Exploring and genotyping *APOE* allele variants in RNA sequencing data for the study of amyotrophic lateral sclerosis

Qi Chang Lin
6101232

Utrecht University

Utrecht Bioinformatics Center

Karolinska Institutet

CMM
CENTER FOR MOLECULAR MEDICINE

Date:	26-08-2022
Supervisor:	Sebastian Lewandowski
Host:	Harris Lab, Department of Clinical Neuroscience, Karolinska Institutet, Sweden
Examiner:	Jan Veldink
University:	Utrecht University
Master’s programme:	Bioinformatics and Biocomplexity
Profile:	General Research Profile (Minor Research Project)

Exploring and genotyping *APOE* allele variants in RNA sequencing data for the study of amyotrophic lateral sclerosis

—

Qi Chang Lin

(6101232)

Table of Contents

Layman abstract – Dutch

Amyotrofe laterale sclerose (ALS) is de op twee na meest voorkomende neurodegeneratieve ziekte die voorkomt bij mensen. Een kenmerk van de ziekte is dat men niet zeker weet wat de algemene oorzaak is die de ziekte veroorzaakt. Wel weten we, na jaren aan onderzoek, dat er veel genen zijn gevonden die geassocieerd worden met de ziekte. Deze genen worden ook wel ALS genen genoemd. Veel van deze genen waren gevonden met betrekking tot zenuwen, associaties met andere celtypes werden vaak achterwege gelaten. In de laatste jaren komen juist de andere celtypes van het brein in de schijnwerpers. Dit patroon dat juist deze alternatieve celtypes te maken hebben met het ziektebeeld zien wij bijvoorbeeld terug in de ziekte van Alzheimer. Hier was een gen gevonden in microglia, een type afweercel in het brein, dat een groot risico geeft om de ziekte te krijgen als het een specifieke mutatie heeft. Dit gen heet *APOE* en kent verschillende vormen. Elk vorm van het gen wordt geclassificeerd door twee mutaties en er bestaan in totaal 3 vormen van het gen. Deze vormen worden ook wel genotypen genoemd.

In ons onderzoek wilden wij kijken of de verschillende genotypen van *APOE* een effect hebben op ALS. Dus wij wilden zien of het hebben van elk van deze genotypen de ziekte juist verergert of verbetert. Een tegenslag in ons onderzoek was echter dat veel datasets geen informatie bieden of bepaalde patiënten een bepaald genotype hebben. Wij hadden daarom zelf een manier ontwikkeld om te gebruiken op publiek toegankelijke RNA-data voor het bepalen van de genotypen van de patiënten. Hierbij lieten wij zien dat, mits de data van hoge kwaliteit is, we met een hoge zekerheid konden zeggen dat een patiënt ook een bepaald genotype heeft. Met onze manier bieden wij onderzoekers een alternatief aan om genotype informatie te genereren van alleen RNA-data, zonder dat zij extra DNA-data nodig hebben. Met de hoeveelheid RNA-data dat al publiekelijk te vinden is, biedt onze methode ook een manier aan om de zee van ongebruikte data op een extra wijze te analyseren. Uiteindelijk hopen wij met het werk dat wij doen de wetenschap te innoveren en ontdekkingen te versnellen, allemaal om de patiënt zo veel mogelijk van dienst te kunnen zijn.

Abstract

Amyotrophic lateral sclerosis (ALS) is the third most occurring neurodegenerative disease in humans and is defined by its heterogeneous clinical presentation. Over the years, many genes related to ALS were discovered by the rapid advancement of sequencing technology. Most of these studies were mainly focused on neuronal involvement. Only recently, non-neuronal cells joined the spotlight of the ALS field. In Alzheimer's disease (AD), the involvement of these non-neuronal cells had been well defined, with the genotype of the *APOE* gene being the biggest risk factor for AD onset. We propose that *APOE* genotype could perhaps also have an effect on ALS. However, *APOE* involvement in ALS is not well defined and thus most public datasets do not readily offer genotype information on this gene. Here we present our method of utilizing a well-established variant calling pipeline on publicly available RNA sequencing data to genotype *APOE* allele status. We observed that if proper quality control steps on publicly available data were taken, we could robustly genotype *APOE* providing that read depth around the region of interest was high enough. After validation with a pre-genotyped dataset, we found that our method managed to reach an accuracy of 0.9667 in determining the correct genotype. Our results show that even if there are no microarray or whole genome sequencing data available for genotyping, we can use RNA sequencing data alone to accurately genotype the polymorphic variants of our gene of interest. We could possibly open up many doors for researchers in the field who are limited to RNA sequencing and need genotype information. Adding to it, it could also contribute to researchers in the exploration of already available datasets and reuse data that otherwise would be less informative. With this, we ultimately hope to contribute to the acceleration of scientific discovery and indirectly help patients towards a better clinical outcome.

Introduction

Amyotrophic lateral sclerosis (ALS), also known as Lou Gehrig's disease or Charcot disease, is the third most prevalent neurodegenerative disease next to Alzheimer's disease (AD) and Parkinson disease (PD) (Vahsen et al., 2021). The disease is characterized by gradual loss of upper and lower motor neurons, leading to a diverse set of health complications and eventually death (Hardiman et al., 2017; Kim et al., 2020; Vahsen et al., 2021). Incidence and survival rates of the disease vary based on ancestral origin. Incidence of ALS around the world ranges from 0.7 to 3 per 100,000 individuals depending on the population (Hardiman et al., 2017). Median overall survival of patients with ALS is estimated to be around 24 to 48 months, with rare events of some patients having the disease for more than a decade before death (Turner et al., 2010).

Depending on the location of onset – bulbar-onset or limb-onset – the disease progresses differently. A third of the patients present the disease with bulbar-onset. Typical symptoms during the early stages of bulbar-onset ALS consist of dysphagia and dysarthria, followed by loss of motor functions in the limbs and body in the later stages of the disease. Limb-onset patients generally present the disease in a reversed manner. Several other symptoms typically presented with ALS include spasticity, hypersalivation, pain, muscle cramps, deep venous thrombosis, mood alterations, cognitive impairment and respiratory impairment. The latter is the main factor leading to death in most patients as a result of respiratory failure. Most of the manifestations can be treated symptomatically, however ALS itself remains without a cure to this day (Hardiman et al., 2017).

From the etiological perspective, ALS can be categorized into two different groups. The first group contains patients with familial ALS (fALS) and the second group are patients with sporadic ALS (sALS). The group of fALS patients only represents around 10% of the total amount of ALS cases, but nonetheless these patients played a big role in etiological studies on the disease. Most of the major causative genes have been found with the help of genetic studies in families with a history of fALS (Kim et al., 2020). Combined with the rise of advanced sequencing methods, the small list of genes associated with ALS has been expanded towards several dozens of disease-associated genes in the past decade (Abel et al., 2012; Lill et al., 2011; McCann et al., 2021). These genes include *SOD1*, *C9ORF72*, *TARDBP* (TDP-43), *FUS/TLS*, *OPTN*, *TBK1*, *GRN*, *NEK1* and *C21ORF2*. Although an abundance of disease-associated genes has been found, they are only found in a select number of cases of fALS and the etiology of sALS remains largely unknown (Kim et al., 2020).

Clinical and genetic differences in each patient are defining characteristics of ALS, marking ALS as a heterogeneous disease. Despite this heterogeneous classification, only neuronal involvement in the disease was broadly studied, while non-neuronal cells received less attention. Now, these cells are becoming more and more interesting to study as their involvement in ALS becomes increasingly elucidated. Especially the involvement of glia cells has been described for ALS, marking the relationship of the interconnected glia cells – microglia, astrocytes and oligodendrocytes – with neuronal damage mechanisms (Geloso et al., 2017; Madore et al., 2020; Trias et al., 2019; Vahsen et al., 2021). Also, involvement of perivascular fibroblasts has been shown to have an association with disease onset, highlighting the importance of vascular cells in ALS context (Månberg et al., 2021).

In contrast to ALS, the involvement of non-neuronal cells in AD has been described extensively. Particularly microglial effects on AD-progression have been observed in patients. Apolipoprotein E (*APOE*) is the main genetic culprit for this association (Belloy et al., 2019; Yamazaki et al., 2019). The presence of a particular allele of the *APOE* gene greatly increases or decreases the chance to develop AD (Montagne et al., 2021). *APOE* in humans can present itself with three different alleles: *APOE* ε2, *APOE* ε3 and *APOE* ε4. *APOE* ε3 is the most common allele amongst the three, followed by *APOE* ε4 and lastly *APOE* ε2. *APOE* ε1 does exist, however the occurrence of this allele has only been reported three times all around the world (Seripa et al., 2007). Due to its limited occurrence, the three other alleles are mostly the matter of discussion in literature when *APOE* is mentioned. The isoforms are defined by the single nucleotide polymorphisms (SNPs) rs429358 (T>C) and rs7412 (C>T). These polymorphisms result in changes for amino acid (AA) 112 and 158 of the protein, defining the different isoforms (Belloy et al., 2019). In AD context, having a single *APOE ε4* allele increases the risk of AD by 4-fold and a homozygous carrier has a 12-fold increase in risk of AD (Montagne et al., 2021). Furthermore, carrying the *APOE* ε4 allele also accelerates the timing of disease onset. In contrast, carrying the *APOE* ε2 allele seems to be having the opposite effects of carrying *APOE* ε4 in terms of developing AD (Belloy et al., 2019). *APOE* genotype dependency on the developmental risk of disease has also been shown in cardiovascular diseases (Bennet et al., 2007; Montagne et al., 2021) and even tumors (Ostendorf et al., 2020).

Here, we want to see whether we can observe an effect in ALS-context caused by differences in *APOE* genotype. To facilitate this, we turned to the public domain. One of the biggest available RNA sequencing (RNA-Seq) datasets representing ALS patient tissue transcriptomes was deposited by Tam et al. (Tam et al., 2019). The dataset (GEO: GSE124439) contains 176 samples of ALS patients, neurological controls and non-neurological controls. The only downside to the dataset was that the dataset does not contain any *APOE* genotyping information of the samples. To solve this issue, we tried to genotype *APOE* ourselves from the raw RNA-Seq data provided by the authors. We used the two defined SNPs for the classification of the *APOE* alleles to find the genotype of the samples. Furthermore, the dataset also presented itself with a gender bias, in which gender represented more variance in the gene expression data than the disease status. To solve this observed phenomenon, we assessed gene expression in the sex chromosomes. We found an indication that the discordances were possibly alignment errors, which could be easily solved with removal in subsequent analysis steps.

Ultimately, the goal was to determine *APOE* genotypes using a robust variant calling pipeline to find high confidence SNPs and validate the calls using a dataset where the *APOE* genotype is well defined.

Material and Methods

Software and Algorithms

The computations and data handling were enabled by resources in project SNIC 2022/22-143 provided by the Swedish National Infrastructure for Computing (SNIC) at UPPMAX.
Allocated resources: 10x 1000 core-h/month and 128 GB local (Crex) storage.
RStudio was executed locally on a MacBook Pro (14-inch, 2021), Apple M1 Pro, 16 GB with macOS Monterey.

Name	Version	Source	Identifier
DESeq2	1.34.0	Love et al., 2014	RRID:SCR_015687
GATK	4.2.0.0	Broad Institute	RRID:SCR_001876
GEOquery	2.62.2	Davis & Meltzer, 2007	RRID:SCR_000146
ggbeeswarm	0.6.0	Erik Clarke & Scott Sherril-Mix	https://cran.r-project.org/web/packages/ggbeeswarm/ vignettes/usageExamples.pdf
ggpubr	0.4.0	Alboukadel Kassambara	RRID:SCR_021139
how_are_we_stranded_here	1.0.1	Beth Signal	https://github.com/betsig/how_are_we_stranded_here/
HTSeq	0.12.4	Anders et al., 2014	RRID:SCR_005514
kallisto	0.44.0	Bray et al., 2016	RRID:SCR_016582
Picard	2.23.4	Broad Institute	RRID:SCR_006525
R	4.1.2	R Project	RRID:SCR_001905
R (UPPMAX)	4.1.1	R Project	RRID:SCR_001905
R_packages (UPPMAX)	4.1.1	UPPMAX	https://www.uppmax.uu.se/support/user-guides/r_packages-module-guide/
RSeQC	2.6.4	Wang et al., 2016	RRID:SCR_005275
rstatix	0.7.0	R Project	RRID:SCR_021240
RStudio	2021.09.1 - 372	R Studio	RRID:SCR_000432
samtools	1.14	Wellcome Sanger Institute	RRID:SCR_002105
STAR	2.7.9a	Alex Dobin	RRID:SCR_004463
Synapse Python Client	2.3.1	Synapse	https://pypi.org/project/synapseclient/
tidyverse	1.3.1	R Studio	RRID:SCR_019186
Trim Galore	0.6.1	Brabaham Institute	RRID:SCR_011847
VariantAnnotation	1.40.0	Obenchain et al., 2014	https://bioconductor.org/packages/release/bioc/ html/VariantAnnotation.html
viridis	0.6.2	https://github.com/ sjmgarnier/viridis/	RRID:SCR_016696

Data and scripts availability

Raw gene expression data generated by Tam et al. (Tam et al., 2019) was downloaded from GEO (GEO: GSE124439) on 07-03-2022. Raw sequencing files (Paired-end FASTQs) of

GSE124439 (n=176) were acquired through the European Nucleotide Archive (ENA) with project-ID PRJNA512012 (https://www.ebi.ac.uk/ena/browser/view/PRJNA512012) and downloaded through cURL using FTP-links provided by ENA on 23-03-2022. Raw sequencing files (Paired-end FASTQs) of ROSMAP batch 1 (n=120; Synapse ID = syn8612097) and batch 3 (n=100; Synapse ID = syn21589959) were acquired through Synapse with the Synapse Python Client from the cloud servers of Synapse on 10-04-2022 and 31-03-2022 respectively.

Scripts used in the study are deposited in the following GitHub-repository:
https://github.com/munytre/2022_KI_Lewandowski

Calculation of TPMs

The transcript lengths were acquired through Ensembl Biomart (Ensembl V105 and GRCh38.p13) and included the lengths of the UTRs of each individual gene. TPMs were calculated according to the following equation (Zhao et al., 2021):

(1)

$$TPM_i = \frac{\frac{q_i}{l_i}}{\Sigma_j \left(\frac{q_j}{l_j}\right)} * 10^6$$

$$where\ q_i = Reads\ mapped\ to\ gene,$$
$$l_i = Average\ transcript\ length\ of\ gene,$$
$$and\ \Sigma_j \left(\frac{q_j}{l_j}\right) = Sum\ of\ mapped\ reads\ to\ gene\ normalized\ by\ transcript\ length$$

Variant calling pipeline

Raw fastq files were downloaded prior to quality control, alignment, counting and variant calling. Sources from which the data were acquired differ depending on the dataset as described previously.

General resources

The following resources were used throughout the pipeline (Schematic overview in Supplementary 1 & 2):

FASTA

Human reference genome (GRCh38) made with all the unmasked autosomes (1-22), sex chromosome X and MT chromosome. Individual FASTAs were acquired from http://ftp.ensembl.org/pub/release-105/fasta/homo_sapiens/dna/ and concatenated into one FASTA with zcat.

Annotation file

Ensembl (V105) for Homo sapiens was acquired from http://ftp.ensembl.org/pub/release-105/gtf/homo_sapiens/Homo_sapiens.GRCh38.105.gtf.gz.

STAR Index

For each data set a new STAR index was made according to the read length of the reads found in the FASTQs following the *max readlength-1* rule as indicated by the authors of STAR.

- For PRJNA512012: --sjdbOverhang = 124
- For ROSMAP batch 1: --sjdbOverhang = 100

- For ROSMAP batch 3: --sjdbOverhang = 149

TrimGalore

The fastq files were first processed with TrimGalore to select high-quality reads. This wrapper script runs Cutadapt and FastQC, to remove possible sequencing adapters, filter low-quality bases, and obtains other statistics. HTML-files generated by FastQC in this step were used to further assess the quality of the data. Validated files resulting from TrimGalore were then used as input for STAR.

STAR

A full alignment of the samples to the human reference genome annotated with Ensembl (V105) was conducted with STAR (Dobin et al., 2013) using twopassMode Basic.

HTSeq-Count

We then proceeded to index the sorted BAMs with samtools prior to read counting. The tool of choice for read counting was HTSeq-Count from HTSeq. In the tool we set --*order* to pos, as the BAMs were sorted by coordinates, and --*stranded* to reverse, as the reads were RF/fr-firststrand stranded. The strandedness for the datasets were confirmed with how_are_stranded_here from https://github.com/betsig/how_are_we_stranded_here.

GATK

The BAM files were processed with a modified GATK pipeline for variant calling. We built the pipeline following the developer's instructions and implemented the modifications in the pipeline indicated next.

AddOrReplaceReadGroups

The first step in our GATK pipeline was to assign read group information to the sorted BAM, which is necessary for the base recalibration step later in the pipeline. We used the defaults setting from AddOrReplaceReadGroups from Picard for this step and the resulting BAM with assigned read group was subsequently indexed with samtools index.

MarkDuplicates

The indexed BAM with assigned read groups was then put through MarkDuplicates from Picard to remove duplicated reads. Compared to recommended settings to only mark the duplicated reads in the BAM, we removed the duplicated reads from the entire BAM.

SortSam

The output BAM from MarkDuplicates was sorted based on coordinates and indexed with Picard SortSam after removal of duplicated reads.

SplitNCigarReads

The sorted BAM was then used in SplitNCigarReads to remove Ns from the reads aligned to the reference. As RNA-Seq data has a lot of Ns indicating regions where introns are located, this could disturb up the variant call. To prevent problems arising from these gaps filled with Ns in the reads SplitNCigarReads was used to split reads with Ns in the cigar.

BaseRecalibrator & ApplyBQSR

The output BAM from SplitNCigarReads was used in BaseRecalibrator to generate a recalculation table. This recalculation table was then used to recalculate QC scores for all reads to remove errors resulting from systematic biases that could arise prior or during sequencing. The recalculation of the QC score in the BAM was done by ApplyBQSR. The VCF from dbSNP (V146, hg38) was used as the reference for known sites and was acquired from the GATK reference bundle located on UPPMAX. The naming of the chromosomes in the VCF was modified to accompany the Ensembl nomenclature of chromosome names by removing "chr". BaseRecalibrator ran two times, one prior ApplyBQSR and one after. This was done to assert the base recalibration was successful in the following step.

AnalyzeCovariates

The two recalibration tables from BaseRecalibrator (prior and after ApplyBQSR) were run through AnalyzeCovariate to generate a pdf to assess the quality of recalibration.

HaplotypeCaller

The output BAM from ApplyBQSR was used in HaplotypeCaller to do the actual variant calling. The variant call was limited to the region 19:44500000-45000000 to spare resources. *APOE* is located on chromosome 19:44905791-44909393 (1-based position), so running the entire genome through HaplotypeCaller would take several hours compared to the limited region. Furthermore, we did not use the recommended settings of --standard-min-confidence-threshold-for-calling = 20 to call RNA-Seq data. (The minimum phred-scaled confidence threshold at which variants should be called.) Instead, we used the tools default setting of 30.

VariantFiltration

At last, the generated VCF containing the calls were hard filtered with VariantFiltration to eliminate low quality calls. We used the same filtering criteria used in the WDL that was provided by GATK to make calls in RNA-Seq data (https://github.com/gatk-workflows/gatk4-RNA-Seq-germline-snps-indels/blob/master/gatk4-rna-best-practices.wdl), which restricted the filtering to FisherStrand (FS) and QualityByDepth (QD). This resulted in calls with FS > 30.0 and QD < 2.0 to be filtered out from the final VCF. Further analyses on the final files were done in R.

Results

Gender separation

Gene expression data (GEO: GSE124439) from Tam et al. (Tam et al., 2019) was inspected before we started genotyping the *APOE* gene from its raw RNA-Seq data. During initial inspection of the raw count table – GSE124439_RAW.tar – provided by the authors, the entire dataset (raw count matrix) was deemed unusable for our purposes due to gender effects as detailed next (Szczepińska & Lewandowski, unpublished).

As part of this study, we scrutinized this dataset. Instead of edgeR, used in the initial inspection, we used DESeq2 for the analysis. The raw count matrix generated by Tam et al. was processed through the wrapper function *DESeq()* from DESeq2, with experimental design = ~ sample_group_ch1. Each group in sample_group_ch1 represented disease status of the patient (i.e. ALS Spectrum; Non-Neurological Control; Other Neurological Controls). Hidden expression biases were studied with a PCA (Principle Component Analysis) using *plotPCA()* from DESeq2. In particular, the two components explaining most of the sample's variance in expression (PC1 and PC2) were used to assess the quality of the data. From this analysis, samples were separated on PC2 into two distinct clusters, each an admixed of different sample groups. Upon inspection, the two clusters separated distinctively with gender (Figure 1). When corrected for gender in the experimental design (~ sample_group_ch1 + Gender), the observed separation of clusters disappeared. This suggests that gender had more effect on the expression differences between samples than the disease context in terms of cases (ALS Spectrum) versus controls (Non-Neurological Control and Other Neurological Controls). We evaluated the contribution of genes in the X and Y chromosomes (i.e., sex-linked genes) in the observed expression variance (Supplementary 3). Upon removal of the Y-linked

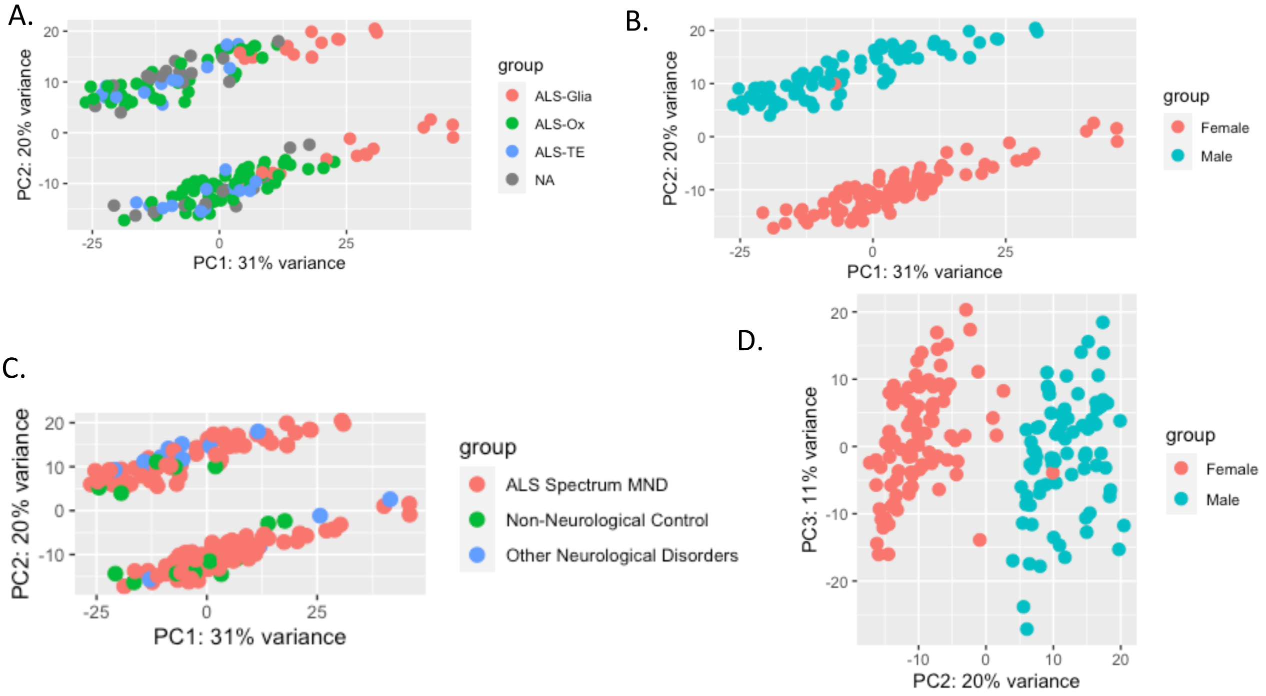


Figure 1. Principle component analyses generated from DESeq2 variance stabilized transformed counts from the full GSE124439 raw count table. A) PCA with PC1 and PC2, samples are colored by ALS classification determined by Tam et al. NAs are non-ALS samples. B) PCA with PC1 and PC2, samples are colored by gender. C) PCA with PC1 and PC2, samples are colored by sample group. D) PCA with PC2 and PC3, samples are colored by gender.

genes, the separation of samples by gender disappeared on the PC2 (13% variance), but remained in PC3 (6% variance, Figure 2). Nevertheless, when the X-linked genes were also removed, the separation by gender in the PC3 (5% of the variance) was also normalized (Figure 3).

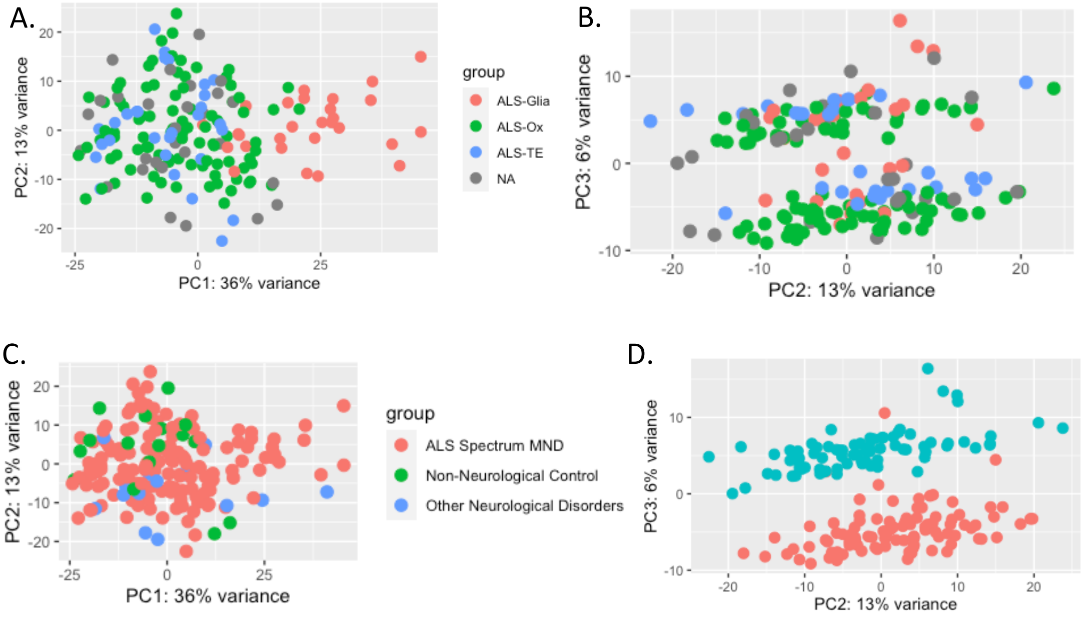


Figure 2. Principle component analyses generated from DESeq2 variance stabilized transformed counts from the full GSE124439 raw count table with Y-linked genes removed. A) PCA with PC1 and PC2, samples are colored by ALS classification Tam et al. NAs are non-ALS samples. B) PCA with PC2 and PC3, samples are colored by ALS classification. NAs are non-ALS samples. C) PCA with PC1 and PC2, samples are colored by sample group. D) PCA with PC2 and PC3, samples are colored by gender.

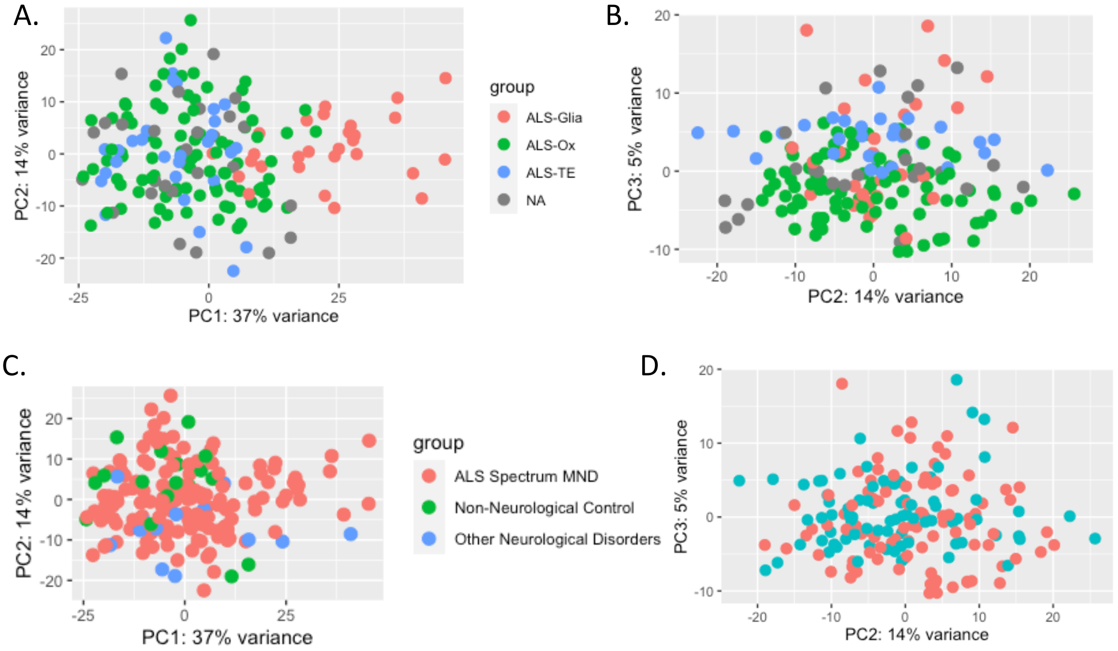


Figure 3. Principle component analyses generated from DESeq2 variance stabilized transformed counts from the full GSE124439 raw count table with both X-linked and Y-linked genes removed. A) PCA with PC1 and PC2, samples are colored by ALS classification Tam et al. NAs are non-ALS samples. B) PCA with PC2 and PC3, samples are colored by ALS classification. NAs are non-ALS samples. C) PCA with PC1 and PC2, samples are colored by sample group. D) PCA with PC2 and PC3, samples are colored by gender.

To determine the reason why gender contributed to a large percentage of the variance observed in the dataset, we proceeded to evaluate the gene counts for several sex-linked genes per sample for each sex.

The groups used for the different comparisons were:

Group 1 (Chromosome X): *AKAP17A, ASMT, ASMTL, CD99, CRLF2, CSF2RA, DHRSX, GTPBP6, IL3RA, IL9R, P2RY8, PLCXD1, PPP2R3B, SHOX, SLC25A6, VAMP7, WASH6P, ZBED1*
Group 2 (Chromosome Y): *AKAP17A, ASMT, ASMTL, CD99, CRLF2, CSF2RA, DHRSX, GTPBP6, IL3RA, IL9R, P2RY8, PLCXD1, PPP2R3B, SHOX, SLC25A6, VAMP7, WASH6P, ZBED1*
Group 3 (Chromosome X-specific with homologue shared by chromosome Y): *AMELX, DDX3X, EIF1AX, KDM5C, KDM6A, NLGN4X, PCDH11X, RPS4X, TBL1X, TGIF2LX, TMSB4X, USP9X, VCX3A, VCX3B, VCX, VCX2, ZFX*
Group 4 (Chromosome Y-specific with homologue shared by chromosome X): *AMELY, DDX3Y, EIF1AY, KDM5D, UTY, NLGN4Y, PCDH11Y, RPS4Y1, RPS4Y2, TBL1Y, TGIF2LY, TMSB4Y, USP9Y, VCY, VCY1B, ZFY*
Group 5 (Chromosome Y-specific): *BPY2, BPY2B, BPY2C, CDY1, CDY1B, CDY2A, CDY2B, DAZ1, DAZ2, DAZ3, DAZ4, HSFY1, HSFY2, PRY, PRY2, PRYP3, RBMY1A1, RBMY1B, RBMY1D, RBMY1E, RBMY1F, RBMY1J, SRY, TSPY1, TSPY10, TSPY2, TSPY3, TSPY4, TSPY8, TSPY9P*

Groups 1 and 2 contain genes in homologue regions of chromosomes X and Y, and because both sex chromosomes contain the coding region for these genes, reads mapping to them can originate in males from either or both chromosomes. Groups 3 and 4 represent genes that are paralogues to each other and originate from the different sex chromosomes. Group 5 contains genes that are Y-linked; thus, they can only originate from the Y-chromosome. Normalized counts (TPMs) (Wagner et al., 2012) were used to qualitatively compare the counts for these genes over the samples. Because we were missing the transcript length for each individual gene, and we did not know which specific transcript per gene was expressed, we used the average transcript length of all transcripts for each gene to calculate the TPMs.

The expression difference between genders was computed for each gene group and its significance determined using Benjamini-Hochberg corrected Wilcoxon and Welch's t-tests (Figure 4). Both statistical methods were used to account for possible differences in data distributions. Genes for groups 1 and 2 are presented as one group (i.e., Group 1) as both represent identical genes in homologue regions of the chromosomes X and Y, respectively. We observed for every group, except for the comparison of group 1 and 2, that there was a significant difference in TPMs for each gender (FDR<0.01). Unexpectedly, expression was detected for genes in groups 4 and 5 for female subjects. Both groups include genes exclusive to the Y-chromosome, and thus male specific. This discordance could be a result of the misalignment of reads due to the high similarity of sequences between genes of groups 3 and 4, paralogue to each other. For group 5 we also detected gene expression in female subjects, which we speculate could be a result of either misalignment or an error made prior to sequencing.

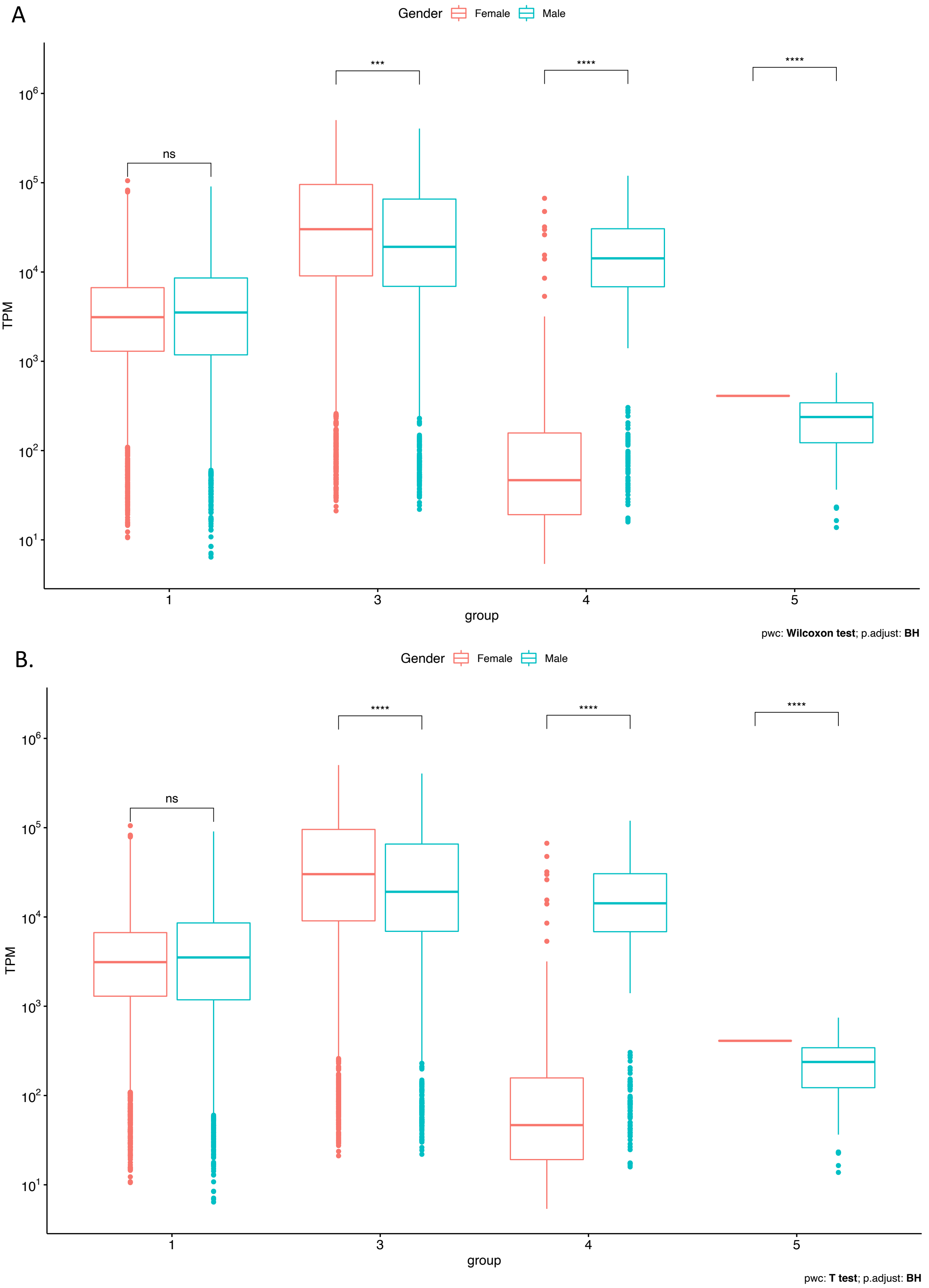


Figure 4. TPMs per gene group are represented in TPMs and separated by gender. Group 1 in the figure encompasses group 1 and 2 as described in an earlier section. A) Pairwise Wilcoxon tests corrected for multiple tests with Benjamini-Hochberg. B) Pairwise Welch's t-test corrected for multiple tests with Benjamini-Hochberg. TPMs of all groups, except the combined first group, are significant different (FDR<0.01) between both genders. In group 4 and 5, expression from Y-linked genes were detected among female patients, suggesting the occurrence of errors during alignment or preparation of the biological samples.

In the end, we made the decision to use the dataset for further analysis, although we found the discordances in the genes located on the sex-linked chromosomes. By excluding the sex-linked chromosomes in subsequent analyses, we ensured that the observed discordances will not affect the newly generated results.

Genotype assignment

Variant calling

After curing the dataset for further usage, we proceeded to infer the genotype of *APOE* from the raw RNA-Seq reported by Tam et al. (Supplementary 4). The two single nucleotide polymorphisms (SNPs) rs429358 (T>C) and rs7412 (C>T) were used to classify the different genotypes (Table 1). Genotype of *APOE* was assigned based on the found SNP(s) by a modified GATK variant calling pipeline (van der Auwera and O'Connor, 2020).

rs429358	**rs7412**		**Allele**		**To observe**
C	T		ε1		Both SNPs
T*	T		ε2		rs7412
T*	C*		ε3		NA
C	C*		ε4		rs429358

Table 1. Schematic overview of the SNPs to classify APOE into its different isoforms. The SNPs rs429358 (T>C) and rs7412 (C>T) define nucleotide changes that result in amino acid changes in *APOE*, resulting in different observed genotypes of the proteins. In AD, the ε4 allele is linked to increased risk of developing the disease. *Reference base for the SNP.

Validation

We assessed the accuracy of our genotyping pipeline by using RNA-Seq data and SNP data generated by The Religious Orders Study and Memory and Aging Project (ROSMAP) Study found on the AD Knowledge Portal (Bennett et al., 2018; Greenwood et al., 2020). All samples in ROSMAP were genotyped for *APOE* with an independent genotyping method where Agencourt Bioscience Corporation made use of high-throughput DNA sequencing to determine polymorphisms in codon 4 of *APOE* and its resulting genotype (de Jager et al., 2018). Thus, running these samples through our pipeline and comparing the predicted genotypes with the originally assigned genotypes provided the necessary validation for our own calls.

We randomly choose 100 and 120 samples from batch 3 and batch 1 respectively with a restriction on the number of selected samples per genotype and ran it through the pipeline. The latter batch contained more samples, as we wanted to include more samples with uncommon genotypes that were available in batch 1 in comparison to batch 3.

ROSMAP batch 3

We initially started with the selection of samples from ROSMAP batch 3 (Table 2 & 3; Supplementary 5). The reason to use this specific batch was that this batch was the only batch of the entire ROSMAP dataset with readily provided raw FASTQs. Other batches contained BAM files and to keep the pipeline consistent, the choice was made to keep the input to FASTQs. We proceeded with the selection of 100 random samples based on metadata provided by Synapse describing the *APOE* genotype of each patient and the availability of bulk

RNA-Seq data. We used the following files to make the selection: ROSMAP_assay_RNA-Seq_metadata.csv, ROSMAP_biospecimen_metadata.csv and ROSMAP_clinical.csv. All files can be requested from Synapse. We furthermore limited the tissue origin to dorsolateral prefrontal cortex to keep the end-results consistent.

Genotype	Amount
ε2ε3	34
ε2ε4	2
ε3ε3	160
ε3ε4	48
ε4ε4	8

Table 2. Total amount of available samples found on Synapse for each *APOE* genotype in ROSMAP batch 3.

Genotype	Amount
ε2ε3	18
ε2ε4	2
ε3ε3	40
ε3ε4	33
ε4ε4	7

Table 3. Total amount of samples per *APOE* genotype selected from ROSMAP batch 3 on Synapse.

After running the selected samples through the pipeline, we discovered that from the 100 samples SNPs have been detected in only 4 samples. We proceeded to look at the number of reads supporting each count and discovered that the value for each called variant was very low. The 4 samples RISK_226, RISK_253, RISK_17_rerun and RISK_390 had DP-values (Read Depth) of 2, 4, 5 and 40 respectively. Due to these low values, we proceeded to look at the counts generated by HTSeq-Count to see what the overall raw counts were for *APOE*. Overall, the number of raw counts for the *APOE* gene looked normal and ranged from 32 to 3878. It is noticeable that only a few samples had counts over 3000 and RISK_390 (raw count = 3129) with a variant call supported by 40 reads was one of them. In the end, due to the low amount of predicted genotype calls in ROSMAP batch 3, we explored ROSMAP batch 1. The latter had considerably higher detected raw counts for APOE when we looked at the raw count data generated by The RNA-Seq Harmonization Study (Synapse ID = syn21241740).

ROSMAP batch 1

After looking around the AD Knowledge Portal, we found that The RNA-Seq Harmonization Study regenerated raw FASTQs for ROSMAP batch 1 (Synapse ID = syn8612097). This meant that we could use ROSMAP batch 1 in our pipeline, instead of the initially provided BAMs (Synapse ID = syn4164376). Because batch 1 contained a lot more uncommon samples than batch 3, we selected for 120 samples to include in the validation set instead of 100 samples (Table 4 & 5; Supplementary 6). In this case, we also included all ε2ε2 samples in the pipeline as well. Sample selection was again random, but with the restriction on the number of samples per genotype. The selection was also based on the files described in the section for

ROSMAP batch 3. Limiting tissue selection was furthermore not necessary, as all samples in batch 1 originated from the dorsolateral prefrontal cortex.

Genotype	Amount
ε2ε2	5
ε2ε3	82
ε2ε4	16
ε3ε3	386
ε3ε4	140
ε4ε4	6

Table 4. Total amount of available samples found on Synapse for each *APOE* genotype in ROSMAP batch 1.

Genotype	Amount
ε2ε2	5
ε2ε3	30
ε2ε4	5
ε3ε3	40
ε3ε4	35
ε4ε4	5

Table 5. Total amount of samples per *APOE* genotype selected from ROSMAP batch 1 on Synapse.

Following our expectations, running batch 1 through the pipeline generated a substantial amount of called variants (Supplementary 7). In total, rs7412 was found in 38 samples and rs429358 was found in 45 samples. From all 120 samples, 4 samples were predicted incorrectly (Table 6). The resulting accuracy of our genotyping efforts using RNA-Seq data was calculated to be 0.9667.

Type	Amount
True Positives (TP)	76
False Positives (FP)	2
True Negatives (TN)	40
False Negatives (FN)	2

Table 6. Overview of predictions of ROSMAP batch 1 samples according to our variant calling pipeline. In where TP = Number of samples with called SNPs that agree with the genotype, FP = Number of samples with called SNPs that do not agree with the genotype, TN = Number of samples with no called SNPs that have a wild-type genotype and FN = Number of samples with no called SNPs that do not have a wild-type genotype.

Comparison of datasets

To find out the reasons why ROSMAP batch 3 was ineffective as validation set, we proceeded with a comparison of the three datasets (PRJNA512012, ROSMAP batch 1 and ROSMAP batch 3) on multiple levels (Figure 5). We saw from the RNA-Seq data that ROSMAP batch 3 had an overall lower amount of uniquely mapped reads, coverage and reads aligned to *APOE*. This indicated that ROSMAP batch 3 was of lower sequencing quality than the other two datasets and therefore could have impacted the variant calling. This was further confirmed when we compared the read depth supporting the SNPs called (rs7412 and rs429358) in *APOE* for each of the three datasets, where ROSMAP batch 3 had substantial lower read depth for the calls made compared to the other two datasets (Figure 6).

Reproduction runs

To verify the technical reproducibility of our results, we ran 2 random samples from each dataset through the pipeline for a second time (Table 7). This was done for PRJNA512012 and ROSMAP batch 1. ROSMAP batch 3 was excluded from the reproduction runs, as the batch did not detect a substantial number of variants. Thus, to save resources, the decision was made to only reproduce the results of the other two datasets.

Dataset	Sample	Output files of second run
PRJNA512012	SRR8375325	Identical output
	SRR8375392	Identical output
ROSMAP batch 1	331_120501	Identical output
	583_120522	Identical output

Table 7. Chosen samples for reproduction runs and the result of the second runs.

The results from HTSeq-Count and GATK from each of the samples were compared for the different runs. The count-file and the filtered VCF from GATK were loaded into R for each sample and compared using *setequal()* from dplyr (tidyverse). To see, if samples from the same dataset were equal due to technical errors, the samples were also cross compared. At the end, the reproduced results of each sample from each dataset were identical.

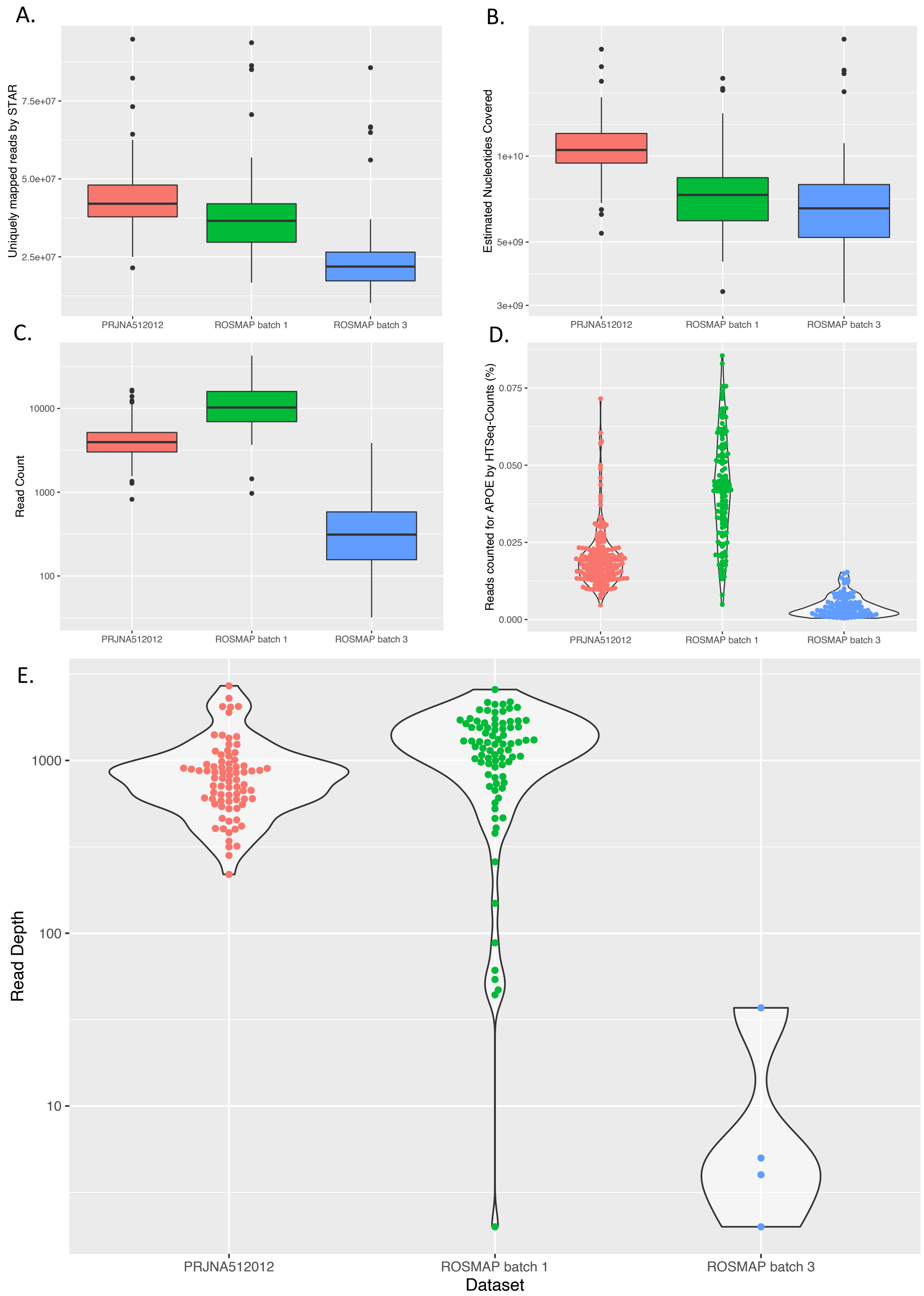


Figure 5. Comparison of PRJNA512012, ROSMAP batch 3 and ROSMAP batch 1. A) Uniquely mapped reads, B) Estimated Nucleotides Covered (n_{Reads} * Read Length), C) Raw read counts generated by HTSeq-Count for *APOE*, D) Percentage of reads aligned to protein coding genes counted for *APOE* by HTSeq-Count for each sample are plotted per dataset (PRJNA512012, ROSMAP batch 3 and ROSMAP batch 1). E) Comparison of read depth for variants called in APOE (rs7412 and rs429358) across the different datasets.

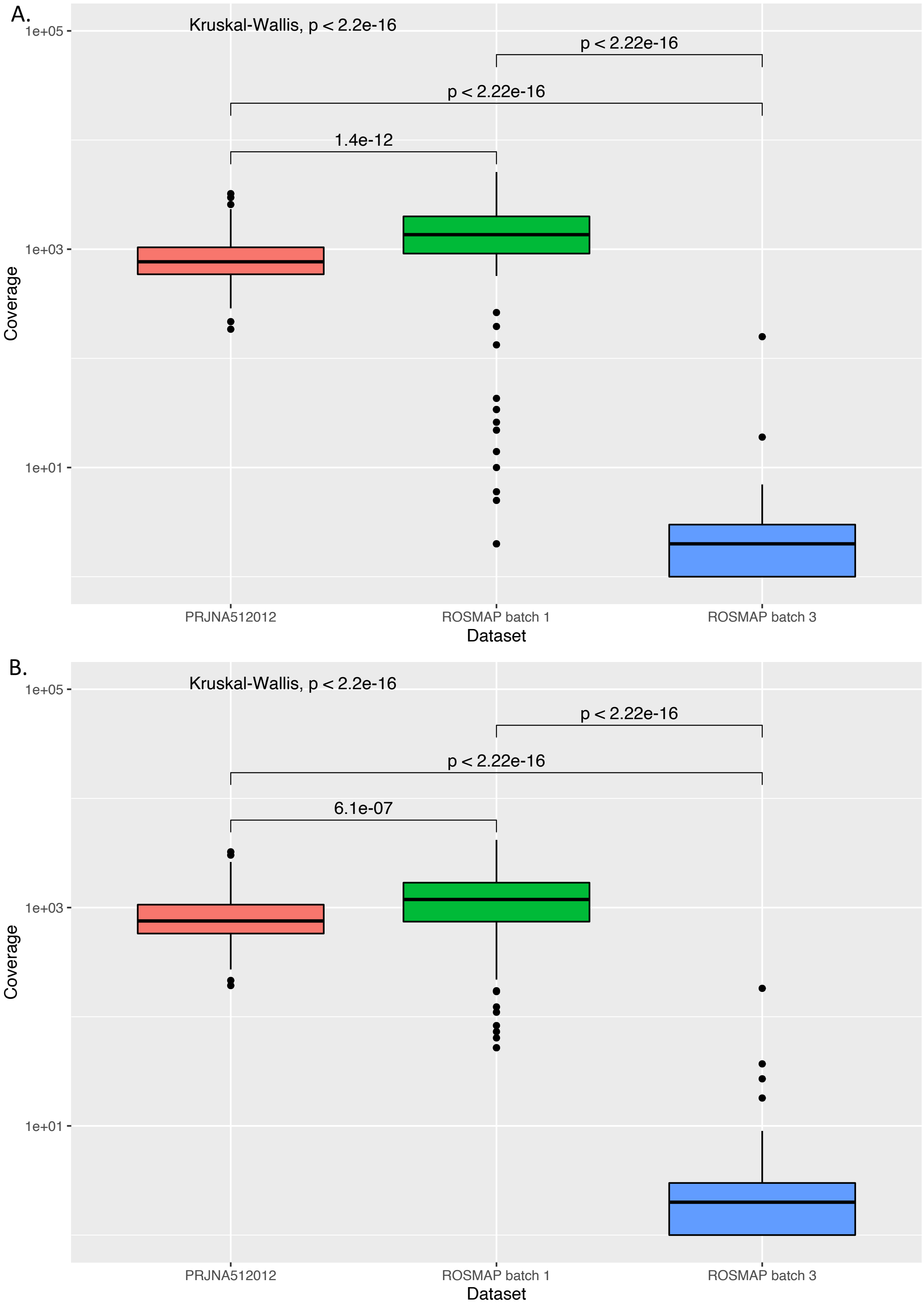


Figure 6. Comparison of PRJNA512012, ROSMAP batch 3 and ROSMAP batch 1 – Coverage of SNP locations. Coverage of SNP location was assessed after read duplication removal for all three datasets with samtools mpileup. Coverage (read depth) of the different datasets are compared with Kruskal-Wallis test. ROSMAP 1 = ROSMAP batch; ROSMAP3 = ROSMAP batch 3. A) Coverage for chr19: 44908822 (rs7412). B) Coverage for chr19: 44908684 (rs429358). All chromosomal locations shown here are 1-based.

Discussion

After decades of research on ALS, we start to understand the disease better. From the anatomical level up down to the molecular level, we start to connect the dots leading to novel treatments and discoveries on the neurodegenerative disease. Particularly on the molecular level, the field has progressed a lot with the developments in sequencing technology. More and more genetic alterations are being linked to cause ALS, in which multiple novel genes are classified ALS-genes (Kim et al., 2020). With the steady decrease in sequencing cost over the years many groups have acquired a substantial amount of sequencing data, particularly RNA-Seq data and its resulting gene expression data. Most of the gene expression data is used to just answer strictly gene expression related questions. However, the biological function can also be affected by specific allele variants as exemplified by *APOE* contribution to Alzheimer's disease and vascular dysfunction (Belloy et al., 2019). Therefore, we aimed to re-genotype the *APOE* variants for future genotype stratification of gene expression in publicly available ALS datasets.

Using the biggest publicly available ALS RNA-Seq dataset (Tam et al., 2019), we managed to demonstrate the utility of *APOE* allele variant calls from raw RNA-Seq data. We showed with our results that data that was available without the *APOE* genotype, could be re-genotyped from RNA-Seq data if proper steps regarding quality control are taken before actual analysis of the data. Initially, the gender effects that were observed in gene expression were a substantial source of variance in principle component analysis. However, with multiple PCAs we showed that gender effects on the variability of the data could be normalized if we do not include gene expression data from the sex chromosomes. Although the number of coding genes in the sex chromosomes are not high in numbers and removal of the effect caused by gender can be done in other ways, we still made the decision to remove the sex chromosomal genes. Gene expression that was biologically unlikely to be found on the sex chromosomes, where Y-linked genes were found to be expressed in females (Figure 4), gave us reasons to believe that the removal of genes originating from the sex chromosomes is necessary if we wanted to fully utilize the dataset for further analyses. Even though we did not further investigate why we found these biologically unlikely gene expression values, we believe that the observations could be a cause of technical issues. We propose that these could either be originating from the sequencer itself or the algorithms that were used to align the sequences. In addition, the sex chromosomes have the same evolutionary origin and thus share a high similarity (Ross et al., 2005; Webster et al., 2019). The many homozygous regions, which include gene encoding regions, could easily introduce problems for alignment algorithms (Olney et al., 2020).

It is worth to mention that removal of the sex chromosomes in the alignment step is not recommended. As stated by the authors of STAR, forced misalignment of reads could happen when a suitable reference is missing (Dobin et al., 2013). Thus, instead of removing all sex chromosomal genes at the alignment step, we made the choice to remove it only during the analysis steps. The X-chromosome was included in the reference genome for the STAR aligner. Contrasting this, the Y-chromosome was excluded from the reference genome. One of the reasons is that the highly homologue regions of both sex chromosomes are the regions that are most robustly assembled, the other regions are variable, repetitive and difficult to resolve (Kuderna et al., 2019). Thus, leaving out the Y-chromosome and using the X chromosome to prevent possible forced misalignment of reads to the reference genome, would provide us with an alignment of higher quality. This is with the assumption that reads originating from the Y-linked genes are not disturbing the alignments of reads on the autosomes and disturbing gene expression rates of the genes located there.

On another note, we showed that if raw RNA sequencing data was of high quality, we can accurately determine the genotype of a gene of interest by assessing present SNPs in the data. Usually, genotyping of a gene is done with either microarrays, whole genome sequencing (WGS) or whole exome sequencing (WES). These techniques have the advantage of having a high coverage at the regions of interest, thereby providing high quality SNP calls if the certain base differs from the reference. RNA sequencing has the downside of only providing reads at locations in the genome where transcription actually happens (Jehl et al., 2021; Sims et al., 2014). Consequently, this means that if a gene of interest is not expressed and thus not transcribed, we cannot find SNPs at all due to the low coverage of that gene.

To improve our genotyping efforts in future perspective, we need to find an optimum of parameters to use in the GATK pipeline. It is unavoidable that coverage rates in different samples of our particular gene of interest will differ. We need to find a general threshold of the read depth that allows us to make SNP calls of high accuracy. One possible way to do this is by using an already genotyped dataset and test different threshold regarding read depth of the region of interest. We can artificially reduce read depth of a particular region, for which a module exists in GATK (DownsampleSam from Picard), through multiple iterations and find a suitable threshold in which SNPs are accurately called. To test the threshold found, we can utilize several different datasets and validate the robustness of the threshold set.

With the found genotypes, we can assess differential expressed genes between the genotypes and other characteristics related to the allele of interest. In particular, for *APOE*, we can look at datasets without readily available *APOE* genotyping in a different context than Alzheimer's disease. For our interest, this method provides us with extra analytical options to assess the effects of *APOE* genotype on ALS disease variability. As survival risk could be a parameter to associate the *APOE* genotype with, we could consider doing survival analysis with the acquired information in the future. However, due to the limitations of our genotyping method, the downside of the mentioned survival analysis will be that the results are an indication of the observed effect than an actual observation of causation. A more robust approach needs to be developed if we want to see a significant association effect.

Furthermore, although the validation dataset (ROSMAP batch 1) had a similar level of coverage of the SNPs, one could argue that the disease origin of the dataset may not be the best comparison material for our purpose. As ROSMAP is originally a study on Alzheimer's disease and the dataset we are interested in, PRJNA512012, is focused on ALS. To rectify this and have a formal validation endeavor, we could use a reference ALS dataset where RNA-Seq and WGS were performed. To see if our pipeline produces the same unbiased result, we can perform the SNP calling and genotyping on the raw RNA-Seq data and compare it with the results of the WGS called SNPs and genotyping. To further increase the validity of the unbiased approach, we can also increase the number of samples in the validation set to e.g., an arbitrary number of 5x the amount of test samples. However, the type of disease should not be a matter of concern towards the variant calls and the subsequent genotyping. As we were only assessing the accuracy of our variant calls, a validation set with a comparable tissue source and coverage was needed to minimize variability of possible biological effects. ROSMAP fulfilled these criteria, as it has a known allele status and a similar tissue source.

To conclude, we managed to use public RNA-Seq data for *APOE* genotyping and to reach high accuracy in calling the actual SNPs defining *APOE* genotypes. Further investigations are needed to validate our approach of genotyping via RNA-Seq data. This approach could be promising, considering the robustness of variant call accuracy achieved in this project. If we manage to find a suitable threshold to accurate determine appropriate read depth levels of our gene of interest, we can try to apply our genotyping method to a vast number of RNA-Seq data available in the public domain. In essence, our approach could reach past the limits that exist on public data and open up novel possibilities that could contribute to our understanding of human biology.

Glossary of abbreviations

AD	Alzheimer’s Disease
ALS	Amyotrophic Lateral Sclerosis
APOE	Apolipoprotein E
DP	Read Depth
fALS	Familial Amyotrophic Lateral Sclerosis
GATK	Genome Analysis Toolkit
PD	Parkinson Disease
RNA-Seq	RNA Sequencing
ROSMAP	The Religious Orders Study and Memory and Aging Project
sALS	Sporadic Amyotrophic Lateral Sclerosis
SNP	Single Nucleotide Polymorphism
WGS	Whole Genome Sequencing
WES	Whole Exome Sequencing

Acknowledgements

The results published here are in whole or in part based on data obtained from the AD Knowledge Portal.

In addition, I want to thank all members of the Harris lab[1] for having me in the lab, especially Sebastian Lewandowski[1] for the amazing support and supervision during my stay at the Karolinska Institutet. Furthermore, I want to thank my collaborator Oscar C Bedoya Reina[2] for his time and help during my project, as well as Anna Szczepińska[1] and Roman Andreev[4] for their input in the project. At last, I want to thank Jan Veldink[5] for his time and effort to act as my examiner from my home university.

[1] Department of Clinical Neuroscience, Karolinska Institutet, Centre for Molecular Medicine, Karolinska Hospital, Stockholm, Sweden.
[2] Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Stockholm, Sweden.
[4] Eidgenössische Technische Hochschule Zürich, Zürich, Switzerland.
[5] UMC Utrecht Brain Center, University Medical Center Utrecht, Department of Neurology, Utrecht University, Utrecht, the Netherlands.

References

Abel, O., Powell, J.F., Andersen, P.M., Al-Chalabi, A., 2012. ALSoD: A user-friendly online bioinformatics tool for amyotrophic lateral sclerosis genetics. Human Mutation 33, 1345–1351. https://doi.org/10.1002/humu.22157

Belloy, M.E., Napolioni, V., Greicius, M.D., 2019. A Quarter Century of APOE and Alzheimer's Disease: Progress to Date and the Path Forward. Neuron. https://doi.org/10.1016/j.neuron.2019.01.056

Bennet, A.M., di Angelantonio, E., Ye, Z., Wensley, F., Dahlin, A., Ahlbom, A., Keavney, B., Collins, R., Wiman, B., de Faire, U., Danesh, J., 2007. Association of Apolipoprotein E Genotypes With Lipid Levels and Coronary Risk. JAMA 298, 1300. https://doi.org/10.1001/jama.298.11.1300

Bennett, D.A., Buchman, A.S., Boyle, P.A., Barnes, L.L., Wilson, R.S., Schneider, J.A., 2018. Religious Orders Study and Rush Memory and Aging Project. Journal of Alzheimer's Disease 64, S161–S189. https://doi.org/10.3233/JAD-179939

de Jager, P.L., Ma, Y., McCabe, C., Xu, J., Vardarajan, B.N., Felsky, D., Klein, H.U., White, C.C., Peters, M.A., Lodgson, B., Nejad, P., Tang, A., Mangravite, L.M., Yu, L., Gaiteri, C., Mostafavi, S., Schneider, J.A., Bennett, D.A., 2018. Data descriptor: A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. Scientific Data 5. https://doi.org/10.1038/sdata.2018.142

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21. https://doi.org/10.1093/bioinformatics/bts635

Geloso, M.C., Corvino, V., Marchese, E., Serrano, A., Michetti, F., D'Ambrosi, N., 2017. The dual role of microglia in ALS: Mechanisms and therapeutic approaches. Frontiers in Aging Neuroscience. https://doi.org/10.3389/fnagi.2017.00242

Greenwood, A.K., Montgomery, K.S., Kauer, N., Woo, K.H., Leanza, Z.J., Poehlman, W.L., Gockley, J., Sieberts, S.K., Bradic, L., Logsdon, B.A., Peters, M.A., Omberg, L., Mangravite, L.M., 2020. The AD Knowledge Portal: A Repository for Multi-Omic Data on Alzheimer's Disease and Aging. Current Protocols in Human Genetics 108. https://doi.org/10.1002/cphg.105

Hardiman, O., Al-Chalabi, A., Chio, A., Corr, E.M., Logroscino, G., Robberecht, W., Shaw, P.J., Simmons, Z., van den Berg, L.H., 2017. Amyotrophic lateral sclerosis. Nature Reviews Disease Primers. https://doi.org/10.1038/nrdp.2017.71

Jehl, F., Degalez, F., Bernard, M., Lecerf, F., Lagoutte, L., Désert, C., Coulée, M., Bouchez, O., Leroux, S., Abasht, B., Tixier-Boichard, M., Bed'hom, B., Burlot, T., Gourichon, D., Bardou, P., Acloque, H., Foissac, S., Djebali, S., Giuffra, E., Zerjal, T., Pitel, F., Klopp, C., Lagarrigue, S., 2021. RNA-Seq Data for Reliable SNP Detection and Genotype Calling: Interest for Coding Variant Characterization and Cis-Regulation Analysis by Allele-Specific Expression in Livestock Species. Frontiers in Genetics 12. https://doi.org/10.3389/fgene.2021.655707

Kim, G., Gautier, O., Tassoni-Tsuchida, E., Ma, X.R., Gitler, A.D., 2020. ALS Genetics: Gains, Losses, and Implications for Future Therapies. Neuron. https://doi.org/10.1016/j.neuron.2020.08.022

Kuderna, L.F.K., Lizano, E., Julià, E., Gomez-Garrido, J., Serres-Armero, A., Kuhlwilm, M., Alandes, R.A., Alvarez-Estape, M., Juan, D., Simon, H., Alioto, T., Gut, M., Gut, I., Schierup, M.H., Fornas, O., Marques-Bonet, T., 2019. Selective single molecule sequencing and assembly of a human Y chromosome of African origin. Nature Communications 10. https://doi.org/10.1038/s41467-018-07885-5

Lill, C.M., Abel, O., Bertram, L., Al-Chalabi, A., 2011. Keeping up with genetic discoveries in amyotrophic lateral sclerosis: The ALSoD and ALSGene databases. Amyotrophic Lateral Sclerosis 12, 238–249. https://doi.org/10.3109/17482968.2011.584629

Madore, C., Yin, Z., Leibowitz, J., Butovsky, O., 2020. Microglia, Lifestyle Stress, and Neurodegeneration. Immunity. https://doi.org/10.1016/j.immuni.2019.12.003

Månberg, A., Skene, N., Sanders, F., Trusohamn, M., Remnestål, J., Szczepińska, A., Aksoylu, I.S., Lönnerberg, P., Ebarasi, L., Wouters, S., Lehmann, M., Olofsson, J., von Gohren Antequera, I., Domaniku, A., de Schaepdryver, M., de Vocht, J., Poesen, K., Uhlén, M., Anink, J., Mijnsbergen, C., Vergunst-Bosch, H., Hübers, A., Kläppe, U., Rodriguez-Vieitez, E., Gilthorpe, J.D., Hedlund, E., Harris, R.A., Aronica, E., van Damme, P., Ludolph, A., Veldink, J., Ingre, C., Nilsson, P., Lewandowski, S.A., 2021. Altered perivascular fibroblast activity precedes ALS disease onset. Nature Medicine 27, 640–646. https://doi.org/10.1038/s41591-021-01295-9

McCann, E.P., Henden, L., Fifita, J.A., Zhang, K.Y., Grima, N., Bauer, D.C., Chan Moi Fat, S., Twine, N.A., Pamphlett, R., Kiernan, M.C., Rowe, D.B., Williams, K.L., Blair, I.P., 2021. Evidence for polygenic and oligogenic basis of Australian sporadic amyotrophic lateral sclerosis. Journal of Medical Genetics 58, 87–95. https://doi.org/10.1136/jmedgenet-2020-106866

Montagne, A., Nikolakopoulou, A.M., Huuskonen, M.T., Sagare, A.P., Lawson, E.J., Lazic, D., Rege, S. v., Grond, A., Zuniga, E., Barnes, S.R., Prince, J., Sagare, M., Hsu, C.-J., LaDu, M.J., Jacobs, R.E., Zlokovic, B. v., 2021. APOE4 accelerates advanced-stage vascular and neurodegenerative disorder in old Alzheimer's mice via cyclophilin A independently of amyloid-β. Nature Aging 1, 506–520. https://doi.org/10.1038/s43587-021-00073-z

Olney, K.C., Brotman, S.M., Andrews, J.P., Valverde-Vesling, V.A., Wilson, M.A., 2020. Reference genome and transcriptome informed by the sex chromosome complement of the sample increase ability to detect sex differences in gene expression from RNA-Seq data. Biology of Sex Differences 11. https://doi.org/10.1186/s13293-020-00312-9

Ostendorf, B.N., Bilanovic, J., Adaku, N., Tafreshian, K.N., Tavora, B., Vaughan, R.D., Tavazoie, S.F., 2020. Common germline variants of the human APOE gene modulate melanoma progression and survival. Nature Medicine 26, 1048–1053. https://doi.org/10.1038/s41591-020-0879-3

Ross, M.T., Grafham, D. v., Coffey, A.J., Scherer, S., McLay, K., Muzny, D., Platzer, M., Howell, G.R., Burrows, C., Bird, C.P., Frankish, A., Lovell, F.L., Howe, K.L., Ashurst, J.L., Fulton, R.S., Sudbrak, R., Wen, G., Jones, M.C., Hurles, M.E., Andrews, T.D., Scott, C.E., Searle, S., Ramser, J., Whittaker, A., Deadman, R., Carter, N.P., Hunt, S.E., Chen, R., Cree, A., Gunaratne, P., Havlak, P., Hodgson, A., Metzker, M.L., Richards, S., Scott, G., Steffen, D., Sodergren, E., Wheeler, D.A., Worley, K.C., Ainscough, R., Ambrose, K.D., Ansari-Lari, M.A., Aradhya, S., Ashwell, R.I.S., Babbage, A.K., Bagguley, C.L., Ballabio, A., Banerjee, R., Barker, G.E., Barlow, K.F., Barrett, I.P., Bates, K.N., Beare, D.M., Beasley, H., Beasley, O., Beck, A., Bethel, G., Blechschmidt, K., Brady, N., Bray-Allen, S., Bridgeman, A.M., Brown, A.J., Brown, M.J., Bonnin, D., Bruford, E.A., Buhay, C., Burch, P., Burford, D., Burgess, J., Burrill, W., Burton, J., Bye, J.M., Carder, C., Carrel, L., Chako, J., Chapman, J.C., Chavez, D., Chen, E., Chen, G., Chen, Y., Chen, Z., Chinault, C., Ciccodicola, A., Clark, S.Y., Clarke, G., Clee, C.M., Clegg, S., Clerc-Blankenburg, K., Clifford, K., Cobley, V., Cole, C.G., Conquer, J.S., Corby, N., Connor, R.E., David, R., Davies, J., Davis, C., Davis, J., Delgado, O., DeShazo, D., Dhami, P., Ding, Y., Dinh, H., Dodsworth, S., Draper, H., Dugan-Rocha, S., Dunham, A., Dunn, M., Durbin, K.J., Dutta, I., Eades, T., Ellwood, M., Emery-Cohen, A., Errington, H., Evans, K.L., Faulkner, L., Francis, F., Frankland, J., Fraser, A.E., Galgoczy, P., Gilbert, J., Gill, R., Glöckner, G., Gregory, S.G., Gribble, S., Griffiths, C., Grocock, R., Gu, Y., Gwilliam, R., Hamilton, C., Hart, E.A., Hawes, A., Heath, P.D., Heitmann, K., Hennig, S., Hernandez, J., Hinzmann, B., Ho, S., Hoffs, M., Howden, P.J., Huckle, E.J., Hume, J., Hunt, P.J., Hunt, A.R., Isherwood, J., Jacob, L., Johnson, D., Jones, S., de Jong, P.J., Joseph, S.S., Keenan, S., Kelly, S., Kershaw, J.K., Khan, Z., Kioschis, P., Klages, S., Knights, A.J., Kosiura, A., Kovar-Smith, C., Laird, G.K., Langford, C., Lawlor, S., Leversha, M., Lewis, L., Liu, W., Lloyd, C., Lloyd, D.M., Loulseged, H., Loveland, J.E., Lovell, J.D., Lozado, R., Lu, J., Lyne, R., Ma, J., Maheshwari, M., Matthews, L.H., McDowall, J., McLaren, S., McMurray, A., Meidl, P., Meitinger, T., Milne, S., Miner, G., Mistry, S.L., Morgan, M., Morris, S., Müller, I., Mullikin, J.C., Nguyen, N., Nordsiek, G., Nyakatura, G., O'Dell, C.N., Okwuonu, G., Palmer, S., Pandian, R., Parker, D., Parrish, J., Pasternak, S., Patel, D., Pearce, A. v., Pearson, D.M., Pelan, S.E., Perez, L., Porter, K.M., Ramsey, Y., Reichwald, K., Rhodes, S., Ridler, K.A., Schlessinger, D., Schueler, M.G., Sehra, H.K., Shaw-Smith, C., Shen, H., Sheridan, E.M., Shownkeen, R., Skuce, C.D., Smith, M.L., Sotheran, E.C., Steingruber, H.E., Steward, C.A., Storey, R., Swann, R.M., Swarbreck, D., Tabor, P.E., Taudien, S., Taylor, T., Teague, B., Thomas, K., Thorpe, A., Timms, K., Tracey, A., Trevanion, S., Tromans, A.C., d'Urso, M., Verduzco, D., Villasana, D., Waldron, L., Wall, M., Wang, Q., Warren, J., Warry, G.L., Wei, X., West, A., Whitehead, S.L., Whiteley, M.N., Wilkinson, J.E., Willey, D.L., Williams, G., Williams, L., Williamson, A., Williamson, H., Wilming, L., Woodmansey, R.L., Wray, P.W., Yen, J., Zhang, J., Zhou, J., Zoghbi, H., Zorilla, S., Buck, D., Reinhardt, R., Poustka, A., Rosenthal, A., Lehrach, H., Meindl, A., Minx, P.J., Hillier, L.W., Willard, H.F., Wilson, R.K., Waterston, R.H., Rice, C.M., Vaudin, M., Coulson, A., Nelson, D.L., Weinstock, G., Sulston, J.E., Durbin, R., Hubbard, T., Gibbs, R.A., Beck, S., Rogers, J., Bentley, D.R., 2005. The DNA sequence of the human X chromosome. Nature 434, 325–337. https://doi.org/10.1038/nature03440

Seripa, D., Matera, M.G., Daniele, A., Bizzarro, A., Rinaldi, M., Gravina, C., Bisceglia, L., Corbo, R.M., Panza, F., Solfrizzi, V., Fazio, V.M., Forno, G.D., Masullo, C., Dallapiccola, B., Pilotto, A., 2007. The missing ApoE allele. Annals of Human Genetics 71, 496–500. https://doi.org/10.1111/j.1469-1809.2006.00344.x

Sims, D., Sudbery, I., Ilott, N.E., Heger, A., Ponting, C.P., 2014. Sequencing depth and coverage: Key considerations in genomic analyses. Nature Reviews Genetics. https://doi.org/10.1038/nrg3642

Tam, O.H., Rozhkov, N. v., Shaw, R., Kim, D., Hubbard, I., Fennessey, S., Propp, N., Phatnani, H., Kwan, J., Sareen, D., Broach, J.R., Simmons, Z., Arcila-Londono, X., Lee, E.B., van Deerlin, V.M., Shneider, N.A., Fraenkel, E., Ostrow, L.W., Baas, F., Zaitlen, N., Berry, J.D., Malaspina, A., Fratta, P., Cox, G.A., Thompson, L.M., Finkbeiner, S., Dardiotis, E., Miller, T.M., Chandran, S., Pal, S., Hornstein, E., MacGowan, D.J., Heiman-Patterson, T., Hammell, M.G., Patsopoulos, N.A., Butovsky, O., Dubnau, Joshua, Nath, A., Bowser, R., Harms, M., Aronica, E., Poss, M., Phillips-Cremins, J., Crary, J., Atassi, N., Lange, D.J., Adams, D.J., Stefanis, L., Gotkine, M., Baloh, R., Babu, S., Raj, T., Paganoni, S., Shalem, O., Smith, C., Zhang, B., Harris, B.T., Fagegaltier, D., Ravits, J., Dubnau, Josh, Gale Hammell, M., 2019. Postmortem Cortex Samples Identify Distinct Molecular Subtypes of ALS: Retrotransposon Activation, Oxidative Stress, and Activated Glia. Cell Reports 29, 1164-1177.e5. https://doi.org/10.1016/j.celrep.2019.09.066

Trias, E., Beilby, P.R., Kovacs, M., Ibarburu, S., Varela, V., Barreto-Núñez, R., Bradford, S.C., Beckman, J.S., Barbeito, L., 2019. Emergence of microglia bearing senescence markers during paralysis progression in a rat model of inherited ALS. Frontiers in Aging Neuroscience 10. https://doi.org/10.3389/fnagi.2019.00042

Turner, M.R., Brockington, A., Scaber, J., Hollinger, H., Marsden, R., Shaw, P.J., Talbot, K., 2010. Pattern of spread and prognosis in lower limb-onset ALS. Amyotrophic Lateral Sclerosis 11, 369–373. https://doi.org/10.3109/17482960903420140

Vahsen, B.F., Gray, E., Thompson, A.G., Ansorge, O., Anthony, D.C., Cowley, S.A., Talbot, K., Turner, M.R., 2021. Non-neuronal cells in amyotrophic lateral sclerosis — from pathogenesis to biomarkers. Nature Reviews Neurology. https://doi.org/10.1038/s41582-021-00487-8

van der Auwera, G., O'Connor, B., 2020. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra, 1st ed. O'Reilly Media, Inc.

Wagner, G.P., Kin, K., Lynch, V.J., 2012. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory in Biosciences 131, 281–285. https://doi.org/10.1007/s12064-012-0162-3

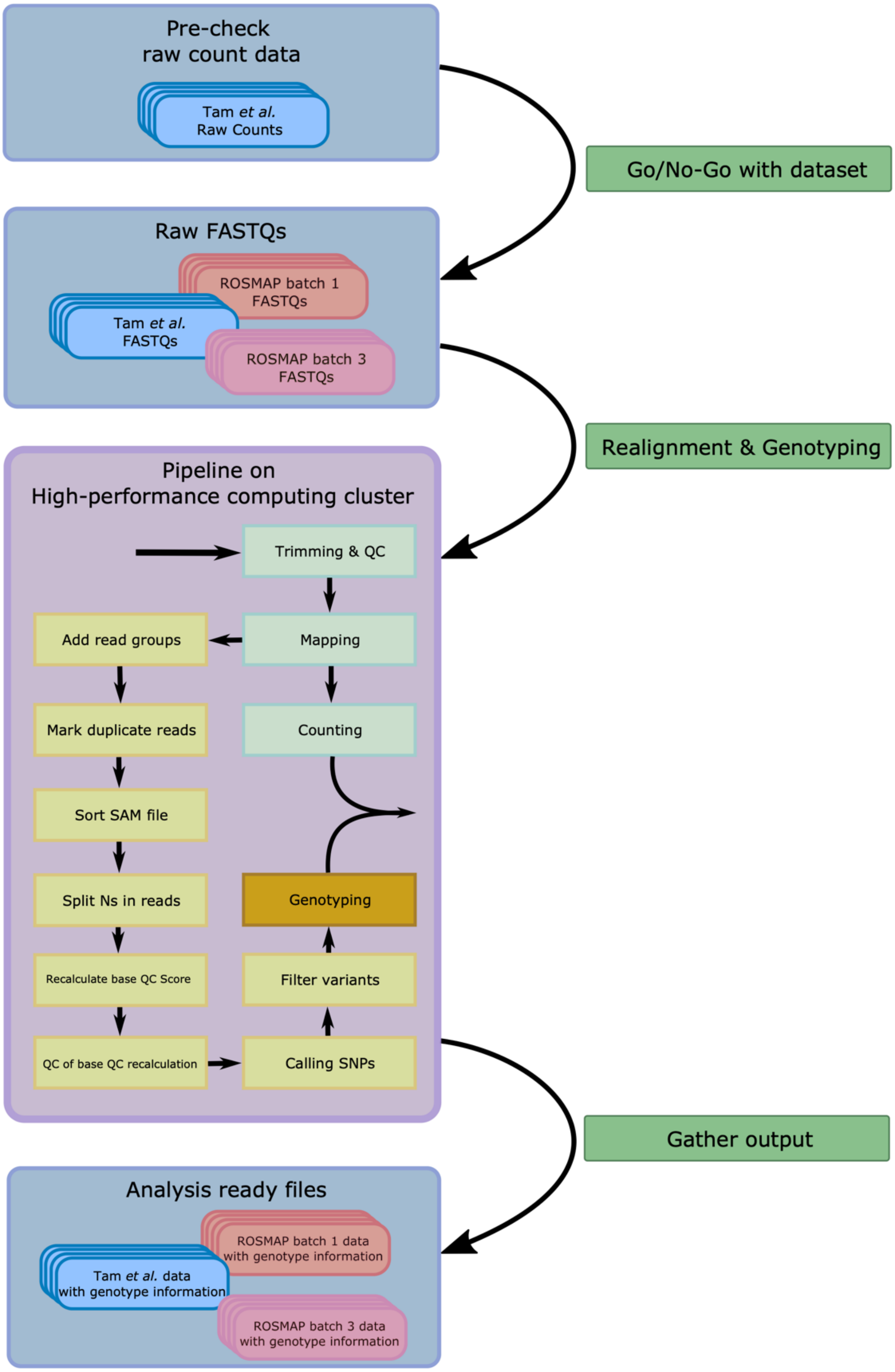
Webster, T.H., Couse, M., Grande, B.M., Karlins, E., Phung, T.N., Richmond, P.A., Whitford, W., Wilson, M.A., 2019. Identifying, understanding, and correcting technical artifacts on the sex chromosomes in next-generation sequencing data. Gigascience 8. https://doi.org/10.1093/gigascience/giz074

Yamazaki, Y., Zhao, N., Caulfield, T.R., Liu, C.C., Bu, G., 2019. Apolipoprotein E and Alzheimer disease: pathobiology and targeting strategies. Nature Reviews Neurology. https://doi.org/10.1038/s41582-019-0228-7

Zhao, Y., Li, M.C., Konaté, M.M., Chen, L., Das, B., Karlovich, C., Williams, P.M., Evrard, Y.A., Doroshow, J.H., McShane, L.M., 2021. TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. Journal of Translational Medicine 19. https://doi.org/10.1186/s12967-021-02936-w

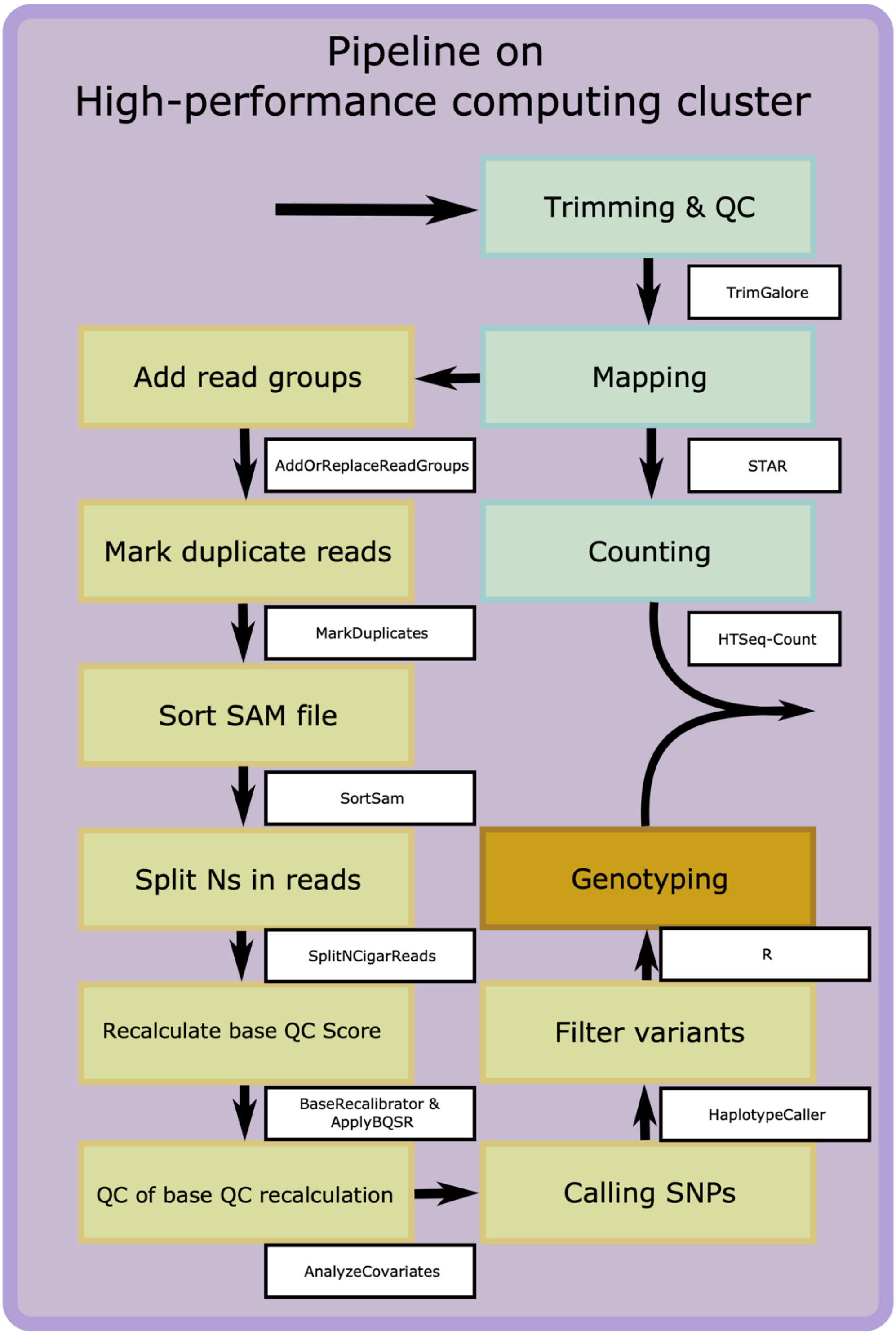
Supplementary

S1. Schematic overview of genotyping pipeline



Raw counts generated by Tam *et al.* were evaluated prior to realignment and genotyping of the raw RNA-Seq data. FASTQs were used as input for the general re-alignment and genotyping pipeline and executed on a high-performance computing cluster. The output files were gathered afterwards for future analyses. A detailed schematic of the realignment and genotyping pipeline is provided in Supplementary 2.

S2. Schematic overview of genotyping pipeline with tools



The pipeline starts with the trimming and quality control of the raw FASTQs of raw RNA-Seq data. After mapping, BAMs are split into two different modules. The green boxes indicate the main module responsible for realignment of the transcriptomics data and counting gene expression. The yellow boxes indicate the modified GATK pipeline used to call SNPs from the realigned reads. Following the SNP detection, we assigned genotype based on found SNPs with a dedicated R-script.

S3. List used to remove genes originating from the sex chromosomes in raw count data

For the version used in the analysis of raw count data obtained from GSE124439 see:
https://github.com/munytre/APOE_supplementary/blob/main/XY_genes.txt
For the version used in the analysis of the re-aligned data from PRJNA512012 see:
https://github.com/munytre/APOE_supplementary/blob/main/XY_genes_ENSEMBL.txt

S4. Genotyped samples of PRJNA512012

The list of 176 samples from PRJNA512012 with their respective genotype predictions can be found on:
https://github.com/munytre/APOE_supplementary/blob/main/PRJNA512012_genotyped.txt

S5. Chosen samples of ROSMAP batch 3

The 100 samples chosen randomly from ROSMAP batch 3 can be accessed via:
https://github.com/munytre/APOE_supplementary/blob/main/ROSMAP_batch_3_samples.txt

S6. Chosen samples of ROSMAP batch 1

The 120 samples chosen randomly from ROSMAP batch 1 can be accessed via:
https://github.com/munytre/APOE_supplementary/blob/main/ROSMAP_batch_1_samples.txt

S7. Genotyped samples of ROSMAP batch 1

The list of 120 samples from ROSMAP batch 1 with their respective genotype prediction and real genotype can be found on:
https://github.com/munytre/APOE_supplementary/blob/main/ROSMAP_batch_1_genotyped.txt