Improving and Estimating Y Chromosome Loss in Blood and Brain Tissues Using High-throughput Sequencing

by

Michael Cory Vermeulen

B.Sc., Queen's University, 2016

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUTE AND POSTDOCTORAL STUDIES

(Medical Genetics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

August 2020

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

Improving and Estimating Y Chromosome Loss in Blood and Brain Tissues Using High-throughput Sequencing

submitted by Michael Cory Vermeulen in partial fulfillment of the requirements for

the degree of Master of Science

in Medical Genetics

Examining Committee:

Dr. Sara Mostafavi, Medical Genetics, Statistics, Computer Science

Co-supervisor

Dr. William Gibson, Medical Genetics

Co-supervisor

Dr. Carolyn Brown, Medical Genetics

Supervisory Committee Member

Dr. Kelly Brown, Pediatrics

Supervisory Committee Member

Abstract

To our knowledge age-related loss of chromosome Y (LOY) in circulating leukocytes is the most common somatic genetic aberration. Many recent epidemiology studies have found robust associations between LOY in leukocytes and age-related diseases such as blood and solid tumour cancers, Alzheimer's disease, and macular degeneration. Despite these associations, the prevalence and mechanisms of LOY in non-hematopoietic cell-types are not well characterized. In response, the need for bioinformatic methods to analyse Y chromosome ploidy across multiple genomic/transcriptomic datatypes has escalated. In the past, the Y chromosome was commonly removed from genomic analyses for several reasons including low gene count, haploidy, lack of biological interest and short-read mapping difficulties. Resultingly, methods for investigating chromosome Y specific trends using next-generation sequencing have suffered and require improvement. The main objective of this thesis was two-fold. First, to improve methods of Y chromosome aneuploidy detection using whole genome sequencing and single-nuclei RNA sequencing. Second, to use these improved methods to provide estimates of loss of Y (LOY) in brain tissue – which had not previously been established in humans. Using genomic characteristics such as mappability, GC content, and read alignment filtering I was able to improve LOY detection in both WGS and single-nuclei RNA-seq. Given high sequence similarity between the X and Y chromosome, strict mappability filtering improves, and smooths read depth estimates of Y chromosome aneuploidy. Using these methods we estimate that of the elderly male population represented in this cohort (median age = 87.5), LOY was found in 13.8% (11/123) of blood samples, 0% (0/159) in prefrontal cortex and 0% (0/78) cerebellum samples. Despite this, we found a significant association between age and reduced Y ploidy in the dorsolateral prefrontal cortex ($R=-0.35$, $p=3.9\times10^{-5}$), suggesting low-frequency LOY may be occurring in the cortex. In single-nuclei data from the dorsolateral prefrontal cortex we found 8.6% of cells lacked a Y chromosome. LOY was enriched in the glial cells, and particularly the microglia where 33% of male cells were affected. Although further evidence is required, LOY within the frontal cortex (and specifically the microglia) may represent an understudied factor in cognitive decline and neurodegeneration.

Lay Summary

Researchers have observed something odd when counting the chromosomes in human white blood cells. Regularly, and more regularly with age, some male white blood cells tend to lack their Y chromosome. Recent studies have found that about 40% of men older than 70 are missing the Y chromosome in a proportion of their white blood cells. Further research has concluded this phenomenon is a sign of damage to DNA and is linked to cancer, Alzheimer's, and other age-related diseases. The goal of my thesis was to improve software used to detect Y chromosome loss to determine if Y loss is also occurring in the brain. In a cohort of elderly men, we found early evidence of Y chromosome loss in brain tissue, specifically in the immune cells of the brain - the microglia. Future studies can use what we have learned to better detect Y loss and investigate it across the body.

Preface

The experiments reported in this thesis were designed in collaboration with my supervisors, Dr. Sara Mostafavi and Dr. William Gibson and conducted by me.

A version of Chapter 2 has been published.

Emma J Graham**, Michael Vermeulen**, Badri Vardarajan, David Bennett, Phil De Jager, Richard V Pearse, Tracy L Young-Pearse, Sara Mostafavi. Somatic mosaicism of sex chromosomes in the blood and brain. Brain Res. 146345 (2019) doi:10.1016/j.brainres.2019.146345.

All data collection, sample preparation, library preparation, and sequencing for data used in Chapter 2 were completed by the ROSMAP consortium (Dr. David Bennett). Data was provided by the ROSMAP consortium and was made available to me by Dr. Mostafavi and Badri Vardarajan. Whole genome sequence processing and read depth binning was completed by Badri Vardarajan. Emma Graham wrote most of the manuscript, was responsible for the transcriptomic analyses and epidemiological analyses, as well as the methods for comparing Y loss between tissues and sequencing batches. Most experiments were designed in collaboration with Emma Graham. I was responsible for the loss of Y analysis of provided read depth data, and filtering to improve aneuploidy detection. In addition, I processed the Affymetrix genotype array data and called Y chromosome loss from that dataset.

Data analysis and project design in Chapter 3 is largely my own, although all data collection, sample preparation, library preparation, and sequencing were completed by Mathys *et al.* and the ROSMAP consortium. Data was provided by the ROSMAP consortium and was made available to me by Dr. Mostafavi through the Synapse Portal. Once sequencing files and phenotype information were provided, I completed most project design, data processing, analysis methods, data visualization, and most biological conclusions, with frequent direction and mentorship from Dr. Mostafavi and Dr. Gibson, and occasional scientific input from Dr. Richard Pearse and Dr. Tracy Young-Pearse. All figures in the thesis were generated by me.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AD – Alzheimer's disease
ANCOVA – analysis of covariance
Ast – astrocyte
CGH - comparative genomic hybridization
CN – estimated copy number
CNV- copy number variant
DLPFC – dorsolateral prefrontal cortex
Ex – excitatory neuron
FACS – fluorescence activated cell sorting
FISH - fluorescence in situ hybridization
G&T-seq – genome and transcriptome sequencing
GEM – The Genome Multitool mapper aligner
GWAS – genome-wide association study
HR – hazard ratio
HSPC – hematopoietic stem cells and precursor cells
In – inhibitory neuron
LOY – mosaic loss of chromosome Y
LRR - log R ratio
MSY – male specific Y region
MVA – Mosaic variegated aneuploidy syndrome
Mic – microglia
OPC – oligodendrocyte progenitor cell
OR – odds ratio
Oli – oligodendrocyte
PAR – pseudoautosomal region
ROSMAP – Religious Orders Study (ROS) and Rush Memory and Aging Project (MAP)
SNV – single-nucleotide variant
SV- structural variant
UMI – unique molecular identifier
WGA – whole genome amplification
WGS – whole genome sequencing
XDR – X-degenerate region
XTR – X-transposed region
chrY – chromosome Y
mLRR – median log R ratio
mLRRY – median log R ratio across male specific Y region
qPCR – quantitative PCR
rRD – relative read depth
scRNA-seq – single-cell RNA sequencing
scWGS – single-cell whole genome sequencing
snRNA-seq – single-nuclei RNA sequencing

Acknowledgements

Firstly, I would like to thank both of my supervisors, Dr. William Gibson and Dr. Sara Mostafavi for their patience, guidance, and expertise. I am proud of the improvements I have made and the scientific knowledge I have gained over the course of my degree, and I thank you for the opportunities you have given me. You have taught me to think like a scientist and have inspired me to pursue a career in genetics. I would also like to acknowledge my committee members Dr. Carolyn Brown and Dr. Kelly Brown for your time and guidance. Also, thank you to all of the individuals involved in the ROSMAP consortium for providing me with organized, high-quality data to investigate.

Secondly, thank you to both the Mostafavi and Gibson lab members, past and present, who provided great scientific and moral support over the years. Special thanks to Emma Graham who helped with a large part of this thesis. Thank you for our helpful discussions both scientific and not.

Lastly, thank you to my parents and Nana for their unconditional love and encouragement towards learning and hard work from a young age. Also thank you to Sara Pieczonka for your devoted and loving support through the completion of this degree.

Dedication

This thesis is dedicated to my parents,

Kent and Debbie,

who have consistently encouraged learning and hard work,

and have supported my goals without exception

1. Introduction

1.1 Post-zygotic mosaicism and aneuploidy in humans

From the fertilization of an oocyte to the death of an organism, genetic alterations of all sizes - from single base pairs to entire chromosomes - are constantly accumulating across its soma. During each cell division an estimated 0.1 to 1 mutations arise and persist as a result of mitotic error.[1] In addition, internal and external mutagens are relentlessly and spontaneously altering DNA. While these errors are usually repaired by well conserved cellular mechanisms, inevitably mutations avoid detection and persist, which leads to genetically distinct cells within the organism. If these cells survive, grow and replicate, an organism can consist of two or more genomically distinct populations of cells - a phenomenon referred to as somatic mosaicism. While a majority of these mutations are benign or detrimental to cell growth, others can improve cellular fitness and/or proliferative capabilities which can lead to a clonal expansion and an amplification of the mutation.[2] Some mosaic clonal expansions can induce cellular and tissue dysfunction that can promote maladies such as cancer[3], neurodegenerative disease[4], and heart disease.[5]

Interestingly, for much of the 20th century the genome was largely considered to be uniform across all cells in an organism.[6] Although the impact of post-zygotic mutation on cancer development and genetic disease was known, both were often regarded as rarities and associated with specific disease states.[7] As baseline sporadic mutation rates, mitosis error rates, and expected lifetime cell divisions were modelled in humans, many began to speculate that the genome may naturally vary between the cells of an organism.[8] Visual evidence provided by mosaic arrangements of skin disorders provided further evidence supporting the theory that widespread, benign somatic mosaicism was a possibility.[9,10] However, there often lacked avenues to experiment and quantify its existence. Early low-throughput karyotyping efforts and fluorescence in situ hybridization (FISH) studies suggested mosaic structural variation within an organism was a relatively common occurrence in blood.[11–14] But it was not until recent deep sequencing efforts[15–20], especially those focusing on the genome of single cells that it was shown all tissues of multicellular organisms largely consist of genetic mosaics.[21–23] Evidence suggests that post-zygotic mosaicism is widespread in a significant proportion of the cells in all tissues through unique single-

nucleotide variants (SNV), copy-number variations (CNV), and other structural variants (SV).[16] We now know somatic mosaicism is an important factor influencing disease pathology and phenotypic variability, and represents a hidden, relatively unexplored confounder in genetic testing.[24]

1.1.1 Cell-specific rates of mosaicism

The propensity for a tissue to develop mutations and clonal expansion relies on factors that include cellular turnover rate (cell lifespan), exposure to environmental mutagens and age.[16] Tissues that are shielded from environmental exposure and composed of fully differentiated, non-mitotic cells (e.g. skeletal muscle, cardiac muscle, adipose and brain tissue) often show a reduced mutational load. Whole blood, skin, liver, intestines, esophageal mucosa and lung are frequently mutated at higher rates, which can partially be explained by direct environmental exposure, and/or rapid cell turnover.[16] For example, skin cells are regularly exposed to UV radiation and other carcinogens, and as expected skin has some of the highest rates of somatic mutation and clonal mosaicism.[16] Further, a tissues propensity for mutation is exacerbated with age. With time, environmental exposure is accumulated, cellular mechanisms to repair mutations are degraded and therefore rates of mutation increase. Age is a particularly significant factor in the blood, as hematopoietic stem cells (HSPCs) rapidly turnover billions of cells a day.[3] Mutations inevitably accumulate in HSPCs over time and their progenitors populate the circulation with increasingly mutated daughter cells that can be selected for and amplified in clonal mosaic expansions.[25] Because of tissue-specific environments and architecture, rates of somatic mosaicism between tissues vary in response to age and other factors.

1.1.2 Somatic aneuploidy

In humans, a majority of post-zygotic mutations consist of SNVs and small insertions and deletions. However, a small proportion of somatic mutations are comprised of copy changes to whole chromosomes - referred to as aneuploidy.[20,26] Aneuploidy commonly arises from nondisjunction events, which is defined as the failure of homologous chromosomes to properly segregate during mitosis or meiosis. A large body of evidence

has concluded that aneuploidy is detrimental at both the organismal and cellular level.[27] In humans, aneuploidy is the leading cause of spontaneous abortions[28], and is present in ~68% of solid tumors.[29] Also, with exception of trisomy chr21 (Down's syndrome), trisomy chr13 (Patau syndrome), trisomy chr18 (Edward's syndrome) and sex chromosome abnormalities, all other constitutive aneuploidies are embryonically lethal, and rarely survivable.[30] At the cellular level, experiments in aneuploid yeast[31], mice[32] and other organisms[27] have found general fitness consequences including defects in cell cycle progression, proliferative disadvantages, increased sensitivity to cellular stress and impaired metabolic properties. Gene dosage imbalances associated with extensive gene duplication or deletion are rarely tolerated in healthy cells and are often negatively selected against. In contrast, aneuploidy is a hallmark of the cancer genome and is frequently tolerated in most neoplasms.[29] Although cancer-specific, recurrent single-chromosome aneuploidies are rare, sporadic aneuploidy is commonly observed as a consequence of general genomic instability. Cancers containing multiple aneuploidies are often more aggressive and prone to reoccurrence, metastasis and drug resistance.[33] This suggests in some cases aneuploidy provides additional genetic diversity which can allow for an increased adaptive potential, while heightened genomic instability can lead to an increased probability of acquiring additional tumor promoting alterations.[33] Although constitutive aneuploidy is severely detrimental, somatic aneuploidy in cancer provides evidence that in diseased states a mosaic of cells can withstand the presence of aneuploidy, persist and thrive.

1.1.3 Aging and somatic mosaicism

Aging is a process of gradual cellular and organismal deterioration that is aided through a combination of normal physiological processes and environmental factors. One of these processes is the accumulation of somatic DNA aberrations.[34] Somatic mutation accelerates aging through the disruption of critical genes and dysregulation of expression, which can lead to further genomic instability and therefore risk of neurodegeneration and other age-related diseases.[35] Chromosomal aneuploidy is an indication of an unstable genome and is considered a hallmark of aging.[34] One obvious model to investigate the aging process and its relation to genome instability is through rare human diseases with progeroid phenotypes – phenotypes resembling premature aging. Most progeroid

diseases involve the disruption of DNA repair (ERCC6, ERCC8; Cockayne syndrome), telomere maintenance (TERT, TERC, CTC1; Dyskeratosis congenita) or mitotic spindle checkpoint genes (BUB1B, CEP57; Mosaic variegated aneuploidy syndrome (MVA)). MVA is a rare autosomal recessive disease that is particularly relevant for investigating the relationship between aneuploidy and aging. The commonly affected gene in MVA, BUB1B (BUB1 Mitotic Checkpoint Serine/Threonine Kinase B), is a central component of the mitotic spindle checkpoint that primarily delays anaphase until all chromosomes are properly attached to mitotic spindles and prepared for segregation.[36] MVA is characterized by extremely high rates of aneuploidy (~25% of all cells), which leads to several classic progeroid features including short stature, increased risk of cancer, nervous system abnormalities, cataracts, loss of fat, curvature of the spine and premature death.[36] Murine models were developed with differing hypomorphic *Bub1r* mutations, each with varied impact on Bub1b protein levels, allowing for a gradient of Bub1b deficient mice.[37] Mice below ~10% WT Bub1r levels were not viable, however those with increasingly reduced functional Bub1b levels experienced the most severe progeroid phenotypes. Furthermore, in a follow-up study, researchers developed a murine model that overexpressed Bub1b. These mice tended to live longer and experienced less age-related tissue degeneration.[38] MVA and the Bub1b mouse provide further evidence that organisms can persist through high-levels of aneuploidy, while also accumulating detrimental, age-related phenotypes.

1.1.4 Clonal mosaicism in the blood

Studies in humans and mice have found evidence of varying levels of somatic aneuploidy in many cell types including the buccal cells [39], lymphocytes[40], fibroblasts[41], leukocytes[42], neurons[43,44] and hepatocytes.[43] The most commonly studied example of this phenomenon is in human leukocytes, where the frequency of large structural aberrations significantly increases with age. In the bone marrow, ~20,000 self renewing HSPCs and their progenitors give rise to about 10^{11} to 10^{12} new cells daily (10 billion of which are leukocytes) to maintain the required immune cell representation in peripheral circulation.[45] In HSPCs, the combination of accumulating mutations and mitotic errors arising during constant rapid cell turnover, results in a mosaic of genetically distinct

clones in hematopoietic circulation.[46] As clones compete to replicate and expand, those with positively selected traits populate the circulatory system in greater numbers. Many studies have discovered high rates of aberrant clonal mosaicism in the normal blood cells of healthy individuals that increase with age.[16,46–48] These age-associated clonal aberrations consist of mosaics of post-zygotic SNVs, CNVs and SVs that recurrently affect many genes, some of which are linked to cancer and leukemic or pre-leukemic states.[46] In a longitudinal study, autosomal SVs (>1MB) in blood were detected in 3.4% (9 / 264) of healthy subjects >60 years old, compared to 0% (0 / 342) in the younger group (<55 years old).[46] The wealth of data from genome-wide association studies (GWAS) has also been reanalyzed for mosaic SVs (>2MB) with similar conclusions. In an analysis of 26,136 cancer-free controls from 13 GWAS studies, the prevalence of large autosomal mosaic SVs increased from 0.23% in individuals under 50 years old to 1.91% between those 75 to 79 years old.[48] In the same study, when compared to healthy controls, rates of mosaic abnormalities were significantly elevated in 41 individuals who were diagnosed with leukemia within 1 year of blood collection (OR = 35.4, $P = 3.8 \times 10^{-11}$).[48] These results emphasise the importance of age on rates of clonal mosaicism in the blood and the associated risk of disease.

Subsequent studies have shown large SV mosaicism affects the sex chromosomes at much greater rates than the autosomes. Mosaic loss of the X-chromosome occurs at ~4 times the rate seen in autosomes,[49] while mosaic loss of chromosome Y (LOY) is the most common post-zygotic aneuploidy found in males.[42] Recent estimates suggest ~43.6% of men over age 70 have detectable mosaic LOY (>5% of cells affected), compared to just 2.5% at age 40.[42] Across several studies this result has been consistently replicated. LOY has been robustly correlated with smoking[50], and a suite of age-related diseases including Alzheimer's[51], non-hematological tumors[52], macular degeneration[53], immune conditions[54,55], type 2 diabetes[56], heart disease[56] and all-cause mortality (**Table 1.1**).[57] Additionally, individuals harbouring high proportions of LOY cells (>30%) have a elevated risk of recurrent cancer and a reduction in average lifespan.[42] Because LOY is a male-specific aberration, it raises the possibility that compromised Y-linked gene expression (though dysregulated tumor suppressor/oncogene expression, and/or disrupted immunosurveillance) could have a role in the male lifespan gap and male cancer incidence bias.[42,58] Further research into the effect of LOY at a cellular level in multiple tissues is necessary to validate these hypotheses and to fully understand its role in the development of age-related disease.

Association variable	Cohort size	Results	P-value	Technology	Tissue sampled	Reference
All cause mortality	982	HR = 1.91 [95% CI 1.17-3.13]	0.01	SNP-array, WGS	Blood	Forsberg *et al.*, 2014
Alzheimer's disease	606 case, 1005 control	OR = 2.80	0.0184	SNP-array, WGS	Blood	Dumanski *et al.* 2016
Autoimmune thyroiditis	31 case, 88 control	Case = 1.95% LOY [0.56-7.2%] Control = 1.31% LOY [0.2-5.6%]	0.037	SNP-array	Blood	Persani et al.
Cancer mortality	982	HR = 3.29 [95% CI 1.51-7.15]	0.003	SNP-array, WGS	Blood	Forsberg *et al.*, 2014
Longer duration of schizophrenia	146 case, 360 control	OR = 1.11 [95% CI 1.03-1.19]	0.007	qPCR	Blood	Hirata *et al.*, 2018
Macular degeneration	5772 case, 6732 control	HR = 1.33 [95% CI 1.206-1.472]	1.6e-08	SNP-array	Blood	Grassmann et al. 2018
Secondary cardiovascular events	366 case	HR = 2.28 [95% CI 1.06-4.76]	0.02	SNP-array	Blood, atherosclerotic plaque	Haitjema *et al.* 2017
Smoking	634 current, 3507 never 218 current, 892 never 48 current, 439 never	TwinGene: OR = 4.3 [95% CI 2.8-6.7] ULSAM: OR = 2.4 [95% CI 1.6-3.6] PIVUS: OR = 3.5 [95% CI 1.4-8.4]	1.31e-11 0.0006 0.02	SNP-array	Blood	Dumanski *et al.* 2015
Smoking	27,748 current 108,859 never	3.4% LOY current [mLRR-Y <-0.15] 0.92% LOY never [mLRR-Y <-0.15]	7.9e-50	SNP-array	Blood	Loftfield et al., 2018
Smoking	934 current 5410 former 3408 never	- OR = 1.33 [95% CI 1.12-1.57] OR = 2.35 [95% CI 1.82-3.03]	- 0.001 5.55e-11	SNP-array	Blood, buccal	Zhou et al. 2016
Solid tumors mortality	982	HR = 3.62 [95% CI 1.56-8.41]	0.003	SNP-array, WGS	Blood	Forsberg *et al.*, 2014
Testicular germ cell tumor	678 case, 774 control	OR = 0.34 [95% CI 0.10-1.17]	0.09	qPCR	Blood	Machiela *et al.* 2017

Table 1.1 Summary of epidemiological studies investigating associations with mosaic loss of chromosome Y
HR = Hazard ratio; OR = Odds ratio

1.1.5 Somatic mosaicism in the brain

Despite low neuronal turnover rates, a number of studies have found evidence of somatic mosaicism in the brain that likely arises during development and expands with age.[16,22,59] Recent deep sequencing and single-cell sequencing efforts have found that individual neurons commonly possess ~800–2000 unique SNVs[59], and 13–41% of human frontal cortex neurons contain large SVs (>1MB).[22] Mosaic aneuploidy in the brain also appears common. Estimates from early studies using FISH and spectral karyotyping vary widely[21], but recent high-resolution, single-cell whole genome sequencing (scWGS) studies have shown that neuronal aneuploidy ranges from 0.6 to 4.9% across the autosomes.[22,43,44,60] There are several factors leading to the genetic heterogeneity seen in the brain. Extensive cell division during development can lead to mutations in neuronal stem/progenitor cells that are propagated to all future cells in the lineage leading to clonal mosaic expansion.[21] Accumulation of age-related mutation is caused by both exogenous factors including radiation, viruses, and mutagenic chemicals and endogenous factors including cytosine deamination, reactive oxygen species, and defective DNA damage repair.[46] Benign somatic mosaicism in neurons may provide benefits through increased genetic diversity and transcriptional heterogeneity allowing for clonal populations to form specialized networks and functions.[61] However, neuronal specialization is already a tightly regulated process and somatic mutation is a known risk factor for intellectual disability, and neurodegenerative disorders.[61]

1.2 Mechanisms of chromosome Y loss in the blood

Mosaic loss of chromosome Y in the blood likely occurs as a result of segregation errors during rapid haematopoietic replication.[42,52] Clones lacking the Y may infer a selective advantage and proliferate, however there are several theories on what specific mechanisms lead to this selective advantage. Chromosome Y is the smallest chromosome, is not required for cell viability, and only contains 9 genes that are ubiquitously expressed outside of the testes. This makes the Y chromosome the most dispensable human chromosome.[62] As a result, LOY could be a marker of general genomic instability as its loss is more easily tolerated, and is therefore observed at greater frequency. Another theory postulates that an important Y-linked growth suppressing gene is deleted when

the Y is lost and proliferation capabilities benefit.[42] For example, Y-linked genes ZFY and UTY, are considered potential tumor suppressor genes.[63] Both genes have X chromosome homologs (ZFX, UTX) that have tumor suppressor qualities and escape X-inactivation and could provide an explanation for recurrent Y chromosome loss in many cancers. Furthermore, key pseudoautosomal region (PAR) genes such as CD99 and SLC25A6, which are involved in leukocyte migration and apoptosis, respectively, could be subject to dysregulation during Y chromosome loss.[64] However, these competing theories are not mutually exclusive, and it is possible passenger and causal processes are occurring simultaneously.

Chromosome Y abnormalities including microdeletions and nullploidy are frequently observed in many human malignancies[65], although in most cases there is limited evidence that Y loss directly leads to tumor progression.[66,67] In testicular germ cell and prostate tumors, male-specific Y chromosome microdeletions are found in almost all samples, with more Y-linked genes being affected in advanced stages.[67] Furthermore, bladder, renal cell, prostate, head and neck, esophagus, male breast cancer, hematological disorders and several other cancers show significant rates of LOY, although role of LOY in oncogenesis remains unclear [67,68] Interestingly, gain of Y events are observed but much less frequently.[68] Despite correlation between Y loss and cancer progression, patterns are complicated and tissue dependant. In bladder malignancies evidence suggests LOY is a passenger event that does not affect tumorigenesis, however in sporadic colorectal cancers LOH in the PAR region may play a role in tumor progression.[68]

1.2.1 Epidemiological associations and GWAS loci

Despite broad epidemiological associations between LOY and age-related disease, determining whether LOY is a causal force or a passenger biomarker has been elusive. There are two proposed theories used to explain LOY and its association with disease. The first theory suggests that LOY directly causes cellular and tissue dysfunction that results in disease. While evidence supporting this theory is limited, some suggest that cells without a Y may lack a Y-linked tumor suppressor and/or a key immune-related gene which could independently contribute to the pathogenesis of disease. Although the Y chromosome is gene poor, its complete deletion could

directly lead to disease. The second theory suggests that LOY is a biomarker of general genomic instability and a recurrent passenger mutation that signals a decline in the DNA maintenance capabilities of a cell. As the genes involved in DNA repair degrade, LOY prevalence increases alongside other disease-causing aberrations which leads to robust correlations with age-associated disease. However, both of these hypotheses could be true. A recent large UKBiobank GWAS study investigating the genetic risk factors for LOY found 137 novel significantly associated loci in addition to 19 previously replicated loci.[42] LOY-associated variants were enriched near somatic cancer drivers (CHEK2, TERT) , cancer therapy targets (PARP1), cell-cycle control genes (CCND2, CDKN1B), and DNA damage response pathway genes (SETD2, TP53). Genetic risk scores using the 156 LOY-associated variants found that LOY associated genotypes increased risk of prostate cancer (Odds ratio (OR) 1.68 (95% CI 1.33–2.11); $P = 1.9 \times 10{-}5$), testicular germ cell tumours (OR 2.97 (1.45–6.07); $P = 0.003$), and glioma (OR 2.3 (1.34–4.17); $P = 0.004$). These findings and similar previously findings suggest germline variants could predispose individuals to LOY, and these same loci predispose to cancer, suggesting LOY is likely a consequence of general genomic instability.[42,69] To further elucidate the causal force of LOY, Thompson *et al.* tested whether LOY susceptibility loci affected disease risk in females.[42] Because females lack a Y chromosome, any significant association would suggest that LOY-associated loci affect general genomic instability, and not Y chromosome-specific instability. When the LOY loci were given polygenic risk scores for genomic instability-linked, female-specific conditions, both breast cancer (OR=1.25 (1.04-1.49); $P = 0.016$) and later age at menopause ($P = 0.003$) were significantly associated.[70] These results provide evidence that LOY is a biomarker of general genomic instability, however they do not exclude the possibility that LOY could still be a causal factor in disease. Further investigation into LOY-specific mechanisms and expression patterns is required to answer these questions.

1.3 Genetics of chromosome Y

Humans have an XY sex-determining system, with the presence of the *SRY* locus on the Y chromosome determining male sex and development. The human X chromosome is ~156Mb in length and contains 1,480 known genes (841 protein coding, 639 long and short noncoding RNAs; Ensembl release 99). In comparison, the

Y chromosome is ~57.2Mb in length, represents about 2% of the human genome and is considered gene-poor, containing ~173 genes (66 protein coding, 107 long and short noncoding RNAs; Ensembl release 99).[71] For many years the perceived function of chromosome Y was confined to sex determination and spermatogenesis.[72] Modern genetics studies have reinforced these functions, while also uncovering several ubiquitously expressed Y chromosome genes and associations between Y-linked gene expression, the immune system and human disease.[62]

1.3.1 Evolution

Mammalian sex chromosomes evolved from a homologous autosome pair ~180-220 million years ago (mya) after one member, proto-Y, acquired a sex-determining locus (*SRY*).[72] During this time, these proto-X and proto-Y chromosomes were of similar size and recombined across their entire length. Over millions of years the pseudo-sex chromosomes began differentiating into the current human X and Y. Through a series of chrY-specific inversion events, the transfer of genetic information between X and Y was eventually limited to the pseudo-autosomal regions (PAR). Each chromosome began independently evolving, leading to significantly different genetic content and function. Despite the extensive divergence, modern sequence in the PARs is near 100% identical between X and Y, and crossover between these regions is required for proper segregation in male meiosis.[68] PAR1 covers 2.7Mb on the short arm of X and Y, while PAR2 spans 0.32Mb on the long arms, together comprising 4.6% of the Y chromosome and hosting several critical genes (**Figure 1.1**).[72]

1.3.2 Male specific Y region (MSY)

The remaining 95% of chromosome Y is referred to as the male-specific Y region (MSY), and it is composed of both heterochromatic and euchromatic regions. The heterochromatic region comprises much of the q arm of chrY and is largely genetically inert, although the region may play a role in chromatin remodelling.[73] The heterochromatic region largely consists of repetitive sequence from highly repetitive sequence families (DYZ1 and DYZ2), and is polymorphic in length between male populations. DYZ1 is a 3.4KB sequence largely

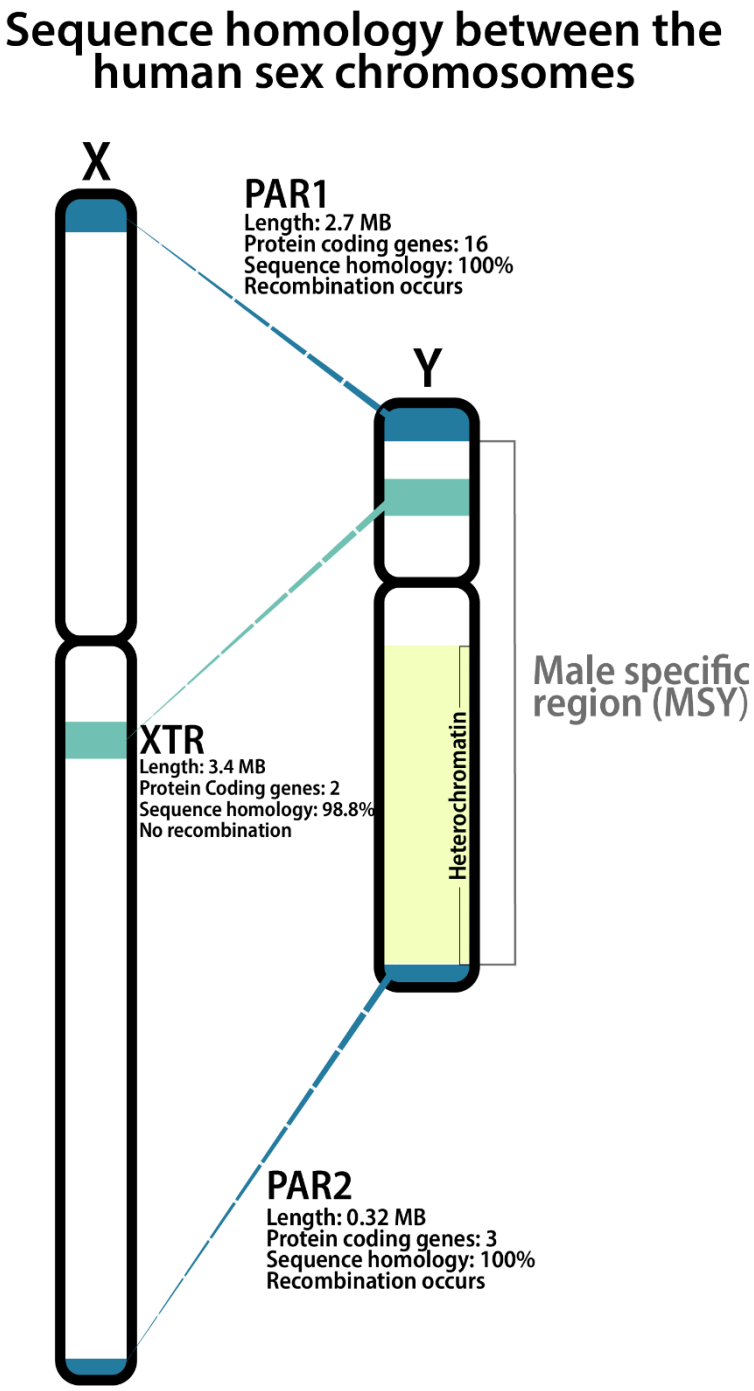


Figure 1.1 The human sex chromosomes share three regions of high sequence homology.
The pseudoautosomal regions (PAR), PAR1 and PAR2 are 100% homologous. The X-transposed region (XTR) shares ~99% sequence homology. The Y chromosome PAR1 is ~2.78 million bases (Mb) spanning from Y:10,001 to 2,781,479. The chromosome Y PAR2 is ~0.33 Mb and spans from Y:56,887,903 to 57,217,415. PAR1 and PAR2 on the Y chromosome are identical in sequence to the X chromosome PAR1 X:10,001 - 2,781,479 and PAR2 X:155,701,383 to 156,030,895. The male specific Y region (MSY) constitutes 95% of the Y sequence and does not recombine with the X.

composed of the 5bp repeat "TTCCA". The region is polymorphic in length between populations, but most males have approximately 3000-4300 copies of DYZ1, constituting about 20% of the entire Y chromosome.[74] The euchromatic MSY can be divided into three main classes: X-transposed region (XTR), X-degenerate region (XDR), and the ampliconic region, each with a unique evolutionary history.[73] The XTR (3.38Mb) shares 98.78% sequence homology with Xq21 and is a result of an X to Y duplication event that occurred 3-4mya after the divergence of humans and chimpanzees.[73,75] The region contains two protein coding genes (PCDH11Y, TGIF2LY), that both possess X-linked homologs.[76] The XDR consists of remnant sequences from the ancient autosomes that X and Y originally evolved from. The XDR contains 16 protein-coding genes (most of which are expressed widely) in addition to many pseudogenes. Most genes in the XDR have X-linked homologs with varying sequence identity (30-96%).[73] The ampliconic region is composed of palindromes and inverted repeat sequence that contains 9 genes, most of which are exclusively expressed in the testes.[73] The palindromic sequence in the MSY enables local recombination in an otherwise non-recombining region. Palindromes enable intrachromosomal gene conversion that likely decrease rates of chrY degeneration and increase ability to adapt. It is complicated, repetitive regions such as the ampliconic region that make Y chromosome sequencing efforts challenging.

1.4 Loss of Y detection methods

The goal of this section is to establish the historical and methodological groundwork of technologies used to karyotype human samples specifically in the context of chromosome Y. The most commonly used and most impactful technologies have been described in detail, however a full summary of all methods has been provided in **Table 1.2**.

Method	mLOY detection	Pros	Cons
Karyotyping	Cells are suspended in metaphase, and condensed chromosomes are stained and visually inspected.	Genome-wide detection, single-cell resolution, low error rate	Labor intensive, very low throughput, requires cell culturing, cells must be actively dividing.
Interphase fluorescent in-situ hybridization (FISH)	Probes with Y-specific sequence are inserted into a single-cell. If the probe binds it fluoresces and is visually detected through microscopy.	Inexpensive, high sensitivity, single cell resolution	Labor intensive, requires technical expertise, low throughput, probe design requires prior knowledge of abnormality of interest.
Quantitative fluorescent PCR (qPCR)	Probes to highly homologous but slightly differing genes, AMELY and AMELX, are used to quantify the genomic content ratio between X and Y chromosomes.	Inexpensive	Not scalable, cannot distinguish deletions and duplications from complete aneuploidy at the probe site, labor intensive.
Single-nucleotide polymorphism array (SNP-array)	Median probe intensity (mLRRY) across probes in the male specific Y region. Allelic ratios in the PAR region can also be used to determine relative X and Y abundance.	Highly scalable, relatively inexpensive for large cohorts, accurate, processing pipelines are well established, can detect deletions, amplifications, LOH and copy neutral LOH, small data files.	Cannot detect balanced rearrangements, lacks single-cell resolution, LOY detection is contingent on >5-10% of cells being LOY, samples usually confined to whole-blood tissues in online databases.
Bulk whole genome sequencing (WGS)	In each sample chrY read depth is measured across genomic intervals and normalized to genome-wide read depth. Allelic ratios of variants in the PAR region can also be used.	Base pair resolution, enhanced coverage, scalable, can detect balanced rearrangements, often available in a wide-range of tissues in online databases.	Expensive, read mapping biases and artefacts are common, data can be difficult to analyze, large amount of data to store
Single-cell WGS (scWGS)	In each cell chrY read depth is measured across genomic intervals and is normalized to cell-specific genome-wide depth.	Single-cell resolution, high chromosomal resolution even at low coverage (1-2%), can survey thousands of genomes from a single sample.	Expensive, whole genome amplification adds additional PCR artefacts, large amount of data to store, library preparation can be labour intensive
Bulk RNA-sequencing (RNA-seq)	Common approaches include measuring read depth across genomic windows of each gene region on a chromosome and comparing to genome-wide average.	Provides additional gene expression information that can be synthesized with aneuploidy/CNV information, RNA-seq data from many tissues is widely available in public databases.	Moderately expensive, using RNA as a proxy for DNA introduces error and variability, highly affected by sequencing depth, chrY expression is limited in many tissues, difficult to detect low-level mosaicism, lacks single-cell resolution
Single-cell RNAseq (scRNA-seq)	Assumes LOY in a cell if there is a complete absence of male-specific Y gene transcripts.	Provides additional cell-type and gene expression information that can be used downstream, single-cell resolution, widely available in public databases.	Expensive, using RNA as a proxy for DNA introduces error and variability, chrY expression is limited in many tissues, dependent on per-cell read depth.

Table 1.2 Summary of methods used to detect Y chromosome loss.

1.4.1 Karyotyping and fluorescent staining

Beginning in the 1960's, in the infancy of cytogenetics, several labs undertook large karyotyping projects, often observing thousands of cells to estimate the prevalence of aneuploidy in tissues of healthy individuals in response to aging.[11,12] Before the use of fluorescent banding techniques, when chromosome identification was difficult and chromosomes were grouped on size and centromere location, Jacobs *et al*. found age associated aneuploidy in female M chromosomes (chr6-12, chrX), and the male Y chromosome (which could be individually identified with "good preparations").[11] In the 1970's chromosomal banding was developed, and individual chromosomes were more easily identified. Several subsequent studies observed age-associated mosaic loss of chromosome Y (45,X0 - LOY) in males and chromosome X in females (45,X0) in circulating leukocytes[77], and bone marrow[7,12,78] at higher rates than autosomes especially in individuals over 60 years of age.

Chromosome Y loss was also consistently observed in various leukemias[79–81] and hematological disorders[82], and investigators were eager to use Y loss as a cancer biomarker. However, the observation of age-associated LOY in both normal aging individuals and those with hematological disease created controversy in the field. It wasn't clear whether Y loss was involved in the malignancy process or was simply a neutral by-product of aging.[83] Many concluded that LOY did not carry a special hazard for hematological disorders and was not an important factor for prognosis.[82,83] However, it was also discovered that LOY was a common occurrence in childhood leukemia which provided evidence against Y-loss exclusively as a result of age.[84] The mechanism and disease implications of chromosome Y loss were still not well understood.

Classical karyotyping is still a commonly used method to observe the ploidy of cell cultures and prenatal samples as it is easily assessable, inexpensive, and provides single-cell resolution. Classical karyotyping requires cells to be in metaphase, and to collect a sufficient number of cells for analysis cells must be cultured and actively replicating. During metaphase, the chromosomes reach peak condensation and can be observed via microscope. In a process known as G-banding, metaphase chromosomes are treated with trypsin, a protease that relaxes the chromatin structure and allows stains such as Giemsa to bind the DNA.[85] G-banding provides additional resolution to the karyotype and allows for the identification of specific chromosomes, and large structural

aberrations. When investigating aneuploid mosaicism, karyotyping does have some limitations and biases. As the number of cells affected by a mosaic abnormality is reduced, a greater number of cells must be assayed in order to detect them. This limits the ability for classical karyotyping to accurately predict low-frequency mosaicism. Also, cells must be cultured and actively replicating which can potentially lead to an overestimation of true aneuploidy events.[86] Nevertheless, karyotyping is still a practical and cost-effective method for scanning the ploidy of single cells across the entire genome.

1.4.2 Fluorescence in-situ hybridization

In the late 1970's classical cytogenetics were improved through the use of *in situ* hybridization of fluorescently labelled probes, commonly referred to as FISH.[87] The technique was revolutionary because it allowed for high-resolution detection, quantification and visualization of user-designed nucleic acid probes. FISH was particularly useful in clinical newborn chromosomal abnormality detection, as the assay was accurate, but also rapid and did not require the 10 to 21-day cell culture required for conventional cytogenetics.[87] The widespread adoption of FISH allowed for larger sample-size aneuploidy surveys. In 1993, across 2490 pooled cases, the association between LOY and age was confirmed again, with an additional finding that the percentage of cells affected by LOY in an individual also increased with age.[88] Oddly, LOY was observed at similar rates in normal, pre-leukemia and acute myeloid leukemia cases, which again supported the notion that LOY was a neutral, age associated aberration that was not causal in disease-related processes.

As a result of the widespread belief that the Y-chromosome was a "genetic wasteland", in combination with technical difficulties detecting its presence compared to autosomes[89], research on the Y slowed in the 1980's and 1990's.[90] FISH evolved into comparative *in situ* hybridization technologies such as comparative genomic hybridization (CGH) and later array-based CGH genome-wide assays that could accurately detect chromosomal abnormalities by separately labelling case and control samples and comparing relative fluorescence. Array-CGH

was primarily used for tumor and embryo karyotyping.[91] To my knowledge, applications towards aneuploidy and LOY in healthy individuals were limited.

1.4.3 Single nucleotide polymorphism arrays (SNP-array)

In the late 2000's, the use of SNP-arrays exploded primarily as a result of the widespread adoption of genome-wide association studies (GWAS).[92] Commercial SNP-array platforms allowed for rapid and accurate genotyping of millions of genetic markers that could be applied to thousands of individuals and resulted in the discovery of many disease-associated loci. The standardization of the platform and increased availability of array-based data led to interest in the prospect of applications in CNV and SV calling.[93] Several aspects of SNP-arrays were well suited to the problem, including dense genomic coverage and allele-specific probe design. At each locus, a SNP-array contains two unique hybridization probes, each specific to the two known alleles at the site (and named A and B). The genotype at this specific SNP can be determined by the ratio of hybridization intensities between probe A and B. SV and CNVs can be detected by comparing the median hybridization intensities across multiple SNPs to a set of diploid reference intensities, resulting in what is known as the log R ratio (LRR). LRR values around 0 are considered copy-neutral, whereas deviations from 0 above or below a threshold are considered either gain or loss events. For example, duplications result in additional genomic content and this is observed on the SNP-array through heightened hybridization intensities in the affected regions compared to the diploid control population. This event produces an LRR significantly greater than 0. Allele-frequency ratios at heterozygous SNP sites can be used as an orthogonal line of evidence. Consistent deviation from expected allelic ratios across genomic segments indicates the existence of copy-number alterations. Through the use of hybridization intensities and allele frequencies on SNP-arrays, SV calling including mosaic SVs (affecting >10% of cells)[57] can be called through several widely available tools.[94,95]

The current state and awareness of mosaic loss of Y research in males can largely be attributed to work from the labs of Lars Forsberg and Jan Dumanski and their initial use of genotyping arrays to investigate mosaic

Y chromosome aneuploidy in longitudinal cohorts containing thousands of aging individuals.[57] The groups from Uppsala University in Sweden have innovated and standardized mosaic aneuploidy detection for the Y chromosome and have brought the surprisingly common aberration to the attention of the medical genetics community. In 2014, Forsberg *et al.* genotyped whole-blood samples from 1153 men extracted between the ages of 70.7-83.6 using Illumina's 2.5MHumanOmni array.[57] Originally, the group was focused on small, acquired, Y-linked structural variants including deletions, gains, duplications, and acquired uniparental disomy. However, to their surprise the most common aberration was the total loss of Y. After removing individuals with previous cancer diagnoses, 8.2% of men showed significant mosaic loss of Y (~>10% of all blood cells lacking a Y chromosome).[57] As blood draws were taken longitudinally in this cohort, the cause of death of many individuals within the study was already known. Using this extensive longitudinal metadata, a continuous measure of LOY severity (median LRR across the Y male specific region of chromosome Y; referred to as mLRRY) was significantly associated with risk of all-cause mortality (HR=1.91, p=0.01), and risk of developing both hematological (HR=3.29; p=0.003) and non-hematological cancers (HR=3.62; p=0.003). The results were replicated in a similar cohort of 1016 Swedish men (PIVUS cohort). This paper reignited interest in LOY and provided strong evidence that LOY in healthy males could potentially be a biomarker for age-related diseases such as cancer.

Subsequent studies using SNP-arrays found robust associations between LOY, environmental exposures [50,96] and disease risk [51,53], providing evidence against the long-held notion that LOY in the blood was a phenotypically neutral event. In 2015, Dumanski *et al.* found that smoking had a dose-dependant association with LOY in blood.[50] Current smokers were found to have a 4-fold risk for LOY in blood compared to non-smokers, and individuals that smoked more frequently had an increased acute risk of LOY that was reduced to a baseline ~20 years following smoking cessation. The following year Dumanski *et al.* found significant associations between LOY and Alzheimer's disease, further suggesting that LOY could have causal effects on disease development [51] These flagship studies paved the way for several epidemiological studies aimed at finding further trait associations with LOY. These studies concluded LOY in blood is associated with increased risk of various

cancers[97], macular degeneration[53], autoimmune thyroiditis[55], biliary cirrhosis[54], major cardiac events[98], obesity[56] and type 2 diabetes (**Table 1.2**).[56] In 2019, 205,011 men from the UK Biobank were surveyed for LOY, making it the largest LOY study to date.[42] Using allele-specific genotyping intensities, strong age-LOY associations were replicated once again and over 20% of the surveyed population showing detectable mosaic loss (>5% of cells with LOY).

1.4.4 Whole genome sequencing (WGS)

The ability to sequence genomes in their entirety has improved the ability to detect somatic variation and has drastically altered the landscape of medical research and clinical genetics. Advancements in microfluidics, fluorescence microscopy, computational power and the completion of the human genome all combined to advance low-throughput Sanger sequencing to modern high-throughput DNA sequencing. Although each brand of short-read sequencer has a different protocol, the resulting output data consists of millions of short 90-300bp reads. Using *in silico* mapping algorithms, each read is mapped back to its original location on the reference genome, essentially re-assembling the genome and providing the full sequence. Over ~48 hours, modern machines (Illumina NovoSeq4000) can sequence 2.5 billion 300bp reads, creating 750GB of genomic data.[99]

For digital karyotyping and SV detection, short-read whole genome sequencing improves on SNP-array technology by providing genome-wide depth at base-pair resolution. At standard coverage (30x), hundreds of millions of mapped reads provide the information required to accurately deduce CNVs, and structural variants (SV) including inversions, translocations and whole chromosome aneuploidies.[100] Most SV detection programs use a combination of read depth analysis and alternative allele fractions.[101–103] Read depth analysis uses fixed or sliding genomic windows to calculate the median read depth across a genomic interval which is then compared to the median read depth genome-wide.[104] Several studies have found that SNP-array mLRR-Y values are highly correlated with LOY estimates derived from WGS.[105] But despite rapid reductions in the cost of sequencing, WGS

remains expensive especially in comparison to SNP-arrays and as a result WGS is commonly used as an orthogonal line of evidence in support of LOY results found using SNP-arrays.[100]

There are also several confounders that need to be addressed with short-read sequencing including systemic GC content, PCR and mapping biases that can distort read depths and lead to errors when calling aneuploidy.[100] When investigating mosaic events these biases become increasingly confounding. GC content bias is caused by a propensity for GC-poor and GC-rich reads to be unrepresented.[106] To correct for this, models can be fit to the distribution of GC % across all reads and read depth values are adjusted. Mapping biases are more difficult to correct but masking known difficulty regions and ambiguous segments of the genome improves the quality of the data.[107]

All of the technologies discussed thus far have either investigated aneuploidy at single-cell resolution in a low-throughput manner (karyotyping, interphase FISH) or provided high-throughput aneuploidy calls using bulk tissue and millions of cells (SNP-array, WGS). When investigating mosaic aneuploidy, the ideal technology would combine the positive aspects of both high- and low- throughput technologies. Single-cell resolution is necessary for observing low-frequency mosaic events. For example, median SNP-array intensities can only detect LOY affecting >10% of cells, as is the case with bulk WGS. On the other hand, high-throughput and other scalable technologies are required to expand sample sizes and test thousands of individuals in order to derive statistically significant conclusions that are representative of the population and are thus generalizable. It is simply not feasible to karyotype thousands of cells from thousands of samples, and bulk technologies cannot detect low-frequency events. However, modern advances to single-cell sequencing technologies do provide these desired characteristics to accurately assess genome-wide mosaicism at all frequencies.

1.4.5 Single-cell WGS sequencing (scWGS)

scWGS analyzes the genome of individual cells through improvements to traditional high-throughput sequencing and is a powerful method for quantifying genetic mosaicism.[108] Advancements in library preparation

allow for DNA from individual cells to be uniquely tagged with oligonucleotide identifiers, multiplexed with the other cells and sequenced together.[108] Additionally, cells are subject to whole genome amplification (WGA) before sequencing which can confound results. Each individual cell only contains ~6 picograms of DNA which needs to be amplified hundreds of times to reach concentrations required for effective sequencing.[108,109] PCR amplification can create read depth biases and amplify technical artefacts which limits CNV calling resolution and can lead to incorrect conclusions when calling CNVs and whole chromosome aberrations.[110] However, even with genomic coverage as low as 0.5-1% per cell, chromosomal copy-number can still be readily calculated at similar or greater resolution than is available from array technologies.[110] Single-cell WGS aneuploidy detection algorithms are similar to those used in bulk WGS, measuring the number of reads mapped to a chromosome and comparing this to the genome average for each cell.

Aneuploidy studies using single-cell WGS in human skin[44], oocytes[111] , sperm [112], liver [44], and brain [22,43,44,60,113] have provided, high-resolution, baseline aneuploidy rates that are significantly lower than those concluded by FISH studies. For example, Knouse *et al.* found that aneuploidy rates in human neurons (2.2%; 95% CI 0.3%–7.9%; n=89) were much less than rates from similar studies using FISH which commonly exceed 20%.[44] Another single-cell study by van den Bos *et* al. found 0.6% of 1482 neurons were aneuploid.[43] In the human liver Knouse *et al*. found 4% (95% CI 1.1%–9.9%; n=62) of cells were aneuploid, while 0% (95% CI 0–6.7%; n=53) of sampled skin cells were aneuploid. [44] Rates of mosaic aneuploidy appear to be over-estimated by FISH, but still relatively common across many tissues.[105] As single-cell sequencing technology improves, studies with greater sample sizes will continue to provide a higher resolution quantification of aneuploidy in humans.

1.4.6 Single-cell RNA sequencing (scRNA-seq)

Although not commonly used for detecting SVs, the transcriptome can provide sufficient information to successfully karyotype samples.[114,115] Single-cell RNAseq is promising as a method to detect aneuploidy as cell-type can be predicted through gene expression markers, allowing for cell-type specific, single-cell resolution

estimates of aneuploidy. Additionally, because of interest in gene expression differences between tissues, much more scRNAseq data has been produced on a wider range of tissues compared to scWGS. When investigating systemic genetic heterogeneity in humans, tissue variety is valuable. Previous studies have shown proof of principle that scRNAseq can detect aneuploidy.[114] Through the use of combined genome and transcriptome sequencing, Griffiths *et al*. benchmarked an approach called *scpoid*. In this method, in each cell, chromosomes showing consistent expression deviation from other cells in the sample are considered aneuploid. The method showed effectiveness when input gene expression showed low variability and high depth. Similar methods have been used to study chromosomal instability in cancer.[116,117] Recently, the *scpoid* method was improved to utilize allele-specific expression and showed utility when detecting aneuploidy in embryos.[118] However, in both of these approaches the sex chromosomes are removed. That being said, single-cell RNAseq has previously been used to detect LOY.[42] Thompson *et* al. used a simple method where cells were labelled LOY if they lacked expression from all genes residing in the male specific Y (MSY). Using this technique, they labelled 15.6% of 13,418 peripheral blood mononuclear cells (PBMC) as LOY (ranging from 7-61% across all individuals), which is roughly comparable to estimates from SNP-arrays. Specifically in B lymphocytes, differential expression analysis between LOY cells (n=277) and normal cells (n=2,459) concluded that TCL1A, a known leukemia driver, was overexpressed in LOY cells (Fold charge (FC): 1.75, $p < 0.0001$).[42] Continuing analyses such as this, while improving the methods and fine tuning them to better incorporate chrY will help understanding of mechanisms and prevalence of LOY across the human body.

Despite the utility of inferring DNA from scRNAseq, there are inherent technical issues and biological realities that complicate LOY detection when using RNA. Mainly, only 9 male-specific Y genes are commonly expressed outside of the testis, and in some cell-types expression of these genes can be difficult to detect.[72] Furthermore, total sequenced read depth per cell depends on a snapshot of transcriptional output of a cell before sequencing. As result, read depth can vary widely between multiple distinct cell types in heterogeneous tissues such as the brain.[119] Because of variability and dependency on read depth, LOY may be overestimated. To limit false positives, sequencing must be deep, and several quality control filters should be in place.[120]

1.5 Sex chromosome specific challenges in NGS

Advances in next generation sequencing technologies, and the improved availability of tools used to analyze the data have drastically enhanced the reliability, cost and practicality of NGS for use in research and clinical settings.[121] Despite these advances, accurately mapping highly repetitive and homologous sequence remains a challenge for short-read sequencing.[122] Repetitive sequences create ambiguities when aligning to the reference genome which can cause incorrect mapping and result in confounded gene expression estimates, variant calling, aneuploidy estimates and other results.[122] Basically, if a read can ambiguously map to several locations in the genome, accurately predicting the true biological original location of this read is not possible. Because of the shared evolutionary origin of the mammalian sex chromosomes, X and Y share a high level of sequence similarity that makes analysis using short-read sequencing more challenging.[123] As mentioned above, chromosome Y also contains an irregular enrichment of repetitive sequence which invokes mapping ambiguities as well. Additionally, chromosome Y exists in a haploid state, meaning there is 50% less genetic material to sequence compared to autosomes, which reduces depth and ability to make accurate variant calls. The increased difficulty of mapping chrY has led to its frequent removal from genomic analyses, potentially leading to a gap of knowledge in human genetics, evolution and disease.[124]

When mapping sex chromosome reads to the reference it is highly recommended to hard mask the homologous PAR on the Y chromosome.[125] Hard-masking is a manipulation of the reference genome where previously determined problem regions are removed and replaced with a placeholder ("N") . When the PAR is hard-masked on either the X or Y, ambiguity is removed in the reference and the PAR is effectively treated as any other diploid autosomal region. Without hard masking the PAR on one of the sex chromosomes, reads can map identically to either the X or Y, leading to poor mapping quality, reduced coverage and poor-quality variant calling.[125] However, most genes and pseudogenes on the Y have X-linked counterparts of high sequence similarity that are not masked in the reference which can result in read mismapping and reduced quality.[125] For example, the XTR regions retains >98% homology between X and Y and reduced mapping quality and increased technically

difficulties are observed in this region.[125] Several tools have been developed to overcome some of the challenges associated with highly repetitive and homologous sequence.[126]

1.6 Objectives and hypothesis

Recent findings of widespread LOY in the leukocytes of aging men and its association with risk of Alzheimer's disease raised interest in determining LOY rates directly in brain tissue. WGS and single-nuclei data from brain were available, but because of the haploid nature, low gene count, and general genetic and evolutionary complexity of the Y chromosome, many published tools designed to estimate mosaic aneuploidy from WGS data remove it before analysis and solely focus on the autosomes. Furthermore, methods to investigate Y chromosome aneuploidy using scRNAseq were largely absent from the literature. My overall objective for this thesis was to use genomic context, genomic characteristics, and manipulation of sequence alignment files to improve and customize mosaic aneuploidy detection for the Y chromosome. This main objective can be described in three main aims:

i) *Improve WGS-based, mosaic Y chromosome detection through the use of genomic characteristics (i.e. mappability, GC content) using standardized DNA array-based LOY values as baseline.*

In the past, several studies have used LOY measures from WGS as an orthogonal line of evidence to complement and further validate the primary array-based estimates of LOY. This is because DNA is commonly extracted from blood samples for array genotyping and LOY is commonly investigated in blood. The dataset available required the primary use of WGS to detect LOY in brain samples. Therefore, the first objective was to implement strict quality control and the use of mappability and GC content to reduce technical variation and improve the biological LOY signal in blood, cerebellum, and cortex samples. Correlation between paired WGS/SNP-array data was used as a baseline.

ii) *Develop a pipeline for filtering and estimating individual cell LOY using low-depth single-nuclei RNAseq.*

The transcriptome has been shown to effectively detect aneuploidy in single-cells. The objective was to develop a pipeline to reliability detect LOY using single-nuclei RNAseq. The pipeline had to overcome low Y-linked gene expression, gene dropout, and highly variable single-nuclei expression levels through the use of aforementioned genomic characteristics, in addition to single-cell sequence alignment file manipulation.

iii) *Determine cell-type specific rates of LOY in the brain.*

The transcriptome of single-cells has been shown to accurately predict cell-type. Using cell-type information inferred from gene expression, my objective was to i) find evidence of LOY in single-nuclei within the brain and, ii) determine the rate of LOY in brain cell-types (i.e. microglia, oligodendrocytes, astrocytes and neurons.)

My main hypotheses are as follows:

i) Filtering regions of high homology, high repetition and abnormal GC content will improve the overall accuracy of mosaic loss of chromosome Y estimates.

ii) As a result of higher turnover rates and rates of replication, glial cells have higher rates of LOY than do neurons.

2. Improving WGS-based mosaic loss of Y detection

2.1 Chapter introduction

Previous study of mosaic loss of chromosome Y (LOY) has largely been confined to blood tissue.[42,51,52,57] However, given links between Alzheimer's disease (AD) and aneuploidy in the brain[39], and recent associations between LOY in the blood and AD[51], I wanted to determine the prevalence of LOY directly in human brain tissue. In order to make reliable conclusions on LOY rates in the brain, existing WGS LOY detection methods needed to be tested and improved. Accordingly, a large portion of the effort in this chapter was devoted to improving the ability to detect mosaic aneuploidy from WGS data, specifically in the context of chromosome Y. The existence of individuals with both SNP-array and WGS data, and/or multiple WGS runs allowed for in-depth quality control, providing confidence in our method and findings. In this chapter I used both SNP-array and WGS data from whole blood, dorsolateral prefrontal cortex, and cerebellar tissue to estimate tissue-specific rates of mosaic loss of chromosome Y (LOY) in elderly men.

2.2 Methods

2.2.1 ROSMAP cohort and data summary

The Religious Orders Study (ROS) and Rush Memory and Aging Project (MAP; together referred to as ROSMAP) are longitudinal studies aimed at characterizing the clinical and pathological features underlying aging, cognitive decline, and neurodegenerative disease. ROS began in 1994 and enrolls nuns, priests, and religious brethren from across the United States. MAP began in 1997 and enrolls individuals from across northeastern Illinois. As of May 29th, 2020, ROSMAP has enrolled 3646 individuals (72.7% female), 62.4% were diagnosed with mild cognitive impairment, 29.4% were diagnosed with dementia and 53.2% of all enrolled individuals have deceased. Although two separate studies, clinical follow-up procedures and sample collection are standardized and made available for joint analysis. At time of enrollment, both studies conduct a clinical evaluation of each individual and take blood samples. Blood sampling continues annually for all MAP

participants, and several hundred ROS participants. Upon death, tissue is extracted from multiple brain regions, the spinal cord, nerves, and muscle. The brain is also autopsied and quantitively evaluated for neuropathology by board-certified neuropathologists. Clinical and neuropathological characterization of these cohorts has been reported elsewhere (Bennett *et al*., 2018).[127] A summary of all information collected on the ROSMAP cohort is provided in De Jager *et al*., 2018; and is available at www.radc.rush.edu.[128]

The data used in this chapter includes 1081 WGS samples (male (M) = 362, female (F) = 719; Illumina HiSeq X), and 1280 samples genotyped via SNP-array (M = 306, F = 974; Affymetrix Human SNP6.0). DNA used for the ROSMAP WGS was extracted from whole blood and several brain regions. Only tissues with a sample size greater than 50 (i.e. dorsolateral prefrontal cortex (DLPFC), whole blood and cerebellum) were analyzed. Of all WGS samples, 363 (M = 129, F = 234) were sampled from whole blood, 458 (M = 152, F = 305) were sampled from the DLPFC and 258 (M = 78, F = 180) were sampled from cerebellum. Because of our interest in chromosome Y, we focused on male samples and female samples were primarily used for quality control. In total, 40 male samples had paired WGS and SNP-array data from whole blood. Additionally, 58 DLPFC and 18 cerebellum WGS samples had overlapping whole blood array data. A visual summary of the data used in the project is provided in **Figure 2.1**.

2.2.2 WGS processing

WGS sequencing and raw sequence file quality control reported below was completed and reported by De Jager *et al.* (2018).[128] Details regarding WGS preparation and sequencing have been included in **Appendix 1**. **Figure 2.2** summarizes the methods used in Chapter 2. The mean depth across all samples was 39.3 (range: 29.0-64.1; **Appendix 2.1**), mean depth across the MSY was 19.5 (range: 7.9-29.7). Between each tissue, genome-wide sequencing depth was not significantly different (**Appendix 2.2**).

All read depth collecting from WGS BAM files was completed by Badri Vardarajan. In the autosomes, read depth values were collected from high mappability regions for individuals of European ancestry specified in

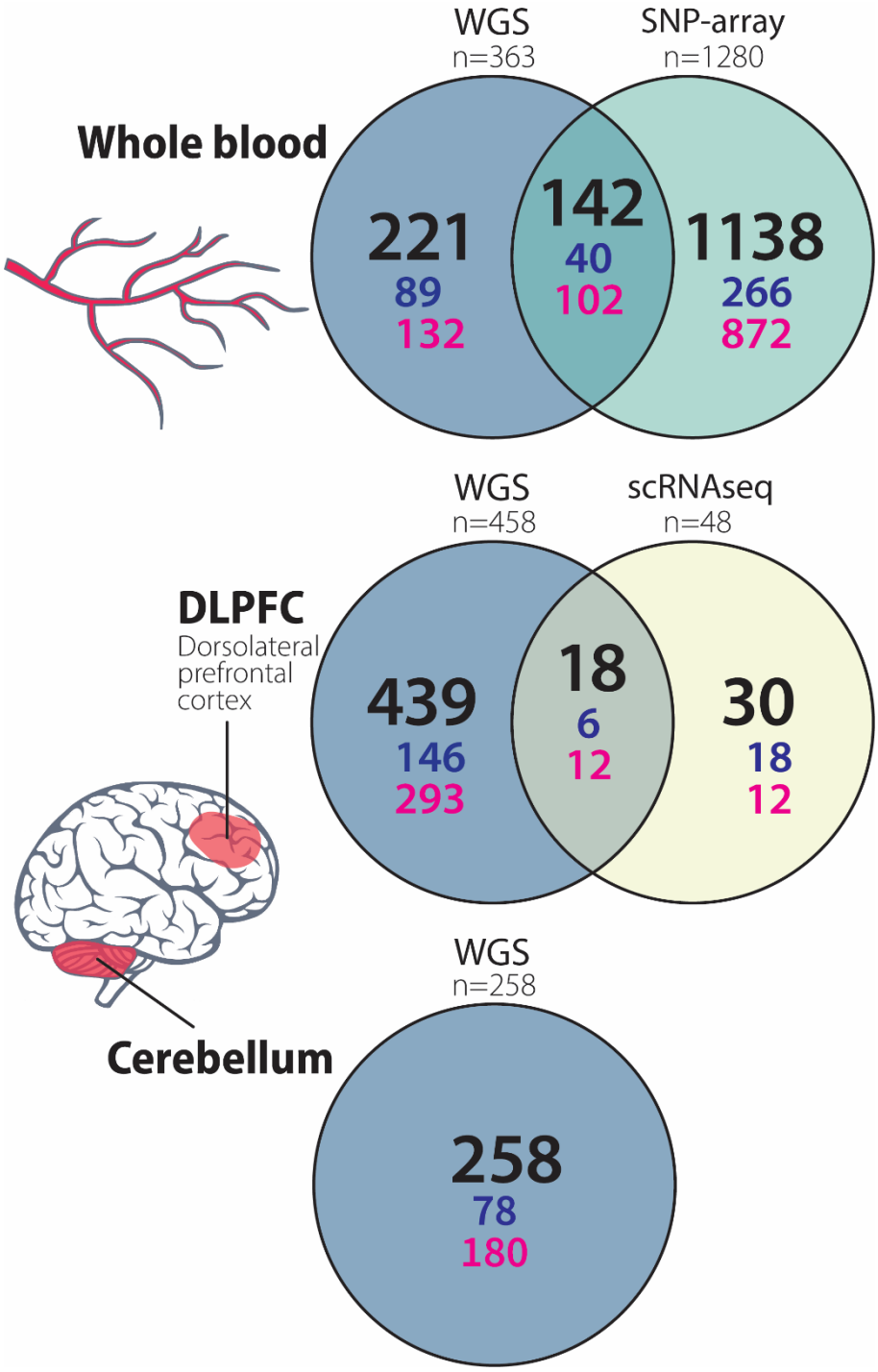


Figure 2.1 Overview of the ROSMAP cohort data used for LOY quantification using SNP-array and WGS technologies. Whole blood samples from males and females were analyzed through WGS (males = 132) and Affymetrix Genotype Array 6.0 (males = 306). Paired data from both WGS and SNP array was available for 40 male samples from blood. Samples from the dorsolateral prefrontal cortex (males = 155, females = 305) and cerebellum (males = 78, females = 180) were analyzed through WGS. Single-nuclei RNA sequencing (snRNA-seq) was performed on DLPFC samples from 24 males and 24 females (detailed in Chapter 3). Female samples were used primarily for quality control.

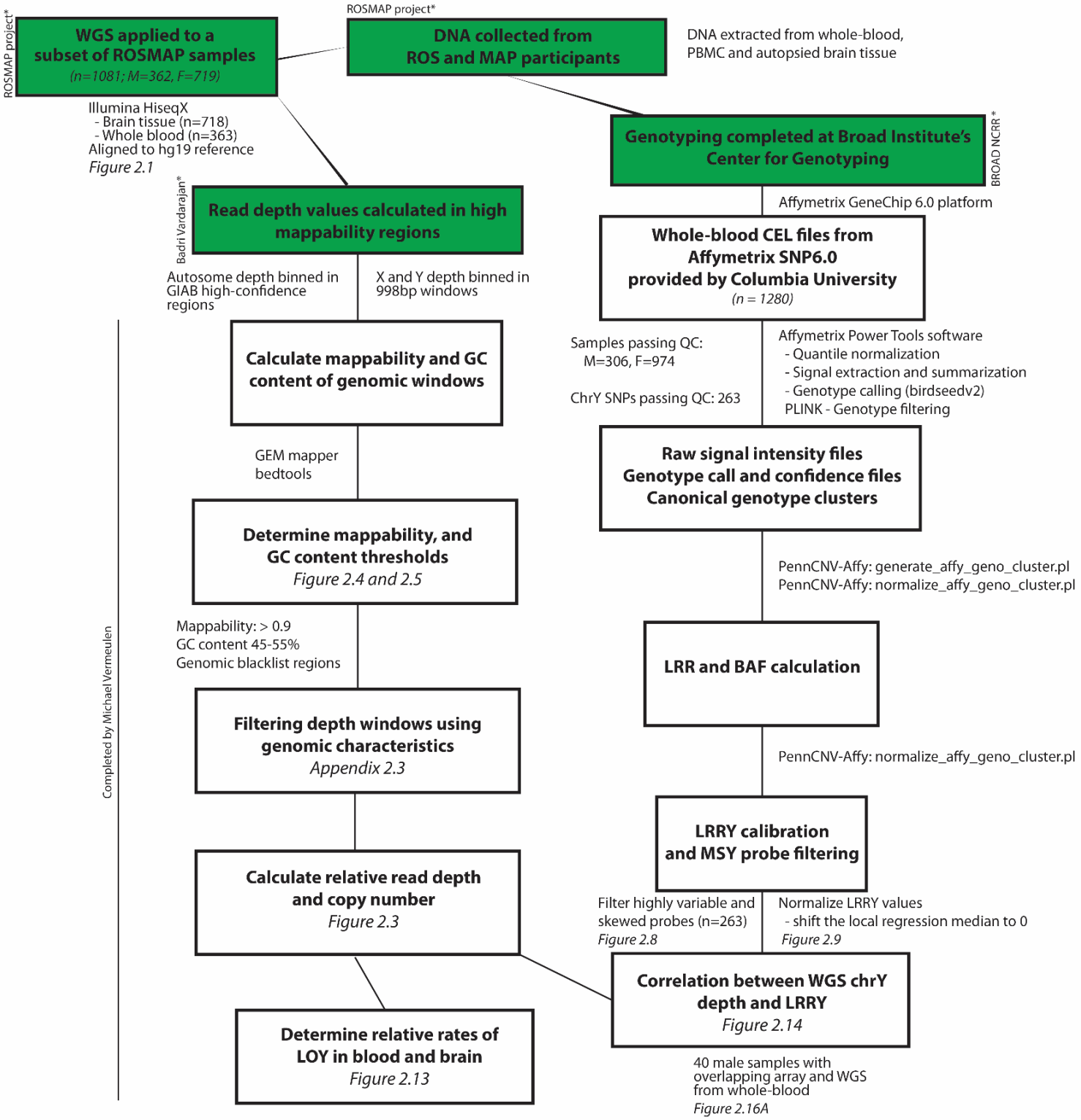


Figure 2.2 Overview of data and methods used for SNP-array and WGS LOY quantification. Flow-chart describing all steps in analysis of 1081 WGS samples (Illumina HiSeq X) and 1280 genotyped samples (Affymetrix SNP6.0 array). Green boxes highlight work that was completed by others. Figures have been referenced where relevant.

the high mappability BED file from the Genome in a Bottle Consortium (GIAB; NA12878, version NISTv3.3.2, build hg37). GIAB high mappability regions are segments of the genome where variants can consistently be mapped with high confidence. These regions have been benchmarked using extensive deep sequencing using multiple platforms and represent regions of the genome where depth measurements are reliable. GATK DepthOfCoverage was used to compute read depth in each genomic window and included reads passing minimum mapping quality (>20) and normalized base quality (> 0) filters. To reduce the effect of PCR artefacts and to conserve computational resources, regions with excess coverage were downsampled to 5000. GIAB high mappability regions are of variable length (range: 1-102,317bp). Window length filters were applied to GIAB high mappability regions, limiting the analysis to regions of length >= 998, as small windows introduce unwanted variability. In the GIAB high mappability file, chromosome Y is not included. During initial analyses we used the average read depth across the entire Y chromosome, effectively treating the Y as a single, large, 59MB genomic window. However, given our primary interest in the Y, to improve read depth resolution and ability to filter, read depth was collected in uniform 998 bp bins across chromosome X and Y.

2.2.2.1 Mappability and blacklist filter

To further improve our ability to estimate mosaic whole chromosome copy number I filtered genomic regions on mappability score (S; also referred to as alignability). The Center for Genomic Regulation (CRG) alignability track provides scores on how uniquely k-mer sequences align to regions in the genome. For example, in the case of the 50mer CRG track, a sliding window of 50bp is applied to the reference genome and each 50bp kmer is mapped back to the genome using the The Genome Multitool (GEM) mapper aligner. Up to 2 mismatches are permitted. For each 50bp window, a mappability score is produced (S = 1 / number of matches to reference genome). Therefore S = 1 represents a unique 50bp match across the entire genome (with two mismatches permitted), whereas 0.5 represents two matches genome-wide and so on. Since the window slides to each base in the reference genome, each base pair is given a mappability score (S). Using the S of each base, the average mappability of genomic regions can be computed. I used a mappability threshold with the goal of eliminating low-

confidence repetitive regions and regions of high sequence homology that could contribute to technical noise and variability when detecting mosaic aneuploidy. This is particularly useful when investigating the sex chromosomes, given their shared sequence homology. Additionally, windows overlapping Duke Excludable Regions and DAC Blacklisted Regions were removed. These files contain regions that have known high multi-mapping to unique mapping ratios and high rates of signal artefacts across multiple cell lines and experiments. Many of the excluded regions do not overlap with genome-wide mappability filters, and ENCODE recommends the use of blacklisted regions, alongside mappability for genome analysis.[107]

2.2.2.2 GC content filter

Genomic regions were also filtered based on GC content in an effort to reduce GC content bias and smooth the read depth signal. GC-rich and GC-poor reads tend to be underrepresented by Illumina sequencing, which can cause unwanted read depth variability. Ideally, GC content bias is accounted for at the sequence read level.[106] However, given the large amount of WGS data used for the project, its storage on external collaborator servers, and the extensive computational resources required to perform this correction, we decided to threshold GC content by genomic region. I input the hg19 reference genome into the *nuc* tool from the bedtools package[129] to calculate the % of guanine and cytosine in each genomic window. For each genomic window, the average GC content was used in tandem with mappability score and blacklist membership to filter poor quality regions, enriching for representative regions of the genome.

2.2.2.3 Relative read depth and estimated copy number

To estimate the mosaic loss or gain of each chromosome in each sample we used a metric we call relative read depth (rRD). Relative read depth values for each chromosome in each individual were computed as the ratio between the mean read depth of all passing genomic windows across the chromosome and the mean read depth across all passing genomic windows genome-wide. Simply, the average read depth across each chromosome is

compared to the average read depth across the genome. Given the non-uniform distribution of bins in the autosomes, a weighted mean based on bin length and median window read depth was used when calculating chromosomal and whole genome depth. In order to prevent auto-normalization, the chromosome undergoing normalization was excluded from the median whole genome read depth calculation. rRD was converted to estimated copy number (CN), by multiplying rRD by 2. This normalizes read depth values to their expected biological ploidy (**Figure 2.3**).

2.2.2.4 Finding GC content and mappability thresholds

The goal of thresholding WGS depth windows on GC content and mappability was to reduce unwanted technical variation and improve correlation with a standardized, orthogonal data source (i.e. SNP-array). To find the optimal combination of thresholds I used two main measures: i) Median absolute deviation (MAD) of estimated copy number across the autosomes in all samples (**Figure 2.4**) and, ii) the correlation between WGS (estimated copy number) and SNP-array (mLRRY) in samples with overlapping data (**Figure 2.5**).

When we initially analyzed relative read depth data across the autosomes without filters, we observed high variation and significant deviation from expected values (**Figure 2.3A**). Although we saw many samples with reduced Y chromosome copy number, several autosomes showed similar patterns of high variability making biological conclusions difficult. We hypothesized that a majority of the observed autosomal variation was technical and could be removed through the selection of high confidence genomic windows that passed GC and mappability filters. To test this, I applied a combination of GC filters and mappability filters to all samples and calculated the MAD of the estimated copy number of each chromosome in each sample (**Figure 2.4**). Using mappability filters from 0 to 1 in 0.1 increments, and GC content filters of 0-1, 0.1-0.9, 0.2-0.8, 0.3-0.7, 0.3-0.6, 0.35-6, 0.35-0.55, 0.4-0.55, 0.45-0.55, it was clear that autosomal CN variability was minimized at GC filters 0.40-0.55 and 0.45-0.55 and mappability filters 0.8 and 0.9. Increasingly strict filters reduced variability until too much information was lost and copy number variability was increased. When I applied the 0.45-0.55 GC/0.9

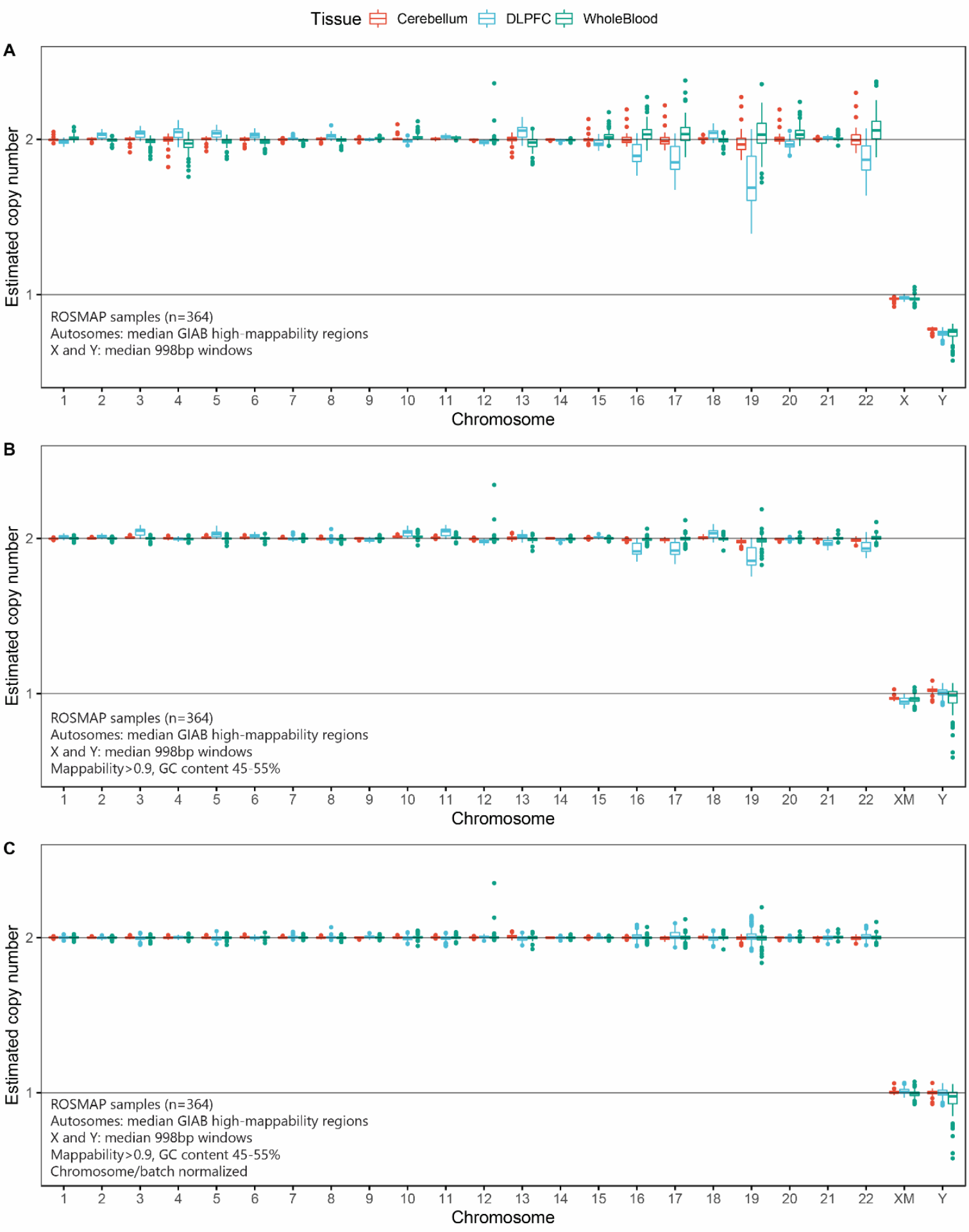

Tissue
Cerebellum
DLPFC
WholeBlood
A
Estimated copy number
2
1
ROSMAP samples (n=364)
Autosomes: median GIAB high-mappability regions
X and Y: median 998bp windows
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
Chromosome
B
Estimated copy number
2
1
ROSMAP samples (n=364)
Autosomes: median GIAB high-mappability regions
X and Y: median 998bp windows
Mappability>0.9, GC content 45-55%
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 XM Y
Chromosome
C
Estimated copy number
2
1
ROSMAP samples (n=364)
Autosomes: median GIAB high-mappability regions
X and Y: median 998bp windows
Mappability>0.9, GC content 45-55%
Chromosome/batch normalized
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 XM Y
Chromosome

Figure 2.3 Estimation of chromosomal content in three tissues across 364 elderly males using WGS. Estimated copy number (CN) is calculated for each chromosome in each sample by dividing chromosomal depth by total genomic depth (**Methods** – Chapter 2) and multiplying by ploidy. Autosomes have an estimated copy number of 2 and the sex chromosomes have an estimated copy number of 1. Each panel represents an additional filtering step within the WGS ploidy detection algorithm which is described in the bottom left corner of each plot. A) Unfiltered ploidy estimation. As expected, the CN for diploid chromosomes is approximately centered at 2, and haploid chromosomes are approximately centered at 1. Without filtering autosomal variance is high which makes biological and statistical conclusions difficult. Additionally, chromosome Y ploidy was significantly below 1, largely because of mapping errors and technical noise (**Figure 2.6**). B) Mappability (>0.9) and GC content filters (0.45-0.55) were added and technical variability was reduced. C) Estimated copy number was further normalized by batch/tissue correcting using the peak of the kernel density function. After correction, the reduction of Y ploidy remains visible.

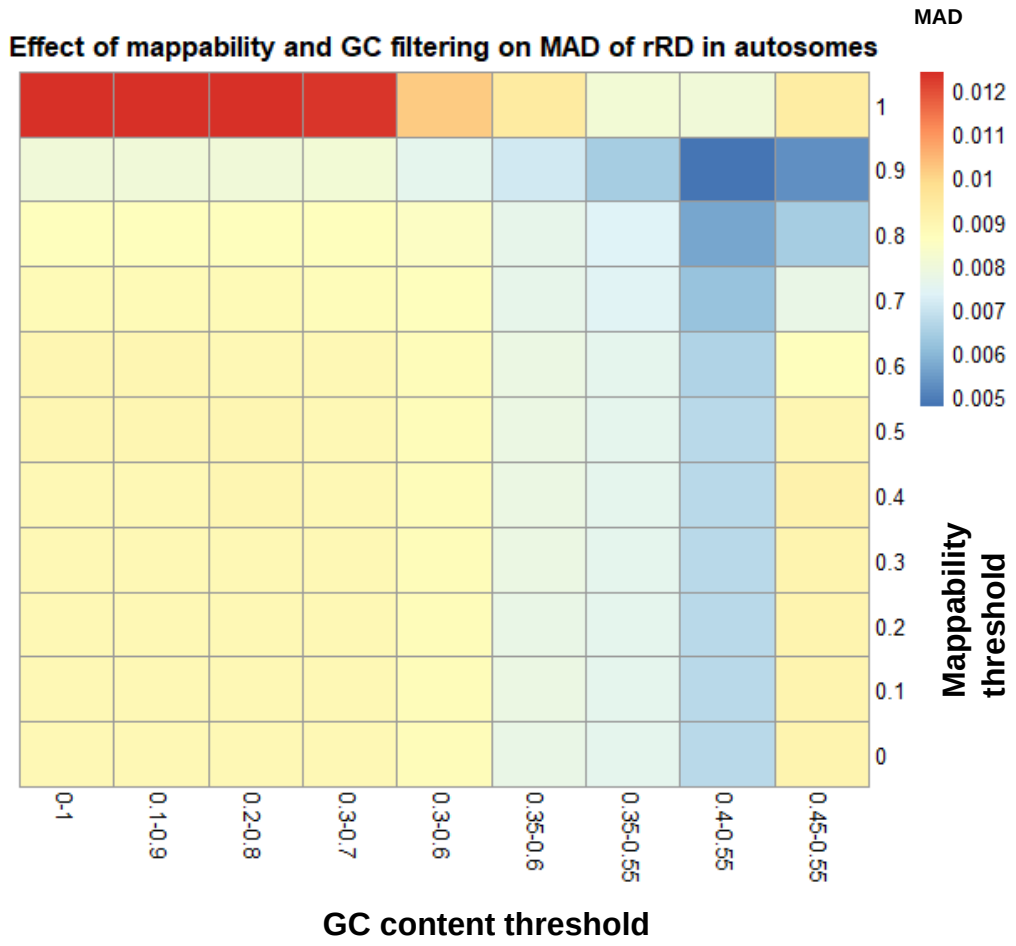


Figure 2.4. Filtering genomic windows using mappability and GC content reduces the variability of copy number estimation across the autosomes. For each combination of mappability and GC content window filters the estimated copy number (CN) was calculated for each chromosome in each sample. For each combination, the median absolute deviation (MAD) was applied to the autosomes. Restricting windows to GC 0.4-0.55 and mappability values greater than 0.9 provided the least overall variability. Reducing copy number variability across the autosomes, provides confidence in chromosome Y ploidy estimates.

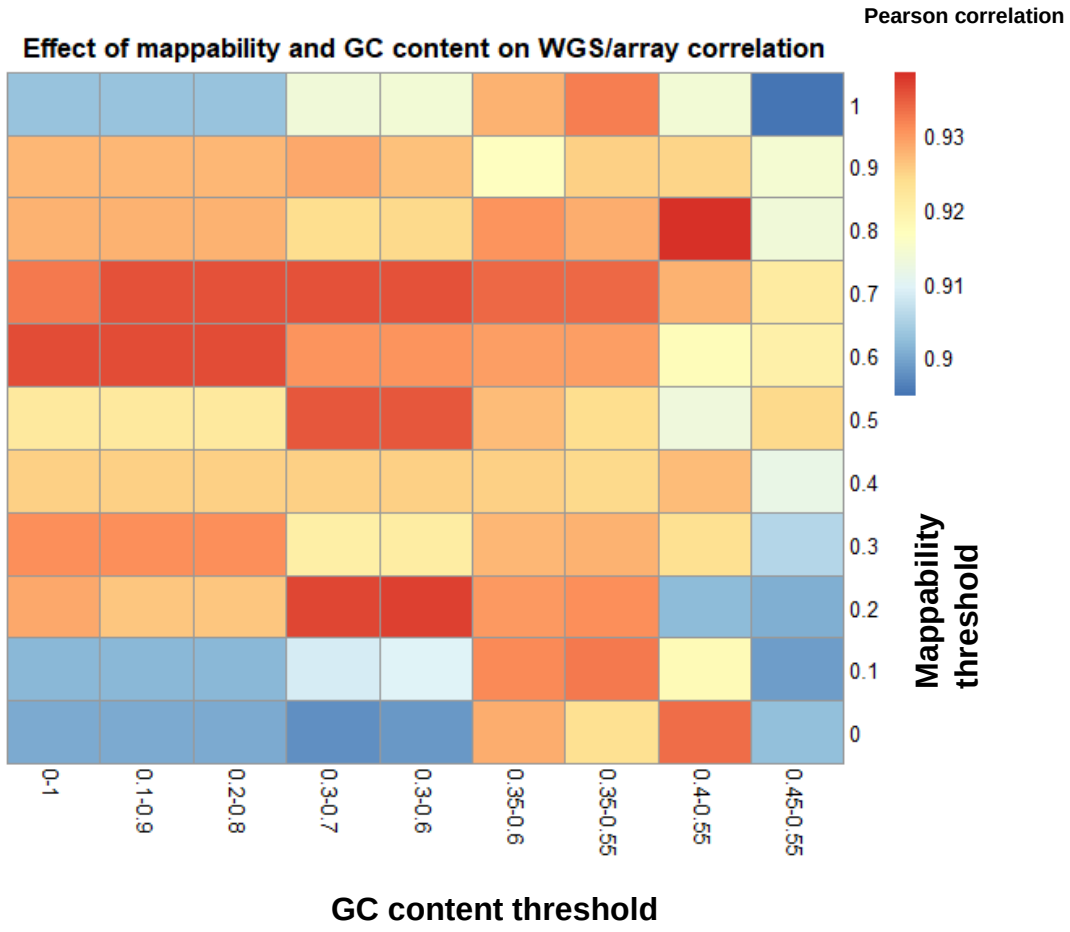


Figure 2.5. Filtering genomic windows using mappability and GC content does not reliably improve correlation between WGS and SNP-array mosaic Y chromosomal copy number estimates. For each combination of GC content and mappability filters, estimated chrY CN from WGS and mLRRY values from SNP-array were tested for correlation (Pearson's). Correlation between the two data sources is high but ranges between 0.90-0.93 in an unpredictable pattern. As data is filtered mores strictly and information is lost, correlations tend to drop. Nevertheless, correlation between array and WGS is high, giving us confidence in the WGS-based copy number estimation methods.

mappability filter (referred to as optimal filter), the autosomal variance was removed and the LOY signal remained, which gave us confidence in the method (**Figure 2.3B**).

Next, the same combination of filters was applied to samples with overlapping array and WGS whole blood data. My hypothesis was that effective filtering enriched the true biological reality and would improve the correlation between the two independent technologies. Given that SNP-array LOY detection is largely standardized, filter combinations were applied to WGS samples. The most effective improvement in correlation between WGS and SNP-array LOY came before filtering when array copy number probes were removed from the analysis (R=0.84 to R=0.89). After this, correlation between technologies was high using all filters (range: 0.89-0.93), but GC and mappability filters did not increase correlation in a reliable pattern. The 0.40-0.55 GC/0.9 mappability filter did display the highest correlation (R=0.935). Ultimately, I chose the 0.45-0.55 GC/0.9 mappability filters was as they resulted in the least estimated ploidy variation amongst autosomes and a high correlation with LOY estimates from array data. When applying the optimal filter, 593 chrY windows (2.6%), and 35,732 (7.2%) autosomal windows remained. These genomic windows were used to calculate rRD (**Appendix 2.3**). When investigating Y depth in individual samples I commonly used the 0.40-0.55 GC/0.9 mappability filter as it provides 1784 windows, as opposed to 593 windows. This increased coverage allows for improved detection of sequencing abnormalities, and large structural variants when visually inspecting.

2.2.2.5 Visual inspection and quality control of Y chromosome WGS data

Additionally, depth across chrY was investigated manually in each sample. For each male sample, I produced panels consisting of increasingly stringent filtering (**Figure 2.6A**). In each case I looked for abnormalities, large structural variations, and poor-quality depth data. When viewing these panels, the importance of mappability filtering was apparent. Consistent, lowly mapped regions in the MSY, XTR regions, and highly mapped regions near the centromeres are removed by the 0.9 mappability filter, allowing for a more accurate measure of depth (**Figure 2.6B**). After manual curation, 4 samples were removed because of highly variable,

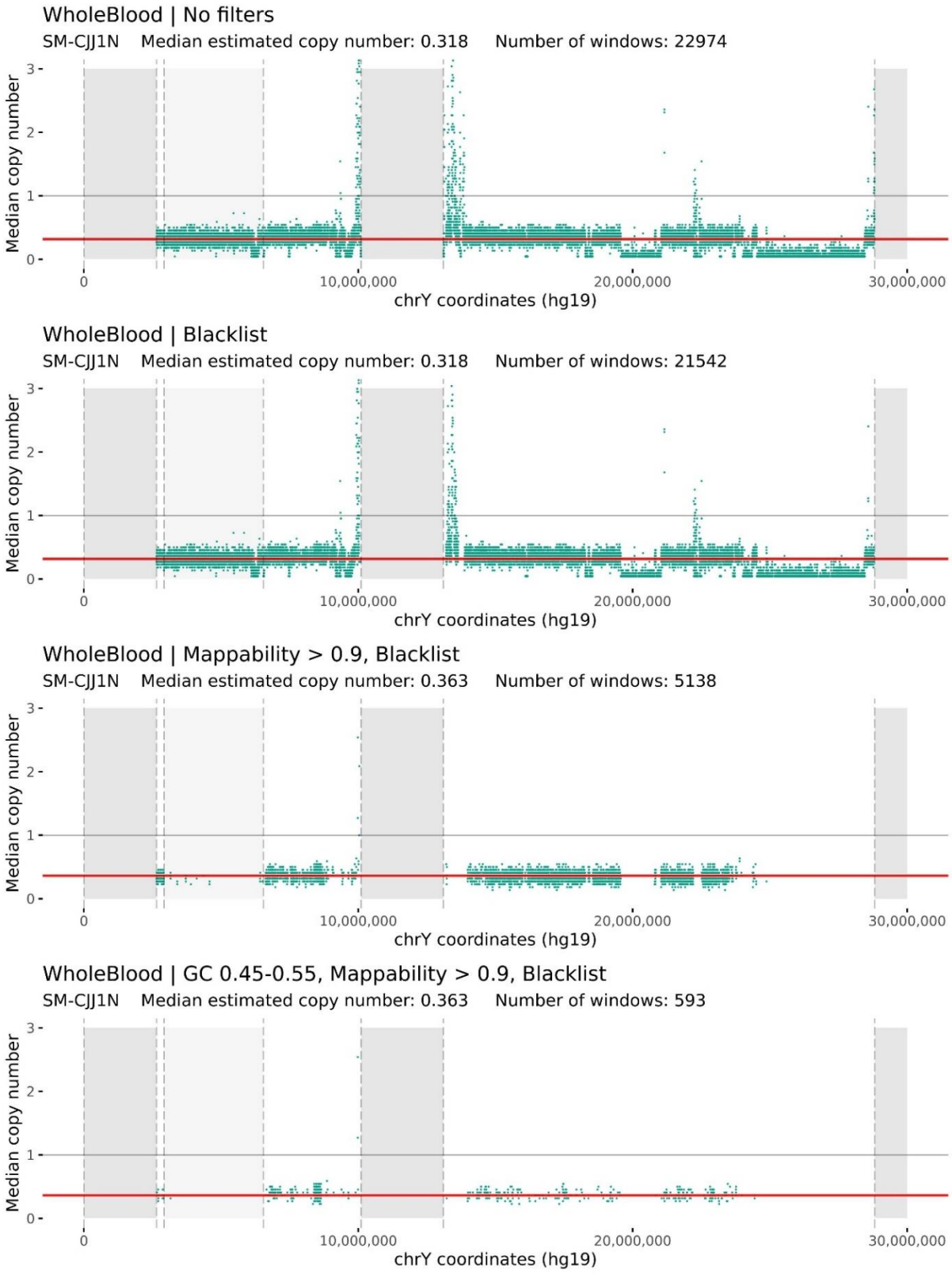


Figure 2.6A. WGS genomic window filtering steps in LOY blood sample. A visualization of the genomic window filtering process. Each panel displays read depth across chromosome Y at increasingly strict filtering steps for a single sample (whole blood) categorized as LOY (CN = 0.363). The gray windows represent notable genetic regions on Y chromosome. Beginning on the left, these include PAR1, XTR, the centromere, and the heterochromatin region which continues for ~28 million bases. The red line represents the median depth across all bins passing filters. A-B) Raw data and blacklist filtered data. High variability especially near the centromere and on the q arm. C) The mappability filter is effective at removing low quality information. D) The GC content filter removes a significant proportion of the data but smooths for GC bias and reduces technical variability. This process was repeated for each sample.

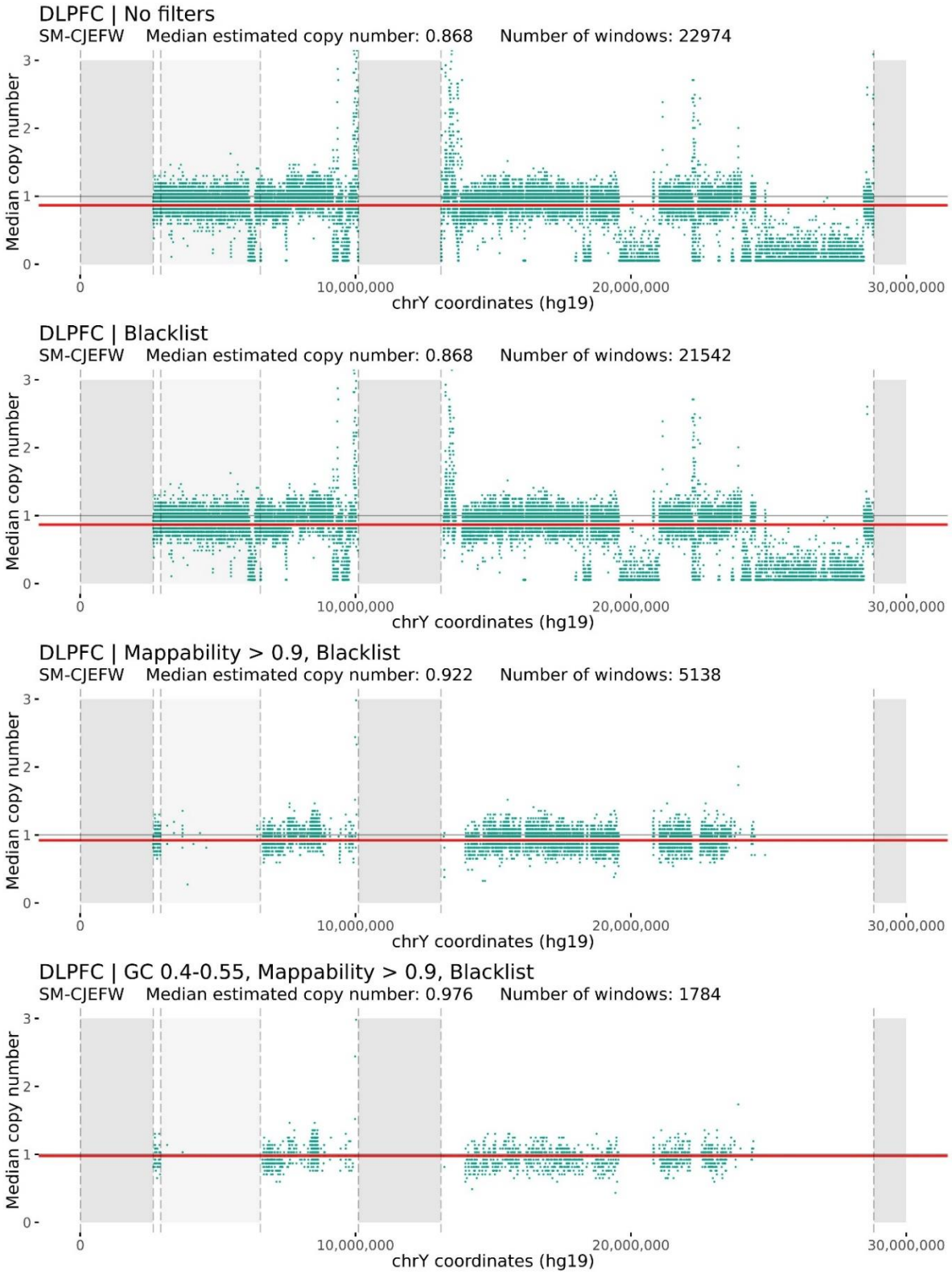


Figure 2.6B. WGS genomic window filtering steps in normal ploidy dorsolateral prefrontal cortex (DLPFC) sample. Example of the filtering process on a normal ploidy sample. Bottom, the 0.4-0.55 GC content filter was used to visually inspect for large structural variation as 1784 windows are used instead of 593 which improves resolution.

poor-quality data across all chromosomes (SM-CJEHE, SM-CJEJ6, SM-CTEE3, SM-CTEE5, **Appendix 2.4**). Furthermore, 6 chrY duplications were discovered, however in each case when the duplication was masked, the estimated chromosomal CN was not significantly affected.

2.2.2.6 Ploidy correction using peak of kernel density estimate

I further normalized CN values in each tissue/chromosome to adjust for tissue specific technical variation and sequencing depth biases. In a method similar to that applied to array mLRRY values (See **SNP-array processing**), estimated CN values were shifted by the difference between the peak of the local regression median and expected chromosomal ploidy. Using the density function in R, I applied a kernel density estimate (using the Shether-Jones bandwidth) independently to the CN values from each chromosome/extraction kit combination (**Figure 2.7; Appendix 2.5**). In each case CN values were shifted by the difference between the peak of the density distribution and expected ploidy (2 for autosomes, 1 for sex chromosomes in males). Copy number peak density normalization reduced variation amongst the autosomes (MAD before = 0.0106, MAD after = 0.0065).

2.2.3 SNP-array processing

All sample preparation and genotyping was completed by the ROSMAP consortium. DNA from ROS and MAP subjects was extracted from whole-blood or lymphocytes and was genotyped using the Affymetrix Genome-Wide HumanSNP Array6.0 platform (Santa Clara, CA) at the Broad Institute's Center for Genotyping (n=1280). All raw genotyping data (Affymetrix probe results files; CEL) were provided by download from Columbia University (Badri Vardarajan). Using provided CEL files, I used Affymetrix Power Tools bundle (v1.20.0) and PennCNV-Affy package to normalize, extract probe signals, generate cluster files, call genotypes and ultimately calculate Log R Ratio (LRR) values (**Figure 2.2**). Specifically, I used the Birdseedv2 algorithm to call genotypes, and used the normalize_affy_geno_cluster.pl program (PennCNV) to generate LRR values. Samples with contrast quality control values < 0.40, median absolute pairwise difference values > 0.35, call rate < 0.95, or

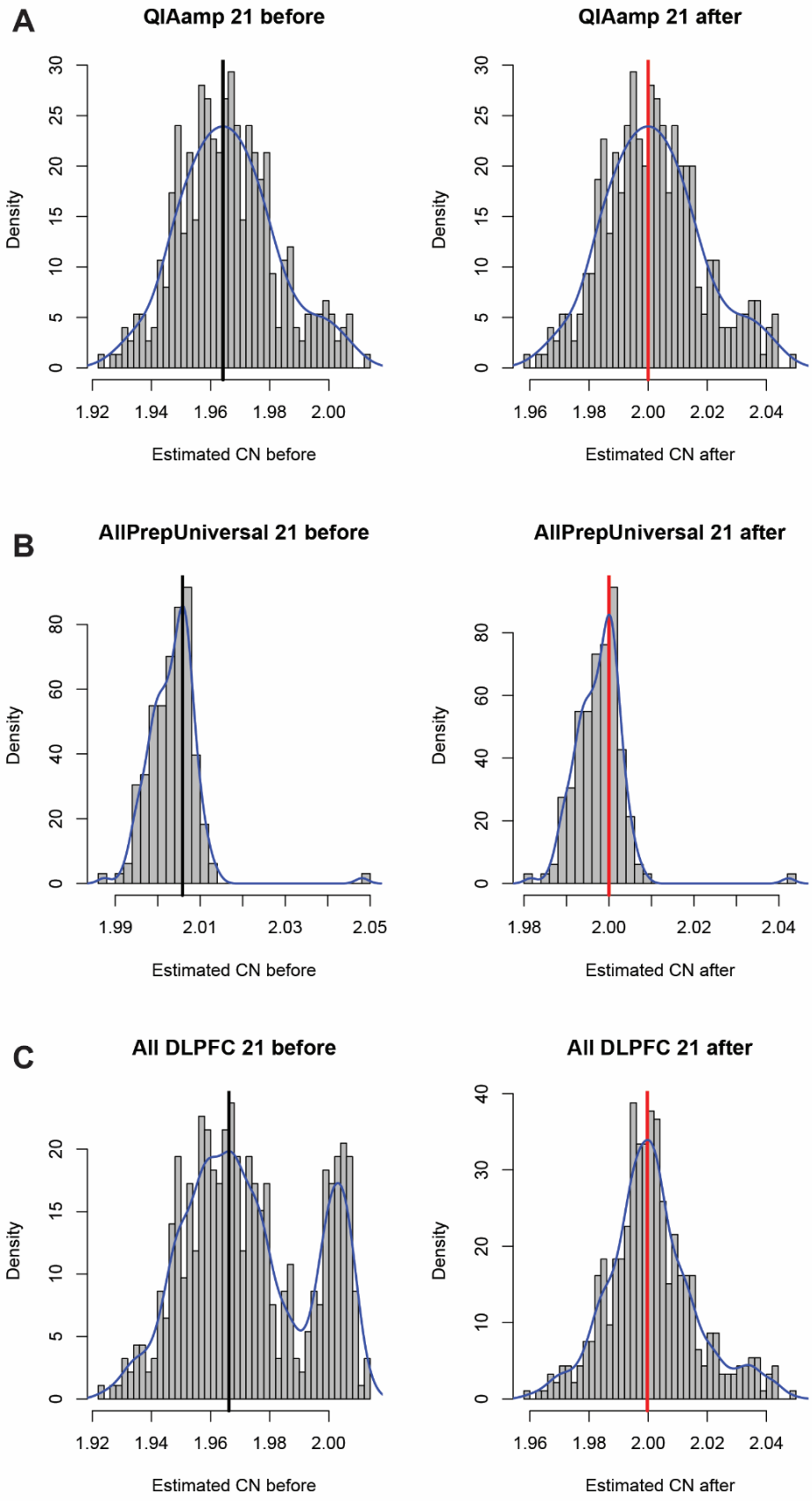


Figure 2.7. Example of chr21 density kernel ploidy correction in dorsolateral prefrontal cortex samples. Density kernel ploidy correction was completed separately for each chromosome/extraction kit combination to normalize depth distributions to the expected ploidy of a chromosome. In this example chr21 is being corrected in a dorsolateral prefrontal cortex (DLPFC) sample. A) First the QIAamp batch is processed. The peak of the density kernel is shifted to the expected ploidy. B) The process is repeated for the AllPrepUniversal batch which has a different distribution. C) When the data is merged, copy number is centered around 2, however the distribution of the data is maintained. The distribution of CN values between kits is normalized.

conflicting sex determination were removed, as per manufacturer instruction. Birdseedv2 files were converted to PLINK format (http://zzz.bwh.harvard.edu/plink/) for SNP QC. SNP probes with genotype call rates < 95%, minor allele frequency < 0.01, mishap test < 1e−9, and Hardy-Weinberg P < 0.001 were removed. After quality control, a total of 306 male and 974 female samples, 757,091 SNP probes and 908,043 copy-number (CN) probes were included for further analysis. In total from chromosome Y this included 8582 CN probes and 271 SNP probes.

Log R Ratio (LRR) in a given chromosome or region was used as a marker of chromosomal content. The Affymetrix Power Tools bundle outputs the LRR as $\log2(R_{observed}/R_{reference})$, where $R_{reference}$ is the probe intensity in a reference population of HapMap individuals. On the Y chromosome, 8797 probes (Affymetrix Genotyping Array 6.0) are located in the male-specific Y region (MSY), and outside the pseudoautosomal regions, PAR1 (Y:10,001–2,649,520; hg19) and PAR2 (Y:59,034,050–59,363,566; hg19). Each SNP is represented by a pair of 2 probe sequences, one for the A and B alleles. Each pair of SNP probes is replicated at least 3 times on the array. CN probes are non-polymorphic sequences represented by a single probe sequence. As a result, CN probe intensity values are more variable (**Appendix 2.6**) and have a different distribution than the SNP probes. Further, previous studies have standardized the use of SNP probes to set mosaic loss thresholds. For these reasons, I removed CN probes and exclusively used SNP probes to determine Y loss. I further filtered SNP probes on chromosome Y using summarized mLRR values across the cohort. The goal was to remove low-confidence probes that showed significant variability and/or consistent outlier means across the cohort. Probes with cohort-wide LRR standard deviation (sd) outside the 99th percentile (>0.806) and/or LRR mean outside the 99th percentile (>0.07) or 1st percentile (<-0.302) were removed (**Figure 2.8**). These filters remove 7 probes, bringing the final Y chromosome SNP probe count to 264 (**Appendix 4**). The median LRR value from these 264 Y chromosome probes is referred to as the mLRRY and is the primary LOY metric derived from SNP-array data.

Previous LOY studies have found that an uncorrected mLRRY value distribution can shift in the positive direction as an effect of the Affymetrix normalization process and its lack of compensation for unexpected, frequent aneuploidy. To correct the normalization bias, mLRR-Y values were corrected as in Forsberg *et al.*

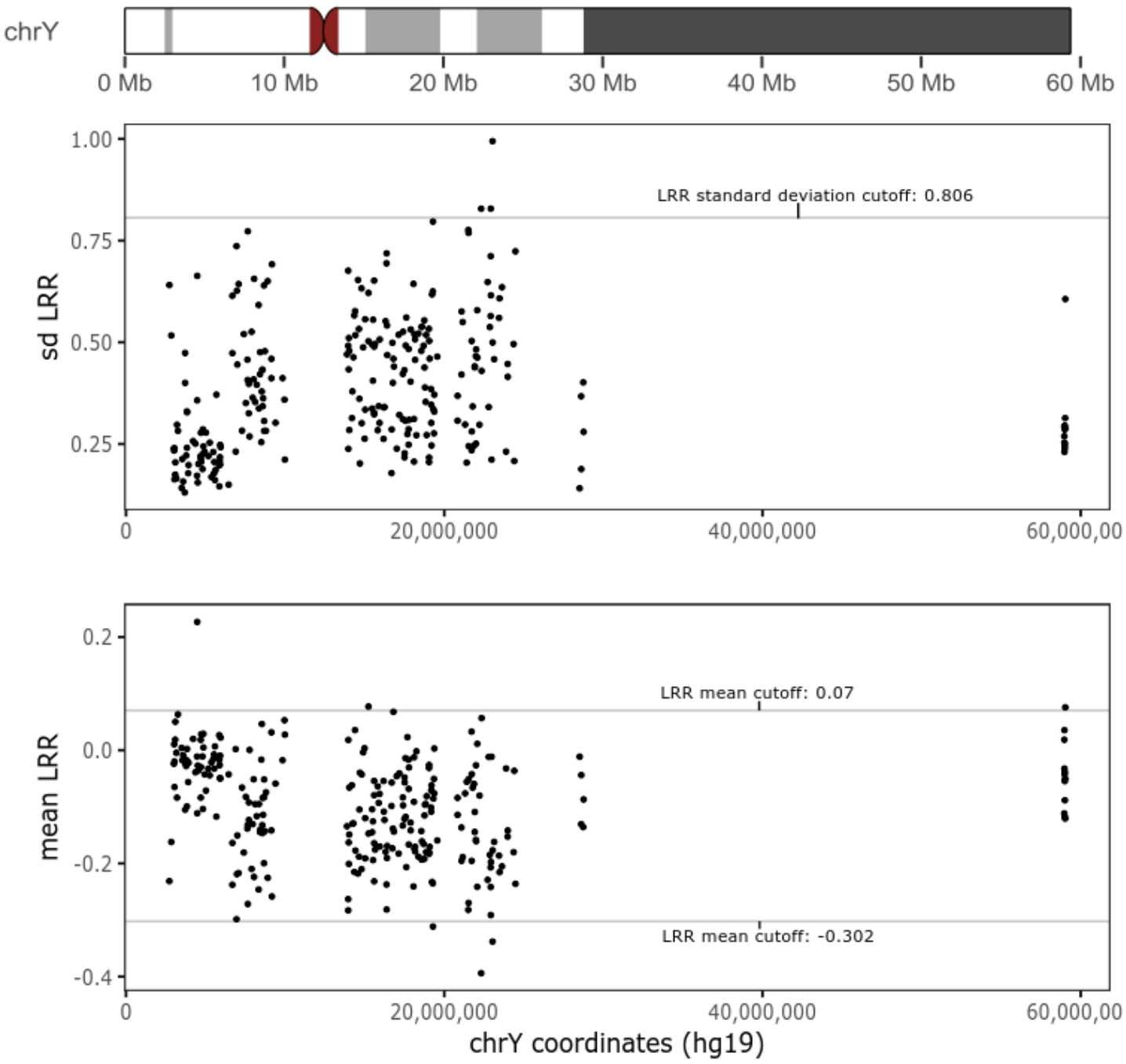


Figure 2.8. Location and summarized log R ratio characteristics of Affymetrix SNP6.0 male-specific Y probes
SNP-probes in the male-specific chrY region (MSY) were filtered using summarized mLRR values. The x-axis represents chrY coordinates (hg19), and each point represents a SNP probe. The goal was to remove low-confidence probes that showed significant LRR variability or outlier means. Probes with cohort-wide LRR standard deviation (sd; top panel) outside the 99th percentile (>0.806) and/or LRR mean outside the 99th percentile (>0.07) or 1st percentile (<-0.302; bottom panel) were removed. These filters removed 7 probes, bringing the final Y chromosome SNP probe count to 264.

(2019).[130] A local regression median was calculated using the density function in R, using the Sheather and Jones (SJ) kernel density estimator. The blue density lines in **Figure 2.9A and B** show the local regression median over the cohort chrY CN distribution while the red line in **Figure 2.9A** denotes the peak of the local regression median (LRRY= 0.025186). This value was subtracted from all mLRRY values, shifting the local regression median to 0. Because gain of Y mosaicism is rarely observed, previous LOY studies have used the normally distributed positive tail of the mLRRY distribution as an estimate of the experimental/technical error distribution. (**Figure 2.9C**). The positive tail of this distribution is assumed to occur in the negative direction as well, so the positive tail is mirrored to create a complete distribution of experimental error. Values outside of the 99% confidence interval of this distribution are considered mosaic gain or loss. Simulations and benchmarking using known mosaic rates suggest this threshold represents the lowest confidently detectable level of mosaicism, which corresponds to ~10% of cells with LOY.[57] The lower limit (i.e. -0.082) was used as the minimum threshold when declaring low-level LOY. The dotted lines in **Figure 2.9B and C** show the limits of the 99% confidence interval of the experimental error distribution.

2.2.4 Transformation of mLRRY values and LOY % estimation

To compare LOY more accurately between genotype array and WGS platforms, mLRRY values were transformed into a metric called rounded SNP-array ratio. Because of the high correlation between WGS copy number and mLRRY, formulas can be applied to transform mLRRY values onto a normalized scale. Following the formulas from Danielsson *et al.* (2019), mLRRY values were transformed.[131] Extensive testing of samples with paired WGS, ddPCR and SNP-array data found that mLRRY values can be transformed to a CN equivalent using the following formula: $Y = (2^{mLRRY})^2$, where Y is the rounded SNP-array ratio.[131] Further, % of LOY cells can be inferred through the following formula: $LOY(\%) = 100*(1-2^{mLRRY})^2$. The LOY threshold at -0.082 mLRRY is equivalent to 0.889 CN and ~10% LOY. Thresholds of increasing LOY severity were set at -0.15 mLRRY which is equivalent to 0.812 CN and ~19% LOY, and -0.315 mLRRY, equivalent to 0.646 CN and 35%

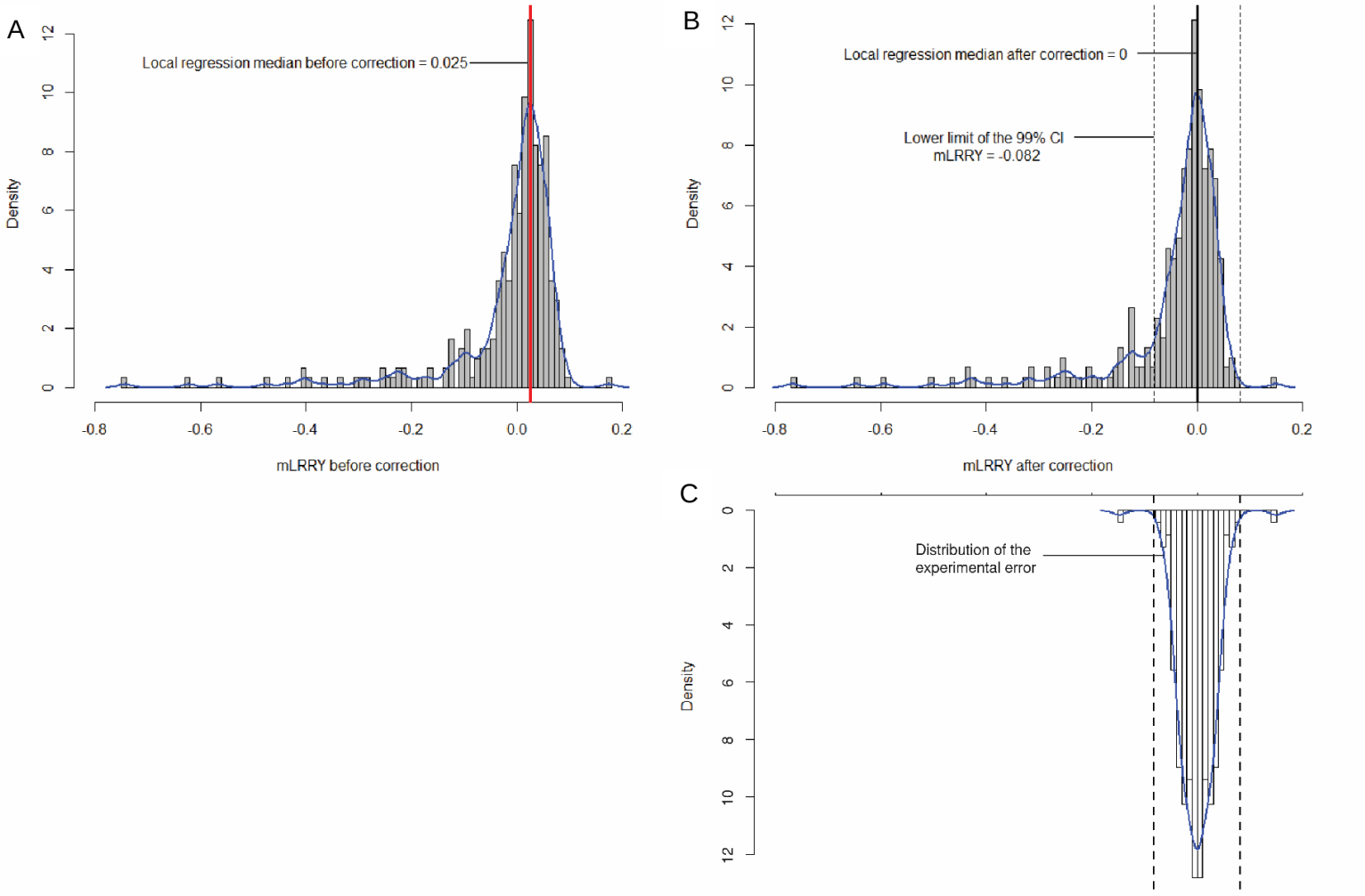


Figure 2.9. Correcting mLRRY values and statistically determining threshold of low-level mosaicism. LOY proportion each sample was estimated using the median log R ratio (mLRRY) across all SNP probes located in the male-specific region of chromosome Y (MSY, Y: 2,694,521–59,034,049, hg19). Samples were genotyped using the Affymetrix™ Genome-Wide Human SNP Array 6.0 which includes 906,600 SNP probes, 271 of which are located within the MSY (264 pass QC). mLRRY values around 0 represent the expected ploidy state. Negative mLRRY values represent a reduction in chrY genomic content and therefore are a proxy for mLOY. The left panel shows uncorrected mLRRY values for all 306 male ROSMAP samples. The Affymetrix normalization algorithm tends to overcorrect mLRR values on the Y as chrY aneuploidy is not adjusted for, hence the distribution is shifted to the right. To correct this bias, data was processed as in Forsberg *et al*. **A-B)** Briefly, the peak of the local regression median was calculated and is represented by the red line. The difference between the peak and 0 was subtracted from all mLRR values, shifting the distribution, and normalizing the values to 0. The dotted lines shows the limits of the 99% confidence interval of the experimental error. **C)** Because gain of Y mosaicism is rarely observed, the distribution of values greater than 0 mLRR is used to estimate experimental error. The lower 99% CI limit (i.e. -0.082) was used as the minimum threshold when determining low-level LOY mosaicism. Simulations and benchmarking by Forsberg *et al.* suggest this low-level mosaicism represents ~10% of cells without chromosome Y.

LOY (**Appendix 2.7**). These thresholds were primarily chosen to compare to previous LOY studies, and the values from our analysis are comparable (**Table 2.1**).

2.2.5 Estimating accuracy of WGS ploidy estimation

In the ROSMAP cohort, 15 individuals were subject to two WGS replicates. In an effort to further establish rates of technical variation in our WGS mosaic ploidy detection method, CN values were compared between these in-sample replicates (**Figure 2.10**). In the autosomes, the absolute mean CN difference between replicates was 0.0173 (n=15, sd =0.021), while for chrY replicates the absolute mean difference was 0.014 (n=5; **Figure 2.10B**). This suggests that technical variation amongst the autosomes is roughly similar to that observed for chromosome Y, and with the exception of chr19 and chrX, CN estimation is highly similar between in sample replicate chromosomes.

2.3 Results

2.3.1 LOY prevalence and associations

Although the main goal of this chapter was not to establish phenotypic and/or disease associations with LOY, I did perform some association tests of known risk factors to provide additional confidence in my methods and data processing pipeline. As expected, mLRRY values were centered around zero and skew in a negative direction (sd = 0.112; **Figure 2.9B**). Negative values represent reduced Y chromosome content in the blood. mLRRY distributions from all other chromosomes did not show negative skew and were centered around 0 (**Appendix 2.8**). Consistent with previous reports, we observed a significant negative correlation between age and mLRRY (R=-0.24, p=1.8x10^{-5}; **Figure 2.11A**) and this effect was specific to chromosome Y (**Figure 2.11B-E**). Age at blood draw was not available, so age of enrollment was used under the assumption that genotyped blood was sampled near initial enrollment. Because of these assumptions the age/LOY correlation is likely underestimated. Unlike previous studies, I was unable to find a relationship between smoking and LOY ("ever":

Study	Cohort	Sample size	MSY probes	mLRRY LOY cutoff	LOY rate	Age range	Genotype array platform
Forsberg *et al.* 2014	PIVUS	488	1690	-0.154	20.5%	70.7-83.6	2.5M Human Omni Express
	ULSAM	1141	2560	-0.08	14.7%	70	2.5M Human Omni
Dumanski *et al.* 2015	PIVUS	488	1690	-0.1182	15.6%	70.7-83.6	2.5M Human Omni Express
	ULSAM	1153	2560	-0.1024	12.6%	70	2.5M Human Omni
	TwinGene	4373	1690	-0.1324	7.5%	48-93	2.5M Human Omni Express
		1323	1690	-0.1324	15.4%	70-93	2.5M Human Omni Express
Loftfield *et al*. 2018	UK Biobank	223,338	691	-0.15	1.7%	37–73	Affymetrix UK BiLEVE and Biobank
		223,338	691	-0.4	0.3%	37–73	
		35,627	691	-0.15	5.2%	65-73	
		35,627	691	-0.4	0.9%	65-73	
Thompson *et al.* 2019	UK Biobank	205,011	691	−0.046	24.0%	40-70	Affymetrix UK BiLEVE and Biobank **used a PAR-LOY method to improve sensitivity*
Dumanski *et al.* 2016	PIVUS	469	1690	-0.1182	21.1%	70	2.5M Human Omni Express
	ULSAM	1138	2560	-0.1024	17.5%	70.7-83.6	2.5M Human Omni
	EADI1	1611	2153	-0.0967	15.4%	65+	Illumina Human610Quad chip
Grassmann et al, 2019	IAMDGC	12504	608	−0.08	16.4%	60-90	Illumina HumanCoreExome
Vermeulen, 2020	**ROSMAP**	**306**	**263**	**-0.082**	**17.0%**	**63-102**	**Affymetrix SNP6.0**
		306	**263**	**-0.15**	**9.4%**	**63-102**	
		306	**263**	**-0.4**	**2.3%**	**63-102**	

Table 2.1 Summary of previous LOY studies using SNP-arrays.

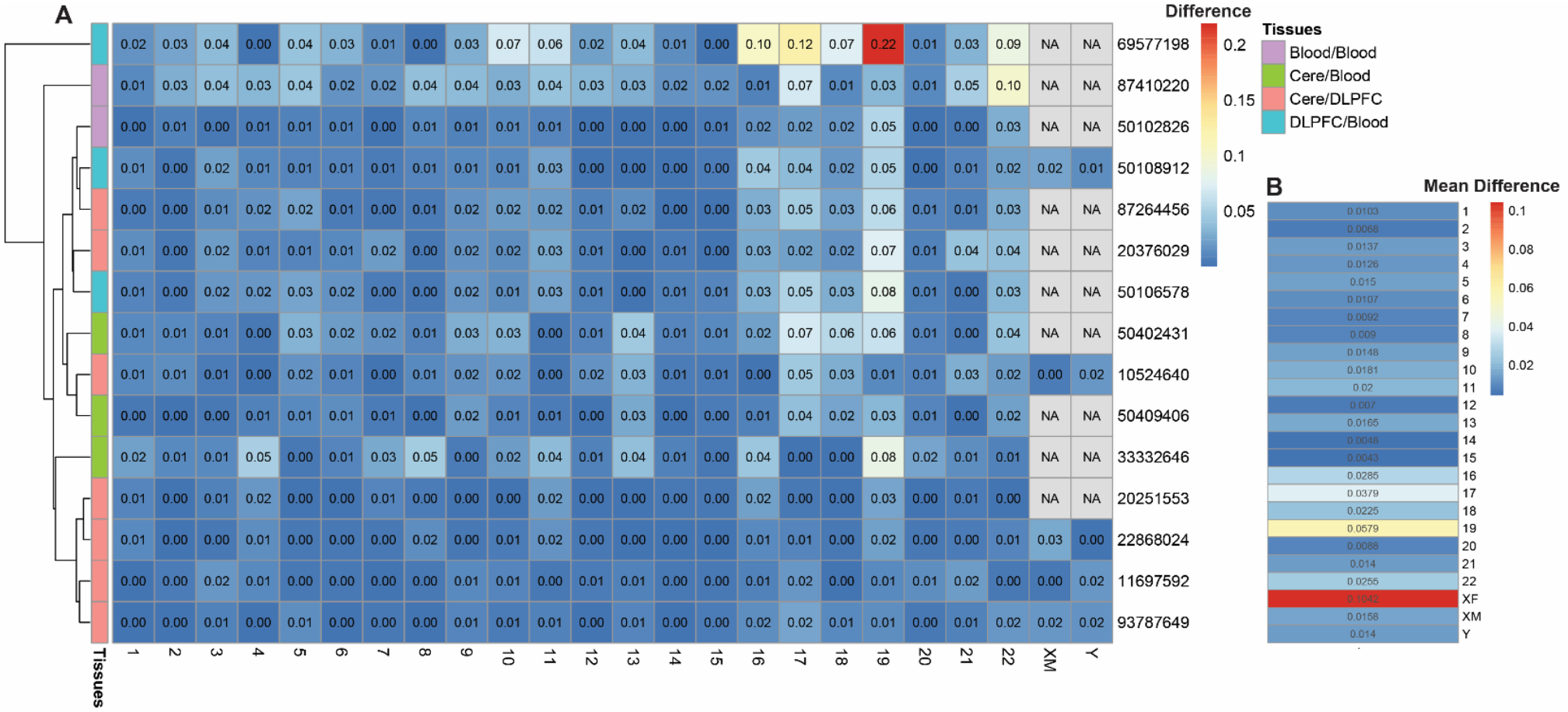


Figure 2.10. Fifteen within sample replicates reveal limited variation in WGS copy number estimation in the autosomes and Y chromosome. 15 samples (5 male, 10 female) were whole genome sequenced twice, which allowed for within sample variance testing of our WGS ploidy detection methods. A) CN was calculated in each chromosome in each sample. The values produced represent the absolute differences between the estimated CN in replicates. Samples are row clustered. B) The absolute mean difference between replicate chromosomes across all 15 samples is summarized. Chromosome Y shows a similar within sample deviation to the autosomes. Chromosome 19 and the female X show elevated variance across most samples. Dorsolateral prefrontal cortex (DLPFC), cerebellum (Cere).

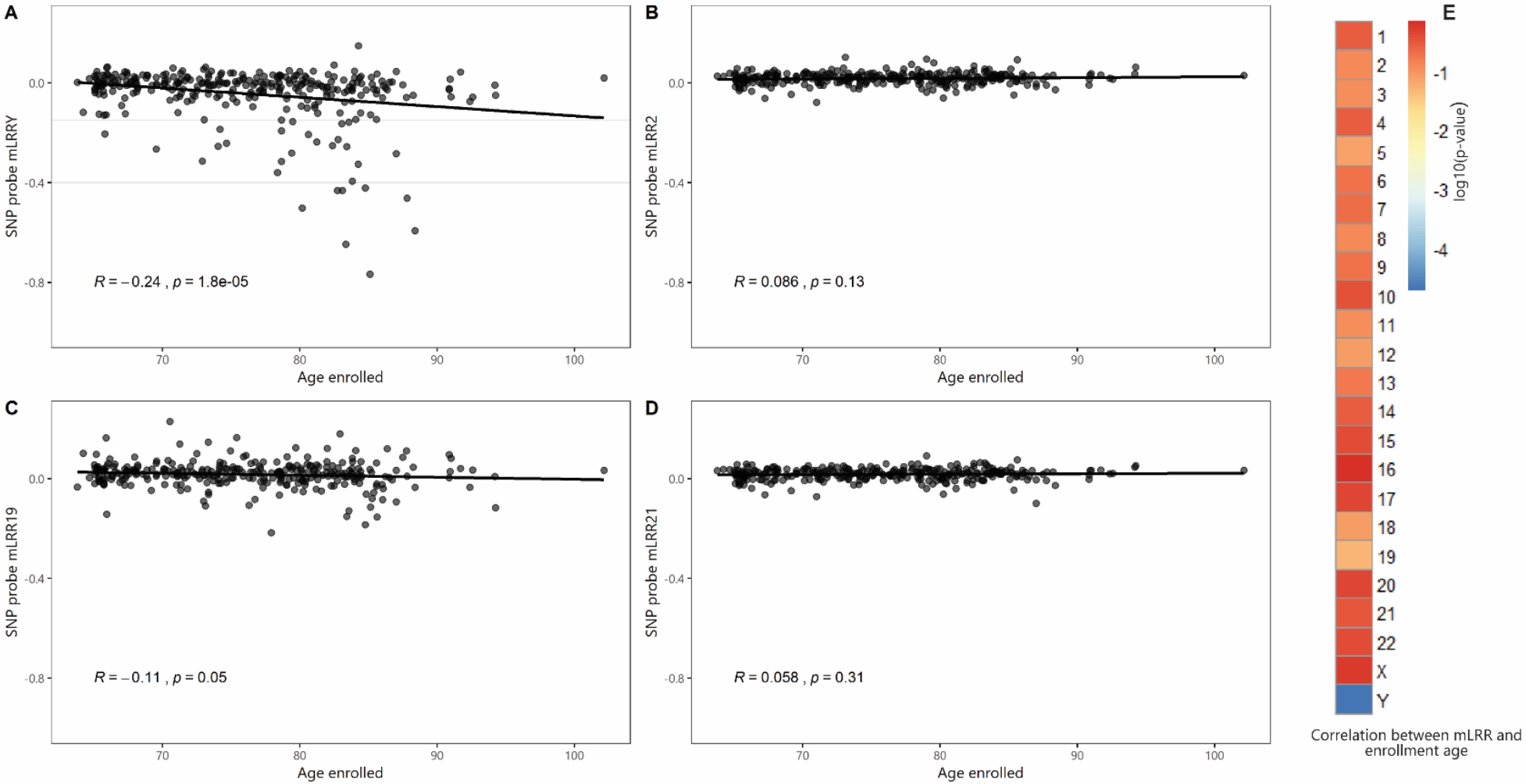
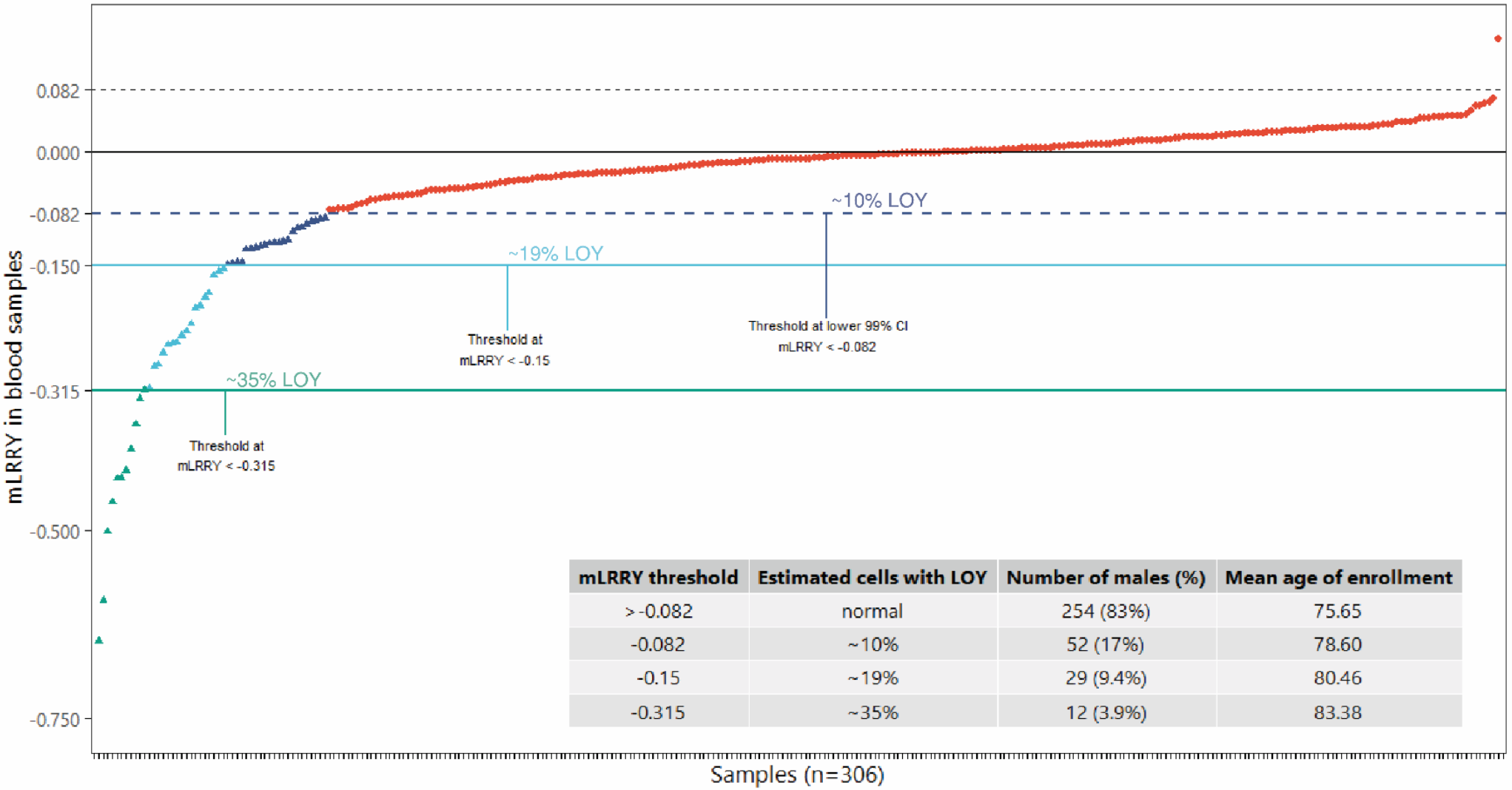


Figure 2.11. Age is correlated with mLRR exclusively on chromosome Y. The mean intensity log-R ratio (mLRR) is a measure of genomic content across a set of probes and is used as a proxy for mosaic aneuploidy. Chromosomes at expected ploidy are represented by an mLRR of around 0. **A)** mLRR values across 264 chromosome Y SNP-probes (mLRRY) are significantly associated with baseline age (p=1.8 x 10^{-5}) in male samples. **B-D)** Associations between age and mLRR values derived from control chromosomes chr2, chr19 and chr21 are not significant. **E)** mLRR-age associations across all other chromosomes are not significant (p>0.05). Age of enrollment was used as the biological age of the sample as specific age at blood draw was not available.

p=0.915, “current”: *p*=0.232). Although this result is unexpected as smoking is one of the main known risk factors for LOY, it can partially be explained by a lag between metadata recording and blood sampling used for genotyping. The correlation between smoking and LOY is transient and dose-dependant and therefore assumptions made on the biological age of samples could be confounding the result. [50]

My main goal in this chapter was to accurately compare and contrast rates of LOY in whole blood and brain tissue in an aging cohort. To do this I estimated the prevalence of LOY across the array and WGS cohorts in addition to the severity of LOY in each sample. For the SNP-array data I used the lower limit of the 99% CI of the experimental error distribution to establish the lowest detectable LOY (from Forsberg *et al.* 2014), and published formulas to convert mLRRY to % LOY cells (Methods; **Appendix 2.7**).[57] In our dataset, an mLRRY value of -0.082 was the lowest detectable LOY threshold and corresponds to ~10% of LOY cells in an individual sample. At this ~10% LOY cut-off, we observed LOY in 52/306 males (17%; mean age = 78.6). Elevated levels of LOY affecting >~19% of cells, were observed in 29/306 samples (9.4%; mean age = 80.5; **Figure 2.12**). These results are comparable to LOY studies using similar arrays and similarly aged cohorts (**Table 2.1).**

WGS estimates of chromosome Y ploidy were distributed around 1 in whole blood, cerebellum and DLPFC (**Figure 2.13C-D**). CN values were normalized to represent biological ploidy and as expected all autosomes were centered around 2, the sex chromosomes centered around 1, and in whole blood chromosome Y showed elevated variance and a skew towards 0 representing LOY (**Figure 2.3**). In blood samples, we observed a significant negative correlation between age of death and chrY CN (R=-0.3, p= 7.4 x 10^{-4}, **Figure 2.14**) but not age at enrollment (R=-0.1, p=0.26). This suggests that blood samples used for whole genome sequencing were likely drawn at a follow up session or during autopsy. Our WGS analysis also failed to find correlations between smoking and LOY (p = 0.767, ANCOVA). In the brain, correlation between age and LOY was brain region specific. LOY was significantly associated with age in the DLPFC (R=0.13, p= 1.13 x 10-4) but not the cerebellum (R=-0.074, p=0.52; **Figure 2.14**). When both DLPFC batches (named by the DNA extraction kit used, i.e. QIAamp and AllPrep) were combined, LOY was significantly correlated with age, however consistent batch



mLRRY threshold	Estimated cells with LOY	Number of males (%)	Mean age of enrollment
> -0.082	normal	254 (83%)	75.65
-0.082	~10%	52 (17%)	78.60
-0.15	~19%	29 (9.4%)	80.46
-0.315	~35%	12 (3.9%)	83.38

Figure 2.12. Summary of mosaic loss of chromosome Y detected via Affymetrix SNP6.0 array for 306 male samples from the ROSMAP cohort. Samples are sorted by increasing median log R ratio on chrY (mLRRY), meaning the samples with the highest LOY rates in blood are plotted farthest to the left. The dotted line denotes the 99% confidence interval of the technical error (**Figure 2.9**). mLRRY values within the 99% CI are considered normal. Using the findings of previous mLOY studies, mLRRY values at the 99% CI represent ~10% of cells missing chrY. The blue line at -0.150 represents ~19% of cells missing LOY, and the green line at -0.315 represents ~35% of cells missing LOY (commonly referred to as severe loss). Note: mLRRY is commonly used as a continuous variable in the study.

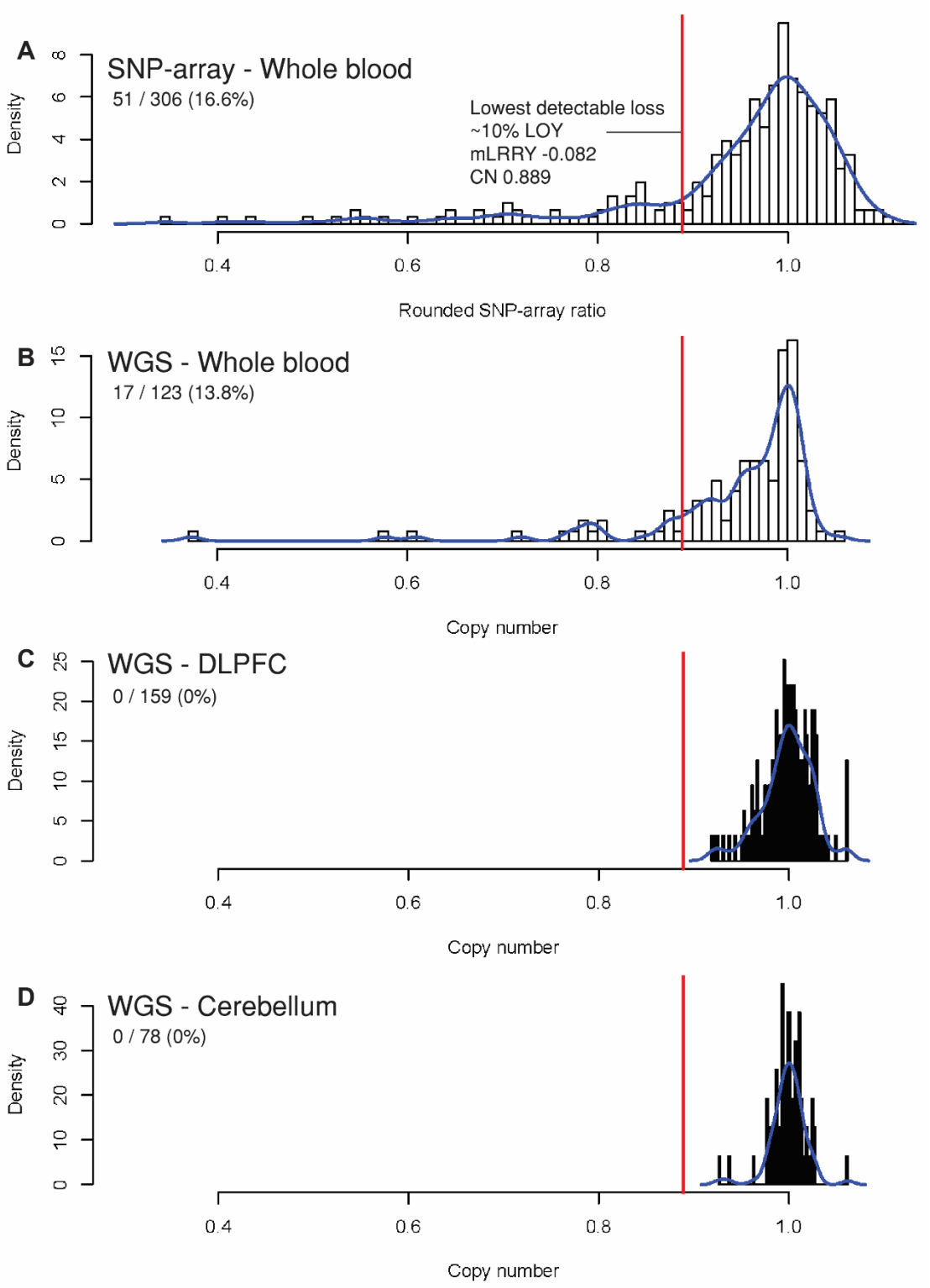


Figure 2.13. LOY is detectable in the blood but not the brain. A-B) Blood samples analyzed for LOY using SNP-array and WGS show similar distributions and estimates of LOY prevalence. The mLRRY unit was transformed and normalized to represent biological ploidy allowing for the ability to directly compare between technologies. In the array dataset, 51/306 (16.6%) of males showed LOY in blood, while 17/123 (13.8%) in WGS showed LOY in blood. We observed 1 gain of Y event. C-D) Using the same LOY stringency (10% LOY cells), 0/159 and 0/78 samples showed LOY in DLPFC and cerebellum, respectively. Although this does not disprove LOY occurrence in the brain, it does provide evidence that LOY is occurring at greater rates in the blood.

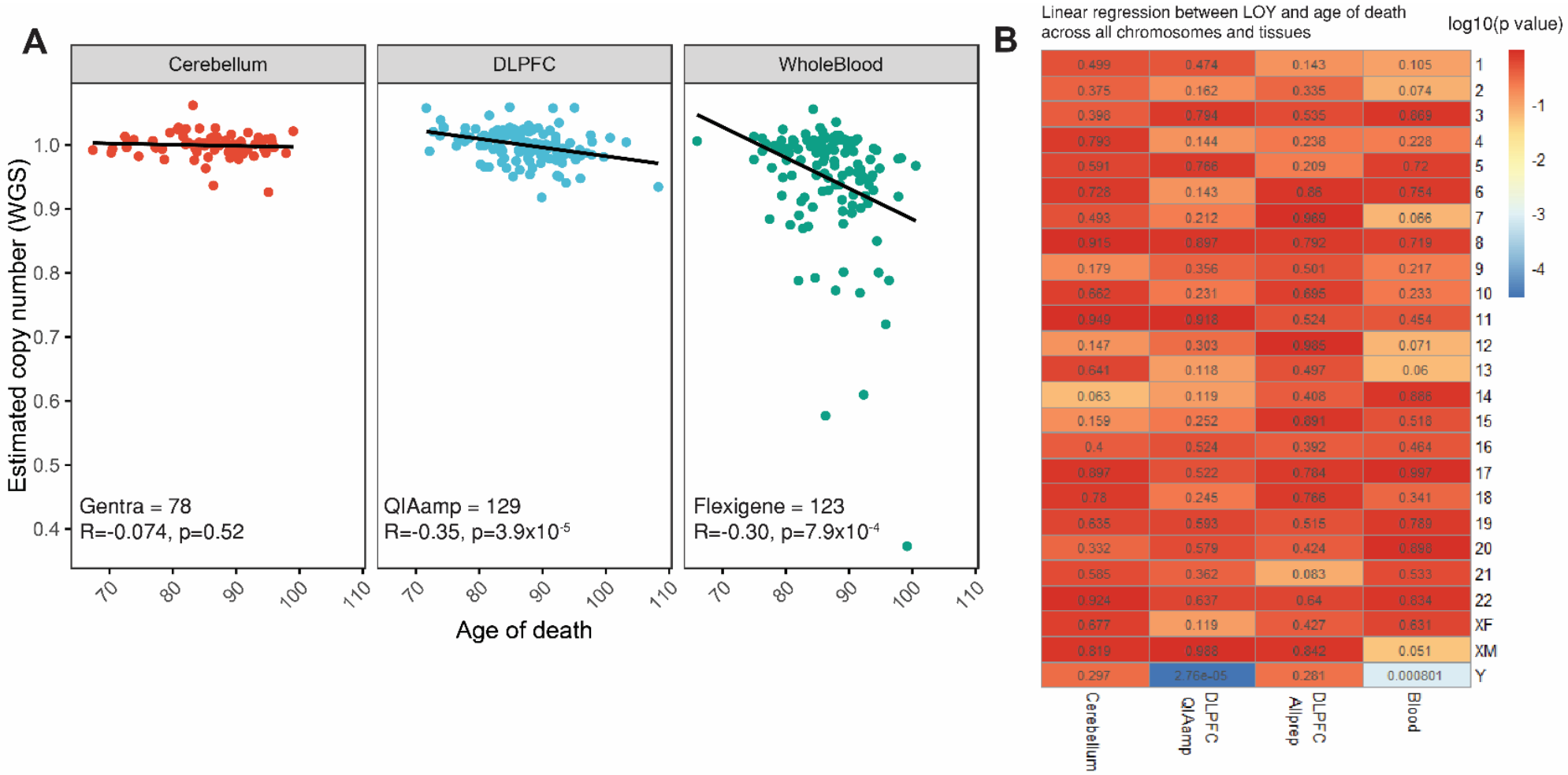


Figure 2.14 LOY is significantly associated with sample age in both whole blood and dorsolateral prefrontal cortex (DLPFC). A) In whole blood and DLPFC tissues age of death is significantly associated with LOY (DLPFC, R=-0.35, *p*=3.9x10-5; whole blood, R=-0.30, *p*=7.9x10-4; Pearson). In the cerebellum this association was not found (R=-0.074, *p*=0.52). Consistent batch specific depth distributions caused technical LOY/age correlations within several autosomes. As a result, the DLFPC batches were separated. B) The Y chromosome is the only chromosome whose estimated copy number correlates with age. Linear models were applied independently to each chromosome in each tissue, controlling for smoking, and other confounders. LOY was associated with age in QIAamp (adj R=-0.33, *p*= 2.76x10-5, n=122) but not AllPrep (adj R=-0.37, *p*= 0.281, n = 20). Dorsolateral prefrontal cortex (DLPFC).

specific depth distributions caused technical LOY/age correlations within several autosomes. As a result, the DLFPC batches were separated and linear models were applied independently (QIAamp; R=-0.33, p= 2.76x10-5, n=122; AllPrep: R=-0.37, p= 0.281, n=20). A 2019 study found a similar negative correlation between age and DLPFC Y content using ddPCR, albeit on a much smaller sample size (n=29).[132]

2.3.2 Comparing LOY between tissues and technology

After quality control and filtering, in samples with overlapping WGS and SNP-array data (n=40) our measures of LOY were highly correlated (R=0.92, Pearson correlation; **Figure 2.15B; Appendix 2.9**). Comparable levels of correlation have been found in similar WGS/SNP-array studies, giving us confidence in our methods and ability to make comparisons between data types and tissues. However, a correlation this strong was unexpected as precise WGS and array blood draw dates were unavailable and were likely taken at different time points which should have introduced additional variance.

After standardizing mLRRY values to rounded SNP-array ratio values between the two technologies were compared directly (**Figure 2.13 and Figure 2.15)**. Using the minimum detectable LOY threshold at 0.889 CN (equivalent to -0.082 mLRRY), LOY was estimated in blood, cerebellum and DLPFC. Using this threshold, 17 of 123 (13.8%) WGS blood samples and 51 of 306 (16.6%) blood samples from SNP-array were LOY. In comparison, 0 of 237 (0%) brain samples showed evidence of LOY (**Figure 2.15B and C**). Chromosome Y CN values from both blood and brain were centered around CN 1, however blood CN showed a negative skew that was not observed in the brain (skewness, blood = -3.15, brain = -0.601; **Figure 2.16A**). Thus, we conclude that LOY occurs more frequently and more severely in the blood compared to the brain.

Comparing rates of LOY between cerebellum and DLPFC tissues was more difficult because of region and batch specific differences in sequencing quality and depth variation. However, after normalization, the two regions could be compared and Y chromosome CN values were not significantly different between cerebellum

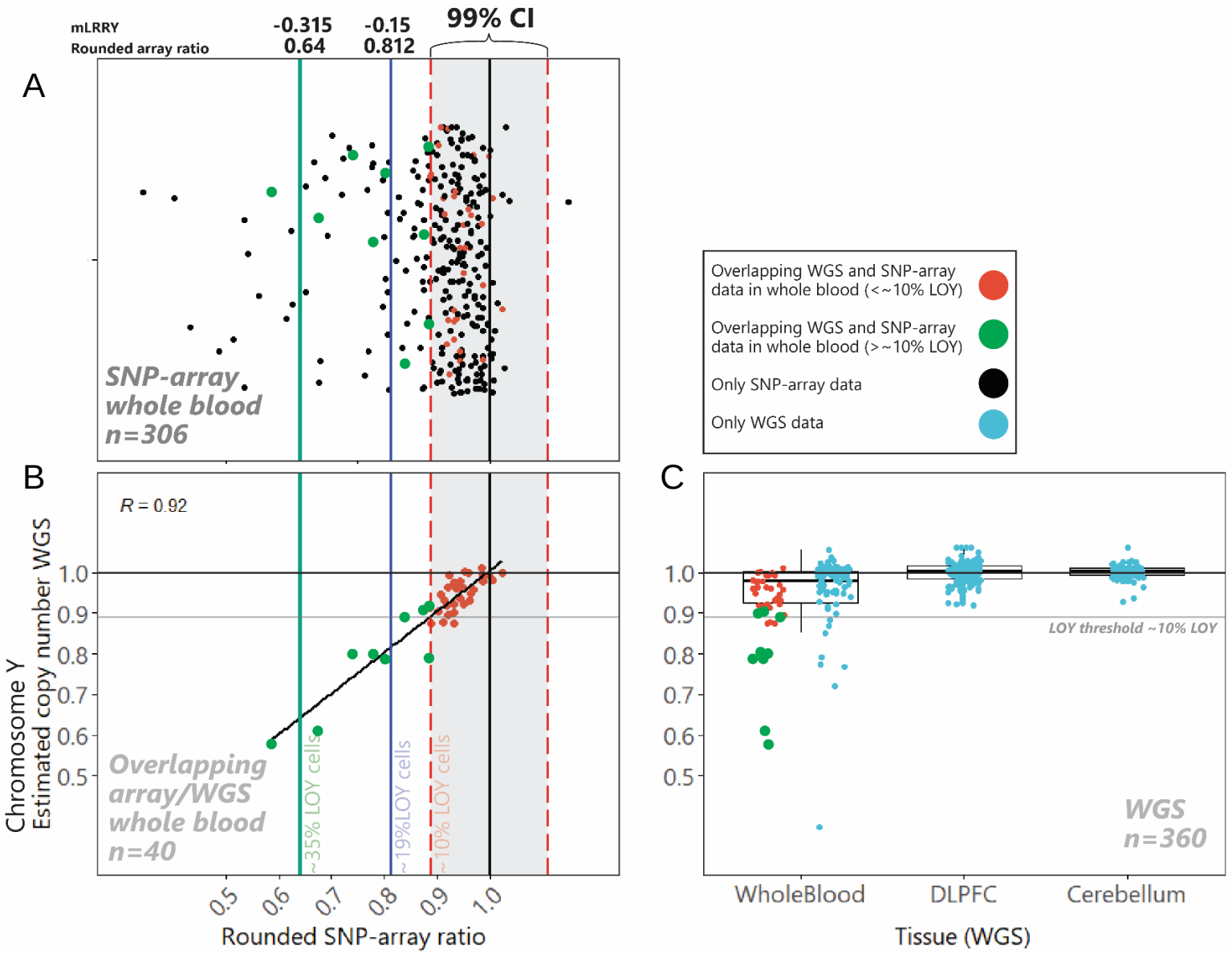


Figure 2.15. Comparing LOY frequency across tissues between WGS and SNP-array. A) Distribution of normalized median Log R Ratio (LRR) values (rounded SNP-array ratio) from the male specific Y region (mLRRY) from all 306 male whole blood samples. The blue lines represent the 99% confidence intervals (CI) of the distribution of experimental variation and represents the lowest detectable mosaic LOY (~10% of cell affected). Each point represents a sample. Each red point has overlapping data between SNP-array and WGS in blood, while each green point represents an overlapping sample with significant LOY detection in array. B) We found a high correlation between array LOY values and estimated copy number (CN) in samples with overlapping data from blood (n=40, *R*=0.93; Pearson correlation). C) Estimated Y chromosome copy number for all WGS samples from blood (n=127), dorsolateral prefrontal cortex (n=159), and cerebellum (n=78). Each red point has overlapping data between SNP-array and WGS in blood (n=40). 13.8% of blood samples and 0% of brain samples exceed the LOY threshold. These results suggest LOY affects <10% of sequenced cells in the cerebellum and dorsolateral prefrontal cortex.

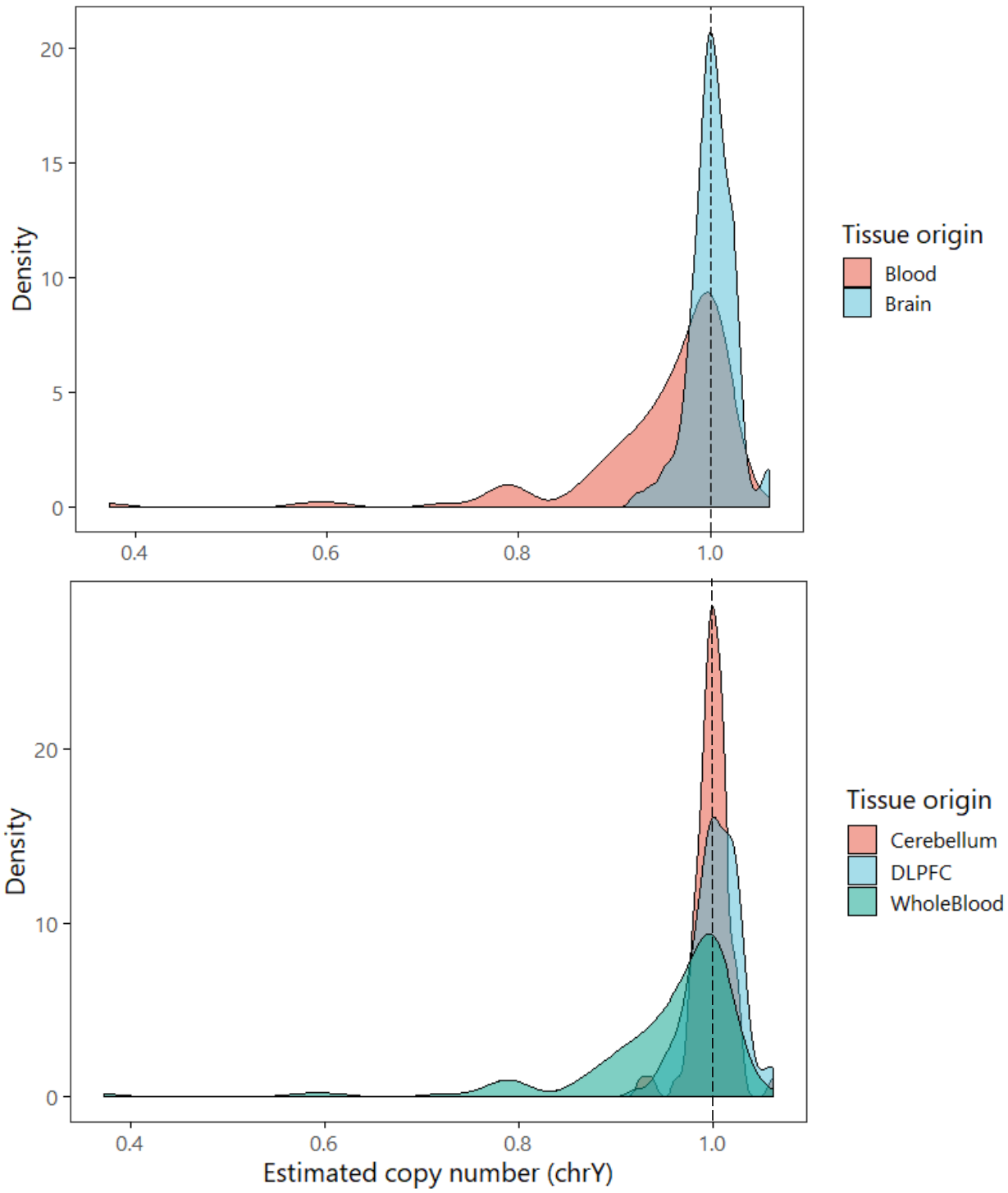


Figure 2.16. Chromosome Y ploidy distributions are negatively skewed in blood but not brain. Top, Chromosome Y CN values from both blood and brain were centered around CN 1, however blood CN showed a negative skew that was not observed in the brain (skewness, blood = -3.15, brain = -0.601). The dotted line highlights the expected ploidy of chromosome Y. Bottom, when distributions are observed by specific tissue the trend remains.

and DLPFC (p=0.16, t-test, **Appendix Figure 2.10**). Differences were within the range of experimental error and given the coarseness of LOY detection using bulk sequencing, a significant difference could not be detected.

With that being said, two additional tests suggest LOY may be occurring in the DLPFC at a frequency below the detectable threshold (~10% LOY cells) of the genotype array. In addition to the 40 paired WGS/array whole blood samples, 74 samples in our cohort were subject to whole blood genotype array and WGS from either cerebellum (n=16) or DLPFC (n=58) tissue. Interestingly, LOY measurements between paired array whole blood and DLPFC WGS samples were significantly correlated (τ =0.29, p=0.0012; Kendall), while in paired whole blood/cerebellum samples they were not (τ =0.22, p=0.24; Kendall; **Figure 2.17**). It is important to note, sample size was much larger for the DLPFC group and correlation directionality was similar. However, it is interesting that paired samples with severe Y loss in blood (>20% loss), all corresponded to Y ploidy levels in the bottom quartile in DLPFC. Also, as reported above, estimated Y chromosome CN in the DLPFC but not cerebellum, was correlated with sample age.

2.4 Chapter summary and conclusion

Mosaic loss of chromosome Y (LOY) has long been observed in blood and bone marrow of elderly men. Recent association studies have found significant correlations between LOY and age-related diseases such as macular degeneration and Alzheimer's disease. Although the role of aneuploidy in psychiatric disorders, and neurodegerative disease remains an active field of research, the characterization of acquired LOY in CNS tissues is unclear and unreported. To further characterize the relative rate of LOY in brain tissue compared to blood we analysed whole genome sequence data (WGS) from an elderly male cohort (median age = 87.5) including dorsolateral prefrontal cortex (WGS; n=159), cerebellum samples (WGS; n=78) and whole blood (WGS; n=123, SNP-array; n=306). Using average sequencing read depth in genomic windows, filtering based on attributes including mappability, GC content, and benchmarking using SNP-array data we were able to establish a reliable algorithm for detecting Y loss using WGS. We concluded that mosaic Y loss that affects >10% of cells occurs

significantly more in blood (13.8%; 17 / 123) compared to brain tissue (0%; 0 / 237). Although LOY was not discovered at our 10% LOY threshold, we found two lines of evidence that support low-frequency LOY in the brain. First, LOY was significant associated with age in the DLPFC (R=0.13, p= 1.13 x 10-4) but not the cerebellum (R=-0.074, p=0.52). Secondly, in samples with paired blood and brain genomic data, we found LOY in blood was significantly associated with LOY in DLPFC (τ =0.29, p=0.0012) but not cerebellum (τ =0.22, p=0.24). Together we conclude there may be low-frequency LOY occurring in the DLPFC that is not observable given the low resolution provided by bulk whole genome sequencing. Further in-depth and higher resolution investigations (such as those in Chapter 3) are required to confirm this observation.

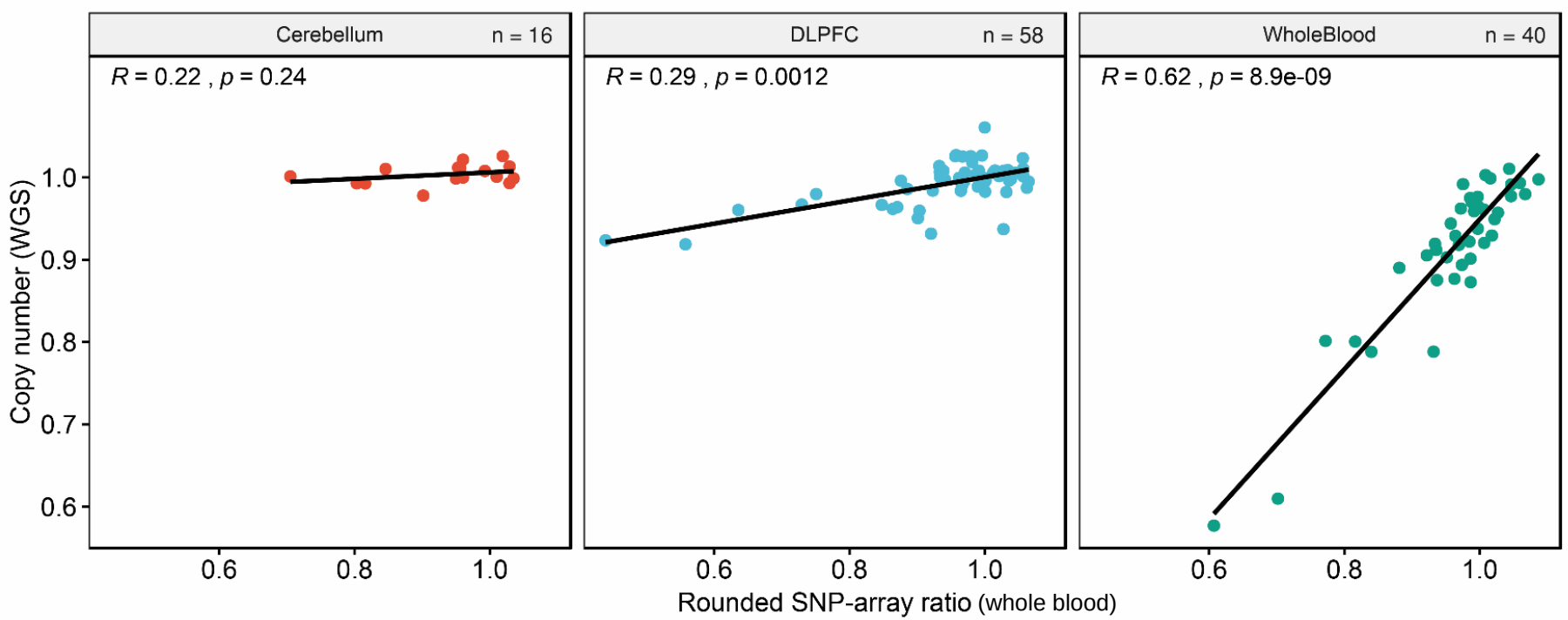


Figure 2.17. Estimated Y chromosome ploidy from dorsolateral prefrontal cortex (DLPFC) and whole blood WGS correlates with ploidy estimates from paired whole blood SNP-array. 114 samples in our cohort were subject to whole blood genotype array and WGS from either cerebellum, DLPFC or whole blood. As expected, paired WGS and array LOY measures from blood were similarly ranked between technologies (τ =0.62, p=8.9x10-9; Kendall rank correlation). Interestingly, paired array blood and DLPFC WGS were also significantly correlated (τ =0.29, p=0.0012), while paired cerebellum samples were not (τ =0.22, p=0.24), although sample size was much smaller (cerebellum/blood, n=16; DLPFC/blood, n = 58; blood/blood, n=40).

3. Loss of Y detection in dorsolateral prefrontal cortex tissue using single-nucleus RNA-seq

3.1 Chapter introduction

Following the bulk WGS analysis of chromosome Y aneuploidy in the dorsolateral prefrontal cortex, I concluded it was reasonable to assume that mosaic loss of Y could be occurring below the detection rate of the methods being used. Although LOY rates in the cortex did not exceed the ~10% LOY cut-off required by the SNP-array detection method, there were two lines of evidence that suggested low-frequency LOY was occurring in cortex cells. First, we observed an age-associated reduction in Y chromosome ploidy in the DLPFC that was not observed in the cerebellum. Secondly, we observed that individuals with high LOY rates in blood (>20% of cells affected) tended to also show reduced Y ploidy in cortex, albeit at a lower level. In many cases, bulk sequencing does not provide the resolution required to accurately detect low-frequency mosaic ploidy loss, so a more in-depth, higher resolution assay was required.

In theory, single-cell sequencing can provide single-cell resolution evidence of Y nullploidy. The cell is the basic unit of life and in the simplest case Y chromosome content can either be detected or not. While single-cell WGS is ideal for the study of genetic mosaicism, single-cell RNAseq can be leveraged as a proxy for genomic content, although the methods to do so are still being developed. Single-cell RNAseq has the additional benefit of transcriptome clustering and cell-type annotation. The transcriptomes of single-cells can be clustered and labelled by cell-type using known gene markers which enables cell-type specificity when estimating aneuploid events. The goal of the work presented in this chapter was to develop methods to detect mosaic loss of chromosome Y using low-depth single-nuclei RNAseq from the dorsolateral prefrontal cortex of elderly men. I used these methods to predict and contrast LOY rates by cell-type.

3.2 Methods

3.2.1 Data characteristics

Sample selection, isolation of nuclei, and droplet-based single-nucleus RNA-seq (snRNA-seq) was completed and reported by Mathys *et al*. (2019).[133] A total of 48 individuals were selected from the ROSMAP cohort. Individuals were specifically selected to represent an equal proportion of sex, and a range of neuropathological characteristics (from no/mild to severe Alzheimer's disease (AD) pathology). A total of 24 control samples were selected (12 male/12 female; **Figure 3.1**). AD-pathology individuals were also balanced between sexes. Both control and affected groups were matched for age (median: 86.7 AD-pathology, 87.1 no-pathology) and years of education (median: 19.5 AD-pathology, 18 no-pathology), as increased age and reduced education have previously been associated with increased risk of AD. A summary of the methods used during this chapter is provided in **Figure 3.2.** Detailed methods regarding library preparation and sequencing can be found in **Appendix 1**.

In total 1.36 billion mapped reads were produced across 70,627 cells (**Figure 3.3C).** Total sequencing depth also differed significantly between samples, ranging from over 48 million reads to 2 million. ROS20 was removed from the study as sequencing depth, sequencing quality and total number of cells were all reduced (**Figure 3.3B-C).**

3.2.2 Single-nuclei RNA-seq alignment

I downloaded snRNA-seq FASTQ files for all 48 samples from the Synapse AMP-AD Knowledge Portal. I used the 10x Genomics Cell Ranger (v3.1) package to align reads to the GRCh38 genome and generate the gene expression matrix. The reference package that 10x Genomics provides (includes the reference genome, transcriptome, and other data files), restricts analysis to spliced, mature mRNA only. Given the high proportion of pre-mRNA in the nucleus, to improve sensitivity, and account for unspliced nuclear transcripts and reads mapping to pre-mRNA molecules I created a custom reference package using the cellranger *mkref* program (**Appendix 5**). I

called gene expression counts using the Cell Ranger *count* program. Cell Ranger *count* performs sequence alignment, quality filtering, cell barcode counting and unique molecular identifier counting (single-cell expression

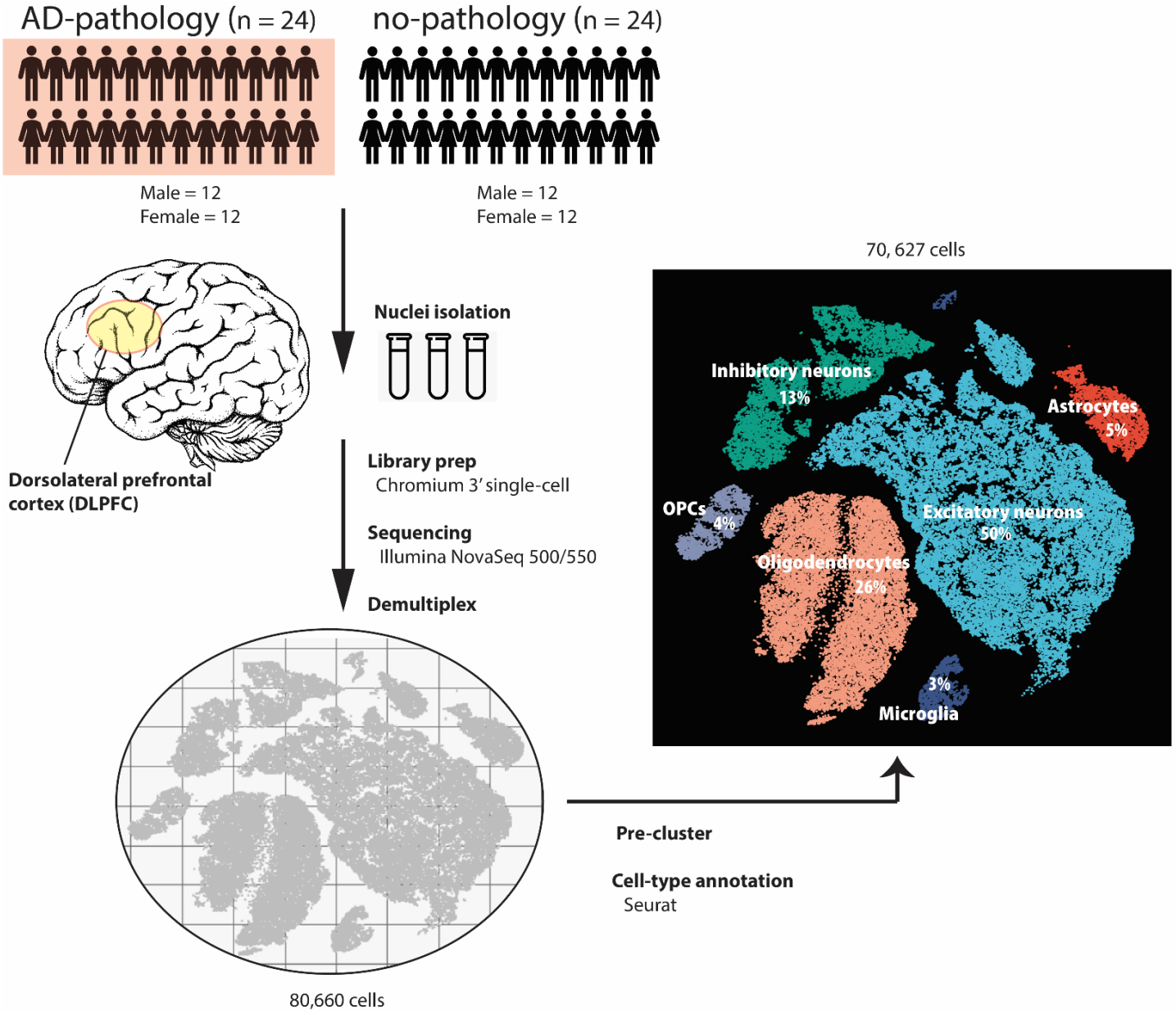


Figure 3.1 Overview of dorsolateral prefrontal cortex single-nuclei RNAseq data processing.
Study cohort details and simplified library preparation/sequencing workflow. 48 samples were selected from the ROSMAP cohort (24 AD-pathology and 24 no-pathology individuals; 24 male and 24 female). 80,600 nuclei were isolated and sequenced. Clustered cells were cell-type annotated using cell-type markers from Lake BB *et al.* (2018). 70,627 cells from 8 major cell types passed quality control, including excitatory neurons, inhibitory neurons, oligodendrocytes, astrocytes, microglia, and oligodendrocyte progenitor cells (OPCs). Pericytes and endothelial cells were removed from the analysis.

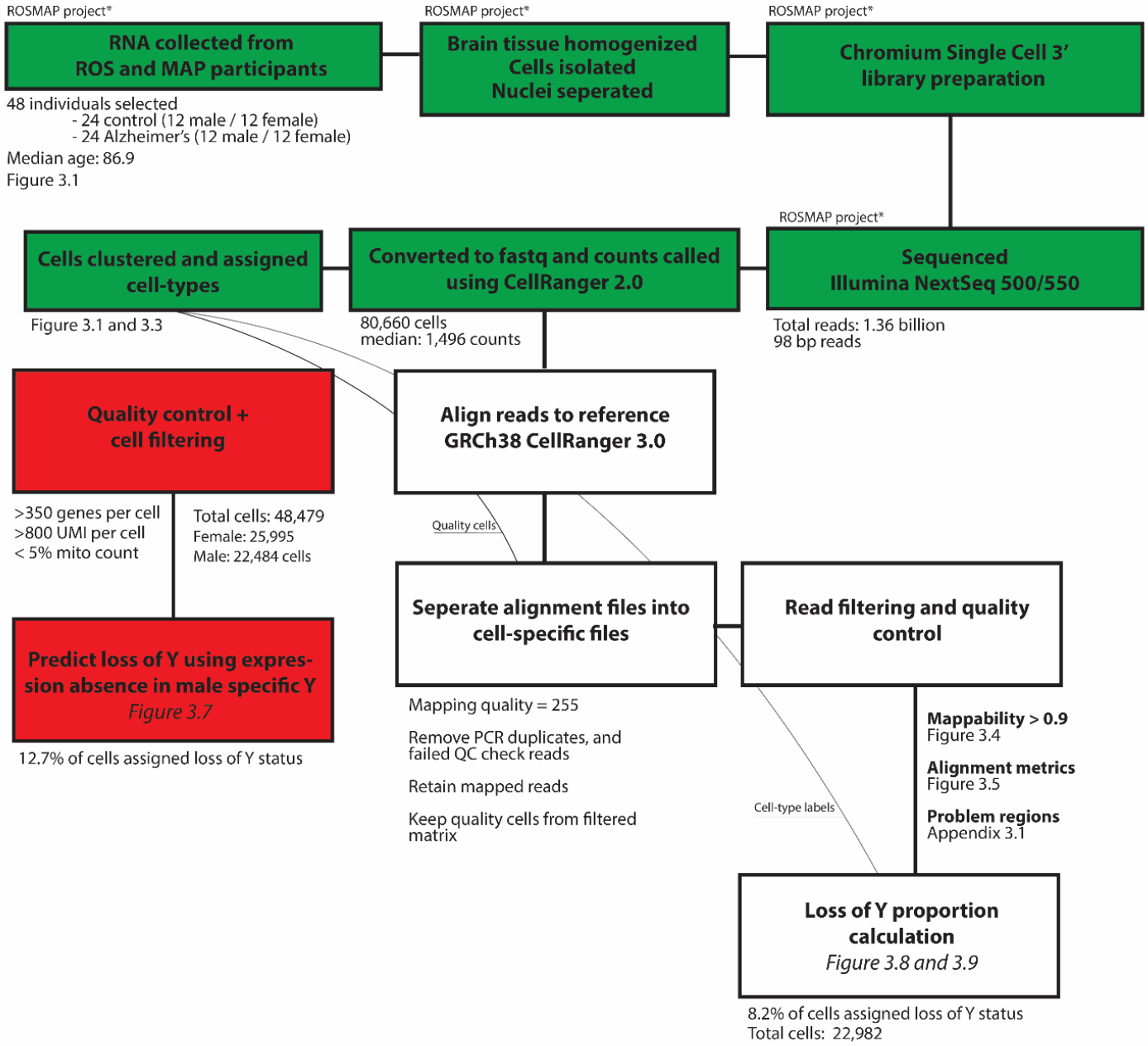


Figure 3.2 Single-nuclei RNAseq LOY detection method overview. A detailed workflow of the single-nuclei RNA-seq loss of Y detection method used in Chapter 3. Green boxes denote work done by the ROSMAP consortium (and Mathys *et al*. 2019) , including all sample selection, preparation, sequencing, and demultiplexing. Red boxes denote work done using previously curated, filtered expression matrices from Mathys et al. (2019). White boxes denote the workflow I developed to improve loss of Y detection sensitivity using raw read filtering.

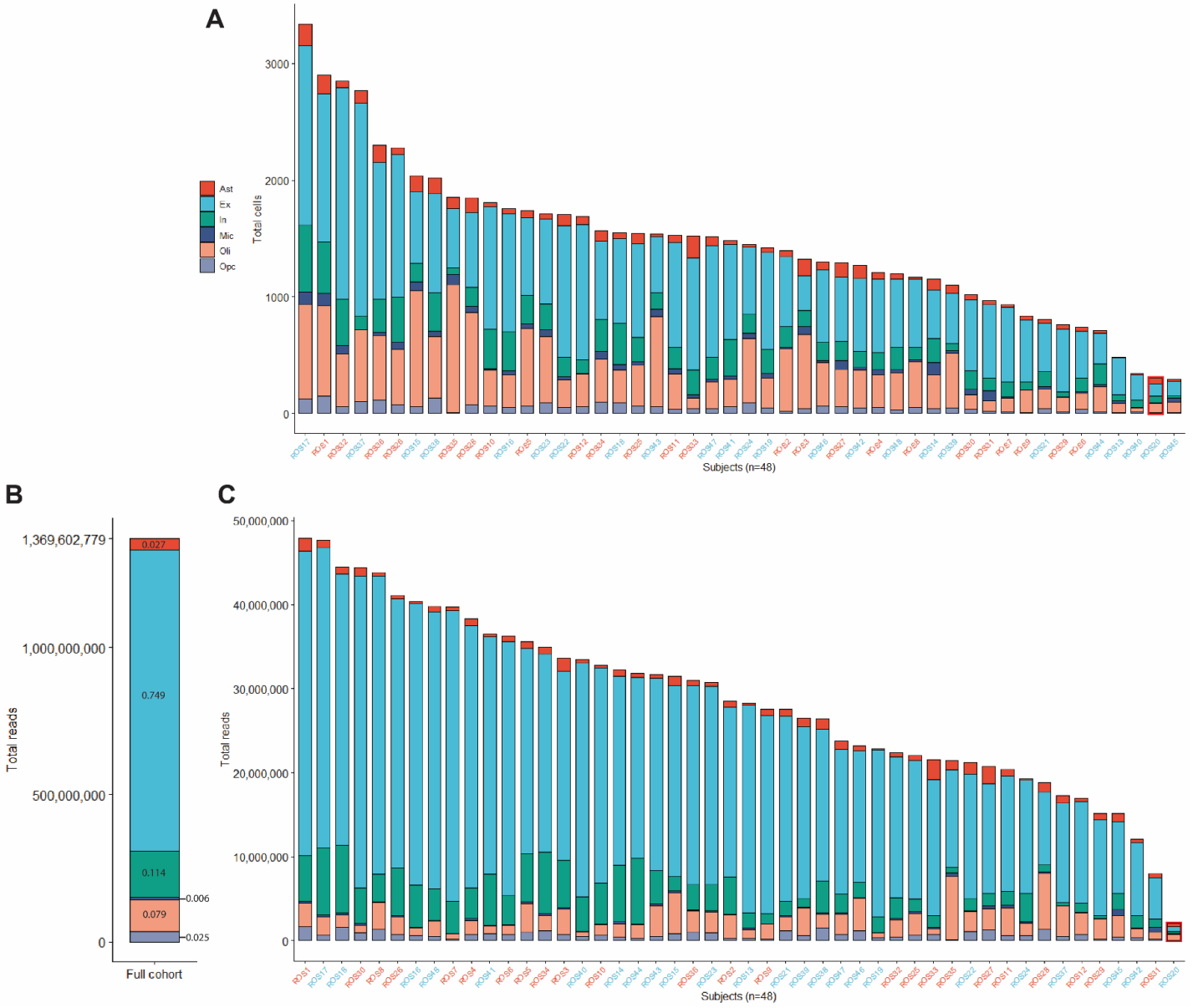


Figure 3.3 Total number of cells and sequencing depth for all 48 dorsolateral prefrontal cortex, snRNA-seq samples. A) The number of cells sequenced for each sample varies (mean = 1,471; sd = 644) and ranges from 297 to 3,364. Red sample labels are female, blue sample labels are males. **B)** Total cumulative sequencing reads by cell-type across all samples. Sequencing output varies between cell-types, which can complicate intra-cell-type comparative analyzes. Approximately 75% of all sequencing reads are derived from excitatory neurons, while only 0.6% are from microglia. Brain cell-types are diverse and have vastly different transcriptional output. A total of 1.36 billion mapped reads were produced across 70,627 cells. **C)** Total sequencing depth also varies significantly between sample, ranging from over 48 million reads to 2 million. ROS20 (highlighted in red box) was removed from the study as sequencing depth, sequencing quality and total number of cells were all significant less than average.

counts). I input FASTQ files were into the program, and cell-level gene expression matrices and binary alignment map files (BAM) were output, both of which were used downstream (**Appendix 5**).

3.2.3 Split sequence alignment files by cell

Cell Ranger *count* produces read alignment files or BAM files (aligned to GRCh38) for each sample. These files contain all reads for all cells by sample. I split alignment files into cell-specific alignment files using bamCleave (https://warwick.ac.uk/fac/sci/systemsbiology/staff/dyer/software/bamcleave/). 10x Genomics' library preparation provides each cell with a unique cell identifier and bamCleave separates reads using these identifiers into individual BAM files. Cell IDs that passed quality control in the already published processed Seurat object were kept, and the rest were removed (see **Expression quality control**). This allows for an increasingly sensitive analysis as cell-specific, raw reads can be observed without strict Cell Ranger count filtering. On the contrary, it also means more false positive reads and PCR artefacts are to be included in the analysis, as the detailed, 10x Genomics specific, Cell Ranger filtering algorithms are bypassed. For this reason, I employed my own read filtering process based on mappability, read quality, false positive peaks, and alignment statistics.

3.2.4 Mappability, problem region and blacklist filtering

First, I filtered reads on standard mapping quality and sequencing quality metrics and counted and stored using SAMtools. Reads were counted genome-wide, and on the X and Y chromosomes. As with previous analyses, the Y chromosome search was confined to the male-specific region (MSY) whereas the X chromosome was confined to the non-PAR X region. Unmapped reads and reads with less than perfect mapping quality (255) were filtered. The details of all passing X and Y reads were stored for further processing (**Appendix 5**).

Next, I filtered reads based on mappability score. As in Chapter 2, the Y chromosome reference genome was split into 1000bp windows. Using the GEM mapping tool, a mappability score was produced for each window across the GRCh38 reference.[126] Reads overlapping multiple windows were assigned proportional mappability scores based on their window membership. Reads within genomic windows exceeding a 0.9

mappability score were kept (**Figure 3.4A**). Reads with reduced mappability are more likely to be errant/ambiguous mapping events. The mappability filter removed 82% of female Y reads (which we assume are false positive evidence of the Y chromosome) and 17% of male Y reads. When applied to the X chromosome, the mappability filter removed 11.4% and 11.6% of female X and male X reads, respectively, suggesting that errant/mis-mapped reads are being removed at a greater rate than true reads. Additionally, I removed reads within ENCODE mappability blacklist regions.

I applied additional filters to X and Y reads based on read alignment statistics provided by the STAR aligner. The number of mismatches (Nm) and alignment score (AS) metrics were used to further select high confidence reads (**Figure 3.5**). The AS is an internal measure of read mapping quality assigned by the STAR aligner. Alignment gaps, mismatches, and other differences from a read mapping perfectly to the reference genome reduces the alignment score. Nm refers to the specific number of base differences between the aligned read and the reference genome. Nm and AS distributions were considerably different between male and female Y reads (**Figure 3.5A-B)**. Reads with AS score greater than 85, and Nm less than 4 (< 4 mismatches) were retained. In total, when combined with mappability filters, I removed >98% of all female reads and just 33% of male reads (**Figure 3.5C**).

Lastly, I removed reads from regions I have called "problem regions", which are derived from peaks of mismapping female reads errantly assigned to the Y chromosome. In all female samples the locations of these mismapped reads were piled up across Y chromosome coordinates. Genomic windows of 1000bp that contained >10 female reads mapping to the Y within intergenic regions were removed in both male and female samples (**Appendix 3.1**). The assumption was that these false positive mapped reads are located in difficult mapping regions of the Y that are not removed by the current filters. These regions likely exhibit mapping difficulties in males and removing these regions should disproportionately remove technical noise and confounding reads. I found that most female reads mapping to chrY were located within repetitive DNA elements, such as Alu SINE elements (**Appendix 3.1**).

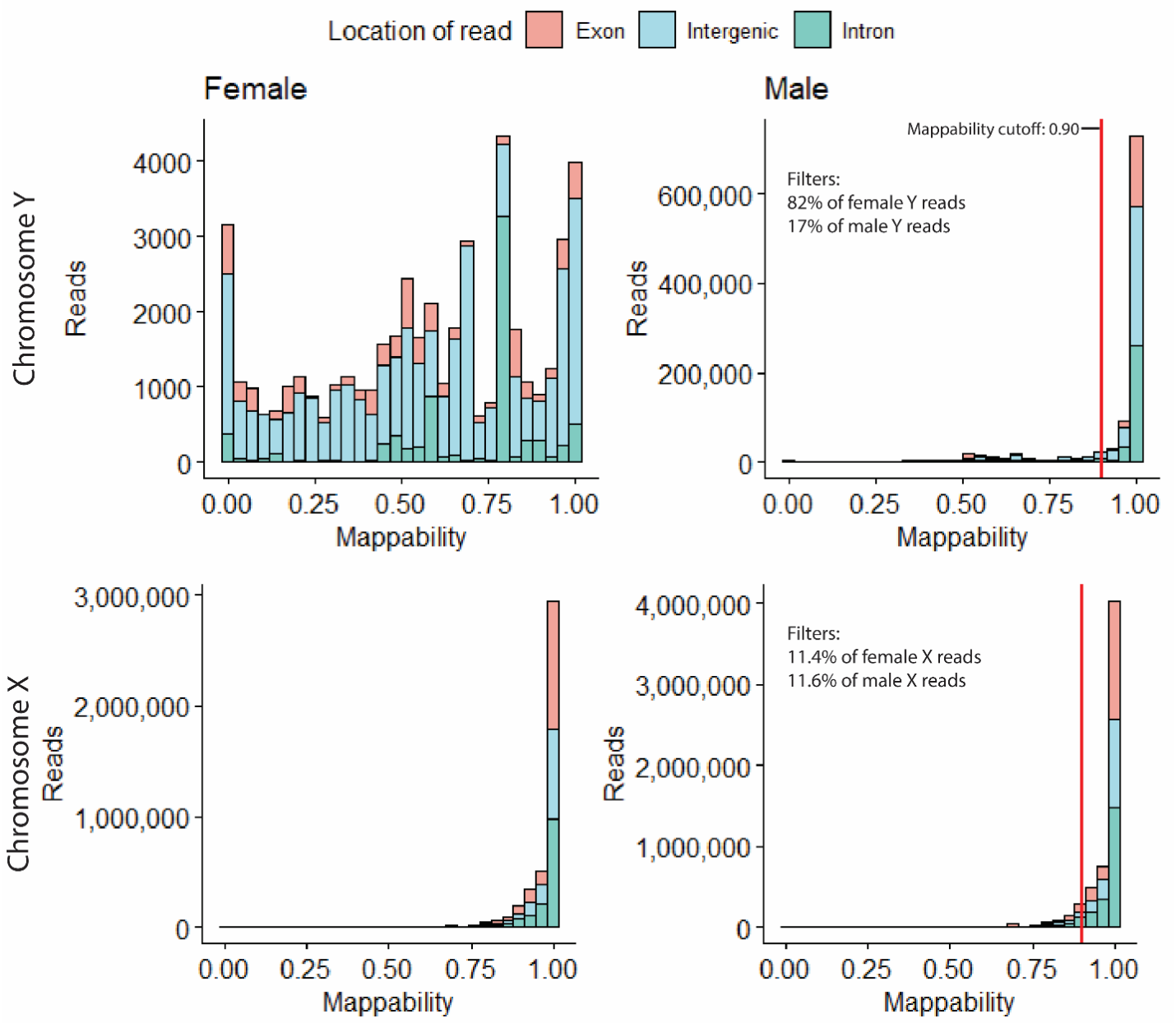


Figure 3.4 Effect of the mappability filter on the sex chromosomes in male and female samples. Top) Histogram of the mappability score given to reads mapping to chrY in male and female samples. Male Y reads are concentrated in highly mappable, unique windows of the genome. Whereas for reads mapping to the Y in females, mappability is uniformly distributed. At a mappability threshold of 0.9, 82% of male reads were retained compared to 17% of female reads. **Bottom)** Histogram of X chromosome read mappability scores for male and females. Mappability distributions are similar between male and female samples, and the mappability threshold of 0.9 removes a similar percentage of reads from each sex (male = 11.6%, female = 11.4%).

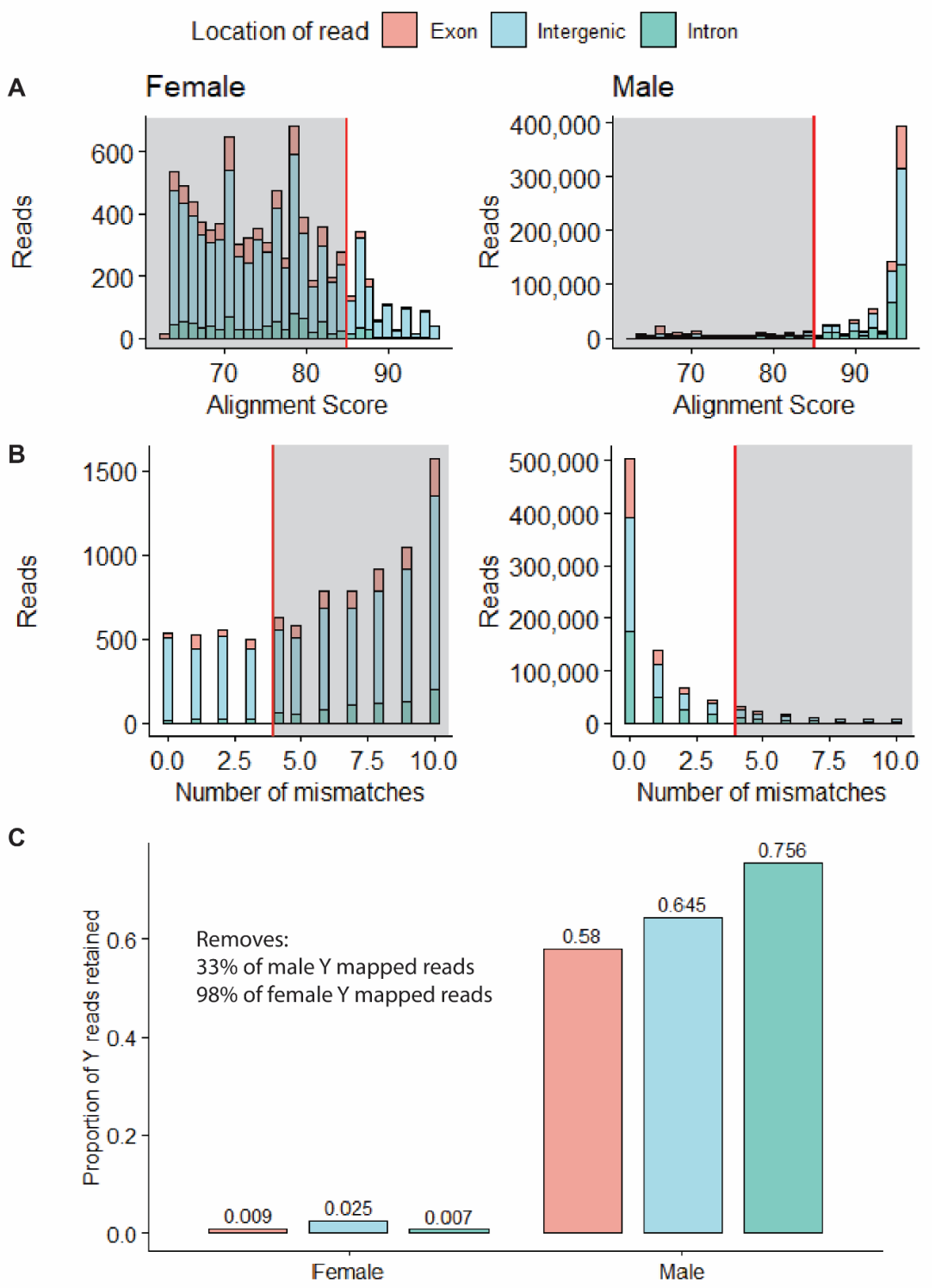


Figure 3.5 Number of alignment mismatches and alignment score further filter errant mapped reads. Reads were further filtered using two alignment metrics: alignment score (AS) and number of mismatches (Nm). **A)** Histograms of AS for reads mapping to the Y in male and female samples. The alignment score (AS) is an internal measure of alignment assigned by the STAR aligner. Alignment gaps, mismatches, and other deviations from a read mapping perfectly to the reference genome reduces the alignment score. Male AS are concentrated near 100, whereas female AS are more variable and commonly below 90. Reads with an AS greater than 85 were retained. **B)** Histogram of Y chromosome read alignment mismatches (Nm). Female reads mapped to the Y have many more mismatches from the reference genome than male reads. Reads with less than 4 mismatches were retained. **C)** After applying mappability, alignment score, and mismatch cut-offs, >98% of female reads were filtered. 33% of male reads were filtered.

3.2.5 Expression quality control

In addition to processing and aligning the fastq files locally, I also used the raw expression matrix dataset containing gene expression values for 80,660 cells provided by Mathys *et al.* (2019) (**Figure 3.2**).[42] The provided dataset also contained cell-type and sub-cluster annotation for each cell (**Figure 3.1** and **Appendix 3.2)**. Because this dataset had been published and peer-reviewed I used these cell-type annotations instead of clustering, deciding on cell markers and determining cluster cell-type myself. However, I did apply custom quality control measures to each cell in the raw dataset to better suit the LOY detection problem. As per published single-cell pipelines, I filtered cells based on low read depth, low detected gene count, and/or a high proportion of mitochondrial gene count. Each of these metrics are indicative of empty droplets (cells lacking a unique oligonucleotide bead), low quality sequencing, and dying or apoptotic cells. In contrast, cells with abnormally high gene counts were also filtered as they may represent doublet droplets (cells with two or more oligonucleotide beads). I adjusted some of these cutoffs to filter cells more stringently with reduced total expression, as sufficient expression is required to accurately predict Y nullploidy. Specific QC steps are detailed below.

i) Library complexity – cells with fewer than 350 expressed genes and greater than 4500 expressed genes were removed.
ii) Sequencing depth – cells with fewer than 800 total unique gene counts were removed.
iii) Mitochondrial gene expression – cells with greater than 15% of total read counts derived from mitochondrial genes were removed.
iv) Gene expression – genes witth expression in less than 3 cells were removed.

After QC, 48,479 cells (male = 22,484) remained with a median value of 1,474 counts per cell. A total of 17,775 genes were represented in the dataset. All single-cell calculations were completed using the R toolkit Seurat (v3.1; http://satijalab.org/seurat/). Once cells and genes were filtered, I log normalized expression counts and multiplied them by a scale factor of 10,000 using the Seurat *NormalizeData* function. I then scaled data so mean expression of each gene across all cells was equal to 0 and expression variance of each gene across all cells

was 1 using the Seurat *ScaleData* function. Scaled data was used for clustering and cluster visualization.

3.2.6 Methods for declaring loss of Y cells

When using single-cell RNAseq to detect Y chromosome nullploidy, the least ambiguous situation would be to observe a complete lack of expression from the MSY region of a single-cell, in addition to expected gene expression rates and read depth across the remainder of the genome. In this hypothetical situation, a cell is transcriptionally normal apart from the absence of Y-linked gene expression. Further confidence in this LOY call could be made if cells of the same cell-type with similar total expression also showed stable Y chromosome expression in most cases. However, one inherent problem of using RNA as a proxy for DNA when investigating chromosome Y, is the underlying possibility that Y chromosome genes are not biologically necessary in a subset of cells. In some cells, Y-linked genes are not expressed and/or detected which could lead to false assumptions of a Y nullploid event. As cell read depth increases, the chance of stochastic Y-linked gene dropout decreases as more of the transcriptome is sequenced. To limit the confounding effect of read depth on LOY calling, I confined analysis to specific quantiles of read depth. This way cells being compared had comparable upper and lower limits of total expression and any biases would be systemic biases (i.e. shared amongst cells being tested).

However, after all the filtering and QC steps, LOY calling can still be difficult and not entirely obvious. The observation of frequent low-level Y chromosome expression in female samples shows that some small proportion of mapped Y chromosome reads are false positives (**Appendix 3.3**). On average, 1 of every 800,000 female filtered reads per cell was located on the Y chromosome. These reads likely consist of PCR artefacts, mis-mapped reads derived from highly similar sequence on the X chromosome and other sources. The existence of these false positives complicates nullploidy detection and increases false negative LOY calls. To deal with some of these technical complications I developed 2 methods of LOY detection:

i) Normalized Y

After read filtering for each cell the MSY read counts are divided by total reads and multiplied by a scale factor of 10000. The logp1 (natural logarithm + 1) is then applied to avoid undefined values. Essentially, for this metric I treated the Y chromosome as a single gene and processed its expression using the same methods as the Seurat *NormalizeData* function.

ii) XY ratio

For this method, in each cell the total number of filtered MSY reads are divided by the total filtered non-PAR X chromosome reads. Since the male X is haploid, and is required for cell viability, it provides a stable comparator for Y chromosome depth.

3.2.6.1 Determining LOY cut-offs using female samples

Generally, I determined LOY thresholds by observing trends in female samples that we assume do not possess a Y chromosome (**Figure 3.6A**). Most female cells (94.5%) have an XY ratio of 0, meaning 0 filtered chrY reads are present in the cell. We assume that female cells lack chromosome Y, and all reads that map back to the Y chromosome must be artefacts and represent technical error. 98% of female cells have a XY ratio below 0.005, and >99% of female cells have a XY ratio below 0.01. The X chromosome is expressed at a similar rate between males and females, so when these XY ratio thresholds are applied to male samples, I assumed most reads being filtered are products of mapping noise and technical error. By applying a <= 0.005 XY filter I should be capturing most cells that biologically lack a chrY. 8.6% of all 22,484 male cells from the DLPFC (from 23 samples) have XY ratios of 0 and are clear LOY cases (**Figure 3.6B**). Ultimately, the XY ratio does not impact LOY proportions greatly at the 0.005 cut-off, as only 9 edge-case cells are reassigned from normal to LOY at this threshold. Given the low read depth of the data, I chose strict cut-offs to limit false positive LOY calls. The same process was done to find a suitable threshold for the normalized Y metric. The cut-off was set at 0.4 which is equivalent to ~1 Y read per 20,000 reads (**Appendix 3.3**). The average male normalized Y was 2.16, while in

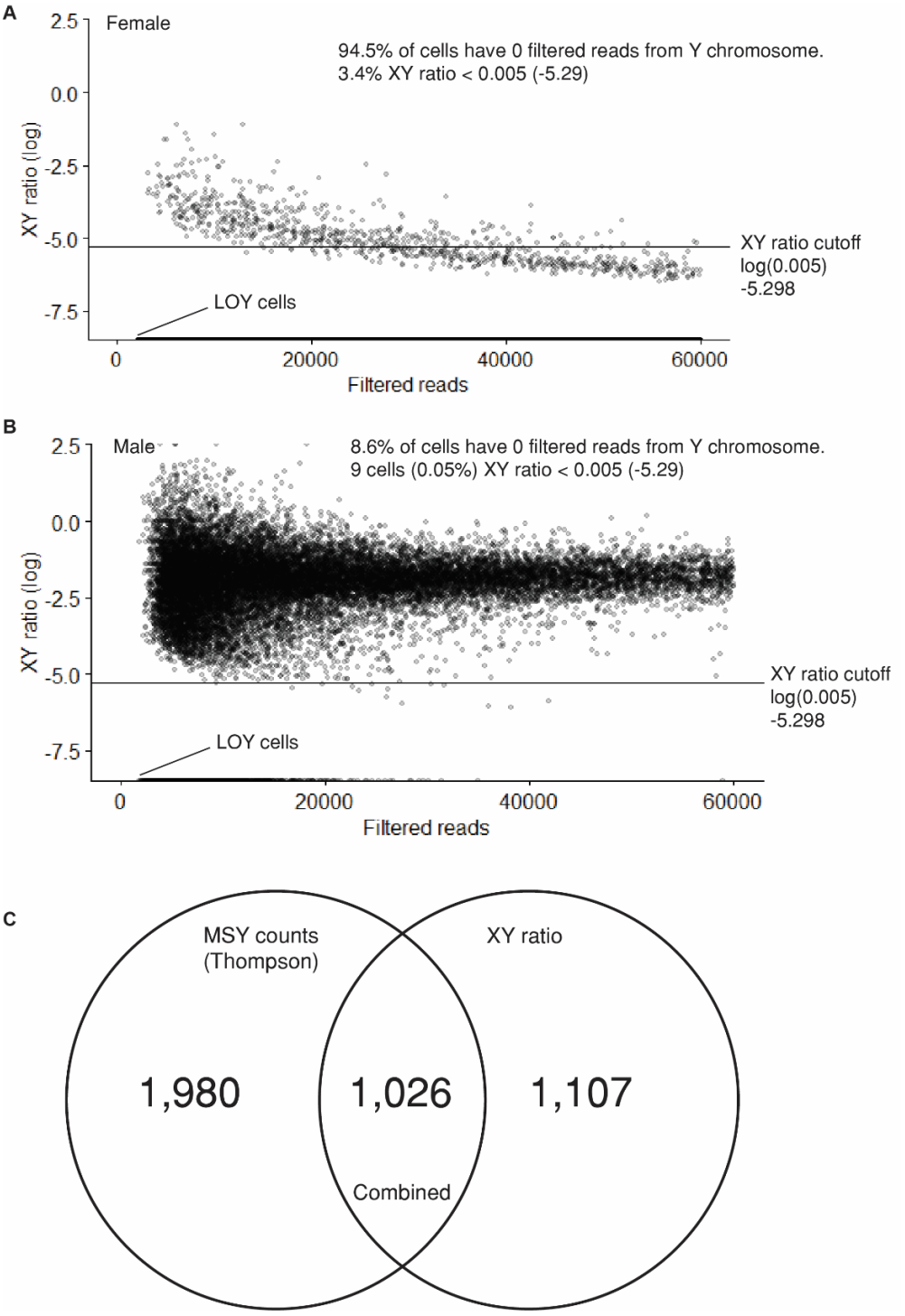


Figure 3.6 Determining the XY-ratio LOY cut-off and comparing to MSY counts method.
A) XY ratio in females is used to model the threshold declaring LOY in males. Each point represents a cell. The logarithmic function has been applied to the XY ratio for visualization. I chose a XY ratio cut-off of 0.005 that detects LOY in 98% of female cells. 94.5% of female cells have 0 Y reads, and 3.5% contain Y reads but are below the threshold. **B)** When the 0.005 XY ratio threshold is applied to males it captures an additional 9 LOY cells, which adds to the 8.6% of cells with 0 Y expression. If a 0.01 XY ratio cut-off is used an additional 47 cells or labelled LOY. **C)** Venn diagram showing the overlap between LOY calls using the MSY gene expression counts (used in Thompson *et al.*) and the XY ratio method. In total 88% of ploidy calls (normal ploidy and LOY calls) were shared between methods. However, only 1,026 XY ratio LOY calls were shared between methods (48.1%).

female samples it was 0.018. The 0.40 cut-off adds an additional 22 LOY edge-case cells to the obvious 0 Y read cells. Surprisingly only 1,026 cells were shared between the MSY counts method and the XY ratio method (**Figure 3.6C**). 1,980 LOY calls were exclusive to the MSY counts method and 1,107 were exclusive to XY ratio. This suggests that each is metric is providing different information.

3.2.6.2 Lack of UMI from MSY region

As of March 2020, the use of single-cell RNAseq to predict Y chromosome loss has only been published once, by Thompson *et al.* in October 2019.[42] Using the 10X Chromium Single Cell 3′ workflow, the group sequenced the transcriptome of 86,160 PBMCs from 19 UK Biobank individuals. Cells with less than 350 expressed genes and 800 unique counts were removed and the LOY status of each cell was determined by a lack of expression from all MSY genes. Using this strategy, 16.5% of cells were assigned LOY status, which was used for downstream differential expression analysis testing. Bulk RNAseq from GTEx shows that Y-linked genes are expressed at similar levels in whole blood and brain tissues (**Appendix 3.4**). Therefore, I replicated this strategy on the Mathys *et al*. [134] single-nuclei DLPFC dataset used in this Chapter. I used various quality control cut-offs in an attempt to calibrate the method given its reduced depth, reduced library complexity and significantly different cell-types (**Table 3.1**). Total expressed gene cut-offs at 250, 350, 500, 800 and 1000 were applied, along with unique count cut-offs at 600, 800, 1000, 2000. Within a cell, LOY status was assigned if all the following Y-linked genes lacked expression: RPS4Y1, ZFY, PCDH11Y, AMELY, TBL1Y, TSPY1, USP9Y, DDX3Y, UTY, TMSB4Y, NLGN4Y, HSFY1, HSFY2, KDM5D, EIF1AY, PRY2, RBMY1J, BPY2, DAZ2, DAZ4. Y-linked genes were included if they showed expression in at least 3 cells in the dataset (**Appendix 3.5**).

	Mathys *et al.* 2019 (used in chapter 3)	**Thompson** *et al.* 2019
Tissue	Frontal cortex	PBMC
Male samples	24	19
Major cell types	8	6
Reads per cell	22,975	64,900
Cells per sample	1,379	4,534
SC sequencing method	single-nucleus	single-cell
Library prepartion	Chromium Single Cell 3′ Reagent Kit v2	Chromium Single Cell 3′ Reagent Kit v2
Sequencer	NovaSeq 6000	NovaSeq 500/550

Table 3.1 Comparison between ROSMAP single-nuclei RNA-seq data (Mathys *et al.* 2019) and UK BioBank single-cell RNA-seq data (Thompson *et al.* 2019). To label cells as LOY, Thompson *et al.* used a lack of expression counts from genes residing in the male specific Y region (MSY). The data used by Thompson *et al.* had a 2.8-fold increase in mean reads per cells compared to the data used in this thesis (Mathys), and therefore their strategy was not directly applicable. Additional filtering and quality control methods were required to replicate the method used by Thompson *et al.*

3.3 Results

3.3.1 Loss of Y detection: replication using MSY counts method (Thompson *et al.*)

I began analyzing the DLPFC single-cell dataset for loss of Y (LOY) using expression counts of genes residing in the male-specific Y (MSY) as a marker of Y chromosome presence. Initially, I replicated quality control cut-offs and thresholds used by Thompson *et al.* to detect LOY in PBMCs.[42] Male cells with <5% mitochondrial RNA (non-apoptotic cells), >350 expressed genes, and >800 UMI were included (n=22,484). Using these quality controls, 2,856 cells were assigned LOY status (12.7%), which is similar to rates found in PBMC (16.5%), but much greater than expected in brain tissue.[42,44,60] When separated by cell type, LOY rates varied significantly, especially between neuronal and non-neuronal cell types (**Figure 3.7A**). Upon further analysis it became clear that LOY assignment using MSY expression counts at these published QC thresholds was highly correlated with cell-specific sequencing depth (**Figure 3.7B-C**). LOY cells had a mean of 1,002 detected genes, and 1,383 total counts, whereas normal cells had a mean of 2,480 detected genes, and 4,599 total counts (**Appendix 3.6**). Cell types with reduced total expression had very high LOY prevalence including microglia (50% LOY), astrocytes (27.6%), and oligodendrocytes (34.1%). Whereas cells with high total expression had low rates of LOY: inhibitory neurons (8.9%), excitatory neurons (5.2%). The large difference in total RNA output and Y-linked gene detection between diverse brain cell-types likely makes simple LOY assignment using MSY counts unreliable when using low-depth single-nuclei data.

After replicating the Thompson *et al.* method[42] it became clear that to produce reliable estimates of LOY, QC would have to be altered and optimized to the specific characteristics of the single-nucleus data. I began by using increasingly strict quality control cut-offs and measuring LOY proportion in each cell-type. LOY rate was highly dependent on both total counts and library complexity (**Appendix 3.7**). Because LOY was so highly associated with read depth and library complexity using this method, I decided a more sensitive method of LOY detection would be beneficial.

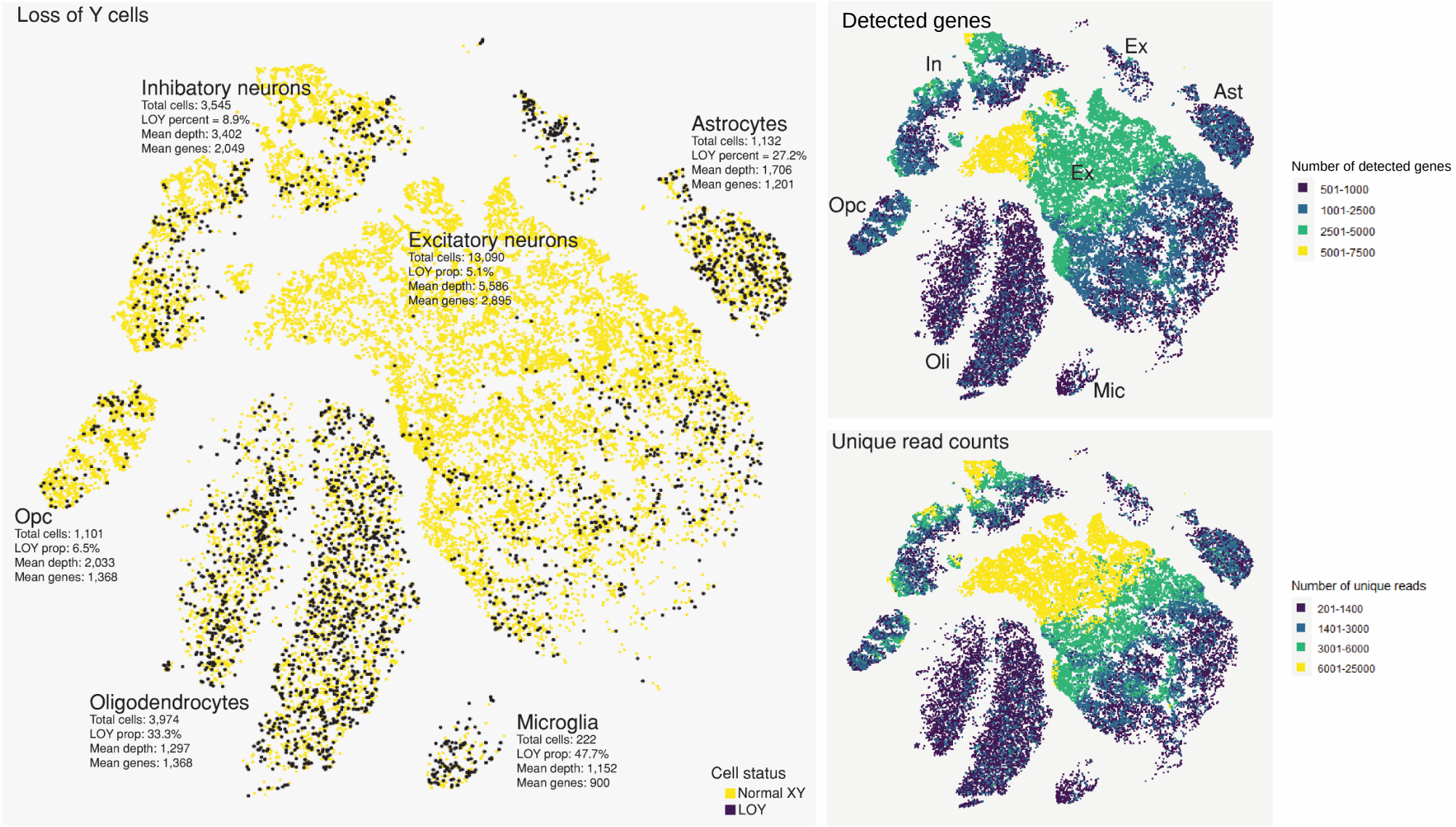


Figure 3.7 T-distributed Stochastic Neighbor Embedding *(t*-SNE*)* clustering of loss of Y labelled cells and correlation with read depth.
Left. MSY gene count method from Thompson *et al.* to detect LOY in PBMCs, visualized using *t*-SNE. Male cells with <5% mitochondrial RNA (non-apoptotic cells), >350 expressed genes, and >800 UMI were included (n=22,484). Using these quality controls, 2,856 cells were assigned LOY status (12.7%). Yellow points represent cells of normal ploidy, while blue points have been labelled LOY. **Right.** Cells assigned LOY are highly concentrated in cells with low library complexity and low unique read counts. Top. Points (cells) are colored based on number of detected genes. Bottom. Points are labelled by read depth.

3.3.2 Sensitive loss of Y detection

In the single-nuclei RNA-seq dataset used in this Chapter, Y-linked gene expression is variable, and subject to frequent gene dropout events where data only captures a small fraction of the transcriptome of each cell. This lack of sensitivity limits the effectiveness of using MSY counts to detect LOY. Cell Ranger, the program used to call gene expression counts from 10x Genomics single-cell data, is conservative when calling gene expression counts. One main goal of the Cell Ranger algorithm is to strictly eliminate PCR artefacts from the amplification process to reliably output biologically true expression counts and limit technical noise. This allowing for accurate clustering, cell-type labelling, pathway analysis and differential expression analysis. The 10x Genomics single-cell assay captures polyadenylated (polyA) transcripts including mRNA, some known long noncoding RNAs and antisense transcripts. When counting unique gene counts, the Cell Ranger algorithm removes intergenic reads, which removes additional evidence of the Y chromosome. Furthermore, Cell Ranger requires reads align to both the genome and transcriptome which further reduces available data. While this is essential for many applications of single-cell transcriptomic analysis, it creates sensitivity problems when trying to detect the presence of a poorly expressed chromosome within low depth data. In order to provide a more accurate assessment of LOY using this dataset, I needed to develop a more sensitive LOY detection method that loosened read inclusion restrictions and used a greater proportion of the available transcriptomic evidence (**Methods**).

3.3.3 Estimated loss of chromosome Y proportions by cell-type

After applying the previous cut-offs of >350 detected genes, >800 UMI, and < 15% mitochondrial gene proportion (increased from 5% as neurons express much more mitochondrial RNA), 22,982 male cells were included. After establishing false positive read tendencies in the female cells (**Methods**), cells with a XY ratio less than 0.005 were labelled LOY cells. Across these cells, my LOY detection method assigned 8.2% of cells as LOY, which was 4.5% less than value determined by the MSY gene expression method. However, again it became apparent that cells with low expression levels were much more likely to be assigned LOY (**Figure 3.8;**

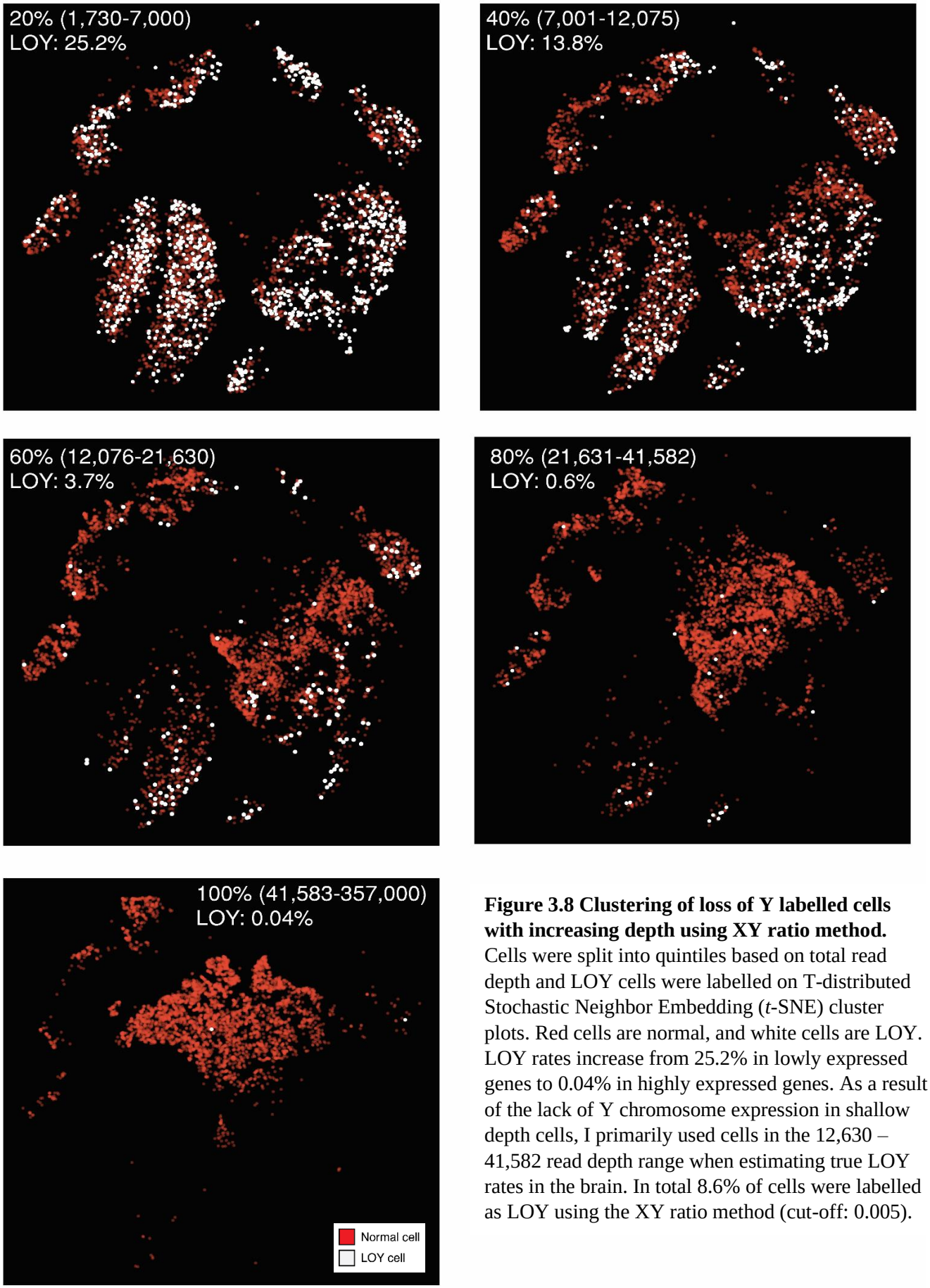


Figure 3.8 Clustering of loss of Y labelled cells with increasing depth using XY ratio method. Cells were split into quintiles based on total read depth and LOY cells were labelled on T-distributed Stochastic Neighbor Embedding (*t*-SNE) cluster plots. Red cells are normal, and white cells are LOY. LOY rates increase from 25.2% in lowly expressed genes to 0.04% in highly expressed genes. As a result of the lack of Y chromosome expression in shallow depth cells, I primarily used cells in the 12,630 – 41,582 read depth range when estimating true LOY rates in the brain. In total 8.6% of cells were labelled as LOY using the XY ratio method (cut-off: 0.005).

Appendix 3.8). In an effort to account for this bias and compare cells of similar depth, I divided cells into five discrete groups based on total filtered cell read depth. In the first 20% (1,730 to 7,000 reads/cell) 25.2% of cells were assigned LOY. As expected, in the subsequent quintiles LOY proportions decreased: 13.8% LOY (40%; 7,001-12,075 reads/cell), 3.7% LOY (60%; 12,076-21,630 reads/cell), 0.6% LOY (80%; 21,631-41,582 reads/cell), 0.04% LOY (100%; 41,583-357,000 reads/cell) (**Figure 3.8**).

Using the discrete read depth groups, I further split LOY proportions by cell type (**Figure 3.9A**). LOY estimates were unstable across all cell types until the 3rd quintile (60%; 21,631-41,582 reads). For example, in the 1st quintile (1,730 to 7,000 reads/cell) 36% of excitatory neurons were LOY, compared to just 3% in the 3rd quintile and <0.01% in the 4th and 5th quintiles. To accurately compare LOY between cell-types I exclusively used cells in the 4th quintile (21,631-41,582 reads/cell), which included 3,444 excitatory neurons, 784 inhibitory neurons, 132 oligodendrocytes, 98 astrocytes, 151 oligodendrocytes progenitor cells and 21 microglia (n=4,630). First, I looked for clustering of LOY cells on the *t*-SNE plot as it is conceivable that LOY cells have similar transcriptomic profiles. However, no noticeable clustering was observed (**Figure 3.9C**). Next, I compared sequencing characteristics such as total number of detected genes, total UMI/cell and filtered X reads/cell between LOY and normal cells within each cell-type (**Figure 3.9E**). In general, LOY cell metrics were slightly reduced compared to normal cells, however in each case values were similar, allowing for an acceptable LOY comparison between cell-types. In this specific group of increased sequencing depth cells (80%; 21,631-41,582 reads/cell), I found LOY rates of 0.23% in excitatory neurons (8 / 3444), 0.12% (1 / 784) in inhibitory neurons, 2.6% in OPCs (4 / 151), 3% in astrocytes (3 / 98), 4.5% in oligodendrocytes (6 / 132), and 28.5% in microglia (6 / 21). Similar trends were found in the 60% quintile as well, albeit at increased rates (**Figure 3.9B**). Although some cell-type sample sizes are small, this data suggests that glial cells and specifically microglia are affected by Y loss at a higher rate than neuronal cells.

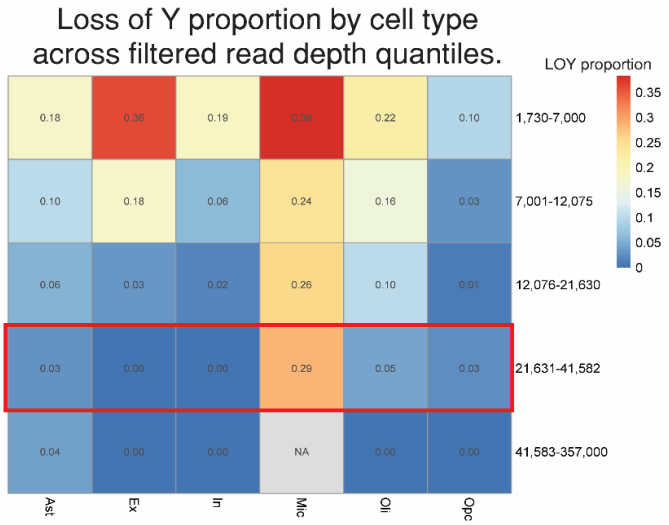
However, I remained skeptical that increased Y loss in glial cells may just be a result of reduced Y-linked gene expression, library complexity and sequencing depth. Neurons express a greater number of transcripts from a

A

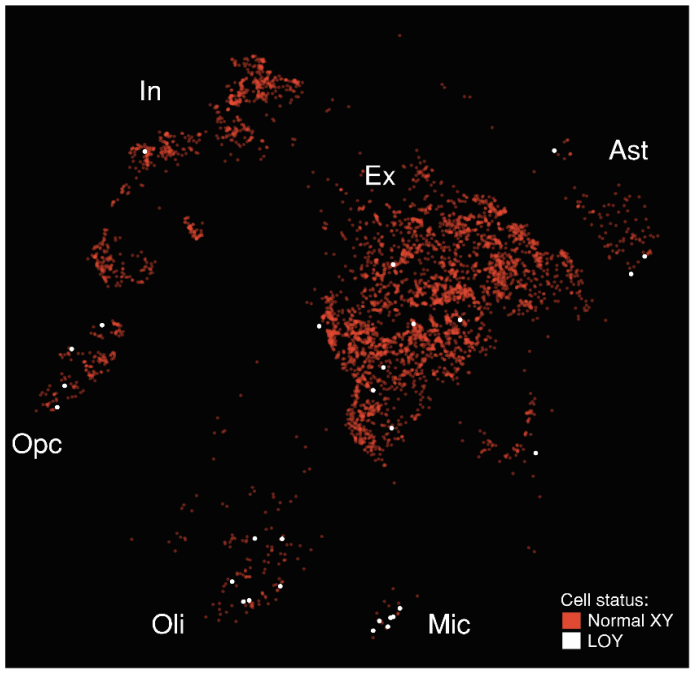
Number of cells within each filtered read depth quantile

Total reads	Percentile	Ex	In	Oli	Ast	Opc	Mic
1,730-7,000	20%	1338	621	1951	302	251	107
7,001-12,075	40%	1716	729	1390	370	337	66
12,076-21,630	60%	2494	956	488	337	324	27
21,631-41,582	80%	3444	784	132	98	151	21
41,583-357,000	100%	4098	455	13	25	38	1

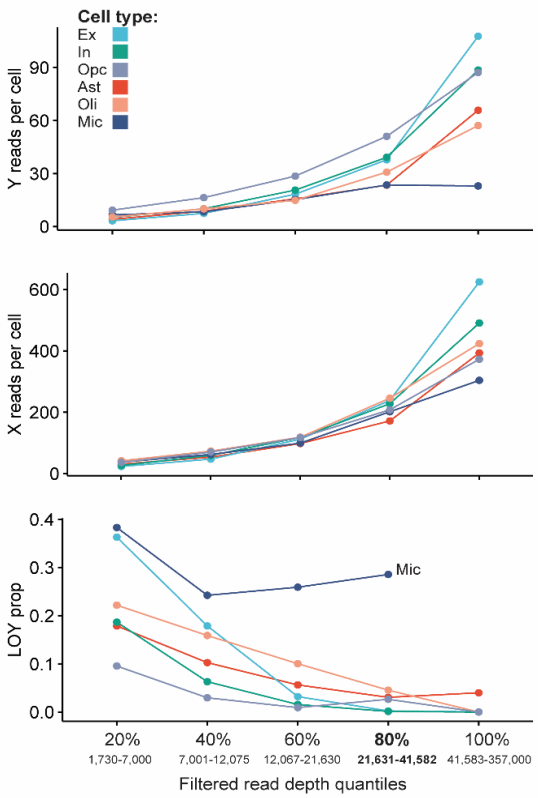
B



C



D



E

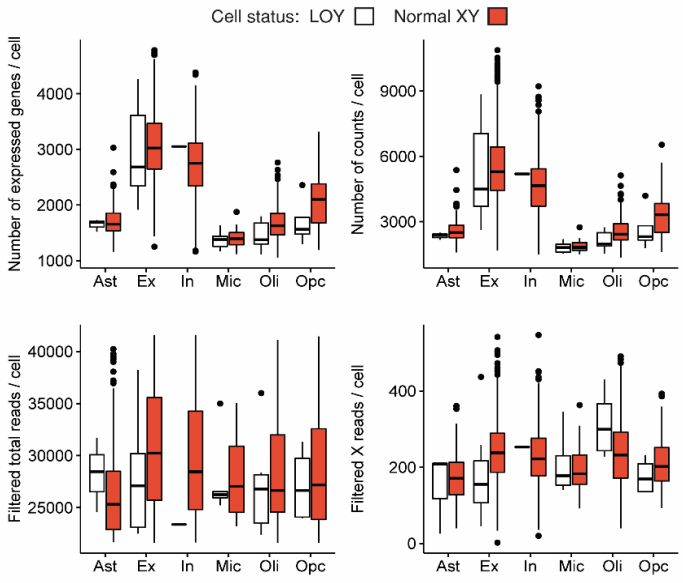
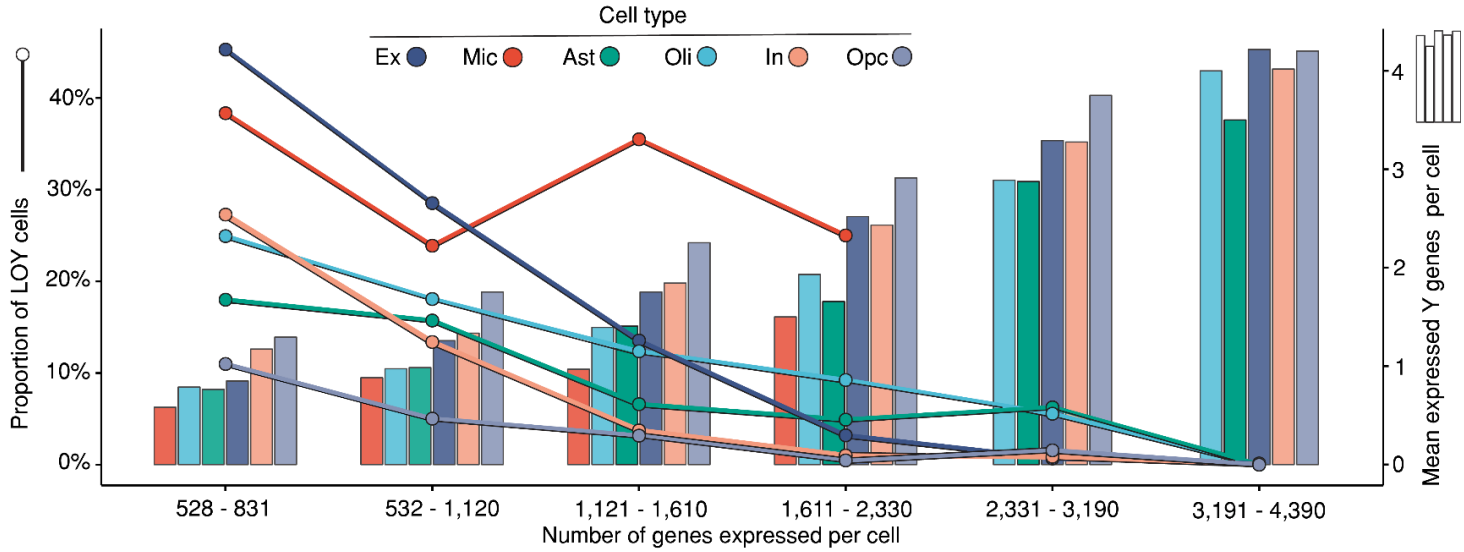


Figure 3.9 Characteristics of LOY cells in the dorsolateral prefrontal cortex across cell-type. Cells were split into quintiles based on total sequencing depth and LOY cells were compared within these ranges. **A)** Total number of cells included in each discrete grouping of cells. The 80% quintile (21,631- 41,582 reads) was used to compare LOY between cells as at this read depth stochastic Y-linked gene dropout is less likely and LOY calls are more confident. **B)** Heatmap showing LOY proportions across cell-type for each of the read depth quintiles. LOY proportions are high in all cell-types in low read depth cells. LOY proportions decrease in all cell-types as read depth increases, with the exception of microglia which maintains high LOY proportions. Glial cells (oligodendrocytes, astrocytes, microglia) all show elevated LOY rates compared to neurons and OPCs. **C)** *t*-SNE plot showing the clustering of LOY cells across the cohort for cells with read depth in the 80% quintile (21,631- 41,582 reads). LOY cells are colored white, while normal cells are colored red. There are no obvious clustering patterns of LOY cells. **D)** X read depth, Y read depth and LOY proportion of each cell-type across sequencing depth. Oligodendrocytes and astrocytes, microglia appear to have an elevated rate of LOY. **E)** Comparison of sequencing characteristics between cells assigned LOY and normal ploidy by cell type. LOY cells are similar to normal cells for most sequencing traits including expressed genes per cell, number of counts per cell, total filtered reads per cell and filtered X reads per cell. Because LOY and normal cell groups are comparable, LOY calls are more meaningful and less likely to be a result of technical differences. Oligodendrocyte (Oli), microglia (Mic), astrocyte (Ast), oligodendrocyte progenitor cell (OPC), excitatory neuron (Ex), inhibitory neuron (In).

more diverse set of genes compared to glia (**Figure 3.7E**). Furthermore, microglia express the smallest set of genes and smallest number of Y-linked genes of all cell types in the study (**Appendix 3.5**). To visualize these relationships, I plotted the proportion of loss of Y cells and average total detected Y-linked genes/cell by cell-type across increasing library complexity (total number of genes detected; **Figure 3.10**). All cells were split into 6 discrete groups based on library complexity. As expected, as general library complexity increased, the average number of detected Y genes increased, and the proportion of cells assigned LOY decreased. For example, in low library complexity excitatory neurons (528-831 detected genes), 0.91 Y genes were detected per cell and 45% were assigned LOY status. In high library complexity excitatory neurons (2,331-4,390 detected genes), >3 Y genes were detected per cell and only 21 of 5750 (0.03%) excitatory neurons were assigned LOY status. Inhibitory neurons and oligodendrocyte progenitor cells follow this same pattern. The glial cells (astrocytes, oligodendrocytes, and microglia) appear to maintain a larger proportion of LOY cells despite increased library complexity and sequencing depth. Microglia deviate from the trends seen in the other cells although it is difficult to make definitive conclusions given the low microglia sample. Nevertheless, microglia LOY rates remain high (35%; 1121-1610 detected genes) even as cell library complexity, and mean Y-linked gene detection increases.

Lastly, I wanted to compare mean depth normalized Y-linked gene expression to LOY proportion. To observe trends more easily, I split the cells into 10 equal fractions based on total read depth (**Appendix 3.9**). As a general trend, LOY proportions decrease as mean normalized Y-linked gene detection increases (**Appendix 3.9** and **Appendix 3.10**). Also, chromosome Y expression is stable and proportionally detected across astrocytes, oligodendrocytes, and microglia as cell read depth increases. For example, oligodendrocyte normalized Y gene expression ranges from 0.10 to 0.12 between 1,730 and 39,700 reads per cell. Across the same range, excitatory neurons normalized Y expression increased from 0.12 to 0.20, and inhibitory neurons increased from 0.15 to 0.40. This suggests that transcriptionally active neurons produce a greater proportion of chromosome Y gene transcripts than less active neurons, whereas glial cells tend to express chromosome Y genes at a similar rate across all transcriptional states.



Ex	383/846 (45%)	267/936 (29%)	176/1300 (14%)	66/2062 (3%)	18/2680 (1%)	4/3070 (0%)
Mic	46/120 (38%)	16/67 (24%)	11/31 (35%)	1/4 (25%)	0	0
Ast	43/239 (18%)	50/318 (16%)	26/394 (7%)	8/163 (5%)	1/16 (6%)	0/2 (0%)
Oli	432/1733 (25%)	258/1428 (18%)	82/663 (12%)	12/130 (9%)	1/18 (6%)	0/2 (0%)
In	84/308 (27%)	59/441 (13%)	26/697 (4%)	9/900 (1%)	6/691 (1%)	0/389 (0%)
Opc	19/173 (11%)	13/259 (5%)	12/379 (3%)	1/216 (0%)	1/64 (2%)	0/10 (0%)

Loss of Y across cell type and library complexity

LOY cells / total cells (LOY %)*

Figure 3.10 Proportion of loss of Y cells (LOY) and mean expressed Y-linked genes by cell type across library complexity. All cells were split into discrete groups based on library complexity (total number of genes expressed). The left axis corresponds to the line plot and displays the proportion of cells assigned LOY status by cell type. The right axis corresponds to the bar plot and displays the mean number of Y genes expressed per cell by cell type. The table below contains total cell counts and LOY counts for each cell type in each discrete library complexity group. As general library complexity increases, the average number of expressed Y genes per cell increases, and the proportion of cells assigned LOY decreases. Although sample sizes are very low, microglia appear to deviate from the trends seen in the other cell types. LOY rates remain high (35%) even as cell library complexity and mean total Y genes expressed increases. Astrocytes and oligodendrocytes also show elevate LOY proportion in cell with increased library complexity and read depth.

3.4 Chapter summary and conclusion

The available data suggests that in the aging dorsolateral prefrontal cortex, the microglia population likely harbors a higher proportion of loss of Y cells than the other cell types. However, greater microglia sample sizes at increased depth are necessary to definitively make this conclusion. Additionally, astrocytes and oligodendrocytes appear to have elevated LOY rates when compared to neurons and oligodendrocyte progenitor cells.

The discrepancy of both read depth and library complexity between diverse brain cell-types makes detecting and comparing Y loss between them difficult. Low-depth single-nucleus RNAseq further complicates the task as the Y chromosome only contains 9 regularly expressed genes in the brain. The LOY detection method conceived during this chapter must be benchmarked using alternative technologies, replicated, and tuned on additional datasets before more confident conclusions can be made. That being said, preliminary evidence suggests that if LOY is occurring in the brain, it is likely occurring in the glial cells and in the microglia specifically.

4. Discussion

4.1 Overview

Recent interest in mosaic loss of chromosome Y in aging men and its association with several negative health outcomes has increased the need for bioinformatic methods to analyze the Y chromosome.[42,135,136] In the past, the Y was commonly removed from genomic analyses for several reasons including low gene count, haploidy, lack of biological interest, short-read mapping difficulties and others.[123] As a result, methods for investigating chromosome Y specific trends using NGS have lagged behind. The main objective of this thesis was to improve methods of Y chromosome aneuploidy detection using whole genome sequencing and single-nuclei RNA sequencing. I then used these methods to provide estimates of loss of Y (LOY) in blood, and brain tissue cell-types.

Previous study suggests that LOY occurs as a result of mitotic replication errors arising through accumulated damage and dysfunction of DNA replication, maintenance, and repair pathways.[42] Therefore, it is reasonable to hypothesize that cell-types that replicate more frequently have a greater potential for mis-segregation and whole chromosome Y loss events. Although LOY rates had not previously been investigated in human brain tissue, given known cellular replenishment rates and selective clonal expansion in the blood, I hypothesized that we would observe less LOY in brain tissue than blood tissue. I further hypothesized that if LOY were occurring in brain tissue, the glial cells, and specifically the microglia[137], would likely be the most affected cell-types. Further understanding of LOY mechanisms and prevalence across human tissues is important as its detection via blood could provide a valuable biomarker of systemic genomic instability and therefore risk of developing cancer, neurodegenerative disease, and other age-related diseases. Given its high prevalence in elderly males, LOY may represent an understudied factor involved in the male immune system function, sex-specific life-span bias, and male cancer incidence bias.[130]

4.2 Summary of findings and limitations

I have split the major findings of this thesis into technical/methodological findings and biological findings. Where relevant, I have added study limitations and directions for future studies.

4.2.1 Technical and methodological

4.2.1.1 Mappability and GC content filtering improve LOY detection using next-generation sequencing data

Past efforts to call copy number alterations and large structural variations using short-read sequencing have corrected for the effect of mappability and GC content biases.[138,139] Because of PCR chemistry during the sequencing process, reads deviating from average GC content ratios are generally unrepresented in raw read counts. Additionally, regions of the reference genome that consist of non-unique and/or ambiguous sequence consistently show depth biases. Because many aneuploidy detection algorithms remove the haploid Y chromosome before analysis [102], as part of this thesis, I used concepts around bias correction from published tools and developed an in-house method for estimating Y chromosome mosaic aneuploidy using WGS data.

When analyzing mosaic aneuploidy of the Y chromosome I found that mappability filtering is of high importance (**Figure 2.6B**). The q-arm of the Y in particular consists of highly repetitive, unreliable sequence that can lead to underestimates of relative chromosomal content. Given that NGS mosaic aneuploidy detection is analyzing data for small deviations from expected in read counts from millions of cells, choosing regions of low technical variance, and high-confidence read mapping is necessary.

Despite our work here, there are several methodological steps that could be applied in the future to further improve the consistency and accuracy of mosaic aneuploidy detection of the sex chromosomes. First, projects involving the sequencing of sex chromosomes should be aligned to a sex-specific reference genome through tools

such as XYalign.[125] Male samples are aligned to the original reference genome, however females should be aligned to a reference with a hard-masked Y chromosome. By removing the Y chromosome from the reference genome when mapping female samples, far fewer mapping ambiguities arise and alignment across the X chromosome is more consistent. Further, before aligning reads from male samples, selectively hard-masking the Y chromosome reference is an important step for reducing mapping difficulties. In particular, PAR1, PAR2, XTR and other highly homologous and/or repetitive regions that introduce technical difficulties should be hard masked within the Y chromosome of the reference genome before sequencing. Filtering these regions before alignment, as opposed to after (as was done in this study), improves mapping quality, alignment scores and likely the consistency of aneuploidy detection using counts.[125] As result of data storage location and time, aligning our WGS samples was unfeasible, but future studies should pay attention to these considerations.

4.2.1.2 Single-nuclei RNAseq is not optimal for LOY detection

Single-nuclei RNA-seq (snRNA-seq) involves the isolation and sequencing of individual nuclei as opposed to whole cells. While each method has its benefits and weaknesses, studies have shown that single-nuclei expression data is similar to single-cell data, and a vast majority of transcripts are captured using both techniques.[140] In addition, snRNA-seq has broadened the capabilities of individual cell genomic research as a greater range of tissues and cell-types can be analyzed, and the variable enzymatic dissociation process common to single-cell library preparation is avoided.[140] Preserved and frozen tissues, as well dissociation adverse and sensitive cells types (eg. neurons, adipocytes) can readily be subject to snRNA-seq, options that are not available using the traditional, whole cell workflow. However, despite the advantages of snRNA-seq there are drawbacks. Mainly, the cytoplasm of the cell is discarded, 10 to 100-fold less cDNA is recovered, and less genes are detected per cell.[141] Ultimately, snRNA-seq data can be applied to a wider range of tissue samples and infers less experimental bias than scRNA-seq but this comes at the expense of reduced total sequenced cDNA, raw sequencing depth, and gene expression counts. When using the transcriptome as a proxy for DNA to detect

aneuploidy, especially that of a poorly expressed chromosome such as chromosome Y, reduced sequencing depth, increased gene dropout and an increasingly sparse expression matrix cause several aneuploidy detection problems.

Perhaps the most consequential drawback of using single-nuclei RNAseq for LOY detection is the lack of Y-linked RPS4Y1 expression (Ribosomal Protein S4 Y-Linked 1). RPS4Y1 resides in the male specific region of the Y chromosome and encodes an S4 component of the 40S ribosomal subunit. In most tissues RPS4Y1 is the highest expressed Y chromosome gene outside of the PAR region (**Appendix 3.4**).[142] Unfortunately, RPS4Y1 transcripts are primarily found in the cytoplasm and therefore single-nuclei RNA-seq struggles to capture its expression in meaningful amounts. Without RPS4Y1, the detection of chromosome Y using the transcriptome is much more dependent on sequencing depth and cell-type specific expression patterns and is increasingly prone to stochastic variation.

As a result of these single-nuclei specific sequencing challenges, I developed an alternative LOY detection method designed to improve sensitivity and overcome the expression "sparsity" problem. Instead of detecting Y chromosome presence using gene expression counts (e.g., using the standard 10x Genomics Cell Ranger algorithm), I used raw reads mapped to the reference genome and applied a custom set of strict filters based on mappability, mapping quality and alignment score. Since Cell Ranger requires mapping to both the genome and transcriptome, novel unannotated transcripts are not included in the analysis which can remove a significant amount of information. Using single-nuclei RNA data from brain tissue elevates the potential impact of these confounders as the human brain represents one of the most complex transcriptomes, and many reads are sequenced as pre-mRNA.[143] Single-nuclei transcriptomes contain a higher percentage of intergenic and intronic transcripts which are poorly annotated compared to those from exons.[143] One snRNA-seq study of the cerebellar cortex found that 7% of reads mapped to non-repetitive intergenic regions, reads that are commonly removed by Cell Ranger.[144] Evidence shows that many of these regions are enriched in conserved microRNA binding sites and may represent alternative brain-specific 3′-UTRs of known genes.[144] The complex brain transcriptome is poorly annotated which could lead to underestimates of LOY using established algorithms. By mapping RNA reads

exclusively to the genome using a splice aware aligner while loosening read filtering restrictions, more Y-linked reads are kept which appears to improve Y chromosome aneuploidy detection sensitivity.

Using my method, the number of filtered Y reads per male cell was 25.7, while females had 0.05 per cell. In comparison, male samples had 3.32 Cell Ranger expression counts from MSY genes per cell, compared to 0.08 per cell in females. This suggests that my method is improving sensitivity without allowing an abundance of false positives, artefacts, and sequencing noise. In total, across all male cells, both LOY methods agree on ~88% of LOY calls. However, one concern is LOY calls do not overwhelmingly overlap between methods. 1,026 of 2,133 (48.1%) of XY ratio LOY calls in the raw read method calls overlapped with the MSY counts method. The methods commonly agree on normal ploidy calls but vary when calling LOY cells. Although the reason for this discrepancy requires more investigation, it is likely due to i) unnecessarily strict filtering applied to exonic reads, ii) variation in per cell sequencing quality and/or, iii) expression count sparsity. Cells with poor sequencing quality are more likely to exhibit poor mapping quality and poor alignment scores which leads to an increased rate of read filtering and a higher probability of being labelled LOY. However, given that both LOY detection methods are derived from the same data, I would expect greater similarly in their conclusions. To improve the overlap between the two methods, I plan to relax exonic read filtering, and further constrain intergenic read filtering. Additionally, in the future, Cell Ranger counts and my sensitive raw read method should be combined and used simultaneously when detecting LOY. Cell Ranger provides highly confident estimates of expression but can suffer from stochastic sparsity problems when calling LOY. The raw read method detailed here likely allows more false negatives but provides and increasingly sensitive estimate of LOY. When combined they may be able to assign LOY calls more confidently, however more benchmarking is required to validate these claims.

4.2.2 Biological

4.2.2.1 Loss of Y occurs at a higher rate in the blood than in the brain.

Using whole genome sequencing (WGS) from 362 elderly male individuals (median age = 87.5) we observed LOY (defined as >10% or more of cells being affected) in 13.8% of blood samples and in 0% of both cerebellum and dorsolateral prefrontal cortex (DLPFC) samples. Additionally, we observed LOY in 16.6% of 306 SNP array blood samples. Recent studies have found similar but slightly greater LOY rates in blood using a 10% affected cell threshold, occurring in about 5-10%, 15-20% and 20-30% of aging men around 60, 70 and 80 years of age, respectively.[50,51,56,135,145] It must be noted that exact chronological age of blood samples used for our WGS and SNP-array blood cohort was unknown. For each sample, the date of study enrollment was treated as the chronological sample age under the assumption that blood was drawn near the beginning of study enrollment. However, this assumption could be leading to errors in sample age. Nevertheless, LOY estimates are broadly in accordance with previous studies in blood (**Table 2.1**). To my knowledge, no previous study has studied the brain for LOY using WGS and therefore we do not have a published benchmark to compare with. However, given low proliferation and replenishment rates in the brain, a LOY prevalence less than 10% was expected. Single-cell WGS studies in the brain have found that 0.7–2.2% of neurons show evidence of autosomal aneuploidy; however it is difficult to extrapolate this to the Y chromosome given its unique behavior in the blood.[43,44,60] Additionally, our single-nuclei sequencing in the DLPFC found that of the 22,484 male cells, 8.6% were declared LOY. However, the true LOY rate is likely lower. When analysis was confined to cells with deeper sequencing (21,631 – 41,582 reads), ~0.6% of cells were declared LOY. For comparison, in PBMCs, Thompson *et al*. labelled 15.6% of 86,160 cells as LOY, with 11.3% of B-lymphocytes showing LOY. However, this study used single-cell RNA sequencing (which included cytoplasmic transcripts) and not single-nuclei.[42]

Since few other studies have analyzed LOY in the brain using next-generation sequencing data we do not have a direct comparator. Several brain aneuploidy studies using FISH and other related methods have published estimates of aneuploidy in humans and mice.[146–149] One study investigating the aging murine brain found age-related, chromosome specific aneuploidy in chromosomes 7, 18 and Y.[146] Elderly mice (28 months) showed LOY

in ~2% of neuronal and glial cells. Interestingly, this study also found higher rates of aneuploidy in the cerebral cortex compared to cerebellum and found that non-neuronal cells were affected by age-related aneuploidy at ~5 times the rate of neuronal cells.[146] In this thesis, both of these conclusions have been replicated. FISH estimates of aneuploidy in the human brain have been notoriously variable, and commonly lack information on the Y chromosome. Most recently, Iourov *et al*. analyzed aneuploidy in chromosomes 13, 18, 21, X and Y using FISH in the cerebral cortex of seven male control patients (mean age = 24.6 ±12.9). Rates of aneuploidy were similar for each chromosome tested (~0.5%), with the exception of the Y chromosome where 0.1% of cells were aneuploid. These results suggest that Y loss is commonly acquired somatically with increasing age and not as a consequence of mitotic errors during brain development. Nevertheless, aneuploid studies using FISH are notoriously variable and larger single-cell, sequencing based aneuploidy studies are required to better elucidate true LOY rates.[105]

4.2.2.2 Age-related loss of Y in the blood and the prefrontal cortex but not cerebellum

In agreement with previous studies we found that LOY rates increased with age in the blood (**Figure 2.17**). Age is the most significant known risk factor for LOY and therefore this association was expected (and largely treated as a quality control assurance). More strikingly, we found two lines of evidence that suggest low-frequency LOY is occurring within the DLPFC at rates below our 10% LOY threshold. First, we found a significant association between LOY and age in the DLPFC (p=3.9x10^{-5}). Interestingly, evidence of age-dependent LOY in the DLPFC has also been published on one other occasion. Through the use of fluorescence qPCR, a 2018 found that LOY in the DLPFC was associated with age, albeit with a small sample size (n=26; p=0.015).[132] Secondly, we found a significant intrasample correlation between LOY rates in blood and DLPFC tissue amongst individuals with genomic data from both tissues. Individuals with severe rates of LOY in the blood (>30% cell affected) also showed reduced Y copy number in the DLPFC. This suggests Y loss is likely occurring at low-frequency in the DLPFC and furthermore, LOY in the blood could be a general biomarker of LOY in other tissues such as the brain.

In the cerebellum we did not observe age-related LOY or a relationship with intrasample LOY rates in blood, which agrees with known age-related changes between brain regions.[150] Prior study has found that cerebellar expression patterns show fewer age-related alterations compared to the cerebral cortex.[150,151] Specifically, many more genes are downregulated with age in the cerebral cortex, which could be a result of accumulated DNA damage. Secondly, a DNA methylation study using an epigenetic biomarker of aging known as "the epigenetic clock", found that the cerebellum is the "youngest" region of the brain in subjects older than 80 years of age. This trend accelerates with age. For example, cerebellar tissue of individuals from 95 to 102 years of age was approximately 10 years younger than expected, whereas frontal cortex tissue was 2 years older than expected.[152] At 110 years of age, the cerebellum was 15 years younger than expected. All together this evidence suggests the human cortex ages at a greater rate than the cerebellum and is it at higher risk of DNA damage. The biological mechanisms leading brain-region specific aging rates are not completely known but are likely a result of differing rates of metabolic activity. Irrespective of age, the cerebellum has a lower metabolic rate than the cortex in both humans and other primates.[151] Tissues with reduced metabolic activity and aerobic respiration rates are expected to produce less reactive species and therefore generally accumulate less DNA damage. In agreement, in both aging humans and mice, cerebellar mtDNA contains fewer deletions and copy number variations than the cortex.[153] Given higher rates of mutation, and early onset of epigenetic aging biomarkers one would expect age-related LOY to occur at a higher rate in the DLPFC than the cerebellum which is what our data suggests.

4.2.2.3 Astrocytes, oligodendrocytes, and especially microglia show elevated rates of Y loss compared to neurons.

Based on our snRNA-seq analysis, this study provides preliminary quantification of LOY rates across 6 cell-types in the DLPFC. To date, there is little information on mutation rates within glial cells in the brain as most previous studies have focused on neurons. Although it was difficult to disentangle cell read depth and cell-type specific Y-linked gene expression patterns from true LOY, we concluded that microglia, oligodendrocytes,

and astrocytes likely have higher LOY prevalence than neurons and oligodendrocytes progenitor cells (OPCs). Specifically, microglia appear to be most frequently affected by LOY (74 of 222 cells: 33% LOY).

In my opinion, LOY rates are overestimated when using all QC passing cells (Global LOY rate = 8.6%). For example, at the lowest read depth quartile, 25% of cells were declared LOY, while in the highest read quartile 0.04% were declared LOY. Considering previous study and known proliferation rates in the brain, 8.6% LOY is likely an overestimation. Because of variable LOY rates in all cell-types in lowly-expressed cells, I also reported a high confidence LOY from cells in the 60-80% percentile of filtered read depth (0.6% LOY; high confidence). Nevertheless, until multi-technology benchmarking is complete, LOY rates determined through our pipeline are purely estimates.

4.2.2.3.1 Microglia

We found that 33% (74 / 222) of all microglia lacked a Y chromosome at all sequencing depths, and 29% (6 / 21) were LOY in the higher confidence depth range, although low sample size and low sequencing depth limit the reliability of these conclusions. To help explain such high incidence of LOY in microglia it is useful to understand their cellular origin, progenitors, and proliferation tendencies. Microglia are the resident macrophages and main immune regulators in the brain.[154] Primarily, microglial cells find and remove damaged cells, apoptotic cells, amyloid plaques, and other cellular debris in the brain.[155] Additionally, microglia monitor neuronal health, exert neuroprotective effects, and promote the regrowth of damaged neural tissue after inflammation.[156] Microglia play important roles in brain development and homeostasis, and when compromised can be key factors in the progression of cognitive decline and neurodegenerative disease.[156] Microglia are glial cells that originate in the yolk sac during the embryonic period and populate the brain mesenchyme early in development.[156] Recent study has found that microglia self-replicate within the brain and rarely rely on replenishment from bone marrow derived monocytic precursors (although peripheral replenishment has been observed in response to severe brain injury, inflammation, and neurodegeneration).[157] In general, resting microglia are long-lived and replicate much

slower than macrophages in the vasculature. A study using 14C levels in genomic DNA found that in healthy conditions, microglia live for an average of 4.2 years, and up to 20 years, with 0.08% of the population replicating each day.[137] This was in comparison to ~10% daily turnover in peripheral monocytes, ~0.001% in oligodendrocytes, and ~0.0001% in cortical neurons.[137] Limited replenishment throughout their lifespan, and the lack of a neuronal progenitor leaves microglia susceptible to age-associated dysfunction through accumulation of DNA damage and cellular dystrophy. Despite slow replication in resting states, in response to neurological injury microglia enter a reactive state that involves a burst of mitotic activity to proportionally respond to the threat.[158] The proliferative activity of microglia often accelerates during normal CNS aging, specifically in response to elevated reactive oxygen species concentrations, physical brain injury, and neuroinflammation.[159,160] Microglial proliferative activity is coupled with subsequent apoptotic activity to return populations to homeostasis.[157] It is possible that long-lived microglial populations accumulate DNA damage that can disrupt DNA maintenance and apoptotic pathways. With increased neurodegeneration, neuroinflammation, and age-related brain injury the microglial replenishment rate increases significantly, and degraded mitotic machinery may lead to elevated incidence of LOY.[159] Microglia have been found constitute ~12% of the cells in the central nervous system and represent a significant population that could influence LOY rates found in bulk tissue.[161] Given these proliferative attributes and the conclusions of our LOY study (microglial LOY rate: 33%) I believe it is reasonable to suggest microglia are the cell type most significantly affected by elevated rates of LOY in the brain.

4.2.2.3.2 Oligodendrocytes

We found that 19.7% (785 / 3,974) of all oligodendrocytes lacked a Y chromosome at all sequencing depths, and 5% (7 / 132) lacked Y in the high-confidence depth range. Oligodendrocytes are glial cells that are responsible for axon myelination and general metabolic support for neurons. Oligodendrocytes are post-mitotic, non-proliferating cells that are produced early in development and are replenished by proliferative OPCs.[162] In response to white matter damage, OPCs proliferate, migrate to occupy the demyelinated region, differentiate into oligodendrocytes, and restore axon myelination.[163] If LOY is accumulating in oligodendrocytes, the primary

origin is likely mis-segregation events during OPC replenishment. In agreement with this theory, when compared to other glial cells OPCs have an increased susceptibility to oxidative damage which is attributed to low cellular antioxidant concentrations, and the highest iron content in the brain, which can invoke elevated free radical formation and lipid peroxidation.[164] However, oligodendrocyte replenishment rates remain low as individuals age, and without a significant selective pressure the potential for large populations of oligodendrocytes to develop LOY seems unlikely.

4.2.2.3.3 Astrocytes

We found that 11.3% (128 / 1,132) of all astrocytes lacked a Y chromosome at all sequencing depth, and 3% (3 / 98) in high-confidence depth range. Astrocytes are glial cells with a star-shaped appearance and are the most common cell-type in the CNS.[165] In addition to key structural roles in the brain, astrocytes have a number of active roles including metabolic support of neurons, neurotransmitter uptake and release, CNS repair, and maintenance of the blood brain barrier.[165] The astrocyte population is produced early in development from progenitor cells in the ventricular and subventricular zones.[166] After initial population, it is accepted that astrocytes do not commonly replicate, however, in response to brain injury and disease, astrocytes enter a reactive state known as astrogliosis.[167] In this state, astrocytes drastically change expression pattens and morphology to aid in CNS repair. Aging also appears to push astrocytes towards a reactive state phenotype, which includes an increasingly proliferative state.[168] If LOY is accumulating in astrocytes, it is most likely to occur in long-lived astrocytes with degraded DNA maintenance pathways, that convert to a reactive state with age, cognitive decline, and neurodegeneration similar to microglia. Further investigation into the association between key astrogliosis markers (e.g. GFAP) and LOY will help clarify this hypothesis.

4.2.2.4 Cellular senescence

As a result of our use of snRNA-seq to estimate LOY, cells with greater sequencing depth and transcriptional output have more expression counts and are more reliable tests for LOY. Cells with less counts often have more dropouts and express fewer genes, increasing the chance of false positive LOY calls – instances where biological cells contain a Y chromosome but sequencing results conclude LOY. Because of this, cells with reduced expression counts are often filtered, and cells with higher expression counts are retained. While this filtering strategy likely removes a large proportion of false positive LOY calls, it is also possible that LOY is concentrated in low expression count cells for biological reasons such as cellular senescence. As proliferative cells replicate across an organism's lifetime, with each passage they inherit fundamental changes that eventually accumulate and lead to a permanent state of cell cycle arrest known as cellular senescence. Senescence is a cellular program often triggered by the DNA damage response and acts as a defense mechanism that prevents cells from acquiring avoidable, potentially malignant mutations.[169] Cellular senescence is a common sign of dystrophic microglia which are associated with aging, cognitive decline, and neurodegenerative diseases.[169] Dystrophic microglia have reduced immunological function, increased cytokine production and an impaired ability to interact with neurons which can lead to elevated neuroinflammation and adverse consequences for neurons.[170] If DNA surveillance detects a Y loss event, it could trigger senescence resulting in reduced transcriptional output in some cell-types. To isolate the effect of sequencing depth and senescence on LOY, a cell-type specific senescence marker score could be given to each cell. This way any associations between them would be discovered. Ultimately, a method that determines if a cell's lack of expression counts is a result of PCR amplification variability, or lack of biological RNA would help separate these effects. Also, deeper per cell sequencing and improved library construction will be useful for more accurate estimates of LOY using scRNA-seq.

4.4 Conclusions and future directions

Across the work in this thesis we improved next-generation sequencing based Y chromosome mosaic aneuploidy detection methods and used them to further the understanding of LOY in the brain. I concluded that LOY occurs at a greater rate in blood than the brain and that LOY increases with age in the dorsolateral prefrontal cortex but not the cerebellum. I also provided an estimate of LOY prevalence in the DLPFC, and provided preliminary evidence that microglia, astrocytes, and oligodendrocytes are the most likely cell-types to accumulate LOY in the cortex. In the future, to improve confidence in our LOY estimates across brain cell-types, our method needs to be subject to replication across larger, deeper sequenced, single-cell datasets and compared to samples with known LOY rates.

Another contribution of our work is in identifying challenges in estimation of LOY with current data types. Future studies seeking to estimate LOY using single-cell RNAseq can hopefully be streamlined and optimized in the following ways:

i) The use of whole cell samples opposed to single-nuclei samples when performing scRNAseq. Cytosolic RPS4Y1 expression is not captured in the nuclei and as a result LOY estimates are susceptible to low-read depth and false positive LOY calls.

ii) Illumina Smart-seq2 plate-based library preparation instead of 10x Genomics Chromium droplet-based preparation. Smart-seq2 captures more genes per cell, less dropouts, and reduced variability amongst lowly expressed genes at the expense of total cells sequenced.[171] Smart-seq2 expression data is likely more compatible with accurate LOY detection but a direct comparison is required.

4.4.1 Future directions

In my opinion, loss of Y has several important implications on male health, many of which await discovery. The recent finding that LOY is detectable in ~40% of males greater than 70 years of age shines light on

the prevalence and magnitude of this biological phenomenon.[42] Below I have synthesized my understanding of LOY and the technologies used to detect it to produce intuitions for future LOY studies.

4.4.1.1 Benchmark scRNA-seq LOY accuracy using tissues with known LOY rates.

Previously research has yet to test the accuracy of scRNA-seq for quantifying LOY. As has been done with SNP array, WGS and qPCR methods of detecting Y loss, scRNA-seq LOY detection could be applied to samples with known rates of LOY.[131] An ideal study design would involve PBMC extraction from individuals with previously determined LOY prevalence. Samples could be subject to a combination of SNP array, WGS, qPCR, and single-cell RNAseq. The concordance between the three former methods of LOY detection (eg. SNP array, WGS, qPCR) has been established, and this study would confirm the accuracy of single-cell transcriptomics for LOY estimation compared to the known, reliable methods. A study benchmarking scRNA-seq-based LOY detection was recently released as a pre-print.[172]

4.4.1.2 Single-cell WGS and G&T-seq

To my knowledge, very few studies of glial somatic mosaicism have been published to date. Given how long glial cells live and replenish themselves, combined with their complicated roles in neurodegeneration, inflammation, and aging, I believe a further understanding of somatic mutation in glial cells would benefit the scientific community. A single-cell WGS study using FACS sorted astrocytes, microglia and oligodendrocyte progenitor cells could shine light on both LOY and general mosaicism in glial populations.

The use of genome and transcriptome sequencing (G&T seq) has previously been used to benchmark aneuploidy detection using single-cell transcriptomics.[114] Simply, cells are isolated, lysed, and single-cell RNA and DNA are separated and sequenced independently. Although, physical separation of DNA and RNA increases the risk of contamination, with improvements to DNA amplification G&T-seq allows for direct, within-cell

detection of the genome and transcriptome. Such an assay would allow for benchmarking of single-cell RNA LOY detection accuracy and would have the additional benefit of accurately characterizing the transcriptome of LOY cells. It would also provide answers for the cellular senescence problem highlighted above. The total transcriptional variance of LOY cells (labelled through reliable DNA-based methods) could be compared to cells of expected ploidy.

4.4.1.3 Repeat analysis across a variety of other cell types

If single-cell RNAseq is confirmed as a reliable technology for detecting Y loss, a wider range of human tissues could be surveyed with confidence. A major benefit of using scRNAseq is that cell-type can be confirmed post-hoc using expression markers. Currently, only the blood[42,50,56], bone marrow[12], buccal cells[135], various tumors[173] and now the brain[174] have been investigated across large cohorts for population estimates of LOY. Given that LOY mechanisms involve DNA maintenance degradation and cellular replication it seems unlikely that LOY is a phenomenon isolated to leukocytes. LOY analyses on highly proliferative colonic crypts and small intestinal epithelium would be interesting considering the association between the development of colorectal cancer and LOY in the blood. These are some of the most proliferative tissues in the human body and observing definitive, age-associated LOY in them would expand the scope of Y chromosome loss as a human disease process.

In my opinion the most logical future direction is to investigate LOY more rigorously in the microglia, and other glial cells. Because of technical reasons, the data used in this thesis did not confidently determine the burden of LOY in the microglia. Repeating our analysis on isolated, reasonably deep sequenced microglia would more definitively answer this question and provide evidence for another factor that may affect neurodegerative disease processes. Recently, isolated microglia from the DLPFC of elderly individuals was subjected to 10x Genomics droplet-based RNA-seq. As a next step I plan to repeat these analyzes on the isolated microglia data to further elucidate LOY rates and find differentially expressed genes between normal and LOY microglial cells.

References

1. Milo, R. & Phillips, R. *Cell Biology by the Numbers*. (Garland Science, 2015).
2. Martincorena, I. Somatic mutation and clonal expansions in human tissues. *Genome Med.* **11**, (2019).
3. Jaiswal, S. *et al.* Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
4. Park, J. S. *et al.* Brain somatic mutations observed in Alzheimer's disease associated with aging and dysregulation of tau phosphorylation. *Nat. Commun.* **10**, 1–12 (2019).
5. Jaiswal, S. *et al.* Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. *N. Engl. J. Med.* **377**, 111–121 (2017).
6. Brown, T. A. *The Human Genome*. (Wiley-Liss, 2002).
7. Pierre, R. V. & Hoagland, H. C. Age-associated aneuploidy: Loss of Y chromosome from human bone marrow cells with aging. *Cancer* **30**, 889–894 (1972).
8. Hall, J. G. Review and hypotheses: somatic mosaicism: observations related to clinical genetics. *Am. J. Hum. Genet.* **43**, 355–363 (1988).
9. Davis, D. G. & Shaw, M. W. An Unusual Human Mosaic for Skin Pigmentation. *N. Engl. J. Med.* **270**, 1384–1389 (1964).
10. Biesecker, L. G. & Spinner, N. B. A genomic view of mosaicism and human disease. *Nat. Rev. Genet.* **14**, 307–320 (2013).
11. Jacobs, P. A., Brunton, M., Court Brown, W. M., Doll, R. & Goldstein, H. Change of Human Chromosome Count Distributions with Age: Evidence for a Sex Difference. *Nature* **197**, 1080–1081 (1963).
12. O'Riordan, M. L., Berry, E. W. & Tough, I. M. Chromosome studies on bone marrow from a male control population. *Br. J. Haematol.* **19**, 83–90 (1970).

13. Delhanty, J. D. A. *et al.* Detection of aneuploidy and chromosomal mosaicism in human embryos during preimplantation sex determination by fluorescent *in situ* hybridisation, (FISH). *Hum. Mol. Genet.* **2**, 1183–1185 (1993).
14. Mukherjee, A. B., Alejandro, J., Payne, S. & Thomas, S. Age-related aneuploidy analysis of human blood cells in vivo by fluorescence in situ hybridization (FISH). *Mech. Ageing Dev.* **90**, 145–156 (1996).
15. Martincorena, I. *et al.* Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
16. Yizhak, K. *et al.* RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science* **364**, (2019).
17. Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).
18. Krimmel, J. D. *et al.* Ultra-deep sequencing detects ovarian cancer cells in peritoneal fluid and reveals somatic TP53 mutations in noncancerous tissues. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 6005–6010 (2016).
19. Abyzov, A. *et al.* One thousand somatic SNVs per skin fibroblast cell set baseline of mosaic mutational load with patterns that suggest proliferative origin. *Genome Res.* **27**, 512–523 (2017).
20. Morimoto, Y. *et al.* Deep sequencing reveals variations in somatic cell mosaic mutations between monozygotic twins with discordant psychiatric disease. *Hum. Genome Var.* **4**, 1–6 (2017).
21. Rohrback, S. *et al.* Submegabase copy number variations arise during cerebral cortical neurogenesis as revealed by single-cell whole-genome sequencing. *Proc. Natl. Acad. Sci.* **115**, 10804–10809 (2018).
22. McConnell, M. J. *et al.* Mosaic Copy Number Variation in Human Neurons. *Science* **342**, 632–637 (2013).
23. Zhang, L. *et al.* Single-cell whole-genome sequencing reveals the functional landscape of somatic mutations in B lymphocytes across the human lifespan. *Proc. Natl. Acad. Sci.* **116**, 9014–9019 (2019).
24. Forsberg, L. A., Gisselsson, D. & Dumanski, J. P. Mosaicism in health and disease — clones picking up speed. *Nat. Rev. Genet.* **18**, 128–142 (2017).

25. Loh, P.-R. *et al.* Insights about clonal hematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* **559**, 350–355 (2018).

26. Andriani, G. A., Vijg, J. & Montagna, C. Mechanisms and consequences of aneuploidy and chromosome instability in the aging brain. *Mech. Ageing Dev.* **161**, 19–36 (2017).

27. Torres, E. M., Williams, B. R. & Amon, A. Aneuploidy: Cells Losing Their Balance. *Genetics* **179**, 737–746 (2008).

28. Jia, C.-W. *et al.* Aneuploidy in Early Miscarriage and its Related Factors. *Chin. Med. J. (Engl.)* **128**, 2772–2776 (2015).

29. Duijf, P. H. G., Schultz, N. & Benezra, R. Cancer cells preferentially lose small chromosomes. *Int. J. Cancer* **132**, 2316–2326 (2013).

30. Nagaoka, S. I., Hassold, T. J. & Hunt, P. A. Human aneuploidy: mechanisms and new insights into an age-old problem. *Nat. Rev. Genet.* **13**, 493–504 (2012).

31. Sheltzer, J. M. *et al.* Aneuploidy Drives Genomic Instability in Yeast. *Science* **333**, 1026–1030 (2011).

32. Zhu, J., Tsai, H.-J., Gordon, M. R. & Li, R. Cellular Stress Associated with Aneuploidy. *Dev. Cell* **44**, 420–431 (2018).

33. Pfau, S. J. & Amon, A. Chromosomal instability and aneuploidy in cancer: from yeast to man. *EMBO Rep.* **13**, 515–527 (2012).

34. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The Hallmarks of Aging. *Cell* **153**, 1194–1217 (2013).

35. Shepherd, C. E., Yang, Y. & Halliday, G. M. Region- and Cell-specific Aneuploidy in Brain Aging and Neurodegeneration. *Neuroscience* **374**, 326–334 (2018).

36. Hanks, S. *et al.* Constitutional aneuploidy and cancer predisposition caused by biallelic mutations in BUB1B. *Nat. Genet.* **36**, 1159–1161 (2004).

37. Baker, D. J. *et al.* BubR1 insufficiency causes early onset of aging-associated phenotypes and infertility in mice. *Nat. Genet.* **36**, 744–749 (2004).

38. Baker, D. J. *et al.* Increased expression of BubR1 protects against aneuploidy and cancer and extends healthy lifespan. *Nat. Cell Biol.* **15**, 96–102 (2013).

39. Thomas, P. & Fenech, M. Chromosome 17 and 21 aneuploidy in buccal cells is increased with ageing and in Alzheimer's disease. *Mutagenesis* **23**, 57–65 (2008).

40. Nowinski, G. P. *et al.* The frequency of aneuploidy in cultured lymphocytes is correlated with age and gender but not with reproductive history. *Am. J. Hum. Genet.* **46**, 1101–1111 (1990).

41. Mukherjee, A. B. & Thomas, S. A longitudinal study of human age-related chromosomal analysis in skin fibroblasts. *Exp. Cell Res.* **235**, 161–169 (1997).

42. Thompson, D. J. *et al.* Genetic predisposition to mosaic Y chromosome loss in blood. *Nature* **575**, 652–657 (2019).

43. van den Bos, H. *et al.* Single-cell whole genome sequencing reveals no evidence for common aneuploidy in normal and Alzheimer's disease neurons. *Genome Biol.* **17**, 116 (2016).

44. Knouse, K. A., Wu, J., Whittaker, C. A. & Amon, A. Single cell sequencing reveals low levels of aneuploidy across mammalian tissues. *Proc. Natl. Acad. Sci.* **111**, 13409–13414 (2014).

45. Pinho, S. & Frenette, P. S. Haematopoietic stem cell activity and interactions with the niche. *Nat. Rev. Mol. Cell Biol.* **20**, 303–320 (2019).

46. Forsberg, L. A. *et al.* Age-related somatic structural changes in the nuclear genome of human blood cells. *Am. J. Hum. Genet.* **90**, 217–228 (2012).

47. Osorio, F. G. *et al.* Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. *Cell Rep.* **25**, 2308-2316.e4 (2018).

48. Jacobs, K. B. *et al.* Detectable clonal mosaicism and its relationship to aging and cancer. *Nat. Genet.* **44**, 651–658 (2012).

49. Machiela, M. J. *et al.* Female chromosome X mosaicism is age-related and preferentially affects the inactivated X chromosome. *Nat. Commun.* **7**, 11843 (2016).

50. Dumanski, J. P. *et al.* Smoking is associated with mosaic loss of chromosome Y. *Science* **347**, 81–83 (2015).

51. Dumanski, J. P. *et al.* Mosaic Loss of Chromosome Y in Blood Is Associated with Alzheimer Disease. *Am. J. Hum. Genet.* **98**, 1208–1219 (2016).

52. Forsberg, L. A. Loss of chromosome Y (LOY) in blood cells is associated with increased risk for disease and mortality in aging men. *Hum. Genet.* **136**, 657–663 (2017).

53. Grassmann, F. *et al.* Y chromosome mosaicism is associated with age-related macular degeneration. *Eur. J. Hum. Genet.* **27**, 36–41 (2019).

54. Lleo, A. *et al.* Y chromosome loss in male patients with primary biliary cirrhosis. *J. Autoimmun.* **41**, 87–91 (2013).

55. Persani, L. *et al.* Increased loss of the Y chromosome in peripheral blood cells in male patients with autoimmune thyroiditis. *J. Autoimmun.* **38**, J193–J196 (2012).

56. Loftfield, E. *et al.* Predictors of mosaic chromosome Y loss and associations with mortality in the UK Biobank. *Sci. Rep.* **8**, 12316 (2018).

57. Forsberg, L. A. *et al.* Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. *Nat. Genet.* **46**, 624–628 (2014).

58. Dorak, M. T. & Karpuzoglu, E. Gender Differences in Cancer Susceptibility: An Inadequately Addressed Issue. *Front. Genet.* **3**, (2012).

59. Lodato, M. A. *et al.* Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359**, 555–559 (2018).

60. Cai, X. *et al.* Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. *Cell Rep.* **8**, 1280–1289 (2014).

61. McConnell, M. J. *et al.* Intersection of diverse neuronal genomes and neuropsychiatric disease: The Brain Somatic Mosaicism Network. *Science* **356**, (2017).

62. Maan, A. A. *et al.* The Y chromosome: a blueprint for men's health? *Eur. J. Hum. Genet.* **25**, 1181–1188 (2017).

63. Davoli, T. *et al.* Cumulative Haploinsufficiency and Triplosensitivity Drive Aneuploidy Patterns to Shape the Cancer Genome. *Cell* **155**, 948–962 (2013).

64. Dumanski, J. P. *et al. Immune cells lacking Y chromosome have widespread dysregulation of autosomal genes*. http://biorxiv.org/lookup/doi/10.1101/673459 (2019) doi:10.1101/673459.

65. Nathanson, K. L. *et al.* The Y Deletion gr/gr and Susceptibility to Testicular Germ Cell Tumor. *Am. J. Hum. Genet.* **77**, 1034–1043 (2005).

66. Arnold, A. P. Y chromosome's roles in sex differences in disease. *Proc. Natl. Acad. Sci.* **114**, 3787–3789 (2017).

67. Kido, T. & Lau, Y.-F. C. Roles of the Y chromosome genes in human cancers. *Asian J. Androl.* **17**, 373–380 (2015).

68. Bianchi, N. O. Y chromosome structural and functional changes in human malignant diseases. *Mutat. Res. Mutat. Res.* **682**, 21–27 (2009).

69. Wright, D. J. *et al.* Genetic variants associated with mosaic Y chromosome loss highlight cell cycle genes and overlap with cancer susceptibility. *Nat. Genet.* **49**, 674–679 (2017).

70. Titus, S. *et al.* Impairment of BRCA1-related DNA double-strand break repair leads to ovarian aging in mice and humans. *Sci. Transl. Med.* **5**, 172ra21 (2013).

71. Bachtrog, D. Y chromosome evolution: emerging insights into processes of Y chromosome degeneration. *Nat. Rev. Genet.* **14**, 113–124 (2013).

72. Jobling, M. A. & Tyler-Smith, C. Human Y-chromosome variation in the genome-sequencing era. *Nat. Rev. Genet.* **18**, 485–497 (2017).

73. Skaletsky, H. *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825–837 (2003).

74. Yadav, S. K., Kumari, A., Javed, S. & Ali, S. DYZ1 arrays show sequence variation between the monozygotic males. *BMC Genet.* **15**, 19 (2014).

75. Ross, M. T. *et al.* The DNA sequence of the human X chromosome. *Nature* **434**, 325–337 (2005).

76. Veerappa, A. M., Padakannaya, P. & Ramachandra, N. B. Copy number variation-based polymorphism in a new pseudoautosomal region 3 (PAR3) of a human X-chromosome-transposed region (XTR) in the Y chromosome. *Funct. Integr. Genomics* **13**, 285–293 (2013).

77. Jarvik, L. F. & Kato, T. Chromosome examinations in aged twins. *Am. J. Hum. Genet.* **22**, 562–573 (1970).

78. Walker, L. M. S. The Chromosomes of Bone-Marrow Cells of Haematologically Normal Men and Women. *Br. J. Haematol.* **21**, 455–461 (1971).

79. Lawler, S. D., Lobb, D. S. & Wiltshaw, E. Philadelphia-chromosome positive bone-marrow cells showing loss of the Y in males with chronic myeloid leukaemia. *Br. J. Haematol.* **27**, 247–252 (1974).

80. Shiffman, N. J., Stecker, E., Conen, P. E. & Gardner, H. A. Males with chronic myeloid leukemia and the 45, XO, Ph1 chromosome pattern. *Can. Med. Assoc. J.* **110**, 1151–1154 (1974).

81. Sandberg, A. A. & Sakurai, M. The missing Y chromosome and human leukaemia. *Lancet Lond. Engl.* **1**, 375 (1973).

82. Rowley, J. D. Nonrandom chromosomal abnormalities in hematologic disorders of man. *Proc. Natl. Acad. Sci. U. S. A.* **72**, 152–156 (1975).

83. Sandberg, A. A., Sakurai, M. & Holdsworth, R. N. Chromosomes and causation of human cancer and leukemia. VIII. DMS chromosomes in a neuroblastoma. *Cancer* **29**, 1671–1679 (1972).

84. Padre-Mendoza, T., Farnes, P., Barker, B. E., Smith, P. S. & Forman, E. N. Y Chromosome Loss in Childhood Leukaemias. *Br. J. Haematol.* **41**, 43–48 (1979).

85. Bates, S. E. Classical Cytogenetics: Karyotyping Techniques. in *Human Pluripotent Stem Cells: Methods and Protocols* (eds. Schwartz, P. H. & Wesselschmidt, R. L.) 177–190 (Humana Press, 2011). doi:10.1007/978-1-61779-201-4_13.

86. Sheltzer, J. M. & Amon, A. The Aneuploidy Paradox: Costs and Benefits of an Incorrect Karyotype. *Trends Genet. TIG* **27**, 446–453 (2011).

87. Huber, D., Voith von Voithenberg, L. & Kaigala, G. V. Fluorescence in situ hybridization (FISH): History, limitations and what to expect from micro-scale FISH? *Micro Nano Eng.* **1**, 15–24 (2018).

88. Loss of the Y chromosome from normal and neoplastic bone marrows. United Kingdom Cancer Cytogenetics Group (UKCCG). *Genes. Chromosomes Cancer* **5**, 83–88 (1992).

89. Hu, D. G., Webb, G. & Hussey, N. Aneuploidy detection in single cells using DNA array-based comparative genomic hybridization. *Mol. Hum. Reprod.* **10**, 283–289 (2004).

90. Quintana-Murci, L. & Fellous, M. The human Y chromosome: the biological role of a "functional wasteland". *J. Biomed. Biotechnol.* **1**, 18–24 (2001).

91. Bejjani, B. A. & Shaffer, L. G. Application of Array-Based Comparative Genomic Hybridization to Clinical Diagnostics. *J. Mol. Diagn. JMD* **8**, 528–533 (2006).

92. Xavier, R. J. & Rioux, J. D. Genome-wide association studies: a new window into immune-mediated diseases. *Nat. Rev. Immunol.* **8**, 631–643 (2008).

93. Yau, C. & Holmes, C. C. CNV discovery using SNP genotyping arrays. *Cytogenet. Genome Res.* **123**, 307–312 (2008).

94. Colella, S. *et al.* QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res.* **35**, 2013–2025 (2007).

95. Wang, K. *et al.* PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–1674 (2007).

96. Wong, J. Y. Y. *et al.* Outdoor air pollution and mosaic loss of chromosome Y in older men from the Cardiovascular Health Study. *Environ. Int.* **116**, 239–247 (2018).

97. Noveski, P. *et al.* Loss of Y Chromosome in Peripheral Blood of Colorectal and Prostate Cancer Patients. *PLoS ONE* **11**, (2016).

98. Haitjema, S. *et al.* Loss of Y Chromosome in Blood Is Associated With Major Cardiovascular Events During Follow-Up in Men After Carotid Endarterectomy. *Circ. Cardiovasc. Genet.* (2017) doi:10.1161/CIRCGENETICS.116.001544.

99. Knetsch, C. W., van der Veer, E. M., Henkel, C. & Taschner, P. DNA Sequencing. in *Molecular Diagnostics* (eds. van Pelt-Verkuil, E., van Leeuwen, W. B. & te Witt, R.) 339–360 (Springer Singapore, 2019). doi:10.1007/978-981-13-1604-3_8.

100. Kosugi, S. *et al.* Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol.* **20**, 117 (2019).

101. Yau, C. OncoSNP-SEQ: a statistical approach for the identification of somatic copy number alterations from next-generation sequencing of cancer genomes. *Bioinforma. Oxf. Engl.* **29**, 2482–2484 (2013).

102. Boeva, V. *et al.* Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* **28**, 423–425 (2012).

103. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. - PubMed - NCBI. https://www.ncbi.nlm.nih.gov/pubmed/21324876.

104. Kong, Y. *et al.* Mosaic Chromosomal Aneuploidy Detection By Sequencing (MAD-seq). *bioRxiv* (2017) doi:10.1101/142299.

105. Andriani, G. A. *et al.* A direct comparison of interphase FISH versus low-coverage single cell sequencing to detect aneuploidy reveals respective strengths and weaknesses. *Sci. Rep.* **9**, 1–7 (2019).

106. Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* **40**, e72–e72 (2012).

107. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci. Rep.* **9**, 1–5 (2019).

108. Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.* **17**, 175–188 (2016).

109. Bäumer, C., Fisch, E., Wedler, H., Reinecke, F. & Korfhage, C. Exploring DNA quality of single cells for genome analysis with simultaneous whole-genome amplification. *Sci. Rep.* **8**, 1–10 (2018).

110. Bakker, B., van den Bos, H., Lansdorp, P. M. & Foijer, F. How to count chromosomes in a cell: An overview of current and novel technologies. *BioEssays News Rev. Mol. Cell. Dev. Biol.* **37**, 570–577 (2015).

111. Hou, Y. *et al.* Genome analyses of single human oocytes. *Cell* **155**, 1492–1506 (2013).

112. Lu, S. *et al.* Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing. *Science* **338**, 1627–1630 (2012).

113. Evrony, G. D. *et al.* Single-Neuron Sequencing Analysis of L1 Retrotransposition and Somatic Mutation in the Human Brain. *Cell* **151**, 483–496 (2012).

114. Griffiths, J. A., Scialdone, A. & Marioni, J. C. Mosaic autosomal aneuploidies are detectable from single-cell RNAseq data. *BMC Genomics* **18**, (2017).

115. Weissbein, U., Schachter, M., Egli, D. & Benvenisty, N. Analysis of chromosomal aberrations and recombination by allelic bias in RNA-Seq. *Nat. Commun.* **7**, (2016).

116. Fan, X. *et al.* Spatial transcriptomic survey of human embryonic cerebral cortex by single-cell RNA-seq analysis. *Cell Res.* **28**, 730–745 (2018).

117. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).

118. Starostik, M. R., Sosina, O. A. & McCoy, R. C. Single-cell analysis of human embryos reveals diverse patterns of aneuploidy and mosaicism. *bioRxiv* 2020.01.06.894287 (2020) doi:10.1101/2020.01.06.894287.

119. Nguyen, Q. H., Pervolarakis, N., Nee, K. & Kessenbrock, K. Experimental Considerations for Single-Cell RNA Sequencing Approaches. *Front. Cell Dev. Biol.* **6**, (2018).

120. Rizzetto, S. *et al. Impact of sequencing depth and read length on single cell RNA sequencing data: lessons from T cells*. http://biorxiv.org/lookup/doi/10.1101/134130 (2017) doi:10.1101/134130.

121. Schwarze, K., Buchanan, J., Taylor, J. C. & Wordsworth, S. Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genet. Med.* **20**, 1122–1130 (2018).

122. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* **13**, 36–46 (2011).

123. Anderson, K., Cañadas-Garre, M., Chambers, R., Maxwell, A. P. & McKnight, A. J. The Challenges of Chromosome Y Analysis and the Implications for Chronic Kidney Disease. *Front. Genet.* **10**, 781 (2019).

124. Alvarez-Cubero, M. J. *et al.* Methodology for Y Chromosome Capture: A complete genome sequence of Y chromosome using flow cytometry, laser microdissection and magnetic streptavidin-beads. *Sci. Rep.* **8**, (2018).

125. Webster, T. H. *et al.* Identifying, understanding, and correcting technical artifacts on the sex chromosomes in next-generation sequencing data. *GigaScience* **8**, (2019).

126. Marco-Sola, S., Sammeth, M., Guigó, R. & Ribeca, P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods* **9**, 1185–1188 (2012).

127. Bennett, D. A. *et al.* Religious Orders Study and Rush Memory and Aging Project. *J. Alzheimers Dis. JAD* **64**, S161–S189 (2018).

128. De Jager, P. L. *et al.* A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. *Sci. Data* **5**, 1–13 (2018).

129. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

130. Forsberg, L. A. *et al.* Mosaic loss of chromosome Y in leukocytes matters. *Nat. Genet.* **51**, 4–7 (2019).

131. Danielsson, M. *et al.* Longitudinal changes in the frequency of mosaic chromosome Y loss in peripheral blood cells of aging men varies profoundly between individuals. *Eur. J. Hum. Genet.* **28**, 349–357 (2020).

132. Kimura, A. *et al.* Loss of chromosome Y in blood, but not in brain, of suicide completers. *PloS One* **13**, e0190667 (2018).

133. Swiech, L. *et al.* In vivo interrogation of gene function in the mammalian brain using CRISPR-Cas9. *Nat. Biotechnol.* **33**, 102–106 (2015).

134. Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).

135. Zhou, W. *et al.* Mosaic loss of chromosome Y is associated with common variation near TCL1A. *Nat. Genet.* **48**, 563–568 (2016).

136. Forsberg, L. A. Back to the drawing board—loss of chromosome Y (LOY) in leukocytes is associated with age-related macular degeneration. *Eur. J. Hum. Genet.* **27**, 17 (2019).

137. Réu, P. *et al.* The Lifespan and Turnover of Microglia in the Human Brain. *Cell Rep.* **20**, 779–784 (2017).

138. Eisfeldt, J., Nilsson, D., Andersson-Assarsson, J. C. & Lindstrand, A. AMYCNE: Confident copy number assessment using whole genome sequencing data. *PLOS ONE* **13**, e0189710 (2018).

139. Jiang, Y. *et al.* CODEX2: full-spectrum copy number variation detection by high-throughput DNA sequencing. *Genome Biol.* **19**, 202 (2018).

140. Ding, J. *et al.* Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.* 1–10 (2020) doi:10.1038/s41587-020-0465-8.

141. Bakken, T. E. *et al.* Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. *PLOS ONE* **13**, e0209648 (2018).

142. Vawter, M. P. *et al.* Gender-specific gene expression in post-mortem human brain: localization to sex chromosomes. *Neuropsychopharmacol. Off. Publ. Am. Coll. Neuropsychopharmacol.* **29**, 373–384 (2004).

143. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

144. Guohua Xu, A. *et al.* Intergenic and Repeat Transcription in Human, Chimpanzee and Macaque Brains Measured by RNA-Seq. *PLoS Comput. Biol.* **6**, (2010).

145. Forsberg, L. A. *et al.* Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. *Nat. Genet.* **46**, 624–628 (2014).

146. Faggioli, F., Wang, T., Vijg, J. & Montagna, C. Chromosome-specific accumulation of aneuploidy in the aging mouse brain. *Hum. Mol. Genet.* **21**, 5246–5253 (2012).

147. Iourov, I. Y., Vorsanova, S. G., Liehr, T. & Yurov, Y. B. Aneuploidy in the normal, Alzheimer's disease and ataxia-telangiectasia brain: Differential expression and pathological meaning. *Neurobiol. Dis.* **34**, 212–220 (2009).

148. Westra, J. W. *et al.* Aneuploid mosaicism in the developing and adult cerebellar cortex. *J. Comp. Neurol.* **507**, 1944–1951 (2008).

149. Schad, C. R. *et al.* Application of fluorescent in situ hybridization with X and Y chromosome specific probes to buccal smear analysis. *Am. J. Med. Genet.* **66**, 187–192 (1996).

150. Kumar, A. *et al.* Age-associated changes in gene expression in human brain and isolated neurons. *Neurobiol. Aging* **34**, 1199–1209 (2013).

151. Fraser, H. B., Khaitovich, P., Plotkin, J. B., Pääbo, S. & Eisen, M. B. Aging and Gene Expression in the Primate Brain. *PLOS Biol.* **3**, e274 (2005).

152. Horvath, S. *et al.* The cerebellum ages slowly according to the epigenetic clock. *Aging* **7**, 294–306 (2015).

153. Kazachkova, N., Ramos, A., Santos, C. & Lima, M. Mitochondrial DNA Damage Patterns and Aging: Revising the Evidences for Humans and Mice. *Aging Dis.* **4**, 337–350 (2013).

154. Li, Q. & Barres, B. A. Microglia and macrophages in brain homeostasis and disease. *Nat. Rev. Immunol.* **18**, 225–242 (2018).

155. Neumann, H., Kotter, M. R. & Franklin, R. J. M. Debris clearance by microglia: an essential link between degeneration and regeneration. *Brain* **132**, 288–295 (2009).

156. Bachiller, S. *et al.* Microglia in Neurological Diseases: A Road Map to Brain-Disease Dependent-Inflammatory Response. *Front. Cell. Neurosci.* **12**, (2018).

157. Huang, Y. *et al.* Repopulated microglia are solely derived from the proliferation of residual microglia after acute depletion. *Nat. Neurosci.* **21**, 530–540 (2018).

158. Remington, L. T., Babcock, A. A., Zehntner, S. P. & Owens, T. Microglial Recruitment, Activation, and Proliferation in Response to Primary Demyelination. *Am. J. Pathol.* **170**, 1713–1724 (2007).

159. Gomez-Nicola, D. & Perry, V. H. Microglial dynamics and role in the healthy and diseased brain: a paradigm of functional plasticity. *Neurosci. Rev. J. Bringing Neurobiol. Neurol. Psychiatry* **21**, 169–184 (2015).

160. Kamphuis, W., Orre, M., Kooijman, L., Dahmen, M. & Hol, E. M. Differential cell proliferation in the cortex of the APPswePS1dE9 Alzheimer's disease mouse model. *Glia* **60**, 615–629 (2012).

161. Lawson, L. J., Perry, V. H., Dri, P. & Gordon, S. Heterogeneity in the distribution and morphology of microglia in the normal adult mouse brain. *Neuroscience* **39**, 151–170 (1990).

162. Bergles, D. E. & Richardson, W. D. Oligodendrocyte Development and Plasticity. *Cold Spring Harb. Perspect. Biol.* **8**, (2016).

163. Maki, T., Liang, A. C., Miyamoto, N., Lo, E. H. & Arai, K. Mechanisms of oligodendrocyte regeneration from ventricular-subventricular zone-derived progenitor cells in white matter diseases. *Front. Cell. Neurosci.* **7**, (2013).

164. Giacci, M. K. *et al.* Oligodendroglia Are Particularly Vulnerable to Oxidative Damage after Neurotrauma In Vivo. *J. Neurosci.* **38**, 6491–6504 (2018).

165. Miller, S. J. Astrocyte Heterogeneity in the Adult Central Nervous System. *Front. Cell. Neurosci.* **12**, (2018).

166. Guizzetti, M., Kavanagh, T. J. & Costa, L. G. Measurements of astrocyte proliferation. *Methods Mol. Biol. Clifton NJ* **758**, 349–359 (2011).

167. Li, K., Li, J., Zheng, J. & Qin, S. Reactive Astrocytes in Neurodegenerative Diseases. *Aging Dis.* **10**, 664–675 (2019).

168. Palmer, A. L. & Ousman, S. S. Astrocytes and Aging. *Front. Aging Neurosci.* **10**, (2018).

169. Martínez-Cué, C. & Rueda, N. Cellular Senescence in Neurodegenerative Diseases. *Front. Cell. Neurosci.* **14**, (2020).

170. Streit, W. J., Khoshbouei, H. & Bechmann, I. Dystrophic microglia in late-onset Alzheimer's disease. *Glia* **68**, 845–854 (2020).

171. Wang, X., He, Y., Zhang, Q., Ren, X. & Zhang, Z. Direct Comparative Analysis of 10X Genomics Chromium and Smart-seq2. *bioRxiv* 615013 (2019) doi:10.1101/615013.

172. Dumanski, J. P. *et al.* Loss of Y in leukocytes, dysregulation of autosomal immune genes and disease risks. *bioRxiv* 673459 (2019) doi:10.1101/673459.

173. Bianchi, N. O. Y chromosome structural and functional changes in human malignant diseases. *Mutat. Res. Mutat. Res.* **682**, 21–27 (2009).

174. Graham, E. J. *et al.* Somatic mosaicism of sex chromosomes in the blood and brain. *Brain Res.* 146345 (2019) doi:10.1016/j.brainres.2019.146345.

Appendix 1
ROSMAP library preparation and sequencing methods

WGS library preparation and sequencing.

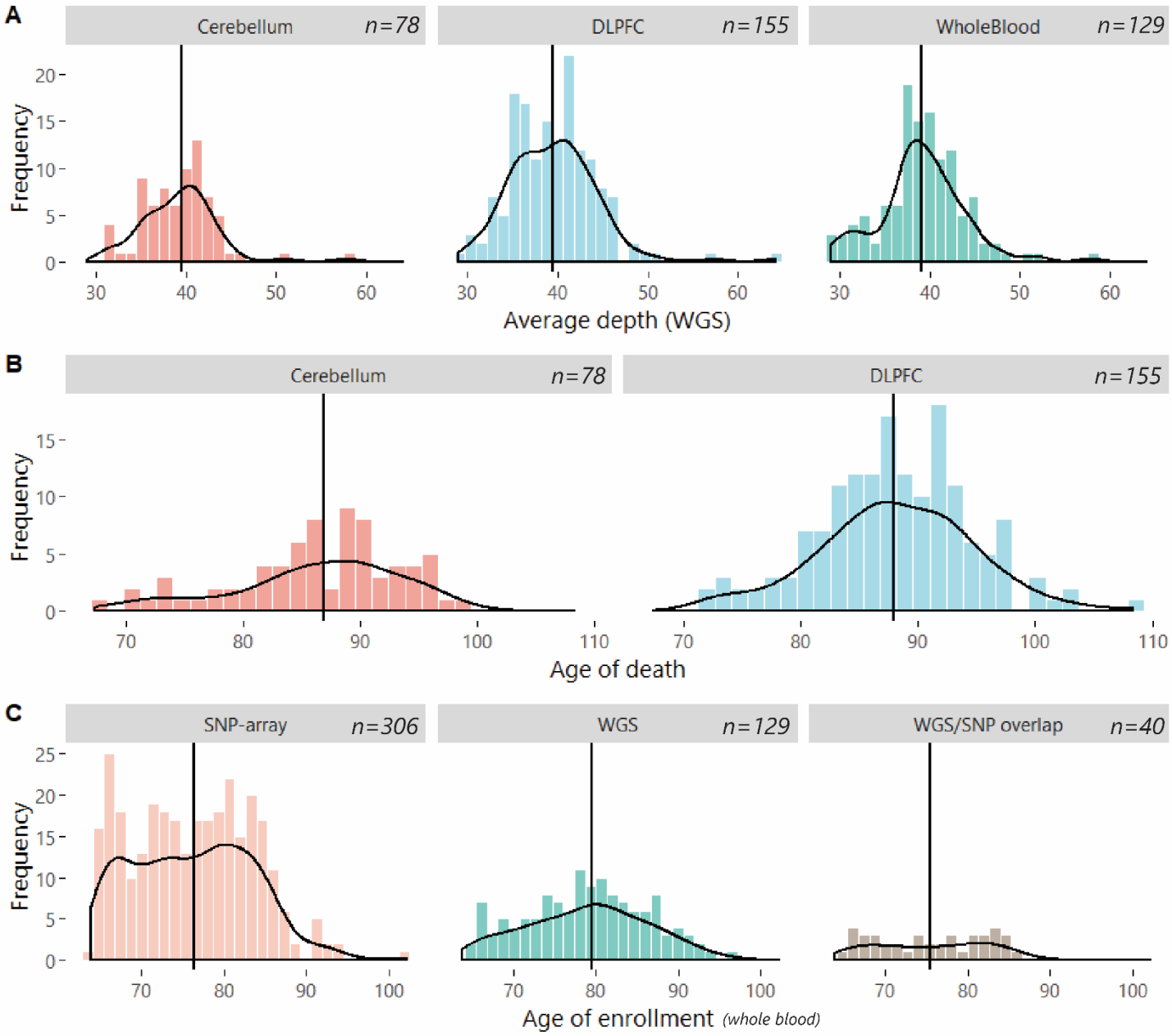
As detailed in De Jager et al. (2018). DNA was extracted from whole blood, DLPFC and cerebellum. DNA libraries were generated using the KAPA Library Preparation Kit. Final libraries were evaluated using fluorescent-based assays including qPCR with the Universal KAPA Library Quantification Kit and Fragment Analyzer (Advanced Analytics) or BioAnalyzer (Agilent 2100). Libraries were sequenced on an Illumina HiSeq X sequencer (v2.5 chemistry) using 2 × 150 bp cycles. Paired-end sequencing reads were aligned to GRCH37 using BWAmem (version 0.7.15). Duplicate reads were marked using Picard MarkDuplicates (version 2.4.1). GATK IndelRealigner (version 3.5) used to improve the consistency of read alignments in regions that contain insertions and deletions, and base quality score recalibration was performed using the GATK BQSR (version 3.5).

Single-nuclei library preparation and sequencing.

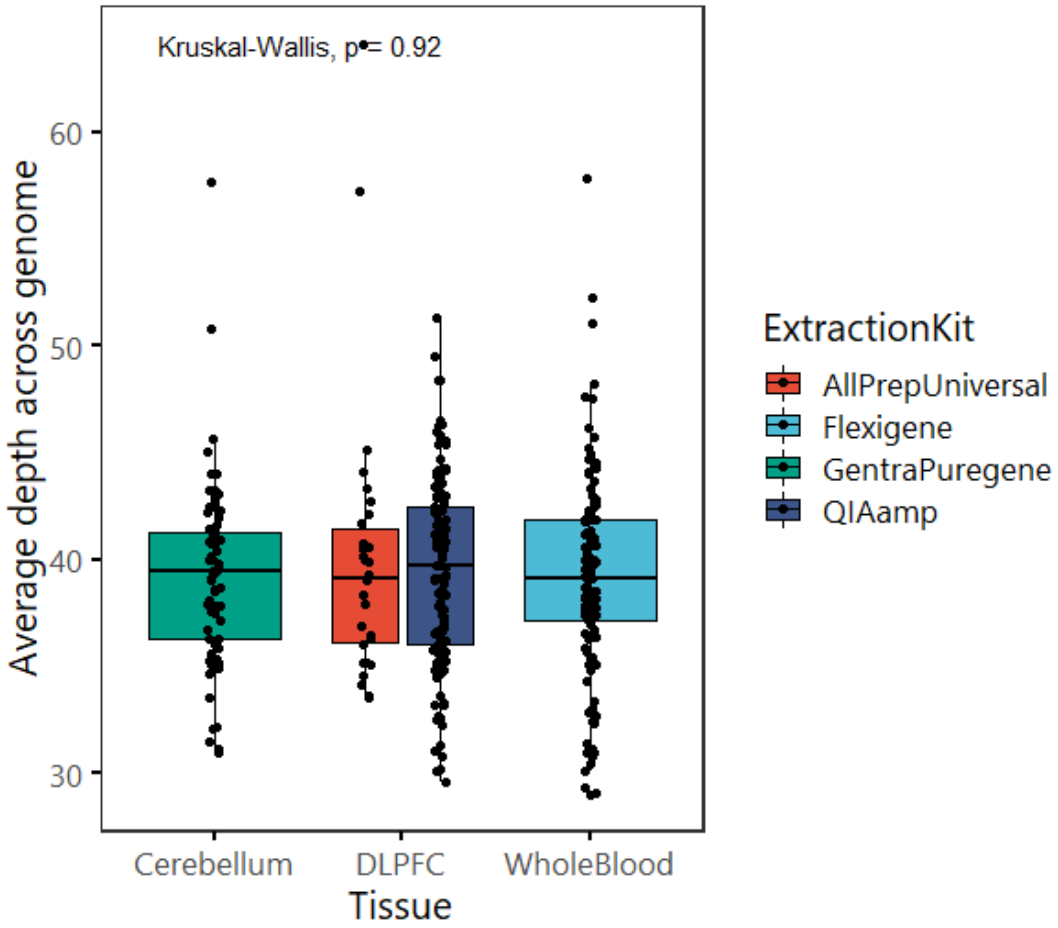
As detailed in Mathys *et* al. (2019). The protocol for the isolation of nuclei from frozen post-mortem brain tissue was adapted from Swiech et al. (2014). Brain tissue was homogenized using a tissue homogenization buffer and filtered through a 40mm cell strainer to isolate individual cells. The nuclei were separated by ultracentrifugation at 9000 RPM, washed and subject to several subsequent rounds of centrifugation. The nuclei were counted and diluted to a concentration of 1000 nuclei per microliter in PBS.

snRNA-seq libraries were prepared using the Chromium Single Cell 3' Reagent Kits v2 according to the manufacturer's protocol (10x Genomics). The generated scRNA-seq libraries were sequenced using Illumina NextSeq 500/550 High Output v2 kits (150 cycles). Raw base call (BCL) files generated by Illumina sequencers were demultiplexed and converted to FASTQ files using cellranger mkfastq (Cell Ranger software 2.0.0).

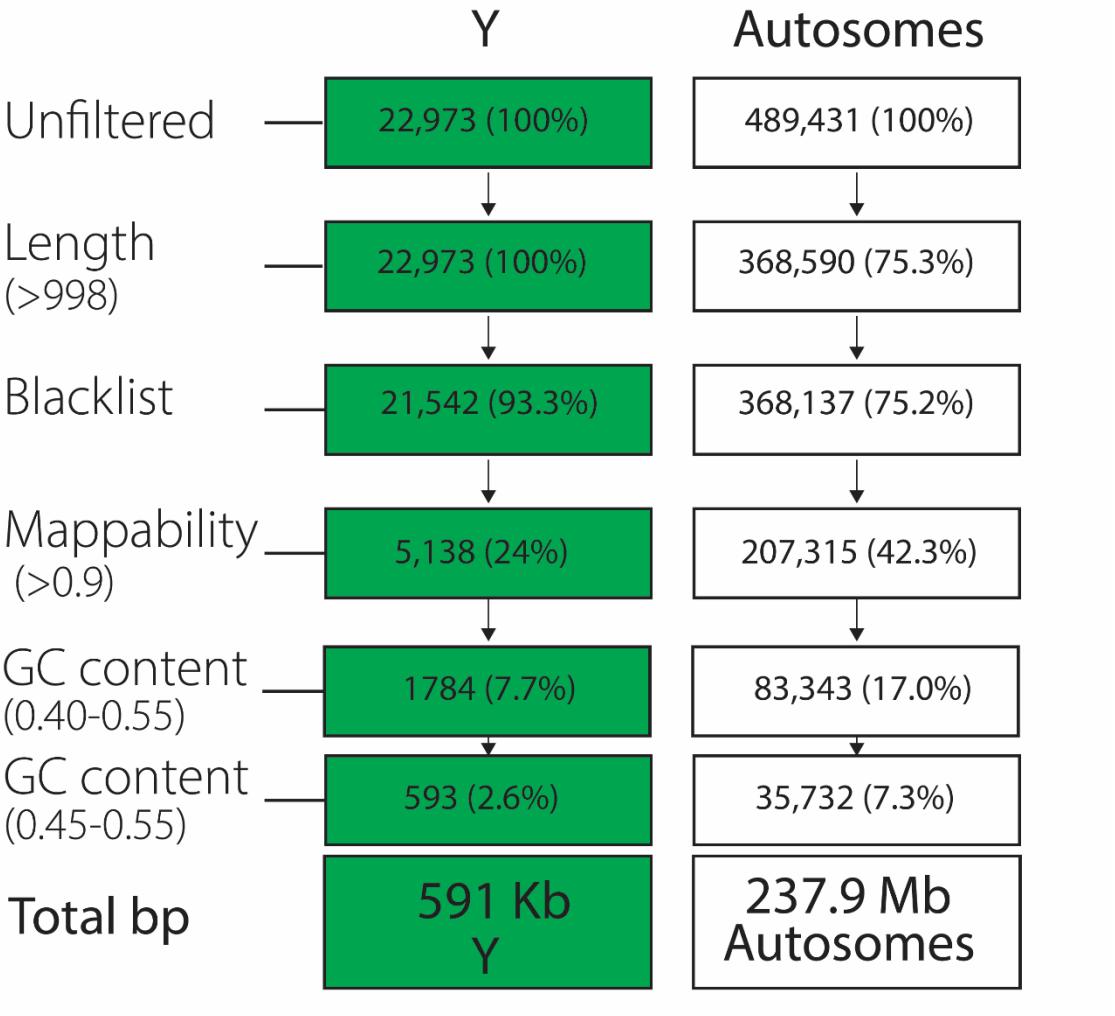
Appendix 2
Chapter 2 - Additional figures



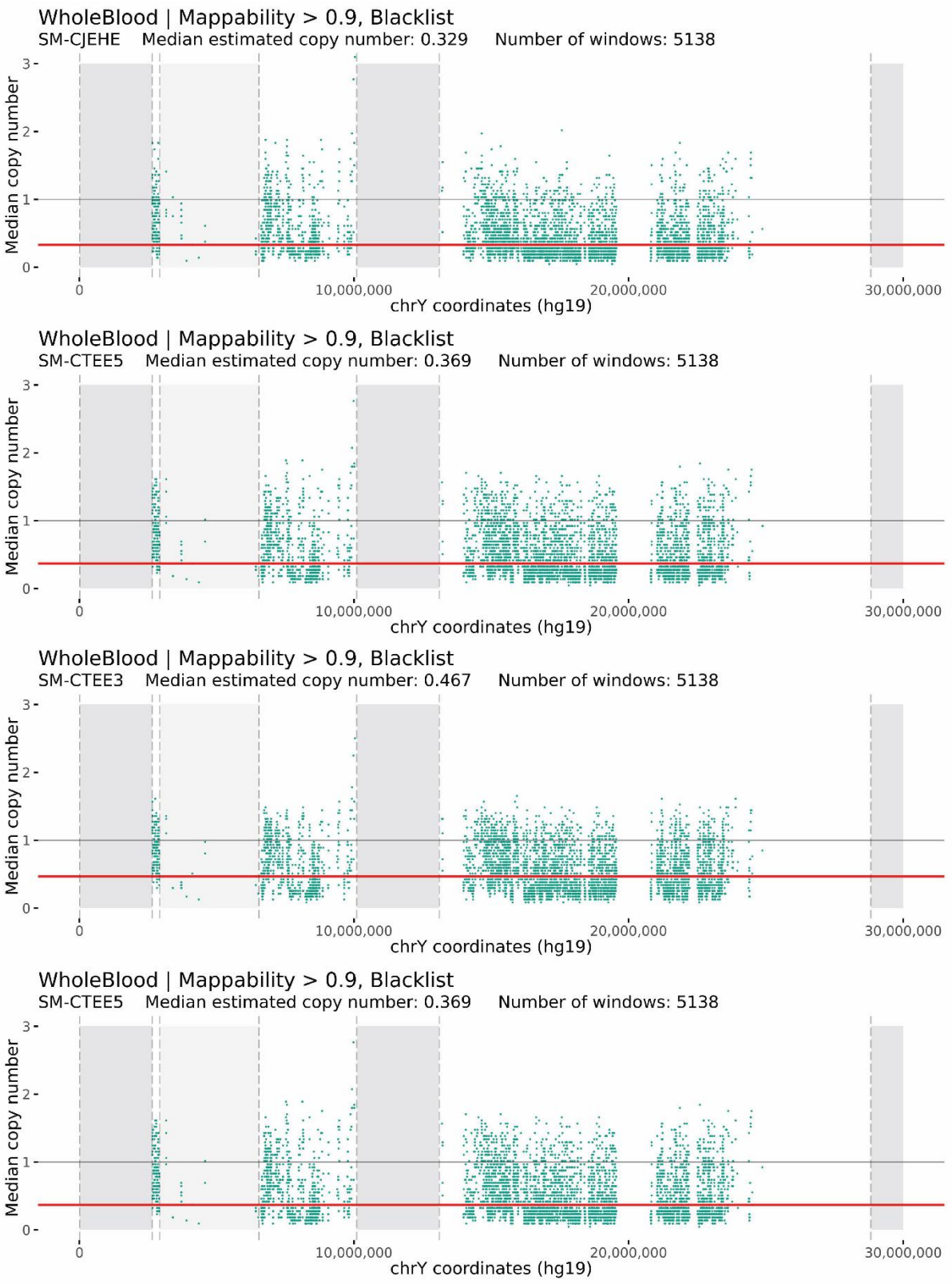
Appendix Figure 2.1 Overview of sequencing depth and age distributions within each tissue. A) WGS depth distributions in each tissue. The median in each tissue is represented by the black line. B) Age of death and age of sampling in both cerebellum and DLPFC tissue (WGS). C) Baseline age or enrollment age distributions of blood samples from SNP-array, WGS, and overlapping (WGS/SNP-array). Date of sampling was not provided, so baseline age was used as an estimation. Dorsolateral prefrontal cortex (DLPFC).



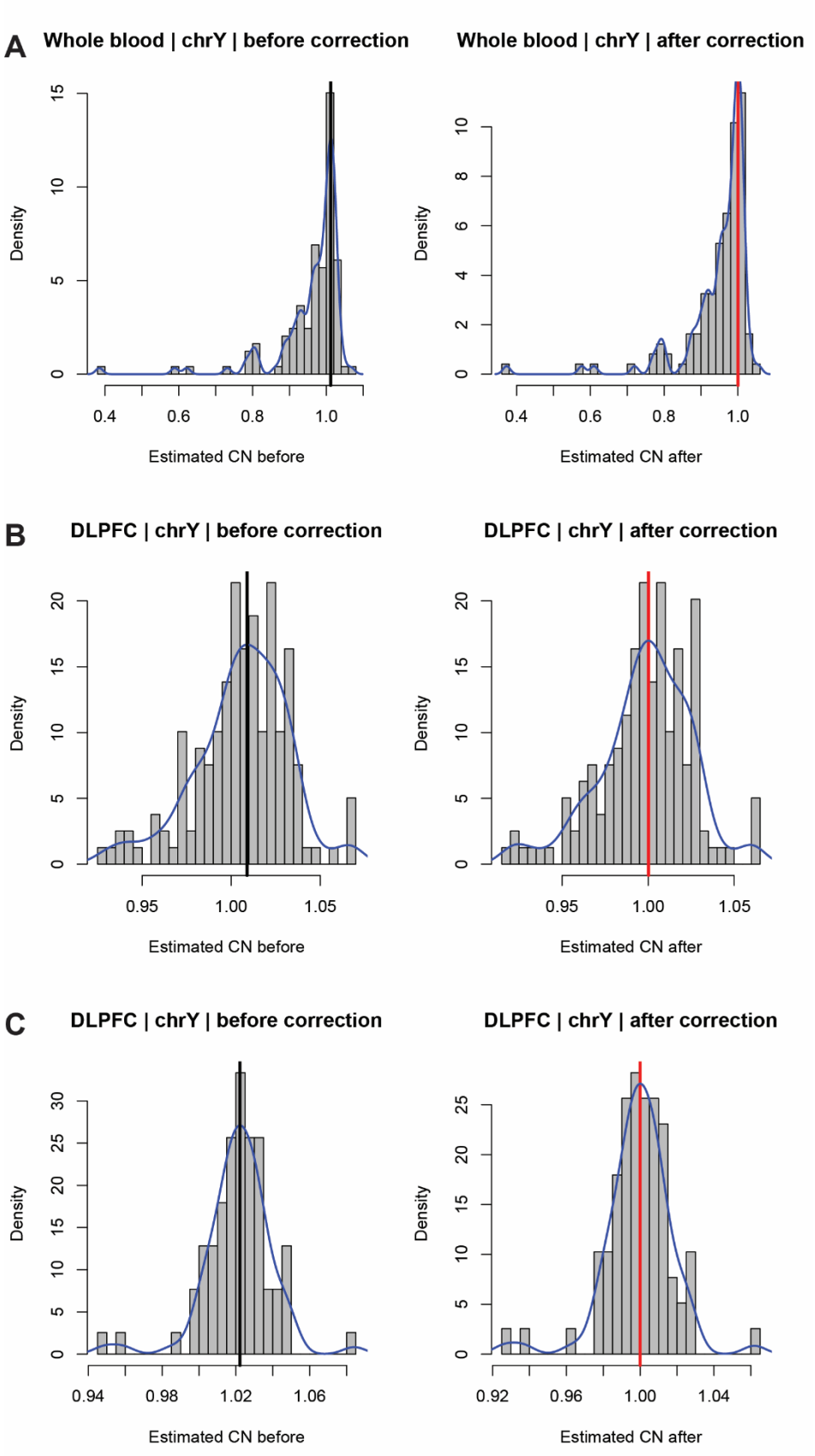
Appendix Figure 2.2 WGS read depth is not significantly different between tissues and extraction kits. Each point represents an individual WGS sample. Samples are divided by tissue and colored by DNA extraction kit.



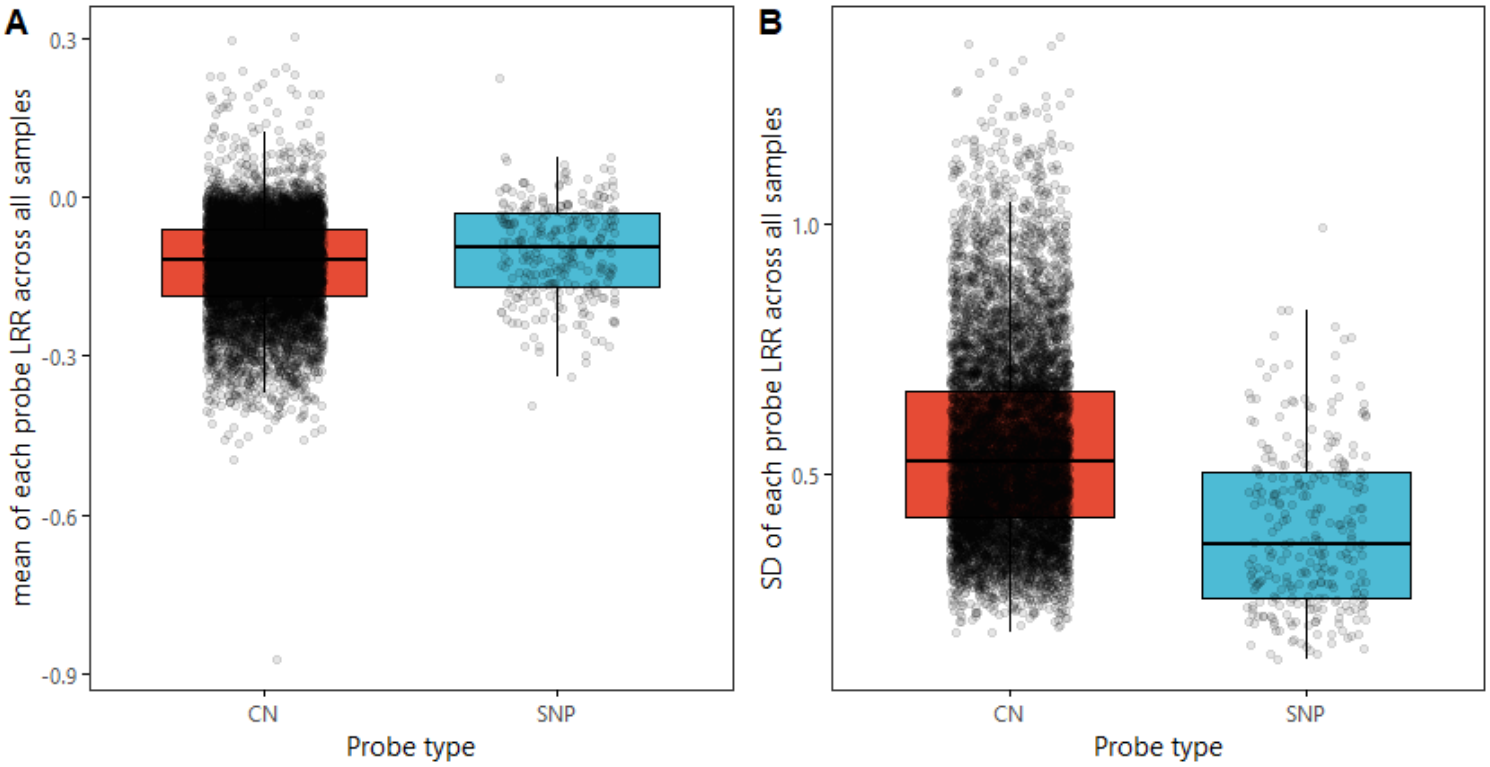
Appendix Figure 2.3 Genomic window filtering process and number of windows remaining. The total number of windows remaining in the study for both the Y chromosome and the sum of the autosomes. After filtering, 2.6% (591 Kb) of chromosome Y is used to determine mosaic ploidy. In contrast, 7.3% of the autosomes were used to calculate ploidy.



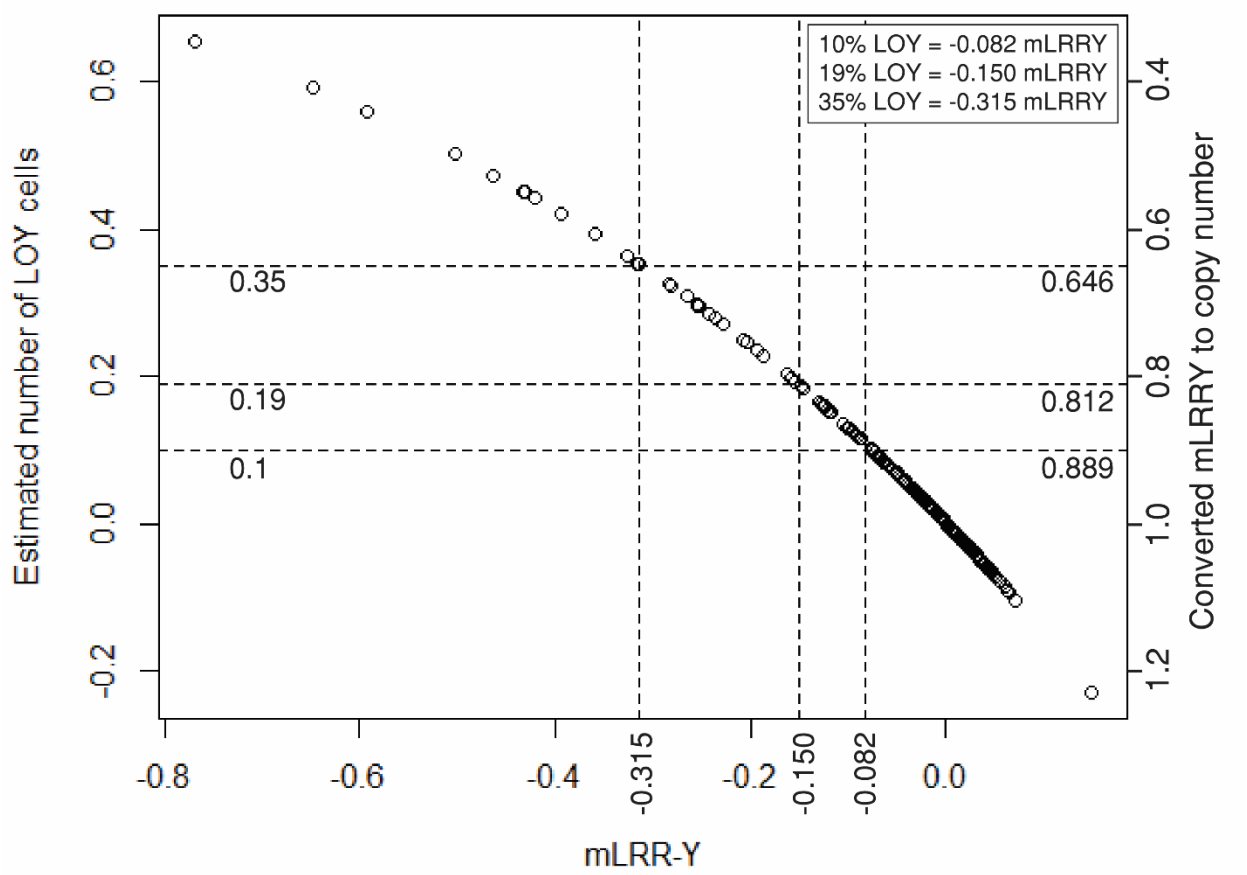
Appendix Figure 2.4 Visual inspection of the Y chromosome in each sample found four highly variable, poorly sequenced samples. Each of these samples had unusually high depth variation genome wide. Originally these samples were considered LOY. Care must be taken to remove poorly sequenced samples as mean depth is often reduced in these samples.



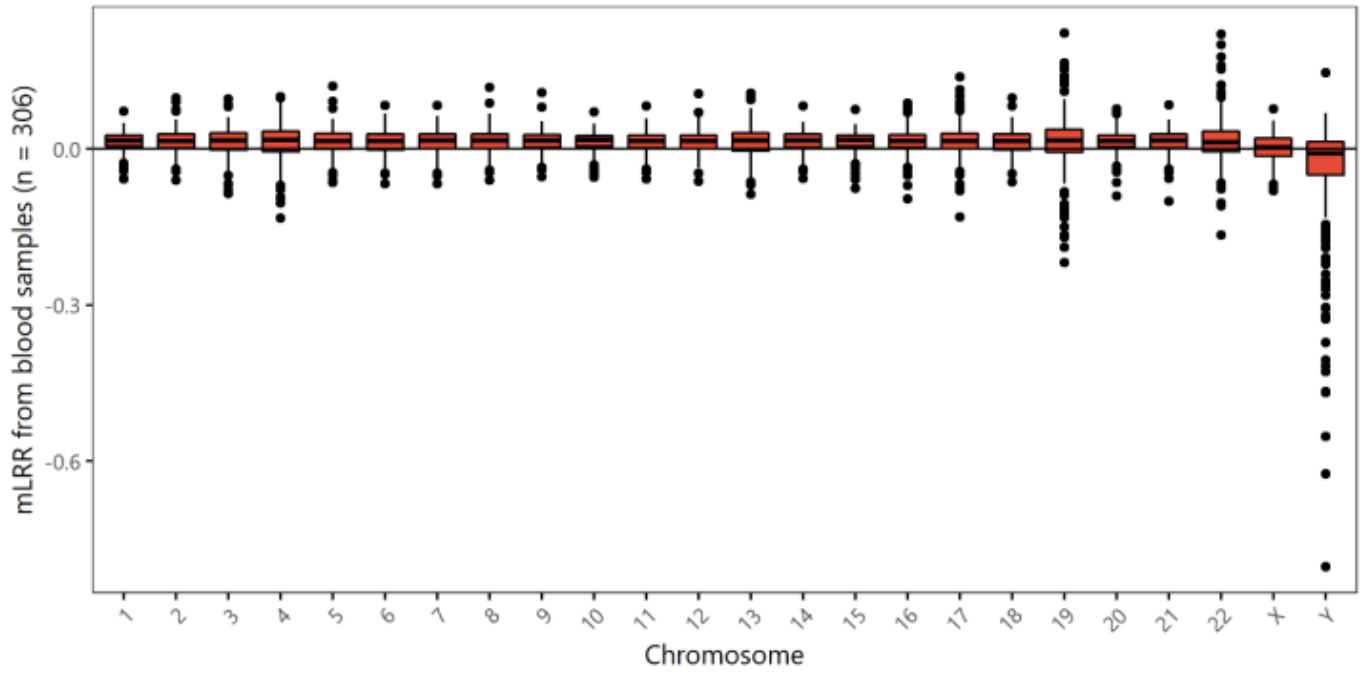
Appendix Figure 2.5 Y chromosome ploidy distribution before and after density kernel estimation correction in each tissue.



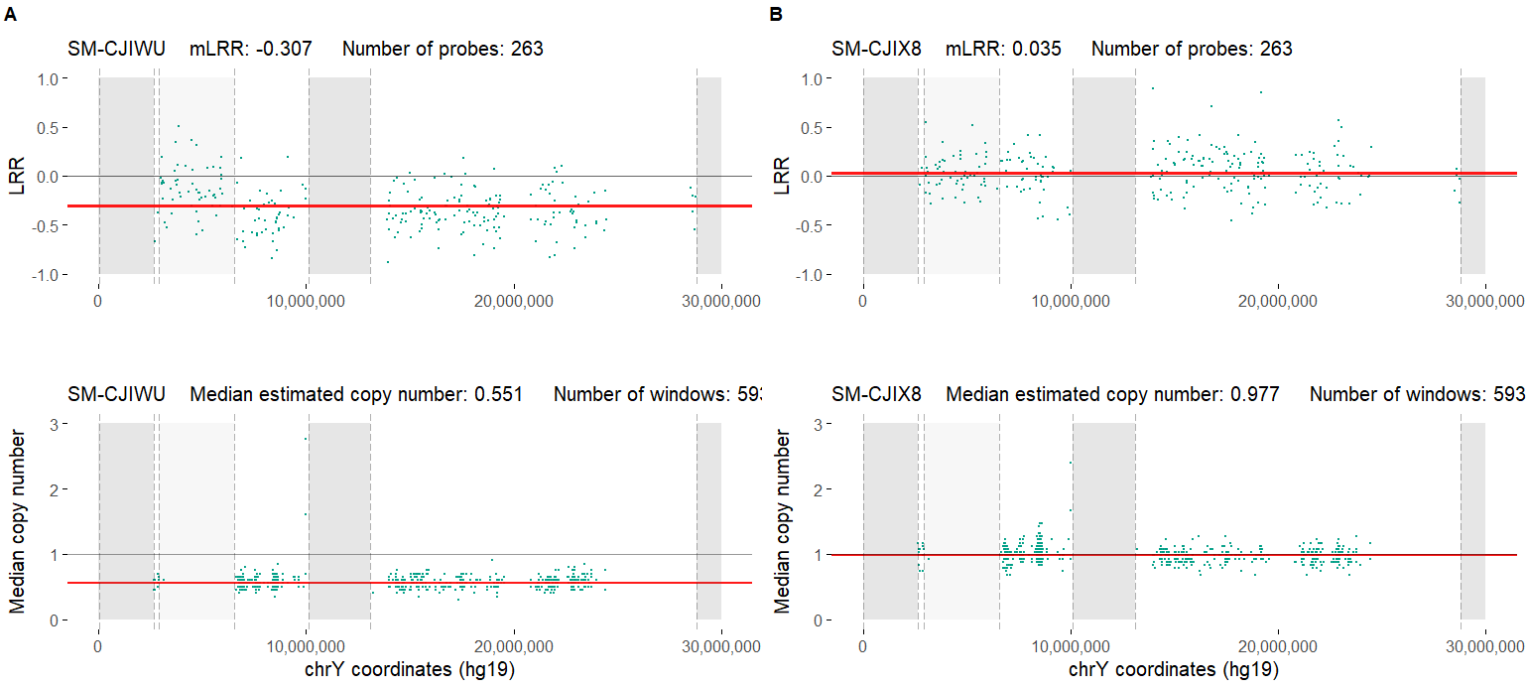
Appendix Figure 2.6 Affymetrix SNP6.0 SNP-array copy number probes are more variable than SNP probes. The Affymetrix SNP6.0 array provides 8582 copy number probes and 271 SNP probes in the male-specific region of chromosome Y. The CN probes provide denser genomic coverage but are more variable than SNP probes. SNP probe processing and normalization has been standardized for LOY detection and therefore CN probes were removed.



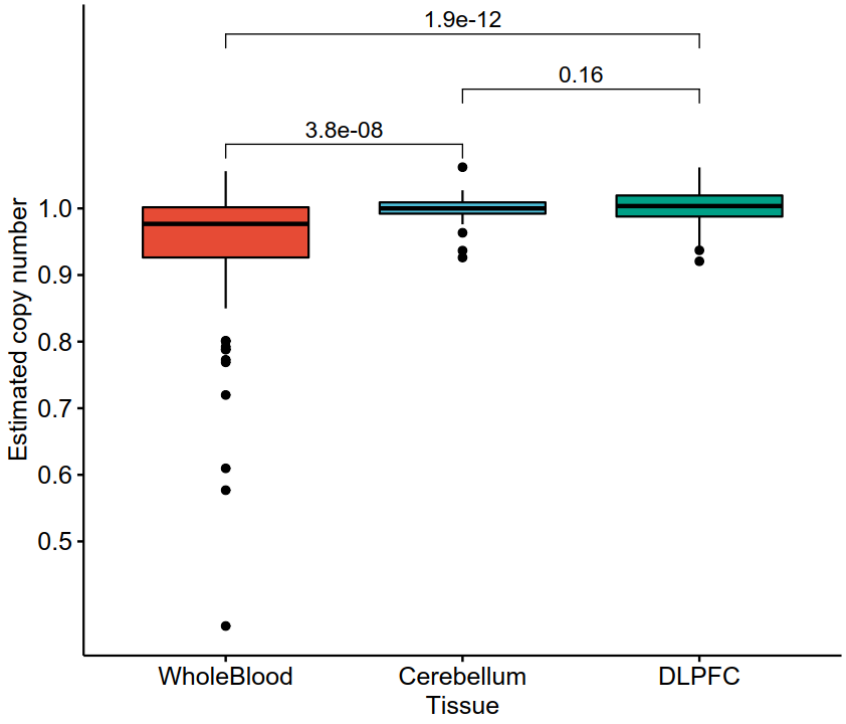
Appendix Figure 2.7 Transformation of mLRRY to rounded SNP-array ratio. To compare LOY more accurately between genotype array and WGS platforms, mLRRY values were transformed into a metric called rounded SNP-array ratio. Because of the high correlation between WGS copy number and mLRRY, formulas can be applied to transform mLRRY values onto an identical scale. Following the formulas from Danielsson *et al.* (2019), mLRRY values were transformed. Extensive testing of samples with paired WGS, ddPCR and SNP-array data found that mLRRY values can be transformed to a CN equivalent using the following formula: $Y = (2^{mLRRY})^2$, where Y is the rounded SNP-array ratio (axis on right). Further, % of LOY cells can be inferred through the following formula: $LOY(\%) = 100*(1-2^{mLRRY})^2$ (axis on left). The LOY threshold at -0.082 mLRRY is equivalent to 0.889 CN and ~10% LOY. Thresholds of increasing LOY severity were set at mLRRY -0.15 which is equivalent to 0.812 CN and ~19% LOY, and -0.315 mLRRY, equivalent to 0.646 CN and 35% LOY. These thresholds were primarily chosen to compare to previous LOY studies, and the values from our analysis are comparable.



Appendix Figure 2.8 Distribution of intensity-based ploidy estimates across the genome in 306 elderly males. As expected, mLRR values for each chromosome are centered around 0. mLRR of 0 represents expected ploidy. Chromosome Y shows significant variation and deviation towards 0, which is not seen in any other chromosome.

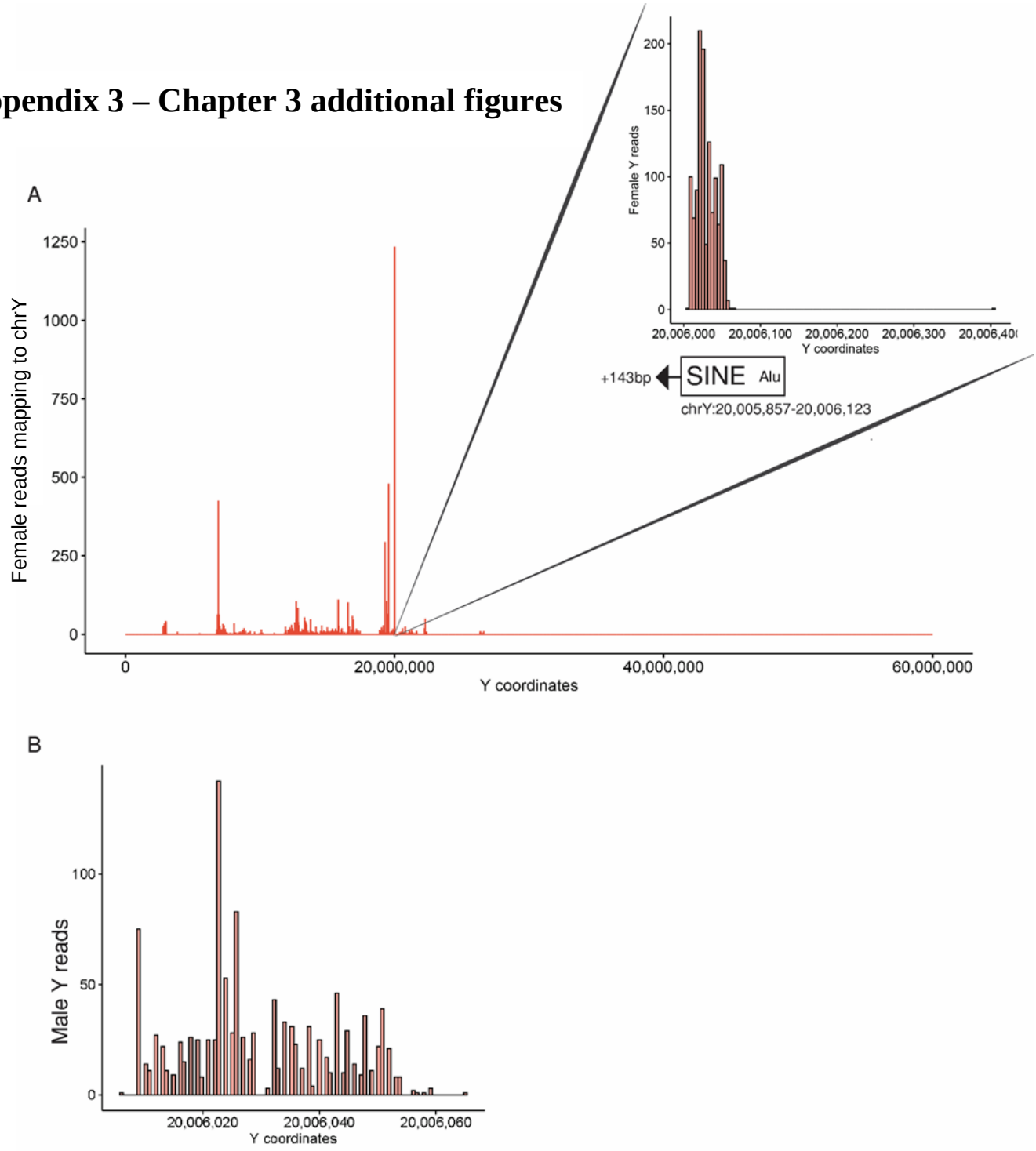


Appendix Figure 2.9 Visualized example of SNP-array and WGS concordance in two paired samples. A) Top, shows each individual probe and its mLRR value across the MSY in a LOY sample (SNP-array). The red line represents the mLRR value. Bottom, WGS sequencing depth across the Y in the same sample. Chromosome Y genomic content is reduced to a similar degree using both technologies. B) Another example in an individual that does not have LOY. Grey segments represent notable genetic regions on the Y chromosome. From left to right, PAR1, XTR (light grey), centromere, heterochromatin.

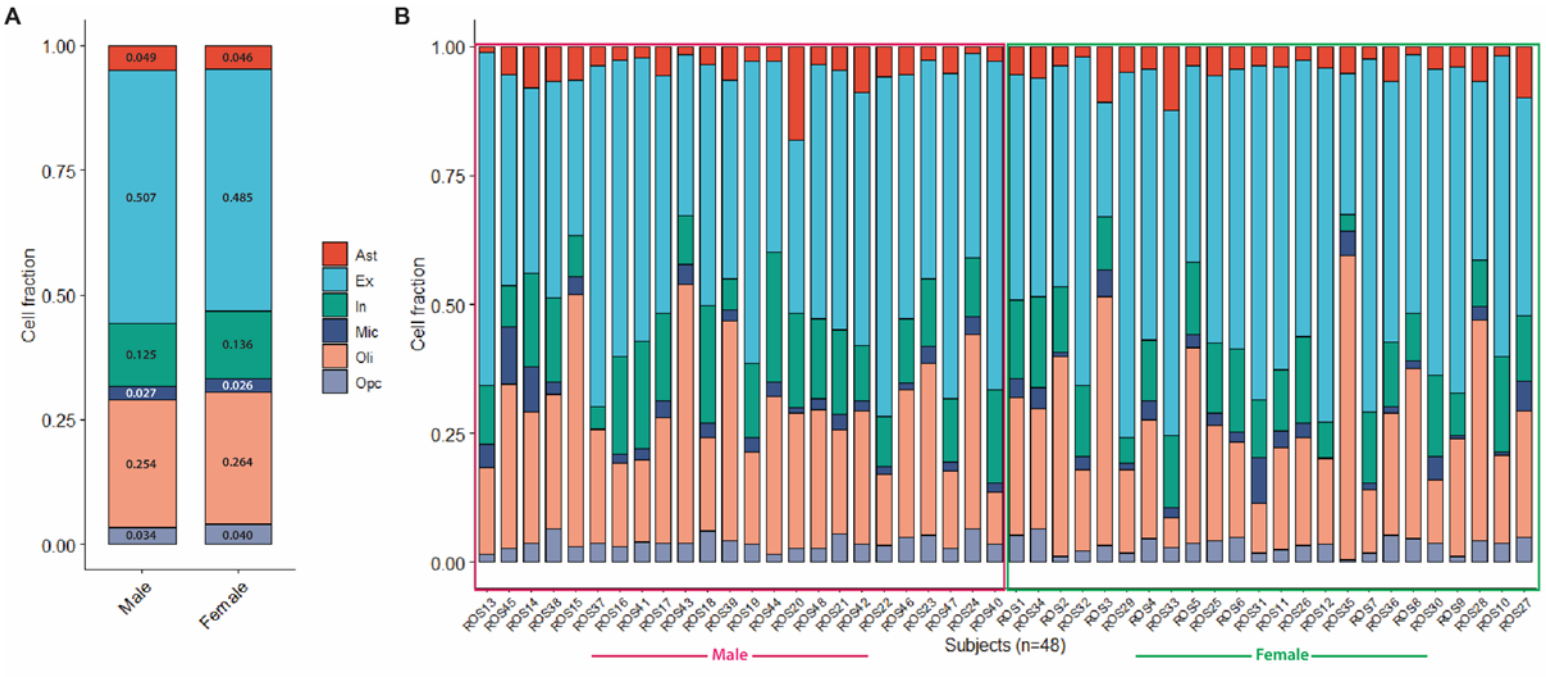


Appendix Figure 2.10 WGS quantified LOY is not significantly different between cerebellum and dorsolateral prefrontal cortex (DLPFC) tissues. Chromosome Y ploidy mean was significantly different between whole blood and both cerebellum (p=3.8x10-8; t-test) and DLPFC tissue (p=1.9x10^{-12}; t-test). Chromosome Y ploidy was not significantly different between cerebellum and DLPFC (p=0.16; t-test).

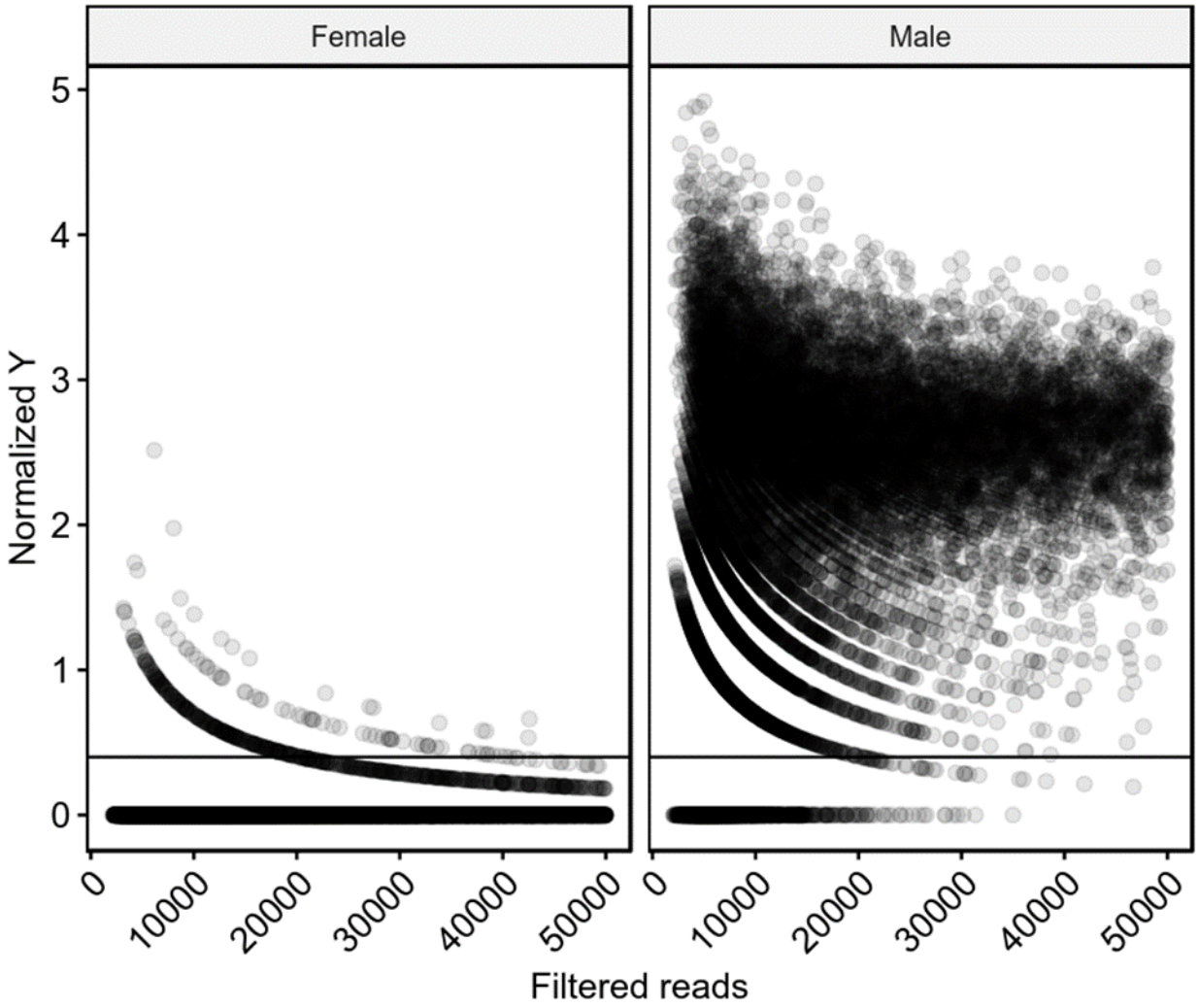
Appendix 3 – Chapter 3 additional figures



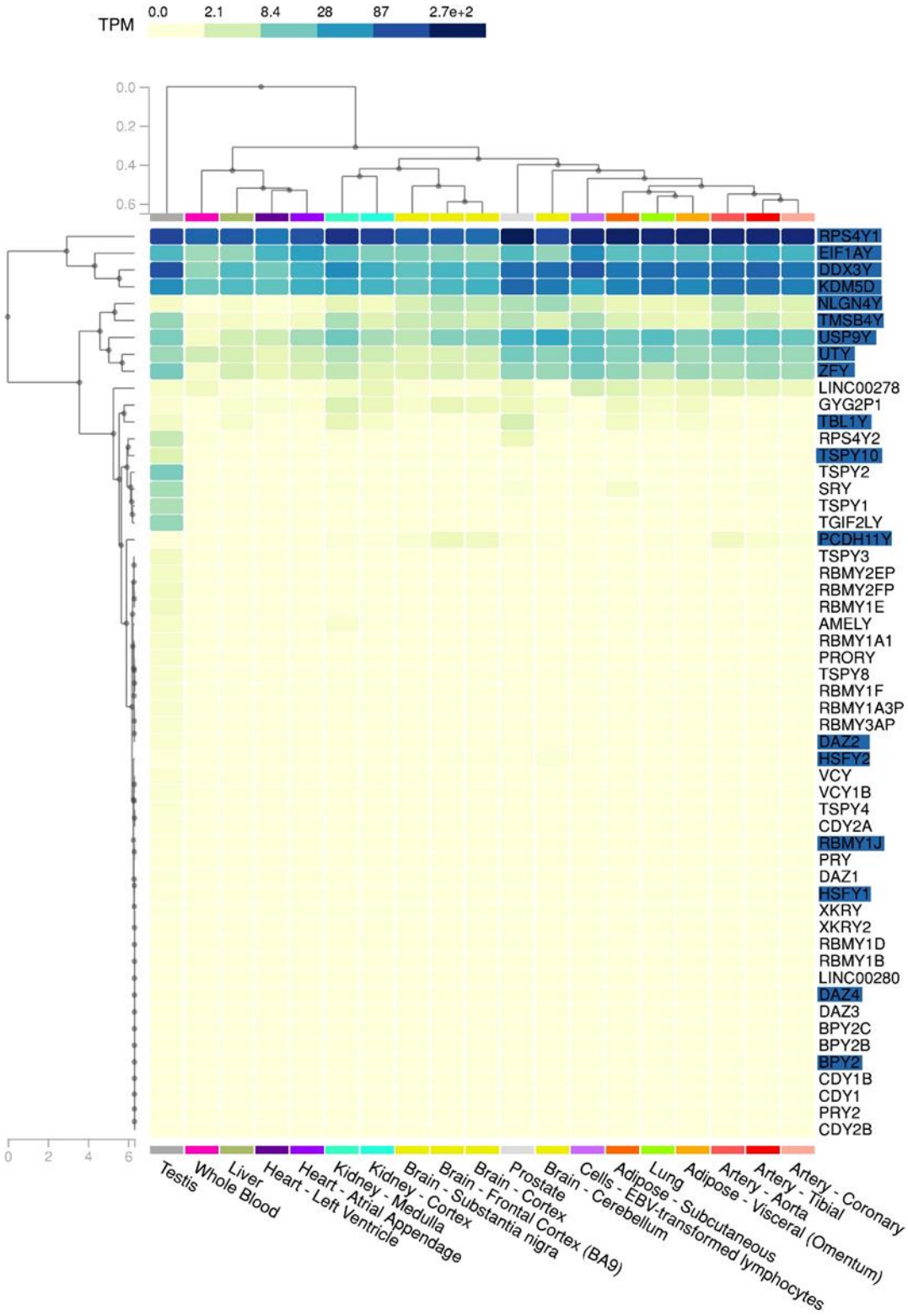
Appendix 3.1. Y chromosome sequencing problem regions in female samples. When using exonic, as well as intronic and intergenic RNA-seq reads as evidence of DNA, technical artefacts are bound to occur. Since females do not have a Y chromosome, we can assume all female Y reads are errant reads. Regions of the genome where these errant reads collect likely contain elements that reliably lead to erroneous mapping in both male and female samples. **A)** Landscape of female reads mapping to the Y across the single-nuclei RNA-seq cohort after filtering (n=24; reads=7013). ~30% of female reads mapping to the Y fall within 3 1kb windows. The insert shows a magnified example of one of these windows (intergenic). The false positive reads overlap with an interspersed nuclear element (Alu). This is observed across most enriched errant read windows. **B)** The queried window above is also observed in male samples. Male reads are mapped in the same location as in the females. We assume these reads are technical artefacts.



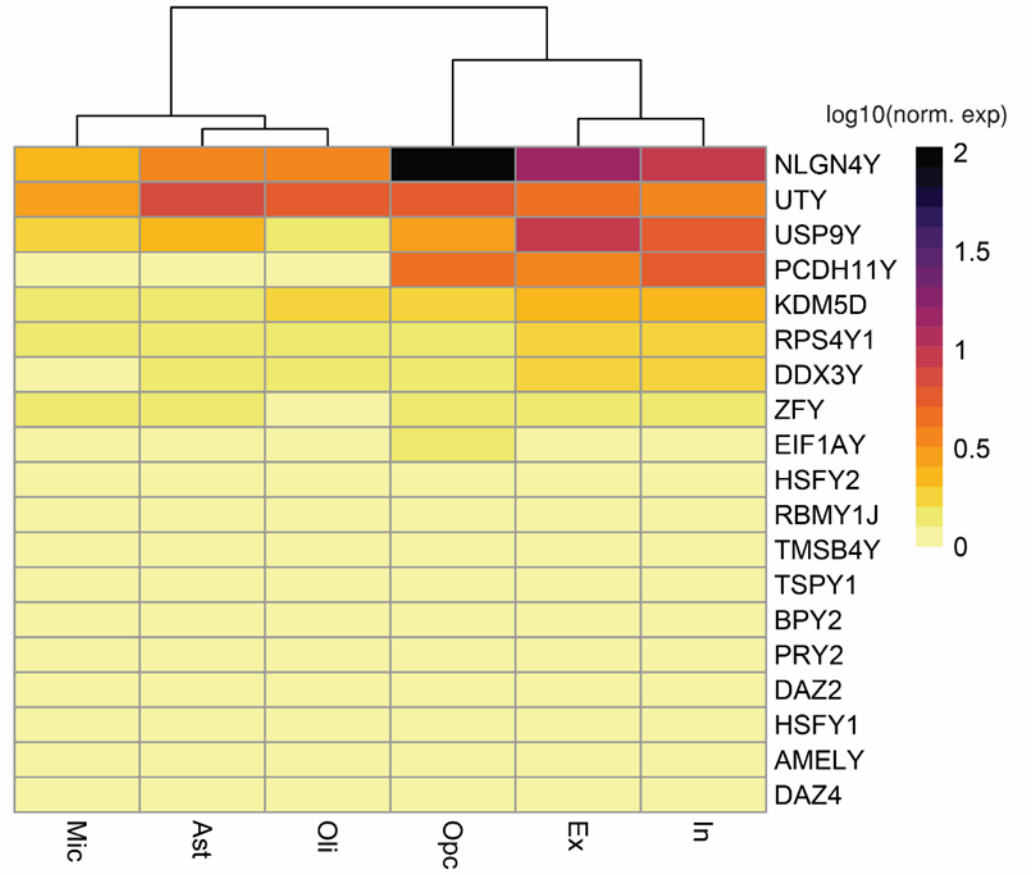
Appendix 3.2. Sequenced cell-type proportions by sample and sex for 48 DLPFC single-nucleus RNA-seq samples. A) Cell-type proportions of isolated cells across 24 male and 24 female samples (n=48). Cell-type proportions are replicated between the sexes. **B)** Cell-type proportions vary slightly across samples, but for the most part are highly similar. Cluster cell-type assignments were completed by Mathys *et al.* 2019. Astrocyte (Ast), excitatory neuron (Ex), inhibitory neuron (In), microglia (Mic), oligodendrocyte (Oli), oligodendrocyte progenitor cell (Opc).



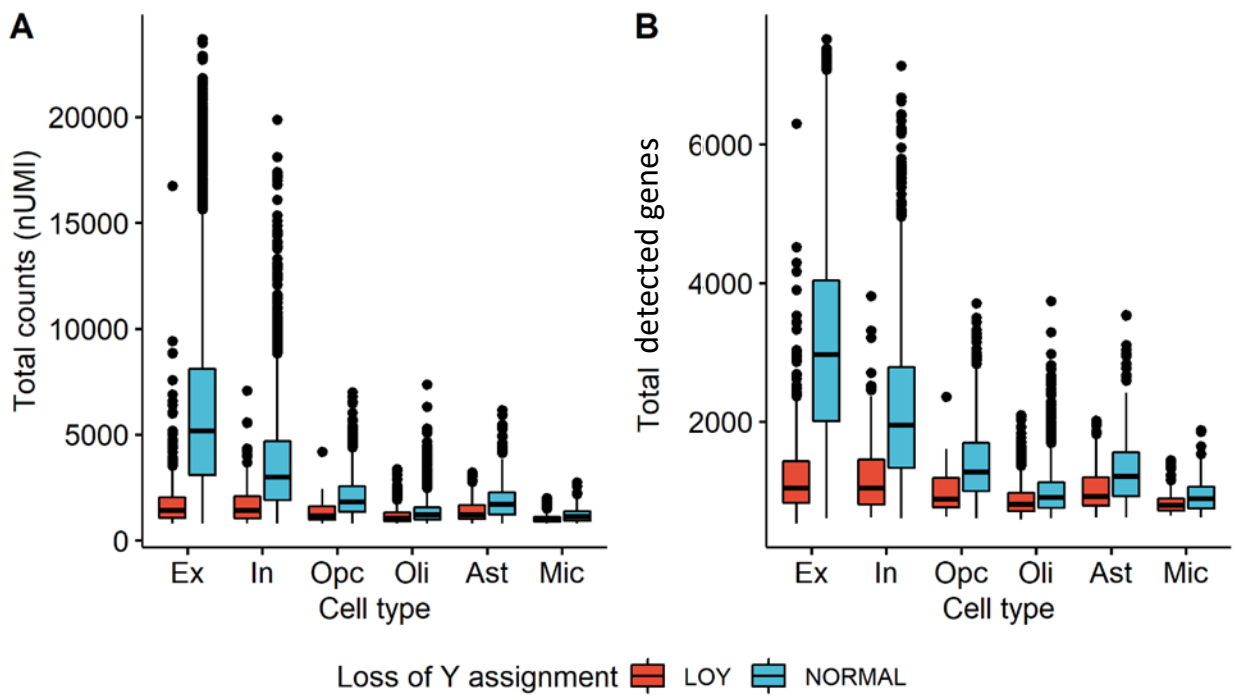
Appendix 3.3. Determining the normalized Y read metric cut-off. After read filtering, male specific Y (MSY) region reads are normalized by cell by total reads and multiplied by a scale factor of 10000. The logp1 (natural logarithm + 1) is then applied to avoid undefined values. Essentially the MSY region is treated as a single gene and normalized as such. The LOY cut-off was set at 0.4 which is equivalent to ~1 Y read per 20,000 reads. The average male normalized Y was 2.16, while in female samples it was 0.018. The 0.40 cut-off adds an additional 22 LOY cells to the obvious 0 Y read LOY cells. Each point represents 1 cell.



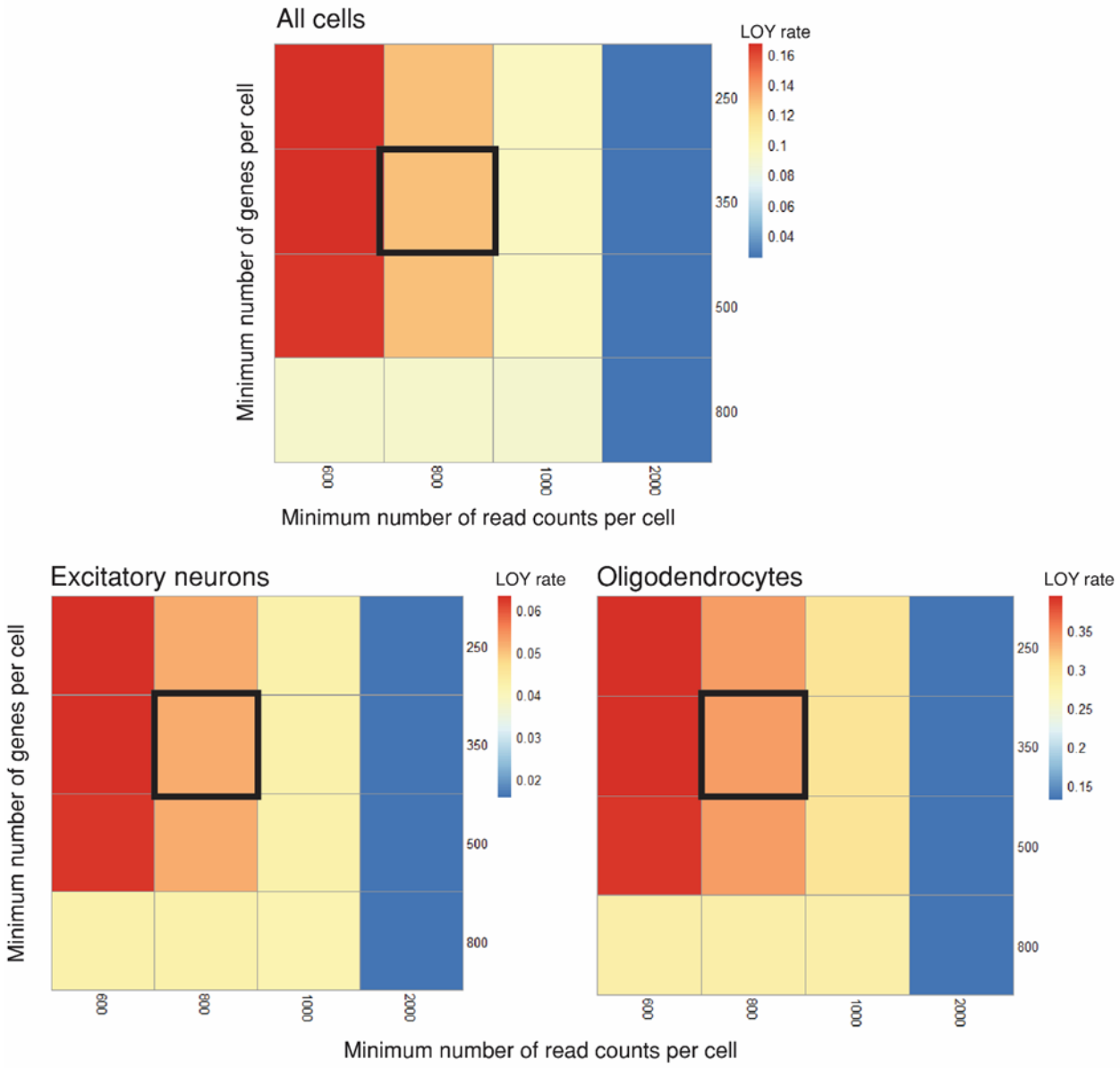
Appendix 3.4. Chromosome Y gene expression across GTEx tissues. 9 Y-linked genes are ubiquitously expressed across human tissues. Additionally, it does not appear that Y-linked gene expression differs significantly between cortex tissue and other tissues such as whole blood. This plot was produced via the GTEx multi-gene query tool (www.gtexportal.org/home/multiGeneQueryPage). Gene names in blue are expressed in at least 3 cells in the single-nuclei dorsolateral prefrontal cortex dataset. All expression values are in transcripts per million (TPM).



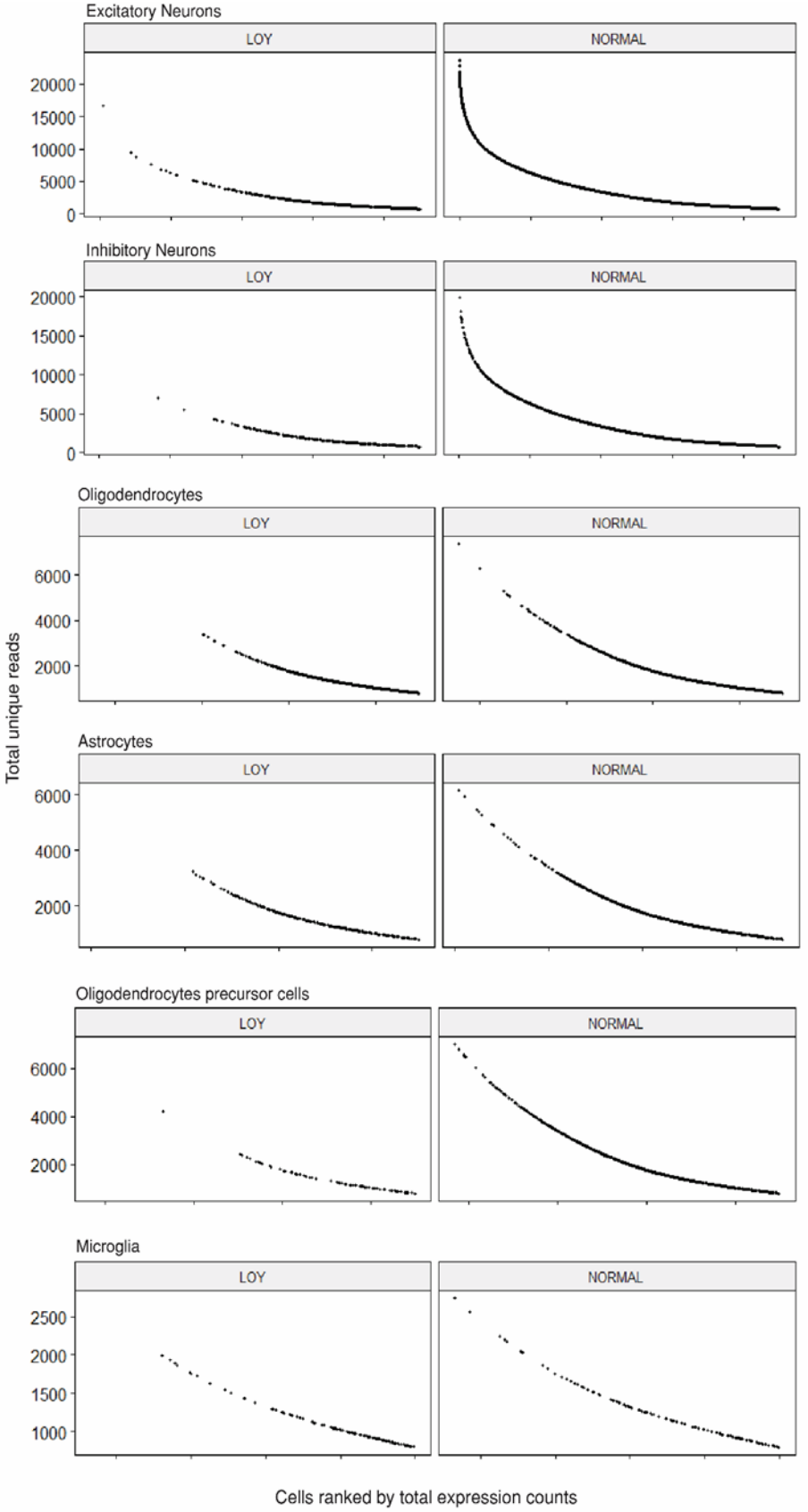
Appendix 3.5. Chromosome Y gene expression across dorsolateral prefrontal cortex cell-types. 8 Y-linked genes are commonly expressed across the major dorsolateral prefrontal cortex cell-types in this dataset. NLGN4Y is the most highly expressed Y-linked gene across all cell-types in the single-nuclei dataset which differs from the bulk RNAseq GTEx data where RPS4Y1 and EIF1AY are the top expressed Y genes in the cortex.



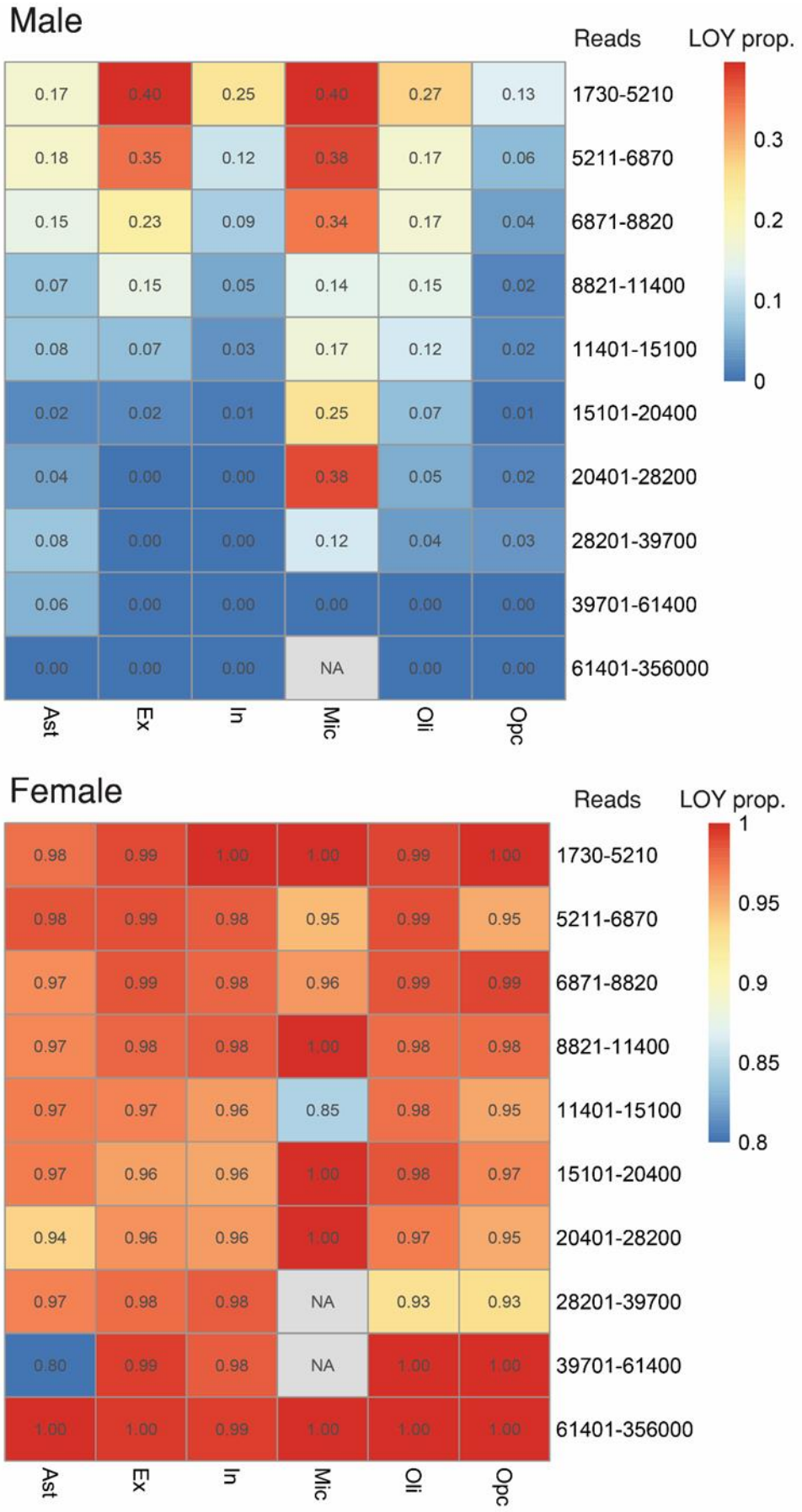
Appendix 3.6. Difference in expression counts and total detected genes between normal and LOY cells using the MSY gene expression counts. A-B) LOY cells have significantly fewer total reads and detected genes. Using the method from Thompson *et al.* enriches for low-expressing, low-depth cells and not necessarily cells lacking a Y chromosome.



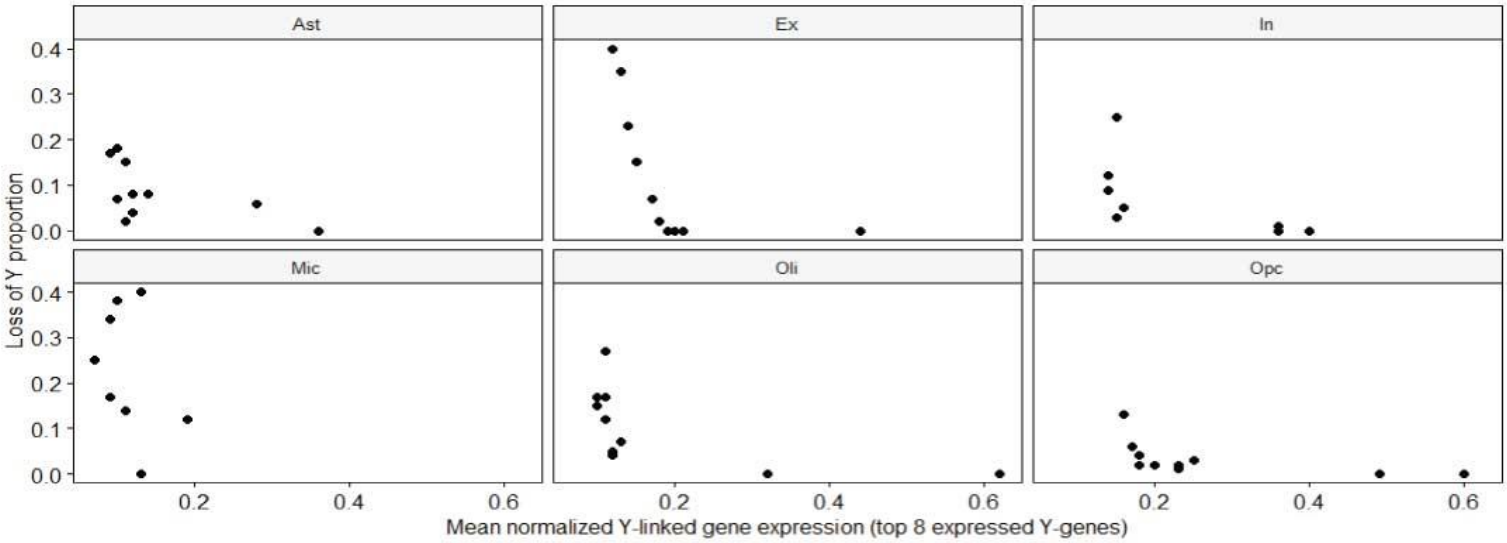
Appendix 3.7. Loss of Y rates across increasingly strict quality control thresholds using MSY gene expression counts method from Thompson *et al*. (2019). Across each cell-type, as thresholds on library complexity and total counts become stricter, LOY rates are reduced. The highlighted box is cut-off used throughout Chapter 3.



Appendix 3.8. LOY cells are concentrated in low read depth cells. For each cell-type, cells are ranked (left to right) by total cell read depth and split by LOY and normal ploidy assignment. Each point is a cell. All cell-types except microglia, have a concentration of LOY cells in amongst the lowest expressed cells, although microglia expression is reduced compared to the other cell-types.



Appendix 3.9. Loss of Y proportions and mean Y gene expression by cell-type across increasing read depth. Top, cells were split into 10 proportional bins based on cell read depth. LOY assignments were made using the XY ratio cut-off of 0.005. Astrocytes, oligodendrocytes, and microglia have elevated rates of Y loss. Bottom, female LOY rates with increasing cell depth.



Appendix 3.10. Loss of Y proportions by mean normalized Y chromosome gene expression by cell type. Single-cell data was split into ten proportional groups based on read depth. For each cell-type, LOY proportion and mean normalized Y chromosome gene expression were compared. Values correspond with **Appendix 3.9**. As total Y gene expression increases, LOY proportion decreases.

Appendix 4
Affymetrix SNP6.0 male specific Y probes used to detect median log R ratio

Affymetrix Male-specific Y probes used to estimate LOY.

Affymetrix_probe_ID	Chr	Genomic_position (hg19)
SNP_A-8655052	Y	2722506
SNP_A-8469844	Y	2846401
SNP_A-8567242	Y	3006232
SNP_A-8289946	Y	3022396
SNP_A-8327510	Y	3044030
SNP_A-8647397	Y	3044113
SNP_A-8537291	Y	3089809
SNP_A-8346521	Y	3100347
SNP_A-8656150	Y	3165517
SNP_A-8572255	Y	3201251
SNP_A-8352925	Y	3266087
SNP_A-8543663	Y	3496739
SNP_A-8369873	Y	3539834
SNP_A-8289954	Y	3586616
SNP_A-8638351	Y	3691575
SNP_A-8687827	Y	3713309
SNP_A-8330311	Y	3713606
SNP_A-8600263	Y	3777756
SNP_A-8349310	Y	3826212
SNP_A-8370743	Y	3827532
SNP_A-8680255	Y	3837764
SNP_A-8364206	Y	3904491
SNP_A-8465374	Y	3931646
SNP_A-8353700	Y	4216073
SNP_A-8304227	Y	4357591
SNP_A-8425444	Y	4462368
SNP_A-8407064	Y	4472623
SNP_A-8432696	Y	4508627
SNP_A-8289971	Y	4548849
SNP_A-8289973	Y	4696023
SNP_A-8289974	Y	4696181
SNP_A-8306261	Y	4703804
SNP_A-8320257	Y	4721636
SNP_A-8319760	Y	4805510
SNP_A-8360538	Y	4834225

SNP_A-8393997	Y	4863907
SNP_A-8404563	Y	4866464
SNP_A-8549683	Y	4939870
SNP_A-8491902	Y	5022306
SNP_A-8648299	Y	5171911
SNP_A-8523311	Y	5259959
SNP_A-8380633	Y	5364159
SNP_A-8561782	Y	5483659
SNP_A-8289995	Y	5532303
SNP_A-8489786	Y	5542746
SNP_A-8309056	Y	5586774
SNP_A-8514827	Y	5632507
SNP_A-8355881	Y	5677457
SNP_A-8592465	Y	5854793
SNP_A-8318909	Y	5866261
SNP_A-8338258	Y	5890420
SNP_A-8347325	Y	5904411
SNP_A-8587589	Y	5913967
SNP_A-8504756	Y	5916713
SNP_A-8697464	Y	5943453
SNP_A-8568887	Y	6449216
SNP_A-8290003	Y	6677619
SNP_A-8650667	Y	6681479
SNP_A-8615420	Y	6892243
SNP_A-8439420	Y	6943522
SNP_A-8521579	Y	6965215
SNP_A-8520572	Y	6995523
SNP_A-8550793	Y	7073423
SNP_A-8520821	Y	7292720
SNP_A-8361499	Y	7401836
SNP_A-8569177	Y	7527958
SNP_A-8518383	Y	7628484
SNP_A-8502405	Y	7642823
SNP_A-8568099	Y	7714986
SNP_A-8555082	Y	7716262
SNP_A-8693123	Y	7766712
SNP_A-8360329	Y	7891188
SNP_A-8383167	Y	7963031
SNP_A-8507691	Y	8021340
SNP_A-8713092	Y	8046731
SNP_A-8573170	Y	8108722
SNP_A-8566304	Y	8214827
SNP_A-8589367	Y	8334875

SNP_A-8290006	Y	8353707
SNP_A-8573547	Y	8418927
SNP_A-8436917	Y	8424089
SNP_A-8372352	Y	8502236
SNP_A-8397097	Y	8533735
SNP_A-8611867	Y	8558969
SNP_A-8290008	Y	8576009
SNP_A-8329411	Y	8590752
SNP_A-8528768	Y	8614138
SNP_A-8511719	Y	8667179
SNP_A-8290010	Y	8679843
SNP_A-8290011	Y	8680661
SNP_A-8531040	Y	8728974
SNP_A-8344561	Y	8796078
SNP_A-8330416	Y	8906357
SNP_A-8627092	Y	9131385
SNP_A-8488238	Y	9131437
SNP_A-8290012	Y	9170545
SNP_A-8363005	Y	9392948
SNP_A-8572231	Y	9841700
SNP_A-8514449	Y	9958663
SNP_A-8364073	Y	9984932
SNP_A-8537348	Y	13887941
SNP_A-8457681	Y	13964228
SNP_A-8346213	Y	13964829
SNP_A-8638845	Y	13974233
SNP_A-8333138	Y	13992338
SNP_A-8676343	Y	14001232
SNP_A-8290014	Y	14028148
SNP_A-8581506	Y	14031334
SNP_A-8536948	Y	14197867
SNP_A-8523431	Y	14231292
SNP_A-8566334	Y	14286528
SNP_A-8293961	Y	14345705
SNP_A-8290016	Y	14392807
SNP_A-8498269	Y	14416216
SNP_A-8584255	Y	14577177
SNP_A-8640091	Y	14641193
SNP_A-8396630	Y	14664631
SNP_A-8322132	Y	14698928
SNP_A-8550103	Y	14804077
SNP_A-8570506	Y	14813991
SNP_A-8707909	Y	14901633

SNP_A-8699812	Y	14993358
SNP_A-8290020	Y	15026424
SNP_A-8306235	Y	15027529
SNP_A-8438204	Y	15232812
SNP_A-8564434	Y	15234830
SNP_A-8436578	Y	15467824
SNP_A-8507820	Y	15472863
SNP_A-8520820	Y	15517851
SNP_A-8390409	Y	15526751
SNP_A-8290024	Y	15590342
SNP_A-8485881	Y	15590674
SNP_A-8422313	Y	15591537
SNP_A-8570141	Y	15594523
SNP_A-8423339	Y	15651438
SNP_A-8290262	Y	15818409
SNP_A-8582170	Y	15879017
SNP_A-8465701	Y	15935524
SNP_A-8499470	Y	15944828
SNP_A-8680540	Y	16185081
SNP_A-8604154	Y	16202980
SNP_A-8474085	Y	16242316
SNP_A-8378741	Y	16315153
SNP_A-8676803	Y	16368310
SNP_A-8290265	Y	16377198
SNP_A-8475261	Y	16401405
SNP_A-8290266	Y	16415916
SNP_A-8615173	Y	16683871
SNP_A-8290267	Y	16699334
SNP_A-8649280	Y	16742224
SNP_A-8642739	Y	16773870
SNP_A-8676185	Y	16804852
SNP_A-8314584	Y	16836079
SNP_A-8637392	Y	17011456
SNP_A-8295774	Y	17132580
SNP_A-8458096	Y	17174741
SNP_A-8470535	Y	17394111
SNP_A-8543515	Y	17398598
SNP_A-8655113	Y	17412198
SNP_A-8607809	Y	17495914
SNP_A-8589485	Y	17502468
SNP_A-8290269	Y	17510288
SNP_A-8418205	Y	17559652
SNP_A-8688109	Y	17570599

SNP_A-8501699	Y	17614366
SNP_A-8676818	Y	17686886
SNP_A-8385899	Y	17755905
SNP_A-8290274	Y	17762668
SNP_A-8691942	Y	17766762
SNP_A-8614867	Y	17782178
SNP_A-8565299	Y	17881230
SNP_A-8586236	Y	17891241
SNP_A-8545763	Y	18066156
SNP_A-8380638	Y	18097251
SNP_A-8558524	Y	18101521
SNP_A-8697932	Y	18117193
SNP_A-8290275	Y	18167403
SNP_A-8290276	Y	18167479
SNP_A-8483248	Y	18248698
SNP_A-8433730	Y	18381735
SNP_A-8290277	Y	18561042
SNP_A-8290278	Y	18578476
SNP_A-8705225	Y	18596847
SNP_A-8295953	Y	18700150
SNP_A-8290279	Y	18747493
SNP_A-8543184	Y	18759669
SNP_A-8304713	Y	18786174
SNP_A-8290280	Y	18831084
SNP_A-8603634	Y	18860537
SNP_A-8496337	Y	18888200
SNP_A-8439125	Y	19038302
SNP_A-8310517	Y	19045552
SNP_A-8572233	Y	19048602
SNP_A-8580553	Y	19054889
SNP_A-8413183	Y	19077394
SNP_A-8420219	Y	19166861
SNP_A-8368215	Y	19179335
SNP_A-8519988	Y	19179540
SNP_A-8537039	Y	19198212
SNP_A-8592406	Y	19233673
SNP_A-8594074	Y	19279765
SNP_A-8329434	Y	19315988
SNP_A-8653199	Y	19349615
SNP_A-8411857	Y	19370916
SNP_A-8417266	Y	19372700
SNP_A-8295972	Y	19563894
SNP_A-8372528	Y	20834703

SNP_A-8332143	Y	20837553
SNP_A-8611029	Y	21080707
SNP_A-8595327	Y	21088297
SNP_A-8452521	Y	21159055
SNP_A-8290281	Y	21309376
SNP_A-8324813	Y	21409706
SNP_A-8352188	Y	21528257
SNP_A-8498938	Y	21535086
SNP_A-8591520	Y	21717208
SNP_A-8290282	Y	21722998
SNP_A-8348747	Y	21730257
SNP_A-8608816	Y	21784286
SNP_A-8412369	Y	21867787
SNP_A-8468019	Y	21917313
SNP_A-8469015	Y	21917832
SNP_A-8555638	Y	21983827
SNP_A-8331392	Y	21990257
SNP_A-8675562	Y	22003770
SNP_A-8657328	Y	22072097
SNP_A-8521142	Y	22072340
SNP_A-8604166	Y	22214221
SNP_A-8511463	Y	22346168
SNP_A-8474444	Y	22725379
SNP_A-8717529	Y	22796697
SNP_A-8682494	Y	22866703
SNP_A-8344910	Y	22914378
SNP_A-8573174	Y	22918577
SNP_A-8355942	Y	22934109
SNP_A-8651773	Y	22972939
SNP_A-8465368	Y	23040647
SNP_A-8298161	Y	23134896
SNP_A-8329413	Y	23443971
SNP_A-8525311	Y	23473201
SNP_A-8539904	Y	23631629
SNP_A-8561218	Y	23883529
SNP_A-8390603	Y	23984056
SNP_A-8496894	Y	23993156
SNP_A-8371607	Y	24359931
SNP_A-8635714	Y	24401940
SNP_A-8365189	Y	24475669
SNP_A-8681669	Y	28509790
SNP_A-8582662	Y	28606269
SNP_A-8383267	Y	28612323

SNP_A-8401720	Y	28733101
SNP_A-8433021	Y	28758193
SNP_A-8716074	Y	58969307
SNP_A-8645689	Y	58970351
SNP_A-8684492	Y	58977087
SNP_A-8627722	Y	58977828
SNP_A-8690438	Y	58997679
SNP_A-8636108	Y	59001292
SNP_A-8623399	Y	59001873
SNP_A-8509720	Y	59002462
SNP_A-8525858	Y	59005350
SNP_A-8455959	Y	59020166
SNP_A-8377945	Y	59020401
SNP_A-8406582	Y	59030720

Appendix 5
Single-nuclei processing code segments

In order to capture pre-mRNA transcripts that are common in the nuclei, a pre-mRNA transcriptome is required. The command below is applied to the transcriptome and outputs a pre-mRNA transcriptome. Essentially all transcripts (including intronic transcripts) are set to exonic. (Method advised by 10x Genomics)

```
awk 'BEGIN{FS="\t"; OFS="\t"} $3 == "transcript"{ $3="exon"; print}' \
       GRCh38-3.0.0.gtf > GRCh38-3.0.0.premrna.gtf
```

This modified GTF file (Gene transfer format; used to hold information about gene structure) is processed into a Cell Ranger package, along with the reference genome. This package is used for sequence alignment using Cell Ranger count.

```
cellranger mkref \
--genome=GRCh38-3.0.0_premrna \
--fasta=GRCh38_genome.fa \
--genes=GRCh38-3.0.0.premrna.gtf
```

```
cellranger count \
  --id=$subject \
  --fastqs=$fastqs \
  --sample=$sample_name \
  --transcriptome=$reference_genome
```

Below is the command used to filter single-nuclei RNAseq reads. The meaning of each flag is listed:

```
samtools view -F 1536 -q 255 $sample_cell_bam Y:2781480-56887902
```

-F flag filtering reads inversely based on a SAM flag value. In this case the flag 1536 refers to reads that fail quality checks, and PCR duplicate reads.

-q flag for filtering reads based on mapping quality (MAPQ). 255 is the maximum mapping quality provided by the STAR aligner. Therefore, only reads with maximum mapping quality are kept.

Y:2781480-56887902 - location of the male specific Y region (hg38)