Cell Genomics, Volume *4*

Supplemental information

Quantification of escape from X chromosome inactivation with single-cell omics data reveals heterogeneity across cell types and tissues

Yoshihiko Tomofuji, Ryuya Edahiro, Kyuto Sonehara, Yuya Shirai, Kian Hong Kock, Qingbo S. Wang, Shinichi Namba, Jonathan Moody, Yoshinari Ando, Akari Suzuki, Tomohiro Yata, Kotaro Ogawa, Tatsuhiko Naito, Ho Namkoong, Quy Xiao Xuan Lin, Eliora Violain Buyamin, Le Min Tan, Radhika Sonthalia, Kyung Yeon Han, Hiromu Tanaka, Ho Lee, Asian Immune Diversity Atlas Network, Japan COVID-19 Task Force, The BioBank Japan Project, Tatsusada Okuno, Boxiang Liu, Koichi Matsuda, Koichi Fukunaga, Hideki Mochizuki, Woong-Yang Park, Kazuhiko Yamamoto, Chung-Chau Hon, Jay W. Shin, Shyam Prabhakar, Atsushi Kumanogoh, and Yukinori Okada

Supplemental figures

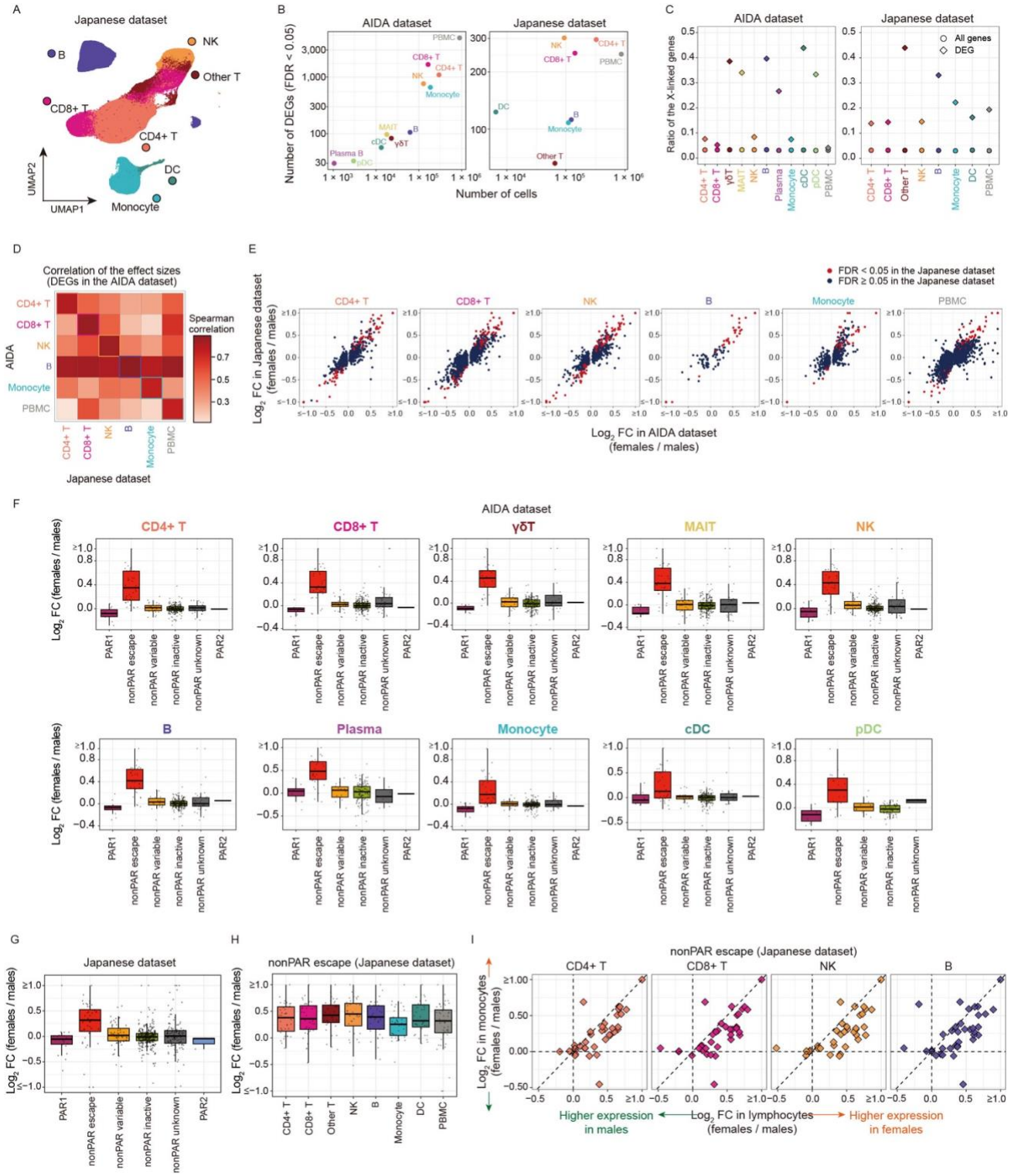

Figure S1. The results of the pseudobulk differentially expressed gene analysis are consistent between the AIDA and Japanese datasets, related to Figure 1.

A, UMAP of the Japanese dataset. **B,** The relationship between the number of cells (x-axis) and the number of the significant DEGs (FDR < 0.05; y-axis) for the AIDA (left) and Japanese (right) datasets. The colors indicate the cell types. **C,** The ratios of the X-linked genes among

all genes (circle) and DEGs (rhombus) are indicated for the AIDA (left) and Japanese (right) datasets. The colors indicate the cell types. **D,** Spearman correlations between the effect sizes in the DEG analysis for the AIDA (y-axis) and Japanese (x-axis) datasets for DEGs (AIDA dataset) in the major cell types. **E,** Scatter plots represent the effect sizes of the significant DEGs detected in the AIDA dataset in the DEG analysis for the AIDA (x-axis) and Japanese (y-axis) datasets. The colors of the points represent whether the genes are significant DEGs in the Japanese dataset. **F,** Box plots represent log2 fold-changes of the gene expression between sexes for each cell type in the AIDA daatset. Genes are grouped according to the XCI status annotated in the previous study. The boxplot indicates the median values (center lines) and IQRs (box edges), with the whiskers extending to the most extreme points within the range between (lower quantile − [1.5 × IQR]) and (upper quantile + [1.5 × IQR]). **G,** A box plot represents log2 fold-changes of the gene expression between sexes in the Japanese dataset. Genes are grouped according to the XCI status annotated in the previous study. The boxplot indicates the median values (center lines) and IQRs (box edges), with the whiskers extending to the most extreme points within the range between (lower quantile − [1.5 × IQR]) and (upper quantile + [1.5 × IQR]). **H,** A box plot represents log2 fold-changes of the escapee gene expression between sexes across cell types in the Japanese dataset. The boxplot indicates the median values (center lines) and IQRs (box edges), with the whiskers extending to the most extreme points within the range between (lower quantile − [1.5 × IQR]) and (upper quantile + [1.5 × IQR]). **I,** Scatter plots represent pairwise comparisons of the log2 fold-changes of the escapee gene expression between sexes in the Japanese dataset. The y-axes represent the log2 fold-changes in monocytes and the x-axes represent the log2 fold-changes in lymphocytes. The dashed lines represent $x = 0$, $x = y$, and $y = 0$. DEG, differentially expressed genes; AIDA, Asian Immune Diversity Atlas; FC, fold-changes; FDR, false discovery ratio; IQR, interquartile range; PAR, pseudoautosomal region; PBMC, peripheral blood mononuclear cells; scRNA-seq, single-cell RNA-seq; UMAP, Uniform manifold approximation and projection; XCI, X chromosome inactivation.

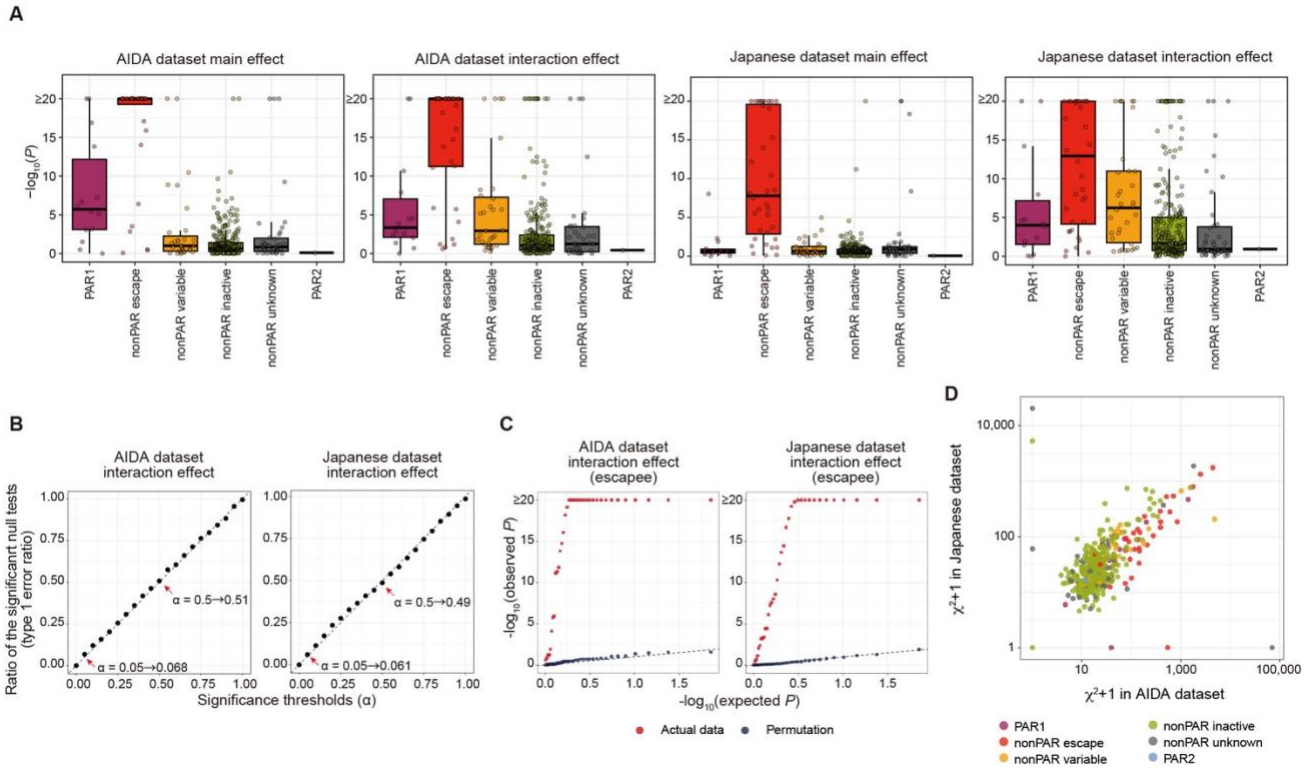

Figure S2. The results of the single-cell level differentially expressed gene analysis are consistent between the AIDA and Japanese datasets, related to Figure 1.

A, A box plot represents P-values for the sex term (AIDA dataset), sex × cell state term (AIDA dataset), sex term (Japanese dataset), and sex × cell state (Japanese dataset) in the single-cell level DEG analysis. Genes are grouped according to the XCI status annotated in the previous study. The boxplot indicates the median values (center lines) and IQRs (box edges), with the whiskers extending to the most extreme points within the range between (lower quantile − [1.5 × IQR]) and (upper quantile + [1.5 × IQR]). **B,** Ratio of significant tests under cell state permutation in the AIDA and Japanese datasets. Each dot represents the ratio of the significant tests (y-axis) at the given alpha threshold (x-axis). **C,** Q-Q plots for the P-values for the sex × cell state term of escapee genes (left, AIDA dataset; right, Japanese dataset). The color of the dots represents whether the test is performed for the actual data or under cell state permutation. **D,** A scatter plot represents the relationship between the χ2 statistics for the sex × cell state term in the AIDA (x-axis) and Japanese dataset (y-axis). The colors of the points represent the XCI status annotated in the previous study. AIDA, Asian Immune Diversity Atlas; DEG, differentially expressed genes; IQR, interquartile range; PAR, pseudoautosomal region; scRNA-seq, single-cell RNA-seq; XCI, X chromosome inactivation.

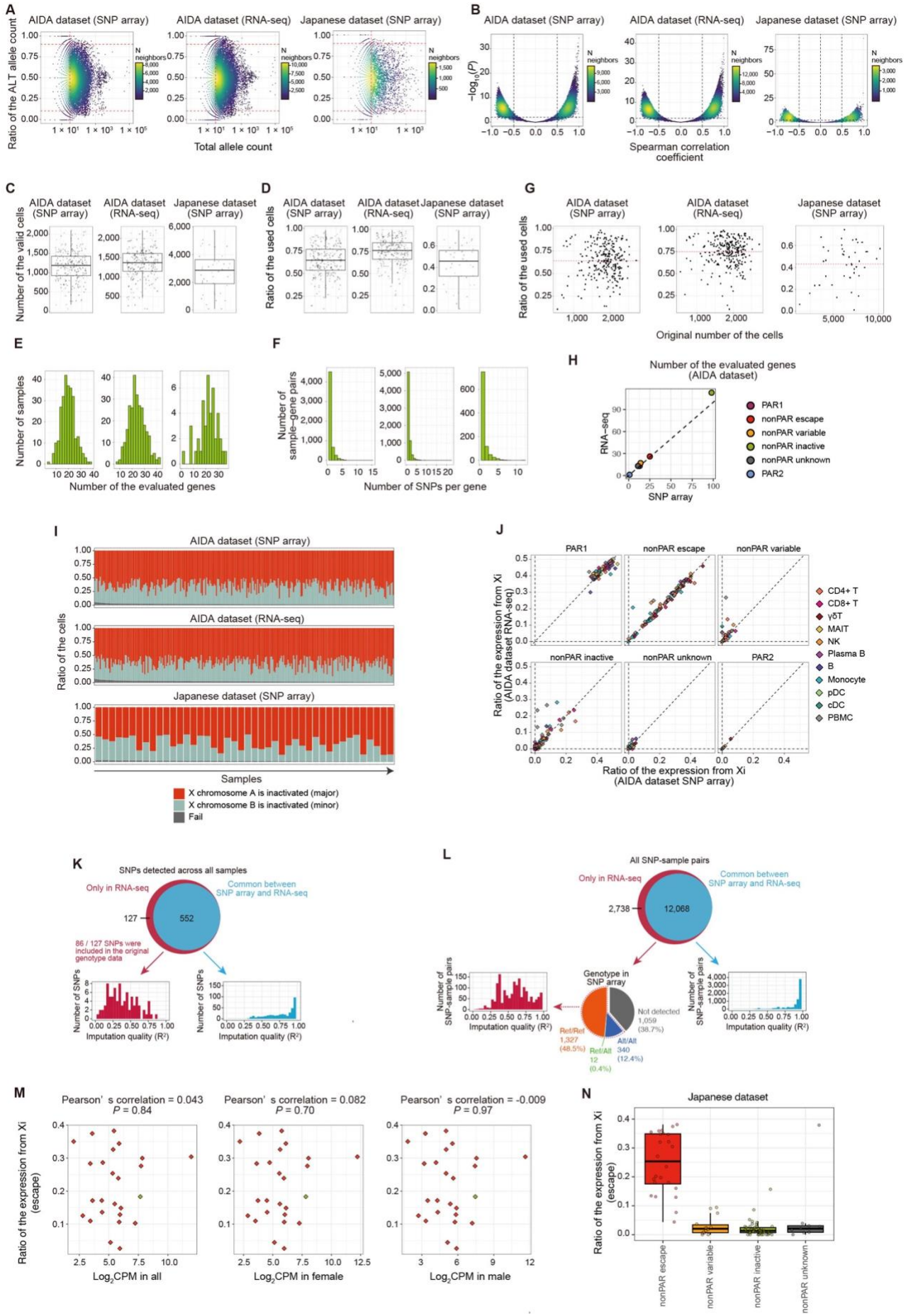
A
AIDA dataset (SNP array)
AIDA dataset (RNA-seq)
Japanese dataset (SNP array)
Ratio of the ALT allele count
Total allele count
N neighbors
B
AIDA dataset (SNP array)
AIDA dataset (RNA-seq)
Japanese dataset (SNP array)
−log10(P)
Spearman correlation coefficient
C
AIDA dataset (SNP array)
AIDA dataset (RNA-seq)
Japanese dataset (SNP array)
Number of the valid cells
D
AIDA dataset (SNP array)
AIDA dataset (RNA-seq)
Japanese dataset (SNP array)
Ratio of the used cells
G
AIDA dataset (SNP array)
AIDA dataset (RNA-seq)
Japanese dataset (SNP array)
Ratio of the used cells
Original number of the cells
E
Number of samples
Number of the evaluated genes
F
Number of sample–gene pairs
Number of SNPs per gene
H
Number of the evaluated genes (AIDA dataset)
RNA-seq
SNP array
PAR1
nonPAR escape
nonPAR variable
nonPAR inactive
nonPAR unknown
PAR2
I
AIDA dataset (SNP array)
AIDA dataset (RNA-seq)
Japanese dataset (SNP array)
Ratio of the cells
Samples
X chromosome A is inactivated (major)
X chromosome B is inactivated (minor)
Fail
J
PAR1
nonPAR escape
nonPAR variable
nonPAR inactive
nonPAR unknown
PAR2
Ratio of the expression from Xi (AIDA dataset RNA-seq)
Ratio of the expression from Xi (AIDA dataset SNP array)
CD4+ T
CD8+ T
γδT
MAIT
NK
Plasma B
B
Monocyte
pDC
cDC
PBMC
K
SNPs detected across all samples
Only in RNA-seq
Common between SNP array and RNA-seq
127
552
86 / 127 SNPs were included in the original genotype data
Number of SNPs
Imputation quality (R²)
L
All SNP-sample pairs
Only in RNA-seq
Common between SNP array and RNA-seq
2,738
12,068
Number of SNP-sample pairs
Imputation quality (R²)
Genotype in SNP array
Not detected
1,059
(38.7%)
Ref/Ref
1,327
(48.5%)
Alt/Alt
340
(12.4%)
Ref/Alt
12
(0.4%)
M
Pearson' s correlation = 0.043
P = 0.84
Pearson' s correlation = 0.082
P = 0.70
Pearson' s correlation = -0.009
P = 0.97
Ratio of the expression from Xi (escape)
Log2CPM in all
Log2CPM in female
Log2CPM in male
N
Japanese dataset
Ratio of the expression from Xi (escape)
nonPAR escape
nonPAR variable
nonPAR inactive
nonPAR unknown

Figure S3. Quantification of escape from XCI by scLinaX, related to Figure 2.

A, The relationships between the ratio of the ALT allele counts (y-axis) and the total allele counts (x-axis) for each SNP for each dataset. The points represent the pairs of sample and SNP, and the colors of the points represent the density of the neighboring points. The red dashed lines indicate the thresholds for QC. **B,** The relationships between the $-\log_{10}$(P-value) (y-axis) and correlation coefficients (x-axis) of the Spearman correlation tests for the pseudobulk ASE profiles generated during scLinaX workflow. The points represent the pairs of two pseudobulk profiles, and the colors of the points represent the density of the neighboring points. The dashed lines indicate the thresholds for QC. **C,D,** The boxplots represent the number of valid cells (c) and the ratio of the used cells (cells for which any of the reference SNPs are detected; d) for each sample. The boxplot indicates the median values (center lines) and IQRs (box edges), with the whiskers extending to the most extreme points within the range between (lower quantile − [1.5 × IQR]) and (upper quantile + [1.5 × IQR]). **E,F,** The histograms represent the number of genes detected in ≥3 individuals (E) and transcribed SNPs with ≥10 UMI coverages per sample-gene pair (F) **G,** The relationship between the original number of the cells (x-axis) and the ratio of the used cells (y-axis). The points represent samples. The red dashed lines indicate the mean ratio of the used cells across samples. **H,** Number of the genes evaluated in the scLinaX analysis of the AIDA dataset with the SNP data based on the SNP array (x-axis) and called from scRNA-seq data (y-axis) in PBMC. **I,** A bar plot represents the ratio of the cells that different X chromosomes are inactivated or are removed from the analysis due to the bi-allelic expression of the reference SNPs. The colors of the dots represent the XCI status annotated in the previous study. **J,** Plots represent the ratio of the expression from Xi in the AIDA dataset calculated from the SNP data derived from SNP array data (x-axis) and scRNA-seq data (y-axis) in all of the cell types. Genes are grouped according to the XCI status annotated in the previous study. **K**, A Venn diagram represents SNPs detected across all samples in the AIDA dataset. The red area indicates the SNPs called only in the scRNA-seq data-based analysis. The blue area indicates the SNPs detected commonly in the methods with SNP array data and only with scRNA-seq data. The red and blue histograms indicate the imputation quality (R^2) of the SNPs included in the red and blue areas, respectively. **L**, A Venn diagram represents SNP-sample pairs included in the analysis with the AIDA dataset. The red area indicates the heterozygous SNP-sample pairs detected only in the scRNA-seq data-based analysis. The blue area indicates the SNP-sample pairs detected both in the methods with SNP array data and only with scRNA-seq data. The pie chart represents the SNP array-based genotype of the heterozygous SNP-sample pairs detected only in the scRNA-seq data-based analysis. The red and blue histograms indicate the imputation quality (R^2) of the SNPs included in the red and blue areas, respectively. **M,** Plots represent the relationship between the mean log2 CPM across all, female, or male samples (x-axis) and the ratio of the expression from Xi (y-

axis) in the AIDA dataset. Genes that are annotated as escape genes or shows evidence of escape in the scLinaX analysis (*SEPTIN6*) are indicated. **N,** A box plot represents the estimated ratio of the expression from Xi in the Japanese dataset. Genes are grouped according to the XCI status annotated in the previous study. The boxplot indicates the median values (center lines) and IQRs (box edges), with the whiskers extending to the most extreme points within the range between (lower quantile − [1.5 × IQR]) and (upper quantile + [1.5 × IQR]). AIDA, Asian Immune Diversity Atlas; ALT, alternative allele; ASE, allele-specific expression; CI, confidence interval; CPM, counts per million; DEG, differentially expressed genes; FC, fold-changes; FDR, false discovery ratio; IQR, interquartile range; PAR, pseudoautosomal region; QC, quality control; REF, reference allele; SNP, single nucleotide polymorphism; Xa, active X chromosome; XCI, X chromosome inactivation; Xi, inactive X chromosome.

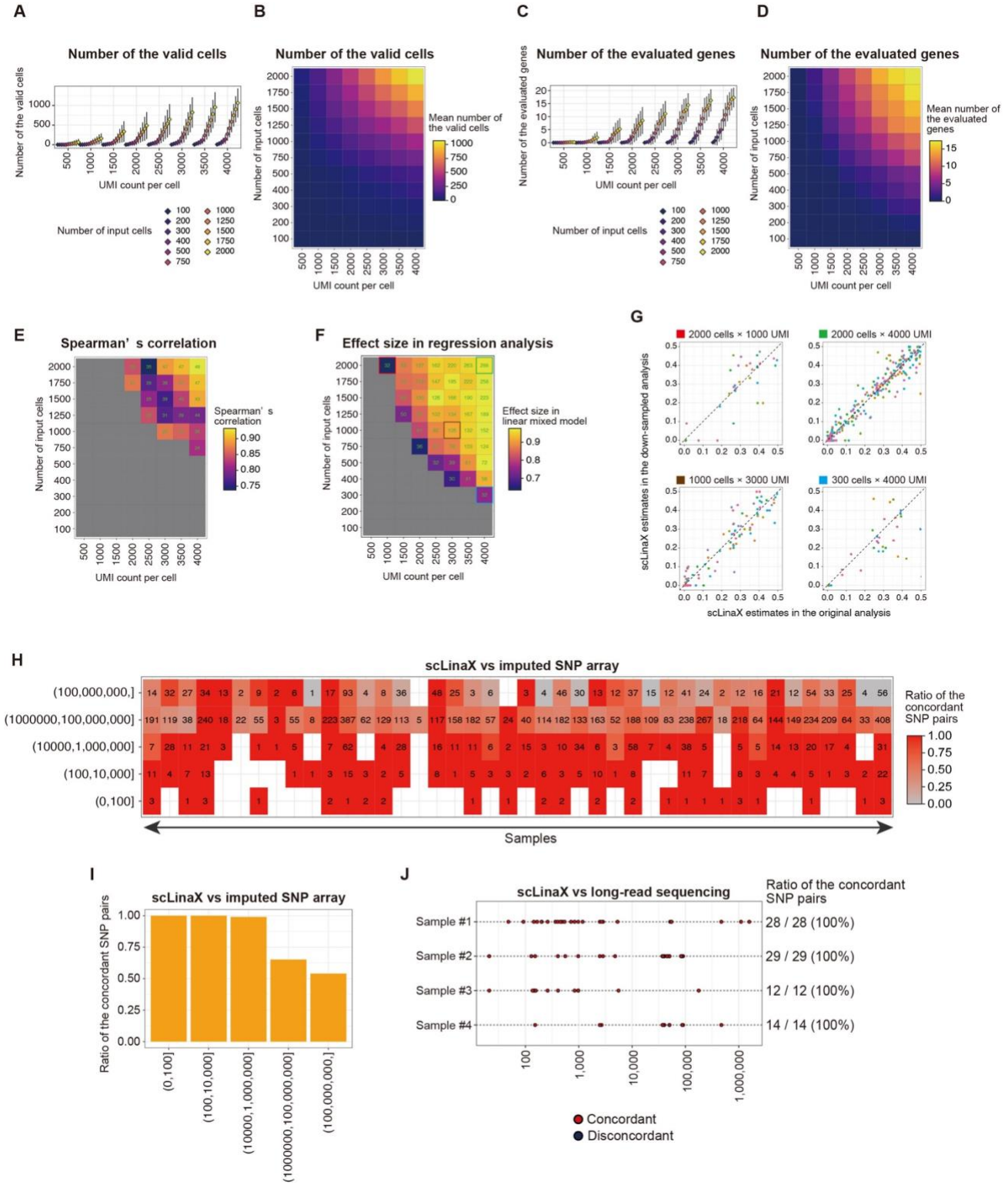

Figure S4. Evaluation of the performance of scLinaX, related to Figure 2.

A,B, Boxplots (A) and heatmap (B) indicate the number of cells that are mapped with the inactivated X chromosome in the scLinaX analysis with different cell numbers and UMI counts. The same set of 22 AIDA samples with a sufficient number of cells and UMI counts are utilized across conditions. Boxplots indicate the median values (center lines) and IQRs (box edges), with the whiskers extending to the most extreme points within the range between (lower

quantile − [1.5 × IQR]) and (upper quantile + [1.5 × IQR]). The color of the heatmap indicates the mean across the samples. **C,D,** Boxplots (C) and heatmap (D) indicate the number of genes that are evaluated in the scLinaX analysis with different cell numbers and UMI counts. The same set of 22 AIDA samples with a sufficient number of cells and UMI counts are utilized across conditions. Boxplots indicate the median values (center lines) and IQRs (box edges), with the whiskers extending to the most extreme points within the range between (lower quantile − [1.5 × IQR]) and (upper quantile + [1.5 × IQR]). The color of the heatmap indicates the mean across the samples. **E,** A heatmap indicates the Spearman's correlation of the scLinaX estimates of the genes between the full dataset and down-sampled dataset. The indicated numbers represent the number of genes used for the correlation tests. The color of the heatmap indicates the Spearman's correlation. **F,** A heatmap indicates the effect sizes in the regression analysis with the following linear mixed model; scLinaX estimates in the original data (per individual) ~ scLinaX estimates in the down-sampled data (per individual) + (1 | individual). The indicated numbers represent the number of gene-sample pairs used for the regression analysis. The color of the heatmap indicates the effect sizes of the scLinaX estimates in the down-sampled data (per individual). **G,** Scatter plots indicate the relationship between the scLinaX estimates in the original data (x-axis) and the down-sampled data (y-axis) in the four representative conditions colored in (F). Different colors of dots represent gene-sample pairs derived from different individuals. The dashed lines indicate $y = x$. **H,** Comparison of the phase information between scLinaX and imputed SNP array data (SHAPEIT4 + Minimac4) in the Japanese dataset. The x-axis indicates each sample and y-axis indicates distance between the pair of SNPs. The color of the tiles indicates the ratio of the SNP pairs which have concordant phase information. The indicated numbers represent the number of the SNP pairs. **I,** A bar plot indicates the ratio of the pairs of SNPs that have concordant phase information between scLinaX and imputed-SNP array data. Pairs of SNPs are stratified according to the distances between the SNPs. **J,** Comparison of the phase information inferred from scLinaX and physical phasing with long-read sequencing in the Japanese dataset. Each dot indicates pairs of SNPs that are phased with both scLinaX and long-read sequencing. Phase information was concordant for all of the pairs of SNPs. AIDA, Asian Immune Diversity Atlas; IQR, interquartile range; UMI, Unique Molecular Identifier.

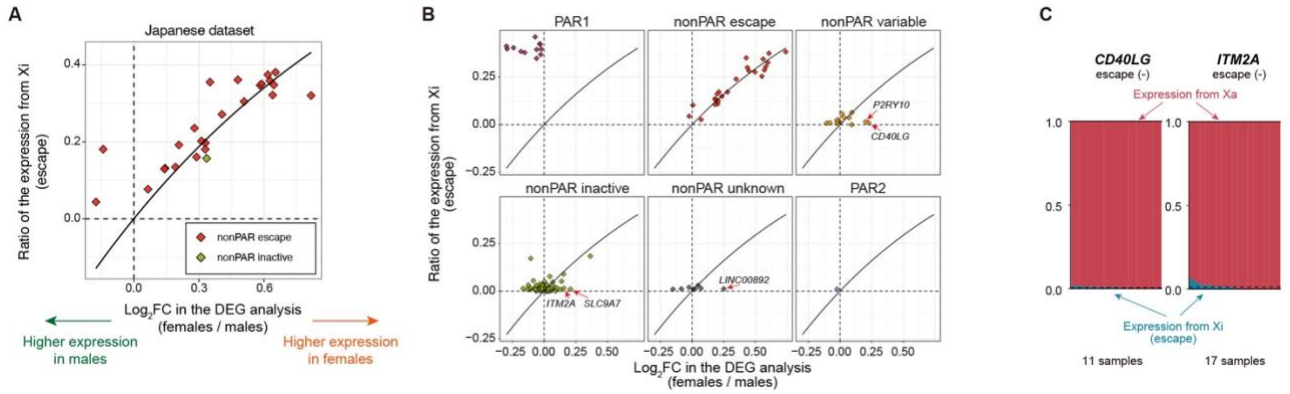

Figure S5. Quantification of escape from XCI by scLinaX, related to Figure 2.
A, A plot represents the relationship between the log2 fold-changes in the DEG analysis (x-axis) and the ratio of the expression from Xi (y-axis) in the Japanese dataset. Genes that are annotated as escape genes or shows evidence of escape in the scLinaX analysis (*SEPTIN6*) are indicated. The curved line indicates the theoretical relationship under the assumption that differential gene expression between sexes is solely due to the expression from Xi and total gene expression in males and Xa-derived gene expression in females are at the same level. Pearson's correlation = 0.86 with a 95% confidence interval of 0.70-0.94. **B,** Plots represent the relationship between the log2 fold-changes in the DEG analysis (x-axis) and the ratio of the expression from Xi (y-axis) in the AIDA dataset for genes with different XCI statuses. The curved line indicates the theoretical relationship under the assumption that differential gene expression between sexes is solely due to the expression from Xi and total gene expression in males and Xa-derived gene expression in females are at the same level. Genes that showed relatively strong deviation from the theoretical relationship are labeled. **C,** Plots represent the ratio of the expression from Xa and Xi at an individual level for the *CD40LG* and *ITM2A* gene. The dashed horizontal line represents the mean ratio of the expression from Xi across samples. Since SNPs on the *ITM2A* gene were included in the initial analysis of the Japanese dataset, scLinaX analysis removing reference SNPs on the *ITM2A* genes was specifically performed for making the plot (right).

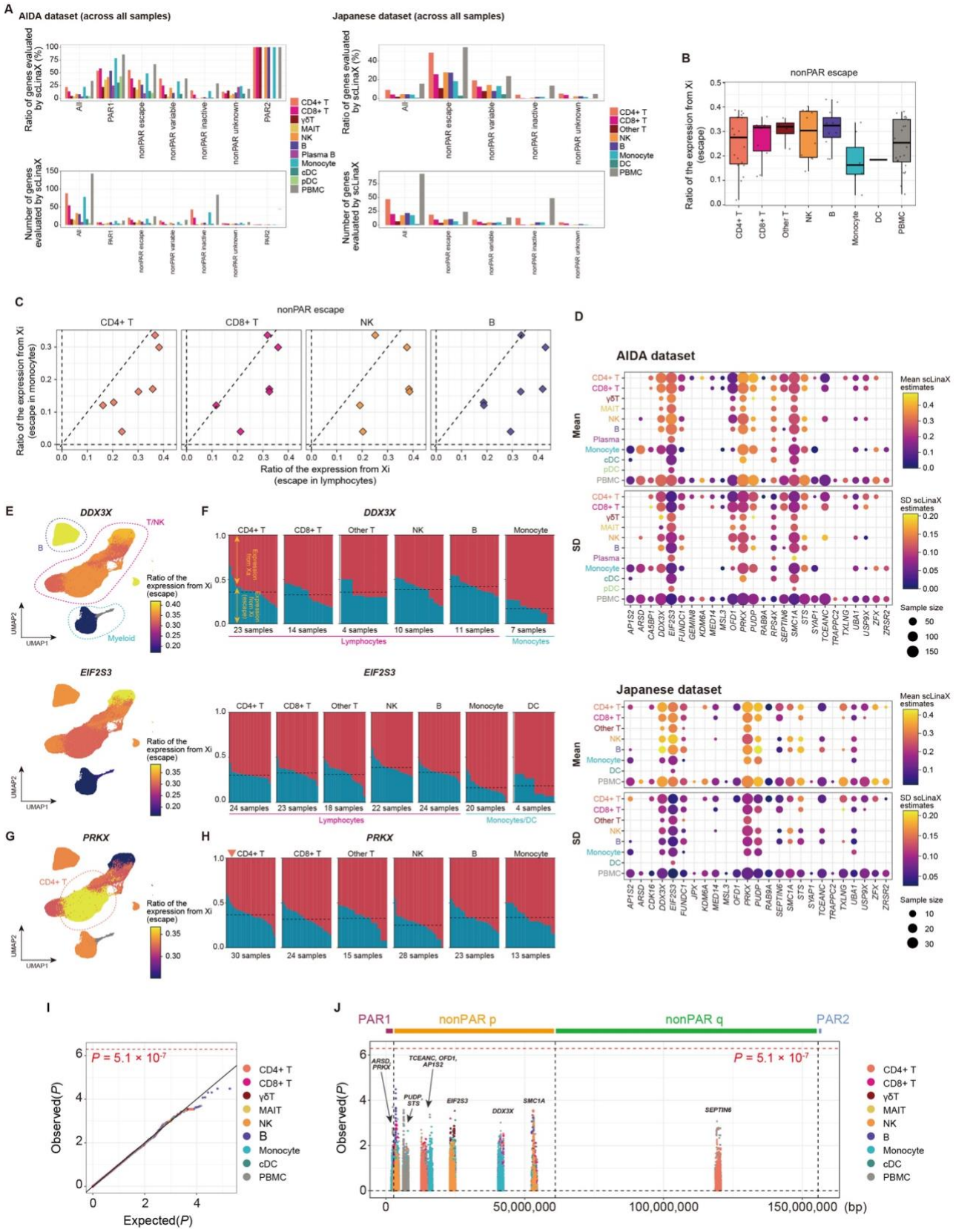

Figure S6. The scLinaX-based quantification of escape from XCI across immune cell types and escape QTL analysis, related to Figure 3.

A, Bar plots represent ratio (top) and number (bottom) of genes evaluated by scLinaX among

the expressing genes for AIDA (left) and Japanese (right) datasets. The color of the bars indicates the cell types. **B,** A box plot represents the estimated ratio of the expression from Xi for escapee genes across cell types in the Japanese dataset. The boxplot indicates the median values (center lines) and IQRs (box edges), with the whiskers extending to the most extreme points within the range between (lower quantile − [1.5 × IQR]) and (upper quantile + [1.5 × IQR]). **C,** Scatter plots represent pairwise comparisons of the ratio of the expression from Xi for escapee genes in the Japanese dataset. The y-axes represent the ratio of the expression from Xi in monocytes and the x-axes represent the ratio of the expression from Xi in lymphocytes. The dashed lines represent $x = 0$, $x = y$, and $y = 0$. **D,** Dot plots represent the mean and SD of the ratio of the expression from Xi across cell types (y-axis) for escapee genes (x-axis). The color of the dots represents the mean or SD of the ratio of the expression from Xi. The size of the dots represents the number of the evaluated samples. Results for the AIDA (top) and Japanese (bottom) datasets are indicated. **E,** UMAPs of the Japanese dataset colored according to the ratio of the expression from Xi estimated for each cell type. Examples of genes that show a higher ratio of expression from Xi in lymphocytes than monocytes are indicated. Cell types whose ratio of the expression from Xi could not be estimated are colored grey. **F,** Plot represents the ratio of the expression from Xa and Xi at an individual level for each cell type in the Japanese dataset. Examples of genes that show a higher ratio of expression from Xi in lymphocytes than monocytes, *DDX3X* and *EIF2S3* genes, are indicated. The dashed horizontal line represents the mean ratio of the expression from Xi across samples for each cell type. **G,** A UMAP of the Japanese dataset colored according to the ratio of the expression from Xi estimated for each cell type. The *PRKX* gene, which shows a unique pattern of heterogeneity of escape across cell types, is indicated. Cell types whose ratio of the expression from Xi could not be estimated are colored grey. **H,** Plot represents the ratio of the expression from Xa and Xi at an individual level for each cell in the Japanese dataset. The *PRKX* gene, which shows a unique pattern of heterogeneity of escape across cell types, is indicated. The dashed horizontal line represents the mean ratio of the expression from Xi across samples for each cell type. **I,J,** Q-Q plot (I) and Manhattan plot (J) for the P-values of the escapee QTL analyses. The color of the dots represents the cell type for which escape QTL is evaluated. The red dashed lines indicate the significance threshold with Bonferroni correction ($\alpha = 0.05$). IQR, interquartile range; QTL, quantitative trait locus; UMAP, Uniform manifold approximation and projection; Xa, active X chromosome; XCI, X chromosome inactivation; Xi, inactive X chromosome.

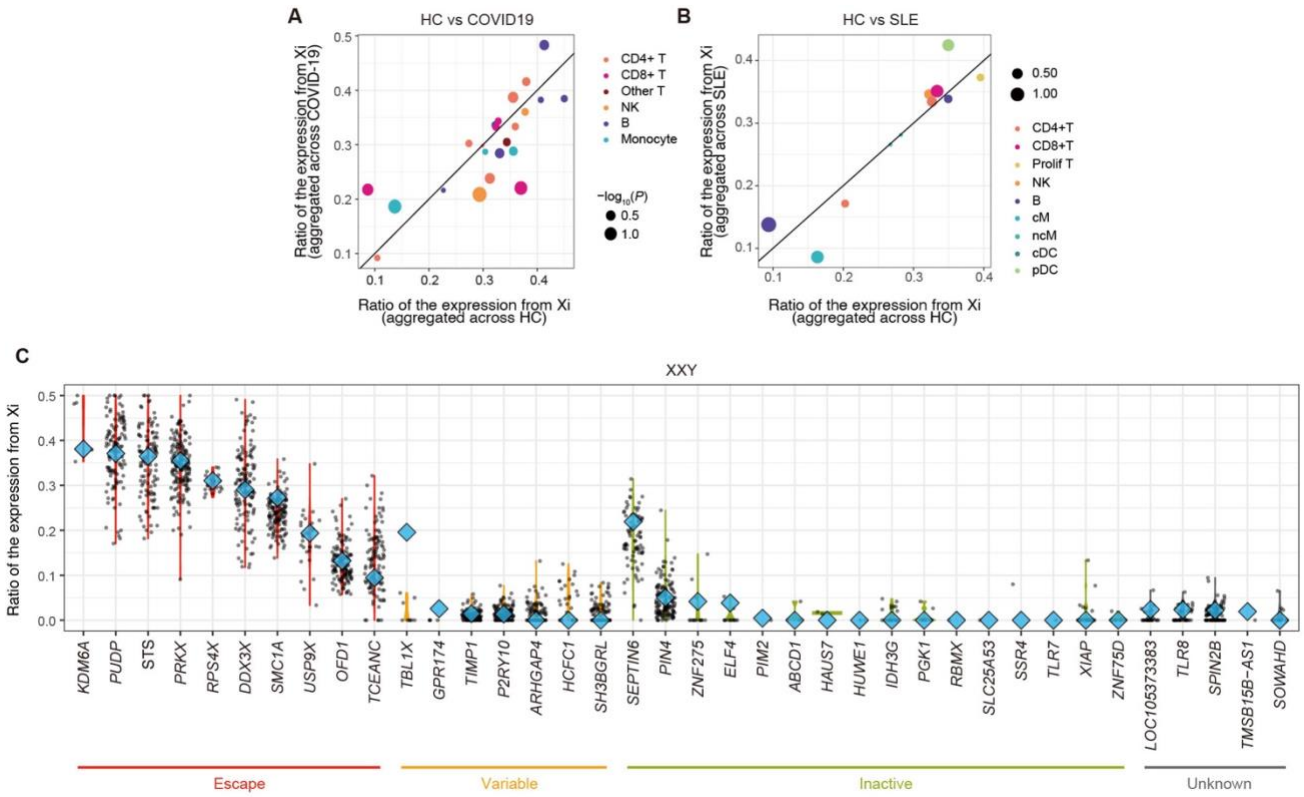

Figure S7. Evaluation of escape in disease conditions, related to Figure 3.

a,b, The ratio of the expression from Xi in healthy subjects (x-axis) and disease patients (y-axis; a, COVID-19; b, SLE). Each point represents a pair of genes and cell types. The colors and sizes of the points indicate the cell type and P-values. The line represents *x* = *y*. **c,** The ratio of the expression from Xi in a male sample with a karyotype of XXY is indicated as blue rhombuses. A violin plot represents the ratio of the expression from Xi in the AIDA datasets. AIDA, Asian Immune Diversity Atlas; COVID-19, coronavirus disease of 2019; HC, healthy control; SLE, systemic lupus erythematosus; SNP, single nucleotide polymorphism; Xi, inactive X chromosome.

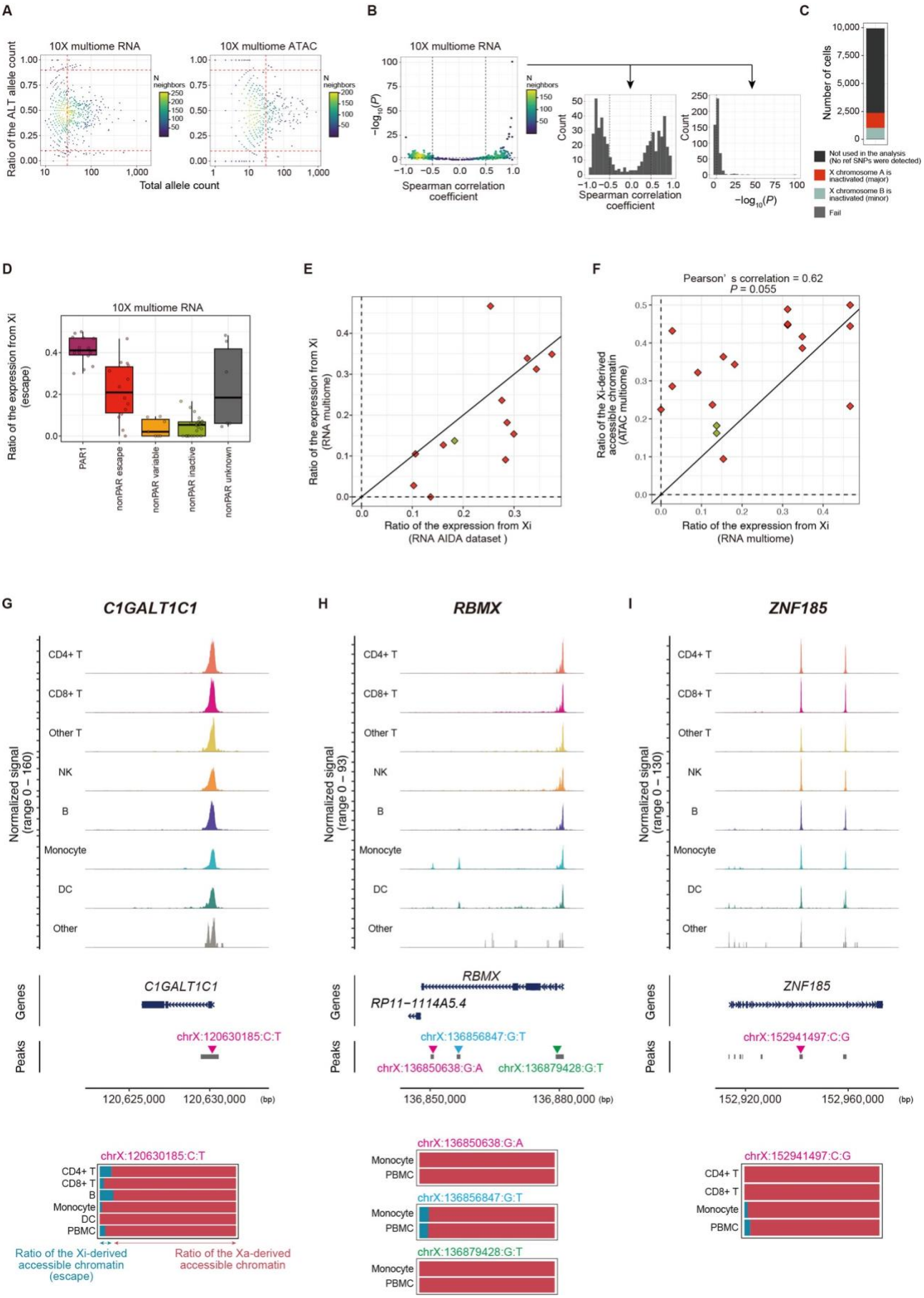
A
10X multiome RNA
10X multiome ATAC
Ratio of the ALT allele count
Total allele count
N neighbors
B
10X multiome RNA
−log10(P)
Spearman correlation coefficient
Count
C
Number of cells
Not used in the analysis (No ref SNPs were detected)
X chromosome A is inactivated (major)
X chromosome B is inactivated (minor)
Fail
D
10X multiome RNA
Ratio of the expression from Xi (escape)
PAR1
nonPAR escape
nonPAR variable
nonPAR inactive
nonPAR unknown
E
Ratio of the expression from Xi (RNA multiome)
Ratio of the expression from Xi (RNA AIDA dataset)
F
Pearson' s correlation = 0.62
P = 0.055
Ratio of the Xi-derived accessible chromatin (ATAC multiome)
Ratio of the expression from Xi (RNA multiome)
G
C1GALT1C1
H
RBMX
I
ZNF185
CD4+ T
CD8+ T
Other T
NK
B
Monocyte
DC
Other
Normalized signal (range 0 – 160)
Normalized signal (range 0 – 93)
Normalized signal (range 0 – 130)
Genes
Peaks
C1GALT1C1
chrX:120630185:C:T
120,625,000
120,630,000
(bp)
RBMX
RP11−1114A5.4
chrX:136856847:G:T
chrX:136850638:G:A
chrX:136879428:G:T
136,850,000
136,880,000
ZNF185
chrX:152941497:C:G
152,920,000
152,960,000
PBMC
Ratio of the Xi-derived accessible chromatin (escape)
Ratio of the Xa-derived accessible chromatin

Figure S8. Application of scLinaX-multi to the 10X multiome dataset, related to Figure 4.

A, The relationships between the ratio of the ALT allele counts (y-axis) and the total allele counts (x-axis) for each SNP for each modality. The points represent the pairs of sample and SNP, and the colors of the points represent the density of the neighboring points. The red dashed lines indicate the thresholds for QC. **B,** The relationships between the $-\log_{10}$(P-value) (y-axis) and correlation coefficients (x-axis) for the Spearman correlation tests for the pseudobulk ASE profiles generated during scLinaX-multi workflow (left). The points represent the pairs of two pseudobulk profiles, and the colors of the points represent the density of the neighboring points. The dashed lines indicate the thresholds for QC. Histograms for the $-\log_{10}$(P-value) (middle) and correlation coefficients (right) of the Spearman correlation tests are also indicated. **C,** A bar plot represents the number of the cells that are not used for the analysis (black), different X chromosomes are inactivated (red/blue), or are removed from the analysis due to the bi-allelic expression of the reference SNPs (grey). **D,** A box plot represents the estimated ratio of the expression from Xi for the gene expression data of the multiome dataset. Genes are grouped according to the XCI status annotated in the previous study. The boxplot indicates the median values (center lines) and IQRs (box edges), with the whiskers extending to the most extreme points within the range between (lower quantile − [1.5 × IQR]) and (upper quantile + [1.5 × IQR]). **E,** A plot represents the concordance of the ratio of the expression from Xi between the AIDA dataset (x-axis) and the multiome dataset (RNA; y-axis). Genes that are annotated as escape genes or shows evidence of escape in the scLinaX analysis (*SEPTIN6*) are indicated. The black line indicates $x = y$. **F,** A plot represents the relationship between the ratio of the expression from Xi (multiome, RNA-level, x-axis) and the ratio of the accessible chromatin derived from Xi (y-axis) for each peak–nearest gene pair. Genes that are annotated as escape genes or shows evidence of escape in the scLinaX analysis (*SEPTIN6*) are indicated. The black line indicates $x = y$. **G,H,I,** The results of the scLinaX-multi for peaks around the representative non-escapee genes, namely *C1GALT1C1* (G), *RBMX* (H), and *ZNF185* (I). Normalized tag counts across cell types are indicated with peak information (top). The ratio of the accessible chromatin derived from Xa and Xi across cell types is indicated as bar plots (bottom) with information on which SNPs are used for the analysis. AIDA, Asian Immune Diversity Atlas; ALT, alternative allele; ASE, allele-specific expression; ATAC, Assay for Transposase-Accessible Chromatin; IQR, interquartile range; PAR, pseudoautosomal region; QC, quality control; REF, reference allele; SNP, single nucleotide polymorphism; Xa, active X chromosome; XCI, X chromosome inactivation; Xi, inactive X chromosome.

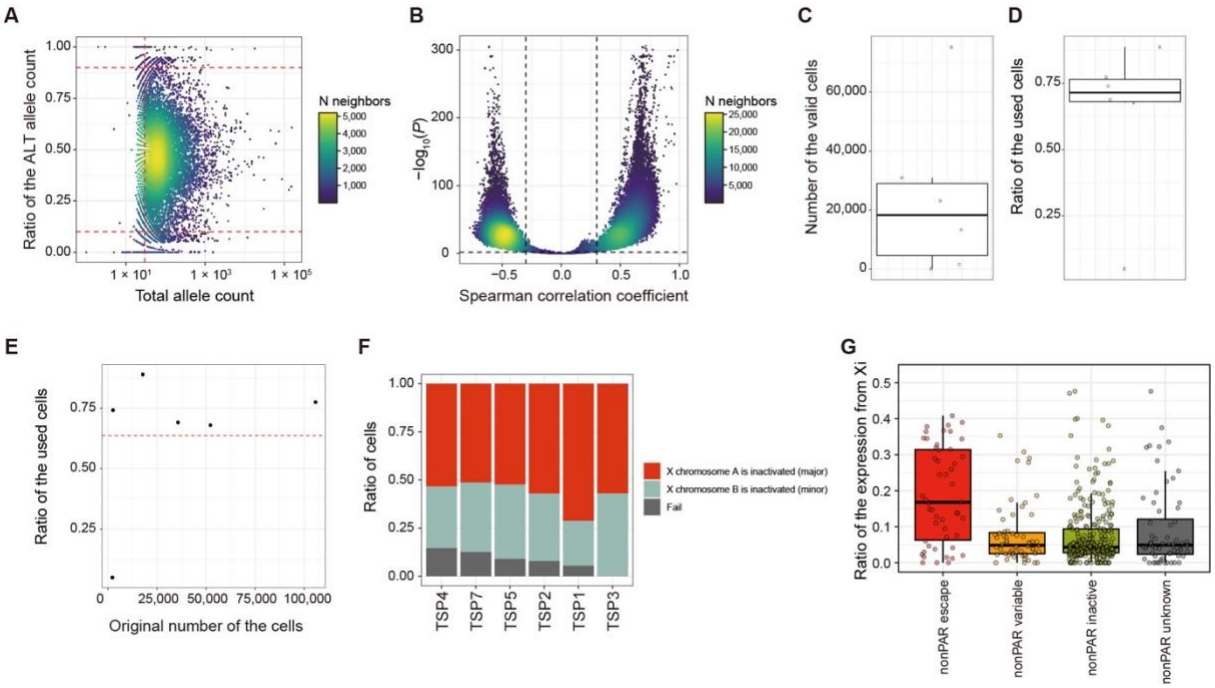

Figure S9. Application of scLinaX to the Tabula Sapiens dataset, related to Figure 5. **A,** The relationships between the ratio of the ALT allele counts (y-axis) and the total allele counts (x-axis) for each SNP. The points represent the pairs of sample and SNP, and the colors of the points represent the density of the neighboring points. The red dashed lines indicate the thresholds for QC. **B,** The relationships between the -$\log_{10}$ (P-value) (y-axis) and correlation coefficients (x-axis) of the Spearman correlation tests for the pseudobulk ASE profiles generated during scLinaX workflow. The points represent the pairs of two pseudobulk profiles, and the colors of the points represent the density of the neighboring points. The dashed lines indicate the thresholds for QC. **C,D,** The boxplots represent the number of valid cells (c) and the ratio of the used cells (cells for which any of the reference SNPs are detected; d) for each sample. The boxplot indicates the median values (center lines) and IQRs (box edges), with the whiskers extending to the most extreme points within the range between (lower quantile − [1.5 × IQR]) and (upper quantile + [1.5 × IQR]). **E,** The relationship between the original number of the cells (x-axis) and the ratio of the used cells (y-axis). The points represent samples. The red dashed lines indicate the mean ratio of the used cells across samples. **F,** A bar plot represents the ratio of the cells that different X chromosomes are inactivated or are removed from the analysis due to the bi-allelic expression of the reference SNPs. **G,** A box plot represents the estimated ratio of the expression from Xi. Genes are grouped according to the XCI status annotated in the previous study. The boxplot indicates the median values (center lines) and IQRs (box edges), with the whiskers extending to the most extreme points within the range between (lower quantile − [1.5 × IQR]) and (upper quantile + [1.5 × IQR]). ALT, alternative allele; ASE, allele-specific expression; IQR, interquartile range; PAR, pseudoautosomal region; QC, quality control; REF, reference allele;

SNP, single nucleotide polymorphism; Xa, active X chromosome; XCI, X chromosome inactivation; Xi, inactive X chromosome.

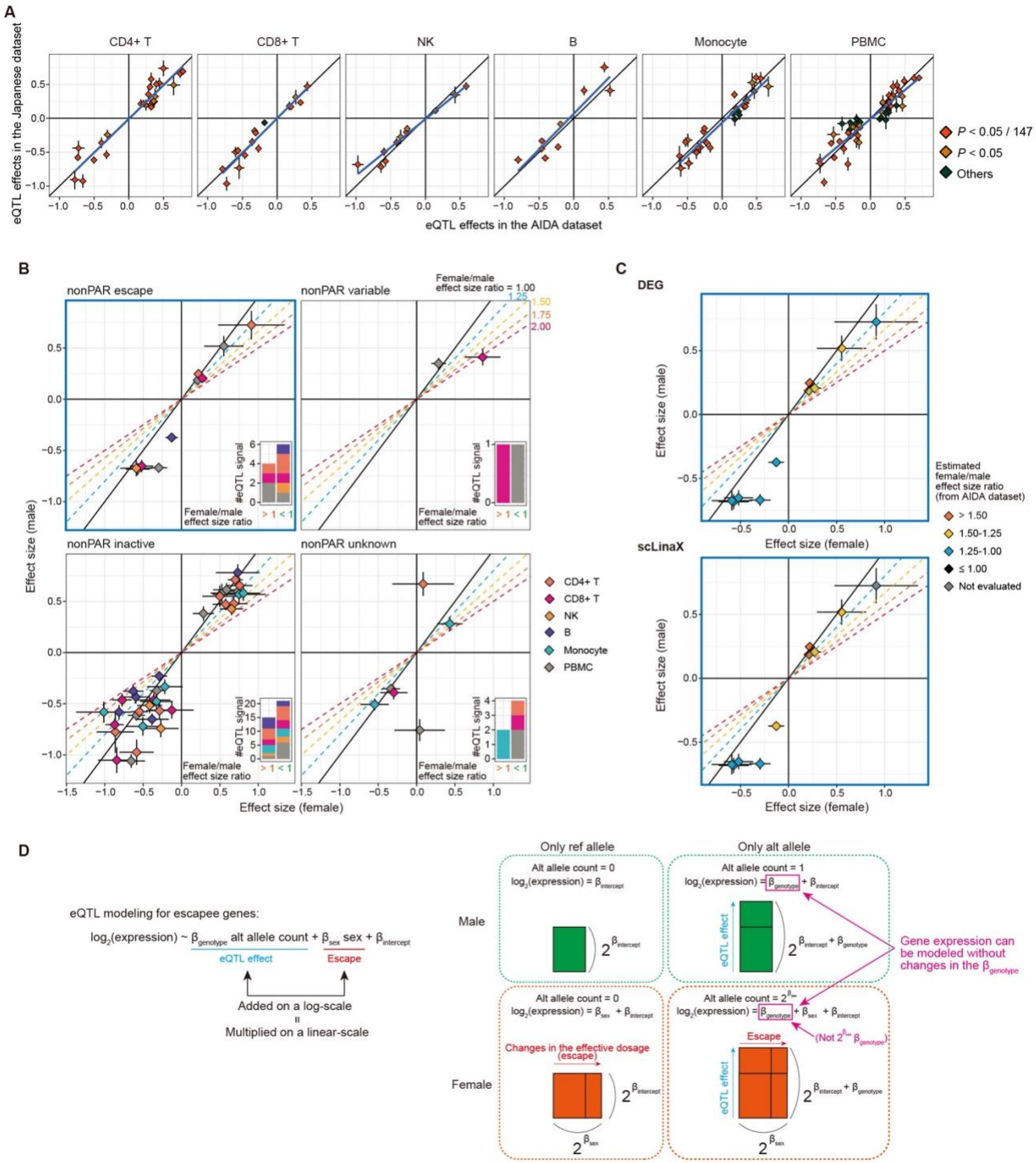

Figure S10. Comparison of the eQTL effect sizes between sexes, related to Figure 6. **A,** The relationship between the eQTL effect sizes between the AIDA dataset (x-axis) and Japanese dataset (y-axis) for the significant eQTL signals detected in the AIDA dataset. The color of the points represents the significance of the eQTL effects in the Japanese dataset ($P < 5 \times 10^{-8}$). The blue line indicates the regression line. **B,** Scatter plots represent the effect sizes of the significant eQTL signals ($P < 5 \times 10^{-8}$; AIDA dataset) in the female-only (x-axis) and male-only (y-axis) analyses with the Japanese dataset, separately for each XCI status.

The error bars indicate standard errors. The color of the plots indicates the cell type in which the eQTL signals are identified. The oblique lines correspond to the female/male effect size ratios described in the plots. The attached bar plots indicate the number of eQTL signals that have larger effect sizes in females (left) and males (right). **C,** The scatter plots for escapee genes (B, upper left) were colored according to the estimated female/male effect size ratio based on the DEG analysis (top) and scLinaX analysis (bottom) with the AIDA dataset. Genes that are not evaluated in the scLinaX analyses are colored grey. **D,** A schematic illustration of the effect of the gene expression normalization method on the eQTL analysis of the escapee genes. DEG, differentially expressed genes; eQTL, expression quantitative trait locus; PAR, pseudoautosomal region; XCI, X chromosome inactivation.

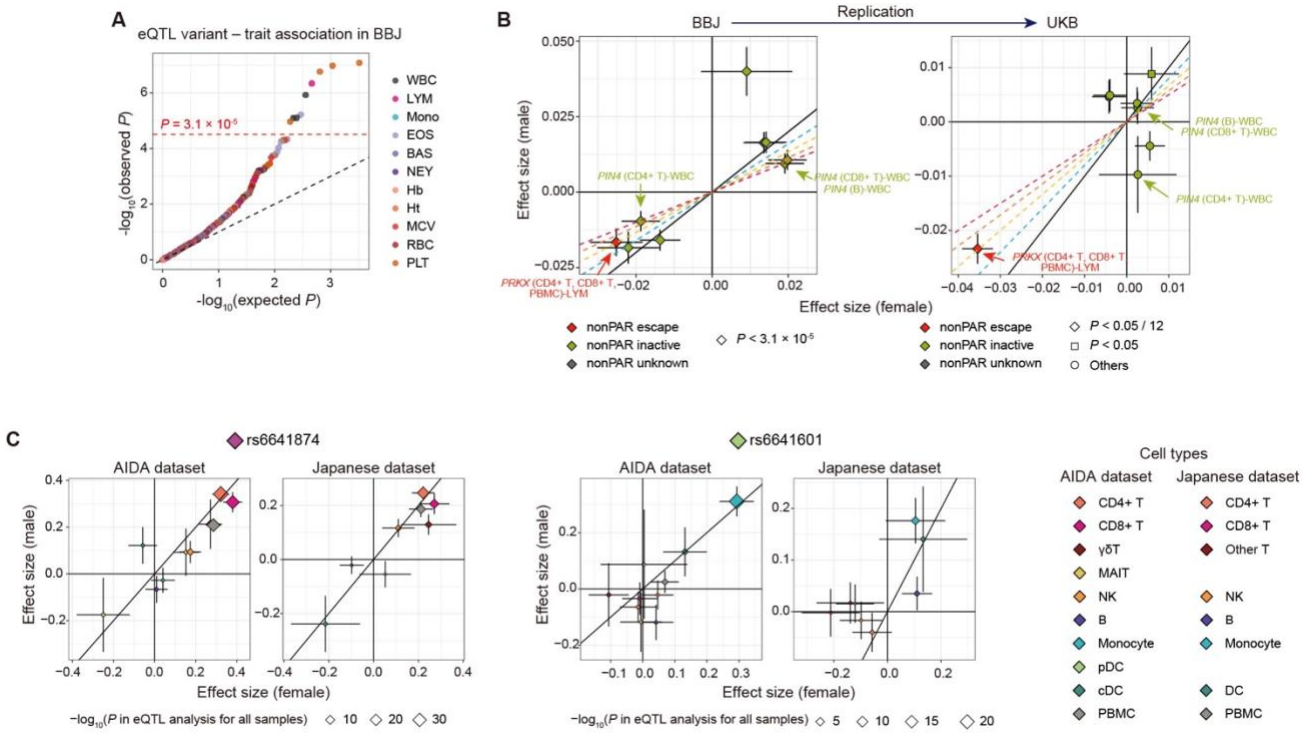

Figure S11. Comparison of the blood-related trait QTL analysis effect sizes between sexes, related to Figure 6.

A, A Q-Q plot represents the association between the significant eQTL variants detected in the AIDA dataset analysis and the blood-related traits in the BBJ cohort. The color of the dots represents the blood-related phenotypes. The red dashed line represents the significance threshold under a multiple-test correction. **B,** The comparisons of the blood-related trait QTL analysis effect sizes between sexes in the BBJ (left) and UKB (right) cohort. The variant–phenotype associations which satisfy the significance threshold in (A) are indicated. The color of the plots represents the XCI status annotated in the previous study. The shapes of the points in the right panel (UKB) indicate whether the significant variant–phenotype association in BBJ is replicated with the UKB dataset. The error bars indicate standard errors. **c,** Scatter plots represent the effect sizes of the two eQTL signals (rs6641874–*PRKX* and rs6641601–*PRKX*) in the female-only (x-axis) and male-only (y-axis) analyses with the AIDA and Japanese dataset. The error bars indicate standard errors. The color of the points indicates the cell type in which the eQTL signals are identified. The size of the points indicates the P-values in the eQTL analysis with all samples. BBJ, BioBank Japan; PAR, pseudoautosomal region; QTL, quantitative trait locus; UKB, UK Biobank.

Supplemental data

Data S1. Overview of the scLinaX method, related to Figure 2.

In Step 1, cells expressing each reference SNP are grouped, and pseudobulk allele-specific expression (ASE) profiles are generated. scLinaX has the option to remove known escapee genes from the pseudobulk ASE profiles (throughout this paper, this option was set as active). The definitions of alleles 1 and 2 are different across cells depending on which allele of the reference SNP is expressed by each cell. In Step 2, the correlation between pseudobulk ASE profiles, which are tied to single reference SNPs, is evaluated. When allele 1 is defined as expressing alleles of a reference SNP, the allele counts of other SNPs should be biased toward allele 1 if the reference allele of the SNPs is on the same X chromosome as the reference allele of the reference SNP. For SNPs with the reference allele on a different X chromosome from the reference allele of the reference SNP, the allele count should be biased toward allele 2. Therefore, positive and negative correlation means that the reference alleles of the reference SNPs are on the same strand and different strands, respectively. Based on the results of the correlation analysis, alleles of the reference SNPs are grouped based on which X chromosome these alleles are on. In Step 3, cells are grouped based on which groups of the reference SNPs are expressed. In Step 4, pseudobulk ASE profiles from cells expressing any of the reference SNPs are generated. The definition of alleles 1 and 2 are different across cells depending on which group of the reference SNP allele is expressed by each cell. The ratio of the expression from Xi is defined as the ratio of allele counts from the alleles with a lower allele count.

Data S2. Accuracy of genotype calling from scRNA-seq data, related to Figure 2.

We compared genotype information derived from SNP array and scRNA-seq data with that obtained solely from scRNA-seq data. We examined SNPs in heterozygous status in at least one sample across the AIDA dataset (Figure S3K) or sample-SNP pairs (**Figure S3L**).

Consequently, when ASE analysis was conducted solely on the scRNA-seq data, 127 additional QC-passed SNPs and 2,738 sample-SNP pairs remained compared to when SNP array data was also utilized. Most of these SNPs and sample-SNP pairs were either undetected or had low imputation quality in the SNP array data, while some exhibited relatively high imputation quality (R2 > 0.7), suggesting that genotype calls from scRNA-seq data were generally accurate but might contain occasional errors. To conservatively and accurately create a catalog of escape for each cell type, we prioritized analyses using both SNP array data and scRNA-seq data when genotype data were available throughout this study. However, genotype calls from scRNA-seq were usually accurate, and scLinaX analysis solely based on scRNA-seq data yielded consistent results compared to analyses based on both scRNA-seq and SNP array data (**Figure S3A-J**). These results suggested that scLinaX can be applied to various datasets even when they do not have paired genotype dataset.

Data S3. Case-control comparison of escape, related to Figure 3.

Although no significant association was detected in the current analysis, there were some potentially interesting results. For example, escape of *TMSB4X* in B cells was stronger in SLE patients than in healthy controls (scLinaX estimates were 0.094 and 0.138, respectively for HC and SLE; $P = 0.052$; **Table S7**). *TMSB4X* was located near the *TLR7* whose escape had been extensively studied in the context of SLE[S1,S2]. Since escape can sometimes happen in clusters of neighboring genes[S3], increase of escape of *TMSB4X* may potentially suggest the aberrant escape of *TLR7* in SLE, while further analysis would be warranted.

Data S4. Overview of the scLinaX-multi method, related to Figure 4.

The input of the scLinaX-multi is single-cell multiome ATAC + Gene Expression data. In Step 1, cells are grouped based on which X chromosome is inactivated by applying scLinaX to the gene expression information of the 10x multiome data. In Step 2, pseudobulk allele-specific

chromatin accessibility profiles are generated by summing up the allele-specific chromatin accessibility data of each single cell. The definition of alleles 1 and 2 is different across cells dependent on which X chromosome is inactivated in each cell. The ratio of the Xi-derived accessible chromatin is defined as a ratio of allele counts from the alleles with a lower allele count.

Supplemental references

S1. Wang, J., Syrett, C.M., Kramer, M.C., Basu, A., Atchison, M.L., and Anguera, M.C. (2016). Unusual maintenance of X chromosome inactivation predisposes female lymphocytes for increased expression from the inactive X. Proceedings of the National Academy of Sciences *113*, E2029–E2038. https://doi.org/10.1073/pnas.1520113113.

S2. Souyris, M., Cenac, C., Azar, P., Daviaud, D., Canivet, A., Grunenwald, S., Pienkowski, C., Chaumeil, J., Mejía, J.E., and Guéry, J.-C. (2018). TLR7 escapes X chromosome inactivation in immune cells. Science Immunology *3*, eaap8855. https://doi.org/10.1126/sciimmunol.aap8855.

S3. Balaton, B.P., and Brown, C.J. (2016). Escape Artists of the X Chromosome. Trends in Genetics *32*, 348–359. https://doi.org/10.1016/j.tig.2016.03.007.

Asian Immune Diversity Atlas (AIDA) Network

Atlas assembly authors are arranged by area of contribution and ordered alphabetically by last name.

Single-cell experimental dataset generation leads: Varodom Charoensawan[1,2,3,4,5,6,7], Chung-Chau Hon[8,9], Partha P. Majumder[10,11], Ponpan Matangkasombut[3,12], Woong-Yang Park[13], Shyam Prabhakar[14,15,16], Jay W. Shin[14,17]

Cohort and sample collection leads: Piero Carninci[18,19], John C. Chambers[15], Marie Loh[14,15], Manop Pithukpakorn[6,20], Bhoom Suktitipat[2,5], Kazuhiko Yamamoto[21]

Overall study design and protocol development: Deepa Rajagopalan[14], Nirmala Arul Rayan[14], Shvetha Sankaran[14]

Sample isolation and processing, single-cell experimental data generation: Juthamard Chantaraamporn[1,2,3], Ankita Chatterjee[10], Supratim Ghosh[22], Kyung Yeon Han[13], Damita Jevapatarakul[3,12], Sarintip Nguantad[1,2,3], Sumanta Sarkar[22], Narita Thungsatianpun[3,12]

Sample isolation and processing: Mai Abe[21], Seiko Furukawa[21], Gyo Inoue[21], Keiko Myouzen[21], Jin-Mi Oh[13], Akari Suzuki[21]

Single-cell experimental data generation: Yoshinari Ando[17,18], Miki Kojima[18], Tsukasa Kouno[17], Jinyeong Lim[13], Arindam Maitra[22], Le Min Tan[14], Prasanna Nori Venkatesh[14]

Single-cell experimental data generation and analysis: Murim Choi[23], Jong-Eun Park[24]

Single-cell data analysis up to cell type annotation: Eliora Violain Buyamin[14], Kian Hong Kock[14], Quy Xiao Xuan Lin[14], Jonathan Moody[8], Radhika Sonthalia[14]

Genotype QC and imputation, GWAS summary statistics: Kazuyoshi Ishigaki[25], Masahiro Nakano[21,26], Yukinori Okada[27,28,29,30,31,32], Yoshihiko Tomofuji[27,28,29]

Affiliations

1) Department of Biochemistry, Faculty of Science, Mahidol University, Bangkok 10400, Thailand

2) Integrative Computational BioScience (ICBS) Center, Mahidol University, Nakhon Pathom 73170, Thailand

3) Systems Biology of Diseases (SyBiD) Research Unit, Faculty of Science Mahidol University, Bangkok 10400, Thailand
4) Research Department, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok 10700, Thailand
5) Department of Biochemistry, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok 10700, Thailand
6) Siriraj Genomics, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok 10700, Thailand
7) School of Chemistry, Institute of Science, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand
8) Laboratory for Genome Information Analysis, RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan
9) Graduate School of Integrated Sciences for Life, Hiroshima University, 1-3-3-2 Kagamiyama, Higashihiroshima, Hiroshima 739-0046, Japan
10) John C. Martin Centre for Liver Research and Innovations, Sonarpur, Kolkata 700150, India
11) Indian Statistical Institute, 203 B.T. Road, Kolkata 700108, India
12) Department of Microbiology, Faculty of Science, Mahidol University, Bangkok 10400, Thailand
13) Samsung Genome Institute, Samsung Medical Center, Seoul 06351, Republic of Korea
14) Genome Institute of Singapore (GIS), Agency for Science, Technology and Research (A*STAR), 60 Biopolis Street, Genome, Singapore 138672, Republic of Singapore
15) Nanyang Technological University, Lee Kong Chian School of Medicine, Clinical Sciences Building, Level 18, 11 Mandalay Road, Singapore 308232, Republic of Singapore
16) Cancer Science Institute of Singapore, National University of Singapore, 14 Medical Drive, Singapore 117599, Republic of Singapore
17) Laboratory for Advanced Genomics Circuit, RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan
18) Laboratory for Transcriptome Technology, RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan
19) Genomics Research Center, Fondazione Human Technopole, Viale Rita Levi-Montalcini, 1 - Area MIND, Milano, Lombardy 20157, Italy
20) Department of Medicine, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok 10700, Thailand
21) Laboratory for Autoimmune Diseases, RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan
22) Biotechnology Research and Innovation Council - National Institute of Biomedical Genomics, Kalyani, West Bengal 741251, India

23) Department of Biomedical Sciences, Seoul National University College of Medicine, Seoul 03080, Republic of Korea
24) Graduate School of Medical Science and Engineering, KAIST, Daejeon, Republic of Korea
25) Laboratory for Human Immunogenetics, RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan
26) Department of Allergy and Rheumatology, Graduate School of Medicine, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8654, Japan
27) Laboratory for Systems Genetics, RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan
28) Department of Statistical Genetics, Graduate School of Medicine, Osaka University, 2-2 Yamadaoka, Suita, Osaka 565-0871, Japan
29) Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives, Osaka University, 2-2 Yamadaoka, Suita, Osaka 565-0871, Japan
30) Department of Genome Informatics, Graduate School of Medicine, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8654, Japan
31) Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC), Osaka University, 3-1 Yamadaoka, Suita, Osaka 565-0871, Japan
32) Premium Research Institute for Human Metaverse Medicine (WPI-PRIMe), Osaka University, Suita 565-0871, Japan