Extensive cellular heterogeneity of X inactivation revealed by single-cell allele-specific expression in human fibroblasts.

Marco Garieri, Georgios Stamoulis, Xavier Blanc, Emilie Falconnet, Pascale Ribaux, Christelle Borel, Federico Santoni and Stylianos E. Antonarakis

www.pnas.org/cgi/doi/10.1073/pnas.1806811115

Materials and Methods

Samples

We established 6 different cell lines from 5 female individuals: 5 primary fibroblast cell lines and one lymphoblastoid cell line (Table S1). We captured 935 single-cell fibroblasts and 48 lymphoblastoid single cells. Lymphoblastoid cells obtained from one of the five female individuals (Figure 1A, Table S1) (1, 2).

Cell growth

Cells were cultured in DMEM GlutaMAX™ (Life Technologies) supplemented with 10% fetal bovine serum (Life Technologies) and 1% penicillin/streptomycin/fungizone mix (Amimed, BioConcept) at 37°C in a 5% CO_2 atmosphere as described (1).

Whole Genome Sequencing

Genomic DNA was extracted for all five individuals using a QIAamp DNA Blood Mini Kit (Qiagen) and fragmented by Covaris to peak sizes of 300–400 bp. Libraries were prepared with TruSeq DNA kit (Illumina) using 1 μg of gDNA and sequenced on an Illumina HiSeq 2000 machine with 2 x 100-bp as previously described(1). All experiments were performed using the manufacturer's protocols. All samples provided with an whole genome average coverage around 25x. For each individual, raw whole genome DNA sequences were analyzed using an in-house pipeline previously described. Briefly, we used the Burrows-Wheeler Aligner (BWA mem v.0.7.10) (3) to align the sequencing reads (fastq) to the human reference genome (GRCh37/hg19). We used SAMtools v.1.4 (4) to remove paired-end duplicates and pile up the remaining reads. BCFtools v.1.4 was used to call the SNVs and Annovar (2016Feb01) (5) for the annotation. SNVs with quality score <100 where excluded from the analysis.

Similarly to Santoni et al. (2), we only used uniquely mapped reads for SNV calling and, in general, variants falling inside repeated regions such as segmental duplications or repeats (according to RepeatMasker) were filtered out (6).”

Single-cell capture

Single-cell capture was performed using the C1 single-cell auto prep system (Fluidigm) following the manufacturer’s procedure(1). The integrated fluidic circuit used for the study is the C1™ Single-Cell mRNA Seq IFC, 17–25 μm with a capacity of 96 chambers. During the capture, all 96 chambers were inspected under an inverted phase contrast microscope; only chambers containing a non-damaged single cell were considered for downstream analysis.

Single-cell RNA-sequencing

SMARTer Ultra Low RNA kit for Illumina sequencing (version 2, Clontech) was used for the cell lysis and cDNA synthesis. Libraries were prepared with 0.3 ng of pre-amplified cDNA using Nextera XT DNA kit (Illumina) as described (1). Libraries were sequenced on an Illumina HiSeq2000 sequencer as 100 bp single-ended reads. RNA sequences were mapped with GEM (7). Uniquely mapping reads were extracted by filtering for mapping quality (MQ>=150). An in-house algorithm was used to quantify RPKM expression using GENCODE v26. Cells with less than 1 million uniquely mapped reads were excluded from further analysis (Figure S2).

Allele-specific expression and classification of escapee genes

For each gene on the X chromosome, the aggregate monoallelic ratio (AR) per cell was calculated by averaging the allelic ratio of the reads covering the respective heterozygous sites (AR = sum of number of reads from the active X allele / total SNV reads; $0 \leq AR \leq 1$).

Since we do not have the availability of parental genotype for all the individuals, we designed an algorithm to estimate the active X allele per site based on the assumption that the active X allele is, on average, more transcribed than the inactive X. We validated this assumption comparing the prediction of our algorithm with the phasing of twins' X alleles based on parental information (more details in Methods S1). According to this metric, inactivated genes cluster around AR=1 while known escapees appear as been uniformly distributed between $0.5 \leq AR \leq 0.95$ (linear phase of the cumulative distribution, Figure S12). As support of this observation, AR distribution of autosomal genes clearly indicates AR=0.95 as the threshold separating biallelically expressed genes from monoallelic expressed genes (Figure S13). Therefore, we consider a gene as escapee in the relaxed set when the aggregate AR is ≤ 0.95 in at least 1 individual and as escapee in the robust set when the aggregate AR is ≤ 0.95 in at least 2 individuals. To reduce the effect of allele dropout, we only consider for the analysis SNV sites covered by at least 5 reads in at least three cells. To reduce sampling bias effects (8) a gene is included in the analysis if detectable in more than 5 different cells and/or SNVs per sample.

Haplotype and multiple cells (doublets) detection

For each cell, the expressed haplotype was estimated by calculating the allelic ratio of each heterozygous site in the X chromosome as identified by whole genome sequencing, excluding sites in the PAR regions (PAR1: chrX:60001-2699520, PAR2: chrX:154931044-155260560) and in known escapee genes (see section Annotation of known escapee genes). The estimated haplotype of each cell was compared to all others through pairwise correlation based hierarchical clustering procedure. A comparison of cells expressing concordant and discordant haplotypes results in a correlation near 1 and -1 respectively. Doublets

simultaneously expressing both haplotypes cluster around an absolute correlation of 0.5 are identified and excluded from further analysis.

Annotation of the escapee genes

First, we curated a list of 190 previously observed escapee genes in different cell types and tissues according to the literature (9) (10-13) (14, 15)(Dataset S2). Specifically, we investigated the status of 115 known escapee genes with available informative heterozygous sites and being expressed in fibroblast and lymphoblast cell lines (Table S3). Second, we have appended the results published in two studies (16, 17) in Dataset S3. Genes detected as escapees in our studies, in absence of citation, have been labeled as novel escapee genes. Genes found as escapee in our study and found subject to inactivation in other studies have been labeled variable escapee genes.

Cell cycle phase assignment

G1, S, and G2M cell cycle stage related gene markers were obtained from CycleBase (18) Cells not expressing *MKI67* have been considered to be in G0 (19) The remaining cells were assigned to their respective cell cycle phase according to the expression of CycleBase genes with Cyclone (20)**.**

ACCESSION NUMBERS

Newly generated RNA and DNA sequencing data are deposited in the European Genome-phenome Archive (EGA, https://www.ebi.ac.uk/ega/) for controlled accesses; the study accession number is (EGASxxxx, to be determined).

Methods S1

In order to group the cells according to their expressed allele (the active chromosome), we applied correlation based hierarchical clustering according to whole X chromosome allelic specific expression (ASE) but excluding variants in known escapees (see Figure 2). At this step we group the cells expressing the inactivated genes (but from the **active** chromosome) from the same X chromosome (we label the two haplotypes as A and B). In other words, we know whether a pair of cells shares the same (AA or BB) or a different phase (AB or BA). We define the allelic ratio per cell c_i per informative site s_j as: $AR(c_i, s_j)$ = (reads in c_i covering s_j from the active X chromosome) / (total number of reads covering s_j).

Given the random inactivation of one of the two alleles, we first have to identify (phasing) the active allele in each cell and for each site. The inactive genes express only from the active X therefore phasing their respective sites is straightforward. On the other hand escapees express from both alleles, thus, following the observations in (21), we make the reasonable assumption that the expected expression from the active X is higher than the expression from the inactive X.

More specifically i) we consider the haplotypes of each cell according to A or B clustering; ii) we assign the phase to each site that maximizes the allelic expression in the highest number of cells.

Let's clarify with an example:

		C1	C2	C3	C4	C5
		A	**B**	**B**	**A**	**B**
s1	**a1**	8	0	4	10	9
	a2	3	10	23	2	7

Here we measure the amount of reads covering the site s_1 in 5 cells. According to clustering they express the active haplotypes reported in the first row. Accordingly the two possible active alleles are enlightened in yellow and blue. In order to choose which of the two is the active one we proceed as follows. The yellow allele presents with higher expression in 4 out of 5 cells (C1,C2,C3,C4) while the blue allele has higher expression in 1 out of 5 cells (C5). Therefore, for the site s_1, we consider as the active X allele the yellow one: C1(A=8), C1(B=10), C1(B=23), C1(A=10), C1(B=7).

In other sites we might find a non-decidable configuration. For example (the notation here is Cell(haplotype,a1,a2); a1 and a2 are not phased yet):

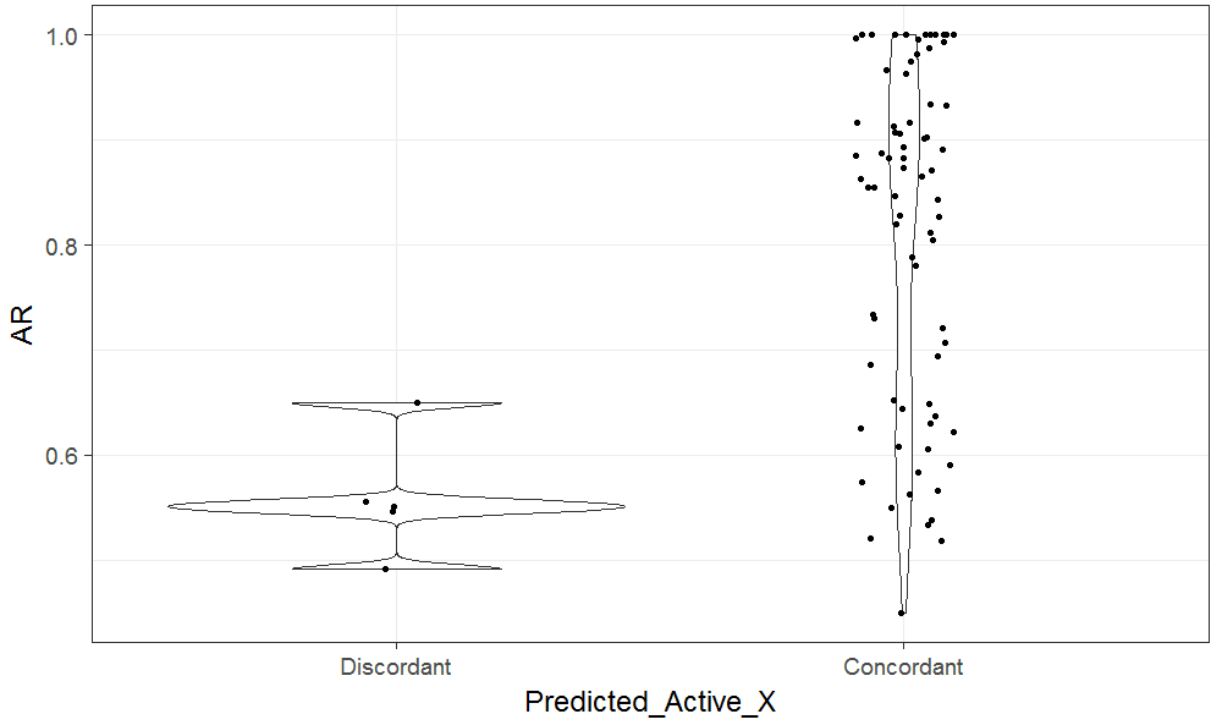
C1(A,4,5), C2(B,2,8), C3(B,7,1), C4(A,6,3)

Both A and B have higher expression in 2 out of 4 cells.

In this cases, we consider as active X alleles the configuration that maximises the two possible Allelic Ratios AR1 = 4+8+1+6/(4+5+2+8+7+1+6+3) = 19/36 and AR2 = 5+2+7+3/36 = 17/36. That is, max(AR1,AR2) = AR1 implies C1(A=4), C2(B=8), C3(B=1), C4(A=6).

To validate our assumption, we consider 77 sites in 12 escapees as detected in Individual 3 and 4 (Twins), the only individuals for which we have parental genotyping.

Figure 1 report on the x-axis the concordance between the prediction made with our approach and the effective active phase as calculated from the parents (respective normalized AR (if AR<0.5 then AR=1-AR) in the y-axis). Only 5 sites out of 79 are discordant (sensitivity = 94%). It is worth noting that the 5 discordant sites present a normalized AR ~ 0.5. When the reads are almost equally distributed between the alleles, our phase assignment method based on maximal expression is, obviously, quite noisy. However, in this case, the error is negligible (AR ~ 1-AR) and does not affect the detection of the escapees.



References

1. Borel C*, et al.* (2015) Biased allelic expression in human primary fibroblast single cells. *Am J Hum Genet* 96(1):70-80.
2. Santoni FA*, et al.* (2017) Detection of Imprinted Genes by Single-Cell Allele-Specific Gene Expression. *Am J Hum Genet* 100(3):444-453.
3. Li H & Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26(5):589-595.
4. Li H*, et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078-2079.
5. Wang K, Li M, & Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38(16):e164.
6. Panousis NI, Gutierrez-Arcelus M, Dermitzakis ET, & Lappalainen T (2014) Allelic mapping bias in RNA-sequencing is not a major confounder in eQTL studies. *Genome Biol* 15(9):467.
7. Marco-Sola S, Sammeth M, Guigo R, & Ribeca P (2012) The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat Methods*.
8. Deng Q, Ramskold D, Reinius B, & Sandberg R (2014) Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* 343(6167):193-196.
9. Carrel L & Willard HF (2005) X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* 434(7031):400-404.
10. Johnston CM*, et al.* (2008) Large-scale population study of human cell lines indicates that dosage compensation is virtually complete. *PLoS Genet* 4(1):e9.
11. Park C, Carrel L, & Makova KD (2010) Strong purifying selection at genes escaping X chromosome inactivation. *Molecular biology and evolution* 27(11):2446-2450.
12. Yasukochi Y*, et al.* (2010) X chromosome-wide analyses of genomic DNA methylation states and gene expression in male and female neutrophils. *Proc Natl Acad Sci U S A* 107(8):3704-3709.
13. Sharp AJ*, et al.* (2011) DNA methylation profiles of human active and inactive X chromosomes. *Genome research* 21(10):1592-1600.
14. Zhang Y*, et al.* (2013) Genes that escape X-inactivation in humans have high intraspecific variability in expression, are associated with mental impairment but are not slow evolving. *Molecular biology and evolution* 30(12):2588-2601.

15. Cotton AM*, et al.* (2013) Analysis of expressed SNPs identifies variable extents of expression from the human inactive X chromosome. *Genome Biol* 14(11):R122.
16. Balaton BP, Cotton AM, & Brown CJ (2015) Derivation of consensus inactivation status for X-linked genes from genome-wide studies. *Biol Sex Differ* 6:35.
17. Tukiainen T*, et al.* (2017) Landscape of X chromosome inactivation across human tissues. *Nature* 550(7675):244-248.
18. Santos A, Wernersson R, & Jensen LJ (2015) Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes. *Nucleic Acids Res* 43(Database issue):D1140-1144.
19. Schonk DM*, et al.* (1989) Assignment of the gene(s) involved in the expression of the proliferation-related Ki-67 antigen to human chromosome 10. *Human genetics* 83(3):297-299.
20. Scialdone A*, et al.* (2015) Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods* 85:54-61.
21. Reinius B*, et al.* (2016) Analysis of allelic expression patterns in clonal somatic cells by single-cell RNA-seq. *Nature genetics* 48(11):1430-1435.

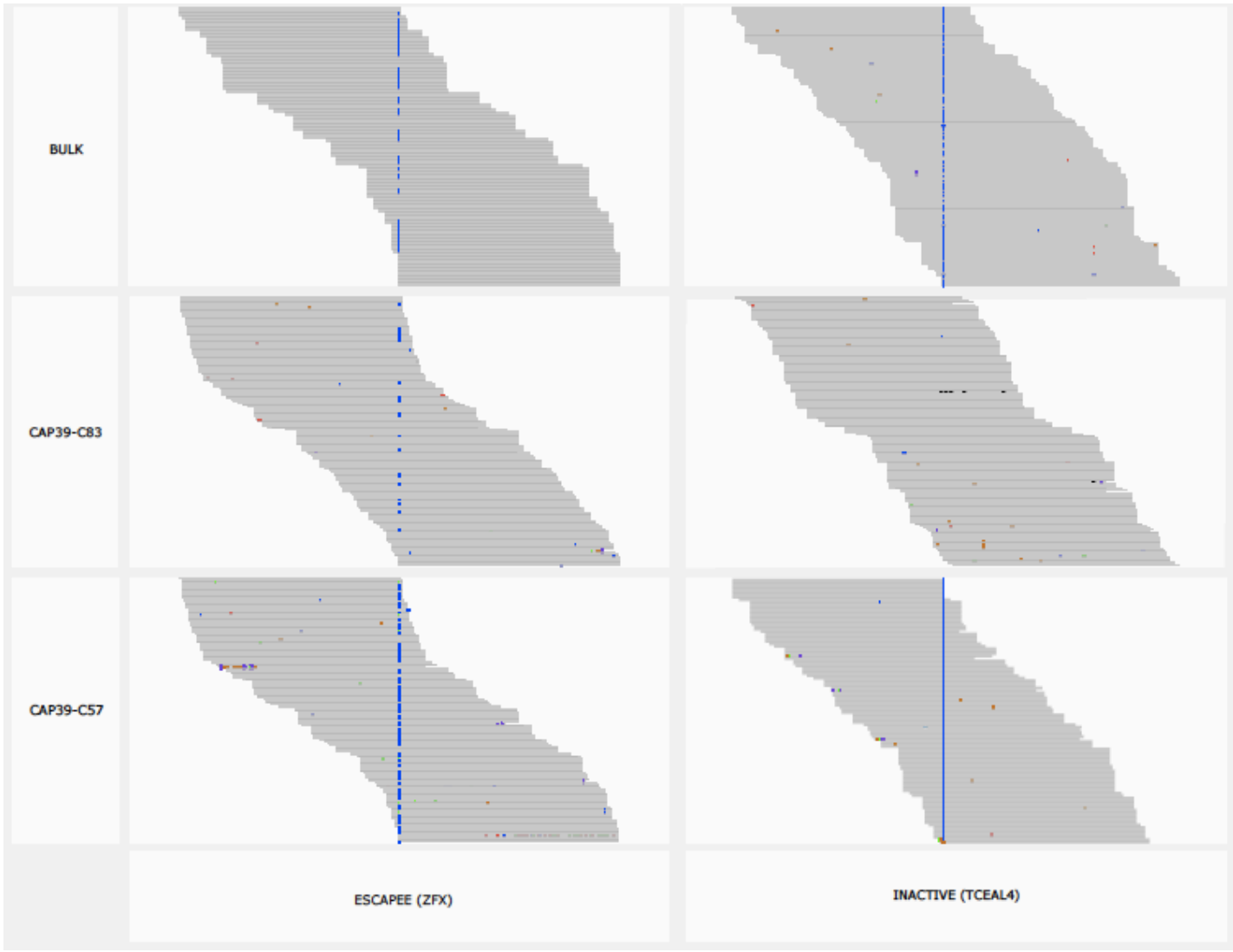


Figure S1: Distribution of reads mapping over two heterozygous sites (the blue vertical line) belonging to an escapee (left, ZFX) and an inactive gene (right, TCEAL4), respectively. Each horizontal line represents a single read mapping at nucleotide level: in grey are the reference alleles (nucleotides); in blue the alternative alleles; other colors are sequencing errors. In bulk, both sites appear as biallelically expressed. The escapee ZFX site is biallelically expressed at single cell level too. Inactive TCEAL4 site is monoallelically expressed from opposite haplotypes in the two single cells.

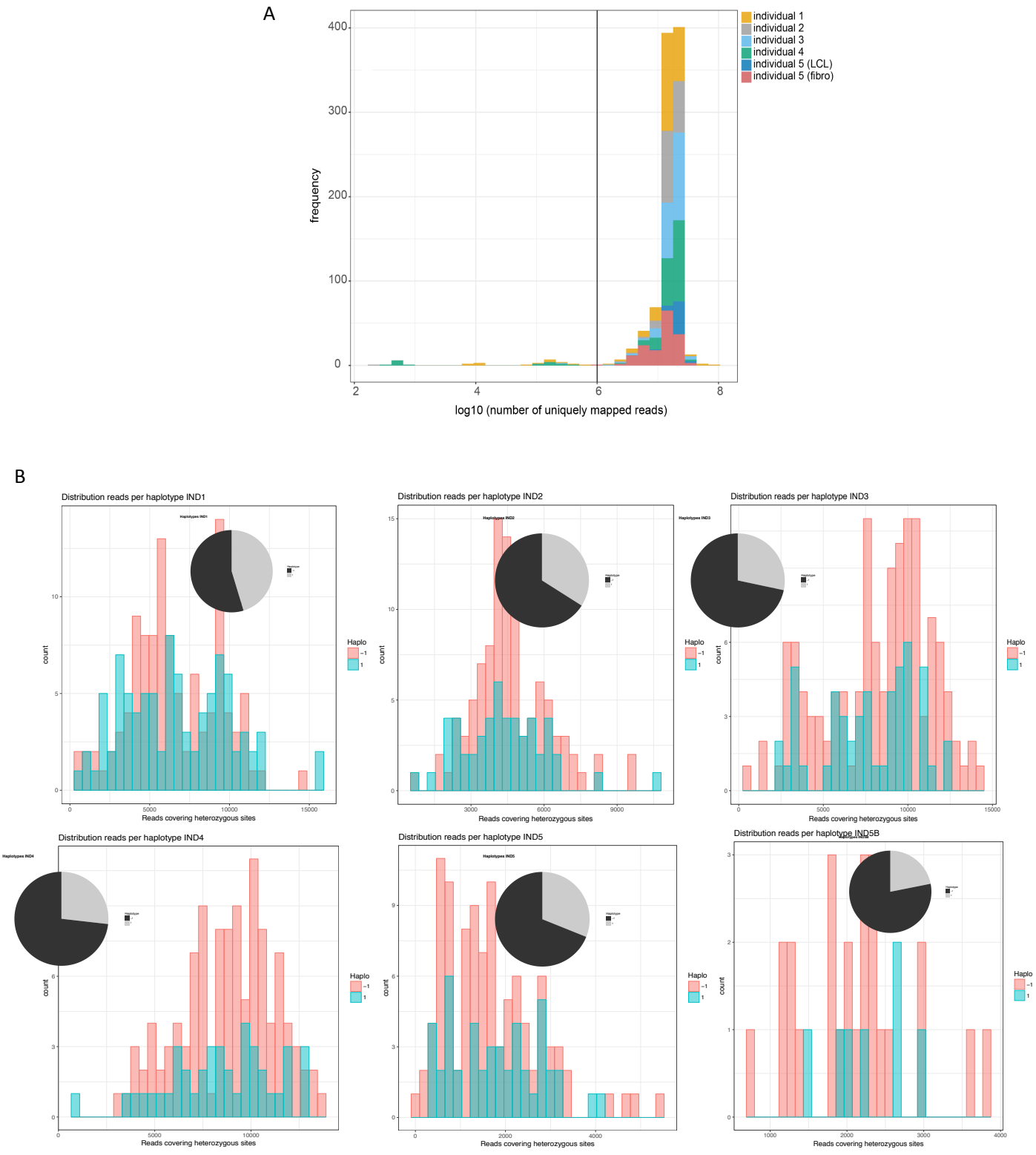


Figure S2: A) Distribution of number of uniquely mapped reads per individual (color coded) per single cell. Acceptance threshold is set at N=10^6 (vertical black line). B) Distribution of haplotype assigned reads and

respective fraction of haplotype expressing cells per individual. Ind5 and Ind5B indicate fibroblast and lymphoblastoid cells from Ind5, respectively.

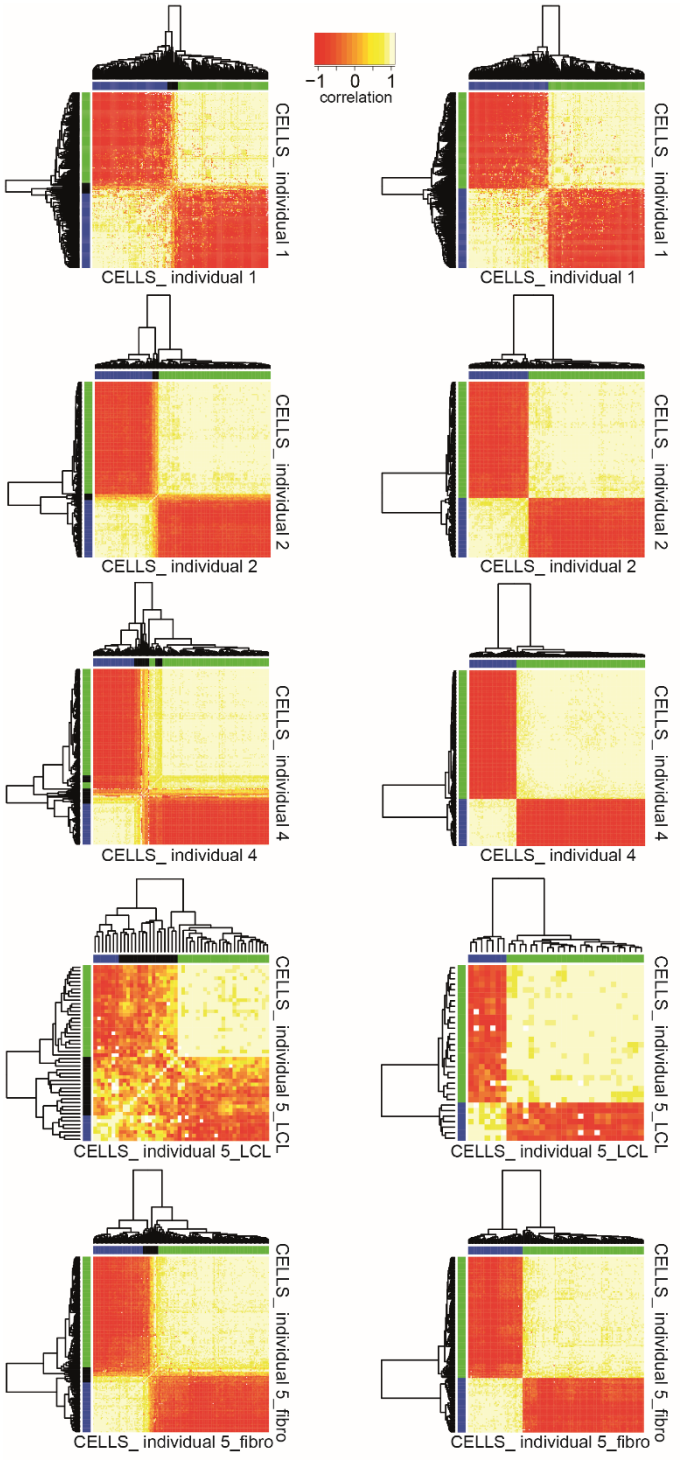


Figure S3: Hierarchical clustering based identification (left) and elimination (right) of confounding doublets (see text for details).

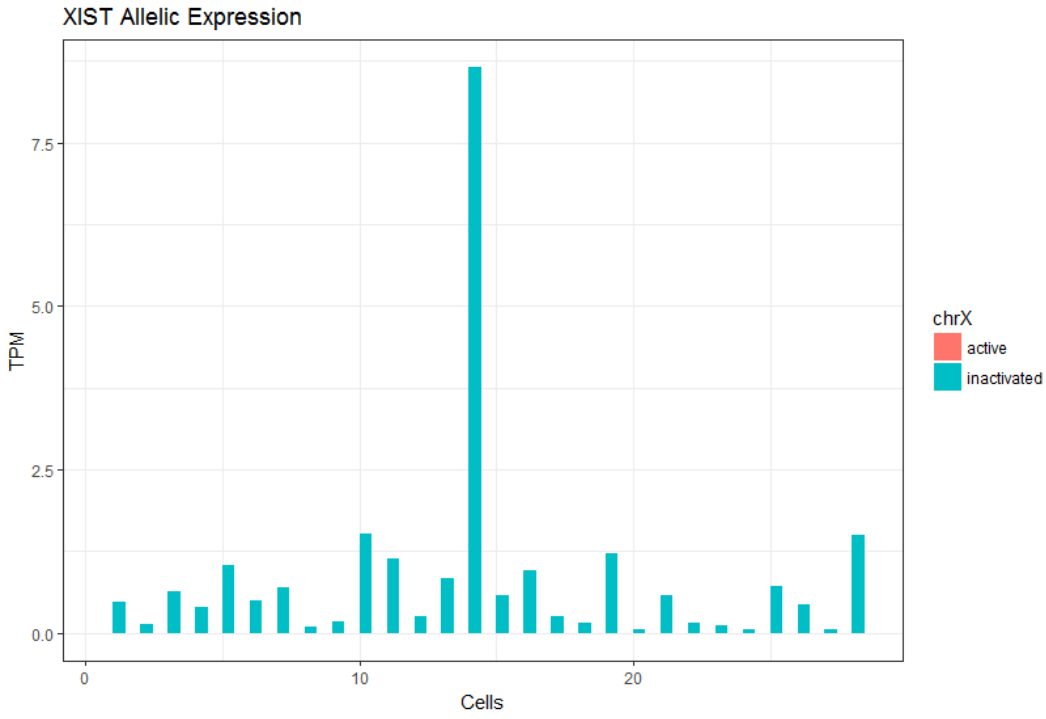


Figure S4. *XIST* phased Allelic Expression (active/inactive alleles) detected in 29 cells of two twin individuals (individual 3 and individual 4) participating in the study. *XIST* expression is detectable from the inactive X chromosome (blue) while *XIST* expression from the active X chromosome is not detectable (red). Alleles were phased using the available parental genotypes.

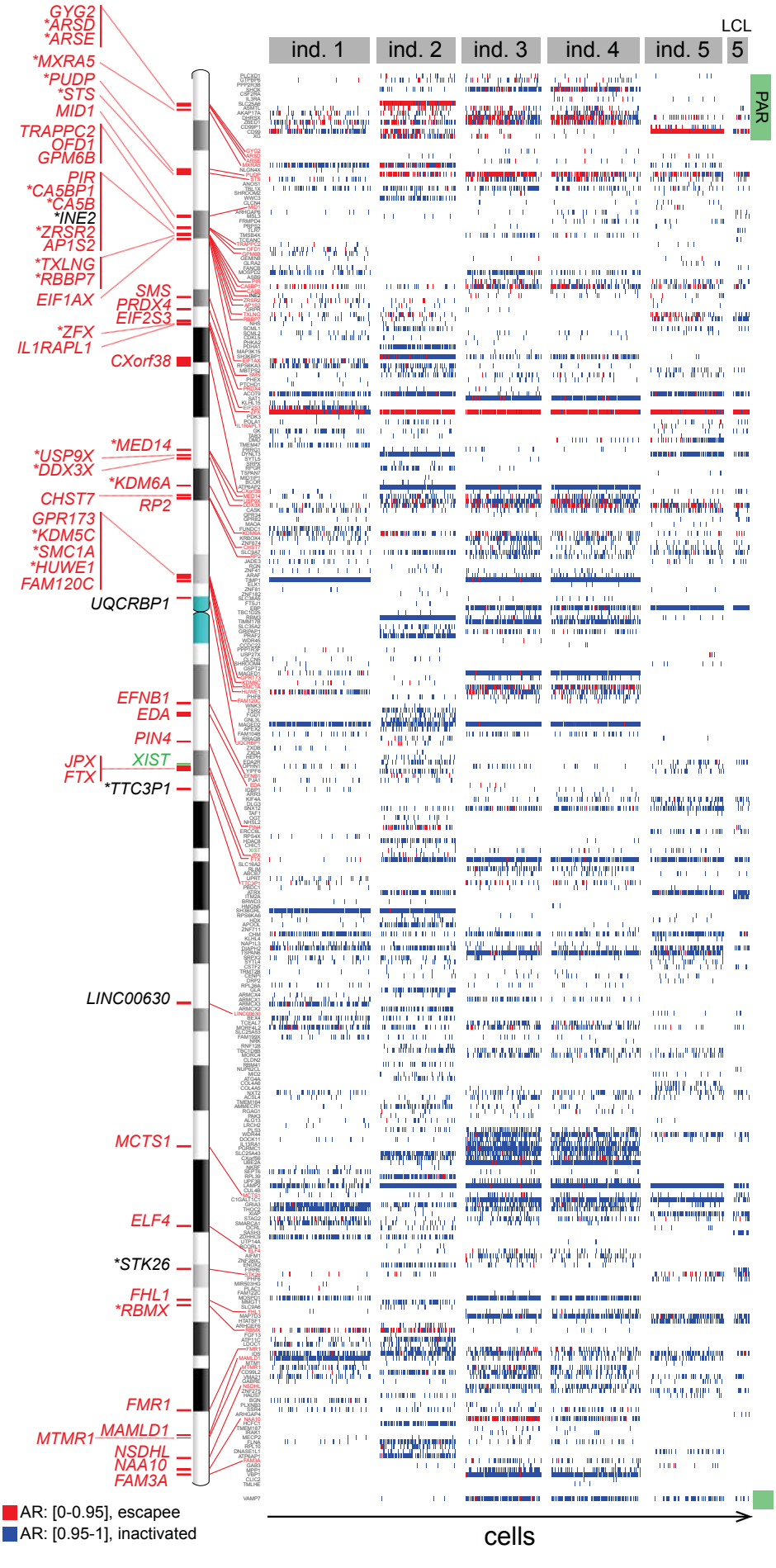

LCL
ind. 1
ind. 2
ind. 3
ind. 4
ind. 5
5
PAR
GYG2
*ARSD
*ARSE
*MXRA5
*PUDP
*STS
MID1
TRAPPC2
OFD1
GPM6B
PIR
*CA5BP1
*CA5B
*INE2
*ZRSR2
AP1S2
*TXLNG
*RBBP7
EIF1AX
SMS
PRDX4
EIF2S3
*ZFX
IL1RAPL1
CXorf38
*MED14
*USP9X
*DDX3X
*KDM6A
CHST7
RP2
GPR173
*KDM5C
*SMC1A
*HUWE1
FAM120C
UQCRBP1
EFNB1
EDA
PIN4
JPX
XIST
FTX
*TTC3P1
LINC00630
MCTS1
ELF4
*STK26
FHL1
*RBMX
FMR1
MTMR1
MAMLD1
NSDHL
NAA10
FAM3A
AR: [0-0.95], escapee
AR: [0.95-1], inactivated
cells

Figure S5. Single-cell allelic ratio profiles with respect to the active haplotype for genes on female X chromosome. For each individual, the allelic ratio for each gene (fibroblasts or lymphoblasts) is reported for each cell along the x-axis (rectangles with AR $\geq$ 0.95 (blue) an AR $\leq$ 0.95 (red)) according to the genomic location of genes in the human X chromosome (y-axis). 55 identified escapee genes in at least one individual are annotated on the left of the X chromosome ideogram. Known escapees are shown in red; novel escapees in black; escapees from the robust set with an asterisk. *XIST* is shown in green. PAR, pseudoautosomal regions; LCL, lymphoblastoid cells.

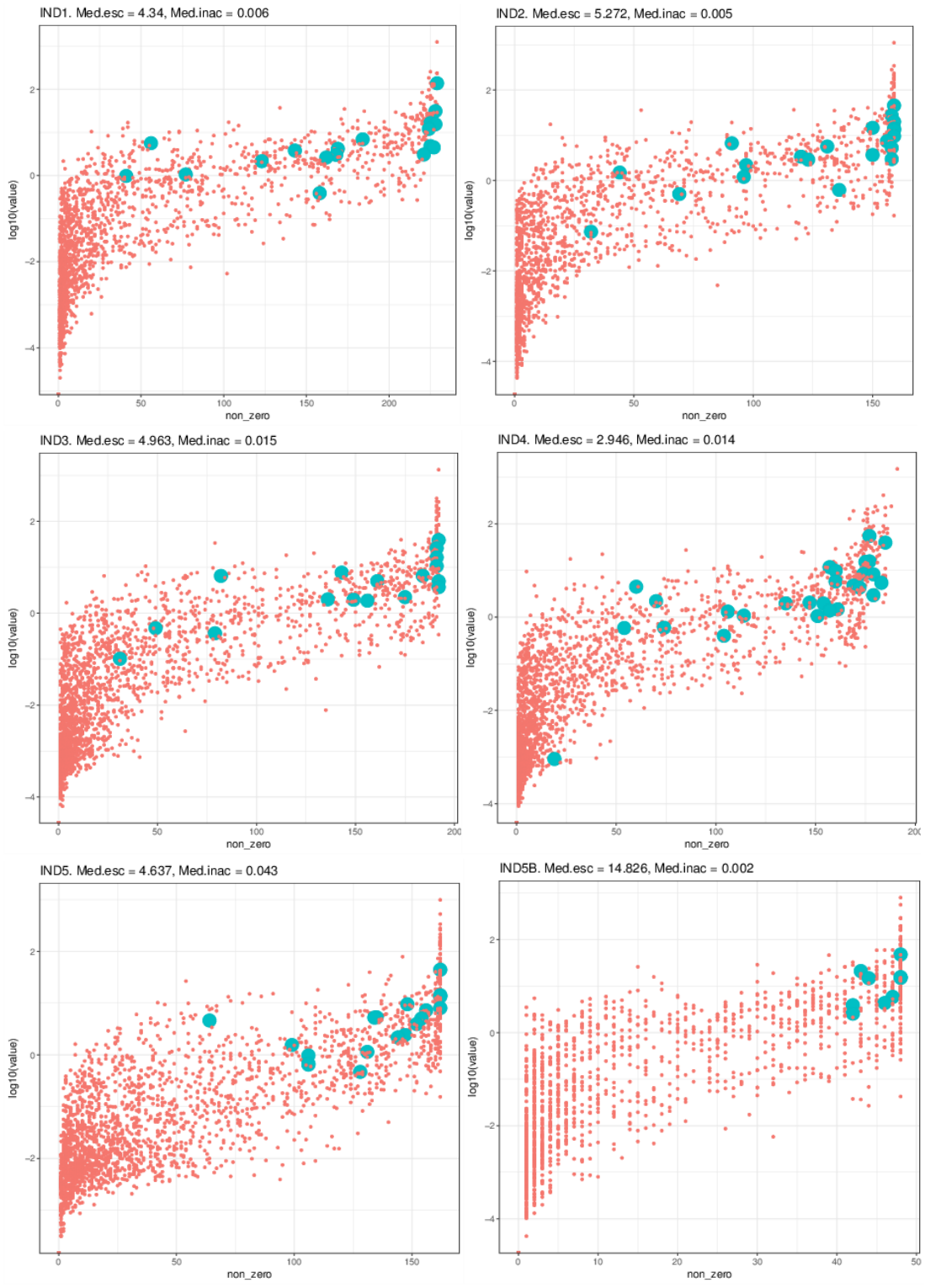


Figure S6. Individual scatter plots with log10 mean gene expression (y-axis) vs. the number of cells (where the respective gene is detectable >1 RPKM) (x-axis). Each dot corresponds to X-linked genes. Dots corresponding to escapee genes are light blue coloured and bigger in size. In all the individuals escapees mostly localize among the most expressed genes (median expression values of escapees and inactive genes are reported in the figure header). Ind5 and Ind5B indicate fibroblast and lymphoblastoid cells from Ind5, respectively.

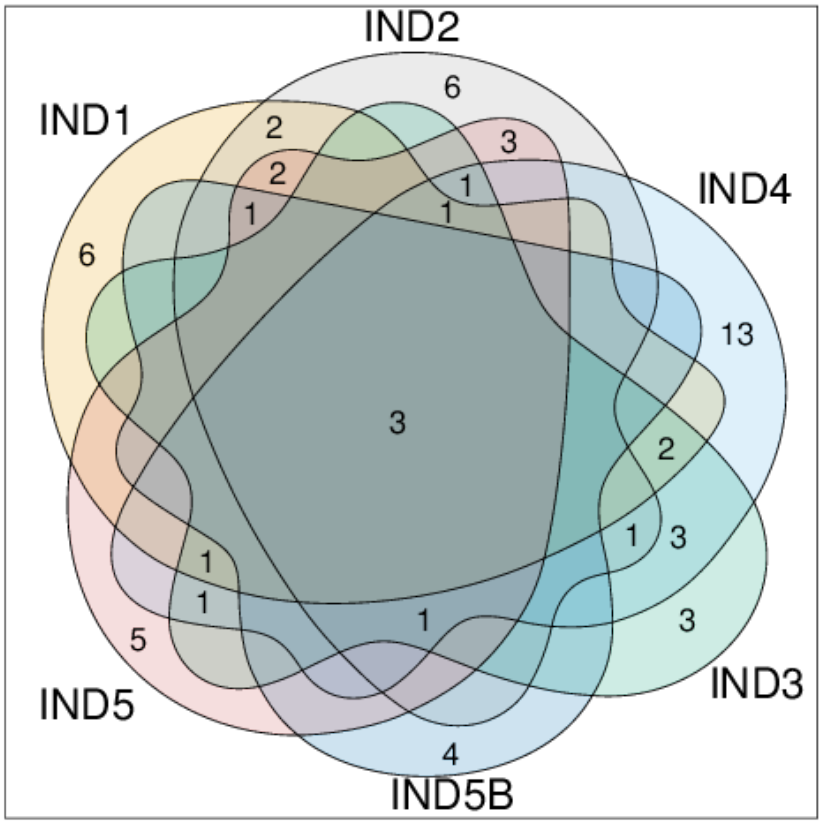


Figure S7. Venn diagram with the number of common detected escapees in the five individuals of the study. Ind5 and Ind5B indicate fibroblast and lymphoblastoid cells from Ind5, respectively.

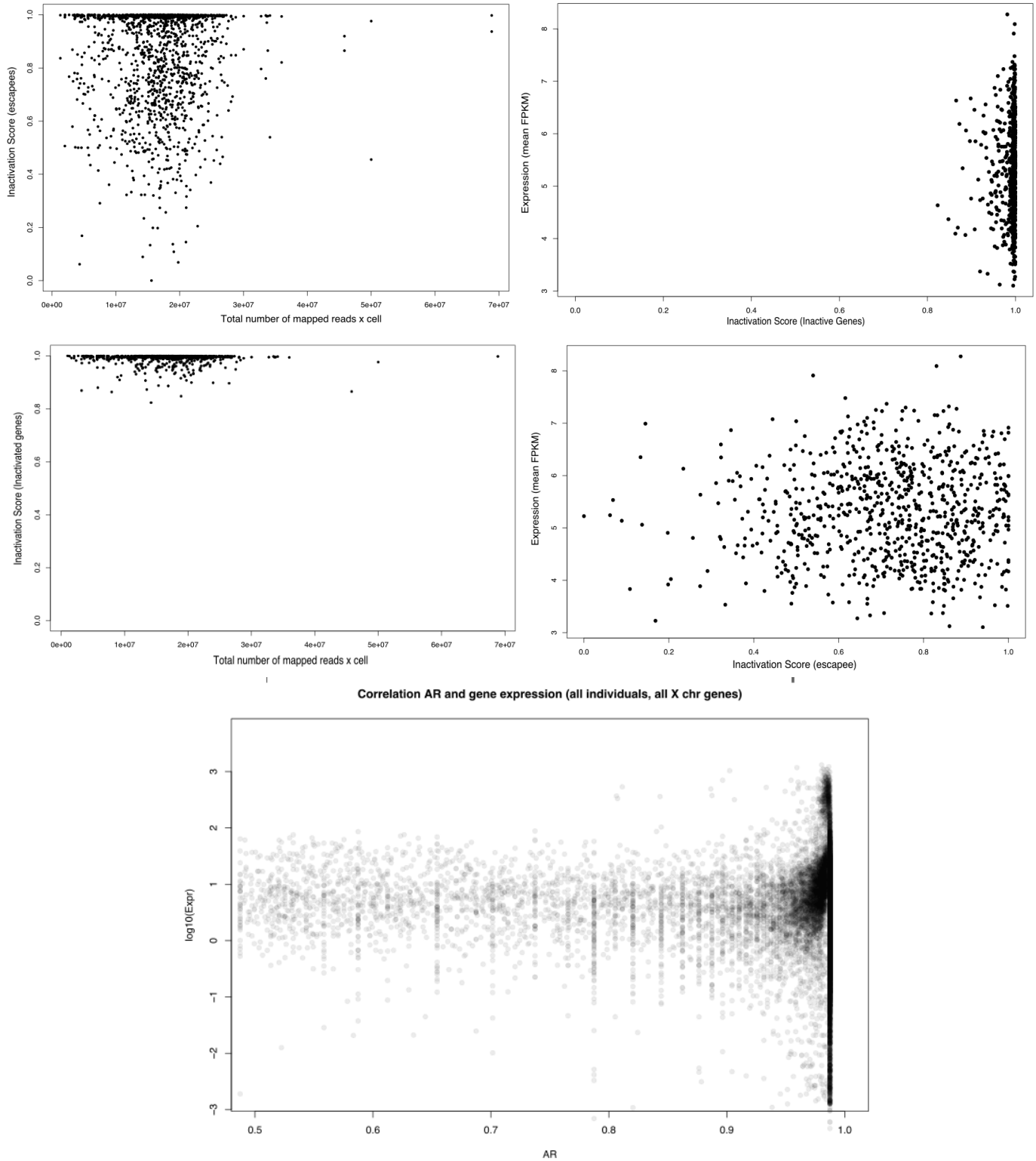
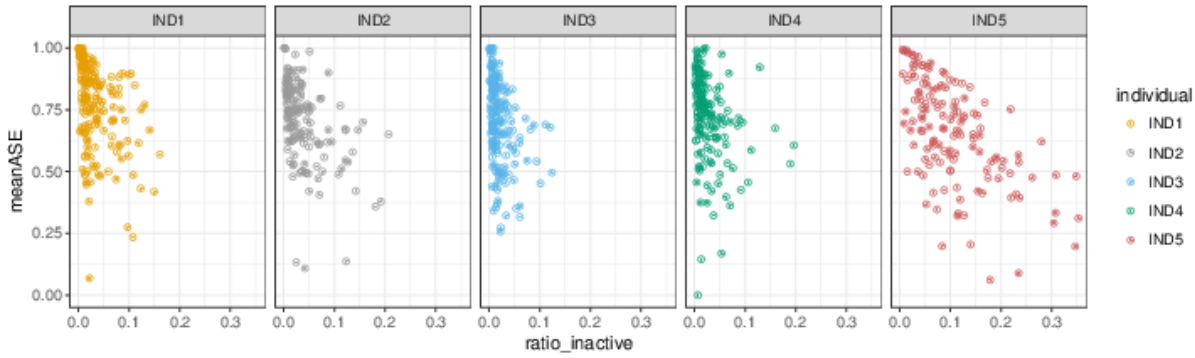
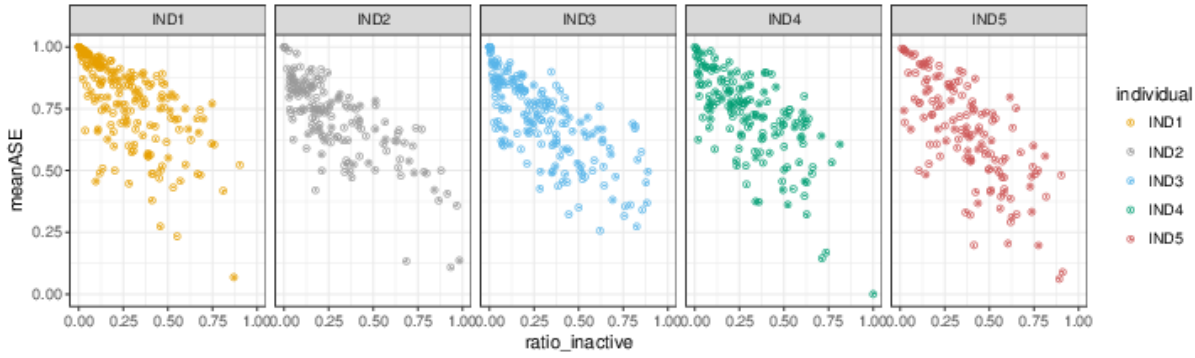


Figure S8. Scatter plots showing a) and b) the inactivation score and sequencing depth per cell for both escapee and inactivated genes and c) and d) between inactivation score and average gene expression per cell for escapee and inactivated genes. In all cases no significant association has been detected by linear regression (all R^2 <.0001; all p-values >0.2). e) Scatter plot representing mirrored Allelic Ratios ([0.5;1] monoallelic for reference or alternative =1; perfectly biallelic =0.5) x gene x cell versus expression x gene x cell. We observed a marginal weak positive correlation (Pearson rho = 0.03; p = 0.057), which is expected given that the majority of genes in the X chromosome (>80%) are monoallelically expressed.



IND	Correlation	pvalue
1	-0.4297624	7.523878e-10
2	-0.4713651	1.142752e-09
3	-0.3720871	3.405064e-07
4	-0.2832346	4.251378e-04
5	-0.5959510	1.950469e-14



IND	Correlation	pvalue
1	-0.6669009	1.501902e-25
2	-0.7609841	1.324647e-29
3	-0.7401815	5.547639e-32
4	-0.6912535	8.887710e-23
5	-0.7245293	2.073303e-23

Figure S9. For all individuals. a) Proportion of reads mapping on the informative alleles of the inactive X chromosome (considering all genes beside the ones in the PAR) vs. the Inactivation Score (meanASE). b) Proportion of reads mapping on the informative alleles of the escapee genes (Xist is not included) vs. the Inactivation Score. Correlation coefficients and respective p-values are reported in the related adjacent tables. Ind5 refers to fibroblast cells.

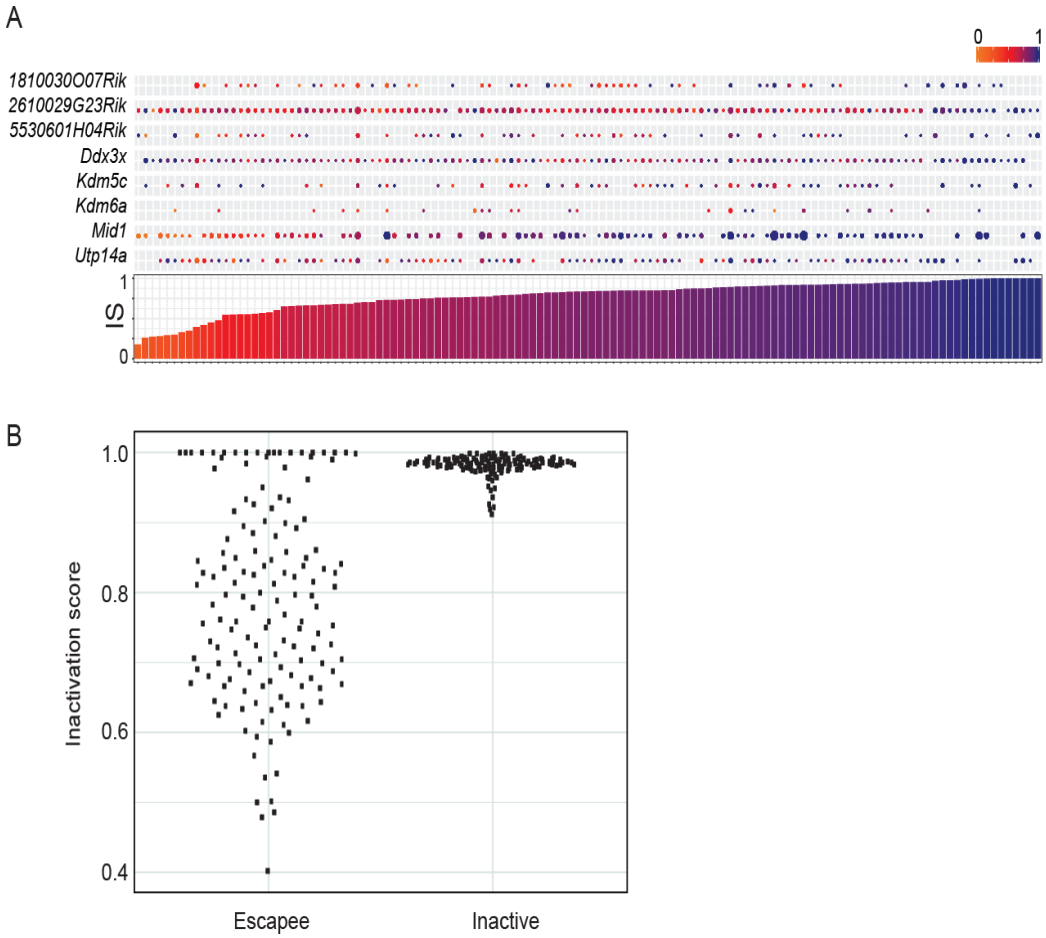


Figure S10. Single cell ASE profile of known escapee genes in mouse fibroblasts. A) Composite figure of individual allelic ratios per gene per cell **(Top of the panel)**. Allelic Ratio profile of robust escapee genes (listed the rows) with a detectable expression in single cells (ordered along the columns) is shown. Every dot represents the allelic ratio of the respective gene in a cell. AR ranges from 0 (light orange) to 1 (dark blue). The size of the dot is proportional to the respective number of reads detected per cell. (%) is the percentage of cells where the respective gene is escaping XCI. **(Bottom of the panel).** Bar plot representing the Inactivation Score (see text for details) per cell. IS ranges from 0 (light orange) to 1 (dark blue). B) Mouse cells ranked by the Inactivation Score calculated on all escapees in the robust set (left - Escapee) and on all inactive genes (right - Inactive). Each dot represents a fibroblast cell.

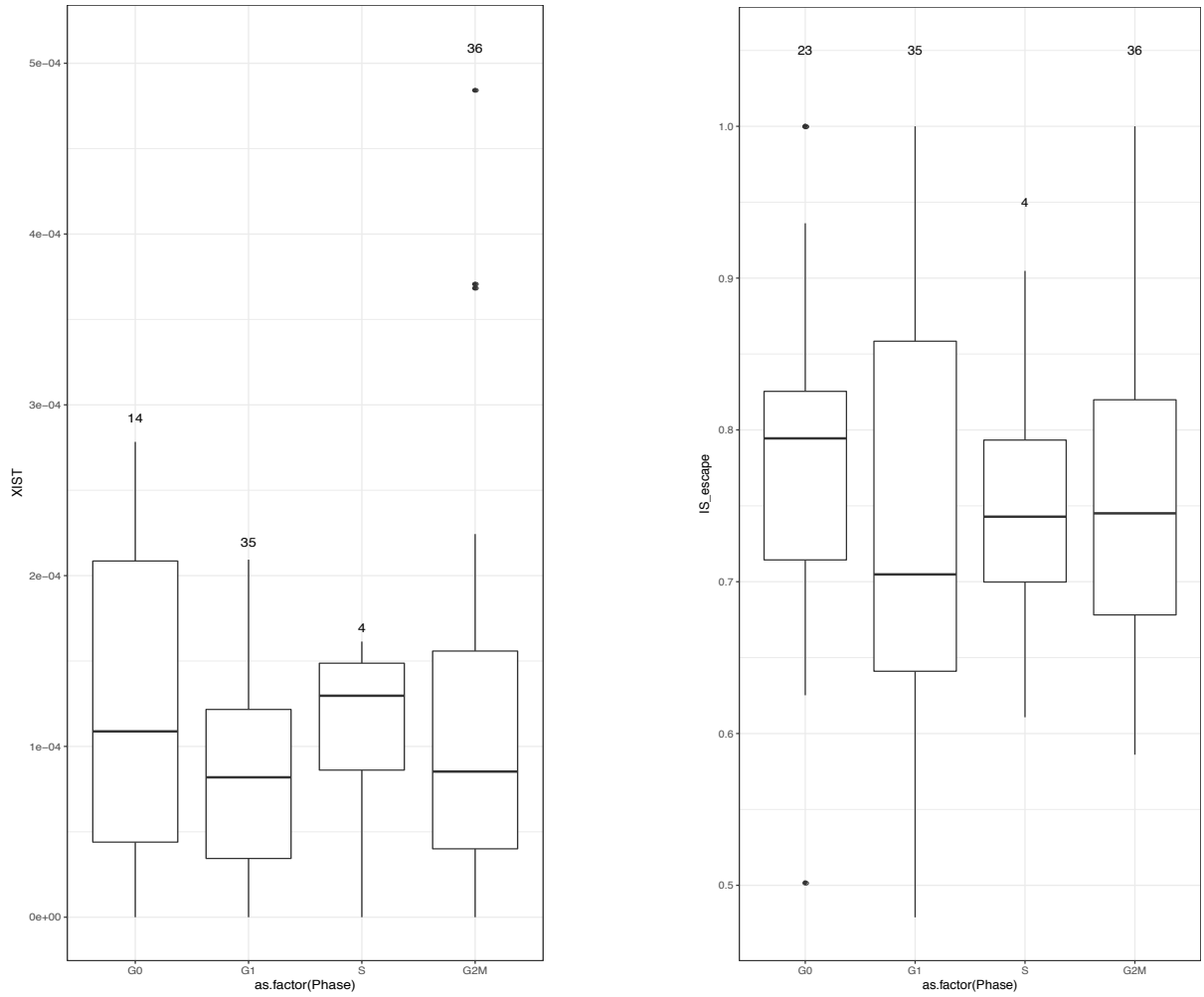


Figure S11. *XIST* expression and IS dependency on cell cycle phases in mouse cells. Distribution of *XIST* expression (left) and of Inactivation Score (right) according to G0, G1, S, G2M cell cycle phases (n=number of cell per phase). Due to the small amount of cells, no pairwise phase comparison reaches statistical significance.

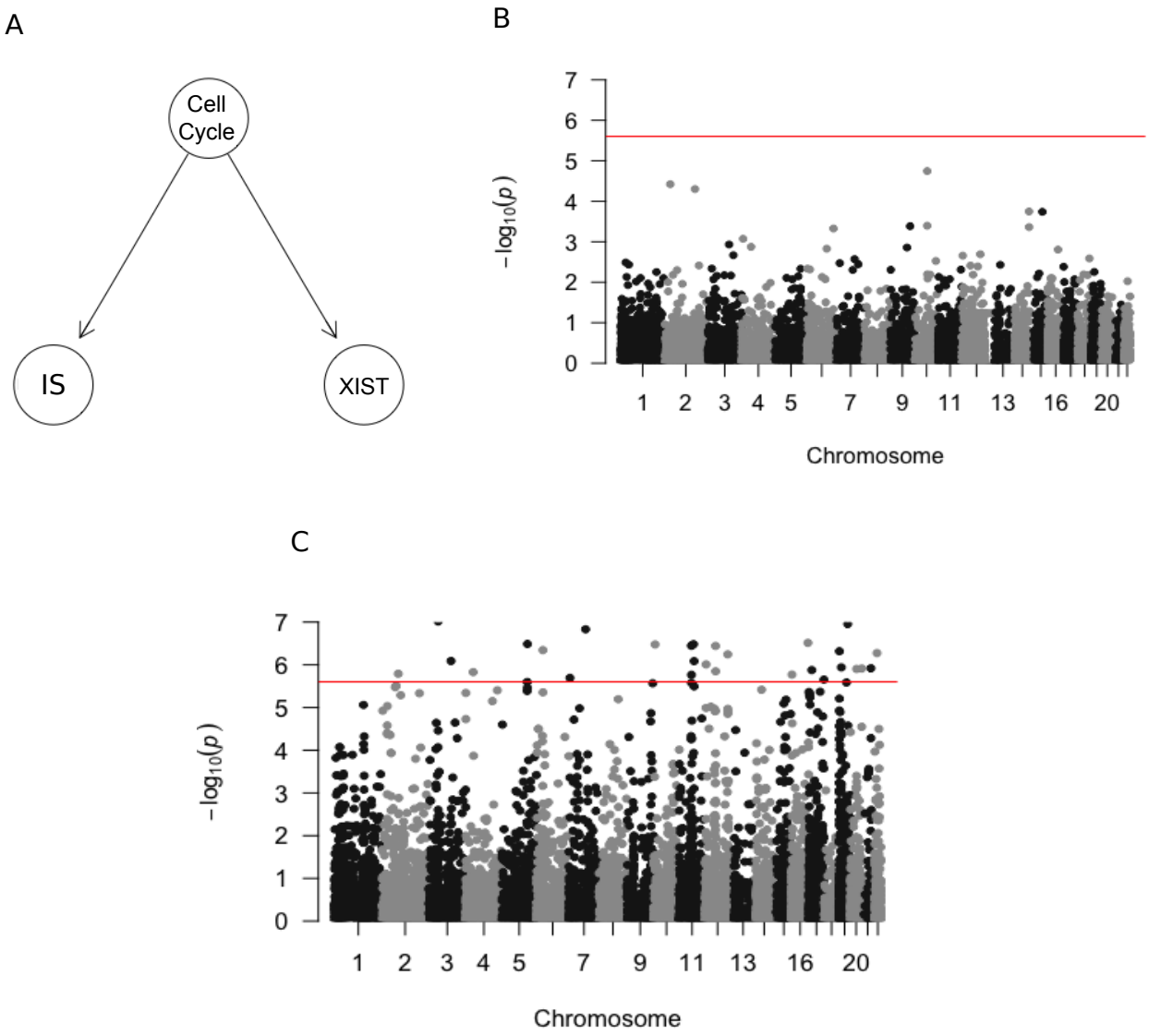


Figure S12. A) Bayesian network modelling the most likely (directional) interaction among IS, *XIST* and cell cycle phases (G0, G1, S, G2M). B) Correlation between autosomal gene expression and IS. Whole genome distribution p-values of Pearson correlation >0, C) same as B with Pearson correlation <0. Horizontal red line indicates Bonferroni corrected threshold for statistical significance (α_{THR}=2.5E-5).

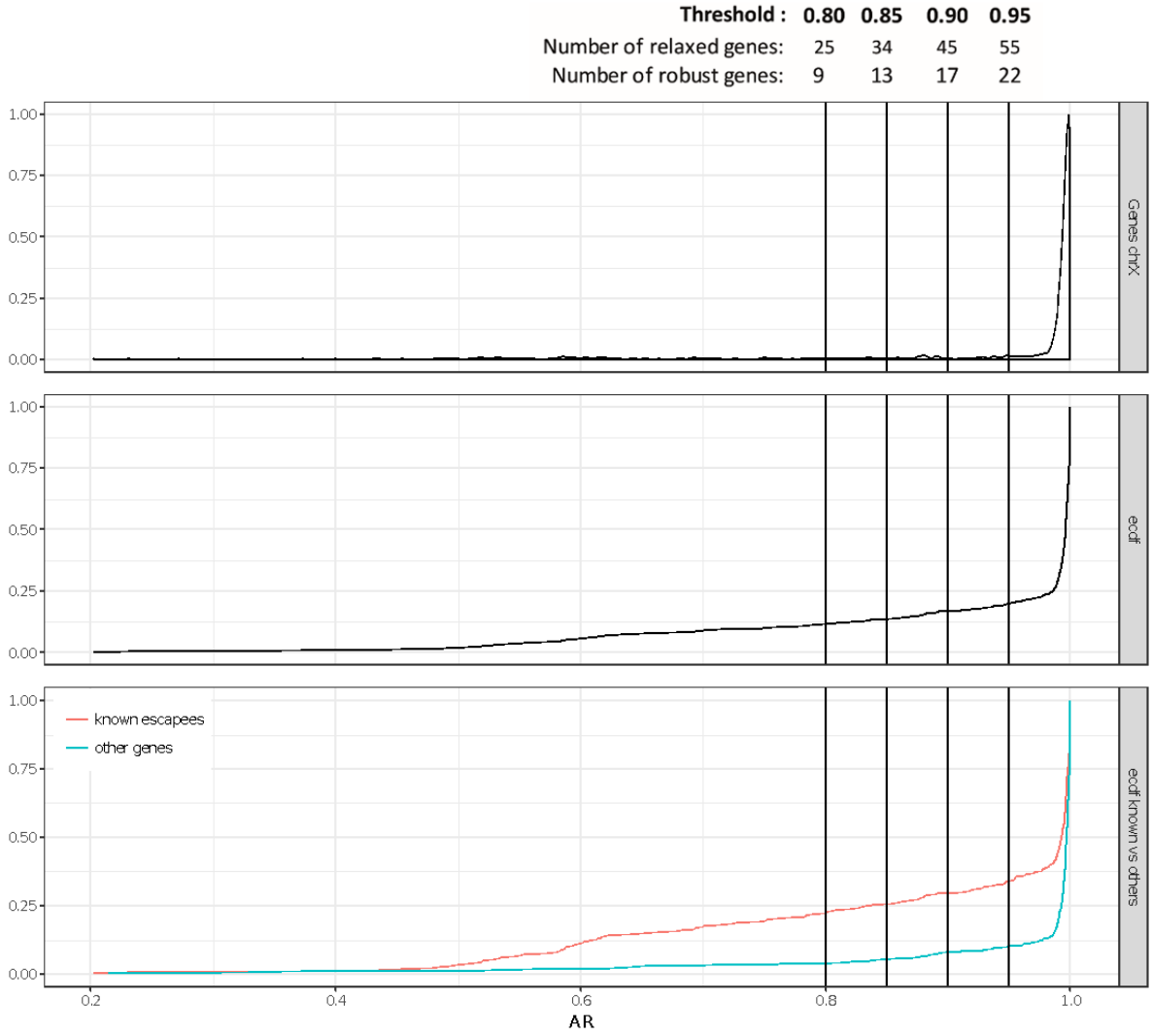


Figure S13. A) Allelic Ratio density distribution per gene per individual. B) Cumulative density distribution. C) Decomposition of the cumulative distribution for known escapees (red) and remaining genes (blue). The linear phase of the cumulative distribution ($0.5 \leq AR \leq 0.95$) is dominated by known escapees. Vertical black lines represent AR threshold at 0.95, 0.90, 0.85, 0.80. The number of genes considered as escapee in the interval ($0 < AR <$ Threshold) is linearly decreasing with the threshold (reported above the black vertical lines).

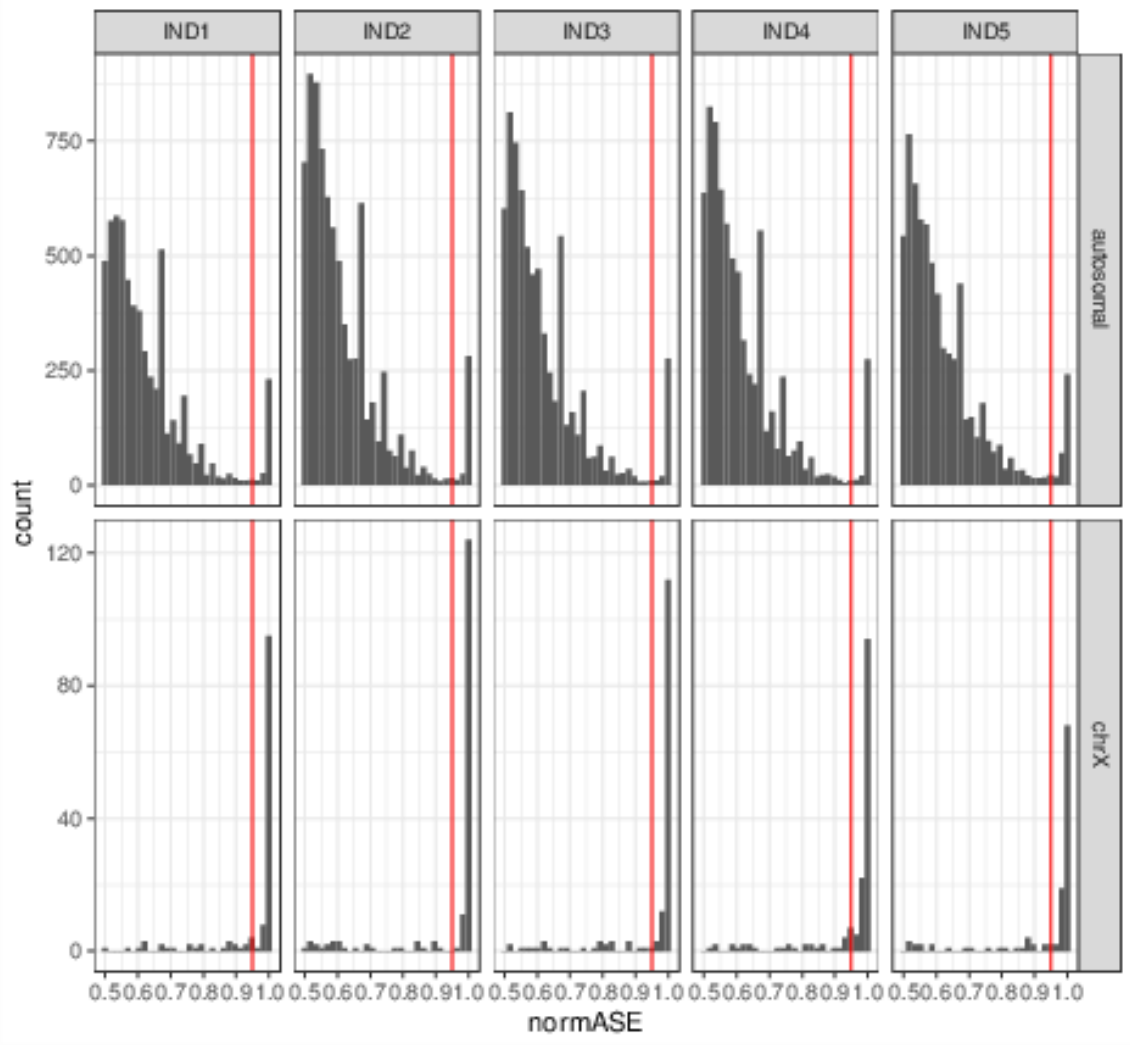


Figure S14. Allelic Ratio distribution of all genes per individual in autosomal and X chromosomes. Vertical red lines indicate AR = 0.95 (for the ease of representation, AR is normalized (normASE) between 0.5 and 1 according to the formula AR = 1-AR if AR< 0.5). Ind5 refers to fibroblast cells.

Table S1: Sample information including number of sequenced single cells and number of single cells after QC.

Sample name	Sample	Reference	Karyotype	Description	Number of sequenced single-cells	Number of single-cells after doublets removal
Individual 1	AG13074	Coriell	46, XX/47, XX, +18; 42.2%/57.8%	primary skin fibroblast	229	203
Individual 2	GM02596	Coriell	46, XX/47, XX, +8; 44.5%/55.5%	primary skin fibroblast	160	153
Individual 3	T1DS	(Dahoun et al. 2008)	47, XX, +21	primary skin fibroblast	192	180
Individual 4	T2N	(Dahoun et al. 2008)	46, XX	primary skin fibroblast	192	153
Individual 5	UCF1014	GenCord collection, (Borel et al. 2015)	46, XX	primary skin fibroblast	162	148
Individual 5	UCB1014	GenCord collection, (Borel et al. 2015)	46, XX	Epstein-Barr Virus-transformed B lymphoblastoid cell line (LCL)	48	32