Title: A whole-organism landscape of X-inactivation in humans

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

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

Contact: bjorn.gylemo@liu.se , maike.bensberg@liu.se, colm.nestor@liu.se

Classification: Biological Sciences, Genetics

Keywords: X-chromosome inactivation, mosaicism, allele-specific expression, escape from XCI

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 (completely skewed) for X-inactivation (nmXCI) has proven a powerful and natural genetic system for profiling X-inactivation in humans. By combining allele-resolution data for one previously reported and two newly identified nmXCI females, we directly determined X-inactivation status of 380 X-linked genes across 30 normal tissues, including 198 genes for which XCI status is directly determined for the first time. Our findings represent a substantial 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. In addition, our study reveals nmXCI as a common feature of the human female population, with profound consequences for the penetrance and expressivity of X-linked traits in humans.

Introduction

During embryonic development in humans, female cells (46, XX) inactivate a single X-chromosome to balance dosage of X-linked gene expression between females (XX) and males (XY). The process of X-chromosome inactivation (XCI) begins during the peri-implantation stage in humans and is initiated by expression of the long non-coding RNA (lncRNA) *XIST* from the X inactivation center (XIC) of one X-chromosome in each female cell. *XIST RNA* coats the X-chromosome which in turn recruits protein complexes including chromatin modifiers which establish a facultative heterochromatin state. Silencing histone modifications and the deposition of DNA methylation ultimately result in the transcriptionally silent inactive X-chromosome (Xi) (**Supp Fig. 1A**). The selective expression of *XIST* from only one X-chromosome needs to be tightly regulated. Whereas the exact mechanism by which a given X-chromosome in a female cell is selected for inactivation is unknown, the XIC in humans includes several lncRNAs that have been implicated in the regulation of *XIST* expression and X-inactivation (**Supp Fig. 1A**) (1). Although the initial choice of X-chromosome for inactivation is random, the inactivated X-chromosome is stably inherited in a clonal fashion throughout all subsequent cell divisions (2). Interestingly, the process of X-inactivation (XCI) is incomplete; approximately 15% of genes escape XCI and remain expressed from both X-chromosomes (3). These 'escape' genes are consequently more highly expressed in females than in males, although the degree of sex-biased expression of escape genes can vary and may be greater for those genes that lack a functional homologue on the Y-chromosome (4). The sex-biased expression of escape genes has been suggested to mediate sex differences in both the frequency and severity of several diseases (5, 6). In addition, genes in the pseudoautosomal regions (PAR1 and PAR2) of the X-chromosome also remain expressed from the Xi and are thus always bi-allelically expressed. PAR1 and PAR2 are short regions of homology between the X- and Y-chromosomes that are essential for pairing of the sex chromosomes during meiosis in males (7). In contrast to escape genes, which tend to be more highly expressed in female cells, genes in the PARs have been suggested to show male-biased expression (8).

82 Consequently, the terms PAR and nonPAR are used to distinguish between the biallelic
83 expression (escape) of genes within the PARs and those located elsewhere on the X-
84 chromosome.
85
86 Human females are typically 'mosaic' for X-inactivation; female tissues contain a mixture of
87 cells that have inactivated either the maternal or paternal X-chromosome (**Fig. 1A**). Mosaicism
88 renders analysis of XCI in primary human tissues highly complex or unfeasible. Consequently,
89 our knowledge of XCI and escape across human tissues is surprisingly sparse and typically
90 based on indirect measures of XCI, such as DNA methylation, sex-biased expression or from
91 observations in cell lines and animal models (4, 9). In a seminal study, Brown and colleagues
92 reported XCI status for 639 genes by integrating data from three published studies that each
93 used different indirect approaches and model systems to infer XCI status (3). This resource
94 continues to serve as a valuable reference for inferred XCI status in humans.
95
96 Interestingly, rare cases of humans in which the same parental X-chromosome has been
97 inactivated in all cells (non-mosaic XCI, nmXCI) have been reported (8, 10-12). Organism-wide
98 nmXCI can arise from a constitutional genetic variant that directly interferes with inactivation
99 of a specific parental X-chromosome, such as mutations in the promoter region of the *XIST*
100 gene (10) and X-autosome translocations (13, 14). Disruption of the XIC locus or autosomal
101 regions involved in the choice of X-chromosome to be inactivated have similarly been
102 suggested to result in direct nmXCI, so called 'primary' skewing (15). Alternatively, selection
103 against any deleterious constitutive genetic variant would result in clonal expansion of cells
104 and indirectly result in nmXCI; 'secondary' skewing (**Fig. 1A**). As nmXCI females lack the
105 confounding effect of mosaicism, they represent a powerful system to study human XCI by
106 enabling direct determination of XCI status from bulk tissue samples. Indeed, using a single
107 nmXCI female identified in the Genotype-Tissue Expression (GTEx) database (16), Tukiainen
108 and colleagues generated the first comprehensive reference map of XCI status across human

109 tissues based on direct determination of allele-specific expression of X-linked genes (8).
110 However, as the ability to determine allele-specific expression is limited to the presence of an
111 informative (heterozygous) genetic variant in the gene of interest, the XCI status of just 186
112 genes was determined across tissues in this one nmXCI female.
113
114 In this study, we identify two additional unrelated nmXCI females in the GTEx database. By
115 combining allele-resolution data for the one previously reported nmXCI female and two newly
116 identified nmXCI females, we directly determined X-inactivation status of 380 X-linked genes
117 across 30 normal tissues, including 15 tissues in which XCI has not previously been
118 characterized. This unique dataset allowed investigation of tissue-specific escape from XCI in
119 humans and generation of the most extensive multi-tissue map of human X-inactivation to
120 date.
121
122 **Results**
123 *Non-mosaic X-inactivation is common in human females*
124 Whereas the frequency of nmXCI females in the general population was originally thought to
125 be less than 1:500 (11), more recent work has suggested a frequency of as high as 1:50 (12).
126 To identify nmXCI females we screened a single tissue of all 285 female donors (**Table S1**,
127 **Table S2**) in the GTEx database (v8 release) (16).Calculating the allelic expression (AE, how
128 much the ratio of Xa/Xi reads deviates from the expected 0.5) of all X-linked genes outside of
129 the PARs (i.e. the nonPAR AE) allows assessment of skewing in female tissues. Briefly, allelic
130 expression was determined by calling heterozygous SNPs on the X-chromosome using
131 genomic data (whole exome sequencing (WES) or whole genome sequencing (WGS)) and
132 subsequently counting the number of each base at every heterozygous SNP (hetSNP) position
133 in RNA-seq data. The allelic expression can then be calculated as, abs(0.5 - (reference reads
134 / total reads)), where a value of zero indicates complete biallelic expression and a value of 0.5
135 indicates complete monoallelic expression. We specifically exclude PAR genes from this

136 calculation as they are not subject to XCI, and consequently, their Xa/Xi read count ratio is not
137 significantly affected by XCI skewing. Three female samples showed an exceedingly high
138 degree of skewing (median chr X nonPAR allelic expression > 0.475 i.e. less than 2.5% of
139 reads coming from the 'inactive' allele), consistent with expression from a single parental
140 chromosome, nmXCI (**Fig. 1B**). Extending allelic expression analysis to all available tissues
141 across the three candidate females (**Table S3**) confirmed their complete nmXCI status (**Fig.**
142 **1C**). These three nmXCI individuals included the single previously identified nmXCI female,
143 UPIC (8), confirming the accuracy of our screening approach (**Fig. 1B, Fig. 1C**).
144
145 nmXCI may result from deleterious genetic mutations causing primary (10) or secondary
146 (acquired) skewing (15) or from stochastic non-random XCI (12). Whereas no small-scale
147 mutations in the XIC were identified in any of the nmXCI females, trisomy 17p was observed
148 in UPIC (**Fig. 1D, Supp Fig. 1B**). How trisomy 17p could result in non-random X-inactivation
149 is unclear, but an unbalanced 17p:X translocation would result in strong selection for silencing
150 of the abnormal X and consequential nmXCI (17). However, expression of genes on 17p was
151 consistently higher than those on 17q across all tissues in UPIC, suggesting the duplicated
152 17p was not silenced (**Supp Fig. 1C**). Indeed, further investigation of donor metadata in GTEx
153 revealed that UPIC had been diagnosed as mosaic for trisomy 17p by traditional karyotyping
154 but did not provide a detailed karyotype. Interestingly, whereas all three females were nmXCI,
155 females nmXCI-1 (GTEx ID: 13PLJ) and nmXCI-2 (GTEx ID: ZZPU) showed greater variation
156 in allelic expression across tissues than UPIC (**Fig. 1C)**, suggesting that their nmXCI may have
157 resulted from clonal selection driven by constitutive genetic variants (secondary skewing).
158
159 *A landscape of X-inactivation across normal human tissues*
160 The ability to detect allele-specific expression is dependent on the presence of a heterozygous
161 genetic variant (SNP) in the gene of interest. Using WES and WGS for both newly discovered
162 nmXCI donors we identified 389 heterozygous SNPs across 316 X-linked genes (count and

163 overlap between nmXCI-1 and nmXCI-2 of heterozygous SNPs and genes in **Fig. 2A,**
164 distribution of genes covered along the X-chromosome in **Fig. 2B, Table S4**). Allelic expression
165 ratios were highly consistent between nmXCI females and allowed us to identify known escape
166 (PAR and nonPAR), inactive and variable (8) XCI genes (**Fig. 2C**). Next, we applied our
167 analytical workflow to the previously identified nmXCI sample, UPIC, observing that the allelic
168 expression levels seen using our analysis pipeline largely matched the findings of Tukiainen
169 *et al* (**Fig. 2D, Supp Fig. 2A**) (8). Notably, our inclusion of additional WGS data of UPIC and
170 identification of 76 further SNPs allowed determination of allelic expression of 36 further genes
171 in this individual (**Fig. 2E**). The use of updated gene-models, correction of reference alignment
172 bias and more stringent quality filtering resulted in exclusion of 35 genes for which allelic
173 expression was erroneously reported (**Fig. 2E, Supp Fig. 2B**).
174
175 Next, we integrated data from all three nmXCI individuals (nmXCI-1, nmXCI-2 and UPIC),
176 allowing direct allelic expression detection of 380 X-linked genes across 30 tissues, including
177 15 tissues for which XCI status has not been directly assessed previously. While 195 genes
178 and 17 tissues could only be examined in a single individual, 185 genes and 13 tissues were
179 examined across multiple nmXCI females, allowing for selected inter-tissue and inter-individual
180 comparisons (**Fig. 3A, Fig. 3B**). Classification of allelic imbalance (mono-allelic | bi-allelic) of
181 individual alleles was initially determined by the binomial test ($P_{ADJ} < 0.01$), according to best
182 practice (18). However, upon visual inspection we identified several well characterised escape
183 and inactive genes that may have been mis-classified as variable escape (**Fig. 3C**).
184 Inappropriate classification of allelic imbalance results from the well documented sensitivity of
185 the binomial test to data over-dispersion, which frequently occurs with read count data (18,
186 19). Consequently, we further manually curated the allelic expression status of all 380 genes
187 using the empirical guideline of allelic ratio > 0.4 as indicating mono-allelic expression and the
188 potential consequences of high and low read counts. Our three criteria for manual re-
189 classification were (i) *low power*, indicating genes with a low statistical power but consistent

190 escape pattern, (ii) *low read count*, indicating genes with a consistent escape pattern but with
191 non-significant escape in a single tissue, and (iii) *over-estimation*, genes in which a high read
192 counts inflated the binomial p-value. (**Fig. 3C, Supp Fig. 3, Table S4, Table S5**). This manual
193 curation resulted in the re-classification of allelic expression status of 32 genes (**Table S5,**
194 **Supp Fig. 3**). Our classifications of XCI status based on direct determination of allele-specific
195 expression represent one of the most extensive and high-confidence maps of X-inactivation to
196 date (**Fig. 3D, Fig. 3E, Supp Fig. 4**). Allelic expression for all genes independent of XCI status
197 was highly correlated between individuals (**Supp Fig. 5A**), and by including three nmXCI
198 individuals in our analysis we were able to determine cross-tissue XCI status for 13 genes
199 which were only covered in one tissue in the previous assessment based on UPIC alone (**Fig.**
200 **4A**). Interestingly, whereas we reveal *EGFL6*, *TSPAN6* and *CXorf38* as variable escape genes,
201 these were previously classified (Balaton *et al*) as inactive, variable and escape, respectively.
202 Finally, we compared our classification with the classification of X-linked escape status from
203 donor UPIC as determined by Tukiainen and colleagues (4, 9, 12). As expected, our
204 classification was largely consistent with that of Tukiainen *et al*, but also revealed the possible
205 misclassification of XCI status of several genes and further added the XCI status for 198 genes
206 for which XCI status had not previously been directly assessed (**Fig. 4B**). Our direct
207 assessment of allelic expression further allowed for validation of XCI status previously
208 determined by indirect measures (3, 8). We confirmed XCI status for many previously classified
209 inactive and escape genes whereas most reported variable escape genes are re-classified as
210 inactive based on our analysis (**Supp Fig. 5B, Supp Fig 5C**). However, as the XCI
211 classifications reported here are largely based on inter-tissue comparisons within individuals,
212 most previous classifications were based on inter-individual comparisons of XCI, confounding
213 direct comparisons between studies.
214

215 **Discussion**

216 Here we describe an extensive multi-tissue map of human X-inactivation directly determined
217 from allele-specific expression of X-linked genes. This data represents a doubling of the
218 number of X-linked genes for which XCI status has been directly determined across multiple
219 normal tissues and individuals. Our results suggest that XCI is stable both within and across
220 individuals and that tissue-specific escape from X-inactivation is rare (11.6% of total genes (44
221 variable genes / 380 total covered X-linked genes)) and often challenging to characterize.
222 Variability in X-linked gene expression is important to characterize, as XCI variability can
223 directly modify expressivity of X-linked pathogenic variants. This provides a mechanism for
224 sex-biased expression differences in certain biological contexts and tissues, which in turn can
225 modify disease-risk.
226
227 nmXCI females allow for direct determination of allele-specific expression of X-linked genes,
228 even in bulk tissue samples. Our study further highlights the power of such natural genetic
229 systems to shed light on XCI in humans. Whereas the origin of nmXCI in females is typically
230 unknown, nmXCI females likely harbor one or more mutations that (i) directly affect the choice
231 of X-chromosome to be inactivated (primary skewing) or (ii) result in selection against cells
232 carrying an active mutated X-chromosome (secondary/acquired skewing). As such, the X-
233 linked genetics of nmXCI females is clearly atypical and could result in aberrant establishment
234 and maintenance of XCI in these females. Whereas the highly similar patterns of XCI observed
235 across the three genetically independent nmXCI females analyzed here suggests that such
236 affects are subtle, additional multi-tissue studies of XCI in mosaic females is required to confirm
237 the wider applicability of our findings. Finally, our identification of two additional nmXCI
238 females among the 285 females in the GTEx database further substantiates recent claims of
239 extreme skewing of XCI as a frequent and major modifier of trait penetrance and expressivity
240 in the human population (12, 20).
241
242

243 **Materials and Methods**

244 *Allele-specific expression analysis for initial identification of nmXCI females*

245 For the initial screen to identify nmXCI females in GTEx pre-processed SNPs were called from
246 WES data from GTEx (GTEx_Analysis_2017-06-
247 05_v8_WholeExomeSeq_979Indiv_VEP_annot.vcf.gz, see
248 https://gtexportal.org/home/methods for specifics on data processing). Downloaded data was
249 lifted over from hg19 to hg38 with LiftoverVcf (GATK v.4.1.9.0) and bcftools (v1.10.2) (21) was
250 utilized to include only hetSNPs on the X-chromosome. One RNA-seq sample per female was
251 downloaded from AnVIL using the Gen3 client, as aligned BAM files (see
252 https://gtexportal.org/home/methods for specifics on data processing). Tissues used for the
253 screen were selected based on abundance in the entire data set (muscle, n=239; skin-lleg,
254 n=21; thyroid, n=1; adipose-subc, n=2; artery-tibi, n=2; esophagus-muco, n=2; nerve, n=2;
255 adipose-visc, n=1; ovary, n=1; uterus, n=1; see **Table S7** for full list of abbreviations). Following
256 data download, allelic expression (AE) analysis of allele counts for hetSNPs was retrieved from
257 RNA-seq data using GATK ASEReadCounter (v.4.1.9.0), requiring a minimum base quality of
258 20 in the RNA-seq data. To remove potentially spurious sites, conservative filters were applied
259 to the data: variant call read depth in WES >= 20 per allele, minor allele read count > 10% of
260 the total read count, and RNA-seq read depth > 10 reads. If a gene carried multiple hetSNPs,
261 only the one with the highest RNA-seq read count was included. AE data from the screen can
262 be found in **Table S2** and **Table S3**. XCI skew of each individual participant was assessed by
263 calculating the single-tissue median AE of all nonPAR chr X genes not detected previously as
264 variable (8) (**Table S1**) in all 285 women in GTEx. Median chr X nonPAR AE of higher than
265 0.475 (equates to < 2.5% of reads coming from the inactive X-chromosome) indicates
266 expression from a single parental chromosome and was used as a cutoff to distinguish
267 between mosaic and non-mosaic XCI (nmXCI) females. As a control, UPIC was included. UPIC
268 is a female which has previously been shown to have nmXCI (8). As UPIC is not included in
269 GTEx_Analysis_2017-06-05_v8_WholeExomeSeq_979Indiv_VEP_annot.vcf.gz, SNP calling

270 was performed using WES data and AE analysis was performed, ultimately leveraging her AE
271 data as a positive control.
272
273 *Detailed analysis of allele-specific expression and XCI status in nmXCI females*
274 For further assessing XCI status of the three females with nmXCI (high median chr X nonPAR
275 AE), the remaining RNA-seq tissue samples available for all three donors were downloaded
276 from GTEx and processed as described above.
277 We further downloaded WES and whole-genome sequencing (WGS) data available for the
278 three nmXCI females, UPIC, 13PLJ (nmXCI-1) and ZZPU (nmXCI-2) and performed SNP
279 calling. Briefly, WES BAM and WGS CRAM files for each donor were downloaded from AnVIL
280 using the Gen3 client. The files were sorted and converted into fastq files using Samtools (v1.9)
281 (21). Raw fastq reads were quality trimmed using FastP (v.0.20.0) (22) with default settings.
282 BWA-MEM (23) was used for mapping reads to the human genome build 38 (hg38,
283 GCA_000001405.15_GRCh38_no_alt_analysis_set.fna) using default settings. Sambamba
284 v0.8.2 (24) was used to mark duplicate reads. The GATK best practice germline short variant
285 discovery pipeline was used to process the aligned reads, using base quality score
286 recalibration and local realignment at known insertions and deletions (indels) (GATK v.4.2.6.1).
287 Indels and SNPs were called jointly across all samples, for both WGS and WES data. Default
288 filters were applied to indel and SNP calls using the variant quality score recalibration (VQSR)
289 approach of GATK. All RNA-seq samples available for each participant (**Table S6**) were
290 downloaded from AnVIL as aligned BAM files using the Gen3 client. Samtools (v1.9) (21) was
291 used to sort and convert the BAM files to fastq files and quality trimming was done with FastP
292 (v.0.20.0) (22) with default settings. RNA-seq reads were aligned with STAR 2-pass mode
293 (v2.7.10a) (25) with GCA_000001405.15_GRCh38_no_alt_analysis_set.fna and --sjdbGTFfile
294 gencode.v40.annotation.gtf index. WASP filtering was performed to reduce reference bias
295 (26). Briefly, GATK SelectVariants was used to extract hetSNPs (that passed the VQSR
296 filtering) for each individual from either the WGS or WES VCFs, and subsequently passed to

297 STAR via --varVCFfile and–waspOutputMode. Reads failing WASP filtering were removed.
298 Following data pre-processing, AE analysis of the allele counts for hetSNPs was retrieved from
299 RNA-seq data using GATK ASEReadCounter. Heterozygous variants that passed VQSR
300 filtering were first extracted for each sample from WES and WGS VCFs using GATK
301 SelectVariants. Following, sample-specific VCFs and RNA-seq BAMs were input to
302 ASEReadCounter requiring minimum base quality of 20 in RNA-seq data. ASEReadCounter
303 outputs from WGS and WES pipelines were joined. If overlapping sites were identified, the
304 hetSNP read counts were merged and the AE counts from the assay with the highest total read
305 count was kept. To remove potentially spurious sites, conservative filters were applied to the
306 data (variant call read depth >= 10 per allele and minor allele read count > 10% of the total
307 variant read count for WES/WGS data, and an RNA-seq read depth of more than 7 reads). If
308 a gene carried multiple hetSNPs, the hetSNP detected in the highest number of tissues was
309 kept. If two hetSNPs were detected in the same number of tissues the one with the highest
310 RNA-seq read count was kept. The 'pituitary' sample from nmXCI-1 was excluded as it had a
311 lower median chr X nonPAR AE than all other tissues from the same individual. Furthermore,
312 the 'lymphoblasts' sample was excluded from donor UPIC since it is unclear how EBV-
313 transformation of lymphocytes affects the X-chromosome. The final ASE table for all X-linked
314 genes in the nmXCI females can be found in **Table S4**.
315
316 *X-inactivation status categorization*
317 Tissue-specific X-inactivation status categorization was performed as previously described (8).
318 Briefly, XCI status of genes was assessed by comparing the allelic read count ratios at each
319 filtered X-chromosomal hetSNP, in each tissue individually. Whether there was significant
320 expression from the inactive X-chromosome (escape) at each hetSNP was tested with a one-
321 sided binomial test where the reads from Xi compared to the total read count was expected to
322 be significantly greater than 0.025 or 2.5% (hypothesized probability of success = 0.025). FDR
323 correction was applied to p values from the binomial test for each of tissues separately using

324 the rstatix R package version 0.7.2 (Kassambara A. 2023). q values < 0.01 were considered
325 significant and indicative of XCI escape for the hetSNP and tissue.
326 To assess across-tissue XCI status of genes, we leveraged the q values obtained above. If a
327 given hetSNP showed significant Xi expression in all tissues in which the gene is expressed,
328 the SNP was categorized as 'escape (across tissues)', a hetSNP with non-significant Xi
329 expression across all tissues in which the gene is expressed was classified as 'inactive (across
330 tissues)'. If a gene was significant in more than one but not in all tissues in which it is expressed
331 the gene was classified as 'variable'. Genes for which we had AE data in only one tissue were
332 either labeled "inactive (data for one single tissue)" or "escape (data for one single tissue)"
333 based on the binomial test. Finally, genes residing in the pseudo-autosomal region (PAR) were
334 labeled 'PAR'. Upon inspection of the classification results several well characterized escape
335 genes and inactive genes were mis-classified as variable escape. Consequently, we manually
336 curated allelic expression status of all 380 genes using the empirical guideline of allelic ratio
337 >0.4 as indicating mono-allelic expression while also considering the consequences of high
338 and low read counts. Our three criteria for manual re-classification were low read count, low
339 power and over-estimation, indicating genes with a low statistical power but consistent escape
340 pattern (low power), consistent escape pattern but with non-significant escape in single tissue
341 (low read count) and over-estimation due to high read counts inflating the binomial p-value
342 (over-estimation). The list of genes that were manually curated, their change in XCI status and
343 the criteria for their manual classification can be found in **Supp Fig. 3** and **Table S5**.
344
345 *Copy number analysis*
346 Copy number analysis was performed using the R-package, QDNASeq (27). Briefly, whole
347 genome sequencing data for all three nmXCI females was mapped to the human reference
348 genome (hg38) by BWA-MEM (23) using default parameters. Raw copy numbers were then
349 estimated by counting the number of reads mapped to non-overlapping bins of 15 kb unless
350 stated otherwise.

351

352 *Investigating chromosome 17p and 17q arm expression*

353 Transcriptome mapping was performed using Salmon v1.9.0 (--gcBias --seqBias --
354 numBootstraps 100) (28) using known Refseq transcripts (NM_* and NR_*) from the
355 GRCh38.p12 assembly (GCF_000001405.38_GRCh38.p12_rna.fna.gz) as a reference.
356 Library type options were IU. Raw fastq reads were quality trimmed using FastP (v.0.20.0) with
357 default settings. Abundance estimates were converted to h5 format using Wasabi
358 (https://github.com/COMBINE-lab/wasabi). Sleuth was used for data normalization and
359 normalized gene abundances (29, 30). Expression matrix (TPM) can be found in **Table S8**.

360

361

362 **References**

363 1. O. Rosspopoff *et al.*, Species-specific regulation of XIST by the JPX/FTX orthologs. *Nucleic*
364 *Acids Res* **51**, 2177-2194 (2023).
365 2. C. Patrat, J. F. Ouimette, C. Rougeulle, X chromosome inactivation in human development.
366 *Development* **147** (2020).
367 3. B. P. Balaton, A. M. Cotton, C. J. Brown, Derivation of consensus inactivation status for X-
368 linked genes from genome-wide studies. *Biol Sex Differ* **6**, 35 (2015).
369 4. L. Carrel, H. F. Willard, X-inactivation profile reveals extensive variability in X-linked gene
370 expression in females. *Nature* **434**, 400-404 (2005).
371 5. M. Souyris *et al.*, TLR7 escapes X chromosome inactivation in immune cells. *Sci Immunol* **3**
372 (2018).
373 6. M. J. Mousavi, M. Mahmoudi, S. Ghotloo, Escape from X chromosome inactivation and
374 female bias of autoimmune diseases. *Mol Med* **26**, 127 (2020).
375 7. A. Helena Mangs, B. J. Morris, The Human Pseudoautosomal Region (PAR): Origin, Function
376 and Future. *Curr Genomics* **8**, 129-136 (2007).
377 8. T. Tukiainen *et al.*, Landscape of X chromosome inactivation across human tissues. *Nature*
378 **550**, 244-248 (2017).
379 9. A. M. Cotton *et al.*, Analysis of expressed SNPs identifies variable extents of expression from
380 the human inactive X chromosome. *Genome Biol* **14**, R122 (2013).
381 10. R. M. Plenge *et al.*, A promoter mutation in the XIST gene in two unrelated families with
382 skewed X-chromosome inactivation. *Nat Genet* **17**, 353-356 (1997).
383 11. J. M. Amos-Landgraf *et al.*, X chromosome-inactivation patterns of 1,005 phenotypically
384 unaffected females. *Am J Hum Genet* **79**, 493-499 (2006).
385 12. E. Shvetsova *et al.*, Skewed X-inactivation is common in the general female population. *Eur J*
386 *Hum Genet* **27**, 455-465 (2019).
387 13. B. P. Favilla *et al.*, Spread of X-chromosome inactivation into autosomal regions in patients
388 with unbalanced X-autosome translocations and its phenotypic effects. *Am J Med Genet A*
389 **185**, 2295-2305 (2021).

14. E. Hatchwell, D. Robinson, J. A. Crolla, A. E. Cockwell, X inactivation analysis in a female with hypomelanosis of Ito associated with a balanced X;17 translocation: evidence for functional disomy of Xp. *J Med Genet* **33**, 216-220 (1996).
15. P. Clerc, P. Avner, Random X-chromosome inactivation: skewing lessons for mice and men. *Curr Opin Genet Dev* **16**, 246-253 (2006).
16. G. T. Consortium, The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-585 (2013).
17. P. H. King, R. Waldrop, J. R. Lupski, L. G. Shaffer, Charcot-Marie-Tooth phenotype produced by a duplicated PMP22 gene as part of a 17p trisomy-translocation to the X chromosome. *Clin Genet* **54**, 413-416 (1998).
18. S. E. Castel, A. Levy-Moonshine, P. Mohammadi, E. Banks, T. Lappalainen, Tools and best practices for data processing in allelic expression analysis. *Genome Biol* **16**, 195 (2015).
19. J. Chen *et al.*, A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. *Nat Commun* **7**, 11101 (2016).
20. A. L. Roberts *et al.*, Age acquired skewed X chromosome inactivation is associated with adverse health outcomes in humans. *Elife* **11** (2022).
21. P. Danecek *et al.*, Twelve years of SAMtools and BCFtools. *Gigascience* **10** (2021).
22. S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884-i890 (2018).
23. H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595 (2010).
24. A. Tarasov, A. J. Vilella, E. Cuppen, I. J. Nijman, P. Prins, Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032-2034 (2015).
25. A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
26. B. van de Geijn, G. McVicker, Y. Gilad, J. K. Pritchard, WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat Methods* **12**, 1061-1063 (2015).
27. I. Scheinin *et al.*, DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. *Genome Res* **24**, 2022-2032 (2014).
28. R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, C. Kingsford, Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417-419 (2017).
29. H. Pimentel, N. L. Bray, S. Puente, P. Melsted, L. Pachter, Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat Methods* **14**, 687-690 (2017).
30. L. Yi, H. Pimentel, N. L. Bray, L. Pachter, Gene-level differential analysis at transcript-level resolution. *Genome Biol* **19**, 53 (2018).

435 **Data and code availability**

436 All data and analysis scripts are included in this manuscript or are available on GitHub
437 https://github.com/ColmNestorLab/tissue_XCI.

438

439 **Author contributions**

440 C.E.N and B.G designed research; B.G. performed research; C.E.N., B.G. and M.B. wrote the
441 paper; M.B. provided insights and discussions.

442

443 **Declarations of interest**

444 The authors declare no competing interests.

445

446 **Funding sources**

447 The project was funded by the Swedish Research Council (2020-01277_VR) and the
448 European Research Council (XX-Health-101045171).

449

450 **Acknowledgements**

451 We thank Shadi Jafari and Sandra Hellberg for helpful discussions and input on the
452 manuscript. Work in the lab of Colm Nestor was funded by the Swedish Research Council
453 (2020-01277_VR) and the European Research Council (XX-Health-101045171).

454

455

456

457

458

459

460

461

462 **Figure legends**

463 **Fig. 1. Identification of females with non-mosaic X-inactivation.** (**A**) The patterns of X-
464 chromosome inactivation (XCI) in women resulting in mosaic (right female) or non-mosaic XCI
465 (nmXCI). The presence of genetic variants can result in nmXCI females by (i) directly
466 determining which X-chromosome can be inactivated (primary skewing, left female) or by (ii)
467 imparting a selective advantage to a small number of cells (secondary skewing, middle
468 female). Xa, active chr X; Xi, inactive chr X, Xm, maternal chr X, Xp, paternal chr X. (**B**) Single-
469 tissue median allelic expression (AE) and standard error of all nonPAR genes on chromosome
470 X (chr X) not previously classified as variable in all 285 women in GTEx. (**C)** Allelic expression
471 per tissue of nonPAR chr X genes not previously classified as variable in mosaic females
472 (median allelic expression < 0.475) and three females identified as non-mosaic, nmXCI-1,
473 nmXCI-2 and UPIC (median allelic expression > 0.475). Boxplot indicating median, 25th and
474 75th percentile. (**D**) Copy number as log2 ratio of chromosome 17 (chr 17) for nmXCI females,
475 UPIC, nmXCI-1 and nmXCI-2. Trisomy 17p in UPIC is highlighted.

476

477 **Fig. 2. Characterization of two novel human nmXCI females. (A)** Overlap of genic
478 heterozygous SNPs (hetSNPs) (upper) and genes with hetSNP (lower) across the two novel
479 nmXCI females (nmXCI-1 and nmXCI-2). Genic hetSNPs were identified using both WES and
480 WGS for each individual. (**B)** Distribution of assessed genes across the X-chromosome. Genes
481 located in the pseudoautosomal region 1 (PAR1) and PAR2 are highlighted in green. **(C)** Allelic
482 expression per tissue for well-characterised PAR, (*AKAP17A*), escape (*PUDP, DDX3X*),
483 inactive (*APOOL*) and variable (*PRKX*) genes (8). Boxplot indicating median, 25th and 75th
484 percentile. **(D)** Spearman correlation of allelic expression values using our analysis approach
485 (Gylemo) and the Tukiainen *et al* analysis pipelines for female UPIC. **(E)** Overlap of genic
486 hetSNPs (left) and genes (right) identified by our analysis (Gylemo) and the Tukiainen *et al*
487 analysis pipeline in female UPIC.

488

489 **Fig. 3. An extended landscape of X-inactivation in humans. (A)** Overlap of genic
490 heterozygous SNPs (hetSNPs) (upper) and genes with hetSNP (lower) across the three
491 females (nmXCI-1, nmXCI-2 and UPIC) with non-mosaic X-inactivation (nmXCI). **(B)** Tissues
492 covered in each nmXCI female. Tissues not covered in UPIC are indicated in bold. **(C)**
493 Examples of genes classified by the binomial test alone and after manual curation. Allelic
494 expression across tissues is shown with XCI status based on the binomial test indicated as
495 inactive (grey circle) or escape (red triangle). **(D)** Allelic expression of X-linked gene categories
496 (lower) and number of genes included in each category (upper). Genes classified as escape
497 and inactive are separated based on whether allelic expression was determined across
498 multiple tissues or in a single sample (data for one single tissue). Boxplot indicating median,
499 25th and 75th percentile. **(E)** Heatmap showing the allelic expression of all genes that show
500 constitutive or variable escape from XCI. Black asterisks within the tiles indicate a significant
501 expression from the inactive X-chromosome (i.e. escape, FDR-corrected binomial q-value <
502 0.01). The 'consensus call' tile is the assigned XCI status across tissues and individuals for
503 each gene with genes classified as variable including both intra- and interindividual variation.
504 Red asterisks indicate genes in which manual curation of XCI status was performed. Grey tiles
505 indicate missing data. **(B, E)** Tissue abbreviations can be found in **Table S7.**
506
507 **Fig. 4. Classification and novel XCI assessment of X-linked genes. (A)** Allelic expression
508 across all available tissues in all three nmXCI females for genes which were only covered in
509 one tissue in the previous assessment based on UPIC alone (8) **(B)** Alluvial plot showing
510 classification of escape status of X-linked genes based on our analysis (Gylemo) compared to
511 a previous assessment based on UPIC alone (Tukiainen) (8). Genes classified as escape and
512 inactive are separated based on whether allelic expression was determined across multiple
513 tissues or in only one sample (data for one single tissue).
514
515

516 **Supplementary figure legends**

517 **Supp Fig. 1. (A)** Expression, *cis*-binding and spreading of the long non-coding RNA *XIST*
518 initiates the epigenetic process of remodelling an active X-chromosome (chr X) into an inactive
519 X-chromosome (Xi). After initiation of X-inactivation activating histone marks (histone 3 lysine
520 27 acetylation; H3K27ac) are removed while silencing histone modifications (trimethylation of
521 lysine 27 on histone 3; H3K27me3 and ubiquitination of lysine 119 on histone 2A; H2AK119Ub)
522 and DNA methylation are deposited on the future inactive X chromosome. This process leads
523 to transcriptional silencing of all genes on the Xi except for genes residing in the
524 pseudoautosomal regions (PARs, green) as well as a subset of non-PAR genes that 'escape'
525 X-inactivation (red) and are continually expressed from the Xi. **(B)** Copy number as log2 ratio
526 using 500kb bins of the whole genome for nmXCI females, UPIC, nmXCI-1 and nmXCI-2.
527 Trisomy 17p in UPIC is highlighted. **(C)** Gene expression from 17p and 17q across all tissues
528 in nmXCI-2 and UPIC. TPM: transcriptions per million. 17p: chromosome 17, p arm. 17q:
529 chromosome 17, q arm. Tissue abbreviations can be found in **Table S7**.

530

531 **Supp Fig. 2. (A)** Spearman correlation of allelic expression values using our analysis approach
532 (Gylemo) and Tukiainen *et al* analysis pipelines for female UPIC in all tissues available. **(B)**
533 List of genes that were included in Tukiainen's analysis of allelic expression in UPIC but were
534 excluded from our analysis including reason for exclusion.

535

536 **Supp Fig. 3.** Read counts for minor and major alleles for all genes where the XCI status was
537 manually curated. The manual curation result and the reason for manual curation is stated
538 above each gene and summarized in the table in the bottom right of the figure. Asterisks
539 indicate significant escape based on the binomial test (FDR-corrected binomial q value < 0.01).

540

541 **Supp Fig. 4.** Heatmap showing the allelic expression of all X-linked genes assayed in this
542 study. Black asterisks within the tiles indicate a significant inactive X-chromosome expression

543 (i.e. escape, FDR-corrected binomial q value < 0.01). The 'consensus call' tile is the assigned
544 XCI status across tissues for each gene with genes classified as variable including both intra-
545 and interindividual variation. Red asterisks before gene names indicate genes in which manual
546 curation was performed to assign XCI status. Grey tiles indicate missing data.
547
548 **Supp Fig. 5. (A)** Spearman correlation of allelic expression between nmXCI females. **(B, C)**
549 Alluvial plots showing direct classification of escape status based on our analysis (Gylemo)
550 compared to classification of escape status of X-linked genes as consensus calls for X-
551 inactivation based on data from multiple previous studies employing indirect measures to
552 determine escape; reported by **(B)** Tukiainen *et al* (8) (in their Suppl. Table.1, Combined XCI
553 status) and **(C)** Balaton *et al* (in their Table S1, Balaton consensus calls) (3).
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569

570 **Supplementary table legends**

571

572 **Table S0:** Text file describing each Supplementary Tables in detail.

573

574 **Table S1:** Genes used in screening to identify non-mosaic XCI (nmXCI) females. Compiled
575 from Supplementary Table 13 in Tukiainen et al, removing genes with a variable escape status
576 in UPIC (XCI across tissues != "Partial, heterogeneous") and genes residing in the
577 pseudoautosomal region (Region != "PAR").

578

579 **Table S2:** Allelic expression of the genes selected in Table S1 for all 285 females analyzed
580 from GTEx.

581

582 **Table S3:** Allelic expression of the genes selected in Table S1 for all tissues available for the
583 nmXCI females: UPIC, nmXCI-1 (13PLJ) and nmXCI-2 (ZZPU).

584

585 **Table S4:** Allelic expression data of 380 genes on the X-chromosome for each of the nmXCI
586 females UPIC: nmXCI-1 (13PLJ) and nmXCI-2 (ZZPU).

587

588 **Table S5:** List of genes which were manually curated, the directionality of the manual curation,
589 the reasoning behind the manual curation.

590

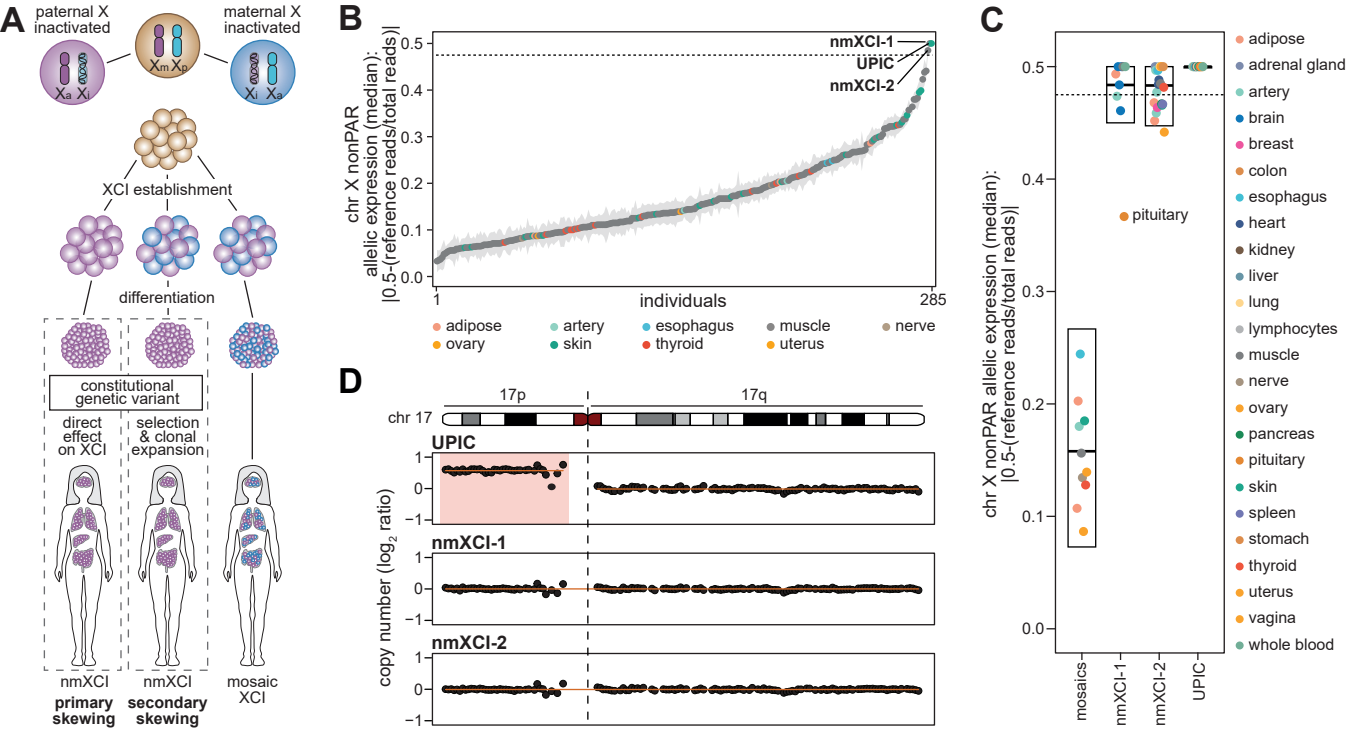
591 **Table S6:** Sample- and tissue-IDs for all data analyzed for nmXCI females: UPIC, nmXCI-1
592 (13PLJ) and nmXCI-2 (ZZPU).

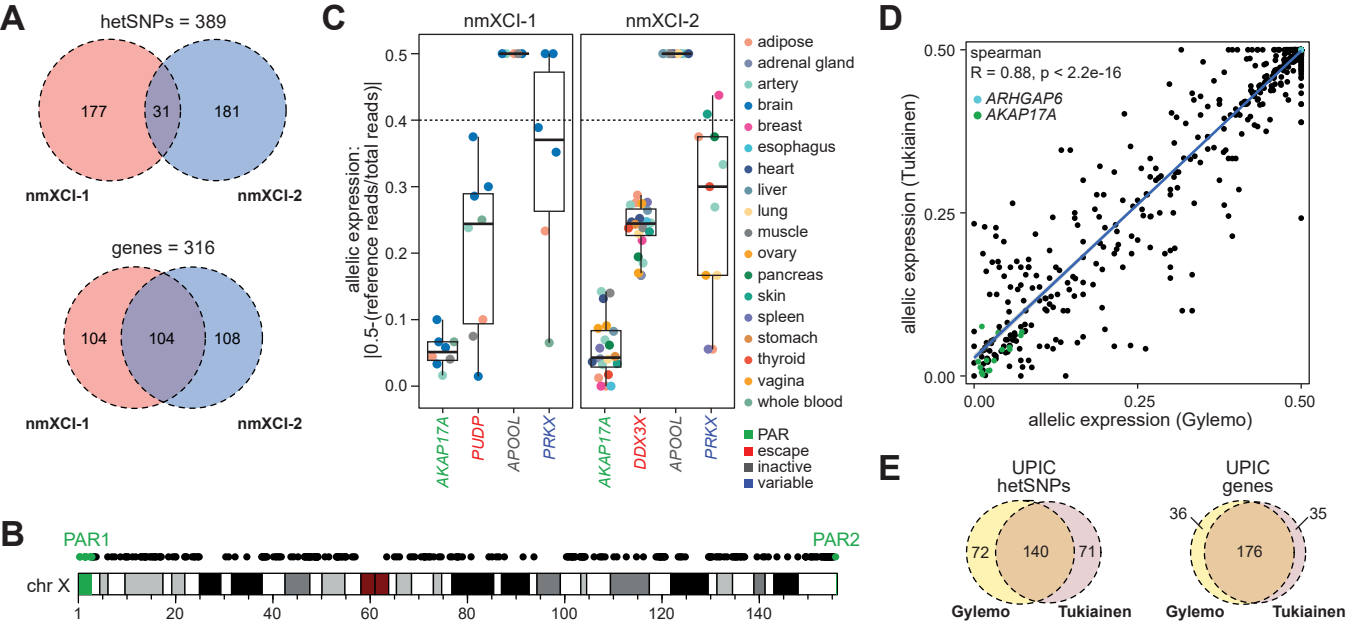
593

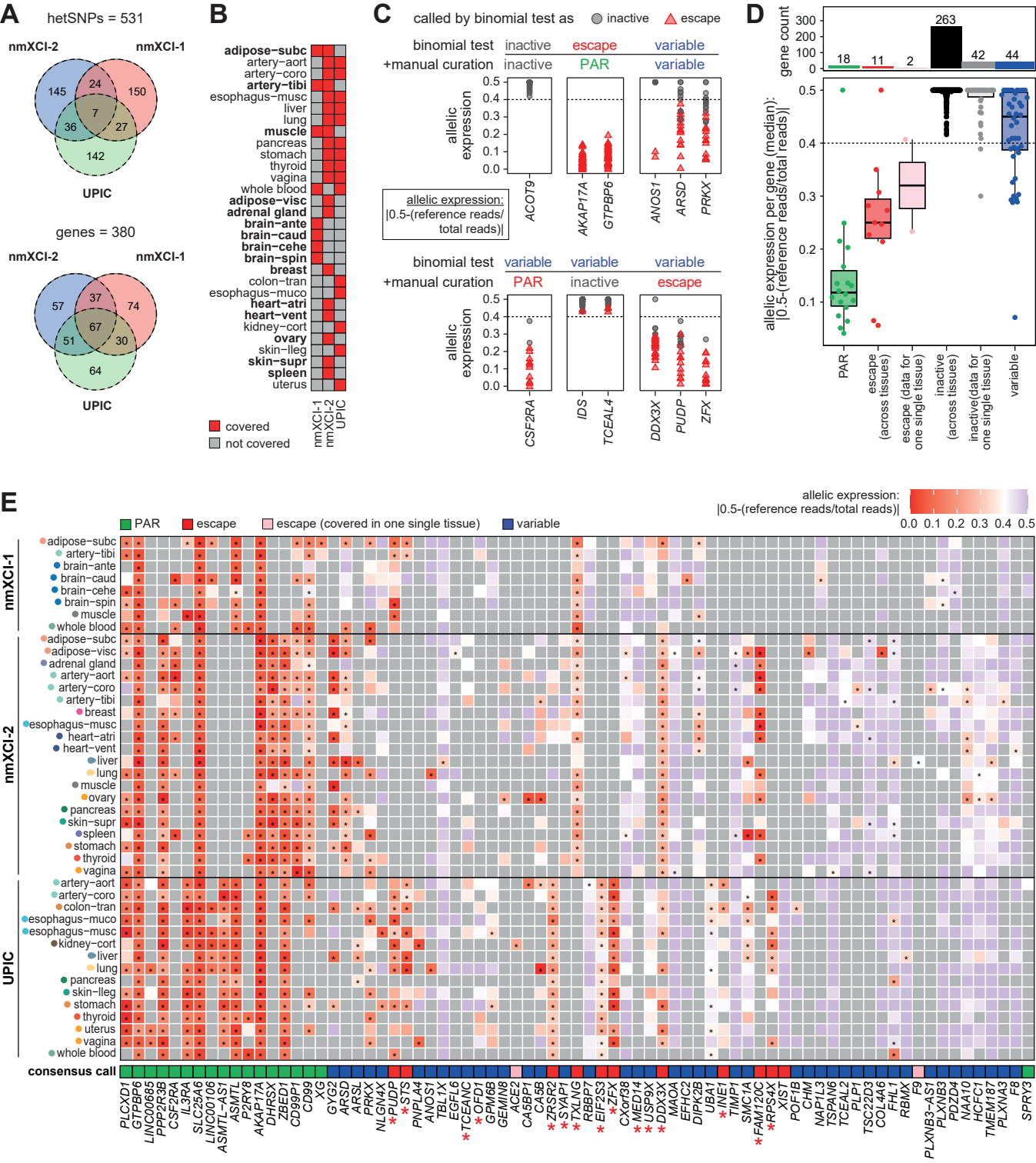
594 **Table S7:** Tissue abbreviations.

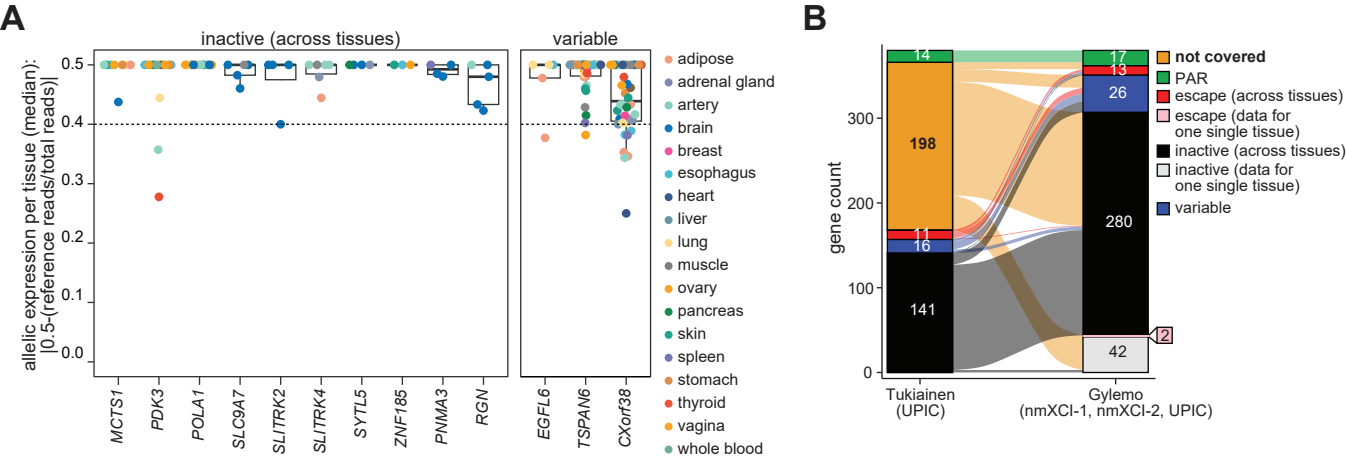
595

596 **Table S8:** Gene expression (TPM) matrix of all tissues for nmXCI females: UPIC, nmXCI-1
597 (13PLJ) and nmXCI-2 (ZZPU).


A
paternal X inactivated
maternal X inactivated
$X_m X_p$
$X_a X_i$
$X_i X_a$
XCI establishment
differentiation
constitutional genetic variant
direct effect on XCI
selection & clonal expansion
nmXCI
primary skewing
nmXCI
secondary skewing
mosaic XCI
B
chr X nonPAR allelic expression (median): |0.5-(reference reads/total reads)|
nmXCI-1
UPIC
nmXCI-2
individuals
1
285
adipose
ovary
artery
skin
esophagus
thyroid
muscle
uterus
nerve
C
chr X nonPAR allelic expression (median): |0.5-(reference reads/total reads)|
pituitary
mosaics
nmXCI-1
nmXCI-2
UPIC
adipose
adrenal gland
artery
brain
breast
colon
esophagus
heart
kidney
liver
lung
lymphocytes
muscle
nerve
ovary
pancreas
pituitary
skin
spleen
stomach
thyroid
uterus
vagina
whole blood
D
17p
17q
chr 17
UPIC
nmXCI-1
nmXCI-2
copy number (log_2 ratio)


A
hetSNPs = 389
177
31
181
nmXCI-1
nmXCI-2
genes = 316
104
104
108
nmXCI-1
nmXCI-2
C
nmXCI-1
nmXCI-2
allelic expression:
|0.5-(reference reads/total reads)|
0.5
0.4
0.3
0.2
0.1
0.0
AKAP17A
PUDP
APOOL
PRKX
AKAP17A
DDX3X
APOOL
PRKX
adipose
adrenal gland
artery
brain
breast
esophagus
heart
liver
lung
muscle
ovary
pancreas
skin
spleen
stomach
thyroid
vagina
whole blood
PAR
escape
inactive
variable
D
spearman
R = 0.88, p < 2.2e-16
ARHGAP6
AKAP17A
allelic expression (Tukiainen)
allelic expression (Gylemo)
0.00
0.25
0.50
B
PAR1
PAR2
chr X
1
20
40
60
80
100
120
140
E
UPIC
hetSNPs
72
140
71
Gylemo
Tukiainen
UPIC
genes
36
176
35
Gylemo
Tukiainen


A
hetSNPs = 531
nmXCI-2
nmXCI-1
145
24
150
7
36
27
142
UPIC
genes = 380
nmXCI-2
nmXCI-1
57
37
74
67
51
30
64
UPIC
B
adipose−subc
artery−aort
artery−coro
artery−tibi
esophagus−musc
liver
lung
muscle
pancreas
stomach
thyroid
vagina
whole blood
adipose−visc
adrenal gland
brain−ante
brain−caud
brain−cehe
brain−spin
breast
colon−tran
esophagus−muco
heart−atri
heart−vent
kidney−cort
ovary
skin−lleg
skin−supr
spleen
uterus
nmXCI-1
nmXCI-2
UPIC
covered
not covered
C
called by binomial test as
inactive
escape
binomial test
inactive
escape
variable
+manual curation
inactive
PAR
variable
allelic expression
allelic expression:
|0.5-(reference reads/
total reads)|
ACOT9
AKAP17A
GTPBP6
ANOS1
ARSD
PRKX
binomial test
variable
variable
variable
+manual curation
PAR
inactive
escape
CSF2RA
IDS
TCEAL4
DDX3X
PUDP
ZFX
D
gene count
18
11
2
263
42
44
allelic expression per gene (median):
|0.5-(reference reads/total reads)|
PAR
escape
(across tissues)
escape (data for
one single tissue)
inactive
(across tissues)
inactive(data for
one single tissue)
variable
E
allelic expression:
|0.5-(reference reads/total reads)|
PAR
escape
escape (covered in one single tissue)
variable
nmXCI-1
nmXCI-2
UPIC
consensus call


A
inactive (across tissues)
variable
allelic expression per tissue (median):
|0.5-(reference reads/total reads)|
0.5
0.4
0.3
0.2
0.1
0.0
MCTS1
PDK3
POLA1
SLC9A7
SLITRK2
SLITRK4
SYTL5
ZNF185
PNMA3
RGN
EGFL6
TSPAN6
CXorf38
adipose
adrenal gland
artery
brain
breast
esophagus
heart
liver
lung
muscle
ovary
pancreas
skin
spleen
stomach
thyroid
vagina
whole blood
B
gene count
300
200
100
0
14
198
11
16
141
17
13
26
280
42
2
not covered
PAR
escape (across tissues)
escape (data for one single tissue)
inactive (across tissues)
inactive (data for one single tissue)
variable
Tukiainen (UPIC)
Gylemo (nmXCI-1, nmXCI-2, UPIC)