SOMATIC COPY NUMBER VARIANT LOAD IN NEURONS OF HEALTHY CONTROLS AND ALZHEIMER'S DISEASE PATIENTS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ZELİHA GÖZDE TURAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
BIOLOGY

JANUARY 2023

Approval of the thesis:

SOMATIC COPY NUMBER VARIANT LOAD IN NEURONS OF HEALTHY CONTROLS AND ALZHEIMER'S DISEASE PATIENTS

submitted by **ZELİHA GÖZDE TURAN** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Biology Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Ayşe Gül Gözen
Head of Department, **Biology** _____

Prof. Dr. Mehmet Somel
Supervisor, **Department of Biological Sciences, METU** _____

Assist. Prof. Dr. Idil Yet
Co-supervisor, **Dep. of Bioinformatics, Hacettepe University** _____

Examining Committee Members:

Assoc. Prof. Dr. Çağdaş Devrim Son
Department of Biological Sciences, METU _____

Prof. Dr. Mehmet Somel
Department of Biological Sciences, METU _____

Assist. Prof. Dr. Aybar Can Acar
Health Informatics, METU _____

Assoc. Prof. Dr. İsmail Kudret Sağlam
Department of Molecular Biology and Genetics, Koç University _____

Assist. Prof. Dr. Arda Çetinkaya
Department of Medical Genetics, Hacettepe University _____

Date: 09.01.2023

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Zeliha Gözde Turan

Signature :

ABSTRACT

SOMATIC COPY NUMBER VARIANT LOAD IN NEURONS OF HEALTHY CONTROLS AND ALZHEIMERS DISEASE PATIENTS

Turan, Zeliha Gözde
Ph.D., Department of Biology
Supervisor: Prof. Dr. Mehmet Somel
Co-Supervisor: Assist. Prof. Dr. Idil Yet

January 2023, 156 pages

The possible role of somatic copy number variations (CNVs) in Alzheimer's disease (AD) aetiology has been controversial. Although cytogenetic studies suggested increased CNV loads in AD brains, a recent single-cell whole-genome sequencing (scWGS) experiment, studying frontal cortex brain samples, found no such evidence. Here we readdressed this issue using low-coverage scWGS on pyramidal neurons dissected via both laser capture microdissection (LCM) and fluorescence activated cell sorting (FACS) across five brain regions: entorhinal cortex, temporal cortex, hippocampal CA1, hippocampal CA3, and the cerebellum. Among reliably detected somatic CNVs identified in 1301 cells obtained from the brains of 13 AD patients and 7 healthy controls, deletions were more frequent compared to duplications. Interestingly, we observed slightly higher frequencies of CNV events in cells from AD compared to similar numbers of cells from controls (4.1% vs. 1.4%, or 0.9% vs. 0.7%, using different filtering approaches), although the differences were not statistically significant. On the technical aspects, we observed that LCM-isolated cells show higher within-cell read depth variation compared to cells isolated with FACS. To re-

duce within-cell read depth variation, we proposed a principal component analysis-based denoising approach that significantly improves signal-to-noise ratios. Lastly, we showed that LCM-isolated neurons in AD harbour slightly more read depth variability than neurons of controls, which might be related to the reported hyperploid profiles of some AD-affected neurons.

Keywords: Single-cell whole-genome sequencing, Copy number variation, Alzheimer's disease, Brain, Laser capture microdissection, Fluorescence-activated cell sorting, Denoising

ÖZ

SAĞLIKLI KONTROLLERİN VE ALZHEİMER HASTALARININ NÖRONLARINDA KOPYA SAYISI VARYASYONU YÜKÜNÜN BELİRLENMESİ

Turan, Zeliha Gözde
Doktora, Biyoloji Bölümü
Tez Yöneticisi: Prof. Dr. Mehmet Somel
Ortak Tez Yöneticisi: Dr. Öğr. Üyesi. Idil Yet

Ocak 2023 , 156 sayfa

Alzheimer hastalığı (AH) etiyolojisinde somatik kopya sayısı varyasyonlarının (KSV) olası rolü tartışmalıdır. Sitogenetik çalışmalar AH beyinlerinde artan CNV yükünü rapor etsede, son zamanlarda yapılan ve tekil hücre tüm genom dizilemesi yöntemi kullanarak frontal korteks beyin örneklerini inceleyen çalışma böyle bir bulgu rapor etmemiştir. Bu çalışmada, beş farkli beyin bölgesinde (entorhinal korteks, temporal korteks, hipokampal CA1, hipokampal CA3 ve beyincik) hem lazerle kesme ve yakalama (LCM) hem de floresan ile aktive edilen hücre sınıflandırılması (FACS) yöntemiyle diseke edilen piramidal nöronlar üzerinde düşük kapsamlı tekil hücre tüm genom dizilemesi yöntemi kullanarak bu soruyu yeniden ele aldık. 13 AH hastasının ve 7 sağlıklı kontrolün beyinlerinden elde edilen 1301 hücrede güvenilir şekilde saptanan somatik KSV'ler arasında, silinmeler, kazanimlara kıyasla daha sıktır. İlginç bir şekilde, kontrollerden alınan benzer sayıda hücreye kıyasla (farklı filtreleme yaklaşımları kullanılarak %4,1'e karşı %1.4 veya %0,9'a karşı %0,7), farklılıklar istatistiksel olarak olmasa da, AH'deki hücrelerde CNV olaylarının biraz daha yük-

sek sıklıkta gözlemledik. Teknik açıdan, LCM ile izole edilmiş hücrelerin, FACS ile izole edilmiş hücrelere kıyasla hücre içi okuma derinliği varyasyonunun daha yüksek olduğunu gözlemledik. Hücre içi okuma derinliği varyasyonunu azaltmak için, sinyal-gürültü oranlarını önemli ölçüde iyileştiren temel bileşen analizine dayalı gürültü giderme yaklaşımı önerdik. Son olarak, AH'deki LCM ile izole edilmiş nöronların, kontrol nöronlarından biraz daha fazla okuma derinliği değişkenliği barındırdığını gösterdik; bu, AH'den etkilenen bazı nöronların rapor edilen hiperploid profilleriyle ilgili olabilir.

Anahtar Kelimeler: Tekil hücre tüm genom dizilemesi, Kopya sayısı varyasyonu, Alzheimer hastalığı, Beyin, Lazer yakalayıcı mikro diseksiyon, Floresan ile aktive edilen hücre sınıflandırılması, Gürültü azaltma

To everyone who has struggled to fit in

ACKNOWLEDGMENTS

First of all, I would like to thank my lovely supervisor Mehmet Somel for his patience, support and guidance during my PhD. After nine years working together, I still admire his knowledge and unique thought processes. Starting from day one, he treats his students as colleagues. The most important thing that I learnt from him was that hierarchy does not have to exist in academia. I hope I will be as patient with my own students as he was with me.

I want to thank my co-advisor İdil Yet for her support, encouragement and motivation. I also thank her for sharing her experience with me. I really appreciate her positive attitude throughout my PhD.

I am grateful to my dissertation monitoring committee members, İsmail Kudret Sağlam and Çağdaş Devrim Son for their valuable times they spent on my thesis. Their perspectives helped me to look the concepts quite differently. Also, I would like to thank Aybar Can Acar and Arda Çetinkaya for their valuable times they spent on my thesis.

I would like to acknowledge several people for their efforts: Poorya Parvizi, Ulaş Işıldak and Etka Yapar. Without your contributions, I could not have completed this work. I want to thank all my colleagues at the METU CompEvo group for their support, help and being there to answer my questions. I am really lucky to be part of such a great team.

I am grateful to the Ecology and Evolutionary Biology Society (Turkey) for their efforts in filling the gaps in science despite the limited resources.

I would like to thank Uwe Ueberham and Thomas Arendt for data generation. I also want to thank Scientific and Technological Research Council of Turkey (ERANET-RUS, 215Z495), ERA.NetRus grant "significans" (coordinator: Thomas Arendt, part-

ners: Ivan Iourov, Mehmet Somel, Charles Duyckaerts) and Alzheimer Forschung Initiative (AFI, 17021, Uwe Ueberham) for supporting this project. I would like to thank YÖK PhD fellowship.

I would like to thank my spouse, Anıl Turan, for his encouragement throughout my PhD. He kept me continue in the times of crisis. Without his constant source of support, I might not have finished this work. I would like to thank Bahar Karagoz and my family for all their support and love. Without you, none of this would have been possible.

I owe special thanks to my dear friends Meriç Erdolu, Melih Yıldız, Dilek Koptekin, Ahmet Yetkin Alici, Gökçen Deniz Özen, Hamit İzgi, İrem Güner, Ulaş Işıldak, Melfi Güner-Işıldak and Poorya Parvizi for their love and friendships. Knowing that you are there for me is a great support, and I am really excited to find out what the climax of our next adventure will be.

Words can not express my love for my cat babies, Yoda, Anakin and Leia. I love every minute that I spend with you. I love moving together from room to room and I love sharing our food, usually mine. You are amazing and when I look at your faces, I can not stop thinking about how beautiful you are. I want to thank all of you for adopting me and making my world better.

TABLE OF CONTENTS

ABSTRACT v

ÖZ vii

ACKNOWLEDGMENTS x

TABLE OF CONTENTS xii

LIST OF TABLES xvii

LIST OF FIGURES xxii

LIST OF ABBREVIATIONS xxviii

CHAPTERS

1 INTRODUCTION 1

1.1 Alzheimer's disease 3

1.2 Types of Alzheimer's disease 3

1.3 Stages of Alzheimer's Disease 4

1.4 Diagnosis of Alzheimer's Disease 5

1.5 Contributors of Alzheimer's Disease 7

1.6 Origin of CNVs and aneuploidies in the brain 10

1.7 Experimental methods to study aneuploidy 11

1.8 The role of aneuploidy in the healthy brain 13

1.9 The role of aneuploidy in AD brain 15

1.10 Research Objectives 18

2 MATERIAL AND METHODS 21

2.1 Wet lab experiments 21

2.1.1 Ethics approval and consent to participate 21

2.1.2 Tissue sources 22

2.1.3 Methods for single-cell isolation 22

2.1.4 scWGS library preparation and sequencing 23

2.2 Data analyses pipeline for scWGS data 23

2.2.1 NGS data analysis 24

2.2.1.1 Read quality control and alignment 24

2.2.1.2 Filtering 24

2.2.1.3 Coverage 25

2.2.2 CNV prediction and cell elimination 25

2.3 Published datasets 30

2.4 Statistical modeling of CNV frequencies and index of dispersion (IOD) levels 30

2.5 Copy number statistics 31

2.6 Principal component analysis (PCA) 32

2.7 Data access 33

3 RESULTS 35

3.1 Summary of the dataset 35

3.2 LCM-isolated cells show a high frequency of depth variability 38

3.3 No difference in CNV frequency between AD and control in the "uncorrected-filtered" dataset 42

3.4 A PCA-based denoising approach minimizes within-cell depth variability . . . 44

3.5 Subchromosomal CNVs are enriched in deletions in the PCA-corrected data . . . 50

3.6 No significant difference between AD and control after PCA-correction or in the van den Bos dataset . . . 51

4 DISCUSSION . . . 57

4.1 The sources of variability among LCM-isolated cells . . . 57

4.2 PCA-based denoising: advantages and caveats . . . 59

4.2.1 Possible application of PCA-based denoising approach in other fields of biology . . . 60

4.3 Limitations and Possible Improvements . . . 61

5 CONCLUSION . . . 65

REFERENCES . . . 67

A INFORMATION ABOUT THE SIGNIFICANT CNVS IN THE UNCORRECTED DATA . . . 95

B INFORMATION ABOUT THE SIGNIFICANT CNVS IN THE CORRECTED DATA . . . 97

C INFORMATION ABOUT THE SIGNIFICANT CNVS IN VAN DEN BOS DATA . . . 99

D THE EFFECTS OF DIAGNOSIS AND COVERAGE ON THE FREQUENCY OF CNVS IN THE UNCORRECTED-FILTERED DATA . . . 101

E THE EFFECTS OF DIAGNOSIS, BRAIN REGIONS, SEX AND COVERAGE ON THE FREQUENCY OF CNVS IN THE UNCORRECTED-FILTERED DATA . . . 103

F THE EFFECTS OF CHROMOSOME, DIAGNOSIS, BRAIN REGIONS, SEX AND COVERAGE ON THE FREQUENCY OF CNVS IN THE UNCORRECTED-FILTERED DATA . . . 105

G THE EFFECTS OF CHROMOSOME, DIAGNOSIS, BRAIN REGIONS AND COVERAGE ON THE FREQUENCY OF CNVS IN THE UNCORRECTED-FILTERED DATA 107

H THE EFFECTS OF CHROMOSOME, DIAGNOSIS, BRAIN REGIONS AND COVERAGE ON THE FREQUENCY OF CNVS IN THE UNCORRECTED-FILTERED DATA 109

I THE EFFECTS OF CHROMOSOME, DIAGNOSIS AND COVERAGE ON THE FREQUENCY OF CNVS IN THE UNCORRECTED-FILTERED DATA 111

J THE EFFECTS OF CHROMOSOME, DIAGNOSIS AND COVERAGE ON THE FREQUENCY OF CNVS IN THE UNCORRECTED-FILTERED DATA 113

K THE EFFECTS OF DIAGNOSIS AND COVERAGE ON THE FREQUENCY OF CNVS IN THE PCA-CORRECTED DATA 115

L THE EFFECTS OF CHROMOSOME, DIAGNOSIS, BRAINS REGIONS, COVERAGE AND SEX ON THE FREQUENCY OF CNVS IN THE PCA-CORRECTED DATA 117

M THE EFFECTS OF CHROMOSOME, DIAGNOSIS, BRAINS REGIONS, COVERAGE AND SEX ON THE FREQUENCY OF CNVS IN THE PCA-CORRECTED DATA 121

N THE EFFECTS OF CHROMOSOME, DIAGNOSIS, BRAINS REGIONS AND COVERAGE ON THE FREQUENCY OF CNVS IN THE PCA-CORRECTED DATA 123

O THE EFFECTS OF CHROMOSOME, DIAGNOSIS, BRAINS REGIONS AND COVERAGE ON THE FREQUENCY OF CNVS IN THE PCA-CORRECTED DATA 127

P THE EFFECTS OF CHROMOSOME, DIAGNOSIS AND COVERAGE ON THE FREQUENCY OF CNVS IN THE PCA-CORRECTED DATA . . 129

Q THE EFFECTS OF CHROMOSOME, DIAGNOSIS AND COVERAGE ON THE FREQUENCY OF CNVS IN THE PCA-CORRECTED DATA . . 131

R THE EFFECTS OF DIAGNOSIS AND COVERAGE ON THE FREQUENCY OF CNVS IN VAN DEN BOS DATA 133

S THE EFFECTS OF CHROMOSOME, DIAGNOSIS AND COVERAGE ON THE FREQUENCY OF CNVS IN VAN DEN BOS DATA 135

T THE EFFECTS OF CHROMOSOME, DIAGNOSIS AND COVERAGE ON THE FREQUENCY OF CNVS IN VAN DEN BOS DATA 137

U THE EFFECTS OF CHROMOSOME, DIAGNOSIS, BRAIN REGIONS AND COVERAGE ON THE FREQUENCY OF CNVS IN THE UNCORRECTED DATA 139

V THE EFFECTS OF CHROMOSOME, DIAGNOSIS, BRAIN REGIONS, COVERAGE AND SEX ON THE FREQUENCY OF CNVS IN THE UNCORRECTED DATA 143

W THE EFFECTS OF CHROMOSOME, DIAGNOSIS AND COVERAGE ON THE FREQUENCY OF CNVS IN THE UNCORRECTED DATA . . . 147

X THE EFFECTS OF CHROMOSOME, DIAGNOSIS, BRAIN REGIONS AND COVERAGE ON THE FREQUENCY OF CNVS IN THE UNCORRECTED DATA 149

Y THE EFFECTS OF CHROMOSOME, DIAGNOSIS, BRAIN REGIONS, COVERAGE AND SEX ON THE FREQUENCY OF CNVS IN THE UNCORRECTED DATA 151

Z THE EFFECTS OF CHROMOSOME, DIAGNOSIS AND COVERAGE ON THE FREQUENCY OF CNVS IN THE UNCORRECTED DATA . . . 153

CURRICULUM VITAE 155

LIST OF TABLES

TABLES

Table 1.1 Summary of the scWGS studies in healthy brain. 19

Table 3.1 Summary of clinical and demographic variables of samples analyzed. 36

Table A.1 Id: The id of the fastq file; Chr: The name of the chromosome that CNV is located; Start, End: The start and end position of the CNV; z2_score: Z-score for the deviation from the expectation; z1_score: Z-score for each CNV that reflects the deviation in read count distribution; Diagnosis: The state of the individual. 95

Table B.1 Id: The id of the fastq file; Chr: The name of the chromosome that CNV is located; Start: The start position of the CNV; End: The end position of the variation; Cnv: The integer copy number state of the copy number variation; z2_score: Z-score for the deviation from the expectation (1-somy or 3-somy); z1_score: Z-score for each CNV that reflects the deviation in read count distribution; Diagnosis: The state of the individual that cell was taken from. 97

Table C.1 Id: The id of the fastq file; Chr: The name of the chromosome that CNV is located; Start: The start position of the CNV; End: The end position of the variation; Cnv: The integer copy number state of the copy number variation; z2_score: Z-score for the deviation from the expectation (1-somy or 3-somy); z1_score: Z-score for each CNV that reflects the deviation in read count distribution; Diagnosis: The state of the individual that cell was taken from. 99

Table D.1 The table shows the results of the GLMM that diagnosis and coverage were added as fixed factors to predict the frequency of CNV. 101

Table E.1 The table shows the results of the GLMM that diagnosis, brain regions, sex and coverage were added as fixed factors to predict the frequency of CNV. 103

Table E.1 Table E (continued) 104

Table F.1 The table shows the results of the GLMM that chromosome, diagnosis, brain regions, sex and coverage were added as fixed factors to predict the frequency of CNV. 105

Table F.1 Table F (continued) 106

Table G.1 The table shows the results of the GLMM that interaction of chromosome and diagnosis, brain regions, sex and coverage were added as fixed factors to predict the frequency of CNV. 107

Table G.1 Table G (continued) 108

Table H.1 The table shows the results of the GLMM that chromosome, diagnosis, brain regions and coverage were added as fixed factors to predict the frequency of CNV. 109

Table H.1 Table H (continued) 110

Table I.1 The table shows the results of the GLMM that interaction of chromosome and diagnosis and coverage were added as fixed factors to predict the frequency of CNV. 111

Table I.1 Table I (continued) 112

Table J.1 The table shows the results of the GLMM that chromosome, diagnosis and coverage were added as fixed factors to predict the frequency of CNV. 113

Table J.1 Table J (continued) 113

Table K.1 The table shows the results of the GLMM that diagnosis and coverage were added as fixed factors to predict the frequency of CNV. 115

Table L.1 The table shows the results of the GLMM that interaction of chromosome and diagnosis, brain regions, sex and coverage were added as fixed factors to predict the frequency of CNV. 117

Table L.1 Table L (continued) 118

Table L.1 Table L (continued) 119

Table M.1 The table shows the results of the GLMM that chromosome, diagnosis, brain regions, sex and coverage were added as fixed factors to predict the frequency of CNV. 121

Table M.1 Table M (continued) 122

Table N.1 The table shows the results of the GLMM that interaction of chromosome and diagnosis, brain regions and coverage were added as fixed factors to predict the frequency of CNV. 123

Table N.1 Table N (continued) 124

Table N.1 Table N (continued) 125

Table O.1 The table shows the results of the GLMM that chromosome, brain regions and coverage were added as fixed factors to predict the frequency of CNV. 127

Table O.1 Table O (continued) 128

Table P.1 The table shows the results of the GLMM that interaction of chromosome and diagnosis and coverage were added as fixed factors to predict the frequency of CNV. 129

Table P.1 Table P (continued) 130

Table Q.1 The table shows the results of the GLMM that chromosome, diagnosis and coverage were added as fixed factors to predict the frequency of CNV. 131

Table Q.1 Table Q (continued) 132

Table R.1 The table shows the results of the GLMM that diagnosis and coverage were added as fixed factors to predict the frequency of CNV. 133

Table S.1 The table shows the results of the GLMM that interaction of chromosome and diagnosis and coverage were added as fixed factors to predict the frequency of CNV. 135

Table S.1 Table S (continued) 136

Table T.1 The table shows the results of the GLMM that chromosome, diagnosis and coverage were added as fixed factors to predict the frequency of CNV. 137

Table T.1 Table T (continued) 138

Table U.1 The table shows the results of the GLMM that interaction of chromosome and diagnosis, brains regions and coverage were added as fixed factors to predict the frequency of CNV. 139

Table U.1 Table U (continued) 140

Table U.1 Table U (continued) 141

Table V.1 The table shows the results of the GLMM that interaction of chromosome and diagnosis, brains regions, sex and coverage were added as fixed factors to predict the frequency of CNV. 143

Table V.1 Table V (continued) 144

Table V.1 Table V (continued) 145

Table W.1 The table shows the results of the GLMM that interaction of chromosome and diagnosis and coverage were added as fixed factors to predict the frequency of CNV. 147

Table W.1 Table W (continued) 148

Table X.1 The table shows the results of the GLMM that chromosome, diagnosis, brain regions and coverage were added as fixed factors to predict the frequency of CNV. 149

Table X.1 Table X (continued) 150

Table Y.1 The table shows the results of the GLMM that chromosome, diagnosis, brain regions, sex and coverage were added as fixed factors to predict the frequency of CNV. 151

Table Y.1 Table Y (continued) 152

Table Z.1 The table shows the results of the GLMM that chromosome, diagnosis and coverage were added as fixed factors to predict the frequency of CNV. 153

Table Z.1 Table Z (continued) 154

LIST OF FIGURES

FIGURES

Figure 1.1 Pathological hallmarks of AD. The upper panel shows a healthy brain. The lower panel illustrates the AD-affected brain. The figure was modified and taken from https://biorender.com/. 4

Figure 1.2 Structure of the human brain. **(A)** lateral view of left hemisphere, **(B)** medial view of right hemisphere, and **(C)** sagittal view of one hemisphere. The figure was taken from [1]. 5

Figure 1.3 Steps of the amyloid cascade hypothesis. The figure was taken from [2]. 8

Figure 2.1 The pipeline of NGS data analysis and CNV detection. 23

Figure 2.2 Steps of circular binary segmentation. Figure was copied from https://github.com/melissayan/vnowchi, and to increase the resolution, redrawed with https://biorender.com/. 28

Figure 2.3 Workflow of *Ginkgo*. Figure was taken from [3]. 29

Figure 3.1 Schematic of the workflow. The pipeline of NGS data analysis and CNV detection. Figure was created with https://biorender.com/ 35

Figure 3.2 Sample information. **(A)** Bar plot showing the number of cells that have been sequenced for each individual. Brain regions are illustrated in different colors (see the colour key on the top of the figure). **(B)** Dot plot showing age of AD (pink) and control (green) individuals. **(C)** The table summarises sex, diagnoses and Braak level of the individuals. The Braak stages of AD patients ranged between III and VI. 37

Figure 3.3 Images from a frozen hippocampal brain slice stained with cresyl-violet showing a pyramidal cell before (left) and after (right) laser capture microdissection-based isolation process using the PALM device. Circles in (left) indicate positions where two pyramidal cells have already been isolated just prior to the picture being taken. Scale Bar, 50 µm 38

Figure 3.4 Median copy number profile of chromosome X or 21 across datasets. Median of the distributions: McConnell Female: 1.94, McConnell Male: 0.96; van den Bos Female: 1.91, van den Bos Female: 0.94, van den Bos Trisomy: 3.00; This paper (Uncorrected) Female: 1.60, This paper (Uncorrected) Male: 1.00. 39

Figure 3.5 Comparison between different cell isolation methods. Box plots showing the distribution of coverage **(A)**, mapping proportion **(B)**, index of dispersion **(C)** among FACS-isolated, LCM-isolated and LCM-isolated blank samples. P-values were calculated using Kruskal–Wallis test among groups and Wilcoxon rank-sum test between groups. 41

Figure 3.6 The distribution of index of dispersion ("IOD") for LCM-isolated cells ($n = 1542$) according to **(A)** diagnoses (Alzheimer's disease ("AD"), control), **(B)** brain regions (hippocampal subfields CA1 ("Hippocampal CA1"), hippocampal subfields CA3 ("Hippocampal CA3"), Entorhinal cortex, Cerebellum, Temporal cortex). For each brain region, we tested whether AD diagnosis was predictive of IOD using a linear mixed-effects (lme) model. Individuals were added as a random factor. Across all tested brain regions, differences were only marginally significant ($p = 0.069$). **(C)** The distribution of IOD across individuals ($n = 20$). Box plots were ordered by the median. Y-axes illustrate the IOD values on the log10 scale. 43

Figure 3.7 Examples of CNVs that (upper panels) failed to pass and (lower panels) passed the filtering criteria (Z_1-scores $\geq$2 and Z_2-scores $\leq$0.5) in the uncorrected-filtered data **(A)** and PCA-corrected data **(B)**. Absolute values of the Z_1- and Z_2 scores for the CNV were indicated on the plot. Mentioned CNV was marked with a red star. X-axes show chromosomes and Y-axes illustrate the CN profile of chromosomes estimated by $Ginkgo$. Each grey dot represents the scaled read counts per bin. Amplifications (CN>2) are shown in red; deletions (CN<2) in blue; disomy (CN=2) in black. 45

Figure 3.8 Overview of cell and CNVs elimination steps in the uncorrected and corrected data. *If CNVs' breakpoints were within the three base pairs window around each other, those were discarded. Figure was created with `https://biorender.com/` 46

Figure 3.9 The heatmap shows the genome-wide copy number profile of cells analyzed in the uncorrected data ($n = 15$) with at least one reliable CNV. CNVs, brain regions and diagnoses are illustrated in different colors (see the colour key on the left of the figure). Each row shows a cell and each column shows a chromosome. 46

Figure 3.10 Boxplots showing the distribution of median CN of chromosome 1 (chr1, upper part of the figure) and chromosome 21 (chr21, lower part of the figure) across bins ($n = 440$ and $n = 68$ for chr1 and chr21, respectively). Each point corresponds to the median CN of each cell. Minimum ("Min"), median ("Med"), maximum ("Max") and standard deviation ("sd") of each distribution were shown on the boxplot. Cells that deviated from the [1.9-2] range were excluded from the analyses to be consistent with our filtering criteria (except for the uncorrected datasets). This study [Uncorrected ($n = 1337$), Uncorrected-filtered ($n = 588$), PCA-corrected ($n = 1301$)]: blue; van den Bos 2016 ($n = 1468$): brown; McConnell 2013 ($n = 109$): purple. 47

Figure 3.11 A principal components analysis (PCA) was performed using the normalized read counts across autosomal bins ($n = 5243$) in published datasets and this study. Because they dominated the PCs, cells deviating from the [1.9-2] range were not included in the analyses. The number of cells for each dataset were indicated on the plot. X-axes illustrate PC1 and PC3 that explain 18.4% and 1.3% of the total variance, respectively. Y-axes show PC2 and PC4 that explain 2.8% and 0.9% of the total variance, respectively. 48

Figure 3.15 Properties of published datasets and this study. **(A)** Violin plot showing the distribution of coverage among different datasets. This study, including only LCM: blue; van den Bos 2016: brown; McConnell 2013: purple. **(B)** Bar plot showing the number of CNVs per cell across datasets. The datasets from our study include cells from both AD and control (Uncorrected, Uncorrected-filtered, PCA-corrected): blue; van den Bos 2016 (including cells from both AD and control): brown; McConnell 2013: purple). The number of cells that were used to calculate CNVs per cell was shown on the X-axis label. 49

Figure 3.16 Distributions of the number of CN before ($n = 3,521$) and after correction ($n = 1298$) across autosomes. For illustration purposes, the bar plot includes up to 10-somy. 50

Figure 3.17 Bar plots represent the number of CNVs per cell for AD and control groups. Blue area: 1-somy (deletions); pink area: 3-somy (duplications). 51

Figure 3.12 Accuracy of the PCA-based denoising approach. The sample ID is ERR1378095. Examples of CN estimates of cells using uncorrected data **(A)** and using data after PCA-based correction **(B-F)**. The x-axes show chromosomes and the y-axes show the CN profile of chromosomes estimated by $Ginkgo$. Each grey dot represents the scaled and normalized read counts per bin. Amplifications (CN>2) are shown in red; deletions (CN<2) in blue; disomy (CN=2) in black. 52

Figure 3.13 A PCA-based denoising approach helps eliminate technical noise. **(A-C)** Examples of CN estimates of cells using uncorrected data (upper panels) and using data after PCA-based correction (lower panels). The x-axes show chromosomes and the y-axes show the CN profile of chromosomes estimated by $Ginkgo$. Each grey dot represents the scaled and normalized read counts per bin. Amplifications (CN>2) are shown in red; deletions (CN<2) in blue; disomy (CN=2) in black. 53

Figure 3.14 CN profiles of three cells deviating from the range of 1.9 and 2 after correction. The upper panels show the uncorrected data, and the lower panels show the corrected data. X-axes show chromosomes and Y-axes illustrate the CN profile of chromosomes estimated by $Ginkgo$. Amplifications (CN>2) are shown in red; deletions (CN<2) in blue; disomy (CN=2) in black. 54

Figure 3.18 The heatmap shows the genome-wide copy number profile of cells analyzed in the corrected data ($n = 10$) with at least one reliable CNV. CNVs, brain regions and diagnoses are illustrated in different colors (see the colour key on the left of the figure). Each row shows a cell and each column shows a chromosome. 55

Figure 3.19 Information about the samples of uncorrected and corrected data. Distribution of samples according to sex (female, male) **(A)**, diagnoses (Alzheimer's Disease, control) **(B)**, brain regions (temporal cortex, cerebellum, entorhinal cortex, hippocampal subfields CA1, hippocampal subfields CA3) **(C)** and Braak level (stage 3, 4, 5 and 6) **(D)**. 55

Figure 4.1 Copy number profile clonal lung tumours. X-axis shows chromosomes, Y-axis shows low ratio of the read counts. The plot was taken from [4]. 62

LIST OF ABBREVIATIONS

CB	Cerebellum
CN	Copy Number
SoS	Sum of squares
CA	Cornu Ammonis
BAM	Binary SAM file
EC	Entorhinal Cortex
IOD	Index Of Dispersion
AD	Alzheimer's Disease
HMM	Hidden Markov Model
SRA	Sequence Read Archive
CNV	Copy Number Variation
SAM	Sequence Alignment/Map
BWA	Burrows-Wheeler Alignment
NGS	Next Generation Sequencing
BED	Browser Extensible Data file
CBS	Circular Binary Segmentation
PCA	Principal Component Analysis
FISH	Fluorescence In Situ hybridization
FACS	Fluorescence Activated Cell Sorting
scWGA	Single Cell Whole Genome Analysis
scWGS	Single Cell Whole Genome Sequencing
LOWESS	Locally WEighted Scatterplot Smoothing

CHAPTER 1

INTRODUCTION

The outline of the thesis is as follows.

Chapter 1, introduction, gives general information about Alzheimer's disease (AD), including pathological and clinical hallmarks and types of AD (early-onset, late-onset). Also I answered the following questions:

- What are AD stages and how are AD stages determined?
- How does AD affect different brain regions?
- What are the neurochemical markers and neuroimaging techniques to diagnose AD?
- What are the contributors to AD and how do they contribute to the AD phenotype?
- What is aneuploidy and where does it stem from in post-mitotic neurons?
- What are the experimental methods to count aneuploidy in mitotic and post-mitotic cells?
- What is the role of aneuploidy in healthy and AD brains?
- What was the aim of this study?

Chapter 2, method, gives detailed information about web-lab experiments that were conducted at the Paul-Flechsig-Institute (Leipzig University, Germany), and dry-lab analyses. The steps/parameters of the bioinformatic analyses were described in detail, including:

- Quality control and alignment of single-cell whole genome sequencing reads to the reference genome
- Post-processing of alignment files
- Coverage calculation
- Steps of the CNV prediction and cell elimination: binning, coverage normalization, GC bias correction, segmentation, post hoc adjustment.
- Analyses of published datasets
- Statistical models to test the difference between AD and control
- PCA-based denoising approach and its pseudocode

Chapter 3, results, contains the results of the bioinformatic analyses, including:

- Summary of the produced data and clinical and demographic properties of samples
- Comparison between FACS-isolated and LCM-isolated cells
- Testing the difference between AD control using different cell filters and CNV thresholds
- The results of the PCA-based denoising approach

Chapter 4, discussion, discuss the findings of the study. These are:

- What would be the biological and technical origins of the high noise in the data?
- The possible approaches that could be taken to disentangle biological variation from technical one
- Pros and cons of PCA-based denoising approach
- Limitations and possible improvements of the study
- The reason for using the $Ginkgo$ algorithm

Chapter 5 is conclusion that summarizes the entire work.

1.1 Alzheimer's disease

Alzheimer's disease (AD) is a neurodegenerative disease of multifactorial aetiology, with numerous genetic and environmental factors each explaining a small proportion of variance in disease onset and progression [5].

AD affects more than 20 million people worldwide and it is the most common cause of irreversible dementia [6]. Indeed, the increasing human lifespan is expected to be reflected in the high incidence of AD within the next 50 years [7].

1.2 Types of Alzheimer's disease

AD is pathologically characterized by the accumulation of extracellular amyloid β (Aβ) plaques and intracellular neurofibrillary tangles (Figure 1.1). These pathological alterations are clinically reflected as memory loss and executive function deficit, followed by death after an average of nine years of diagnosis [8].

Even though they are clinically indistinguishable, there are two types of AD. Familial or early-onset Alzheimer's disease (FAD) accounts for ~5% of AD cases and affects individuals below the age of 60. FAD is mainly associated with causative mutations in three genes: *APP* (on chromosome 21) encoding for amyloid precursor protein, *PSEN1* (on chromosome 14) encoding for Presenilin 1, and *PSEN2* (on chromosome 1) encoding for Presenilin 2 [9, 10]. AD-causing mutations on these genes are highly penetrant and inherited in an autosomal dominant manner [10].

Sporadic or late-onset Alzheimer's disease (LOAD) manifests itself after the age of 60 and accounts for over 95% of AD cases [11]. The estimated heritability of the disease changes between 40 and 80% [12, 13, 14]. Studies suggested that genetic variants showing low penetrance with high prevalence are the main contributors to LOAD. For example, genome-wide association studies (GWAS) have identified dozens of genetic loci that could be associated with AD [15].

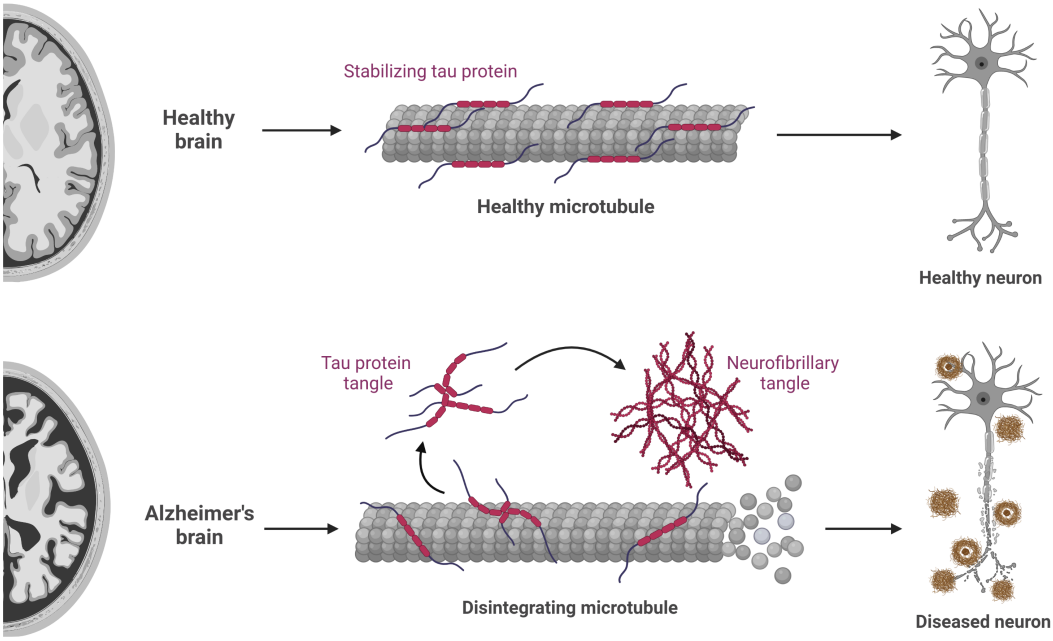


Figure 1.1: Pathological hallmarks of AD. The upper panel shows a healthy brain. The lower panel illustrates the AD-affected brain. The figure was modified and taken from `https://biorender.com/`.

GWAS can be a powerful method to identify single nucleotide polymorphisms (SNPs) associated with human diseases and traits. SNPs hitting protein-coding regions show their impact on phenotype by changing protein function and can be characterized by high penetrance and predictable inheritance patterns as in FAD. Most GWAS-identified SNPs (~88%), nevertheless, lie in non-coding parts of the DNA, and the transition from association to causation is hampered by both the linkage disequilibrium structure of the genome (correlation among SNPs) and their small effect size on phenotype [16]. GWAS-identified loci will be discussed in the section 1.5.

1.3 Stages of Alzheimer's Disease

The severity of AD can be determined using the location of the neurofibrillary tangles in the brain, which is called Braak staging (I–VI) [17].

AD does not affect each brain region equally. At Braak stages I and II, neurofib-

rillary tangles could be found in the entorhinal cortex and hippocampus. At Braak stages III and IV, neurofibrillary tangles are distributed throughout the limbic system and temporal cortex. At the latest stages of the disease, Braak stages V-VI, tangles spread through the primary sensory cortices [18, 19].

Accumulation of tangles also manifests itself as impairment in cognitive and executive functions. For example, the hippocampus, entorhinal cortex and cingulate gyrus are the heavily affected brain regions (Figure 1.2) [20]. The hippocampus and entorhinal cortex are part of the temporal lobe and have a role in memory formation. The cingulate gyrus is a part of the limbic lobe and involves emotional regulation. By Braak stages III-IV, these three brain regions are already affected. Thus, individuals with AD experience mood changes, memory impairment and face recognition deficits [18].

Beyond neurofibrillary tangles, Aβ deposition is also used to determine the severity of AD (stage A-C). Aβ deposition starts in the frontal and temporal lobes and spreads through the other part of the cerebral cortex [18, 21].

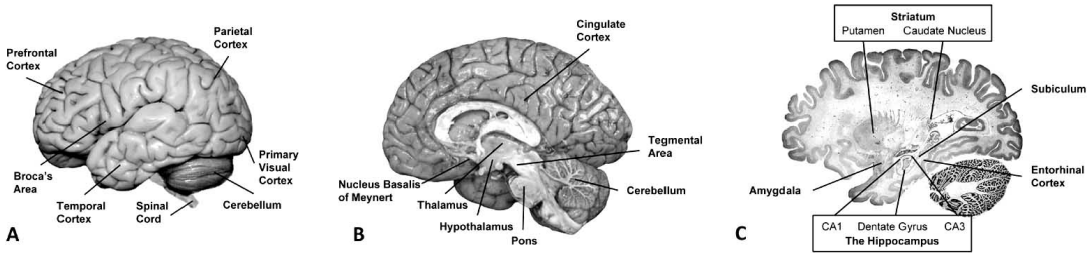


Figure 1.2: Structure of the human brain. **(A)** lateral view of left hemisphere, **(B)** medial view of right hemisphere, and **(C)** sagittal view of one hemisphere. The figure was taken from [1].

1.4 Diagnosis of Alzheimer's Disease

Neuropathological lesions of AD (i.e. accumulation of Aβ extracellular plaques and intracellular neurofibrillary tangles) usually manifest themselves earlier than clinical

symptoms [22, 23]. To take advantage of potential therapies, diagnosing AD in the earlier stages of the disease is crucial [24, 25]. For this aim, several neurochemical markers and neuroimaging techniques have been employed [26].

Neurochemical markers can be used to identify pathological lesions. For example, reduced Aβ42 levels in the cerebrospinal fluid (CSF) is the biomarker of the Aβ plaques [24]. Similarly, an increased concentration of hyperphosphorylated tau protein in the CSF is the marker of neurofibrillary tangles [27].

Magnetic resonance imaging (MRI) is also widely used to detect hallmarks of neurodegeneration. For example, atrophy of the hippocampus and entorhinal cortex is an important structural marker of AD [28, 29, 30, 31]. Other structural changes that neuroimaging techniques could detect are reduction in the cortical grey matter and in the cortical thickness of the temporal regions [32, 33, 34]. Beyond structural changes, one functional magnetic resonance imaging (fMRI) study showed that AD patients had reduced functional connectivity in the posterior cingulate cortex (have a role in retrieving autobiographical memories) and other parts of the brain [35].

In contrast to fMRI, positron emission tomography (PET) is an invasive technique that measures the metabolic activity of the cerebral cortex using fluorodeoxyglucose (FDG). FDG behave similarly to glucose and accumulate in the brain according to the glucose metabolic rate. PET studies showed that AD patients had decreased metabolic rates of glucose in the temporal and parietal association cortices [36]. On the other hand, the brain regions least affected by AD (i.e. primary motor cortex, visual cortex, cerebellum) show no sign of impairment in glucose metabolic rate [37].

In summary, multiple biomarkers and anatomical changes can be used to study AD development. It is possible that combination of markers (i.e. Aβ42 and hyperphosphorylated tau in the CSF, cerebral glucose metabolism, hippocampal volumetry) will increase the probability of detecting people with mild cognitive impairment who are likely to develop AD [27].

1.5 Contributors of Alzheimer's Disease

One of the leading contributors to AD is the aggregation of Aβ peptides as a form of amyloid plaques in the extracellular spaces in the brain [38]. The *APP* gene on chromosome 21 encodes for amyloid precursor protein (APP). The Aβ peptide is the cleavage product of APP by secretases (α-secretase, β-secretase and γ-secretase) [39]. In the non-amyloidogenic pathway, cleavage of APP is mediated by α-secretase and γ-secretase. The product of the non-amyloidogenic pathway is a soluble peptide. In the amyloidogenic pathway, on the other hand, Aβ40 and Aβ42 are the derivatives of the APP by β-and γ-secretases [40, 41, 42]. These are hydrophobic species and contribute to formation of senile plaques. Moreover, Aβ42 isoform has a tendency to aggregate more than Aβ40 [43, 21, 44, 38]. For example, missense mutations in *APP*, *PSEN1* and *PSEN2* are the leading causes for the familial AD [45, 46, 47, 48]. The way that these three genes contribute to AD pathology is by increasing the ratio of Aβ42/Aβ40 and thus promoting the aggregation of insoluble Aβ plaques [49, 50, 51].

The role of Aβ in AD could be explained by the amyloid cascade hypothesis (Figure 1.3) which was proposed by Hardy and Higgins [52, 53]. According to the amyloid cascade hypothesis, overproduction and accumulation of Aβ42 start the sequence of pathogenic events that result in AD. Aβ42 oligomers can stimulate the activation of microglia and astrocytes. Failure to remove Aβ42 oligomers might lead to synaptic damage, following chronic immune responses. Aβ42 also propagates the hyperphosphorylation of tau in the downstream of the cascade. Hyperphosphorylated tau forms neurofibrillary tangles, which are another pathological lesion in AD. The last step of the cascade is neuronal dysfunction that eventually causes dementia [54, 53, 38, 55, 2].

Another hallmark of AD is neuroinflammation, characterized by the increased number of astrocytes or microglial cells (also called gliosis) [56, 57, 58, 59, 60]. Microglia are a member of the innate immune system and have several roles, including the engulfment of damaged neurons [61, 62, 63, 64]. In AD pathology, microglia help to mitigate the burden of senile plaques by degrading Aβ (reviewed in [65, 66, 67]. However, the burden of senile plaque causes chronic activation of innate immune

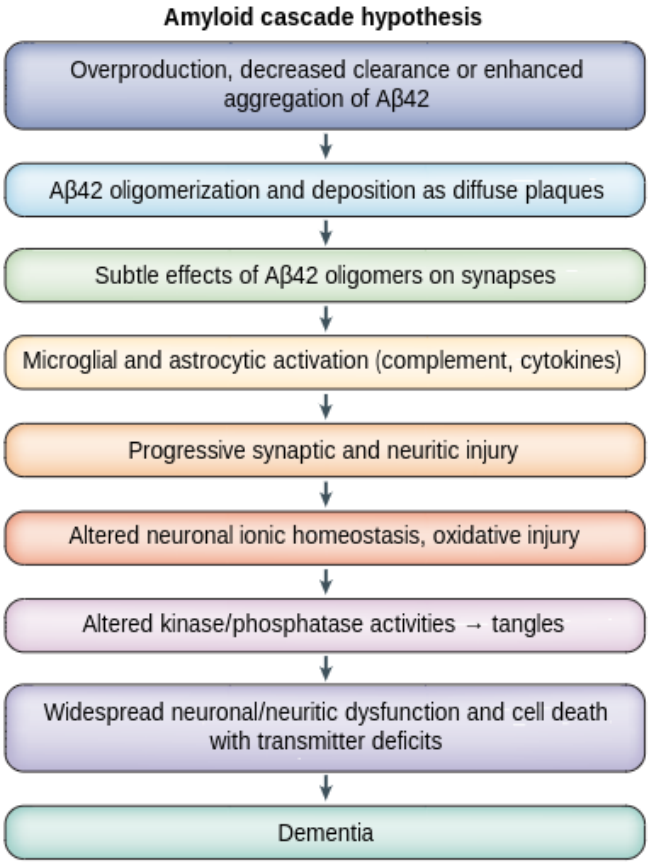


Figure 1.3: Steps of the amyloid cascade hypothesis. The figure was taken from [2].

response and exacerbates neuroinflammation, followed by impairment in microglial phagocytosis [68, 69, 70, 71]. The crucial role of microglia in AD progression has been identified by GWAS.

GWAS identified ~30 risk loci for AD and most of them were related to the immune response [72, 73]. One of them is the *TREM2* (triggering receptor expressed on myeloid cells 2) gene, which is mainly expressed by microglia in the brain and encodes proteins for TREM2 receptors. TREM2 receptors are located on the surface of the microglia. Aβ peptides are one of the ligands of TREM2 receptors and trigger the chain of events that can result in the degradation or compaction of Aβ peptides. It was identified that loss-of-function variants of the TREM2 gene are associated with the increased risk of developing AD by approximately 2-4 folds [74, 75].

Apart from *TREM2*, the apolipoprotein E (*APOE*) gene is another GWAS-identified genetic risk factor for developing LOAD [76, 77, 78, 79, 80, 81]. *APOE* is mainly expressed by astrocytes and microglial cells in the brain, and it is a polymorphic gene that has three alleles ($\varepsilon 2$, $\varepsilon 3$, and $\varepsilon 4$) [82, 83]. APOE protein contributes to AD pathology by slowing down Aβ clearance and accelerating Aβ deposition [84, 85, 86, 87]. The burden of Aβ changes in a dose-dependent manner. For example, individuals carrying the $\varepsilon 4$ allele of APOE in the homozygous state have been identified to show 15-fold increased susceptibility to AD and have the highest plaque density in their brain compared to other genotypes [88, 89, 90]. On the other hand, the $\varepsilon 2$ allele of APOE plays a protective role against AD [88].

The formation of neurofibrillary tangles (consists of hyperphosphorylated tau) is another neuropathological hallmark of AD and is triggered by Aβ as in neuroinflammation. Microtubules play an important role in cellular functions such as division and axonal stability [91]. The assembly and stabilization of microtubules are mediated by microtubule-associated proteins. In the brain, tau is the microtubule-associated protein (encoded by *MAPT* gene on chromosome 17) which is responsible for maintaining axonal stability [92, 93, 94, 95]. Phosphorylation of tau is crucial to bind microtubules. On the other hand, when it is in the hyperphosphorylated stage, tau detaches from microtubules [96], and the hyperphosphorylated tau aggregate into neurofibrillary tangles. Also, tau could disassociate from the microtubule-tau complex that causes instability of microtubules and end up with impairment in axonal transport [97, 98]. The formation of neurofibrillary tangles is not specific to AD and could also be observed in several neurodegenerative diseases like Parkinson's [99].

On the other hand, one of the less-studied potential contributors of AD is aneuploidies and copy number variations (CNVs) in the brain, which can be defined as the gain or loss of whole chromosomes or chromosomal segments. The role of aneuploidies and CNVs in AD will be discussed in the following sections.

1.6 Origin of CNVs and aneuploidies in the brain

Aneuploidy refers to the gain or loss of the whole chromosome and CNVs can be defined as the gain or loss DNA segment ($\geq$ 1 kilobase pair) [100, 101]. Early studies using cytogenetic techniques reported the relatively high frequency of aneuploid cells in the human brain (4-40%) [102, 103, 104]. This raises the question about the origin of aneuploidy in postmitotic neurons.

In 2001, Rehen and colleagues estimated that ~33% of neural progenitor cells (NPCs) are aneuploid in embryonic mice [105]. NPCs have mitotic capacity; and during the proliferation of NPCs, several mitotic errors, like chromosome missegregation, could generate aneuploid NPCs. Then, aneuploid NPCs can give rise to postmitotic neurons [105, 106].

CNV formation in the brain mainly occurs due to an error in the DNA repair mechanism which is called nonhomologous end joining [107, 108, 109]. On the other hand, there could be several chromosome segregation defects that lead to aneuploidy in the brain which are lagging chromosomes, supernumerary centrosomes, and non-disjunction [106].

- Lagging chromosomes: When a kinetochore is attached to microtubules coming from both spindle poles, it is known as merotelic kinetochore attachment. Due to the merotelic kinetochore attachment, chromosomes could lag behind in the spindle midzone during anaphase which could end up gain or loss of chromosome(s) [106, 110].

- Supernumerary centrosomes: Before anaphase, centrosomes separate into two groups and form bipolar spindle poles. Sometimes, there could be more than two centrosomes, which are called supernumerary centrosomes. Cells with supernumerary centrosomes could suffer from merotelic kinetochore attachments, which violates bipolar cell division, leading to aneuploid progeny [106, 111].

- Non-disjunction: Due to the segregation error in sister chromatids, both of them could move to the one pole of the cell. Non-disjunction results in a two aneu-

ploid progeny: one with monosomy and another with trisomy [106].

Previous studies showed that neural aneuploidy can appear on each chromosome with equal probability, without any preferences [105, 112, 113, 114, 112, 115]. On the other hand, we and others reported more deletions than duplications [105, 103, 104, 116, 117, 118, 114, 119]. For example, Rehen and colleagues calculated that ~33% of neuronal progenitors were aneuploid in embryonic mice brains [105] using cytogenetic techniques. The percentage of hypoploid cells was striking which was 98%. Similarly, in the 2014 single-cell study, Cai and colleagues identified more losses than gains in the adult human brain. The authors also examined the frequency of CNVs in human lymphoblasts and they found approximately equal numbers of loss and gains. The observed trend in the brain could be explained by the following. Chromosome segregation defects tend to occur in the type of lagging chromosomes and supernumerary centromeres rather than nondisjunction in the developing mice brains. And these two chromosome missegregation events predominantly produce hypoploid cells [106]. Indeed, nondisjunction is not only a rare event during neurogenesis but also produces hypoploid and hyperploid cells at a similar rate [106].

1.7 Experimental methods to study aneuploidy

There are several methods to study copy number variations in cells: karyotype analysis, spectral karyotyping (SKY), interphase fluorescence in situ hybridization (i-FISH), metaphase fluorescence in situ hybridization (metaphase-FISH) and single-cell whole genome sequencing (scWGS) [120].

Karyotype analysis, SKY and metaphase-FISH are proper for cells with mitotic capacity [121]. In karyotype analysis, cultured cells are arrested in metaphase to enumerate condensed chromosomes. SKY and metaphase-FISH take advantage of the stability of DNA molecules. First, fluorescently-labelled probe DNA and complementary target sequences are denatured to single-stranded DNAs. Then, both molecules are annealed to form a stable double-stranded molecule. Due to the limited number of fluorescent labels, one can use metaphase-FISH to analyze a limited number of chromosomes per cell [121, 122]. On the other hand, SKY allows us to study all

chromosomes by employing 24 chromosome-specific probes [123, 124, 125].

It is generally accepted that mature neurons in healthy brains may carry somatic CNVs/ aneuploidies, but the frequency of such events is uncertain. i-FISH and scWGS are the methods for studying aneuploidy in non-dividing cells like neurons. However, the frequency of aneuploidy that was reported by these two methods varies dramatically. Studies using i-FISH (4-40%) [102, 103, 104] reported much higher aneuploidy frequency than studies using scWGS (<4%) [107, 126, 127].

Andriani and colleagues compared two methods to understand the underlying reasons for the difference in the reported aneuploidy frequency [128]. One possible explanation for the overestimation of aneuploidy by i-FISH is that i-FISH could suffer from probe hybridization problems and/or probe clustering [129, 130]. To overcome these problems, the authors suggested a 2-probe/chromosome i-FISH assay rather than one probe/chromosome i-FISH assay. In the 2-probe/chromosome i-FISH assay, the authors used two probes which are labelled by different fluorophores. If two probes report the same events (gains or losses), then loci are called aneuploid. Even though 2-probe/chromosome assay reduces the false positives compared to 1-probe/chromosome, still this method is not proper for calling aneuploidy across the genome. This is because similar to metaphase-FISH, i-FISH allows only studying a few chromosomes at a time.

To test the sensitivity scWGS, the authors used mock aneuploid cells which are either polyploid (i.e. 4n, 8n) or have complex aneuploidy (aneuploidy in a polyploid background, i.e. chromosome 2 gain in 8n ploidy background). The reason for testing complex aneuploidy is to mimic the biology of aneuploidy. For example, in the liver most hepatocytes are polyploid [131]. The authors showed that scWGS fail to identify polyploidy (4n and 8n are predicted as 2n) with a 0% sensitivity. On the other hand, scWGS sensitivity was 33.3% accurate in predicting complex aneuploidy (i.e. predicting aneuploidy in the polyploid background). Andriani and colleagues suggested the improvement in the DNA amplification step in scWGS and the usage of the combination of two methods as possible solutions.

1.8 The role of aneuploidy in the healthy brain

Over the last decade, advances in next-generation sequencing (NGS) technologies gave fresh impetus to somatic CNV analyses by allowing variants to be determined at the single-cell level across species [132]:

- McConnell *et al.*, 2013: The authors sequenced 110 frontal cortex neurons from three young humans (20-26 years) [107]. The authors showed that 41% of the analyzed cells include at least one CNV (2.9-75 Mb) and 2.7% of them have aneuploidy. The authors also reported that deletions occur approximately twice as frequently as duplications. These results implied that neurons with mosaic CNVs are common in the human frontal cortex. The data coverage was ~0.1X and DNA was amplified using the GenomePlex Single Cell Whole Genome Amplification Kit (Sigma).

- Cai *et al.*, 2014: The authors analyzed 97 neurons from three neurotypical human cerebral cortex at ~0.08X coverage [116]. The authors did not report any neurons having aneuploidy. Among 19 neurons from one individual, 68% (13/19) harboured megabase-scale CNVs (median 2.3 Mb; 1.7-17 Mb), with an average of 3.4 CNVs per neuron. In addition, two cells (6%) shared the same CNVs which supports the idea that the CNVs might stem from progenitor cells during development. The authors also reported more deletions than duplications, consistent with McConnell *et al.*, 2013 [107]. In this study, DNA was amplified by GenomePlex (Sigma).

- Knouse *et al.*, 2014: In this single-cell study, the authors first tested the sensitivity and specificity of their methods using trisomy 16 mouse embryos [127]. Using *HMMcopy*, the authors reported that trisomy 16 was detected across all samples without any false positive or negative calls. Then, to investigate the prevalence of aneuploidy in the mouse brain, the authors sequenced 19 neurons from the mouse cerebral cortex, and found no aneuploidy (predicted by *HMMcopy*). The authors stated that labelling neurons with NeuN could result in high sequence variability. The reason for this is that the labelling steps including fixation and immunostaining might affect the whole genome amplification.

Then, 89 neurons from four neurotypical human frontal cortex were sequenced at ~0.1X coverage. The age of the individuals ranged between 48 and 70 years. The authors only reported the prevalence of aneuploidy in the human frontal cortex, as 2.2%. In this study, DNA was amplified using GenomePlex (Sigma) and ~10% of the cells showing high variation in read depth were removed from the study.

- Knouse *et al.*, 2016: The authors compared the accuracy of different CNV prediction algorithms (*DNAcopy* and *HMMcopy*, see 2.2.2) and stated that CNVs larger than 5 Mb could reliably be called at ~0.1X coverage [133]. The authors analyzed 105 neurons from four neurotypical human frontal cortex at ~0.1X coverage. The prevalence of CNVs in the human brain was estimated as 9%. Most CNVs were deletions (90%) and located near the telomeres. Compared to keratinocytes from the skin, the authors reported high read depth variability in neurons that could be explained by differences in chromatin structure. The data from the previous studies were also reanalyzed (McConnell: 41%, Cai: 68%), and the authors reported a lower rate of CNV frequency than the original studies (McConnell: 17%, Cai: 10%) [107, 116]. This result indicates how chosen methods affect the reported CNV frequency.

- van den Bos *et al.*, 2016: van den Bos and colleagues first tested the accuracy of their pipeline using trisomy 21 data from individuals having Down's syndrome ($n = 36$ single cells) [126]. Trisomy 21 were detected across all samples without any false positive or negative calls. Analyzing 589 neurons from 6 control individuals, the authors reported aneuploid prevalence at 0.7% and concluded that aneuploid cells are not common in the healthy brain. To avoid PCR biases, the authors skipped the DNA amplification steps. Even though there was no such step, 40% of cells showing high read depth variation were discarded from the analyses.

- Vitak *et al.*, 2017: Vitak and colleagues implemented an experimental method, single-cell combinatorial indexed sequencing (SCI-seq), to produce thousands of low-coverage single-cell libraries in parallel [134]. An advantage of SCI-seq is that it does not require labour-intensive cell isolation steps. Thus, SCI-seq may be preferable for low-pass scWGS. The authors also prepared 552 single-

cell libraries from two rhesus macaques (4 and 9 years), and used different thresholds to report an aneuploidy rate of ~10% (min: 3.1%, max: 25.0%).

- Rohrback *et al.*, 2018: Rohrback and colleagues determined how the frequency of CNVs changes during mouse brain development using 658 cells from 43 mice [135]. The authors claimed that DNA amplification with GenomePlex (Sigma) introduces noise in the data, so instead, they used transposase-based amplification (TbA) to amplify genomic DNA. This approach made it possible to assess CNVs as small as 0.25 Mb. The authors reported 5394 events in 488 cells (~11 per cell), most of which were smaller than 5 Mb. Consistent with previous studies, deletions dominated the frequency of CNVs [107, 116, 133].

- Chronister *et al.*, 2019: Chronister and colleagues sequenced 589 neurons from the neocortex of five neurotypical human individuals [136]. The authors estimated CNV-carrying neurons at around 30% (n = 197) in young adults and 10% (n = 392) in old adults. In other words, CNV-bearing neuron frequencies decrease from young to old adulthood. This result might indicate that CNV-bearing neurons may be eliminated throughout a lifetime in neurotypical individuals. Consistent with previous studies, there were more deletions than duplications among somatic mutations [107, 116, 133, 135]. The authors also reported that CNV frequencies in neurons (4%–23.1%) are higher than in non-neuronal cells (4.7%–8.7%).

1.9 The role of aneuploidy in AD brain

Over the last two decades, a number of FISH and cytogenetic-based studies investigated CNV frequencies in AD and healthy control brains [137, 102, 117, 138, 139, 113, 140]. Several of these reported extra copies of chromosomes in neurons of the AD brain [137, 140, 117, 138, 139, 113]. This, in turn, implies that chromosomal imbalance might contribute to AD pathogenesis via altered gene expression levels. An example of such imbalance is seen in individuals with Down's syndrome (DS); carrying an extra copy of chromosome 21 appears to facilitate aggregation of amyloid-β (Aβ) plaques in the brains of DS individuals similar to the AD phenotype [46, 138, 141].

There are various explanations for why post-mitotic neurons in AD brains could carry high frequencies of somatic CNV [142]. According to one view, the high CNV burden in the AD brain originates from neurogenesis in the embryonic period. This excessive somatic mutation may be pathogenic and manifest itself as increased AD risk during aging [143]. However, Abascal *et al.* recently analyzed single cells from granulocytes, smooth muscle, neurons, skin, colon, bronchus and bladder in humans, and showed that somatic mutation (single nucleotide change or indel) accumulation in cells with mitotic capacity and in post-mitotic neurons follow similar trajectories. That is, mutational processes (possibly also including CNVs) appear to occur in a time-dependent manner rather than being division-dependent [144]. Accordingly, CNVs in AD brains may have accumulated during individuals' lifetimes. However, this scenario also appears inconsistent with the observation that CNV-bearing neuron frequencies decrease from young to old adulthood [136]. Another view suggests that AD itself might cause dysregulation in neurons, and AD-affected mature neurons might re-enter the cell cycle, resulting in increased CNV load [117, 145], which may then be eliminated at later stages of AD, thus causing neurodegeneration [139].

In a 2016 study, van den Bos and colleagues used scWGS to compare the prevalence of aneuploidy in neurons from healthy control and AD patients [126]. Analyzing 1482 neurons from 10 AD patients and 6 control individuals, the authors reported aneuploid prevalence at 0.7% and 0.6% for control and AD neurons, respectively, and concluded that aneuploid cells are not more common in the AD brain.

These findings by van den Bos and colleagues implied that CNVs might have no relationship to AD pathogenesis, in contrast with earlier findings from FISH and cytometry studies. However, the study by van den Bos and colleagues had a number of limitations. One was that the authors only estimated aneuploidy (full chromosome gain or loss), while large CNVs, which could also contribute to pathogenesis, had remained uncharacterized. Another limitation was that only one brain region was examined, the frontal cortex, while atrophy of the medial temporal lobe and specifically the hippocampus is generally considered to be a strong predictor of AD [146]. The study also did not distinguish among neuron types that may carry sensitivity to AD

differentially [147]. Thirdly, the study discarded a large fraction of cells (39%) for showing high within-cell variability in genome coverage, although it was unclear to what extent these represented pure technical error versus cells with complex karyotypes. Finally, only NeuN positive neurons were included in the experiment, which substantially restricts the significance of this study due to different reasons:

1. Recently, up to 30% of cortical neurons have been reported being NeuN-negative following diffuse brain injury, which may be related to certain neurons being particularly vulnerable to membrane disruption [148], a process which was also recently associated with AD [149, 150].

2. Considerable or even complete loss of NeuN immunoreactivity was also reported for neurons affected by ischemic insults (middle cerebral artery occlusion) without significant cell loss [151] or in neurons that just entered the cell death process [152]. Interestingly, these neuronal populations are of special interest because energy and nutritional deficiency and cell loss are essential characteristics of the AD brain [153].

3. The intensity of NeuN staining is reported to be lower in AD samples [154], and further due to many NeuN negative cortical neurons in FTLD-TDP (frontotemporal lobar degeneration with TDP-43 inclusions) patients, Yousef *et al.* suggested NeuN staining as an indicator of healthy neurons [155]. However, if NeuN reflects a neuron's health, any selection of NeuN positive cells would lead to a substantial bias for studying any neurodegenerative disease.

These methodological issues could potentially explain the discrepancies between the findings by van den Bos *et al.* and those based on FISH and cytogenetic studies [137, 117, 140, 138, 139, 113]. Notably, a recent technical comparison between FISH and scWGS using mock aneuploid cells reported a tendency of the latter to severely underestimate aneuploidy [128]. It is thus possible that both neurons with CNV and nuclei thereof display altered physicochemical properties. This may result in selection bias against abnormal nuclei with high CNV loads when using the fluorescence activated cell/nuclei sorting (FACS, FANS) isolation method (exerting mechanical stress [156]) and high hydrodynamic pressure [157], applied by van den Bos and colleagues, and artificially inflate euploidy frequencies. Moreover, besides restriction to

NeuN positive cells, usage of only intact nuclei could preclude or bias AD neurons with nuclear envelope stress or rupture [158].

1.10 Research Objectives

Here we generated and analyzed scWGS data to establish the frequency of CNVs (both full chromosome aneuploidies and sub-chromosomal CNVs) in five different brain regions that differ in vulnerability to AD in healthy brains and brains of AD patients [20]. We employed two different single-cell isolation methods, laser capture microdissection (LCM) and FACS, to isolate neuronal nuclei. LCM, despite being technically challenging, has the advantages of allowing for specific neuron types to be chosen, and being neutral towards normal and abnormal nuclei. We further employed a principal component analysis-based denoising approach to eliminate false positive CNV calls that might result from either systematic experimental biases or repetitive regions in the human genome. Finally, we analyzed published datasets to replicate our main results and check the sensitivity and specificity of our bioinformatics pipeline.

Overall, the aim of the study was to answer the following questions:

- Do the frequency of whole chromosome aneuploidy/ CNV differ in the brains of healthy control and AD?
- Are there region-specific differences in the prevalence and distribution of aneuploidy/CNV between healthy controls and AD?
- Are there differences between the frequency of gains and losses?
- Does aneuploidy/CNV preferentially affect certain chromosomes?
- Are there any differences between LCM and FACS in terms of technical noise?

Table 1.1: Summary of the scWGS studies in healthy brain.

Study	WGA	Brain region	Number of indiv.	Number of cells	Findings
McConnell *et al.*, 2013 [107]	GenomePlex	Human frontal cortex	3	110	2.7% neuronal aneuploidy
					41% CNV (range between 2.9-75 Mb)
Cai *et al.*, 2014 [116]	GenomePlex	Human cerebral cortex	3	97	No aneuploidy
					~68% CNV (range between 1.7-17 Mb)
Knouse *et al.*, 2014 [127]	GenomePlex	Human frontal lobe	4	89	~2.2% neuronal aneuploidy
					CNVs not studied
Knouse *et al.*, 2016 [133]	GenomePlex	Human frontal lobe	4	105	No aneuploidy
					9% neuronal CNV
van den Bos *et al.*, 2016 [126]	Strand-seq	Human frontal cortex	6	589	0.7% neuronal aneuploidy
					CNVs not studied
Vitak *et al.*, 2017 [134]	SCI-seq	Rhesus macaque frontal cortex	2	552	~10% neuronal aneuploidy
					CNVs not studied
Rohrback *et al.*, 2018 [135]	TbA	Mouse cerebral cortex	43	658	~2% neuronal aneuploidy
					~11 CNVs per cell
Chronister *et al.*, 2019 [136]	PicoPLEX	Human neocortex	5	879	~30% CNVs in young
					~10% CNVs in old

WGA: whole genome amplification, indiv: individual, SCI-seq: single-cell combinatorial indexed sequencing, TbA: transposase-based amplification

CHAPTER 2

MATERIAL AND METHODS

2.1 Wet lab experiments

All experiments were conducted at the Paul-Flechsig-Institute (Leipzig University, Germany). People who contributed the data generation were Uwe Ueberham, Thomas Arendt, Vincent Richter, Jana Bochmann, Sarah-Kristin Waterholter, Sabrina Leclere-Turbant and Charles Duyckaerts.

2.1.1 Ethics approval and consent to participate

The samples were obtained from brains collected in a Brain Donation Program of the Brain Bank NeuroCEB run by a consortium of Patients Associations: ARSEP (association for research on multiple sclerosis), CSC (cerebellar ataxias), LECMA (European league against Alzheimer disease) and France Parkinson.

The consents, that have been validated by the Ethical Committee Ile de France 6, were signed by the patients themselves or their next of kin in their name, in accordance with the French Bioethical Laws.

The Brain Bank NeuroCEB has been declared at the Ministry of Higher Education and Research and has received approval to distribute samples (agreement AC-2013-1887). The autopsy protocol has been approved by the Biomedicine Agency as requested by the French Law.

2.1.2 Tissue sources

Frozen postmortem human brain tissues -temporal cortex, hippocampal subfields cornu ammonis (CA) 1, hippocampal subfields cornu ammonis (CA) 3, cerebellum (CB) and entorhinal cortex (EC)- from a total of 13 AD patients and 7 non-demented age-matched controls were obtained from the GIE NeuroCEB Brain Bank (France).

AD cases were diagnosed according to the National Institute of Aging and Reagan Institute Criteria [159] and immunohistochemically processed for tau and amyloid pathologies [17, 160]. Control cases were non-demented individuals who died without known neurological disorders.

Post-mortem delays and mean ages of control and AD cases were not significantly different. The average age of death was for control cases ($n = 7$) 71.57 years (± 5.13 years SEM) and for AD cases ($n = 13$) 70.15 years (± 3.63 years SEM) ($p = 0.822$). The average post-mortem delays were 31.14 hours (± 7.10 hours SEM) for control cases and 26.17 hours (± 4.08 hours SEM) ($p = 0.52$).

2.1.3 Methods for single-cell isolation

Fluorescence-activated cell sorting (FACS): Neuronal nuclei were extracted following the protocol described in [161]. Briefly, frozen brain samples were thawed in the hypotonic lysis buffer. Neuronal nuclei were stained with propidium iodide and sorted using BD FACSAria II SORP (BD Biosciences). Genomic DNA was then isolated and amplified as described below (see 2.1.4).

Laser capture microdissection (LCM): Frozen brain samples at –80°C were thawed to -20°C, sliced using CryoCut Freezing Microtome at 30 µm thickness, and mounted on a membrane slide (Carl Zeiss). After staining with cresyl violet, single cells were cut out and placed into an adhesive cap by PALM MicroBeam (Carl Zeiss). Neurons of the individual 5603 were collected using both FACS ($n = 12$) and LCM ($n = 64$).

2.1.4 scWGS library preparation and sequencing

Genomic DNA was amplified using WGA4 (GenomePlex® Single Cell Whole Genome Amplification Kit) and then purified using the MinElute PCR Purification Kit (Qiagen). The specific adapters were added to the DNA via Phusion® PCR followed by purification with the MinElute PCR Purification Kit (Qiagen). Sample quality was evaluated using agarose gel electrophoresis. Sequencing was performed on the HiSeq2500 platform (Illumina) with paired-end 100 bp (PE100) or 150 bp (PE150) modes.

2.2 Data analyses pipeline for scWGS data

Analyses were conducted by myself with the help of Poorya Parvizi, Ulaş Işıldak and Etka Yapar. Data analyses pipeline for scWGS data consist of two main parts: **(1)** NGS data analyses and **(2)** CNV detection (Figure2.1).

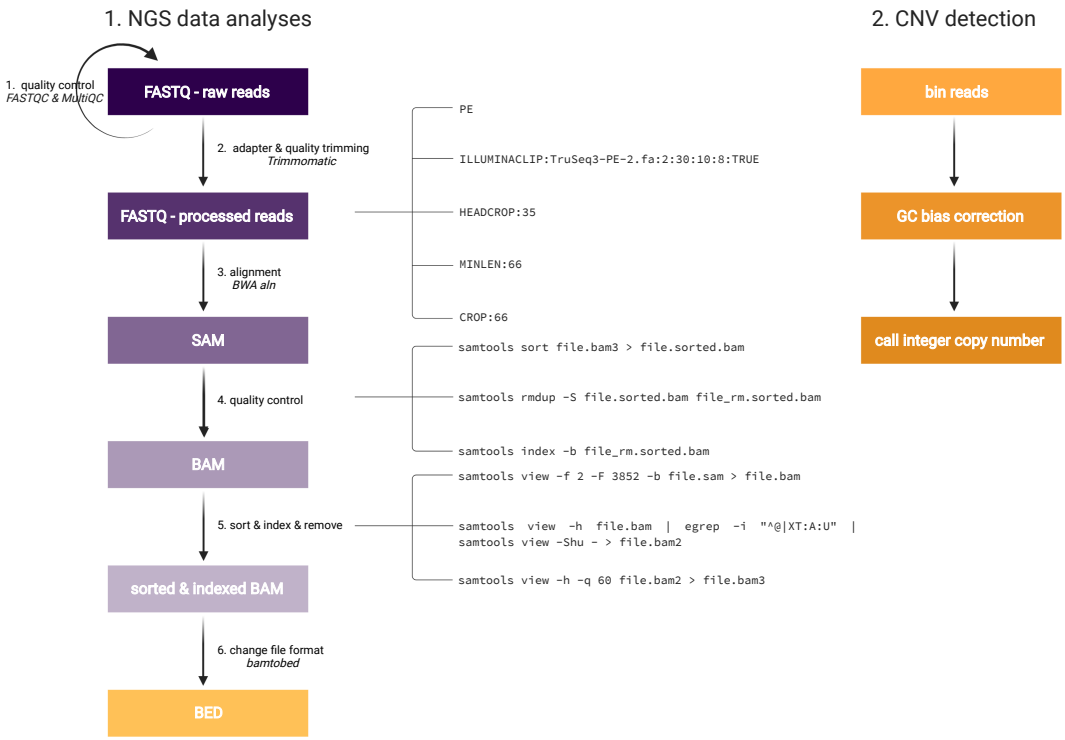


Figure 2.1: The pipeline of NGS data analysis and CNV detection.

2.2.1 NGS data analysis

2.2.1.1 Read quality control and alignment

The $FastQC$ tool (version 0.11.9) was used to check the quality of the raw Illumina reads. The results of $FastQC$ were summarized using $MultiQC$ (version 1.9) [162]. The mean sequence lengths of the reads (ranging between 101 and 151) were inspected using the output of the $MultiQC$ ($general\_stats\_table$). To avoid biases that would affect the interpretation of the results, all reads were trimmed to a length of 66 (the longest possible length in all reads). Illumina adapter and low-quality bases (the first 35 bp) were removed using $Trimmomatic$ [163] with the following parameters: "ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10:8:TRUE HEADCROP:35 MINLEN:66 CROP:66".

The quality of the trimmed reads was checked again using $FastQC$ and $MultiQC$. Adapter-trimmed paired-end FASTQ files were mapped to the hg19 human reference genome (`/ftp://ftp.ensembl.org/pub/release-75/fasta/homo_sapiens/dna/`) using **B**urrows-**W**heeler **A**lignment (BWA v.0.7.17) [164] with "aln" and "sampe" options.

2.2.1.2 Filtering

The output of the BWA aligner in **S**equence **A**lignment/**M**ap (SAM) format was further processed by $SAMtools$ v1.10 [165] to obtain high-quality uniquely aligned reads. The applied steps are as follows:

(1) keep reads mapped in proper pair and discard reads marked with SAM flag 3852:
`samtools view -f 2 -F 3852 -b file.sam > file.bam`

(2) extract uniquely mapped reads from BAM files:
`samtools view -h file.bam | egrep -i "^@ | XT:A:U" | samtools view -Shu - > file.bam2` [166]

(3) obtain reads having MAPQ scores 60:

```
samtools view -h -q 60 file.bam2 > file.bam3
```

(4) sort BAM files:

```
samtools sort file.bam3 > file.sorted.bam
```

(5) filter out PCR duplicates:

```
samtools rmdup -S file.sorted.bam file_rm.sorted.bam
```

(6) index BAM files:

```
samtools index -b file_rm.sorted.bam
```

(7) convert BAM file into BED format using the $Bedtools$ "bamToBed" command ($Bedtools$ v2.27.1) [167].

2.2.1.3 Coverage

$Bedtools$ v2.27.1 algorithm "genomeCoverageBed" was used to obtain coverage of the bases on each BAM file.

2.2.2 CNV prediction and cell elimination

CNV calling was performed using $Ginkgo$ [3]. We had three main reasons for using $Ginkgo$ over its most commonly used alternative, $HMMcopy$ [168]. First, a recent study [169] performed benchmarking on $Ginkgo$ and two other widely used methods $HMMcopy$ and $CopyNumber$, and found that $Ginkgo$ was the most accurate algorithm for inferring the absolute copy number profiles (although $HMMcopy$ was superior in identifying breakpoints and running time). Second, $Ginkgo$ provided the advantage of outputting data with normalised coverages per cell, which we could use in our PCA-based denoising method, and further in estimating the genome-wide copy number of each cell, which we used to filter cells for high levels of variability in read depth. Third, our tests on the sensitivity and specificity of $Ginkgo$ using trisomy-21 in DS and monosomy-X in males in published data [126, 107] revealed 100% and 94% detection rates across the two published datasets. Overall, apart from its advantages, $Ginkgo$ enable us to develop PCA-based method denoising approach.

CNV prediction involves several steps. These steps and the parameters used in $Ginkgo$ [3] will be discussed as follows (Figure 2.3):

- Binning: Uniquely mapped reads are binned into the non-overlapping fixed- or variable-size genomic windows. The aim of the binning is to reduce the effect of the amplification biases [3]. Also, working with a lower resolution is more manageable than working on the entire genome [168].

  The use of variable-length bins has several advantages. The first one is to avoid false positive deletions in the regions that correspond to blacklisted areas (repetitive or low-mappability regions) in the human genome [170]. Variable-size bins have wider intervals in the regions that span blacklisted areas [170]. In this way, each bin also has the same number of reads with a constant variance [3]. Considering that the read counts in each bin are used to predict CNVs, having the same number of reads with a constant variance is important. Due to the above-mentioned advantages of variable-length bins, we used variable length bins with an average size of 500-kb.

- Coverage normalization: $Ginkgo$ performs coverage normalization; the read count in each bin is divided by the mean read count across bins per sample [3]. This scaling step is crucial for our PCA-based denoising approach, which will be discussed in the following section, to work accurately.

- GC bias correction: GC extreme regions (GC-poor < 40% or GC-rich > 60%) could form secondary structures that hamper the efficiency of DNA polymerase during PCR amplifications [171, 172, 173, 3, 174, 175, 176, 177]. Inefficient amplification of GC extreme regions can be reflected as uneven coverage or even no coverage of reads in Illumina sequencing, known as GC content bias. $Ginkgo$ fit locally weighted linear regression (LOWESS) between GC content and log-scaled read counts in each bin to correct GC bias [3]. Then, GC-normalized values will used for segmentation.

- Segmentation: The aim of the segmentation is to divide the genome into non-overlapping regions that have different copy numbers [169]. The bound-

aries of the segments are called breakpoints. To identify change points or breakpoints where copy number transitions occur, most segmentation methods are based on the idea that chromosomal aberrations, for human deviation from two, can be distinguished by abrupt changes in the read count distribution [178].

One method that can be used to identify breakpoints in the genome is circular binary segmentation (Figure 2.2) [179, 180]. $Ginkgo$ uses CBS which is implemented in the $DNAcopy$ package in R [180]. To identify breakpoints, CBS compares the log values of the read count in each potential segment with the regions outside the segment using a t-test. CBS runs per chromosome and splits them into subsegments recursively. The start and end positions of the chromosome are connected and the connection is called 0. X and Y are the breakpoints of the segment, where $0 < X < Y$. If the read count distribution is statistically significant between $X \rightarrow Y$ and $Y \rightarrow 0 \rightarrow X$, the region is segmented as $X \rightarrow Y$, $Y \rightarrow 0$ and $0 \rightarrow X$ [179, 178, 181].

- Post hoc adjustment: CBS only determines the breakpoints of the segment. Post hoc adjustment is required to determine the integer copy number states. $Ginkgo$ takes the following steps to do it [3]:

  (1) Variable size of 500-kb bins that correspond to 5578 genomic windows; and suppose $DNAcopy$ estimated the first breakpoint to be between bin 1 and bin 25, then $Ginkgo$ takes the median of the normalized read counts (coverage- and GC-normalized) between bin 1 and bin 25 (called segment X). Then, the read counts for all bins in segment X are replaced with the median bin count of segment X.

  (2) To determine the copy number state of each sample, $Ginkgo$ next scales the segmented bin counts so that the mean bin count equals the ploidy of the sample. Thus, the values from the previous step are divided by the mean for each sample. The resulting value is called the raw copy-number profile (RCNP).

  (3) RCNP is multiplied with the possible genome-wide copy number of the sample, ranging from 1.5 to 6 with increment of 0.5 [1.50, 1.55, . . ., 5.95, 6.00]. The product of the RCNP and genome-wide copy number is called

the scaled copy-number profile (SCNP).

(4) To determine the genome-wide copy number of the sample, the sum-of-squares (SoS) error between the SCNP and the RCNP is then calculated for each multiplier. The multiplier with the smallest SoS error will be the genome-wide copy number of the sample.

(5) The rounded value of the SCNP is called the final integer copy-number profile (FCNP), such that each segment is represented with an integer value.

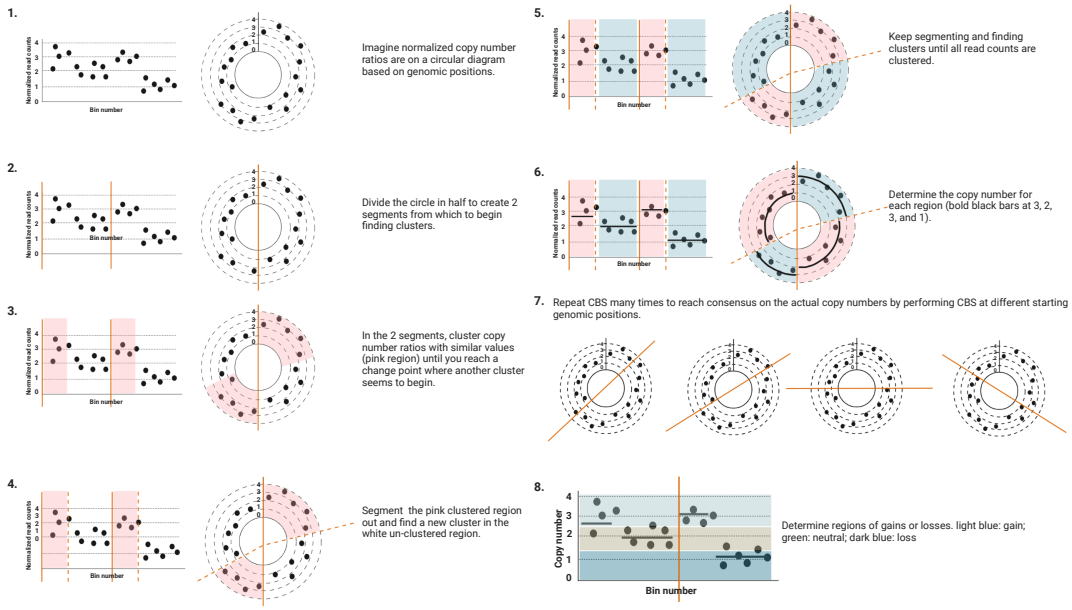


Figure 2.2: Steps of circular binary segmentation. Figure was copied from `https://github.com/melissayan/vnowchi`, and to increase the resolution, redrawed with `https://biorender.com/`.

The command-line version of *Ginkgo* was downloaded from `https://github.com/robertaboukhalil/ginkgo`. The tool was run under the following settings: **(1)** variable size of 500 kb bins [166] based on simulations of 76 bp reads aligned with BWA, **(2)** independent segmentation method, **(3)** ward and euclidean options for the clustering method and clustering distance metric, respectively. Before the segmentation step, GC correction was performed by *Ginkgo* using the R function "LOWESS" (see [3]). For segmentation, *Ginkgo* uses

the CBS algorithm implemented in $DNAcopy$ in R [182]. $DNAcopy$ runs with the following parameters: `alpha=0.0001, undo.SD=1, min.width=5` [133]. We also run $HMMcopy$ as described in [168] (using the parameter $e = 0.995$).

The number of reads was divided into the variable size of 500 kb bins that correspond to 5578 genomic windows. Only cells with >50,000 reads were kept in downstream analyses (approximately nine reads per window), resulting in $n = 1337$ cells.

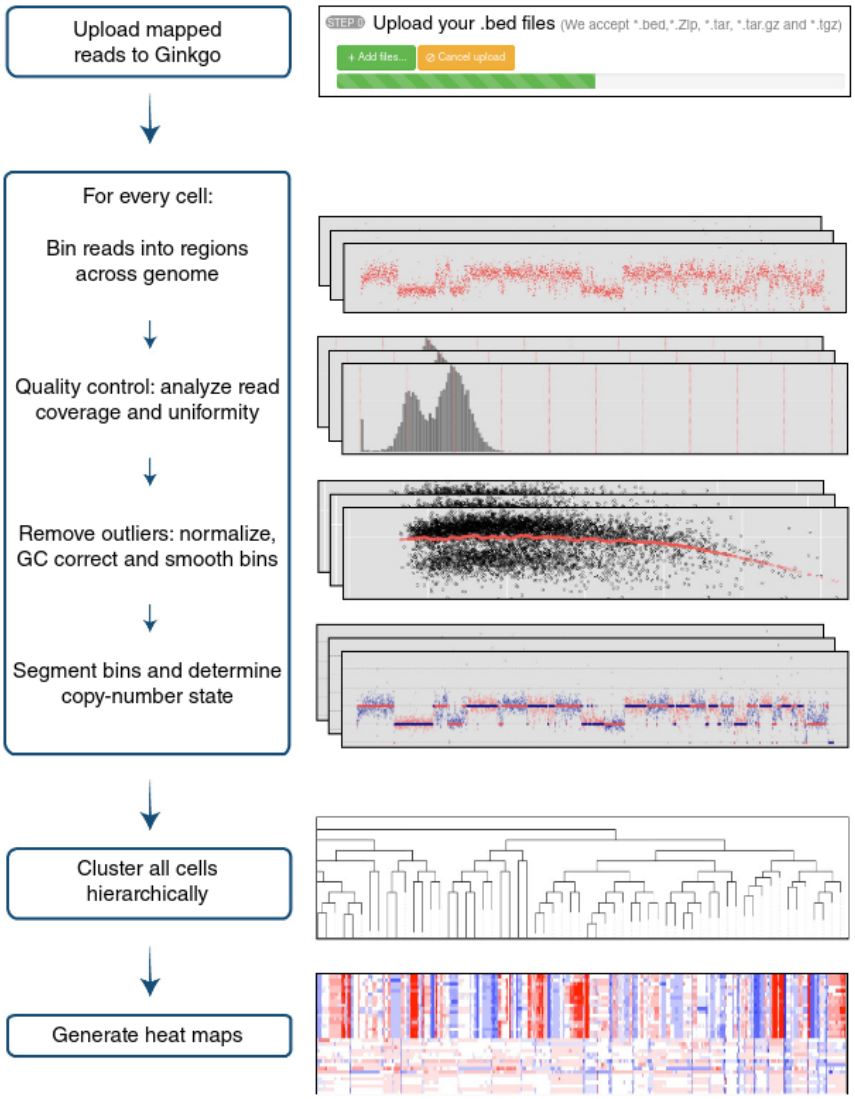


Figure 2.3: Workflow of $Ginkgo$. Figure was taken from [3].

2.3 Published datasets

The van den Bos 2016 dataset: Data was downloaded from EBI ArrayExpress with the accession numbers E-MTAB-4184 and E-MTAB-4185 [126]. Only the cells that were reported as having good quality libraries were included in the analysis (AD:883; control:586; Down's syndrome:34). Adapter sequences were trimmed with the following parameters: "ILLUMINACLIP:adapter.fa:2:30:10:8:TRUE MINLEN:51". Single end reads were aligned to the hg19 human reference genome using BWA with "aln" and "samse" options. The remaining steps are the same as those described in sections Filtering, except that here we used the SAM flag 3844 (because this dataset was single-end sequenced) and used MAPQ scores 20 (because this dataset did not have enough reads which having the MAPQ 60). Note that due to the missing sample information in the database, the number of analyzed cells in this work does not match what van den Bos and colleagues reported in their original publication.

The McConnell 2013 dataset: FASTQ files of 110 cells were downloaded from the NCBI SRA database with accession number SRP030642 [107]. Adapter sequence was trimmed with the following parameters: "ILLUMINACLIP:adapter.fa:2:30:10:8:TRUE MINLEN:39". Paired-end reads were aligned to the hg19 human reference genome using BWA with "aln" and "sampe" options. The remaining steps are the same as those described in sections 2.2.1.2.

2.4 Statistical modeling of CNV frequencies and index of dispersion (IOD) levels

When modelling CNV frequencies, our null hypothesis was no difference in the frequency of CNVs in the AD brain when compared to healthy controls. The overdispersed and zero-dominated nature of the response variable (the frequency of CNVs) suggest that the data should be fitted using a zero-inflated negative binomial model. For this aim, the "glmmadmb" function (package: glmmADMB) [183] in R 3.6.3 with the following parameters was used: "zero-inflated = TRUE" and "family = nbinom1". The fixed factors of the model were diagnoses (AD and control), chromosomes (autosomes), sex (male and female), brain regions (temporal cortex,

hippocampus CA1, hippocampus CA3, cerebellum, entorhinal cortex) and coverage. Individual effects were added as a random factor. Note that sex could not be used as a fixed factor in the van den Bos 2016 dataset because cells that remained after filtering only belonged to females.

We also compared the difference between AD and control in terms of CNV frequency using *HMMcopy* estimates. The fixed factors of the model were diagnoses (AD vs. control) and coverage per cell. The individual effect was added as a random factor.

When modelling the index of dispersion (IOD, the ratio between the variance of read coverage and the mean), we used the same approach as above. IOD (the response variable) were predicted using diagnoses (AD and control), brain regions (temporal cortex, hippocampus CA1, hippocampus CA3, cerebellum, entorhinal cortex) and coverage as an explanatory variables using "glmmadmb" function (package: glmmADMB) [183] in R 3.6.3. Individual effects were added as a random factor. The distribution of the IOD was right-skewed and the model was run with "family = gamma" parameter.

To compare the IOD across different brain regions, "lme" function (package: "nlme") in R 3.6.3 with diagnoses as fixed effects and the individual as a random effect was used.

2.5 Copy number statistics

After reads were mapped into the bins, read counts in each bin were divided by the mean read counts across bins for each cell. This value corresponds to the normalized read counts as calculated by $Ginkgo$ (see [3]).

A Z_1-score for each CNV was calculated using the normalized read counts. It was calculated as the cell mean (mean normalized read counts across autosomes) minus the CNV mean (mean read counts between CNV boundaries) divided by the standard

deviation (sd) of CNV:

$$Z_1\text{-}score = \frac{mean_{cell} - mean_{CNV}}{sd_{CNV}}$$

The Z_2-score of each CNV was calculated by calculating the difference between the $Ginkgo$-estimated integer copy number state (1 or 3) and the observed normalized read count, dividing by the standard deviation (sd) of the normalized read counts:

$$Z_2\text{-}score = \frac{estimated\_state_{CNV} - mean(observed\_readcount_{CNV})}{sd_{CNV}}$$

CNVs with two standard deviations below or above the cell's mean and CNVs with Z_2-score smaller than or equal to 0.5 were kept in the analysis. Using these combinations, monosomy X ($\geq$90% of the chromosome's length) was correctly predicted in 58.1% (217 of 373) of males in the uncorrected data.

2.6 Principal component analysis (PCA)

To remove experimental noise from the data, the following steps were applied for every cell:

(1) One cell (x) at a time was discarded from the analysis. For the remaining cells, PCA was applied on the normalized read counts using the "prcomp" function with the parameter "scale.=TRUE" in R 3.6.3.

(2) n PCs that explained at least 90% of the variance in total was chosen.

(3) To remove the effect of the chosen PCs from the focal cell x, a linear regression model with normalized read counts from cell x as a response, and the n PCs as explanatory variables was constructed using the R "lm" function.

(4) Residuals from this model were calculated.

(5) To prevent errors during a lowess fit of GC content (log transformation of negative residuals produces NaNs), plus one was added to the residuals. If there still remained values less than or equal to zero, those values were replaced with the smallest positive number for the focal cell x.

(6) The resulting value was set as a new value of the focal cell x, and *Ginkgo* was run with the new values.

The pseudocode of the method is as follows:

Data = SegNorm	Coverage normalized read counts, except cell *X* $m \times n$ matrix of m bins and n cells
DataX = prcomp(Data)	PCA was applied on the normalized read counts
PCs = which(cumsum(DataX) > 90%)	Choose n PCs that explained at least 90% variation
Res = residuals(lm(cell *X*))	Fit a linear model to capture all the pattern in the data and get residuals
Res = Res+1	Add 1 to the residuals to prevent error during lowess fit of GC content
if(min(Res)≤0){ Res[Res ≤ 0] = posMin(Res[Res > 0])}	If there, replace values less than or equal to zero with the smallest positive number

Also, PCA of the normalized read counts across different datasets was performed in *R* 3.6.3 using the "prcomp" function with the parameter "scale.=FALSE".

2.7 Data access

All data from this study have been submitted to the European Nucleotide Archive (ENA) repository under accession number PRJEB51941. The codes and additional information can be found in the Github repository (`https://github.com/zgturan/brain_CN`).

CHAPTER 3

RESULTS

3.1 Summary of the dataset

We used scWGS to determine the frequency of CNVs in the temporal cortex, hippocampal subfields cornu ammonis (CA) 1, hippocampal subfields cornu ammonis (CA) 3, cerebellum (CB) and entorhinal cortex (EC) of 13 AD patients and 7 age-matched healthy controls (Table 3.1). We collected samples from multiple regions from 8 donors, and only the temporal cortex from the rest. Overall, we maintained a balanced distribution with respect to condition, donor, and brain regions. Figure 3.2A shows the distribution of cells according to each individual and brain region.

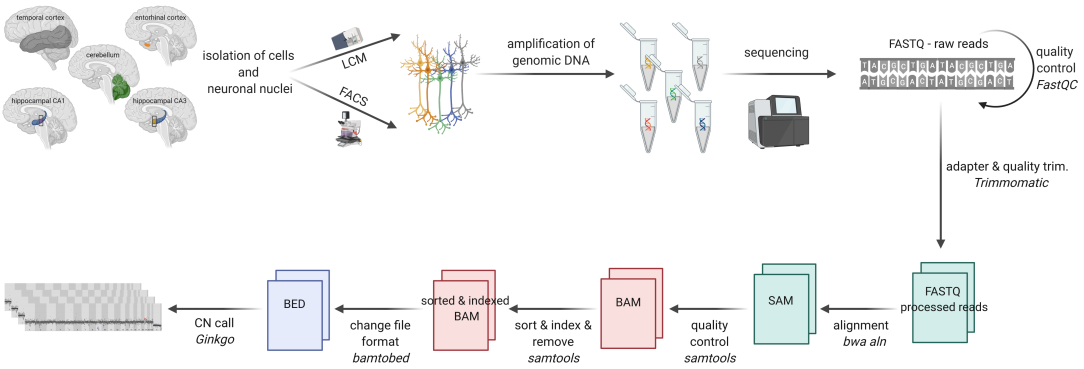


Figure 3.1: Schematic of the workflow. The pipeline of NGS data analysis and CNV detection. Figure was created with `https://biorender.com/`

Neuronal nuclei were isolated using either FACS (sorted with propidium iodide, $n = 12$) or LCM (sorted with cresyl violet, $n = 1552$), the latter performed on frozen brain slices (Figure 3.3). LCM-isolated non-neuronal "blank" regions were used as

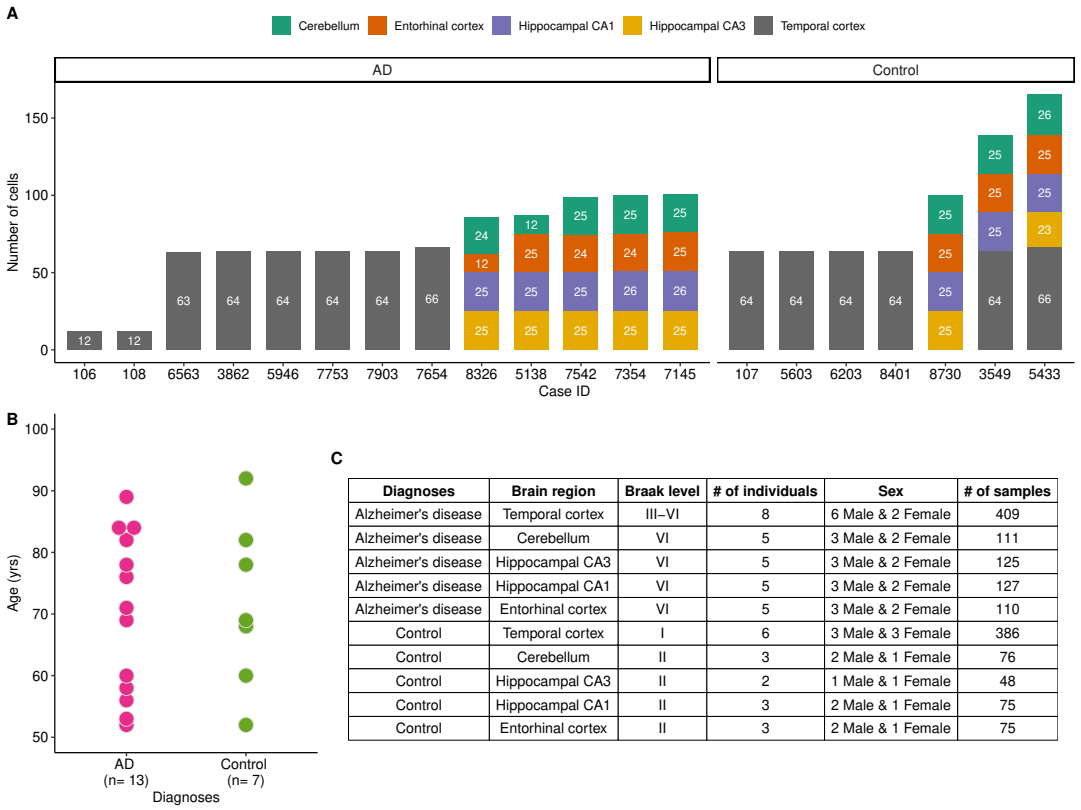
negative control ($n = 10$). The LCM method, although more difficult to implement than FACS, was chosen to ensure the selection of nuclei of pyramidal neurons for sequencing, known to be particularly sensitive to AD [147]. For technical comparison, neurons of a single individual were collected both using FACS ($n = 12$) and LCM ($n = 64$). scWGS libraries were prepared using GenomePlex whole-genome amplification and specific adapters were inserted using Phusion® PCR. Paired-end reads were mapped to the human reference genome, followed by stringent filtering to obtain uniquely mapped reads. This resulted in a median of 276,446 reads, corresponding to a coverage of 0.006X per LCM-isolated cell (range: [133 - 1,909,016] reads and [0.000003X - 0.04X] coverage) (Figure 3.5A).

Table 3.1: Summary of clinical and demographic variables of samples analyzed.

Case ID	Diagnoses	Age	Sex	Braak level	Post-mortem delay (h)
106	AD	52	male	4	58
108	AD	84	male	3	30
6203	control	78	male	0	23
3549	control	69	male	0	6
3862	AD	58	male	3	5.5
6563	AD	82	male	6	25
7753	AD	71	male	4	30
8401	control	60	female	0	28
5603	control	92	female	1	21
5946	AD	89	female	6	10
7903	AD	84	female	4	38
107	control	68	male	0	48
5433	control	52	female	2	29
7654	AD	78	male	6	8.5
7145	AD	56	female	6	26
7354	AD	76	male	6	27
7542	AD	53	male	6	Not determined
8730	control	82	male	2	63

Table 3.1 (continued)

Case ID	Diagnoses	Age	Sex	Braak level	Post-mortem delay (hour)
5138	AD	60	female	6	26
8326	AD	69	male	6	30



C

Diagnoses	Brain region	Braak level	# of individuals	Sex	# of samples
Alzheimer's disease	Temporal cortex	III–VI	8	6 Male & 2 Female	409
Alzheimer's disease	Cerebellum	VI	5	3 Male & 2 Female	111
Alzheimer's disease	Hippocampal CA3	VI	5	3 Male & 2 Female	125
Alzheimer's disease	Hippocampal CA1	VI	5	3 Male & 2 Female	127
Alzheimer's disease	Entorhinal cortex	VI	5	3 Male & 2 Female	110
Control	Temporal cortex	I	6	3 Male & 3 Female	386
Control	Cerebellum	II	3	2 Male & 1 Female	76
Control	Hippocampal CA3	II	2	1 Male & 1 Female	48
Control	Hippocampal CA1	II	3	2 Male & 1 Female	75
Control	Entorhinal cortex	II	3	2 Male & 1 Female	75

Figure 3.2: Sample information. **(A)** Bar plot showing the number of cells that have been sequenced for each individual. Brain regions are illustrated in different colors (see the colour key on the top of the figure). **(B)** Dot plot showing age of AD (pink) and control (green) individuals. **(C)** The table summarises sex, diagnoses and Braak level of the individuals. The Braak stages of AD patients ranged between III and VI.

CNVs were predicted using the *Ginkgo* algorithm, which uses circular binary segmentation (CBS) to estimate deletion or duplication events [3]. Negative controls ($n = 10$) and FACS-isolated neurons ($n = 12$) were analyzed separately and are not included in the main results. *Ginkgo* was run on our dataset with $n = 1542$

cells, while in parallel, two published scWGS datasets were also analyzed: one by van den Bos and colleagues ("van den Bos 2016"), comprising $n = 1469$ cells from healthy and AD brains (median coverage 0.005X), and another by McConnell and colleagues ("McConnell 2013"), comprising $n = 110$ cells from healthy brains (median coverage 0.047X). Note that the van den Bos 2016 dataset includes only 61% of cells produced in that study, because data from cells filtered for high noise levels were not published and thus could not be included here.

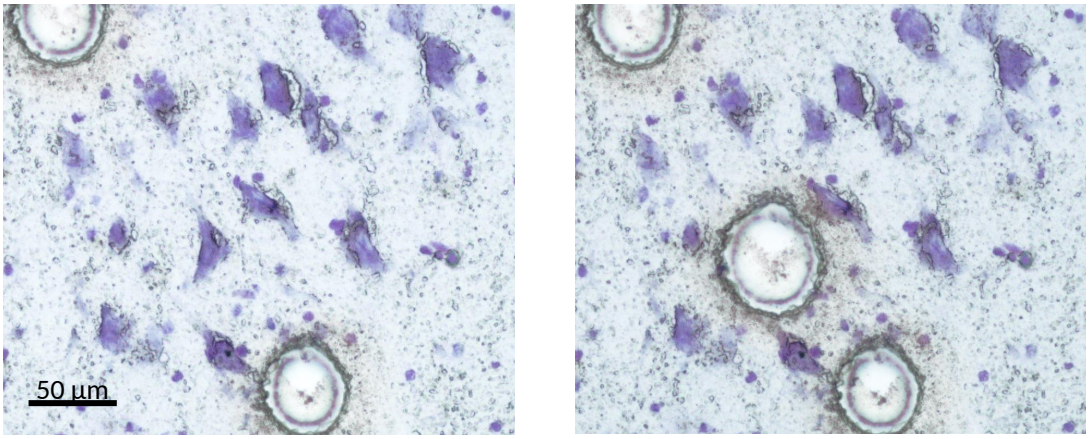


Figure 3.3: Images from a frozen hippocampal brain slice stained with cresyl-violet showing a pyramidal cell before (left) and after (right) laser capture microdissection-based isolation process using the PALM device. Circles in (left) indicate positions where two pyramidal cells have already been isolated just prior to the picture being taken. Scale Bar, 50 µm

3.2 LCM-isolated cells show a high frequency of depth variability

We first evaluated the sensitivity and specificity of our bioinformatics pipeline on scWGS data using trisomy-21 in DS and monosomy-X in males in published data. Analyzing $n = 34$ neuronal nuclei from DS individuals [126], trisomy-21 was correctly predicted across all samples without any false positive or false negative calls. In addition, monosomy-X was accurately predicted in 94.2% (338 of 359) of cells from males across the two published datasets [126, 107] (Figure 3.4).

Ginkgo includes an algorithm that uses the distribution of read depth across the genome to infer the average DNA copy number of each cell, which is estimated within a range of 1.5 to 6. It would be expected that the majority of human neurons would carry on average two copies of each autosome, although high frequencies (10-35%) of hyperploid neurons have also been reported, especially in AD brains [139].

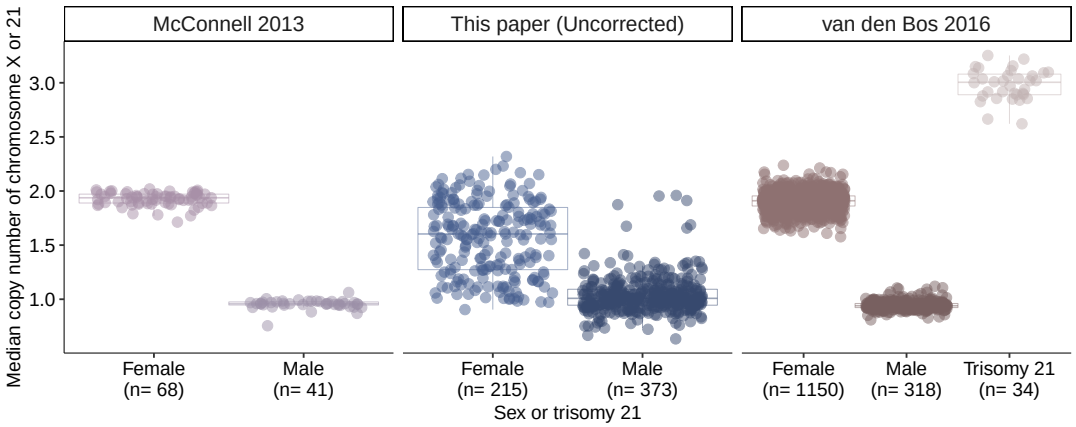


Figure 3.4: Median copy number profile of chromosome X or 21 across datasets. Median of the distributions: McConnell Female: 1.94, McConnell Male: 0.96; van den Bos Female: 1.91, van den Bos Female: 0.94, van den Bos Trisomy: 3.00; This paper (Uncorrected) Female: 1.60, This paper (Uncorrected) Male: 1.00.

Applying *Ginkgo* on the two published datasets, we found that for 99.9% (1577 of 1579) of cells the estimated average copy number lies within [1.9-2]. Using the same algorithm on our dataset, however, only 45% (687 of 1542) of the cells had average copy numbers estimated within the [1.9-2] range; i.e. 55% were non-euploid. Although hyperploid neurons have been described in control brains at ~10% frequency using FISH [139], the observed non-euploidy estimates suggest that our dataset carries particularly high levels of variability in read depth. These differences, in turn, could be related to the LCM protocol used, as the published scWGS experiments had used FACS.

To investigate this possibility, we compared the quality metrics of cells we had

collected using FACS or LCM for this study. These metrics were mapping proportion (the number of mapped reads/ the total number of reads), coverage, and index of dispersion (IOD, the ratio between the variance of read coverage and the mean).

FACS-isolated cells had higher sequencing coverage and mapping proportions than the LCM-isolated ones (Wilcoxon two-sided rank-sum test, $p < 0.0001$ and $p < 0.001$ for coverage and mapping proportion, respectively) (Figure 3.5A-B). Note that the difference in coverage variability between FACS and LCM has not been reported elsewhere. In addition, FACS-isolated cells had low IOD values, indicating less variation in sequence depth than the rest of the samples (Kruskal–Wallis test, $p = 1.5e - 07$) (Figure 3.5C).

Because our LCM and FACS samples originated from different brain regions with different cell type proportions, we also asked whether such differences could explain the observed LCM vs. FACS differences. To rule out this possibility, we compared the index of dispersion value of the cells that were taken from the temporal cortex of the same individual using FACS ($n = 12$) and using LCM ($n = 64$). We found a significant difference in the direction of higher variability in LCM (Wilcoxon rank-sum test $p < 0.001$), indicating that the observed variability between LCM and FACS can not be simply explained by differences in cell type proportion among brain regions. We note that the higher noise observed in LCM data was not solely due to higher genome coverage, as the FACS-based data from the van den Bos 2016 dataset had a median coverage comparable to ours (0.005X vs. 0.006X), but did not show comparable variability as in our LCM data.

These differences in IOD between LCM and FACS could be potentially explained by the higher sensitivity of the LCM procedure to experimental noise, compared to FACS. Alternatively, they could partly represent abnormal nuclei selected out in FACS but captured by LCM.

We next investigated the possibility that underlying variation may be caused by technical and/or biological factors. For this, we used a generalized linear mixed model (GLMM) to explain IOD (the response variable) per LCM-isolated cell

($n = 1542$) as a function of diagnosis (AD vs. control), genome coverage, and brain region as fixed factors, and individual as a random factor.

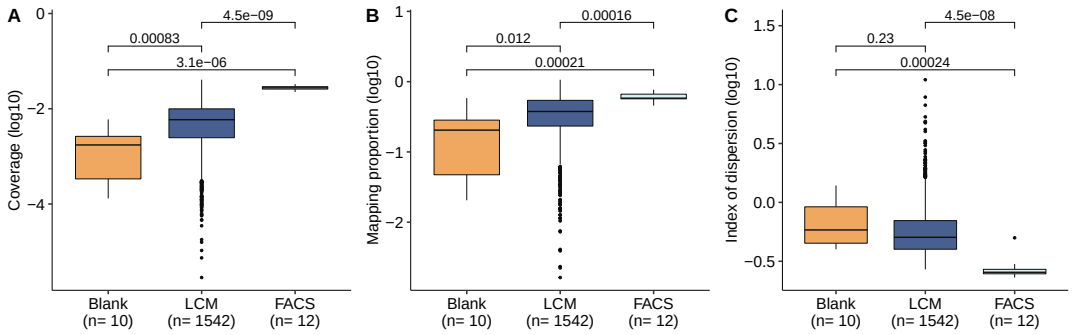


Figure 3.5: Comparison between different cell isolation methods. Box plots showing the distribution of coverage **(A)**, mapping proportion **(B)**, index of dispersion **(C)** among FACS-isolated, LCM-isolated and LCM-isolated blank samples. P-values were calculated using Kruskal–Wallis test among groups and Wilcoxon rank-sum test between groups.

We found that coverage has a significant negative effect on IOD, as may be expected ($z = -21.06$, $p < 0.0001$). Compared to the cerebellum, the region least affected by neurodegenerative diseases [20], we found a significantly high IOD for the entorhinal cortex ($z = 2.61$, $p < 0.05$), hippocampal CA1 ($z = 3.34$, $p < 0.001$) and hippocampal CA3 ($z = 3.75$, $p < 0.001$), but not for the temporal cortex ($z = -0.28$, $p = 0.78$) (Figure 3.6B).

Finally, neurons from control individuals have slightly less IOD than AD patients ($z = -1.93$, $p = 0.054$) (Figure 3.6). This result might suggest a tendency for neurons of AD patients to carry more variable DNA content and is consistent with cytometry analyses reporting a high occurrence of hyperploid neurons in the AD brain [139]. Although these findings imply a role of biological factors in read count variation within cells, it still remains possible that confounding technical factors influence our data. Given this uncertainty about the source of variability, we continued the analyses by filtering our dataset to remove the most variable cells.

3.3 No difference in CNV frequency between AD and control in the "uncorrected-filtered" dataset

We then used $Ginkgo$ to call CNV events from the "uncorrected-filtered" dataset ($n = 882$ cells from 13 AD patients, and $n = 660$ cells from 7 healthy controls). We found 19,608 events in 882 cells from AD patients (22.2 per cell), and 14,844 events in 660 cells from healthy controls (22.5 per cell).

We then tested the observed frequency difference between AD and control using a GLMM with a negative binomial error distribution. The response variable (the frequency of CNVs) was predicted using a combination of fixed factors, including diagnoses, chromosomes, brain regions, sex and coverage. The individual effect was added as a random factor. We found no statistically significant difference between AD and control across all tested combinations (GLMM, $p \geq 0.17$; Table U-Z).

CNV estimation from low coverage scWGS data is known to be highly sensitive to technical noise, and a large proportion of the called CNV events likely represent false positives. We thus decided to filter both cells and CNV events in our dataset to obtain a more reliable dataset [136, 127, 116].

We started by removing the most highly variable cells among the LCM-isolated ones ($n = 1542$) using the following criteria. First, 13% (205 of 1542) of the cells with a low number of reads (<50,000) were discarded from the analysis. Second, as most cells are expected to be diploid, and also given that the Ginkgo-estimated copy number (CN) profiles of 99% of cells in the McConnell 2013 and van den Bos 2016 datasets were observed to lie between [1.9-2], we excluded those cells with CN values beyond this range (54% excluded, 726 of 1337). Third, we filtered out 23 of the remaining 611 cells (4%) that showed extreme CNV intensity, which we defined as three or more chromosomes of a cell carrying predicted CNVs that cover >70% of their length (Figure 3.8). Information about the remaining cells ($n = 588$) is provided in Figure 3.19.

From these 588 cells, we called 3521 CNVs (~5.9 events per cell) in the uncorrected

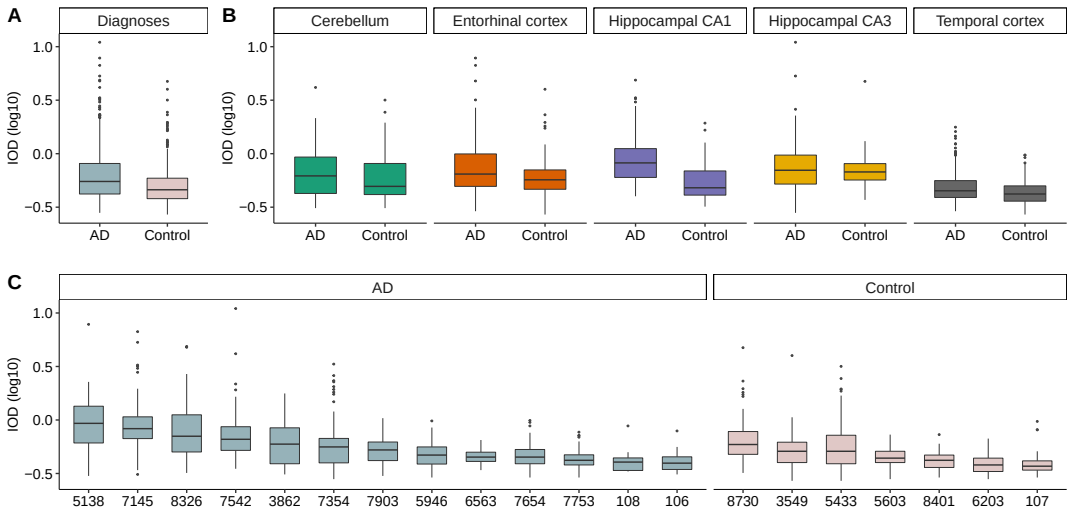


Figure 3.6: The distribution of index of dispersion ("IOD") for LCM-isolated cells ($n = 1542$) according to **(A)** diagnoses (Alzheimer's disease ("AD"), control), **(B)** brain regions (hippocampal subfields CA1 ("Hippocampal CA1"), hippocampal subfields CA3 ("Hippocampal CA3"), Entorhinal cortex, Cerebellum, Temporal cortex). For each brain region, we tested whether AD diagnosis was predictive of IOD using a linear mixed-effects (lme) model. Individuals were added as a random factor. Across all tested brain regions, differences were only marginally significant ($p = 0.069$). **(C)** The distribution of IOD across individuals ($n = 20$). Box plots were ordered by the median. Y-axes illustrate the IOD values on the log10 scale.

data, which we call the "uncorrected-filtered" dataset. We further applied a number of conservative filtering criteria to remove potential false positives: **(1)** We only included megabase scale CNVs ($\geq$10 Mb), considering that detection of small events with low coverage data will be unreliable. **(2)** We limited the analyses to 1-somy and 3-somy events, assuming that most somatic CNVs involving chromosomes or chromosome segments would involve loss or duplication of a single copy. **(3)** We only included CNVs with unique boundaries across all analysed cells, assuming that somatic CNV breakpoint boundaries should be generally randomly distributed across the human genome. **(4)** We removed CNVs on the proximal portion of the chr19 p-arm, where frequently observed duplications were previously reported as low coverage sequencing artifacts [166]. **(5)** To ensure the reliability of the CNV signal, we calculated a standard Z-score for each CNV that reflects the deviation in

read count distribution in that region compared to the rest of the cell (which we call Z_1, see Methods), and only accepted CNVs with absolute values of Z_1-scores $\geq$2. **(6)** We reasoned that read counts in a real CNV should be closely clustered around expected integer values (e.g. 1 or 3). To assess this, we calculated a Z-score for the deviation from the expectation (called Z_2), and only accepted events with absolute values of Z_2-scores $\leq$0.5 (Figure 3.8, Figure 3.7).

After CNV filtering, we found 12 CNV events across 295 cells in 13 AD individuals and 4 CNV events across 293 cells in 7 controls (Table A). Among the 295 pyramidal neurons analyzed from the 13 AD patients, we found 10 deletions (3.39% per cell) and 2 duplications (0.68% per cell). These events ranged in size from about 10.14 to 77.01 Mb (median: 19.31 Mb) and were observed in the temporal cortex and the entorhinal cortex.

Of the 293 neurons from 7 control brains, 1 deletion (0.34% per cell) and 3 duplications (1.02% per cell) were detected in the temporal cortex with a size range of 10.81 to 54.67 Mb (median: 14.51 Mb). Again testing the CNV frequency differences between AD and control brains using a GLMM, we found no statistically significant effect (GLMM, $p \geq 0.88$) (Figure 3.9, Figure 3.17, Table D-J).

We also implemented an alternative algorithm, $HMMcopy$ [168], to predict CNVs (see Methods). Overall, 75% (12/16) of the $HMMcopy$ predictions overlapped with the CNV events that we found after filtering the uncorrected $Ginkgo$ predictions. Comparing predicted CNV event frequencies between AD and control we again found no significant difference ($z = -1.34$, $p = 0.18$).

3.4 A PCA-based denoising approach minimizes within-cell depth variability

To gain further insight into within-cell variability in our dataset (the uncorrected-filtered version) compared to the two published scWGS datasets, we calculated the median CN of chr1 and chr21 (the largest and smallest chromosomes) across all three. We still found conspicuously higher within-cell variation in our dataset, despite having discarded highly variable cells (Figure 3.10).

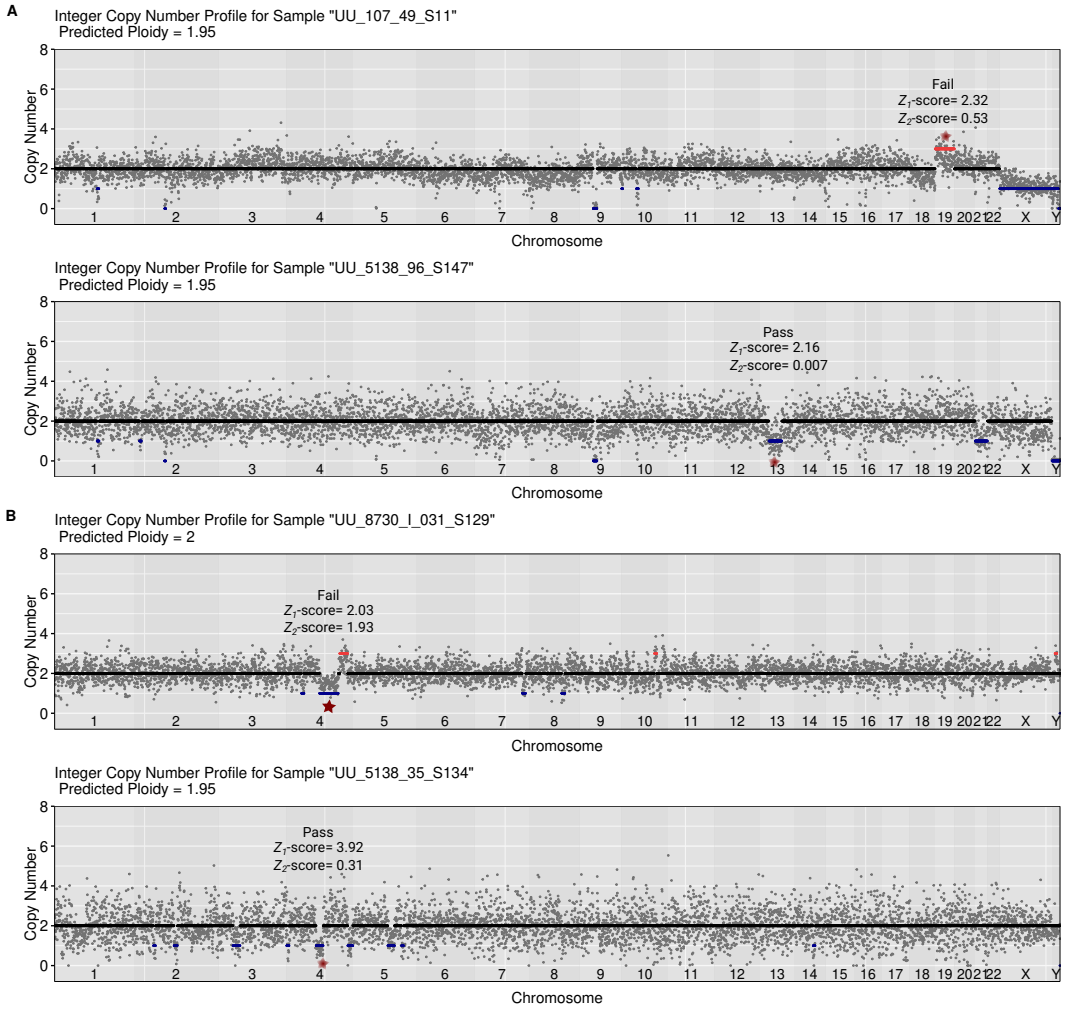


Figure 3.7: Examples of CNVs that (upper panels) failed to pass and (lower panels) passed the filtering criteria (Z_1-scores $\geq$2 and Z_2-scores $\leq$0.5) in the uncorrected-filtered data **(A)** and PCA-corrected data **(B)**. Absolute values of the Z_1- and Z_2 scores for the CNV were indicated on the plot. Mentioned CNV was marked with a red star. X-axes show chromosomes and Y-axes illustrate the CN profile of chromosomes estimated by *Ginkgo*. Each grey dot represents the scaled read counts per bin. Amplifications (CN>2) are shown in red; deletions (CN<2) in blue; disomy (CN=2) in black.

We then used the autosomal normalized read counts to perform a PCA on the uncorrected-filtered data and published datasets. We also included blank (negative control) samples and FACS-isolated cells to illustrate how reads counts from these two groups relate to others. According to the PCA, LCM-isolated uncorrected-

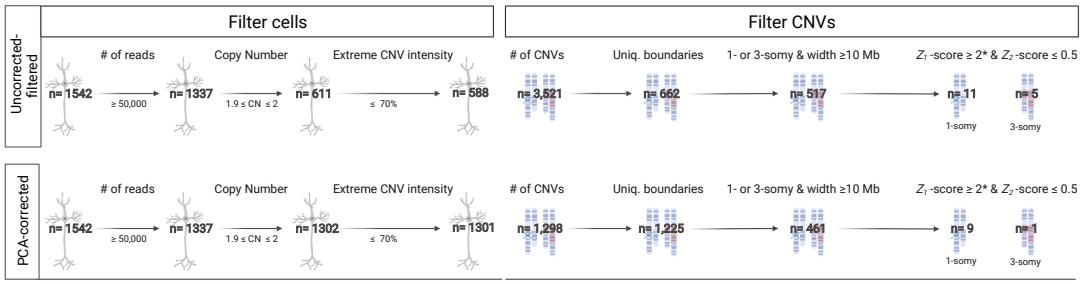


Figure 3.8: Overview of cell and CNVs elimination steps in the uncorrected and corrected data. *If CNVs' breakpoints were within the three base pairs window around each other, those were discarded. Figure was created with `https://biorender.com/`

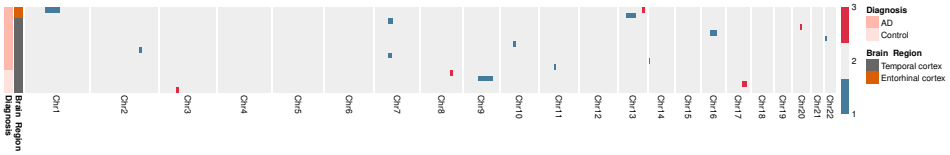


Figure 3.9: The heatmap shows the genome-wide copy number profile of cells analyzed in the uncorrected data ($n = 15$) with at least one reliable CNV. CNVs, brain regions and diagnoses are illustrated in different colors (see the colour key on the left of the figure). Each row shows a cell and each column shows a chromosome.

filtered data and blank samples were separated from the published datasets and FACS-isolated cells (Figure 3.11). This result might also highlight distinct profiles of LCM-isolated cells.

We then sought an approach that could reduce this elevated within-cell variability in read depth, assuming it is of technical origin and possibly related to the LCM procedure. Experimental biases could involve cross-contamination across cells during isolation, or biases that arise during DNA amplification. Although the former should be mainly random, the latter may follow systematic patterns, such as some chromosome segments being more or less prone to be amplified.

We thus devised a procedure for removing putative patterns of systematic read depth variation across cells (see Methods). The algorithm starts by choosing a focal

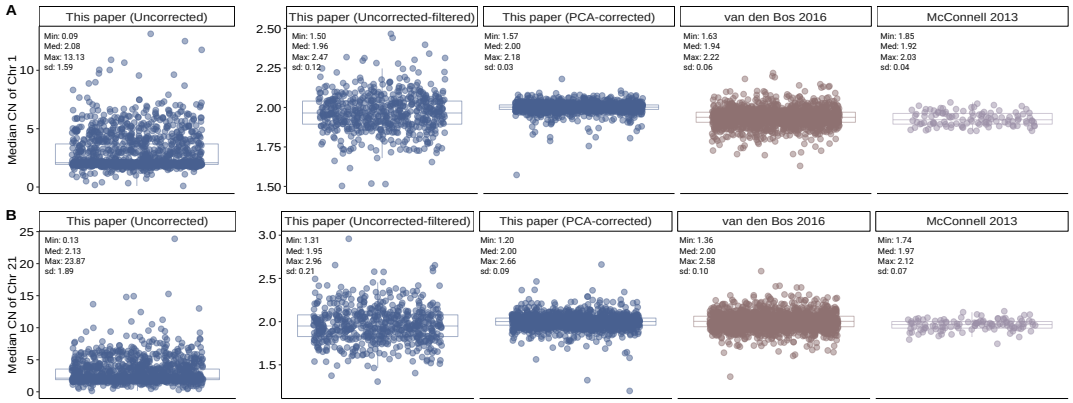


Figure 3.10: Boxplots showing the distribution of median CN of chromosome 1 (chr1, upper part of the figure) and chromosome 21 (chr21, lower part of the figure) across bins ($n = 440$ and $n = 68$ for chr1 and chr21, respectively). Each point corresponds to the median CN of each cell. Minimum ("Min"), median ("Med"), maximum ("Max") and standard deviation ("sd") of each distribution were shown on the boxplot. Cells that deviated from the [1.9-2] range were excluded from the analyses to be consistent with our filtering criteria (except for the uncorrected datasets). This study [Uncorrected ($n = 1337$), Uncorrected-filtered ($n = 588$), PCA-corrected ($n = 1301$)]: blue; van den Bos 2016 ($n = 1468$): brown; McConnell 2013 ($n = 109$): purple.

cell x in the dataset, and calculating principal components (PCs) from the normalized read counts per autosome across the rest of the cells (except cell x). It then collects all PCs explaining $\geq$90% of the variance. Treating these as representatives of systematic variation, it removes their values from the normalized read counts of cell x using multiple regression analysis. These steps are performed on all cells individually, creating a "denoised" dataset. The final dataset contains residuals from the multiple regressions instead of the normalized read counts. Notably, this procedure should remove experimentally-induced variation in read depth shared among cells, and also any recurrently occurring somatic CNVs. Rare somatic CNVs, instead, would be mostly unique to each cell and randomly distributed in the genome, and thus would not be affected.

To test the accuracy of the PCA-based denoising approach, we used published Down's syndrome data [126]. To prevent the algorithm from capturing the common

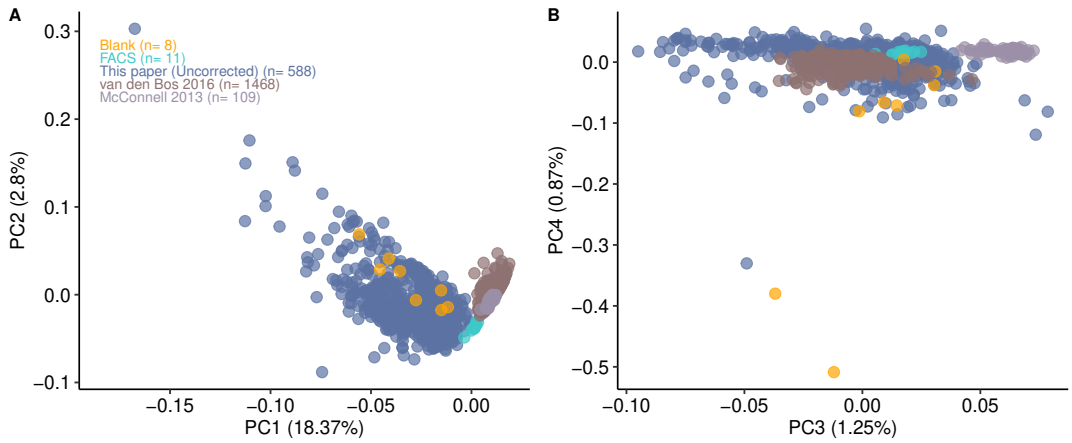


Figure 3.11: A principal components analysis (PCA) was performed using the normalized read counts across autosomal bins ($n = 5243$) in published datasets and this study. Because they dominated the PCs, cells deviating from the [1.9-2] range were not included in the analyses. The number of cells for each dataset were indicated on the plot. X-axes illustrate PC1 and PC3 that explain 18.4% and 1.3% of the total variance, respectively. Y-axes show PC2 and PC4 that explain 2.8% and 0.9% of the total variance, respectively.

variation in trisomy 21, we only included one cell with DS at a time. The common variation was removed in 10% intervals, starting from 50% (Figure 3.12). When we remove 50% and 60% of the variations, we can capture trisomy 21 in 97% (33 of 34) of the cells. However, removing more than 60% of the variation results in disomy in chromosome 21. We will look for the reasons for the loss of the aneuploidy signal.

After filtering cells with a low number of reads ($n = 205$) and denoising our dataset with this approach, CN and CNV prediction were performed using *Ginkgo*. We further compared the results between the PCA-corrected and "uncorrected-filtered" datasets. Examples of cells having "noisy" profiles before and after correction are shown in Figure 3.13, which suggests a dramatic reduction in within-cell variability. Beyond visual inspection, we also analyzed three statistics. First, we studied the CN profile of cells after PCA correction. We found 97% (1302 of 1337) now lie between 1.9 and 2. This result is comparable to the two published datasets described above and much higher than uncorrected data (45%). Second, we

calculated the number of CNV events per cell (sum of the number of CNV/ number of cells) across datasets. In the van den Bos 2016 and McConnell 2013 datasets, we estimated 5.6 and 8.1 CNVs per cell, respectively (Figure 3.15E). In our dataset, in the uncorrected version, we found 23.9 CNVs per cell, in the "uncorrected-filtered" data 6.0 CNVs per cell, and in the PCA-corrected data, we estimated on average 1.0 CNV event per cell. The denoising leads to lower CNV estimates in our data, which is more conservative and possibly more realistic than the higher estimates without correction (Figure 3.16). Third, we estimated the standard deviation in CN among cells for chr1 and chr21. For chr1 and chr21, the standard deviations in the PCA-based data were 4 and 2.3 times lower than in the "uncorrected-filtered" data, respectively, and comparable to CN standard deviations in the two published datasets (Figure 3.10).

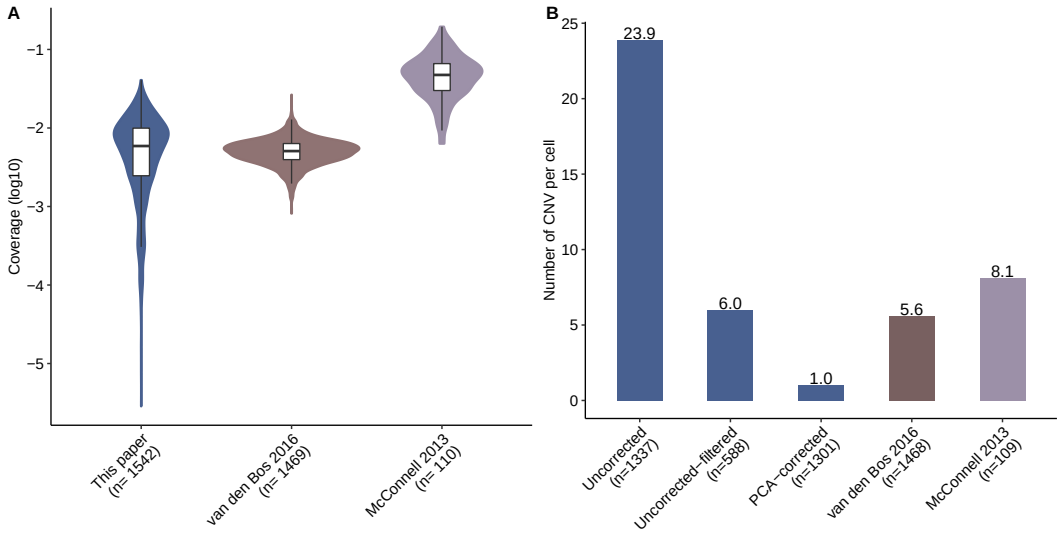


Figure 3.15: Properties of published datasets and this study. **(A)** Violin plot showing the distribution of coverage among different datasets. This study, including only LCM: blue; van den Bos 2016: brown; McConnell 2013: purple. **(B)** Bar plot showing the number of CNVs per cell across datasets. The datasets from our study include cells from both AD and control (Uncorrected, Uncorrected-filtered, PCA-corrected): blue; van den Bos 2016 (including cells from both AD and control): brown; McConnell 2013: purple). The number of cells that were used to calculate CNVs per cell was shown on the X-axis label.

3.5 Subchromosomal CNVs are enriched in deletions in the PCA-corrected data

Based on these three statistics, we decided to study this PCA-corrected version of our dataset. For downstream analysis, we further eliminated cells that deviated from the ploidy range of [1.9-2] (2.6%, 35 of 1337) or showed extreme CNV intensity (0.08%, 1 of 1302) (Figure 3.8A). We thus created a denoised dataset of 1301 pyramidal neurons from 20 individuals (Figure 3.19).

Estimating CNVs in this dataset using $Ginkgo$, we found 1298 CNVs in total (~1 event per cell). To remove false positives, we also performed the same CNV prediction and downstream analyses on our PCA-corrected data (Figure 3.8A). After these steps, we found a total of 9 deletion events (0.7% per cell) and 1 duplication event (0.08% per cell) across 1301 cells in 20 individuals among all tested brain regions (except for the hippocampal CA1 where no CNV event was found). This excess of deletions is unexpected under the null hypothesis of equal expectation of duplication and deletions (two-sided binomial test $p = 0.021$), but consistent with previous observations of more deletions than duplications among somatic mutations [136, 133, 107, 116].

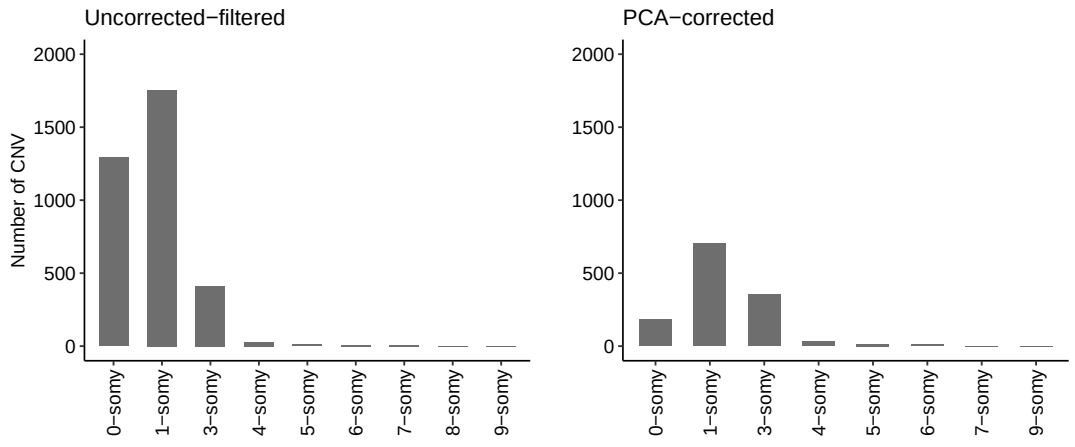


Figure 3.16: Distributions of the number of CN before ($n = 3,521$) and after correction ($n = 1298$) across autosomes. For illustration purposes, the bar plot includes up to 10-somy.

3.6 No significant difference between AD and control after PCA-correction or in the van den Bos dataset

Studying CNV frequencies with respect to diagnosis, we found 6 CNV events across 688 cells in 13 AD individuals and 4 CNV events across 613 cells in 7 controls (Fig. 6E). Performing the formal test for the hypothesis of AD versus control differences with this data, we again found no significant difference between the groups (GLMM, $p \geq 0.80$; Table K-Q). Information about the CNVs and cells can be found in Table B.

We also repeated the same analysis on the van den Bos 2016 dataset, from which originally only aneuploidy was reported. Here we identified 11 CNV events across 883 cells in 10 AD individuals and 3 CNV events across 585 cells in 6 controls (Table C). The difference was in the same direction as in our dataset, but again not significant (GLMM, $p \geq 0.79$) (Figure 3.17, Table R-T).

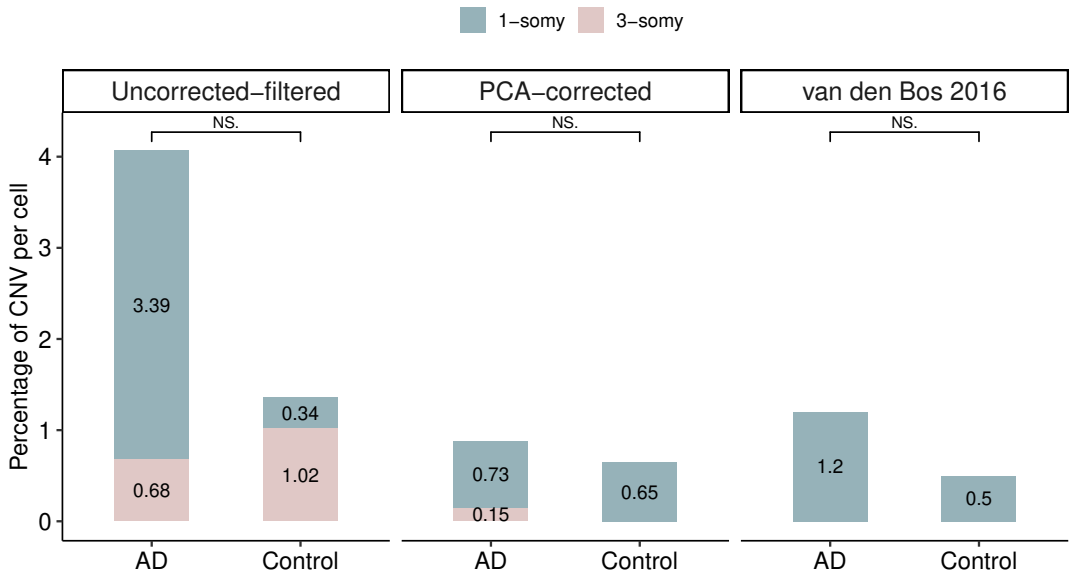


Figure 3.17: Bar plots represent the number of CNVs per cell for AD and control groups. Blue area: 1-somy (deletions); pink area: 3-somy (duplications).

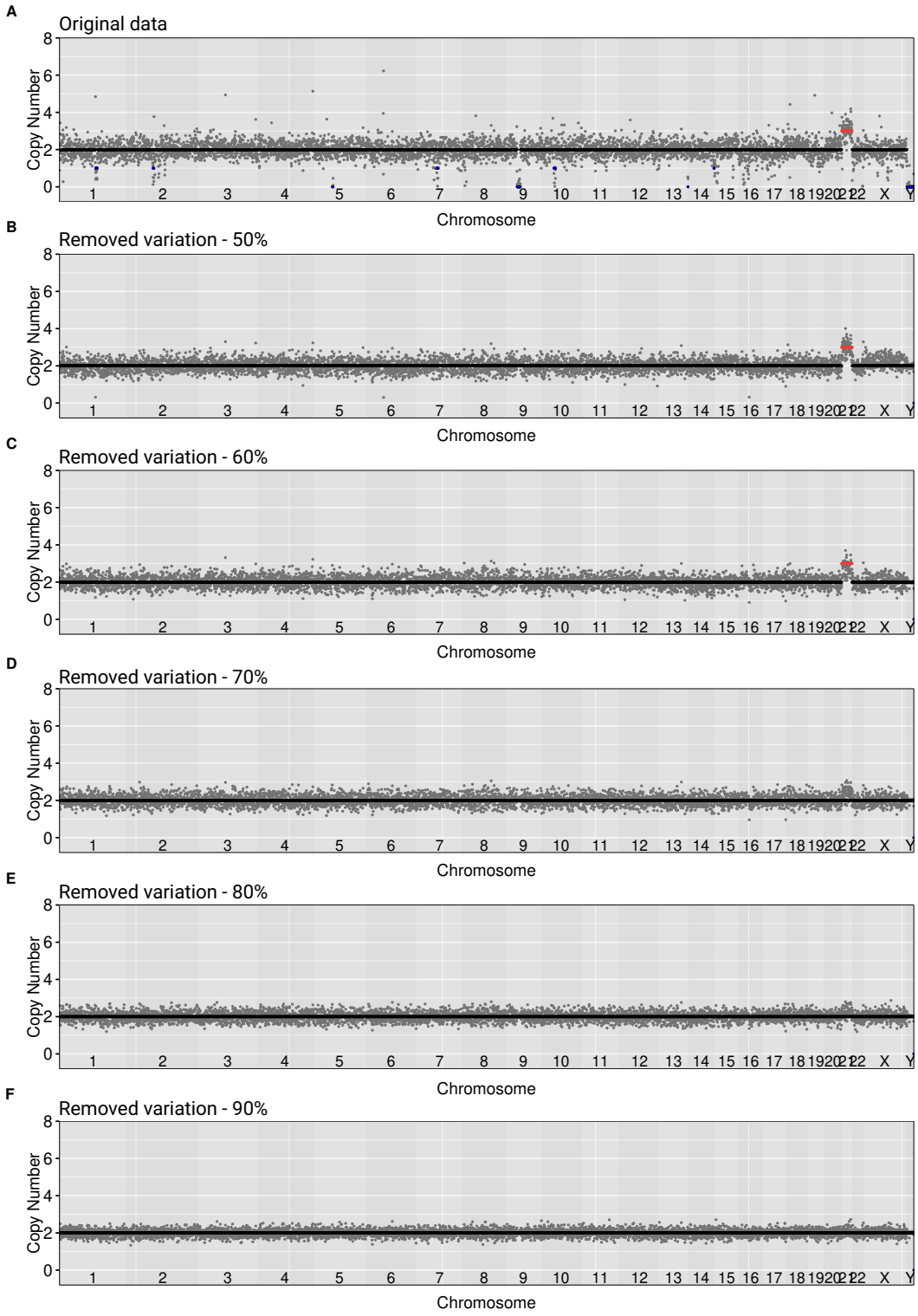


Figure 3.12: Accuracy of the PCA-based denoising approach. The sample ID is ERR1378095. Examples of CN estimates of cells using uncorrected data **(A)** and using data after PCA-based correction **(B-F)**. The x-axes show chromosomes and the y-axes show the CN profile of chromosomes estimated by *Ginkgo*. Each grey dot represents the scaled and normalized read counts per bin. Amplifications (CN>2) are shown in red; deletions (CN<2) in blue; disomy (CN=2) in black.

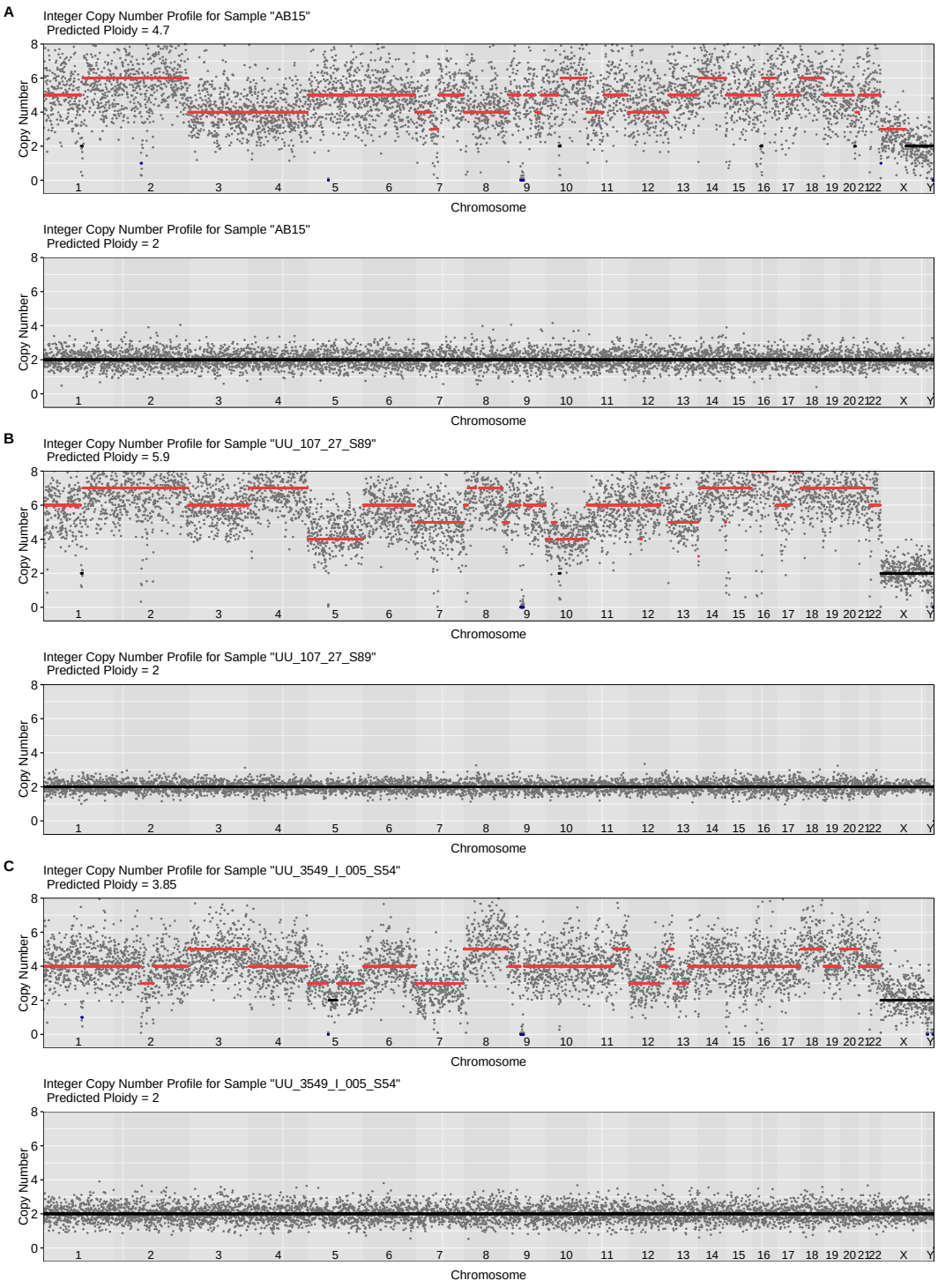


Figure 3.13: A PCA-based denoising approach helps eliminate technical noise. **(A-C)** Examples of CN estimates of cells using uncorrected data (upper panels) and using data after PCA-based correction (lower panels). The x-axes show chromosomes and the y-axes show the CN profile of chromosomes estimated by *Ginkgo*. Each grey dot represents the scaled and normalized read counts per bin. Amplifications (CN>2) are shown in red; deletions (CN<2) in blue; disomy (CN=2) in black.

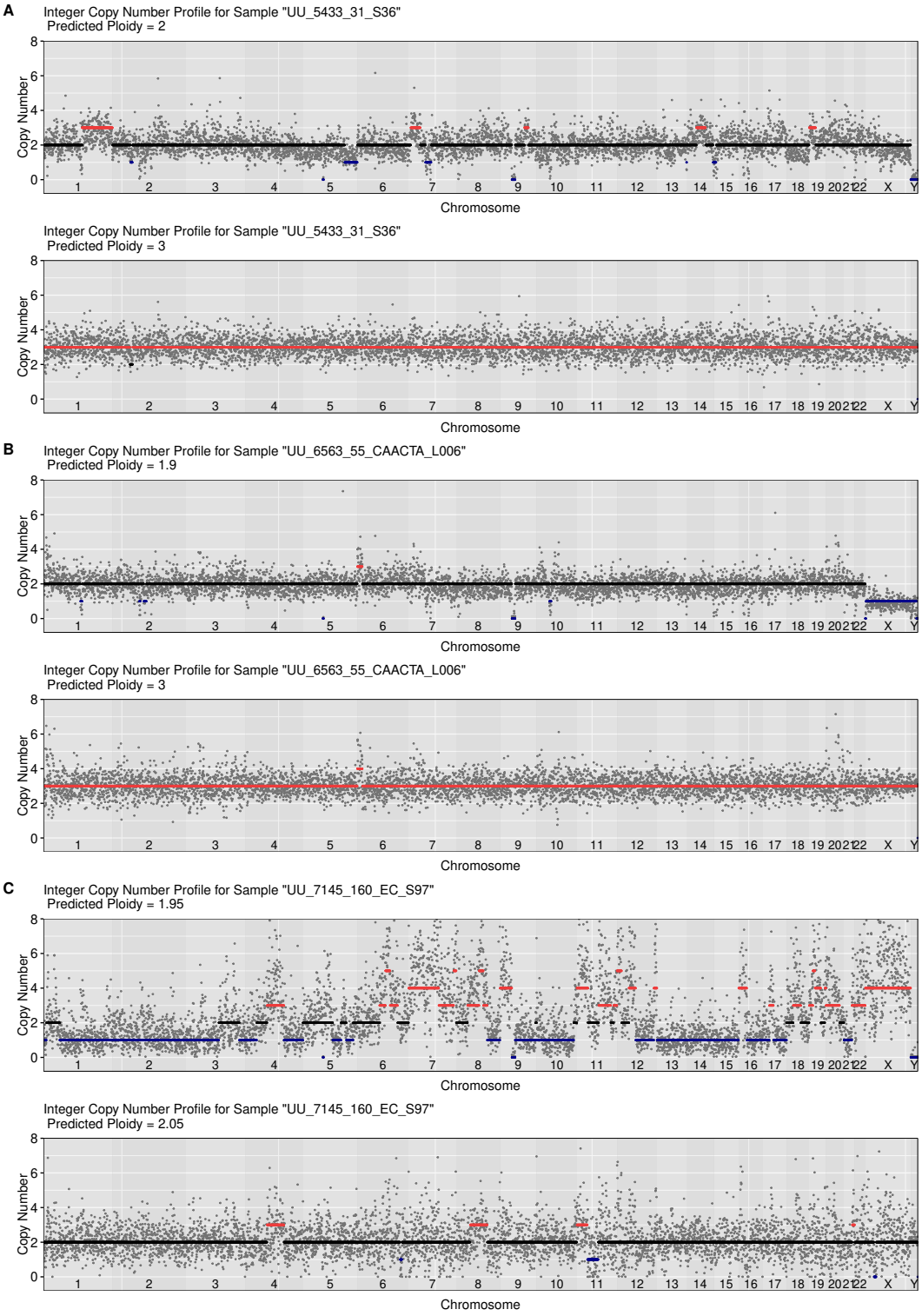


Figure 3.14: CN profiles of three cells deviating from the range of 1.9 and 2 after correction. The upper panels show the uncorrected data, and the lower panels show the corrected data. X-axes show chromosomes and Y-axes illustrate the CN profile of chromosomes estimated by *Ginkgo*. Amplifications (CN>2) are shown in red; deletions (CN<2) in blue; disomy (CN=2) in black.

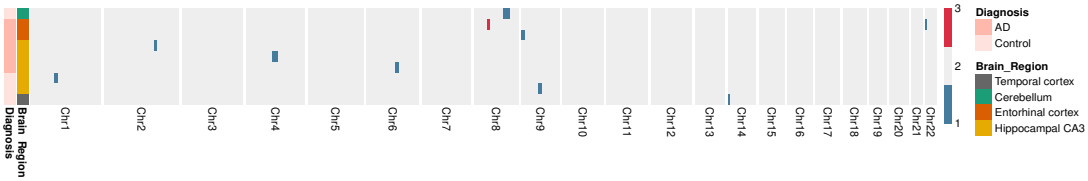


Figure 3.18: The heatmap shows the genome-wide copy number profile of cells analyzed in the corrected data ($n = 10$) with at least one reliable CNV. CNVs, brain regions and diagnoses are illustrated in different colors (see the colour key on the left of the figure). Each row shows a cell and each column shows a chromosome.

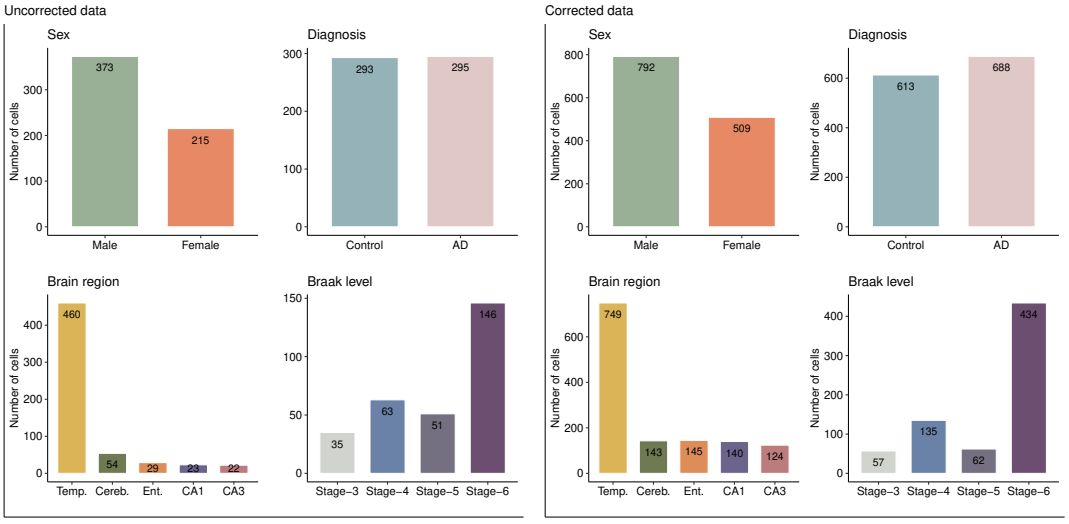


Figure 3.19: Information about the samples of uncorrected and corrected data. Distribution of samples according to sex (female, male) **(A)**, diagnoses (Alzheimer's Disease, control) **(B)**, brain regions (temporal cortex, cerebellum, entorhinal cortex, hippocampal subfields CA1, hippocampal subfields CA3) **(C)** and Braak level (stage 3, 4, 5 and 6) **(D)**.

CHAPTER 4

DISCUSSION

The sequence of the human genome was completed in 2003 [184]. The cost of sequencing has been declining rapidly since the beginning of the human genome project. The availability of the vast amount of data and the desire to identify biological phenomena, including the role of copy number variations in complex diseases, gave fresh impetus to scientific research. Studies using cytogenetic techniques reported a high frequency of CNVs in the AD brain [137, 140, 117, 138, 139, 113]. On the other hand, the scWGS study showed that there was not any difference between the AD brain and healthy control in terms of CNVs [126]. In this work, we produced scWGS data to solve this discrepancy.

Here we discuss technical aspects and the biological outcomes of our work.

4.1 The sources of variability among LCM-isolated cells

To the best of our knowledge, this is the first study to use LCM to collect neuronal nuclei for scWGS. Our results showed that LCM-isolated cells showed significantly higher within-cell read depth variation compared to FACS-isolated ones (Figure 3.5C). One random source of high variation could be cross-contamination of LCM-isolated cells during the isolation [185], which in turn might be reflected in the downstream analysis as duplications. In line with this possibility, we found that the number of duplications ($\geq$3-somy) is higher than the number of deletions (0- and 1-somy) in the uncorrected data (deletion to duplication ratio: 0.39). We applied several elimination steps to remove "noisy" cells and to filter nominally false positive

CNVs. After these elimination steps, the deletion to duplication ratio increased to 6.46 in the uncorrected-filtered data.

In addition to filtering the uncorrected data, we devised a PCA-based denoising approach to remove systematic variation across the genome, which could be experimentally-induced, but could also reflect convergent somatic CNVs shared among different individuals. Segments systematically deviating from the genome average have also been described in other neuronal scWGS datasets [136]. Our results showed that PCA-based denoising can strongly reduce within-cell variance in CN among cells (Figure 3.13). If the noise that was removed is experimentally-induced, then our result means that this noise was partly shared among cells and not entirely random. One source of systematic bias might be genome-wide variation in the propensity to DNA degradation and/or DNA amplification, perhaps due to GC content, chromatin structure or nuclear location of chromosomal segments [133]. Such biases would be shared among cells and effectively removed by PCA.

Beyond technical biases, biological factors could also explain the higher read-depth variability in LCM-isolated than FACS-isolated neurons. Chronister and colleagues recently reported that CNV frequencies in neurons (4%–23.1%) are higher than non-neuronal cells (4.7%–8.7%) [136]. Moreover, cytological studies suggested that AD brains harbour hyperploid neurons more frequently than healthy controls [139]. Consistent with the latter report, we found that neurons from AD patients tend to have higher IOD than control individuals. Also, the cerebellum, which is relatively spared from AD, had lower IOD than the entorhinal cortex and hippocampal areas (but not the temporal cortex). This might be interpreted as a reflection of biological factors on the read-depth variation which is captured efficiently in LCM data. Indeed, if the FACS procedure eliminated cells having abnormal karyotypes, this would result in a cell population with artificially uniform and "clean" ploidy levels. In conclusion, we predict that although random factors (e.g. contamination) and systematic biases most likely contribute to relatively high variation in LCM-collected scWGS data, biological variation may also be a contributor.

To disentangle biological variation from technical variation, the following approaches could be taken. First, apart from our work, there is only one study which tested the difference between AD and control in terms of CNVs [126]. In that study, the authors discarded the large proportion of cells (39%) having high within-cell variability (data is not available) [126]. Analyzing the discarded cells from this study and comparing the read-depth variability between AD and control, as we did in our data, would have given a better understanding of the origin of the variability.
Second, the biological phenomenon underlying the hypothesis being tested should be considered, as we did in the PCA-based denoising approach. CNVs are expected to be randomly distributed in the genome without any common patterns across different individuals or datasets. Thus, taken into account, PCA could have been applied to the other published datasets. Then we tried to answer whether the variation that was captured in our dataset is correlated with the others. If it has a technical origin, we expect to find a high correlation among PCs that are captured in different datasets. The same approach could have been applied to FACS-isolated cells in our data.
Third, one can test the difference between AD and control, as we did in our work, using different thresholds to eliminate cells with high variability. Even though this approach might not help differentiate technical noise from biological variation, it will demonstrate the robustness of the results.

4.2 PCA-based denoising: advantages and caveats

scWGS is a promising method for predicting CNVs with limited sequencing per cell. However, as in our study, within-cell variation that may represent false-positive CNVs hinder analyses in low coverage data. Our PCA-based denoising method can be used as a practical solution for *in silico* cleaning of such data. The approach is based on the idea that somatic CNVs are randomly distributed in the human genome and are particular to each cell. One possible drawback of this approach is that if some neurons from the same individual share the same CNV due to shared developmental ancestry, our method will eliminate such real signals. A more subtle approach could take into account possible clonal relatedness among cells [186].

The PCA-based denoising is expected to have removed any CNVs and aneuploidies that are shared among neurons (instead of being cell-specific), due to common origin in the same individual or due to recurrent mutations. Therefore our results only pertain to single cell-specific CNVs.

Another drawback could arise if certain genomic regions are predisposed to undergo copy number changes; in that case, our method may cause overcorrection. In our dataset, observing an unexpectedly high frequency of CNVs (23.9 events) per cell in the uncorrected version, we chose to remove $\geq$90% of the common variance. After applying PCA-based correction, the CNV rate per cell decreased by 96%. This, in turn, resulted in a lower number of CNVs per cell in the corrected data, even compared to published datasets (see Figure 3.15B, van den Bos 2016: 5.6, McConnell 2013: 8.1 CNVs, PCA-corrected data: 1.0 per cell). This difference might be attributable to the overcorrection of normalized read counts.

4.2.1 Possible application of PCA-based denoising approach in other fields of biology

In cancer studies, identifying whether different tumours in the same patient develop from a single ancestral cell or share a clonal origin (also known as metastasis) are of paramount importance. If the second tumour originates from a different founder cell, that is known as the second primary cancer [187, 188, 4]. Origin of recurrence has clinical implications by defining which treatment would be applied. For example, assume a patient with head and neck cancer developed lung cancer after treatment [189, 190]. If lung cancer arises from different precursor cells, then surgery is one commonly used method to properly treat second primary lung cancers. Otherwise, if lung cancer originated from the same ancestral cell that means that can also spread to other parts of the body [187].

Genetic fingerprints of tumour cells are used to identify the origin of tumour cells including loss of heterozygosity (LOH) and analysis of copy number variations. It is expected that clonal tumours share common somatic mutations like

chromosomal aberrations [191, 192, 193]. One possible outcome of chromosomal aberrations (i.e. chromosome loss) is the loss of heterozygosity (LOH) which is characterized by heterozygous non-cancerous cells for a given loci and homozygotes tumour cell [188, 194, 195, 196, 197]. LOH occurs frequently in the regions that are associated with tumour suppressor genes. This means chromosome loss promotes tumorigenesis by affecting the expression profile of genes that are related to cell proliferation [187, 198].

For example, Bollet and colleagues hypothesized that loss of tumour suppressor genes can be observed in both clonal and primary seconder cancer. However, if the two tumour cells share identical breakpoints for deletion that cover the tumour suppressor genes, these two cells should be clonally related [199]. Figure 4.1 shows the copy number profile of two lung tumours that have the same origin from the same patient. They have the same deletions on chromosomes 1, 3, 6 and 22 [200, 4].

Overall, contrary to CNV analysis in the brain, cancer genomics seeks for a common breakpoint to relate two different tumour cells. We note that our PCA-based approach could also be used to detect recurrent breakpoints in single-cell cancer genomics. Because clonal cancer cells would also inherit the same CNVs, shared CNV breakpoints identified in PCAs can be used to study clonal evolution.

4.3 Limitations and Possible Improvements

Our study has several limitations:

- This study only focused on relatively large ($\geq$10 Mb) CNVs for sake of sensitivity. However, smaller CNVs may still be much more common and could have contributions to neurodegenerative disease. Future studies on somatic genomic variation in AD might therefore focus on a smaller scale ($<$10 Mb) CNVs, for which improvement of experimental protocols and/or the use of higher coverage data appears to be needed [169].
- Our analysis of published data from van den Bos et al. (2016) could not include a large fraction of cells that they had discarded for showing high depth

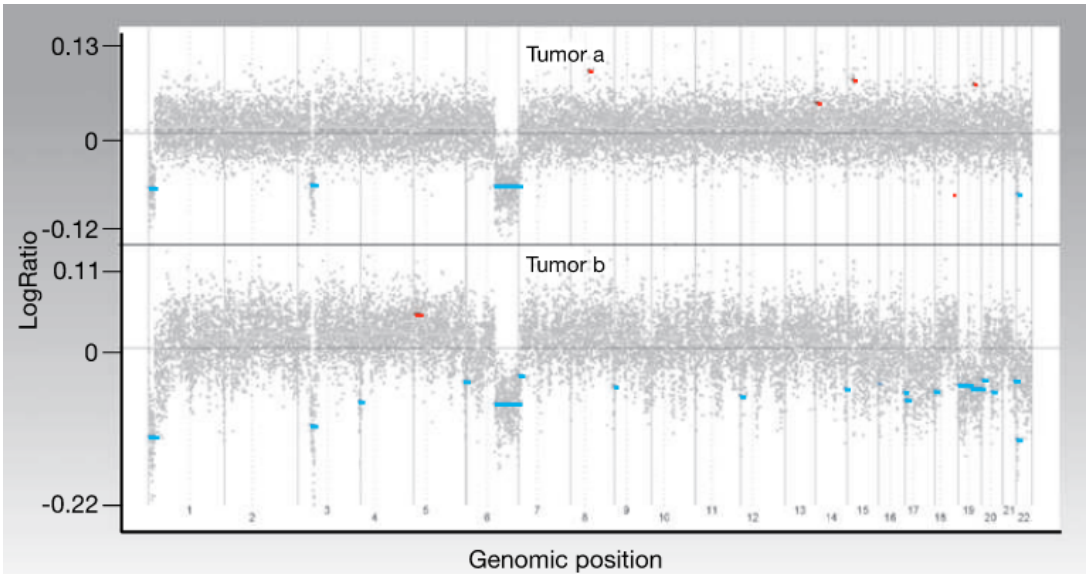


Figure 4.1: Copy number profile clonal lung tumours. X-axis shows chromosomes, Y-axis shows low ratio of the read counts. The plot was taken from [4].

variability [126]. Even though studies exclude the cells assuming their technical origin, the data belonging to excluded cells should also be uploaded to the public repository. This is especially important for the research areas with controversial findings. In our case, cytogenetic-based studies reported the increased frequency of CNVs in AD. One might think that this might be reflected as high within-cell variability in single-cell studies. Considering this, all data from this study were submitted to the ENA repository.

- Recent work has suggested that CNV-bearing neurons may be eliminated through lifetime in neurotypical individuals [136], and work on hyperploid neurons has also suggested selection against hyperploidy during AD progression [139]. This raises the possibility that dynamic elimination may have obscured a possible signal of AD-control difference in neuronal CNV loads, because our sample size did not allow studying disease stage as a separate factor. To study the AD stage as a separate factor, the following steps would have been taken. First, the same number of cells from different stages could be sequenced. After the downstream analysis, the CNVs, which pass filters (we and others applied stringent filters to eliminate false positive calls), for each AD stage would be found. This information could be used to determine the number of sequenced cells in

the next round of sequencing. More cells from the less-represented AD stage can be included.

CHAPTER 5

CONCLUSION

Our main motivation in this study was to describe the relative prevalence of CNVs in the AD brain, where the evidence has been equivocal. Contrary to earlier cytogenetic work [137, 117, 138, 139, 113, 140, 46, 141], an scWGS study had reported no difference in neuronal aneuploidy levels in the frontal cortex of AD patients versus controls [126]. However, the CNV load in different brain regions and relative frequency to the healthy age-matched controls had remained unclear. For example, the entorhinal cortex and hippocampal CA1 have roles in memory formation and learning and are the earliest and most heavily affected regions in AD [1]. On the other hand, hippocampal CA3 is less affected, and neurons in the cerebellum are thought to be relatively spared from neurodegenerative disease [20].

Here we tackled the same question by comparing AD patients and controls using LCM-isolated cells across five different brain regions, either using the raw data ($n = 588$ cells after filtering) or using a denoising approach ($n = 1301$ cells after filtering). To our knowledge, this is the first dataset that includes scWGS data from pyramidal neurons isolated from AD and control brains in multiple brain regions. Although our AD sample contained slightly higher CNV frequencies than the control sample, none of the comparisons was statistically significant. Our analysis of the van den Bos 2016 dataset yielded a qualitatively similar result, also consistent with the original observation of no significant difference in aneuploidy levels in this dataset [126].

Cost-effective and high-throughput Illumina sequencing has dominated the research area for the last decade. Next-generation sequencing technologies produce a short read length ranging from 50 to 1000 bases [201, 202]. The short read length hampers

the mapping and analysis of complex genomic regions including repetitive regions [203, 204, 205]. In addition to that GC extreme regions could not efficiently be amplified during the PCR amplification step and are manifested as uneven coverage of Illumina reads. This in turn is reflected in little knowledge of the genomic regions that have repeat or atypical GC content [171, 172, 173, 3, 174, 175, 176, 177] (see section 2.2.2). To overcome these drawbacks one possible solution could be taking advantage of third-generation sequencing or long-read sequencing. Long-read sequencing was developed by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) [206, 207, 208, 209, 210]. These technologies produce reads ranging between 10 kilobases to several megabases that encompass the repetitive regions of the genome [211]. Apart from spanning repetitive genomic regions, PCR amplification is not required [212]. So this leads to evenly distributed coverage across the genome, even in GC extreme regions. Overall, our results call for further research into the possible role of CNVs in AD pathogenesis using different methods like long-read sequencing.

REFERENCES

[1] X. Wang, M. L. Michaelis, and E. K. Michaelis, "Functional genomics of brain aging and alzheimer's disease: focus on selective neuronal vulnerability," *Curr. Genomics*, vol. 11, pp. 618–633, Dec. 2010.

[2] M. Citron, "Strategies for disease modification in alzheimer's disease," *Nat. Rev. Neurosci.*, vol. 5, pp. 677–685, Sept. 2004.

[3] T. Garvin, R. Aboukhalil, J. Kendall, T. Baslan, G. S. Atwal, J. Hicks, M. Wigler, and M. C. Schatz, "Interactive analysis and assessment of single-cell copy-number variations," *Nat. Methods*, vol. 12, pp. 1058–1060, Nov. 2015.

[4] I. Ostrovnaya, A. B. Olshen, V. E. Seshan, I. Orlow, D. G. Albertson, and C. B. Begg, "A metastasis or a second independent cancer? evaluating the clonal origin of tumors using array copy number data," *Stat. Med.*, vol. 29, pp. 1608–1621, July 2010.

[5] N. A. Bishop, T. Lu, and B. A. Yankner, "Neural mechanisms of ageing and cognitive decline," *Nature*, vol. 464, pp. 529–535, Mar. 2010.

[6] M. Goedert and M. G. Spillantini, "A century of alzheimer's disease," *Science*, vol. 314, pp. 777–781, Nov. 2006.

[7] M. P. Mattson, "Pathways towards and away from alzheimer's disease," *Nature*, vol. 430, pp. 631–639, Aug. 2004.

[8] M. Citron, "Alzheimer's disease: strategies for disease modification," *Nat. Rev. Drug Discov.*, vol. 9, pp. 387–398, May 2010.

[9] G. Thinakaran, "The role of presenilins in alzheimer's disease," *J. Clin. Invest.*, vol. 104, pp. 1321–1327, Nov. 1999.

[10] R. E. Tanzi and L. Bertram, "New frontiers in alzheimer's disease genetics," *Neuron*, vol. 32, pp. 181–184, Oct. 2001.

[11] J. Bali, A. H. Gheinani, S. Zurbriggen, and L. Rajendran, "Role of genes linked to sporadic alzheimer's disease risk in the production of β-amyloid peptides," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, pp. 15307–15311, Sept. 2012.

[12] I. J. Räihä, T. Rajala, L. B. Sourander, J. A. Kaprio, and M. J. Koskenvuo, "Alzheimer's disease in finnish twins," *Lancet*, vol. 347, pp. 573–578, 1996.

[13] M. Gatz, N. L. Pedersen, S. Berg, B. Johansson, K. Johansson, J. A. Mortimer, S. F. Posner, M. Viitanen, B. Winblad, and A. Ahlbom, "Heritability for alzheimer's disease: the study of dementia in swedish twins," *J. Gerontol. A Biol. Sci. Med. Sci.*, vol. 52, pp. M117–25, Mar. 1997.

[14] N. L. Pedersen, S. F. Posner, and M. Gatz, "Multiple-threshold models for genetic influences on age of onset for alzheimer disease: Findings in swedish twins," *Am. J. Med. Genet.*, vol. 105, pp. 724–728, Dec. 2001.

[15] R. Sims, M. Hill, and J. Williams, "The multiplex model of the genetics of alzheimer's disease," *Nat. Neurosci.*, vol. 23, pp. 311–322, Mar. 2020.

[16] M. D. Gallagher and A. S. Chen-Plotkin, "The Post-GWAS era: From association to function," *Am. J. Hum. Genet.*, vol. 102, pp. 717–730, May 2018.

[17] H. Braak and E. Braak, "Neuropathological stageing of alzheimer-related changes," *Acta Neuropathol.*, vol. 82, pp. 239–259, Sept. 1991.

[18] H. Braak and E. Braak, "Frequency of stages of alzheimer-related lesions in different age categories," *Neurobiol. Aging*, vol. 18, pp. 351–357, July 1997.

[19] J. Therriault, T. A. Pascoal, F. Z. Lussier, C. Tissot, M. Chamoun, G. Bezgin, S. Servaes, A. L. Benedet, N. J. Ashton, T. K. Karikari, J. Lantero-Rodriguez, P. Kunach, Y.-T. Wang, J. Fernandez-Arias, G. Massarweh, P. Vitali, J.-P. Soucy, P. Saha-Chaudhuri, K. Blennow, H. Zetterberg, S. Gauthier, and P. Rosa-Neto, "Biomarker modeling of alzheimer's disease using PET-based braak staging," *Nature Aging*, vol. 2, pp. 526–535, June 2022.

[20] J. Xu, S. Patassini, N. Rustogi, I. Riba-Garcia, B. D. Hale, A. M. Phillips, H. Waldvogel, R. Haines, P. Bradbury, A. Stevens, R. L. M. Faull, A. W. Dowsey, G. J. S. Cooper, and R. D. Unwin, "Regional protein expression in

human alzheimer's brain correlates with disease severity," *Communications Biology*, vol. 2, p. 43, Feb. 2019.

[21] C. L. Masters, R. Bateman, K. Blennow, C. C. Rowe, R. A. Sperling, and J. L. Cummings, "Alzheimer's disease," *Nature Reviews Disease Primers*, vol. 1, p. 15056, Oct. 2015.

[22] Z. S. Khachaturian, "Diagnosis of alzheimer's disease," *Arch. Neurol.*, vol. 42, pp. 1097–1105, Nov. 1985.

[23] S. S. Mirra, A. Heyman, D. McKeel, S. M. Sumi, B. J. Crain, L. M. Brownlee, F. S. Vogel, J. P. Hughes, G. van Belle, and L. Berg, "The consortium to establish a registry for alzheimer's disease (CERAD). part II. standardization of the neuropathologic assessment of alzheimer's disease," *Neurology*, vol. 41, pp. 479–486, Apr. 1991.

[24] K. Blennow and H. Hampel, "CSF markers for incipient alzheimer's disease," *Lancet Neurol.*, vol. 2, pp. 605–613, Oct. 2003.

[25] P. T. Lansbury, "Back to the future: the 'old-fashioned' way to new medications for neurodegeneration," *Nat. Med.*, vol. 10, pp. S51–S57, July 2004.

[26] H. Hampel, K. Bürger, S. J. Teipel, A. L. W. Bokde, H. Zetterberg, and K. Blennow, "Core candidate neurochemical and imaging biomarkers of alzheimer's disease," *Alzheimers. Dement.*, vol. 4, pp. 38–48, Dec. 2007.

[27] H. Hampel, K. Buerger, R. Zinkowski, S. J. Teipel, A. Goernitz, N. Andreasen, M. Sjoegren, J. DeBernardis, D. Kerkman, K. Ishiguro, H. Ohno, E. Vanmechelen, H. Vanderstichele, C. McCulloch, H.-J. Möller, P. Davies, and K. Blennow, "Measurement of Phosphorylated Tau Epitopes in the Differential Diagnosisof Alzheimer Disease: A Comparative Cerebrospinal Fluid Study," *Arch. Gen. Psychiatry*, vol. 61, no. 1, pp. 95–102, 2004.

[28] C. R. Jack, Jr, R. C. Petersen, Y. C. Xu, P. C. O'Brien, G. E. Smith, R. J. Ivnik, B. F. Boeve, S. C. Waring, E. G. Tangalos, and E. Kokmen, "Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment," *Neurology*, vol. 52, pp. 1397–1403, Apr. 1999.

[29] A. T. Du, N. Schuff, D. Amend, M. P. Laakso, Y. Y. Hsu, W. J. Jagust, K. Yaffe, J. H. Kramer, B. Reed, D. Norman, H. C. Chui, and M. W. Weiner, "Magnetic resonance imaging of the entorhinal cortex and hippocampus in mild cognitive impairment and alzheimer's disease," *J. Neurol. Neurosurg. Psychiatry*, vol. 71, pp. 441–447, Oct. 2001.

[30] C. Pennanen, M. Kivipelto, S. Tuomainen, P. Hartikainen, T. Hänninen, M. P. Laakso, M. Hallikainen, M. Vanhanen, A. Nissinen, E.-L. Helkala, P. Vainio, R. Vanninen, K. Partanen, and H. Soininen, "Hippocampus and entorhinal cortex in mild cognitive impairment and early AD," *Neurobiol. Aging*, vol. 25, pp. 303–310, Mar. 2004.

[31] P. N. Wang, J. F. Lirng, K. N. Lin, F. C. Chang, and H. C. Liu, "Prediction of alzheimer's disease in mild cognitive impairment: a prospective study in taiwan," *Neurobiol. Aging*, vol. 27, pp. 1797–1806, Dec. 2006.

[32] J. C. Baron, G. Chételat, B. Desgranges, G. Perchey, B. Landeau, V. de la Sayette, and F. Eustache, "In vivo mapping of gray matter loss with voxel-based morphometry in mild alzheimer's disease," *Neuroimage*, vol. 14, pp. 298–309, Aug. 2001.

[33] G. F. Busatto, G. E. J. Garrido, O. P. Almeida, C. C. Castro, C. H. P. Camargo, C. G. Cid, C. A. Buchpiguel, S. Furuie, and C. M. Bottino, "A voxel-based morphometry study of temporal lobe gray matter reductions in alzheimer's disease," *Neurobiol. Aging*, vol. 24, pp. 221–231, Mar. 2003.

[34] J. P. Lerch, J. C. Pruessner, A. Zijdenbos, H. Hampel, S. J. Teipel, and A. C. Evans, "Focal decline of cortical thickness in alzheimer's disease identified by computational neuroanatomy," *Cereb. Cortex*, vol. 15, pp. 995–1001, Nov. 2004.

[35] R. Leech and D. J. Sharp, "The role of the posterior cingulate cortex in cognition and disease," *Brain*, vol. 137, no. 1, pp. 12–32, 2013.

[36] T. Kuwert, P. Bartenstein, F. Grünwald, K. Herholz, R. Larisch, O. Sabri, H. J. Biersack, E. Moser, H. W. Müller-Gärtner, O. Schober, M. Schwaiger, U. Büll,

and W. D. Heiss, "[clinical value of positron emission tomography in neuromedicine. position paper on results of an interdisciplinary consensus conference]," *Nervenarzt*, vol. 69, pp. 1045–1060, Dec. 1998.

[37] G. W. Small, "Positron emission tomography imaging of dementia," *Dialogues Clin. Neurosci.*, vol. 2, pp. 168–169, June 2000.

[38] D. J. Selkoe and J. Hardy, "The amyloid hypothesis of alzheimer's disease at 25 years," *EMBO Mol. Med.*, vol. 8, pp. 595–608, June 2016.

[39] R. MacLeod, E.-K. Hillert, R. T. Cameron, and G. S. Baillie, "The role and therapeutic targeting of α-, β- and γ-secretase in alzheimer's disease," *Future Sci OA*, vol. 1, p. FSO11, Nov. 2015.

[40] B. De Strooper, "Proteases and proteolysis in alzheimer disease: A multifactorial view on the disease process," *Physiol. Rev.*, vol. 90, no. 2, pp. 465–494, 2010.

[41] L. Rajendran and W. Annaert, "Membrane trafficking pathways in alzheimer's disease," *Traffic*, vol. 13, no. 6, pp. 759–770, 2012.

[42] M. S. Uddin, M. T. Kabir, P. Jeandet, B. Mathew, G. Ashraf, A. Perveen, M. N. Bin-Jumah, S. A. Mousa, and M. M. Abdel-Daim, "Novel Anti-Alzheimer's therapeutic molecules targeting amyloid precursor protein processing," *Oxid. Med. Cell. Longev.*, vol. 2020, p. 7039138, Apr. 2020.

[43] C. A. Lane, J. Hardy, and J. M. Schott, "Alzheimer's disease," *Eur. J. Neurol.*, vol. 25, no. 1, pp. 59–70, 2018.

[44] S. Makin, "The amyloid hypothesis on trial," *Nature*, vol. 559, pp. S4–S7, July 2018.

[45] R. Sherrington, E. I. Rogaev, Y. Liang, E. A. Rogaeva, G. Levesque, M. Ikeda, H. Chi, C. Lin, G. Li, K. Holman, T. Tsuda, L. Mar, J.-F. Foncin, A. C. Bruni, M. P. Montesi, S. Sorbi, I. Rainero, L. Pinessi, L. Nee, I. Chumakov, D. Pollen, A. Brookes, P. Sanseau, R. J. Polinsky, W. Wasco, H. A. R. Da Silva, J. L. Haines, M. A. Pericak-Vance, R. E. Tanzi, A. D. Roses, P. E. Fraser, J. M. Rommens, and P. H. St George-Hyslop, "Cloning of a gene bearing missense mu-

tations in early-onset familial alzheimer's disease," *Nature*, vol. 375, pp. 754–760, June 1995.

[46] A. Goate, M.-C. Chartier-Harlin, M. Mullan, J. Brown, F. Crawford, L. Fidani, L. Giuffra, A. Haynes, N. Irving, L. James, R. Mant, P. Newton, K. Rooke, P. Roques, C. Talbot, M. Pericak-Vance, A. Roses, R. Williamson, M. Rossor, M. Owen, and J. Hardy, "Segregation of a missense mutation in the amyloid precursor protein gene with familial alzheimer's disease," *Nature*, vol. 349, pp. 704–706, Feb. 1991.

[47] D. A. Carter, E. Desmarais, M. Bellis, D. Campion, F. Clerget-Darpoux, A. Brice, Y. Agid, A. Jaillard-Serradt, and J. Mallet, "More missense in amyloid gene," *Nat. Genet.*, vol. 2, pp. 255–256, Dec. 1992.

[48] M.-C. Chartier-Harlin, F. Crawford, H. Houlden, A. Warren, D. Hughes, L. Fidani, A. Goate, M. Rossor, P. Roques, J. Hardy, and M. Mullan, "Early-onset alzheimer's disease caused by mutations at codon 717 of the β-amyloid precursor protein gene," *Nature*, vol. 353, pp. 844–846, Oct. 1991.

[49] D. Scheuner, C. Eckman, M. Jensen, X. Song, M. Citron, N. Suzuki, T. D. Bird, J. Hardy, M. Hutton, W. Kukull, E. Larson, E. Levy-Lahad, M. Viitanen, E. Peskind, P. Poorkaj, G. Schellenberg, R. Tanzi, W. Wasco, L. Lannfelt, D. Selkoe, and S. Younkin, "Secreted amyloid beta-protein similar to that in the senile plaques of alzheimer's disease is increased in vivo by the presenilin 1 and 2 and APP mutations linked to familial alzheimer's disease," *Nat. Med.*, vol. 2, pp. 864–870, Aug. 1996.

[50] S. Hecimovic, J. Wang, G. Dolios, M. Martinez, R. Wang, and A. M. Goate, "Mutations in APP have independent effects on Aβ and CTFγ generation," *Neurobiol. Dis.*, vol. 17, pp. 205–218, Nov. 2004.

[51] S. Kumar-Singh, J. Theuns, B. Van Broeck, D. Pirici, K. Vennekens, E. Corsmit, M. Cruts, B. Dermaut, R. Wang, and C. Van Broeckhoven, "Mean age-of-onset of familial alzheimer disease caused by presenilin mutations correlates with both increased abeta42 and decreased abeta40," *Hum. Mutat.*, vol. 27, pp. 686–695, July 2006.

[52] J. Hardy and D. Allsop, "Amyloid deposition as the central event in the aetiology of alzheimer's disease," *Trends Pharmacol. Sci.*, vol. 12, pp. 383–388, Jan. 1991.

[53] J. Hardy and D. J. Selkoe, "The amyloid hypothesis of alzheimer's disease: progress and problems on the road to therapeutics," *Science*, vol. 297, pp. 353–356, July 2002.

[54] F. M. LaFerla, B. T. Tinkle, C. J. Bieberich, C. C. Haudenschild, and G. Jay, "The alzheimer's a beta peptide induces neurodegeneration and apoptotic cell death in transgenic mice," *Nat. Genet.*, vol. 9, pp. 21–30, Jan. 1995.

[55] E. S. Musiek and D. M. Holtzman, "Three dimensions of the amyloid hypothesis: time, space and 'wingmen'," *Nat. Neurosci.*, vol. 18, pp. 800–806, June 2015.

[56] E. Tarkowski, N. Andreasen, A. Tarkowski, and K. Blennow, "Intrathecal inflammation precedes development of alzheimer's disease," *J. Neurol. Neurosurg. Psychiatry*, vol. 74, pp. 1200–1205, Sept. 2003.

[57] D. H. Cribbs, N. C. Berchtold, V. Perreau, P. D. Coleman, J. Rogers, A. J. Tenner, and C. W. Cotman, "Extensive innate immune gene activation accompanies brain aging, increasing vulnerability to cognitive decline and neurodegeneration: a microarray study," *J. Neuroinflammation*, vol. 9, p. 179, July 2012.

[58] B. Zhang, C. Gaiteri, L.-G. Bodea, Z. Wang, J. McElwee, A. A. Podtelezhnikov, C. Zhang, T. Xie, L. Tran, R. Dobrin, E. Fluder, B. Clurman, S. Melquist, M. Narayanan, C. Suver, H. Shah, M. Mahajan, T. Gillis, J. Mysore, M. E. MacDonald, J. R. Lamb, D. A. Bennett, C. Molony, D. J. Stone, V. Gudnason, A. J. Myers, E. E. Schadt, H. Neumann, J. Zhu, and V. Emilsson, "Integrated systems approach identifies genetic nodes and networks in late-onset alzheimer's disease," *Cell*, vol. 153, pp. 707–720, Apr. 2013.

[59] F. Brosseron, M. Krauthausen, M. Kummer, and M. T. Heneka, "Body fluid

cytokine levels in mild cognitive impairment and alzheimer's disease: a comparative overview," *Mol. Neurobiol.*, vol. 50, pp. 534–544, Feb. 2014.

[60] T. Shabab, R. Khanabdali, S. Z. Moghadamtousi, H. A. Kadir, and G. Mohan, "Neuroinflammation pathways: a general review," *Int. J. Neurosci.*, vol. 127, pp. 624–633, Aug. 2016.

[61] F. Ginhoux, S. Lim, G. Hoeffel, D. Low, and T. Huber, "Origin and differentiation of microglia," *Front. Cell. Neurosci.*, vol. 7, 2013.

[62] G. W. Kreutzberg, "Microglia: a sensor for pathological events in the CNS," *Trends Neurosci.*, vol. 19, pp. 312–318, Aug. 1996.

[63] D. P. Schafer, E. K. Lehrman, A. G. Kautzman, R. Koyama, A. R. Mardinly, R. Yamasaki, R. M. Ransohoff, M. E. Greenberg, B. A. Barres, and B. Stevens, "Microglia sculpt postnatal neural circuits in an activity and complement-dependent manner," *Neuron*, vol. 74, pp. 691–705, May 2012.

[64] J. L. Frost and D. P. Schafer, "Microglia: Architects of the developing nervous system," *Trends Cell Biol.*, vol. 26, pp. 587–597, Aug. 2016.

[65] C. Y. D. Lee and G. E. Landreth, "The role of microglia in amyloid clearance from the AD brain," *J. Neural Transm.*, vol. 117, pp. 949–960, Aug. 2010.

[66] K. I. Mosher and T. Wyss-Coray, "Microglial dysfunction in brain aging and alzheimer's disease," *Biochem. Pharmacol.*, vol. 88, no. 4, pp. 594–604, 2014.

[67] M. T. Heneka, M. J. Carson, J. E. Khoury, G. E. Landreth, F. Brosseron, D. L. Feinstein, A. H. Jacobs, T. Wyss-Coray, J. Vitorica, R. M. Ransohoff, K. Herrup, S. A. Frautschy, B. Finsen, G. C. Brown, A. Verkhratsky, K. Yamanaka, J. Koistinaho, E. Latz, A. Halle, G. C. Petzold, T. Town, D. Morgan, M. L. Shinohara, V. H. Perry, C. Holmes, N. G. Bazan, D. J. Brooks, S. Hunot, B. Joseph, N. Deigendesch, O. Garaschuk, E. Boddeke, C. A. Dinarello, J. C. Breitner, G. M. Cole, D. T. Golenbock, and M. P. Kummer, "Neuroinflammation in alzheimer's disease," *Lancet Neurol.*, vol. 14, no. 4, pp. 388–405, 2015.

[68] G. Krabbe, A. Halle, V. Matyash, J. L. Rinnenthal, G. D. Eom, U. Bernhardt, K. R. Miller, S. Prokop, H. Kettenmann, and F. L. Heppner, "Functional im-

pairment of microglia coincides with beta-amyloid deposition in mice with alzheimer-like pathology," *PLoS One*, vol. 8, p. e60921, Apr. 2013.

[69] S. E. Hickman, E. K. Allison, and J. El Khoury, "Microglial dysfunction and defective beta-amyloid clearance pathways in aging alzheimer's disease mice," *J. Neurosci.*, vol. 28, pp. 8354–8360, Aug. 2008.

[70] F. A. Edwards, "A unifying hypothesis for alzheimer's disease: From plaques to neurodegeneration," *Trends Neurosci.*, vol. 42, pp. 310–322, May 2019.

[71] W. J. Streit, Q.-S. Xue, J. Tischer, and I. Bechmann, "Microglial pathology," *Acta Neuropathologica Communications*, vol. 2, p. 142, Sept. 2014.

[72] A. G. Efthymiou and A. M. Goate, "Late onset alzheimer's disease genetics implicates microglial pathways in disease risk," *Mol. Neurodegener.*, vol. 12, p. 43, May 2017.

[73] L. Bertram and R. E. Tanzi, "Alzheimer disease risk genes: 29 and counting," *Nat. Rev. Neurol.*, vol. 15, pp. 191–192, Apr. 2019.

[74] R. Guerreiro, A. Wojtas, J. Bras, M. Carrasquillo, E. Rogaeva, E. Majounie, C. Cruchaga, C. Sassi, J. S. K. Kauwe, S. Younkin, L. Hazrati, J. Collinge, J. Pocock, T. Lashley, J. Williams, J.-C. Lambert, P. Amouyel, A. Goate, R. Rademakers, K. Morgan, J. Powell, P. St George-Hyslop, A. Singleton, J. Hardy, and Alzheimer Genetic Analysis Group, "TREM2 variants in alzheimer's disease," *N. Engl. J. Med.*, vol. 368, pp. 117–127, Nov. 2012.

[75] T. Jonsson, J. K. Atwal, S. Steinberg, J. Snaedal, P. V. Jonsson, S. Bjornsson, H. Stefansson, P. Sulem, D. Gudbjartsson, J. Maloney, K. Hoyte, A. Gustafson, Y. Liu, Y. Lu, T. Bhangale, R. R. Graham, J. Huttenlocher, G. Bjornsdottir, O. A. Andreassen, E. G. Jönsson, A. Palotie, T. W. Behrens, O. T. Magnusson, A. Kong, U. Thorsteinsdottir, R. J. Watts, and K. Stefansson, "A mutation in APP protects against alzheimer's disease and age-related cognitive decline," *Nature*, vol. 488, pp. 96–99, Aug. 2012.

[76] M.-C. Chartier-Hariln, M. Parfitt, S. Legrain, J. Pérez-Tur, T. Brousseau, A. Evans, C. Berr, O. Vldal, P. Roques, V. Gourlet, J.-C. Fruchart, A. Delacourte, M. Rossor, and P. Amouyel, "Apolipoprotein E, 4 allele as a major risk

factor for sporadic early and late-onset forms of Alzheimer's disease: analysis of the 19q13.2 chromosomal region," *Hum. Mol. Genet.*, vol. 3, no. 4, pp. 569–574, 1994.

[77] H. Houlden, R. Crook, H. Backhovens, G. Prihar, M. Baker, M. Hutton, M. Rossor, J. J. Martin, C. Van Broeckhoven, and J. Hardy, "ApoE genotype is a risk factor in nonpresenilin early-onset alzheimer's disease families," *Am. J. Med. Genet.*, vol. 81, no. 1, pp. 117–121, 1998.

[78] E. H. Corder, A. M. Saunders, W. J. Strittmatter, D. E. Schmechel, P. C. Gaskell, G. W. Small, A. D. Roses, J. L. Haines, and M. A. Pericak-Vance, "Gene dose of apolipoprotein E type 4 allele and the risk of alzheimer's disease in late onset families," *Science*, vol. 261, pp. 921–923, Aug. 1993.

[79] A. M. Saunders, W. J. Strittmatter, D. Schmechel, P. H. George-Hyslop, M. A. Pericak-Vance, S. H. Joo, B. L. Rosi, J. F. Gusella, D. R. Crapper-MacLachlan, and M. J. Alberts, "Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic alzheimer's disease," *Neurology*, vol. 43, pp. 1467–1472, Aug. 1993.

[80] W. J. Strittmatter, A. M. Saunders, D. Schmechel, M. Pericak-Vance, J. Enghild, G. S. Salvesen, and A. D. Roses, "Apolipoprotein e: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial alzheimer disease," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 90, pp. 1977–1981, Mar. 1993.

[81] E. M. Reiman, R. J. Caselli, L. S. Yun, K. Chen, D. Bandy, S. Minoshima, S. N. Thibodeau, and D. Osborne, "Preclinical evidence of alzheimer's disease in persons homozygous for the epsilon 4 allele for apolipoprotein E," *N. Engl. J. Med.*, vol. 334, pp. 752–758, Mar. 1996.

[82] R. E. Pitas, J. K. Boyles, S. H. Lee, D. Foss, and R. W. Mahley, "Astrocytes synthesize apolipoprotein E and metabolize apolipoprotein e-containing lipoproteins," *Biochim. Biophys. Acta*, vol. 917, pp. 148–161, Jan. 1987.

[83] S. Grehan, E. Tse, and J. M. Taylor, "Two distal downstream enhancers direct

expression of the human apolipoprotein E gene to astrocytes in the brain," *J. Neurosci.*, vol. 21, pp. 812–822, Feb. 2001.

[84] R. Deane, A. Sagare, K. Hamm, M. Parisi, S. Lane, M. B. Finn, D. M. Holtzman, and B. V. Zlokovic, "apoe isoform-specific disruption of amyloid beta peptide clearance from mouse brain," *J. Clin. Invest.*, vol. 118, pp. 4002–4013, Nov. 2008.

[85] T.-P. V. Huynh, A. A. Davis, J. D. Ulrich, and D. M. Holtzman, "Apolipoprotein E and alzheimer's disease: the influence of apolipoprotein E on amyloid-β and other amyloidogenic proteins: Thematic review series: ApoE and lipid homeostasis in alzheimer's disease," *J. Lipid Res.*, vol. 58, pp. 824–836, May 2017.

[86] D. M. Holtzman, K. R. Bales, T. Tenkova, A. M. Fagan, M. Parsadanian, L. J. Sartorius, B. Mackey, J. Olney, D. McKeel, D. Wozniak, and S. M. Paul, "Apolipoprotein E isoform-dependent amyloid deposition and neuritic degeneration in a mouse model of alzheimer's disease," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 97, pp. 2892–2897, Mar. 2000.

[87] E. Hudry, J. Dashkoff, A. D. Roe, S. Takeda, R. M. Koffie, T. Hashimoto, M. Scheel, T. Spires-Jones, M. Arbel-Ornath, R. Betensky, B. L. Davidson, and B. T. Hyman, "Gene transfer of human apoe isoforms results in differential modulation of amyloid deposition and neurotoxicity in mouse brain," *Sci. Transl. Med.*, vol. 5, p. 212ra161, Nov. 2013.

[88] L. A. Farrer, L. A. Cupples, J. L. Haines, B. Hyman, W. A. Kukull, R. Mayeux, R. H. Myers, M. A. Pericak-Vance, N. Risch, and C. M. van Duijn, "Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and alzheimer disease. a meta-analysis. APOE and alzheimer disease meta analysis consortium," *JAMA*, vol. 278, no. 16, pp. 1349–1356, 1997.

[89] D. E. Schmechel, A. M. Saunders, W. J. Strittmatter, B. J. Crain, C. M. Hulette, S. H. Joo, M. A. Pericak-Vance, D. Goldgaber, and A. D. Roses, "Increased amyloid beta-peptide deposition in cerebral cortex as a consequence of apolipoprotein E genotype in late-onset alzheimer disease," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 90, pp. 9649–9653, Oct. 1993.

[90] G. William Rebeck, J. S. Reiter, D. K. Strickland, and B. T. Hyman, "Apolipoprotein E in sporadic alzheimer's disease: Allelic variation and receptor interactions," *Neuron*, vol. 11, pp. 575–580, Oct. 1993.

[91] P. Barbier, O. Zejneli, M. Martinho, A. Lasorsa, V. Belle, C. Smet-Nocca, P. O. Tsvetkov, F. Devred, and I. Landrieu, "Role of tau as a Microtubule-Associated protein: Structural and functional aspects," *Front. Aging Neurosci.*, vol. 11, p. 204, Aug. 2019.

[92] M. D. Weingarten, A. H. Lockwood, S. Y. Hwo, and M. W. Kirschner, "A protein factor essential for microtubule assembly," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 72, pp. 1858–1862, May 1975.

[93] R. L. Neve, P. Harris, K. S. Kosik, D. M. Kurnit, and T. A. Donlon, "Identification of cDNA clones for the human microtubule-associated protein tau and chromosomal localization of the genes for tau and microtubule-associated protein 2," *Molecular Brain Research*, vol. 1, pp. 271–280, Dec. 1986.

[94] G. B. Witman, D. W. Cleveland, M. D. Weingarten, and M. W. Kirschner, "Tubulin requires tau for growth onto microtubule initiating sites," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 73, pp. 4070–4074, Nov. 1976.

[95] A. Andreadis, W. M. Brown, and K. S. Kosik, "Structure and novel exons of the human tau gene," *Biochemistry*, vol. 31, pp. 10626–10633, Nov. 1992.

[96] Y. Xia, S. Prokop, K.-M. M. Gorion, J. D. Kim, Z. A. Sorrentino, B. M. Bell, A. N. Manaois, P. Chakrabarty, P. Davies, and B. I. Giasson, "Tau ser208 phosphorylation promotes aggregation and reveals neuropathologic diversity in alzheimer's disease and other tauopathies," *Acta Neuropathologica Communications*, vol. 8, p. 88, June 2020.

[97] K. C. Wilhelmsen, T. Lynch, E. Pavlou, M. Higgins, and T. G. Nygaard, "Localization of disinhibition-dementia-parkinsonism-amyotrophy complex to 17q21-22," *Am. J. Hum. Genet.*, vol. 55, pp. 1159–1165, Dec. 1994.

[98] I. D'Souza, P. Poorkaj, M. Hong, D. Nochlin, V. M. Lee, T. D. Bird, and G. D. Schellenberg, "Missense and silent tau gene mutations cause frontotemporal

dementia with parkinsonism-chromosome 17 type, by affecting multiple alternative RNA splicing regulatory elements," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 96, pp. 5598–5603, May 1999.

[99] K. T. Wong, I. V. Allen, S. McQuaid, and R. McConnell, "An immunohistochemical study of neurofibrillary tangle formation in post-encephalitic parkinsonism," *Clin. Neuropathol.*, vol. 15, pp. 22–25, Jan. 1996.

[100] R. Redon, S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, E. K. Cho, S. Dallaire, J. L. Freeman, J. R. González, M. Gratacòs, J. Huang, D. Kalaitzopoulos, D. Komura, J. R. MacDonald, C. R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M. J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, J. Zhang, L. Armengol, D. F. Conrad, X. Estivill, C. Tyler-Smith, N. P. Carter, H. Aburatani, C. Lee, K. W. Jones, S. W. Scherer, and M. E. Hurles, "Global variation in copy number in the human genome," *Nature*, vol. 444, pp. 444–454, Nov. 2006.

[101] Y.-C. Tang and A. Amon, "Gene copy-number alterations: a cost-benefit analysis," *Cell*, vol. 152, pp. 394–405, Jan. 2013.

[102] S. D. Pack, R. J. Weil, A. O. Vortmeyer, W. Zeng, J. Li, H. Okamoto, M. Furuta, E. Pak, I. A. Lubensky, E. H. Oldfield, and Z. Zhuang, "Individual adult human neurons display aneuploidy: detection by fluorescence in situ hybridization and single neuron PCR," *Cell Cycle*, vol. 4, pp. 1758–1760, Dec. 2005.

[103] S. K. Rehen, Y. C. Yung, M. P. McCreight, D. Kaushal, A. H. Yang, B. S. V. Almeida, M. A. Kingsbury, K. M. S. Cabral, M. J. McConnell, B. Anliker, M. Fontanoz, and J. Chun, "Constitutional aneuploidy in the normal human brain," *J. Neurosci.*, vol. 25, pp. 2176–2180, Mar. 2005.

[104] Y. B. Yurov, I. Y. Iourov, V. V. Monakhov, I. V. Soloviev, V. M. Vostrikov, and S. G. Vorsanova, "The variation of aneuploidy frequency in the developing and adult human brain revealed by an interphase FISH study," *J. Histochem. Cytochem.*, vol. 53, pp. 385–390, Mar. 2005.

[105] S. K. Rehen, M. J. McConnell, D. Kaushal, M. A. Kingsbury, A. H. Yang, and

J. Chun, "Chromosomal variation in neurons of the developing and adult mammalian nervous system," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 98, pp. 13361–13366, Nov. 2001.

[106] A. H. Yang, D. Kaushal, S. K. Rehen, K. Kriedt, M. A. Kingsbury, M. J. McConnell, and J. Chun, "Chromosome segregation defects contribute to aneuploidy in normal neural progenitor cells," *J. Neurosci.*, vol. 23, pp. 10454–10462, Nov. 2003.

[107] M. J. McConnell, M. R. Lindberg, K. J. Brennand, J. C. Piper, T. Voet, C. Cowing-Zitron, S. Shumilina, R. S. Lasken, J. R. Vermeesch, I. M. Hall, and F. H. Gage, "Mosaic copy number variation in human neurons," *Science*, vol. 342, pp. 632–637, Nov. 2013.

[108] A. Poduri, G. D. Evrony, X. Cai, and C. A. Walsh, "Somatic mutation, genomic variation, and neurological disease," *Science*, vol. 341, p. 1237758, July 2013.

[109] P.-C. Wei, A. N. Chang, J. Kao, Z. Du, R. M. Meyers, F. W. Alt, and B. Schwer, "Long neural genes harbor recurrent DNA break clusters in neural Stem/Progenitor cells," *Cell*, vol. 164, pp. 644–655, Feb. 2016.

[110] S. Santaguida and A. Amon, "Short- and long-term effects of chromosome mis-segregation and aneuploidy," *Nat. Rev. Mol. Cell Biol.*, vol. 16, pp. 473–485, Aug. 2015.

[111] A. J. Holland and D. W. Cleveland, "Losing balance: the origin and impact of aneuploidy in cancer," *EMBO Rep.*, vol. 13, pp. 501–514, June 2012.

[112] F. Faggioli, T. Wang, J. Vijg, and C. Montagna, "Chromosome-specific accumulation of aneuploidy in the aging mouse brain," *Hum. Mol. Genet.*, vol. 21, pp. 5246–5253, Sept. 2012.

[113] Y. B. Yurov, S. G. Vorsanova, T. Liehr, A. D. Kolotii, and I. Y. Iourov, "X chromosome aneuploidy in the alzheimer's disease brain," *Mol. Cytogenet.*, vol. 7, p. 20, Mar. 2014.

[114] M. J. McConnell, D. Kaushal, A. H. Yang, M. A. Kingsbury, S. K. Rehen, K. Treuner, R. Helton, E. G. Annas, J. Chun, and C. Barlow, "Failed clearance

of aneuploid embryonic neural progenitor cells leads to excess aneuploidy in the atm-deficient but not the trp53-deficient adult cerebral cortex," *J. Neurosci.*, vol. 24, pp. 8090–8096, Sept. 2004.

[115] S. E. Peterson, A. H. Yang, D. M. Bushman, J. W. Westra, Y. C. Yung, S. Barral, T. Mutoh, S. K. Rehen, and J. Chun, "Aneuploid cells are differentially susceptible to caspase-mediated death during embryonic cerebral cortical development," *J. Neurosci.*, vol. 32, pp. 16213–16222, Nov. 2012.

[116] X. Cai, G. D. Evrony, H. S. Lehmann, P. C. Elhosary, B. K. Mehta, A. Poduri, and C. A. Walsh, "Single-Cell, genome-wide sequencing identifies clonal somatic Copy-Number variation in the human brain," *Cell Rep.*, vol. 8, pp. 1280–1289, Sept. 2014.

[117] B. Mosch, M. Morawski, A. Mittag, D. Lenz, A. Tarnok, and T. Arendt, "Aneuploidy and DNA replication in the normal human brain and alzheimer's disease," *J. Neurosci.*, vol. 27, pp. 6859–6867, June 2007.

[118] J. W. Westra, S. E. Peterson, Y. C. Yung, T. Mutoh, S. Barral, and J. Chun, "Aneuploid mosaicism in the developing and adult cerebellar cortex," *J. Comp. Neurol.*, vol. 507, pp. 1944–1951, Apr. 2008.

[119] Z. G. Turan, V. Richter, J. Bochmann, P. Parvizi, E. Yapar, U. Işıldak, S.-K. Waterholter, S. Leclere-Turbant, Ç. D. Son, C. Duyckaerts, İ. Yet, T. Arendt, M. Somel, and U. Ueberham, "Somatic copy number variant load in neurons of healthy controls and alzheimer's disease patients," *Acta Neuropathologica Communications*, vol. 10, p. 175, Nov. 2022.

[120] M. A. Kingsbury, Y. C. Yung, S. E. Peterson, J. W. Westra, and J. Chun, "Aneuploidy in the normal and diseased brain," *Cell. Mol. Life Sci.*, vol. 63, pp. 2626–2641, Nov. 2006.

[121] D. M. Bushman and J. Chun, "The genomically mosaic brain: aneuploidy and more in neural diversity and disease," *Semin. Cell Dev. Biol.*, vol. 24, pp. 357–369, Mar. 2013.

[122] B. Bakker, H. van den Bos, P. M. Lansdorp, and F. Foijer, "How to count chro-

mosomes in a cell: An overview of current and novel technologies," *Bioessays*, vol. 37, pp. 570–577, Mar. 2015.

[123] M. Liyanage, A. Coleman, S. du Manoir, T. Veldman, S. McCormack, R. B. Dickson, C. Barlow, A. Wynshaw-Boris, S. Janz, J. Wienberg, M. A. Ferguson-Smith, E. Schröck, and T. Ried, "Multicolour spectral karyotyping of mouse chromosomes," *Nat. Genet.*, vol. 14, pp. 312–315, Nov. 1996.

[124] M. Macville, T. Veldman, H. Padilla-Nash, D. Wangsa, P. O'Brien, E. Schröck, and T. Ried, "Spectral karyotyping, a 24-colour FISH technique for the identification of chromosomal rearrangements," *Histochem. Cell Biol.*, vol. 108, pp. 299–305, Oct. 1997.

[125] G. Imataka and O. Arisaka, "Chromosome analysis using spectral karyotyping (SKY)," *Cell Biochem. Biophys.*, vol. 62, pp. 13–17, Jan. 2012.

[126] H. van den Bos, D. C. J. Spierings, A. Taudt, B. Bakker, D. Porubský, E. Falconer, C. Novoa, N. Halsema, H. G. Kazemier, K. Hoekstra-Wakker, V. Guryev, W. F. A. den Dunnen, F. Foijer, M. Colomé-Tatché, H. W. G. M. Boddeke, and P. M. Lansdorp, "Single-cell whole genome sequencing reveals no evidence for common aneuploidy in normal and alzheimer's disease neurons," *Genome Biology*, vol. 17, p. 116, May 2016.

[127] K. A. Knouse, J. Wu, C. A. Whittaker, and A. Amon, "Single cell sequencing reveals low levels of aneuploidy across mammalian tissues," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, pp. 13409–13414, Sept. 2014.

[128] G. A. Andriani, E. Maggi, D. Piqué, S. E. Zimmerman, M. Lee, W. Quispe-Tintaya, A. Maslov, J. Campisi, J. Vijg, J. C. Mar, and C. Montagna, "A direct comparison of interphase FISH versus low-coverage single cell sequencing to detect aneuploidy reveals respective strengths and weaknesses," *Sci. Rep.*, vol. 9, p. 10508, July 2019.

[129] J. M. Levsky and R. H. Singer, "Fluorescence in situ hybridization: past, present and future," *J. Cell Sci.*, vol. 116, pp. 2833–2838, July 2003.

[130] L. Hu, K. Ru, L. Zhang, Y. Huang, X. Zhu, H. Liu, A. Zetterberg, T. Cheng, and W. Miao, "Fluorescence in situ hybridization (FISH): an increasingly de-

manded tool for biomarker research and personalized medicine," *Biomarker Research*, vol. 2, p. 3, Feb. 2014.

[131] A. W. Duncan, A. E. Hanlon Newell, L. Smith, E. M. Wilson, S. B. Olson, M. J. Thayer, S. C. Strom, and M. Grompe, "Frequent aneuploidy among normal human hepatocytes," *Gastroenterology*, vol. 142, pp. 25–28, Jan. 2012.

[132] T. Baslan, J. Kendall, L. Rodgers, H. Cox, M. Riggs, A. Stepansky, J. Troge, K. Ravi, D. Esposito, B. Lakshmi, M. Wigler, N. Navin, and J. Hicks, "Genome-wide copy number analysis of single cells," *Nat. Protoc.*, vol. 7, pp. 1024–1041, June 2012.

[133] K. A. Knouse, J. Wu, and A. Amon, "Assessment of megabase-scale somatic copy number variation using single-cell sequencing," *Genome Res.*, vol. 26, pp. 376–384, Jan. 2016.

[134] S. A. Vitak, K. A. Torkenczy, J. L. Rosenkrantz, A. J. Fields, L. Christiansen, M. H. Wong, L. Carbone, F. J. Steemers, and A. Adey, "Sequencing thousands of single-cell genomes with combinatorial indexing," *Nat. Methods*, vol. 14, pp. 302–308, Jan. 2017.

[135] S. Rohrback, C. April, F. Kaper, R. R. Rivera, C. S. Liu, B. Siddoway, and J. Chun, "Submegabase copy number variations arise during cerebral cortical neurogenesis as revealed by single-cell whole-genome sequencing," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 115, pp. 10804–10809, Sept. 2018.

[136] W. D. Chronister, I. E. Burbulis, M. B. Wierman, M. J. Wolpert, M. F. Haakenson, A. C. B. Smith, J. E. Kleinman, T. M. Hyde, D. R. Weinberger, S. Bekiranov, and M. J. McConnell, "Neurons with complex karyotypes are rare in aged human neocortex," *Cell Rep.*, vol. 26, pp. 825–835.e7, Jan. 2019.

[137] Y. Yang, D. S. Geldmacher, and K. Herrup, "DNA replication precedes neuronal cell death in alzheimer's disease," *J. Neurosci.*, vol. 21, pp. 2661–2668, Apr. 2001.

[138] I. Y. Iourov, S. G. Vorsanova, T. Liehr, and Y. B. Yurov, "Aneuploidy in the normal, alzheimer's disease and ataxia-telangiectasia brain: differential ex-

pression and pathological meaning," *Neurobiol. Dis.*, vol. 34, pp. 212–220, May 2009.

[139] T. Arendt, M. K. Brückner, B. Mosch, and A. Lösche, "Selective cell death of hyperploid neurons in alzheimer's disease," *Am. J. Pathol.*, vol. 177, pp. 15–20, May 2010.

[140] P. Thomas and M. Fenech, "Chromosome 17 and 21 aneuploidy in buccal cells is increased with ageing and in alzheimer's disease," *Mutagenesis*, vol. 23, pp. 57–65, Nov. 2007.

[141] H. Potter, A. Granic, and J. Caneus, "Role of trisomy 21 mosaicism in sporadic and familial alzheimer's disease," *Curr. Alzheimer Res.*, vol. 13, no. 1, pp. 7–17, 2016.

[142] T. Arendt, J. Stieler, and U. Ueberham, "Is sporadic alzheimer's disease a developmental disorder?," *J. Neurochem.*, vol. 143, pp. 396–408, May 2017.

[143] H. Potter, "Review and hypothesis: Alzheimer disease and down syndrome–chromosome 21 nondisjunction may underlie both disorders," *Am. J. Hum. Genet.*, vol. 48, pp. 1192–1200, June 1991.

[144] F. Abascal, L. M. R. Harvey, E. Mitchell, A. R. J. Lawson, S. V. Lensing, P. Ellis, A. J. C. Russell, R. E. Alcantara, A. Baez-Ortega, Y. Wang, E. J. Kwa, H. Lee-Six, A. Cagan, T. H. H. Coorens, M. S. Chapman, S. Olafsson, S. Leonard, D. Jones, H. E. Machado, M. Davies, N. F. Øbro, K. T. Mahubani, K. Allinson, M. Gerstung, K. Saeb-Parsy, D. G. Kent, E. Laurenti, M. R. Stratton, R. Rahbari, P. J. Campbell, R. J. Osborne, and I. Martincorena, "Somatic mutation landscapes at single-molecule resolution," *Nature*, vol. 593, pp. 405–410, May 2021.

[145] K. Herrup and T. Arendt, "Re-expression of cell cycle proteins induces neuronal cell death during alzheimer's disease," *J. Alzheimers. Dis.*, vol. 4, pp. 243–247, June 2002.

[146] L. Clerx, I. A. van Rossum, L. Burns, D. L. Knol, P. Scheltens, F. Verhey, P. Aalten, P. Lapuerta, L. van de Pol, R. van Schijndel, R. de Jong, F. Barkhof, R. Wolz, D. Rueckert, M. Bocchetta, M. Tsolaki, F. Nobili, L.-O. Wahlund,

L. Minthon, L. Frölich, H. Hampel, H. Soininen, and P. J. Visser, "Measurements of medial temporal lobe atrophy for prediction of alzheimer's disease in subjects with mild cognitive impairment," *Neurobiol. Aging*, vol. 34, pp. 2003–2013, Aug. 2013.

[147] H. Fu, J. Hardy, and K. E. Duff, "Selective vulnerability in neurodegenerative diseases," *Nat. Neurosci.*, vol. 21, pp. 1350–1358, Oct. 2018.

[148] M. L. Hernandez, T. Chatlos, K. M. Gorse, and A. D. Lafrenaye, "Neuronal membrane disruption occurs late following diffuse brain trauma in rats and involves a subpopulation of NeuN negative cortical neurons," *Front. Neurol.*, vol. 10, pp. 1238–1238, Nov. 2019.

[149] C. A. Sallaberry, B. J. Voss, J. Majewski, J. Biernat, E. Mandelkow, E. Y. Chi, and C. M. Vander Zanden, "Tau and membranes: Interactions that promote folding and condensation," *Frontiers in Cell and Developmental Biology*, vol. 9, 2021.

[150] D. Mrdenovic, I. S. Pieta, R. Nowakowski, W. Kutner, J. Lipkowski, and P. Pieta, "Amyloid β interaction with model cell membranes – what are the toxicity-defining properties of amyloid β?," *Int. J. Biol. Macromol.*, vol. 200, pp. 520–531, Mar. 2022.

[151] I. Unal-Cevik, M. Kilinç, Y. Gürsoy-Ozdemir, G. Gurer, and T. Dalkara, "Loss of NeuN immunoreactivity after cerebral ischemia does not indicate neuronal cell loss: a cautionary note," *Brain Res.*, vol. 1015, pp. 169–174, July 2004.

[152] M. A. Davoli, J. Fourtounis, J. Tam, S. Xanthoudakis, D. Nicholson, G. S. Robertson, G. Y. K. Ng, and D. Xu, "Immunohistochemical and biochemical assessment of caspase-3 activation and DNA fragmentation following transient focal ischemia in the rat," *Neuroscience*, vol. 115, no. 1, pp. 125–136, 2002.

[153] C. G. Ardanaz, M. J. Ramírez, and M. Solas, "Brain metabolic alterations in alzheimer's disease," 2022.

[154] M. Gonzalez-Rodriguez, S. Villar-Conde, V. Astillero-Lopez, P. Villanueva-Anguita, I. Ubeda-Banon, A. Flores-Cuadrado, A. Martinez-Marcos, and

D. Saiz-Sanchez, "Neurodegeneration and astrogliosis in the human CA1 hippocampal subfield are related to hsp90ab1 and bag3 in alzheimer's disease," *Int. J. Mol. Sci.*, vol. 23, p. 165, Dec. 2021.

[155] A. Yousef, J. L. Robinson, D. J. Irwin, M. D. Byrne, L. K. Kwong, E. B. Lee, Y. Xu, S. X. Xie, L. Rennert, E. Suh, V. M. Van Deerlin, M. Grossman, V. M.-Y. Lee, and J. Q. Trojanowski, "Neuron loss and degeneration in the progression of TDP-43 in frontotemporal lobar degeneration," *Acta Neuropathologica Communications*, vol. 5, p. 68, Sept. 2017.

[156] M. C. Chongtham, H. Todorov, J. E. Wettschereck, S. Gerber, and J. Winter, "Isolation of nuclei and downstream processing of cell-type-specific nuclei from micro-dissected mouse brain regions – techniques and caveats," *bioRxiv*, p. 2020.11.18.374223, Jan. 2020.

[157] M. C. Chongtham, T. Butto, K. Mungikar, S. Gerber, and J. Winter, "INTACT vs. FANS for Cell-Type-Specific nuclei sorting: A comprehensive qualitative and quantitative comparison," *Int. J. Mol. Sci.*, vol. 22, May 2021.

[158] Y. Kamikawa, A. Saito, and K. Imaizumi, "Impact of nuclear envelope stress on physiological and pathological processes in central nervous system," *Neurochem. Res.*, Apr. 2022.

[159] "Consensus recommendations for the postmortem diagnosis of alzheimer's disease," *Neurobiol. Aging*, vol. 18, pp. S1–S2, July 1997.

[160] D. R. Thal, U. Rüb, M. Orantes, and H. Braak, "Phases of a beta-deposition in the human brain and its relevance for the development of AD," *Neurology*, vol. 58, pp. 1791–1800, June 2002.

[161] A. Matevossian and S. Akbarian, "Neuronal nuclei isolation from human postmortem brain tissue," *J. Vis. Exp.*, Oct. 2008.

[162] P. Ewels, M. Magnusson, S. Lundin, and M. Käller, "MultiQC: summarize analysis results for multiple tools and samples in a single report," *Bioinformatics*, vol. 32, no. 19, pp. 3047–3048, 2016.

[163] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data Bioinformatics Oxford Academic," *Bioinformatics*, vol. 30, pp. 2114–2120, Aug. 2014.

[164] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, pp. 1754–1760, July 2009.

[165] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup, "The sequence Alignment/Map format and SAMtools," *Bioinformatics*, vol. 25, pp. 2078–2079, June 2009.

[166] K. A. Knouse, J. Wu, and A. Hendricks, "Detection of copy number alterations using single cell sequencing," *J. Vis. Exp.*, Feb. 2017.

[167] A. R. Quinlan and I. M. Hall, "BEDTools: a flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, pp. 841–842, Jan. 2010.

[168] D. Lai and G. Ha, "HMMcopy: A package for bias-free copy number estimation and robust CNA detection in tumour samples from WGS HTS data,"

[169] X. F. Mallory, M. Edrisi, N. Navin, and L. Nakhleh, "Assessing the performance of methods for copy number aberration detection from single-cell DNA sequencing data," *PLoS Comput. Biol.*, vol. 16, pp. e1008012–e1008012, July 2020.

[170] N. Navin, J. Kendall, J. Troge, P. Andrews, L. Rodgers, J. McIndoo, K. Cook, A. Stepansky, D. Levy, D. Esposito, L. Muthuswamy, A. Krasnitz, W. R. McCombie, J. Hicks, and M. Wigler, "Tumour evolution inferred by single-cell sequencing," *Nature*, vol. 472, pp. 90–94, Apr. 2011.

[171] M. A. Quail, M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow, and Y. Gu, "A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina MiSeq sequencers," *BMC Genomics*, vol. 13, p. 341, July 2012.

[172] S. O. Oyola, T. D. Otto, Y. Gu, G. Maslen, M. Manske, S. Campino, D. J. Turner, B. MacInnis, D. P. Kwiatkowski, H. P. Swerdlow, and M. A. Quail,

"Optimizing illumina next-generation sequencing library preparation for extremely at-biased genomes," *BMC Genomics*, vol. 13, p. 1, Jan. 2012.

[173] T. H. Jakobsen, M. A. Hansen, P. Ø. Jensen, L. Hansen, L. Riber, A. Cockburn, M. Kolpen, C. Rønne Hansen, W. Ridderberg, S. Eickhardt, M. Hansen, P. Kerpedjiev, M. Alhede, K. Qvortrup, M. Burmølle, C. Moser, M. Kühl, O. Ciofu, M. Givskov, S. J. Sørensen, N. Høiby, and T. Bjarnsholt, "Complete genome sequence of the cystic fibrosis pathogen achromobacter xylosoxidans NH44784-1996 complies with important pathogenic phenotypes," *PLoS One*, vol. 8, p. e68484, July 2013.

[174] D. G. McDowell, N. A. Burns, and H. C. Parkes, "Localised sequence regions possessing high melting temperatures prevent the amplification of a DNA mimic in competitive PCR," *Nucleic Acids Res.*, vol. 26, pp. 3340–3347, July 1998.

[175] Q. Chou, "Minimizing deletion mutagenesis artifact during taq DNA polymerase PCR by e. coli SSB," *Nucleic Acids Res.*, vol. 20, p. 4371, Aug. 1992.

[176] K. Usdin and K. J. Woodford, "CGG repeats associated with DNA instability and chromosome fragility form structures that block DNA synthesis in vitro," *Nucleic Acids Res.*, vol. 23, pp. 4202–4209, Oct. 1995.

[177] X. Chen, S. V. Mariappan, P. Catasti, R. Ratliff, R. K. Moyzis, A. Laayoun, S. S. Smith, E. M. Bradbury, and G. Gupta, "Hairpins are formed by the single DNA strands of the fragile X triplet repeats: structure and biological implications," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 92, pp. 5199–5203, May 1995.

[178] M. A. van de Wiel, F. Picard, W. N. van Wieringen, and B. Ylstra, "Preprocessing and downstream analysis of microarray DNA copy number profiles," *Brief. Bioinform.*, vol. 12, pp. 10–21, Jan. 2011.

[179] A. B. Olshen, E. S. Venkatraman, R. Lucito, and M. Wigler, "Circular binary segmentation for the analysis of array-based DNA copy number data," *Biostatistics*, vol. 5, pp. 557–572, Oct. 2004.

[180] E. S. Venkatraman and A. B. Olshen, "A faster circular binary segmentation al-

gorithm for the analysis of array CGH data," *Bioinformatics*, vol. 23, pp. 657–663, Mar. 2007.

[181] M.-C. Huang, T.-P. Chuang, C.-H. Chen, J.-Y. Wu, Y.-T. Chen, L.-H. Li, and H.-C. Yang, "An integrated analysis tool for analyzing hybridization intensities and genotypes using new-generation population-optimized human arrays," *BMC Genomics*, vol. 17, p. 266, Mar. 2016.

[182] V. Seshan and A. Olshen, "DNAcopy: DNA copy number data analysis. R package version 1.60.0," 2019.

[183] D. A. Fournier, H. J. Skaug, J. Ancheta, J. Ianelli, A. Magnusson, M. Maunder, A. Nielsen, and J. Sibert, "AD model builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models," *Optim. Methods Softw.*, vol. 27, no. 2, pp. 233–249, 2011.

[184] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, Y. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati,

D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, J. Szustakowki, and International Human Genome Sequencing Consortium, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, pp. 860–921, Feb. 2001.

[185] E. Kummari, S. X. Guo-Ross, and J. B. Eells, "Laser capture microdissection–a demonstration of the isolation of individual dopamine neurons and the entire ventral tegmental area," *J. Vis. Exp.*, p. e52336, Feb. 2015.

[186] R. Wang, D.-Y. Lin, and Y. Jiang, "SCOPE: A normalization and Copy-Number estimation method for Single-Cell DNA sequencing," *Cell Syst*, vol. 10, pp. 445–452.e6, May 2020.

[187] C. B. Begg, K. Eng, A. Olshen, and E. S. Venkatraman, "Statistical evaluation of evidence for clonal allelic alterations in array-CGH experiments," 2007.

[188] I. Ostrovnaya, V. E. Seshan, and C. B. Begg, "Comparison of properties of tests for assessing tumor clonality," *Biometrics*, vol. 64, pp. 1018–1022, Feb. 2008.

[189] T. W. Geurts, P. M. Nederlof, M. W. M. van den Brekel, L. J. van't Veer, D. de Jong, A. A. M. Hart, N. van Zandwijk, H. Klomp, A. J. M. Balm, and M.-L. F. van Velthuysen, "Pulmonary squamous cell carcinoma following head and neck squamous cell carcinoma: metastasis or second primary?," *Clin. Cancer Res.*, vol. 11, pp. 6608–6614, Sept. 2005.

[190] P. P. Leong, B. Rezai, W. M. Koch, A. Reed, D. Eisele, D. J. Lee, D. Sidransky, J. Jen, and W. H. Westra, "Distinguishing second primary tumors from lung metastases in patients with head and neck squamous cell carcinoma," *J. Natl. Cancer Inst.*, vol. 90, pp. 972–977, July 1998.

[191] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," *Cell*, vol. 144, pp. 646–674, Mar. 2011.

[192] F. Mertens, B. Johansson, M. Höglund, and F. Mitelman, "Chromosomal imbalance maps of malignant solid tumors: a cytogenetic survey of 3185 neoplasms," *Cancer Res.*, vol. 57, pp. 2765–2780, July 1997.

[193] F. Mitelman, "Recurrent chromosome aberrations in cancer," *Mutat. Res.*, vol. 462, pp. 247–253, Apr. 2000.

[194] J. Huang, C. Behrens, I. Wistuba, A. F. Gazdar, and J. Jagirdar, "Molecular analysis of synchronous and metachronous tumors of the lung: impact on management and prognosis," *Ann. Diagn. Pathol.*, vol. 5, pp. 321–329, Dec. 2001.

[195] P. K. Ha and J. A. Califano, "The molecular biology of mucosal field cancerization of the head and neck," *Crit. Rev. Oral Biol. Med.*, vol. 14, no. 5, pp. 363–369, 2003.

[196] E. N. Imyanitov, E. N. Suspitsin, M. Y. Grigoriev, A. V. Togo, E. S. Kuligina, E. V. Belogubova, K. M. Pozharisski, E. A. Turkevich, C. Rodriquez, C. J.

Cornelisse, K. P. Hanson, and C. Theillet, "Concordance of allelic imbalance profiles in synchronous and metachronous bilateral breast carcinomas," *International Journal of Cancer*, vol. 100, pp. 557–564, Aug. 2002.

[197] C. Hafner, R. Knuechel, R. Stoehr, and A. Hartmann, "Clonality of multifocal urothelial carcinomas: 10 years of molecular genetic studies," *Int. J. Cancer*, vol. 101, pp. 1–6, Sept. 2002.

[198] "Loss of heterozygosity," in *Encyclopedia of Cancer* (M. Schwab, ed.), pp. 2075–2076, Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

[199] M. A. Bollet, N. Servant, P. Neuvial, C. Decraene, I. Lebigot, J.-P. Meyniel, Y. De Rycke, A. Savignoni, G. Rigaill, P. Hupé, A. Fourquet, B. Sigal-Zafrani, E. Barillot, and J.-P. Thiery, "High-resolution mapping of DNA breakpoints to define true recurrences among ipsilateral breast cancers," *J. Natl. Cancer Inst.*, vol. 100, pp. 48–58, Dec. 2007.

[200] N. Girard, I. Ostrovnaya, C. Lau, B. Park, M. Ladanyi, D. Finley, C. Deshpande, V. Rusch, I. Orlow, W. D. Travis, W. Pao, and C. B. Begg, "Genomic and mutational profiling to assess clonal relationships between multiple non-small cell lung cancers," *Clin. Cancer Res.*, vol. 15, pp. 5184–5190, Aug. 2009.

[201] F. Kraft and I. Kurth, "Long-read sequencing in human genetics," *Med. Genet.*, vol. 31, pp. 198–204, June 2019.

[202] M. L. Metzker, "Sequencing technologies - the next generation," *Nat. Rev. Genet.*, vol. 11, pp. 31–46, Dec. 2009.

[203] L. Feuk, A. R. Carson, and S. W. Scherer, "Structural variation in the human genome," *Nat. Rev. Genet.*, vol. 7, pp. 85–97, Feb. 2006.

[204] A. P. J. de Koning, W. Gu, T. A. Castoe, M. A. Batzer, and D. D. Pollock, "Repetitive elements may comprise over Two-Thirds of the human genome," *PLoS Genet.*, vol. 7, no. 12, pp. 1–12, 2011.

[205] T. J. Treangen and S. L. Salzberg, "Repetitive DNA and next-generation sequencing: computational challenges and solutions," *Nat. Rev. Genet.*, vol. 13, pp. 36–46, Jan. 2012.

[206] M. J. Levene, J. Korlach, S. W. Turner, M. Foquet, H. G. Craighead, and W. W. Webb, "Zero-mode waveguides for single-molecule analysis at high concentrations," *Science*, vol. 299, pp. 682–686, Jan. 2003.

[207] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. Dewinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner, "Real-time DNA sequencing from single polymerase molecules," *Science*, vol. 323, pp. 133–138, Nov. 2008.

[208] J. Shendure, S. Balasubramanian, G. M. Church, W. Gilbert, J. Rogers, J. A. Schloss, and R. H. Waterston, "DNA sequencing at 40: past, present and future," *Nature*, vol. 550, pp. 345–353, Oct. 2017.

[209] M. Jain, H. E. Olsen, B. Paten, and M. Akeson, "The oxford nanopore MinION: delivery of nanopore sequencing to the genomics community," *Genome Biol.*, vol. 17, p. 239, Nov. 2016.

[210] F. J. Rang, W. P. Kloosterman, and J. de Ridder, "From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy," *Genome Biol.*, vol. 19, p. 90, July 2018.

[211] S. Mitsuhashi and N. Matsumoto, "Long-read sequencing for rare human genetic diseases," *J. Hum. Genet.*, vol. 65, pp. 11–19, Jan. 2020.

[212] E. E. Schadt, S. Turner, and A. Kasarskis, "A window into third-generation sequencing," *Hum. Mol. Genet.*, vol. 19, pp. R227–40, Sept. 2010.

Appendix A

INFORMATION ABOUT THE SIGNIFICANT CNVS IN THE UNCORRECTED DATA

Table A.1: Id: The id of the fastq file; Chr: The name of the chromosome that CNV is located; Start, End: The start and end position of the CNV; z2_score: Z-score for the deviation from the expectation; z1_score: Z-score for each CNV that reflects the deviation in read count distribution; Diagnosis: The state of the individual.

Id	Chr	Start	End	Cnv	Diagnosis
UU_107_12_S77	17	58256478	74514052	3	Control
UU_3862_79_L034	7	45295083	65896370	1	AD
UU_5138_96_S147	13	44035004	80109253	1	AD
UU_5433_51_S53	3	56213156	67019640	3	Control
UU_5603_21_L005	9	71032126	125699272	1	Control
UU_5946_28_L001	22	1	24295222	1	AD
UU_6563_49_L006	20	25787596	36796499	3	AD
UU_7354_I_137_S43	13	101594898	111734294	3	AD
UU_7354_I_137_S43	1	73189692	150204341	1	AD
UU_7654_12_S3	7	45808404	63834492	1	AD
UU_7654_15_S13	14	1	22375568	1	AD
UU_7654_26_S21	11	48176543	61353311	1	AD
UU_7753_36_L007	16	28407614	63225072	1	AD
UU_7903_61_L004	10	49428736	59674967	1	AD
UU_87_L006	8	104342980	117105149	3	Control

Appendix B

INFORMATION ABOUT THE SIGNIFICANT CNVS IN THE CORRECTED DATA

Table B.1: Id: The id of the fastq file; Chr: The name of the chromosome that CNV is located; Start: The start position of the CNV; End: The end position of the variation; Cnv: The integer copy number state of the copy number variation; z2_score: Z-score for the deviation from the expectation (1-somy or 3-somy); z1_score: Z-score for each CNV that reflects the deviation in read count distribution; Diagnosis: The state of the individual that cell was taken from.

Id	Chr	Start	End	Cnv	Diagnosis
UU_5138_35_S134	4	86984399	106932571	1	AD
UU_5433_326_S30	8	96140297	119137264	1	Control
UU_5603_52_L006	14	1	24916961	1	Control
UU_7354_I_103_S36	8	43329609	55035801	3	AD
UU_7354_I_103_S36	22	1	23276518	1	AD
UU_7354_II_047_S92	2	164822153	176008128	1	AD
UU_7542_II_097_S48	9	1	11826032	1	AD
UU_8326_7_S83	6	95928146	107672699	1	AD
UU_8730_II_093_S188	1	76770431	88562330	1	Control
UU_8730_II_097_S192	9	74083336	85787269	1	Control

Appendix C

INFORMATION ABOUT THE SIGNIFICANT CNVS IN VAN DEN BOS DATA

Table C.1: Id: The id of the fastq file; Chr: The name of the chromosome that CNV is located; Start: The start position of the CNV; End: The end position of the variation; Cnv: The integer copy number state of the copy number variation; z2_score: Z-score for the deviation from the expectation (1-somy or 3-somy); z1_score: Z-score for each CNV that reflects the deviation in read count distribution; Diagnosis: The state of the individual that cell was taken from.

Id	Chr	Start	End	Cnv	Diagnosis
ERR1378282	22	1	23782366	1	Control
ERR1378629	4	108476285	119736342	1	AD
ERR1378635	13	1	30294913	1	AD
ERR1378719	14	77445367	89115254	1	AD
ERR1378826	3	152127567	172018108	1	AD
ERR1378878	3	89454397	111352570	1	AD
ERR1379126	14	40760956	57126369	1	AD
ERR1379307	11	85385566	101970188	1	AD
ERR1379307	11	47164512	60338499	1	AD
ERR1379401	12	63446848	107326971	1	AD
ERR1379401	3	141979881	153142991	1	AD
ERR1379401	12	10901928	25156069	1	AD
ERR1391220	6	13337848	24533404	1	Control
ERR1391220	4	17484969	54220408	1	Control

Appendix D

THE EFFECTS OF DIAGNOSIS AND COVERAGE ON THE FREQUENCY OF CNVS IN THE UNCORRECTED-FILTERED DATA

Table D.1: The table shows the results of the GLMM that diagnosis and coverage were added as fixed factors to predict the frequency of CNV.

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	-3.034	0.5699	-5.324	<0.0001
DiagnosisControl	-0.1182	0.7604	-0.1554	0.8765
Coverage	3.564	56.66	0.06289	0.9499

Appendix E

THE EFFECTS OF DIAGNOSIS, BRAIN REGIONS, SEX AND COVERAGE ON THE FREQUENCY OF CNVS IN THE UNCORRECTED-FILTERED DATA

Table E.1: The table shows the results of the GLMM that diagnosis, brain regions, sex and coverage were added as fixed factors to predict the frequency of CNV.

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	-20.44	8731	-0.002341	0.9981
ChrChr7	18.84	8731	0.002158	0.9983
ChrChr13	18.84	8731	0.002158	0.9983
ChrChr3	-1.519e-05	12350	-1.23e-09	1
ChrChr9	1.222e-05	12350	9.896e-10	1
ChrChr22	18.15	8731	0.002079	0.9983
ChrChr20	18.15	8731	0.002079	0.9983
ChrChr1	18.15	8731	0.002079	0.9983
ChrChr14	18.15	8731	0.002079	0.9983
ChrChr11	18.15	8731	0.002079	0.9983
ChrChr16	18.15	8731	0.002079	0.9983
ChrChr10	18.15	8731	0.002079	0.9983
ChrChr2	18.15	8731	0.002079	0.9983
ChrChr8	4.26e-06	12340	3.452e-10	1
ChrChr12	-1.047e-06	12350	-8.476e-11	1
ChrChr15	-3.237e-06	12350	-2.621e-10	1
ChrChr18	-4.314e-06	12350	-3.493e-10	1

Table E.1: Table E (continued)

ChrChr19	-1.896e-06	12350	-1.535e-10	1
ChrChr21	-1.88e-07	12350	-1.522e-11	1
ChrChr4	2.063e-06	12350	1.671e-10	1
ChrChr5	-1.565e-06	12350	-1.267e-10	1
ChrChr6	1.092e-05	12350	8.839e-10	1
DiagnosisControl	19.14	8731	0.002192	0.9983
Brain_regionTemporal_Cortex	-0.4751	0.6891	-0.6894	0.4906
SexMale	0.1779	0.5439	0.3271	0.7436
Coverage	17.16	60.73	0.2825	0.7776

Appendix F

THE EFFECTS OF CHROMOSOME, DIAGNOSIS, BRAIN REGIONS, SEX AND COVERAGE ON THE FREQUENCY OF CNVS IN THE UNCORRECTED-FILTERED DATA

Table F.1: The table shows the results of the GLMM that chromosome, diagnosis, brain regions, sex and coverage were added as fixed factors to predict the frequency of CNV.

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	-2.575	1.27	-2.029	0.04249
ChrChr7	0.6932	1.225	0.566	0.5714
ChrChr13	0.6931	1.225	0.5659	0.5715
ChrChr3	3.399e-06	1.414	2.404e-06	1
ChrChr9	5.634e-06	1.414	3.984e-06	1
ChrChr22	4.247e-06	1.414	3.003e-06	1
ChrChr20	1.356e-05	1.414	9.587e-06	1
ChrChr1	8.246e-05	1.414	-5.831e-05	1
ChrChr14	1.795e-05	1.414	1.269e-05	1
ChrChr11	3.455e-05	1.414	2.443e-05	1
ChrChr16	2.769e-05	1.414	1.958e-05	1
ChrChr10	1.251e-05	1.414	8.848e-06	1
ChrChr2	1.713e-06	1.414	1.211e-06	1
ChrChr8	4.492e-05	1.414	3.176e-05	1
ChrChr12	-18.51	10430	-0.001774	0.9986
ChrChr15	-18.51	10430	-0.001774	0.9986
ChrChr18	-18.51	10430	-0.001773	0.9986

Table F.1: Table F (continued)

ChrChr19	-18.51	10430	-0.001773	0.9986
ChrChr21	-18.51	10430	-0.001774	0.9986
ChrChr4	-18.51	10430	-0.001774	0.9986
ChrChr5	-18.51	10430	-0.001774	0.9986
ChrChr6	-18.51	10430	-0.001774	0.9986
DiagnosisControl	-0.111	0.7721	-0.1437	0.8857
Brain_regionTemporal_Cortex	-0.4751	0.6892	-0.6893	0.4906
SexMale	0.1779	0.5439	0.327	0.7437
Coverage	17.17	60.73	0.2828	0.7773

Appendix G

THE EFFECTS OF CHROMOSOME, DIAGNOSIS, BRAIN REGIONS AND COVERAGE ON THE FREQUENCY OF CNVS IN THE UNCORRECTED-FILTERED DATA

Table G.1: The table shows the results of the GLMM that interaction of chromosome and diagnosis, brain regions, sex and coverage were added as fixed factors to predict the frequency of CNV.

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	-22.17	22200	-0.0009987	0.9992
ChrChr7	20.71	22200	0.0009329	0.9993
ChrChr13	20.71	22200	0.0009329	0.9993
ChrChr3	-1.665e-05	31360	-5.309e-10	1
ChrChr9	1.345e-05	31590	4.26e-10	1
ChrChr22	20.02	22200	0.0009017	0.9993
ChrChr20	20.02	22200	0.0009017	0.9993
ChrChr1	20.02	22200	0.0009017	0.9993
ChrChr14	20.02	22200	0.0009017	0.9993
ChrChr11	20.02	22200	0.0009017	0.9993
ChrChr16	20.02	22200	0.0009017	0.9993
ChrChr10	20.02	22200	0.0009017	0.9993
ChrChr2	20.02	22200	0.0009017	0.9993
ChrChr8	4.676e-06	31380	1.49e-10	1
ChrChr12	-1.149e-06	31450	-3.654e-11	1
ChrChr15	-3.553e-06	31440	-1.13e-10	1
ChrChr18	-4.735e-06	31440	-1.506e-10	1

Table G.1: Table G (continued)

ChrChr19	-2.082e-06	31460	-6.616e-11	1
ChrChr21	-2.063e-07	31450	-6.561e-12	1
ChrChr4	2.265e-06	31470	7.196e-11	1
ChrChr5	-1.718e-06	31460	-5.46e-11	1
ChrChr6	1.198e-05	31450	3.81e-10	1
DiagnosisControl	21.02	22200	0.0009469	0.9992
Brain_regionTemporal_Cortex	-0.4378	0.6868	-0.6374	0.5239
Coverage	11.54	58.52	0.1973	0.8436

Appendix H

THE EFFECTS OF CHROMOSOME, DIAGNOSIS, BRAIN REGIONS AND COVERAGE ON THE FREQUENCY OF CNVS IN THE UNCORRECTED-FILTERED DATA

Table H.1: The table shows the results of the GLMM that chromosome, diagnosis, brain regions and coverage were added as fixed factors to predict the frequency of CNV.

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	-2.441	1.189	-2.052	0.04016
ChrChr7	0.693	1.225	0.5658	0.5715
ChrChr13	0.6931	1.225	0.5659	0.5714
ChrChr3	-0.0001457	1.414	-0.000103	0.9999
ChrChr9	-6.131e-05	1.414	-4.335e-05	1
ChrChr22	-0.000116	1.414	-8.202e-05	0.9999
ChrChr20	-1.25e-05	1.414	-8.841e-06	1
ChrChr1	-0.0003361	1.414	-0.0002376	0.9998
ChrChr14	-6.301e-05	1.414	-4.455e-05	1
ChrChr11	-0.0001026	1.414	-7.252e-05	0.9999
ChrChr16	-3.996e-05	1.414	-2.826e-05	1
ChrChr10	-6.213e-05	1.414	-4.394e-05	1
ChrChr2	-0.0001153	1.414	-8.152e-05	0.9999
ChrChr8	-5.347e-05	1.414	-3.781e-05	1
ChrChr12	-15.91	2853	-0.005577	0.9956
ChrChr15	-15.91	2853	-0.005577	0.9956
ChrChr18	-15.91	2853	-0.005577	0.9956

Table H.1: Table H (continued)

ChrChr19	-15.91	2853	-0.005577	0.9956
ChrChr21	-15.91	2853	-0.005577	0.9956
ChrChr4	-15.91	2853	-0.005577	0.9956
ChrChr5	-15.91	2853	-0.005577	0.9956
ChrChr6	-15.91	2853	-0.005577	0.9956
DiagnosisControl	-0.09548	0.7742	-0.1233	0.9018
Brain_regionTemporal_Cortex	-0.4377	0.6868	-0.6373	0.5239
Coverage	11.54	58.52	0.1973	0.8436

Appendix I

THE EFFECTS OF CHROMOSOME, DIAGNOSIS AND COVERAGE ON THE FREQUENCY OF CNVS IN THE UNCORRECTED-FILTERED DATA

Table I.1: The table shows the results of the GLMM that interaction of chromosome and diagnosis and coverage were added as fixed factors to predict the frequency of CNV.

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	-22.72	25500	-0.000891	0.9993
ChrChr7	20.99	25500	0.000823	0.9993
ChrChr13	20.99	25500	0.000823	0.9993
ChrChr3	-1.689e-05	35800	-4.719e-10	1
ChrChr9	1.364e-05	35850	3.806e-10	1
ChrChr22	20.29	25500	0.0007958	0.9994
ChrChr20	20.29	25500	0.0007958	0.9994
ChrChr1	20.29	25500	0.0007958	0.9994
ChrChr14	20.29	25500	0.0007958	0.9994
ChrChr11	20.29	25500	0.0007958	0.9994
ChrChr16	20.29	25500	0.0007958	0.9994
ChrChr10	20.29	25500	0.0007958	0.9994
ChrChr2	20.29	25500	0.0007958	0.9994
ChrChr8	4.744e-06	36340	1.305e-10	1
ChrChr12	-1.166e-06	36070	-3.232e-11	1
ChrChr15	-3.604e-06	36080	-9.991e-11	1
ChrChr18	-4.804e-06	36060	-1.332e-10	1

Table I.1: Table I (continued)

ChrChr19	-2.112e-06	36130	-5.845e-11	1
ChrChr21	-2.093e-07	36070	-5.803e-12	1
ChrChr4	2.298e-06	36050	6.374e-11	1
ChrChr5	-1.743e-06	36100	-4.827e-11	1
ChrChr6	1.216e-05	36060	3.371e-10	1
DiagnosisControl	21.27	25500	0.0008342	0.9993
Coverage	3.589	56.66	0.06334	0.5239
ChrChr7:Diagnosiscontrol	-42.29	48940	-0.0008641	0.9993
ChrChr13:Diagnosiscontrol	-42.29	49170	-0.0008602	0.9993
ChrChr3:Diagnosiscontrol	-0.0001432	35800	-4e-09	1
ChrChr9:Diagnosiscontrol	-3.587e-05	35850	-1.001e-09	1
ChrChr22:Diagnosiscontrol	-41.6	49410	-0.0008419	0.9993
ChrChr20:Diagnosiscontrol	-41.6	49110	-0.0008471	0.9993
ChrChr1:Diagnosiscontrol	-41.6	49440	-0.0008414	0.9993
ChrChr14:Diagnosiscontrol	-41.6	49460	-0.000841	0.9993
ChrChr11:Diagnosiscontrol	-41.6	49360	-0.0008427	0.9993
ChrChr16:Diagnosiscontrol	-41.6	49390	-0.0008422	0.9993
ChrChr10:Diagnosiscontrol	-41.6	49220	-0.0008451	0.9993
ChrChr2:Diagnosiscontrol	-41.6	49530	-0.0008399	0.9993
ChrChr8:Diagnosiscontrol	-8.318e-05	36340	-2.289e-09	1
ChrChr12:Diagnosiscontrol	-21.31	55780	-0.000382	0.9997
ChrChr15:Diagnosiscontrol	-21.31	55650	-0.0003828	0.9997
ChrChr18:Diagnosiscontrol	-21.31	55690	-0.0003826	0.9997
ChrChr19:Diagnosiscontrol	-21.31	55720	-0.0003824	0.9997
ChrChr21:Diagnosiscontrol	-21.31	55620	-0.000383	0.9997
ChrChr4:Diagnosiscontrol	-21.31	55550	-0.0003835	0.9997
ChrChr5:Diagnosiscontrol	-21.31	55540	-0.0003836	0.9997
ChrChr6:Diagnosiscontrol	-21.31	55520	-0.0003837	0.9997

Appendix J

THE EFFECTS OF CHROMOSOME, DIAGNOSIS AND COVERAGE ON THE FREQUENCY OF CNVS IN THE UNCORRECTED-FILTERED DATA

Table J.1: The table shows the results of the GLMM that chromosome, diagnosis and coverage were added as fixed factors to predict the frequency of CNV.

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	-2.716	1.124	-2.417	0.01563
ChrChr7	0.6931	1.225	0.5659	0.5714
ChrChr13	0.6931	1.225	0.5659	0.5714
ChrChr3	-7.826e-05	1.414	-5.534e-05	1
ChrChr9	-2.08e-05	1.414	-1.471e-05	1
ChrChr22	-8.441e-05	1.414	-5.969e-05	1
ChrChr20	-2.588e-05	1.414	-1.83e-05	1
ChrChr1	-3.683e-05	1.414	-2.605e-05	1
ChrChr14	-5.146e-05	1.414	-3.639e-05	1
ChrChr11	-6.307e-05	1.414	-4.46e-05	1
ChrChr16	-3.28e-05	1.414	-2.319e-05	1
ChrChr10	-4.468e-05	1.414	-3.159e-05	1
ChrChr2	-8.541e-05	1.414	-6.039e-05	1
ChrChr8	-3.793e-07	1.414	-2.682e-07	1
ChrChr12	-18.96	13090	-0.001449	0.9988
ChrChr15	-18.96	13090	-0.001448	0.9988
ChrChr18	-18.96	13100	-0.001447	0.9988

Table J.1: Table J (continued)

ChrChr19	-18.96	13100	-0.001448	0.9988
ChrChr21	-18.96	13100	-0.001448	0.9988
ChrChr4	-18.96	13100	-0.001448	0.9988
ChrChr5	-18.96	13090	-0.001448	0.9988
ChrChr6	-18.96	13090	-0.001448	0.9988
Diagnosiscontrol	-0.1182	0.7604	-0.1555	0.8765
Coverage	3.586	56.66	0.06328	0.9495

Appendix K

THE EFFECTS OF DIAGNOSIS AND COVERAGE ON THE FREQUENCY OF CNVS IN THE PCA-CORRECTED DATA

Table K.1: The table shows the results of the GLMM that diagnosis and coverage were added as fixed factors to predict the frequency of CNV.

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	-3.205	0.7247	-4.423	<0.0001
DiagnosisControl	-0.168	0.6474	-0.2595	0.7953
Coverage	40.27	76.4	0.527	0.5982

Appendix L

THE EFFECTS OF CHROMOSOME, DIAGNOSIS, BRAINS REGIONS, COVERAGE AND SEX ON THE FREQUENCY OF CNVS IN THE PCA-CORRECTED DATA

Table L.1: The table shows the results of the GLMM that interaction of chromosome and diagnosis, brain regions, sex and coverage were added as fixed factors to predict the frequency of CNV.

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	-2.268	2.042	-1.11	0.2668
ChrChr8	8.401e-06	1.414	5.941e-06	1
ChrChr14	-17.23	5508	-0.003128	0.9975
ChrChr22	2.302e-06	1.414	1.628e-06	1
ChrChr2	-3.136e-05	1.414	-2.218e-05	1
ChrChr9	-2.622e-05	1.414	-1.854e-05	1
ChrChr6	5.389e-06	1.414	3.81e-06	1
ChrChr1	-17.23	5506	-0.003129	0.9975
ChrChr10	-17.23	5507	-0.003128	0.9975
ChrChr11	-17.23	5507	-0.003128	0.9975
ChrChr12	-17.23	5507	-0.003128	0.9975
ChrChr13	-17.23	5506	-0.003129	0.9975
ChrChr15	-17.23	5507	-0.003128	0.9975
ChrChr16	-17.23	5507	-0.003128	0.9975
ChrChr17	-17.23	5507	-0.003128	0.9975
ChrChr18	-17.23	5507	-0.003129	0.9975
ChrChr19	-17.23	5507	-0.003128	0.9975

Table L.1: Table L (continued)

ChrChr20	-17.23	5507	-0.003128	0.9975
ChrChr21	-17.23	5507	-0.003128	0.9975
ChrChr3	-17.23	5507	-0.003128	0.9975
ChrChr5	-17.23	5507	-0.003128	0.9975
ChrChr7	-17.23	5507	-0.003128	0.9975
DiagnosisControl	-16.91	6127	-0.00276	0.9978
Brain_regionEntorhinal_Cortex	0.6403	2.279	0.2809	0.7788
Brain_regionHippocampus_CA3	0.3698	1.956	0.1891	0.85
Brain_regionTemporal_Cortex	-0.16	1.462	-0.1094	0.9129
SexMale	-0.2137	1.341	-0.1594	0.8733
Coverage	45.62	106.1	0.4299	0.6673
ChrChr8:DiagnosisControl	17.44	6127	0.002847	0.9977
ChrChr14:DiagnosisControl	34.67	8238	0.004208	0.9966
ChrChr22:DiagnosisControl	9.486e-06	8664	1.095e-09	1
ChrChr2:DiagnosisControl	2.255e-05	8664	2.602e-09	1
ChrChr9:DiagnosisControl	17.44	6127	0.002847	0.9977
ChrChr6:DiagnosisControl	-1.985e-05	8663	-2.291e-09	1
ChrChr1:DiagnosisControl	34.67	8238	0.004209	0.9966
ChrChr10:DiagnosisControl	17.23	10270	0.001678	0.9987
ChrChr11:DiagnosisControl	17.23	10270	0.001678	0.9987
ChrChr12:DiagnosisControl	17.23	10270	0.001678	0.9987
ChrChr13:DiagnosisControl	17.23	10270	0.001678	0.9987
ChrChr15:DiagnosisControl	17.23	10270	0.001678	0.9987
ChrChr16:DiagnosisControl	17.23	10270	0.001678	0.9987
ChrChr17:DiagnosisControl	17.23	10270	0.001678	0.9987
ChrChr18:DiagnosisControl	17.23	10270	0.001678	0.9987
ChrChr19:DiagnosisControl	17.23	10270	0.001678	0.9987
ChrChr20:DiagnosisControl	17.23	10270	0.001678	0.9987
ChrChr21:DiagnosisControl	17.23	10270	0.001678	0.9987
ChrChr3:DiagnosisControl	17.23	10270	0.001678	0.9987

Table L.1: Table L (continued)

ChrChr5:DiagnosisControl	17.23	10270	0.001678	0.9987
ChrChr7:DiagnosisControl	17.23	10270	0.001678	0.9987

Appendix M

THE EFFECTS OF CHROMOSOME, DIAGNOSIS, BRAINS REGIONS, COVERAGE AND SEX ON THE FREQUENCY OF CNVS IN THE PCA-CORRECTED DATA

Table M.1: The table shows the results of the GLMM that chromosome, diagnosis, brain regions, sex and coverage were added as fixed factors to predict the frequency of CNV.

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	-2.779	2.058	-1.35	0.177
ChrChr8	0.6932	1.225	0.566	0.5714
ChrChr14	4.769e-06	1.414	3.372e-06	1
ChrChr22	0.0003745	1.414	0.0002648	0.9998
ChrChr2	-2.708e-06	1.414	-1.915e-06	1
ChrChr9	0.6932	1.225	0.566	0.5714
ChrChr6	4.19e-05	1.414	2.963e-05	1
ChrChr1	-6.059e-06	1.414	-4.284e-06	1
ChrChr10	-17.74	7125	-0.00249	0.998
ChrChr11	-17.74	7125	-0.00249	0.998
ChrChr12	-17.74	7124	-0.00249	0.998
ChrChr13	-17.74	7125	-0.00249	0.998
ChrChr15	-17.74	7124	-0.00249	0.998
ChrChr16	-17.74	7125	-0.00249	0.998
ChrChr17	-17.74	7125	-0.00249	0.998
ChrChr18	-17.74	7125	-0.00249	0.998
ChrChr19	-17.74	7125	-0.00249	0.998

Table M.1: Table M (continued)

ChrChr20	-17.74	7124	-0.00249	0.998
ChrChr21	-17.74	7125	-0.00249	0.998
ChrChr3	-17.74	7125	-0.00249	0.998
ChrChr5	-17.74	7124	-0.00249	0.998
ChrChr7	-17.74	7125	-0.00249	0.998
DiagnosisControl	0.1263	1.056	0.1195	0.9048
Brain_regionEntorhinal_Cortex	0.6408	2.279	0.2811	0.7786
Brain_regionHippocampus_CA3	0.37	1.956	0.1892	0.8499
Brain_regionTemporal_Cortex	-0.16	1.462	-0.1094	0.9129
Sexmale	-0.2137	1.341	-0.1594	0.8734
Coverage	45.61	106.1	0.4298	0.6674

Appendix N

THE EFFECTS OF CHROMOSOME, DIAGNOSIS, BRAINS REGIONS AND COVERAGE ON THE FREQUENCY OF CNVS IN THE PCA-CORRECTED DATA

Table N.1: The table shows the results of the GLMM that interaction of chromosome and diagnosis, brain regions and coverage were added as fixed factors to predict the frequency of CNV.

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	-2.14	1.862	-1.149	0.2506
ChrChr8	-3.707e-05	1.414	-2.621e-05	1
ChrChr14	-17.63	6736	-0.002617	0.9979
ChrChr22	-4.331e-05	1.414	-3.062e-05	1
ChrChr2	-3.472e-05	1.414	-2.455e-05	1
ChrChr9	-3.088e-05	1.414	-2.184e-05	1
chrchr6	-1.194e-05	1.414	-8.44e-06	1
ChrChr1	-17.63	6736	-0.002618	0.9979
ChrChr10	-17.63	6735	-0.002618	0.9979
ChrChr11	-17.63	6735	-0.002618	0.9979
ChrChr12	-17.63	6736	-0.002617	0.9979
ChrChr13	-17.63	6735	-0.002618	0.9979
ChrChr15	-17.63	6735	-0.002618	0.9979
ChrChr16	-17.63	6736	-0.002617	0.9979
ChrChr17	-17.63	6736	-0.002617	0.9979
ChrChr18	-17.63	6735	-0.002618	0.9979
ChrChr19	-17.63	6736	-0.002618	0.9979

Table N.1: Table N (continued)

ChrChr20	-17.63	6735	-0.002618	0.9979
ChrChr21	-17.63	6736	-0.002617	0.9979
ChrChr3	-17.63	6736	-0.002617	0.9979
ChrChr5	-17.63	6736	-0.002617	0.9979
ChrChr7	-17.63	6735	-0.002618	0.9979
DiagnosisControl	-17.39	7490	-0.002322	0.9981
Brain_regionEntorhinal_Cortex	0.3666	1.475	0.2486	0.8037
Brain_regionHippocampus_CA3	0.1333	1.268	0.1051	0.9163
Brain_regionTemporal_Cortex	-0.1369	1.454	-0.09413	0.925
Coverage	39.04	96.4	0.4049	0.6855
ChrChr8:DiagnosisControl	17.84	7490	0.002382	0.9981
ChrChr14:DiagnosisControl	35.47	10070	0.003522	0.9972
ChrChr22:DiagnosisControl	5.537e-05	10590	5.228e-09	1
ChrChr2:DiagnosisControl	2.57e-05	10590	2.427e-09	1
ChrChr9:DiagnosisControl	17.84	7490	0.002382	0.9981
ChrChr6:DiagnosisControl	-2.862e-06	10590	-2.702e-10	1
ChrChr1:DiagnosisControl	35.47	10070	0.003522	0.9972
ChrChr10:DiagnosisControl	17.63	12550	0.001404	0.9989
ChrChr11:DiagnosisControl	17.63	12550	0.001404	0.9989
ChrChr12:DiagnosisControl	17.63	12550	0.001405	0.9989
ChrChr13:DiagnosisControl	17.63	12550	0.001405	0.9989
ChrChr15:DiagnosisControl	17.63	12550	0.001405	0.9989
ChrChr16:DiagnosisControl	17.63	12550	0.001405	0.9989
ChrChr17:DiagnosisControl	17.63	12550	0.001404	0.9989
ChrChr18:DiagnosisControl	17.63	12550	0.001405	0.9989
ChrChr19:DiagnosisControl	17.63	12550	0.001404	0.9989
ChrChr20:DiagnosisControl	17.63	12550	0.001404	0.9989
ChrChr21:DiagnosisControl	17.63	12550	0.001405	0.9989
ChrChr3:Diagnosiscontrol	17.63	12550	0.001404	0.9989
ChrChr5:DiagnosisControl	17.63	12550	0.001404	0.9989

Table N.1: Table N (continued)

ChrChr7:DiagnosisControl	17.63	12550	0.001404	0.9989

Appendix O

THE EFFECTS OF CHROMOSOME, DIAGNOSIS, BRAINS REGIONS AND COVERAGE ON THE FREQUENCY OF CNVS IN THE PCA-CORRECTED DATA

Table O.1: The table shows the results of the GLMM that chromosome, brain regions and coverage were added as fixed factors to predict the frequency of CNV.

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	-2.651	1.88	-1.41	0.1585
ChrChr8	0.6932	1.225	0.566	0.5714
ChrChr14	-0.0001017	1.414	-7.192e-05	0.9999
ChrChr22	0.0001801	1.414	0.0001274	0.9999
ChrChr2	-0.0002293	1.414	-0.0001622	0.9999
ChrChr9	0.6931	1.225	0.5659	0.5715
ChrChr6	1.118e-05	1.414	7.904e-06	1
ChrChr1	-7.867e-05	1.414	-5.563e-05	1
ChrChr10	-16.47	3765	-0.004374	0.9965
ChrChr11	-16.47	3765	-0.004373	0.9965
ChrChr12	-16.47	3765	-0.004373	0.9965
ChrChr13	-16.47	3765	-0.004374	0.9965
ChrChr15	-16.47	3765	-0.004374	0.9965
ChrChr16	-16.47	3765	-0.004374	0.9965
ChrChr17	-16.47	3765	-0.004374	0.9965
ChrChr18	-16.47	3765	-0.004374	0.9965
ChrChr19	-16.47	3765	-0.004374	0.9965

Table O.1: Table O (continued)

ChrChr20	-16.47	3765	-0.004374	0.9965
ChrChr21	-16.47	3765	-0.004373	0.9965
ChrChr3	-16.47	3765	-0.004374	0.9965
ChrChr5	-16.47	3765	-0.004374	0.9965
chrchr7	-16.47	3765	-0.004374	0.9965
DiagnosisControl	0.04898	0.9259	0.0529	0.9578
Brain_regionEntorhinal_Cortex	0.3665	1.475	0.2485	0.8038
Brain_regionHippocampus_CA3	0.1337	1.268	0.1054	0.9161
Brain_regionTemporal_Cortex	-0.1377	1.454	-0.09469	0.9246
Coverage	39.1	96.4	0.4056	0.685

Appendix P

THE EFFECTS OF CHROMOSOME, DIAGNOSIS AND COVERAGE ON THE FREQUENCY OF CNVS IN THE PCA-CORRECTED DATA

Table P.1: The table shows the results of the GLMM that interaction of chromosome and diagnosis and coverage were added as fixed factors to predict the frequency of CNV.

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	-1.906	1.166	-1.635	0.102
ChrChr8	-7.308e-05	1.414	-5.167e-05	1
ChrChr14	-19.68	18810	-0.001046	0.9992
ChrChr22	-8.002e-05	1.414	-5.658e-05	1
ChrChr2	-8.633e-05	1.414	-6.105e-05	1
ChrChr9	-3.616e-05	1.414	-2.557e-05	1
ChrChr6	-1.124e-05	1.414	-7.948e-06	1
ChrChr1	-19.68	18760	-0.001049	0.9992
ChrChr10	-19.68	18790	-0.001048	0.9992
ChrChr11	-19.68	18780	-0.001048	0.9992
ChrChr12	-19.68	18790	-0.001047	0.9992
ChrChr13	-19.68	18790	-0.001047	0.9992
ChrChr15	-19.68	18800	-0.001047	0.9992
ChrChr16	-19.68	18800	-0.001047	0.9992
ChrChr17	-19.68	18790	-0.001047	0.9992
ChrChr18	-19.68	18790	-0.001048	0.9992
ChrChr19	-19.68	18800	-0.001047	0.9992
ChrChr20	-19.68	18790	-0.001047	0.9992

Table P.1: Table P (continued)

ChrChr21	-19.68	18800	-0.001047	0.9992
ChrChr3	-19.68	18800	-0.001047	0.9992
ChrChr5	-19.68	18800	-0.001047	0.9992
ChrChr7	-19.68	18800	-0.001047	0.9992
DiagnosisControl	-19.66	20950	-0.0009386	0.9993
Coverage	40.27	76.4	0.5271	0.5981
ChrChr8:DiagnosisControl	19.9	20950	0.00095	0.9992
ChrChr14:DiagnosisControl	39.58	28150	0.001406	0.9989
ChrChr22:DiagnosisControl	9.349e-05	29620	3.156e-09	1
ChrChr2:DiagnosisControl	7.626e-05	29630	2.574e-09	1
ChrChr9:DiagnosisControl	19.9	20950	0.00095	0.9992
ChrChr6:DiagnosisControl	-5.282e-06	29630	-1.783e-10	1
ChrChr1:DiagnosisControl	39.58	28130	0.001407	0.9989
ChrChr10:DiagnosisControl	19.68	35090	0.000561	0.9996
ChrChr11:DiagnosisControl	19.68	35080	0.0005611	0.9996
ChrChr12:DiagnosisControl	19.68	35080	0.000561	0.9996
ChrChr13:DiagnosisControl	19.68	35080	0.0005611	0.9996
ChrChr15:DiagnosisControl	19.68	35080	0.000561	0.9996
ChrChr16:DiagnosisControl	19.68	35080	0.0005611	0.9996
ChrChr17:DiagnosisControl	19.68	35080	0.0005611	0.9996
ChrChr18:DiagnosisControl	19.68	35080	0.0005611	0.9996
ChrChr19:DiagnosisControl	19.68	35090	0.0005609	0.9996
ChrChr20:DiagnosisControl	19.68	35080	0.0005611	0.9996
ChrChr21:DiagnosisControl	19.68	35080	0.000561	0.9996
ChrChr3:DiagnosisControl	19.68	35080	0.0005611	0.9996
ChrChr5:DiagnosisControl	19.68	35080	0.0005611	0.9996
ChrChr7:DiagnosisControl	19.68	35080	0.000561	0.9996

Appendix Q

THE EFFECTS OF CHROMOSOME, DIAGNOSIS AND COVERAGE ON THE FREQUENCY OF CNVS IN THE PCA-CORRECTED DATA

Table Q.1: The table shows the results of the GLMM that chromosome, diagnosis and coverage were added as fixed factors to predict the frequency of CNV.

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	-2.417	1.194	-2.024	0.04292
ChrChr8	0.6932	1.225	0.566	0.5714
ChrChr14	-2.678e-05	1.414	-1.894e-05	1
ChrChr22	5.337e-05	1.414	3.774e-05	1
ChrChr2	-4.205e-05	1.414	-2.974e-05	1
ChrChr9	0.6931	1.225	0.5659	0.5715
ChrChr6	3.948e-06	1.414	2.792e-06	1
ChrChr1	3.775e-05	1.414	2.67e-05	1
ChrChr10	-15.94	2891	-0.005513	0.9956
ChrChr11	-15.94	2891	-0.005513	0.9956
ChrChr12	-15.94	2891	-0.005514	0.9956
ChrChr13	-15.94	2891	-0.005514	0.9956
ChrChr15	-15.94	2891	-0.005514	0.9956
ChrChr16	-15.94	2891	-0.005513	0.9956
ChrChr17	-15.94	2891	-0.005513	0.9956
ChrChr18	-15.94	2891	-0.005513	0.9956
ChrChr19	-15.94	2891	-0.005513	0.9956
ChrChr20	-15.94	2891	-0.005513	0.9956

Table Q.1: Table Q (continued)

ChrChr21	-15.94	2891	-0.005513	0.9956
ChrChr3	-15.94	2891	-0.005513	0.9956
ChrChr5	-15.94	2891	-0.005513	0.9956
ChrChr7	-15.94	2891	-0.005513	0.9956
DiagnosisControl	-0.1679	0.6474	-0.2594	0.7953
Coverage	40.26	76.4	0.5269	0.5982

Appendix R

THE EFFECTS OF DIAGNOSIS AND COVERAGE ON THE FREQUENCY OF CNVS IN VAN DEN BOS DATA

Table R.1: The table shows the results of the GLMM that diagnosis and coverage were added as fixed factors to predict the frequency of CNV.

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	-2.301	1.426	-1.613	0.1067
Diagnosisnormal	0.1033	0.7325	0.141	0.8879
Coverage	200.2	216.2	0.926	0.3544

Appendix S

THE EFFECTS OF CHROMOSOME, DIAGNOSIS AND COVERAGE ON THE FREQUENCY OF CNVS IN VAN DEN BOS DATA

Table S.1: The table shows the results of the GLMM that interaction of chromosome and diagnosis and coverage were added as fixed factors to predict the frequency of CNV.

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	-21.17	8236	-0.002571	0.9979
ChrChr4	18.03	8236	0.00219	0.9983
ChrChr13	18.03	8236	0.00219	0.9983
ChrChr14	18.73	8236	0.002274	0.9982
ChrChr3	19.13	8236	0.002323	0.9981
ChrChr11	18.72	8236	0.002273	0.9982
ChrChr12	18.72	8236	0.002274	0.9982
ChrChr6	4.003e-06	11650	3.437e-10	1
ChrChr1	-7.964e-06	11650	-6.837e-10	1
ChrChr10	-6.203e-06	11650	-5.326e-10	1
ChrChr15	-7.016e-06	11650	-6.024e-10	1
ChrChr16	-5.838e-06	11650	-5.012e-10	1
ChrChr17	-7.026e-06	11650	-6.032e-10	1
ChrChr18	-5.859e-06	11650	-5.03e-10	1
ChrChr19	-7.156e-06	11650	-6.144e-10	1
ChrChr2	-7.32e-06	11650	-6.285e-10	1
ChrChr20	-5.422e-06	11650	-4.655e-10	1
ChrChr21	-7.32e-06	11650	-6.285e-10	1

Table S.1: Table S (continued)

ChrChr5	-4.294e-06	11650	-3.687e-10	1
ChrChr7	-6.69e-06	11650	-5.744e-10	1
ChrChr8	-8.699e-06	11650	-7.468e-10	1
ChrChr9	-1.701e-13	11650	-1.46e-17	1
DiagnosisNormal	19.48	8236	0.002365	0.9981
Coverage	192.5	195.8	0.9831	0.3256
ChrChr4:DiagnosisNormal	-18.03	8236	-0.00219	0.9983
ChrChr13:DiagnosisNormal	-37.48	18660	-0.002009	0.9984
ChrChr14:DiagnosisNormal	-38.18	18660	-0.002046	0.9984
ChrChr3:DiagnosisNormal	-38.58	18660	-0.002068	0.9984
ChrChr11:DiagnosisNormal	-38.18	18660	-0.002046	0.9984
ChrChr12:DiagnosisNormal	-38.18	18660	-0.002045	0.9984
ChrChr6:DiagnosisNormal	-7.814e-05	11650	-6.708e-09	1
ChrChr1:DiagnosisNormal	-19.45	20400	-0.0009536	0.9992
ChrChr10:DiagnosisNormal	-19.45	20400	-0.0009536	0.9992
ChrChr15:DiagnosisNormal	-19.45	20400	-0.0009535	0.9992
ChrChr16:DiagnosisNormal	-19.45	20400	-0.0009535	0.9992
ChrChr17:DiagnosisNormal	-19.45	20400	-0.0009536	0.9992
ChrChr18:DiagnosisNormal	-19.45	20400	-0.0009535	0.9992
ChrChr19:DiagnosisNormal	-19.45	20400	-0.0009535	0.9992
ChrChr2:DiagnosisNormal	-19.45	20400	-0.0009536	0.9992
ChrChr20:DiagnosisNormal	-19.45	20400	-0.0009538	0.9992
ChrChr21:DiagnosisNormal	-19.45	20400	-0.0009536	0.9992
ChrChr5:DiagnosisNormal	-19.45	20400	-0.0009535	0.9992
ChrChr7:DiagnosisNormal	-19.45	20400	-0.0009535	0.9992
ChrChr8:DiagnosisNormal	-19.45	20400	-0.0009537	0.9992
ChrChr9:DiagnosisNormal	-19.45	20400	-0.0009536	0.9992

Appendix T

THE EFFECTS OF CHROMOSOME, DIAGNOSIS AND COVERAGE ON THE FREQUENCY OF CNVS IN VAN DEN BOS DATA

Table T.1: The table shows the results of the GLMM that chromosome, diagnosis and coverage were added as fixed factors to predict the frequency of CNV.

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	-2.976	1.739	-1.711	0.08708
ChrChr4	0.6948	1.26	0.5514	0.5813
ChrChr13	0.03085	1.45	0.02127	0.983
ChrChr14	0.7201	1.266	0.5689	0.5694
ChrChr3	1.037	1.197	0.8665	0.3862
ChrChr11	0.7622	1.281	0.5951	0.5517
ChrChr12	0.713	1.264	0.5643	0.5726
ChrChr6	0.0106	1.446	0.007333	0.9941
ChrChr1	-18.04	8297	-0.002174	0.9983
ChrChr10	-18.04	8298	-0.002173	0.9983
ChrChr15	-18.04	8295	-0.002174	0.9983
ChrChr16	-18.04	8297	-0.002174	0.9983
ChrChr17	-18.04	8297	-0.002174	0.9983
ChrChr18	-18.04	8298	-0.002173	0.9983
ChrChr19	-18.04	8299	-0.002173	0.9983
ChrChr2	-18.04	8296	-0.002174	0.9983
ChrChr20	-18.04	8298	-0.002174	0.9983
ChrChr21	-18.04	8297	-0.002174	0.9983
ChrChr5	-18.04	8296	-0.002174	0.9983

Table T.1: Table T (continued)

ChrChr7	-18.04	8297	-0.002174	0.9983
ChrChr8	-18.04	8298	-0.002174	0.9983
ChrChr9	-18.04	8297	-0.002174	0.9983
DiagnosisNormal	0.1909	0.7024	0.2718	0.7858
Coverage	181.3	205	0.8842	0.3766

Appendix U

THE EFFECTS OF CHROMOSOME, DIAGNOSIS, BRAIN REGIONS AND COVERAGE ON THE FREQUENCY OF CNVS IN THE UNCORRECTED DATA

Table U.1: The table shows the results of the GLMM that interaction of chromosome and diagnosis, brains regions and coverage were added as fixed factors to predict the frequency of CNV.

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	0.4794	0.108	4.437	<0.0001
ChrChr5	-0.05698	0.04189	-1.36	0.1738
ChrChr9	0.2184	0.03907	5.59	<0.0001
ChrChr7	-0.6189	0.04919	-12.58	<0.0001
ChrChr10	-0.1922	0.04326	-4.443	<0.0001
ChrChr2	-0.02477	0.04178	-0.593	0.5532
ChrChr12	-0.6451	0.04972	-12.98	<0.0001
ChrChr15	-0.8027	0.05156	-15.57	<0.0001
ChrChr18	-0.9557	0.05452	-17.53	<0.0001
ChrChr17	-0.8859	0.05328	-16.63	<0.0001
ChrChr4	-0.4701	0.04749	-9.898	<0.0001
ChrChr14	-0.684	0.04995	-13.69	<0.0001
ChrChr13	-0.7818	0.05166	-15.13	<0.0001
ChrChr3	-0.3766	0.04647	-8.105	<0.0001
ChrChr11	-0.5702	0.04886	-11.67	<0.0001
ChrChr20	-0.8581	0.05272	-16.28	<0.0001
ChrChr22	-1.079	0.05636	-19.15	<0.0001

Table U.1: Table U (continued)

ChrChr16	-0.921	0.0538	-17.12	<0.0001
ChrChr19	-0.8119	0.05173	-15.7	<0.0001
ChrChr21	-1.104	0.05676	-19.46	<0.0001
ChrChr8	-0.8159	0.05234	-15.59	<0.0001
ChrChr6	-0.6059	0.04948	-12.25	<0.0001
DiagnosisControl	0.239	0.1755	1.362	0.1733
Brain_regionEntorhinal_Cortex	0.3164	0.02404	13.16	<0.0001
Brain_regionHippocampus_CA1	0.4514	0.02305	19.59	<0.0001
Brain_regionHippocampus_CA3	0.4471	0.02389	18.72	<0.0001
Brain_regionTemporal_Cortex	-0.756	0.03566	-21.2	<0.0001
Coverage	32.47	1.275	25.46	<0.0001
ChrChr5:DiagnosisControl	-0.1399	0.0619	-2.26	0.0238
ChrChr9:DiagnosisControl	-0.168	0.05776	-2.908	0.003633
ChrChr7:DiagnosisControl	-0.2754	0.07526	-3.66	0.0002527
ChrChr10:DiagnosisControl	-0.02414	0.06291	-0.3838	0.7012
ChrChr2:DiagnosisControl	-0.1807	0.06207	-2.912	0.003592
ChrChr12:DiagnosisControl	-0.2696	0.07606	-3.545	0.0003933
ChrChr15:DiagnosisControl	-0.1998	0.07765	-2.573	0.01009
ChrChr18:DiagnosisControl	-0.266	0.08338	-3.191	0.001419
ChrChr17:DiagnosisControl	-0.2804	0.0818	-3.428	0.0006074
ChrChr4:DiagnosisControl	-0.2504	0.07214	-3.471	0.0005176
ChrChr14:DiagnosisControl	-0.2228	0.07565	-2.945	0.003233
ChrChr13:DiagnosisControl	-0.2393	0.0787	-3.04	0.002366
ChrChr3:DiagnosisControl	-0.2689	0.0706	-3.808	0.0001399
ChrChr11:DiagnosisControl	-0.3494	0.07557	-4.624	<0.0001
ChrChr20:DiagnosisControl	-0.2519	0.08048	-3.13	0.001747
ChrChr22:DiagnosisControl	-0.3252	0.08719	-3.73	0.0001918
ChrChr16:DiagnosisControl	-0.2471	0.08196	-3.014	0.002574
ChrChr19:DiagnosisControl	-0.2767	0.07909	-3.499	0.0004668
ChrChr21:DiagnosisControl	-0.2985	0.08741	-3.415	0.0006372

Table U.1: Table U (continued)

ChrChr8:DiagnosisControl	-0.1824	0.07902	-2.308	0.021
ChrChr6:DiagnosisControl	-0.2438	0.07513	-3.245	0.001174

Appendix V

THE EFFECTS OF CHROMOSOME, DIAGNOSIS, BRAIN REGIONS, COVERAGE AND SEX ON THE FREQUENCY OF CNVS IN THE UNCORRECTED DATA

Table V.1: The table shows the results of the GLMM that interaction of chromosome and diagnosis, brains regions, sex and coverage were added as fixed factors to predict the frequency of CNV.

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	0.5497	0.159	3.458	0.0005451
ChrChr5	-0.05698	0.04189	-1.36	0.1738
ChrChr9	0.2184	0.03907	5.59	<0.0001
ChrChr7	-0.6189	0.04919	-12.58	<0.0001
ChrChr10	-0.1922	0.04326	-4.443	<0.0001
ChrChr2	-0.02478	0.04177	-0.5931	0.5531
ChrChr12	-0.6451	0.04972	-12.98	<0.0001
ChrChr15	-0.8027	0.05156	-15.57	<0.0001
ChrChr18	-0.9557	0.05452	-17.53	<0.0001
ChrChr17	-0.8859	0.05328	-16.63	<0.0001
ChrChr4	-0.4701	0.04749	-9.898	<0.0001
ChrChr14	-0.684	0.04995	-13.69	<0.0001
ChrChr13	-0.7818	0.05166	-15.13	<0.0001
ChrChr3	-0.3766	0.04647	-8.105	<0.0001
ChrChr11	-0.5702	0.04886	-11.67	<0.0001
ChrChr20	-0.8581	0.05272	-16.28	<0.0001
ChrChr22	-1.079	0.05636	-19.15	<0.0001

Table V.1: Table V (continued)

ChrChr16	-0.921	0.0538	-17.12	<0.0001
ChrChr19	-0.8119	0.05173	-15.7	<0.0001
ChrChr21	-1.104	0.05676	-19.46	<0.0001
ChrChr8	-0.8159	0.05234	-15.59	<0.0001
ChrChr6	-0.6059	0.04948	-12.25	<0.0001
DiagnosisControl	0.2271	0.1752	1.296	0.1949
Brain_regionEntorhinal_Cortex	0.3163	0.02404	13.16	<0.0001
Brain_regionHippocampus_CA1	0.4514	0.02305	19.59	<0.0001
Brain_regionHippocampus_CA3	0.447	0.02389	18.71	<0.0001
Brain_regionTemporal_Cortex	-0.7557	0.03566	-21.19	<0.0001
SexMale	-0.1021	0.1704	-0.5993	0.549
Coverage	32.47	1.275	25.46	<0.0001
ChrChr5:Diagnosiscontrol	-0.1399	0.0619	-2.26	0.0238
ChrChr9:DiagnosisControl	-0.168	0.05776	-2.908	0.003634
ChrChr7:DiagnosisControl	-0.2754	0.07526	-3.66	0.0002526
ChrChr10:DiagnosisControl	-0.02414	0.06291	-0.3838	0.7012
ChrChr2:DiagnosisControl	-0.1807	0.06207	-2.912	0.003593
ChrChr12:DiagnosisControl	-0.2696	0.07606	-3.545	0.0003933
ChrChr15:DiagnosisControl	-0.1998	0.07765	-2.573	0.01009
ChrChr18:DiagnosisControl	-0.266	0.08338	-3.191	0.001418
ChrChr17:DiagnosisControl	-0.2804	0.0818	-3.428	0.0006074
ChrChr4:DiagnosisControl	-0.2504	0.07214	-3.472	0.0005174
ChrChr14:DiagnosisControl	-0.2228	0.07565	-2.945	0.003234
ChrChr13:DiagnosisControl	-0.2393	0.0787	-3.04	0.002366
ChrChr3:DiagnosisControl	-0.2689	0.0706	-3.809	0.0001397
ChrChr11:DiagnosisControl	-0.3494	0.07557	-4.624	<0.0001
ChrChr20:DiagnosisControl	-0.2519	0.08048	-3.13	0.001747
ChrChr22:DiagnosisControl	-0.3252	0.08719	-3.73	0.0001918
ChrChr16:DiagnosisControl	-0.2471	0.08196	-3.015	0.002574
ChrChr19:DiagnosisControl	-0.2767	0.07909	-3.499	0.0004669

Table V.1: Table V (continued)

ChrChr21:DiagnosisControl	-0.2985	0.08741	-3.415	0.0006373
ChrChr8:DiagnosisControl	-0.1824	0.07902	-2.308	0.021
ChrChr6:DiagnosisControl	-0.2438	0.07513	-3.245	0.001174

Appendix W

THE EFFECTS OF CHROMOSOME, DIAGNOSIS AND COVERAGE ON THE FREQUENCY OF CNVS IN THE UNCORRECTED DATA

Table W.1: The table shows the results of the GLMM that interaction of chromosome and diagnosis and coverage were added as fixed factors to predict the frequency of CNV.

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	0.2814	0.1196	2.353	0.01861
ChrChr5	-0.05783	0.04283	-1.35	0.177
ChrChr9	0.2243	0.03988	5.623	<0.0001
ChrChr7	-0.6209	0.05026	-12.35	<0.0001
ChrChr10	-0.1882	0.04414	-4.263	<0.0001
ChrChr2	-0.03356	0.04285	-0.7834	0.4334
ChrChr12	-0.6482	0.05083	-12.75	<0.0001
ChrChr15	-0.7892	0.05228	-15.09	<0.0001
ChrChr18	-0.9467	0.05535	-17.1	<0.0001
ChrChr17	-0.8778	0.05413	-16.22	<0.0001
ChrChr4	-0.4819	0.04876	-9.883	<0.0001
ChrChr14	-0.6788	0.05087	-13.34	<0.0001
ChrChr13	-0.7777	0.05262	-14.78	<0.0001
ChrChr3	-0.3968	0.04789	-8.286	<0.0001
ChrChr11	-0.577	0.05005	-11.53	<0.0001
ChrChr20	-0.847	0.05352	-15.83	<0.0001
ChrChr22	-1.058	0.05688	-18.61	<0.0001
ChrChr16	-0.9084	0.05455	-16.65	<0.0001

Table W.1: Table W (continued)

ChrChr19	-0.7976	0.05244	-15.21	<0.0001
ChrChr21	-1.084	0.05732	-18.91	<0.0001
ChrChr8	-0.8162	0.05338	-15.29	<0.0001
ChrChr6	-0.6156	0.05073	-12.14	<0.0001
DiagnosisControl	0.1803	0.1996	0.9032	0.3664
Coverage	14.83	1.23	12.06	<0.0001
ChrChr5:DiagnosisControl	-0.1269	0.0636	-1.995	0.04608
ChrChr9:DiagnosisControl	-0.1546	0.05927	-2.608	0.009118
ChrChr7:DiagnosisControl	-0.2704	0.07733	-3.497	0.0004713
ChrChr10:DiagnosisControl	-0.01381	0.06452	-0.2141	0.8305
ChrChr2:DiagnosisControl	-0.1701	0.06401	-2.658	0.007867
ChrChr12:DiagnosisControl	-0.2759	0.07847	-3.515	0.000439
ChrChr15:DiagnosisControl	-0.1897	0.07906	-2.399	0.01644
ChrChr18:DiagnosisControl	-0.2573	0.08501	-3.026	0.002475
ChrChr17:DiagnosisControl	-0.2791	0.0836	-3.339	0.0008412
ChrChr4:DiagnosisControl	-0.2614	0.07486	-3.491	0.0004803
ChrChr14:DiagnosisControl	-0.2118	0.0774	-2.737	0.006199
ChrChr13:DiagnosisControl	-0.2439	0.08084	-3.018	0.002548
ChrChr3:DiagnosisControl	-0.2715	0.07341	-3.698	0.0002172
ChrChr11:DiagnosisControl	-0.3515	0.07803	-4.505	<0.0001
ChrChr20:DiagnosisControl	-0.2509	0.08225	-3.051	0.002281
ChrChr22:DiagnosisControl	-0.3107	0.08819	-3.523	0.0004263
ChrChr16:DiagnosisControl	-0.2476	0.08366	-2.959	0.003086
ChrChr19:DiagnosisControl	-0.268	0.08053	-3.328	0.0008754
ChrChr21:DiagnosisControl	-0.2812	0.0884	-3.182	0.001465
ChrChr8:DiagnosisControl	-0.186	0.08127	-2.289	0.02209
ChrChr6:DiagnosisControl	-0.245	0.07768	-3.154	0.00161

Appendix X

THE EFFECTS OF CHROMOSOME, DIAGNOSIS, BRAIN REGIONS AND COVERAGE ON THE FREQUENCY OF CNVS IN THE UNCORRECTED DATA

Table X.1: The table shows the results of the GLMM that chromosome, diagnosis, brain regions and coverage were added as fixed factors to predict the frequency of CNV.

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	0.5696	0.1064	5.355	<0.0001
ChrChr5	-0.1202	0.03085	-3.895	<0.0001
ChrChr9	0.143	0.02879	4.969	<0.0001
ChrChr7	-0.7395	0.03717	-19.9	<0.0001
ChrChr10	-0.203	0.03144	-6.457	<0.0001
ChrChr2	-0.1058	0.0309	-3.425	0.0006144
ChrChr12	-0.7633	0.03757	-20.32	<0.0001
ChrChr15	-0.8916	0.03857	-23.12	<0.0001
ChrChr18	-1.073	0.04123	-26.02	<0.0001
ChrChr17	-1.008	0.04038	-24.97	<0.0001
ChrChr4	-0.5806	0.03572	-16.25	<0.0001
ChrChr14	-0.7827	0.03749	-20.88	<0.0001
ChrChr13	-0.8874	0.03893	-22.79	<0.0001
ChrChr3	-0.495	0.03498	-14.15	<0.0001
ChrChr11	-0.7212	0.03719	-19.39	<0.0001
ChrChr20	-0.9689	0.0398	-24.34	<0.0001
ChrChr22	-1.22	0.04301	-28.37	<0.0001

Table X.1: Table X (continued)

ChrChr16	-1.03	0.04057	-25.38	<0.0001
ChrChr19	-0.9331	0.03911	-23.86	<0.0001
ChrChr21	-1.234	0.04319	-28.58	<0.0001
ChrChr8	-0.8978	0.03919	-22.91	<0.0001
ChrChr6	-0.7137	0.0372	-19.18	<0.0001
DiagnosisControl	0.04001	0.1709	0.2341	0.8149
Brain_regionEntorhinal_Cortex	0.3157	0.02405	13.13	<0.0001
Brain_regionHippocampus_CA1	0.451	0.02305	19.56	<0.0001
Brain_regionHippocampus_CA3	0.4466	0.02389	18.69	<0.0001
Brain_regionTemporal_Cortex	-0.7574	0.03569	-21.22	<0.0001
Coverage	32.46	1.276	25.44	<0.0001

Appendix Y

THE EFFECTS OF CHROMOSOME, DIAGNOSIS, BRAIN REGIONS, COVERAGE AND SEX ON THE FREQUENCY OF CNVS IN THE UNCORRECTED DATA

Table Y.1: The table shows the results of the GLMM that chromosome, diagnosis, brain regions, sex and coverage were added as fixed factors to predict the frequency of CNV.

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	0.6398	0.1579	4.051	<0.0001
ChrChr5	-0.1202	0.03085	-3.895	<0.0001
ChrChr9	0.143	0.02879	4.968	<0.0001
ChrChr7	-0.7395	0.03717	-19.9	<0.0001
ChrChr10	-0.203	0.03144	-6.457	<0.0001
ChrChr2	-0.1058	0.0309	-3.425	0.0006143
ChrChr12	-0.7633	0.03757	-20.32	<0.0001
ChrChr15	-0.8916	0.03857	-23.12	<0.0001
ChrChr18	-1.073	0.04123	-26.02	<0.0001
ChrChr17	-1.008	0.04038	-24.97	<0.0001
ChrChr4	-0.5806	0.03572	-16.25	<0.0001
ChrChr14	-0.7827	0.03749	-20.88	<0.0001
ChrChr13	-0.8874	0.03893	-22.79	<0.0001
ChrChr3	-0.495	0.03498	-14.15	<0.0001
ChrChr11	-0.7212	0.03719	-19.39	<0.0001
ChrChr20	-0.9689	0.0398	-24.34	<0.0001
ChrChr22	-1.22	0.04301	-28.37	<0.0001

Table Y.1: Table Y (continued)

ChrChr16	-1.03	0.04057	-25.38	<0.0001
ChrChr19	-0.9331	0.03911	-23.86	<0.0001
ChrChr21	-1.234	0.04319	-28.58	<0.0001
ChrChr8	-0.8978	0.03919	-22.91	<0.0001
ChrChr6	-0.7137	0.0372	-19.18	<0.0001
DiagnosisControl	0.02804	0.1706	0.1644	0.8694
Brain_regionEntorhinal_Cortex	0.3157	0.02405	13.13	<0.0001
Brain_regionHippocampus_CA1	0.4509	0.02305	19.56	<0.0001
Brain_regionHippocampus_CA3	0.4466	0.02389	18.69	<0.0001
Brain_regionTemporal_Cortex	-0.7571	0.03569	-21.21	<0.0001
SexMale	-0.102	0.1706	-0.5979	0.5499
Coverage	32.45	1.276	25.44	<0.0001

Appendix Z

THE EFFECTS OF CHROMOSOME, DIAGNOSIS AND COVERAGE ON THE FREQUENCY OF CNVS IN THE UNCORRECTED DATA

Table Z.1: The table shows the results of the GLMM that chromosome, diagnosis and coverage were added as fixed factors to predict the frequency of CNV.

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	0.3672	0.1179	3.115	0.001841
ChrChr5	-0.1142	0.03168	-3.604	0.0003134
ChrChr9	0.156	0.02952	5.283	<0.0001
ChrChr7	-0.7376	0.03814	-19.34	<0.0001
ChrChr10	-0.1938	0.03224	-6.011	<0.0001
ChrChr2	-0.1087	0.03184	-3.414	0.000639
ChrChr12	-0.767	0.03866	-19.84	<0.0001
ChrChr15	-0.8724	0.03925	-22.23	<0.0001
ChrChr18	-1.058	0.042	-25.19	<0.0001
ChrChr17	-0.9979	0.04122	-24.21	<0.0001
ChrChr4	-0.5951	0.03696	-16.1	<0.0001
ChrChr14	-0.7714	0.03832	-20.13	<0.0001
ChrChr13	-0.8835	0.03991	-22.14	<0.0001
ChrChr3	-0.5143	0.03628	-14.17	<0.0001
ChrChr11	-0.7266	0.03832	-18.96	<0.0001
ChrChr16	-1.03	0.04057	-25.38	<0.0001
ChrChr19	-0.9331	0.03911	-23.86	<0.0001
ChrChr21	-1.234	0.04319	-28.58	<0.0001
ChrChr8	-0.8978	0.03919	-22.91	<0.0001

Table Z.1: Table Z (continued)

ChrChr19	-0.9134	0.0398	-22.95	<0.0001
ChrChr21	-1.205	0.04368	-27.59	<0.0001
ChrChr8	-0.8981	0.04023	-22.32	<0.0001
ChrChr6	-0.7221	0.03838	-18.81	<0.0001
DiagnosisControl	-0.01407	0.1951	-0.07212	0.9425
Coverage	14.8	1.23	12.03	<0.0001

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Turan, Zeliha Gözde

EDUCATION

Degree	Institution	Year of Graduation
MSc in Biology	Middle East Technical University	2016
BSc in Biology	Ege University	2011

PUBLICATIONS

Yildiz M, ..., **Turan ZG†** and Somel M†. Under the shadow: Old-biased genes are subject to weak purifying selection at both the tissue and cell type-specific levels. bioRxiv, 2022. †: Equal contributions

Turan ZG, ..., Somel M, Ueberham U. Somatic copy number variant load in neurons of healthy controls and Alzheimer's disease patients. Acta Neuropathologica Communications, 2022.

Turan ZG, Parvizi P, ..., Somel M. Molecular footprint of Medawar's mutation accumulation process in mammalian aging. Aging Cell, 2019.

PRESENTATIONS

Yildiz M, Izgi H, **Turan ZG***, Somel M*. The role of the Medawar's mutation accumulation theory in metazoan ageing in the light of cell type-specific transcrip-

tome changes. 8th Ecology and Evolutionary Biology Symposium. September 2022, Turkey. Poster presentation.

Turan ZG, Parvizi P, Donertas HM, Tung J, Khaitovich P, Somel M. A molecular test of the mutation accumulation theory of aging. 16th Congress of the European Society for Evolutionary Biology (ESEB). August 2017, Netherlands. Poster presentation.

Turan ZG, Parvizi P, Donertas HM, Tung J, Khaitovich P, Somel M. Molecular explanation for mammalian aging. 4th Ecology and Evolutionary Biology Symposium. July 2017, Turkey. Poster presentation.

Turan ZG, Parvizi P, Somel M. Notable decrease in transcriptome conservation during mammalian aging. Biology of Ageing Conference. October 2015, Singapore. Poster presentation.

Turan ZG, Parvizi P, Somel M. Notable decrease in transcriptome conservation during mammalian aging. Annual Meeting of the Society for Molecular Biology and Evolution (SMBE). July 2015, Austria. Poster presentation.

Turan ZG, Somel M. Notable decrease in transcriptome conservation during mammalian aging. Workshop on the Evolution and Ecology. February 2015, Turkey. Oral presentation.

LANGUAGES AND TOOLS

R, Python, Shell, Git, GitHub, LaTeX, MarkDown, Inkscape

COMMUNICATION

English (C1), Turkish (Mother tongue)