LETTER

OPEN

doi:10.1038/nature24265

Landscape of X chromosome inactivation across human tissues

Taru Tukiainen[1,2], Alexandra-Chloé Villani[2,3], Angela Yen[2,4], Manuel A. Rivas[1,2,5], Jamie L. Marshall[1,2], Rahul Satija[2,6,7], Matt Aguirre[1,2], Laura Gauthier[1,2], Mark Fleharty[2], Andrew Kirby[1,2], Beryl B. Cummings[1,2], Stephane E. Castel[6,8], Konrad J. Karczewski[1,2], François Aguet[2], Andrea Byrnes[1,2], GTEx Consortium†, Tuuli Lappalainen[6,8], Aviv Regev[2,9], Kristin G. Ardlie[2], Nir Hacohen[2,3] & Daniel G. MacArthur[1,2]

X chromosome inactivation (XCI) silences transcription from one of the two X chromosomes in female mammalian cells to balance expression dosage between XX females and XY males. XCI is, however, incomplete in humans: up to one-third of X-chromosomal genes are expressed from both the active and inactive X chromosomes (Xa and Xi, respectively) in female cells, with the degree of 'escape' from inactivation varying between genes and individuals[1,2]. The extent to which XCI is shared between cells and tissues remains poorly characterized[3,4], as does the degree to which incomplete XCI manifests as detectable sex differences in gene expression[5] and phenotypic traits[6]. Here we describe a systematic survey of XCI, integrating over 5,500 transcriptomes from 449 individuals spanning 29 tissues from GTEx (v6p release) and 940 single-cell transcriptomes, combined with genomic sequence data. We show that XCI at 683 X-chromosomal genes is generally uniform across human tissues, but identify examples of heterogeneity between tissues, individuals and cells. We show that incomplete XCI affects at least 23% of X-chromosomal genes, identify seven genes that escape XCI with support from multiple lines of evidence and demonstrate that escape from XCI results in sex biases in gene expression, establishing incomplete XCI as a mechanism that is likely to introduce phenotypic diversity[6,7]. Overall, this updated catalogue of XCI across human tissues helps to increase our understanding of the extent and impact of the incompleteness in the maintenance of XCI.

Mammalian female tissues consist of two mixed cell populations, each with either the maternally or paternally inherited X chromosome marked for inactivation. To overcome this heterogeneity, assessments of human XCI have often been confined to the use of artificial cell systems[1] or to samples that have skewed XCI[1,2], that is, preferential inactivation of one of the two X chromosomes; this is common in clonal cell lines but rare in karyotypically normal, primary human tissues[8] (Extended Data Fig. 1and Supplementary Note). Others have used bias in DNA methylation[3,4,9] or in gene expression[5,10] between males and females as a proxy for XCI status. Surveys of XCI are powerful in engineered model organisms, for example, mouse models with completely skewed XCI[11], but the degree to which these discoveries are generalizable to human XCI remains unclear given marked differences in XCI initiation and the extent of escape across species[7]. Here we describe a systematic survey of the landscape of human XCI using three complementary RNA sequencing (RNA-seq)-based approaches (Fig. 1) that together enable the assessment of XCI from individual cells to population level across a diverse range of human tissues.

Given the limited accessibility of most human tissues, particularly in large sample sizes, no global investigation into the impact of incomplete XCI on X-chromosomal expression has been conducted in datasets spanning multiple tissue types. We used the Genotype-Tissue Expression (GTEx) project[12,13] dataset (v6p release), which includes high-coverage RNA-seq data from diverse human tissues, to investigate male–female differences in the expression of 681 X-chromosomal genes that encode proteins or long non-coding RNA in 29 adult tissues (Extended Data Table 1), hypothesizing that escape from XCI should typically result in higher female expression of these genes. Previous work[5,10,14] has indicated that some of the genes that escape XCI (hereafter referred to as escape genes) show female bias in expression, but our analysis benefits from a larger set of profiled tissues and individuals, as well as the high sensitivity of RNA-seq.

To confirm that male–female expression differences reflect incomplete XCI, we assessed the enrichment of sex-biased expression in known XCI categories using 561 genes with previously assigned XCI status, defined as escape ($n = 82$), variable escape ($n = 89$) or inactive ($n = 390$) (Fig. 1 and Supplementary Table 1). Sex-biased expression is enriched in escape genes compared to both inactive genes (two-sided paired Wilcoxon rank-sum test, $P = 3.73 \times 10^{-9}$) and variable escape genes ($P = 3.73 \times 10^{-9}$) (Fig. 2b and Extended Data Fig. 2), with 74% of escape genes showing significant (false-discovery rate (FDR) $q < 0.01$) male–female differences in at least one tissue (Fig. 2a, Extended Data Figs 3, 4 and Supplementary Table 2). In line with two active X-chromosomal copies in females, escape genes in the non-pseudoautosomal, that is, the X-specific, region (nonPAR) predominantly show female-biased expression across tissues (52 out of 67 assessed genes, binomial $P = 6.46 \times 10^{-6}$). However, genes in the pseudoautosomal region PAR1, are expressed more highly in males (14 out of 15 genes, binomial $P = 9.77 \times 10^{-6}$) (Fig. 2a), suggesting that combined Xa and Xi expression in females fails to reach the expression arising from X and Y chromosomes in males (discussed below).

Sex bias of escape genes is often shared across tissues; these genes show a higher number of tissues with sex-biased expression than genes in other XCI categories (Fig. 2a and Extended Data Fig. 2c), a result that is not driven by differences in the breadth of expression of escape and inactive genes (Extended Data Fig. 2e). Also, the direction of sex bias across tissues is consistent (Fig. 2a, c and Extended Data Fig. 2b). Together, these observations indicate that there is global and tight control of XCI, that potentially arises from early lockdown of the epigenetic marks regulating XCI. Previous reports have identified several epigenetic signatures associated with XCI escape in humans and mice[15]; in agreement with these discoveries we show that escape

[1]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [2]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. [3]Center for Immunology and Inflammatory Diseases, Massachusetts General Hospital, Charlestown, Massachusetts 02129, USA. [4]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [5]Department of Biomedical Data Science, Stanford University, Stanford, California 94305, USA. [6]New York Genome Center, New York, New York 10013, USA. [7]Center for Genomics and Systems Biology, Department of Biology, New York University, New York, New York 10003, USA. [8]Department of Systems Biology, Columbia University, New York, New York 10032, USA. [9]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.
†Lists of participants and their affiliations are included in the online version of the paper.

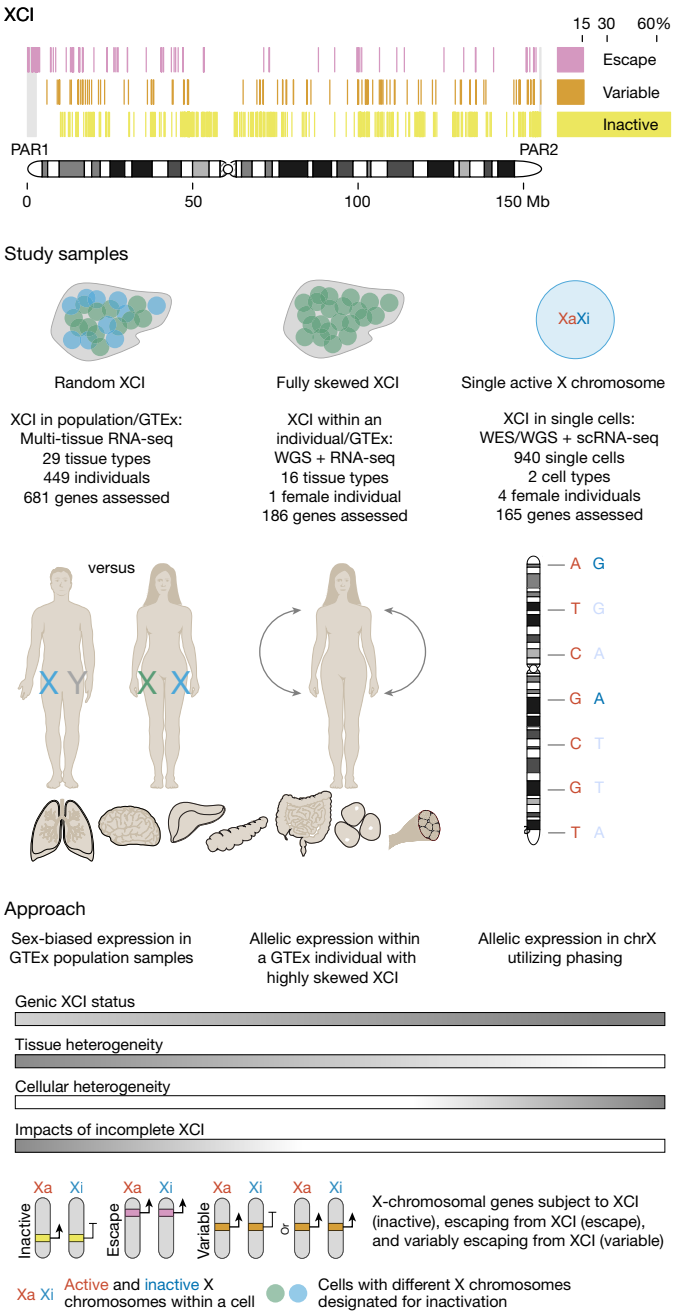


Figure 1 | Schematic overview of the study. Previous expression-based surveys of XCI[1,2] have established the incomplete and variable nature of XCI, but these studies have been limited in the tissue types and samples assessed. To investigate the landscape of XCI across human tissues, we combined three approaches: (1) sex biases in expression using population-level GTEx data across 29 tissue types; (2) allelic expression in 16 tissue samples from a female GTEx donor with fully skewed XCI, and (3) validation using scRNA-seq by combining allelic expression and genotype phasing. WGS, whole-genome sequencing; WES, whole-exome sequencing.

genes are enriched in chromatin states that are related to active transcription (Fig. 2e).

Although sex bias on the X chromosome is broadly specific to escape genes, some genes show unexpected patterns. Eight genes with some previous evidence for inactivation show >90% concordance in effect direction and significant sex bias (Fig. 2d and Supplementary Table 3), suggesting that variable escape can also have considerable population-level effects. For example, *CHM* demonstrates such concordance in sex bias and escape at this gene is confirmed when using single-cell RNA-seq (scRNA-seq; see below). One gene (*RP11-706O15.3*) without an assigned XCI status shows a similar sex bias pattern to escape genes. *RP11-706O15.3* resides between escape and variable escape genes *PRKX* and *NLGN4X* (Fig. 2d), consistent with known clustering of escape genes[1,2]. Some escape genes show more heterogeneous sex bias, for example, *ACE2* (Fig. 2a and Supplementary Discussion). Many such genes lie in the evolutionarily older region of the chromosome[16], in Xq, where escape genes also show higher tissue-specificity and lower expression levels (Extended Data Fig. 5), characteristics that have been linked with higher protein evolutionary rates[17,18].

Although sex bias serves as a proxy for XCI status, it provides only an indirect measurement of XCI. We identified a GTEx female donor with an unusual degree of skewing of XCI (Fig. 3a), in whom the same copy of chromosome X was silenced in approximately 100% of cells across all tissues, but with no X-chromosomal abnormality detected by whole-genome sequencing (Extended Data Fig. 6 and Supplementary Note), providing an opportunity to analyse allele-specific expression (ASE) across 16 tissues to investigate XCI. This approach is analogous to previous surveys in mouse[11] or in human cell lines with skewed XCI[2], but extends the assessment to a larger number of tissues and avoids biases arising from genetic heterogeneity between tissue samples.

Analysis of the X-chromosomal allelic counts (Supplementary Tables 4–6) from this GTEx donor highlights the incompleteness and consistency of XCI across tissues (Fig. 3b). Approximately 23% of the 186 X-chromosomal genes that were assessed show expression from both alleles, indicative of incomplete XCI, matching previous estimates of the extent of escape[1,2]. For 43% of the genes that were expressed from both alleles in this sample, Xi expression is of a similar magnitude between tissues, therefore supporting the observation of a general global and tight control of XCI. However, suggesting some tissue dependence in XCI, the rest of the genes that were expressed from both alleles show variability in Xi expression, including a subset of genes (5.8% of all genes) that appear biallelic in only one of the multiple tissues assayed. While tissue-specific escape is common in mouse[11], limited evidence exists for such a pattern in human tissues other than for neurons[3,4,9]. In our data, one of the genes with the strongest evidence for tissue-specific escape is *KAL1* (Fig. 3f and Supplementary Table 6), the causal gene for X-linked Kallmann syndrome. We show that *KAL1* shows biallelic expression exclusively in the lung (Fig. 3f), in line with the strong female bias detected specifically in lung expression in the analysis described above and in Fig. 2a, suggesting that tissue differences in escape can directly translate into tissue-specific sex biases in gene expression. The predictions of XCI status in this sample not only align with previous assignments (Fig. 3c–f and Supplementary Table 7, for example, *TSR2*, *XIST* and *ZBED1*) but also suggest five new incompletely inactivated genes (Fig. 3g–k and Supplementary Table 5), three of which act in a tissue-specific manner. For instance, *CLIC2*, which in previous studies was shown to either be subject to[2] or variably escape from[1] XCI, shows considerable Xi expression only in skin tissue. Such specific patterns illustrate the need to assay multiple tissue types to fully uncover the diversity in XCI.

The emergence of scRNA-seq methods[19] presents an opportunity to directly assess XCI without the complication of cellular heterogeneity in bulk tissue samples (Fig. 1), as demonstrated recently in mouse studies[20–23] and in human fibroblasts[24] and preimplantation development[25]. To directly profile XCI in human samples, we examined scRNA-seq data in combination with deep genotype sequences from 940 immune-related cells from four females: 198 cells from lymphoblastoid cell lines (LCLs) sampled from three females of African (Yoruba) ancestry, and 742 blood dendritic cells from a female of Asian ancestry[26] (Fig. 1 and Extended Data Table 2). We used ASE to distinguish the expression coming from each of the two X-chromosomal haplotypes in a given cell (Supplementary Table 4). Because the inference of allele-specific phenomena in single cells is complicated by widespread

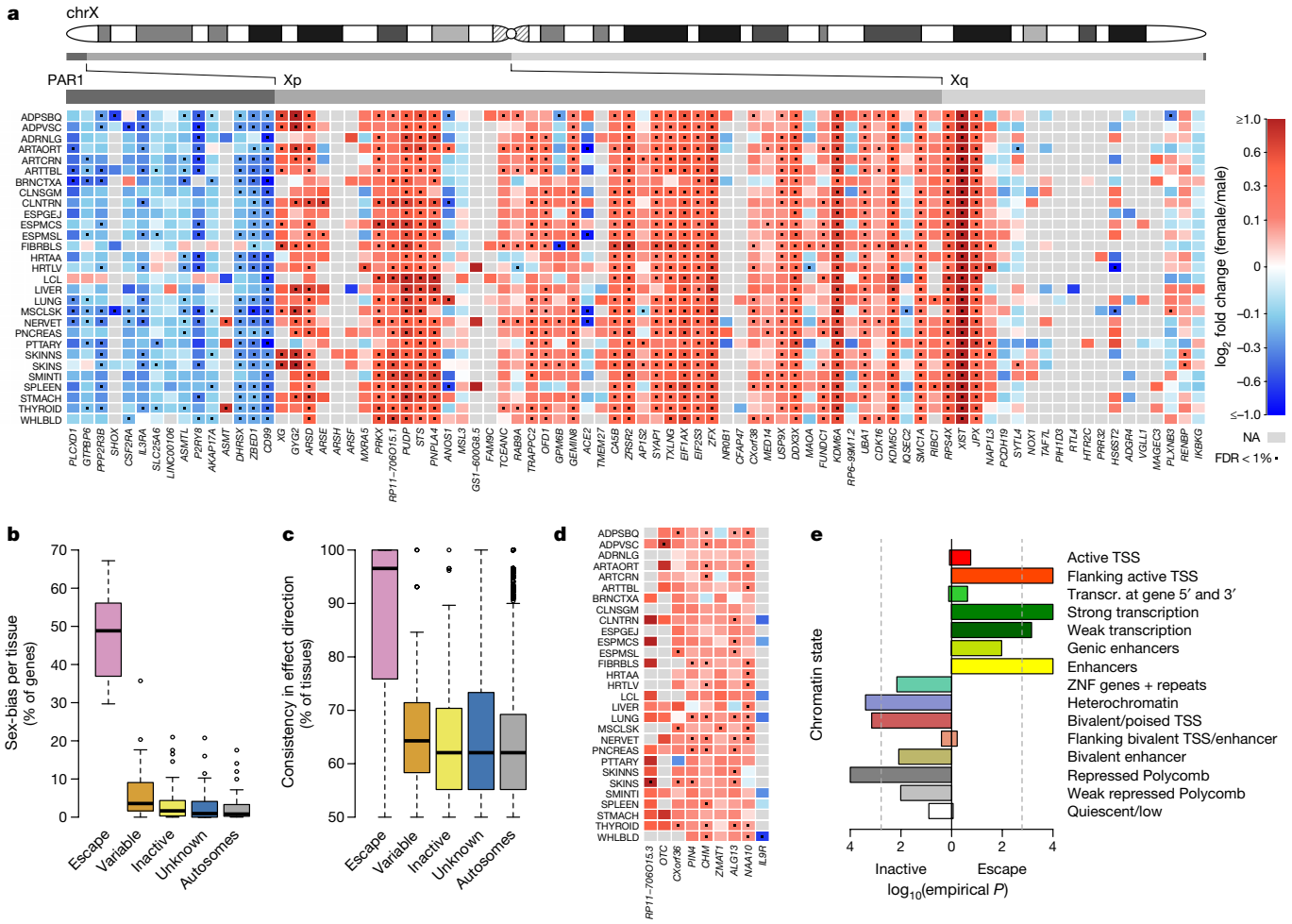


Figure 2 | Assessment of tissue-sharing and population-level impacts of incomplete XCI in GTEx data. a, Male–female expression differences in reported XCI-escaping genes ($n = 82$) across 29 GTEx tissues. Definitions for the abbreviations can be found in Extended Data Table 1. **b**, Proportion of significantly biased (FDR <1%) genes in each tissue by reported XCI status. **c**, Proportion of tissues where the bias direction is shared with the reported XCI status. Genes expressed in at least five tissues are included. **d**, Sex bias pattern of nine genes not classified as full escape genes that follow a similar profile to established escape genes. **e**, Chromatin state enrichment between escape and inactive genes in the Roadmap Epigenomics[31] female samples.

monoallelic expression[21,27–29], besides searching for X-chromosomal sites with biallelic expression (Extended Data Fig. 7), we leveraged genotype phase information to detect sites for which the expressed allele was discordant with the active X chromosome in that cell.

Only 129 (78%) out of the 165 assayed genes (41–98 per sample) were fully inactivated in these data whereas the rest showed incomplete XCI in one or more samples (Fig. 4a, b and Supplementary Tables 8, 9); this is mostly consistent with previous assignments of XCI status to these genes (Fig. 4a and Supplementary Table 10). For instance, single-cell data reveal consistent expression from both X-chromosomal alleles for eleven genes in PAR1, in line with their known escape from XCI (for example, *ZBED1*, Fig. 4c), and replicate the known expression of *XIST* exclusively from Xi (Fig. 4d).

We next assessed whether our approach could extend the spectrum of escape from XCI. For seven genes that have previously been reported as inactivated, the data from single cells pointed to incomplete XCI (Fig. 4e–k and Supplementary Table 11), including *FHL1*, which was also highlighted as a candidate escape gene in the GTEx ASE analysis (Fig. 4e), and *ATP6AP2*, which displays predominantly female-biased expression across GTEx tissues (Fig. 4h). Both of these genes demonstrate significant Xi expression in only a subset of the scRNA-seq samples, a pattern that is consistent with variable escape[1,2]. Between-individual variability exists not only in the presence but also in the degree of expression from Xi (for example, *MSL3*, Fig. 4l). Highlighting the capacity of scRNA-seq to provide information beyond bulk RNA-seq, we identify examples where Xi expression varies considerably between the two X-chromosomal haplotypes within an individual (for example, *ASMTL*; Supplementary Table 12), suggesting *cis*-acting variation as one of the determinants for the level of Xi expression[3]. As a further layer of heterogeneity in Xi expression, we find a unique pattern for *TIMP1*. For this gene, the level of Xi expression across cells is not significant, but exclusive to a subset of cells that express the gene biallelically (Extended Data Fig. 7), pointing to cell-to-cell variability in escape.

Using the ASE estimates from the scRNA-seq and GTEx analyses to infer the magnitude of the incompleteness of XCI, we find that expression from Xi at escape genes rarely reaches levels equal to expression from Xa, Xi expression remaining on average at 33% of Xa expression. However, there is a lot of variability along the chromosome (Extended Data Fig. 8a and Supplementary Discussion), as has previously been demonstrated in specific tissue types[1,2]. Balanced expression dosage between males and females in PAR1 requires full escape from XCI, however, Xi expression remains below Xa expression also in this region (mean Xi to Xa ratio is around 0.80), pointing to partial spreading of XCI beyond nonPAR. In further support that the consistent male bias in PAR1 expression (Fig. 2a) is due to the incompleteness of escape, we observe no systematic up- or downregulation of Y chromosome expression in PAR1 (Extended Data Fig. 8b and Supplementary Discussion).

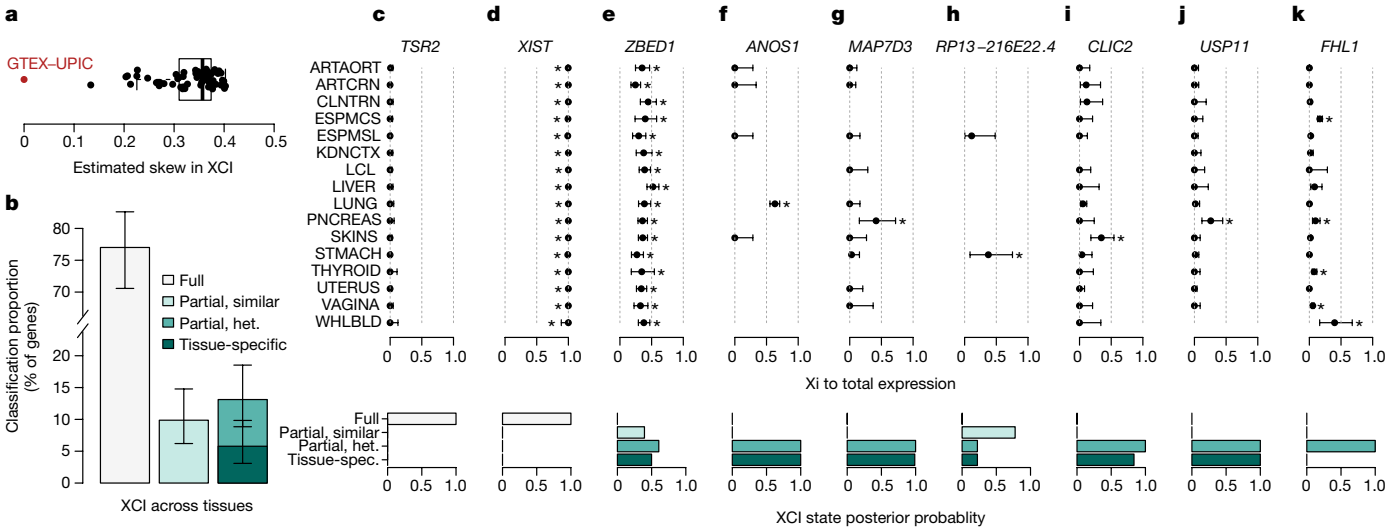


Figure 3 | Assessment of tissue-sharing of XCI in a GTEx donor with a highly skewed XCI. a, Distribution of the skewness of XCI in GTEx female samples ($n = 62$, v3 release). Each data point shows the mean skew in XCI across tissue samples per individual. **b**, Classification of X-chromosomal genes ($n = 186$) into full or incomplete and tissue-shared or heterogeneous XCI based on the analysis of ASE patterns across tissues. Error bars show the 95% credible interval. **c**–**e**, Examples of genes where the ASE-based assessment of XCI status match previously reported assignments (*TSR2*, inactive; *XIST*, escape; *ZBED1*, escape). Note that *XIST* is only expressed monoallelically from Xi, which is unusual for an escape gene. **f**, *KAL1* shows strong evidence for tissue-specific escape. **g**–**k**, Genes without previous or conclusive evidence for escape from XCI that are classified as incompletely inactivated in this sample. In **c**–**k** asterisks indicate that the Xi expression in the given tissue was significant at FDR <1% (one-sided binomial test) and error bars show the 95% confidence interval.

As another consequence of the partial Xi expression, several of the X–Y homologous genes in nonPAR[30] become male-biased when expression from the Y chromosome counterpart is accounted for (Extended Data Fig. 8c).

By combining diverse types and analyses of high-throughput RNA-seq data, we have systematically assessed the incompleteness and heterogeneity in XCI across 29 human tissues (Supplementary Table 13). We establish that scRNA-seq is suitable for surveys of human XCI and present the first steps towards understanding the cellular-level variability in the maintenance of XCI. Our phasing-based approach enables the full use of low-coverage scRNA-seq, however, because any single individual and cell type is only informative for restricted number of genes, larger datasets with more diverse cell types and conditions are required to fully profile XCI. We have therefore used the multi-tissue GTEx dataset to explore XCI in a larger number of X-chromosomal genes and to assess the tissue heterogeneity and impacts of XCI on gene expression differences between the sexes.

These analyses show that incomplete XCI is mostly shared between individuals and tissues, and extend previous surveys by pinpointing several examples of variability in the degree of XCI escape between cells, chromosomes, and tissues. In addition, our data demonstrate that escape from XCI results in sex-biased expression of at least 60 genes,

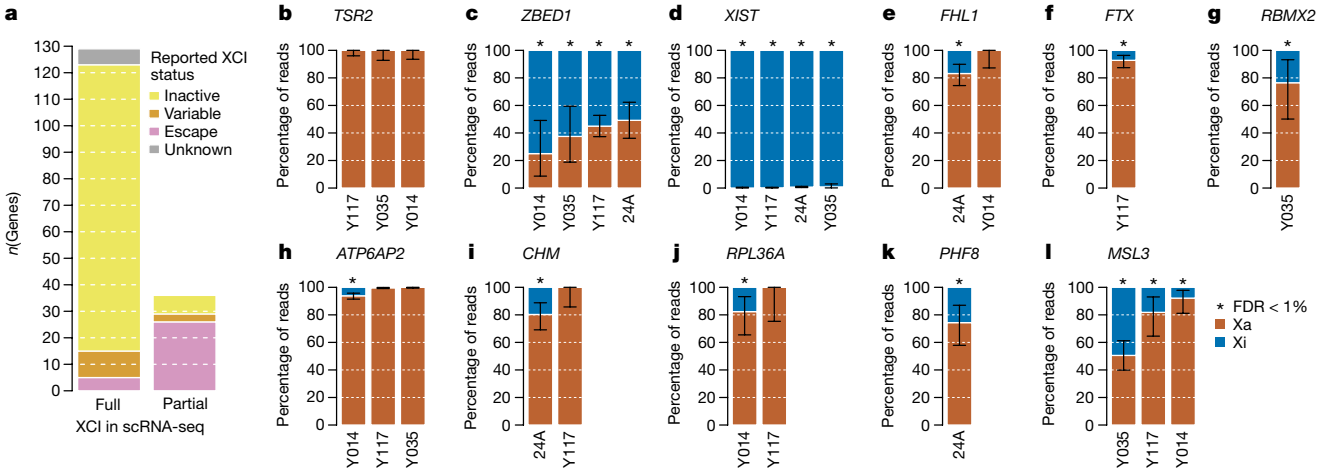


Figure 4 | Analysis of XCI using scRNA-seq. a, Proportion of genes demonstrating full and partial XCI in the ASE analysis in scRNA-seq data, and the concordance with previously reported XCI status. **b**–**l**, Examples of genes with different XCI patterns in scRNA-seq: previously reported inactive gene (**b**), known escape gene in PAR1 (**c**), escape gene with known exclusive expression from Xi (**d**), new candidates for escape genes that demonstrate incomplete XCI in only a subset of samples (**e**–**k**), and a known escape gene that shows escape of varying degrees in the three samples (Pearson's χ^2 test for equal proportions, $P = 3.80 \times 10^{-7}$) (**l**). b–l, *x* axis labels are sample identifiers. Asterisk above a bar indicates that the proportion of Xi expression, that is, blue bar, in a given sample is significantly greater than the expected baseline (FDR <1%, one-sided binomial test). Error bars show the 95% confidence interval.

potentially contributing to sex-specific differences in health and disease (Supplementary Discussion). As a whole, these results highlight the between-female and male–female diversity introduced by incomplete XCI, the biological implications of which remain to be fully explored.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 30 August 2016; accepted 8 September 2017.

1. Carrel, L. & Willard, H. F. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* **434,** 400–404 (2005).
2. Cotton, A. M. *et al.* Analysis of expressed SNPs identifies variable extents of expression from the human inactive X chromosome. *Genome Biol.* **14,** R122 (2013).
3. Cotton, A. M. *et al.* Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation. *Hum. Mol. Genet.* **24,** 1528–1539 (2015).
4. Schultz, M. D. *et al.* Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* **523,** 212–216 (2015).
5. Johnston, C. M. *et al.* Large-scale population study of human cell lines indicates that dosage compensation is virtually complete. *PLoS Genet.* **4,** e9 (2008).
6. Tukiainen, T. *et al.* Chromosome X-wide association study identifies loci for fasting insulin and height and evidence for incomplete dosage compensation. *PLoS Genet.* **10,** e1004127 (2014).
7. 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,** 367–378 (2014).
8. Amos-Landgraf, J. M. *et al.* X chromosome-inactivation patterns of 1,005 phenotypically unaffected females. *Am. J. Hum. Genet.* **79,** 493–499 (2006).
9. Lister, R. *et al.* Global epigenomic reconfiguration during mammalian brain development. *Science* **341,** 1237905 (2013).
10. Zhang, Y. *et al.* Transcriptional profiling of human liver identifies sex-biased genes associated with polygenic dyslipidemia and coronary artery disease. *PLoS ONE* **6,** e23506 (2011).
11. Berletch, J. B. *et al.* Escape from X inactivation varies in mouse tissues. *PLoS Genet.* **11,** e1005079 (2015).
12. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348,** 648–660 (2015).
13. The GTEx Consortium. Genetic effects on gene expression across tissues. https://doi.org/10.1038/nature24277 (2017).
14. Melé, M. *et al.* The human transcriptome across tissues and individuals. *Science* **348,** 660–665 (2015).
15. Balaton, B. P. & Brown, C. J. Escape artists of the X chromosome. *Trends Genet.* **32,** 348–359 (2016).
16. Ross, M. T. *et al.* The DNA sequence of the human X chromosome. *Nature* **434,** 325–337 (2005).
17. Pál, C., Papp, B. & Hurst, L. D. Highly expressed genes in yeast evolve slowly. *Genetics* **158,** 927–931 (2001).
18. Winter, E. E., Goodstadt, L. & Ponting, C. P. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome Res.* **14,** 54–61 (2004).
19. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* **16,** 133–145 (2015).
20. Chen, G. *et al.* Single-cell analyses of X chromosome inactivation dynamics and pluripotency during differentiation. *Genome Res.* **26,** 1342–1354 (2016).
21. 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).
22. Reinius, B. *et al.* Analysis of allelic expression patterns in clonal somatic cells by single-cell RNA-seq. *Nat. Genet.* **48,** 1430–1435 (2016).
23. Wang, M., Lin, F., Xing, K. & Liu, L. Random X-chromosome inactivation dynamics *in vivo* by single-cell RNA sequencing. *BMC Genomics* **18,** 90 (2017).
24. Wainer-Katsir, K. & Linial, M. Single cell expression data reveal human genes that escape X-chromosome inactivation. Preprint at http://www.biorxiv.org/content/early/2016/10/09/079830 (2016).
25. Petropoulos, S. *et al.* Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. *Cell* **165,** 1012–1026 (2016).
26. Villani, A. C. *et al.* Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356,** eaah4573 (2017).
27. Borel, C. *et al.* Biased allelic expression in human primary fibroblast single cells. *Am. J. Hum. Genet.* **96,** 70–80 (2015).
28. Kim, J. K., Kolodziejczyk, A. A., Ilicic, T., Teichmann, S. A. & Marioni, J. C. Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nat. Commun.* **6,** 8687 (2015).
29. Marinov, G. K. *et al.* From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Res.* **24,** 496–510 (2014).
30. Bellott, D. W. *et al.* Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature* **508,** 494–499 (2014).
31. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J. Maller, F. Zhao and M. Lek for technical assistance and P. J. Siponen for assistance with figure design. T.T. was supported by the Academy of Finland (285725), Finnish Cultural Foundation, Orion-Farmos Research Foundation and Emil Aaltonen Foundation. K.J.K. is supported by a NIGMS Fellowship (F32GM115208). This work was supported by NIH grants U54DK105566, R01MH101820 and R01GM104371 to D.G.M. The Genotype-Tissue Expression (GTEx) project was supported by the Common Fund of the Office of the Director of the National Institutes of Health. Additional funds were provided by the NCI, NHGRI, NHLBI, NIDA, NIMH and NINDS. Donors were enrolled at Biospecimen Source Sites funded by NCI\SAIC-Frederick, Inc. (SAIC-F) subcontracts to the National Disease Research Interchange (10XS170), Roswell Park Cancer Institute (10XS171) and Science Care, Inc. (X10S172). The Laboratory, Data Analysis, and Coordinating Center (LDACC) was funded through a contract (HHSN268201000029C) to The Broad Institute; this grant also provided funding to D.G.M. and T.T. Biorepository operations were funded through an SAIC-F subcontract to the Van Andel Institute (10ST1035). Additional data repository and project management were provided by SAIC-F (HHSN261200800001E). The Brain Bank was supported by supplements to University of Miami grants DA006227 and DA033684 and to contract N01MH000028. Statistical Methods development grants were made to the University of Geneva (MH090941 and MH101814), the University of Chicago (MH090951, MH090937, MH101820 and MH101825), the University of North Carolina, Chapel Hill (MH090936 and MH101819), Harvard University (MH090948), Stanford University (MH101782), Washington University St. Louis (MH101810) and the University of Pennsylvania (MH101822).

Author Contributions T.T. and D.G.M. designed the study. A.-C.V. designed and conducted the scRNA-seq experiments. T.T., A.Y., M.A.R., M.A., L.G., M.F. and B.B.C. analysed the data. J.L.M., R.S., S.E.C., A.K., K.J.K., F.A., A.B., T.L., A.R., K.G.A., N.H. and D.G.M. provided tools and reagents. T.T. and D.G.M. wrote the manuscript with input from other authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to T.T. (ttuk@broadinstitute.org) or D.G.M. (danmac@broadinstitute.org).

Reviewer Information *Nature* thanks A. Clark and the other anonymous reviewer(s) for their contribution to the peer review of this work.

GTEx Consortium

Laboratory, Data Analysis & Coordinating Center (LDACC)—Analysis Working Group François Aguet[1], Kristin G. Ardlie[1], Beryl B. Cummings[1,2], Ellen T. Gelfand[1], Gad Getz[1,3], Kane Hadley[1], Robert E. Handsaker[1,4], Katherine H. Huang[1], Seva Kashin[1,4], Konrad J. Karczewski[1,2], Monkol Lek[1,2], Xiao Li[1], Daniel G. MacArthur[1,2], Jared L. Nedzel[1], Duyen T. Nguyen[1], Michael S. Noble[1], Ayellet V. Segrè[1], Casandra A. Trowbridge[1], Taru Tukiainen[1,2]

Statistical Methods groups—Analysis Working Group Nathan S. Abell[5,6], Brunilda Balliu[6], Ruth Barshir[7], Omer Basha[7], Alexis Battle[8], Gireesh K. Bogu[9,10], Andrew Brown[11,12,13], Christopher D. Brown[14], Stephane E. Castel[15,16], Lin S. Chen[17], Colby Chiang[18], Donald F. Conrad[19,20], Nancy J. Cox[21], Farhan N. Damani[8], Joe R. Davis[5,6], Olivier Delaneau[11,12,13], Emmanouil T. Dermitzakis[11,12,13], Barbara E. Engelhardt[22], Eleazar Eskin[23,24], Pedro G. Ferreira[25,26], Laure Frésard[5,6], Eric R. Gamazon[21,27,28], Diego Garrido-Martín[9,10], Ariel D.H. Gewirtz[29], Genna Gliner[30], Michael J. Gloudemans[5,6,31], Roderic Guigo[9,10,32], Ira M. Hall[18,19,33], Buhm Han[34], Yuan He[35], Farhad Hormozdiari[23], Cedric Howald[11,12,13], Hae Kyung Im[36], Brian Jo[29], Eun Yong Kang[23], Yungil Kim[8], Sarah Kim-Hellmuth[15,16], Tuuli Lappalainen[15,16], Gen Li[37], Xin Li[6], Boxiang Liu[5,6,38], Serghei Mangul[23], Mark I. McCarthy[39,40,41], Ian C. McDowell[42], Pejman Mohammadi[15,16], Jean Monlong[9,10,43], Stephen B. Montgomery[5,6], Manuel Muñoz-Aguirre[9,10,44], Anne W. Ndungu[39], Dan L. Nicolae[36,45,46], Andrew B. Nobel[47,48], Meritxell Oliva[36,49], Halit Ongen[11,12,13], John J. Palowitch[47], Nikolaos Panousis[11,12,13], Panagiotis Papasaikas[9,10], YoSon Park[14], Princy Parsana[8], Anthony J. Payne[39], Christine B. Peterson[50], Jie Quan[51], Ferran Reverter[9,10,52], Chiara Sabatti[53,54], Ashis Saha[8], Michael Sammeth[55], Alexandra J. Scott[18], Andrey A. Shabalin[56], Reza Sodaei[9,10], Matthew Stephens[45,46], Barbara E. Stranger[36,49,57], Benjamin J. Strober[35], Jae Hoon Sul[58], Emily K. Tsang[6,31], Sarah Urbut[46], Martijn van de Bunt[39,40], Gao Wang[46], Xiaoquan Wen[59], Fred A. Wright[60], Hualin S. Xi[51], Esti Yeger-Lotem[7,61], Zachary Zappala[5,6], Judith B. Zaugg[62], Yi-Hui Zhou[60]

Enhancing GTEx (eGTEx) groups Joshua M. Akey[29,63], Daniel Bates[64], Joanne Chan[5], Lin S. Chen[17], Melina Claussnitzer[1,65,66], Kathryn Demanelis[17], Morgan Diegel[64], Jennifer A. Doherty[67], Andrew P. Feinberg[35,68,69,70], Marian S. Fernando[36,49], Jessica Halow[64], Kasper D. Hansen[68,71,72], Eric Haugen[64], Peter F. Hickey[72], Lei Hou[1,73], Farzana Jasmine[17], Ruiqi Jian[5], Lihua Jiang[5], Audra Johnson[64], Rajinder Kaul[64], Manolis Kellis[1,73], Muhammad G. Kibriya[17], Kristen Lee[64], Jin Billy Li[5], Qin Li[5], Xiao Li[5], Jessica Lin[5,74], Shin Lin[5,75], Sandra Linder[5,6], Caroline Linke[36,49], Yaping Liu[1,73], Matthew T. Maurano[76], Benoit Molinie[1], Stephen B. Montgomery[5,6], Jemma Nelson[64], Fidencio J. Neri[64], Meritxell Oliva[36,49], Yongjin Park[1,73], Brandon L. Pierce[17], Nicola J. Rinaldi[1,73], Lindsay F. Rizzardi[68], Richard Sandstrom[64], Andrew Skol[36,49,57], Kevin S. Smith[5,6], Michael P. Snyder[5], John Stamatoyannopoulos[64,74,77], Barbara E. Stranger[36,49,57], Hua Tang[5], Emily K. Tsang[6,31], Li Wang[1], Meng Wang[5], Nicholas Van Wittenberghe[1], Fan Wu[36,49], Rui Zhang[5]

NIH Common Fund Concepcion R. Nierras[78]

NIH/NCI Philip A. Branton[79], Latarsha J. Carithers[79,80], Ping Guan[79], Helen M. Moore[79], Abhi Rao[79], Jimmie B. Vaught[79]

NIH/NHGRI Sarah E. Gould[81], Nicole C. Lockart[81], Casey Martin[81], Jeffery P. Struewing[81], Simona Volpi[81]

NIH/NIMH Anjene M. Addington[82], Susan E. Koester[82]

NIH/NIDA A. Roger Little[83]

Biospecimen Collection Source Site—NDRI Lori E. Brigham[84], Richard Hasz[85], Marcus Hunter[86], Christopher Johns[87], Mark Johnson[88], Gene Kopen[89], William F. Leinweber[89], John T. Lonsdale[89], Alisa McDonald[89], Bernadette Mestichelli[89], Kevin Myer[86], Brian Roe[86], Michael Salvatore[89], Saboor Shad[89], Jeffrey A. Thomas[89], Gary Walters[88], Michael Washington[88], Joseph Wheeler[87]

Biospecimen Collection Source Site—RPCI Jason Bridge[90], Barbara A. Foster[91], Bryan M. Gillard[91], Ellen Karasik[91], Rachna Kumar[91], Mark Miklos[90], Michael T. Moser[91]

Biospecimen Core Resource—VARI Scott D. Jewell[92], Robert G. Montroy[92], Daniel C. Rohrer[92], Dana R. Valley[92]

Brain Bank Repository—University of Miami Brain Endowment Bank David A. Davis[93], Deborah C. Mash[93]

Leidos Biomedical—Project Management Anita H. Undale[94], Anna M. Smith[95], David E. Tabor[95], Nancy V. Roche[95], Jeffrey A. McLean[95], Negin Vatanian[95], Karna L. Robinson[95], Leslie Sobin[95], Mary E. Barcus[96], Kimberly M. Valentino[95], Liqun Qi[95], Steven Hunter[95], Pushpa Hariharan[95], Shilpi Singh[95], Ki Sung Um[95], Takunda Matose[95], Maria M. Tomaszewski[95]

ELSI Study Laura K. Barker[97], Maghboeba Mosavel[98], Laura A. Siminoff[97], Heather M. Traino[97]

Genome Browser Data Integration & Visualization—EBI Paul Flicek[99], Thomas Juettemann[99], Magali Ruffier[99], Dan Sheppard[99], Kieron Taylor[99], Stephen J. Trevanion[99], Daniel R. Zerbino[99]

Genome Browser Data Integration & Visualization—UCSC Genomics Institute, University of California Santa Cruz Brian Craft[100], Mary Goldman[100], Maximilian Haeussler[100], W. James Kent[100], Christopher M. Lee[100], Benedict Paten[100], Kate R. Rosenbloom[100], John Vivian[100], Jingchun Zhu[100]

[1]The Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts 02142, USA. [2]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [3]Massachusetts General Hospital Cancer Center and Department of Pathology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [4]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02114, USA. [5]Department of Genetics, Stanford University, Stanford, California 94305, USA. [6]Department of Pathology, Stanford University, Stanford, California 94305, USA. [7]Department of Clinical Biochemistry and Pharmacology, Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel. [8]Department of Computer Science, Johns Hopkins University, Baltimore, Maryland 21218, USA. [9]Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, 08003 Barcelona, Spain. [10]Universitat Pompeu Fabra (UPF), 08002 Barcelona, Spain. [11]Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland. [12]Institute for Genetics and Genomics in Geneva (iG3), University of Geneva, 1211 Geneva, Switzerland. [13]Swiss Institute of Bioinformatics, 1211 Geneva, Switzerland. [14]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [15]New York Genome Center, New York, New York 10013, USA. [16]Department of Systems Biology, Columbia University Medical Center, New York, New York 10032, USA. [17]Department of Public Health Sciences, The University of Chicago, Chicago, Illinois 60637, USA. [18]McDonnell Genome Institute, Washington University School of Medicine, St. Louis, Missouri 63108, USA. [19]Department of Genetics, Washington University School of Medicine, St. Louis, Missouri 63108, USA. [20]Department of Pathology & Immunology, Washington University School of Medicine, St. Louis, Missouri 63108, USA. [21]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee 37232, USA. [22]Department of Computer Science, Center for Statistics and Machine Learning, Princeton University, Princeton, New Jersey 08540, USA. [23]Department of Computer Science, University of California, Los Angeles, California 90095, USA. [24]Department of Human Genetics, University of California, Los Angeles, California 90095, USA. [25]Instituto de Investigação e Inovação em Saúde (i3S), Universidade do Porto, 4200-135 Porto, Portugal. [26]Institute of Molecular Pathology and Immunology (IPATIMUP), University of Porto, 4200-625 Porto, Portugal. [27]Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, 1105 AZ Amsterdam, The Netherlands. [28]Department of Psychiatry, Academic Medical Center, University of Amsterdam, 1105 AZ Amsterdam, The Netherlands. [29]Lewis Sigler Institute, Princeton University, Princeton, New Jersey 08540, USA. [30]Department of Operations Research and Financial Engineering, Princeton University, Princeton, New Jersey 08540, USA. [31]Biomedical Informatics Program, Stanford University, Stanford, California 94305, USA. [32]Institut Hospital del Mar d'Investigacions Mèdiques (IMIM), 08003 Barcelona, Spain. [33]Department of Medicine, Washington University School of Medicine, St. Louis, Missouri 63108, USA. [34]Department of Convergence Medicine, University of Ulsan College of Medicine, Asan Medical Center, Seoul 138-736, South Korea. [35]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland 21218, USA. [36]Section of Genetic Medicine, Department of Medicine, The University of Chicago, Chicago, Illinois 60637, USA. [37]Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, New York 10032, USA. [38]Department of Biology, Stanford University, Stanford, California 94305, USA. [39]Wellcome Trust Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford OX3 7BN, UK. [40]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Oxford OX3 7LE, UK. [41]Oxford NIHR Biomedical Research Centre, Churchill Hospital, Oxford OX3 7LJ, UK. [42]Computational Biology & Bioinformatics Graduate Program, Duke University, Durham, North Carolina 27708, USA. [43]Human Genetics Department, McGill University, Montreal, Quebec H3A 0G1, Canada. [44]Departament d'Estadística i Investigació Operativa, Universitat Politècnica de Catalunya, 08034 Barcelona, Spain. [45]Department of Statistics, The University of Chicago, Chicago, Illinois 60637, USA. [46]Department of Human Genetics, The University of Chicago, Chicago, Illinois 60637, USA. [47]Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, North Carolina 27599, USA. [48]Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina 27599, USA. [49]Institute for Genomics and Systems Biology, The University of Chicago, Chicago, Illinois 60637, USA. [50]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA. [51]Computational Sciences, Pfizer Inc, Cambridge, Massachusetts 02139, USA. [52]Universitat de Barcelona, 08028 Barcelona, Spain. [53]Department of Biomedical Data Science, Stanford University, Stanford, California 94305, USA. [54]Department of Statistics, Stanford University, Stanford, California 94305, USA. [55]Institute of Biophysics Carlos Chagas Filho (IBCCF), Federal University of Rio de Janeiro (UFRJ), 21941902 Rio de Janeiro, Brazil. [56]Department of Psychiatry, University of Utah, Salt Lake City, Utah 84108, USA. [57]Center for Data Intensive Science, The University of Chicago, Chicago, Illinois 60637, USA. [58]Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, California 90095, USA. [59]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, USA. [60]Bioinformatics Research Center and Departments of Statistics and Biological Sciences, North Carolina State University, Raleigh, North Carolina 27695, USA. [61]National Institute for Biotechnology in the Negev, Beer-Sheva, 84105, Israel. [62]European Molecular Biology Laboratory, 69117 Heidelberg, Germany. [63]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08540, USA. [64]Altius Institute for Biomedical Sciences, Seattle, Washington 98121, USA. [65]Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts 02215, USA. [66]University of Hohenheim, 70599 Stuttgart, Germany. [67]Huntsman Cancer Institute, Department of Population Health Sciences, University of Utah, Salt Lake City, Utah 84112, USA. [68]Center for Epigenetics, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA. [69]Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA. [70]Department of Mental Health, Johns Hopkins University School of Public Health, Baltimore, Maryland 21205, USA. [71]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland 21205, USA. [72]Department of Biostatistics, Johns Hopkins

University, Baltimore, Maryland 21205, USA. [73]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [74]Department of Medicine, University of Washington, Seattle, Washington 98195, USA. [75]Division of Cardiology, University of Washington, Seattle, Washington 98195, USA. [76]Institute for Systems Genetics, New York University Langone Medical Center, New York, New York 10016, USA. [77]Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA. [78]Office of Strategic Coordination, Division of Program Coordination, Planning and Strategic Initiatives, Office of the Director, NIH, Rockville, Maryland 20852, USA. [79]Biorepositories and Biospecimen Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, Maryland 20892, USA. [80]National Institute of Dental and Craniofacial Research, Bethesda, Maryland 20892, USA. [81]Division of Genomic Medicine, National Human Genome Research Institute, Rockville, Maryland 20852, USA. [82]Division of Neuroscience and Basic Behavioral Science, National Institute of Mental Health, NIH, Bethesda, Maryland 20892, USA. [83]Division of Neuroscience and Behavior, National Institute on Drug Abuse, NIH, Bethesda, Maryland 20892, USA. [84]Washington Regional Transplant Community, Falls Church, Virginia 22003, USA. [85]Gift of Life Donor Program, Philadelphia, Pennsylvania 19103, USA. [86]LifeGift, Houston, Texas 77055, USA. [87]Center for Organ Recovery and Education, Pittsburgh, Pennsylvania 15238, USA. [88]LifeNet Health, Virginia Beach, Virginia 23453, USA. [89]National Disease Research Interchange, Philadelphia, Pennsylvania 19103, USA. [90]Unyts, Buffalo, New York 14203, USA. [91]Pharmacology and Therapeutics, Roswell Park Cancer Institute, Buffalo, New York 14263, USA. [92]Van Andel Research Institute, Grand Rapids, Michigan 49503, USA. [93]Brain Endowment Bank, Miller School of Medicine, University of Miami, Miami, Florida 33136, USA. [94]National Institute of Allergy and Infectious Diseases, NIH, Rockville, Maryland 20852, USA. [95]Biospecimen Research Group, Clinical Research Directorate, Leidos Biomedical Research, Inc., Rockville, Maryland 20852, USA. [96]Leidos Biomedical Research, Inc., Frederick, Maryland 21701, USA. [97]Temple University, Philadelphia, Pennsylvania 19122, USA. [98]Department of Health Behavior and Policy, School of Medicine, Virginia Commonwealth University, Richmond, Virginia 23298, USA. [99]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton CB10 1SD, UK. [100]UCSC Genomics Institute, University of California Santa Cruz, Santa Cruz, California 95064, USA.

METHODS

GTEx data. The GTEx project[12] collected tissue samples from 554 postmortem donors (187 females, 357 males; age range, 20–70), carried out RNA-seq on 8,555 tissue samples and generated genotyping data for up to 449 donors (GTEx analysis v6p release). More detailed methods can be found in ref. 13. All GTEx data, including RNA, genome and exome sequencing data, used in the analyses described here are available through dbGaP under accession number phs000424.v6.p1, unless otherwise stated. Summary data and details on data production and processing are also available from the GTEx Portal (http://gtexportal.org).

Single-cell samples. For the human dendritic cells samples profiled, the healthy donor (ID: 24A) was recruited from the Boston-based PhenoGenetic project, a resource of healthy subjects that are re-contactable by genotype[32]. The donor was a female Asian individual from China, 25 years of age at the time of blood collection. She was a non-smoker, had a normal BMI (height: 168.7 cm; weight: 56.45 kg; BMI: 19.8), and normal blood pressure (108/74). The donor had no family history of cancer, allergies, inflammatory disease, autoimmune disease, chronic metabolic disorders or infectious disorders. She provided written informed consent for the genetic research studies and molecular testing, as previously reported[26].

Daughters of three parent–child Yoruba trios from Ibadan, Nigeria (that is, YRI trios), collected as part of the International HapMap Project, were chosen for single-cell profiling, both to maximize heterozygosity and due to availability of parental genotypes enabling phasing. DNA and LCLs were ordered from the NHGRI Sample Repository for Human Genetic Research (Coriell Institute for Medical Research): LCLs from B lymphocytes for the three daughters (catalogue numbers: GM19240, GM19199 and GM18518) and DNA extracted from LCLs for all members of the three trios (catalogue numbers for DNA: NA19240, NA19238, NA19239, NA19199, NA19197, NA19198, NA18518, NA18519 and NA18520). These YRI samples are referred to by their family IDs: Y014, Y035 and Y117.

Clinical muscle samples. To assess whether PAR1 genes are equally expressed from X and Y chromosomes, a combination of skeletal muscle RNA-seq data and trio genotyping data from eight male patients with muscular dystrophy, sequenced as part of an unrelated study, was used. Patient cases with available muscle biopsies were referred from clinicians starting April 2013 until June 2016. All patients included for RNA-seq had previously available trio whole-exome sequencing (WES) data, with one sample having additional trio whole-genome sequencing (WGS). Muscle biopsies were shipped frozen from clinical centres by liquid nitrogen dry shipping and, where possible, frozen muscle was sectioned on a cryostat and stained with haematoxylin and eosin to assess muscle quality as well as the presence of overt freeze–thaw artefacts.

Genotyping. The GTEx v6p release includes WGS data for 148 donors, including GTEX-UPIC. WGS libraries were sequenced on the Illumina HiSeqX or Illumina HiSeq2000. WGS data was processed through a Picard-based pipeline, using base quality score recalibration and local realignment at known indels. BWA-MEM aligner was used for mapping reads to the human genome build 37 (hg19). Single-nucleotide polymorphisms (SNPs) and indels (insertions and deletions) were jointly called across all 148 samples and additional reference genomes using HaplotypeCaller v.3.1 of GATK. Default filters were applied to SNP and indel calls using the variant quality score recalibration (VQSR) approach of GATK. An additional hard filter InbreedingCoeff ≤ -0.3 was applied to remove sites that VQSR failed to filter.

WGS for one of the clinical muscle samples was performed on 500 ng to 1.5 μg of genomic DNA using a PCR-Free protocol that substantially increases the uniformity of genome coverage. These libraries were sequenced on the Illumina HiSeq X10 with 151-bp paired-end reads and a target mean coverage of $>30\times$, and were processed similarly to the above description.

The Y117 trio (sample IDs NA19240 (daughter), NA19238 (mother), and NA19239 (father)) was whole-genome-sequenced as part of the 1000 Genomes Project as described previously[33]. The VCF file containing the WGS-based genotypes for SNPs (YRI.trio.2010_09.genotypes.vcf.gz) was downloaded from the FTP site of the project. The genotype coordinates (in human genome build 36) in the original VCF were converted to hg19 using the liftover script (liftOverVCF.pl) and chain files provided as part of the GATK package.

WES was performed using Illumina's capture Exome (ICE) technology (Y035, Y014, 24A) or Agilent SureSelect Human All Exon Kit v.2 exome capture (clinical muscle samples) with a mean target coverage of $>80\times$. WES data was aligned with BWA, processed with Picard, and SNPs and indels were jointly called with other samples using GATK HaplotypeCaller package v.3.1 (24A, clinical muscle samples) or v.3.4 (Y035, Y014). Default filters were applied to SNP and indel calls using the VQSR approach. A modified version of the Ensembl variant effect predictor was used for variant annotation for all WES and WGS data. For trio WES or WGS data the genotypes of the proband were phased using the PhaseByTransmission tool of the GATK toolkit.

Single-cell data preparation and sequencing. For profiling of healthy dendritic cells (DCs), peripheral blood mononuclear cells (PBMCs) were first isolated from fresh blood within 2 h of collection, using Ficoll–Paque density gradient centrifugation as previously described[34]. Single-cell suspensions were stained as per the manufacturer's recommendations with an antibody panel designed to enrich for all known blood DC population for single-cell sorting and scRNA-seq profiling[26]. A total of 24 single cells from four loosely gated populations were sorted per 96-well plate, with each well containing 10 μl of lysis buffer. A total of eight plates were analysed by scRNA-seq.

All LCL cell lines were cultured according to Coriell's recommendations (medium: RPMI 1640, 2 mM L-glutamine, 15% fetal bovine serum (all three from ThermoFisher Scientific)) in T25 tissue culture flask with 10–20 ml medium at 37 °C in 5% carbon dioxide. Cells were split upon reaching a cell density of approximately 300,000–400,000 viable cells per ml. All three lymphoblast cultures were split once before single-cell sorting. Cells were washed with $1\times$ PBS, the pellet was resuspended and stained with DAPI (Biolegend) for viability according to the manufacturer's recommendations.

All single live cells (for both DCs and LCL cell lines) were sorted into a 96-well full-skirted Eppendorf plate chilled to 4 °C, that were pre-prepared with 10 μl TCL buffer (Qiagen) supplemented with 1% β-mercaptoethanol (lysis buffer), using a BD FACS Fusion instrument. Single-cell lysates were sealed, vortexed, spun down at 300*g* at 4 °C for 1 min, immediately placed on dry ice and transferred for storage at −80 °C.

The Smart-Seq2 protocol was performed on single-sorted cells as described[35,36], with some modifications as described in ref. 26 (Supplementary Methods). A total of 768 single DCs isolated from a healthy Asian female individual, along with 96 single cells from GM19240, 48 single cells from GM19199 and 48 single cells from GM18518 were profiled. In brief, single-cell lysates were thawed on ice, purified and reverse-transcribed using Maxima H Minus Reverse Transcriptase. PCR was performed with KAPA HiFi HotStart ReadyMix (KAPA Biosystems) and purified with Agencourt AMPureXP SPRI beads (Beckman-Coulter). The concentration of amplified cDNA was measured on the Synergy H1 Hybrid Microplate Reader (BioTek) using High-Sensitivity Qubit reagent (Life Technologies) and the size distribution of select wells was checked on a High-Sensitivity Bioanalyzer Chip (Agilent). The expected concentration was around $0.5-2\,\mathrm{ng\,\mu l^{-1}}$ with a size distribution that sharply peaked around 2 kb.

Library preparation was carried out using the Nextera XT DNA Sample Kit (Illumina) with custom indexing adapters, allowing up to 384 libraries to be simultaneously generated in a 384-well PCR plate (note that DCs were processed in a 384-well plate whereas LCLs were processed in 96-well plate format). The concentration of the final pooled libraries was measured using the High-Sensitivity DNA Qubit (Life Technologies) and the size distribution was measured on a High-Sensitivity Bioanalyzer Chip (Agilent). The expected concentration of the pooled libraries was $10–30\,\mathrm{ng\,\mu l^{-1}}$ with a size distribution of 300–700 bp. For the DCs, we created pools of 384 cells, whereas 96 LCL samples were pooled at the time. We sequenced one library pool per lane as paired-end 25-bp reads on a HiSeq2500 (Illumina). Barcodes used for indexing are listed in the Supplementary Methods.

RNA-seq in GTEx. RNA sequencing was performed using a non-strand-specific RNA-seq protocol with polyA selection of RNA using the Illumina TruSeq protocol with sequence coverage goal of 50 million 76-bp paired-end reads as has been previously described in detail[12]. The RNA-seq data, except for GTEX-UPIC, was aligned with TopHat v.1.4.1 to the UCSC human genome release version hg19 using the Gencode v.19 annotations as the transcriptome reference. Gene level read counts and reads per kilobase per million reads (RPKMs) were derived using the RNA-SeQC tool[37] using the Gencode v.19 transcriptome annotation. The transcript model was collapsed into a gene model as described previously[12]. Read count and RPKM quantification include only uniquely mapped and properly paired reads contained within exon boundaries.

RNA-seq alignment to personalized genomes. For the four single-cell samples and for GTEX-UPIC RNA-seq, data were processed using a modification of the AlleleSeq pipeline[38,39] to minimize reference allele bias in alignment. A diploid personal reference genome for each of the samples was generated with the vcf2diploid tool[38] including all heterozygous biallelic single-nucleotide variants identified in WES or WGS either together with (YRI samples) or without (GTEX-UPIC, 24A) maternal and paternal genotype information. The RNA-seq reads were then aligned to both parental references using STAR[40] v.2.4.1a in a per-sample two-pass mode (GTEX-UPIC and YRI samples) or v.2.3.0e (24A) using hg19 as the reference. The alignments were combined by comparing the quality of alignment between the two references: for reads aligning uniquely to both references the alignment with the higher alignment score was chosen and reads aligning uniquely to only one reference were kept as such.

RNA-seq of clinical muscle samples. Patient RNA samples derived from primary muscle were sequenced using the GTEx sequencing protocol[12] with sequence coverage of 50 million or 100 million 76-bp paired-end reads. RNA-seq reads were aligned using STAR[40] 2-pass version v.2.4.2a using hg19 as the reference genome. Junctions were filtered after first pass alignment to exclude junctions with less than 5 uniquely mapped reads supporting the event and junctions found on the mitochondrial genome. The value for unique mapping quality was assigned to 60 and duplicate reads were marked with Picard MarkDuplicates (v.1.1099).

Catalogue of X-inactivation status. To compare results from the ASE and GTEx analyses with previous observations on genic XCI status we collated findings from two earlier studies[1,2] that represent systematic expression-based surveys into XCI. Each study catalogues hundreds of X-linked genes and together the data span two tissue types.

Carrel and Willard[1] surveyed in total 624 X-chromosomal transcripts expressed in primary fibroblasts in nine cell hybrids each containing a different human Xi. In order to find the gene corresponding to each transcript, the primer sequences designed to test the expression of the transcripts in the original study were aligned to reference databases based on the Gencode v.19 transcriptome and hg19 using in-house software (unpublished) (Supplementary Methods). In total 553 transcripts primer pairs were successfully matched to X-chromosomal Gencode v.19 reference mapping together with 470 unique X-chromosomal genes (Supplementary Methods). These 470 genes were split into three XCI status categories (escape, variable, inactive) based on the level of Xi expression (that is, the number of cell lines expressing the gene from Xi) resulting in 75 escape, 51 variable escape and 344 inactive genes.

Cotton *et al.*[2] surveyed XCI using allelic imbalance in clonal or near-clonal female LCL and fibroblast cell lines and provided XCI statuses for 508 genes (68 escape, 146 variable escape, 294 subject genes). The data were mapped to Gencode v.19 using the reported gene names and their known aliases (Supplementary Methods), resulting in a list of XCI statuses for 506 X-chromosomal genes.

The results were combined by retaining the XCI status in the original study where possible (that is, same status in both studies or gene unique to one study) and for genes where the reported XCI statuses were in conflict the following rules were applied: (1) a gene was considered 'escape' if it was called escape in one study and variable in the other; (2) 'variable escape' if classified as escape and inactive; and (3) 'inactive' if classified as inactive in one study and variable escape in the other. The final combined list of XCI statuses consisted of 631 X-chromosomal genes including 99 escape, 101 variable escape and 431 inactive genes.

Analysis of sex-biased expression. Differential expression analyses were conducted to identify genes that are expressed at significantly different levels between male and female samples using 29 GTEx v6p tissues with RNA-seq and genotype data available from more than 70 individuals after excluding samples flagged in QC and sex-specific, outlier (that is, breast tissue) and highly correlated tissues[14]. Only autosomal and X-chromosomal protein-coding or long non-coding RNA genes in Gencode v.19 were included, and all lowly expressed genes were removed (Extended Data Table 1 and Supplementary Methods).

Differential expression analysis between male and female samples was conducted following the voom-limma pipeline[41–43] available as an R package through Bioconductor (https://bioconductor.org/packages/release/bioc/html/limma.html) using the gene-level read counts as input. The analyses were adjusted for age, three principal components inferred from genotype data using EIGENSTRAT[44], sample ischaemic time, surrogate variables[45,46] built using the sva R package[47], and the cause of death classified into five categories based on the four-point Hardy scale (Supplementary Methods).

To control the FDR, the qvalue R package was used to obtain q values applying the adjustment separately for the differential expression results from each tissue. The null hypothesis was rejected for tests with q values below 0.01.

XY homologue analysis. A list of Y-chromosomal genes with functional counterparts in the X chromosome, that is, X–Y gene pairs, was obtained from ref. 30, which lists 19 ancestral Y chromosome genes that have been retained in the human Y chromosome. After excluding two of the genes (*MXRA5Y* and *OFD1Y*), which were annotated as pseudogenes in ref 30, and a further four genes (*SRY*, *RBMY*, *TSPY* and *HSFY*) that according to ref. 30 have clearly diverged in function from their X-chromosomal homologues, the remaining 13 Y-chromosomal genes were matched with their X-chromosome counterparts using gene-pair annotations given in ref. 30 or by searching for known paralogues of the Y-chromosomal genes. To test for completeness of dosage compensation of the X–Y homologous genes, the sex-bias analysis in GTEx data was repeated replacing the expression of the X-chromosomal counterpart with the combined expression of the X and Y homologues.

Chromatin state analysis. To study the relationship between chromatin states and XCI, we used chromatin state calls from the Roadmap Epigenomics Consortium[31]. Specifically, we used the chromatin state annotations from the core 15-state model, publicly available at http://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html#core_15state. We followed our previously published method[48] to calculate the covariate-corrected percentage of each gene body assigned to each chromatin state. After pre-processing, we filtered down to the 399 inactive and 86 escape genes on the X chromosome and down to 38 female epigenomes.

To compare the chromatin state profiles of the escape and inactive genes in female samples, we used the one-sided Wilcoxon rank-sum test. Specifically, for each chromatin state, we averaged the chromatin state coverage across the 38 female samples for each gene, and compared that average chromatin state coverage for all 86 escape genes to the average chromatin state coverage for all 399 inactive genes. We performed both one-sided tests, to test for enrichment in escape genes, as well as for enrichment in inactive genes.

Next, we performed simulations to account for possible chromatin state biases, such as the fact that the escape and inactive genes are all from the X chromosome. Specifically, we generated 10,000 randomized simulations where we randomly shuffled the escape or inactive labels on the combined set of 485 genes, while retaining the sizes of each gene set. For each of these simulated escape and inactive gene sets, we calculated both one-sided Wilcoxon rank-sum test P values as described above, and then, we calculated a permutation P value for the real gene sets based on these 10,000 random simulations (Supplementary Methods). Finally, we used Bonferroni multiple hypothesis corrections for our significance thresholds to correct for our 30 tests, one for each of 15 chromatin states, and both possible test directions.

Allele-specific expression. For ASE analysis the allele counts for biallelic heterozygous variants were retrieved from RNA-seq data using GATK ASEReadCounter (v.3.6)[39]. Heterozygous variants that passed VQSR filtering were first extracted for each sample from WES or WGS VCFs using GATK SelectVariants. The analysis was restricted to biallelic SNPs owing to known issues in mapping bias in RNA-seq against indels[39]. Sample-specific VCFs and RNA-seq BAMs were inputted to ASEReadCounter requiring minimum base quality of 13 in the RNA-seq data (scRNA-seq samples, GTEX-UPIC) or requiring coverage in the RNA-seq data of each variant to be at least 10 reads, with a minimum base quality of 10 and counting only reads with unique mapping quality (MQ = 60) (clinical muscle samples).

For downstream processing of the scRNA-seq and GTEX-UPIC ASE data, we applied further filters to the data to focus on exonic variation only and to conservatively remove potentially spurious sites (Supplementary Methods), for example, sites with non-unique mappability were removed, and furthermore, after an initial analysis of the ASE data, we subjected 22 of the X-chromosomal ASE sites to manual investigation. For GTEX-UPIC the X-chromosomal ASE data was limited to only one site per gene in case of multiple ASE sites, by selecting the site with coverage >7 reads in the largest number of tissues, to have equal representation of each gene for downstream analyses.

Assessing ASE across tissues. For the GTEX-UPIC individual, for whom ASE data from up to 16 tissues per each ASE site was available, we applied the two-sided hierarchical grouped tissue model (GTM*) implemented in MAMBA v.1.0.0 (refs 49, 50) to ASE data. The Gibbs sampler was run for 200 iterations with a burn-in of 50 iterations.

GTM* is a Bayesian hierarchical model that borrows information across tissues and across variants, and provides parameter estimates that are useful for interpreting global properties of variants. It classifies the sites into ASE states according to their tissue-wide ASE profiles and provides an estimate of the proportion of variants in each of the five different ASE states (strong ASE across all tissues (SNGASE), moderate ASE across all tissues (MODASE), no ASE across all tissues (NOASE) and heterogeneous ASE across tissues (HET1 and HET0)).

To summarize the GTM* output in the context of XCI, SNGASE was considered to reflect full XCI, MODASE and NOASE were taken together to represent partial XCI with similar effects across tissues, and HET1 and HET0 were considered to reflect partial yet heterogeneous patterns of XCI across tissues. To combine estimates from two ASE states, we summed the estimated proportions in each class and subsequently calculated the 95% confidence intervals for each remaining ASE state using Jeffreys' prior.

Determining XCI status in GTEX-UPIC. In addition to the ASE states provided by the above MAMBA analysis, genic XCI status was assessed by comparing the allelic ratios at each X-chromosomal ASE site in each tissue individually. For each ASE site, the alleles were first mapped to Xa and Xi; the allele with lower combined relative expression across tissues was assumed to be the Xi allele. As an exception, at *XIST* the higher expressing allele was assumed to be the Xi allele. The significance of Xi expression at each ASE observation was tested using a one-sided binomial test, where the hypothesized probability of success was set at 0.025, that is, the

fraction of Xi expression from total expression was expected to be significantly greater than 0.025. To account for multiple testing, a FDR correction was applied, using the qvalue R package, to the *P* values from the binomial test for each of the 16 tissues separately. Observations with *q* values <0.01 were considered significant, that is, indicative of incomplete XCI at the given ASE site and tissue.

Biallelic expression in single cells. Biallelic expression in individual cells in the X chromosome was assessed only at ASE sites covered by the minimum of eight reads. A site was considered biallelically expressed when (1) allelic expression >0.05 and (2) the one-sided binomial test indicated allelic expression to be at least nominally significantly greater than 0.025. Only genes with at least two observations of biallelic expression across all cells within a sample were counted as biallelic.

Phasing scRNA-seq data. We assigned each cell to either of two cell populations distinguished by the parental X-chromosome designated for inactivation using genotype phasing. For the YRI samples, where parental genotype data was available, the assignment to the two parental cell populations was unambiguous for all cells where X-chromosomal sites outside PAR1 or frequently biallelic sites were expressed. For 24A, no parental genotype data were available, and we therefore used the correlation structure of the expressed X-chromosomal alleles across the 948 cells to infer the two parental haplotypes using the fact that in individual cells the expressed alleles at the chrX sites subject to full inactivation (that is, the majority chrX ASE sites), are from the X chromosome active in each cell (Supplementary Methods). In other words, while monoallelic expression in scRNA-seq in the autosomes is largely stochastic in origin, in the X chromosome the pattern of monoallelic expression is consistent across cells with the same parental X chromosome active[22], unless a gene is expressed also from the inactive X. As such, for the phase inference calculations, we excluded all PAR1 sites and all additional sites that were frequently biallelic, to minimize the contribution of escape genes to the phase estimation. After assigning each informative cell to either of the parental cell populations, the reference and alternate allele reads for each ASE site were mapped to active and inactive allele reads within each sample using the actual or inferred parental haplotypes. The data were first combined per variant by taking the sum of active and inactive counts separately across cells, and further similarly combined per gene, if multiple SNPs per gene were available. For 24A the allele expressed at *XIST* was assumed the Xi allele, in line with the exclusive Xi expression in the Yoruba samples confirmed using the information on parental haplotypes.

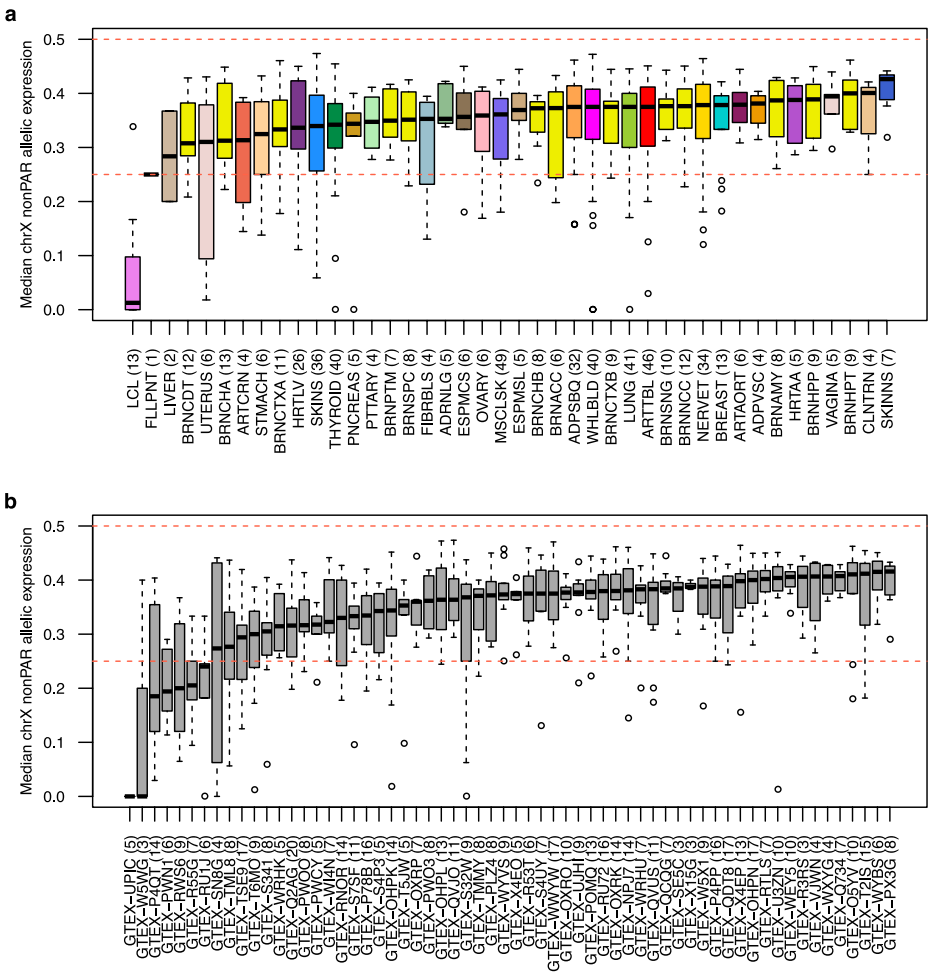
Determining XCI status from scRNA-seq ASE. Before calling XCI status using the Xa and Xi read counts from the phased data aggregated across cells, we excluded all sites without fewer than five cells contributing ASE data at each gene and also all sites with coverage lower than eight reads across cells within each sample. To determine whether the observed Xi expression is significantly different from zero, and therefore indicative of incomplete XCI at the site or gene, we required the Xi to total expression ratio to be significantly (q value <0.01) greater than the hypothesized upper bound for error, 0.025. This threshold was determined using the proportion of miscalled alleles at *XIST* ASE sites (by definition, *XIST* should express only alleles from the inactive chrX) in the two YRI samples, which presented with fully skewed XCI, that is, the same active X chromosome across all assessed cells. The median proportion of miscalled *XIST* alleles was 0, yet one site in one of the samples showed up to 2.5% of other allele calls, and therefore this was chosen as the error margin. FDR correction, conducted using the qvalue R package, was applied to each sample individually. Genes where at least one of the samples showed significant Xi expression were considered partially inactivated, while the remaining were classified as subject to full XCI. Allelic dropout, which is extensive in scRNA-seq[19,28], can lead to biases in allelic ratios in individual cells, that is, in our case resulting in false negatives where true escape genes are classified as inactivated, the used approach is based on using aggregate data across several cells and therefore the XCI status estimates are robust to such errors.

ChrX and chrY expression in PAR1. Using the parental origin of each allele reference and alternate allele read counts at PAR1 ASE sites were assigned to X and Y chromosomes (that is, maternally and paternally inherited alleles, respectively). For each sample, the PAR1 ASE data was summarized by gene by taking the sum of X and Y chromosome reads across all informative ASE sites within each gene. Significance of deviation from equal expression was assessed using a two-sided binomial test.

Manual curation of heterozygous variants from ASE analyses. Twenty-two heterozygous variants assessed in chrX ASE analysis were subjected to manual curation because of results in the XCI analysis that were in conflict with previous assignment of the underlying gene to be subject to full XCI. For each sample, BWA-aligned germline BAM files were viewed in IGV using either WGS or WES data. The presence of a number of characteristics called into question the confidence of the variant read alignments and thus the variant itself (Supplementary Methods). Allele balance that deviated significantly from 50:50 was considered suspect and often coincided with the existence of homology between the reference sequence in the region surrounding the variant and another area of the genome, as ascertained using the UCSC browser self-chain track and/or BLAT alignment of variant reads from within IGV. Other sequence-based annotations added to the VCF by HaplotypeCaller were also evaluated in the interests of examining other signatures of ambiguous mapping. The phasing of nearby variants was also considered. If phased variants occurred in the DNA sequencing data that were not assessed in the ASE analysis, those variants were considered suspect.

Data availability. Gene expression and genotype data from the GTEx v6p release are available in dbGaP (study accession phs000424.v6.p1; http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v6.p1). Raw RNA-seq data for 24A is available through dbGaP accession number phs001294.v1.p1 (https://www.ncbi.nlm.nih.gov/bioproject/?term=phs001294.v1.p1). The authors declare that all data supporting the findings of this study are available within the paper and its Supplementary Information. Source Data for Figs 2–4 are provided with the paper.

32. Xia, Z. *et al.* A 17q12 allele is associated with altered NK cell subsets and function. *J. Immunol.* **188,** 3315–3322 (2012).
33. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
34. Lee, M. N. *et al.* Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* **343,** 1246980 (2014).
35. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10,** 1096–1098 (2013).
36. Trombetta, J. J. *et al.* Preparation of single-cell RNA-seq libraries for next generation sequencing. *Curr. Protoc. Mol. Biol.* **107,** 4.22.1–4.22.17 (2014).
37. DeLuca, D. S. *et al.* RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28,** 1530–1532 (2012).
38. Rozowsky, J. *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.* **7,** 522 (2011).
39. Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E. & Lappalainen, T. Tools and best practices for data processing in allelic expression analysis. *Genome Biol.* **16,** 195 (2015).
40. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
41. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15,** R29 (2014).
42. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43,** e47 (2015).
43. Smyth, G. K. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3,** 1–25 (2004).
44. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38,** 904–909 (2006).
45. Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3,** 1724–1735 (2007).
46. Leek, J. T. & Storey, J. D. A general framework for multiple testing dependence. *Proc. Natl Acad. Sci. USA* **105,** 18718–18723 (2008).
47. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28,** 882–883 (2012).
48. Yen, A. & Kellis, M. Systematic chromatin state comparison of epigenomes associated with diverse properties including sex and tissue type. *Nat. Commun.* **6,** 7973 (2015).
49. Pirinen, M. *et al.* Assessing allele-specific expression across multiple tissues from RNA-seq read data. *Bioinformatics* **31,** 2497–2504 (2015).
50. Rivas, M. A. *et al.* Effect of predicted protein-truncating genetic variants on the human transcriptome. *Science* **348,** 666–669 (2015).



Extended Data Figure 1 | Assessment of skew in XCI in GTEx female samples (v3 analysis release). a, The estimated skew in XCI by tissue across individuals. **b**, The skew in XCI by individual across available tissue samples. The number in brackets after the tissue or sample name indicates the number of individuals or tissues, respectively, contributing to each box plot. Details of the analysis can be found in the Supplementary Note.

a

Proportion of sex-biased expression

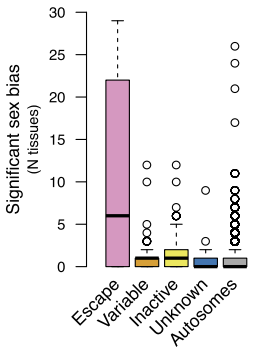
Group 1	Group 2	P-value	Mean (group 1)	Mean (group 2)
Escape	Variable	3.73E-09	0.48	0.06
Escape	Inactive	3.73E-09	0.48	0.04
Escape	Unknown	3.73E-09	0.48	0.03
Escape	Autosomes	3.73E-09	0.48	0.03
Variable	Inactive	2.19E-03	0.06	0.04
Variable	Unknown	9.67E-04	0.06	0.03
Variable	Autosomes	7.58E-06	0.06	0.03
Inactive	Unknown	9.58E-03	0.04	0.03
Inactive	Autosomes	2.43E-05	0.04	0.03
Unknown	Autosomes	1.00E+00	0.03	0.03

b

Consistency in effect direction

Group 1	Group 2	P-value	N (group 1)	Mean (group 1)	N (group 2)	Mean (group 2)
Escape	Variable	3.35E-12	73	0.88	71	0.66
Escape	Inactive	1.68E-23	73	0.88	376	0.64
Escape	Unknown	9.36E-14	73	0.88	74	0.65
Escape	Autosomes	2.46E-27	73	0.88	16350	0.64
Variable	Inactive	1.06E-01	71	0.66	376	0.64
Variable	Unknown	3.94E-01	71	0.66	74	0.65
Variable	Autosomes	7.53E-02	71	0.66	16350	0.64
Inactive	Unknown	6.33E-01	71	0.64	74	0.65
Inactive	Autosomes	9.65E-01	376	0.64	16350	0.64
Unknown	Autosomes	6.48E-01	74	0.65	16350	0.64

c

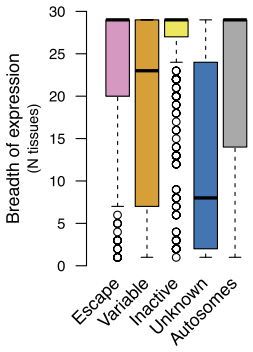


d

Significant sex bias (Number of tissues)

Group 1	Group 2	P-value	N (group 1)	Mean (group 1)	N (group 2)	Mean (group 2)
Escape	Variable	2.52E-10	82	11.02	89	1.08
Escape	Inactive	1.06E-15	82	11.02	390	1.06
Escape	Unknown	2.01E-18	82	11.02	120	0.37
Escape	Autosomes	1.68E-27	82	11.02	19158	0.62
Variable	Inactive	6.48E-01	89	1.08	390	1.06
Variable	Unknown	4.76E-06	89	1.08	120	0.37
Variable	Autosomes	3.43E-03	89	1.08	19158	0.62
Inactive	Unknown	8.01E-10	390	1.06	19158	0.37
Inactive	Autosomes	1.60E-12	390	1.06	19158	0.62
Unknown	Autosomes	4.08E-04	120	0.37	19158	0.62

e

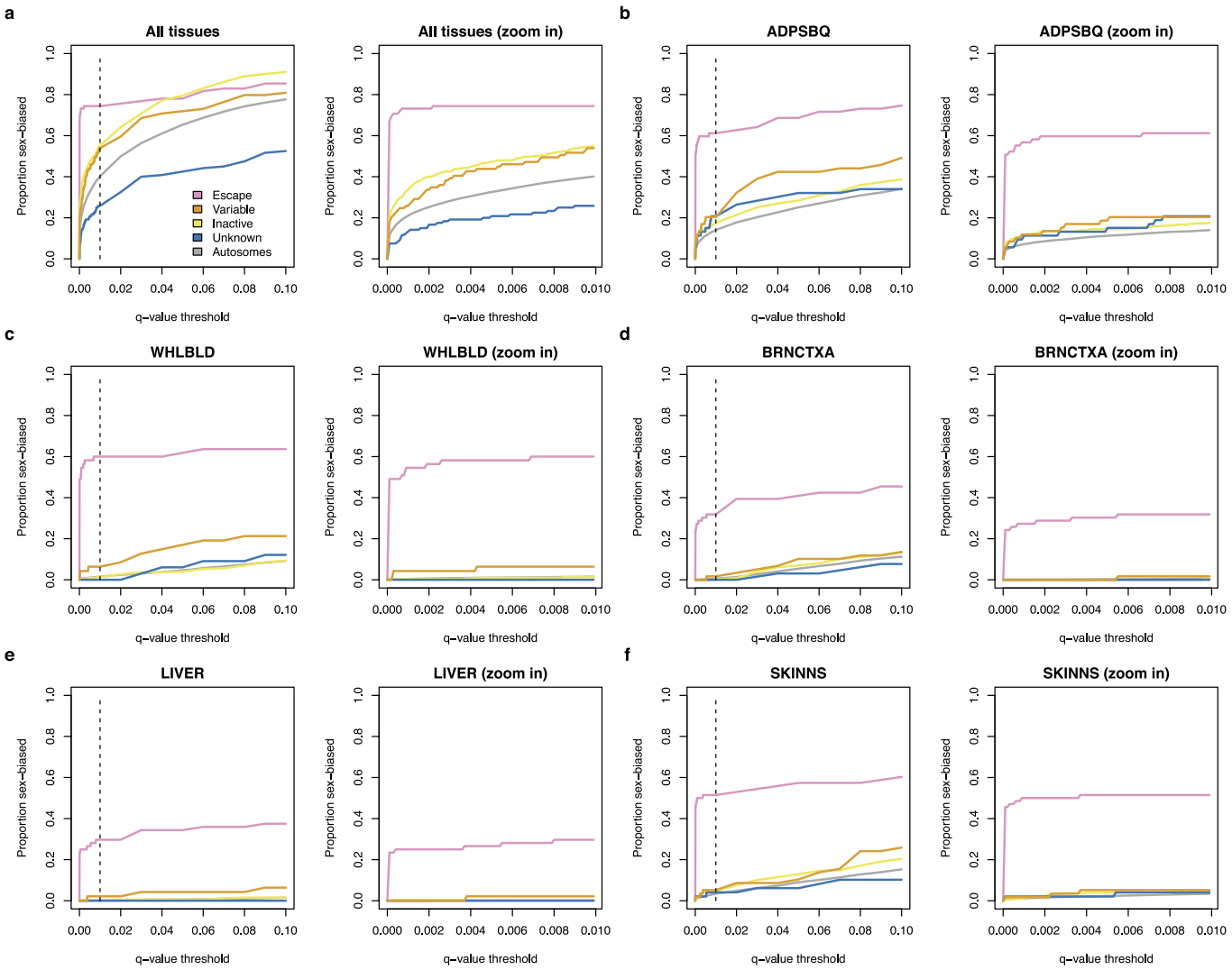


f

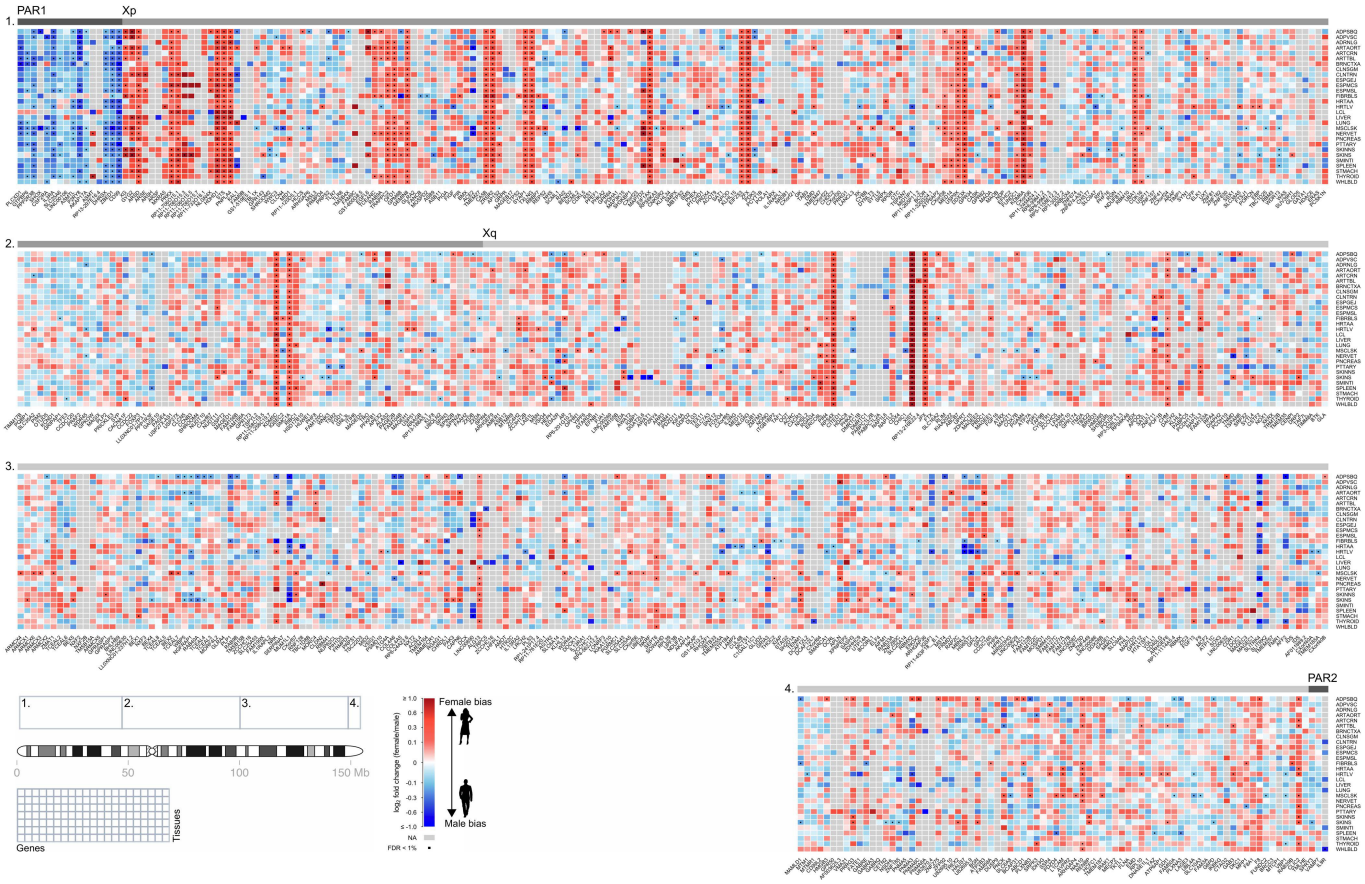
Breadth of expression (Number of tissues)

Group 1	Group 2	P-value	N (group 1)	Mean (group 1)	N (group 2)	Mean (group 2)
Escape	Variable	1.65E-04	82	22.84	89	18.21
Escape	Inactive	3.83E-02	82	22.84	390	26.07
Escape	Unknown	4.32E-11	82	22.84	120	12.03
Escape	Autosomes	2.24E-01	82	22.84	19158	21.78
Variable	Inactive	6.60E-16	89	18.21	390	26.07
Variable	Unknown	9.51E-05	89	18.21	120	12.03
Variable	Autosomes	1.35E-05	89	18.21	19158	21.78
Inactive	Unknown	2.65E-35	390	26.07	19158	12.03
Inactive	Autosomes	1.82E-13	390	26.07	19158	21.78
Unknown	Autosomes	2.75E-22	120	12.03	19158	21.78

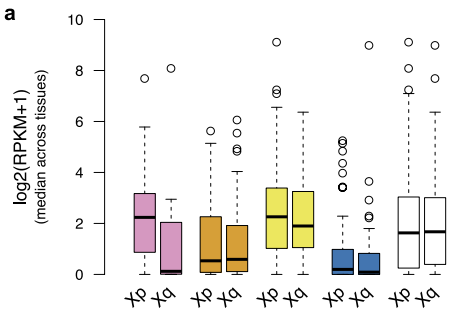
Extended Data Figure 2 | Comparison of expression characteristics between reported genic XCI categories in the GTEx data. a, The statistics for the comparison of the proportion of significantly biased (FDR <1%) genes by reported XCI status. Distributions are illustrated in Fig. 2b. $n = 29$ for all comparisons. **b**, The statistics for the comparison of the consistency in effect sizes across tissues. Distributions are illustrated in Fig. 2c. Only genes expressed in at least five of the 29 tissues are included. **c**, Number of tissues showing significant sex bias (FDR <1%) per gene by reported XCI status. **d**, Statistics for the comparison illustrated in **c**. **e**, Number of tissues in which genes are expressed by reported XCI status. **f**, Statistics for the comparison illustrated in **e**. All *P* values are from two-sided Wilcoxon rank-sum tests, except for a, where a paired, two-sided Wilcoxon rank-sum test was used. Only genes assessed for sex bias in at least one tissue are included, unless otherwise stated.



Extended Data Figure 3 | Change in the proportion of discovered sex-biased genes by XCI category with varying *q* value cut-offs. a, The proportion of sex-biased genes across tissues. Here a gene is classified as sex-biased if the *q* value for association falls below the given threshold in at least one tissue. **b**–**f**, Examples of the change in the proportion of sex-biased expression in individual tissues. The dashed black line indicates the FDR <1% cut-off applied in the analyses to determine sex-biased expression. ADPSBQ, adipose, subcutaneous; WHLBLD, whole blood; BRNCTXA, brain, cortex; SKINNS, skin not sun exposed (suprapubic).

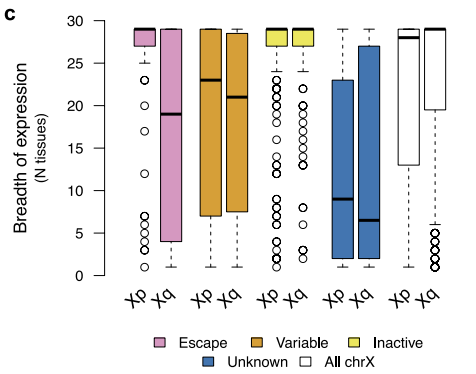


Extended Data Figure 4 | Heat map representation of male–female expression differences in all assessed X-chromosomal genes ($n = 681$) across 29 GTEx tissues. The colour scale displays the direction of sex bias, with red colour indicating higher female expression. Genes that were too weakly expressed to be assessed in a given tissue type in the sex bias analysis are coloured grey. Dots mark the observations where sex bias was significant at FDR <1%.



b

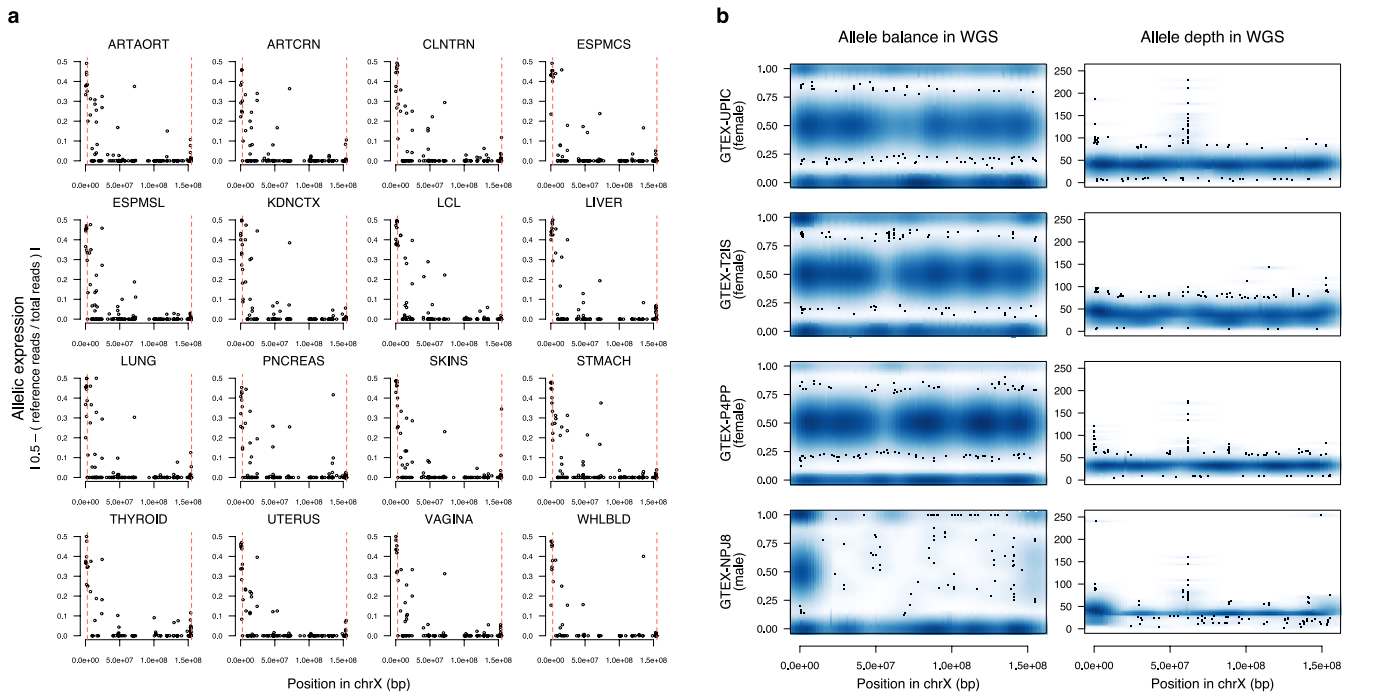
Group	P-value	N (Xp)	Mean (Xp)	N (Xq)	Mean (Xq)
Escape	1.43E-03	63	2.30	19	1.19
Variable	8.74E-01	32	1.35	57	1.32
Inactive	4.02E-01	138	2.27	252	2.42
Unknown	6.34E-01	38	0.77	82	0.74
All chrX	4.08E-01	271	1.95	410	1.87



d

Group	P-value	N (Xp)	Mean (Xp)	N (Xq)	Mean (Xq)
Escape	1.83E-03	63	24.84	19	16.21
Variable	9.21E-01	32	18.13	57	18.26
Inactive	9.57E-01	138	26.19	252	26.00
Unknown	9.28E-01	38	12.29	82	11.91
All chrX	7.50E-02	271	22.97	410	21.66

Extended Data Figure 5 | Comparison of expression characteristics between Xp and Xq, the evolutionary newer and older regions of chrX, respectively, by XCI status and for the whole chromosome. **a**, **b**, The level of median expression across GTEx tissues in $\log_2$ RPKM units. **c**, **d**, The breadth of expression measured as the number of tissues (max = 29) in which genes are expressed (median expression across samples >0.1 RPKM and expressed in more than 10 individuals at >1 counts per million). *P* values are calculated using the Wilcoxon rank-sum test. All genes expressed in at least one tissue are included in the comparisons.

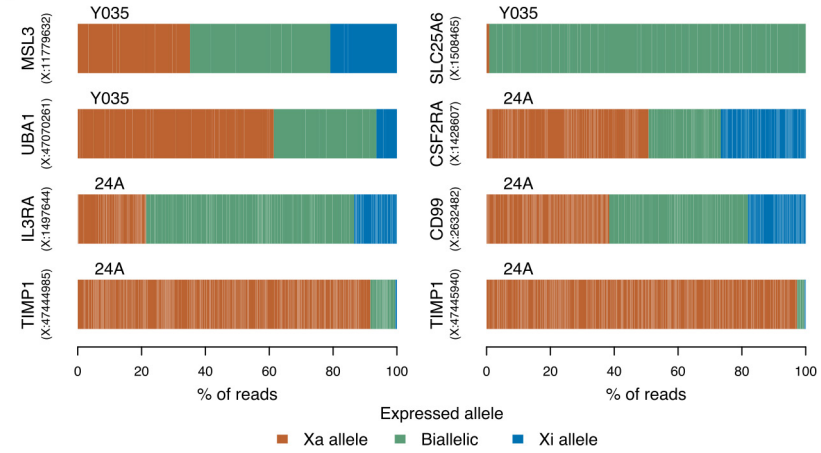


Extended Data Figure 6 | X-chromosomal RNA-seq and WGS data in the GTEx donor with fully skewed XCI (GTEX-UPIC). a, Allelic expression in chrX in 16 RNA-sequenced tissue samples available from the donor. Dashed red lines indicate PAR1 and PAR2 boundaries. **b**, Allele balance and allele depth across chrX in WGS for GTEX-UPIC and two female and one male GTEx WGS samples that were randomly chosen.

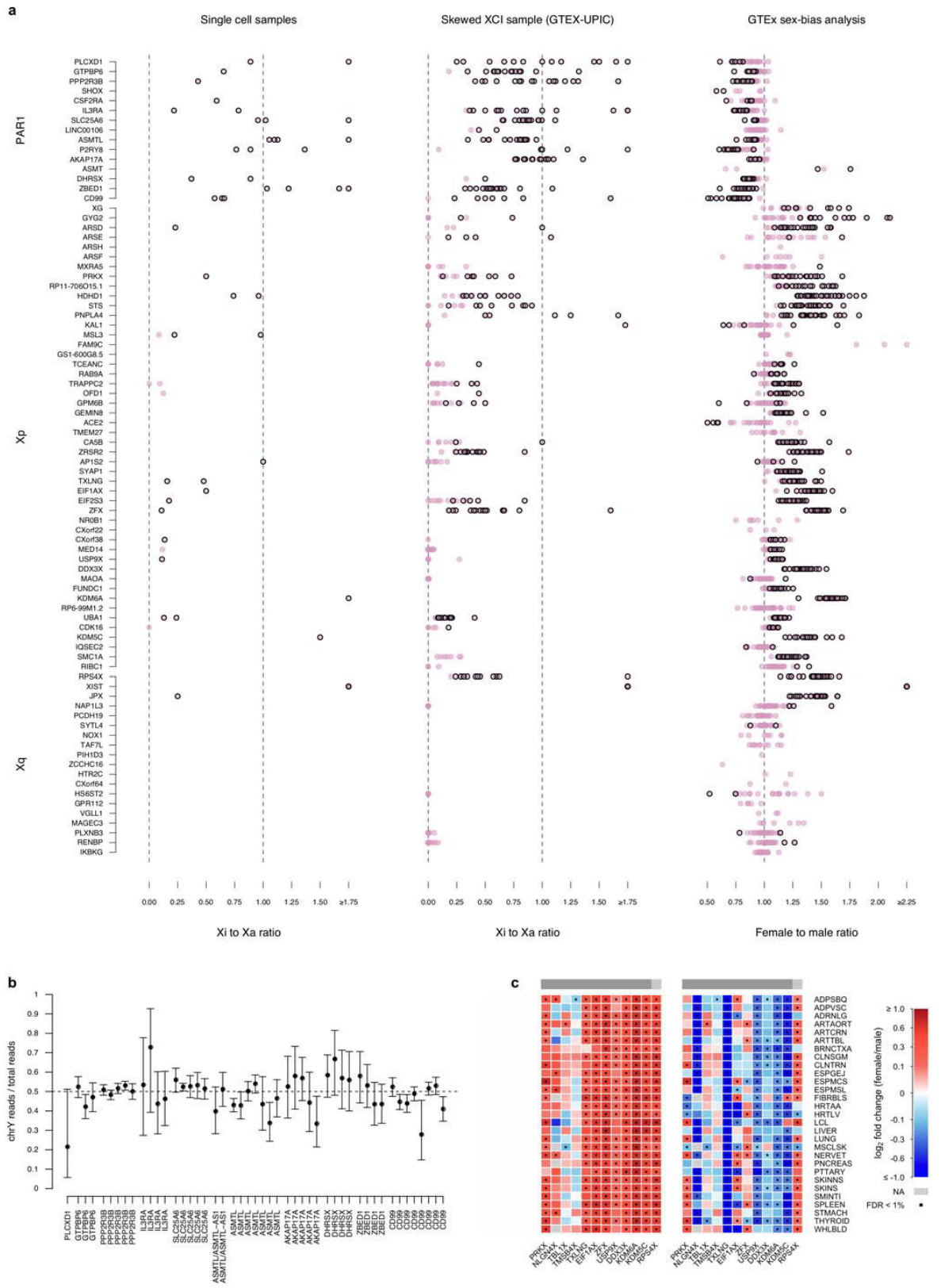
a

ID	Number of genes detected in several cells (ASE site coverage ≥8 reads in individual cells)		
	Biallelic	Monoallelic	Biallelic genes
Y117	2	25	CD99;UBA1
Y035	3	11	MSL3;SLC25A6;UBA1
Y014	3	16	SLC25A6;EIF2S3;ATP6AP2
24A	4	31	CSF2RA;IL3RA;CD99;TIMP1

b



Extended Data Figure 7 | Expressed alleles at biallelically expressed ASE sites in scRNA-seq. a, X-chromosomal genes repeatedly biallelic in scRNA-seq (see Methods for details). **b**, Illustration of the relative expression from the two alleles at all X-chromosomal ASE sites that were repeatedly biallelically expressed across cells in either of the two scRNA-seq samples that showed random XCI (Y035 and 24A). Narrow white lines separate observations from individual cells.



Extended Data Figure 8 | See next page for caption.

Extended Data Figure 8 | Assessment of the level of Xi expression at escape genes and in different regions of the X chromosome. a, The ratio of Xi-to-Xa expression in the single-cell samples (left; each circle represents a sample), in the skewed XCI donor from GTEx (middle; each circle represents a tissue), and the female-to-male ratio in expression (right; each circle represents a tissue) at reported escape genes. Genes are ordered according to their location in the X chromosome with genes in the pseudoautosomal region residing in the top part. A dark border around a circle indicates that there was significant evidence for Xi expression greater than the baseline in the given sample or tissue (left and middle) or significant sex-bias in the given tissue (right). Given some outliers, for example, *XIST*, the Xi-to-Xa ratio is capped at 1.75 and female-to-male ratio at 2.25. **b**, The relative expression arising from the X and Y chromosome at PAR1 genes in skeletal muscle in eight males. The allelic expression at these genes was assigned to the two chromosomes using parental genotypes available for these samples (see Methods for details). The dashed line at 0.5 indicates the point where expression from X and Y chromosomes is equal. The error bars give the 95% confidence intervals for the observed read ratio. **c**, Heat map representation of the change in pattern of sex-bias at 13 X–Y homologous gene pairs (see Methods for details) in nonPAR from only including the X-chromosomal expression (heat map on the left) to accounting for the Y-chromosomal expression (heat map on the right). The colour scale displays the direction of sex-bias with red colour indicating higher female expression. Genes that were too lowly expressed in the given tissue type to be assessed in the sex-bias analysis are coloured grey. Dots mark the observations where sex-bias was significant at FDR $<1\%$. The grey bars on top of the heat maps indicate the location of the gene in the X chromosome: dark grey indicating Xp and lighter grey Xq.

Extended Data Table 1 | Tissues, individuals and genes in the GTEx sex-bias analysis

Tissues		Individuals				Genes analyzed		
Abbreviation	Full name	All	Females	Males	Mean age	All	Autosomes	ChrX
ADPSBQ	Adipose - Subcutaneous	297	186	111	52.15	15,273	14,735	538
ADPVSC	Adipose - Visceral (Omentum)	184	117	67	51.97	15,301	14,765	536
ADRNLG	Adrenal Gland	126	70	56	50.51	14,956	14,435	521
ARTAORT	Artery - Aorta	197	126	71	51.11	14,675	14,137	538
ARTCRN	Artery - Coronary	118	70	48	51.7	14,881	14,350	531
ARTTBL	Artery - Tibial	284	183	101	50.26	14,501	13,981	520
BRNCTXA	Brain - Cortex	92	66	26	57.67	15,339	14,791	548
CLNSGM	Colon - Sigmoid	114	72	42	48.28	15,045	14,524	521
CLNTRN	Colon - Transverse	255	159	96	50.93	15,732	15,181	551
ESPGEJ	Esophagus - Gastroesophageal Junction	124	74	50	53.52	14,770	14,245	525
ESPMCS	Esophagus - Mucosa	169	97	72	48.89	15,137	14,617	520
ESPMSL	Esophagus - Muscularis	126	78	48	50.74	14,879	14,356	523
FIBRBLS	Cells - Transformed fibroblasts	240	150	90	50.2	13,635	13,158	477
HRTAA	Heart - Atrial Appendage	218	137	81	48.62	14,662	14,145	517
HRTLV	Heart - Left Ventricle	159	105	54	53.64	14,075	13,586	489
LCL	Cells - EBV-transformed lymphocytes	190	123	67	50.75	13,067	12,621	446
LIVER	Liver	96	63	33	53.52	14,031	13,556	475
LUNG	Lung	277	181	96	52.06	16,154	15,590	564
MSCLSK	Muscle - Skeletal	361	228	133	51.85	13,623	13,153	470
NERVET	Nerve - Tibial	256	163	93	51.65	15,563	15,020	543
PNCREAS	Pancreas	149	87	62	50.09	14,355	13,861	494
PTTARY	Pituitary	86	64	22	56.37	16,068	15,489	579
SKINNS	Skin - Not Sun Exposed (Suprapubic)	195	128	67	53.06	15,601	15,069	532
SKINS	Skin - Sun Exposed (Lower leg)	300	188	112	52.22	15,746	15,211	535
SMINTI	Small Intestine - Terminal Ileum	77	43	34	47.62	15,594	15,046	548
SPLEEN	Spleen	89	50	39	48.26	14,993	14,469	524
STMACH	Stomach	169	97	72	48.2	15,604	15,057	547
THYROID	Thyroid	278	179	99	52.14	15,974	15,417	557
WHLBLD	Whole Blood	338	213	125	51.64	13,187	12,751	436
Total		449	290	159	52.27	19,839	19,158	681

Extended Data Table 2 | scRNA-seq samples

ID	24A	Y117	Y035	Y014
Ancestry	China, Asia	Yoruba / Nigeria, Africa	Yoruba / Nigeria, Africa	Yoruba / Nigeria, Africa
Design	Singleton	Trio	Trio	Trio
Genotype data	WES	WGS	WES	WES
Number of cells	742	96	48	48
Cell type	Dendritic cells	LCL	LCL	LCL
Sequenced read pairs (mean (range))	1,187,000 (335-7,403,000)	2,547,000 (38,190-5,126,000)	2,571,000 (46,940-5,038,000)	2,436,000 (69,130-5,457,000)
Aligned read pairs* (mean (range))	808,600 (197-5,727,000)	1,471,000 (14,910-3,309,000)	1,459,000 (16,400-2,893,000)	1,391,000 (14,920-3,067,000)
Alignment rate (mean (range))	0.667 (0.271-0.799)	0.545 (0.251-0.645)	0.551 (0.266-0.615)	0.526 (0.175-0.606)
Skew in XCI (% maternal active : % paternal active)	54:46 (373 cells where one parental chromosome active, 315 cells where the other parental chromosome active, 54 cells uninformative for X-chromosomal phasing)	100:0 (90 cells where maternal X chromosome active, 6 cells uninformative for X-chromosomal phasing)	79:21 (37 cells where maternal X chromosome active, 8 cells where paternal X chromosome active, 2 cells uninformative for X-chromosomal phasing)	100:0 (43 cells where maternal X chromosome active, 2 cells uninformative for X-chromosomal phasing)
Notes	Due to the unavailability of parental genotype information, the parental origin of the inferred X-chromosomal haplotypes is unknown			

*Uniquely aligned, properly paired, quality-control passed reads.

nature research

Corresponding author(s): Tukiainen / MacArthur

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

No statistical methods were used to predetermine sample size.

2. Data exclusions

Describe any data exclusions.

The GTEx samples were curated according to pre-established QC criteria as detailed in the accompanying manuscript by Aguet et al. ScRNA-seq data was limited to those cells that were informative for chromosome X allelic expression.

3. Replication

Describe whether the experimental findings were reliably reproduced.

The analyses conducted were exploratory and the results were not replicated in independent data sets. However each analysis included multiple data points (individuals and/or tissues) thus providing further support for the conclusions drawn.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

The experiments were not randomized. The study included no allocation into experimental groups.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

The investigators were not blinded to allocation during experiments and outcome assessment. The study included no allocation into experimental groups.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	RNA-seq alignment: Tophat version v1.4.1, STAR versions 2.4.2a, 2.4.1a or 2.3.0e; RNA-seq QC and quantification: RNA-SeQC; Allelic expression and variant calling: GATK version 3.1 or 3.4. Data processing: R version 3.4.0.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	All unique materials are readily available from the authors or from commercial sources as described in the Online Methods.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	The antibody panels used to enrich for all known blood DC population for single cell sorting and single cell RNA-sequencing (scRNA-seq) are described in Villani et al (Science 2017). All antibodies are commercially available as described in Supplementary Table 14.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	YRI LCLs were obtained from NHGRI Sample Repository for Human Genetic Research (Coriell Institute for Medical Research).
b. Describe the method of cell line authentication used.	None of the cell lines used were authenticated.
c. Report whether the cell lines were tested for mycoplasma contamination.	Coriell Biorepositories declares that their lymphoblastoid cell lines are free of bacterial, fungal or mycoplasma contamination. No other tests were run to test for mycoplasma contamination.
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No commonly misidentified cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	No animals were used

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	24A: Female, Asian ancestry, 25 yo, healthy Y117, Y035 and Y014: Female, African ancestry, age and health status unknown GTEx-UPIC: Female, European ancestry, 21 yo, cause of death asphyxiation See Extended Data Table for information on other GTEx donors