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 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.

55 **Introduction**

56 During embryonic development in humans, female cells (46, XX) inactivate a single X-
57 chromosome to balance dosage of X-linked gene expression between females (XX) and males
58 (XY). The process of X-chromosome inactivation (XCI) begins during the peri-implantation
59 stage in humans and is initiated by expression of the long non-coding RNA (lncRNA) *XIST*
60 from the X inactivation center (XIC) of one X-chromosome in each female cell. *XIST RNA* coats
61 the X-chromosome which in turn recruits protein complexes including chromatin modifiers
62 which establish a facultative heterochromatin state. Silencing histone modifications and the
63 deposition of DNA methylation ultimately result in the transcriptionally silent inactive X-
64 chromosome (Xi) (**Supp Fig. 1A**). The selective expression of *XIST* from only one X-
65 chromosome needs to be tightly regulated. Whereas the exact mechanism by which a given
66 X-chromosome in a female cell is selected for inactivation is unknown, the XIC in humans
67 includes several lncRNAs that have been implicated in the regulation of *XIST* expression and
68 X-inactivation (**Supp Fig. 1A**) (1). Although the initial choice of X-chromosome for inactivation
69 is random, the inactivated X-chromosome is stably inherited in a clonal fashion throughout all
70 subsequent cell divisions (2). Interestingly, the process of X-inactivation (XCI) is incomplete;
71 approximately 15% of genes escape XCI and remain expressed from both X-chromosomes
72 (3). These 'escape' genes are consequently more highly expressed in females than in males,
73 although the degree of sex-biased expression of escape genes can vary and may be greater
74 for those genes that lack a functional homologue on the Y-chromosome (4). The sex-biased
75 expression of escape genes has been suggested to mediate sex differences in both the
76 frequency and severity of several diseases (5, 6). In addition, genes in the pseudoautosomal
77 regions (PAR1 and PAR2) of the X-chromosome also remain expressed from the Xi and are
78 thus always bi-allelically expressed. PAR1 and PAR2 are short regions of homology between
79 the X- and Y-chromosomes that are essential for pairing of the sex chromosomes during
80 meiosis in males (7). In contrast to escape genes, which tend to be more highly expressed in
81 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 an extensive multi-tissue map of human X-inactivation to date.
120
121 **Results**
122 *Non-mosaic X-inactivation is common in human females*
123 Whereas the frequency of nmXCI females in the general population was originally thought to
124 be less than 1:500 (11), more recent work has suggested a frequency of as high as 1:50 (12).
125 To identify nmXCI females we screened a single tissue of all 285 female donors (**Table S1**,
126 **Table S2**) in the GTEx database (v8 release) (16). Briefly, calculating the allelic expression
127 (AE, how much the ratio of Xa/Xi reads deviates from the expected 0.5) of all X-linked genes
128 outside of the PARs (i.e. the nonPAR AE) allows assessment of skewing in female tissues. We
129 specifically exclude PAR genes from this calculation as they are not subject to XCI, and
130 consequently, their Xa/Xi read count ratio is not significantly affected by XCI skewing. Three
131 female samples showed an exceedingly high degree of skewing (median chr X nonPAR allelic
132 expression > 0.475), consistent with expression from a single parental chromosome, nmXCI
133 (**Fig. 1B**). Extending allelic expression analysis to all available tissues across the three
134 candidate females (**Table S3**) confirmed their complete nmXCI status (**Fig. 1C**). These three

135 nmXCI individuals included the single previously identified nmXCI female, UPIC (8), confirming
136 the accuracy of our screening approach (**Fig. 1B, Fig. 1C**).
137
138 nmXCI may result from deleterious genetic mutations causing primary (10) or secondary
139 (acquired) skewing (15) or from stochastic non-random XCI (12). Whereas no small-scale
140 mutations in the XIC were identified in any of the nmXCI females, trisomy 17p was observed
141 in UPIC (**Fig. 1D, Supp Fig. 1B**). How trisomy 17p could result in non-random X-inactivation
142 is unclear, but an unbalanced 17p:X translocation would result in strong selection for silencing
143 of the abnormal X and consequential nmXCI (17). However, expression of genes on 17p was
144 consistently higher than those on 17q across all tissues in UPIC, suggesting the duplicated
145 17p was not silenced (**Supp Fig. 1C**). Indeed, further investigation of donor metadata in GTEx
146 revealed that UPIC had been diagnosed as mosaic for trisomy 17p by traditional karyotyping
147 but did not provide a detailed karyotype. Interestingly, whereas all three females were nmXCI,
148 females nmXCI-1 (GTEx ID: 13PLJ) and nmXCI-2 (GTEx ID: ZZPU) showed greater variation
149 in allelic expression across tissues than UPIC (**Fig. 1C)**, suggesting that their nmXCI may have
150 resulted from clonal selection driven by constitutive genetic variants (secondary skewing).
151
152 *A landscape of X-inactivation across normal human tissues*
153 The ability to detect allele-specific expression is dependent on the presence of a heterozygous
154 genetic variant (SNP) in the gene of interest. Using whole exome sequencing (WES) and whole
155 genome sequencing (WGS) for both newly discovered nmXCI donors we identified 389
156 heterozygous SNPs across 316 X-linked genes (**Fig. 2A, Fig. 2B, Table S4**). Briefly, allelic
157 expression was determined by calling heterozygous SNPs on the X-chromosome using
158 genomic data (WES or WGS) and subsequently counting the number of each base at every
159 heterozygous SNP (hetSNP) position in RNA-seq data. The allelic expression can then be
160 calculated abs(0.5 - (reference reads / total reads)), where a value of zero indicates complete
161 biallelic expression and a value of 0.5 indicates complete monoallelic expression. Allelic

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

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

212 **Discussion**

213 Here we describe an extensive multi-tissue map of human X-inactivation directly determined
214 from allele-specific expression of X-linked genes. This data represents a doubling of the
215 number of X-linked genes for which XCI status has been directly determined across multiple

216 normal tissues and individuals. Our results suggest that XCI is stable both within and across
217 individuals and that tissue-specific escape from X-inactivation is rare and often challenging to
218 characterize. Variability in X-linked gene expression is important to characterize, as XCI
219 variability can directly modify expressivity of X-linked pathogenic variants. This provides a
220 mechanism for sex-biased expression differences in certain biological contexts and tissues,
221 which in turn can modify disease-risk.
222
223 nmXCI females allow for direct determination of allele-specific expression of X-linked genes,
224 even in bulk tissue samples. Our study further highlights the power of such natural genetic
225 systems to shed light on XCI in humans. Whereas the origin of nmXCI in females is typically
226 unknown, nmXCI females likely harbor one or more mutations that (i) directly affect the choice
227 of X-chromosome to be inactivated (primary skewing) or (ii) result in selection against cells
228 carrying an active mutated X-chromosome (secondary/acquired skewing). As such, the X-
229 linked genetics of nmXCI females is clearly atypical and could result in aberrant establishment
230 and maintenance of XCI in these females. Whereas the highly similar patterns of XCI observed
231 across the three genetically independent nmXCI females analyzed here suggests that such
232 affects are subtle, additional multi-tissue studies of XCI in mosaic females is required to confirm
233 the wider applicability of our findings. Finally, our identification of two additional nmXCI
234 females among the 285 females in the GTEx database further substantiates recent claims of
235 extreme skewing of XCI as a frequent and major modifier of trait penetrance and expressivity
236 in the human population (12, 20).
237
238 **Materials and Methods**
239 *Allele-specific expression analysis for initial identification of nmXCI females*
240 For the initial screen to identify nmXCI females in GTEx pre-processed SNPs were called from
241 WES data from GTEx (GTEx_Analysis_2017-06-
242 05_v8_WholeExomeSeq_979Indiv_VEP_annot.vcf.gz, see

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

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

296 SelectVariants. Following, sample-specific VCFs and RNA-seq BAMs were input to
297 ASEReadCounter requiring minimum base quality of 20 in RNA-seq data. ASEReadCounter
298 outputs from WGS and WES pipelines were joined. If overlapping sites were identified, the
299 hetSNP read counts were merged and the AE counts from the assay with the highest total read
300 count was kept. To remove potentially spurious sites, conservative filters were applied to the
301 data (variant call read depth >= 10 per allele and minor allele read count > 10% of the total
302 variant read count for WES/WGS data, and an RNA-seq read depth of more than 7 reads). If
303 a gene carried multiple hetSNPs, the hetSNP detected in the highest number of tissues was
304 kept. If two hetSNPs were detected in the same number of tissues the one with the highest
305 RNA-seq read count was kept. The 'pituitary' sample from nmXCI-1 was excluded as it had a
306 lower median chr X nonPAR AE than all other tissues from the same individual. Furthermore,
307 the 'lymphoblasts' sample was excluded from donor UPIC since it is unclear how EBV-
308 transformation of lymphocytes affects the X-chromosome. The final ASE table for all X-linked
309 genes in the nmXCI females can be found in **Table S4**.
310
311 *X-inactivation status categorization*
312 Tissue-specific X-inactivation status categorization was performed as previously described (8).
313 Briefly, XCI status of genes was assessed by comparing the allelic read count ratios at each
314 filtered X-chromosomal hetSNP, in each tissue individually. Whether there was significant
315 expression from the inactive X-chromosome (escape) at each hetSNP was tested with a one-
316 sided binomial test where the reads from Xi compared to the total read count was expected to
317 be significantly greater than 0.025 or 2.5% (hypothesized probability of success = 0.025). FDR
318 correction was applied to p values from the binomial test for each of tissues separately using
319 the rstatix R package version 0.7.2 (Kassambara A. 2023). q values < 0.01 were considered
320 significant and indicative of XCI escape for the hetSNP and tissue.
321 To assess across-tissue XCI status of genes, we leveraged the q values obtained above. If a
322 given hetSNP showed significant Xi expression in all tissues in which the gene is expressed,

323 the SNP was categorized as 'escape (across tissues)', a hetSNP with non-significant Xi
324 expression across all tissues in which the gene is expressed was classified as 'inactive (across
325 tissues)'. If a gene was significant in more than one but not in all tissues in which it is expressed
326 the gene was classified as 'variable'. Genes for which we had AE data in only one tissue were
327 either labeled "inactive (data for one single tissue)" or "escape (data for one single tissue)"
328 based on the binomial test. Finally, genes residing in the pseudo-autosomal region (PAR) were
329 labeled 'PAR'. Upon inspection of the classification results several well characterized escape
330 genes and inactive genes were mis-classified as variable escape. Consequently, we manually
331 curated allelic expression status of all 380 genes using the empirical guideline of allelic ratio
332 >0.4 as indicating mono-allelic expression while also considering the consequences of high
333 and low read counts. Our three criteria for manual re-classification were low read count, low
334 power and over-estimation, indicating genes with a low statistical power but consistent escape
335 pattern (low power), consistent escape pattern but with non-significant escape in single tissue
336 (low read count) and over-estimation due to high read counts inflating the binomial p-value
337 (over-estimation). The list of genes that were manually curated, their change in XCI status and
338 the criteria for their manual classification can be found in **Supp Fig. 3** and **Table S5**.
339
340 *Copy number analysis*
341 Copy number analysis was performed using the R-package, QDNASeq (27). Briefly, whole
342 genome sequencing data for all three nmXCI females was mapped to the human reference
343 genome (hg38) by BWA-MEM (23) using default parameters. Raw copy numbers were then
344 estimated by counting the number of reads mapped to non-overlapping bins of 15 kb unless
345 stated otherwise.
346
347 *Investigating chromosome 17p and 17q arm expression*
348 Transcriptome mapping was performed using Salmon v1.9.0 (--gcBias --seqBias --
349 numBootstraps 100) (28) using known Refseq transcripts (NM_* and NR_*) from the

350 GRCh38.p12 assembly (GCF_000001405.38_GRCh38.p12_rna.fna.gz) as a reference.
351 Library type options were IU. Raw fastq reads were quality trimmed using FastP (v.0.20.0) with
352 default settings. Abundance estimates were converted to h5 format using Wasabi
353 (https://github.com/COMBINE-lab/wasabi). Sleuth was used for data normalization and
354 normalized gene abundances (29, 30). Expression matrix (TPM) can be found in **Table S8**.

355

356

357 **References**

358 1. O. Rosspopoff *et al.*, Species-specific regulation of XIST by the JPX/FTX orthologs. *Nucleic*
359 *Acids Res* **51**, 2177-2194 (2023).
360 2. C. Patrat, J. F. Ouimette, C. Rougeulle, X chromosome inactivation in human development.
361 *Development* **147** (2020).
362 3. B. P. Balaton, A. M. Cotton, C. J. Brown, Derivation of consensus inactivation status for X-
363 linked genes from genome-wide studies. *Biol Sex Differ* **6**, 35 (2015).
364 4. L. Carrel, H. F. Willard, X-inactivation profile reveals extensive variability in X-linked gene
365 expression in females. *Nature* **434**, 400-404 (2005).
366 5. M. Souyris *et al.*, TLR7 escapes X chromosome inactivation in immune cells. *Sci Immunol* **3**
367 (2018).
368 6. M. J. Mousavi, M. Mahmoudi, S. Ghotloo, Escape from X chromosome inactivation and
369 female bias of autoimmune diseases. *Mol Med* **26**, 127 (2020).
370 7. A. Helena Mangs, B. J. Morris, The Human Pseudoautosomal Region (PAR): Origin, Function
371 and Future. *Curr Genomics* **8**, 129-136 (2007).
372 8. T. Tukiainen *et al.*, Landscape of X chromosome inactivation across human tissues. *Nature*
373 **550**, 244-248 (2017).
374 9. A. M. Cotton *et al.*, Analysis of expressed SNPs identifies variable extents of expression from
375 the human inactive X chromosome. *Genome Biol* **14**, R122 (2013).
376 10. R. M. Plenge *et al.*, A promoter mutation in the XIST gene in two unrelated families with
377 skewed X-chromosome inactivation. *Nat Genet* **17**, 353-356 (1997).
378 11. J. M. Amos-Landgraf *et al.*, X chromosome-inactivation patterns of 1,005 phenotypically
379 unaffected females. *Am J Hum Genet* **79**, 493-499 (2006).
380 12. E. Shvetsova *et al.*, Skewed X-inactivation is common in the general female population. *Eur J*
381 *Hum Genet* **27**, 455-465 (2019).
382 13. B. P. Favilla *et al.*, Spread of X-chromosome inactivation into autosomal regions in patients
383 with unbalanced X-autosome translocations and its phenotypic effects. *Am J Med Genet A*
384 **185**, 2295-2305 (2021).
385 14. E. Hatchwell, D. Robinson, J. A. Crolla, A. E. Cockwell, X inactivation analysis in a female with
386 hypomelanosis of Ito associated with a balanced X;17 translocation: evidence for functional
387 disomy of Xp. *J Med Genet* **33**, 216-220 (1996).
388 15. P. Clerc, P. Avner, Random X-chromosome inactivation: skewing lessons for mice and men.
389 *Curr Opin Genet Dev* **16**, 246-253 (2006).
390 16. G. T. Consortium, The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-585
391 (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).
31. B. P. Balaton, A. M. Cotton, C. J. Brown, Derivation of consensus inactivation status for X-linked genes from genome-wide studies. *Biol Sex Differ* **6** (2015).

434 **Data and code availability**

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

437

438 **Author contributions**

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

441

442 **Declarations of interest**

443 The authors declare no competing interests.

444

445 **Funding sources**

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

448

449 **Acknowledgements**

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

453

454

455

456

457

458

459

460

461 **Figure legends**

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

475

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

487

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

515 **Supplementary figure legends**

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

529

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

534

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

539

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

542 (i.e. escape, FDR-corrected binomial q value < 0.01). The 'consensus call' tile is the assigned
543 XCI status across tissues for each gene with genes classified as variable including both intra-
544 and interindividual variation. Red asterisks before gene names indicate genes in which manual
545 curation was performed to assign XCI status. Grey tiles indicate missing data.
546
547 **Supp Fig. 5. (A)** Spearman correlation of allelic expression between nmXCI females. **(B, C)**
548 Alluvial plots showing direct classification of escape status based on our analysis (Gylemo)
549 compared to classification of escape status of X-linked genes as consensus calls for X-
550 inactivation based on data from multiple previous studies employing indirect measures to
551 determine escape; reported by **(B)** Tukiainen *et al* (8) (in their Suppl. Table.1, Combined XCI
552 status) and **(C)** Balaton *et al* (in their Table S1, Balaton consensus calls) (31).
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568

569 **Supplementary table legends**

570

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

572

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

577

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

580

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

583

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

586

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

589

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

592

593 **Table S7:** Tissue abbreviations.

594

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

Figure 1

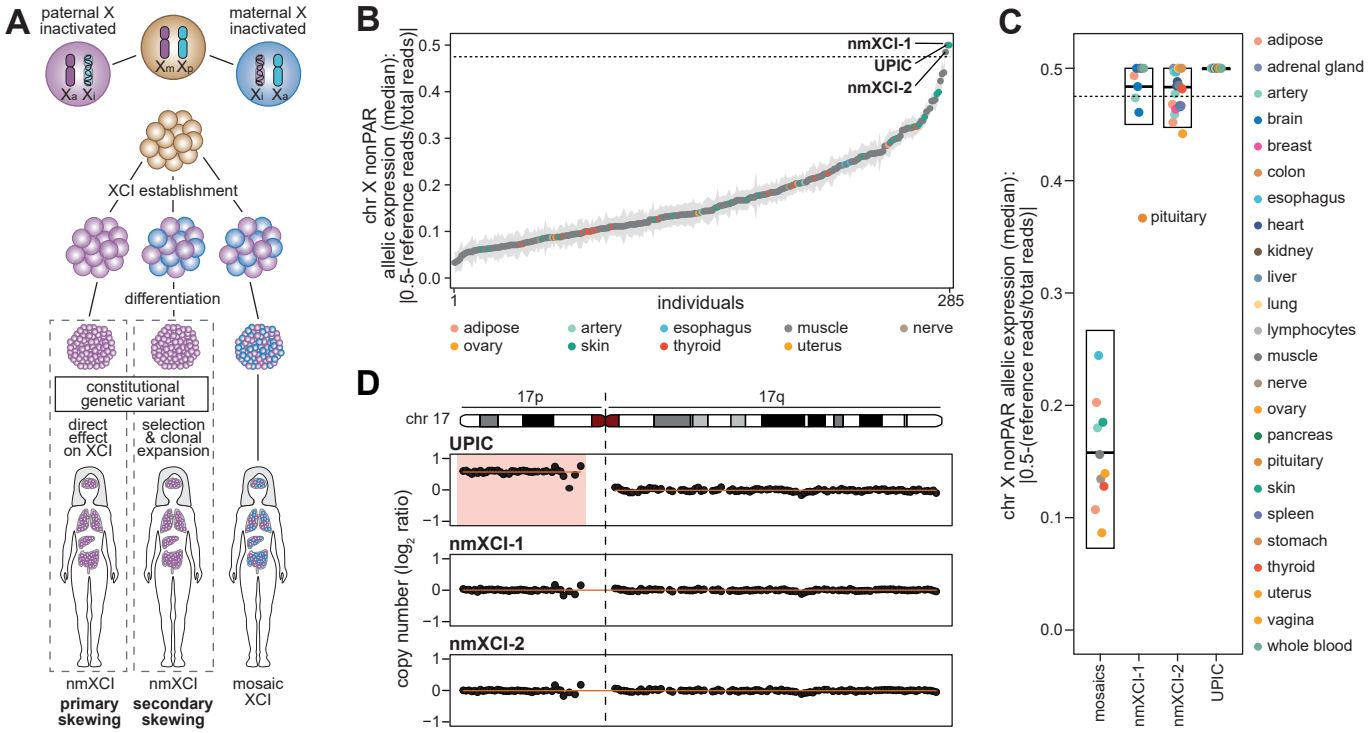

A
paternal X inactivated
maternal X inactivated
XCI establishment
differentiation
constitutional genetic variant
direct effect on XCI
selection & clonal expansion
nmXCI primary skewing
nmXCI secondary skewing
mosaic XCI
B
chr X nonPAR allelic expression (median): |0.5-(reference reads/total reads)|
individuals
nmXCI-1
UPIC
nmXCI-2
adipose
artery
esophagus
muscle
nerve
ovary
skin
thyroid
uterus
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
chr 17
17p
17q
UPIC
nmXCI-1
nmXCI-2
copy number (log2 ratio)

Figure 2

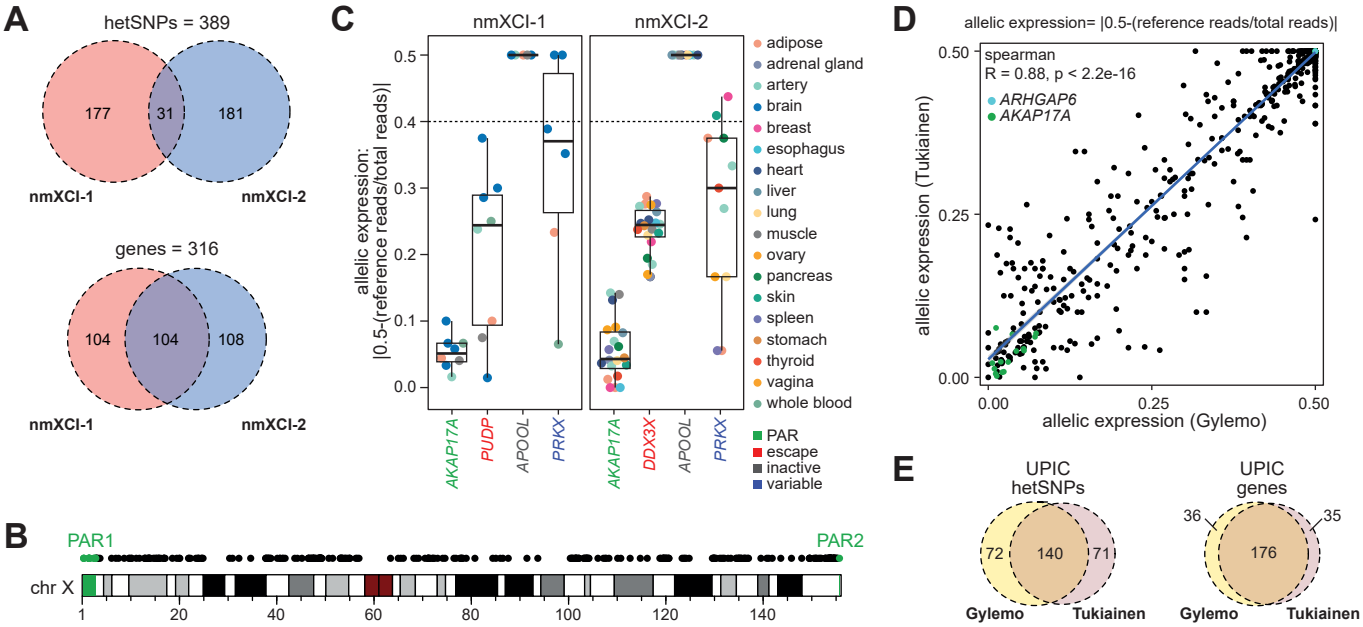

A
hetSNPs = 389
177
31
181
nmXCI-1
nmXCI-2
genes = 316
104
104
108
nmXCI-1
nmXCI-2
B
PAR1
PAR2
chr X
1
20
40
60
80
100
120
140
C
nmXCI-1
nmXCI-2
allelic expression: |0.5-(reference reads/total reads)|
0.5
0.4
0.3
0.2
0.1
0.0
AKAP17A
PUDP
APOOL
PRKX
AKAP17A
DDX3X
APOOL
PRKX
adipose
adrenal gland
artery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

Figure 3

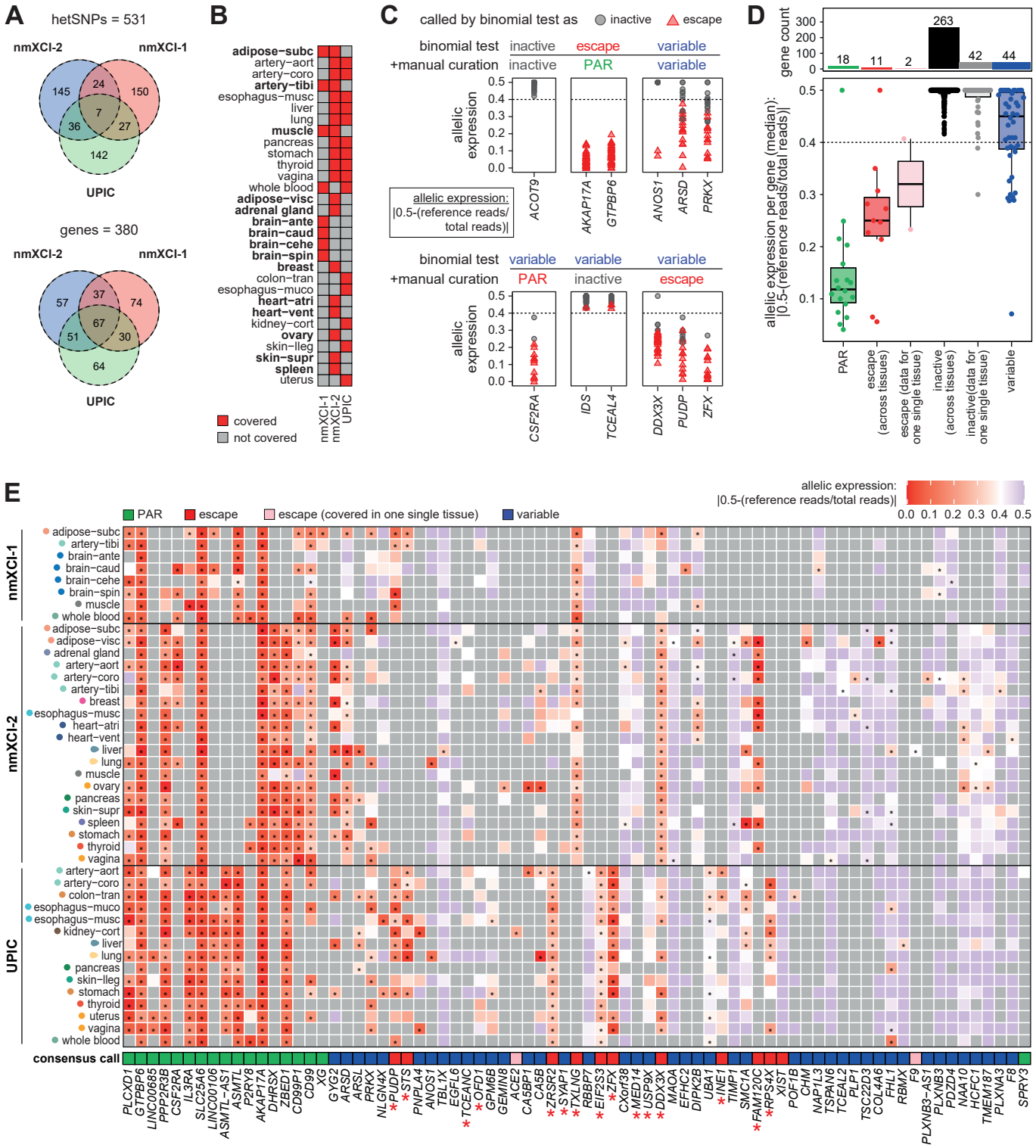
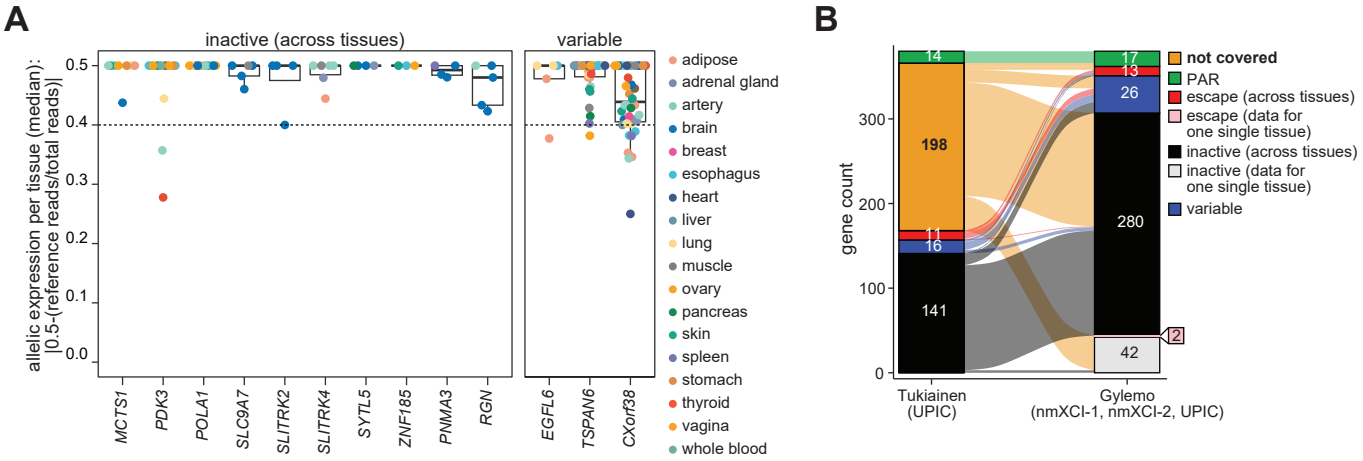
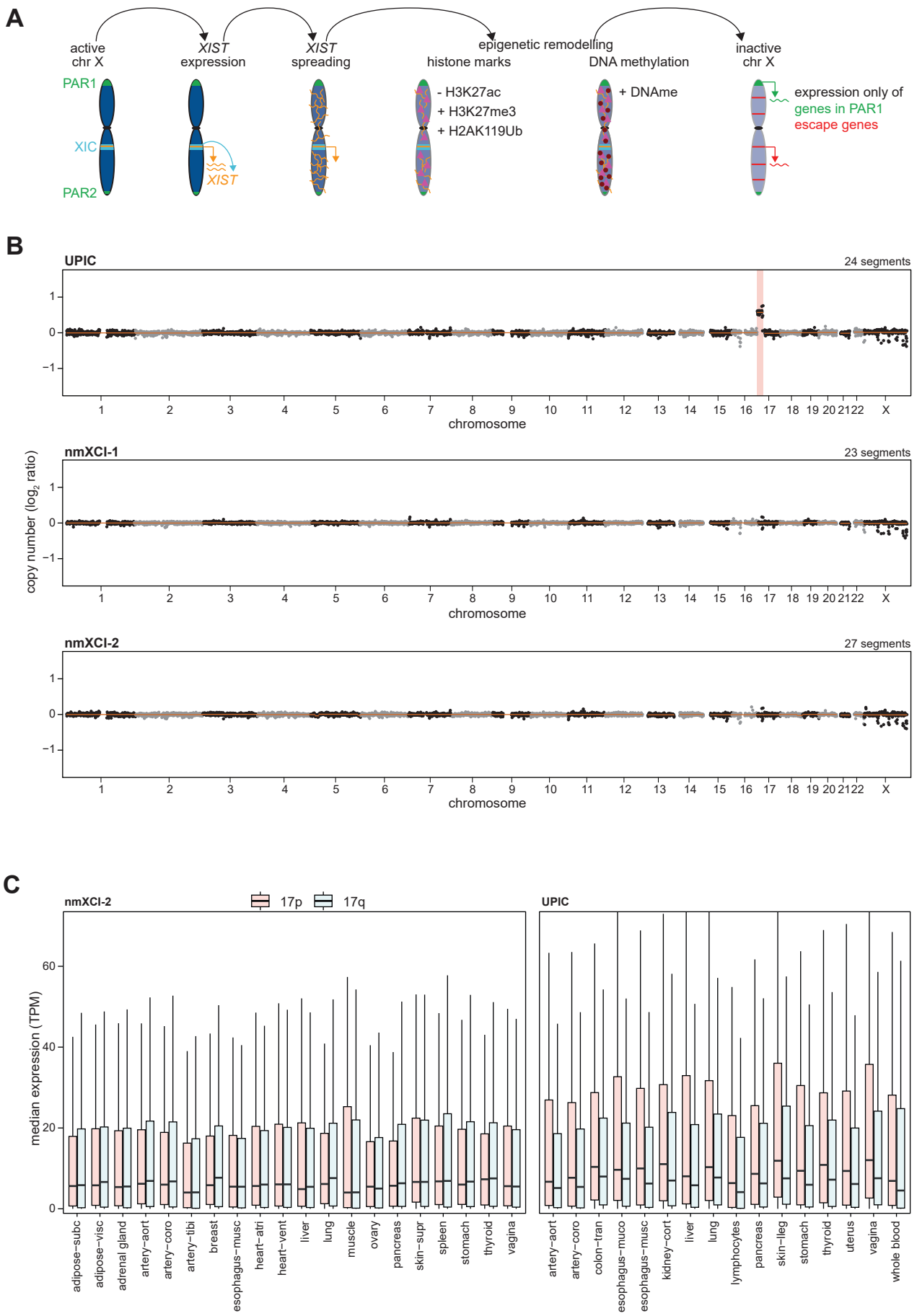

A
hetSNPs = 531
nmXCI-2
nmXCI-1
145
24
150
7
36
27
142
UPIC
genes = 380
nmXCI-2
nmXCI-1
57
37
74
67
51
30
64
UPIC
B
adipose−subc
artery−aort
artery−coro
artery−tibi
esophagus−musc
liver
lung
muscle
pancreas
stomach
thyroid
vagina
whole blood
adipose−visc
adrenal gland
brain−ante
brain−caud
brain−cehe
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
263
18
11
2
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

Figure 4

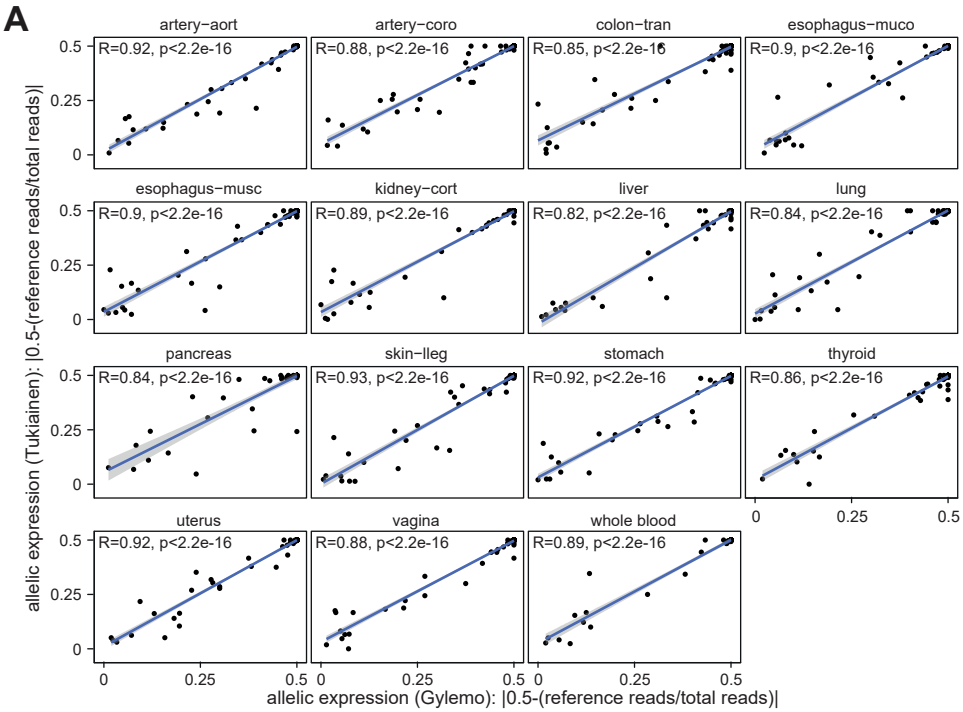

A
inactive (across tissues)
variable
allelic expression per tissue (median): |0.5-(reference reads/total reads)|
0.5
0.4
0.3
0.2
0.1
0.0
MCTS1
PDK3
POLA1
SLC9A7
SLITRK2
SLITRK4
SYTL5
ZNF185
PNMA3
RGN
EGFL6
TSPAN6
CXorf38
adipose
adrenal gland
artery
brain
breast
esophagus
heart
liver
lung
muscle
ovary
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
Tukiainen (UPIC)
Gylemo (nmXCI-1, nmXCI-2, UPIC)
not covered
PAR
escape (across tissues)
escape (data for one single tissue)
inactive (across tissues)
inactive (data for one single tissue)
variable

Supplementary Figure 1

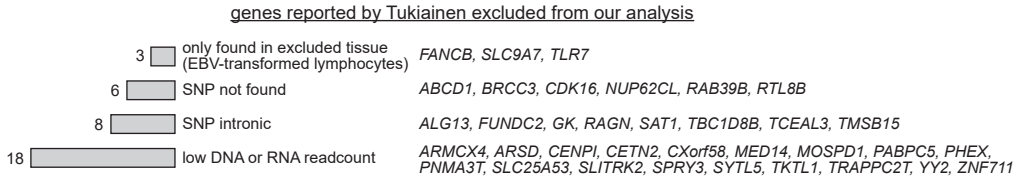


Supplementary Figure 2

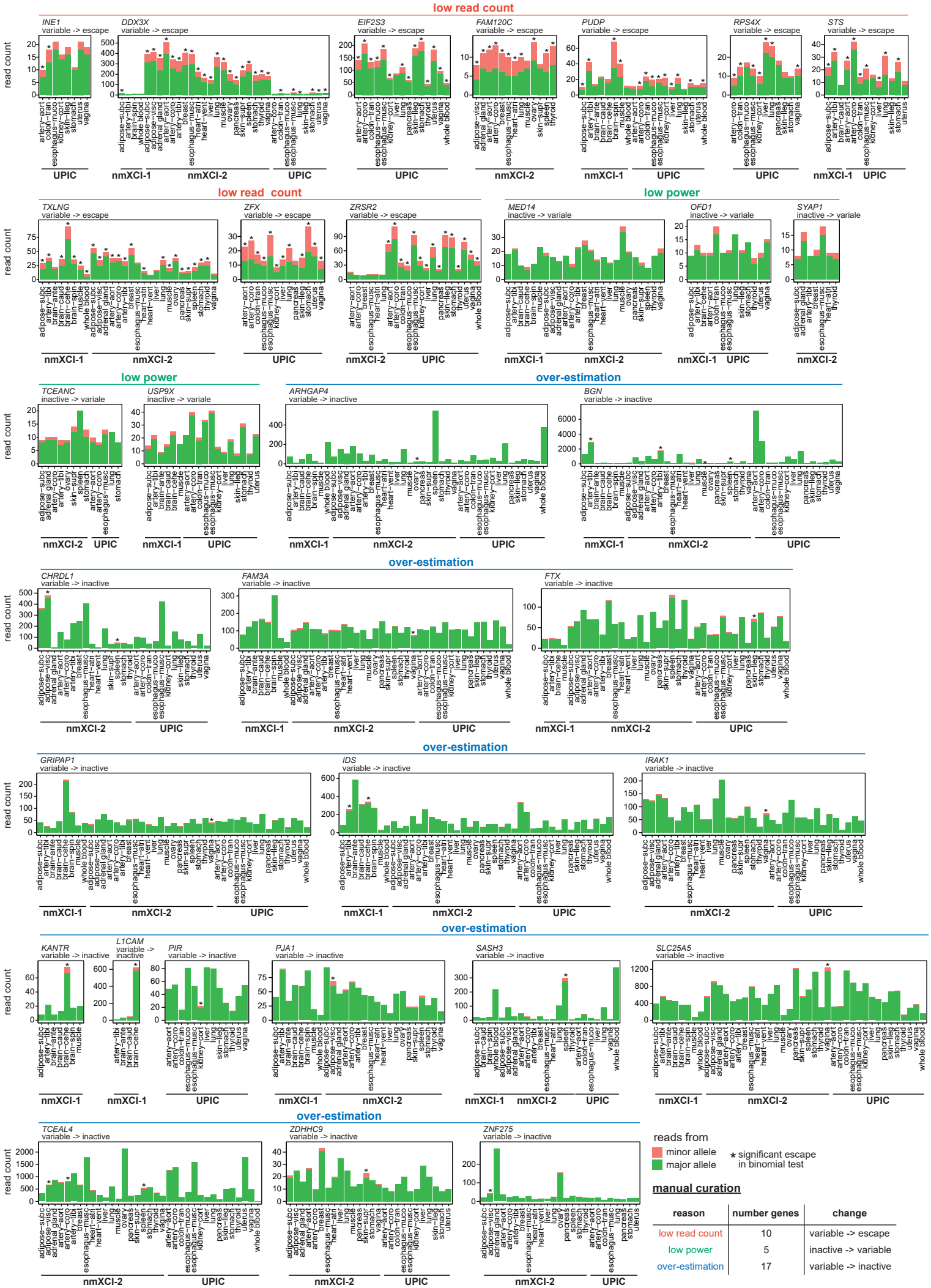
A



B



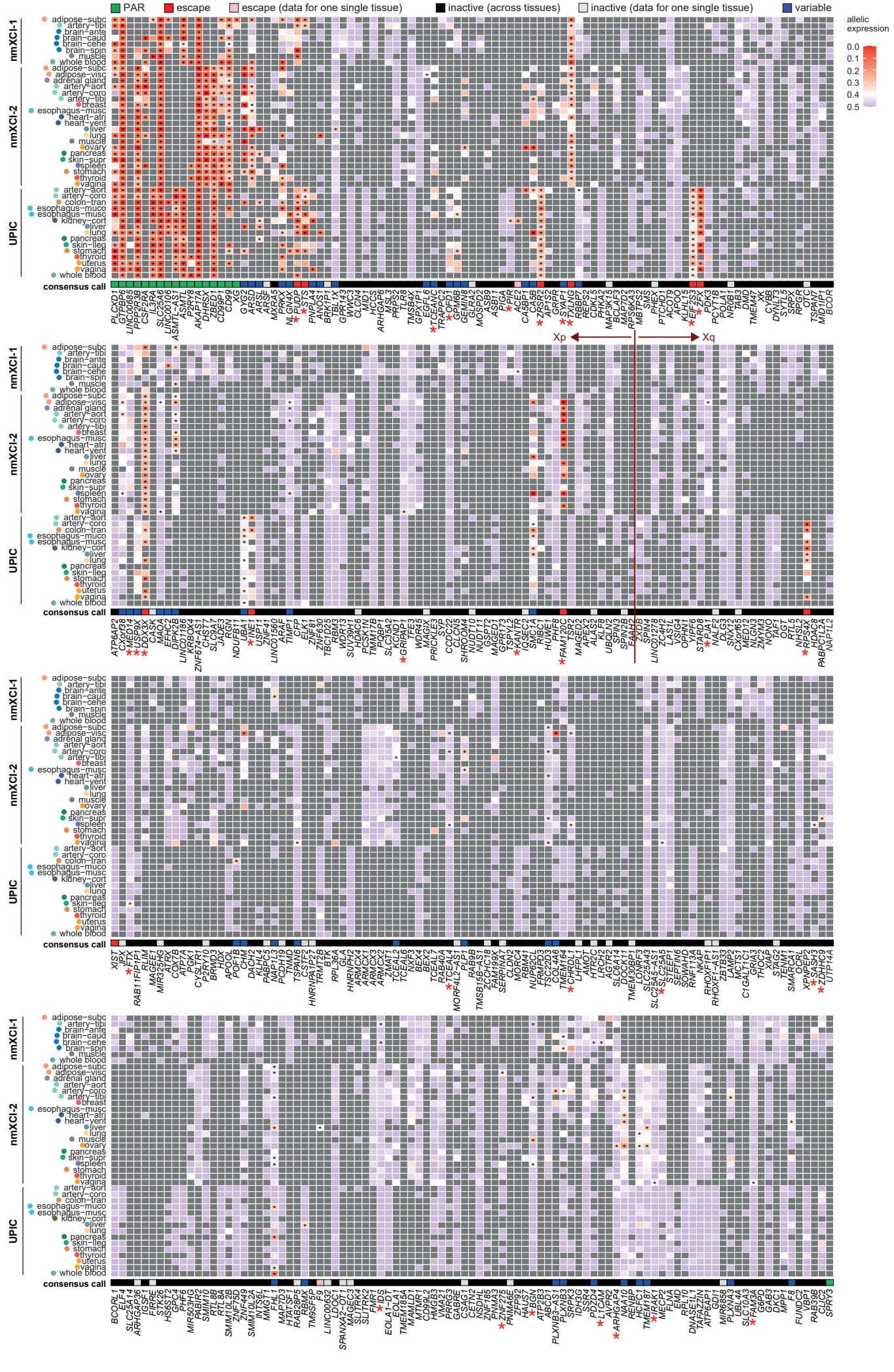
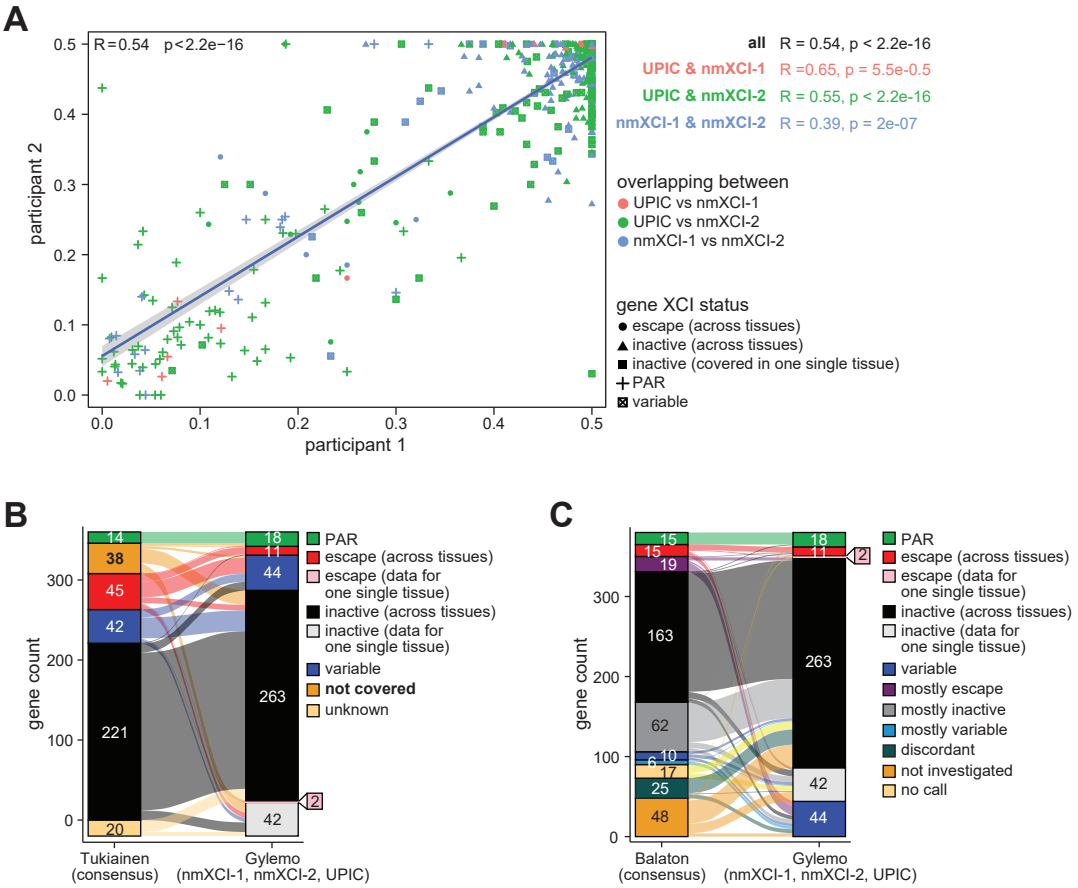
Supplementary Figure 3



manual curation

reason	number genes	change
low read count	10	variable -> escape
low power	5	inactive -> variable
over-estimation	17	variable -> inactive

Supplementary Figure 4



Supplementary Figure 5