Title: A whole organism map of X-inactivation reveals tissue-specific escape as an exceptionally rare event in humans

Authors: Björn Gylemo[1] and Colm E. Nestor[1*]

Email: bjorn.gylemo@liu.se and colm.nestor@liu.se

Affiliations:

[1]Crown Princess Victoria Children´s Hospital, Department of Biomedical and Clinical Sciences (BKV), Linköping University, Linköping, Sweden

*Corresponding author. Email: Colm.Nestor@liu.se

Abstract

As females are mosaic for X-inactivation, direct determination of X-linked allelic expression in bulk tissues is typically unfeasible. Using females that are non-mosaic for X-inactivation (nmXCI) as a powerful and natural genetic system we directly determined X-inactivation status of 401 X-linked genes across 32 normal tissues. We reveal that nmXCI is common in the female population and that contrary to previous reports, variable escape is an exceedingly rare event in humans. As the most complete atlas of XCI in humans to date, our findings represent a major advance in our understanding of human X-inactivation and will serve as a reference for dissecting the genetic origin of sex-bias in human traits.

Keywords

X-chromosome inactivation, allele-specific expression, non-mosaic XCI

Background

Human females are typically 'mosaic' for X-inactivation; female tissues contain a mixture of cells that have inactivated either the maternal (Xm) or paternal X-chromosome (Xp) (Fig. 1A). Mosaicism renders analysis of X-inactivation (XCI) in primary human tissues highly complex

26 or unfeasible. Consequently, our knowledge of XCI across human tissues is surprisingly sparse
27 and typically based on indirect and imprecise measures of XCI, such as DNA methylation, sex-
28 biased expression or from observations in animal models (1,2). Here, we establish a unique set
29 of three unrelated females in the Genotype-Tissue Expression (GTEx) project in which the same
30 parental X-chromosome was inactivated in all cells, thereby removing the confounding effect
31 of mosaicism (3). Such non-mosaic (nmXCI) females enable direct determination of the XCI
32 status even from bulk tissue samples. Using these powerful and natural human genetic models
33 we could directly determine the XCI status of 401 genes across 32 normal human tissues,
34 representing a major advance in our understanding of human X-inactivation.
35

36 **Results and Discussion**

37 *Non-mosaic X-inactivation (nmXCI) is common in human females*

38 Females exhibiting non-mosaic XCI (nmXCI) are a powerful system in which to dissect the
39 XCI status of human X-linked genes, but their frequency in the general population is unknown
40 with current estimates ranging from 1:50 to 1:500 individuals (*4-7*). To identify nmXCI females
41 we first screened allele-specific expression of genes known to undergo X-inactivation (**Table**
42 **S1**) in a single tissue across all 285 female donors (**Table S2**) in the GTEx database (v8 release)
43 (*7*), including a single nmXCI female previously identified in GTEx (donor UPIC) as a control
44 (*7*). Tissues used for the screen were selected based on abundance in the entire data set. Six
45 female samples showed complete mono-allelic expression of inactive genes (allelic expression
46 > 0.4) in the selected tissue, consistent with expression from a single parental chromosome,
47 nmXCI (**Fig. 1B**). Extending allele-specific expression analysis to all available tissues (**Table**
48 **S3**) across the six potential nmXCI females confirmed the complete nmXCI status of three
49 individuals (including the nmXCI control, UPIC) (**Fig. 1C**). The remaining three potential
50 nmXCI females showed a mixture of mosaic and highly skewed XCI across tissues (**Fig. 1C**).

We found no obvious detrimental mutations in the coding or regulatory regions of the *XIST* gene as previously described in a report of familial nmXCI (**see methods**) (*8*), thus the genetic basis of nmXCI in these three individuals is unknown.

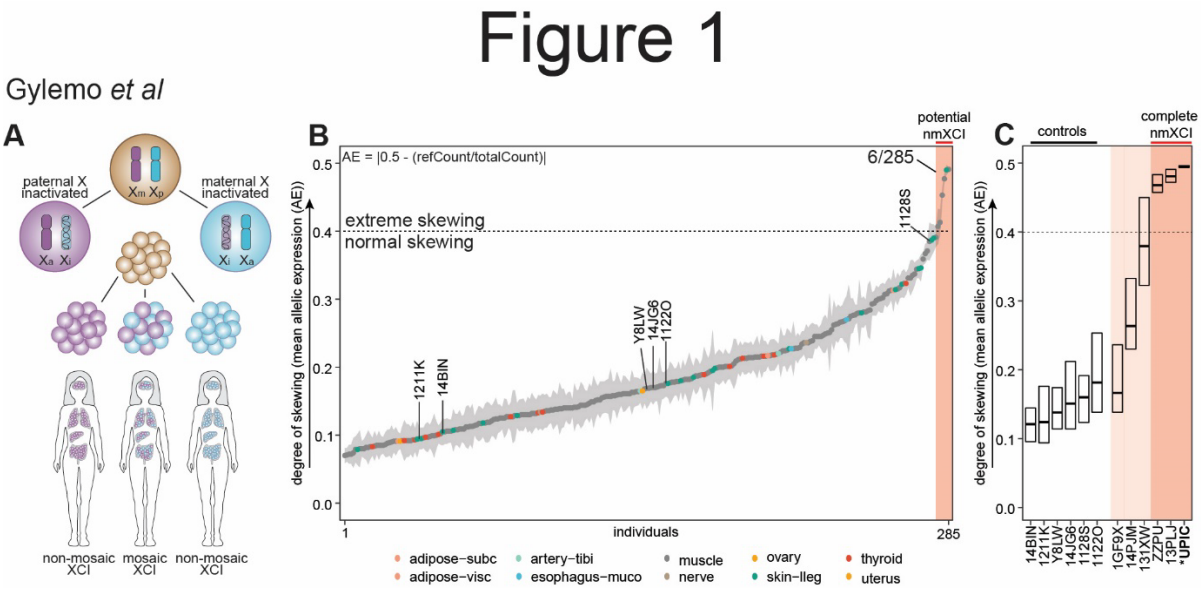

Fig. 1. Identification of females with non-mosaic X-inactivation. (**A**), The patterns of X-chromosome inactivation (XCI) in women resulting in mosaic or non-mosaic XCI. (**B**), Single-tissue allelic expression (AE) of 237 inactive genes (*7*) (see **Table S1**) in all 285 women in GTEx. Tissue abbreviations can be found in **Table S9**. Each dot represents mean AE across all inactive genes in that sample, grey ribbon represents standard error and the red box highlights women with a mean AE > 0.4 (UPIC, 13PLJ, ZZPU, 131XW, 14PJM and 1GF9X). (**C**), Multiple-tissue AE of the 237 inactive genes (7) (see **Table S1**) in the six women with a mean AE > 0.4 and 6 randomly selected females (14BIN, 1211K, Y8LW, 14JG6, 1129S and 1122O). Boxes indicate mean ± 25th and 75th inter quartile range (IQR).

A landscape of X-inactivation across normal human tissues

The ability to detect allele-specific expression is dependent on the presence of a heterozygous variant in the gene of interest. Using both whole exome sequencing and whole genome

69 sequencing for all three nmXCI individuals we identified 756 heterozygous SNPs (**Fig. 2A and**
70 **Table S4**) across 401 X-linked genes (**Fig 2B, Fig. S1A and Table S4**), representing 43% of
71 all protein coding X-linked genes (**Fig. S1B)**. For 177 of these X-linked genes this is the first
72 direct determination of their XCI status across multiple tissues in humans (**Fig. S1A and Table**
73 **S5**). The tissue-specific allelic expression status of X-linked genes was highly similar across all
74 three individuals and independent of the SNP used in each comparison, supporting the
75 robustness of our approach (**Fig. S1C and Fig. S1D**). As expected, genes located in the pseudo-
76 autosomal regions (PARs) of the X-chromosome and well-characterised nonPAR escape genes
77 showed clear bi-allelic expression across tissues (**Fig. 2D and Fig. 2E**), genes undergoing X-
78 inactivation showed mono-allelic expression (**Fig. 2F**) whereas 66 genes exhibited varying
79 degrees of mono- and bi-allelic expression across tissues, 'variable' escape genes (**Fig. 2G**).

80 *Tissue-specific escape from XCI is exceptionally rare in humans*

81 Unexpectedly, whereas PAR, escape and inactive genes identified here showed clear overlap
82 with previous classifications of XCI status, only ~7% of variable escape genes were similarly
83 classified in previous studies of XCI (**Fig. 2H**). We noted that genes previously classified as
84 'variable' escape were significantly less expressed than all other classes of X-linked genes
85 across 50 tissues in 103 randomly selected females in the GTEx, regardless of whether the
86 current or previous classification of XCI was used (**Fig. 2I, Fig. S1E, Fig. S1F and Table S6**).
87 This suggests that classification of many genes as variable escape may have reflected a lower
88 expression level and consequently fewer sequencing reads, which increased risk for inaccurate
89 determination of allelic balance. To explore this phenomenon further we manually examined
90 the SNP quality and read depth for all 66 variable genes across all tissues and individuals.
91 Whereas some known (e.g., *ANOS1*, *CLIC2, NAA10*) and novel (*ARSD, CA5B, TRAPPC2*)
92 genes show clear and reproducible variability in escape across tissues and individuals, most

genes previously classified as variable have expression at the limit of detection and have most likely been miss-classified as a result (**Fig. 2J, Fig. 2K, Fig. S2A and Table S7**).

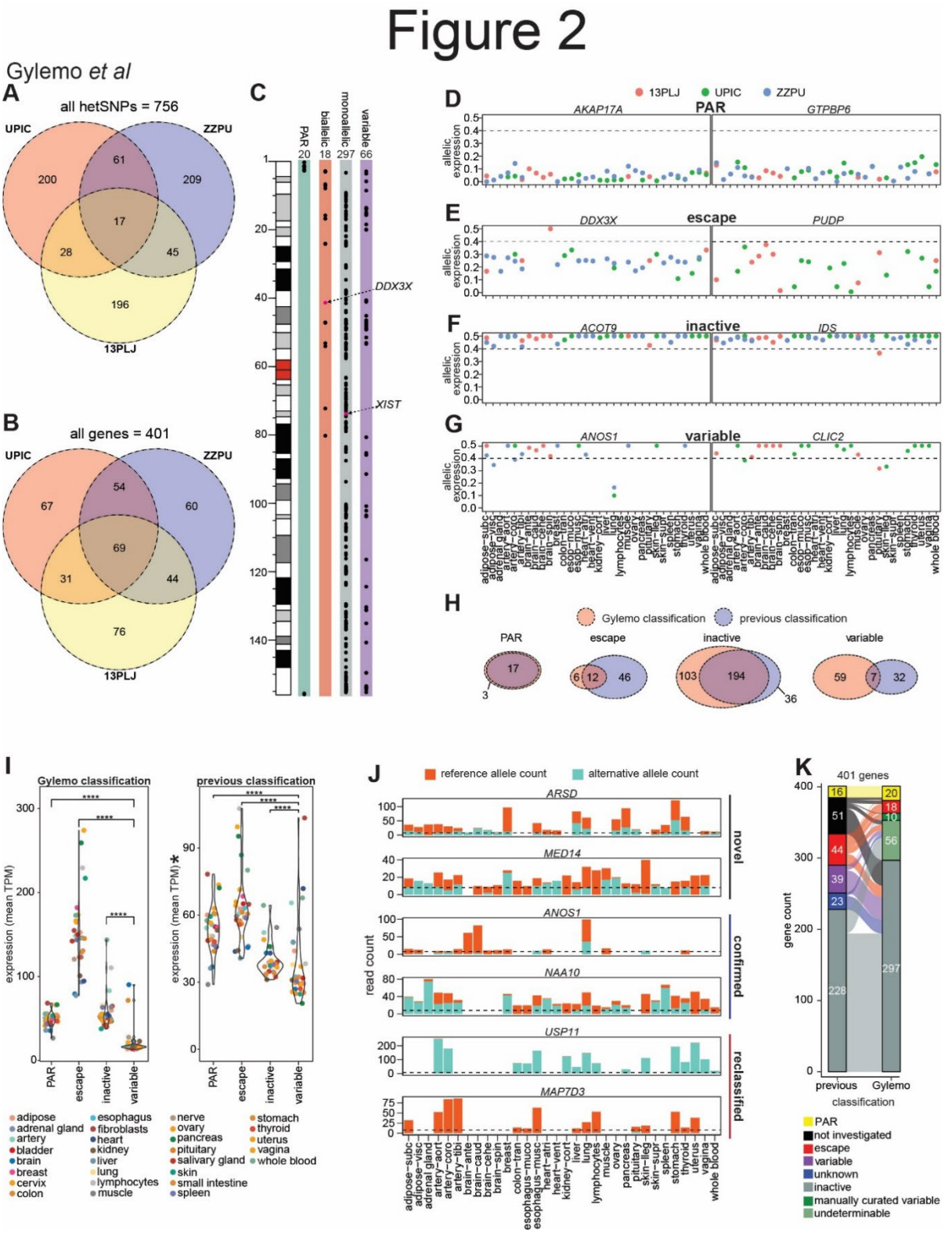

Fig. 2. An extended landscape of X-inactivation in humans. (A,B), Overlap of heterozygous SNPs (hetSNPs) (A) and genes (B) across the three females with non-mosaic X-inactivation

(nmXCI). (**C**), Genomic location and allele-specific classification count of covered X-linked genes. PAR = genes located in the pseudoautosomal region, biallelic = genes with an allelic expression (AE) < 0.4 in at least 85% of individual sample and tissue, monoallelic = genes with an AE > 0.4 in at least 85% of individual sample and tissue, variable = genes with AE < 0.4 in more than 15% and less than 85 % of individual sample and tissues. (**D-G**), AE across tissues in the three nmXCI females for two example genes per category: PAR genes (D), escape genes (E), inactive genes (F), variable genes (G). Data points missing in (E, left panel) and (G, left panel) reflect the lack of a hSNPs present in that gene in ZZPU. (**H**), Number of genes per XCI classification and their overlap with a previous assessment (*7*). (**I**), Mean transcript-per-million (TPM) expression of each gene categories previously described (*7*) across 103 randomly selected females (see **Fig. S1F** for number of samples per tissue). Each dot indicates the mean expression of all genes per tissue within each gene category. Tissues are merged on tissue category, see **Fig. S1E** for all tissues included in GTEx. ****: p-value <= 0.0001, Benjamini-Hochberg adjusted unpaired two-sided Wilcoxon test. (**J**), Reference and alternative allele read counts of example novel and known (*7*) variable genes and variable genes reclassified as inactive genes. **(K)**, Reclassification of previous assessments (*7*) of XCI status of genes. (D-I), tissue abbreviations can be found in **Table S9.**

Conclusions

Here we describe the most comprehensive map of human X-inactivation to date. This data represents a doubling of the number of X-linked genes for which XCI status has been directly determined, across multiple normal tissues and individuals. Our results suggest that XCI is highly stable both within and across individuals, and that previously suggested levels of tissue-specific escape from X-inactivation may have been vastly overestimated. The classification of many genes as variable escape may be a consequence of a low read count, which increases the

risk for inaccurate determination of XCI status. Variability in X-linked gene expression is important to characterise, as XCI variability would directly modify expressivity of X-linked pathogenic variants and provide a mechanism for sex-biased expression differences in certain biological contexts, which in turn can modify disease-risk. Moreover, our identification of additional nmXCI females among the 285 females in the GTEx database further substantiates recent claims of extreme skewing of XCI as a frequent and major modifier of trait penetrance and expressivity in the human population (5,6).

Materials and Methods

Allele-specific expression analysis

For initial screen for identifying nmXCI females in GTEx, pre-processed single-nucleotide polymorphisms (SNP) calling performed from whole-exome sequencing (WES) data from GTEx was downloaded (GTEx_Analysis_2017-06-05_v8_WholeExomeSeq_979Indiv_VEP_annot.vcf.gz, see https://gtexportal.org/home/methods for specifics on how GTEx processed the data), lifted over from hg19 to hg38 with LiftoverVcf (GATK v.4.1.9.0) (*9*), bcftools (v1.10.2) (*10*) was utilized to include only heterozygous SNPs (hetSNPs) on the X-chromosome. One RNA-seq sample per participant (see **Table S2**) was downloaded from AnVIL using the Gen3 client, as aligned BAM files (see https://gtexportal.org/home/methods for specifics on how GTEx processed the data). Tissues used for the screen were selected based on abundance in the entire data set (muscle (n=239), skin-lleg (n=21), thyroid (n=14), adipose-subc (n=2), artery-tibi (n=2), esophagus-muco (n=2), nerve (n=2), adipose-visc (n=1), ovary (n=1), uterus (n=1), (see **Table S9** for tissue abbreviations)). Following data download, allelic expression (AE) analysis of allele counts for the hetSNPs was retrieved from the RNA-seq data using GATK ASEReadCounter (v.4.1.9.0), requiring a minimum base quality of 20 in the RNA-seq data. To

remove potentially spurious sites, conservative filters were applied to the data (variant call read depth of at least 20 per allele, minor allele read count > 10% of the total read count and RNA-seq read depth of more than 10). If a gene carried multiple hetSNPs, only the one with the highest RNA-seq read count was included. To assess the XCI status of each individual participant, the mean AE (allelic expression = |0.5 – reference reads/total reads|) of genes previously shown to be inactive (*7*) was calculated (see **Table S2**). A mean AE of inactive genes > 0.4 was used as a cut-off between mosaic XCI (mXCI) and non-mosaic XCI (nmXCI), with values > 0.4 indicating all or most cells has inactivated the same parental X-chromosome. As a control, UPIC was included. UPIC is a female which has previously been shown to have a nmXCI (*7*). As UPIC is not included in the GTEx_Analysis_2017-06-05_v8_WholeExomeSeq_979Indiv_VEP_annot.vcf.gz, SNP calling was performed using WES data and ASE analysis was performed (see next section for detailed pipeline), ultimately leveraging her ASE data as a positive control. For further assessing XCI status of the 6 samples with a high mean AE, the remaining RNA-seq tissue samples available for those participants were downloaded from GTEx and processed in the same manner as above (see **Table S3**).

Following the initial screen, we sought to refine the resolution of the XCI map. To this end, we downloaded WES and whole-genome sequencing (WGS) data available for the three nmXCI females (UPIC, 13PLJ and ZZPU) and performed SNP calling. Briefly, WES BAM and WGS CRAM files for the participants were downloaded from AnVIL using the Gen3 client. The files were sorted and converted into fastq files using Samtools (v1.9) (*10*). Raw fastq reads were quality trimmed using FastP (v.0.20.0) (*11*) with default settings. BWA-MEM (*12*) was used for mapping reads to the human genome build 38 (hg38, GCA_000001405.15_GRCh38_no_alt_analysis_set.fna) using default settings. Sambamba v0.8.2 (*13*) was used to mark duplicate reads. The GATK best practice germline short variant discovery pipeline was used to process the aligned reads, using base quality score recalibration

173 and local realignment at known indels (GATK v.4.2.6.1). Indels (insertions and deletions) and
174 SNPs were called jointly across all samples, for both WGS and WES data. Default filters were
175 applied to indel and SNP calls using the variant quality score recalibration (VQSR) approach
176 of GATK. All RNA-seq sample available for each participant (**Table S8**) was downloaded from
177 AnVIL using the Gen3 client as aligned BAM files, using Samtools (v1.9) (*10*) to sort and
178 convert the BAM files to fastq files and quality trimming with FastP (v.0.20.0) (*11*) with default
179 settings. RNA-seq reads were aligned with STAR 2-pass mode (v2.7.10a) (*14*) with a
180 GCA_000001405.15_GRCh38_no_alt_analysis_set.fna and --sjdbGTFfile
181 gencode.v40.annotation.gtf index. WASP filtering was performed to reduce the issue of
182 reference bias (*15*). Briefly, GATK SelectVariants was used to extract hetSNPs (that passed the
183 VQSR filtering) for each individual, from either the WGS or WES VCFs, and subsequently
184 passed to STAR via --varVCFfile and–waspOutputMode. Reads failing WASP filtering were
185 removed. Following data pre-processing, AE analysis of the allele counts for hetSNPs were
186 retrieved from RNA-seq data using GATK ASEReadCounter. Heterozygous variants that
187 passed VQSR filtering were first extracted for each sample from WES and WGS VCFs using
188 GATK SelectVariants. Sample-specific VCFs and RNA-seq BAMs were inputted to
189 ASEReadCounter requiring minimum base quality of 20 in the RNA-seq data.
190 ASEReadCounter outputs from WGS and WES pipelines were merged, if overlapping sites
191 were identified the variant read counts were merged. To remove potentially spurious sites,
192 conservative filters were applied to the data (variant call read depth >= 8 per allele, minor allele
193 read count > 10% of the total read count and RNA-seq read depth > 8). If a gene carried multiple
194 hetSNPs, only the one with the highest RNA-seq read count was included. For comparisons to
195 previous gene coverage of UPIC (**Fig. S1A**), supplementary table 5 in Tukiainen *et al* (*7*) was
196 utilized as a prior.

197 <u>X-inactivation status categorization</u>

198 All 401 genes were first classified into PAR or nonPAR depending on their chromosomal
199 location, and further categorized into 'biallelic', 'monoallelic' and 'variable' depending on their
200 AE in each individual and tissue. Genes with an AE < 0.4 in at least 85% of tissues and
201 individuals was classified as 'biallelic', Genes with an AE > 0.4 in at least 85% of tissues and
202 individuals was 'monoallelic' and genes that varied between in their 'monoallelic' or 'biallelic'
203 classification across tissues was classified as 'variable' (i.e., AE > 0.4 in more than 15% but
204 less than 85% of tissues. The genes categorized as 'variable' were subject to manual curation,
205 as it was difficult to accurately define a computational approach to robustly call XCI on
206 hetSNPs with low read coverage. Briefly, genes (i) with a low read count per allele and (ii) for
207 which changes in mono-allelic/bi-allelic classification across tissue were derived from single-
208 digit read counts were classified as 'undeterminable'. Genes for which variation in XCI status
209 across tissues were supported by higher read counts were considered more high-confidence
210 variable genes and thus classified as 'curated variable'.

211 <u>GTEx RNA-seq data</u>

212 GTEx RNA-seq data for was downloaded from AnVIL using the Gen3 client (as aligned BAM
213 files), sorted and converted into fastq files with Samtools (v1.9) (*10*) and quality trimmed with
214 FastP (v.0.20.0) using default settings. Transcriptome mapping was performed using Salmon
215 v1.9.0 (*16*) (--gcBias --seqBias --numBootstraps 100 -l A) using known Refseq transcripts
216 (NM_* and NR_*, excluding all Y-PAR transcripts) from the GRCh38.p12 assembly
217 (GCF_000001405.38_GRCh38.p12_rna.fna.gz) as a reference. Abundance estimates were
218 converted to h5 format using Wasabi (https://github.com/COMBINE-lab/wasabi) and data
219 normalization and gene-level transcript per million (TPM) was calculated using the gene_mode
220 function implemented in Sleuth (v0.30.0) (*17-18*).

221 <u>Statistical analysis</u>

222 For statistical testing between groups a Benjamini-Hochberg adjusted unpaired two-sided
223 Wilcoxon test was used. P-values < 0.05 was deemed significant.

224 Abbreviations

225 ***XCI***: X-chromosome inactivation

226 **Xm**: Maternal X-chromosome

227 **Xp**: Paternal X-chromosome

228 **GTEx:** Genotype-Tissue Expression

229 **nmXCI**: non-mosaic X-chromosome inactivation

230 **PAR**: pseudo-autosomal region

231 **SNP**: single-nucleotide polymorphisms

232 **WES**: whole-exome sequencing

233 **WGS**: whole-genome sequencing

234 **hetSNP**: heterozygous single-nucleotide polymorphisms

235 **AE**: allelic expression

236 **Declarations**

237 Ethics approval and consent to participate

238 Not applicable.

239 Consent for publication

240 Not applicable.

241 Availability of data and materials

242 No raw data was generated as part of this study. All GTEx data are available from 'The database
243 of Genotypes and Phenotypes (dbGaP)' under accession number phs000424. All code for
244 figure generation have been made freely available on GitHub
245 (https://github.com/bgylemo/EpiLiu0987).

246 Competing interests

Authors declare that they have no competing interests.

Funding

European Research Council grant 101045171 XX-Health (CEN)

Swedish Research Council gran 2020-01277_VR (CEN)

Authors' contributions

Conceptualization: CEN

Methodology: CEN, BG

Investigation: BG

Visualization: BG

Funding acquisition: CEN

Project administration: CEN

Supervision: CEN

Writing – original draft: CEN, BG

Acknowledgements

-

Authors' information (optional)

-

References

1. Carrel, L. & Willard, H.F. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* **434**, 400-4 (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. Consortium, G.T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-5 (2013).

271 4. Amos-Landgraf, J.M. *et al.* X chromosome-inactivation patterns of 1,005
272 phenotypically unaffected females. *Am J Hum Genet* **79**, 493-9 (2006).
273 5. Roberts, A.L. *et al.* Age acquired skewed X chromosome inactivation is associated with
274 adverse health outcomes in humans. *Elife* **11**(2022).
275 6. Shvetsova, E. *et al.* Skewed X-inactivation is common in the general female population.
276 *Eur J Hum Genet* **27**, 455-465 (2019).
277 7. Tukiainen, T. *et al.* Landscape of X chromosome inactivation across human tissues.
278 *Nature* **550**, 244-248 (2017).
279 8. Plenge, R.M. *et al.* A promoter mutation in the XIST gene in two unrelated families
280 with skewed X-chromosome inactivation. *Nat Genet* **17**, 353-6 (1997).
281 9. Castel, S.E., Levy-Moonshine, A., Mohammadi, P., Banks, E. & Lappalainen, T. Tools
282 and best practices for data processing in allelic expression analysis. *Genome Biol* **16**,
283 195 (2015).
284 10. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**(2021).
285 11. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ
286 preprocessor. *Bioinformatics* **34**, i884-i890 (2018).
287 12. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler
288 transform. *Bioinformatics* **26**, 589-95 (2010).
289 13. Tarasov, A., Vilella, A.J., Cuppen, E., Nijman, I.J. & Prins, P. Sambamba: fast
290 processing of NGS alignment formats. *Bioinformatics* **31**, 2032-4 (2015).
291 14. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21
292 (2013).
293 15. van de Geijn, B., McVicker, G., Gilad, Y. & Pritchard, J.K. WASP: allele-specific
294 software for robust molecular quantitative trait locus discovery. *Nat Methods* **12**, 1061-
295 3 (2015).

296 16. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A. & Kingsford, C. Salmon provides fast
297 and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417-419 (2017).
298 17. Pimentel, H., Bray, N.L., Puente, S., Melsted, P. & Pachter, L. Differential analysis of
299 RNA-seq incorporating quantification uncertainty. *Nat Methods* **14**, 687-690 (2017).
300 18. Yi, L., Pimentel, H., Bray, N.L. & Pachter, L. Gene-level differential analysis at
301 transcript-level resolution. *Genome Biol* **19**, 53 (2018).