A genome-wide association study of maternal genetic effects in autism spectrum disorder

Vučinić, Kim

UNIVERSITY OF ZAGREB
FACULTY OF SCIENCE
DEPARTMENT OF BIOLOGY

KIM VUČINIĆ

A GENOME-WIDE ASSOCIATION STUDY OF MATERNAL GENETIC EFFECTS IN AUTISM SPECTRUM DISORDER

Master Thesis

Dublin, 2021.

SVEUČILIŠTE U ZAGREBU
PRIRODOSLOVNO-MATEMATIČKI FAKULTET
BIOLOŠKI ODSJEK

KIM VUČINIĆ

ANALIZA UTJECAJA MAJČINOG GENOTIPA NA RAZVOJ POREMEĆAJA IZ SPEKTRA AUTIZMA KOD DJECE METODOM GENOMSKE ASOCIJACIJE

Diplomski rad

Dublin, 2021.

This thesis was created in the Shields group at University College Dublin, under the supervision of Professor Denis Shields, co-supervision of doctoral student Catherine Mahoney, and Professor Kristijan Vlahoviček. The thesis is submitted for grading to the Department of Biology at the Faculty of Science, University of Zagreb, with the aim of obtaining a Master's degree in molecular biology.

This thesis is dedicated to my family. My mother, who always believed in me, especially when I did not, and our dog Tèba, who was always up for cuddles. I wish you were both here to celebrate this moment.

I am very grateful to Denis, who gave me an opportunity to work with him and Catherine. Thank you for giving me freedom in conducting this study, helping me when I was stuck, and immense support in this crazy year.

Another important person for this thesis was Catherine with whom I shared struggles involving the demonic dataset (I guess every dataset is demonic in its own ways). Thank you for all your help, especially for explaining everything regarding statistics that puzzled me.

I would like to thank professor Kristian Vlahoviček for easy communication and quick solving of problems relating to tedious bureaucracy.

There are no words to express my gratitude for having my sister from another mother, Annamaria, in my life. Good thing that I do not have to, she is the one that understands! Thank you.

Juraj, thank you for accepting me the way I am. If you endured me during the lockdown and this thesis, nothing can stop you. Also, your family is great and I would like to thank them for all the help and for accepting me as part of your family.

Everybody reading this thesis should be thankful to Annamaria and Juraj for making this thesis grammatically correct and more readable. However, some mistakes are due to happen, after all, they are just humans.

I am very grateful to all my family and friends for supporting me during these years. To my grandma and her life stories, to Kristijan - my baking buddy (no bread was made during writing of this thesis), to Jeff and his photos of the cutest pets, to Klara and Iva - our little group of mutual support, to Simon for all the chats and in the end printing this thesis, to Stefan with all the suggestions, your quirky stickers and teaching me how to cook (trying to), to Anastazija for our daily chats, and the list goes on. Forgive me if I did not mention you, but this is already too long.

At last but not least, I would like to thank all my mentors, professors and colleagues for the last six years.

BASIC DOCUMENTATION CARD

University of Zagreb
Faculty of Science
Department of Biology

Master Thesis

A Genome-wide Association Study Of Maternal Genetic Effects In Autism Spectrum Disorder

Kim Vučinić
Division of Molecular Biology, Horvatovac 102A, 10 000 Zagreb, Croatia

Around 2% of children in the world are diagnosed with autism spectrum disorder (ASD), a family of conditions influenced by both genetic and environmental factors. Findings suggest that maternal genetic composition affects the development of ASD in offspring. The subject matter is underresearched and the replication of findings is problematic. The aim of this study was to identify potential candidate loci in mothers that might increase the risk of development of ASD in offspring. A case-control genome-wide association study was performed on the SPARK dataset (n = 27,290) that consists of quad, trio and duo families where one or multiple individuals can be affected. Affected children's mothers were used as cases and affected children's fathers as controls. Maternal genetic effects were also modelled with log-linear models proposed by Weinberg (1999). In addition, permutation tests were used to assess the significance of the Bayes factor that was used to rank SNPs according to their degree of association. The results indicate involvement of maternal genetic effect in the offspring's development. The finding of variants associated with increased risk of developing ASD is important for understanding the aetiology of ASD, which could lead to better personalized medicine in the future.

(40 pages, 8 figures, 7 tables, 97 references, original in English)
Thesis deposited in the Central Biological Library
Keywords: GWAS, case-control study, ASD, maternal genetic effects, Bayesian statistics, permutations
Supervisor: Professor Denis Shields
Assistant Supervisor: Professor Kristian Vlahoviček
Reviewers: Professor Kristian Vlahoviček
Asst. Prof. Sofia Ana Blažević
Asst. Prof. Damjan Franjević
Substitution: Asst. Prof. Rosa Karlić
Thesis accepted:

TEMELJNA DOKUMENTACIJSKA KARTICA

Sveučilište u Zagreb
Prirodoslovno matematički fakultet
Biološki odsjek Diplomski rad

Analiza utjecaja majčinog genotipa na razvoj poremećaja iz spektra autizma kod djece metodom genomske asocijacije

Kim Vučinić
Zavod za molekularnu biologiju, Horvatovac 102A, 10 000 Zagreb, Hrvatska

Otprilike 2% djece u svijetu je dijagnosticirano s poremećajem iz spektra autizma, neurorazvojnim poremećajem koji ima genetske i okolišne uzroke. Istraživanja ukazuju na mogućnost utjecaja majčinog genotipa na razvoj autizma kod djeteta. Problematika je nedovoljno istražena te je replikacija potencijalnih kandidata u većini slučajeva neuspjela. Cilj ovog diplomskog rada je bio identificirati potencijalne genske markere u majkama koji povećavaju rizik razvoja poremećaja iz spektra autizma kod djece. U ovom radu, genomska analiza je provedena na skupu podataka SPARK (n = 27 290) koji se sastoji od četveročlanih, tročlanih i dvočlanih obitelji gdje je jedan ili više pojedinaca dijagnosticirano s poremećajem iz spektra autizma. Eksperiment je dizajniran tako da su majke slučajevi, a očevi kontrole u genomskoj analizi. Utjecaj majčinog genotipa je modeliran Weinberg log-linearnim modelom. Permutacijom je određena značajnost genetičkih markera koji su rangirani ovisno o jačini asocijacije po Bayesovom faktoru. Rezultati ukazuju na utjecaj majčinog genotipa na razvoj poremećaja kod djece. Istraživanje genetičkih markera asociranih s povećanim rizikom za razvoj poremećaja iz spektra autizma je bitno za bolje razumijevanje etiologije autizma, što može poboljšati personaliziranu medicinu u budućnosti.

(40 stranica, 8 slika, 7 tablica, 97 literaturnih navoda, jezik izvornika: engleski)
Rad je pohranjen u Središnjoj biološkoj knjižnici
Ključne riječi: GWAS, case-control study , ASD, utjecaj majčinog genotipa, Bayes statistika, permutacije
Voditelj: prof. dr. sc. Denis Shields
Neposredni voditelj: prof. dr. sc. Kristian Vlahoviček
Ocjenitelji: prof. dr. sc. Kristian Vlahoviček
doc. dr. sc. Sofia Ana Blažević
izv. prof. dr. sc. Damjan Franjević
Zamjena: doc. dr. sc. Rosa Karlić
Rad prihvaćen:

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ABF	approximate Bayes factor
ASD	autism spectrum disorder
BF	Bayes factor
FWER	family-wise error rate
FDR	false discovery rate
IBD	identity by descent
LD	Linkage disequilibrium
LRT	likelihood ratio test
PCA	principal component analysis
PO	prior odds
QC	quality control
QQ	quantile-quantile
RR	relative risk
SNP	single nucleotide polymorphism
SPARK	Simons Foundation Powering Autism Research for Knowledge
TDT	transmission disequilibrium testing
USA	United States of America
WGS	whole genome sequencing

1 INTRODUCTION

Each human being is unique, yet the whole population shows a spectrum of traits that contribute to human individuality. The percentage of shared DNA between human individuals is 99.99%. Nevertheless, there is sufficient genetic variation in 0.01% to determine the susceptibility of having a specific phenotype, such as a disease or a trait like eye colour.

1.1 HUMAN GENETICS

1.1.1 *The significance of a change*

Many types of genetic variation exist but they are usually divided into point mutations, indels and structural variants depending on the length of affected DNA. Genetic variations can occur in the coding or non-coding region of a gene or the intergenic regions. Most variations do not have biological significance because a small percentage of DNA in the human genome are coding regions that translate into proteins. Moreover, a change in the coding region can result in unchanged amino acid sequence, or a changed amino acid sequence but unchanged protein function.

Having said that, if it is biologically significant, a change in the DNA can cause a severe impact depending on the location. On the molecular level, an alteration of amino acid sequence can cause a truncated or nonfunctional protein. Moreover, mutations in non-protein coding regions can cause numerous errors like incorrect splicing, altered mRNA stability, degradation of mRNA, changed affinity of transcription factor binding to the protein and many other errors.

1.1.2 *Rare versus complex diseases*

Most of the aforementioned genetic variations can lead to disease on an individual level. Rare diseases, such as cystic fibrosis, are traced back to rare genetic variants, or often called mutations, in a single gene. The disease is expressed regardless of the environment because of the strong effect of a mutated gene. The underlying biological mechanisms of rare diseases are mostly well studied and the heritability is explained with Mendel's laws.

However, the majority of diseases can not be explained with genetic mutations in one gene. Common diseases such as heart disease, cancer, diabetes and psychiatric

disorders, which have a high incidence in developed countries, have interaction of numerous genes and environmental factors. Additionally, unlike in the case of rare diseases, there is no clear segregation of the phenotype in complex diseases because it does not follow Mendel's laws.

1.1.3 *Genetic effects*

Trying to unravel the mechanism behind a complex disease can be challenging as there are other underlying mechanisms in play. As stated by Buyske (Buyske, 2008), observed associations in cases can be confounded by a maternal genetic effect instead of being a genetic effect of offspring. Other underlying mechanisms should be considered as well, such as maternal and paternal imprinting, maternal-fetal interactions, and parental indirect effects. The terminology can be ambiguous, but the following sections will explain some of the mentioned phenomena in more detail.

1.1.3.1 *Parental genetic effects*

The parental genetic effect, in some papers also known as transgenerational effects (Tsang et al., 2013; Connolly and Heron, 2015; Connolly, Anney, et al., 2017), is an indirect effect where a parent's genotype affects child's phenotype. Parent's possession of a certain allele can cause increased or decreased risk of developing a disease, regardless if a child has said allele variant or not. One of the schoolbook examples is a coiling direction of snail shells in species *Lymnaea peregra*. The right coiling of the snail shell is a dominant trait. If a mother is homozygous or heterozygous for right coiling, an offspring will have a right-coiled shell even if it was homozygous for left coiling. However, if the mother is homozygous for left coiling and an offspring is heterozygous for right coiling, the offspring would still have left coiling of the shell (Boycott et al., 1931). It is hypothesized that the mothers' genotype creates the environment that affects the first cleavage of the zygote ultimately changing the coiling of the shell in a proband.

Most of the research has focused on maternal genetic effects even though paternal are also possible (Lawson, Cheverud, and Wolf, 2013; Crean and Bonduriansky, 2014). However, mothers have more opportunities to affect the offspring's environment, especially in mammals through intrauterine milieu (Allen J Wilcox, Clarice R Weinberg, and Rolv Terje Lie, 1998), where it directly affects fetus development.

Although statistically, it could be easier if maternal effect and offspring's genotype were independent, in reality, that is not the case. It is confounded as offspring inherits alleles from its mother and thereby shares it with her (Buyske, 2008; Wolf and Wade, 2009).

1.1.3.2 *Parent-of-Origin effects*

The most well-known parent-of-origin effect is genomic imprinting, where the offspring's phenotype depends on the origin of the variant allele. In this phenomenon, an allele from only one parent is expressed, contrary to the simultaneous expression from both alleles. It is called maternal or paternal imprinting based on whose allele is imprinted. Mutations in imprinted genes have more severe consequences as there is no substitute allele. Moreover, as imprinted genes are functionally haploid, it can lead to the expression of a recessive allele, if a dominant allele is silenced (Jirtle and Weidman, 2007).

The genomic imprinting mechanism is to a certain degree explained with epigenetic changes, mostly DNA methylation, but it can also be due to histone modification, the formation of heterochromatin or regulation with noncoding RNA and RNAi. In the case of DNA methylation, imprinted genes are methylated making them inactive. Both parental germ cells have chromosomes with different marks, or often called imprints, that have to be maintained during fertilization and development of an embryo. However, the imprints are later erased and reimplemented in the offspring's gametogenesis depending on the sex of the individual (MacDonald, 2012).

The first imprinting-related diseases discovered in human population were Angelman and Prader-Willi syndrome. Offspring with Angelman syndrome often have problems with speech, development, and balance, as well as intellectual disability and a specific laugh ("happy puppet syndrome"). Individuals with Prader-Willi have symptoms like obesity, intellectual disability, short stature and hypogonadism (Pavlica, 2020). Both diseases are characterized with deletions on the long arm (q arm) of the chromosome 15. The difference is in the loss of parental genes. In Prader-Willi, the paternal genes are deleted and maternal are silenced, while in Angelman, the maternal genes are lost and paternal are silenced (Lobo, 2008).

However, other effects could lead to the appearance of a parent-of-origin effect in the absence of imprinting. One of them is the existence of a genetic difference between reciprocal heterozygotes and the other is parental indirect genetic effects that were covered in the previous subsection (Lawson, Cheverud, and Wolf, 2013).

1.2 GENOME WIDE ASSOCIATION STUDIES

With the rapid development of sequencing technologies and decreasing cost of sequencing a human genome, GWAS has become a more available method in the research. The aim of GWAS is to identify the genetic risk for a particular disease in the population by measuring and analysing DNA variations across the human genome. It is proven to be especially useful in identifying genetic risks in common complex diseases. Finding associations with the disease can help with a

better biological understanding of it, and in the future it could help with detecting, preventing and treating the disease.

1.2.1 *Study design*

An appropriate study design has to be chosen depending on the aim of the study to maximize the power to detect the association. Study designs for GWAS can roughly be divided into two categories: population-based and family-based studies (Evangelou et al., 2006; Kazma and J. N. Bailey, 2011).

The population-based design consists of collecting unrelated individuals that can be affected or unaffected. A case-control study is the most popular design for binary traits (Zondervan, 2011), where cases are individuals with the outcome/disease and controls are without the outcome/disease. Usually, samples are taken from the same ethnic population to avoid bias caused by admixture. Since two groups are defined at the beginning of the experiment, this type of control-case study can be found in the literature under the term retrospective study (Montreuil, Bendavid, and Brophy, 2005; Song and Chung, 2010; Kraft and D. G. Cox, 2008). An alternative is a prospective cohort where a cohort of random unaffected individuals is monitored through a long period of time (Melamed and Robinson, 2018). Relevant data is collected based on the hypothesized cause of the disease under investigation. This approach is better for non-biased selection of control group, however, it is very time-consuming and expensive (Cardon and Bell, 2001).

An alternative approach is a family-based design which is a type of case-only analysis where the affected individuals and their relatives are sampled (Kraft and D. G. Cox, 2008). The most common family types are trios (parents and affected offspring), duos (mother or father and affected offspring) or nuclear families (parents, affected offspring and siblings) (Kazma and J. N. Bailey, 2011). The advantage of this design is robustness to population stratification and the ability to identify parent-of-origin effects. However, family-based samples can be hard to collect, especially if it is a late-onset disease, where it might be impossible to genotype parents of the affected individual (Hirschhorn and Daly, 2005; Cardon and Bell, 2001).

1.2.2 *The mechanics of GWAS*

In general, the sample, which is usually saliva, is taken from individuals. DNA is isolated and genotyped. There are three genotyping options depending on the volume of DNA of interest and cost: whole genome sequencing (WGS), reduced representation sequencing (RRS) and SNP arrays.

SNP arrays are still primarily used for GWAS, despite the fact that they do not cover all genetic variants as compared to WGS. There are a few studies that used WGS (Gilly et al., 2018; Höglund et al., 2019; Berg et al., 2019) to increase power to detect variants associated with the disease (Gilly et al., 2018). However, the price

difference between these two options is still considerable for most studies, especially when similar results can be achieved with SNP arrays and imputed genotype data. Therefore, SNP arrays are more beneficial in terms of given accuracy and cost.

The most commonly used genotyping technologies are Illumina (San Diego, CA) and Affymetrix (Santa Clara, CA) (DiStefano and Taverna, 2011). Illumina method uses oligonucleotides attached to beads. The bead represents a locus or an allele. The sample fragment hybridizes with oligonucleotides on the bead and it gets prolonged for one nucleotide each step, which is detected as a specific colour. Similarly, Affymetrix method uses spots, where each spot has attached oligonucleotides representing a locus or an allele. The allele on a sample DNA is detected by differential hybridization (Bush and Moore, 2012). To conclude, Illumina method is more expensive, but oligonucleotide specificity is higher than in Affymetrix method and it is more flexible in changing the composition of the beadpool. However, Affymetrix method is less expensive which can be important in high-volume studies (DiStefano and Taverna, 2011).

After successfully obtaining a large number of variants (500,000 - 1,000,000 or more) and carrying out pre-processing and quality control (QC) steps, the association analysis can be carried out. For each variant, the test is performed to identify if there is an association between that variant and the phenotype. An appropriate statistical test is used depending on the number of phenotype groups. If the trait is quantitative, the generalized linear model is used. The generalized linear model tests if the genotype groups are independent of one another, if the trait is normally distributed, and that there is no significant variance within each group. The association in dichotomous traits, that is found in case-control studies, is tested with contingency tables or logistic regression. Contingency tables (e.g., χ^2 test) test whether an observed measure drastically deviates from the expected measure under the null hypothesis. Logistic regression tries to predict how likely it is that a subject is a case for a given genotype. Usually, logistic regression is the preferred method because it can incorporate covariates in the analysis and give more accurate results.

1.2.3 *Advantages and Disadvantages of GWAS*

Since the first GWAS was published in 2002 (Ozaki et al., 2002), the method has dramatically improved and has become more rigorous. It has proven to be very successful in finding new associations between novel variants and traits. Furthermore, those findings can be used to investigate potentially relevant genes that can lead to the discovery of novel biological mechanisms, such as the discovery of variant in *ATG161L1* gene in Crohn disease (Hampe et al., 2007; Murthy et al., 2014).

Genetic variants found by GWAS can have practical use in medicine. These findings can identify individuals at a higher risk, prevent or monitor, and detect diseases earlier, which could lead to the better patients outcome. For example, it was found that two variants in the *LOXL1* gene contribute to 99% of exfoliation

glaucoma in the population (Thorleifsson et al., 2007). That finding could be used for predicting the risk of developing the disease. Its usefulness was also proven in disease classification and subtyping, for example in diagnosing diabetes subtypes (Thanabalasingham et al., 2011), and in a selection of drug candidates (Nelson et al., 2015).

The application of GWAS findings is not restricted to disease aetiology, on the contrary, its use can be various. For instance, it can be utilized in estimation of birth location (Hoggart et al., 2012), forensic analysis (Homer et al., 2008), it can help in determining structure of the population (Jakkula et al., 2008) or it can help in determining history of a population (Reich et al., 2009; Yunusbaev et al., 2019), and many more.

However, when conducting GWAS, it is important to be aware of its limitations. The first problem is the number of tests. Each association test between a single nucleotide polymorphism (SNP) and a trait is an independent test, and in GWAS there are hundred to thousand SNPs tested. In order to account for multiple hypothesis testing, a high level of significance is needed to minimize the number of false positives. Usually, a strict Bonferroni correction is used, leading to an underpowered study. There are other options, such as Šidák correction, methods controlling for false discovery rate (FDR), permutation testing, and a more rigorous Bayesian approach. This methods will be described in more detail in Multiple Testing Correction and Bayesian statistics.

Secondly, the method explains only a small part of the overall heritability of complex diseases, although it identifies numerous variants associated with them. Explaining all heritability of disease is challenging because there are many components to complex traits, such as detection of common and rare variants with small effects, detection of ultra-rare variants, complex interactions (gene-gene and gene-environment) and so on. Gene-gene interaction, also called epistasis, is an interaction between genes that influences a phenotype. Epistasis, which was proven in model organisms to be a key component in genetic architecture (Mackay, 2014), is problematic because it does not have methodological consensus (Ritchie and Van Steen, 2018), the methods lack the statistical power to detect it and it is computationally burdensome due to all interactions that should be investigated (Mackay, 2014). Until it is possible to address all the components, there will be unexplained heritability in a disease. The small effect size and missing heritability limit the use of GWAS in medicine (Ritchie and Van Steen, 2018).

Thirdly, when interpreting the results, the investigator should be aware that a significant signal is not necessarily a causal variant. It might be in linkage disequilibrium with one or more causal variants (Tam et al., 2019), which requires further investigation.

With all that in mind, GWAS has its advantages and disadvantages, similarly to other methods. Nonetheless, it is proven as a useful tool in understanding the genetic architecture in complex traits when the study is thoughtfully planned.

1.3 AUTISM SPECTRUM DISORDER

One of the common complex diseases that affects around 1% of the population is autism spectrum disorder (ASD). The term "spectrum" refers to a wide range of different types and severities of symptoms that individuals with ASD have. Individuals are usually diagnosed early in life because symptoms appear within the first two years of life. Males are affected more than females, a recent meta-analysis (Loomes, Hull, and Mandy, 2017) reported a male:female ratio of 3:1. However, it can still happen that individuals are diagnosed later in life if they have less severe symptoms. The disease affects an individual's behaviour and ability to properly communicate with others. Some signs that the child could be autistic are repetitive behaviour, avoiding eye contact, not showing interest in objects around them, having speech problems and unusual body movements.

1.3.1 *Current insight*

Only a small number of causal genetic effects is known for ASD, despite extensive research. To unravel different aspects of the disease, various kinds of studies were designed. In the next few sections, studies relevant to this thesis will be mentioned.

1.3.1.1 *Twin studies*

Twin studies give a better understanding about the contribution of genetic and environmental factors to the aetiology of the disease. The estimated heritability of ASD is between 64% and 91% in twin studies (Tick et al., 2016). It was shown that monozygotic twins have a higher probability of developing ASD in both twins than dizygotic twins (A. Bailey et al., 1995; Rosenberg et al., 2009), which led to the conclusion that ASD is mostly due to strong genetic effects and less due to environment. However, there are two studies (Hallmayer et al., 2011; Frazier et al., 2014) that reported a stronger influence of environmental factors than previously estimated. Tick et al. (2016) analyzed all twin studies published before 2016, including Hallymayer et al.(2011) and Frayier et al. (2014), and concluded that aetiology of ASD occurs mostly due to strong genetic factors than environmental although the effect of environmental factors should not be neglected.

1.3.1.2 *Environmental factors*

Shared environmental effects become more significant when the prevalence rate is lowered. Tick et.al (2016.) demonstrated that by lowering prevalence rate from 5% to 1%, environmental effects rise from 7% to 35% (Tick et al., 2016). The prevalence rate of ASD is around 1% in the population and it is probable that some environmental effects contribute to the disease. In the last decade, studies revealed several environmental risk factors for ASD, such as parents' age (Wu et al., 2017), maternal

diabetes (Xu et al., 2014; Wan et al., 2018) and obesity (Reynolds et al., 2014), as well as complications during pregnancy and birth (for example neonatal hypoxia (Modabbernia et al., 2016), mother's valproate usage during pregnancy (Christensen et al., 2013) and preterm birth(Agrawal et al., 2018)).

1.3.1.3 *Genetic factors*

The genetics of ASD is heterogeneous, same as the disease itself. It varies in type of variation (from SNPs to CNVs), type of heritability (additive, dominant or recessive), frequency of variation (common, rare and ultra-rare variants), and origin of mutation (*de novo* or inherited). One study showed that common variants contribute more to the heritability of ASD on a population-level, while rare variants contribute on the individual level (Gaugler et al., 2014). In addition, another study reported that common variants with a small effect have additive effect in ASD (Klei et al., 2012).

Mutations in specific genes are consistently reported for ASD. Genes that are often mentioned include oxytocin receptor (*OXTR*), serotonin transporter (*SLC6A4*) and the gamma-aminobutyric acid (*GABA*) that could be used as clinical targets in the future. Genes associated with monogenic autism are often mentioned as well, such as fragile X mental retardation 1 (*FMR1*), tuberous sclerosis 1 and 2 (*TSC1, TSC2*) and methyl CpG binding protein 2 (*MECP2*) (Yoo, 2015). Furthermore, it is important to mention synaptic genes because mutations in these genes lead to synaptic dysfunction and are linked to ASD and other complex neurological disorders (Giovedí et al., 2014; X. Wang et al., 2018). One of the first synaptic genes that have been associated with ASD were neuroligin genes, *NLGN3* and *NLGN4X* (Jamain et al., 2003). Expression of these genes produces cell surface proteins that are involved in cell-cell interactions at the postsynaptic site. Since then, many other synaptic genes were also identified including SH3 and multiple ankyrin repeat domains (*SHANK*), contacting associated protein-like 2 (*CNTNAP2*), neurexin (*NRXN*) and many more (Giovedí et al., 2014; Yoo, 2015).

1.3.1.4 *Maternal genetic effects*

As stated in Genetic effects, detected associations can be confounded with underlying mechanisms. Involvement of both fetal and maternal genes, as well as interactions between the two, were implicated in pre-eclampsia (Cnattingius et al., 2004), spina bifida (Jensen et al., 2006), neural tube defects (Brody et al., 2002), and schizophrenia (Palmer et al., 2006). Maternal uterus is an environmental factor that affects normal fetus development. In twin studies of schizophrenia, the higher concordance for schizophrenia was noticed in monozygotic twins that shared placenta than in the twins that did not (Davis, Phelps, and Bracha, 1995). As these two neurological disorders share many common risk alleles (Psychiatric Genomics Consortium et al., 2013), it is necessary to investigate how genetic variation in mother affects the fetal environment in ASD.

The number of studies that have investigated maternal genetic effects in ASD is limited. In two small candidate gene studies, the frequency of transmission from maternal grandparents to mothers, and transmission from mothers to offspring was examined with transmission disequilibrium testing (TDT). It is a family-based association test that compares the frequency of transmitted and non-transmitted variant alleles from heterozygous parents to affected offspring. Comparing mothers with children diagnosed with ASD and mothers with healthy children, it was observed that glutathione S-transferase P1 gene (*GSTP1*) and human leukocyte antigen - DR isotype 4 gene (*HLA-DR4*) had significant transmission disequilibrium from maternal grandparents to case mothers. Transmission disequilibrium was not seen from parents to offspring. This supported the hypothesis that these genes have a risk allele in mothers that acts during pregnancy and contributes to the development of ASD in children (Williams et al., 2007; Johnson et al., 2009).

Two separate studies, Tsang et al. (2013.) and Yuan and Dougherty (2014.), conducted a case-control GWAS where they investigated maternal genetic effects. Tsang et. al (2013.) were one of the first to have a case-control study investigating maternal genetic effects where a relevant phenotype was having an autistic child. Case-control tests were conducted separately for offspring and mothers, where they compared autistic versus healthy individuals. For mothers, that meant comparing mothers that have an autistic child versus mothers that do not. Additionally, they investigated maternal-fetal interactions (transgenerational epistasis). Yuan and Dougherty (2014.) had a two-step design for association analysis. First, they compared mothers and fathers who were genotyped on the same technology. Because the association signals could possibly be confounded by sex, they had a second step. In second association analysis, a comparison was made between mothers that have autistic children and ones that do not have. A combined p-value of variants, that passed the threshold of both association tests, had to pass a Bonferroni threshold to be considered significant. However, there were no significant findings nor replications in both studies.

Furthermore, a third study, Shah (2012.), conducted a case-control GWAS in which they compared mothers versus fathers in trio families. Variant rs17743708 in *ABCC11* gene was found to be over-represented among mothers with children diagnosed with ASD. They also split mothers into two phenotypes: "strict" and "non-strict" based on the phenotype of affected offspring. The strict phenotype had a higher frequency of "A" allele in this SNP which could indicate that this variant could be used in determining the risk and severity of a disorder. However, the finding was not replicated (Naisha Shah et al., 2012).

A recent study investigated parent-of-origin effects with the use of Bayesian threshold proposed by Wakefield where they identified multiple regions that were previously associated with ASD. Interesting findings were variants in *SHANK3* gene and *WBSCR17* gene that were significant for the maternal genetic effect. Even so, they were unable to replicate these findings because cohorts were intrinsically different (Connolly, Anney, et al., 2017). Autism Genome Project dataset consists of

simplex (one affected offspring) and multiplex (multiple affected offspring) families, making it more enriched for common variants, while Simon Simplex Collection dataset has only simplex families, that are enriched in *de novo* and rare variants. All studies advocate that further research is needed in order to better understand the extent of the effect that maternal genetic effects have on ASD.

2 THE GOALS OF RESEARCH

A major part of variation is passed on from one generation to the next. Each biological parent passes on half of their genetic material to the offspring. However, parents can affect children even with genes that are not passed on. For example, mothers' genotypes can affect the development of children through intrauterine environment.

The main interest of this thesis is to identify potential candidate loci in mothers that might increase the risk of developing ASD in offspring. In the exploratory phase, a case-control GWAS is employed to see if there are any differences between alleles in mothers versus fathers from trio families with affected offspring. The graphical representation of design is shown in figure 1.

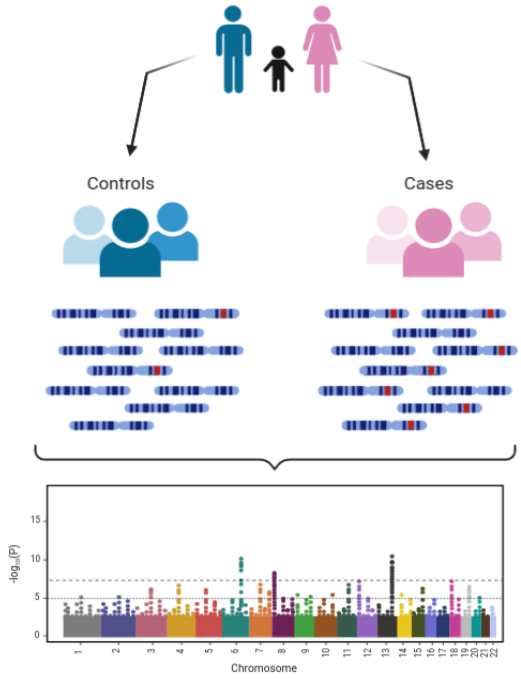

Figure 1: Graphical representation of GWAS design.
The study design consists of trio families: unaffected parents and an affected child. In GWAS, the prevalence of maternal genotypes is compared to paternal as a crude way of investigating maternal genetic effects in children with ASD. The results will be shown with a Manhattan plot.

Mothers are cases where the phenotype of interest is having a child with ASD. Fathers are used as controls as it is assumed that paternal genetic effects are

independent from maternal. An additional benefit is that parents are genotyped on the same genotyping technology at the same time, which avoids systematic bias.

GWAS serves as a preliminary method to infer possible maternal genetic effects by looking for difference between paternal and maternal genomic information. Assuming an additive model for complex disease, the association analysis performs a Cochran-Armitage trend test with 1 degree of freedom. The primary method for inferring maternal genetic effects will be a more powerful modelling approach, Weinberg log-linear method, that will estimate maternal genetic effects based on counts of family types. This method can easily be modified to allow for dominant and recessive effects, as well as imprinting effects. The results of modelling will be reported with Bayes factor rather than p-value because it compares probabilities of null and alternative hypothesis removing uncertainty found in a frequentist approach. Additionally, this allows for values to be compared between studies of different sample sizes.

Discovery of a new variant associated with the autism spectrum disorder could help with a better understanding of the disorders' mechanism and help with prevention, detection, and treatment of ASD in the future.

3

MATERIALS AND METHODS

3.1 DATASET

Version 4 of the the SPARK dataset was obtained from SFARIBase. The whole genome genotyping data is available for 27,290 individuals. Families consist of quads (parents, affected offspring, and affected or/and unaffected siblings), trios (parents and affected offspring) and duos (one parent and affected offspring). Families can be multiplex or simplex based on several individuals in a family that have a disease. Participants were registered online or in person at 25 clinical sites across the United States of America (USA) where they completed questionnaires about medical history and social behaviours. Hence, the case status is self-reported. Saliva was collected using the OGD-500 kit (DNA Genotek) and sent to a laboratory (Baylor Mirac Genetics Laboratories in Houston, TX or PreventionGenetics in Marshfield, WI). Each participant was genotyped with a SNP array (Infinium Global Screening Array-24) (Feliciano et al., 2018). The participants in the SSC were excluded by Simons Foundation Powering Autism Research for Knowledge (SPARK).

3.2 QUALITY CONTROL

It is necessary to do quality control in order to be sure that the results are reliable because GWAS is subjected to biases and confounding due to artefacts in the data. Problems can arise due to technical errors, like handling or processing of samples, or due to hidden structure in the data, for example, individuals from different populations or related individuals. The quality control steps were adapted based on QC steps proposed by Anderson et al. (2010) and Marees et al.(2018). and carried out in PLINK program. Visualization of the QC steps was done in R.

3.2.1 *Missingness*

In order to obtain high quality DNA without a systematic bias, missingness is checked on a sample and variant level. For each individual, the program calculates the number of variants that are missing, and for each variant, it calculates the number of individuals that do not have the information for that variant. The thresholds for PLINK2 flags `--mind` and `--geno` were set to 0.05, which means it excluded all individuals that are missing more than 5% of genotypes and all variants with missing genotype rate over 5%.

3.2.2 *Gender inconsistencies*

The discrepancies between assigned gender and gender determined based on an individual's genotype indicates DNA sample swap. Samples need to have variants on sex chromosomes to determine the gender. A priori checking the gender, the short regions of homology between X and Y chromosome, called pseudoautosomal regions - PAR1 and PAR2, should be removed, otherwise male heterozygous genotypes would confuse the algorithm. Flag `--check-sex` calculates homozygosity rates for each individual. Calls that have F-value smaller than 0.2 are labelled as females, and calls with F value larger than 0.8 are labelled as males. Those thresholds are data-dependent and should be adjusted accordingly. Individuals that are marked as PROBLEM should be removed or the gender should be updated.

From this point on, quality control is performed only on autosomal chromosomes.

3.2.3 *Minor Allele Frequency*

Suppose a variant that has two alleles, A and a, where A is more common (major allele) and a is less common (minor allele). In that case, the minor allele frequency is estimated with the following formula:

$$p_a = \frac{observed\ minor\ allele\ count}{total\ observations} = \frac{(n_{Aa} + 2n_{aa})}{2N}, \tag{1}$$

where n_{Aa} is the number of people with Aa genotype, n_{aa} is the number of people with aa genotype, and N is the number of individuals.

Variants with a low MAF in a population (frequency < 0.01) are called rare variants. The power to detect variant-phenotype association for rare variants is low and genotyping rare variants is prone to errors. Because of those reasons, it is common to exclude them in GWAS quality control step by removing all variants that have MAF lower than a given threshold. Depending on the study sample size, the threshold can be more or less stringent. It is advised to use a MAF threshold of 0.01 for sample size equal or more than $100,000$ samples, and 0.05 for the sample size of $10,000$ samples. In this thesis, a threshold of 0.01 was used, even though power for rare variants in the association analysis is lower. The reasoning is having more variants for investigating maternal genetic effects which will be reported with Bayes factor that takes MAF and power into account. This filter considers only parents.

3.2.4 *Hardy-Weinberg Equilibrium (HWE)*

In the ideal situation, where there is no selection, migration, or inbreeding and individuals from a large population mate at random, The Hardy-Weinberg equilibrium law indicates that genotype frequencies should be determined by allele frequencies.

If a variant has allele frequencies, p and q = (1-p), the genotype frequencies should be: $p_{AA} = p^2$, $p_{Aa} = 2pq$ and $p_{aa} = q^2$.

Deviation from HWE is tested by χ^2 or exact test by comparing observed and expected genotype counts. The deviation from HWE can occur due to a genotyping error, population stratification, the association between the allele and disease in cases, or if any of the assumptions are violated. In PLINK, the flag `--hwe` $1e05$ was used with a modifier *'keep-fewhet'*. If this modifier is not used in quality control steps, the test fails due to significant population stratification. PLINK calculates a deviation from HWE due to a small number of heterozygous and removes normal variants from the analysis. Same as MAF filter, it considers only parents.

3.2.5 *Heterozygosity*

The number of heterozygote genotypes in an individual can be an indicator of sample contamination or inbreeding and as such it is used as a quality control step. In this experiment, data consists of trio families in which case inbreeding is inevitable, but it is still possible to check for inbreeding between parents. The heterozygosity rate is used as a measure of heterozygosity. Heterozygosity was calculated only for parents with the following formula:

$$\textit{heterozygosity rate} = \frac{N - O}{N}, \tag{2}$$

where N is the number of non-missing genotypes and O is the number of observed homozygous genotypes for an individual. Excess of heterozygosity indicates sample contamination and a lack of it indicates inbreeding.

The distribution of the mean heterozygosity plot should be inspected for the extremities of the data. The established criteria for removing individuals with an excessive or reduced proportion of heterozygosity is to remove all individuals that lie outside three standard deviations of the mean.

3.2.5.1 *Variant pruning*

Before doing heterozygosity check in PLINK, it is necessary to prune variants to generate a list of non-correlated variants that will be used for calculating heterozygosity rate. Variant pruning is carried out with `--indep-pairwise` flag that takes 3 arguments. The first argument is the window size expressed in a number of variants, the second is the number of variants that the window will be shifted for and the third is the r^2 threshold. In this analysis, standard parameters were used. The window size was set to 50 variants, for each calculation, the window was shifted for 5 variants and the r^2 threshold was set to 0.2. The program outputs two files, prune.in and prune.out, where it lists all variant IDs that are in LD with each other, and the ones that are not, respectively. After removing variants in LD, heterozygosity can be calculated and plotted.

3.2.6 *Relatedness*

One of the sources of bias can be relatedness between individuals in a case-control study. Duplicates, twins, and first and second-degree cousins can cause bias by over-representing their genotypes in the study. In this experiment, relatedness was checked between parents by calculating identity by descent (IBD) which determines how much of identical nucleotide sequence in the segment is shared between two individuals. The more closely related individuals are, more segments are shared.

Before proceeding with calculating IBD, variant pruning is done with the same list obtained for heterozygosity check. Flag `--genome` creates pairs between all individuals and calculates IBD for each pair. The threshold for removing one person from a pair is 0.185, that is a halfway between thresholds for third and second-degree cousins.

3.3 POPULATION STRATIFICATION

All samples are from the USA, which is a melting pot of cultures. The ancestry origins from the individuals can introduce systematic bias in the data, called population stratification. That underlying structure in the population can cause spurious results in association analysis. It is known that the allele frequencies are distinct in each subpopulation and it is caused by non-random mating. However, this dataset consists of trio families, where the difference between parents' genetic information is compared. Individuals in families tend to come from a similar genetic background and provide strong protection against population stratification. Nevertheless, it is still possible that there is some confounding of genetic background, either between families or if parents have different origins. Correcting for it avoids any possible confounding. The correction can be done either by using principal components or by doing separate analysis for each subpopulation, which leads to smaller sample size and lower power to detect the association.

In order to compare SPARK dataset with the 1000Genomes reference dataset, it was necessary to run the same quality control steps on the reference and extract variants that are common between the data of interest and the reference data. After obtaining a list of common variants, both datasets were subsetted for those variants and then merged. Running principal component analysis (PCA) with flag `--pca approx 10` in PLINK gives back 10 principal components, which can be visualized in R.

Sample ethnic origins were deduced with the first three principal components that were calculated in PLINK. All samples were retained for association analysis and covariates were used to correct for population stratification.

3.4 ASSOCIATION ANALYSIS

Logistic regression with covariates was performed to see if there are any variants in mothers that could be associated with increased risk of a child developing ASD. It compares observed and expected genotype frequencies between cases (mothers) and controls (fathers) by assuming additive genetic model. It means that for previously mentioned variant example, it is expected that genotype *Aa* will have increased disease risk r, while genotype *aa* will have 2r. The additive model is often assumed in complex diseases when doing GWAS because the underlying genetic model is often unknown. The test will increase power if there is a trend but it does not affect the distribution of test statistic under the null hypothesis. The test results can still be meaningful if the trend assumption is not satisfied. This model is tested with a Cochran-Armitage test for trend, which is a modified Pearson χ^2 test that can incorporate three possible genotypes frequencies.

3.4.0.1 *Multiple Testing Correction*

As already stated in Advantages and Disadvantages of GWAS, it is important to address the problem of multiple testing. For the association test, the results were corrected with PLINK's flag `--adjust` that returns adjusted p-values with several correction methods.

Bonferroni, Šidák single-step, Holm-Bonferroni, and Holm-Šidák step-down procedures correct for the family-wise error rate (FWER) which is a type I error or the probability of making one or more false discoveries. Bonferroni is the most strict procedure. The adjusted p-value threshold is calculated by dividing a desired significant level by a number of hypotheses, $\alpha_{BONF} = \frac{\alpha}{number\ of\ tests}$. A large number of studies use a proposed p-value threshold of $5x10^{-8}$ which is calculated based on 1 million hypotheses and a significant level of 0.05. Most studies are underpowered when using this correction because there are many false negatives due to this strict threshold. A higher power can be achieved with Šidák single-step procedure where p-value threshold is equal to $\alpha_{SID} = 1 - (1 - \alpha)^{\frac{1}{number\ of\ tests}}$. Moreover, the tests need to be independent, which is not often the case because of the linkage disequilibrium between variants. If that is violated, the threshold becomes too conservative. Both Bonferroni and Sidak corrections have a less strict form detonated with Holm prefix that uses step down testing. In this method, p-values are sorted from highest to lowest. The highest p-value is compared with calculated Bonferroni or Sidak p-value threshold, the second highest is compared again but the time number of tests in the formulas is n-1, for the third is n-2 and so on. The lowest p-value is then found among the ones that pass the threshold, and all p-values that are higher or equal to the threshold are considered significant. The number of tests for each hypothesis is equal to $n + 1 - k$, where k is the index or rank of the sorted p-value. In this

way, null hypotheses $H_1, ..., H_{k-1}$ are rejected, and $H_k, ..., H_n$ are accepted. With this adjustment, there is a lower increase of type II errors than in traditional procedures.

There are two more methods used by PLINK, Benjamini & Hochberg and Benjamini & Yekutieli procedures. Both of these methods control for FDR. If an FDR value is equal to 0.05, it means that 5% of detected positive results are truly negative. Contrary to the step-down, these methods use the step-up procedure. All p-values are sorted in the ascending order and then compared if they are smaller than $P_{HB} = \frac{k}{n}\alpha$. Among the p-values that satisfy the criterion, the largest is chosen. All p-values that are smaller or equal to the largest one are considered significant. Benjamin & Yekutieli is similar to Benjamini & Hochberg but more conservative. The formula is similar, the only difference is $\alpha = \frac{\alpha}{\sum_k^{i=1} \frac{1}{i}}$.

3.5 MENDELIAN INCONSISTENCIES

After the association analysis, the dataset of only European trio families was created. In families that had multiple affected children, one child was picked randomly for inclusion. In each family, it was necessary to check for inconsistencies in Mendelian inheritance. A Mendelian error arises if an offspring has an allele that could not have been inherited from parents based on their alleles for that variant. For example, if both parents are homozygous, *AA*, a child should also be homozygous. If an offspring is heterozygous *AB* or homozygous for different allele BB, it is detected as a Mendelian error. The reasons for these errors can be numerous, it can be due to the wrong assignment of a family relatedness, genotyping errors, or *de novo* mutations. Flag `--me` in PLINK needs two parameters, the first determines the percentage of Mendelian errors that has to be found in a family to discard it from the analysis and the second indicates the percentage of Mendelian errors in a variant that determines if a variant will be kept or discarded. The thresholds used in this analysis were 0.05 and 0.05.

Furthermore, before modelling the maternal genetic effects, missingness and MAF were checked again. The number of removed variants and individuals at each step of QC can be found in table 1.

3.6 LOG-LINEAR MODELING

In 1998, Wilcox, Weinberg and Lie proposed a method for detecting the contribution of offspring's, maternal and combined genetic effects in trio families. The families consist of a father (F), a mother (M) and an affected child (C). Each individual can carry 0, 1 or 2 risk alleles for a biallelic gene. With that in mind, there are 15 possible family combinations and six mating types which are listed in table 2. The mating types are defined by the number of risk alleles carried by parents. For example, if one parent has 1 allele and the other has 0 then they fall under mating type 5.

Table 1: Quality control steps for SPARK dataset

	No. of individuals	No. of variants
Before QC	26 673	16 754 504
Missingness <95 %	26 673	16 753 240
Gender inconsistencies	26 666	-
MAF <1%	-	147 794
HWE <0.00001	-	144 585
Heterozygosity[1]	14 063	-
Relatedness[1]	13 853	-
PCA[2]	23 937	-
Mendelian errors	23 937	137 660
Missingness <95%	23 937	137 652
MAF <1%	-	126 532
Final	23 937	126 532

[1] showing number of individuals only for parents

[2] after performing PCA, non European families were removed

The premise is that cases will show a higher relative risk (RR) than controls because they carry the risk allele. When investigating maternal genetic effect, a mother is a case and it is expected to see the over-representation of a risk allele compared to the null model where similar counts are expected between parents within each mating type. Offspring's genetic effects are represented by parameters R1 and R2, where R1 is the relative risk with one copy of the allele and R2 is the relative risk with two copies of the allele compared to no copies. Maternal genetic effects are depicted in the same way with S1 and S2. These parameters can be estimated using the maximum likelihood approach where the theoretical multinomial distributions written in table 2 are fitted to the observed triad genotype counts for each variant.

The general equation for modeling maternal and offspring's genotypes is written in equation 3.

$$\begin{aligned} ln[E(n_{F,M,C}))] = \gamma_j + \beta_1 I_{\{C=1\}} + \beta_2 I_{\{C=2\}} \\ + \alpha_1 I_{\{M=1\}} + \alpha_2 I_{\{M=2\}} + ln(2) I_{\{F=M=C=1\}}, \end{aligned} \tag{3}$$

where γ_j represents six parental mating types which serve as stratification parameters, and I are "dummy" variables where values are 0 or 1 depending on whether the child or mother has one or two risk alleles. For example, if a mother has one risk allele, then $I_{\{M=1\}} = 1$ and $I_{\{C=1\}} = I_{\{C=2\}} = I_{\{M=1\}} = 0$. The model is easily modified for modeling dominant and recessive effects by addition of two variables ($\beta_1 = \beta_2$, $\alpha_1 = \alpha_2$) or omitting the single-allele variables ($\beta_1 = 0$, $\alpha_1 = 0$).

Table 2: Theoretical frequencies for offspring's and maternal effects in trio families for a diallelic SNP. Combined tables from the paper Allen J Wilcox, Clarice R Weinberg, and Rolv Terje Lie (1998).

No. of Risk Alleles (MFC) [a,b]	Mating Type	Theoretical Frequency For Offspring's Effects[c,e]	Theoretical Frequency For Maternal Effects[d,e]
222	1	R2 γ_1	S2 γ_1
212	2	R2 γ_2	S2 γ_2
211	2	R1 γ_2	S2 γ_2
122	2	R2 γ_2	S1 γ_2
121	2	R1 γ_2	S1 γ_2
201	3	R1γ_3	S2 γ_3
021	3	R1 γ_3	γ_3
112	4	R2 γ_4	S1 γ_4
111	4	2 R1 γ_4	2 S1 γ_4
110	4	γ_4	S1 γ_4
101	5	R1 γ_5	S1 γ_5
100	5	γ_5	S1 γ_5
011	5	R1 γ_5	γ_5
010	5	γ_5	γ_5
000	6	γ_6	γ_6

[a] Family trio combinations of mothers (M), fathers (F) and affected offspring (C).

[b] Genotypes are depicted as 0, 1 and 2 depending on the individual's number of risk alleles

[c] R1 and R2 are offspring relative risks associated with one or two copies of risk allele

[d] S1 and S2 are maternal relative risks associated with one or two copies of risk allele

[e] γ_n is the parameter for nth mating type

$ln(2)I_{\{F=M=C=1\}}$ is an offset that covers special case where all three family members are heterozygous and it is unknown from which parent the child inherited each allele.

Furthermore, they expanded the model to allow for the imprinting effects (C. Weinberg, A. Wilcox, and R. Lie, 1998), as shown in 4 equation.

$$\begin{aligned} ln[E(n_{F,M,C}))] = \gamma_j + \beta_1 I_{\{C=1\}} + \beta_2 I_{\{C=2\}} + \alpha_1 I_{\{M=1\}} + \alpha_2 I_{\{M=2\}} \\ + \varepsilon_1 I_{\{F\}} + \varepsilon_2 I_{\{M\}} + ln(2) I_{\{F=M=C=1\}}, \end{aligned} \tag{4}$$

where ε_n are parameters for paternal and maternal imprinting.

After the modelling and acquiring parameters' estimations, the relative risks can be calculated by taking the exponent of the parameter. For example, e^{β_1} would be a relative risk for a mother having one allele. In the additive log-linear model, estimations of maternal and offspring parameters are independent and it is possible to separately calculate the effects of maternal or offspring's genotype on the risk of developing a disease.

The general equation 3 was implemented in function colEMlrt as a part of R package "trio" by Schwender et al.(2020). However, the function was modified as a part of this thesis in order to be able to choose the parameters for modelling and to get more summary statistics from the tests such as α and β estimates from the equation. With this modification, it is possible to choose modelling of only maternal effects which was previously not possible, and to calculate alternative statistics like approximate Bayes factor (ABF).

3.7 LIKELIHOOD RATIO TEST

The goal is to compare the null hypothesis, $H_0 : \beta = 0$, there are no maternal genetic effects, to alternative hypothesis, $H_1 : \beta \neq 0$, there are maternal genetic effects. Usually, nested models are compared and the goodness-of-fit is tested with likelihood ratio test (LRT). For example, to investigate maternal effects, a comparison would be made between the full model depicted in (3) and a null model which would be a general model with omitted maternal parameters. The values from the test do not have a meaning by themselves, but by comparing the difference between log likelihoods of the nested models (5), it can be determined which hypothesis is more likely. The likelihood ratio statistic converges to χ^2 distribution and, with correct degrees of freedom, p-values can be obtained.

$$D = -2ln\left(\frac{likelihood\ for\ null\ model}{likelihood\ for\ alternative\ model}\right) \tag{5}$$

The most used statistic for deciding which variants are significant in GWAS is the p-value. However, there are many drawbacks that come with it, the biggest being that it neglects the power of tests, which is a function of sample size and MAF. With p-values, only one threshold is set for all sample sizes (Stephens and Balding, 2009).

3.8 BAYESIAN STATISTICS

There is an alternative where instead of p-values, significance can be determined with Bayes factor (BF). This measure takes power into the account, it can easily be compared between variants in the study and with variants from different studies, and it easily incorporates biological information. Bayes theorem is a ratio between the probability of data under null to the probability of data under alternative hypothesis:

$$BF = \frac{Pr(data|H_0)}{Pr(data|H_A)} \tag{6}$$

The formula is similar to LRT equation 5, but instead of comparing two parameters in the model, it compares two different models. If BF is equal to 1, it means that observed data is equally likely for null and alternative hypothesis. A larger BF means there is more evidence for null hypothesis, and the smaller it is, there is more evidence for alternative hypothesis.

To decide on the significance, the posterior odds on the null hypothesis are required. The intuitive form for posterior odds is a multiplication of BF and prior odds (PO):

$$Posterior\ odds\ of\ H_0 = BF\ x\ PO\ of\ H_0, \tag{7}$$

where $PO = \frac{\pi_0}{1-\pi_0}$ represents prior odds of no association, in which π_0 denotes a prior probability that the null is true, and $1 - \pi_0$ a prior probability that alternative is true. If posterior odds on H_0 are smaller than the ratio between type II and type I errors (ratio of cost), the null hypothesis is rejected and it is concluded that an association is significant.

3.8.1 *Approximate Bayes factor*

After specifying PO and ratio of the cost (R), a simpler form of Bayes factor can be used. The formula for the approximate Bayes factor against null hypothesis is:

$$ABF = \sqrt{\frac{V+W}{V}} exp\left(-\frac{z^2}{2}\frac{W}{(V+W)}\right), \tag{8}$$

where z is the Wald Z score ($Z = \frac{\beta}{s.e.\beta}$), β is the maximum-likelihood estimate of the parameter, $\sqrt{V}$ is the standard error of the parameter and W is the prior variance.

In the full model, standard errors can be inflated if a larger number of variables is used. Therefore, the estimates and standard errors of the parameters were taken from simplified models in which maternal and offspring's parameters were modelled independently. For prior odds, the assumption is that 500 out of 1,000,000 common variants are significant contributors to ASD. In that case, the prior probability for H_0 being true is $\pi_0 = 1 - \frac{500}{1{,}000{,}000} = 0.9995$ which leads to $PO = \frac{0.9995}{1-0.9995} = 1,999$. The

ratio of the cost was set to 10, as in Connolly et al. (2017). It is sensible to say that type II errors are 10 times worse than type I errors because false negative findings are lost in the true negative findings and can not be followed up. Furthermore, the same prior variance was decided to be used in this thesis, $W = 0.42^2$, because ASD families are also investigated in this thesis and the odds ratios findings are relevant in the same way. The results will only be reported for variants for which MAF is equal to or greater than 0.05 because the power to detect a non-null association for a relative risk of 1.5 for MAF smaller than 0.05 is very weak - between 0.2 and 0.4 (Wakefield, 2008).

3.9 PERMUTATION TESTING

The Bayes factor is a measure of the strength of evidence for a genotype-phenotype association. This measure by itself should be enough for deciding if an association is present or not. However, it is calculated based on the subjective prior belief which can be questioned. One way of addressing this issue is with "Bayes/non-Bayes compromise" (Good, 1992). This method uses permutation tests to control for type I error (number of false positives). Permutation tests randomize the case and control statuses to calculate all possible values of the used test statistic under the null hypothesis. The test statistic distribution of the permuted datasets is used to determined how likely it is to observe the original association by chance. In the SPARK dataset, case and control statuses are maternal and paternal IDs. Permuted datasets are created by randomly assigning parental IDs inside a family. The Bayes factor is obtained for each permuted dataset. The p-value is then calculated as a ratio of permuted datasets for which the Bayes factor is less than or equal to the one obtained from the observed data and the total number of permutations (eq 9).

$$p_{perm} = \frac{\sum_{n=1}^{N_{tot}} \left(BF_{j,n} \leq BF_{j,original}\right)}{N_{tot}}, \tag{9}$$

where N_{tot} is the number of permutations, $BF_{j,n}$ is the Bayes factor of j-th variant calculated in nth permutation, and $BF_{j,original}$ is the calculated Bayes factor of j-th variant in the observed data.

It is important to note that permutation tests assume that the labels are exchangeable. The SPARK dataset has a family structure that needs to be preserved, which constrains the possible permutations and reduces the power. However, permutation tests can still be used and offer an advantage when mathematical and biological properties are not well understood.

P-value obtained with this approach is valid regardless of chosen priors. However, the permutation procedures are quite time-consuming in case of large datasets. Because of that, in this thesis, only 500 permutations were made. The R script for permutation can be found in the Appendix.

3.10 OTHER ANALYSIS

3.10.1 *Transmission disequilibrium test*

The TDT detects association between a variant and a disease if there is an unequal transmission of alleles between heterozygous parents to affected offspring. Two types of TDT were carried out. In classic TDT, the number of transmitted and non-transmitted alleles from heterozygous parents to affected offspring are compared with equation 10.

$$TDT_{stat} = \frac{(a-b)^2}{a+b}, \tag{10}$$

where a is the number of transmitted and b is the number of non transmitted minor alleles.

The second TDT is a parent-of-origin TDT, where separate TDTs for parents are calculated. For each parent, it calculates transmitted and non transmitted alleles from heterozygous parents to the affected offspring. It gives additional information if the observed excess of the minor allele of the variant could be due to parental imprinting.

In both tests, TDT statistic follows one degree of freedom χ^2 distribution.

3.10.2 *Linkage Disequilibrium*

The association signal can be seen because there is a true association of that variant or the variant is in linkage disequilibrium with a causal variant. Linkage disequilibrium (LD) is the correlation between neighbouring variants, where their observed alleles are more, or less, associated in the population than is expected for independent variants.

The commonly used measures for LD are squared Pearson correlation coefficient, r^2 and coefficient of linkage disequilibrium, D. The latter measure is more intuitive. The loci are in linkage equilibrium if the haplotype frequencies are equal to the product of their allele frequencies. D is calculated by subtracting the product of allele frequencies from haplotype frequency. The value of D can be negative depending on the allele frequency of loci, but standardizing it solves the issue. On the other hand, r^2 takes frequencies of loci into account when it is calculated. The standardized D and r^2 both range from 0 to 1, where 1 is linkage disequilibrium and 0 is linkage equilibrium.

The LD analysis is done in PLINK with flag `--r2` on statistically significant variants to identify haplotype blocks. The results are plotted in R for variants that have $r^2 > 0.2$.

4

RESULTS

The SPARK dataset consists of 27,290 individuals genotyped for around 17 million variants on Infinium technology. After the QC steps and the creation of trio families, the final number was 13,206 individuals or 4,420 trio families (parents and affected offspring) and 126,532 variants.

4.1 PRINCIPAL COMPONENT ANALYSIS

All parents were compared to the 1000Genomes reference to infer their ancestry. In figure 2, the PCA plot shows the first two principal components separating individuals between three groups: East Asian, African and others, which includes European, American and South Asian.

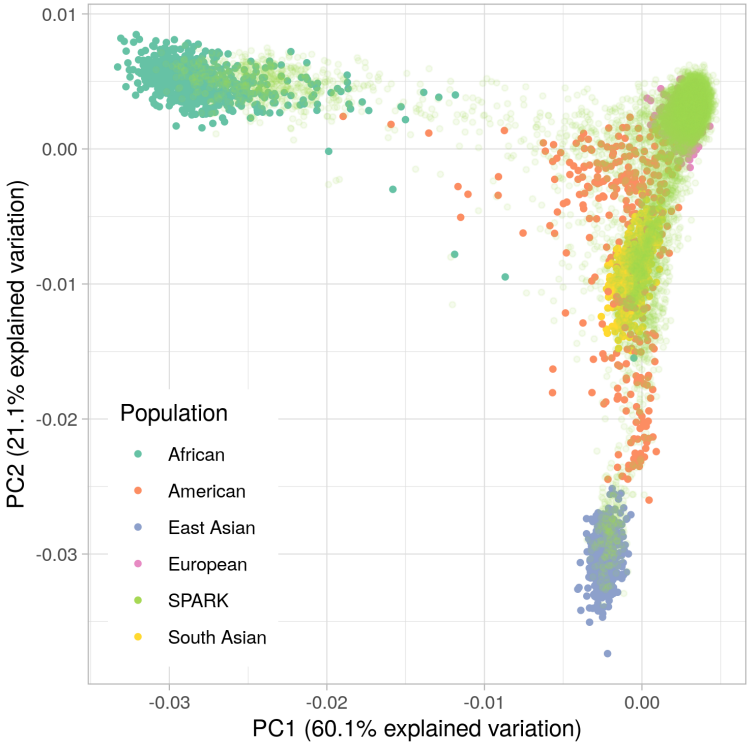

Figure 2: PCA plot of first two PCs.
The first two principal components of the 1000Genomes and SPARK dataset stratified by populations. Each population has a distinct colour. The distribution of parents is shown in light green.

The second PCA plot shown in Figure 3 indicates a clear separation of South and East Asians from the rest of populations.

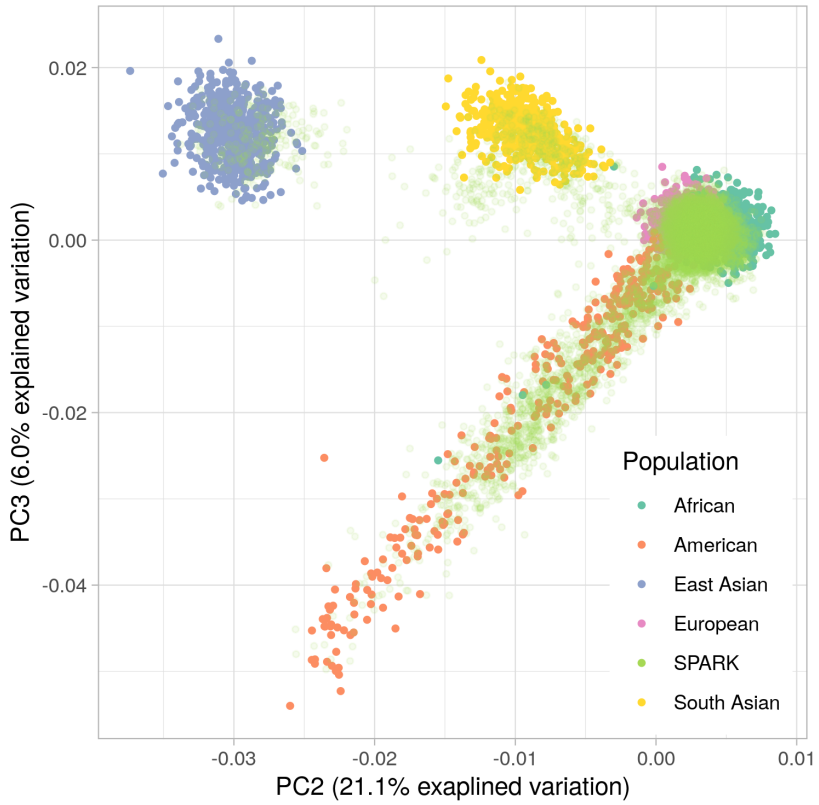

Figure 3: PCA plot of second and third PCs.
Second and third PCs of SPARK individuals against 1000Genomes population reference panels of African, American, Asian and European populations. The distribution of SPARK parents is shown in light green.

In both figures 2 and 3, the SPARK dataset is coloured in light green. The transparency of the colour indicates the density of individuals from SPARK dataset. A strong, light green colour represents a majority of SPARK dataset that overlaps with European individuals from 1000Genomes. Pale, light green dots distributed in other population groups are SPARK individuals that have non-European origin.

4.2 ASSOCIATION ANALYSIS

The maternal genotypes were compared to the paternal as a preliminary method to investigate if any maternal variants are associated with offspring's risk of developing ASD. Association of 144,585 variants was tested with logistic regression and the population structure was adjusted with the first three PCs. Obtained p-values were used to asses the success of removing bias with quantile-quantile (QQ) plot. In figure 4, the QQ plot shows that most variants follow the null hypothesis, indicating a

successful quality control, without the results getting affected by the population structure. Additionally, the genomic inflation factor calculated in PLINK is 1.048, confirming a small probability of false positives due to the population stratification, relatedness, or genotyping errors. Covariates were also included in the genomic inflation factor calculation.

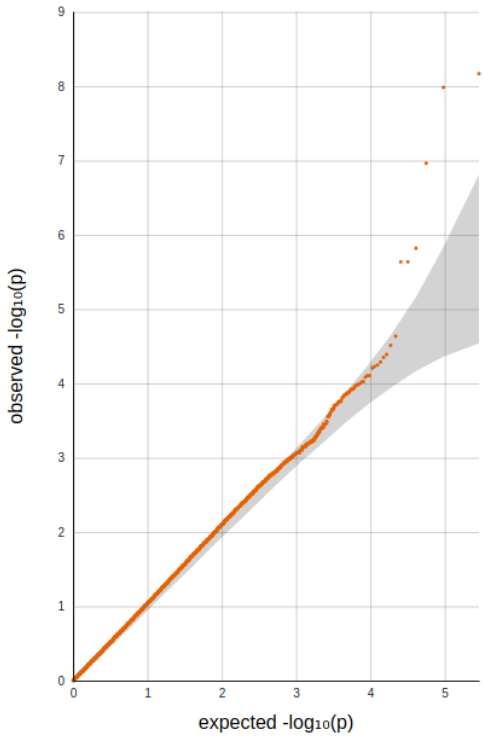

Figure 4: QQ plot of the association results.
The plot shows expected versus observed p-values from association test where maternal genotypes were compared against paternal genotypes. The association was tested with logistic regression and population structure was adjusted with first three principal components.

The result of the association analysis is visualized with a Manhattan plot shown in figure 5. Most GWAS use a fixed Bonferroni threshold of 5×10^{-8} (Kanai, Tanaka, and Okada, 2016), that is shown in figure 5 as a red line. The threshold is based on 1 million tests, and the SPARK dataset had 126,532 variants after the QC steps were performed. The difference between fixed and normal Bonferroni correction, that takes the number of tests into account, can be seen between the figure 5, where only two variants are significant, and table 3, where three variants are significant.

As can be seen in table 3, almost all tests for correction of multiple hypothesis testing problem consider three variants to be significant, except FDR Benjamini - Yekutieli method. The most significant association is a synonymous variant rs782320706 in *FCGBP* gene on chromosome 19. The other two variants are mapped to an intronic region of Ankyrin Repeat Domain 36 gene (*ANKRD36*).

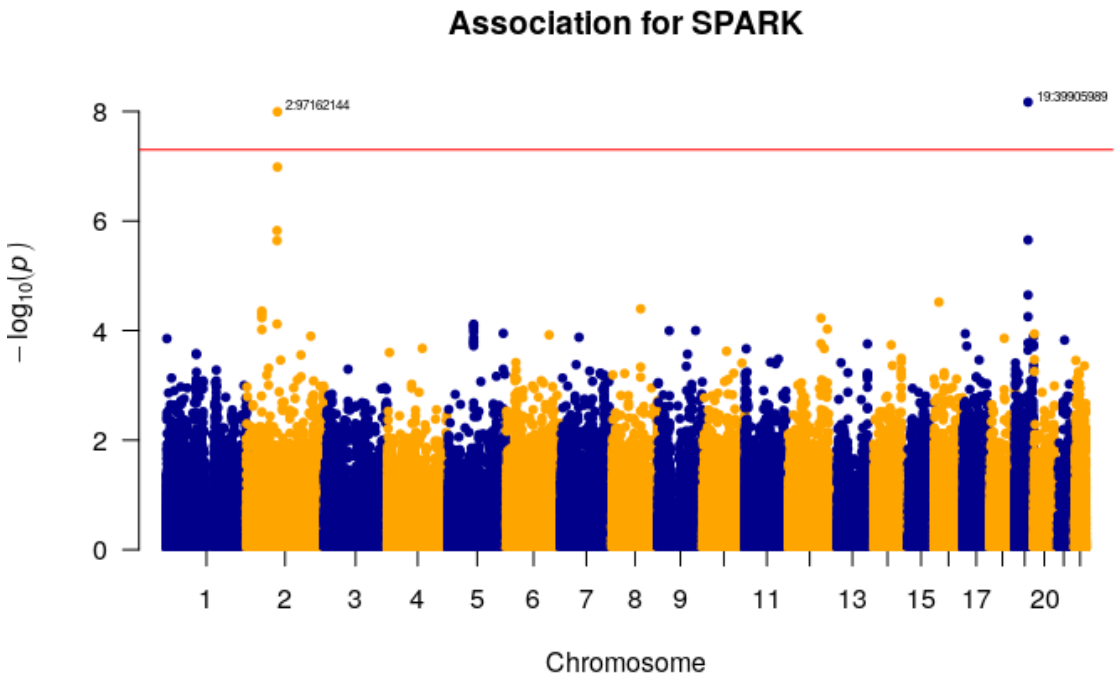

Figure 5: Manhattan plot of SPARK dataset.
Manhattan plot showing results of association analysis of SPARK dataset mothers versus fathers. The variants are ordered by genomic position on the x-axis. The y-axis is the negative logarithm to base 10 of the association p-values for each variant. The red line is a strict Bonferroni threshold (5×10^{-8}). Two most significant variants are labelled.

Table 3: Results of association analysis between mothers and fathers of affected offspring.
Top results from association analysis with unadjusted and adjusted p-values. The significant p-values are written in bold.

SNP	Region	Gene	MAF	p-value	Bonf*	Holm*	Sidak SS*	Sidak SD*	FDR_BH*	FDR_BY*
rs782320706	19q13.2	FCGBP	0.782	6.76e-09	**9.78e-4**	**9.78e-4**	**9.78e-4**	**9.78e-4**	**7.34e-04**	**9.15e-03**
rs371783424	2q11.2	ANKRD36	0.020	1.02e-08	**1.47e-03**	**1.47e-03**	**1.47e-03**	**1.47e-03**	**7.30e-04**	**9.15e-03**
rs1275549550	2q11.2	ANKRD36	0.022	1.04e-07	**1.50e-02**	**1.50e-02**	**1.49e-02**	**1.49e-02**	**4.99e-03**	6.22e-02
rs760504696	2q11.1	ANKRD36C	0.012	1.50-06	0.22	0.22	0.20	0.20	0.0542	0.68
rs200671922	19q13.2	FCGBP	0.824	2.22e-06	0.32	0.32	0.28	0.28	0.0549	0.68
rs77708223	2q11.1	ANKRD36C	0.010	2.28e-06	0.33	0.33	0.28	0.28	0.0549	0.68

* Bonf - Bonferroni correction, Sidak SS and SD - Šidák single-step and Holm-Šidák step-down adjusted p-value FDR_BH - Benjamini & Hochberg (1995) step-up false discovery control, FDR_BY - Benjamini & Yekutieli (2001) step-up false discovery control

However, the minor allele frequency of variants in *ANKRD36* and *ANKRD36C* is between 0.01 and 0.025, which can produce a false positive association signal. In order to lower the number of false positives when investigating maternal effects, the log-linear modelling and Bayesian factors were adopted and used in the next steps.

4.3 MATERNAL EFFECTS

The log-linear modelling was used to investigate the maternal effects as a better method for distinguishing effect types. The method can easily incorporate recessive and dominant genetic model, but an additive model was used in this thesis. The maternal effects were fitted separately from the offspring's. The evidence categories proposed by Jeffreys were used to decide which variants were significant based on the ABF. P-values were not used as they anticipate stronger effects with a lower MAF, although the relationship between MAF and effect size has not been shown, making p-values more prone to false positives (Wakefield, 2009). Hence, the biggest difference between the Bayes and the p-values approaches will be for variants with a low MAF. The table 4 shows the categories and the observed number of variants in the SPARK dataset. It was found that 147 out of 126,532 variants have a maternal genetic effect. In figure 6, the distribution of the negative logarithm of the

Table 4: The evidence category for Bayes factor by Jeffreys (1961.) and observed counts.

Bayes factor	$-log_{10}$BF	Strength of evidence for H_1	Observed
< 0.0001	> 4	Decisive evidence for H_1	147
$0.0001 - 0.001$	$3 - 4$	Very strong evidence for H_1	104
$0.001 - 0.01$	$2 - 3$	Strong evidence for H_1	249
$0.01 - 0.1$	$1 - 2$	Substantial evidence for H_1	787
$0.1 - 0.32$	$0.5 - 1$	Anecdotal evidence for H_1	1,859

approximate Bayes factor is plotted. The peak denotes a majority of variants that show evidence for the null hypothesis ($-log_{10}ABF < 0$).

Figure 7 plots the variants with the decisive evidence. The colour in the plot depicts the effect of a mother having one risk allele. The estimated maternal effects for significant variants range from 0.76 to 1.29. As can be seen, most of the variants have a maternal genetic effect smaller than 1.

The top 10 significant variants ranked by ABF are shown in table 5. For these variants, no association with offspring's genotypes was observed based on ABF.

The variant with the most evidence of having maternal genetic effects was on chromosome 17 in the pregnancy specific beta-1-glycoprotein 1 gene (*PSG1*, rs12459171, $S_1 = 0.86$). Additionally, the variant with the strongest observed maternal effect (S1 = 1.29) is rs116948313 in scinderin gene (*SCIN*). It is of interest for investigation into the maternal gene and its possible effects on offspring. The counts for the variant with the strongest maternal effect are shown in table 6, where an over-representation of allele "T" in mothers compared to fathers can be seen.

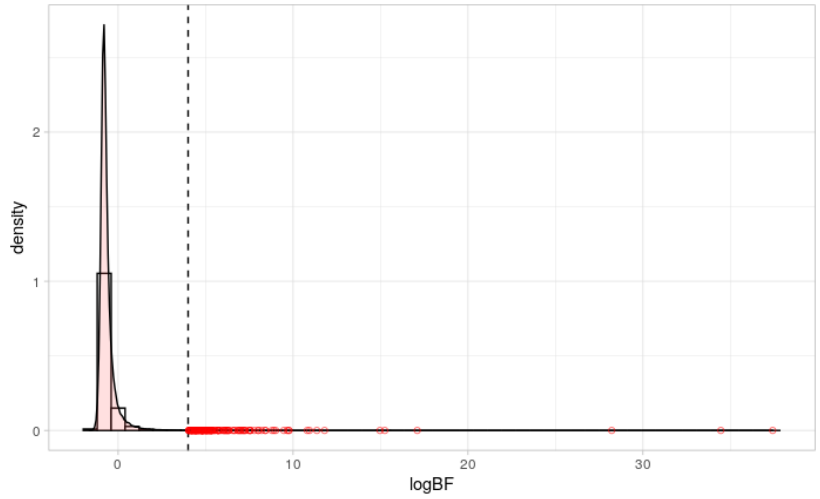

Figure 6: Density plot of ABFs.
The plot shows a distribution of calculated ABF for each variant. The threshold for decisive evidence for H_1 is shown as a dashed line. Significant variants are shown as small red circles.

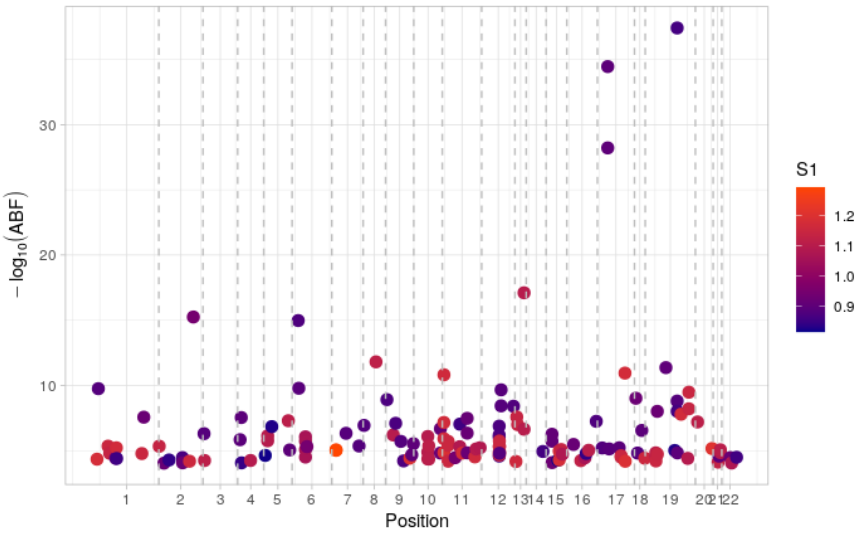

Figure 7: Significant results for maternal genetic effects.
The plot shows the significant results for maternal effects. On x-axis is the position of a variant in base pairs, on y-axis is the negative logarithm of the approximate Bayes factor. The effect of a mother having one risk allele (S1) is colour coded as shown in the legend.

Table 5: Top 10 ranked variants based on ABF.
The p-value for an observation with 500 permutations should be given as ≤ 0.002, if it is seen in no permutations, and as ≤ 0.004 if seen in two etc.

rsID	Region	*Gene*	MAF	S1	$-log_{10}$ABF	p-value
rs12459171	19q13.2	*PSG1*	0.11	0.86	37.42	0.04
rs1129235	17p11.2	*TRPV2*	0.39	0.899	34.45	0.09
rs8121	17p11.2	*TRPV2*	0.39	0.90	28.21	0.06
rs2275843	13q34	*COL4A1*	0.16	1.10	17.10	0.06
rs4535042	2q35	*ATIC*	0.31	0.94	15.24	0.41
rs45617437	6p22.1	*MOG*	0.05	0.88	14.97	0.03
rs66493340	8q21.2	*CNBD1*	0.45	1.12	11.80	0.03
rs2072496	19p13.11	*JAK3*	0.10	0.90	11.36	0.02
rs11658510	17q25.1	*COG1*	0.29	1.17	10.94	0.00
rs77173309	11p15.5	*PIDD1*	0.13	1.16	10.81	0.05

Table 6: The counted number of families for variant rs116948313 in complete trio families.
The mothers' or fathers' genotypes that contribute to the risk allele in a child are marked bold. There are more families where a mother contributes to risk allele than families where the father contributes.

Maternal Genotype	Paternal Genotype	Offspring Genotype	Number of families
TT	**TT**	TT	0
TT	**TC**	TT	0
TT	TC	TC	0
TC	**TT**	TT	0
TC	**TT**	TC	0
TT	CC	TC	8
CC	**TT**	TC	9
TC	**TC**	TT	15
TC	**TC**	TC	28
TC	TC	CC	11
TC	CC	TC	199
TC	CC	CC	230
CC	**TC**	TC	162
CC	TC	CC	170
CC	CC	CC	3570

4.3.1 *Permutation*

In order to control for FDR, mother and fathers were shuffled in each family. Log-linear modelling was repeated on permuted datasets to see for each variant how often the same or more significant BF will be obtained with random assignments of parents. After 500 permutations, the p-value was considered significant for 115 out of 147 variants. Calculated p-values of the top significant variants can be found in the table 5. As it can be seen, variant rs4535042 was not significant after 500 permutations. However, the variant with the strongest maternal effect was still significant with p-value is 0.00.

4.4 OTHER ANALYSIS

4.4.1 *Transmission disequilibrium test*

The observed association with maternal genetic effects could likely be confounded by the child's genotype or maternal imprinting effects. An offspring and a mother share a common allele, which makes it possible to observe the same pattern of risks between them. As for maternal imprinting, the excess of the maternal risk allele would indicate that the observed maternal effects could likely be maternal imprinting effects.

TDT without a special design can not detect maternal genetic effects, but it is a powerful method for detecting the effects of offspring's genotype and the maternal imprinting. TDT and parent-of-origin TDT were performed for the top 10 significant variants and the variant with the strongest maternal genetic effects. The results are shown in table 7. The classic TDT was performed to see if there is any association with offspring's genotype, while parent-of-origin TDT was used to see if the detected maternal effect could have been camouflaged maternal imprinting.

In table 7, it can be seen that obtained p-values of classic TDT are insignificant, therefore the offspring's genotype does not contribute to the risk of developing ASD for these variants. This has also been seen with the results of log-linear modelling where the effect of offspring's genotypes were modelled. Additionally, it can be seen that variants with a lower MAF have lower TDT p-value (rs116948313, which has MAF = 0.051, has the lowest TDT p-value of 0.505).

All variants are insignificant for parent-of-origin TDT, meaning there is no excess of maternal or paternal alleles being transmitted to an offspring. Therefore, there is no evidence that observed maternal genetic effects are in fact camouflaged maternal imprinting. The results show a true maternal genetic effect.

Table 7: TDT results for rs116948313 and top 10 significant variants.
Counts of alleles and p-values of classic TDT and parent-of-origin TDT for 10 significant variants and the variant with the strongest maternal effect are shown. The variant with the strongest maternal genetic effect is divided with a horizontal line from the rest of the variants. T is a number of transmitted and NT are non transmitted alleles. TDT was separately calculated only for parents, only mother and only fathers.

rsID	T:NT	$pval_{TDT}$	pat (T:NT)	$pval_{TDTp}$	mat (T:NT)	$pval_{TDTm}$	$pval_{POO}$
rs116948313	419:450	0.293	191:195	0.839	228:255	0.219	0.505
rs12459171	797:789	0.841	419.5:419.5	1.000	377.5:369.5	0.770	0.831
rs1129235	2143:2134	0.891	1090:1092	0.949	1054:1042	0.793	0.816
rs8121	2141:2134	0.915	1090:1090	0.9829	1052:1044	0.8613	0.889
rs2275843	1193:1192	0.984	572.5:575.5	0.929	620.5:616.5	0.910	0.887
rs4535042	1861:1866	0.935	944:950	0.890	917:916	0.981	0.910
rs45617437	440:449	0.763	230.5:240.5	0.645	209.5:208.5	0.961	0.725
rs66493340	2145:2131	0.831	1062:1048	0.744	1082:1084	0.983	0.805
rs2072496	818:828	0.805	426.5:432.5	0.838	391.5:395.5	0.887	0.969
rs11658510	1721:1705	0.785	834.5:814.5	0.622	886.5:890.5	0.924	0.674
rs77173309	1002:982	0.653	478:454	0.432	524:528	0.902	0.511

4.4.2 *Linkage Disequilibrium*

The LD plot in figure 8 shows a number of small and one large haplotype blocks on chromosome 12. Variants, that are in LD on chromosome 12, are all found in LDL Receptor Related Protein 1 gene (*LPR1*). The decisive evidence of those variants in that gene could be caused by linkage disequilibrium with a causal variant that has a maternal genetic effect.

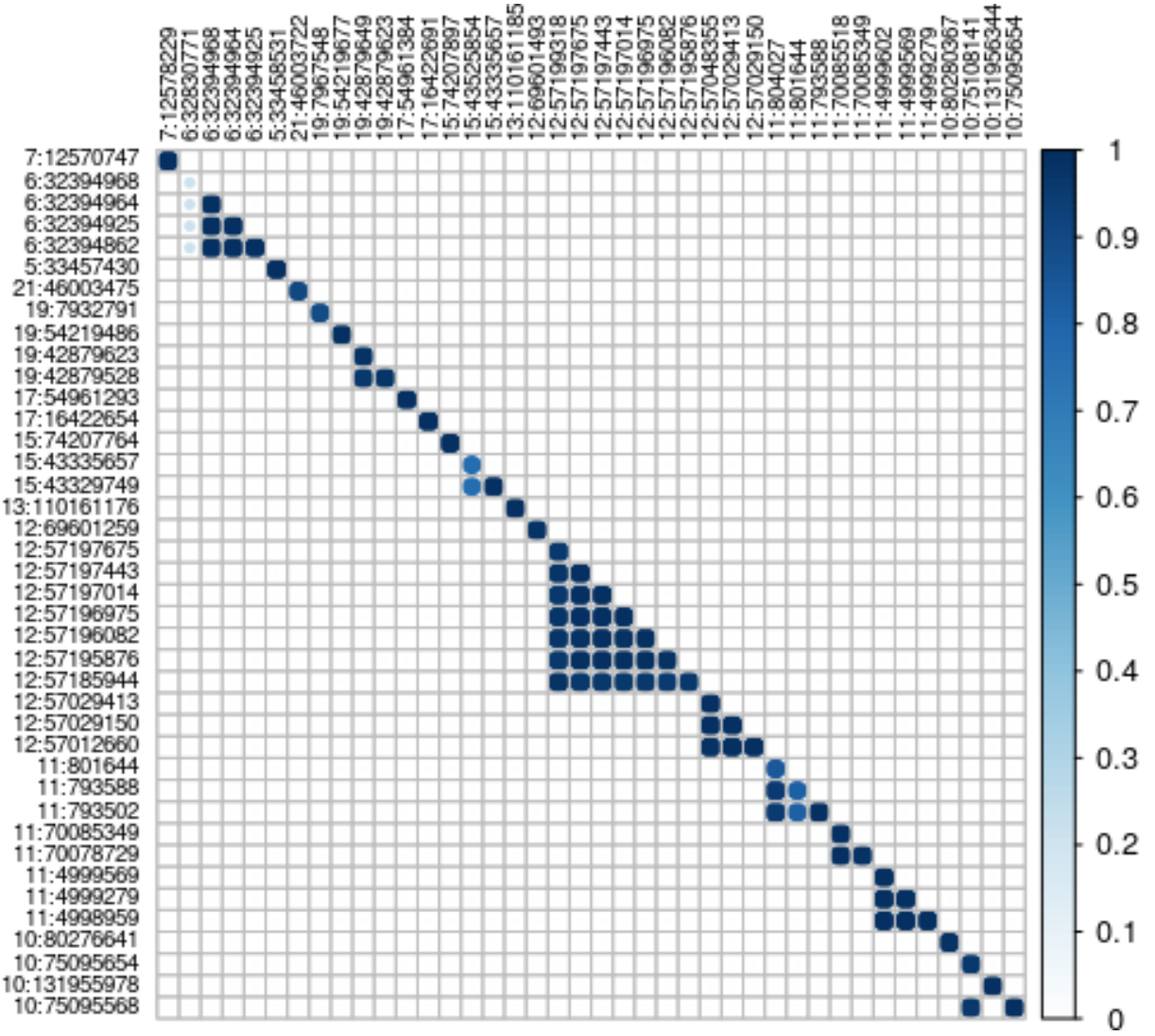

Figure 8: Linkage disequilibrium plot.
Linkage disequilibrium (LD) was tested between the variants that are significant based on the approximate Bayes factor to see the range of LD between hits. The LD plot shows pairs of significant variants that have $r^2 > 0.2$. The location of variants is given as a `chromosome:base pair` position on that chromosome. The strength of linkage disequilibrium is shown with a colour gradient.

5

DISCUSSION

Autism spectrum disorder is a complex disorder with genetic and environmental components where only some of the causal factors are known. One part of this enigma is the effect of maternal genotype on the offspring. The hypothesis is that some variants in mother increase the risk of offspring developing ASD, regardless of the mother's allele being inherited by the offspring. In this thesis, maternal genetic effects were investigated in SPARK dataset.

In this thesis, a preliminary association test was carried out between mothers and fathers to see if there are any variants that exhibit maternal genetic effects. The most significant results were variants in *FCGBP* and *ANKRD36* genes. The minor allele frequency of variants in *ANKRD36* gene were around 0.015, which is very close to cutoffs for rare variants. It is possible that these are false positives, even though a case-control design with trio families increases power to detect an association in variants that have a lower MAF (Tsang et al., 2013).

Unlike other studies, where they used TDT (Tsang et al., 2013; Yuan and Dougherty, 2014) or TDT and association test (Naisha Shah et al., 2012) to investigate maternal effects, I used log-linear modelling which does not need grandparents or healthy controls. Instead, it can use complete and incomplete trios (C. Weinberg, 1999), and in some cases can be more powerful than a TDT (C. Weinberg, A. Wilcox, and R. Lie, 1998). Additionally, I have decided to report the approximate Bayesian factor proposed by Wakefield rather than the p-value. The reason is because it considers prior odds of finding an association, the knowledge about the effect size, and power, which is influenced by MAF and sample size (Wakefield, 2008; Wakefield, 2012).

In *ANKRD36* and *ANKRD36C* genes, no variants had any evidence for maternal genetic effects based on ABF. Variant rs371783424 had estimated S1 = 1.55 but the Bayes factor was 1, meaning there is no evidence for null nor alternative hypothesis. A large maternal genetic effect in variants with low MAF could be anticipated if the sample size for the variant is too small, but ABF accounts for sample size, and in this case makes rs371783424 insignificant. In *FCGBP* gene, there was one variant that had anecdotal evidence for the alternative hypothesis. These results showed us that the significant variants from association analysis are most likely to be false positives when power, prior knowledge of the effect size and prior odds are considered.

Based on ABF, 147 out of 126,532 variants had decisive evidence for the alternative hypothesis. The prior odds for ABF are chosen based on researchers' belief which can be questioned. Connolly, Anney, et al. (2017) investigated the effect of the chosen prior odds for the Bayesian factor on their ASD dataset and decided that $\pi_0 = 0.9995$ is the right choice. Having similar data, I decided on the same prior

odds. Furthermore, the significance of these results was tested with permutation tests. For the top results, variant rs4535042 was not significant after 500 permutations as can be seen in table 4. Out of the initial 147 variants, 115 remained significant after 500 permutations. When using Bayes factor, up to two false discoveries per 10 true discoveries are expected, depending on the sample size (Wakefield, 2012). For 147 variants, the expected number of false positives would be 25, but in this case it is 32. However, it should be noted that only 500 permutations were done. A minimum of 1,000 of permutations would give us more precise results. The genes discussed in the following paragraphs are all significant after 500 permutations.

The most significant variant was detected in *PSG1* gene that has not been directly linked to ASD. However, a study showed association with pre-eclampsia where they observed lower levels of serum protein PSG1 in women with pre-eclampsia compared to healthy pregnant women (Temur et al., 2020). This pregnancy complication is associated with an increased risk of ASD in offspring who are 1.25 times more likely to have ASD compared to offspring whose mother had a successful pregnancy (Jenabi et al., 2019; Maher et al., 2020). PSG1 is a member of pregnancy-specific beta-1-glycoproteins that are secreted from trophoblast cells of the placenta (Bohn, 1971) and act as immunomodulators to prevent the maternal immune system from rejecting the fetus. Additionally, they induce an immune response in case of an infection or an inflammation at the placenta-uterine border. PSG1 induces growth factors: transforming growth factor beta 1 (TGFB1) and vascular endothelial growth factor A (VEGFA). These growth factors have roles in immune tolerance, trophoblast invasion, and vascular development throughout the pregnancy to ensure the proper development of a fetus (Ha et al., 2010). If the vascular structure is underdeveloped and not enough trophoblasts invade the uterine arteries, the fetal brain does not get an adequate amount of nutrients to develop properly, which could result in neurodevelopmental disorders.

The strongest maternal genetic effect was observed in scinderin gene (*SCIN*), which is highly expressed in placenta (source: The Human Protein Atlas). It is mentioned in one meta-study as one of known CNVs associated with ASD (Ch'ng et al., 2015), but no other mention was found in autism-related papers. The SCIN is a Ca^{2+}-dependent F-actin filament-severing protein that modifies the microfilament network in plasma membranes. By modifying the plasma membrane, it has a role in secretion by controlling a cytoskeleton and regulating the pool of vesicles ready to be released, as well as the rate of exocytosis. It was shown that the increased severing activity of scinderin can increase serotonin release from platelets (Marcu et al., 1996). Interestingly, different studies observed elevated levels of serotonin in the blood of ASD individuals (Schain and Freedman, 1961; Abramson et al., 1989; Piven et al., 1991), making it a primary candidate biomarker for identifying ASD (Gabriele, Sacco, and Persico, 2014). It might be possible that the observed variant in mothers changes the scinderin activity triggering the secretion of a larger number of serotonin molecules. During pregnancy, mothers' serotonin molecules

could enter the brain of a developing fetus when the blood brain barrier (BBB) is still permeable. A higher concentration of serotonin causes loss of serotonin terminals that leads to decreased oxytocin levels and increased production of a calcitonin-gene related peptide (CGRP). These physiological changes have been observed in previous studies of ASD (Hadjikhani, 2010).

One of the interesting findings was the region with high LD on chromosome 12. The length of the region is 11,731 base pairs long, which is not extensive for Caucasians. The regions of LD in Caucasians can span up to 100,000 base pairs (Zhu et al., 2003), but the LD window sized used in this thesis was 1,000 base pairs. The *LRP1* gene has various *de novo*, common, and rare variants associated with multiple neurodevelopmental disorders, one of them being ASD. It was shown that the accumulation of the truncated gene leads to more severe autism (Torrico et al., 2019). LRP1 is a cell surface protein with various functions in different pathways, one of them being lipid metabolism. Terrand et al. (2009) showed that this protein stimulates the Wnt5a pathway that is important for cell proliferation and differentiation in the development of a fetus and adults, as well as accumulation of cholesterol. Defective LRP1 will not stimulate Wnt5a pathway, which will result in cholesterol not being properly stored. Moreover, it is highly expressed in placenta, where it has a role in lipid transport and serves as a hem receptor. Normal placental transfer of lipids is essential for fetal development. *LRP1* gene has been associated with ASD in offspring and it should be considered as a candidate gene for maternal genetic effects as well.

Unfortunately, replicating the findings from other studies was not completely successful, which could be due to a different methodology, sample size, or test statistics, or the observed associations were false. A study done by Naisha Shah et al. (2012) used autism genome project (AGP) dataset, that consists of multiplex and simplex families, where the association with maternal genetic effects was found in *ABCC11* gene. Variants found in *ABCC11* gene had substantial evidence ($log_{10}BF = 1.74,\ S1 = 1.085$) in our dataset. The main findings from Connolly, Anney, et al. (2017) were not replicated. They used SSC dataset that consisted of only simplex families. Only *GNB1L* gene had variants with strong evidence for maternal genetic effects. Some evidence was found for *CNTN4*, *CNTNAP2*, *MACROD2*, *LAMA1*, *SDK1*, and *NFIA* genes, which had maternal genetic effects in the Tsang et al. (2013) paper that used Early Marker for Autism (EMA) dataset. However, it was mostly substantial or strong, except for nine variants in *CNTN4* gene that had decisive evidence and the maternal effect was around 1.2. A good replication dataset for this study would be a dataset consisting of multiplex and simplex families from Europe or North America with similar sample size. ASD datasets are limited, but among the ones mentioned here, AGP fulfils most of the criteria making it the most appropriate replication cohort.

There are several limitations to this study. First is the choice of MAF threshold in QC. During the literature search, a MAF threshold of 0.01 seemed appropriate based

on sample size and number of variants, especially because the ABF statistic takes MAF into the account during calculations. However, a paper by Wakefield shows a drastic decrease in power to detect association for variants that have MAF lower than 0.1 (Wakefield, 2008). For log-linear modelling, I have decided to correct it and take only those variants that pass the 0.05 MAF threshold. Secondly, maternal imprinting was only investigated with TDT to see if there is any excess of maternal alleles being transmitted to offspring. In the future, it would be preferable to incorporate the imprinting in the `WeinbergLRT` function. In addition, the maternal-fetal interactions were not investigated. Thirdly, the replication of the results on an independent population was not done which is considered as a golden standard in GWAS. Replicating the same procedure on an independent population with the same study design rules out the bias and gives statistical evidence for the observed association.

Nevertheless, these results indicate that there are maternal genetic effects involved in offspring's development that increase the risk for ASD. The strongest maternal genetic effect was 1.29, which is a relatively weak effect, but multiple genes are showing weak maternal genetic effects indicating the complexity of this disorder. The results also show how the same genes can be risk factors for multiple neurological disorders, making them harder to use in diagnostic purposes.

The next steps in interpreting the results of this study would be replication in an independent population. The replicated variants would indicate that mutations in those genes in mothers are risk factors for having an offspring with ASD. These findings can help in better understanding of the mechanisms in mothers that are involved in the development of ASD in offspring. In the future, it could be used in the prevention or early diagnosis during pregnancy.

6

CONCLUSION

A combination of log-linear modelling and Bayesian statistics was used in the detection of maternal genetic effects in SPARK dataset. Out of 126,531 variants, 147 variants in mothers were identified as potential risk factors in the development of ASD in offspring, but the results ought to be replicated on an independent dataset.

The most significant variant was found in *PSG1* gene, which is linked to pre-eclampsia in women. Although it is an indirect association with ASD, this pregnancy complication could create a deficient environment and affect normal fetus development.

The *SCIN* gene has exhibited the strongest maternal genetic effect. It is hypothesized that the SCIN protein could disrupt normal transport of serotonin molecules *in utero* and affect the child's development.

Furthermore, numerous variants in *LRP1* gene showed consistent strong maternal genetic effects, as they are probably in a linkage disequilibrium with the causal variant. The LRP1 protein is important in lipid transport and the results indicate its involvement in the normal fetus development.

The maternal imprinting for top results was checked with parent-of-origin TDT, which gave an assurance that the observed effects are maternal genetic effects and not maternal imprinting.

The effect of the identified maternal variants on the expression of proteins and their specific role in molecular pathways is unknown, but the results indicate possible changes to *in utero* environment, which could lead to abnormal fetus development.

The overlap between risk factors for ASD and other neurodevelopmental disorders is considerable, making the prediction of a specific disorder difficult. Discovery of more risk factors could improve predictions, as well as treatment in the future. In order to achieve that, further research in the architecture of neurodevelopmental disorders is warranted.

g

APPENDIX

In listings 1 and 2, the modified parts of function colEMlrt from trio package are shown. In listing 3, an original code is shown for permuting labels from fam file.

If you are interested in getting a part of commands or all commands and Unix scripts that were used, contact me at kvucinic@stud.biol.pmf.hr .

Listing 1: WeinbergLRT function, part 1

```
WeinbergLRT <- function (mat.snp, model = c("general", "dominant", "recessive"
    ), child = TRUE, maternal = FALSE, parentMissing = c("father", "mother", "
    either"), iter = 40, eps = 10^-16) {

...
 estR1 <- seR1 <- estR2 <- seR2 <- estS1 <- seS1 <- estS2 <- seS2 <- rep.int(
     NA, n.snp)

# matrix for mating types
X.null <- matrix(c(rep(c(0, 1, 0), c(0, 1, 14)), c(0, 1, 1, 0, 1, 0, 1, rep(0,
     8)), c(0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0), rep(c(0, 1, 0, 1, 0),
     c(5, 1, 3, 1, 5)), c(0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 1,  0)), 15)

# matrix for S1 and S2
M.full <- matrix(c(c(0, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 1, 1, 0, 0), c(0, 0, 0,
    0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1)), 15)

# matrix for R1 and R2 depending on the model
if (child) {
        if (type == "general") {
                Z.full <- matrix(c(rep(0:1, c(11, 4)), rep(c(0, 1, 0),
                c(4, 7, 4))), 15)
                df <- 2
        }
        else if (type == "recessive") {
                Z.full <- rep(0:1, c(11, 4))
                df <- 1
        }
        else if (type == "dominant") {
                Z.full <- rep(0:1, c(4, 11))
                df <- 1
        }
}else{df <- 1} # maternal
# full model. null is Xnull
if (maternal == T & child == T) {
X.full <- cbind(X.null, M.full, Z.full)
}
else if (maternal == F && child == T) {
X.full <- cbind(X.null, Z.full)
}
else if (maternal == T && child == F) {
X.full <- cbind(X.null, M.full)
}
else {print("There are no nested models to be compared")}

...
```

Listing 2: WeinbergLRT function, part 2

```
# EM algorithm
        ...
# log-linear modelling
        ...
if (child) {
 if (maternal) {
        estR1[i] <- summary(ll.full)$coefficients[10,1]
        seR1[i] <- summary(ll.full)$coefficients[10,2]
        estR2[i] <- summary(ll.full)$coefficients[9,1]
        seR2[i] <- summary(ll.full)$coefficients[9,2]
        estS1[i] <- summary(ll.full)$coefficients[7,1]
        seS1[i] <- summary(ll.full)$coefficients[7,2]
        estS2[i] <- summary(ll.full)$coefficients[8,1]
        seS2[i] <- summary(ll.full)$coefficients[8,2]
        df <- 2
 }else{
        estR1[i] <- summary(ll.full)$coefficients[8,1]
        seR1[i] <- summary(ll.full)$coefficients[8,2]
        estR2[i] <- summary(ll.full)$coefficients[7,1]
        seR2[i] <- summary(ll.full)$coefficients[7,2]
        estS1[i] <- seS1[i] <- estS2[i] <- seS2[i] <- 0
        df <- 1
 }
# child = FALSE
}else{
 if (maternal) {
        estS1[i] <- summary(ll.full)$coefficients[7,1]
        seS1[i] <- summary(ll.full)$coefficients[7,2]
        estS2[i] <- summary(ll.full)$coefficients[8,1]
        seS2[i] <- summary(ll.full)$coefficients[8,2]
        estR1[i] <- seR1[i] <- estR2[i] <- seR2[i] <-0
        df <- 1
 }else{
 estR1[i] <- seR1[i] <- estR2[i] <- seR2[i] <- estS1[i] <- seS1[i] <- estS2[i]
     <- seS2[i] <- df <- 0
  }
 }
}
stat <- ll.red_dev - ll.full_dev
pval <- pchisq(stat, df, lower.tail = FALSE)
names(ll.red_dev) <- names(ll.full_dev) <- names(pval) <- names(stat) <-
    colnames(mat.snp)
out <- data.table(SNP = colnames(mat.snp), estR1 = estR1, seR1 = seR1, estR2 =
     estR2, seR2 = seR2,
estS1 = estS1, seS1 = seS1, estS2 = estS2, seS2 = seS2, ll.red = ll.red_dev /
    -2,
ll.full = ll.full_dev/-2, stat = stat, pval = pval)
return(out)
}
```

Listing 3: Permutation script

```
library(data.table)

fam <- fread("sorted_trios_QC.fam")
colnames(fam) <- c("FID", "IID", "FATID", "MOTID", "SEX", "PHENO")

for (task in 1:500) {
        no_trios <- length(unique(fam$FID))
        coin <- sample(1:2, no_trios, replace = TRUE)

        for (i in 1:length(coin)) {
                fam[(1+i*3-3), SEX := coin[i]]
                fam[(2+i*3-3), SEX := ifelse(coin[i] == 1, 2,1) ]
                fam[(3+i*3-3),  ':=' (FATID = ifelse(coin[i] == 1, fam[(1+i*
                    3-3), IID],fam[(2+i*3-3), IID]), MOTID = ifelse(coin[i] ==
                    1, fam[(2+i*3-3), IID],fam[(1+i*3-3), IID]))]
        }

        new_fam <- fam[order(FID, FATID, MOTID, SEX)]

        name_file_indiv <- paste0("/path/fam_perm_indiv", sep = "_", task, ".
            csv")
        name_file_par <- paste0("/path/fam_perm_par", sep = "_", task, ".csv")
        name_file_sex <- paste0("/path/fam_perm_sex", sep = "_", task, ".csv")

        fwrite(new_fam[,1:2], file = name_file_indiv, row.names = FALSE, col.
            names = FALSE, sep = "\t")
        fwrite(new_fam[FATID != 0 & MOTID != 0, 1:4], file = name_file_par,
            row.names = FALSE, col.names = FALSE, sep = "\t")
        fwrite(new_fam[,c(1:2,5)], file = name_file_sex, row.names = FALSE,
            col.names = FALSE, sep = "\t")
}
```

BIBLIOGRAPHY

Abramson, Ruth K, Harry H Wright, Richard Carpenter, William Brennan, Osvaldo Lumpuy, Elisabeth Cole, and S Robert Young (1989). "Elevated blood serotonin in autistic probands and their first-degree relatives." In: *Journal of Autism and Developmental Disorders* 19.3, pp. 397–407.

Agrawal, Sachin, Shripada C Rao, Max K Bulsara, and Sanjay K Patole (2018). "Prevalence of autism spectrum disorder in preterm infants: a meta-analysis." In: *Pediatrics* 142.3, e20180134.

Bailey, Anthony, An Le Couteur, I Gottesman, P Bolton, E Simonoff, E Yuzda, and M Rutter (1995). "Autism as a strongly genetic disorder: evidence from a British twin study." In: *Psychological medicine* 25.1, pp. 63–77.

Berg, Sanne van den, Jérémie Vandenplas, Fred A van Eeuwijk, Aniek C Bouwman, Marcos S Lopes, and Roel F Veerkamp (2019). "Imputation to whole-genome sequence using multiple pig populations and its use in genome-wide association studies." In: *Genetics Selection Evolution* 51.1, pp. 1–13.

Bohn, H (1971). "Detection and characterization of pregnancy proteins in the human placenta and their quantitative immunochemical determination in sera from pregnant women." In: *Archiv fur Gynakologie* 210.4, pp. 440–457.

Boycott, Arthur Edwin, C Diver, SL Garstang, and FM Turner (1931). "II. The inheritance of sinistrality in Limnæa peregra (Mollusca, Pulmonata)." In: *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character* 219.462-467, pp. 51–131.

Brody, Lawrence C, Mary Conley, Christopher Cox, Peadar N Kirke, Mary P McKeever, James L Mills, Anne M Molloy, Valerie B O'Leary, Anne Parle-McDermott, John M Scott, et al. (2002). "A polymorphism, R653Q, in the trifunctional enzyme methylenetetrahydrofolate dehydrogenase/methenyltetrahydrofolate cyclohydrolase/formyltetrahydrofolate synthetase is a maternal genetic risk factor for neural tube defects: report of the Birth Defects Research Group." In: *The American Journal of Human Genetics* 71.5, pp. 1207–1215.

Bush, William S and Jason H Moore (2012). "Genome-wide association studies." In: *PLoS Comput Biol* 8.12, e1002822.

Buyske, Steven (2008). "Maternal genotype effects can alias case genotype effects in case–control studies." In: *European Journal of Human Genetics* 16.7, pp. 784–785.

Cardon, Lon R and John I Bell (2001). "Association study designs for complex diseases." In: *Nature Reviews Genetics* 2.2, pp. 91–99.

Ch'ng, Carolyn, Willie Kwok, Sanja Rogic, and Paul Pavlidis (2015). "Meta-analysis of gene expression in autism spectrum disorder." In: *Autism Research* 8.5, pp. 593–608.

Christensen, Jakob, Therese Koops Grønborg, Merete Juul Sørensen, Diana Schendel, Erik Thorlund Parner, Lars Henning Pedersen, and Mogens Vestergaard (2013). "Prenatal valproate exposure and risk of autism spectrum disorders and childhood autism." In: *Jama* 309.16, pp. 1696–1703.

Cnattingius, Sven, Marie Reilly, Yudi Pawitan, and Paul Lichtenstein (2004). "Maternal and fetal genetic factors account for most of familial aggregation of preeclampsia: a population-based Swedish cohort study." In: *American journal of medical genetics Part A* 130.4, pp. 365–371.

Connolly, Siobhan, Richard Anney, Louise Gallagher, and Elizabeth A Heron (2017). "A genome-wide investigation into parent-of-origin effects in autism spectrum disorder identifies previously associated genes including SHANK3." In: *European Journal of Human Genetics* 25.2, pp. 234–239.

Connolly, Siobhan and Elizabeth A Heron (2015). "Review of statistical methodologies for the detection of parent-of-origin effects in family trio genome-wide association data with binary disease traits." In: *Briefings in Bioinformatics* 16.3, pp. 429–448.

Crean, Angela J and Russell Bonduriansky (2014). "What is a paternal effect?" In: *Trends in ecology & evolution* 29.10, pp. 554–559.

Davis, James O, Jeanne A Phelps, and H Stefan Bracha (1995). "Prenatal development of monozygotic twins and concordance for schizophrenia." In: *Schizophrenia bulletin* 21.3, pp. 357–366.

DiStefano, Johanna K and Darin M Taverna (2011). "Technological issues and experimental design of gene association studies." In: *Disease Gene Identification*. Springer, pp. 3–16.

Evangelou, Evangelos, Thomas A Trikalinos, Georgia Salanti, and John PA Ioannidis (2006). "Family-based versus unrelated case-control designs for genetic associations." In: *PLoS Genet* 2.8, e123.

Feliciano, Pamela, Amy M Daniels, LeeAnne Green Snyder, Amy Beaumont, Alexies Camba, Amy Esler, Amanda G Gulsrud, Andrew Mason, Anibal Gutierrez, Amy Nicholson, et al. (2018). "SPARK: a US cohort of 50,000 families to accelerate autism research." In: *Neuron* 97.3, pp. 488–493.

Frazier, Thomas W, Lee Thompson, Eric A Youngstrom, Paul Law, Antonio Y Hardan, Charis Eng, and Nathan Morris (2014). "A twin study of heritable and shared environmental contributions to autism." In: *Journal of autism and developmental disorders* 44.8, pp. 2013–2025.

Gabriele, Stefano, Roberto Sacco, and Antonio M Persico (2014). "Blood serotonin levels in autism spectrum disorder: a systematic review and meta-analysis." In: *European Neuropsychopharmacology* 24.6, pp. 919–929.

Gaugler, Trent, Lambertus Klei, Stephan J Sanders, Corneliu A Bodea, Arthur P Goldberg, Ann B Lee, Milind Mahajan, Dina Manaa, Yudi Pawitan, Jennifer Reichert, et al. (2014). "Most genetic risk for autism resides with common variation." In: *Nature genetics* 46.8, pp. 881–885.

Gilly, Arthur, Daniel Suveges, Karoline Kuchenbaecker, Martin Pollard, Lorraine Southam, Konstantinos Hatzikotoulas, Aliki-Eleni Farmaki, Thea Bjornland, Ryan Waples, Emil VR Appel, et al. (2018). "Cohort-wide deep whole genome sequencing and the allelic architecture of complex traits." In: *Nature communications* 9.1, pp. 1–9.

Giovedí, Silvia, Anna Corradi, Anna Fassio, and Fabio Benfenati (2014). "Involvement of synaptic genes in the pathogenesis of autism spectrum disorders: the case of synapsins." In: *Frontiers in pediatrics* 2, p. 94.

Good, Irving John (1992). "The Bayes/non-Bayes compromise: A brief review." In: *Journal of the American Statistical Association* 87.419, pp. 597–606.

Ha, Cam T, Julie A Wu, Ster Irmak, Felipe A Lisboa, Anne M Dizon, James W Warren, Suleyman Ergun, and Gabriela S Dveksler (2010). "Human pregnancy specific beta-1-glycoprotein 1 (PSG1) has a potential role in placental vascular morphogenesis." In: *Biology of reproduction* 83.1, pp. 27–35.

Hadjikhani, Nouchine (2010). "Serotonin, pregnancy and increased autism prevalence: is there a link?" In: *Medical hypotheses* 74.5, pp. 880–883.

Hallmayer, Joachim, Sue Cleveland, Andrea Torres, Jennifer Phillips, Brianne Cohen, Tiffany Torigoe, Janet Miller, Angie Fedele, Jack Collins, Karen Smith, et al. (2011). "Genetic heritability and shared environmental factors among twin pairs with autism." In: *Archives of general psychiatry* 68.11, pp. 1095–1102.

Hampe, Jochen, Andre Franke, Philip Rosenstiel, Andreas Till, Markus Teuber, Klaus Huse, Mario Albrecht, Gabriele Mayr, Francisco M De La Vega, Jason Briggs, et al. (2007). "A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1." In: *Nature genetics* 39.2, pp. 207–211.

Hirschhorn, Joel N and Mark J Daly (2005). "Genome-wide association studies for common diseases and complex traits." In: *Nature reviews genetics* 6.2, pp. 95–108.

Hoggart, Clive J, Paul F O'Reilly, Marika Kaakinen, Weihua Zhang, John C Chambers, Jaspal S Kooner, Lachlan JM Coin, and Marjo-Riitta Jarvelin (2012). "Fine-scale estimation of location of birth from genome-wide single-nucleotide polymorphism data." In: *Genetics* 190.2, pp. 669–677.

Höglund, Julia, Nima Rafati, Mathias Rask-Andersen, Stefan Enroth, Torgny Karlsson, Weronica E Ek, and Åsa Johansson (2019). "Improved power and precision with whole genome sequencing data in genome-wide association studies of inflammatory biomarkers." In: *Scientific reports* 9.

Homer, Nils, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig (2008). "Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays." In: *PLoS Genet* 4.8, e1000167.

Jakkula, Eveliina, Karola Rehnström, Teppo Varilo, Olli PH Pietiläinen, Tiina Paunio, Nancy L Pedersen, Ulf deFaire, Marjo-Riitta Järvelin, Juha Saharinen, Nelson

Freimer, et al. (2008). "The genome-wide patterns of variation expose significant substructure in a founder population." In: *The American Journal of Human Genetics* 83.6, pp. 787–794.

Jamain, Stéphane, Hélène Quach, Catalina Betancur, Maria Råstam, Catherine Colineaux, I Carina Gillberg, Henrik Soderstrom, Bruno Giros, Marion Leboyer, Christopher Gillberg, et al. (2003). "Mutations of the X-linked genes encoding neuroligins NLGN3 and NLGN4 are associated with autism." In: *Nature genetics* 34.1, pp. 27–29.

Jenabi, Ensiyeh, Manoochehr Karami, Salman Khazaei, and Saeid Bashirian (2019). "The association between preeclampsia and autism spectrum disorders among children: a meta-analysis." In: *Korean journal of pediatrics* 62.4, p. 126.

Jensen, Liselotte E, Analee J Etheredge, Karen S Brown, Laura E Mitchell, and Alexander S Whitehead (2006). "Maternal genotype for the monocyte chemoattractant protein 1 A (-2518) G promoter polymorphism is associated with the risk of spina bifida in offspring." In: *American Journal of Medical Genetics Part A* 140.10, pp. 1114–1118.

Jirtle, Randy L and Jennifer R Weidman (2007). "Imprinted and more equal." In: *Am Sci* 95, pp. 143–149.

Johnson, William G, Steven Buyske, Audrey E Mars, Madhura Sreenath, Edward S Stenroos, Tanishia A Williams, Rosanne Stein, and George H Lambert (2009). "HLA-DR4 as a risk allele for autism acting in mothers of probands possibly during pregnancy." In: *Archives of pediatrics & adolescent medicine* 163.6, pp. 542–546.

Kanai, Masahiro, Toshihiro Tanaka, and Yukinori Okada (2016). "Empirical estimation of genome-wide significance thresholds based on the 1000 Genomes Project data set." In: *Journal of human genetics* 61.10, pp. 861–866.

Kazma, Rémi and Julia N Bailey (2011). "Population-based and family-based designs to analyze rare variants in complex diseases." In: *Genetic epidemiology* 35.S1, S41–S47.

Klei, Lambertus, Stephan J Sanders, Michael T Murtha, Vanessa Hus, Jennifer K Lowe, A Jeremy Willsey, Daniel Moreno-De-Luca, W Yu Timothy, Eric Fombonne, Daniel Geschwind, et al. (2012). "Common genetic variants, acting additively, are a major source of risk for autism." In: *Molecular autism* 3.1, pp. 1–13.

Kraft, Peter and David G Cox (2008). "Study designs for genome-wide association studies." In: *Advances in genetics* 60, pp. 465–504.

Lawson, Heather A, James M Cheverud, and Jason B Wolf (2013). "Genomic imprinting and parent-of-origin effects on complex traits." In: *Nature Reviews Genetics* 14.9, pp. 609–617.

Lobo, I (2008). "Genomic imprinting and patterns of disease inheritance." In: *Nat. Educ* 1, p. 66.

Loomes, Rachel, Laura Hull, and William Polmear Locke Mandy (2017). "What is the male-to-female ratio in autism spectrum disorder? A systematic review and

meta-analysis." In: *Journal of the American Academy of Child & Adolescent Psychiatry* 56.6, pp. 466–474.

MacDonald, William A (2012). "Epigenetic mechanisms of genomic imprinting: common themes in the regulation of imprinted regions in mammals, plants, and insects." In: *Genetics research international* 2012.

Mackay, Trudy FC (2014). "Epistasis and quantitative traits: using model organisms to study gene–gene interactions." In: *Nature Reviews Genetics* 15.1, pp. 22–33.

Maher, Gillian M, Christina Dalman, Gerard W O'Keeffe, Patricia M Kearney, Fergus P McCarthy, Louise C Kenny, and Ali S Khashan (2020). "Association between preeclampsia and autism spectrum disorder and attention deficit hyperactivity disorder: an intergenerational analysis." In: *Acta Psychiatrica Scandinavica* 142.4, pp. 348–350.

Marcu, Monica G, Li Zhang, Kerstin Nau-Staudt, and José-Trifaró Trifaró (1996). "Recombinant scinderin, an F-actin severing protein, increases calcium-induced release of serotonin from permeabilized platelets, an effect blocked by two scinderin-derived actin-binding peptides and phosphatidylinositol 4, 5-bisphosphate." In: *Blood* 87.1, pp. 20–24.

Melamed, Alexander and Julian N Robinson (2018). "A study design to identify associations: Study design: observational cohort studies." In: *BJOG: An International Journal of Obstetrics & Gynaecology* 125.13, pp. 1776–1776.

Modabbernia, Amirhossein, Josephine Mollon, Paolo Boffetta, and Abraham Reichenberg (2016). "Impaired gas exchange at birth and risk of intellectual disability and autism: a meta-analysis." In: *Journal of autism and developmental disorders* 46.5, pp. 1847–1859.

Montreuil, Bernard, Yves Bendavid, and James Brophy (2005). "What is so odd about odds?" In: *Canadian journal of surgery* 48.5, p. 400.

Murthy, Aditya, Yun Li, Ivan Peng, Mike Reichelt, Anand Kumar Katakam, Rajkumar Noubade, Merone Roose-Girma, Jason DeVoss, Lauri Diehl, Robert R Graham, et al. (2014). "A Crohn's disease variant in Atg16l1 enhances its degradation by caspase 3." In: *Nature* 506.7489, pp. 456–462.

Nelson, Matthew R, Hannah Tipney, Jeffery L Painter, Judong Shen, Paola Nicoletti, Yufeng Shen, Aris Floratos, Pak Chung Sham, Mulin Jun Li, Junwen Wang, et al. (2015). "The support of human genetic evidence for approved drug indications." In: *Nature genetics* 47.8, pp. 856–860.

Ozaki, Kouichi, Yozo Ohnishi, Aritoshi Iida, Akihiko Sekine, Ryo Yamada, Tatsuhiko Tsunoda, Hiroshi Sato, Hideyuki Sato, Masatsugu Hori, Yusuke Nakamura, et al. (2002). "Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction." In: *Nature genetics* 32.4, pp. 650–654.

Palmer, Christina GS, Hsin-Ju Hsieh, Elaine F Reed, Jouko Lonnqvist, Leena Peltonen, J Arthur Woodward, and Janet S Sinsheimer (2006). "HLA-B maternal-fetal genotype matching increases risk of schizophrenia." In: *The American Journal of Human Genetics* 79.4, pp. 710–715.

Pavlica, Mirjana (Nov. 30, 2020). URL: http://www.genetika.biol.pmf.unizg.hr/pogl17.html.

Piven, Joseph, Guochuan Tsai, Eileen Nehme, Joseph T Coyle, Gary A Chase, and Susan E Folstein (1991). "Platelet serotonin, a possible marker for familial autism." In: *Journal of Autism and Developmental Disorders* 21.1, pp. 51–59.

Psychiatric Genomics Consortium, Cross-Disorder Group of the et al. (2013). "Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis." In: *The Lancet* 381.9875, pp. 1371–1379.

Reich, David, Kumarasamy Thangaraj, Nick Patterson, Alkes L Price, and Lalji Singh (2009). "Reconstructing Indian population history." In: *Nature* 461.7263, pp. 489–494.

Reynolds, Lauren C, Terrie E Inder, Jeffrey J Neil, Roberta G Pineda, and Cynthia E Rogers (2014). "Maternal obesity and increased risk for autism and developmental delay among very preterm infants." In: *Journal of Perinatology* 34.9, pp. 688–692.

Ritchie, Marylyn D and Kristel Van Steen (2018). "The search for gene-gene interactions in genome-wide association studies: challenges in abundance of methods, practical considerations, and biological interpretation." In: *Annals of translational medicine* 6.8.

Rosenberg, Rebecca E, J Kiely Law, Gayane Yenokyan, John McGready, Walter E Kaufmann, and Paul A Law (2009). "Characteristics and concordance of autism spectrum disorders among 277 twin pairs." In: *Archives of pediatrics & adolescent medicine* 163.10, pp. 907–914.

Schain, Richard J and Daniel X Freedman (1961). "Studies on 5-hydroxyindole metabolism in autistic and other mentally retarded children." In: *The Journal of pediatrics* 58.3, pp. 315–320.

SCIN expression in placenta (Feb. 5, 2021). URL: https://www.proteinatlas.org/ENSG00000006747-SCIN/tissue.

SFARI Base (Nov. 18, 2020). URL: https://www.sfari.org/resource/sfari-base/.

Shah, Naisha et al. (2012). "Maternal Genetic Effects in Autism Spectrum Disorder."

Song, Jae W and Kevin C Chung (2010). "Observational studies: cohort and case-control studies." In: *Plastic and reconstructive surgery* 126.6, p. 2234.

SPARK registration (Nov. 18, 2020). URL: https://sparkforautism.org/.

Stephens, Matthew and David J Balding (2009). "Bayesian statistical methods for genetic association studies." In: *Nature Reviews Genetics* 10.10, pp. 681–690.

Tam, Vivian, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré, and David Meyre (2019). "Benefits and limitations of genome-wide association studies." In: *Nature Reviews Genetics* 20.8, pp. 467–484.

Temur, Muzaffer, Gülçin Serpim, Sabiha Tuzluoğlu, Fatma Nurgül Taşgöz, Elif Şahin, and Emin Üstünyurt (2020). "Comparison of serum human pregnancy-specific beta-1-glycoprotein 1 levels in pregnant women with or without preeclampsia." In: *Journal of Obstetrics and Gynaecology* 40.8, pp. 1074–1078.

Terrand, Járome, Váronique Bruban, Li Zhou, Wanfeng Gong, Zeina El Asmar, Petra May, Kai Zurhove, Philipp Haffner, Claude Philippe, Estelle Woldt, et al. (2009). "LRP1 controls intracellular cholesterol storage and fatty acid synthesis through modulation of Wnt signaling." In: *Journal of Biological Chemistry* 284.1, pp. 381–388.

Thanabalasingham, G, N Shah, M Vaxillaire, T Hansen, T Tuomi, D Gašperíková, Magdalena Szopa, E Tjora, TJ James, P Kokko, et al. (2011). "A large multi-centre European study validates high-sensitivity C-reactive protein (hsCRP) as a clinical biomarker for the diagnosis of diabetes subtypes." In: *Diabetologia* 54.11, pp. 2801–2810.

Thorleifsson, Gudmar, Kristinn P Magnusson, Patrick Sulem, G Bragi Walters, Daniel F Gudbjartsson, Hreinn Stefansson, Thorlakur Jonsson, Adalbjorg Jonasdottir, Aslaug Jonasdottir, Gerdur Stefansdottir, et al. (2007). "Common sequence variants in the LOXL1 gene confer susceptibility to exfoliation glaucoma." In: *Science* 317.5843, pp. 1397–1400.

Tick, Beata, Patrick Bolton, Francesca Happé, Michael Rutter, and Frühling Rijsdijk (2016). "Heritability of autism spectrum disorders: a meta-analysis of twin studies." In: *Journal of Child Psychology and Psychiatry* 57.5, pp. 585–595.

Torrico, Bàrbara, Alex D Shaw, Roberto Mosca, Norma Vivó-Luque, Amaia Hervás, Noèlia Fernàndez-Castillo, Patrick Aloy, Mònica Bayés, Janice M Fullerton, Bru Cormand, et al. (2019). "Truncating variant burden in high-functioning autism and pleiotropic effects of LRP1 across psychiatric phenotypes." In: *Journal of psychiatry & neuroscience: JPN* 44.5, p. 350.

Tsang, Kathryn M, Lisa A Croen, Anthony R Torres, Martin Kharrazi, Gerald N Delorenze, Gayle C Windham, Cathleen K Yoshida, Ousseny Zerbo, and Lauren A Weiss (2013). "A genome-wide survey of transgenerational genetic effects in autism." In: *PloS one* 8.10, e76978.

Wakefield, Jon (2008). "Reporting and interpretation in genome-wide association studies." In: *International Journal of Epidemiology* 37.3, pp. 641–653.

– (2009). "Bayes factors for genome-wide association studies: comparison with P-values." In: *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* 33.1, pp. 79–86.

– (2012). "Commentary: Genome-wide significance thresholds via Bayes factors." In: *International journal of epidemiology* 41.1, pp. 286–291.

Wan, Hongquan, Chunguo Zhang, He Li, Shuxin Luan, and Chang Liu (2018). "Association of maternal diabetes with autism spectrum disorders in offspring: a systemic review and meta-analysis." In: *Medicine* 97.2.

Wang, Xinyuan, Kimberly M Christian, Hongjun Song, and Guo-li Ming (2018). "Synaptic dysfunction in complex psychiatric disorders: from genetics to mechanisms." In: *Genome medicine* 10.1, p. 9.

Weinberg, CR (1999). "Allowing for missing parents in genetic studies of case-parent triads." In: *The American Journal of Human Genetics* 64.4, pp. 1186–1193.

Weinberg, CR, AJ Wilcox, and RT Lie (1998). "A log-linear approach to case-parent–triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting." In: *The American Journal of Human Genetics* 62.4, pp. 969–978.

Wilcox, Allen J, Clarice R Weinberg, and Rolv Terje Lie (1998). "Distinguishing the effects of maternal and offspring genes through studies of "case-parent triads"." In: *American journal of epidemiology* 148.9, pp. 893–901.

Williams, Tanishia A, Audrey E Mars, Steven G Buyske, Edward S Stenroos, Rong Wang, Marivic F Factura-Santiago, George H Lambert, and William G Johnson (2007). "Risk of autistic disorder in affected offspring of mothers with a glutathione S-transferase P1 haplotype." In: *Archives of pediatrics & adolescent medicine* 161.4, pp. 356–361.

Wolf, Jason B and Michael J Wade (2009). "What are maternal effects (and what are they not)?" In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 364.1520, pp. 1107–1115.

Wu, S, F Wu, Y Ding, J Hou, J Bi, and Z Zhang (2017). "Advanced parental age and autism risk in children: a systematic review and meta-analysis." In: *Acta Psychiatrica Scandinavica* 135.1, pp. 29–41.

Xu, Guifeng, Jin Jing, Katherine Bowers, Buyun Liu, and Wei Bao (2014). "Maternal diabetes and the risk of autism spectrum disorders in the offspring: a systematic review and meta-analysis." In: *Journal of autism and developmental disorders* 44.4, pp. 766–775.

Yoo, Heejeong (2015). "Genetics of autism spectrum disorder: current status and possible clinical applications." In: *Experimental neurobiology* 24.4, pp. 257–272.

Yuan, Han and Joseph D Dougherty (2014). "Investigation of Maternal Genotype Effects in Autism by Genome-Wide Association." In: *Autism Research* 7.2, pp. 245–253.

Yunusbaev, Ural, Albert Valeev, Milyausha Yunusbaeva, Hyung Wook Kwon, Reedik Mägi, Mait Metspalu, and Bayazit Yunusbayev (2019). "Reconstructing recent population history while mapping rare variants using haplotypes." In: *Scientific reports* 9.1, pp. 1–9.

Zhu, Xiaofeng, Denise Yan, Richard S Cooper, Amy Luke, Morna A Ikeda, Yen-Pei C Chang, Alan Weder, and Aravinda Chakravarti (2003). "Linkage disequilibrium and haplotype diversity in the genes of the renin–angiotensin system: findings from the family blood pressure program." In: *Genome research* 13.2, pp. 173–181.

Zondervan, Krina T (2011). "Genetic association study design." In: *Analysis of Complex Disease Association Studies*. Elsevier, pp. 25–48.

BIOGRAPHY

I was born 1995 in Houilles, France. After finishing Andrije Mohorovičić high school in Rijeka, I have enrolled in a Bachelor programme in Biology at the Faculty of Science, University of Zagreb. In 2018, I acquired my Bachelor's Degree in Biology, and started to pursue a Master's Degree in Molecular biology.

During my studies, I was involved in various student activities. I have been a member of Biology Students Association (BIUS), where I co-founded a Section for Anatomy, Morphology and Preparation (AMP), attended meeting from other sections and participated in field trips. In my second year of Bachelor programme, I organized a Symposium of Life Sciences' Students (SiSB) at the Faculty of Science and, almost a year later , I helped with organising a Symposium of Biology Students in Europe (SymBioSE) in Zagreb, Croatia. Also, I have been helping as a Undergraduate Teaching Assistant in the course Genetics, doing laboratory work in different groups and participated in multiple events like Night of Biology and Scientific Picnic.