EXAMINATION OF PRETERM BIRTH – RELATED SINGLE NUCLEOTIDE GENE POLYMORPHISMS IN NEWBORNS and MATERNAL GENOME

Master Thesis

Systems biology master program

Vilnius university

STUDENT NAME: Martynas Kairys

STUDENT NUMBER: 2116141

SUPERVISOR: Dr. Alina Urnikytė

SUPERVISOR DECISION: ……

FINAL GRADE ……

DATE OF SUBMISSION: 15 05 2023

CONTENTS

LIST OF ABBREVIATIONS

Single nucleotide polymorphism – SNP

Copy number variation – CNV

Whole – genome sequencing – WGS

Preterm premature rupture of membranes – PPROM

eQTL – expression quantitative trait loci

INTRODUCTION

Preterm birth, defined as the delivery of a baby before the completion of 37 weeks of gestation, is a major public health concern worldwide. It is estimated that around 15 million babies are born prematurely each year, making preterm birth a leading cause of neonatal morbidity and mortality. For a child, preterm birth can result in short and long–term health problems, which include respiratory distress, neurological deficits, developmental delays, and increased risk of mortality. Currently, the risk factors for preterm birth are known. These include poor air and water quality, lack of accessible health, lack of food, certain viruses, infection, and previous complications with pregnancy. Despite advancements in perinatal care, the preterm birth percentage remains at about the same level of 10% approximately.

It is now speculated that genetics also play a part in preterm birth, which has given scientists a new field of research in preterm birth. Research on polymorphisms linked to preterm birth is being conducted in several populations, but no specific risk variant or gene has been identified. Studies show that different populations have different risk variants associated with preterm birth. In Lithuania, research like this has not yet been completed. Because gathering data of preterm-born babies is complicated, this paper focuses on checking, whether healthy, term-born Lithuanian babies have the same variants, as other risk variants from other populations are being reported.

The majority of preterm births, around 70%, are spontaneous with about a half being without any apparent cause. Genetic factors are one of the significant risks for preterm birth (Sheikh *et al.*, 2016). The focus of this review is on single nucleotide polymorphisms, that are reported to have an association with preterm birth.

AIM AND TASKS

Aim:

To analyze and evaluate single nucleotide gene polymorphisms related to preterm birth in healthy Lithuanian newborn and maternal genomes.

Tasks:

1. Perform a literature search for candidate genes associated with preterm birth.
2. Parse and transform whole genome variation data, gathered from the Lithuanian population to appropriate formats for subsequent analysis.
3. Annotate genomic variation data in newborn and maternal cohorts.
4. Identify DNA variants in the fetal and maternal genomes that can alter the risk of spontaneous preterm birth.

Single Nucleotide Polymorphism

Single nucleotide polymorphisms, also known as SNPs (pronounced “snips”) and sometimes called single nucleotide variants (SNVs), are the most common genomic variation amongst humans. SNP is a genomic variant, at a single base position in the DNA strand. Statistically, one SNP is found in every 500 - 700 bases, meaning that the average person has around 2.5 to 3.5 million SNPs in his/her genetic code (Lavebratt and Sengul, 2006; Shen *et al.*, 2013). Scientists have found around 100 million SNPs in different populations (Spencer, Zhang and Pfeifer, 2015). If the SNP is found in more than 1% of the population, it is called “polymorphism”. If the occurrence is less than 1% it is called a “mutation” (Keats and Sherman, 2013).

Mostly, these variants are benign and do not cause any damage to the organism. Nonetheless, these variants can also cause a disease or a significant enough change, that it causes all sorts of pathologies. Because of this, SNPs are widely used as biological markers to find specific genes that are associated with diseases (U.S. National Library of Medicine, 2020). SNPs can be inherited, meaning that all cells of the organism have a different nucleotide in the DNA chain. Also, SNPs can be somatic, meaning that only some part of the cells will have a different nucleotide, compared to the reference genome. Somatic SNPs usually affect one tissue, and the most common phenotype associated with somatic polymorphism is cancer (Spencer, Zhang and Pfeifer, 2015).

SNPs can be classified using different classification methods. One of these is based on their position in the DNA chain. These are: gene coding SNPs, intergenic SNPs and perigenic SNPs. Intergenic SNPs are SNPs that occur in a strand of DNA between two genes, while perigenic SNPs occur in a non-coding region (Vallejos-Vidal *et al.*, 2020). Gene coding SNPs are rarely found, compared to other types of SNPs. Nonetheless, this group plays a big part in the development of genetic diseases. Gene coding SNPs also can be further divided into synonymous and non-synonymous SNPs according to the change in the amino acid sequence (S. Yang *et al.*, 2020). If the amino acid sequence is changed, then the SNP is called non-synonymous, if it is not changed, then the SNP is synonymous. Synonymous change is further classified as missense (when the change in the sequence results in an amino acid sequence change) and nonsense change (when a change affects the length of the amino acid sequence).

Also, polymorphisms can be classified differently, according to the effect done to the DNA chain. Using the later classification SNPs can be classified into single-base deletion, insertion, transition, and transversion. Transition is a process in which purine switches to pyrimidine or vice versa, meaning that A or G switches to T or C. Transversion is a similar process to transition, the only difference is that the base structure or the core ring does not switch, meaning that pyrimidine switches to pyrimidine (A switches to G or vice versa) and purine switches to purine (T switches to C or vice versa) (Butler, 2012).

Preterm Birth

If the baby is delivered alive before the 37th week of pregnancy, the birth is considered preterm. In affluent nations, preterm birth, also known as premature births, is the main cause of perinatal mortality and morbidity. Preterm births are further classified as early gestation births (<34 weeks) and late preterm births (34 – 37 weeks of gestation) (Chawanpaiboon *et al.*, 2019). Premature births account for 75% of perinatal mortality, and more than 50% of long–term morbidity (McCormick, 1985). Even though preterm delivered babies survive with the help of current medicine and incubators, this can still lead to neurodevelopmental impairments, and respiratory and gastrointestinal complications (Goldenberg *et al.*, 2008).

Around 15 million babies are born preterm each year, with the global percentage of preterm births being 11%. Preterm birth accounts for up to 18% of all deaths of children under the age of 5 and as much as 35% of all deaths among newborns (Walani, 2020). Even though, babies' chance of survival, who are born after the 34th week of gestation is regarded as the same as a baby, born after 37 weeks. Preterm birth could lead to various complications. According to statistics, around 1 in 10 babies that are born prematurely will have a permanent disability, such as blindness, deafness, lung diseases, or palsy (*Premature birth statistics*, no date). According to statistics provided by CDC, preterm births in the USA rose by 0.4% in 2021, from 10.1% in 2020 to 10.5% in 2021 (*Preterm Birth*, 2022).

Precursors leading to preterm birth are: induced or preterm labor caesarean section, spontaneous preterm labor with intact membranes, and preterm premature rupture of membranes (PPROM). Induced labor and preterm labor caesarean section are classified together and called indicated preterm birth. Indicated preterm birth is defined as a life birth, before 37 weeks of gestation and induced for medical reasons. These include pre-eclampsia, intrauterine growth restriction or fetal distress. However, indicated preterm births account for only 30% of all preterm births. Other 70% are spontaneous preterm births. These include PPROM and spontaneous preterm labor with intact membranes (Goldenberg *et al.*, 2008). Physiological events that trigger spontaneous preterm birth are largely unknown. Additional risk factors include decidual haemorrhage, mechanical factors, hormonal changes, and infections (Di Renzo, Roura and European Association of Perinatal Medicine-Study Group on Preterm Birth, 2006). PPROM is a complication of pregnancy, in which the amniotic sac ruptures. This could lead to leakage of amniotic fluid, which surrounds and protects the fetus in the uterus. PPROM could have consequences for both mother and her child, such as infection, preterm birth, fetal distress, and developmental problems (Goldenberg *et al.*, 2008).

However, the exact mechanism of preterm labor and why it happens is unclear. Despite risk factors being reported, most preterm births lack a risk factor (Vogel *et al.*, 2018). However, risk factors in general have been described. Firstly, the ethnicity of a mother is thought to be one of the factors. According to research, black, African – American and African – Caribbean women are

consistently reported as giving preterm birth more often (Fiscella, 1996; Goldenberg *et al.*, 1996; Vogel *et al.*, 2018). Why these women are affected more often than white or Asian women is currently unknown. The age of the mother is also thought to be related to preterm birth, with adolescent women and advanced maternal-age women having a higher percentage. The interval between pregnancies is also regarded as a risk factor. Placental, uterine, and fetal conditions, such as placenta previa, fetal birth defects, and placental abruption, are also regarded as risk factors. These are biological risk factors. There are more of them, including infections, pregnancy with multiples, and previous pregnancies resulting in preterm birth (Vogel *et al.*, 2018; Griggs *et al.*, 2020). Without biological risk factors, there are also socioeconomic and environmental factors. Socioeconomic factors are usually most perceived in third–world countries, and include lack of quality health care, racism, chronic stress, lack of food, low income, poverty, and so on (Griggs *et al.*, 2020). Lastly, environmental factors also play a part in preterm births. According to some researchers increase in preterm birth percentage is associated with the population and industrialization of the country (Goldenberg *et al.*, 2008). Environmental factors include air and water quality, smoking, alcohol use, substance use, poor diet habits, long working hours, and long periods of standing (Griggs *et al.*, 2020).

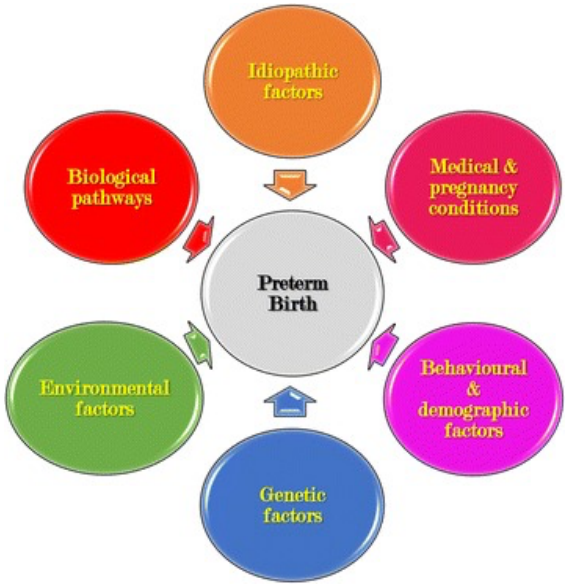

Figure 1. Risk factors associated with preterm birth.
Source: (Sheikh et al.*, 2016).*

Genetic Association with Preterm Birth

Genetics also plays an important role in preterm birth. The number of studies on genetic polymorphism in relation to preterm birth has dramatically increased since the IOM report in 2007 (*Evidence-Based Medicine and the Changing Nature of Health Care*, 2008). Genomic variations and mutations of both mother and child are important in preterm births. Genetics may contribute to

around 40% of preterm births (Svensson *et al.*, 2009). However, this information is dispersed and not available on a single platform or database.

Maternal genomic variation

Preterm births are a complex trait, that cannot be explained with a variation or mutation in a single gene. SNPs found in the maternal genome have been identified in various chromosomes and genes, which shows that preterm birth is a process involving a lot of different biological pathways.

A study by Zhang et al., 2015 has shown, that SNP rs17053026 is the most statistically significantly associated with preterm births. This SNP is located on chromosome 3 and exhibits an odds ratio of 0.44. The affected gene is *DCP1A*, which encodes an RNA-decapping enzyme. This enzyme plays a role in the transcription regulation (Zhang *et al.*, 2015). The same study has shown 4 different SNPs (rs6989497, rs6987111, rs6989156, and rs7823365) located in chromosome 8, with all being found in a single gene *CCDC25*. This gene encodes a protein called Coiled-Coil Domain Containing 25, which is involved in cell division and proliferation. Moreover, this protein has connections with various cancer forms (L. Yang *et al.*, 2020; Siriphak *et al.*, 2021). However, the functionality of this protein in preterm birth is poorly understood, but it looks to be highly conserved across species (Zhang *et al.*, 2015). The odds ratio of 5.22 has been reported for SNP rs12066169, which is located on chromosome 1 encompassing *PAX7* gene. This gene is a member of paired box family of transcription factors, which is involved in fetal development and cancer growth (Zhang *et al.*, 2015). Gene–based analysis showed that the most significant genes are *CCDC25*, *TMEM2*, and *MYPN*. As mentioned before, 4 SNPs were shown in *CCDC25*, and *MYPN* gene contained one SNP, rs6480306. *MYPN* gene codes myopalladin protein, which is found in muscles and is responsible for tethering proteins at the Z-disc and communication between sarcomere and nucleus in cardiac and skeletal muscles (Filomena *et al.*, 2021). However, study noted that none of the found SPNs have reached a genome – wide significance threshold of 5×10^{-8} (Zhang *et al.*, 2015). GWAS research, carried out in a Danish/Norwegian population did not identify any of the SNPs described in previously mentioned study. This shows that different ethnicities could differ in genomic variations associated with preterm birth. After Bonferroni adjustment for multiple testing, none of the SNPs remained significant with a p-value less than 0.05. However, before Bonferroni adjustment, the most promising finding was a SNP in *FRMD7* gene, rs2747022. Before adjustment, the G allele at this SNP was associated with an increased risk of spontaneous preterm birth in Danish and Norwegian data. After including previously tested Argentinian families, the result remained the same (Myking *et al.*, 2013). *FRMD7* encodes FERM domain – containing protein 7, which is involved in the development and function of nerve cells in the retina and is associated with X – linked idiopathic congenital nystagmus. It has also been shown that this protein is found in various parts of the brain in embryos (Tarpey *et al.*, 2006). Furthermore, it has been thought that *FRMD7* coded protein is important for neurite

development and neuronal differentiation (Betts-Henderson *et al.*, 2010). SNP that was closest to significance was rs7892483, which had strong linkage disequilibrium to four other SNPs (rs5972070, rs5972071, rs5973734 and rs5973741). All these SNPs are located in a gene desert, that has not been previously associated with any disease. Several loci associated with diseases have been linked to gene deserts and it is thought that these regions may contain regulatory elements, that can modulate gene expression (Myking *et al.*, 2013).

Maternal genomic variation is now thought to be one of the risk factors associated with preterm birth. Studies of this field usually find different genes and SNPs that are statistically significant, however only on rare occasions these findings are found to be significant on a genome level. Nonetheless, with more studies being conducted, future findings could show significance and help foresee upcoming preterm births according to genetic risk variants.

Fetal genomic variation

Even though mother's genetics play an important role in preterm birth, fetus's genome is also analysed. Most of the studies done on preterm birth topic cover maternal and fetal genomic variations. This suggests, that fetus's genome is also playing part in time of birth.

Statistically significant SNPs have been found in fetus's genomes as well. One of the most significant neonatal SNP was rs17527054, which is located on chromosome 6 comprising the area of major histocompatibility complex (MHC) (Zhang *et al.*, 2015). This complex is related to adaptive immune system and cell surface proteins. The same study also reported a second significant neonatal SNP, which was rs3777722, located in the *RNASET2* gene (Zhang *et al.*, 2015). This gene is located on chromosome 6 as well, and codes a ribonuclease, which's variants have been associated with malignancies and leukoencephalopathy. Other identified SNPs have not reached significance threshold, however they are also of interest. For example, rs184270 is located close to *KCNH7* gene, which codes a member of the potassium voltage – gated channel. Other SNPs have been found in *SMAD9, TGF- β*, *NOL10, RSPO2. L3MBTL3* and *INPP1* genes. Interestingly, study's top 3 SNPs were all located in the *NOL10* gene, on chromosome 2. This gene encodes a nucleolar protein and is highly conservative across species (Zhang *et al.*, 2015). Pairs of SNPs have been reported in *RSPO2* and *RREB1* genes. RSPO2 product is a secreted protein that is connected to WNT signaling pathway, and has been proposed to regulate craniofacial pattering (Jin *et al.*, 2011). Mutations in this gene is connected to limb and lung development and in dogs, a variant of analogic gene is associated with moustache and eyebrow thickness (Cadieu *et al.*, 2009). Product of *RREB1* encodes a zinc finger transcription factor and is located on chromosome 6. This protein is part in a Ras signaling pathway. Mutations in this gene are associated with type 2 diabetes associated end – stage kidney disease (Bonomo *et al.*, 2014). Study found that the most significant genes are

RNASET2, MFSD6, L3MBTL3. In addition, SNP rs3777722 overlaps with *RNASET*2, and SNP rs4429972 overlaps with *L3MBTL2* gene. Already mentioned Danish/Norwegian study also found SNPs and genes in neonatal genome, that are associated with preterm birth. The study found that before Bonferroni correction, SNPs rs4239992, rs17328647, rs5953790 and rs2485792s had statistical significance. All these SNPs are closely linked with the *MIR505*/*ATPC11* gene region (Myking *et al.*, 2013). *MIR505* is an RNA gene, that is affiliated with miRNA class. Esophagus Adenocarcinoma and Borna Disease have been associated with *MIR505* (*MIR505 Gene*, 2023). miRNAs molecules can have important regulatory functions in gene expression, and might be involved in preterm birth (Bartel, 2004). However, none of these SNPs were found to be chromosome – wide significant in the replication study. In addition, SNP rs6652393 was found to be most promising among males, but not in females. This would indicate a sex – specific effect. The SNP is located in *IL1RAPL2* gene, which is mapped to chromosome X, and it codes a protein of the IL – 1 receptor family (Myking *et al.*, 2013). Another study found that this gene is expressed in skin, liver, placenta and fetal brain tissues and only weak expression has been detected in adult brain (Born *et al.*, 2000; Ferrante *et al.*, 2001).

Preterm – premature rupture of membranes could be one of the factors leading to preterm birth. Looking more into PPROM from the genetic perspective could suggest risk variants of PPROM, which can lead to preterm birth. Study conducted in Richmond, Virginia, USA, found that some gene mutations were only found in cases of PPROM. The genes found specifically were *CARD6, NLRP10* and *NOD2*. In addition, SNP in *CARD6* gene (rs number is not provided in the study), changed the C allele to G allele, which generated a stop codon. This led to protein being truncated at position 560, instead of 1037 (Modi *et al.*, 2017). The gene in question, *CARD6*, is a caspase recruitment domain family member 6 protein. This protein is involved in negative regulation of the innate immune response (Strauss *et al.*, 2018). Mutation was found in two cases of PPROM, and none were found in the term pregnancy control group. *NLRP10* mutation was found in one case of PPROM. This mutation led to protein truncating from 655 amino acid long, to only being 103. Nonsense mutation was also found in *DEFB1* protein (SNP rs5743490). This mutation truncated protein to only four amino acids, so no active peptide was made. This gene encodes beta – defensin 1, which is an antimicrobial factor, that is produced by epithelial cells (Modi *et al.*, 2017). According to the study, most of the variations found were nonsense mutations, where because of SNP, the regular codon is changed to a stop codon. In some cases, active peptides are not made, in other cases they lack a specific domain, that is also necessary for correctly functioning peptide.

As it can be seen from these findings, fetal genome should not be overlooked. Not only maternal genome is very important in a term birth, but fetal genome also plays a part. Until today, a single risk variant cannot be identified with a 100% certainty. However, studies in different populations show different results, which is one of the reasons to research ethnically different populations. Variants that were found in Danish/Norwegian population were not present in the study

completed in the USA and vice versa. It could be argued that different populations may have different risk variants that are a contributing factor to preterm birth.

Whole – genome sequencing

As the sequencing of the DNA has become cheaper with introduction of 3rd generation sequencing, whole – genome sequencing (WGS) is gaining popularity. This is a technique used to sequence a complete genome of an organism. The goal of this process is to determine whole, or nearly whole, genome (van El *et al.*, 2013).

Completion of WGS cannot be done without certain preparation. Firstly, because a whole DNA molecule is too big for sequencers to analyse, it must be broken down. This is done by cutting DNA into smaller pieces, so the instrument can correctly determine nucleotides. These smaller DNA molecules are then are marked using small DNA tags, otherwise known as bar codes (*Whole Genome Sequencing*, 2022). Barcoded molecules are then analysed using sequencing instruments. This generated data is then analysed using various scientific software, to determine the full genome sequence of an organism. In later years 3rd generation sequencing techniques are used the most. This is done, because 3rd generation sequencers provide longer reads than any generations before, making this generation the most suitable for WGS (Mukhopadhyay, 2009).

WGS has opened a new way to study various diseases in humans. Cancer research is the one that received most advancement using WGS. Cancer tumour WGS analyses led researchers in discovering valuable information, with structural variation of tumour genome being the most explored topic (Zhao, Jones and Jones, 2019). Research with cancer is not the only field that can be studied with WGS. Because of its broad application spectrum, this approach found its way into the clinic. WGS can also be used to identify bacterial strains from isolates, which is helping to pinpoint medicine that would be effective against certain diseases. In addition, WGS approach can be used to determine pathogen's phenotype, which is also helpful in prescribing specific medicine, that could help the patient (Balloux *et al.*, 2018).

With increasing popularity of WGS and its widening application spectrum, this sequencing type could become a useful addition to clinics. This approach could help doctors and scientists determine the pathogen that is affecting a patient, as well as prescribe medicine, that works best against certain types of pathogens. With sequencing technologies moving forward and becoming cheaper, WGS should become more popular approach in diagnostics.

Previous Research

As can be seen from previously described topics, genome analysis of fetal and maternal genomes is important for the understanding of preterm birth. However, genome is different across populations. Furthermore, different populations may have different variants that are associated with preterm birth (Huang, Shu and Cai, 2015).

Comparing the study of Danish/Norwegian (Myking *et al.*, 2013) population with the study conducted in the USA (Zhang *et al.*, 2015) clear differences can be seen. Neither of these studies identified common genes or polymorphism. Each study concluded that different genes and SNPs were statistically significant after performed analysis. This leads to the belief, that population difference is important in studies that are very closely related to population genetic differences. Most of the results after statistical adjustments were insignificant. Genetics are now included in the risk factors of preterm birth. To understand the involvement of it deeper, more extensive research is needed. In addition, different organism systems are also concluded to be involved. Some studies report that innate immunity and inflammation-related gene polymorphisms are associated with preterm birth, others point to vascular or metabolic-related genes (Sheikh *et al.*, 2016). It is now widely believed that genetics indeed play a big role in preterm birth, in addition, the synergy between maternal and fetal genomes is also very important. Analysis of variants associated with preterm birth should be done for both maternal and fetal genomes, to acquire the most accurate results possible.

Similar studies have never been conducted in Lithuania. According to statistics, Lithuanian mothers undergo preterm birth less frequently, compared to other European countries (5.9% of all births in 2008) (Murphy and McLoughlin, 2015). According to statistics published by Lithuanian Hygiene Institute, in 2021, 23,379 children were born. Of these, 1,159 were born preterm, which is approximately 5% (Basys *et al.*, 2022). Consequences of preterm birth range from no difference, between preterm born baby and term born baby, to mental and physiological abnormalities. Because of this, understanding which genes and variants could be associated with preterm birth in the Lithuanian population is an important task to complete.

METHODS

The methods section should be detailed to the level enabling an informed person reading the "Methods" section to repeat the project. Usually methods are structured into data, methodology or algorithms, versioned software used, statistical methods employed, data and software availability and bioethics approval if applicable.

Whole–genome sequencing data

Data used for the analysis was of the project ANELGEMIA (No. S-MIP-20-34), provided by the supervisor of the Master's thesis. Data was a single not annotated VCF file for 25 Lithuanian newborns, and PLINK format MAP and PED files for 25 Lithuanian mothers. These files contained whole genome sequencing variation data. Sequencing and quality control analysis was completed by CeGaT company, located in Tubingen, Germany. 100 ng DNA was sequenced paired–end using Illumina NovaSeq 6000 Sequencing System. Analysis of sequenced genomes was performed using DRAGEN platform, using BWA – MEM and GATK – HC software. Reads were mapped to the reference genome hg19. The quality of the FASTQ files was assured using FastQC software (Urnikyte *et al.*, 2022).

Additional data, a list of candidate genes, was collected by the author of the Master's thesis. Gene list was collected from literature sources and NCBI Genome Data Viewer website. Search term for genes associated with preterm birth was "preterm neonates" (link to the website: https://www.ncbi.nlm.nih.gov/genome/gdv/browser/genome/?id=GCF_000001405.40). Studies from different populations were also used for gene collection. Used literature sources include analysis from Danish/Norwegian, Argentinian population and USA population (Myking *et al.*, 2013; Zhang *et al.*, 2015; Sheikh *et al.*, 2016; Modi *et al.*, 2017). The final candidate gene list contains 51 genes for newborns and 34 genes for mothers. The full lists of genes can be found in Appendix 1 and Appendix 2.

Algorithm design

Analysis parts of the algorithm were written using a makefile. Makefile is a file describing files, rules and commands, which generate final output from a given input executing a single command. Apart from annotation, other parts of the analysis, such as filtering, vcf file acquisition, conversion to map and ped files, Hardy – Weinberg equilibrium calculations and frequency calculations were done in a single makefile. To reproduce the same results of the Master's thesis, it

is sufficient enough to have a written makefile, the same input, and a directory tree. The complete makefile can be found in Appendix 3.

Genomic variant annotation

Received VCF file had to be annotated. This was done using software called ANNOVAR (Wang, Li and Hakonarson, 2010). ANNOVAR is a software introduced in 2010, and still supported and updated to this day. Primarily it is used for single nucleotide variant and insertion/deletion annotation. The program is written using Perl modules and can be used easily after downloading and unzipping the file. Before performing annotation, the person, using ANNOVAR, must decide, what databases will be used for annotation. These databases must be downloaded to the machine a person is working with (Wang, Li and Hakonarson, 2010).

ANNOVAR can use a VCF file as input. This type of file is a standard file format for storing variation data. VCF files are a standard output of variant calling softwares and annotation tools. Each VCF file has a header, which encodes the metadata. There are 8 mandatory columns in the VCF file. These include chromosome number, position, in which the polymorphism was found, the ID of specific polymorphism, reference allele, alternative allele, quality score, filter column and info column.

ANNOVAR can be acquired using the following command line:

```
wget
http://www.openbioinformatics.org/annovar/download/annovar.latest.tar.gz
```

The downloaded file must be unzipped, which can be done using the following command line:

```
gunzip annovar.latest.tar.gz
```

After unzipping, folder annovar can be accessed. Inside this folder, perl scripts and downloaded databases can be found, along with example files, which can be used for training purposes.

For this work, dbSNP, version 1.50 was used, along with refGene database. Databases were downloaded to ZMGKSERV2 using the following command lines:

```
./annovar/annotate_variation.pl –buildver hg19 –downdb –webfrom annovar
refGene humandb
./annovar/annotate_variation.pl –buildver hg19 –downdb –webfrom annovar
avsnp150 humandb
```

Downloaded databases are found in humandb folder, located in main ANNOVAR folder.

After databases are downloaded, annotation can be started. Annotation can be performed using 2 different perl scripts, first one is called annotate_variation.pl, another one is called table_annovar.pl. These scripts do the same thing, however, table_annovar.pl can output the results

into a CSV file format or some other format suitable for the user. Because of this feature, table_annovar.pl was used. The command for annotation of the newborns file:

```
perl ./annovar/table_annovar.pl ANE25_newborns_ok_2023_modchr_rsID.vcf ./annovar/humandb/ –buildver hg19 –out ANE25_newborns_annotted –nastring . –operation f,g –vcfinput –protocol avsnp150,refGene
```

Similar command was used for the annotation of the mothers vcf file. The command used was as follows:

```
perl ./annovar/table_annovar.pl ANE.women.data_vcf.vcf ./annovar/humandb/ –buildver hg19 –out ANE.women_annotted –nastring . –operation f,g –vcfinput –protocol avsnp150,refGene
```

Both command lines generates various outputs. Files contains information for all found variations, genes, SNPs, insertions/deletions, chromosome, location of polymorphism, and other information from given databases. This file can be used for further analysis.

Candidate gene filtering

Filtering, according to collected candidate genes was performed. This was done using a self-written makefile of the author. Makefile is a file describing files, rules and commands, on how to get files from certain input. In this case, annotated file was one input file, list of genes found in literature, that are associated with premature birth, was the second and third inputs. First, it was essential to filter genes associated with premature birth from annotated WGS files. This was done using “awk” and “if” commands in the makefile:

```
$(GENE_DIR_N)/%.txt:
	awk -v x="$(*F)" -F '\t' '{if ($$8 == x) {print $$0}}' $(ANNOTTED_FILE_N) > $@
```

Figure 2. Command in the makefile used to get entry lines according to gene name.

Note that _N at the end of the directory shows, that this specific rule was used to acquire files from newborns’ data. In the makefile, _M shows that the command is used to acquire files from mothers’ data. This is continuous throughout the whole makefile.

These commands generated files under ./outputs/newborns/gene/ and ./outputs/mothers/gene directories and were given a name of the gene, for example, AAA.txt. The file contained all entries, where gene column was equal to provided gene name from the gene list. Furthermore, SNPs list would be beneficial for further analysis. Because of this, rs columns were filtered out from the file generated using previous command, which contained only the rs numbers

of polymorphisms found within that gene. This file was saved under ./outputs/newborns/rs_gene/ and ./outputs/mothers/rs_gene/ directories and were given the name of the gene it was found in. Command used for this purpose:

```
$(RS_DIR_N)/%.txt: $(GENE_DIR_N)/%.txt
        awk '{print $$6}' $< | sed '/\./d' > $@
```

Figure 3. Command in makefile used to get rs numbers.

For further analysis, vcf files were needed. Provided vcf files had to be filtered out according to rs numbers. This was done using following command:

```
$(VCF_DIR_N)/%.vcf: $(RS_DIR_N)/%.txt
        vcftools --vcf $(RAW_FILE_N) --snps $< --out $@ --recode
        mv $@.recode.vcf $@
```

Figure 4. Command in makefile used to filter out the original vcf file and save new file with rows specific to rs number.

mv command on the third line of the picture renames the files. This is done because VCFtools software (Danecek *et al.*, 2011) automatically adds .recode. part in the file name.

The rule takes a single file from ./outputs/newborns/rs_gene/ or ./outputs/mothers/rs_gene directory, and uses the list of rs numbers to filter out starting vcf file, to acquire a vcf files that only have polymorphisms of the gene of interest. The output of this command can be found under ./outputs/mothers/vcf_gene/ and ./outputs/newborns/vcf_gene.

VCF files conversion to MAP and PED files

Further analysis is performed by using PLINK v1.07 (Purcell *et al.*, 2007) software. For completion of the analysis, PLINK format MAP and PED files are necessary. Conversion, of VCF files to MAP and PED files, was done using a makefile rule shown in Figure 5.

```
$(MAP_PED_DIR_N)/%: $(VCF_DIR_N)/%.vcf
        vcftools --vcf $< --out $@ --plink
```

Figure 5. Conversion of VCF file to MAP and PED files using vcftools software in makefile.

Following the completion of a rule, MAP and PED files are generated to ./outputs/newborns/map_ped_gene/ and ./outputs/mothers/map_ped_gene directories, with file name being gene name (for example AAA.map and AAA.ped). In addition, logs of command execution are created, which can be found inside the same directories.

Hardy – Weinberg Equilibrium analysis

Hardy – Weinberg equilibrium states that allele and genotype frequencies in a population will remain constant from generation to generation in the absence of other evolutionary influences. The equilibrium was calculated using PLINK v1.07 software by executing the following rule in makefile:

```
$(HWE_DIR_N)/%: $(MAP_PED_DIR_N)/%
	plink1 --file $< --hardy --out $@
```

Figure 6. Hardy – Weinberg equilibrium calculation rule in makefile.

This command generates 3 files as an output. 1st one is a file containing Hardy – Weinberg equilibrium results and has the extension of .hwe. 2nd file is a list of subjects whose sex was not specified. This file has an extension of .nosex. 3rd generated file is the logs file. All output files can be found in ./outputs/newborns/hwe_gene and ./outputs/mothers/hwe_gene directories for newborns and mothers accordingly. All 3 files are named after the gene, for example, AAA.hwe, AAA.nosex, and AAA.log.

For further analysis steps, hwe files were merged into a single file using cat command in the command line. These files were then transferred to the personal computer of the thesis author, for more convenient analysis processes.

Frequency analysis

Frequency is described as the relative frequency of an allele at a particular locus in a population. This number can be expressed as a fraction or a percentage. Factors affecting frequency include evolutionary factors: mutations, gene drift, migration, and natural selection. To calculate frequencies PLINK v2.0 software (Chang *et al.*, 2015) was used. The command was executed in makefile using the syntax provided in Figure 7.

```
$(FREQ_DIR_N)/%: $(VCF_DIR_N)/%.vcf
	test -s $< | plink2 --vcf $< --freq --out $@ || true
```

Figure 7. Frequency calculation rule in makefile.

Because PLINK v2.0 cannot take files with no variations as an input, this command is a bit more complicated than the others. The first part of the command is the plink2 command, which would throw an error if the input file has no variations. However, the second part of the command (|| true) lets the command continue, even though the first part of the command failed. Meaning that if input vcf has no variations, then the second line will “force” the following command to continue.

Command outputs 2 files for each input file. The first file is a frequencies file, and has an extension of .afreq. The second file is the logs file. Both of these files can be found under ./outputs/newborns/freq and ./outputs/mothers/freq.

Data analysis

Frequency calculation gives information about the variants frequency inside a given dataset. To check if these frequencies are similar to the other population frequencies, online tools can be used. One of these tools is SNPNexus (Oscanoa *et al.*, 2020). This is a web-based variant annotation tool, designed to simplify the SNP annotation. To start the SNPNexus, a user must enter some personal details, including full name, institution, institutional email, and dataset name. After this, the Human Assembly must be chosen. In this work, GRCh37/hg19 assembly was used. After the assembly is chosen, a user can select one of 5 options for data input. The way used in this research was an upload of variations in a vcf file. All SNPs can be put in SNPNexus at the same time. Because of this, concatenation of the files is possible. Firstly, vcf files were concatenated into a single file using the following commands:

```
grep –v ‘^#’ *.vcf > mothers_full.vcf
grep –v ‘^#’ *.vcf > newborns_full.vcf
```

This command finds lines starting with #, reverses the logic, so only lines without # are printed, and moves those lines to an output file. Both files had file names at the start of each line, which is not acceptable for SNPNexus. These names were removed using following commands:

```
sed ‘s/^[^:]*://’ mothers_full.vcf > mothers.vcf
sed ‘s/^[^:]*://’ newborns_full.vcf > newborns.vcf
```

These commands completed the file preparation for SNPNexus. For more simple usage of SNPNexus the prepared files were downloaded to a personal computer via Remote Desktop Manager app. The personal computer is running macOS Monterey Version 12.6.5.

Genetic data analysis in mothers cohort

For mothers’ data, the files provided by the supervisor were in PLINK MAP and PED format. For further analysis with the same pipeline, they had to be converted to a single VCF file. This was done using PLINK v1.07 software with the following command:

```
plink --file ANE.women.data --recode vcf --out ANE.women.data_vcf
```

This created a single vcf file, named ANE.women.data_vcf.vcf, and a log file. The same analysis pipeline as described for newborns was repeated.

Statistical analysis

Calculation of Hardy – Weinberg equilibrium, using PLINK, automatically calculates statistical p–value. PLINK uses chi–squared test to calculate the p–value of Hardy - Weinberg equilibrium. Results of Hardy – Weinberg equilibrium were then transferred to the personal computer of the thesis author and analysed using Microsoft Excel software. Firstly, rows with TEST column value "ALL" were filtered. These rows were again filtered according to the P column (which is a p–value). Results were considered statistically significant if the p–value was less than 0.05.

Versioned Software Used

- For annotation, ANNOVAR software was used (Wang, Li and Hakonarson, 2010). The exact version is not provided. Last update to software itself (not the additional databases), was carried out on 24th of October, 2019.
- Makefile was executed using a GNU Make software, version 4.2.1. GNU Make is a free software that can be installed using a command line. Specifically this analysis was ran on ZMGKSERV2 of VU MF BMI ZMGK, which is a server with Linux operating system installed.
- For conversion of VCF files to map and ped files, vcftools v 0.1.16 software was used (Danecek *et al.*, 2011). VCFtools is a software than can be installed to a machine from cloning a repository in GitHub using the following command line:

  ```
  git clone https://github.com/vcftools/vcftools.git
  ```

  When the repository is downloaded, compilation of VCFtools is completed by executing the following commands:

  ```
  ./autogen.sh
  ./configure
  make
  sudo make install
  ```

  In case a person is working on a maintained server, only the administrator of the server can install vcftools to the machine (indicated by sudo command)
- For calculations of Hardy – Weinberg equilibrium, PLINK v1.07 software was used (Purcell *et al.*, 2007). PLINK v1.07 can be downloaded using the following website: https://zzz.bwh.harvard.edu/plink/download.shtml. Here a user also has to choose correct operating system. After downloading, command `plink` must be executed for full compilation of the software.

- For frequency calculation PLINK v2.0 was used (Chang *et al.*, 2015). This software can be downloaded in the following website: https://www.cog-genomics.org/plink/2.0/. Here user has to choose correct operating system.
- SNPNexus is a web-based tool used for SNP annotation (Dayem Ullah *et al.*, 2018). Tool can be accessed by using the following website: https://www.snp-nexus.org/v4/. Website link shows, that the version of SNPNexus is v4.

RESULTS

Results section will be split into smaller sections. One is for newborns' data results, the other is for mothers' data results.

Newborns cohort

Annotation analysis results

After annotation was performed, 10 files were generated. One of the generated files was with the extension .avinput. This is a file that ANNOVAR generates by default. It is a file that is generated from input vcf file, so that ANNOVAR software can read it, annotate it, and analyse it. In addition, logs file is also generated during annotation. Other files can be categorized into 2 groups: one is for refGene database, the other one is for avsnp150 database. Avsnp150 database files contain file which has only SNPs, the other file is where all SNPs are dropped. refGene database files contain separate log, just for annotation with this specific database as well as exonic variant function file, variant function file and a sequence of genes that were determined to have variation (something similar to reference genome). The file of interest is a vcf and txt files (safe files with different extensions), having both refGene and avsnp150 annotations. Both files were used. txt file was used for gene filtering and rs number acquisition, and vcf file was used to generate separate vcf files for polymorphisms of a specific gene.

Used txt file contained 8565435 lines, where every line is an information about specific allele and a polymorphism. VCF file, contained 7602755 lines, with same information. The number of lines differ because of specific formatting of the files. Both files were used in the further analysis process.

Looking at log files, the annotation was completed successfully.

Filtering and VCF file acquisition results

The initial annotated file contained 7,602,516 polymorphisms. Filtering of the VCF file with candidate gene list showed that polymorphisms were not found in 22 genes. 29 other genes left contained from 10 to 2574 polymorphisms. The complete number of polymorphisms found in genes can be seen in Table 1.

Table 1. Number of polymorphisms found in genes associated with preterm birth.

Gene name(s)	Polymorphisms found
ACE, AMOT, APOE, ATP11C, COMT, DACH2, FRMD7, IGF1, IL1RAPL2, LPHN2, MIR505, MMP9, MTHRF, NOS2A, NOS3A, PLAC1, SMAD9, SORL1, TGFB1, TLR7, UTP14A, VDR	0
ADD1	286
ADIPOQ	64
COL24A1	2,291
CRP	13
CXCL8	18
EGR1	11
IL10	14
IL1B	24
IL6	30
INPP1	166
KCNH7	1,276
L3MBTL3	472
LEP	64
MAN1A1	913
MBL2	47
NOL10	852
NPPB	10
PPARG	589
RNASET2	214
RREB1	533
RSPO2	836
SP3	332
SPOCK3	2,574
TFAP2A	71
TLR4	83
TMEM229A	10
TNF	10
VEGFA	68

The gene with the most variations was *SPOCK3* gene, which encodes a member of calcium–binding proteoglycan proteins, that contain thyroglobulin type-1 and Kazal – like domains. The gene with the least variations found was *NPPB*, which encodes a secreted protein which functions as a cardiac hormone. Looking at the filtering of rs numbers, all of the detected polymorphisms already had an rs number, meaning that all of the found polymorphisms were already described in the avsnp150 database.

After generating the VCF files, all of the files had at least 231 lines. This is because VCF files have specific headers, which accumulate to that amount of lines. The gene files in VCF format, that had no polymorphisms detected, were populated with a VCF file header. However, looking at empty files, it can be seen, that after the header is finished no other line of polymorphisms can be seen (Figure 7.)

```
##bcftools_annotateVersion=1.14+htslib-1.14
##bcftools_annotateCommand=annotate -c ID -a /opt/home/alina/ANE_analize/dbSNP1>
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  wgs_S34>
```

Figure 7. VCF file of a gene (ACE gene), that had no polymorphisms detected.

VCF files, of genes, that had detected polymorphisms had additional lines after the header lines. An example of this can be seen in Figure 8.

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	wgs_S34
5	148205372	rs2053044	A	G	172.04	PASS	.		
5	148205503	rs758739682	G	GT	47.82	PASS	.		
5	148205505	rs780755859	CA	C	46.94	PASS	.		

Figure 8. VCF file of a gene (ADRB2 gene), that had polymorphisms detected.

Looking at the files, it can be concluded that the acquisition of VCF files has been completed successfully. The FILTER column of each polymorphism is filled with the value “PASS”. The acquired VCF files have been used for the further analysis process.

Hardy – Weinberg Equilibrium Calculation analysis

Hardy – Weinberg calculations yielded 58 files. The completed command generates 2 files. Knowing this, there were 29 genes, which generated 2 files each. This is consistent result throughout the research, where genes with no polymorphisms detected generate empty files or files with no information, besides the header. The generated hwe file can be seen in Figure 9.

GNU nano 4.8 ADRB2.hwe

CHR	SNP	TEST	A1	A2	GENO	O(HET)	E(HET)	P
5	rs2053044	ALL	A	G	0/16/5	0.7619	0.4717	0.01538
5	rs2053044	AFF	A	G	0/0/0	-nan	-nan	NA
5	rs2053044	UNAFF	A	G	0/0/0	-nan	-nan	NA
5	rs758739682	ALL	G	GT	0/1/0	1	0.5	1
5	rs758739682	AFF	G	GT	0/0/0	-nan	-nan	NA
5	rs758739682	UNAFF	G	GT	0/0/0	-nan	-nan	NA
5	rs780755859	ALL	CA	C	0/1/0	1	0.5	1
5	rs780755859	AFF	CA	C	0/0/0	-nan	-nan	NA
5	rs780755859	UNAFF	CA	C	0/0/0	-nan	-nan	NA
5	rs750087033	ALL	TCC	T	0/1/0	1	0.5	1
5	rs750087033	AFF	TCC	T	0/0/0	-nan	-nan	NA
5	rs750087033	UNAFF	TCC	T	0/0/0	-nan	-nan	NA
5	rs12654778	ALL	G	A	0/14/5	0.7368	0.4654	0.03984
5	rs12654778	AFF	G	A	0/0/0	-nan	-nan	NA
5	rs12654778	UNAFF	G	A	0/0/0	-nan	-nan	NA
5	rs146649973	ALL	C	G	0/1/0	1	0.5	1
5	rs146649973	AFF	C	G	0/0/0	-nan	-nan	NA
5	rs146649973	UNAFF	C	G	0/0/0	-nan	-nan	NA
5	rs11168070	ALL	G	C	0/16/5	0.7619	0.4717	0.01538

Figure 9. Generated Hard – Weinberg equilibrium (hwe) file.

Each file contains 9 columns. CHR column specifies, in which chromosome the polymorphism is found, SNP denotes the rs number of the SNP that is being analysed. TEST column specifies whether the statistical analysis was conducted on affected, or unaffected people (case-control studies). Because this is not a case–control study, the results for AFF and UNAFF rows are NA or -nan. A1 column show what is the minor allele, and A2 specifies the major allele. GENO column represents the number of homozygous and heterozygous subjects. For example, in the second row of the file, GENO column is filled with 0/16/5, this means that 16 people had homozygous genotype of GG, and 5 people were heterozygous. O(HET) and E(HET) columns represent the observed and expected heterozygosity. In such cases as can be seen in row 5, only 1 person had the SNP of number rs758739682. Because the sample size is not sufficient for this specific polymorphism, both heterozygosity calculations are not correct. The last column is the Hardy – Weinberg equilibrium P value, calculated by using chi–squared test with one degree of freedom.

After analysing the data, it was found that out of 27,960 entry lines, only 1,050 of the found polymorphisms had p–value of less than 0.05. According to Hardy – Weinberg equilibrium calculations, 1,050 polymorphisms were not in equilibrium. In almost all of the cases, expected heterozygosity was lower than observed heterozygosity, suggesting that there is an excess of heterozygotes in the population. SNPs were located on chromosomes 1 through 9. Out of these 1050 polymorphisms, 49 were found as having an association with preterm birth or preterm labor according to GAD database. These polymorphisms were found in chromosomes 1, 2, 3, 6, and 9, involving *IL10, IL1B, PPARG, TLR4,* and *VEGFA* genes. *IL10* gene codes interleukin-10 protein, which is an anti-inflammatory cytokine. Similarly, *IL1B* codes for interleukin-1 beta protein, which is also a cytokine. This cytokine is an important mediator of the inflammatory response in a person's

organism. *TLR4* is also associated with immune system, more specifically, with innate immune system. This gene codes for a transmembrane protein, which is fundamentally important in pathogen recognition and activation of innate immunity. As mentioned in the literature analysis, genes associated with the immune system were found by other researchers (Engel *et al.*, 2005; Speer *et al.*, 2006). *PPARG* gene codes for glitazone reverse insulin resistance receptor, which regulates fatty acid storage and glucose metabolism. Looking at the newest literature, this gene or its coded protein is not mentioned in any of the populations as having a heavy impact on preterm birth. However, earlier research showed that there might be a connection between polymorphisms in *PPARG* gene and preterm birth (Meirhaeghe *et al.*, 2007). Lastly, *VEGFA* gene codes vascular endothelial growth factor A, which shows high activity in vascular endothelial cells. As mentioned in the literature analysis, some researchers have shown that genes, associated with vascular activity, may affect preterm birth (Sheikh *et al.*, 2016).

As analysis has shown, 49 polymorphisms were not in Hardy – Weinberg equilibrium and were associated with preterm birth in other researches. Observed heterozygosity was higher for these 49 polymorphisms, suggesting that there could be an excess of these heterozygotes in the Lithuanian population. However, all the subjects were born healthy and not preterm. Possibly, found polymorphisms, in heterozygous state, may not be a risk factor for preterm birth in the Lithuanian population. However, further analysis is needed for definitive conclusion.

Frequency calculation analysis

Frequency calculations yielded 102 files. This is because every file from 51 initial genes generated 2 files each, one is the result file, and the other is the log file. The files with no variations also generated temporary .psam files, which are not used in the analysis, because they contain only ID of the sample. 29 files have the necessary information. A sample of .afreq file can be seen in Figure 10.

#CHROM	ID	REF	ALT	ALT_FREQS	OBS_CT
4	rs35780153	C	T	0.59375	32
4	rs1019911648	G	C	0.5	4
4	rs999958800	C	T	0.5	2
4	rs1877723	C	T	0.59375	32
4	rs16843452	C	T	0.625	16
4	rs12503220	G	A	0.55	20
4	rs397823169	A	AT	0.5	8
4	rs1024112114	G	C	0.5	2
4	rs2096760	C	T	0.59375	32
4	rs73189448	A	G	0.625	16
4	rs766678687	CTTT	C	0.5	4
4	rs776294620	CTTTCTTT	CCTTT,C	0.125,0.375	8
4	rs769626222	T	TC	0.5	8
4	rs762733079	CTTT	C	0.5	2
4	rs371676635	T	TC,TCTTTC	0.35,0.3	20

Figure 10. Example of frequency file (.afreq).

The file contains chromosome number, rs number of the SNP, reference allele, alternative allele, frequency of the alternative allele appearing in the dataset, and OBS_CT column. OBS_CT column, as described in PLINK manual is observed count column. This column is the observed count of minor allele. It is calculated using the formula provided below:

$$OBS\_CT = minor\ homozygote\ count + heterozygote\ count\ \times\ 2\ .$$

If the polymorphism contains more than one nucleotide, or more alternative alleles found, they are also listed in the columns provided.

SNPNexus frequency data analysis

SNPNexus tool was used to acquire information about other populations, as well as additional information from various databases, such as ClinVar, COSMIC, CADD, and GAD. Dataset used data only from those genes, which were described in literature most often and the most likely to have an effect on preterm birth. These genes were *VEGFA, ADIPOQ, IL6, IL10* and *TNF.*

Analysis was separated into 4 different approaches. Each approach is used to analyse data from different sources. These sources are 1000 Genomes Project, HapMap, gnomAD, and genome gnomAD. Starting with the 1000 Genomes project, contained data from 5 different super populations. These are American, African, East Asian, European, and South Asian super populations. Data was analysed using Microsoft Excel program. Out of 144 found polymorphisms by SNPNexus, 87 polymorphisms found were associated with preterm birth. All these polymorphisms had a higher frequency in the European super population, compared to the calculated frequency in Lithuanians. However, only 22 polymorphisms were in Hardy – Weinberg disequilibrium. All of these SNPs had a lower calculated frequency, compared to European super population data. Polymorphisms rs833067 (C > T), rs1413711 (T > C), and rs833070 (T > C) had a lower calculated frequency than African, admixed American and East Asian super populations. Polymorphisms were checked in the eQTL database. All entries did not find any significant eQTLs. However, for all three, same sQTLs were listed. Tissues associated with these polymorphisms are skeletal muscle tissue and heart tissue. All three polymorphisms can be found on chromosome 6 and are located in *VEGFA* gene. As mentioned before, this gene codes a protein, which shows a high activity in vascular endothelial cells. In addition, rare polymorphisms, such as rs3024492 (T > A) were also found to be in disequilibrium and in association with preterm birth. According to frequency calculations, the Lithuanian population frequency was 0.556 compared to the European super population frequency of 0.222. Other super populations had even lower values. rs3024492 is found on chromosome 1 and is located near *IL10* gene. This gene was previously described in Hardy – Weinberg equilibrium analysis. The SNP rs3024492 is reported as an eQTL for *IL10* in whole blood (p = 0.00011), according to the Genotype

– Tissue Expression (GTEx) portal (https://gtexportal.org/home/; accessed on 14/05/23).

HapMap approach contains a lot more populations. In total, there are 12 separate populations. These include Toscani in Italy population, Yoruba in Ibada population, Masai in Kinyawa, Kenya population, Mexican ancestry in Los Angeles population, Luhya in Webuye, Kenya population, Japanese in Tokyo population and many more. However, HapMap project was completed in 2009 and has not been updated since. The data from this project is still useful in some situations, however, it cannot show the newest data available. After HapMap analysis, only 55 polymorphisms were matched and analysed. Out of these, 36 were associated with preterm birth and 9 were in Hardy – Weinberg disequilibrium. As in 1000 Genomes project data, all polymorphisms were found on chromosome 1 in *IL10* gene or on chromosome 6, in *VEGFA* gene. The first thing that can be seen after analysis, is that Unrelated Han Chinese in Beijing population did not have any of the found polymorphisms. This column was filled with the value “None” for all the rows. Polymorphism rs833070 (T > C), according to HapMap data, had a lower calculated frequency, than any other population except Toscani population in Italy. Another interesting find is that polymorphism rs2146323 (C > A) was found in 3 out of 12 populations. These included Utah residents with ancestry from northern and western Europe, Japanese in Tokyo and Yoruba in Ibadan, Nigeria populations. This polymorphism is located on chromosome 6 near *VEGFA* gene. In the same gene, polymorphism rs833069 (T > C) can be found. It's frequency varies greatly among populations, from 0.212 in Gujurati Indians in Houston, to 0.7386 in Luhya in Webuye population. In calculated frequency for Lithuanians this value is 0.5, which is equal to population is South Western US, having African ancestry. SNP rs833069 is reported as an eQTL for *VEGFA* in pancreas ($p = 7.5 \cdot 10^{-11}$), thyroid ($p = 2.8 \cdot 10^{-9}$), adrenal gland ($p = 0.0000057$), esophagus – muscularis ($p = 2.2 \cdot 10^{-7}$) and esophagus - gastroesophageal junction ($p = 0.0000098$), according to GTEx database.

Third approach was using gnomAD database. This project was launched in 2016 and is stil updated. gnomAD database contains information from African/African American population, Latino population, East Asian population, Finnish population, Non – Finnish European population, South Asian population and Other populations. gnomAD database found only 11 polymorphisms from the given list. Out of the found polymorphisms, 8 were associated with preterm birth. Out of these, only 2 were in Hardy – Weinberg disequilibrium. Both of these SNPs were found in chromosome 6, near *VEGFA* gene. In addition, both polymorphisms had a higher frequency, compared to other gnomAD populations. Other 6 genes associated with preterm birth, were found to be in Hardy – Weinberg equilibrium. SNP rs2069860 (A > T), which can be found on chromosome 7, in *IL6* gene, had a lower frequency, compared to other populations in the gnomAD.

Frequency analysis showed that out of the found polymorphisms, only a small part is related to preterm birth and is not in equilibrium according to Hardy – Weinberg calculations. Some of the mentioned polymorphisms had a frequency that was similar to database one. In addition, it must noted that these calculations may have an inaccuracy, because of the small sample size. However,

findings are similar to the literature sources, showing polymorphisms connected to the immune system and vascular system.

SNPNexus phenotype data analysis

Along with frequency data, SNPNexus also can search other databases, including ClinVar, PolyPhen, SIFT databases. All of these databases had only a few polymorphisms found in them, however, this information is also very important.

Starting with the ClinVar database, 3 entries were found. The first one was rs62625753 (G > A), which can be found in chromosome 3 in *ADIPOQ* gene. Clinically this SNP was classified as likely benign. Another found polymorphism was rs2010963 (C > G), which can be found in *VEGFA* gene in chromosome 6. This polymorphism was classified as a risk factor with the phenotype being microvascular complications of diabetes I. And lastly, rs2069860 (A > T) was classified as benign, with no phenotypes found. This SNP can be found in chromosome 7, in *IL6* gene.

PolyPhen database found 2 entries from this dataset. Those being rs62625753 and rs2069860 (A > T). As can be seen, these are 2 out of 3 polymorphisms that were found in ClinVar database. Even though rs2069860 was classified as benign, it cannot be said that the results were identical for rs62625753. PolyPhen found that this polymorphism is probably damaging. Amino acid change is present during this polymorphism. Amino acid changes from glycine to serine. In addition, SIFT database results are identical to PolyPhen database results. SIFT database classifies rs62625753 as deleterious, and rs2069860 as tolerated polymorphism.

The last analysed database was GAD database. This is a genetic association database, which collects genetic association studies from scientific literature. With the input data of newborns, GAD database returned 70,425 entries. A lot of input variants generated more than one result from the database, some of them generating more than a thousand. GAD distinguished 18 classes of diseases out of newborns' data. Times of class occurrence can be seen in Figure 11.

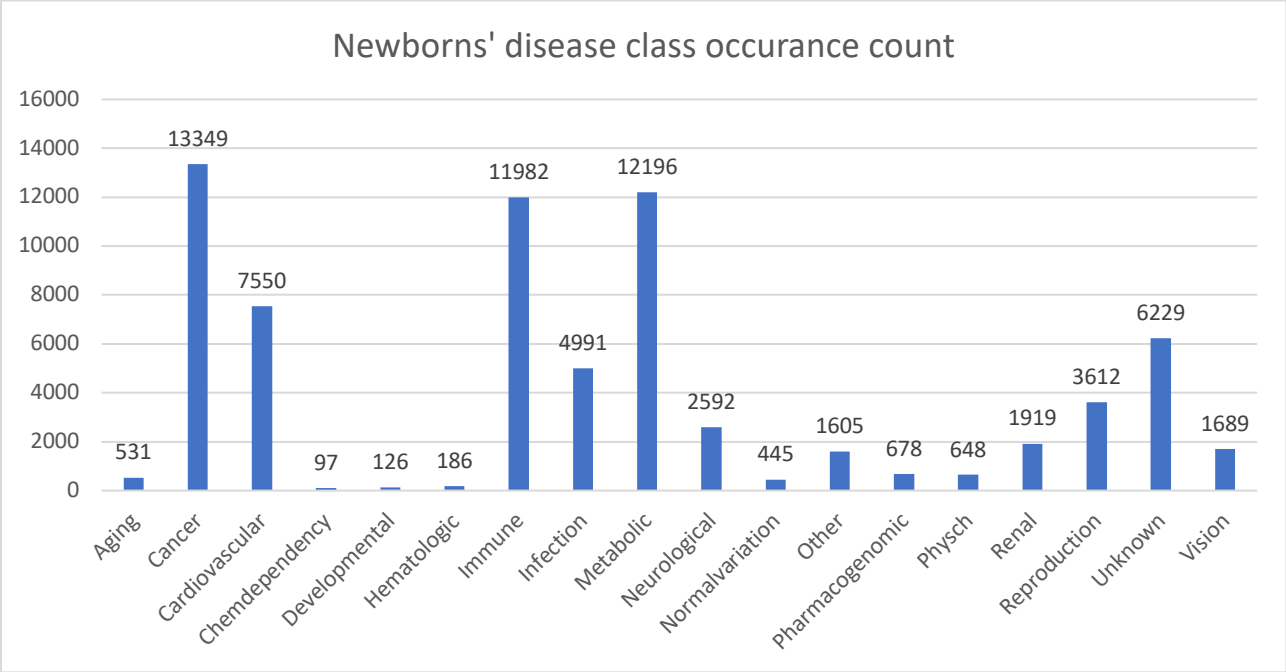

Figure 11. Class occurrence count of newborns' data from GAD database.

As can be seen from the figure, the cancer class had the most associations with input polymorphisms. The result could be explained by the vast amount of cancer types and research magnitude. Another interesting find is with immune disseases class. As mentioned in the literature analysis, immune system polymorphisms play part in preterm birth, both from maternal and fetal genomes perspective. Few polymorphisms found were associated with chemodependency, which is a type of substance use disorder. These polymorphisms were found to be associated with alcoholism, smoking behavior, and cirrhosis. Other disease class includes various disease or conditions, that cannot be classified in other classes. These include Epstein – Barr virus infection, cystic fibrosis, heart transplant complications. In this class, pregnancy loss and preterm delivery are also classified. Out of 1,605 associations found with other disease classes, 26 included recurrent pregnancy loss or preterm delivery. An interesting note is that this is newborns' data. These findings could suggest that these polymorphisms are associated with preterm delivery of newborn's baby, or that they are associated with preterm birth of analysed newborn.

Mothers cohort

Annotation results

Just like using newborns' data, 10 files were generated. The files had the same characteristics as previously described: ANNOVAR input file, log file, separate database files for each database used, VCF and TXT files of interest. Even though the command used for annotation was almost the same, the resulting files had fewer lines than newborns' data. For comparison, the

txt file of newborns had 8,565,435 lines, while mothers' data contained 1,508,949 lines. Even though this file was converted from map and ped files, it did not impact any of the variants that were described in map and ped files. Looking at the log of conversion from map and ped to vcf files, 1,508,984 variants were found in 25 people. Meaning that when annotation is performed 1 top line is added to the total line count. Looking at the logs of annotation, it can be seen that annotation for the mothers' data was performed successfully and can be used in further analysis processes.

Filtering and VCF file acquisition results

The annotated file contained 1,508,984 polymorphisms. Gene list for mothers had 34 genes listed. Out of these, only 6 genes had 0 polymorphisms detected. 28 genes had from 1 polymorphism to 349 polymorphisms detected in one gene. A complete number of polymorphisms can be found in Table 2.

Table 2. The number of polymorphisms found in mothers' genes.

Gene name(s)	Polymorphisms found
CFS2, CYP51A1, HMGCR, RLN2, S100A9, ADRB2	0
ADD1	10
COL1A2	18
COL4A3	35
COL5A1	101
COL5A2	128
CR1	37
DEFA5	4
EBF1	123
EDN1	4
EEFSEC	147
EGR1	3
GSTM1	1
HLA-DQA1	349

HPGD	26
OXTR	12
PPARG	24
PTGER3	120
PTGS1	7
PTGS2	2
SERPINE1	8
SFTPD	9
SP3	7
TFAP2A	2
TLR4	3
VEGFA	10
WNT4	19

Gene with the most polymorphisms was *HLA-DQA1* gene, with 349 polymorphisms found. This gene is a major histocompatibility complex gene. It is found in chromosome 6. Protein encoded by this gene is one of two proteins that are required for DQ heterodimer, which is an essential cell surface receptor for the function of the immune system. Polymorphisms in this gene could lead to dysfunction of immune system. Immune system abnormalities are described in literature as influencing preterm birth. *GSTM1* gene had only 1 polymorphism found. This gene codes for glutathione S – transferase Mu 1. It can be found on chromosome 1 and is functions in the detoxification of electrophilic compounds, including carcinogens, therapeutic drugs, toxins, products of oxidative stress. Looking at files, containing rs numbers, all the polymorphisms had the rs number.

All VCF files had at least 28 lines. This is because of the headers in VCF files. Comparing it to newborns' data, this time the amount of header lines was a lot smaller. It could be because of the formatting of primary MAP and PED files and their conversion to VCF file. Looking at files with exactly 28 lines, no information after the header is present in the file. It can be seen in figure 12.

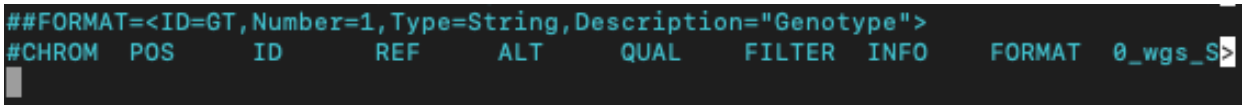

Figure 12. Example of an empty VCF file in mothers' data (ADRB2 gene vcf file is shown in the picture).

Acquired VCF files were used in further analysis steps.

Hardy – Weinberg Equilibrium Calculations

This analysis step generated 90 files. This is because empty files still generated log outputs. According to this, 6 out of 90 files are log files from empty VCF files. Genes with found polymorphisms yielded 3 files each: an hwe file, nosex file, and a log file. An example of an acquired HWE file can be seen in Figure 13.

CHR	SNP	TEST	A1	A2	GENO	O(HET)	E(HET)
3	rs6801826	ALL	C	G	0/5/20	0.2	0.18
3	rs6801826	AFF	C	G	0/0/0	-nan	-nan
3	rs6801826	UNAFF	C	G	0/0/0	-nan	-nan
3	rs4678001	ALL	T	C	0/13/8	0.619	0.4274
3	rs4678001	AFF	T	C	0/0/0	-nan	-nan
3	rs4678001	UNAFF	T	C	0/0/0	-nan	-nan
3	rs10048957	ALL	G	A	0/5/20	0.2	0.18
3	rs10048957	AFF	G	A	0/0/0	-nan	-nan
3	rs10048957	UNAFF	G	A	0/0/0	-nan	-nan

Figure 13. Example of HWE file (ADCY5 gene). In the picture P column is not present.

Just like in newborns' data, it contained the same structure and meaning. After analysis, it was found that out of 4,242 polymorphisms found, only 195 polymorphisms had a lower p–value than 0.05. Just as in newborns' data, these polymorphisms were found to be in disequilibrium, suggesting an excessive amount of heterozygotes in the population. 27 out of 195 polymorphisms were found to be associated with preterm birth. These polymorphisms were found in 4 genes: *COL4A3, COL5A1,* and *OXTR*. Most of these polymorphisms were found in *COL5A1* gene, which codes for collagen alpha 1. It is known that mutations in this gene can cause Ehlers–Danlos syndrome, which is a group of 13 genetic connective tissue disorders. The newest literature does not consider connective tissue problems as having an effect on preterm birth, however, earlier literature has shown some connection (Menon *et al.*, 2009). *COL4A3* gene also codes for collagen protein, more specifically collagen alpha 3. This protein is a major structural component of the basement membranes. Only 1 polymorphism was found to be in disequilibrium and associated with preterm birth in this gene (rs10933170 (T > G)). According to GTEx portal, polymorphism rs10933170 was listed as an eQTL for *COL4A3* in testis. Two polymorphisms were found in association with preterm birth, located in *OXTR* gene. Interestingly, this gene codes for the oxytocin receptor, which is a receptor for hormone and neurotransmitter oxytocin. The newest literature does not list *OXTR* gene polymorphisms as a risk factor for preterm birth, however, older literature sources claimed that polymorphisms in this gene are associated (Romero *et al.*, 2010).

Out of 4,242 polymorphisms found, 195 SNPs had higher observed heterozygosity, than expected heterozygosity, suggesting an excess of heterozygotes in the population. Analysis showed that 27 of these polymorphisms were associated with preterm birth, however knowing that mothers' gave birth on term, suggests that for Lithuanian population these polymorphisms, in heterozygous state, may not have an impact on preterm birth.

Frequency calculations

Frequency calculations using mothers' data yielded 62 files. Out of these 34 were log files, and 28 were .afreq files, used for the frequency analysis. Because this was an identical command line, the generated result structure remained the same. Just like in newborns' data, a file contained chromosome number, rs number, reference allele, alternative allele, frequency of the alternative allele and OBS_CT column, which has the same meaning as in the newborns' data.

SNPNexus frequency data analysis

Just as in newborns' data analysis, mothers' data was also put through SNPNexus analysis. Same information was collected regarding the frequency data from different populations using different databases. Databases are the same as in newborns' data: 1000 Genomes Project, HapMap, gnomAD and specific genome gnomAD. Just as in newborns' data, genes, that were the most mentioned and most likely to have an effect were chosen. This time it was only three genes: *EBF1, ADCY5* and *WNT4*.

The data had 186 polymorphisms, however neither of the frequency data databases found any frequency result, with uploaded VCF file. Even though mothers' dataset had more polymorphisms than newborns' data. Same case was with phenotype data. Neither of the databases, ClinVar, SIFT or PolyPhen found any entries in their databases.

For this reason, different approach was used for the analysis with SNPNexus. Instead of taking vcf files, polymorphism IDs (rs numbers) were concatenated to a single file, and at the beginning of each line "dbsnp" and a tabulation was added. This was done, because of SNPNexus' specifications of the input. Same genes were used for the analysis as was done with VCF files.

With polymorphism IDs data, population data was acquired and analysed. Starting with 1000 Genomes Project super population data, out of the 186 polymorphisms, 176 were found on the 1000 Genomes Project database. However, none of the found polymorphisms had any association with preterm birth. In addition, only 6 polymorphisms were found to be in disequilibrium. These polymorphisms were found in chromosomes 1, 3, and 5. Most of these polymorphisms had a higher calculated frequency than other populations. 1 polymorphism was located in chromosome 1 in the *WNT4* gene, 2 were found on chromosome 3 in *ADCY5* gene and 3 were found on chromosome 5 in *EBF1* gene. Out of 176 polymorphisms, 2 had associations with birth weight. These SNPs were rs9844212 (G > C) and rs11923649 (G > A). Both were found in the *ADCY5* gene and were in Hardy – Weinberg disequilibrium. Variant rs9844212 (G > C) had a calculated frequency of 0.667, which

was higher than any other population frequency. European population frequency was listed as 0.591. *ADCY5* gene is coding for adenylyl cyclase type 5 protein, which was shown to interact with RGS2 protein. RGS2 protein is thought to have protective effects against myocardial hypertrophy and atrial arrhythmias (Nunn *et al.*, 2010). The other SNP, rs11923649, had a lower calculated frequency than Eastern Asian population, which had a frequency of 0.6845. The calculated frequency was equal to 0.652, which was higher than European frequency.

Analysis with HapMap data yielded 117 polymorphisms. None of these polymorphisms were found in unrelated Han Chinese in Beijing population data. 4 out of the 117 polymorphisms were found to be in Hardy – Weinberg disequilibrium in the Lithuanian data. 1 polymorphism was found on chromosome 3 in *ADCY5* gene, and 3 were found on chromosome 5 in *EBF1* gene. Out of these, none had any associations with preterm birth, however rs9844212 (G > C) was associated with birth weight. This polymorphism was described only in Utah residents with ancestry from northern and western Europe population, as well as Japanese in Tokyo population and Yoruba in Ibadan population. All frequencies listed in the database were lower than calculated. Other 3 polymorphisms found had no association with preterm birth or birth weight but were found to be in Hardy – Weinberg disequilibrium. One of these polymorphisms, rs31196 (C > A), was found only in Japanese in Tokyo population. Frequency in the database was listed as 0.0135, where the calculated frequency in the Lithuanian population was 0.651.

gnomAD database yielded 3 polymorphisms. These were rs4482616 (A > G), rs9844212(G > C) and rs9855969 (A > G), all found on chromosome 3. Out of these polymorphisms, rs9844212 was the only one that had lower frequency in all the gnomAD populations compared to calculated Lithuanian mothers' frequency. The other 2 polymorphisms had lower calculated frequency than the East Asian population.

None of the identified polymorphisms were related to preterm birth directly. However, as mentioned, few of them had associations with birth weight. In addition, these polymorphisms were found in genes, which in literature are associated with preterm birth. Knowing, that the data is from the mothers who gave birth on term, probably these SNPs are not associated with preterm birth in the Lithuanian population. However, further investigation is required to confirm the findings.

SNPNexus phenotype data analysis

Mothers' data did not yield as much information as newborns' data. Looking at the phenotypical analysis, only ClinVar database found entries. COSMIC and SIFT database search did not yield any results. Polymorphism found in ClinVar was rs9855969 (A > G), which was classified as benign. Phenotypes included dyskinesia with facial myokymia.

GAD databse found 1,589 associated phenotypes. These were separated into 8 classes, a lot less than newborns' data. Class occurrence count can be seen in figure 14.

In the contrast to newborns' dataset, the cancer class was not present in the mothers'

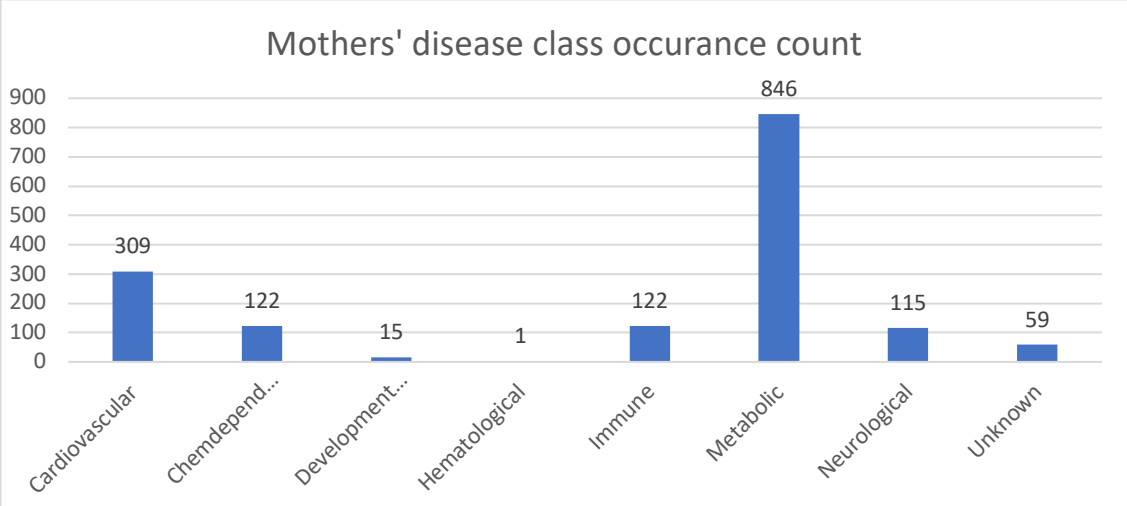

Figure 14. Disease class occurrence count in mothers' dataset.

dataset. Most associations were found with metabolic diseases, resulting in 53.2% of all associated polymorphisms. An interesting find is under the unknown disease class. This class was composed of 2 phenotypes: endometriosis and birth weight/diabetes mellitus type 2. Both phenotypes are associated with pregnancy complications. Endometriosis is a condition, where tissue like lining of the uterus grows outside of it, for example in fallopian tubes. This increases the risk of ectopic pregnancy because embryo can get caught in the fallopian tube.

DISCUSSION

The aim of the thesis was to analyse and evaluate of single nucleotide gene polymorphisms related to preterm birth in healthy Lithuanian newborn and maternal. Variant annotation, and calculations of Hardy – Weinberg equilibrium and SNP frequency was performed.

Hardy – Weinberg equilibrium results showed that observed heterozygosity was higher than the expected heterozygosity. This could be because the model did not accurately follow the described model. According to literature sources, the discrepancies could have happened because of the small sample size (Cox and Kraft, 2006). For most of the cases, observed heterozygosity was higher than expected heterozygosity, which indicates an excess of heterozygotes. For newborns, 1050 polymorphisms were shown to be in disequilibrium, and for mothers – 195. To accurately determine, whether the polymorphism is in equilibrium in the population, 2 to 3 generations (Lachance, 2016) need to be tested for the same SNP, then the Hardy – Weinberg equilibrium must be calculated. Comparing the results to the literature data, it was found that a small number of SNPs, that were in disequilibrium were related to preterm birth. Most of the findings support previously

described polymorphisms, and correlate to systems, which are thought to be involved in preterm birth.

Frequency calculations were compared against 3 different databases (1000 Genomes Project, HapMap and gnomAD) for each dataset. For the newborns' dataset, calculated variants' frequencies were varying closely to different European populations frequencies and some of them were associated with preterm birth and were in Hardy – Weinberg disequilibrium. However, most frequencies were higher compared to the European populations. This could be, because of the small dataset. In addition, different European populations have different ancestries, which could also explain the difference between calculated frequency and frequency, listed in the database. Variants comprise genes that are strongly debated to be associated with preterm birth. For mothers' dataset, none of the polymorphisms found across 4 databases were associated with preterm birth. However, variants were found, that were associated with birth weight. Compared to databases, some SNPs found in the datasets contradict database entries (International HapMap Consortium, 2003; Fairley *et al.*, 2020; Chen *et al.*, 2022). This is because of the differences in populations and their ancestry.

One of the limitations of the thesis was the sample size. Smaller datasets mean less conclusive results, however, the collection of bigger datasets is complicated, because of the nature of the research. Most of the time, after giving birth, mothers are not interested in participating in research, no matter if the baby was born on term or preterm. In addition, the data was given by the supervisor out of the ANELGEMIA project. Furthermore, data was collected from pure Lithuanians, of at least 3rd generation. Because our country's history is intertwined with surrounding nations, finding pure Lithuanians is also a challenge.

The second limitation of the project was the nature of the data. Data was collected from healthy term–born babies, and not from preterm-born babies. Analyzing the variations, associated with preterm birth with collected data, can only contradict previous findings. If the variation is found in a healthy Lithuanian newborn or mother, who has given birth at term, it could be argued that specific polymorphism is not a cause of preterm birth, or there must be other polymorphisms in a combination.

CONCLUSIONS

1. Candidate genes were gathered from literature sources, including less likely candidates as well. Candidate gene list for newborns included 51 genes, for mothers – 34.
2. Whole genome sequencing variation data was annotated successfully, which allowed for parsing and transformation to appropriate formats for further analysis.
3. Hardy – Weinberg equilibrium calculations and analyses identified that polymorphisms departed from Hardy – Weinberg equilibrium in both cohorts. Newborns' data included 1,050 polymorphisms departing from the equilibrium, mothers' – 195.
4. Calculated SNP frequencies were compared against databases, with findings showing that various polymorphisms, associated with preterm birth, are present in healthy Lithuanian newborns cohort. Most of the found variants had a higher frequency compared to different European populations.
5. More extensive research should be completed, including preterm – born babies' and their mothers' data into the analysis.

ACKNOWLEDGEMENTS

Supervisor of master's thesis – Dr. Alina Urnikytė, for consultations and knowledge sharing during preparation of the thesis.

Lecturer during Systems Biology course – Dr. Saulius Gražulis, for great explanation of Makefile syntax, which was one of the main points in the master's thesis.

REFERENCES

1. Balloux, F. *et al.* (2018) 'From Theory to Practice: Translating Whole-Genome Sequencing (WGS) into the Clinic.', *Trends in microbiology*, 26(12), pp. 1035–1048. doi: 10.1016/j.tim.2018.08.004.
2. Bartel, D. P. (2004) 'MicroRNAs: genomics, biogenesis, mechanism, and function.', *Cell*, 116(2), pp. 281–97. doi: 10.1016/s0092-8674(04)00045-5.
3. Basys, V. *et al.* (2022) 'Gimimų medicininiai duomenys', *Lithuanian Institute of Hygiene*. Available at: https://www.hi.lt/uploads/pdf/leidiniai/Statistikos/Gimimu/gimimai_2021.pdf.
4. Betts-Henderson, J. *et al.* (2010) 'The nystagmus-associated FRMD7 gene regulates neuronal outgrowth and development.', *Human molecular genetics*, 19(2), pp. 342–51. doi: 10.1093/hmg/ddp500.
5. Bonomo, J. A. *et al.* (2014) 'The ras responsive transcription factor RREB1 is a novel candidate gene for type 2 diabetes associated end-stage kidney disease.', *Human molecular genetics*, 23(24), pp. 6441–7. doi: 10.1093/hmg/ddu362.
6. Born, T. L. *et al.* (2000) 'Identification and Characterization of Two Members of a Novel Class of the Interleukin-1 Receptor (IL-1R) Family', *Journal of Biological Chemistry*, 275(39), pp. 29946–29954. doi: 10.1074/jbc.M004077200.
7. Butler, J. M. (2012) 'Single Nucleotide Polymorphisms and Applications', in *Advanced Topics in Forensic DNA Typing*. Elsevier, pp. 347–369. doi: 10.1016/B978-0-12-374513-2.00012-9.
8. Cadieu, E. *et al.* (2009) 'Coat variation in the domestic dog is governed by variants in three genes.', *Science (New York, N.Y.)*, 326(5949), pp. 150–3. doi: 10.1126/science.1177808.
9. Chang, C. C. *et al.* (2015) 'Second-generation PLINK: rising to the challenge of larger and richer datasets.', *GigaScience*, 4(1), p. 7. doi: 10.1186/s13742-015-0047-8.
10. Chawanpaiboon, S. *et al.* (2019) 'Global, regional, and national estimates of levels of preterm birth in 2014: a systematic review and modelling analysis.', *The Lancet. Global health*, 7(1), pp. e37–e46. doi: 10.1016/S2214-109X(18)30451-0.
11. Chen, S. *et al.* (2022) 'A genome-wide mutational constraint map quantified from variation in 76,156 human genomes', *bioRxiv*. doi: https://doi.org/10.1101/2022.03.20.485034.
12. Cox, D. G. and Kraft, P. (2006) 'Quantification of the power of Hardy-Weinberg equilibrium testing to detect genotyping error.', *Human heredity*, 61(1), pp. 10–4. doi: 10.1159/000091787.
13. Danecek, P. *et al.* (2011) 'The variant call format and VCFtools.', *Bioinformatics (Oxford, England)*, 27(15), pp. 2156–8. doi: 10.1093/bioinformatics/btr330.
14. Dayem Ullah, A. Z. *et al.* (2018) 'SNPnexus: assessing the functional relevance of genetic variation to facilitate the promise of precision medicine.', *Nucleic acids research*, 46(W1), pp. W109–W113. doi: 10.1093/nar/gky399.

15. van El, C. G. *et al.* (2013) 'Whole-genome sequencing in health care: recommendations of the European Society of Human Genetics.', *European journal of human genetics : EJHG*, 21(6), pp. 580–4. doi: 10.1038/ejhg.2013.46.
16. Engel, S. A. M. *et al.* (2005) 'Risk of spontaneous preterm birth is associated with common proinflammatory cytokine polymorphisms.', *Epidemiology (Cambridge, Mass.)*, 16(4), pp. 469–77. doi: 10.1097/01.ede.0000164539.09250.31.
17. *Evidence-Based Medicine and the Changing Nature of Health Care* (2008) *Evidence-Based Medicine and the Changing Nature of Health Care*. Washington, D.C.: National Academies Press. doi: 10.17226/12041.
18. Fairley, S. *et al.* (2020) 'The International Genome Sample Resource (IGSR) collection of open human genomic variation resources.', *Nucleic acids research*, 48(D1), pp. D941–D947. doi: 10.1093/nar/gkz836.
19. Ferrante, M. I. *et al.* (2001) 'IL1RAPL2 maps to Xq22 and is specifically expressed in the central nervous system.', *Gene*, 275(2), pp. 217–21. doi: 10.1016/s0378-1119(01)00659-x.
20. Filomena, M. C. *et al.* (2021) 'Myopalladin knockout mice develop cardiac dilation and show a maladaptive response to mechanical pressure overload.', *eLife*, 10. doi: 10.7554/eLife.58313.
21. Fiscella, K. (1996) 'Race, perinatal outcome, and amniotic infection.', *Obstetrical & gynecological survey*, 51(1), pp. 60–6. doi: 10.1097/00006254-199601000-00022.
22. Goldenberg, R. L. *et al.* (1996) 'Medical, psychosocial, and behavioral risk factors do not explain the increased risk for low birth weight among black women.', *American journal of obstetrics and gynecology*, 175(5), pp. 1317–24. doi: 10.1016/s0002-9378(96)70048-0.
23. Goldenberg, R. L. *et al.* (2008) 'Epidemiology and causes of preterm birth.', *Lancet (London, England)*, 371(9606), pp. 75–84. doi: 10.1016/S0140-6736(08)60074-4.
24. Griggs, K. M. *et al.* (2020) 'Preterm Labor and Birth: A Clinical Review.', *MCN. The American journal of maternal child nursing*, 45(6), pp. 328–337. doi: 10.1097/NMC.0000000000000656.
25. Huang, T., Shu, Y. and Cai, Y.-D. (2015) 'Genetic differences among ethnic groups.', *BMC genomics*, 16(1), p. 1093. doi: 10.1186/s12864-015-2328-0.
26. International HapMap Consortium (2003) 'The International HapMap Project.', *Nature*, 426(6968), pp. 789–96. doi: 10.1038/nature02168.
27. Jin, Y.-R. *et al.* (2011) 'The canonical Wnt signaling activator, R-spondin2, regulates craniofacial patterning and morphogenesis within the branchial arch through ectodermal-mesenchymal interaction.', *Developmental biology*, 352(1), pp. 1–13. doi: 10.1016/j.ydbio.2011.01.004.
28. Keats, B. J. B. and Sherman, S. L. (2013) 'Population Genetics', in *Emery and Rimoin's Principles and Practice of Medical Genetics*. Elsevier, pp. 1–12. doi: 10.1016/B978-0-12-383834-6.00015-X.

29. Lachance, J. (2016) *Hardy-Weinberg Principle - an overview | ScienceDirect Topics*, *Encyclopedia of Evolutionary Biology*.
30. Lavebratt, C. and Sengul, S. (2006) 'Single nucleotide polymorphism (SNP) allele frequency estimation in DNA pools using Pyrosequencing.', *Nature protocols*, 1(6), pp. 2573–82. doi: 10.1038/nprot.2006.442.
31. McCormick, M. C. (1985) 'The contribution of low birth weight to infant mortality and childhood morbidity.', *The New England journal of medicine*, 312(2), pp. 82–90. doi: 10.1056/NEJM198501103120204.
32. Meirhaeghe, A. *et al.* (2007) 'A possible role for the PPARG Pro12Ala polymorphism in preterm birth.', *Diabetes*, 56(2), pp. 494–8. doi: 10.2337/db06-0915.
33. Menon, R. *et al.* (2009) 'Racial disparity in pathophysiologic pathways of preterm birth based on genetic variants.', *Reproductive biology and endocrinology : RB&E*, 7(1), p. 62. doi: 10.1186/1477-7827-7-62.
34. *MIR505 Gene* (2023) *GeneCaRNA*. Available at: https://www.genecards.org/cgi-bin/carddisp.pl?gene=MIR505.
35. Modi, B. P. *et al.* (2017) 'Mutations in fetal genes involved in innate immunity and host defense against microbes increase risk of preterm premature rupture of membranes (PPROM).', *Molecular genetics & genomic medicine*, 5(6), pp. 720–729. doi: 10.1002/mgg3.330.
36. Mukhopadhyay, R. (2009) 'DNA sequencers: the next generation.', *Analytical chemistry*, 81(5), pp. 1736–40. doi: 10.1021/ac802712u.
37. Murphy, M. and McLoughlin, G. (2015) 'Born too soon: Preterm birth in Europe trends, causes and prevention', *Entre Nous*.
38. Myking, S. *et al.* (2013) 'X-chromosomal maternal and fetal SNPs and the risk of spontaneous preterm delivery in a Danish/Norwegian genome-wide association study.', *PloS one*. Edited by W. Yan, 8(4), p. e61781. doi: 10.1371/journal.pone.0061781.
39. Nunn, C. *et al.* (2010) 'RGS2 inhibits beta-adrenergic receptor-induced cardiomyocyte hypertrophy.', *Cellular signalling*, 22(8), pp. 1231–9. doi: 10.1016/j.cellsig.2010.03.015.
40. Oscanoa, J. *et al.* (2020) 'SNPnexus: a web server for functional annotation of human genome sequence variation (2020 update).', *Nucleic acids research*, 48(W1), pp. W185–W192. doi: 10.1093/nar/gkaa420.
41. *Premature birth statistics* (no date) *Tommy's*. Available at: https://www.tommys.org/pregnancy-information/premature-birth/premature-birth-statistics#:~:text=These statistics show the percentage,White%3A 7.4%25 (Accessed: 15 May 2023).
42. *Preterm Birth* (2022) *CDC*. Available at: https://www.cdc.gov/reproductivehealth/maternalinfanthealth/pretermbirth.htm.
43. Purcell, S. *et al.* (2007) 'PLINK: a tool set for whole-genome association and population-

based linkage analyses.', *American journal of human genetics*, 81(3), pp. 559–75. doi: 10.1086/519795.

44. Di Renzo, G. C., Roura, L. C. and European Association of Perinatal Medicine-Study Group on Preterm Birth (2006) 'Guidelines for the management of spontaneous preterm labor.', *Journal of perinatal medicine*, 34(5), pp. 359–66. doi: 10.1515/JPM.2006.073.
45. Romero, R. *et al.* (2010) 'Identification of fetal and maternal single nucleotide polymorphisms in candidate genes that predispose to spontaneous preterm labor with intact membranes.', *American journal of obstetrics and gynecology*, 202(5), pp. 431.e1–34. doi: 10.1016/j.ajog.2010.03.026.
46. Sheikh, I. A. *et al.* (2016) 'Spontaneous preterm birth and single nucleotide gene polymorphisms: a recent update.', *BMC genomics*, 17(Suppl 9), p. 759. doi: 10.1186/s12864-016-3089-0.
47. Shen, H. *et al.* (2013) 'Comprehensive characterization of human genome variation by high coverage whole-genome sequencing of forty four Caucasians.', *PloS one*. Edited by P. Awadalla, 8(4), p. e59494. doi: 10.1371/journal.pone.0059494.
48. Siriphak, S. *et al.* (2021) 'Kallikrein-11, in Association with Coiled-Coil Domain Containing 25, as a Potential Prognostic Marker for Cholangiocarcinoma with Lymph Node Metastasis.', *Molecules (Basel, Switzerland)*, 26(11), p. 3105. doi: 10.3390/molecules26113105.
49. Speer, E. M. *et al.* (2006) 'Role of single nucleotide polymorphisms of cytokine genes in spontaneous preterm delivery.', *Human immunology*, 67(11), pp. 915–23. doi: 10.1016/j.humimm.2006.08.291.
50. Spencer, D. H., Zhang, B. and Pfeifer, J. (2015) 'Single Nucleotide Variant Detection Using Next Generation Sequencing', in *Clinical Genomics*. Elsevier, pp. 109–127. doi: 10.1016/B978-0-12-404748-8.00008-3.
51. Strauss, J. F. *et al.* (2018) 'Spontaneous preterm birth: advances toward the discovery of genetic predisposition.', *American journal of obstetrics and gynecology*, 218(3), pp. 294-314.e2. doi: 10.1016/j.ajog.2017.12.009.
52. Svensson, A. C. *et al.* (2009) 'Maternal effects for preterm birth: a genetic epidemiologic study of 630,000 families.', *American journal of epidemiology*, 170(11), pp. 1365–72. doi: 10.1093/aje/kwp328.
53. Tarpey, P. *et al.* (2006) 'Mutations in FRMD7, a newly identified member of the FERM family, cause X-linked idiopathic congenital nystagmus.', *Nature genetics*, 38(11), pp. 1242–4. doi: 10.1038/ng1893.
54. U.S. National Library of Medicine (2020) 'What are single nucleotide polymorphisms (SNPs)? - Genetics Home Reference - NIH', *U.S. National Library of Medicine*. Available at: https://medlineplus.gov/genetics/understanding/genomicresearch/snp/.
55. Urnikyte, A. *et al.* (2022) 'Inherited and De Novo Variation in Lithuanian Genomes: Introduction to the Analysis of the Generational Shift.', *Genes*, 13(4), p. 569. doi:

10.3390/genes13040569.

56. Vallejos-Vidal, E. *et al.* (2020) ‘Single-Nucleotide Polymorphisms (SNP) Mining and Their Effect on the Tridimensional Protein Structure Prediction in a Set of Immunity-Related Expressed Sequence Tags (EST) in Atlantic Salmon (Salmo salar)’, *Frontiers in Genetics*, 10. doi: 10.3389/fgene.2019.01406.
57. Vogel, J. P. *et al.* (2018) ‘The global epidemiology of preterm birth.’, *Best practice & research. Clinical obstetrics & gynaecology*, 52, pp. 3–12. doi: 10.1016/j.bpobgyn.2018.04.003.
58. Walani, S. R. (2020) ‘Global burden of preterm birth.’, *International journal of gynaecology and obstetrics: the official organ of the International Federation of Gynaecology and Obstetrics*, 150(1), pp. 31–33. doi: 10.1002/ijgo.13195.
59. Wang, K., Li, M. and Hakonarson, H. (2010) ‘ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.’, *Nucleic acids research*, 38(16), p. e164. doi: 10.1093/nar/gkq603.
60. *Whole Genome Sequencing* (2022) *Centers for Disease Control and Prevention*. Available at: https://www.cdc.gov/pulsenet/pathogens/wgs.html (Accessed: 28 December 2022).
61. Yang, L. *et al.* (2020) ‘DNA of neutrophil extracellular traps promotes cancer metastasis via CCDC25.’, *Nature*, 583(7814), pp. 133–138. doi: 10.1038/s41586-020-2394-6.
62. Yang, S. *et al.* (2020) ‘Insights on SNP types, detection methods and their utilization in Brassica species: Recent progress and future perspectives.’, *Journal of biotechnology*, 324, pp. 11–20. doi: 10.1016/j.jbiotec.2020.09.018.
63. Zhang, H. *et al.* (2015) ‘A genome-wide association study of early spontaneous preterm delivery.’, *Genetic epidemiology*, 39(3), pp. 217–26. doi: 10.1002/gepi.21887.
64. Zhao, E. Y., Jones, M. and Jones, S. J. M. (2019) ‘Whole-Genome Sequencing in Cancer.’, *Cold Spring Harbor perspectives in medicine*, 9(3), p. a034579. doi: 10.1101/cshperspect.a034579.

SUMMARY

Examination of Preterm Birth – Related Single Nucleotide Gene Polymorphisms in Newborns and Maternal Genome

Martynas Kairys

Systems biology master studies

Vilnius University, Faculty of Medicine, Institute of Biomedical Sciences, Department of Human and Medical Genetics.

Aim of the project was to analyze and evaluate single nucleotide gene polymorphisms related to preterm birth in healthy Lithuanian newborn and maternal genomes. The topic was chosen because preterm birth is still a big concern in healthcare, even with advancement of medicine. According to statistics, 13.4 million children were born preterm in 2020. This results in approximately 1 in 10 babies. Preterm born babies could face various medical conditions.

Analysis was done using a makefile written by the author. Makefile is a set of rules and commands written in a specific language. Data for the research was provided by the supervisor of the thesis. Data was annotated, filtered, and prepared for further analysis. Analysis was done by calculating Hardy – Weinberg equilibriums and frequencies for the single nucleotide polymorphisms. Hardy – Weinberg equilibrium calculations were used to check which polymorphisms are imbalanced, while frequency calculations and comparison to other populations provided information about prevalence.

Results for Hardy – Weinberg equilibrium showed that there is an excess of heterozygotes in Lithuanian population for some polymorphisms, associated with preterm birth. Variant frequency calculations enabled comparative analysis to other populations. It was found that most of the found SNPs had a higher frequency compared to different European populations. It was found that, both groups had variants, that are associated with preterm birth, according to literature sources. In most cases, frequencies of these variants were higher compared to analysed populations. More extensive research on this topic should be completed, including preterm – born babies' and their mothers' data.

SUMMARY IN LITHUANIAN

Su neišnešiotumu susijusių vieno nukleotido polimorfizmų tyrimas naujagimių ir jų motinų genomuose

Martynas Kairys
Sistemų biologijos magistras
Vilniaus Universitetas, medicinos fakultetas, biomedicinos institutas, žmogaus ir medicininės genetikos katedra.

Darbo tikslas buvo išanalizuoti ir įvertinti vieno nukleotido polimorfizmų sąsają su priešlaikiniu gimdymu sveikuose lietuvių naujagimių ir jų motinų genomuose. Tema buvo pasirinkta, nes priešlaikinis gimdymas vis dar yra opi problema visame pasaulyje, net su patobulėjusiomis medicinos priemonėmis. Statistikos duomenimis, 2020 metais 13,4 milijono vaikų gimė per anksti. Apytiksliai, tai yra 1 iš 10 vaikų. Per anksti gimę vaikai dažnai susiduria su įvairiomis medicininėmis problemomis.

Analizė buvo atlikta naudojant autoriaus parašytą makefile. Tai yra failas, savyje laikantis komandas ir taisykles, kurios yra parašytos specifine makefile kalba. Genetiniai duomenys tyrimui buvo pateikti darbo vadovo. Duomenys buvo anotuoti, filtruoti bei konvertuoti į reikiamus formatus. Analizė atlikta apskaičiuojant Hardžio – Vainbergo pusiausvyrą, bei polimorfizmų dažnius. Hardžio – Vainbergo pusiausvyra galėjome nustatyti variantų nukrypimą nuo pusiausvyros, o dažnių rezultatai leido palyginti variantų dažnius su kitomis populiacijomis.

Apskaičiavus Hardžio – Vainbergo pusiausvyrą, buvo nustatyta, kad naujagimių grupėje yra variantų, susijusių su priešlaikiniu gimdymu, kurie nėra pusiausvyroje. Atlikus dažnių palyginamąją analizę, nustatėme, kad didesnė dalis rastų genetinių variantų turėjo didesnį dažnį lyginant su kitomis Europos populiacijomis. Abejose analizuotose grupėse buvo nustatyti genų variantai, kurie, pagal literatūrinius šaltinius, yra siejami su nešnešiotumu. Šių variantų dažniai, lyginant su duomenų bazėmis, buvo didesni nei kitose analizuotose populiacijose. Remiantis sudarytais genų sąrašais, reikėtų atlikti tolimesnę analizę, į ją įtraukiant neišnešiotų naujagimių ir jų motinų duomenis.

APPENDICES

Appendix 1. Fetal genes list.

VEGFA
ADIPOQ
IL6
IL10
TNF
TLR4
IL1B
MMP9
CRP
IGF1
CXCL8
MTHRF
NPPB
ACE
TGFB1
LEP
MBL2
APOE
VDR
PPARG
ADRB2
NOS2A
NOS3A
ADD1
COMT
EGR1
TFAP2A
SP3
FRMD7
COL24A1
UTP14A
PLAC1
MIR505
IL1RAPL2
TLR7
ATP11C
AMOT
DACH2
RNASET2
LPHN2
MAN1A1
L3MBTL3
INPP1
SMAD9
TMEM229A
RREB1
SPOCK3
RSPO2
SORL1
KCNH7
NOL10

Appendix 2. Maternal genes list.

EBF1
EEFSEC
ADCY5
WNT4
OXTR
RLN2
PTGER3
PTGS1
PTGS2
HPGD
S100A9
COL4A3
COL1A2
EDN1
COL5A2
COL5A1
VEGFA
ANGPT1
ADRB2
ADD1
SERPINE1
CYP51A1
HMGCR
GSTM1
PPARG
TLR4
DEFA5
HLA-DQA1
CR1
SFTPD
CSF2
EGR1
TFAP2A
SP3

Appendix 3. Makefile used in the thesis.

```
SHELL := /bin/bash
INPUT_DIR = inputs
OUTPUT_DIR = outputs
NEWBORN_DIR = $(OUTPUT_DIR)/newborns
MOTHER_DIR = $(OUTPUT_DIR)/mothers
VCF_INPUT_DIR = vcf_files
GENE_DIR_N = $(NEWBORN_DIR)/gene
GENE_DIR_M = $(MOTHER_DIR)/gene
RS_DIR_N = $(NEWBORN_DIR)/rs_gene
RS_DIR_M = $(MOTHER_DIR)/rs_gene
VCF_DIR_N = $(NEWBORN_DIR)/vcf_gene
VCF_DIR_M = $(MOTHER_DIR)/vcf_gene
MAP_PED_DIR_N = $(NEWBORN_DIR)/map_ped_gene
MAP_PED_DIR_M = $(MOTHER_DIR)/map_ped_gene
```

```
HWE_DIR_N = $(NEWBORN_DIR)/hwe_gene
HWE_DIR_M = $(MOTHER_DIR)/hwe_gene
FREQ_DIR_N = $(NEWBORN_DIR)/freq
FREQ_DIR_M = $(MOTHER_DIR)/freq
ANNOTTED_FILE_N = $(INPUT_DIR)/ANE25.newborns_annotted.hg19_multianno.txt
ANNOTTED_FILE_M = $(INPUT_DIR)/ANE.women_annotted.hg19_multianno.txt
RAW_FILE_N =
$(VCF_INPUT_DIR)/newborns/ANE25_newborns_ok_2023_modchr_rsID_copy.vcf
RAW_FILE_M = $(VCF_INPUT_DIR)/mothers/ANE.women.data_vcf.vcf
GENE_LIST_FILE_N = $(INPUT_DIR)/genes_newborn.list
GENE_LIST_FILE_M = $(INPUT_DIR)/genes_mother.list
GENE_LIST_N = $(shell grep -v "\#" $(GENE_LIST_FILE_N))
GENE_LIST_M = $(shell grep -v "\#" $(GENE_LIST_FILE_M))
GENE_FILES_N = $(GENE_LIST_N:%=$(GENE_DIR_N)/%.txt)
GENE_FILES_M = $(GENE_LIST_M:%=$(GENE_DIR_M)/%.txt)
RS_FILES_N = $(GENE_FILES_N:$(GENE_DIR_N)/%.txt=$(RS_DIR_N)/%.txt)
RS_FILES_M = $(GENE_FILES_M:$(GENE_DIR_M)/%.txt=$(RS_DIR_M)/%.txt)
VCF_FILES_N = $(RS_FILES_N:$(RS_DIR_N)/%.txt=$(VCF_DIR_N)/%.vcf)
VCF_FILES_M = $(RS_FILES_M:$(RS_DIR_M)/%.txt=$(VCF_DIR_M)/%.vcf)
MAP_PED_FILES_N = $(VCF_FILES_N:$(VCF_DIR_N)/%.vcf=$(MAP_PED_DIR_N)/%)
MAP_PED_FILES_M = $(VCF_FILES_M:$(VCF_DIR_M)/%.vcf=$(MAP_PED_DIR_M)/%)
HWE_FILES_N = $(MAP_PED_FILES_N:$(MAP_PED_DIR_N)/%=$(HWE_DIR_N)/%)
HWE_FILES_M = $(MAP_PED_FILES_M:$(MAP_PED_DIR_M)/%=$(HWE_DIR_M)/%)
FREQ_FILES_N = $(VCF_FILES_N:$(VCF_DIR_N)/%.vcf=$(FREQ_DIR_N)/%)
FREQ_FILES_M = $(VCF_FILES_M:$(VCF_DIR_M)/%.vcf=$(FREQ_DIR_M)/%)

all: $(RS_FILES_N) $(GENE_FILES_N) $(VCF_FILES_N) $(MAP_PED_FILES_N)
$(HWE_FILES_N) $(GENE_FILES_M) $(RS_FILES_M) $(VCF_FILES_M)
$(MAP_PED_FILES_M) $(HWE_FILES_M) $(FREQ_FILES_N) $(FREQ_FILES_M)

$(GENE_DIR_N)/%.txt:
      awk -v x="$(*F)" -F '\t' '{if ($$8 == x) {print $$0}}'
$(ANNOTTED_FILE_N) > $@

$(GENE_DIR_M)/%.txt:
      awk -v x="$(*F)" -F '\t' '{if ($$8 == x) {print $$0}}'
$(ANNOTTED_FILE_M) > $@

$(RS_DIR_N)/%.txt: $(GENE_DIR_N)/%.txt
      awk '{print $$6}' $< | sed '/\./d' > $@

$(RS_DIR_M)/%.txt: $(GENE_DIR_M)/%.txt
      awk '{print $$17}' $< | sed '/\./d' > $@

$(VCF_DIR_N)/%.vcf: $(RS_DIR_N)/%.txt
      vcftools --vcf $(RAW_FILE_N) --snps $< --out $@ --recode
      mv $@.recode.vcf $@

$(VCF_DIR_M)/%.vcf: $(RS_DIR_M)/%.txt
      vcftools --vcf $(RAW_FILE_M) --snps $< --out $@ --recode
      mv $@.recode.vcf $@

$(MAP_PED_DIR_N)/%: $(VCF_DIR_N)/%.vcf
      vcftools --vcf $< --out $@ --plink

$(MAP_PED_DIR_M)/%: $(VCF_DIR_M)/%.vcf
      vcftools --vcf $< --out $@ --plink

$(HWE_DIR_N)/%: $(MAP_PED_DIR_N)/%
      plink1 --file $< --hardy --out $@
```

```
$(HWE_DIR_M)/%: $(MAP_PED_DIR_M)/%
	plink1 --file $< --hardy --out $@

$(FREQ_DIR_N)/%: $(VCF_DIR_N)/%.vcf
	plink2 --vcf $< --freq --out $@ || true

$(FREQ_DIR_M)/%: $(VCF_DIR_M)/%.vcf
	plink2 --vcf $< --freq --out $@ || true
```