Methods:

Allele-specific X-linked gene expression profile generation from RNA-seq

Bulk RNA-seq data from each clone derived from NPCs and cardiomyocytes were processed independently from a series of bam files, one for each clone. Only reads uniquely mapped to the X chromosome were retained.

SNP read counts for X chromosome reads were compiled using **cellsnp-lite**[1]. The tool was run with minimum read count (`minCOUNT`) set to 5, minimum minor allele frequency (`minMAF`) set to 0.05, and minimum mapping quality score (`minMAPQ`) set to 20. Allele frequency information was sourced from the cellsnp-lite documentation [https://cellsnp-lite.readthedocs.io/en/latest/main/data.html#data-list-of-common-snps], which contains 36.6 million SNPs called from the 1000 genome project[2] with a minor allele frequency (MAF) > 0.0005 in hg38.

To derive allele-specific counts, we performed the following steps. Note that SNP processing and filtering were performed using scripts sourced from the **scLinaX**[3] preprocessing vignette [https://ytomofuji.github.io/scLinaX/articles/scLinaX_preprocessing_example.html]:

1. **SNP Count Processing** (`RCODE_process_cellsnp.r`): Raw SNP count matrices generated by cellsnp-lite were processed to extract allele-specific counts for each SNP and cell. The script loaded allele depth (ALT), total read depth (DP), and other allele counts (OTH), then computed reference allele counts (REF = DP - ALT). SNPs were annotated with their genomic coordinates and reference/alternative alleles from the associated VCF file. The processed data was reformatted into a long-format table where each row represents a unique SNP-clone pair, containing columns for the SNP ID, chromosome, SNP position, reference (REF) and alternative (ALT) base pairs, clone barcode, reference allele count, alternative allele count, and other allele count, with missing values imputed as zero

2. **SNP Quality Control** (`RCODE_QC.r`): The processed SNP count data was filtered based on functional annotations obtained from **ANNOVAR**[4]. SNPs were matched to their corresponding genomic annotations, retaining only those located in exonic, intronic, UTR, ncRNA, or splicing regions of a gene. Variants mapping to multiple genes or lacking gene annotations were removed, and the final QC-passed SNP dataset was saved for downstream analysis.

3. **SNP Replicate Aggregation**: SNP counts were summed across technical replicates (denoted by rep1, rep2). In all further analyses, technical replicates were treated as one clone. Finally, annotated SNP dataframes from all clones and cell types were concatenated into a single, comprehensive dataframe for downstream analyses, and each gene was labeled with their X inactivation status sourced sourced from a compilation of studies by Balaton et al, Tukiainen et al, and Oliva et al[5][6][7]. In total, there were 14,039 SNPs from 30 clones across 8 donors and 2 cell types.

XCI ratio calculation and comparison between genotypes

To estimate the proportion of genes that are subject to or escape X chromosome inactivation (XCI) and compare XCI patterns across genotypes, an analytical pipeline was implemented as follows.

First, to reduce the effect of sequencing error, an error proportion for each single nucleotide polymorphism (SNP) was calculated by dividing the lowest read count among the reference (REF), alternate (ALT), and other (OTH) allele counts by the total read count across these alleles. SNPs with an error proportion greater than 0.01 were excluded from further analysis.

Next, the SNP-level XCI proportion was determined by assuming that for each SNP, the allele with the highest read count corresponds to the active X chromosome (Xa), while the allele with the second highest read count corresponds to the inactive X chromosome (Xi). The proportion of Xi reads was calculated as Xi / (Xa + Xi). An exception was made for the *XIST* gene, where the equation was reversed such that the Xi proportion was calculated as Xa / (Xa + Xi), given its known expression pattern from the Xi. Additionally, SNPs with a total read count (Xa + Xi) below the minimum read count threshold (min_read_count = 20) were excluded.

To ensure a robust approximation of the Xi proportion on the gene level, each gene's Xi proportion was calculated as the Xi proportion of the SNP with the highest total read count (Xa + Xi). Subsequently, Xi proportions were averaged across clones from the same donor to derive a donor-specific Xi proportion. Genes were retained for further analysis only if they had data from at least three donors per genotype (XX and XXY); genes that did not meet this criterion were excluded.

To assess differences in Xi proportions between genotypes, a two-sample t-test was performed using the `ttest_ind` function from the **scipy**[8] package with default parameters. The resulting p-values were then adjusted for multiple comparisons using the Benjamini-Hochberg correction, implemented via the `multipletests` function from the **statsmodels**[9] package. All results were compiled into a dataframe. Finally, gene locations were annotated based on Release 47 (GRCh38.p14) from GENCODE[10].