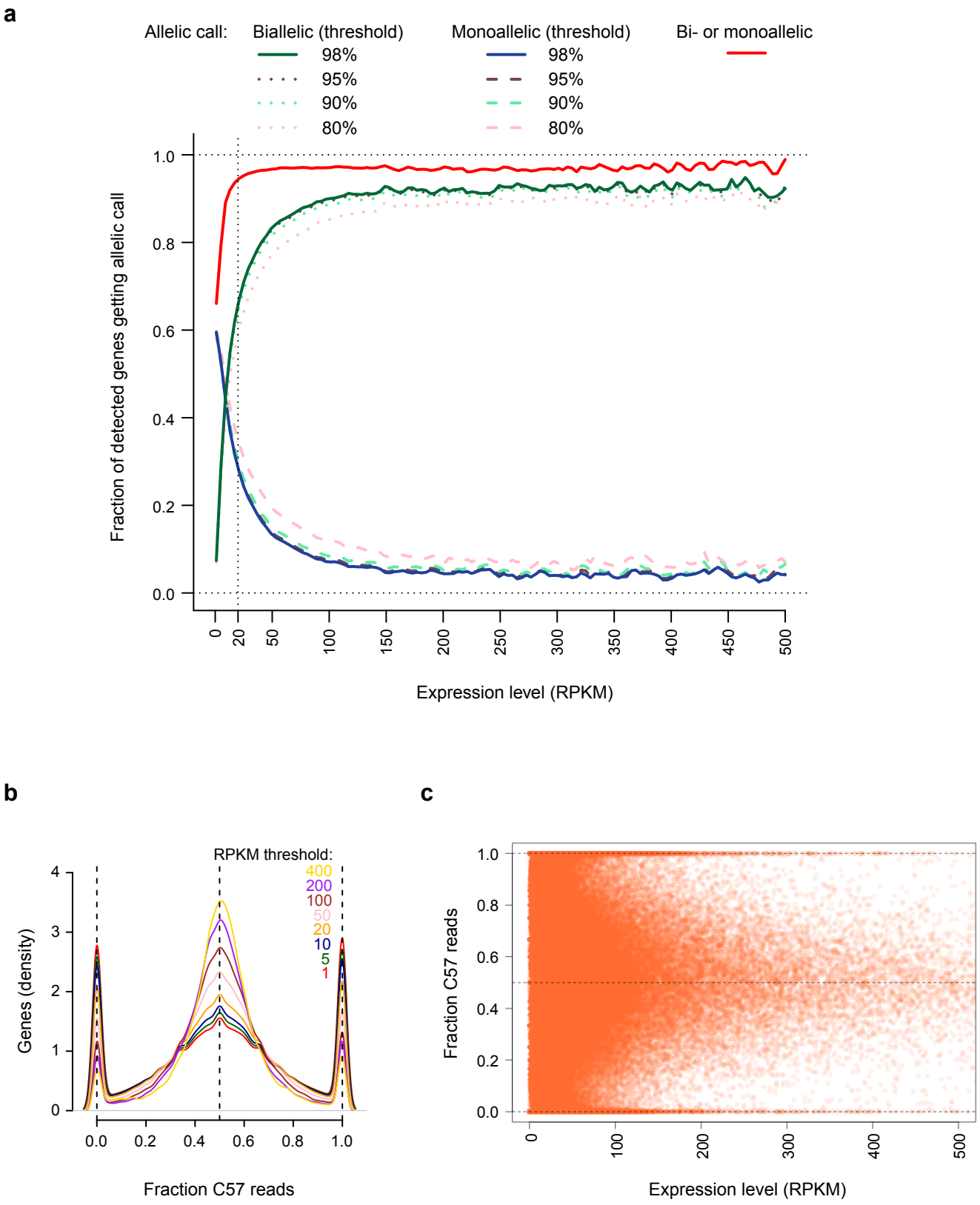
Supplementary Figure 2



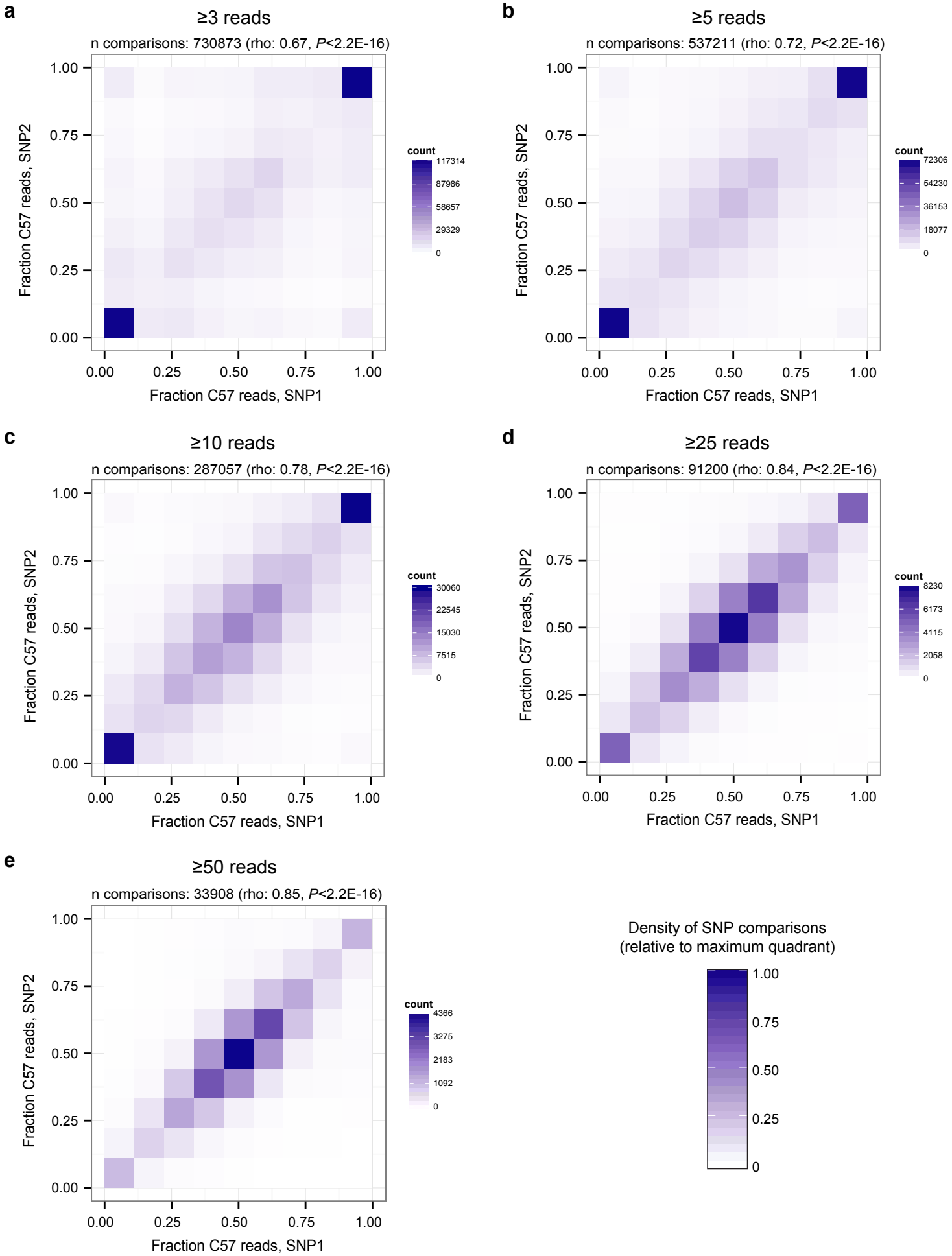
Supplementary Figure 3



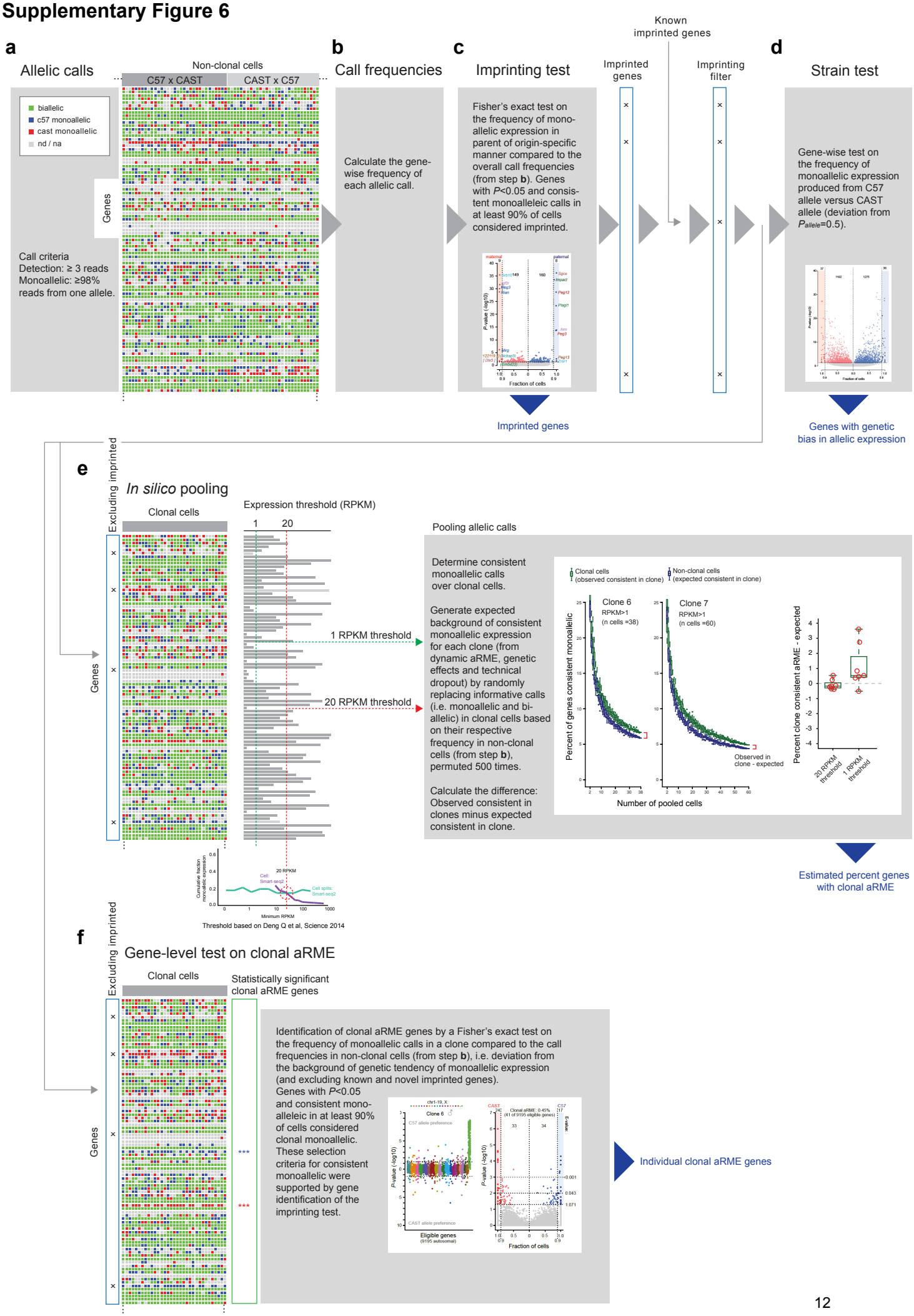
Supplementary Figure 4



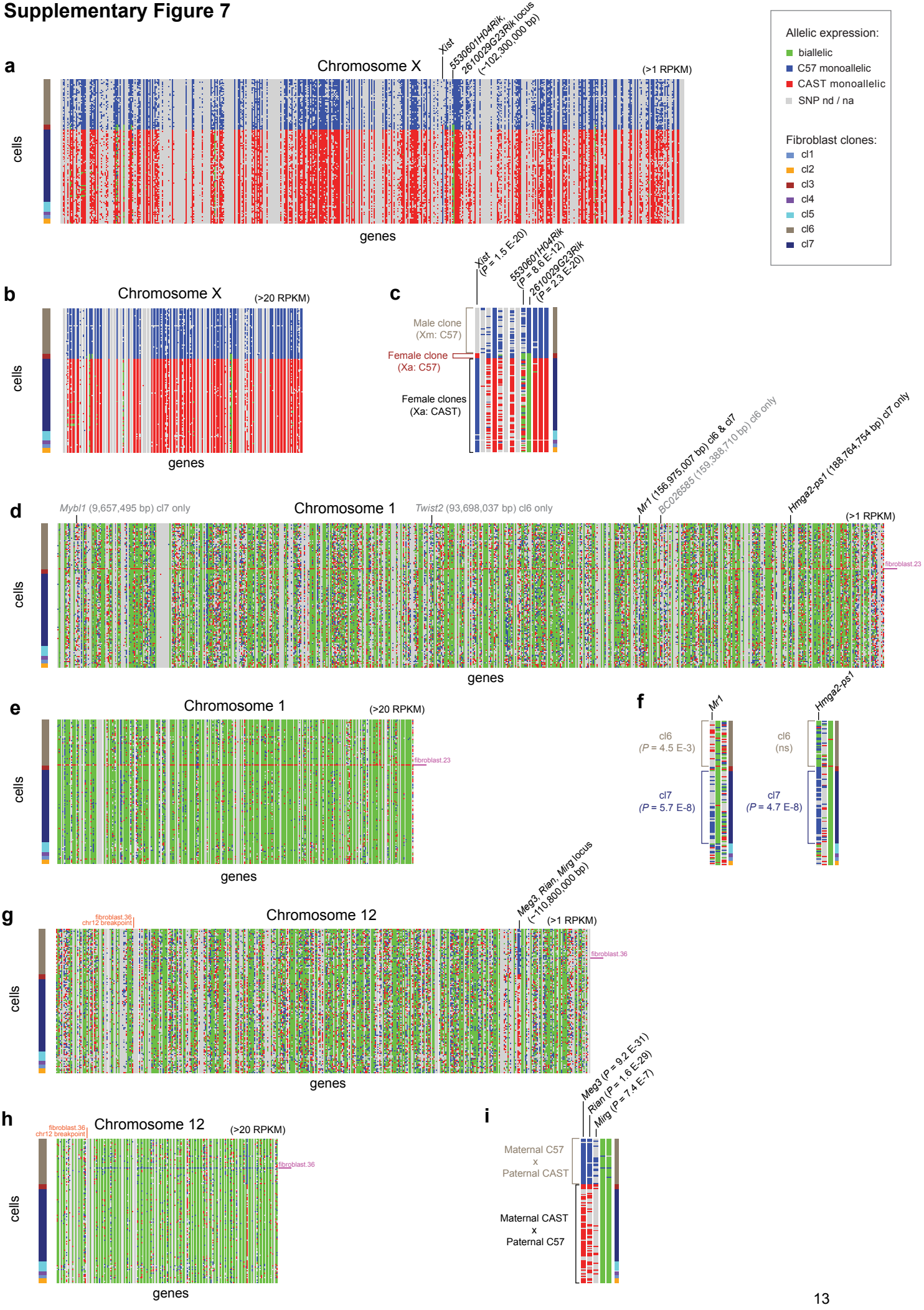
Supplementary Figure 5



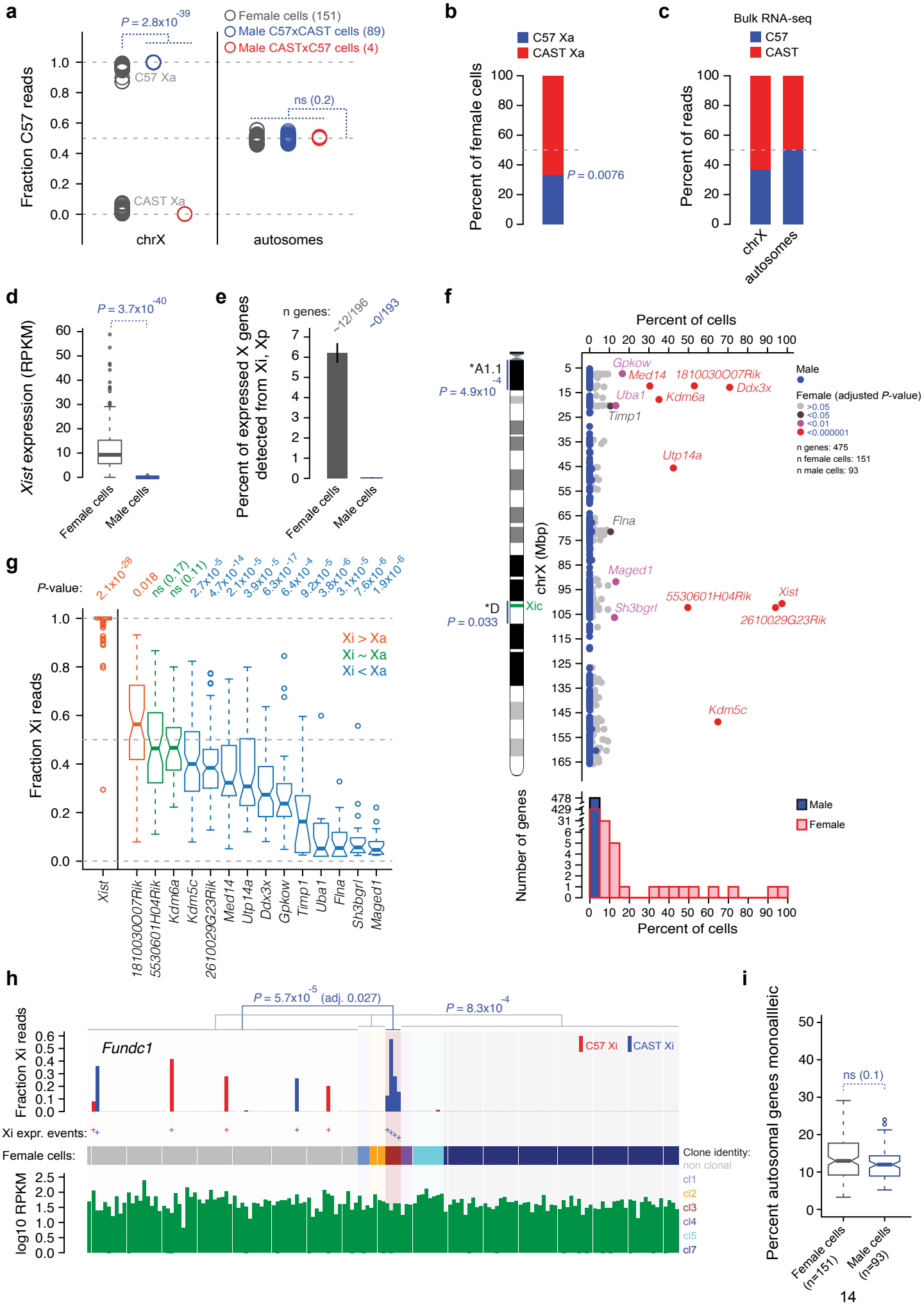
Supplementary Figure 6



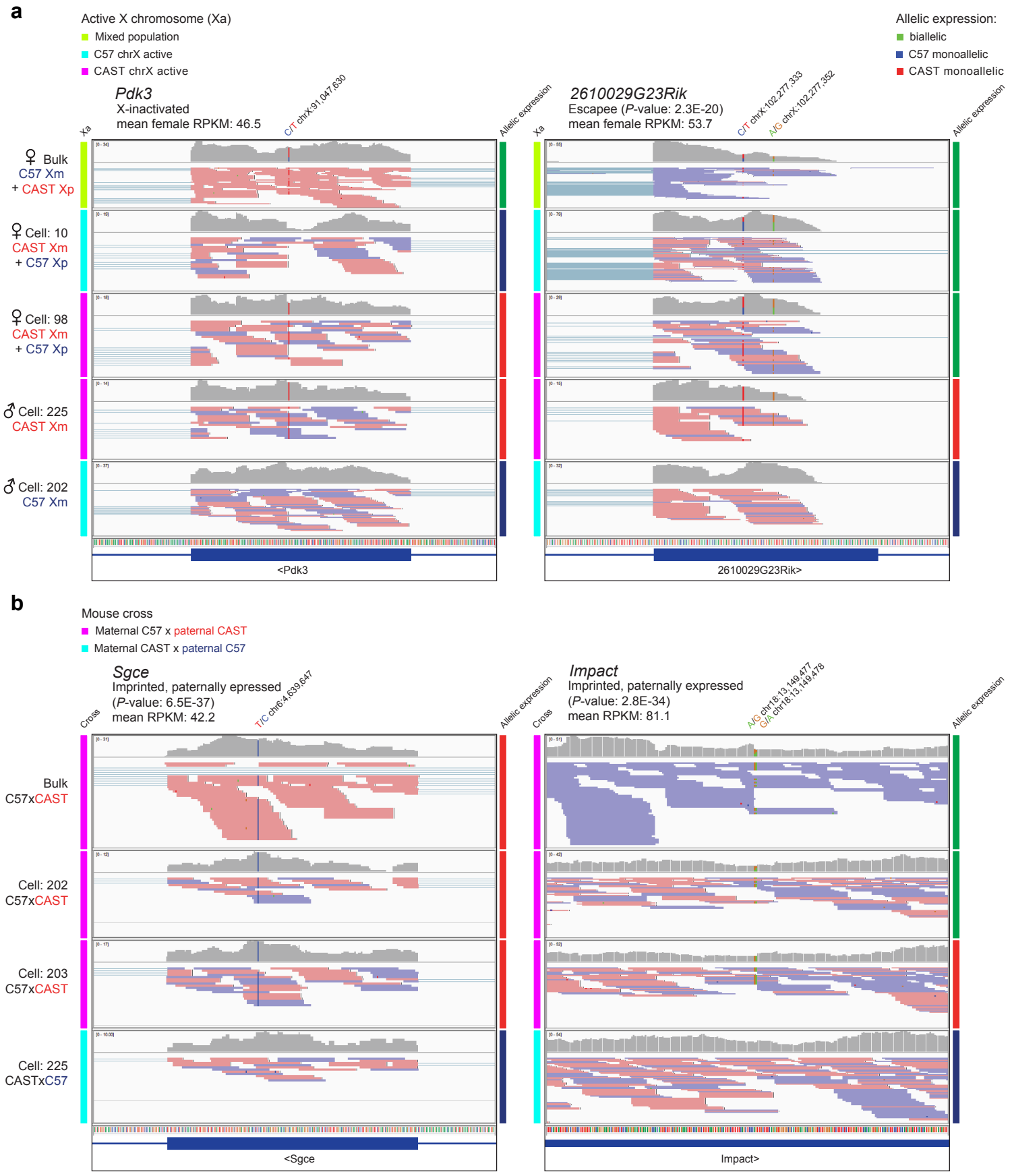
Supplementary Figure 7



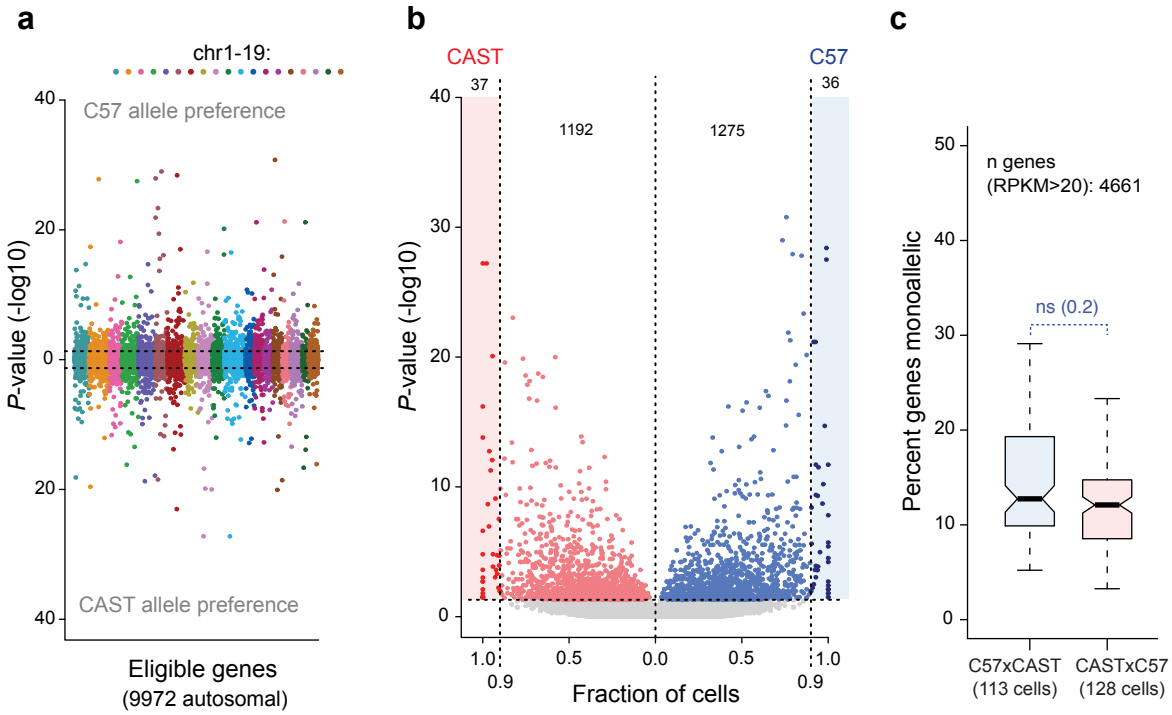
Supplementary Figure 8



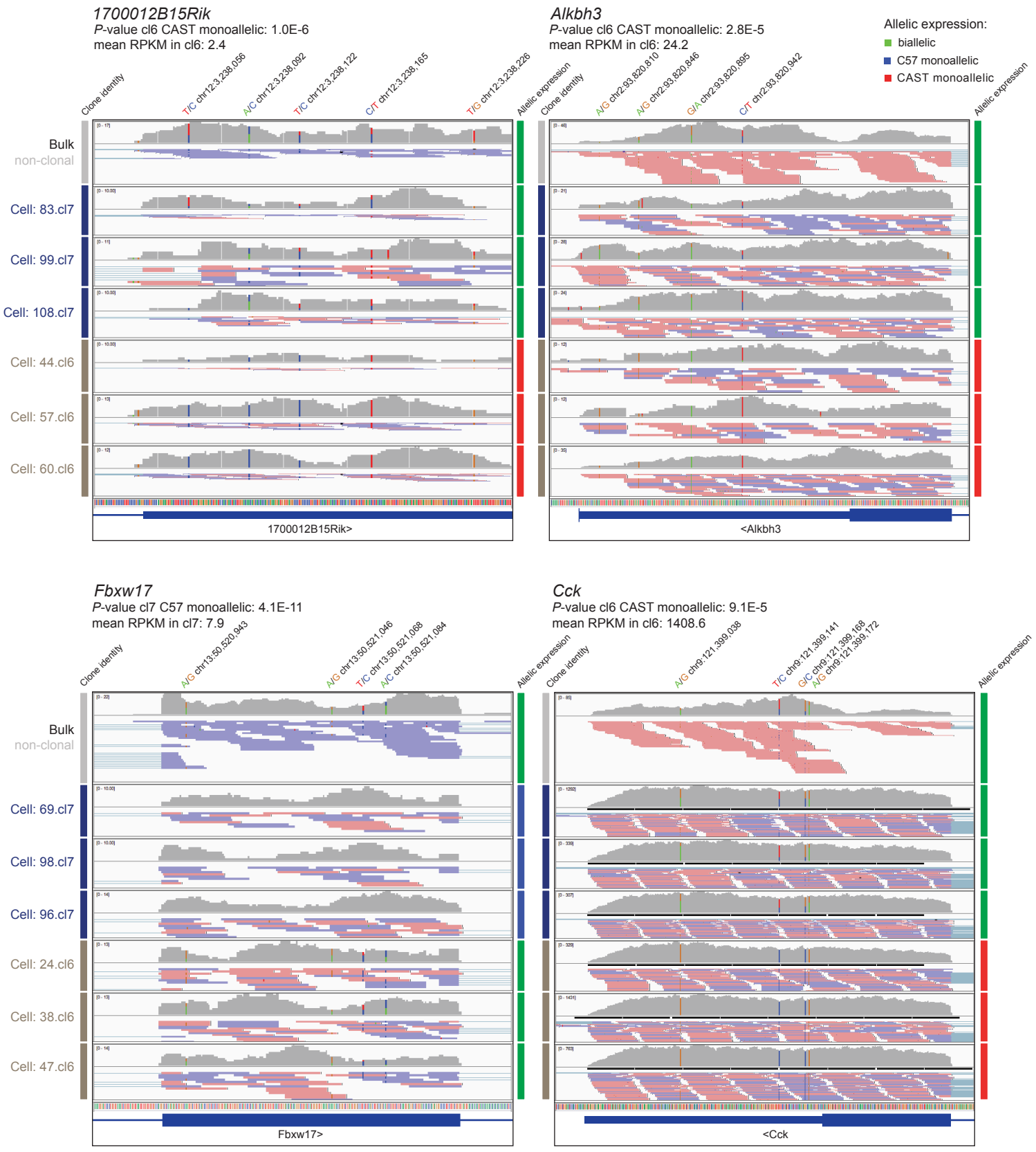
Supplementary Figure 9



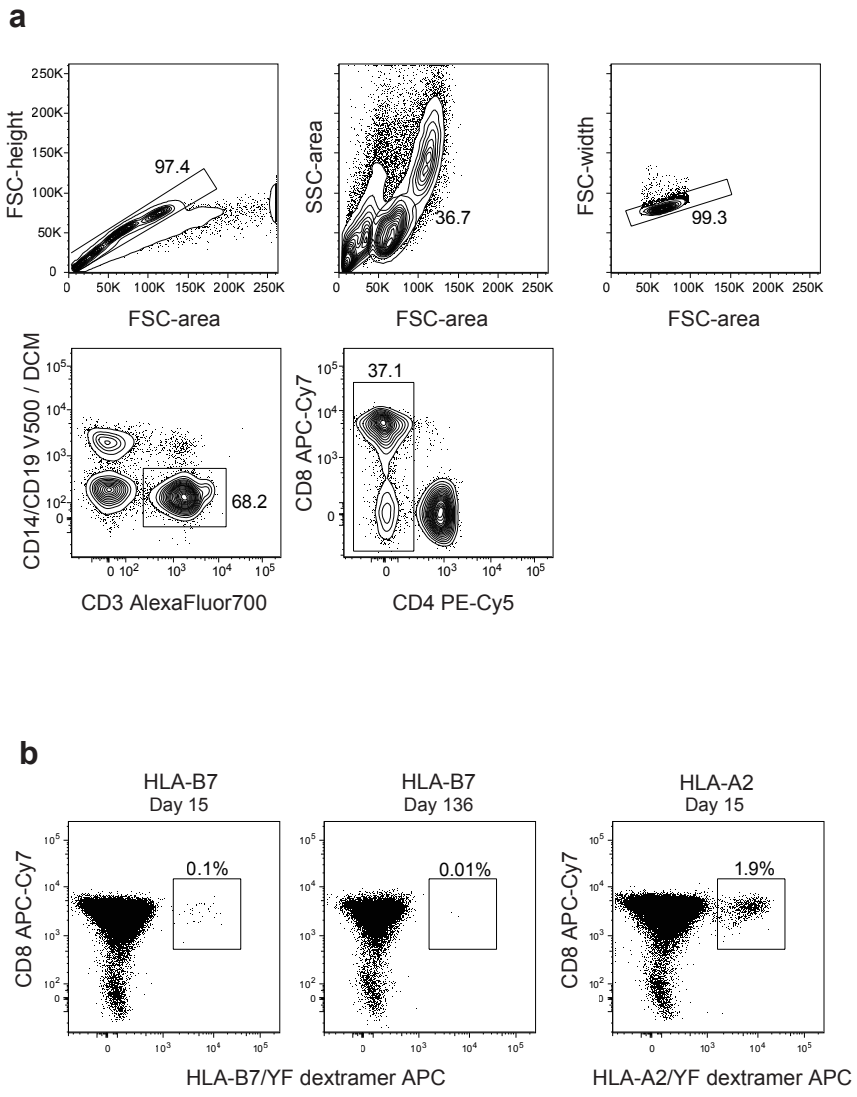
Supplementary Figure 10



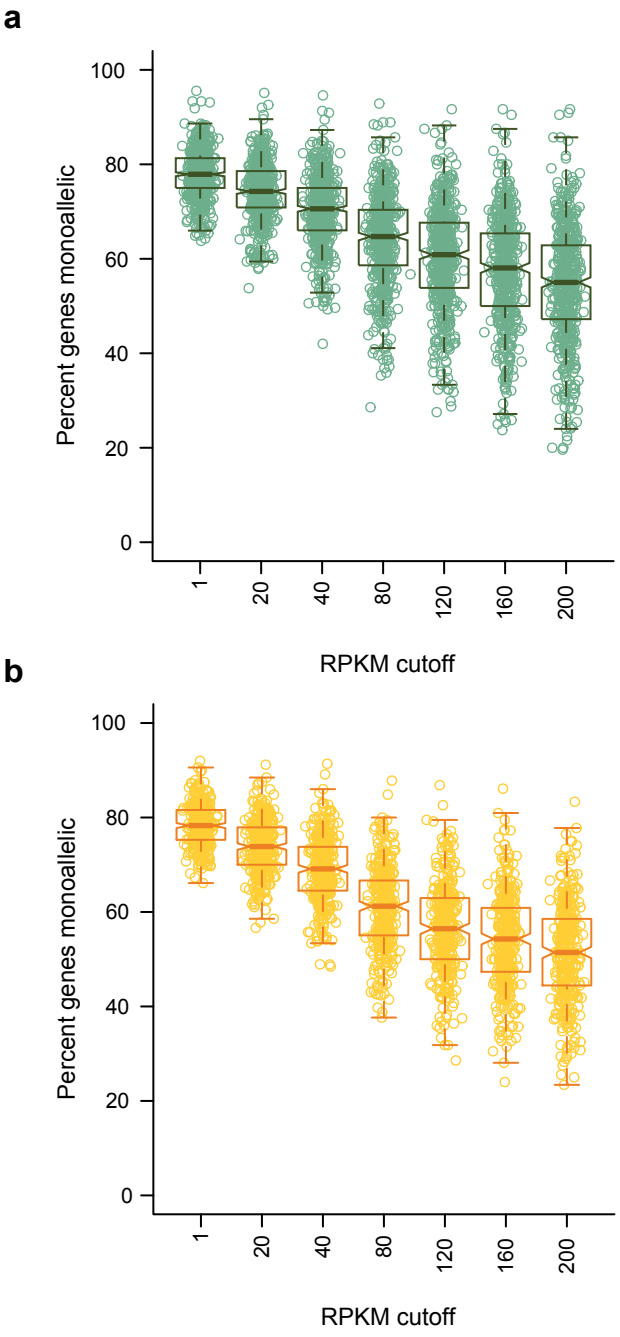
Supplementary Figure 11



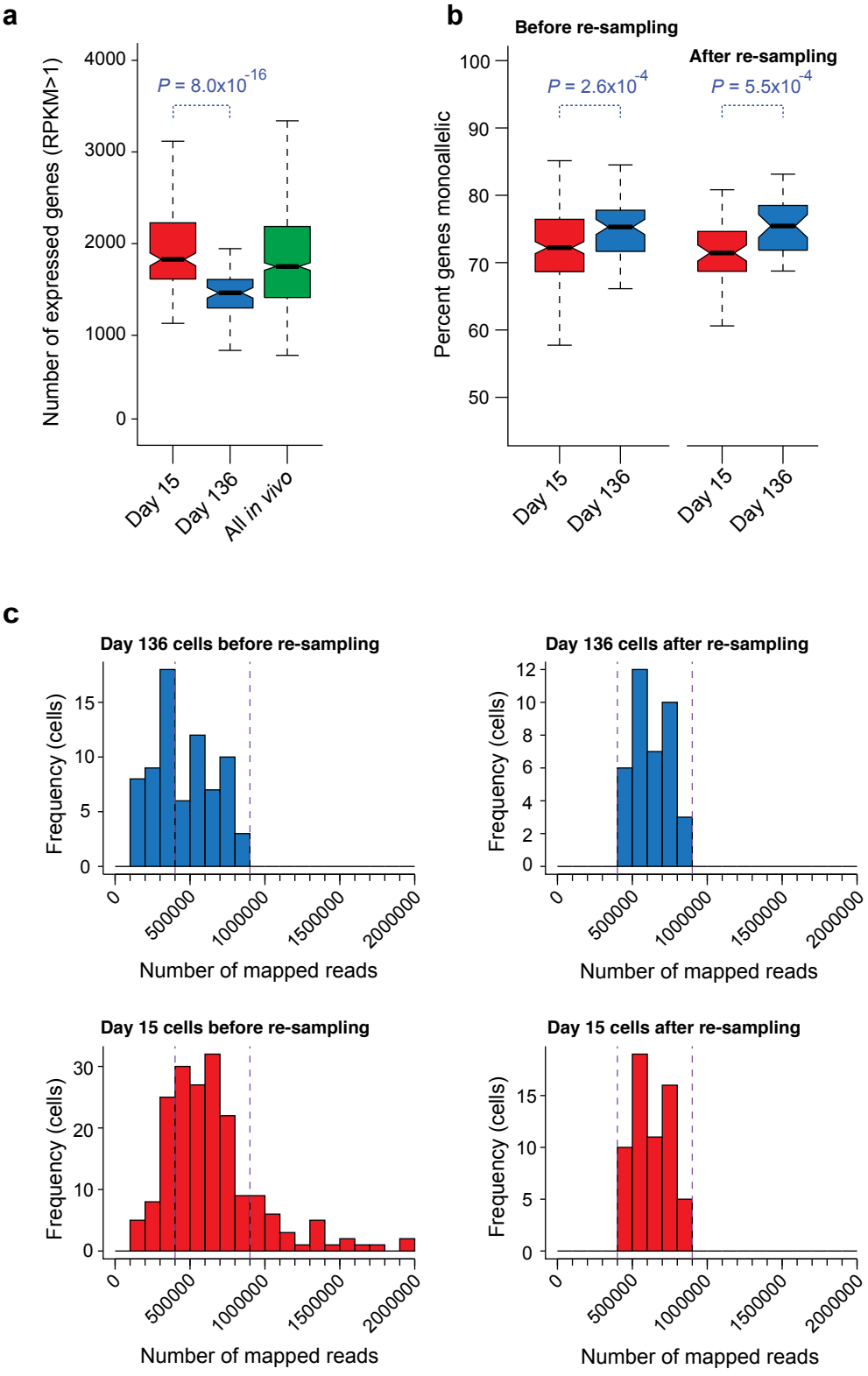
Supplementary Figure 12



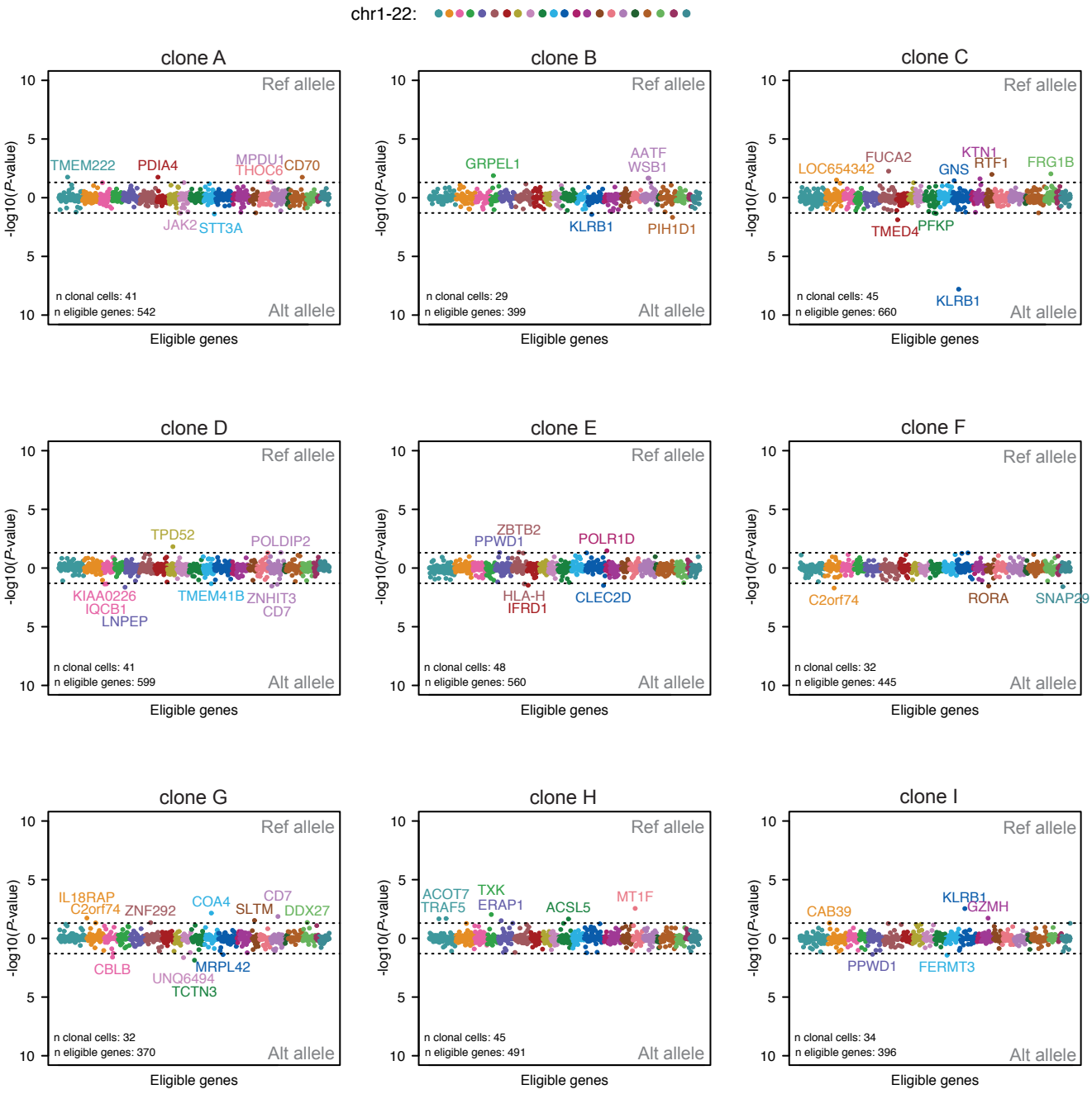
Supplementary Figure 13



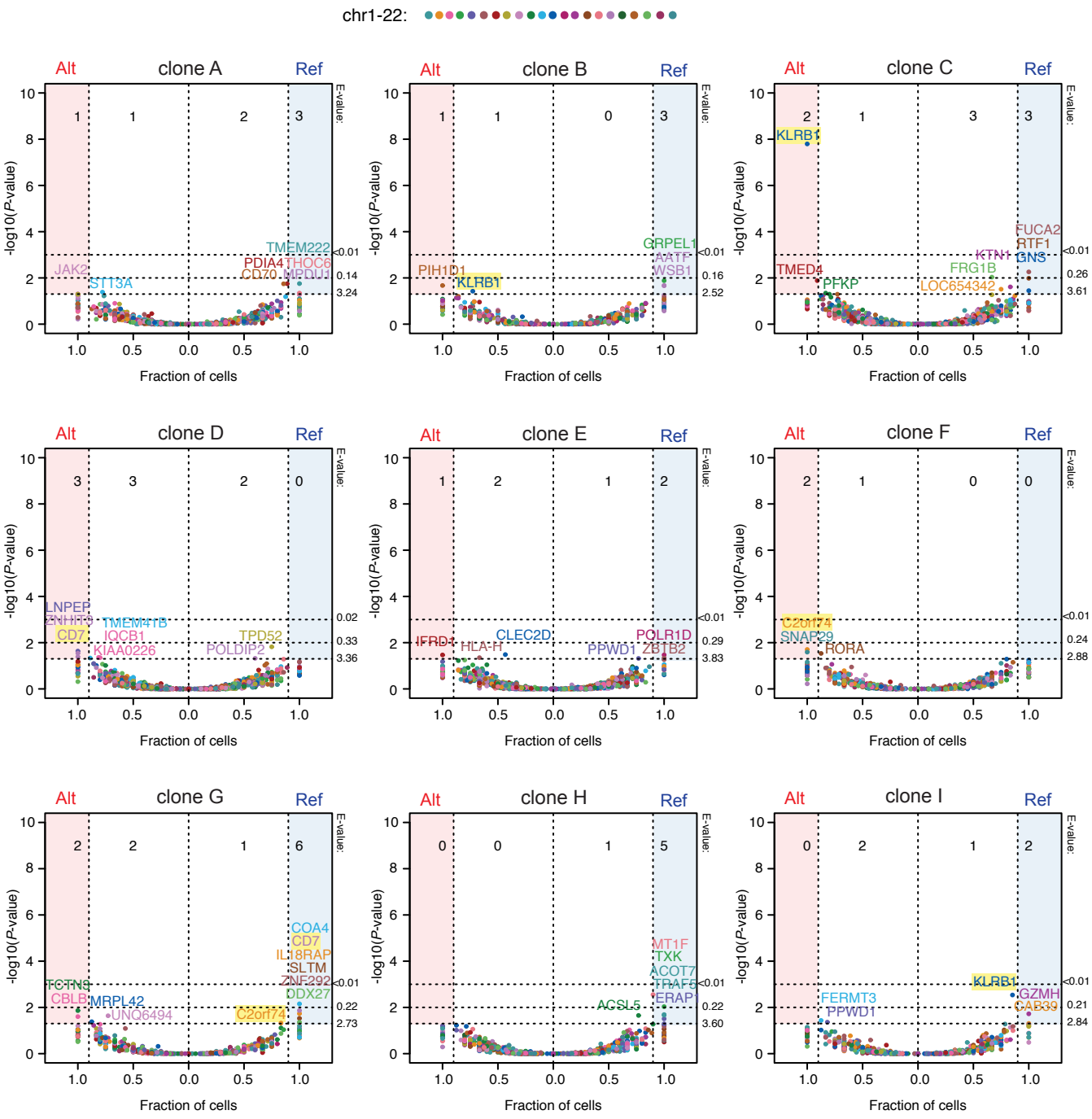
Supplementary Figure 14



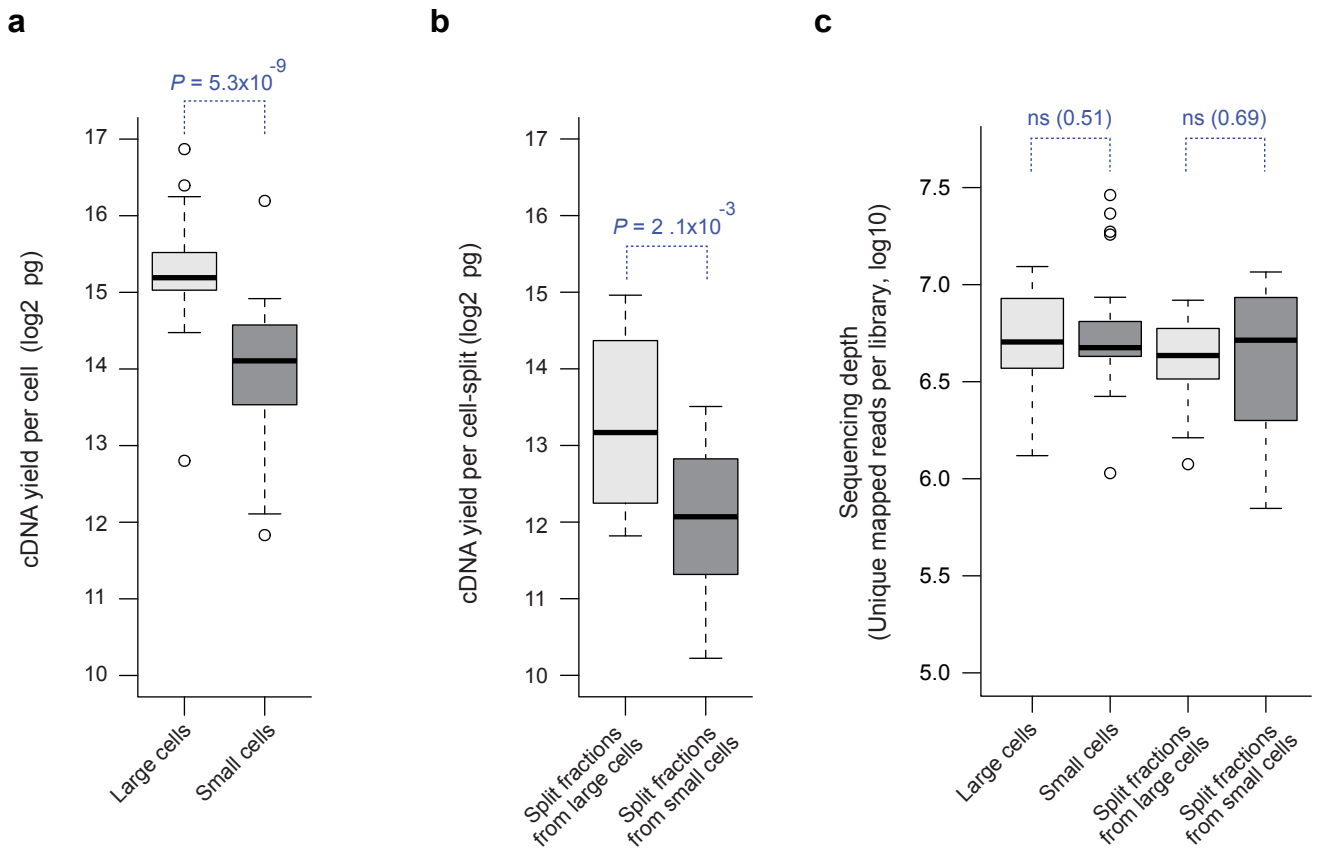
Supplementary Figure 15



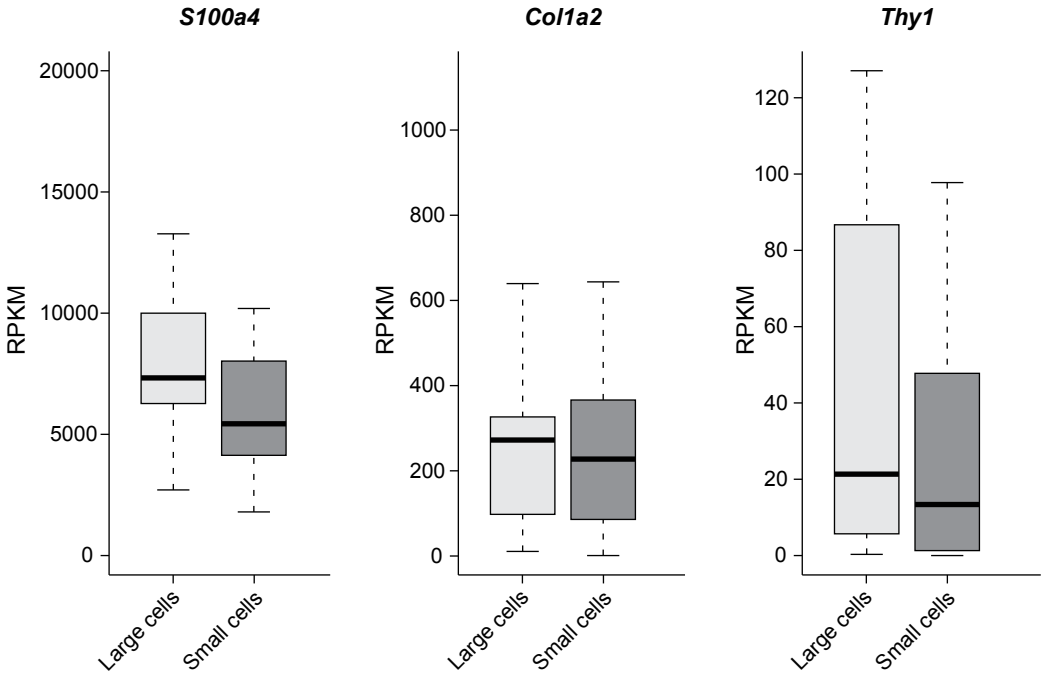
Supplementary Figure 16



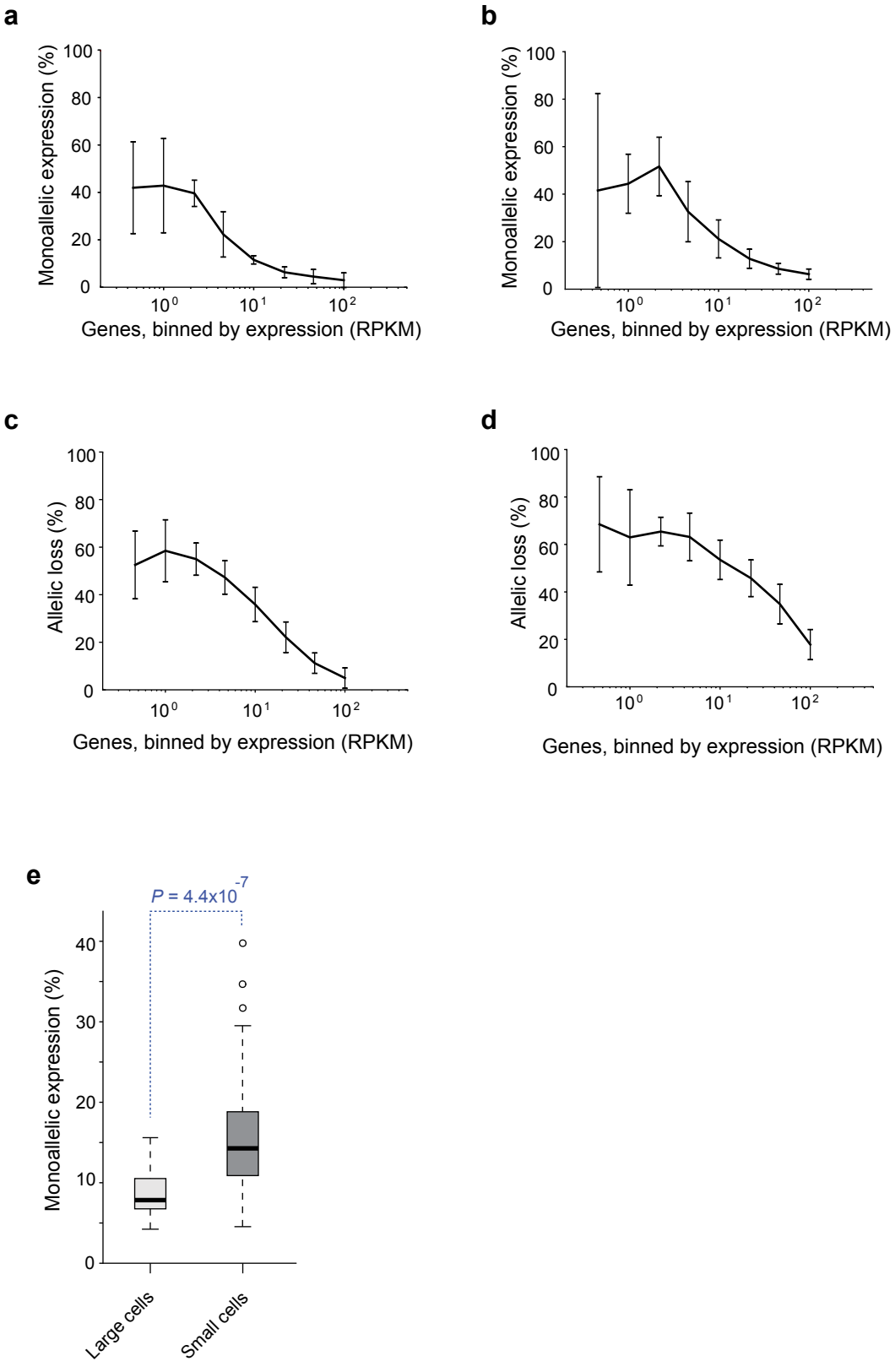
Supplementary Figure 17



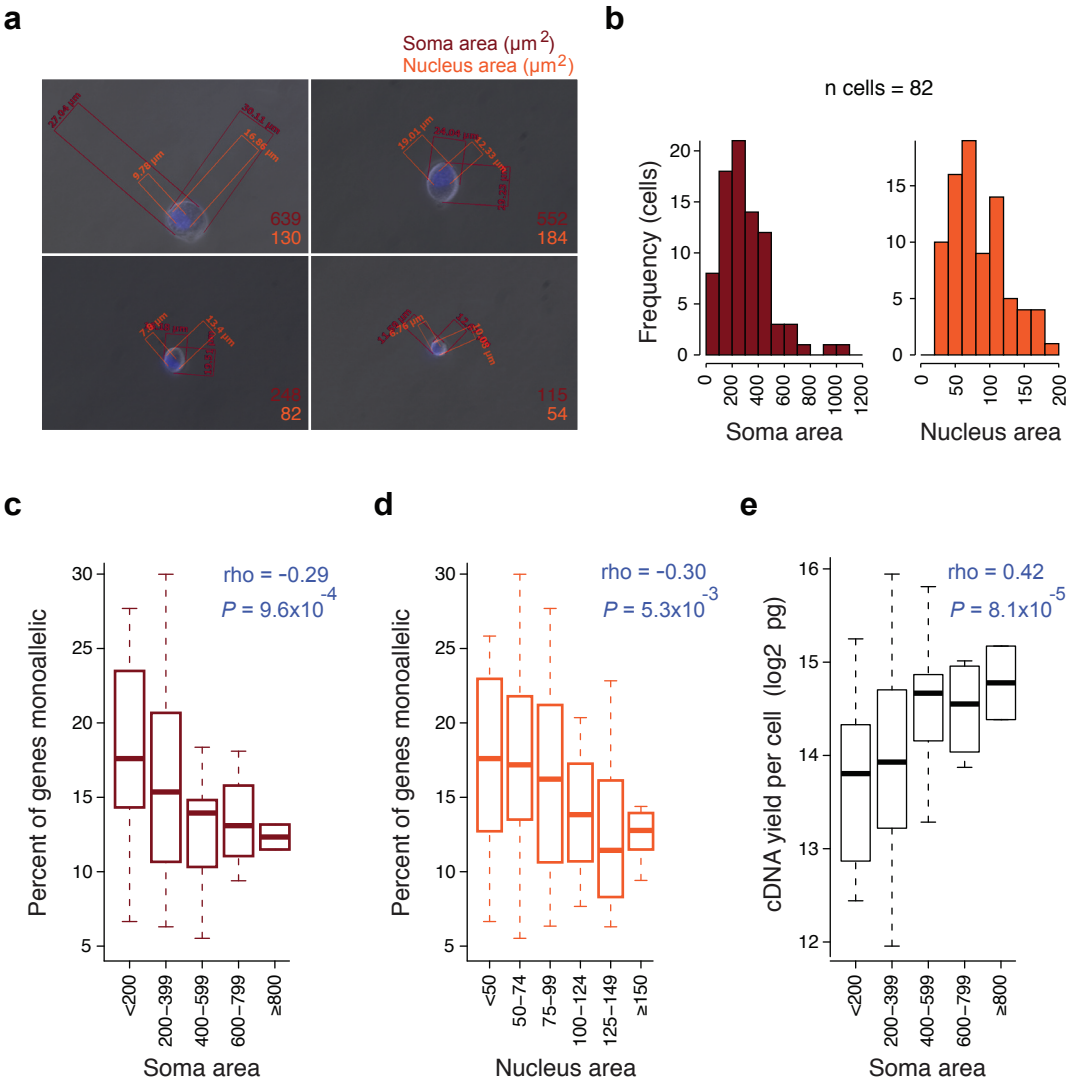
Supplementary Figure 18



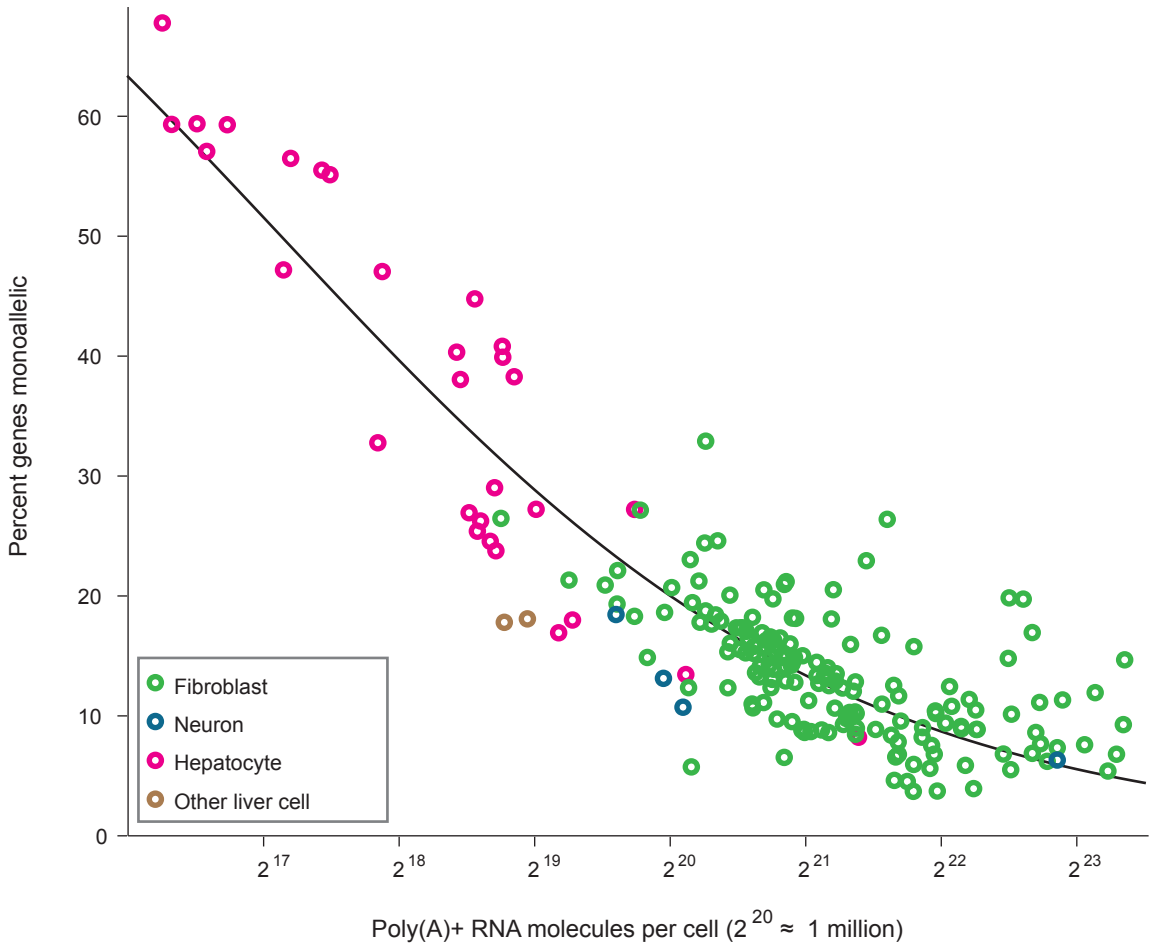
Supplementary Figure 19



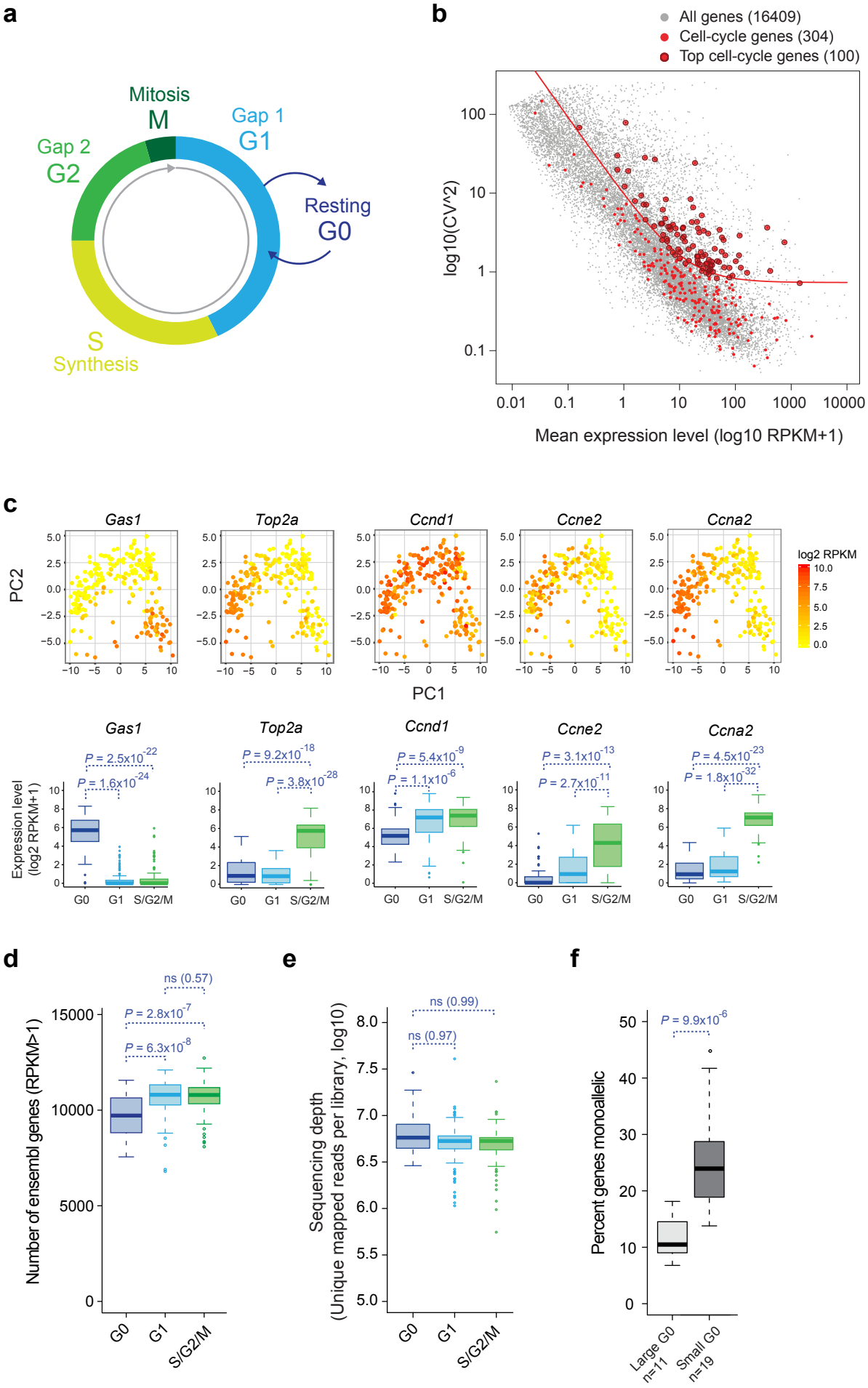
Supplementary Figure 20



Supplementary Figure 21



Supplementary Figure 22



Supplementary Note 1

Exploring properties of genes that escape random X-chromosome inactivation using single-cell RNA-seq

X-chromosome inactivation[1] (XCI) is a process that balances RNA levels of X-chromosome gene expression between females (which carry two X chromosomes per cells) and males (which carry only one X chromosome per cells). In the inner cell mass of the early female mouse embryo, one of the two parental X chromosomes becomes randomly selected to undergo transcriptional silencing, while the remaining X-chromosome copy stays transcriptionally active[2]. This process is governed by elements and genes located in the X-inactivation centre[3]. After the initial choice of active and inactive X chromosome has been resolved, the X-active (Xa) and X-inactive (Xi) chromosome status is maintained through mitotic cell divisions, so that the resulting female individual consists of a "mosaic" of cells with either the maternal (Xm) or paternal (Xp) X chromosome being active. Central to the process of XCI is the long non-coding RNA *Xist*[4-6], which is exclusively (or at least near-exclusively) transcribed from, and covers, the inactive X chromosome[7], and is essential for normal XCI[8,9]. Although most genes on Xi are though to be completely silenced, a few genes are known to escape[10,11] XCI, and thus to some degree avoid the effects of the silencing modifications on Xi. Such genes, as well as the mechanisms and properties of escape, are interesting to explore not only since escape might cause sex difference in RNA levels, but also since this might reveal basic properties of active and inactive transcriptional gene states. Studies on the allelic expression of escapee genes have either characterized expression in individual cells, using single- or few-gene approaches such as RNA-FISH or RT-PCR, or studied gene expression transcriptome-wide over populations of cells with the same parental XCI status[12]. Recently, methods to indentify escapees based on their deviation in allelic expression ratios compared to the overall allelic expression ratios of X-chromosome genes in tissues from out-crossed mice have also been developed[13]. X-chromosome-wide analyses of escape from random XCI in individual cells are however lacking.

In this Supplementary Note we took advantage of our single-cell RNA-seq data from mouse fibroblasts (C57BL/6J x CAST/EiJ, reciprocal cross) to characterize basic properties of genes escaping XCI in individual cells.

We first calculated the overall fraction C57- and CAST-derived reads for the X chromosome and for autosomal chromosomes in individual female and male fibroblasts (See **Supplementary Note 2** for details on RNA-sequencing and the analysis of allelic expression). This demonstrated the pattern of random XCI in female cells, since individual cells had the majority of their X-chromosome expression derived consistently from either the C57 or the CAST alleles (**Supplementary Figure 8a**, left panel). Based on the identity of the predominantly expressed X-chromosome copy, we classified female cells as either "C57 Xa" or "CAST Xa", having the C57 or the CAST X chromosome active respectively. Female cells had on average 97.7% of their X-chromosome expression (percent of mapped reads) derived from the Xa. As expected, male cells expressed a single maternal X chromosome (mean 99.9% and 99.8% of X-chromosome reads mapped to Xm in the C57 x CAST and CAST x C57 crosses, respectively). Although most X-gene expression in female cells derived from Xa, we detected significant transcription occurring from the inactive X chromosome in female cells (P=2.8x10^{-39}, two-sided Wilcoxon test versus male cells), as expected due to the presence of genes escaping XCI. The fraction reads derived from the C57 and CAST alleles on autosomal chromosomes demonstrated the lack of any strong mapping bias towards the C57 or CAST genomes (mean 49.8% C57-derived reads, P=0.2, two-sided Wilcoxon test; **Supplementary Figure 8a**, right panel). The X-chromosome status ("C57 Xa" or "CAST Xa") was conserved over mitotic division of cells, as observed in clonally expanded female cells (**Supplementary Figure 7a-b**). We noticed that a significant majority (66%) of the female cells had the CAST X chromosome active (P=0.0076, binominal test; considering only non-clonal cells to avoid pseudo replicates due to clonal XCI conservation) (**Supplementary Figure 8b**). This number was in good agreement with the fraction CAST-derived X-chromosome expression (63% of reads) observed in bulk RNA sequencing of a female primary fibroblasts culture (**Supplementary Figure 8c**). Our observations by single-cell RNA-seq thereby support the notion of a "strong" X-controlling element contained in the CAST X chromosome[14,15].
As expected, *Xist* was expressed exclusively in female cells (**Supplementary Figure 8d**). We counted the number of genes with expression from Xi in each female cell, and found that on average 6% (12/196) of the expressed genes (RPKM>1) in the individual cells had Xi expression (average 0/193

genes in male cells) (**Supplementary Figure 8e**); a number that is in the upper range of the 3–6% of X-linked genes found to consistently escape XCI in cell lines[12,16] or in the brain[17], as analyzed over large populations of cells (masking possible heterogeneity in escape due to high cell numbers). To identify genes with consistent escape over many fibroblast cells, we calculated the frequency of female cells with escape events (i.e. expression from Xi at the time point of sampling) for each gene, and as a background of technical false positive detections for each gene we calculated Xp detection frequencies in male cells (**Supplementary Figure 8f**). This identified 15 genes (3.2%) with frequent escape out of the 475 expressed genes available for testing: *Xist*, *1810030O07Rik*, *5530601H04Rik*, *Kdm6a*, *Kdm5c*, *2610029G23Rik*, *Med14*, *Utp14a, Ddx3x*, *Gpkow*, *Timp1*, *Uba1*, *Flna*, *Sh3bgrl*, and *Maged1* (P<0.05, Benjamini-Hochberg adjusted, Fisher's exact test) (**Supplementary Table 2**). Most of these genes have been shown to escapee XCI in previous studies[12,13,18,19]. We found that cytoband A1.1 and D had significant enrichment of escapee genes given their gene densities (EASE Score < 0.05) (**Supplementary Figure 8f**). We explored whether the 15 identified frequently escaping genes tended to have different expression output from the Xa and Xi alleles, specifically for cells with simultaneous (biallelic) presence of RNA from both the Xa and Xi at the time point of sampling. With the exception of *Xist* (which was nearly exclusively detected from Xi), only the gene *1810030O07Rik* had a slightly but significantly higher expression from the Xi allele (P=0.018, two-sided Wilcoxon test) (**Supplementary Figure 8g**). *5530601H04Rik* and *Kdm6a* had equal expression output from Xi and Xa, and all other escapees had significantly lower output from the Xi allele as compared to Xa (P=6.3x10^{-17} to 6.4x10^{-4}, two-sided Wilcoxon test). These observations suggest that diverse escapee genes have both different tendencies to produce Xi transcription events as well as different levels of Xi-to-Xa output during such escape events, generally having a decreased expression tendency and decreased expression output from Xi. As demonstrated in **Supplementary Figure 8f**, many genes along the X chromosome escaped XCI in only a small fraction of all the sampled fibroblast cells. Employing the clonal cells, we sought to explore whether the ability to produce "sporadic" escape events might be mitotically inheritable. We identified one gene, *Fundc1*, that escaped XCI in only 9% (6/69) of female non-clonal cells but in 100% (4/4) of cells in one of the clones (cl3). Given *Fundc1*'s escape frequency in non-clonal cells, the probability of independent escape events occurring in 4/4 clonal cells was calculated to P=5.7x10^{-5} (binominal distribution, P=0.027 after adjustment for multiple testing considering all genes and clones). Similarly, the probability of *Fundc1* independently never escaping in any cell of the other female clones (0/78 cells) was 8.3x10^{-4}. These data suggest that for some X-chromosome genes with low frequency of escape events over the population, such as for the gene *Fundc1*, both the ability to escape from XCI as well as the ability to maintain a silent state on Xi represent mitotically transmittable traits. Thus, clonal variability in escapee-gene composition contributes to cellular heterogeneity in allelic gene expression in females. In the specific case of *Fundc1*, our result may explain why *Fundc1* is expressed ~10% higher in female mouse tissues compared to those of males[20]. Our analyses demonstrate the usefulness of employing single-cell RNA-seq to study escape from XCI.

References to Supplementary Note 1

1. Lyon, M.F. Gene action in the X-chromosome of the mouse (Mus musculus L.). *Nature* **190**, 372-3 (1961).
2. Augui, S., Nora, E.P. & Heard, E. Regulation of X-chromosome inactivation by the X-inactivation centre. *Nat Rev Genet* **12**, 429-42 (2011).
3. Avner, P. & Heard, E. X-chromosome inactivation: counting, choice and initiation. *Nat Rev Genet* **2**, 59-67 (2001).
4. Borsani, G. *et al*. Characterization of a murine gene expressed from the inactive X chromosome. *Nature* **351**, 325-9 (1991).
5. Brockdorff, N. *et al*. Conservation of position and exclusive expression of mouse Xist from the inactive X chromosome. *Nature* **351**, 329-31 (1991).
6. Brown, C.J. *et al*. A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature* **349**, 38-44 (1991).
7. Clemson, C.M., McNeil, J.A., Willard, H.F. & Lawrence, J.B. XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure. *J Cell Biol* **132**, 259-75 (1996).
8. Marahrens, Y., Panning, B., Dausman, J., Strauss, W. & Jaenisch, R. Xist-deficient mice are defective in dosage compensation but not spermatogenesis. *Genes Dev* **11**, 156-66 (1997).
9. Penny, G.D., Kay, G.F., Sheardown, S.A., Rastan, S. & Brockdorff, N. Requirement for Xist in X chromosome inactivation. *Nature* **379**, 131-7 (1996).

10. Balaton, B.P. & Brown, C.J. Escape Artists of the X Chromosome. *Trends Genet* **32**, 348-59 (2016).
11. Deng, X., Berletch, J.B., Nguyen, D.K. & Disteche, C.M. X chromosome regulation: diverse patterns in development, tissues and disease. *Nat Rev Genet* **15**, 367-78 (2014).
12. Yang, F., Babak, T., Shendure, J. & Disteche, C.M. Global survey of escape from X inactivation by RNA-sequencing in mouse. *Genome Res* **20**, 614-22 (2010).
13. Berletch, J.B. *et al*. Escape from X inactivation varies in mouse tissues. *PLoS Genet* **11**, e1005079 (2015).
14. Chadwick, L.H., Pertz, L.M., Broman, K.W., Bartolomei, M.S. & Willard, H.F. Genetic control of X chromosome inactivation in mice: definition of the Xce candidate interval. *Genetics* **173**, 2103-10 (2006).
15. Thorvaldsen, J.L., Krapp, C., Willard, H.F. & Bartolomei, M.S. Nonrandom X chromosome inactivation is influenced by multiple regions on the murine X chromosome. *Genetics* **192**, 1095-107 (2012).
16. Calabrese, J.M. *et al*. Site-specific silencing of regulatory elements as a mechanism of X inactivation. *Cell* **151**, 951-63 (2012).
17. Wu, H. *et al*. Cellular resolution maps of X chromosome inactivation: implications for neural development, function, and disease. *Neuron* **81**, 103-19 (2014).
18. Reinius, B. *et al*. Female-biased expression of long non-coding RNAs in domains that escape X-inactivation in mouse. *BMC Genomics* **11**, 614 (2010).
19. Splinter, E. *et al*. The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. *Genes Dev* **25**, 1371-83 (2011).
20. Reinius, B. *et al*. Abundance of female-biased and paucity of male-biased somatically expressed genes on the mouse X-chromosome. *BMC Genomics* **13**, 607 (2012).

Supplementary Note 2

Full Methods: *Experimental procedures*

Derivation of primary mouse fibroblasts
Primary fibroblasts were derived from adult CAST/EiJ x C57BL/6J or C57BL/6J x CAST/EiJ mice (approval by the Swedish Board of Agriculture, Jordbruksverket: N343/12) by skinning, mincing and culturing tail explants in fibroblast medium (DMEM high glucose (Invitrogen 41965-039), 10% ES cell FBS (Gibco Ref. 16141-079), 1% Penicillin/Streptomycin (Invitrogen 15140-114), 1% Non-essential amino acids (Invitrogen 11140-035), 1% Sodium-Pyruvate (Invitrogen 11360-039), 0.1mM b-Mercaptoethanol (Sigma)). For one clone (cl6) the fibroblasts were derived from minced E14.5 skin instead of tails, using the same procedure. After removal of explants, the culture was passaged twice to attain a pure fibroblast culture. Cells from passage 2-4 were dissociated (TrypLE Express, Gibco Ref. 12604-013), diluted to very low cell density in medium, and cells were manually picked by mouth pipetting under 10x or 20x magnification using a thin glass capillary in picking volumes of 0.5μl medium. Cells for RNA-seq were transferred to RNAse-free PCR tubes containing 2μl Smart-seq2 lysis buffer (1.9μl 0.2% TritonX-100; 0.1μl Recombinant RNase inhibitor (TaKaRa 40U/μl Ref. 2313A) per reaction) and snap-frozen on dry ice. Information on the derivation of clonal fibroblasts is found in the section *Monoclonal fibroblasts*.

Dissociation of cells from mouse *in vivo* tissues
Liver and brain cortex cells were dissociated from C57BL/6J x CAST/EiJ E14.5 embryos by cutting the tissues into small pieces and incubation in TrypLE at 37°C for 30 min followed by passing through a cell strainer (70μm), and single cells were picked in the same way as mouse fibroblasts.

Measuring the size of primary mouse fibroblasts
For some fibroblasts, we annotated the dissociated cell soma size as "large" (diameter of dissociated cell: 25-35μm) or "small" (diameter of dissociated cell: 10-20μm) while picking the cells under the microscope. We processed these cells exactly the same as the other fibroblasts, which were not annotated by size, and the libraries from these size-annotated cells were used for the analyses presented in **Fig. 4a-b**. For another set of fibroblasts (CAST/EiJ x C57BL/6J) we stained the nuclei with Hoechst 33342, and photographed the cells under a fluorescence microscope (Zeiss Axiovert 200M) before picking. The cells' soma and nuclei areas were measured using AxioVision 4.8.2 (Carl Zeiss MicroImaging). The sequencing libraries from these cells (listed in **Supplementary Table 1d**) were used in the analyses presented in **Supplementary Fig. 20**. These Hoechst-stained cells were not considered in other analyses since Hoechst 33342 binds the minor groove of double-stranded DNA, which could affect gene expression.

Single-cell RNA-sequencing: Reverse transcription
Single-cell Smart-seq2 libraries were prepared as earlier described[1]. Briefly, 2.1μl priming buffer mix (1μl 10mM dNTPs; 1μl 10μM Smarter oligo-dT primer; 0.1μl ERCC spike-in RNA (Ambion, Ref. 4456740, see the section *Exogenous spike-in RNA*) per reaction) was added to the cell lysate and incubated for 3 min at 72°C in a thermal cycler, and then put on ice. Reverse transcription (RT) was performed by the addition of 5.6μl Smart-seq2 RT mix (0.5μl SuperScriptII (Invitrogen Cat. 18064-014); 2μl 5x SuperScriptII buffer; 0.5μl 100mM DTT; 2μl 5M betaine; 0.1μl 1mM $MgCl_2$; 0.25μl Recombinant RNase inhibitor (TaKaRa 40U/μl Ref. 2313A); 0.1μl 100μM LNA strand switch primer; 0.15μl water, per reaction) and incubation (90 min 42°C; 10 "strand-switch" cycles of (2 min 50°C; 2 min 70°C); 4°C) in a thermal cycler.

Single-cell RNA-sequencing: cDNA amplification
cDNA amplification was performed by the addition of 15μl of Smart-seq2 PCR mix (12.5μl 2x KAPA HiFi HotStart ReadyMix (KAPA Biosystems Ref. KK2602); 0.25μl 10μM ISPCR primers; 2.25μl water, per reaction) and incubation (3 min 98°C; 18 cycles of (20 s 98°C; 15s 67°C; 6 min 72°C); 5 min 70°C; 4°C) in a thermal cycler. The cDNA was purified using 0.7:1 volume of AMPure XP beads (Beckman Coulter, Ref. A63882) according to the manufacturer's protocol (No. PT5163-1). The purified cDNA was inspected on an Agilent 2100 Bioanalyzer to determine cDNA concentration and size distribution, using Agilent High Sensitivity DNA chips (Ref. 5067-4626).

Single-cell RNA-sequencing: Tagmentation and sequencing
Successful cDNA libraries were tagmented, using the Nextera XT DNA kit (Illumina Cat. FC-131-1024, and in accordance with the provided protocol No. 15031942) or using our in-house-generated Tn5 (which produces sequencing libraries of indistinguishable characteristics to those generated using the Nextera XT kit)[2]. For tagmentation using our in-house Tn5, 1ng of cDNA in 5μl water was mixed with 15μl tagmentation mix (1μl of Tn5; 2μl 10x TAPS $MgCl_2$ Tagmentation buffer; 5μl 40% PEG8000; 7μl water, per reaction) and incubated 8 min at 55°C in a thermal cycler. Tn5 was inactivated and released from the DNA by the addition of 5μl 0.2% SDS and 5 min incubation at room temperature. Sequence library amplification was performed using 5μl Nextera XT Index primers (Illumina, Ref. 15032356) and 15μl PCR mix (1μl KAPA HiFi DNA polymerase (KAPA Biosystems Ref. KK202); 10μl 5x KAPA HiFi buffer; 1.5 μl 10mM dNTPs; 2.5μl water, per reaction), and incubation (3 min 72°C; 30 s 95°C; 10 cycles of (10 s 95°C; 30 s 55°C; 30 s 72°C); 5 min 72°C; 4°C) in a thermal cycler. More details about the tagmentation procedure and the synthesis of Tn5 are available in our recent publication[2]. DNA sequencing libraries were purified using 1:1 volume of AMPure XP beads (Beckman Coulter, Ref. A63882) according to the manufacturer's protocol (No. PT5163-1, version PR0X3693), inspected on an Agilent 2100 Bioanalyser, using Agilent High Sensitivity DNA chips (Ref. 5067-4626), and DNA concentration was measured using a Qubit 2.0 Fluorometer (Invitrogen) with the Qubit dsDNA High Sensitivity Assay kit (Molecular Probes, Ref. Q32854). Pools of samples for multiplexing, with unique Illumina barcode for each cell, were prepared according to the Nextera XT DNA Sample Preparation Guide (No. 15031942 page 46, llumina). DNA sequencing was performed on an Illumina HiSeq2000. All mouse single-cell RNA-seq libraries used in the analyses of aRME are listed in **Supplementary Table 1** and deposited to GEO (GSE75659).

Generation of libraries from split-cell lysates
Split-cell libraries were prepared as for single cells described above, with the following modifications: Single cells were lysed in 4μl lysis buffer instead of 2μl, and the lysate was thoroughly homogenized by pipetting up and down. 2μl of the homogenized lysate was then transferred to two separate tubes, each containing 0.5μl lysis buffer (to facilitate complete release of the lysate), and processed into sequencing libraries. Only split-pairs that generated similar Agilent Bioanalyzer profiles were selected for sequencing.

Exogenous spike-in RNA
Spike-in RNA (ERCC, Ambion, Ref. 4456740) was diluted in 10mM Tris-HCl pH7.5, and 0.1μl of diluted (1:109,658, stock:final concentration) ERCC was added to each cell's RT priming buffer (corresponding to 56,848 ERCC molecules). For T-cell libraries spike-in RNA was diluted 1:1,200,000 and 0.1μl per reaction was added to the RT priming buffer (corresponding to 5,195 ERCC molecules). The analysis of spike-in RNA and calculation of copy numbers is described under the section *Computational procedures*.

Monoclonal fibroblasts
Single fibroblasts (from passage 2-4) were picked manually by mouth pipetting using a thin glass capillary. Each individual cell was placed within a marked area in the bottom of a gelatinized culture dish, observing the release of a single cell per dish under 10-20x magnification. The medium was the same fibroblast medium as described in the section *Derivation of primary mouse fibroblasts*. The culture dish and medium was equilibrated for temperature and CO_2 in the cell incubator previous to cell seeding (we found that this equilibration was important for obtaining vital monoclonal cultures, and ~1/10 seeded cells produced dividing clonal lines). To avoid losing many of the clonal cells in the relatively large volume of medium in the culture dish during dissociation, we performed the dissociation and picking as follows: We removed the medium, then washed the dish with PBS and added TrypLE, keeping focus on the cells under a microscope at 10-20x magnification. We then carefully picked the cells by mouth pipetting and gently blowing "puffs" of TrypLE on the cells under the microscope, so that the cells were released from the bottom of the dish one at a time. For one clone (cl7) we also picked multi-cell samples of 5 or 15 cells that were jointly lysed and processed. The processing of the clonal cells into sequencing libraries was performed as described in the preceding sections.

Human subjects for isolation of T-cells
Human volunteers were recruited to participate in an ongoing study examining the longitudinal immune response to the yellow fever vaccine YFV 17D (approved by the Regional Ethical Review

Board in Stockholm, Sweden: 2008/1881-31/4, 2013/216-32 and 2014/1890-32). All subjects gave written informed consent prior to the study. A single subject (subject ID: YFV2001(male)) was selected for in depth analysis based on being both HLA-A2 and HLA-B7 positive and having T-cell responses against YFV epitopes presented on both HLA types. Peripheral blood was collected at day 15 and day 136 after vaccination. Peripheral blood mononuclear cells (PBMCs) were isolated by density centrifugation (Lymphoprep, Stem Cell Technologies, Ref. 07801) and stored in liquid nitrogen in 90% FCS and 10% dimethylsulfoxide (DMSO) until sorting.

Single-cell sorting of human CD8$^+$ T-cells
Cryopreserved PBMCs from the donor were thawed and immediately re-suspended in cold FACS buffer (PBS supplemented with 2% bovine serum and 2mM EDTA). CD8$^+$ T-cells were enriched by negative selection (Miltenyi Human CD8 negative selection kit, Ref. 130-096-495), and were subsequently incubated for 15 min with either HLA-B07:02/RPIDDRFGL-dextramers (day 15 and day 136 cells) or HLA-A02:01/LLWNGPMAV-dextramers (day 15 cells) (Immudex, Denmark), followed by incubation with a panel of surface antibodies to detect live CD3$^+$CD8$^+$ T-cells (CD3-Alexa Flour 700 (UCHT1, BD Biosciences), CD8-APC Cy7 (SK1, BD Biosciences), CD4-PE-Cy5 (RPA-T4, Ebiosciences), CD14-Horizon V500 (MΦP9, BD Biosciences), CD19-Horizon V500 (HIB19, BD Biosciences), and Live/Dead Fixable Aqua dead cell stain kit (Invitrogen, Cat. L34957)). After 30 minutes of incubation at 4°C, the cells were washed twice and resuspended in cold FACS buffer for sorting. The cells were sorted using single-cell mode on a BecktonDickinson ARIA III cell sorter (equipped with a 405nm violet laser, 488nm blue laser, 561nm yellow/green laser, and 633nm red laser) into 96 well PCR plates (ThermoScientific, Ref. AB-0800) containing 4ul Smart-seq2 lysis buffer (as described for mouse cells). Negative wells were left on the plates to monitor potential contamination, and index sorting mode was used to validate cellular identity in each well after sorting.

***Ex vivo* expansion of human primary T-cells**
CD8$^+$ T-cells were isolated by negative selection from thawed cryopreserved PBMCs and were stained with the HLA-A2/LLWNGPMAV dextramer and antibodies as described above. Single live CD8$^+$CD3$^+$HLA-A2 dextramer$^+$ cells were sorted directly into 96 well U-bottom plates containing 2mg/ml peptide (LLWNGPMAV), 20U/ml IL-2, and 50.000 irradiated (40Gy) CD3-depleted autologous PBMCs in complete T-cell media (RPMI 1640 with 10% pooled heat-inactivated human AB sera, 2mM L-glutamine, 50mM 2-mercaptoethanol, 10mM Hepes, 1mM sodium pyruvate, 100U/ml penicillin and 100mg/ml streptomycin) and were cultured for 18 days. Every 4-5 days half of the media was replaced with fresh T-cell media containing 50U/ml IL-2 and 2mg/ml peptide, and the wells were visually inspected for proliferation. After 18 days, nine expanded clones were selected for single cell sorting for RNA-seq based on the presence of sufficient cell numbers (>100) and no evidence of contaminating cells (all live cells were CD8$^+$ and bound the HLA-A2/LLWNGPMAV dextramer). Single cell RNA-seq libraries were generated as described below for the *in vivo* studies on YFV-specific CD8$^+$ T-cells.

Single-cell RNA–seq on human T-cells
We prepared single-cell RNA-seq libraries as described above for mouse cells**,** with the following modifications: dNTP and OligodT were included in the lysis buffer, ERCC spike-ins were used at 1:1,200,000 dilution to prevent ERCC reads from overinflating the sequencing libraries of the T-cells, which contained substantially lower amounts of mRNA compared to fibroblasts, and the primary PCR during cDNA amplification had 22 cycles instead of 18 cycles, again to account for the lower RNA content in T-cells. The samples were sequenced using an Illumina HiSeq2000 on high output mode with 2x125bp sequencing. All T-cell single-cell RNA-seq libraries used in the analyses are listed in **Supplementary Table 1** and deposited to GEO (GSE75659).

Exome sequencing
Exome capture was performed on YFV2001 using the Nextera Exome Sequencing kit (Nextera Rapid Capture Expanded Exome Kit, Illumina, Ref. FC-140-1000) in accordance with the manufacturer's protocol. Genomic DNA was isolated from unsorted donor PBMCs during CD8$^+$ T-cell sorting (QIAGEN, DNA isolation kit, Cat. 69581). Library preparation was performed using reagents contained in the Nextera Exome Sequencing kit, and the samples were sequenced using an Illumina HiSeq2000.

Bulk sequencing of cultured fibroblasts
To compare the sensitivity of single-cell RNA-seq in the detection of monoallelic gene expression with bulk-RNA sequencing, we performed deep sequencing of bulk RNA, extracted using Trizol, from a female C57BL/6J x CAST/EiJ fibroblast culture (~2 million cells, passage 3). For the library preparation, we used Illumina TruSeq Stranded mRNA Library Prep Kit (according to the manufacturer's instructions), and sequencing was performed on an Illumina HiSeq2000. Using these data (16,092,712 unique trascriptome-mapped reads), we identified known imprinted genes (which are to be expressed from the same parental allele throughout the culture). The following criteria were used to identify known imprinted genes expressed in fibroblasts: *i)* Being listed as known imprinted in the GeneImprint repository (www.geneimprint.com). *ii)* Being expressed in fibroblasts (RPKM>1 in the bulk). *iii)* Having at least 75% expression bias towards one parental allele in bulk. This identified 10 previously known imprinted genes (listed in **Supplementary Table 3b** along with results from single-cell RNA-seq), all of them being expressed from the anticipated parental allele. Mapping of the sequence reads to the CAST and C57 genomes were performed in the same way as for single cells.

Full Methods: *Computational procedures*

Alignment of reads and allelic calling for mouse cells
We aligned reads using RNA-STAR[3] independently towards the mm9 genome assembly (which is of the C57 genotype) and an in-house generated CAST mm9 assembly, in which we had replaced C57 bases and SNPs with CAST bases, using CAST SNPs identified in the Sanger mouse strain genome sequencing project. To filter out false positive SNPs before making allelic calls, we required that each SNP occurred with both genotypes in the total set of sequenced fibroblasts. We used this filtered SNP set and counted C57- and CAST-informative bases for each gene, using the samtools mpileup command. To further eliminate potentially false SNP calls, occurring at low frequency, we only conceded the allelic calls for genes with at least 3 allele-informative reads. The allelic expression of each gene in each cell was called as "biallelic", "reference monoallelic", "alternative monoallelic" or "not detected" (**Supplementary Fig. 6a**). A gene was considered monoallelically expressed in a cell if one allele contributed 98% or more of the genotype-informative bases. Thus we allowed for up to 2% of reads to be aligning to the other allele without calling it "biallelic", in order to tolerate a small degree of PCR and sequencing errors, known to affect high-expressed genes in particular[4]; considerably more stringent than previous studies RNA-seq studies[5-8] that allowed for up to 20% of reads coming from the "silent" allele. We demonstrated that using these to criteria (98% or more of the reads from one allele and at least 3 reads) were robust and resulted in accurate calls through analyses of allelic calls for genes on the X chromosome in male cells (i.e. cells with only one parental X chromosome), and the results are presented in **Supplementary Fig. 3a-b**. We re-analyzed the data using a minimum of 10 allelic counts threshold (instead of a minimum of 3) for gene and cell inclusion, and confirmed that conclusions remained using this threshold. We verified that multiple SNPs within the same genes and cells provided coherent allelic expression estimates (**Supplementary Fig. 5**). Fibroblasts expressed on average 10,702 genes. The number of allele-informative autosomal genes, eligible in the gene-level test for fibroblasts, was 8,264-9,195 genes, and the number of allele-informative genes with a mean RPKM >20 in fibroblasts was 4,514.

Expression levels
We calculated RPKM values (reads per kilobase and million mapped reads) using rpkmforgenes[9] version 7 Feb 2014 with uniquely mapped reads; Ensembl annotation (from UCSC genome browser, last modification date 11 April 2012) for mouse, and hg19 for human; length compensation for unmappable positions[10] and the settings -fulltranscript -mRNAnorm -rmnameoverlap and -bothendsceil.

Percent monoallelically expressed genes
Percent monoallelically expressed genes for each cell was calculated as the number of genes with monoallelic calls divided by the total number of genes with an allele call (i.e. having ≥3 allele-informative reads) multiplied by 100, including genes with mean expressions above 1 or 20 RPKM over the group of cells considered (thresholds are stated in the figure legends). We used the 20 RPKM threshold in all analyses where we estimated total aRME in cells, since at this threshold the false negatives and false positives monoallelic calls are in balance both in previous experiments [4] and in the split-cell experiments on the mouse fibroblasts. The lower 1 RPKM threshold was used when performing tests on clonal aRME as genes with clonal effects were expressed at these low levels. For analyses of split-cell experiments, where the cell lysate was split in two prior to cDNA conversion, amplification and sequencing, we similarly calculated allelic calls for all genes in each half-cell library.

Analyses of allelic calls in split-cell experiments
Here we describe how we infer monoallelic expression and allelic dropouts in a cell before its lysate were split into two fractions and tubes using the analytical model we recently developed[4]. In brief, we compared allelic calls for each gene in the two paired libraries and determined the fraction of genes for which the splitted cells gave coherent calls (e.g. biallelic in both, maternal monoallelic in both and so forth), and the fraction of genes with different calls (e.g. biallelic in one and monoallelic in the second sample etc.). We introduced equations that model the observed allelic calls and that account for allelic losses (e.g. from biallelic to monoallelic or not detected; from monoallelic to not-detected) as detailed in **fig. S26** in Deng et al.[4]. We solved the equations to find the fraction of monoallelic expression (x), the allelic losses (z) and the fraction biallelic expression (y). From these variables, we calculated the fraction of expressed genes that were monoallelic as x/(x+y) and these are reported in **Supplementary Fig. 19a-b**. Since z would depend on the starting number of transcripts, we inferred x, y, z for different expression levels (i.e. RPKM bins) and we plotted the inferred variables as a function of the expression levels (**Supplementary Fig. 19c-d**). Importantly, the analytical model solves the allelic dropouts and monoallelic expression per gene expression bin, and it is a threshold-independent method that can account for varying dropout levels depending on RNA contents of cells prior to lysis and splits. Therefore the approach allows for varying allelic dropouts in the analyses of each split-cell pair, and we observed that smaller fibroblasts obtained slightly higher dropout levels than larger fibroblasts (**Supplementary Fig. 19c-d**).

Corrections for chromosomal aneuploidies
We surveyed the allelic calls along the chromosomes to detect large-scale chromosomal aberrations, including aneuploidies. These were identified as chromosomes, or parts of chromosomes, where a large number of genes were all deviating in allelic expression in favor of CAST or C57 genotypes. We screened for chromosomal aberrations in the cells of each clone, by comparing the monoallelic calls towards each genotype for each chromosome. Chromosomes were flagged if abs(monoallelic calls CAST – monoallelic calls C57) > 50% and were then visually inspected. To this end, we plotted monoallelic calls for genes ordered along chromosomes and confirmed that the bias towards a genotype was confined to a region of the chromosome (partial aneuploidy) or whether it spanned the full chromosome. For example, clone (cl7) had acquired an aneuploidy of chr3 and chr4 and these chromosomes were therefore not considered in the analysis of this clone. The criterion also filtered out the following chromosomes in individual cells; Clone 5: chr2 (cell 17). Clone 6: chr1 (cell 23), chr2 (cell 52), chr4 (cell 30 and 47), chr5 (cell 36 and 45), chr7 (cell 27, 30, 47, 57 and 60), chr8 (cell 23), chr12 (cell 36), chr13 (cell 57), chr16 (cell 27 and 54), chr19 (cell 27). Examples of such cells are shown in **Supplementary Fig. 7**.

Inferring number of RNA molecules per cell from spike-ins
Using our dilutions (described above) we had on average added 56,848 spike-in molecules (or 5,195 for T-cells) to each cell lysate (http://tools.lifetechnologies.com/downloads/ERCC_Controls_Analysis.txt). To derive the total number of polyA$^+$ molecules per cell (used for **Supplementary Fig. 21**), we divided the number of added ERCC spike-in molecules by the total RPKM for spike-ins in each a sample and multiplied by the total RPKM for genes. The conversion from RPKM to RNA molecules was approximately 1:1 in mouse fibroblasts. Similar conversion from RPKMs to RNA molecules were obtained in the split-cell experiments in which the fraction of underlying monoallelic expression and stochastic allelic dropouts plateau at approximately 1-2 RPKMs, supporting that those estimates in mouse primary fibroblasts corresponded to ~1 RNA molecule. Note that these conversions will depend on cell type and single-cell RNA-seq assay, and for example, analyzing the same cells with Smart-seq and Smart-seq2 will lead to different conversions between RPKMs and RNA molecules (see **fig. S28** in Deng et al.[4]).

Identification of genes with imprinted or strain-biased allelic expression
To identify imprinted genes, we translated the CAST/EiJ (CAST) and C57BL/6J (C57) calls (see the section *Alignment of reads and allelic calling for mouse cells*) to maternal (mat) and paternal (pat) calls, according to the parental genotypes ($CAST_{mat}$ x $C57_{pat}$ or $C57_{mat}$ x $CAST_{pat}$). To calculate gene-level *P*-values for imprinted allelic expression, we applied Fisher's exact test on the frequency of monoallelic expression in a parental-specific manner compared to the overall call frequencies (**Supplementary Fig. 6c**). To consider a gene imprinted, and for inclusion in **Supplementary Table 3a**, the criteria were coherent monoallelic parental-specific calls in at least 90% of the cells and $P<0.05$ for the paternal bias in monoallelic expressions. We identified genes with strain-biased allelic expression, likely resulting

from genetic strain-related differences, using the observed frequencies C57 or CAST monoallelic in cells (**Supplementary Fig. 6b**), and we evaluated the frequency deviation from *P*=0.5 using a binominal test (**Supplementary Fig. 6d**). Known (www.geneimprint.com) and novel imprinted genes were filtered out before performing the strain analyses.

Calling the identity of the active X chromosome
To call the identity of the active X chromsome (Xa) in fibroblasts (C57 Xa or CAST Xa) we calculated the fraction X-chromosome-derived reads from C57 and CAST in each cell (shown in **Supplementary Figure 8a**). This separated the female cells into two distinct groups, either having the majority of X-chromosome reads derived from the C57 or the CAST copy, and the Xa identity in each cell was called according to this separation.

Analysis of clonal random monoallelic expression
Clone-consistent random monoallelic expression was investigated using two different approaches that estimated the total clonal aRME effect and identified individual genes with clonal aRME expression in a clone. First, to estimate the percentage of genes with clonal aRME in mouse primary fibroblasts or human *in vivo* T-cells we compared the observed frequency of clone-consistent monoallelic calls over the cells of each clone separately to the expected clone-consistent aRME deriving from dynamic aRME through "*in silico* pooling" of cellular allelic calls (**Supplementary Fig. 6e**). Observed clone-consistent monoallelic expression was computed as the percentage of detected genes (i.e. genes with at least one allelic call in the cells of a clone) with allele-consistent monoallelic expression in the clonal cells. The expected background frequency of clone-consistent monoallelic expression (being consistent monoallelic over the cells by random chance alone, due to transcriptional bursting and/or technical dropout of RNA species) was computed through *in silico* pooling of non-clonal cells and/or allelic calls. We evaluated the background frequencies derived from *in silico* pooling of either cells or allelic calls and both approaches yielded similar background estimates. Importantly, both randomization approaches retain important gene-level features within the data, such as the relationship between expression levels of genes and the frequency at which biallelic or monoallelic expression are observed. Moreover, we randomly sampled cells or allelic calls from clonally non-related cells sharing as many cellular features as possible with clonal cells. For example, when evaluating clonal aRME in a clone, we sampled randomly from non-clonal fibroblast of simular cellular sizes to not introduce bias. For all the analyses presented in the study, we used randomized allelic calls, rather than cells, from non-clonal cells by performing 500 random samplings, since pooling of randomized calls also conserve possible clone-specific patterns of expressed genes. In each round, we replaced informative allelic calls per gene (i.e. calls assigned as biallelic, reference monoallelic, or alternative monoallelic) in the cells of each clone, with a random allele-informative call from the same gene in the non-clonal cells, according to the frequency with which the allelic calls occurred in the non-clonal cell population (**Supplementary Fig. 6b**). The clonal aRME (reported in **Supplementary Fig. 6e**, **Fig. 1e** and **Fig. 3c**) was thus defined as the observed percent of genes with clone-consistent monoallelic expression minus the median percent genes with clone-consistent monoallelic expression from the 500 random samplings of allelic calls from non-clonal cells.

Gene-level test to identify genes with clonal aRME
To identify individual genes with clonal aRME in fibroblasts and *ex vivo* T-cells we applied a gene-level test (**Supplementary Fig. 6f**), assessing the statistical significance of allelic call frequencies in a clone compared to the overall call frequencies in non-clonal cells as background (i.e. taking the genetic biases in allelic expression into account as the background expectation). We classified genes as subjected to clonal aRME based on *i)* Fisher's exact test *P*-value below 0.05 (one-sided test for either over-representation of reference-monoallelic calls or over-representation of alternative-monoallelic calls) *ii)* required consistent monoallelic calls in at least 90% of the cells with an allelic call for the particular gene and clone. The rationale for requiring 90% of cells to have clone-consistent calls was based on analyses of known imprinted genes expressed in fibroblasts (see the shape of the distribution in **Fig. 2b**), and to avoid false negatives from errors affecting individual cells throughout experiments. The number of eligible genes, stated for each clone-specific analysis as well as for the imprinting and strain analyses, denote the number of genes for which the *P*-value could theoretically have been called significant using the Fischer's exact test on the available call frequencies (e.g. genes with less than 3 allelic calls in a clone were not considered to be eligible). All genes with $P<0.05$ in the imprinting test, as well as previously known imprinted genes (www.geneimprint.com) were filtered out in both the pooling and the gene-level analyses for fibroblasts. To account for the multiple testing performed above, we computed the expected number of false positives (E-values) using the identical criteria specified above,

but on randomly sampled non-clonal cells. We calculated the expected number of false positive clonal aRME genes in the gene-level by 1000 randomizations per clone and reported the mean number of false clonal aRME genes at a given *P*-value threshold from these randomization as E-values (in **Fig. 2c-d**, **Fig. 3d** and **Supplementary Fig. 16**.

Identification of T-cell *in vivo* clones

T-cell clones were identified based on cDNA sequences corresponding to the TCR-alpha (TCRα) and TCR-beta (TCRβ) chains. To identify the sequences, reads were analyzed using the MiTCR platform[11] and a list of putative TCRα and TCRβ sequences was identified for each single cell RNA-seq library. The reads were filtered to remove unproductive sequences corresponding to either erroneous calls or PCR artifacts (as annotated in the MiTCR output). After this initial screening, a secondary pipeline was developed in which publicly available sequences in the NCBI and Ensemble gene repositories of TCRα and TCRβ related gene components were used as a reference to screen for reads aligning to these regions. Reads were aligned to this reference and then assembled using Velvet[12] and submitted to the International ImMunoGeneTics (IMGT) TCR sequence identifying platform[13]. The resulting hits were compared with the MiTCR screen to validate the individual sequences. In most cases either one or both of the TCR chains could be identified with both methods, and only cells with a high degree of certainty, presenting both the TCRα and TCRβ sequences were considered for analysis of their clonal identity. Because both TCR chains are generated as separate recombination events that include random nucleotide insertions and alternative gene segment usage, it is highly unlikely that two distinct T-cells will arise bearing identical both TCRα and TCRβ chains at the nucleotide level[14]. Cells with identical TCRα and TCRβ sequences were consequently classified as clones. The identification of HLA-B7 clones was considerable more successful than HLA-A2 clones, likely reflecting a smaller (~20-fold) starting population of HLA-B7 cells prior to T-cell activation, which is in agreement with earlier reports[15], thus resulting in higher numbers of sampled mitotically related cells at time points after vaccination. Clones with at least three cells were considered for further analysis of clonal aRME, resulting in 32 individual *in vivo* T-cell clones.

Identification of heterozygous SNPs in human donor T-cells

To generate a list of high-confidence SNPs in the male donor, we considered only heterozygous bases supported in the exome data and present in dbSNP (build 138) reference database. The criteria for inclusion were: at least three reads of each SNP variant in the exome data and at most 50-fold difference between the two, and that both SNP bases were validated by the single cell RNA-seq data (allelic call for the SNP in at least 1% of the cells). Allelic calling was performed in the same way as for mouse cells (i.e. ≥ 3 allele-informative reads to call an SNP expressed, and considered monoallelically expressed if one SNP had at least 50-fold more reads than the other SNP within a cells). To filter out SNPs expressed with strong genetic bias, e.g. in imprinted genes, and possible remaining false-positive SNPs, we further discarded SNPs with a detected bias for one base (reference or alternative) beyond 70% over all cells (**Supplementary Fig. 3c**), and all X-chromosome genes. This resulted in the inclusion of 1,010 confirmed autosomal SNPs expressed in the human donor's T-cells, corresponding to 806 unique autosomal genes post filtering. For genes with multiple SNPs we used the most informative SNP (i.e. the SNP with highest number of reads). T-cells expressed (RPKM>1) on average 1,846 genes, and the number of expressed genes varied between the acute and memory phase (**Supplementary Fig. 14a**). The number of eligible allele-informative genes in *ex vivo*-expanded T-cell clones ranged between 370 and 660 (**Supplementary Fig. 15**), and the number of allele-informative genes with a mean RPKM >20 was 399.

Allelic calling for human cells

For the analysis of allelic expression in T-cells, we used the same criteria and analyses as for fibroblasts (but using human reference/alternative SNPs rather than mouse CAST/C57 SNPs), considering genes with a mean expression of at least 1 or 20 RPKM over the group of T-cells considered, as stated in the figure legends. To confirm that an effect of sequence depth was not producing the differences in percent monoallelically expressed genes observed between day 15 (more transcriptionally active) and day 136 (less transcriptionally active) after vaccination, we re-sampled cells from the day 15 and day 136 populations so that the re-sampled sets had the same population distributions of number of mapped reads per cell (excluding ERCC reads) and observed that the difference in percent monoallelically expressed genes remained (**Supplementary Fig. 14b-c**).

GO analyses of genes correlating with cellular size and mRNA content
We used DEseq[16] to identify genes that were differentially expressed between large (diameter of dissociated cell: 25-35μm, n=22 cells) and small (diameter of dissociated cell: 10-20μm, n=46 cells) fibroblasts. This identified 137 genes with adjusted *P*-value < 0.05 (listed in **Supplementary Table 10**), that were used for the GO analysis presented in **Supplementary Table 7**. To identify genes that scaled with the total amount of mRNA in the cells, we calculated the Spearman correlation between the number of spike-in ERCC reads and the number of reads for each gene, normalized to the total number of reads per cell (to account for sequence depth). This analysis identified a large number of significant genes (complete list of genes and correlations in **Supplementary Table 11**) from which we used the top 1000 positively correlated genes (Spearman correlations to total mRNA content: 0.59 to 0.33, *P*-values: $5.5x10^{-17}$ to $5.8x10^{-5}$) and the top 1000 negatively correlated genes (Spearman correlations to total mRNA content: -0.51 to -0.21, *P*-values: $2.2x10^{-12}$ to $5.8x10^{-3}$) for the GO analysis presented in **Supplementary Table 8** and **Supplementary Table 9** respectively. GO analyses were performed using DAVID[17,18], using genes expressed in fibroblasts (genes with mean expression >1 RPKM) as the background set.

Cell-cycle classification on fibroblast cells
The cell-cycle phase of individual fibroblasts cells was identified by regressing the cell-cycle genes identified by Whitfield[19]. A fit was performed using the R function glmgam.fit(), as previously described by Brennecke et al.[20]. The 100 cell-cycle genes presenting the greatest variability were used for principal component analysis whereby three phases of the cell cycle could be separated.

References to Supplementary Note 2

1. Picelli, S. *et al*. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10,** 1096–1098 (2013).
2. Picelli, S. *et al*. Tn5 transposase and tagmentation procedures for massively-scaled sequencing projects. *Genome Res*. gr.177881.114 (2014). doi:10.1101/gr.177881.114
3. Dobin, A. *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
4. Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343,** 193–196 (2014).
5. Gendrel, A.-V. *et al*. Developmental Dynamics and Disease Potential of Random Monoallelic Gene Expression. *Developmental Cell* **28,** 366–380 (2014).
6. Eckersley-Maslin, M. A. *et al*. Random monoallelic gene expression increases upon embryonic stem cell differentiation. *Developmental Cell* **28,** 351–365 (2014).
7. Jeffries, A. R. *et al*. Stochastic Choice of Allelic Expression in Human Neural Stem Cells. *STEM CELLS* **30,** 1938–1947 (2012).
8. Li, S. M. *et al*. Transcriptome-wide survey of mouse CNS-derived cells reveals monoallelic expression within novel gene families. *PLoS ONE* **7,** e31751 (2012).
9. Ramsköld, D., Wang, E. T., Burge, C. B. & Sandberg, R. An Abundance of Ubiquitously Expressed Genes Revealed by Tissue Transcriptome Sequence Data. *PLoS Comput Biol* **5,** e1000598 (2009).
10. Storvall, H., Ramsköld, D. & Sandberg, R. Efficient and comprehensive representation of uniqueness for next-generation sequencing by minimum unique length analyses. *PLoS ONE* **8,** e53822 (2013).
11. Bolotin, D. A. *et al*. MiTCR: software for T-cell receptor sequencing data analysis. *Nat. Methods* **10,** 813–814 (2013).
12. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*. **18,** 821–829 (2008).
13. Lefranc, M.-P. *et al*. IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res*. **37,** D1006–12 (2009).
14. Paul, W. E. Fundamental Immunology. 1304 (2012).
15. Blom, K. *et al*. Temporal dynamics of the primary human T cell response to yellow fever virus 17D as it matures from an effector- to a memory-type response. *J. Immunol*. **190,** 2150–2158 (2013).
16. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol*. **11,** R106 (2010).
17. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*. **37,** 1–13

(2009).
18. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4,** 44–57 (2009).
19. Whitfield, M. L. *et al*. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell* **13,** 1977–2000 (2002).
20. Brennecke, P. *et al*. Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10,** 1093–1095 (2013).