Supplemental Materials

Mosaic loss of Chromosome Y in aged human microglia.

Michael C. Vermeulen¹, Richard Pearse², Tracy Young-Pearse², Sara Mostafavi³

¹ Center for Molecular Medicine and Therapeutics, University of British Columbia, Vancouver, Canada
² Department of Neurology, Harvard Medical School, Boston, USA
³ Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, USA

Corresponding Author
Sara Mostafavi, Ph.D.
saramos@cs.washington.edu
185 E Stevens Way NE
Seattle, WA 9819
Table of Contents

SUPPLEMENTAL METHODS ........................................................................................................3
SUPPLEMENTAL RESULTS ........................................................................................................6
SUPPLEMENTAL FIGURES .........................................................................................................8
  SUPPLEMENTARY FIGURE S1 .................................................................................................8
  SUPPLEMENTARY FIGURE S2 .................................................................................................9
  SUPPLEMENTARY FIGURE S3 .................................................................................................11
  SUPPLEMENTARY FIGURE S4 .................................................................................................13
  SUPPLEMENTARY FIGURE S5 .................................................................................................14
  SUPPLEMENTARY FIGURE S6 .................................................................................................15
  SUPPLEMENTARY FIGURE S7 .................................................................................................16
  SUPPLEMENTARY FIGURE S8A ...............................................................................................18
  SUPPLEMENTARY FIGURE S8B ...............................................................................................19
  SUPPLEMENTARY FIGURE S9 .................................................................................................20
  SUPPLEMENTARY FIGURE S10 ...............................................................................................21
  SUPPLEMENTARY FIGURE S11 ...............................................................................................22
  SUPPLEMENTARY FIGURE S12 ...............................................................................................23
  SUPPLEMENTARY FIGURE S13 ...............................................................................................24
  SUPPLEMENTARY FIGURE S14 ...............................................................................................25
  SUPPLEMENTARY FIGURE S15 ...............................................................................................26
  SUPPLEMENTARY FIGURE S16 ...............................................................................................27
  SUPPLEMENTARY FIGURE S17 ...............................................................................................45
  SUPPLEMENTARY FIGURE S18 ...............................................................................................47
  SUPPLEMENTARY FIGURE S19 ...............................................................................................48
  SUPPLEMENTARY FIGURE S20 ...............................................................................................49
  SUPPLEMENTARY FIGURE S21 ...............................................................................................50
  SUPPLEMENTARY FIGURE S22 ...............................................................................................51

SUPPLEMENTAL TABLES ..........................................................................................................52
SUPPLEMENTAL CODE ............................................................................................................52
SUPPLEMENTAL DATA .............................................................................................................52
REFERENCES ............................................................................................................................53
Supplemental Methods

Single-cell RNAseq dataset curation

**ROSMAP study datasets**

Single-cell microglia data from ‘Olah’ and ‘Mathys’ originated from fresh autopsy and frozen tissues from two longitudinal aging studies: Rush Religious Order Study (ROS) and Rush Memory and Ageing Project (MAP). Raw FASTQ files were obtained via Synapse from https://www.synapse.org/#!Synapse:syn12514624 (‘Olah’) and https://www.synapse.org/#!Synapse:syn18485175 (‘Mathys’). A summary of all information collected on the ROSMAP cohort is available at www.radc.rush.edu.

**Single-cell pre-processing**

*Including intronic UMIs to improve LOY estimation sensitivity*

Accompanying the biological findings in this study, technical improvements regarding LOY detection using scRNA-seq were made. Mainly, we found that including introns for UMI counting significantly increased detected genes with minimal impact on clustering. In 10x Chromium 3’ single-cell assays, to uniquely barcode each individual RNA molecule, poly(dT) primers are designed to prime poly(A) mRNA tails. Poly(dT) primers can also prime internal, often intronic, poly(A) tracts resulting in intronic UMIs from mature RNA (Technical note: 10x Genomics CG000376 - Rev A; https://assets.ctfassets.net/an68im79xiti/awNZTarmwqmwxKcvDv9ww/b05500661e36290b8ce59689ff889ea8/CG000376_TechNote_Antisense_Intronic_Reads_SingleCellGeneExpression_RevA.pdf). Because intronic UMIs are commonly derived from a biological molecule and are correlated with exonic UMI counts, they represent a valuable information source that improves transcriptional sensitivity for calling LOY with scRNA-seq.

**Initial scRNA-seq dataset preprocessing and annotation**

All single-cell and single-nuclei processing was performed using Seurat (v4.0.2). A basic template is provided in *Supplemental Code*. Filtered expression matrices were normalized using the `NormalizeData` function using the log normalize method and a scale factor of 10000. The top 2500 variable genes were determined using `FindVariableFeatures` using the `vst` method. Gene expression matrices were scaled and centered using `ScaleData` where the effects of nUMI, percent mitochondrial UMI and percent ribosomal UMI variables were regressed out. Principal components analysis (PCA) was performed using `RunPCA` and UMAP using `RunUMAP`. All ribosomal genes, mitochondrial genes, male–specific genes, female–specific genes (XIST) were removed from the variable features list prior to PCA calculation. PC selection for UMAP and clustering was determined using the elbow method. Clustering was performed using `FindNeighbors` (default settings) and `FindClusters` using 0.1 resolution. Datasets were visually inspected for sample and batch–specific clustering patterns. If biases were observed, batch correction was applied using Harmony (v1.0.0) (Korsunsky et al. 2019). Cell type annotations were determined through an iterative approach of increasing resolution using both automatic annotation through SingleR (Aran et al. 2019) and traditional annotation using differentially expressed canonical marker genes (*Supplemental Fig. S21; Supplemental Code*). Differentially expressed genes in each cluster were determined using MAST (Finak et al. 2015) via `FindAllMarkers`. Latent variables in each MAST model included nUMI, nGenes, ribosomal UMI pct expressed, donor, and brain region (when multiple tissues were present in an individual study).

**Microglia subcluster processing and annotation**

For microglia subcluster analysis, marker gene lists were derived from previous publications (*Supplemental Table S8*). Each member of a specific subtype classification (i.e., H1, H2) is not consistently distinct between datasets. Leiden clusters defined using `FindClusters` were each labelled with a microglia subtype based on gene set enrichment using `AddModuleScore`. Homeostatic microglia (H) were primarily determined using C3, CD74, CX3CR1 and P2RY12.
Neuron-associated homeostatic microglia subtypes additionally expressed CTNNA2, NAV3<sup>hi</sup> and GRID2. Disease associated microglia (DAM) marker genes included APOE, LPL, CD63, CD83, MYO1E, SPP1 and DSCAM. Inflammatory subtypes (I) were commonly enriched for age-associated microglia modules and were determined using CD163, TGFBI, and DYPD. Cytokine response microglia (CRM) were primarily determined using CCL3, CCL4 and IL1B. Cycling microglia were annotated using MKI67, CENPP, BRIP1, TOP2A and the KEGG cell-cycle gene set. Interferon response microglia (IRM) were annotated using IFI27, IFI44L, IFIT1, IFIT2 and IFIT3. As a result of high environmental sensitivity, single-cell isolation can induce a stressed phenotype in microglia (Thrupp et al. 2020). Stressed microglia were labelled using dissociation-associated microglia genes from Thrupp et al. including FOS, JUN, JUNB and DUSP. CNS-associated macrophages, which include perivascular and meningeal macrophages, were annotated using CD163<sup>hi</sup>, F13A1 and LYVE1. Monocytes were annotated using VCAN, CCR2, and FCN1. Blood-derived tumor-associated macrophages were annotated using KNYU, VAV3, and TGFBI in addition to a tumor associated macrophage marker gene set (Müller et al. 2017). For Fig. 4, and Supplemental Fig. S15, microglia, and astrocytes from the GSE160936 dataset (Smith et al. 2022) were isolated by brain region of origin (SSC or EC) and processed as above. Through the Seurat workflow, canonical correlation analysis (CCA) was used to integrate microglia populations. We followed the scRNA-seq integration vignette provided by the Satija Lab (https://satijalab.org/seurat/articles/integration_introduction.html) and used default settings. Density contouring for LOY and non-LOY nuclei/cells was done using the dittoDimPlot function from the dittoSeq package (Bunis et al. 2020). Contours were added using the ‘do.contour’ flag. Kernel gene-weighted density estimation was performed using the Nebulosa package (Alquicira-Hernandez et al. 2021).

**Determining technical confounders of donor-level LOY**

We found that the sequencing depth and feature diversity in a single-cell transcriptome were strongly correlated with LOY estimates (Supplemental Fig S7B). Limiting the dataset to high-information cells/nuclei dampens this effect. Setting sequencing thresholds at >=3000 UMI and >=1000 genes significantly reduces association between LOY estimates and UMI depth. This suggests that with sufficient sequencing depth single-cell GEX LOY estimates become increasingly stable and less prone to stochastic MSY gene dropout. To determine additional technical confounders to LOY across donors, we regressed a set of potential confounders against LOY proportion. Linear regression was performed for each cell type using the lm function including pairwise tests between LOY percentage and each of the following variables: total cell/nuclei number, MSY gene sparsity score, normalized Chr Y expression, 10x chemistry/version, mean UMI depth, and donor age.

**Male–specific Y Chromosome expression sparsity filtering and adjustment**

Within a cell type population/cluster, single-cell LOY estimates are strongly associated with the sparsity of male–specific Y (MSY) gene expression within Chromosome Y expressing cells (Supplemental Fig. S7). To remove the technical confounder of insufficient MSY expression we filtered and adjusted our data using a metric we call MSY sparsity score. To calculate this score, within each cell type cluster we first subset for cells with evidence of Y Chromosome presence (MSY UMI >= 1) and iteratively calculated the percentage of cells expressing each detected MSY gene using Seurat DotPlot. To generate the raw sparsity value, percent expressed values of each MSY gene were summed. Sparsity values were determined for each cell type cluster/population in each donor and were later subjected to scaling and centering. When making LOY frequency estimates cell type populations with MSY sparsity score < -1.75 were removed (Supplemental Fig. S4B).

We further used MSY sparsity scores to adjust LOY estimates. Using LOY percentage estimates and MSY sparsity scores from a collection of 1.4 million male cells (Supplemental Data S1) we calculated the non-linear least squares best fit using a 2-parameter exponential decay model (DRC.expoDecay) (Supplemental Fig. S7; Supplemental Code). After fitting the brain LOY data to this model, positive residuals were taken as adjusted LOY percentage estimates, and negative residuals were set to 0. Simply, our adjusted LOY estimate represents the deviation above expected LOY given the sparsity of the Chr Y-proficient cells in their cluster/population.
Removed samples and datasets

Datasets from accessions syn16780177, syn22130805, PRJNA434002 were considered but were removed because of low sequencing depth and feature diversity. Subject MD5893 from GSE126836 was considered low quality and was removed.

Sex chromosome disorder bulk RNAseq analysis (GSE126712)

The GSE126712 (Zhang et al. 2020) dataset contains PBMC transcriptomes from Turner syndrome (45,XO; \(n=13\)), Klienfelter’s syndrome (47,XXY; \(n=14\)) and control (46,XX, \(n=15\); 46,XY, \(n=13\)) (Supplemental Fig S22). Bulk RNAseq expression matrices were downloaded from GEO accession GSE126712 and processed using edgeR. Normalization factors were determined using TMM and expression values were converted to CPM and log normalized. Genes with low read counts were removed using filterByExpr with default settings. Differential expression analysis was performed using limma/voom. The design matrix was constructed using sex chromosome ploidy (i.e., XO = 1; XX/XY = 2; XXY = 3). Multiple statistical testing was controlled using FDR < 0.01 (Benjamini-Hochberg). DE genes were considered if the absolute value logFC > 0.3. Plots were generated using the ggplot2, ggpubr and pheatmap packages.

Single nuclei multiome ATAC and gene expression analysis

The multimodal LOY analysis was performed on sample data provided by 10x Genomics (https://www.10xgenomics.com/resources/datasets/fresh-frozen-lymph-node-with-b-cell-lymphoma-14-k-sorted-nuclei-1-standard-2-0-0). The dataset contains joint scATAC-seq and scRNA-seq for ~14,000 nuclei from an intra-abdominal lymph node tumor from a male patient diagnosed with non-Hodgkin lymphoma. Filtered feature barcode matrices (HDF5) and fragment files were downloaded from 10x Genomics and processed using both Seurat (v4.0.2) (Hao et al. 2021) and Signac (v1.2.1) (Stuart et al. 2021), following the Signac 10x scATAC vignette (https://satijalab.org/signac/articles/pbmc_vignette.html). Briefly, ATAC data was normalized using TF-IDF and features in the 75th quartile were considered top features. Peaks were identified using Cellranger and peaks within blacklist regions were removed (blacklist_hg38_unified). Gene expression pre-processing was performed as above, and dimension reduction and clustering were performed solely using the transcriptome. HG38 gene annotations from standard chromosomes were added from EnsDb.Hsapiens.v86.
Supplemental Results

*Investigation of LOY in AD and control snRNA-seq from EC and SSC (Smith et al.)*

*Corresponds with Figure 4.*

Consistent with our initial LOY analysis, across all Smith et al. samples and brain regions LOY nuclei were most common in the microglia (1917 / 10222; 15.79%; range (donor) = 1.84%–56.98%) and rare in neuronal (4 / 1136; 0.15%; range = 0%-0.86%) and astrocyte (182 / 30,659; 0.59%; range = 0.19%-2.41%) populations (Fig. 6). 48,748 QC’d nuclei were retained including 32,122 astrocyte (65.9%), 13,222 microglia (27.1%), and 1,304 neuronal nuclei (2.67%) from 4 AD and 4 control donors. All samples displayed high per cell UMI counts (median = 6,386; IQR = 6,788.7), gene counts (median = 2,758; IQR = 1824.7), sufficient nuclei sample size (mean nuclei per sample = 3,585.1; SD = 1,439.2), and robust Y Chromosome expression in biological replicates, therefore providing a quality opportunity to compare LOY in glial cells between AD and control samples (Fig. 6; Supplementary Fig. S1). Endothelial cells (48 / 410; 10.48%), OPCs (27 / 323; 7.77%), pericytes (42 / 571; 6.85%) and oligodendrocytes (9 / 157; 5.73%) also showed elevated LOY rates, however sample sizes were considerably smaller and prone to variation. LOY microglia were observed in every subject but were primarily seen in microglia populations of 4 samples from 2 AD donors: A096 (Age: 87; SSC LOY: 40.3%; EC LOY: 47.5%) and A163 (Age: 83; SSC LOY: 46.2%; EC LOY: 61.8%; Supplemental Fig. S26). These 2 subjects harbored 70.3% of the LOY population in the Smith et al. dataset.

We and others (Dumanski et al. 2021; Mattisson et al. 2021) have found cells/nuclei with fewer UMI counts and greater expression sparsity have an increased probability of false-positive LOY calls through random dropout of all Y Chromosome genes. To confirm that LOY estimates were not a result of low sequencing depth we called LOY using increasing minimum UMI thresholds (Supplemental Results Fig. S1A). In both the EC and SSC samples, as the minimum UMI threshold was increased from 1000 to 6000, microglia LOY frequency remained stable, while in astrocyte and neuronal populations LOY frequencies approached 0. Furthermore, in astrocyte and neuronal populations, LOY nuclei have significantly lower mean UMI and feature counts than non-LOY/normal nuclei (Supplemental Results Fig. S1B,C), while in microglia these attributes are not significantly different between the groups. Together, this suggests in the Smith et al. dataset that we are observing biologically meaningful LOY nuclei in the microglia, whereas in astrocyte and neuronal populations LOY labelled cells/nuclei are more likely to be a result of technical and stochastic factors.
Supplemental Results Figure S1: Attributes compared between LOY and non-LOY nuclei within each cell type cluster (GSE160936). (A) LOY frequencies in each cell type as nUMI thresholds are increased for (Left) entorhinal cortex (EC) and (Right) somatosensory cortex (SSC) derived nuclei. Each colored line represents an individual cell type cluster. LOY frequencies are quite stable as UMI thresholds increase suggesting LOY classification is not overly confounded by sequencing depth. In both brain regions, with more UMI information astrocytes and neurons approach LOY frequencies of 0 while microglia maintain frequencies > 10%. B-C, Violins and boxes show (B) UMI counts and (C) detected genes (nGenes) in LOY (red) and non-LOY (grey) nuclei for each cell type cluster. Data from both EC and SSC samples are included.
Supplemental Figures

Supplemental Figure S1: (A) Marker genes used to annotate Leiden-based clusters with cell types. *GPM6A* and *PAX5* were used to differentiate cancerous and non-cancerous B cells. (B) Comparison of LOY and non-LOY/normal chromatin accessibility-based gene activity across all chromosomes in B tumor nuclei (n = 1,895). LOY was classified using the gene expression (GEX) assay. B tumor nuclei were included if they contained >1500 GEX UMI, >1000 GEX features, >10000 ATAC UMI and > 1000 ATAC genes. Gene activity scores were calculated using genes with > 0.2 mean normalized expression and > 0 gene activity across all nuclei from the chromosome of interest. Tests of difference between LOY and normal groups for each chromosome are provided along the x-axis. Wilcoxon test: (ns) P > 0.01, (*) P < 0.01, (*** P < 0.0001.
Microglia reprocessing
- Microglia clusters isolated
- Batch correction
- Subtype annotation

Analyses
- Cell-type and subtype LOY proportion
- Differential expression
- Infiltrating monocyte signature

Download files
- GEO Synapse HCA
- Single-nuclei RNA-seq
- Single-cell RNA-seq
- FASTQ files
- 10X 3’v2
- 10X 3’v3
- 10X 5’v1

Alignment and demultiplex
- cellranger count
- include-introns

Count reads mapping to intronic regions
- Preferred over pre-mRNA reference.

Single-cell data processing
- Normalization and QC
- Seurat 4.0.2 workflow
- Doublet removal
- scDeBiFinder
- Batch correction
- Harmony
- Cell type annotation
- SingleR
- Canonical markers
- Published signatures

Adjust LOY estimates
- MSY sparsity
- Supplementary code S1

Call loss of chromosome Y
- MSY sparsity score
- LOY percentage
- Adjusted LOY is equal to residual values above curve

Commonly expressed male specific Y genes
- USP9Y
- UTY
- RPS4Y1
- DDXY
- KDM5D
- ZFY
- TTTY13
- NLGN4Y
- LINC00278
- EIF1AY
- TTTY10

Multi-modal loss of Y
- Process ATAC+GEX single-nuclei data from B-cell lymphoma sample.

Call loss of chromosome Y
- Using ATAC-seq

Infiltrating monocyte signature
- LOY vs non-LOY
- LOY events likely occur more frequently within the CNS, but can be introduced by invading monocytes.

Male cells lacking expression from all common MSY genes are defined as LOY.
Supplemental Figure S2: Bioinformatic workflow for single-cell and single-nuclei pre-processing and LOY analysis. Flow diagram illustrates the study design and bioinformatic steps used to analyze LOY in the brain. A graphical abstract of key findings, and associated figures is also provided. All single-cell and single-nuclei RNAseq data collected for this study was generated using 10x Genomics Chromium library preparation. Single-cell (sc) and single-nuclei (sn) RNAseq data was subject to slightly different alignment workflows. For scRNA-seq, to improve transcriptional sensitivity, introns were retained for UMI counting. This required all single-cell data sets to be aligned through our local pipeline using Cellranger (v5.0.0) and the “--include introns” flag. An internal 10x technical note (Interpreting Intrinsic and Antisense Reads in Single Cell Gene Expression Data - Rev A) (10x Genomics 2021) found that including introns for single-cell data improves nUMI and nGene counts with minimal impact on clustering. Since snRNA-seq data is commonly processed with introns included, we did not align all snRNA-seq datasets locally. Once expression matrices were generated, all subsequence pre-processing was identical between snRNA-seq and scRNA-seq datasets (Seurat).
**Supplemental Figure S3:** Visualization of chromatin accessibility and gene expression in LOY nuclei for two MSY genes (*UTY* and *ZFY*). Multi-modal coverage plots showing chromatin accessibility and gene expression data for LOY and non-LOY/normal nuclei for two MSY genes (A) *UTY* and (B) *ZFY*. In each panel the top shows normalized chromatin accessibility peaks for LOY and normal nuclei, which correspond with the genomic coordinates (hg38) and gene model at the bottom. The right panel contains violin plots of normalized gene expression for the gene of interest. LOY was classified using snRNA-seq and was determined by the absence of all MSY gene expression. All included nuclei were filtered using gene expression and ATAC sequencing metrics which included: >1500 UMI counts, >1000 features (GEX), and >10000 UMI counts, >1000 features (ATAC).
Supplemental Figure S4: QC threshold values for sample size and MSY sparsity score. Scatterplots displaying quality control threshold values for (A) cell/nuclei population sample size and (B) MSY sparsity score (in Chr Y proficient/Chr Y wt cells). Each point represents a donor/cell type population and is colored by labelled cell type. Each donor/cell type population required >= 100 cells/nuclei and >= -1.75 MSY sparsity score to be included in LOY proportion estimation. More information on MSY sparsity score can be found in Supplemental Fig. S7, Supplemental Methods and Supplemental Code.
Supplemental Figure S5: Donor age distribution for each brain cell type population. Included donor/cell type populations required >= 100 cells/nuclei (>=3000 UMI; >=1000 Genes) and >-1.75 MSY sparsity score (Supplemental Methods; Supplemental Fig. S7). Mean donor age and sample size for each cell type are provided at the top of each histogram. The total number of QC passing donors without age data is provided in parentheses. Samples lacking age data were not included in analyses involving age adjustment.
Supplemental Figure S6: Effect of MSY sparsity adjustment for each major brain cell type. Boxplots of mean LOY frequencies of each donor across brain cell types using (A) unadjusted and (B) adjusted LOY estimates. LOY enrichment was observed in microglia populations before and after adjustment using MSY sparsity (Supplemental Fig. S7). In oligodendrocytes and astrocytes MSY expression was often sparse and lacked feature diversity. For this reason, oligodendrocyte and astrocyte LOY estimation is often inflated when using raw values but is corrected after MSY sparsity adjustment.
Exponential Decay Model (2 params)

1.4 million male nuclei/cells

10x Genomics

>3000 nUMI, >1000 nGene, >100 cells

Samples = 535

Cell type populations = 1401

Tissues: Blood, liver, spleen, lung, esophagus, cord blood, brain, bone marrow

A

B

C

D

E

Raw LOY by cell-type

Adjusted LOY by cell-type

MSY sparsity adjustment
**Supplemental Figure S7:** Adjusting LOY values using MSY gene sparsity. Summary of LOY adjustment strategy using MSY sparsity. (A) A 2-parameter exponential decay model was fit to LOY percentage and MSY sparsity using data from single-cell and single-nuclei data (droplet-based 10x Genomics) ([Supplemental Data S1; Supplemental Code](#)). Single-cell data included 535 donors and was derived from brain, blood, liver, spleen, lung, esophagus, cord blood, and bone marrow. All primary tumor samples were removed. (B) Heatmap showing Pearson correlations between LOY percentage and technical confounders before (Right) and after (Left) MSY sparsity adjustment. Pearson correlation: (ns) $P > 0.05$, (*) $P < 0.01$, (**) $P < 0.001$, (***) $P < 0.0001$. (C) To determine adjusted LOY estimates, our brain data was fit to the decay model and residuals were retained. Positive residuals were set as adjusted LOY estimates. Donor/cell types with negative residuals were given adjusted LOY values of 0. (D-E) Boxplots displaying the log transformed (D) raw LOY estimates and (E) MSY sparsity adjusted LOY estimates.
Supplemental Figure S8A: Cell type–specific associations between LOY estimates and age. Scatterplots displaying the correlation between unadjusted LOY percentage and age for control donors (Top) and all donors (Bottom) across five major brain cell types. Points represent an individual donor/cell type population and are colored based on neurological diagnosis. Each scatter plot contains a linear model with the formula: LOY ~ Donor age. The associated Pearson correlation and p-value are provided in the first line of each header. Adjusted p-values are derived from a multiple linear regression model with the formula: LOY ~ median_nUMI + 10x chemistry + MSY sparsity score + donor_organism.age.
**Supplemental Figure S8B**: Cell type–specific associations between LOY estimates and age. Scatterplots displaying the correlation between adjusted LOY percentage and age in control donors (Top) and all donors (Bottom) across five major brain cell types. Points represent an individual donor/cell type population and are colored based on neurological diagnosis. LOY estimates were adjusted using MSY sparsity scores (Methods). Each scatter plot contains a linear model with the formula: LOY ~ Donor age. The associated Pearson correlation and p-value are provided in the first line of each header. Adjusted p-values are derived from a multiple linear regression model with the formula: LOY ~ median_nUMI + 10x chemistry + MSY sparsity score + donor_organism.age.
Supplemental Figure S9: Disease–specific LOY across five major brain cell types. LOY frequencies facetted by neurodegerative diagnosis and brain cell type. Black points represent median LOY frequency, while black lines represent the interquartile range. Each colored point represents the LOY percentage estimate of a cell type population in an individual donor.
Supplemental Figure S10: Comparing LOY estimates between AD, HD and control donors. (A,B) LOY percentage estimates between (A) AD or (B) HD and control donors for each major brain cell type. Non-neurodegerative (Non-ND) refers to donors without a neurodegenerative diagnosis. Adjusted $P$-values are derived from a multiple linear regression model with the formula: $\text{LOY} \sim \text{median}_\text{nUMI} + 10x \text{chemistry} + \text{MSY sparsity score} + \text{donor}_\text{organism}.\text{age} + \text{diagnosis}$. Donor sample size for each group is provided in parenthesis on the x-axis. (C,D) Boxplots displaying adjusted LOY percentages in for (C) microglia and (D) OPCs between control, AD and other neurodegenerative diseases.
Supplemental Figure S11: Cumulative PAR and mLATE gene expression across 28 donors (A) Cumulative, normalized PAR expression compared between LOY and normal population for each major brain cell type. Each point represents an individual donor population with > 25 cells (Paired Wilcoxon test) (C) Heatmap showing significance of PAR expression loss between LOY and normal populations within each brain cell type across 16 cohorts. Significance values were determined using Wilcoxon tests between PAR module scores for all LOY and normal nuclei in each group. Populations with > 25 cells were included. Wilcoxon test: (ns) $P > 0.05$; (*) $P < 0.05$, (**) $P < 0.01$, (***) $P < 0.001$. (B) Heatmap illustrating microglia DE (average logFC) for individual subjects with > 50 LOY cells/nuclei. Additional columns on the left show cumulative PAR expression and neurodegenerative diagnosis, respectively. Significance is provided within each cell. False-discovery rate: (*) FDR < 0.05, (**) FDR < 0.01, (***) FDR < 0.001.
**Supplemental Figure S12**: Differential expression analysis in pooled microglia between LOY and normal single-cell populations. (A) Volcano plot showing 193 differentially expressed genes in LOY microglia (Bonferroni, $P < 0.1$ & abs(logFC) > 0.10). LOY was classified using Chr Y genes so all male-specific Y genes were removed from visualization. The following 19 samples were included in this analysis: GSM4886748, GSM4886749, GSM4886750, GSM4886751, GSM4886758, GSM4886760, GSM4886761, GSM5106119, GSM5106120, GSM5550451, GSM5550475, GSM5550481, GSM5550491, GSM5550495, GSM5550501, GSM5550503, Microglia_MO_MCI3, pPDsHSrSNxi1963, pPDsHSrSNxi4775. DE genes were called using MAST with latent variables including sample, nUMI, nFeature, and percent.mt. (B) Enriched gene sets for 193 mLATE genes from Metascape ($P < 0.05$; Bonferroni) (**Supplemental Table S5**).
Supplemental Figure S13: Expression of recurrent microglia LOY DE genes and SLIT/ROBO genes across all brain cell types. Boxplots illustrate cell type–specific expression of selected mLATE, SLIT/ROBO signaling pathway genes and infiltrating monocyte marker genes (CCL2). Points overlayed on each boxplot are colored by sample brain region origin. OL Oligodendrocyte, CAM CNS-associated macrophage, OPC oligodendrocyte progenitor cell, EC Entorhinal cortex, OC Occipital cortex, OTC Occipitotemporal cortex, PFC Pre-frontal cortex, SSC Somatosensory cortex.
Supplemental Figure S14: Cell type marker genes and proportions for GSE160936 single-nuclei dataset from AD and control brain tissue. Corresponds to Figure 4. (A) Dot plot displaying markers genes used to annotate Leiden clusters with cell types. (B) Donut chart of relative cell type proportions ($n=48,748$). Astrocytes and microglia nuclei clusters represent over 90% of all nuclei in from the GSE160936 dataset.
Supplemental Figure S15: LOY frequencies in astrocytes across the EC and SSC (GSE160936). (A) UMAP reduction and cell type annotations of 48,748 single-nuclei brain transcriptomes from GSE160936. Markers used to annotate clusters include C3 and CSF1R (microglia), AQP4 and GFAP (astrocyte), RBFOX3 and GAD1 (neuron), and MBP (oligodendrocyte). (B) UMAP plot showing LOY classification of each male nucleus (nUMI > 3000 and nGene > 1000). LOY nuclei (red) and non-LOY/normal nuclei (grey). Astrocyte nuclei from SSC (C) and EC (F) were isolated, CCA integrated, and clustered. (C,F). Subclusters were annotated using high (yellows) and low (blues) GFAP expression. (D,G), UMAP reductions of SSC and EC split by AD diagnosis; Nuclei are colored by LOY (red) and non-LOY/normal (grey). LOY nuclei frequencies are provided at the bottom of each UMAP plot. (E,H), Violin boxplots showing UMI counts (nUMI) and detected genes (nGenes) for LOY and normal nuclei. The boxes represent the 25th percentile, median, and 75th percentile. The whiskers extend to the furthest value that is no more than 1.5 times the inter-quartile range. Significant differences between LOY and non-LOY groups were assessed using the non-parametric Wilcoxon test, resulting in the p-values shown for each comparison. All tests were two-sided. ns < 0.05, *P < 0.05, **P < 0.01, ***P < 0.001.
**Supplemental Figure S16A:** Microglia subtype characterization and LOY summary (GSE160936). Microglia from single-nuclei RNAseq data (10x 3’ v3) from GEO accession GSE160936 were isolated and reprocessed. (A) UMAP plot showing microglia subtype clusters for 15,791 male and 7,040 female nuclei. Each point represents a nuclei transcriptome and is coloured by subtype annotation (nUMI > 1500 & nGene > 800). (B) UMAP plot illustrating LOY status for each included male microglia (LOY: 15.9%; n = 1942; nUMI > 3000; nGene > 1000). (C-D) Density contours for (C) LOY nuclei and (D) non-LOY nuclei illustrate clustering density patterns for each group. (E-F) Heatmaps displaying (E) published microglia gene set module scores and (F) marker genes/DE genes used to annotate each subcluster. Gene sets used are provided in **Supplemental Table S8.** H1-3 Homeostatic microglia, DAM1-2 Disease associated microglia, I1-2 Inflammatory microglia, CRM Cytokine release microglia.
Supplemental Figure S16B: Microglia subtype characterization and LOY summary (syn12514624). Microglia from single-cell RNAseq data (10x 3’ v2) from Synapse accession syn12514624 were isolated and reprocessed. (A) UMAP plot showing microglia subtype clusters for male \( (n = 4,580) \) and female \( (n = 6,741) \) nuclei. Each point represents a nucleus and is coloured by subtype annotation \( (\text{nUMI} > 1500 \& \text{nGene} > 800) \). (B) UMAP plot illustrating LOY status for each included male microglia \( (\text{LOY \%: 22.7\%; nUMI} > 3000 \& \text{nGene} > 1000) \). (C-D) Density contours for (C) LOY nuclei and (D) non-LOY nuclei illustrate clustering density patterns for each group. (E-F) Heatmaps displaying (E) published microglia gene set module scores and (F) marker genes/DE genes used to annotate each subcluster. Gene sets used are provided in Supplemental Table S8. H1-4 Homeostatic microglia, DAM1 Disease associated microglia, II Inflammatory microglia, CRM Cytokine release microglia.
Supplemental Figure S16C: Microglia subtype characterization and LOY summary (GSE178265).
Microglia from single-nuclei RNAseq data (10x 3' v3) from GEO accession GSE178265 were isolated and reprocessed. (A) UMAP plot showing microglia subtype clusters for male (\(n = 8,835\)) nuclei. Each point represents a nucleus and is coloured by subtype annotation (nUMI > 1500 & nGene > 800). (B) UMAP plot illustrating LOY status for each included male microglia (LOY %: 3.7%; \(n = 328\)). (C-D) Density contours for (C) LOY nuclei and (D) non-LOY nuclei illustrate clustering density patterns for each group. (E-F) Heatmaps displaying (E) published microglia gene set module scores and (F) marker genes/DE genes used to annotate each subcluster. Gene sets used are provided in Supplemental Table S8. H1-4 Homeostatic microglia, DAM1 Disease associated microglia, I1 Inflammatory microglia, CRM Cytokine release microglia.
**Supplemental Figure S16D:** Microglia subtype characterization and LOY summary (GSE137444). Microglia from single-cell RNAseq data (10x 3’ v2) from GEO accession GSE137444 were isolated and reprocessed. (A) UMAP plot showing microglia subtype clusters for male ($n = 11,756$) and female nuclei ($n = 29,284$). Each point represents a cell and is coloured by subtype annotation (nUMI > 1500 & nGene > 800). (B) UMAP plot illustrating LOY status for each included male microglia (LOY: 233/7039 (3.2%); nUMI > 3000 & nGene > 1000). (C-D) Density contours for (C) LOY nuclei and (D) non-LOY nuclei illustrate clustering density patterns for each group. (F-E) Heatmaps displaying (E) published microglia gene set module scores and (F) marker genes/DE genes used to annotate each subcluster. Gene sets used are provided in **Supplemental Table S8.**

H1-3 Homeostatic microglia, DAM Disease associated microglia, IRM Interferon response microglia, CRM1-2 Cytokine release microglia.
**Supplemental Figure S16E**: Microglia subtype characterization and LOY summary (GSE148822). Microglia from single-nuclei RNAseq data (10x 3’ v3) from GEO accession GSE148822 were isolated and reprocessed. (A) UMAP plot showing microglia subtype clusters for 15,737 male nuclei. Each point represents a nucleus and is coloured by subtype annotation (nUMI > 1500 & nGene > 800). (B) UMAP plot illustrating LOY status for each included male microglia transcriptome (LOY: 3.7%; nUMI > 3000 & nGene > 1000). (C-D) Density contours for (C) LOY nuclei and (D) non-LOY nuclei illustrate clustering density patterns for each group. (E-F) Heatmaps displaying (E) published microglia gene set module scores and (F) marker genes/DE genes used to annotate each subcluster. Gene sets used are provided in **Supplemental Table S8**. H1: Homeostatic microglia, DAM1: Disease associated microglia, I1: Inflammatory microglia, CAM: CNS-associated macrophage.
Supplemental Figure S16F: Microglia subtype characterization and LOY summary (GSE174332). Microglia from single-nuclei RNAseq data (10x 3’ v3) from GEO accession GSE174332 were reprocessed. (A) UMAP plot showing microglia subtype clusters for 10,695 male and 5826 female nuclei. Each point represents a nucleus and is coloured by subtype annotation (nUMI > 1500 & nGene > 800). (B) UMAP plot illustrating LOY status for each included male microglia (LOY: 3.9%; nUMI > 3000 & nGene > 1000). (C-D) Density contours for (C) LOY nuclei and (D) non-LOY nuclei illustrate clustering density patterns for each group. (E-F) Heatmaps displaying (E) published microglia gene set module scores and (F) marker genes/DE genes used to annotate each subcluster. Gene sets used are provided in Supplemental Table S8.

H1-3 Homeostatic microglia, DAM1-3 Disease associated microglia, I Inflammatory microglia, CAM CNS-associated macrophage.
**Supplemental Figure S16G**: Microglia subtype characterization and LOY summary (GSE174367). Microglia from single-nuclei RNAseq data (10x 3’ v3) from GEO accession GSE174367 were reprocessed. (A) UMAP plot showing microglia subtype clusters for 5964 male and 4115 female nuclei. Each point represents a nucleus and is coloured by subtype annotation (nUMI > 1500 & nGene > 1000). (B) UMAP plot illustrating LOY status for each included male microglia (LOY: 9.3%, n = 557). Density contours for (C) LOY nuclei and (D) non-LOY nuclei illustrate clustering density patterns for each group. Heatmaps displaying (E) published microglia gene set module scores and (F) marker genes/DE genes used to annotate each subcluster. Gene sets used are provided in Supplemental Table S8. H1-3 Homeostatic microglia, DAM Disease associated microglia, I Inflammatory microglia, CAM CNS-associated macrophage.
Supplemental Figure S16H: Microglia subtype characterization and LOY summary (GSE183068). Microglia from single-nuclei RNAseq data (10x 3’ v3) from GEO accession GSE183068 were reprocessed. (A) UMAP plot showing microglia subtype clusters for 10,695 male nuclei. Each point represents a nucleus and is coloured by subtype annotation (nUMI > 1500 & nGene > 1000). (B) UMAP plot illustrating LOY status for each included male microglia (LOY: 11.8%; nUMI > 3000 & nGene > 1000). Density contours for (C) LOY nuclei and (D) non-LOY nuclei illustrate clustering density patterns for each group. Heatmaps displaying (E) published microglia gene set module scores and (F) marker genes/DE genes used to annotate each subcluster. Gene sets used are provided in Supplemental Table S8.

H1-2 Homeostatic microglia, DAM1-3 Disease associated microglia, I Inflammatory microglia, CAM CNS-associated macrophage, CRM Cytokine response microglia, TAM Tumor-associated macrophage.
**Supplemental Figure S16:** Microglia subtype characterization and LOY summary (GSE167494). Microglia from single-nuclei RNAseq data (10x 3’ v3) from GEO accession GSE167494 (GSE167490 and GSE167492) were combined and reprocessed. (A) UMAP plot showing microglia subtype clusters for 1982 female and 1590 male nuclei. Each point represents a nucleus and is coloured by subtype annotation (nUMI > 1500 & nGene > 1000). (B) UMAP plot illustrating LOY status for each included male microglia (LOY: 20.9%). Density contours for (C) LOY nuclei and (D) non-LOY nuclei illustrate clustering density patterns for each group. Heatmaps displaying (E) published microglia gene set module scores and (F) marker genes/DE genes used to annotate each subcluster. Gene sets used are provided in **Supplemental Table S8.** H1-3 Homeostatic microglia, DAM1 Disease associated microglia, CAM CNS-associated macrophage.
Supplemental Figure S17: LOY prevalence in microglia subtypes. Heatmaps showing subtype LOY proportions split by (A) cohort and (B) individual sample (>3000 nUMI & >1000 nGene). Subtypes were included in (A) if they contained >= 50 cells and were included in (B) if they contained >= 25 cells. Grey squares denote subtypes that failed to pass the sample size threshold. (C-F), LOY microglia are enriched in the pro-inflammatory cluster of syn12514624 (Olah et al. 2020). (C) Y Chromosome and (E) PAR module scores are significantly reduced in the pro-inflammatory cluster (I1). (F) LOY density visualized using kernel density estimation (syn12514624). (G-J) LOY is enriched within a cluster of blood-derived tumor-associated microglia/macrophages in GSE183068 (Sayed et al. 2021). (G) UMAP plot displaying clustering patterns microglia/macrophages. Density of (H) LOY cells and (J) TAMs using density kernel estimation. (I) Barplot displaying LOY proportions by subtype (GSE183068; >3000 nUMI; >1000 nGene).
**Supplemental Figure S18**: Infiltrating monocytes and microglia have similar LOY frequencies (GSE160936). UMAP reductions displaying 12,139 male microglia, monocyte, and CNS associated macrophage (CAM) nuclei from GSE160936 (from the somatosensory cortex and entorhinal cortex). Each point represents an individual nucleus and is colored by (A) cluster and (B) LOY status. (C-E) Three microglia UMAP plots displaying kernel gene-weighted density estimation for three known markers of infiltrating monocytes in the CNS (CCR2, VCAN and FCN1) from dataset GSE160936 (Smith et al. 2022). Monocyte expression was confined to one small cluster and was detected in 674 nuclei (2.7%; male and female). Text below each plot displays LOY proportions for male nuclei with and without detected CCR2, VCAN and FCN1 expression. Total sample size is provided in parentheses. Nuclei were included if they contained > 3000 nUMI and >1000 nGenes.
Supplemental Figure S19: PAR gene expression by cell type across 30 brain datasets. Cell type–specific normalized expression (log2) of the top 12 commonly expressed pseudoautosomal region (PAR) genes across 29 brain datasets. Points overlayed on each boxplot are colored by derived brain region and represent a donor/cell type population. OL Oligodendrocyte, CAM CNS-associated macrophage, OPC oligodendrocyte progenitor cell, EC Entorhinal cortex, OC Occipital cortex, OTC Occipitotemporal cortex, PFC Pre-frontal cortex.
Supplemental Figure S20: Male–specific Y (MSY) gene expression by brain cell type. Cell type–specific normalized expression (log2) of the top 12 most expressed male–specific Y (MSY) genes across 30 brain datasets. Points overlayed on each boxplot are colored by brain region and represent a donor/cell type population. OL Oligodendrocyte, CAM CNS-associated macrophage, OPC oligodendrocyte progenitor cell, EC Entorhinal cortex, OC Occipital cortex, OTC Occipitotemporal cortex, PFC Pre-frontal cortex.
Supplemental Figure S21: Canonical gene markers used to label cell type clusters. Points overlaid on each boxplot are colored by brain region and represent a donor/cell type population. OL Oligodendrocyte, CAM CNS-associated macrophage, OPC oligodendrocyte progenitor cell, EC Entorhinal cortex, OC Occipital cortex, OTC Occipitotemporal cortex, PFC Pre-frontal cortex.
Supplemental Figure S22: PBMC differential expression analysis between XO, XY/XX and XXY donors. Reprocessed data from GEO accession GSE126712. Differential expression analysis using limma voom found 610 DE genes (247 upregulated and 363 downregulated) were associated with increasing sex chromosome ploidy (FDR < 0.01 and absolute value logFC > 0.3). XO, XX/XY and XXY donors have 1, 2 and 3 sex chromosome copies, respectively. 

(A) Line plot illustrating summarized PAR gene expression as sex chromosome ploidy increases. Errors bars represent standard error.

(B-C) Boxplots displaying normalized expression by karyotype for (B) nine expressed PAR genes and (C) two recurrently upregulated microglia loss of Y associated transcriptional effect (mLATE) genes in microglia (ROBO1 and RBFOX2). 

(D)
Heatmap illustrating mLATE gene expression in sex chromosome disorder PBMCs. Autosomal feature overlap between LOY microglia DE and sex chromosome ploidy DEs in PBMCs was significant (Fisher’s exact test, $P < 0.001$). Columns are colored by sex chromosome karyotype.

**Supplemental Tables**

Supplementary Table S1 – Included brain scRNA-seq dataset summary  
Supplementary Table S2 – Brain cell type LOY summary by donor  
Supplementary Table S3 - Summary of 193 LATE genes in microglia and comparison with previously published data.  
Supplementary Table S4 - Genomic region enrichment of microglia LATEs by donor  
Supplementary Table S5 – microglia LATE gene set enrichment results (Metascape)  
Supplementary Table S6 - Summary of subtype–specific microglia LOY estimates by cohort  
Supplementary Table S7 - Summary of LOY estimates by monocyte marker status  
Supplementary Table S8 – Microglia gene sets used in this study

**Supplemental Code**

Functions.R – contains all helper R functions for LOY analysis.  
call_LOY.R – contains R functions used to call LOY in Seurat objects.  
LOY_adjust.R – contains R function to adjust LOY using MSY sparsity.

All functions and scripts used to process data and produce figures can be found at:  
https://github.com/michaelcvermeulen/microglia-loss-of-y

**Supplemental Data**

Supplementary Data S1 – Donor–specific LOY summary files for each cell type across a diverse set of human tissues with increasing UMI depth thresholding. Used for model fitting and MSY sparsity LOY adjustment. All included sc/snRNA-seq datasets in this file are droplet-based 10x Genomics prepared datasets.

Supplementary Data S2 – Donor–specific LOY summary files in brain cell types using 1500 and 3000 UMI depth thresholds. Used for Figure 3 and Figure 4.
References

Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, Chak S, Naikawadi RP, Wolters PJ, Abate AR, et al. 2019. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* 20: 163–172.

Bunis DG, Andrews J, Fragiadakis GK, Burt TD, Sirota M. 2020. dittoSeq: universal user-friendly single-cell and bulk RNA sequencing visualization toolkit. *Bioinformatics* 36: 5535–5536.

Dumanski JP, Halvardson J, Davies H, Rychlicka-Buniowska E, Mattisson J, Moghadam BT, Nagy N, Węglarczyk K, Bukowska-Strakova K, Danielsson M, et al. 2021. Immune cells lacking Y chromosome show dysregulation of autosomal gene expression. *Cell Mol Life Sci.* https://doi.org/10.1007/s00018-021-03822-w (Accessed April 28, 2021).

Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, Slichter CK, Miller HW, McElrath MJ, Prlic M, et al. 2015. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* 16: 278.

Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner M, Loh P, Raychaudhuri S. 2019. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* 16: 1289–1296.

Mattisson J, Danielsson M, Hammond M, Davies H, Gallant CJ, Nordlund J, Raine A, Edén M, Kilander L, Ingelsson M, et al. 2021. Leukocytes with chromosome Y loss have reduced abundance of the cell surface immunoprotein CD99. *Sci Rep* 11: 15160.

Müller S, Kohanbash G, Liu SJ, Alvarado B, Carrera D, Bhaduri A, Watchmaker PB, Yagnik G, Di Lullo E, Malatesta M, et al. 2017. Single-cell profiling of human gliomas reveals macrophage ontogeny as a basis for regional differences in macrophage activation in the tumor microenvironment. *Genome Biol* 18: 234.

Olah M, Menon V, Habib N, Taga MF, Ma Y, Yung CJ, Cimpean M, Khairallah A, Coronas-Samano G, Sankowski R, et al. 2020. Single cell RNA sequencing of human microglia uncovers a subset associated with Alzheimer’s disease. *Nat Commun* 11: 6129.

Sayed FA, Kodama L, Fan L, Carling GK, Udeochu JC, Le D, Li Q, Zhou L, Wong MY, Horowitz R, et al. 2021. AD-linked R47H-TREM2 mutation induces disease-enhancing microglial states via AKT hyperactivation. *Sci Transl Med* 13: eabe3947.

Smith AM, Davey K, Tsartsalis S, Khozoie C, Fancy N, Tang SS, Liaptsi E, Weinert M, McGarry A, Muirhead RCJ, et al. 2022. Diverse human astrocyte and microglial transcriptional responses to Alzheimer’s pathology. *Acta Neuropathol (Berl)* 143: 75–91.

Thrupp N, Sala Frigerio C, Wolfs L, Skene NG, Fattorelli N, Poovathingal S, Fourny Y, Matthews PM, Theys T, Mancuso R, et al. 2020. Single-Nucleus RNA-Seq Is Not Suitable for Detection of Microglial Activation Genes in Humans. *Cell Rep* 32: 108189.

Zhang X, Hong D, Ma S, Ward T, Ho M, Pattini R, Duren Z, Stankov A, Shrestha SB, Hallmayer J, et al. 2020. Integrated functional genomic analyses of Klinefelter and Turner syndromes reveal global network effects of altered X chromosome dosage. *Proc Natl Acad Sci* 117: 4864–4873.