Differential Expression Analysis of Chromosome X RNA-seq Data

Devon Allies

2026-05-17

Table of contents

List of Figures

List of Tables

Overview

This analysis evaluates transcriptional variations on Chromosome X (chrX) between male and female individuals to uncover sex-specific dynamics. This study utilizes a cohort of 12 paired-end RNA Sequencing (RNAseq) samples obtained from distinct geographical populations. By adjusting for population structure, this analysis isolates robust sex-linked genetic expression underlying chromosome-wide regulation mechanisms.

Methods

Data Processing

Raw paired-end sequencing reads were processed using a reproducible, custom Nextflow DSL2 pipeline hosted at github.com/DevonAllies/rnaseq-pipeline. The computational workflow was systemically executed using the following steps:

Step	Tool	Purpose
Quality Control (QC)	FastQC	Assess raw read quality
Adapter trimming	Trim Galore	Remove sequencing adapters
Alignment	HISAT2	Splice-aware alignment to chrX
Assembly	StringTie	Per-sample transcript quantification
Merging	StringTie merge	Consensus transcriptome across samples
Re-estimation	StringTie	Re-quantify against merged Gene Transfer Format (GTF)
QC report	MultiQC	Aggregate quality metrics

The pipeline execution was controlled via the baseline terminal command:

```
nextflow run rnaseq.nf -profile local --hisat2_prefix indexes/chrX_tran
```

Differential Expression

Downstream Differential Expression (DE) modeling was performed at the gene level using the Ballgown R package. To isolate sexual differences, phenotypic sex was treated as the primary covariate. Population origin was strictly embedded as a confounding adjustment variable to account for ancestry-driven batch effects. Statistical significance was controlled using False Discovery Rate (FDR) threshold of $q - value < 0.05$.

Setup and Sample Phenotypes

Sample Population Metrics

Table 2: Sample phenotype information

ids	sex	population
ERR188044_chrX	male	YRI
ERR188104_chrX	male	YRI
ERR188234_chrX	female	YRI
ERR188245_chrX	female	GBR
ERR188257_chrX	male	GBR
ERR188273_chrX	female	YRI
ERR188337_chrX	female	GBR
ERR188383_chrX	male	GBR
ERR188401_chrX	male	GBR
ERR188428_chrX	female	GBR
ERR188454_chrX	male	YRI
ERR204916_chrX	female	YRI

The sample group balances phenotypic sex and regional cohorts across 12 distinct repository entries. The baseline experimental structure is detailed below:

- **YRI Cohort:** Comprises 3 male samples (ERR188044_chrX, ERR188104_chrX, ERR188454_chrX) and 3 female samples (ERR188234_chrX, ERR188273_chrX, ERR204916_chrX).
- **GBR Cohort:** Comprises 3 male samples (ERR188257_chrX, ERR188383_chrX, ERR188401_chrX) and 3 female samples (ERR188245_chrX, ERR188337_chrX, ERR188428_chrX).

Exploratory Analysis

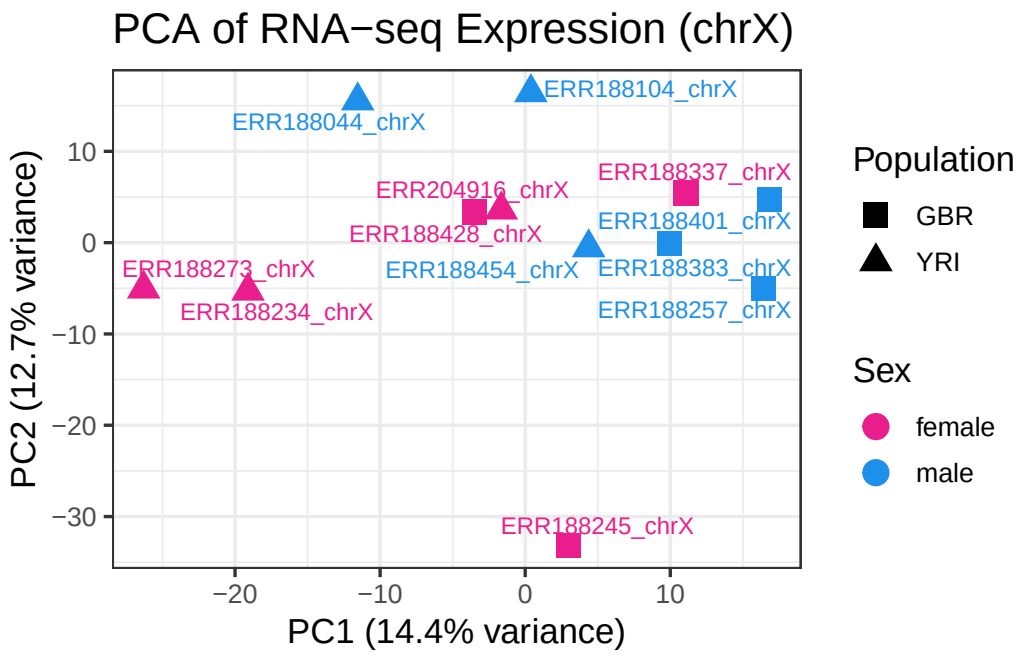

Figure 1: PCA of transcript-level FPKM expression across all chrX samples

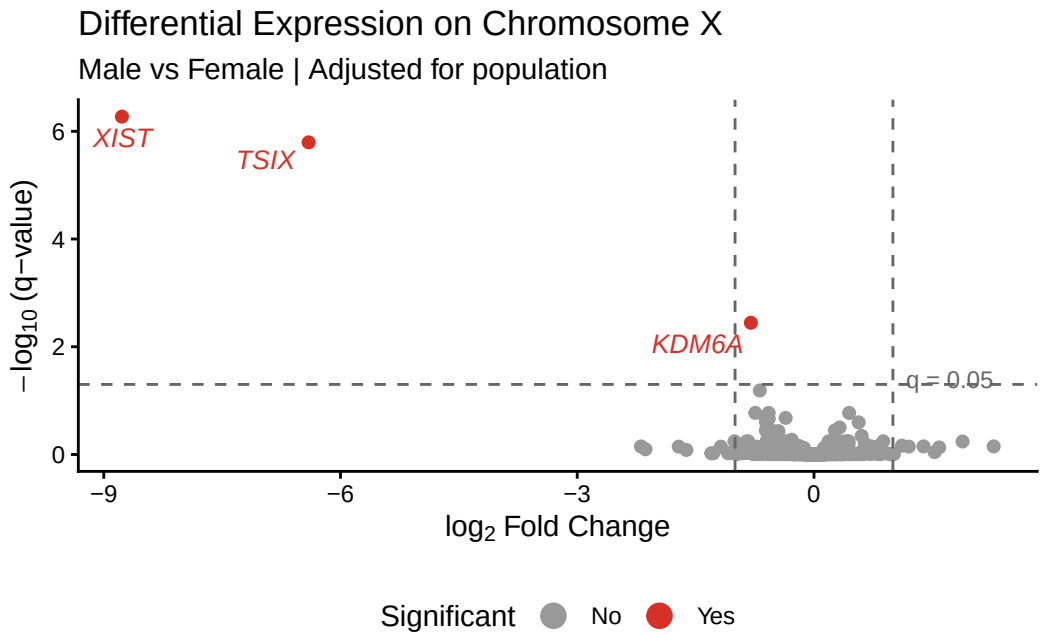

Figure 2: Volcano plots of differential gene expression on chrX. Fold change is calculated male relative to female, adjusted for population (YRI and GBR). Dashed horizontal line indicates q = 0.05 significance threshold.

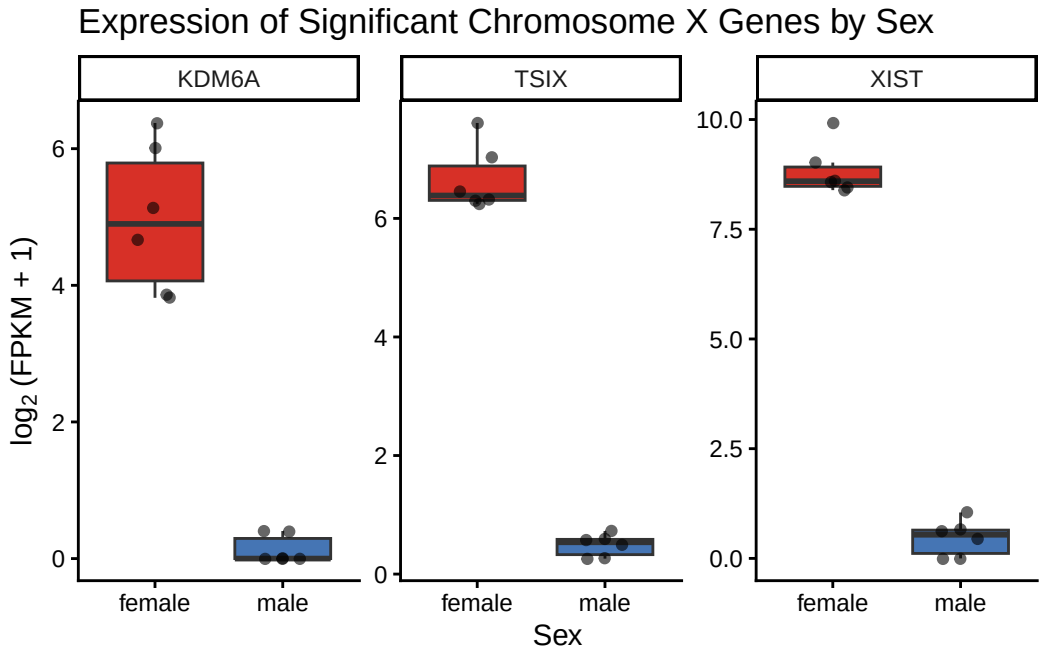

Figure 3: Expression of significant Chromosome X genes by sex (FPKM)

Results

Transcript Filtering

An initial set of 2,343 transcripts was extracted from the StringTie quantification step. To eliminate uninformative background noise and improve statistical power, transcripts with low variance across samples were pruned. Applying a filtration cutoff threshold of $rowVar > 1$ successfully yielded 1,095 high-confidence transcripts retained for exploratory and differential expression processing.

Exploratory Variance Analysis

To understand global relationship structures within the filtered datasets, a Principal Component Analysis (PCA) was calculated on transcript-level Fragments Per Kilobase of transcript per Million mapped reads (FPKM) expression data.

As plotted in Figure 1, the first two principal components explained 27.1% of total variance (PC1: 14.4%, PC2: 12.7%). The distribution patterns reveal a modest separation based on biological sex along the axis trajectories. Concurrently, clear macro-clusters demonstrate underlying population structure, showing distinct spatial partitioning between European (GBR) and African (YRI) cohorts. This population grouping confirms the necessity of utilizing nested model adjustments within the F-test framework to avoid ancestry-driven false-positive discoveries.

Differential Expression

Table 3: Significant differentially expressed transcripts (q < 0.05)

	geneNames	geneIDs	feature	id	fc	pval	qval
1058	TSIX	MSTRG.331	transcript	1058	0.012	9.1e-09	5.0e-06
1059	XIST	MSTRG.332	transcript	1059	0.002	6.9e-09	5.0e-06
435	KDM6A	MSTRG.148	transcript	435	0.032	3.1e-05	1.1e-02

Table 4: Significant differentially expressed genes (q < 0.05)

feature	id	fc	pval	qval
gene	MSTRG.332	0.002	9.3e-10	5.3e-07
gene	MSTRG.331	0.012	5.6e-09	1.6e-06
gene	MSTRG.233	0.574	1.9e-05	3.6e-03

Statistical testing on the curated genomic features revealed a highly localized set of differentially expressed targets. As mapped out in Figure 2, only three specific genes crossed the designated horizontal significance threshold of $q = 0.05$. Fold-change computations reflect expression levels of males relative to females; consequently, negative $\log_2$ fold-change parameters pinpoint strong female-biased gene expression patterns.

The three significant genes identified are:

- **XIST (MSTRG.332):** Shows an extreme female bias with a fold change value of 0.00229 and a significance score ($q = 5.3 \times 10^{-7}$).

- **TSIX (MSTRG.331):** Displays an aligned female bias with an intermediate fold change metric of 0.01181 ($q = 1.6 \times 10^{-6}$).
- **KDM6A (MSTRG.233):** Reaches significant separation parameters with an adjusted fold change value of 0.57437 ($q = 3.6 \times 10^{-3}$).

Gene Abundance Quantification

To validate individual candidate patterns, expression profiles were plotted using $\log_2$ (FPKM + 1) metrics. The resulting boxplots (Figure 3) illustrate clear group distributions:

- **XIST and TSIX Expression:** Both features exhibit definitive binary regulation profiles. Abundance levels are highly elevated in females (with median $\log_2$ values resting between 6.0 and 9.0) but drop down to zero baseline noise across all male samples. This is consistent with their known roles: *XIST* drives female-specific X-Chromosome Inactivation (XCI), while *TSIX* acts as its antisense regulator partner.
- **KDM6A Expression:** Unlike the regulatory Ribonucleic acid (RNA), *KDM6A* is actively transcribed in both sexes. However, female samples present a statistically elevated expression ceiling compared to male profiles. This matches its classification as an XCI-escape gene that regularly maintains dual-dosage activity in female tissue.

Discussion and Conclusion

Biological Implications of Female-Biased Expression

The differential expression analysis of human Chromosome X transcripts revealed a highly localized regulatory signature. Only three genes - *XIST*, *TSIX*, and *KDM6A* - exhibited statistically significant expression differences between sexes ($q < 0.05$). All three candidates demonstrated explicit female-biased abundance. This tight localization underscores the stringent evolutionary conservation governing sex-chromosome dosage mechanisms in humans.

Mechanistic Synergy of XIST and TSIX

The identification of XIST (MSTRG.332) and TSIX (MSTRG.331) as the most statistically significant differentially expressed features aligns with established models of XCI (Lyon, 1961; Galupa and Heard, 2018). In female mammals, dosage compensation between XX and XY individuals is achieved by transcriptionally silencing one of the two female X chromosomes.

- **XIST Function:** This long non-coding RNA (lncRNA) is transcribed exclusively from the future inactive X chromosome (X_i). It physically coats the chromosome in cis, recruiting repressive chromatin-modifying complexes to initiate heterochromatin silencing.
- **TSIX Function:** Operating as an antisense lncRNA partner, TSIX is transcribed from the opposite strand (Lee and Lu, 1999) and overlaps the XIST locus. It acts as a negative regulator of XIST. Prior to inactivation, TSIX expression on the future active X chromosome (X_a) represses XIST accumulation, preserving transcription.

The binary expression profiles captured in the FPKM boxplots (Figure 3) — where both lncRNAs show near-zero baseline expression across all male samples — reflect this strict functional biology. Because males possess only a single, active X chromosome, the XIST/TSIX inactivation locus is never activated.

Dosage Escape by KDM6A

The significant female bias observed in KDM6A (MSTRG.233) highlights a distinct biological phenomenon: XCI escape. KDM6A encodes a histone demethylase responsible for removing repressive H3K27me3 marks, thereby promoting transcriptionally active chromatin states.

Unlike XIST and TSIX, KDM6A maintains active transcription in male individuals. However, its elevated expression in females confirms that it escapes complete epigenetic silencing (Balaton, Cotton and Brown, 2015) on the inactive X chromosome. Because KDM6A is transcribed from both the active and inactive X chromosomes in females, it avoids single-dosage constraint, resulting in a higher steady-state messenger RNA (mRNA) abundance relative to males. This dual-dosage profile may contribute to sex-specific differences in immune responses and oncogenic susceptibility, as KDM6A plays a key role in chromatin remodeling.

Conclusion

By controlling for confounding population structures via Ballgown's nested F-test framework, this study isolated true sex-linked transcriptional signals from geographic background noise. The identification of XIST, TSIX, and KDM6A validates the RNAseq processing pipeline, as these targets represent the canonical core of human X-chromosome regulation. The clear separation of population cohorts in the exploratory PCA highlights the necessity of accounting for genetic ancestry in multi-population expression studies to prevent false-positive discoveries.

List of Acronyms

DE Differential Expression
FDR False Discovery Rate
FPKM Fragments Per Kilobase of transcript per Million mapped reads
GTF Gene Transfer Format
PCA Principal Component Analysis
QC Quality Control
RNA Ribonucleic acid
RNAseq RNA Sequencing
TPM Transcripts Per Million
XCI X-Chromosome Inactivation
chrX Chromosome X
lncRNA long non-coding RNA
mRNA messenger RNA

Reference

Balaton, B.P., Cotton, A.M. and Brown, C.J., 2015. Derivation of consensus inactivation status for X-linked genes from genome-wide studies. *Biology of Sex Differences*, 6, p.35. https://doi.org/10.1186/s13293-015-0053-7.

Galupa, R. and Heard, E., 2018. X-chromosome inactivation: A crossroads between chromosome architecture and gene regulation. *Annual Review of Genetics*, 52, pp.535–566. https://doi.org/10.1146/annurev-genet-120116-024611.

Lee, J.T. and Lu, N., 1999. Targeted mutagenesis of Tsix leads to nonrandom X inactivation. *Cell*, 97(2), pp.249–260. https://doi.org/10.1016/S0092-8674(00)80735-0.

Lyon, M.F., 1961. Gene action in the X-chromosome of the mouse. *Nature*, 190, pp.372–373. https://doi.org/10.1038/190372a0.