Transcription and potential functions of a novel *XIST* isoform in male peripheral glia

Kevin O'Leary [1], Meng-Yen Li [2], Kevyn Jackson [1], Lijie Shi [1], Elena Ezhkova [2], Bernice E. Morrow [1,3], Deyou Zheng [1,4,5,6,*]

1. Department of Genetics, Albert Einstein College of Medicine, Bronx, NY, USA, 10461
2. Department of Stem Cell Biology and Regenerative Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA, 10029
3. Departments of Obstetrics and Gynecology, and Pediatrics, Albert Einstein College of Medicine, Bronx, NY, USA, 10461
4. Department of Neurology, Albert Einstein College of Medicine, Bronx, NY, USA, 10461
5. Department of Neuroscience, Albert Einstein College of Medicine, Bronx, NY, USA, 10461
6. Data Science Institute, Albert Einstein College of Medicine, Bronx, NY, USA, 10461

Supplemental Materials, including Supplemental Methods, List of 10 Supplemental Tables, and 20 Supplemental Figures.

Supplemental Methods

Overexpression of sXIST in the HEK-293T cell line and RNA sequencing

Cell cultures: HEK-293T cells (derived from female embryonic tissue) were cultured in DMEM (Corning, 10-013-CV) with 10% FBS and 1% pen/strep solution (Corning, 30-002-Cl) in 5% CO_2 at 37°C.

Plasmids, lentivirus generation and infection: GFP (pLV-eGFP) was a gift from Pantelis Tsoulfas (Addgene plasmid # 36083) and the full-length human *XIST* (G1A) was a gift from Jeanne Lawrence (Addgene plasmid # 24690). We initially tried PCR amplification of the entire *sXIST* RNA sequence for cloning and transfection, but we were unsuccessful. We then opted to clone the last 3,323 base pair (bp) sequence of *XIST* exon 6 (ChrX:73,820,656-73,823,979(-)). The PCR product was generated using G1A as template and a pair of primers (5'-CTAGGGATCCACTACATGCCCTAGGATATAA-3' and 5'-CTAGACCGGTTTTTCAAAACAGTATATTT-3'). The product was subcloned into pLV-eGFP vector. For ectopic GFP and GFP-s*XIST* expression, HEK-293T cells were transfected with 10μg of pLV-eGFP and pLV-eGFP-s*XIST* lentiviral construct, 10μg packing plasmids (psPAX2, pMD2.G, from Addgene plasmids #12259 and #12260), and 40μl jetPEI transfection regent (Polyplus, 89129-916) according to the manufacturer's instructions. Cells were replaced with fresh DMEM medium after 24 hours post-transfection. Lentivirus-containing DMEM was collected 48 hours post-transfection and filtered through a 0.45μm filter. HEK-293T was cultured with lentivirus-containing media and supplemented with 4μg/ml polybrene (Millipore Sigma, H9268). Virus-containing media were replaced 24 hours post-infection with fresh DMEM, and cells were cultured for an additional 24 hours. Virus-transfected cells were selected with 5μg/mL puromycin (Millipore Sigma, P8833) for 3 days.

RNA purification and bulk mRNA-seq library preparation: GFP and GFP-s*XIST*-expressed (*XIST* OE) cells were collected directly into RLT Plus buffer (QIAGEN, 1053393), and total RNA was isolated with the RNeasy Plus Mini Kit (QIAGEN, 74136) and RNase-free DNase I treatment (QIAGEN, 79254) according to the manufacturer's instructions. RNA quality was measured using a Bioanalyzer and Agilent RNA 6000 Pico kit (Agilent, 5067-1513), and samples with

RNA integrity numbers (RIN) > 8 were used for library preparation. Libraries were constructed from 100ng of total RNA using the Universal Plus mRNA-Seq with NuQuant® (Tecan, 0520-A01). Following the manufacturer's instructions of mRNA elution, mRNA was subjected to fragmentation at 94°C for 8 minutes. First strand, second-strand cDNA synthesis, end repair, adapter ligation, and amplification were carried out following the manufacturer's instructions. The concentration and quality of the libraries were determined using Qubit (Invitrogen, Q32854), Bioanalyzer and Agilent High Sensitivity DNA kit (Agilent, 5067-4626). The libraries were sequenced at GENEWIZ (Azenta) on the Illumina NovaSeq platform, obtaining 150-nucleotide paired-end reads.

Extended information on the collection of human scRNA-seq data:

Mehdiabadi *et al*. performed snRNA-seq on 10 healthy and diseased fetal and child hearts to understand the fetal gene program in pediatric dilated cardiomyopathy (Mehdiabadi et al. 2022). We downloaded the data from GEO (GSE185100) for the left ventricles of healthy children, aged 4 (male), 10 (female), and 14 (male) years-old. Sim *et al.* conducted snRNA-seq to understand age and sex-dependent changes in gene expression during development (Sim et al. 2021). From GEO (GSE156707), we obtained data from left ventricular samples collected from a 35 year-old female, a 41 year-old male, and a 42 year-old male (Sim et al. 2021). The third dataset was downloaded from the Heart Cell Atlas (https://www.heartcellatlas.org/), which were generated originally by Kanemaru *et al*. from a variety of heart regions (apex, left atrium, left ventricle, right atrium, right ventricle, and intraventricular septum) using both scRNA-seq and snRNA-seq (Kanemaru et al. 2023). We used data from 14 heatlhy individuals (7 males and 7 females) between the ages of 40 and 75 years-old.

Differential expression analysis, density plots, gene set enrichment analysis, and overrepresentation analysis

From DESeq2, each gene had associated *p* values, adjusted *p* values, $\log_2$fold change values, etc. We obtained the chromosomal location of each gene using biomaRt and further grouped genes on X/Y Chromosomes to known PAR (Weng et al. 2016) and X escapee genes (Wainer Katsir and Linial 2019). For cilia-related SPC genes in **Figure 5**, we chose to use genes within the “cilium movement” pathway to get the maximal number of cilia-related genes, which was 10. After filtering for genes with normalized DESeq2 expression counts between 1 and the 99.9th percentile value for our overexpression data and 100 for all other data (to eliminate outliers), we created density plots based on log fold change ($\log_2$FC) values from DESeq2 comparisons for individual groups of genes. We also ranked genes by $\log_2$FC to run gene set enrichment analysis (GSEA) using the clusterProfiler R package (Xu et al. 2024) (gseGO function). We followed the same differential expression analysis procedure with bulk RNA-seq data. For overrepresentation analysis of genes that were more strongly correlated with *XIST* in males, we used the ToppFun (Chen et al. 2009) API. Due to the high number of redundant significant GO terms for genes with stronger negative correlation with *XIST* in males, we selected 5 non-redundant, significant (FDR < 0.05) GO terms (eliminating all but one muscle system-related term) for plotting. For genes with greater positive correlation with *XIST* in males, we simply selected the top 5 significant GO terms. For both gene sets, we also selected the top 5 significant terms for categories “cytoband” and “MicroRNA” for plotting.

Transcription factor regulation prediction using SCENIC

For the human heart and skeletal muscle data, we selected the glial cells and used the single-cell regulatory network inference and clustering (SCENIC) R package (Aibar et al. 2017) to predict transcription regulatory programs in the glial subpopulations. SCENIC uses GENIE3 (Huynh-Thu et al. 2010) to infer co-expression modules between TFs and target genes followed by Rcis Target (Aibar et al. 2017) to determine whether the TF binding motif is enriched among co-expressed genes. AUCell (Aibar et al. 2017) is then used to create regulon activity scores for each cell. After running SCENIC, we created a heatmap to show the regulon activity for all regulons with a score > 0.01 for either myelinating or non-myelinating glia. We extracted transcription factors predicted to regulate *XIST* and plotted their gene set activity using AUCell (Aibar et al. 2017).

Supplemental Tables and Figures.

Supplemental Tables:

Supplemental Table S1: Summary of public datasets used in this study

Supplemental Table S2: Number and percentage of human male and female cells expressing *XIST* for all identified skeletal muscle and heart cell types

Supplemental Table S3: Human CELLxGENE data used to determine male and female cell types with a high percentage of *XIST*-positive cells across tissues

Supplemental Table S4: All TFs predicted by JASPAR to bind to the region ChrX:73,841,364-73,841,611 (hg38)

Supplemental Table S5: Full DESeq2 output from XIST OE data

Supplemental Table S6: Summary of cardiomyopathy DEGs (by disease type) and their overlap with SPC and SNC genes

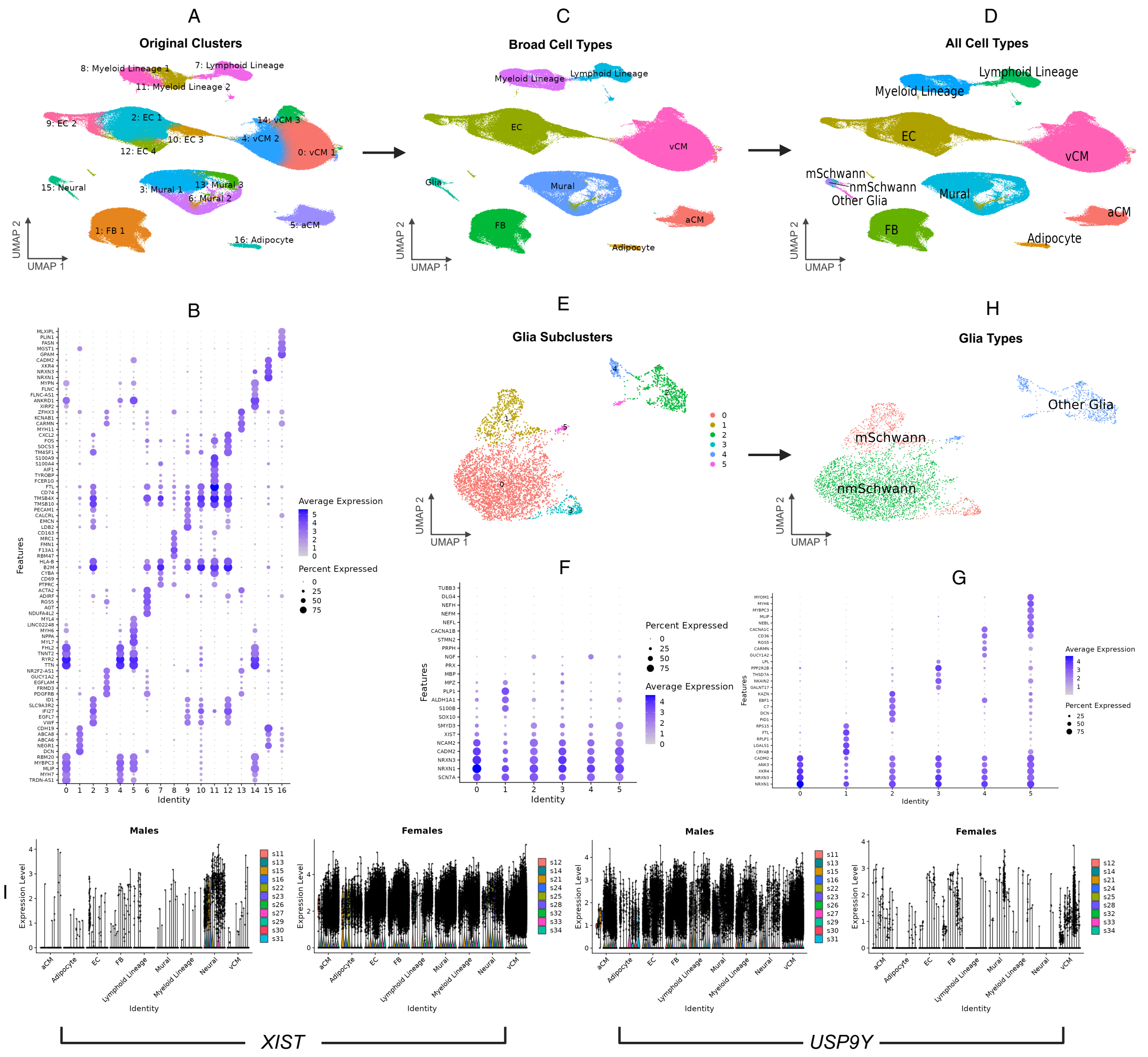
Supplemental Table S7: Data used to determine which polyneuropathy cell types exhibited a significant change in *XIST* expression

Supplemental Table S8: Number and percentage of mouse predicted male and female cells expressing *Xist* for identified peripheral nerve cell types

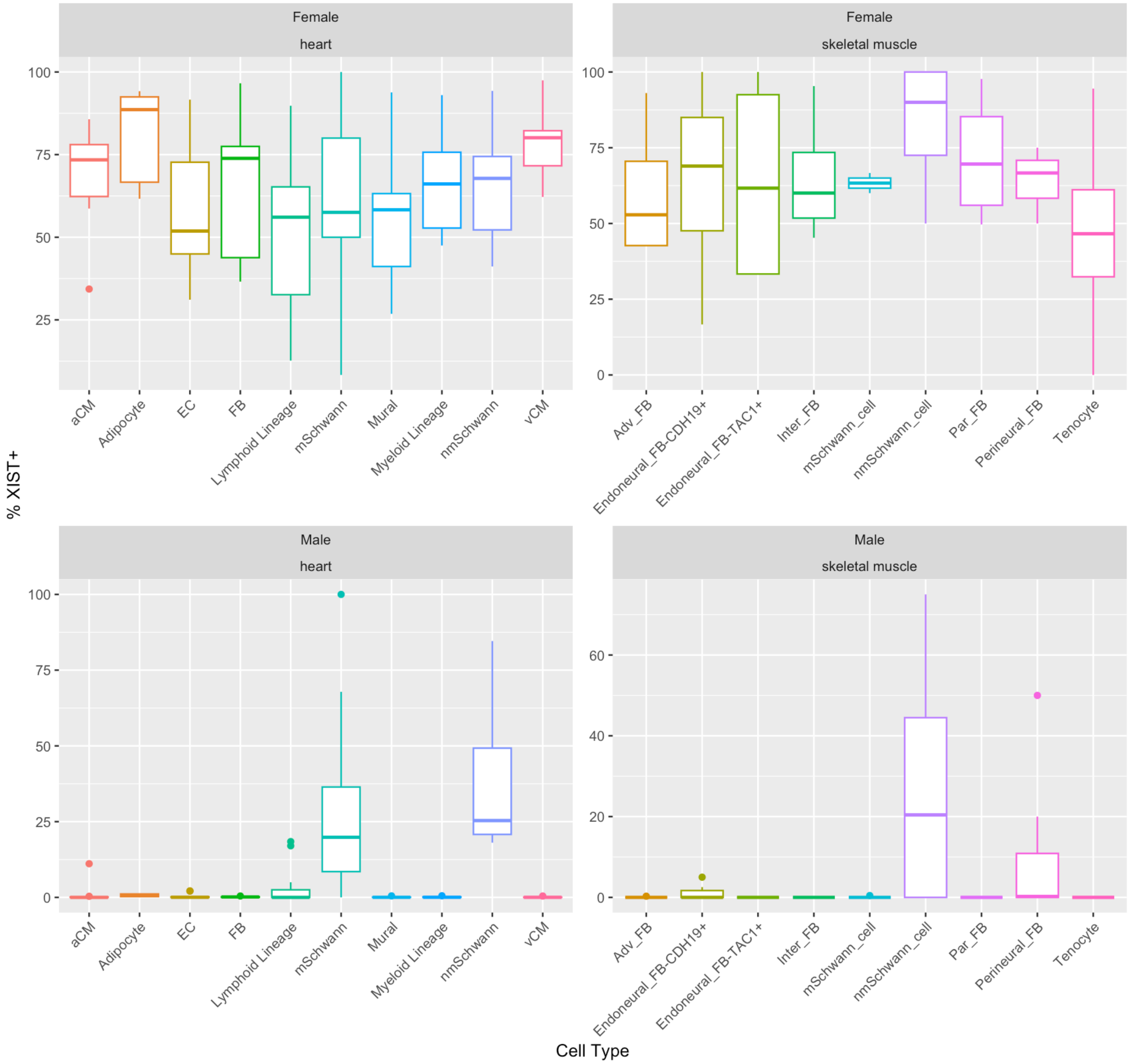
Supplemental Table S9: Summary of differences in Xist expression between mouse non-myelinating (nm) and myelinating (m) Schwann cells (SCs) by predicted sex.

Supplemental Table S10: Mouse CELLxGENE data used to determine male and female cell types with a high percentage of *Xist*-positive cells across tissues

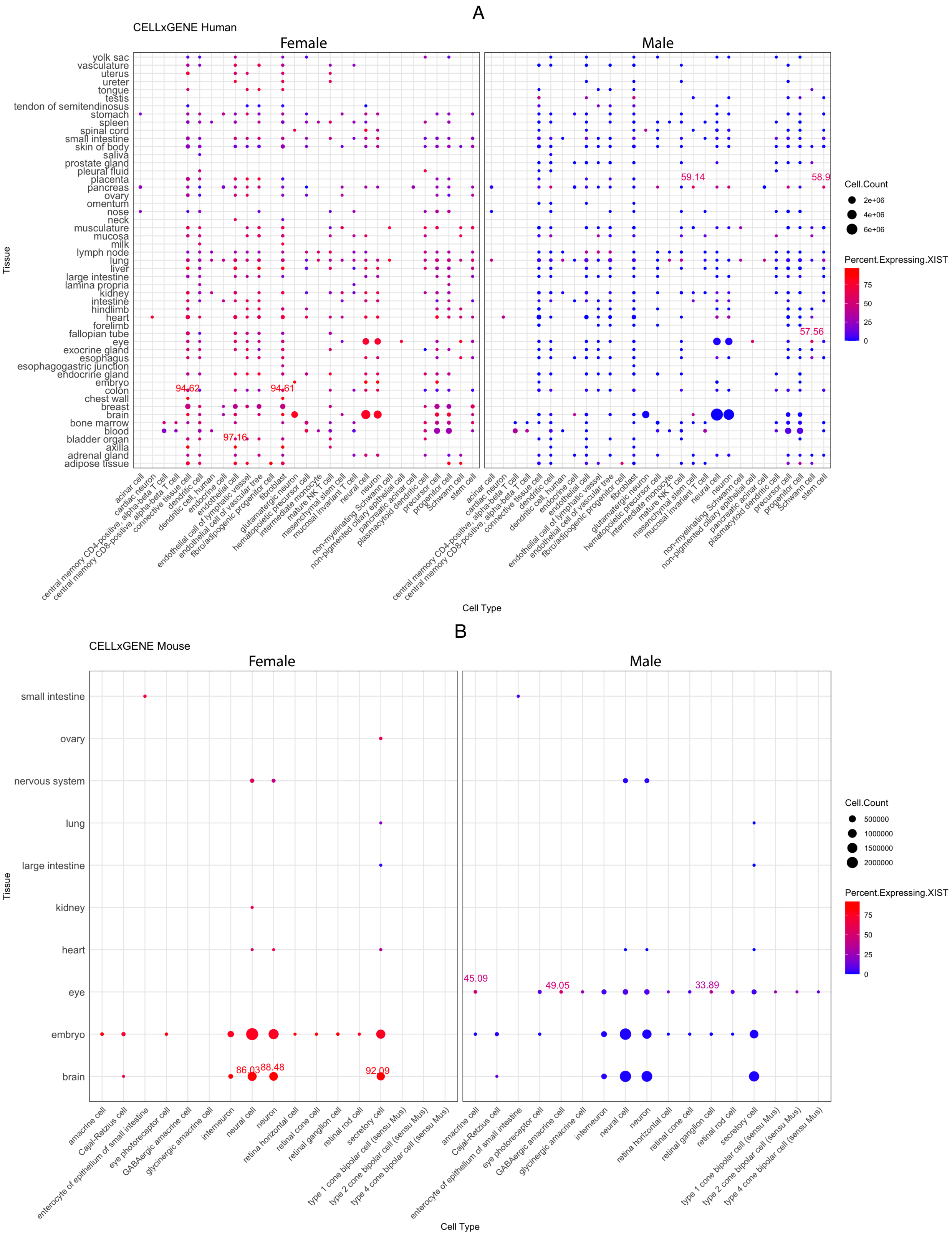
Supplemental Figures:



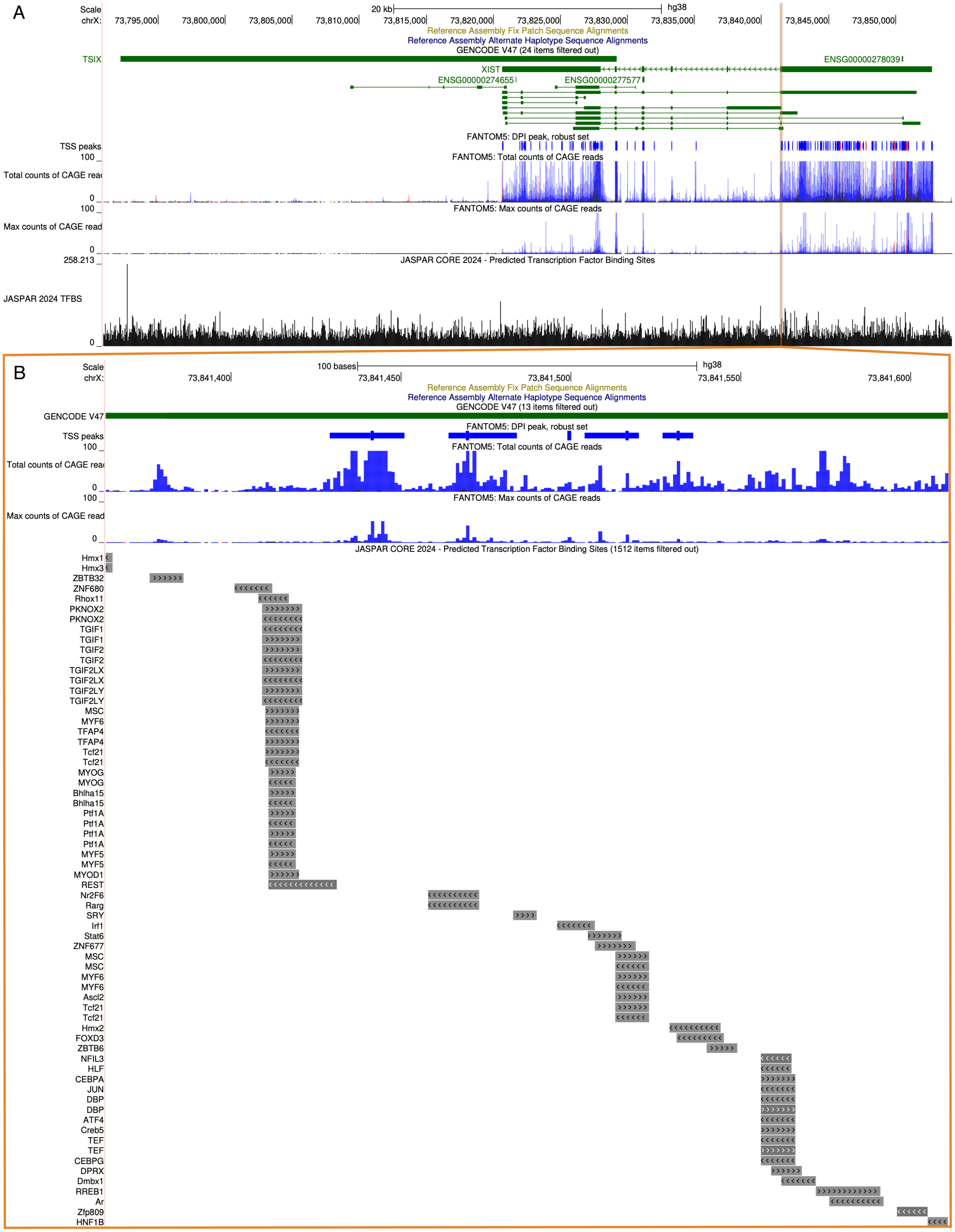
Supplemental Figure S1: Cell type clustering and identification of integrated sc/snRNA-seq data from human hearts. A) Original Seurat clusters **B)** top markers for each Seurat cluster used to create broad cell types in **C. D)** All broad cell types along with identified non-myelinating Schwann cells, myelinating Schwann cells, and other glia. **E)** clustered glial cell subtypes based on markers in **F** and **G**. **H)** finalized glial subtypes. **I)** Sample sex was confirmed by plotting *XIST* and *USP9Y* expression by cell type.



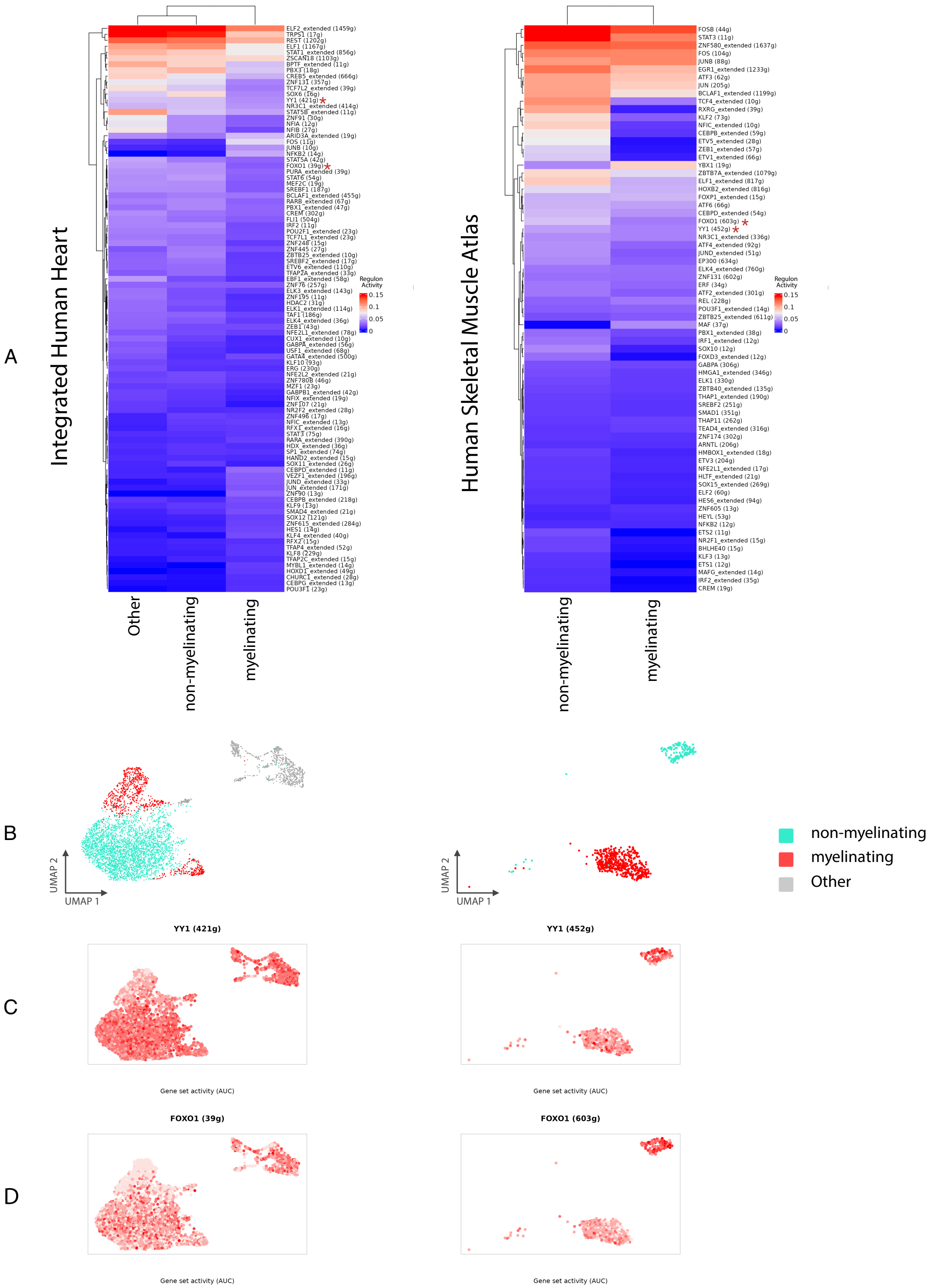
Supplemental Figure S2: Percentage of *XIST*+ cells across cell types from integrated heart and skeletal muscle datasets. Percentages were calculated for each individual then presented as a boxplot.



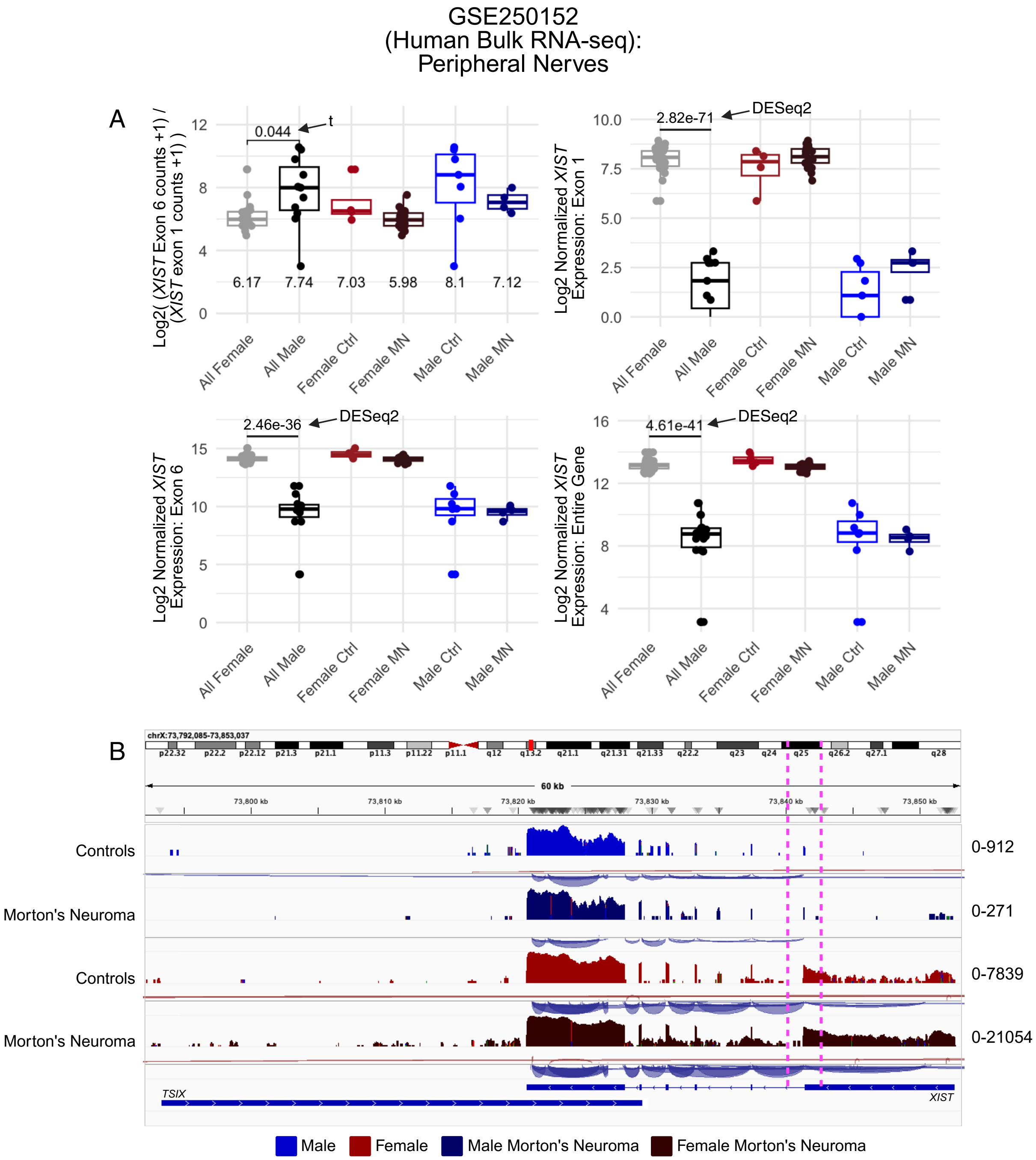
Supplemental Figure S3: Percentage of cells that are *XIST*+ by cell type across all human (A) and mouse (B) datasets in CELLxGENE



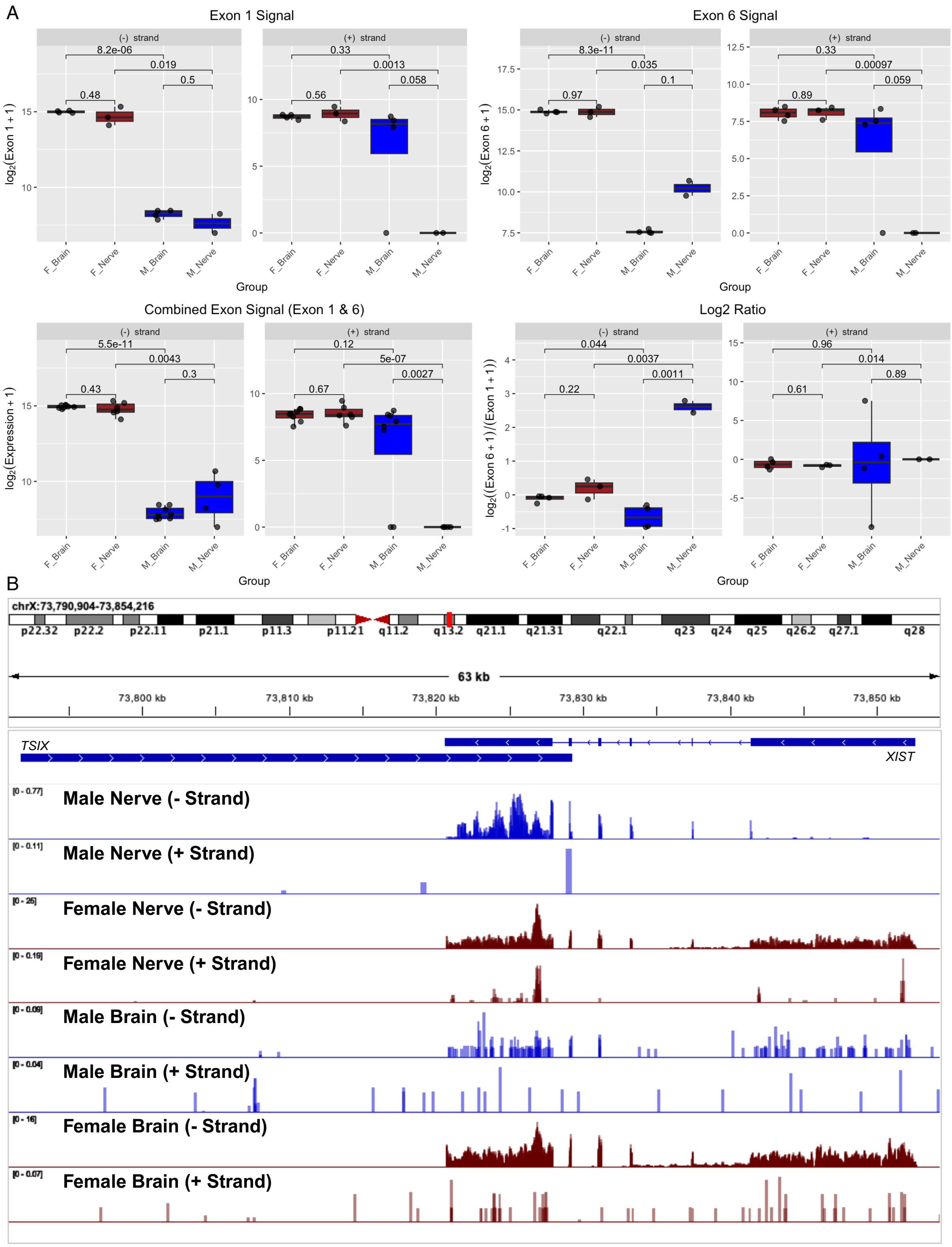
Supplemental Figure S4: *XIST* locus in UCSC genome browser (hg38). A) ChrX:73,790,905-73,854,216, representing the entire *XIST* locus and part of *TSIX* with FANTOM5 and JASPAR tracks displayed. **B)** ChrX:73,841,364-73,841,611, representing the glia-specific peak called from scATAC-seq heart data, with predicted transcription factor (TF) binding sites from JASPAR. Note that only a fraction of predicted TF binding sites is shown here.



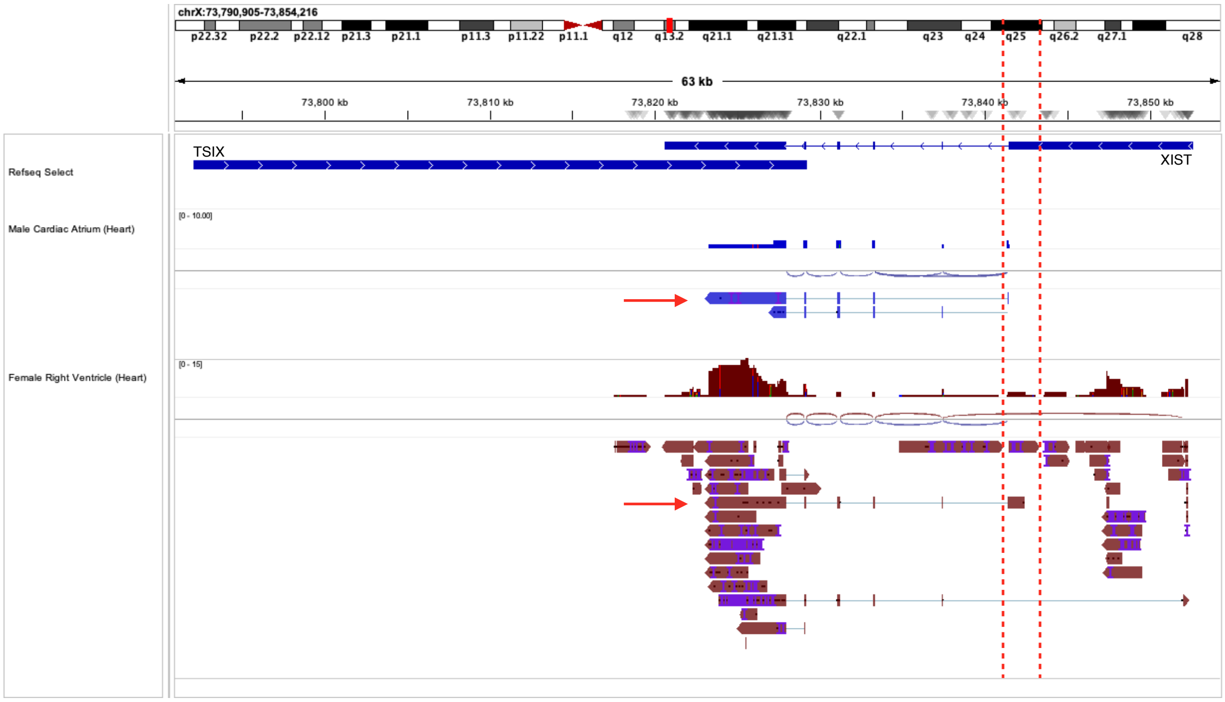
Supplemental Figure S5: SCENIC results from integrated heart and skeletal muscle datasets scRNA-seq datasets. For both integrated and skeletal muscle datasets: **A)** Heatmap of regulon activity for regulons with an activity score > 0.01. **B)** UMAP with non-myelinating, myelinating, and other Schwann cells colored. **C)** Gene set activity for *YY1*. **D)** gene set activity for *FOXO1*.



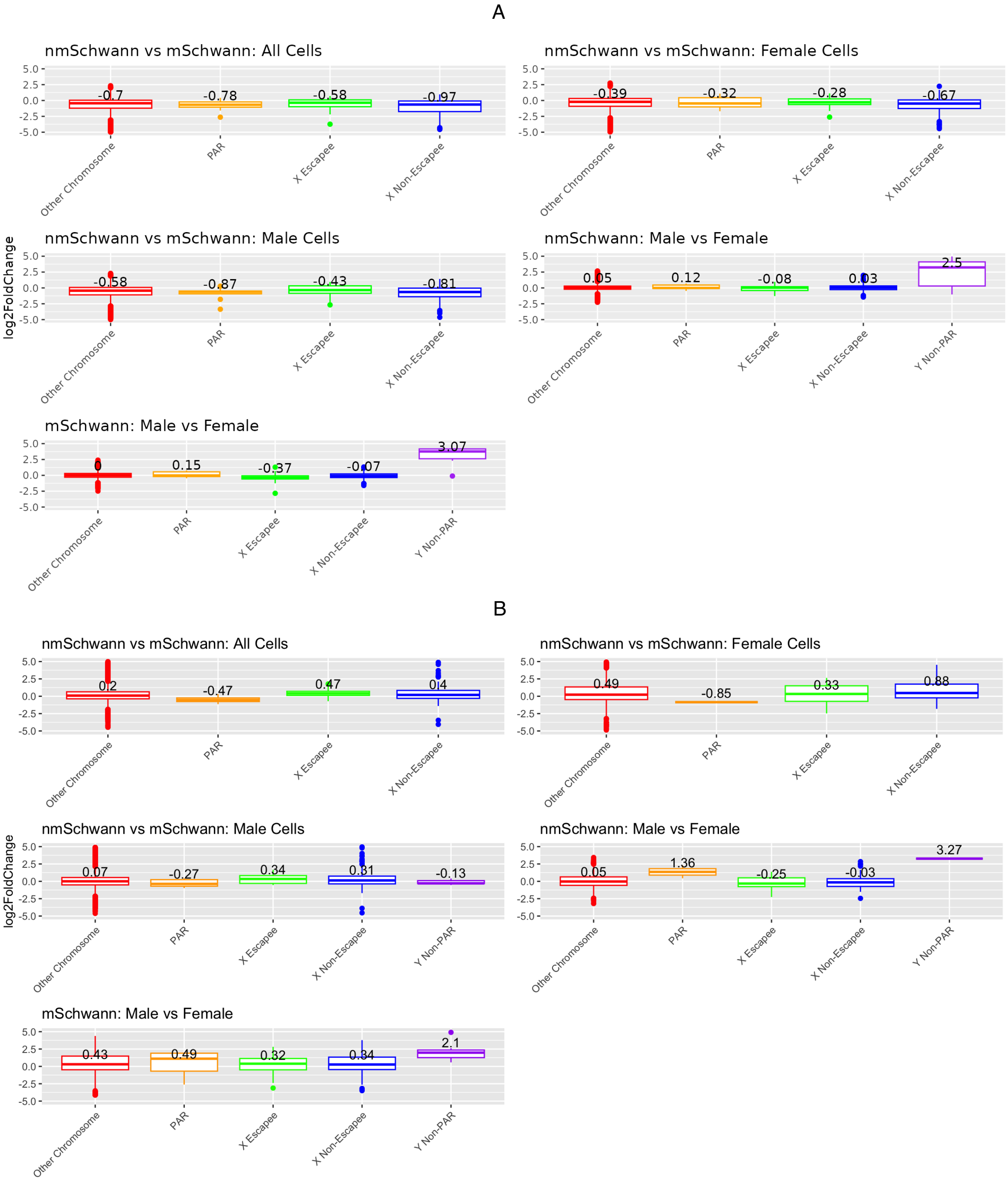
Supplemental Figure S6: Characterization of *XIST* exon 1 and 6 reads from peripheral nerve bulk RNA-seq. **A)** Log$_2$ normalized exon 6 to exon 1 read proportions and expression. **B)** RNA-seq tracks from male (blue) and female (brown) Morton's Neuroma and control peripheral nerve samples. The 3' end of exon 1 is indicated by pink dashed lines. Data ranges for each track are shown on the right.



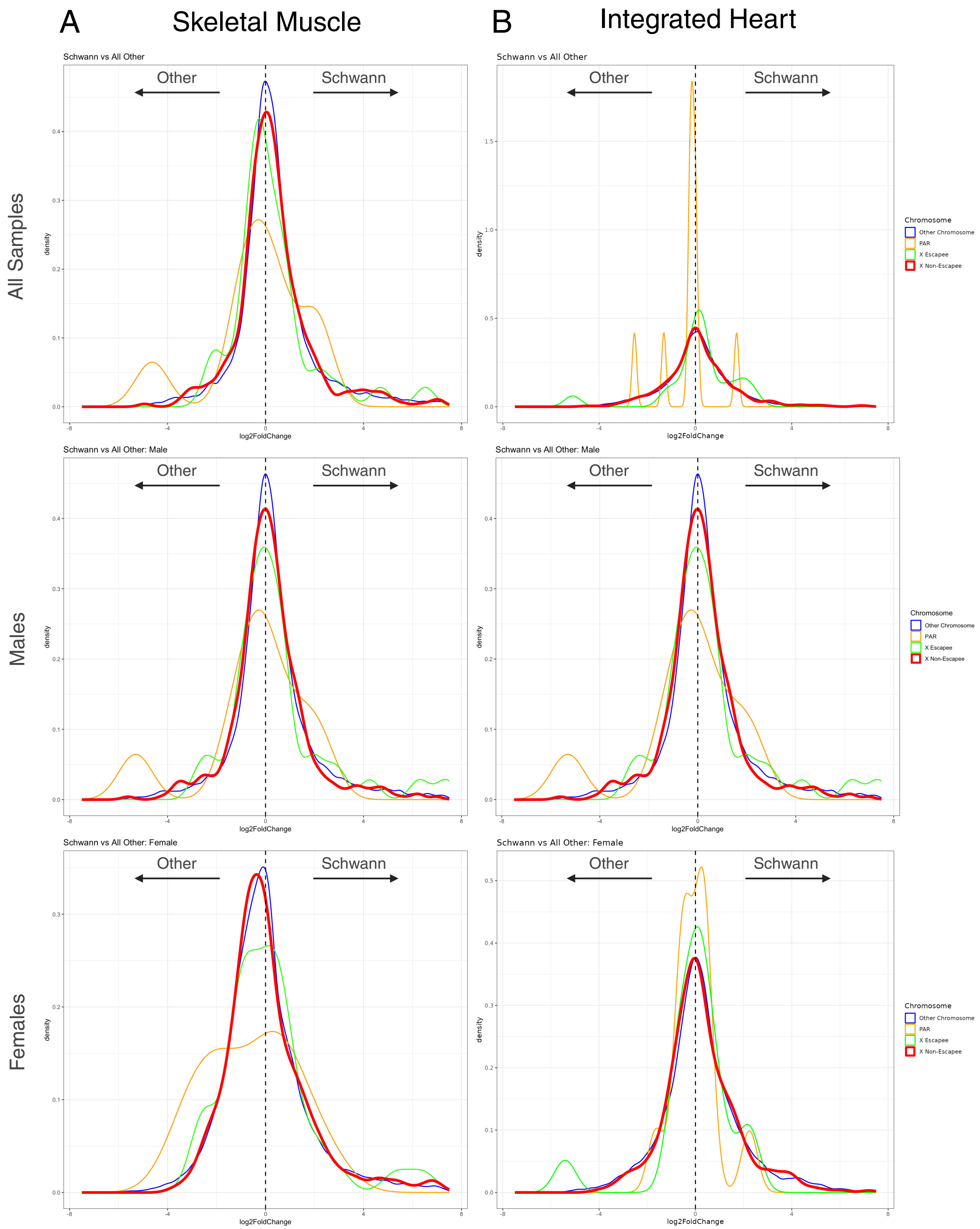
Supplemental Figure S7: **ENCODE nerve and brain bulk RNA-seq data by strand**. **A)** Signal from *XIST* exon 1, exon 6, exon 1 & 6, and exon 6 to exon 1 ratio by tissue, sex, and strand. Total signal was determined by multiplying the average signal using bigWigSummary by the length of the region. **B)** BigWig RNA-seq tracks for male and female brain and nerve samples by strand. Tracks for each sample, strand, and tissue combination were overlayed using IGV.



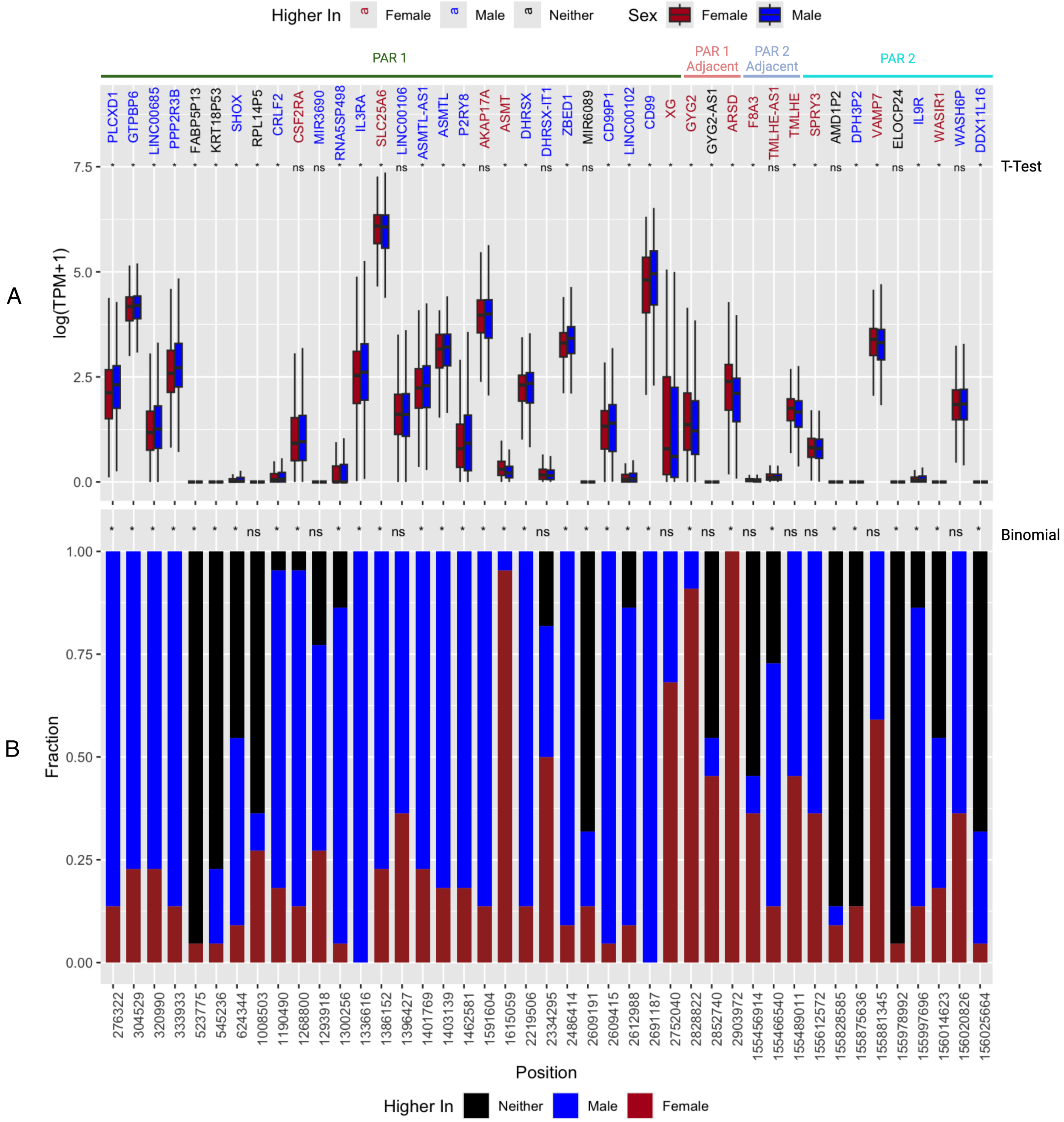
Supplemental Figure S8: Long-read RNA sequencing transcripts over the *XIST* locus in male and female heart samples. Two ENCODE long-read RNA-seq heart samples, one male and one female, showed a transcript that extends from the 3' end of *XIST* exon 1 through most of *XIST* exon 6.



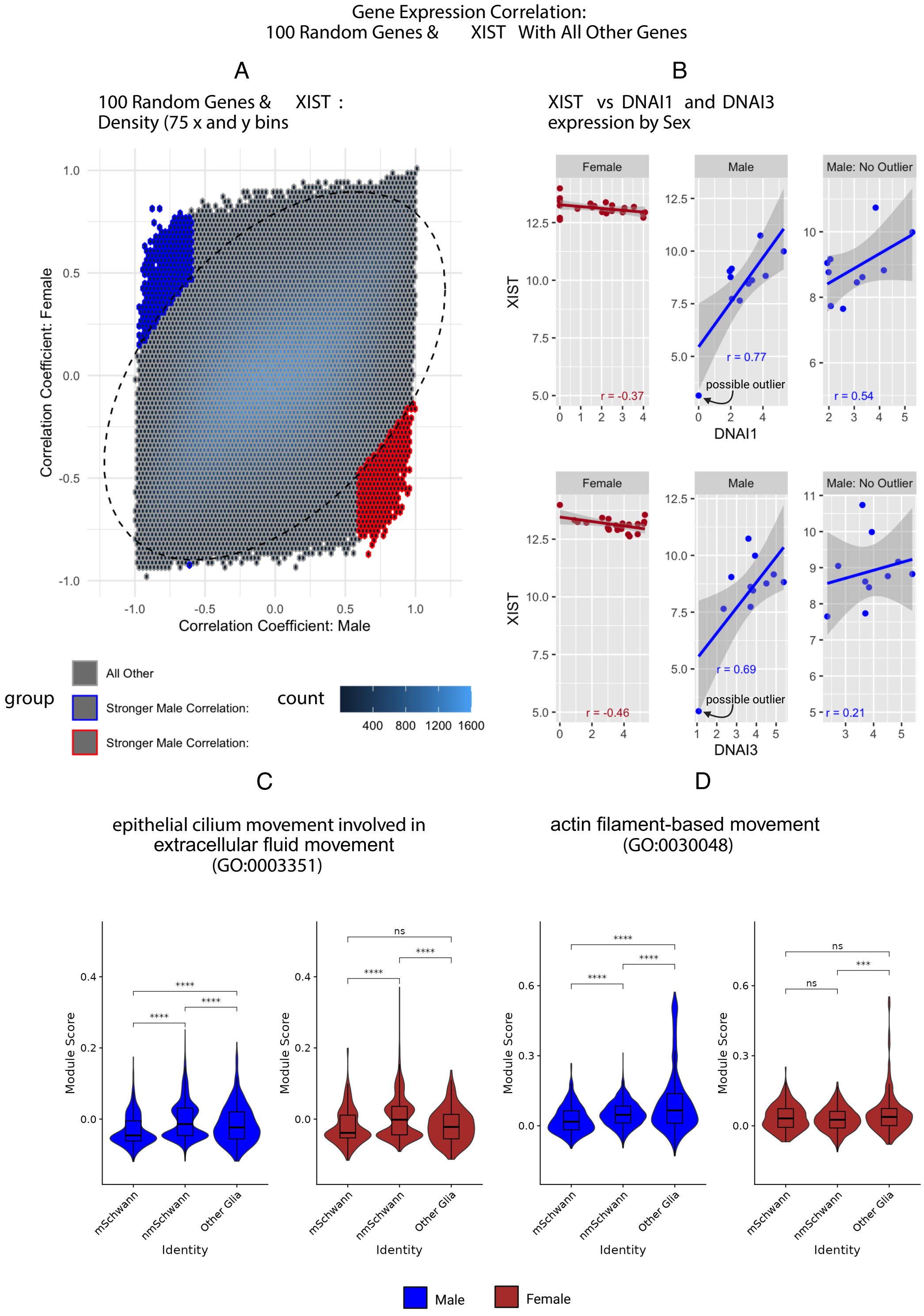
Supplemental Figure S9: Mean $\log_2$FC values for a variety of comparisons from integrated heart (A) and skeletal muscle (B) datasets by gene groups.



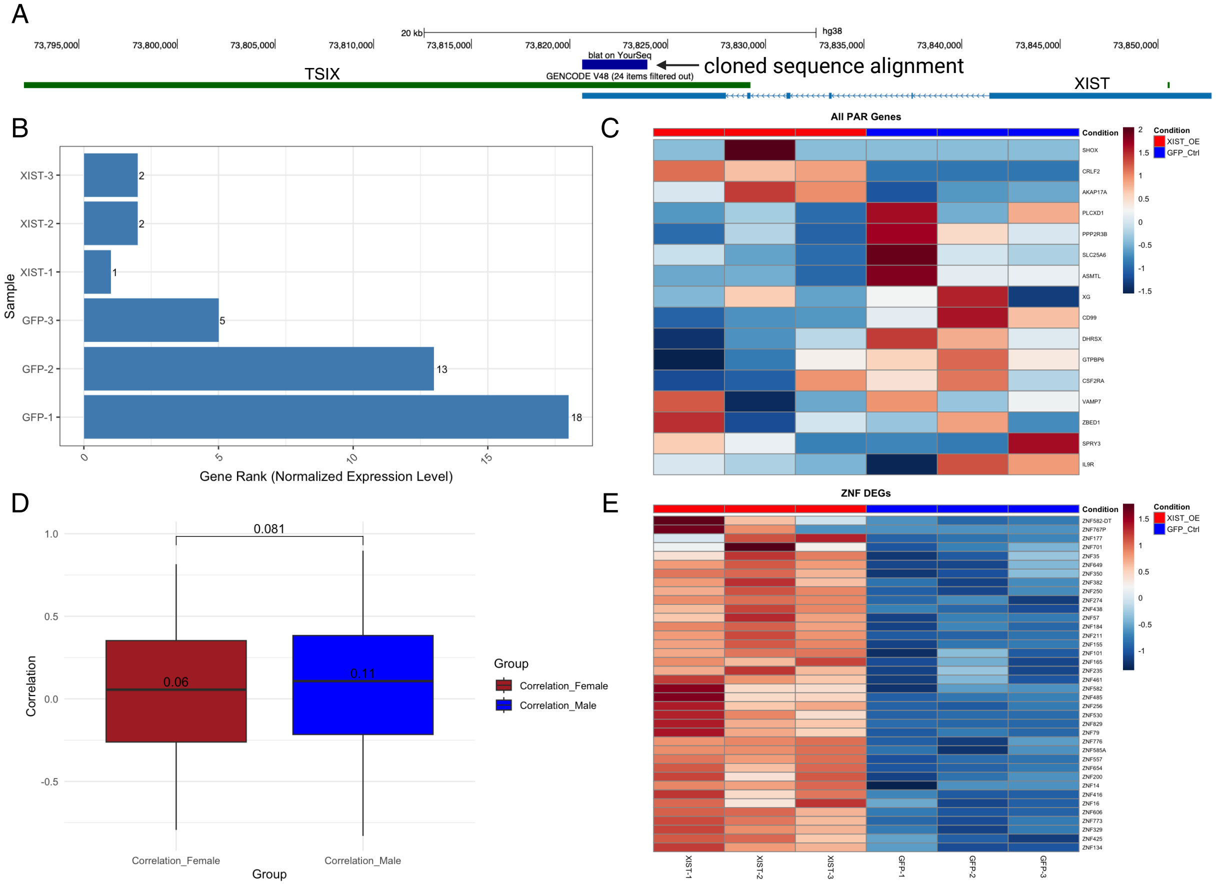
Supplemental Figure S10: Density plots showing log$_2$FC values for Schwann vs all other cells. Higher log$_2$FC values indicate higher expression in Schwann cells while lower values indicate higher expression in all other cells. **A** (left) density plots are from scRNA-seq skeletal muscle data while **B** (right) are from scRNA-seq integrated heart data.



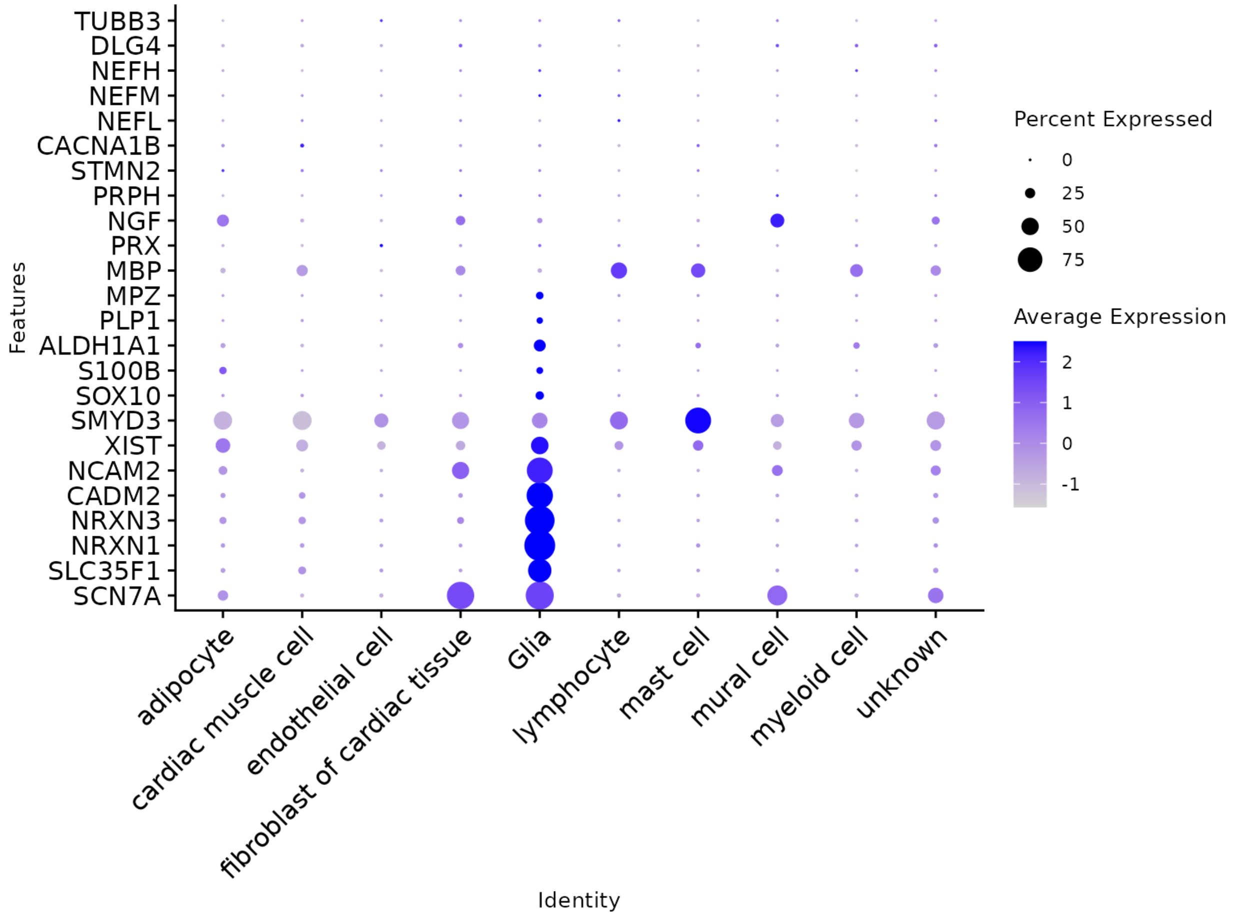
Supplemental Figure S11: Assessment of pseudoautosomal region (PAR) gene expression differences between male and female bulk RNA-seq GTEx data. A) Log(TPM+1) across all tissues and samples for each gene in PAR 1 and 2, as well as their three closest flanking genes. Significant differences between males and females are marked by asterisks. If higher in males, the text for the gene is blue. If higher for females, the text is brown. **B)** Proportion of tissues for which each gene in **A** was significantly higher in males (blue), females (brown), or neither (black). For example, *PLCXD1* was higher in males for about 80% of tested tissues whereas *AMD1P2* showed no difference between males and females for about 80% of tested tissues.



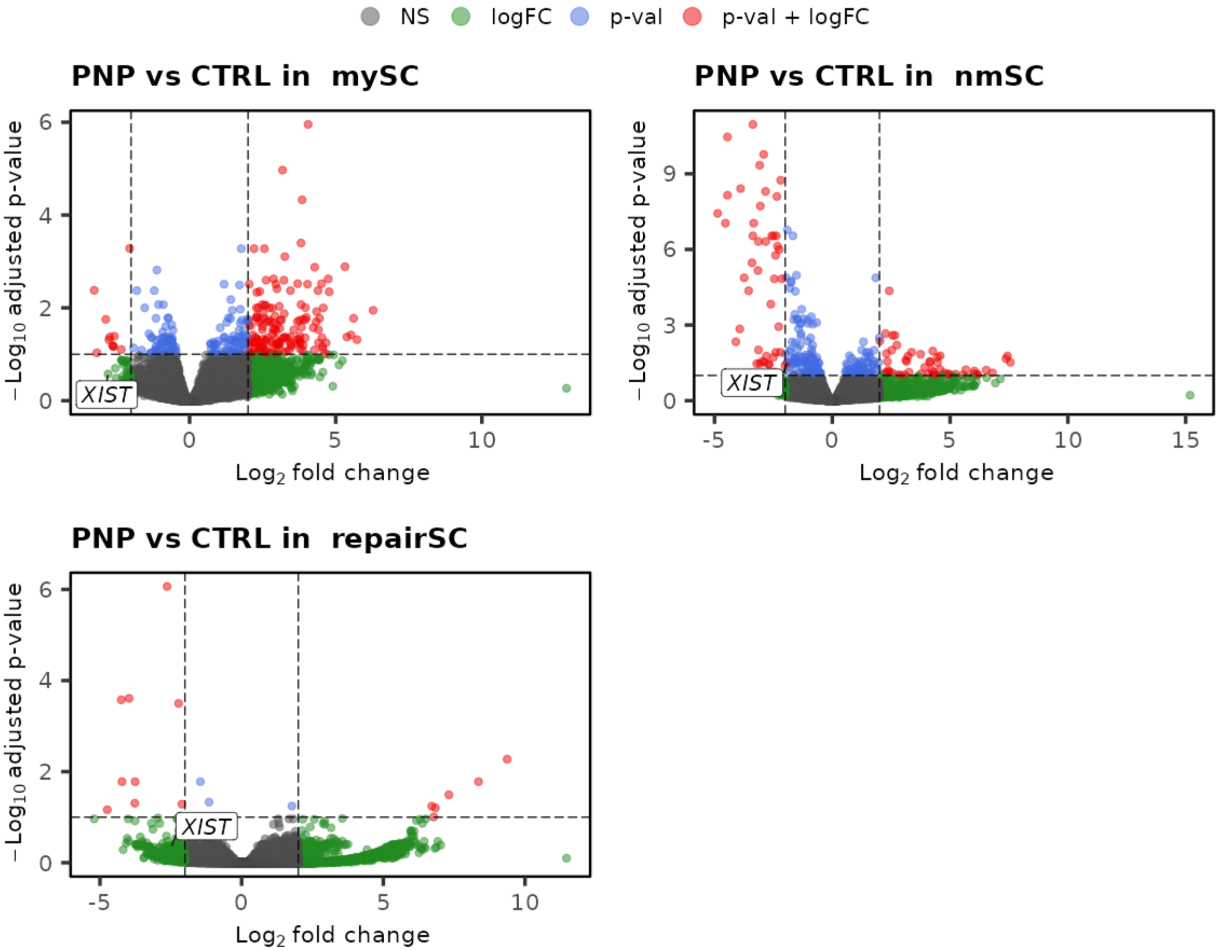
Supplemental Figure S12: Gene expression correlation of 100 random genes with *XIST* and module scores for integrated heart data. A) 100 random genes were correlated with *XIST* separately for males and females. For each gene, their female vs male Pearson correlation values were plotted. An ellipse was fit around 99% of these points and was used as the predicted correlation agreement range for **Figure 4A**. **B)** *XIST* vs *DNAI1/3* expression correlation with correlation coefficients shown. *DNAI1/3* were chosen due to their strong, positive correlations with *XIST* in males. **C)** module score visualization across glial cell types using genes in the GO:0003351 gene set. **D)** module score visualization across glial cell types using genes in the GO:0030048 gene set.



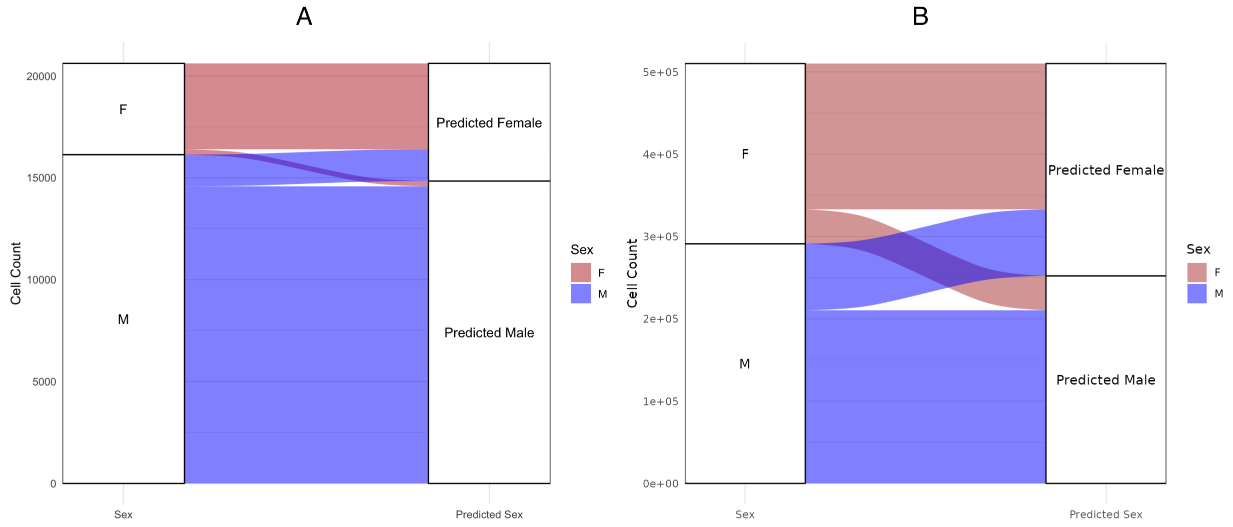
Supplemental Figure S13: *XIST* OE gene ranks, pseudoautosomal region (PAR) gene expression, and cloned sequence location. A) For each sample, *XIST*'s rank based on normalized expression level is shown. *XIST* expression was in the top two for all three *XIST* OE samples. **B)** PAR gene expression scaled across rows, showing a general decrease in these genes' expression in the *XIST* OE group. **C)** The cloned *XIST* exon 6 sequence aligns to the second half of exon 6 (Chr X:73,820,656-73,823,979). **D)** Box plot of Zinc finger (ZNF) gene expression correlations with *XIST* from Morton's neuroma data. **E)** ZNF genes with adjusted DESeq2 p-value < 0.05 and | fold change | > 1.5.



Supplemental Figure S14: Markers used to reclassify cells labeled as "neural" to "glia" from cardiomyopathy data.



Supplemental Figure S15: Volcano plots following differential expression testing between cells from patients with polyneuropathy and controls. *XIST* is labeled in each panel.

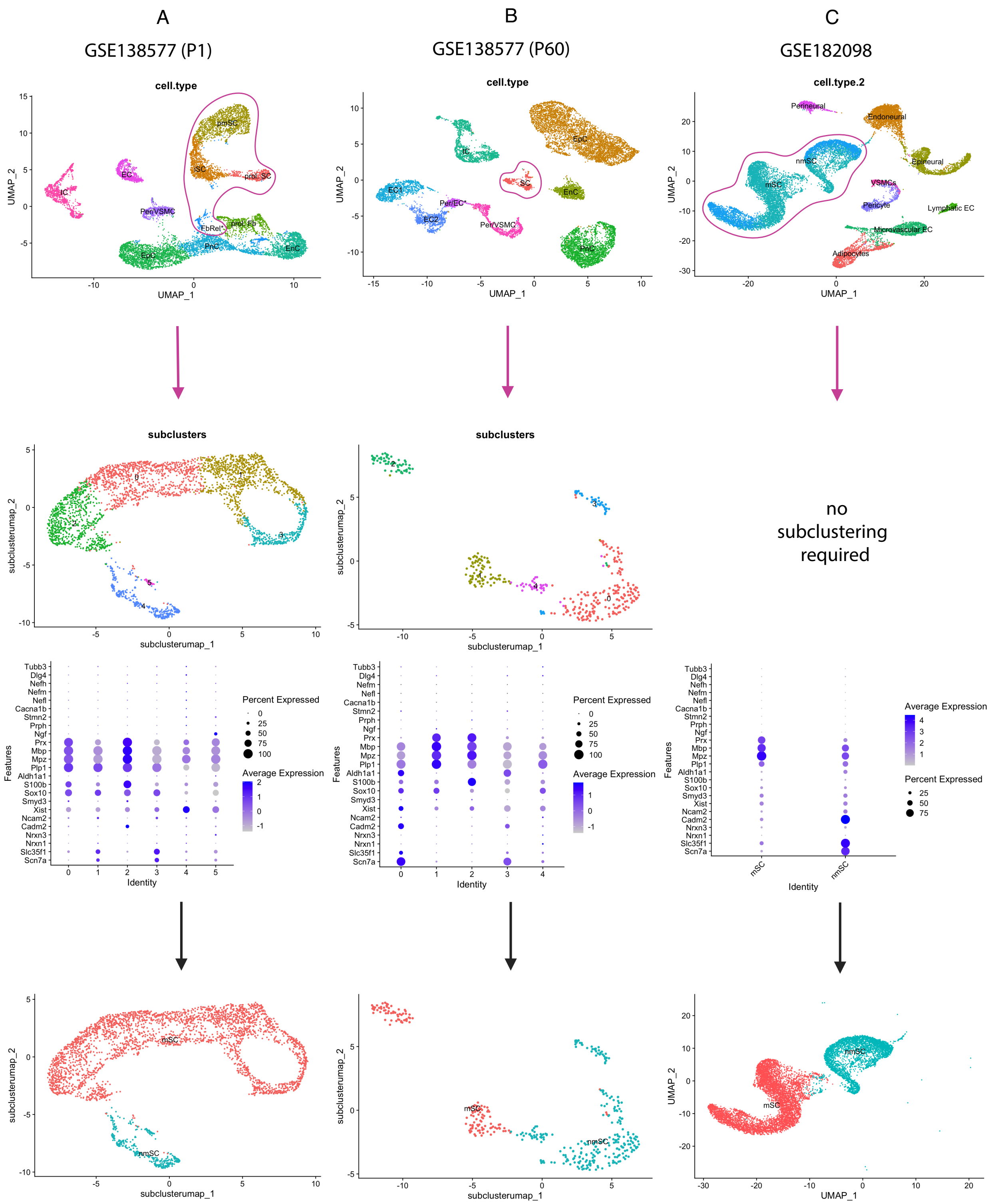


20,611 human skeletal muscle cells:
91.18% Accurate

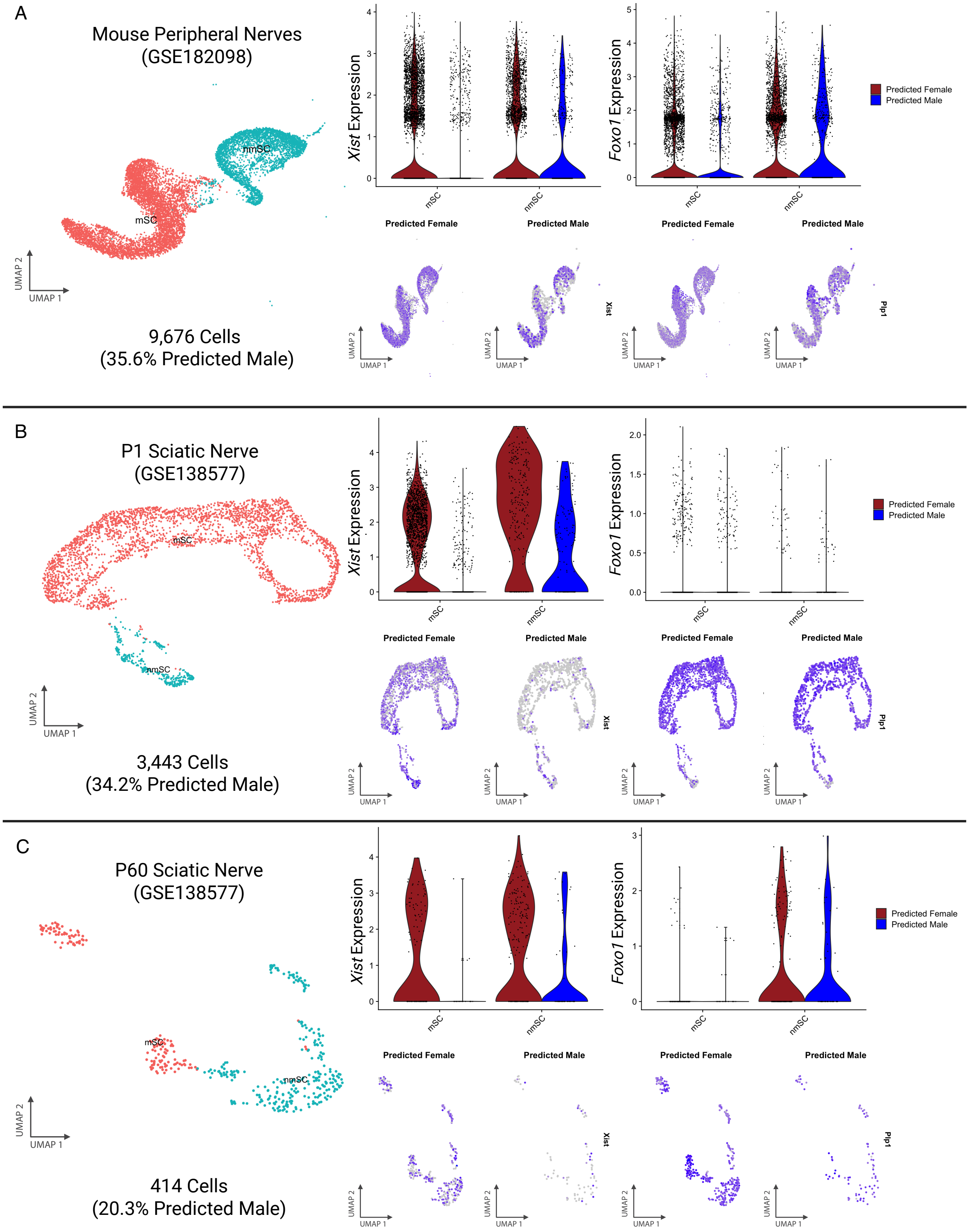
510,171 human heart cells:
75.95% Accurate

Average accuracy = 83.57%

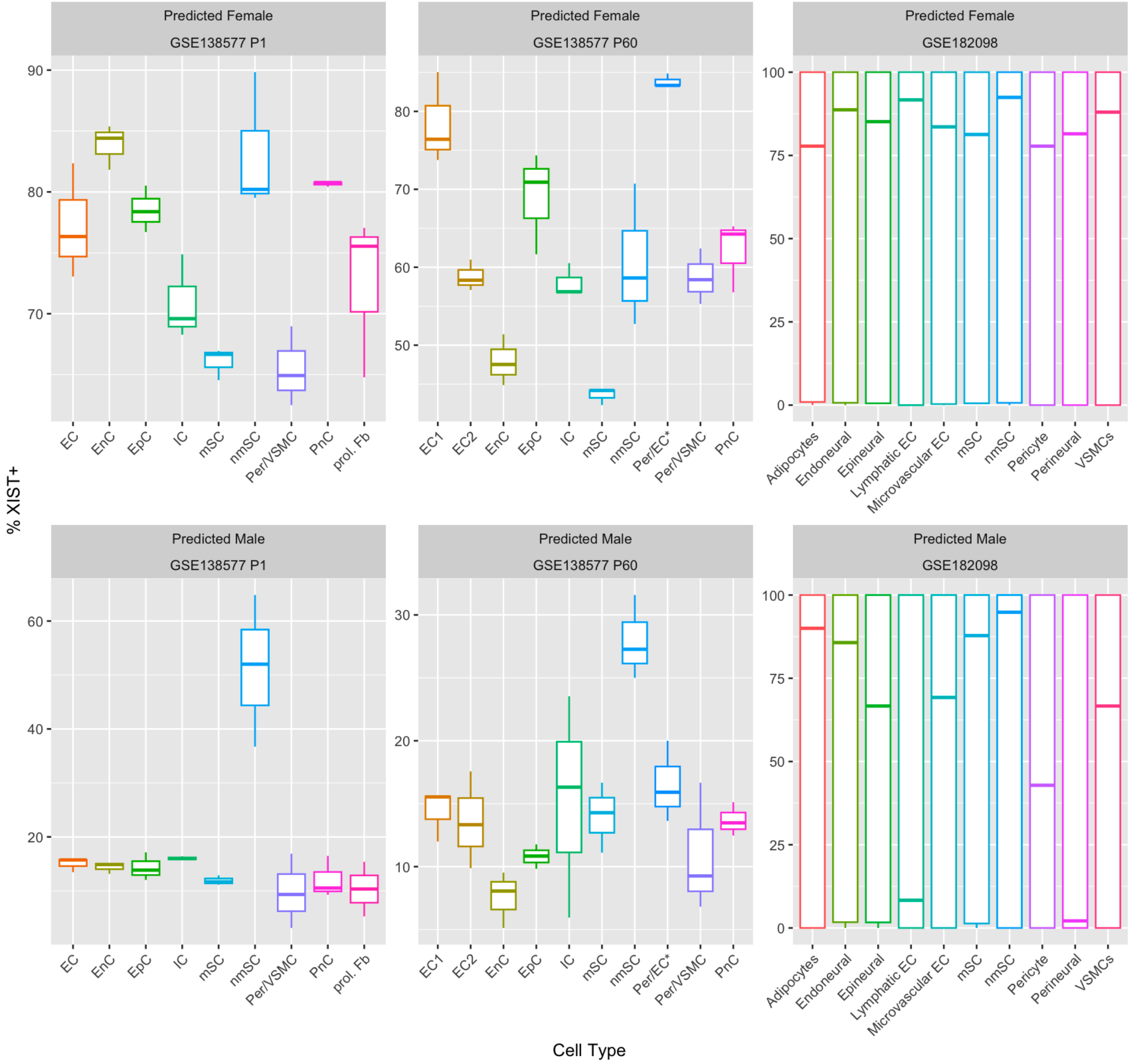
Supplemental Figure S16: Assessment of cell sex prediction accuracy from human data. A-B) Alluvial plot showing the number of cells with known sex that align with predicted sex assignment. **A)** Sex prediction accuracy from 20,611 human skeletal muscle cells. **B)** Sex prediction accuracy from 510,171 human heart cells.



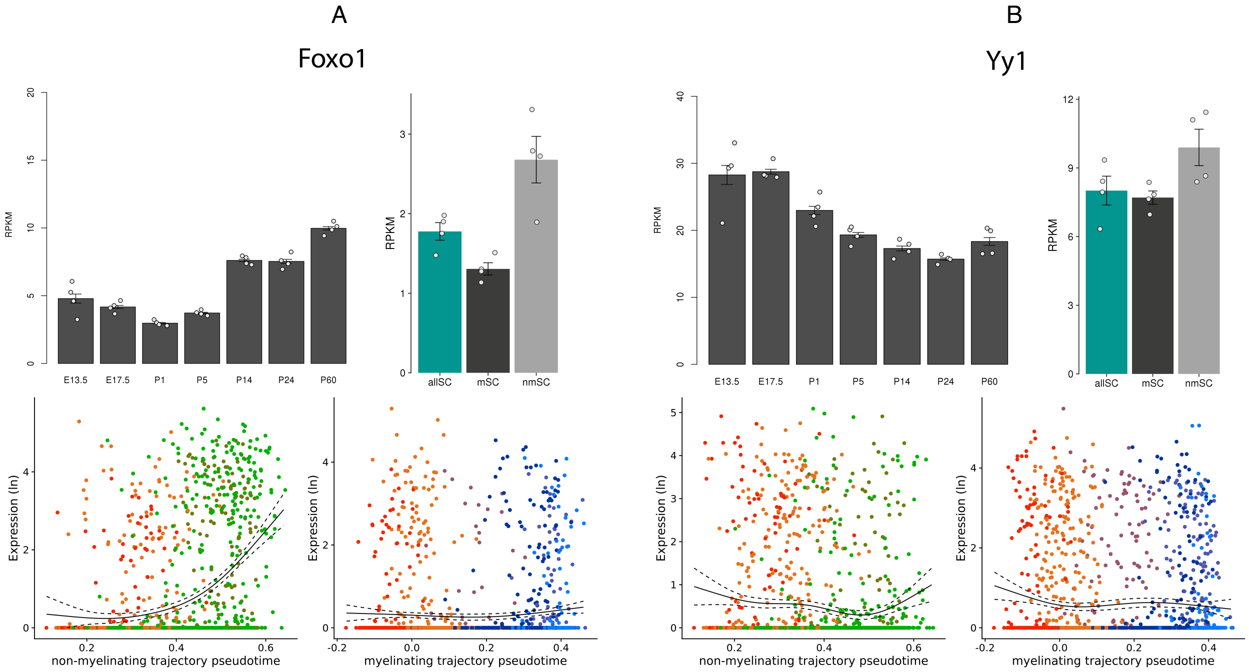
Supplemental Figure S17: Determining myelinating and non-myelinating Schwann cells in mouse scRNA-seq data. A-B) Schwann cells (SC) were selected and reclustered, followed by identification of non-myelinating Schwann cells (nnSC) and myelinating Scwhann cells (mSC) using the same markers as those used for human heart and skeletal muscle datasets. **C)** mSCs and nmSCs were already annotated, thus reclustering was not required.



Supplemental Figure S18: *XIST* expression in mouse Schwann cells. A-C) UMAP and *Xist/Foxo1* expression from peripheral (GSE182098) and sciatic (GSE138577) nerves, grouped by Schwann cell type and split by predicted sex (based on Y Chromosome expression).



Supplemental Figure S19: Percentage of *Xist*+ cells across cell types from scRNA-seq mouse datasets. Percentages were calculated for each individual then presented as a boxplot.



adapted from https://snat.ethz.ch/

Supplemental Figure S20: Summary of *Foxo1* (A) and *Yy1* (B) expression in Schwann cell types through developmental stage and pseudotime from snat.ethz.ch (bulk RNA-seq).

References

Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, Rambow F, Marine J-C, Geurts P, Aerts J et al. 2017. SCENIC: single-cell regulatory network inference and clustering. *Nature Methods* **14**: 1083-1086.

Chen J, Bardes EE, Aronow BJ, Jegga AG. 2009. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res* **37**: W305-311.

Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. 2010. Inferring regulatory networks from expression data using tree-based methods. *PLoS One* **5**.

Kanemaru K, Cranley J, Muraro D, Miranda AMA, Ho SY, Wilbrey-Clark A, Patrick Pett J, Polanski K, Richardson L, Litvinukova M et al. 2023. Spatially resolved multiomics of human cardiac niches. *Nature* **619**: 801-810.

Mehdiabadi NR, Boon Sim C, Phipson B, Kalathur RKR, Sun Y, Vivien CJ, Ter Huurne M, Piers AT, Hudson JE, Oshlack A et al. 2022. Defining the Fetal Gene Program at Single-Cell Resolution in Pediatric Dilated Cardiomyopathy. *Circulation* **146**: 1105-1108.

Sim CB, Phipson B, Ziemann M, Rafehi H, Mills RJ, Watt KI, Abu-Bonsrah KD, Kalathur RKR, Voges HK, Dinh DT et al. 2021. Sex-Specific Control of Human Heart Maturation by the Progesterone Receptor. *Circulation* **143**: 1614-1628.

Wainer Katsir K, Linial M. 2019. Human genes escaping X-inactivation revealed by single cell expression data. *BMC Genomics* **20**: 201.

Weng S, Stoner SA, Zhang DE. 2016. Sex chromosome loss and the pseudoautosomal region genes in hematological malignancies. *Oncotarget* **7**: 72356-72372.

Xu S, Hu E, Cai Y, Xie Z, Luo X, Zhan L, Tang W, Wang Q, Liu B, Wang R et al. 2024. Using clusterProfiler to characterize multiomics data. *Nature Protocols* **19**: 3292-3320.