



**Faculty of  
Medicine**

Systems Biology Master Program

Department of Human and Medical Genetics

Faculty of Medicine

Natalija Ošmianskaja

Master Thesis

Priešlaikinio gimdymo viso genomo asociacijos tyrimas naudojant genotipavimo  
duomenis

Genome-Wide Association Study of Preterm Birth Using Genotyping Data

|                                  |                                    |
|----------------------------------|------------------------------------|
| Supervisor                       | Alina Urnikytė, PhD                |
| Head of the department or clinic | Prof. Algirdas Utkus, PhD          |
| Advisor                          | Svetlana Dauengauer-Kirlienė, MD   |
|                                  | Vilnius, 2026-05-14                |
| Student's email                  | natalija.osmianskaja@mf.stud.vu.lt |

# CONTENTS

|                                                                                      |           |
|--------------------------------------------------------------------------------------|-----------|
| <b>LIST OF ABBREVIATIONS .....</b>                                                   | <b>4</b>  |
| <b>INTRODUCTION.....</b>                                                             | <b>5</b>  |
| Aim of the study .....                                                               | 6         |
| Objectives.....                                                                      | 6         |
| <b>LAY DESCRIPTION .....</b>                                                         | <b>7</b>  |
| <b>LITERATURE REVIEW .....</b>                                                       | <b>8</b>  |
| Preterm birth.....                                                                   | 8         |
| Risk factors for PTB.....                                                            | 9         |
| Heritability of PTB.....                                                             | 10        |
| Maternal and fetal genetic variants in PTB.....                                      | 12        |
| PTB and genome-wide association study (GWAS) .....                                   | 16        |
| Research in Lithuania.....                                                           | 20        |
| <b>METHODS .....</b>                                                                 | <b>22</b> |
| Samples .....                                                                        | 22        |
| Phenotype markers .....                                                              | 23        |
| Input data.....                                                                      | 24        |
| The workflow of the study .....                                                      | 25        |
| Data quality control.....                                                            | 26        |
| Manhattan Plot Visualization.....                                                    | 35        |
| QQ Plot Visualization .....                                                          | 36        |
| Identification and prioritization of overlapping SNPs across association models..... | 36        |
| Annotation.....                                                                      | 38        |
| <b>RESULTS .....</b>                                                                 | <b>40</b> |
| Study cohort .....                                                                   | 40        |
| <b>DISCUSSION .....</b>                                                              | <b>83</b> |
| <b>CONCLUSIONS .....</b>                                                             | <b>88</b> |
| <b>RECOMMENDATION.....</b>                                                           | <b>89</b> |

|                                                                                         |           |
|-----------------------------------------------------------------------------------------|-----------|
| <b>ACKNOWLEDGEMENTS .....</b>                                                           | <b>90</b> |
| <b>REFERENCES.....</b>                                                                  | <b>91</b> |
| <b>SUMMARY .....</b>                                                                    | <b>97</b> |
| <b>SUMMARY IN LITHUANIAN.....</b>                                                       | <b>98</b> |
| <b>APPENDICES .....</b>                                                                 | <b>99</b> |
| Appendix 1. Candidate variant list investigated in the study (GRCh38) .....             | 99        |
| Appendix 2. Covariate file covariate.txt, used for logistic regression with PLINK ..... | 100       |
| Appendix 3. Sample file sample.txt, used for SNPTTEST association analysis.....         | 103       |

## LIST OF ABBREVIATIONS

|       |                                        |
|-------|----------------------------------------|
| BPD   | Bronchopulmonary dysplasia             |
| CHR   | Chromosome                             |
| CI    | Confidence interval                    |
| CNV   | Copy number variation                  |
| cPVL  | Cystic periventricular leukomalacia    |
| DEG   | Differentially expressed gene          |
| GA    | Gestational age                        |
| GWAS  | Genome-wide association study          |
| IVH   | Intraventricular hemorrhage            |
| MAF   | Minor allele frequency                 |
| OR    | Odds ratio                             |
| PPROM | Preterm premature rupture of membranes |
| PTB   | Preterm birth                          |
| QQ    | Quantile-quantile (plot)               |
| ROP   | Retinopathy of prematurity             |
| SNP   | Single-nucleotide polymorphism         |
| SPB   | Spontaneous preterm birth              |
| VEPTB | Vascular endothelial preterm birth     |
| WGS   | Whole-genome sequencing                |

# INTRODUCTION

Preterm birth (PTB), which is the delivery of the infant before the completion of 37 weeks of gestation, is one of the most prevalent causes of morbidity and mortality in preterm neonates. Preterm infants born before the completion of the 34th week of gestation are more susceptible to serious health complications that affect various body organs, such as respiratory distress syndrome, bronchopulmonary dysplasia, retinopathy of the premature infant, intraventricular hemorrhage and immunological disorders (Jakuskiene *et al.*, 2011). Although substantial advances have been made in the prevention and care of preterm births, the biological basis of PTB and the health complications associated with them is still not well understood.

Several environmental, maternal and clinical risk factors for PTB have been identified, including infections, maternal health conditions and pregnancy-related health complications. However, it is estimated that about two-thirds of PTBs occur due to unknown causes (Vogel *et al.*, 2018). Increasing evidence suggests a genetic contribution to PTB (Solé-Navais *et al.*, 2023).

Genome-wide association study (GWAS) of single nucleotide polymorphism (SNP) became significant tool in the identification of genetic variations that may be linked with complex traits or diseases. There have been many genome-wide association studies that have sought to understand the genetic predisposition to PTB. These studies have identified genetic variations that may play a role in inflammation, placenta function and fetal development. However, few studies have sought to understand the role of genetic variations that may play a role in organ-specific complications and immune response among extremely or very preterm-born neonates. In addition, genomic studies that incorporate the genotype of neonates with their phenotypes remain limited.

Investigating genomic variation in preterm neonates and comparing it with term-born controls may contribute to a better understanding of potential genetic factors involved in neonatal adaptation and vulnerability to complications. Even in studies with limited sample sizes, such analyses may help characterize genetic variation in the studied population and generate hypotheses for future larger-scale investigations.

This study investigates genomic variation in preterm and term-born neonates and evaluates potential relationships between identified genetic variants and neonatal complications, including respiratory distress syndrome, intraventricular haemorrhage (IVH), bronchopulmonary dysplasia

(BPD) and other clinical outcomes. The analysis is based on genotyping data from neonates born between 2022 and 2024 at Vilnius University Hospital Santaros Klinikos, combined with clinical phenotype information obtained from medical records and parental surveys.

This study represents one of the first attempts in Lithuania to analyze genomic variation in preterm neonates together with detailed clinical phenotype information of both the child and parents, providing an initial exploration of potential genetic factors influencing neonatal outcomes.

### Aim of the study

The aim of this study is to investigate genomic variation in preterm and term-born neonates and assess potential associations between genetic variants, PTB and neonatal complications, as well as to explore their possible functional relevance.

### Objectives

1. Perform a case-control GWAS to investigate genetic variation and evaluate potential associations between SNP variants and PTB.
2. Annotate genomic variation data in preterm and term-born neonates to identify potentially functional variants and explore biological mechanisms associated with PTB.
3. Assess potential associations between genetic variants and neonatal complications.

## LAY DESCRIPTION

PTB defined as birth before 37 weeks of gestation, remains one of the most important challenges in perinatal medicine. Infants born too early are more likely to experience respiratory problems, developmental issues and other health complications. Although advances have been made in the medical care of preterm infants, the causes of PTB remain unclear. Increasing scientific evidence suggests that genetic factors may play an important role in PTB; however, these factors are still not fully understood.

Understanding the biological mechanisms leading to PTB is important not only for researchers but also for clinicians and families affected by this condition. Better knowledge of genetic influences could help explain why some pregnancies are more vulnerable and may contribute to improved prediction, prevention and treatment strategies in the future.

In this study, genetic data of infants born prematurely were analyzed and compared to infants born at term. The genome-wide approach was employed to detect small genetic differences that are more common in premature infants. These genetic differences were analyzed using bioinformatics approaches to determine whether they could play a role in biological processes that are relevant to pregnancy and early development.

The present study is one of the first attempts to collect and analyze genetic data on PTB in Lithuania. Although no statistically significant associations were identified, this work provides a foundation for future studies investigating the genetic basis of PTB in the Lithuanian population.

# LITERATURE REVIEW

## Preterm birth

PTB, defined by the World Health Organization (WHO) as delivery before the completion of 37 weeks of gestation, is a significant global health concern, impacting approximately 10% of all live births each year.

An estimated 13.4 million neonates were born preterm in 2020, accounting for approximately 9% of all births worldwide. In Southern Asia and sub-Saharan Africa, the highest rate 13.2% of PTB was recorded, while it was found to be less than 8% in Northern America, Europe, Australia, New Zealand. Approximately 900 000 children died in 2019 due to complications of PTBs. Many survivors face a lifetime of disability, including learning disabilities and visual and hearing problems (Ohuma *et al.*, 2023).

The degree of prematurity is associated with an increased risk of mortality and morbidity. The highest risk is observed in infants born extremely preterm, that is, before 28 weeks of gestation, followed by infants born very preterm, between 28 and 32 weeks of gestation and newborns born moderate to late preterm, between 32 and 37 weeks of gestation (Teune *et al.*, 2011).

Late preterm newborn mortality rate is nine times that of full term infants and preterm neonates were exposed to a greater risk of intercurrent conditions during the neonatal period. These intercurrent conditions were inversely related to gestational age (GA) (Araújo *et al.*, 2012).

PTB is linked to a higher risk of death from non-communicable diseases such as cardiovascular diseases, chronic lung diseases and diabetes during young adulthood (Risnes *et al.*, 2021).

Children and adults born moderate to late preterm generally have average intelligence, however, they tend to achieve slightly lower scores on intelligence tests than individuals born at term. In addition, moderate to late PTB has been associated with an increased risk of psychiatric morbidity, including substance-use disorders, mood disorders and particularly psychotic disorders (Fernández De Gamarra-Oca *et al.*, 2021).

PTB is a major public health problem because it is one of the leading causes of infant death, is a major contributor to long-term health, developmental and psychiatric problems and places a significant burden on health-care systems and societies worldwide.

## Risk factors for PTB

Most PTBs occur spontaneously due to spontaneous preterm birth (SPB) or preterm premature rupture of membranes (PPROM). However, PTB may also result from medically indicated delivery, when early birth is undertaken because continuation of pregnancy poses a risk to maternal or fetal health (Vogel *et al.*, 2018).

Medically indicated PTBs most commonly result from conditions such as preeclampsia, eclampsia and intrauterine growth restriction. In contrast, SPBs are considered a multifactorial syndrome arising from diverse underlying mechanisms, including infection or inflammation, vascular pathology and uterine overdistension. Established risk factors for SPB include a history of preterm delivery, Black race, periodontal disease and low maternal body mass index. Among predictive markers a shortened cervical length and elevated cervicovaginal fetal fibronectin levels are the strongest indicators of SPB risk (Goldenberg *et al.*, 2008).

Several studies were focused on finding associations between PTB and mother's health status.

A 2013 analysis of association between parent age and health status and clinical consequences experienced by a newborn (Kozuki *et al.*, 2013) stated that higher PTB risk observed among adolescent mothers <18 years of age and women of advanced maternal age  $\geq 35$ .

PTB risk is associated with cigarette smoking before and during pregnancy in several studies (Liu *et al.*, 2020), (Kondracki and Hofferth, 2019), (Soneji and Beltrán-Sánchez, 2019).

Association between PTB and mother's health status and risk factors before pregnancy were found (Haas *et al.*, 2005). Multiple indicators of poor pre-pregnancy health were linked to a higher likelihood of preterm delivery, including low body mass index (OR 2.38; 95% CI 1.04-5.48), a prior diagnosis of chronic hypertension (OR 3.12; 95% CI 1.94-5.02), limited physical activity before pregnancy (OR 2.31; 95% CI 1.41-3.77) and smoking prior to conception (OR 2.20; 95% CI 1.29-3.75). By contrast, depressive symptoms before pregnancy showed no significant association with preterm delivery risk.

Infections that have been linked to PTB include syphilis (Qin *et al.*, 2014), HIV (Xiao *et al.*, 2015), hepatitis C (Huang *et al.*, 2015), malaria (Desai *et al.*, 2007), bacterial vaginosis (Haahr *et*

*al.*, 2016), Chlamydia trachomatis (Olson-Chen, Balaram and Hackney, 2018), chorioamnionitis (Galinsky *et al.*, 2013), urinary tract infections (Cunnington, Kortsalioudaki and Heath, 2013).

Several studies have confirmed that obesity is associated with an increased risk of SPB (Marchi *et al.*, 2015), (Torloni *et al.*, 2009), (McDonald *et al.*, 2010).

Factors affecting PTB also include environment: water and air quality, use of alcohol, substances or opioid, poor diet habits, long working hours and long periods of standing (Griggs *et al.*, 2020).

A wide range of maternal, infectious, lifestyle and environmental factors, including extremes of maternal age, poor pre-pregnancy health, smoking, obesity, infections and adverse environmental exposures contribute to PTB risk, highlighting its multifactorial and complex etiology.

### Heritability of PTB

PTB is a complex condition with a multifactorial etiology involving the interplay of environmental, clinical and biological factors. In addition to well-established maternal, fetal and obstetric risk factors, growing evidence indicates that genetics plays a significant role in determining the risk of preterm delivery. Familial clustering of PTB and differences in risk across populations suggest a heritable component that cannot be fully explained by environmental exposures alone.

A genetic epidemiologic study of 630,000 Swedish families (Svensson *et al.*, 2009) investigated relative contributions of maternal and fetal genetic factors to PTB for nearly one million Swedish births. Using familial clustering analyses and mixed models, the authors found that PTB risk was significantly increased among sisters but not among families linked through brothers, indicating a strong maternal genetic influence. Approximately 25% of the variation in PTB was attributed to maternal genetic factors, while fetal genetic effects were minimal and these maternal effects were independent of established maternal risk factors.

The possibility that the risk of PTB may be transmitted across generations was suggested several decades ago (Porter *et al.*, 1997). Women who were themselves born preterm have an approximately 20% higher risk of delivering preterm, with the magnitude of risk increasing as their own GA at birth decreases.

Furthermore, women whose sisters delivered preterm infants have an approximately 80% higher risk of experiencing PTB themselves (Svensson *et al.*, 2009).

Another study (Boyd *et al.*, 2009) evaluated the influence of PTB within both maternal and paternal family lines on a woman's risk of PTB using nationwide Danish registry data from over 1.1 million singleton births. The risk of PTB was strongly increased among women with a previous PTB but was not influenced by a partner's history of preterm children. Increased risk was observed when PTBs occurred in the woman's mother, full sisters, or maternal half-sisters, whereas no association was found with paternal relatives or the partner's family, suggesting that PTB risk is predominantly transmitted through the female lineage, potentially reflecting maternal genetic, imprinting, or mitochondrial effects.

While large epidemiological studies of Scandinavian and European-American populations suggest a significant maternal genetic contribution to PTB, other studies have also provided evidence that fetal genetic factors contribute to PTB risk alongside maternal genetic effects.

Heritability estimates suggest that maternal genetic factors contribute approximately 17%-36% to PTB risk, whereas children of twins (COT) studies have not demonstrated a significant contribution of fetal genetic factors. However, analyses using a Markov Chain Monte Carlo (MCMC) approach suggested that the estimated genetic contribution varied depending on how PTB was modeled. Maternal genetic factors accounted for 27.1% of GA variance when analyzed as a continuous trait, whereas fetal genetic factors were associated with PTB liability when the condition was modeled as a binary outcome (York *et al.*, 2009).

Analysis of increased risk of PPRM in African Americans compared to European Americans revealed genetic variation in genes encoding pro- and anti-inflammatory cytokines and their corresponding receptors. The genes involved in pro-inflammatory signaling are implicated in extracellular matrix metabolism, as inflammatory cytokines promote the expression of matrix-degrading metalloproteinases, which may contribute to membrane weakening and rupture (Anum *et al.*, 2009).

PTB shows evidence of familial aggregation, indicating that genetic factors contribute to its risk, alongside environmental and clinical influences. Epidemiological and family-based studies have

demonstrated that women with a personal or family history of PTB, particularly through the maternal lineage, have a higher likelihood of delivering preterm.

### Maternal and fetal genetic variants in PTB

Genetic variation is common differences in DNA sequence within a population that contributes to individual variability in biological traits and susceptibility to disease. The most frequent type of genetic variations is the SNP, which involves a change in a single base pair in the genome. Although most variants are biologically neutral, some can alter gene expression or protein function and thereby influence physiological processes and disease risk.

Identifying genetic variants associated with an increased risk of PTB typically involves a systematic approach that includes the selection of candidate genes based on their known roles in biological pathways relevant to the pathophysiology of PTB.

Large parental-allele analysis found fifteen loci acting via the maternal genome, seven via both maternal and fetal genomes and two via the fetal genome alone for gestational duration, showing a direct fetal contribution (Solé-Navais *et al.*, 2023).

The study performed GWAS meta-analysis of 18 797 PTB and 260 246 in-time deliveries mostly from Northern Europe countries. Sixteen of twenty two loci did not overlap with any previously reported gestational duration-associated locus.

The results show that both maternal and fetal genomes contribute to gestational duration and the risk of PTB. Some loci act mainly through maternal genetic variation, while others act through fetal genetics or depend on the parent from whom the allele is inherited. This highlights the complex biological mechanisms regulating the timing of birth (Figure 1 a, b).

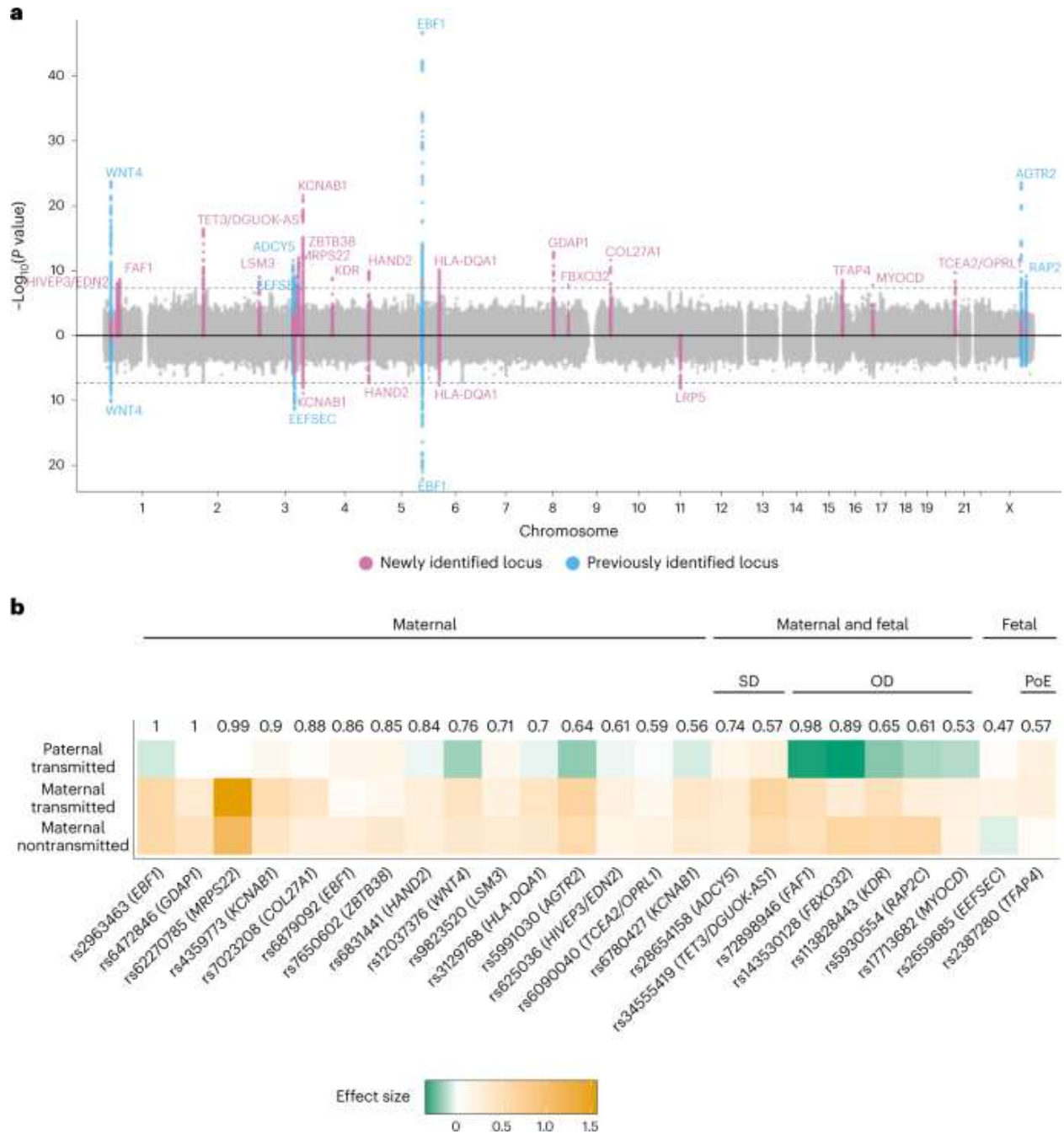


**Figure 1.** Genetic loci associated with gestational duration and PTB and their maternal and fetal genetic effects (Solé-Navais *et al.*, 2023)

(a) Miami plot showing genome-wide association results for gestational duration (upper panel) and PTB (lower panel). The x axis represents chromosomal position across the genome, while the y axis shows the statistical significance of association ( $-\log_{10}$  p-value). The dashed horizontal line

indicates the genome-wide significance threshold. Colored markers highlight loci identified in the analysis, with different colors distinguishing previously identified and newly identified loci.

(b) Heatmap illustrates the relative contributions of maternal and fetal genetic effects for the identified loci based on transmitted and non-transmitted parental alleles. Rows represent paternal transmitted, maternal transmitted and maternal non-transmitted alleles. Colors indicate the direction and magnitude of the genetic effect for each variant.

Fifteen variants related to *EBF1*, *GDAP1*, *MRPS22*, *KCNAB1*, *COL27A1*, *ZBTB38*, *HAND2*, *WNT4*, *LSM3*, *HLA-DQA1*, *AGTR2*, *HIVEP3/EDN2*, *TCEA2/OPRL1*, *KCNAB1* were identified with higher probability of maternal effect, seven (*FAF1*, *FBXO32*, *KDR*, *RAP2C*, *MYOCD*, *ADCY5*, *TET3/DGUOK-AS1*) of both maternal and fetal, while two (*TFAP4*, *EEFSEC*) having a fetal-only effect.

Genetic variation rs28654158 at the *ADCY5* locus had both maternal and fetal effects on gestational duration and probably can be related to the fetal insulin hypothesis.

SNP rs7594852 was attributed as fetal only. This variant is related to *CKAP2L* intron located on 2 chromosome in a genomic region that includes the pro-inflammatory cytokine genes *IL1A* and *IL1B*, which are known to play an important role in inflammatory signaling (Liu *et al.*, 2019).

The most comprehensive GWAS was conducted to investigate loci across six genes *EBF1*, *EEFSEC*, *AGTR2*, *WNT4*, *ADCY5* and *RAP2C*, which are associated with gestational duration (Zhang *et al.*, 2017). Three of these six genes (*EBF1*, *EEFSEC* and *AGTR2*) showed strong associations with PTB in a European ancestry cohort of 43 568 women. Additionally the study reported that maternal SNPs accounted for approximately 23% of the observed phenotypic variance in PTB, indicating that other genetic, environmental, or gene environment factors likely contribute to the remaining unexplained variation.

One of PTB associated genes is *EBF1* (early B-cell factor 1) which is essential for normal B cell development. Genetic variation rs2963463 was later confirmed to be significantly associated with PTB in another study (Mladenović *et al.*, 2024). Women with the CC genotype had a three to four times lower risk of PTB.

Another finding of Zhang et al. is that *EEFSEC* (eukaryotic elongation factor, selenocysteine tRNA-specific) significantly associated with PTB risk, is involved in the incorporation of selenocysteine into selenoproteins. Selenoproteins, including glutathione peroxidases and thioredoxin reductases, play essential roles in cellular homeostasis by maintaining redox balance, providing antioxidant defense and modulating inflammatory responses. *EEFSEC* rs201450565 variation was linked with PTB, however, it also linked to prostate cancer (Gudmundsson *et al.*, 2009) and the risk of hypospadias (Paternoster *et al.*, 2011).

According to Zhang et al., *AGTR2* (angiotensin II receptor type 2) the coding gene located closest to a cluster of x chromosome rs201386833 SNP, that reached genome wide significance in the analysis of PTB and has been suggested to play a role in the modulation of uteroplacental circulation with robust replication in Nordic cohorts.

Another study (Knijnenburg *et al.*, 2019) was focusing on multi-omics approach for whole-genome sequencing (WGS), RNA sequencing (RNA-seq) and DNA methylation data collected from 791 American family trios, including 270 families with PTB and investigated using whole-genome sequencing, RNA sequencing and DNA methylation data.

The study identified differentially expressed genes and differentially methylated genomic loci, then used expression and methylation quantitative trait locus analyses to link genetic variants to changes in gene expression and DNA methylation. Two genes associated with PTB *RAB31* and *RBPJ*, were consistently identified across all three data modalities, including whole-genome sequencing, RNA sequencing and DNA methylation profiling. Pathway analyses associated vascular endothelial preterm birth (VEPTB) with several biological pathways, notably epidermal growth factor receptor and prolactin signaling, inflammation- and immunity-related pathways, chemokine signaling, interferon- $\gamma$  signaling and Notch1 signaling.

Associations between *RAB31* and *RBPJ* and PTB have been identified in previous studies.

*RAB31* shows higher expression at term rather than at PTB that has been linked to placental abruption (Denis *et al.*, 2014).

A variant rs117652912 located in *RAB31* intron, was found in European and mixed ancestry mothers. *RAB31* is a one of the RAS oncogene family members, responsible for coding of a small GTPase binding protein (Winn *et al.*, 2007).

*RBPJ* is upregulated and hypermethylated in VEPTB and functions as a transcriptional regulator in the Notch signaling pathway. Heterozygous mutations in *RBPJ* that impair DNA binding have been identified as a cause of Adams-Oliver syndrome, a rare developmental disorder characterized by skin, limb and vascular abnormalities (Hassed *et al.*, 2012).

Genetic studies demonstrate that SNPs contribute to PTB risk through both maternal and fetal genomes, with maternal genetic effects playing a predominant role. Large GWAS and parental allele analyses have identified numerous novel loci associated with gestational duration and PTB, highlighting pathways related to immunity, inflammation, vascular function, endocrine signaling and oxidative stress. Multi-omics approach also supported the role of genes such as *EBF1*, *EEFSEC*, *AGTR2*, *RAB31* and *RBPJ* in the complex genetic basis of PTB.

### PTB and genome-wide association study (GWAS)

GWAS is a genetics approach used to identify common genetic variants across the genome that are statistically associated with a trait or disease, such as PTB or gestational duration. GWAS compares the frequency of millions of SNPs between individuals with different phenotypes, for example cases vs controls, or across a continuous trait. Variants that show significant frequency differences are interpreted as being associated with the trait, pointing to genomic regions and biological pathways involved in disease risk (Uffelmann *et al.*, 2021) (Figure 2).

GWAS usually starts with definition of phenotype and selection of cases and controls or a continuous outcome and recruitment of a sufficiently large and well characterized population.

Genome-wide SNP genotyping is commonly performed using microarrays, although sequencing approaches, including whole-genome sequencing (WGS), whole-exome sequencing (WES) and targeted sequencing, can also be used to identify genetic variants.

Quality control is applied to remove low quality samples and variants, for example missingness, Hardy-Weinberg equilibrium violations, population outliers.

Imputation is used to identify genetic variants that were not directly measured during genotyping, particularly when only a subset of genome-wide variants is available (e.g., from SNP arrays). Using reference panels, such as the 1000 Genomes Project (The 1000 Genomes Project

Consortium *et al.*, 2015), imputation increases genomic coverage and improves comparability across studies.

Each SNP is tested for association with the phenotype using regression models, adjusting for covariates such as ancestry, sex and other relevant clinical factors.

Strong genome-wide significance threshold (typically  $p\text{-value} < 5 \times 10^{-8}$ ) is applied to control for false-positive findings due to testing millions of variants.

Associated loci are mapped to genes and pathways, sometimes followed by functional analyses such as gene expression analysis, epigenomics, protein function studies or pathway and network analyses to understand biological mechanisms. (Uffelmann *et al.*, 2021)

Collaboration and data sharing have been central to the success of genome-wide association studies. An early commitment to open science was formalized in 1997 through the Bermuda Principles, which stated that all human genomic sequence data generated by publicly funded large-scale sequencing projects should be made freely available in the public domain (Cook-Deegan and McGuire, 2017).

GWAS have shifted PTB genetics beyond candidate gene approaches by identifying reproducible maternal and fetal risk loci associated with gestational duration and SPB. The majority of genetic signals arises from the maternal genome and large sample sizes are critical for robust discovery (Zhang *et al.*, 2017).

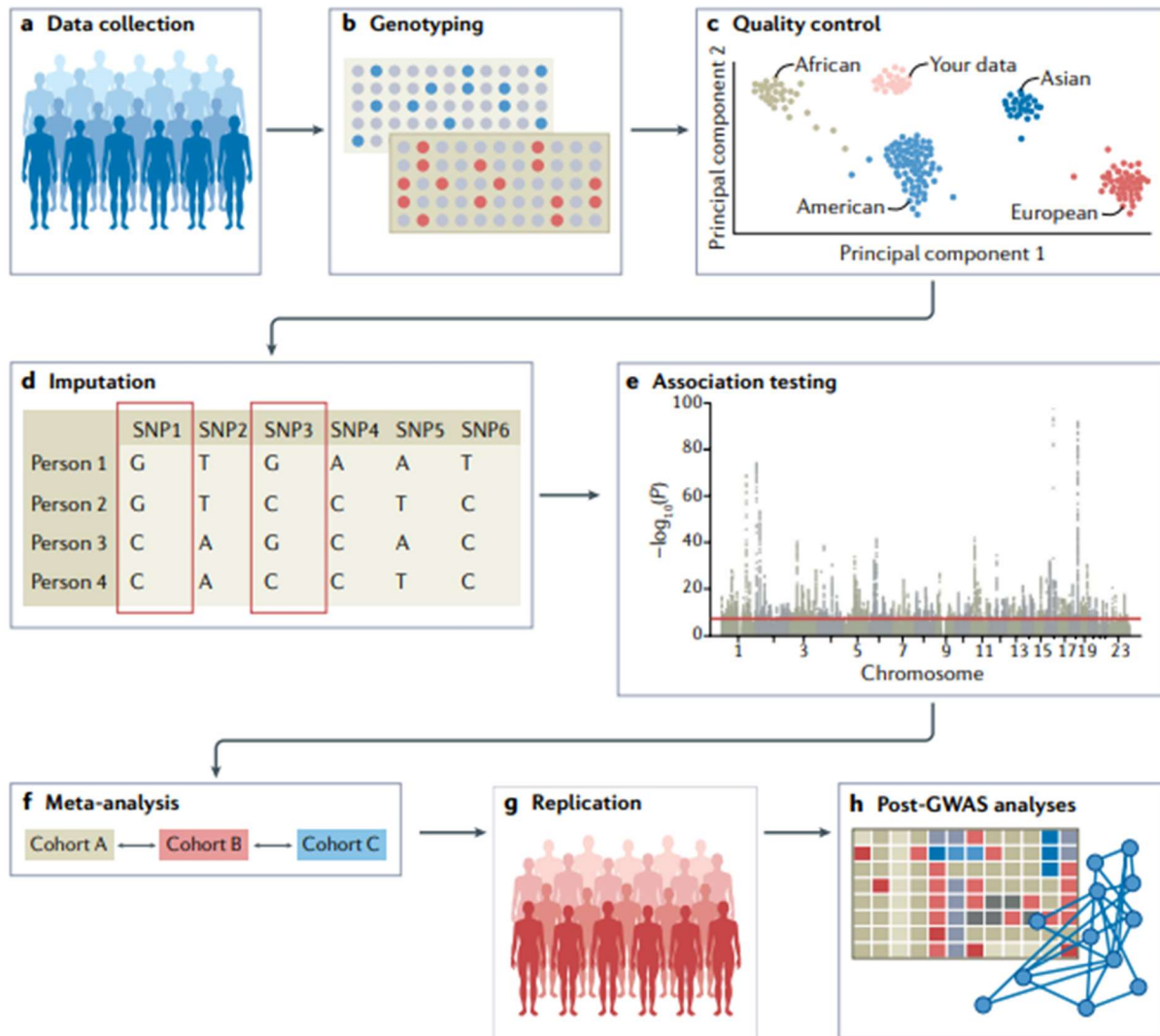


**Figure 2.** Overview of the main steps in a GWAS (adapted from Uffelmann et al., 2021). The workflow includes data collection (a), genotyping (b), quality control including population structure assessment (c), genotype imputation (d), association testing (e), meta-analysis (f), replication of findings (g) and post-GWAS analyses (h).

Major GWAS findings in PTB and gestational duration include:

- *EBF1*, *EEFSEC*, *AGTR2* association with SPB (Zhang *et al.*, 2017)
- Fifteen variants showed a higher probability of maternal effects, including *EBF1*, *GDAP1*, *MRPS22*, *KCNAB1*, *COL27A1*, *ZBTB38*, *HAND2*, *WNT4*, *LSM3*, *HLA-DQA1*, *AGTR2*, *HIVEP3/EDN2* and *TCEA2/OPRL1*, seven exhibited both maternal and fetal effects: *FAF1*, *FBXO32*, *KDR*, *RAP2C*, *MYOCD*, *ADCY5*, *TET3/DGUOK-AS1* and two were associated exclusively with fetal effects: *TFAP4* and *EEFSEC* (Solé-Navais *et al.*, 2023)
- *ZBTB38*, *HAND2*, *TET3*, *KCNAB1*, *COL27A1*, *GNAQ*, *KCNN2* associations with PTB were observed (Pasanen *et al.*, 2022)

Maternal vs fetal GWAS comparisons disclosed that fetal GWAS have rarer hits. Only few loci were identified with evidence indicating that the effect is driven by the fetal rather than the maternal genotype. For example, a locus on chromosome 2q13 was robustly associated with gestational duration and replicated in over 9,000 additional infants. Lead variant, rs7594852, affects binding of the *HIC1* transcriptional repressor and maps to a region containing 97 genes involved in pro-inflammatory pathways central to parturition, underscoring the public health relevance of understanding genetic regulation of gestational timing (Liu *et al.*, 2019)

A fetal focused GWAS across ancestries identified two genome-wide loci but no replication, suggesting fetal risk is likely due to many common/rare variants and gene-environment interplay rather than single SNPs. The most significant variant rs17591250 is an intergenic SNP on chromosome 1. The second significant variant corresponds to rs1979081 which is also intergenic, but located on chromosome 8 (Rappoport *et al.*, 2018)

Region specific GWAS results performed on Indian (Bhattacharjee *et al.*, 2023), Korean (Hur *et al.*, 2023), Peruvian (Juvinao-Quintero *et al.*, 2023) and Japanese (Hasegawa *et al.*, 2023) populations generally show no or few genome-wide hits, but replicate some European loci, e.g., *WNT4*, and point to pathways in inflammation, progesterone/steroid metabolism, ion channels and myometrial contraction.

The most important biological pathways related to PTB highlighted by GWAS are:

- Uterine development and function: *WNT4*, *HAND2*, *TBX5*, *RYR2* associated with calcium handling and contraction (Zhang *et al.*, 2017), (Pasanen *et al.*, 2022), (Hur *et al.*, 2023)
- Progesterone/steroid metabolism pathways: *AGTR2*, *ADCY5*, *GC* (Zhang *et al.*, 2017), (Pasanen *et al.*, 2022), (Juvinao-Quintero *et al.*, 2023)
- Immune/inflammatory signaling: 2q13 IL-1 cluster, *TLR4* *NFKB*-related genes, other inflammation networks (Liu *et al.*, 2019), (Bacelis *et al.*, 2016), (Brubaker *et al.*, 2016)

GWAS is a powerful approach used to identify common genetic variants associated with PTB by analyzing millions of SNPs across large populations. GWAS have transformed PTB genetics by moving beyond candidate gene studies to identify reproducible maternal and fetal risk loci, revealing that most detectable genetic signals arise from the maternal genome, while fetal effects are fewer and more difficult to replicate. These studies have highlighted key biological pathways involved in uterine function, hormonal regulation and immune and inflammatory signaling, underscoring the polygenic and multifactorial nature of PTB.

## Research in Lithuania

Lithuania appears to have relatively low PTB rates, with outcomes improving over time but remaining challenging for the most immature infants. In 2024 5.7% of all births in Lithuania were reported as PTB (Institute of Hygiene, 2026).

The study at Kaunas Medical University Hospital followed 338 very preterm infants ( $\leq 32$  weeks,  $\leq 1500$  g) admitted 2003-2005 reveals that mortality is 53.5% for extremely preterm (22-27 weeks), while for very preterm infants (28-32 weeks) is 2.9%. Among survivors, 78.2% were discharged without major short-term complications, while the remaining experienced specific morbidities, including bronchopulmonary dysplasia (BPD) 6.3%, severe retinopathy of prematurity (ROP) 4.2%, grade III-IV intraventricular hemorrhage (IVH) 10.9% and cystic periventricular leukomalacia (cPVL) 8.0% (Jakuskiene *et al.*, 2011).

Analysis from neighboring Latvia reports a stable PTB rate of 5.5% for 2000-2023 period among the lowest globally. Lithuania is cited alongside Latvia and Estonia with similarly low rates: ~5.4% in Lithuania and 5.6% in Estonia, versus a European median of 6.9% (Andzane *et al.*, 2025). Study

does not detail specific neonatal complications, focusing instead on trends in PTB rates and changes in the proportions of spontaneous versus indicated PTBs over time.

However, neither study from the Baltic countries conducted analyses assessing genotype data in relation to parental and child phenotypic characteristics.

The present study addresses this gap by analysing genome-wide genetic variation in Lithuanian preterm infants and integrating these data with clinical and phenotypic information. The aim is to identify genetic variants potentially associated with PTB and to provide initial insights into the genetic architecture of this condition in the Lithuanian population. This work represents one of the first attempts to investigate genetic susceptibility to PTB in Lithuania and may contribute to a better understanding of biological mechanisms underlying the timing of birth.

## METHODS

### Samples

The study cohort consisted of twenty-nine preterm neonates (cases) and twenty-four term-born (controls) neonates.

The study included preterm neonates delivered between 2022 and 2024 at Vilnius University Hospital Santaros Klinikos who had available biological samples, parental consent and complete clinical information required for genetic analysis. DNA was extracted from neonatal umbilical cord blood samples collected at birth. Prematurely born neonate genetic data was genotyped using Illumina Infinium Global Diversity 1.8 M Array (Illumina, San Diego, CA, USA). The data have not been previously published.

Data from twenty-four healthy term-born infants were obtained from the “Inherited and De Novo Variation in Lithuanian Genomes study” (Urniškytė *et al.*, 2022). Healthy term-born neonates in the control group underwent WGS using the Illumina NovaSeq 6000 platform (Illumina, San Diego, CA, USA). Sequencing reads were aligned to the human reference genome GRCh38 and SNPs were detected using a variant-calling pipeline.

Information on parental demographic and health characteristics, pregnancy and childbirth history and results of the newborn’s clinical examination and laboratory and instrumental tests were collected. The data was fully pseudo anonymized and no individual participants could be identified.

Genotype and phenotype data was provided for this study by VU MF Translational Health Research Institute Populational Genomic Laboratory. This research is a part of the ANELGEMIA project (No. S-MIP-20-34), which has received ethical approval from Vilnius Regional Bioethics Committee (No. 2020/6-1243-724, 22-06-2022).

## Phenotype markers

Clinical outcomes and maternal characteristics were categorized as phenotype markers based on established diagnostic criteria and information obtained from clinical records for both case and control groups.

The following phenotype variables were included in the analysis.

Maternal:

- Gestational diabetes (GD) / preeclampsia: maternal pregnancy complications diagnosed during pregnancy according to established clinical criteria and recorded as present or absent.  
GD is a form of diabetes that is first diagnosed during pregnancy and is characterized by elevated blood glucose levels resulting from impaired glucose tolerance. Preeclampsia is a pregnancy-specific hypertensive disorder characterized by new onset high blood pressure after 20 weeks of gestation, often accompanied by proteinuria or other signs of organ dysfunction, such as liver, kidney, or hematological abnormalities.
- SPB and PPROM: SPB defined as delivery before 37 completed weeks of gestation following spontaneous onset of labor. Preterm premature rupture of membranes defined as rupture of the fetal membranes before the onset of labor and before 37 weeks of gestation.
- Latency period: time interval between rupture of the fetal membranes and delivery, recorded in hours.
- Maternal age: age of the mother at delivery, recorded in years.
- Number of pregnancies: total number of pregnancies reported for the mother (gravidity).
- Previous pregnancy complications: presence of complications in previous pregnancies, including miscarriage, preeclampsia, or other obstetric complications.
- Maternal ancestry: maternal ethnic background (Lithuanian / non-Lithuanian), recorded based on self-reported ethnicity.

- Maternal family history: presence of pregnancy-related or genetic conditions within the maternal family.

#### Neonatal:

- Gestational age (GA): duration of pregnancy at delivery, recorded in weeks and determined based on obstetric assessment using the last menstrual period and ultrasound examination.
- Birth weight: the weight of the neonate measured immediately after birth, recorded in grams (g).
- Sex: neonatal biological sex recorded as male or female.
- Weight category: birth weight classification based on gestational-age-adjusted growth standards, categorizing neonates as small, appropriate or large for gestational age.
- Exact birth-weight percentile for GA and sex: the precise percentile rank of the newborn's weight compared with a reference population of neonates with the same GA and sex.

#### Input data

SNP genotype data for the case group obtained using microarray based genotyping and SNP variant data for the control group derived from WGS were provided for the study by VU MF Translational Health Research Institute Populational Genomic Laboratory in PLINK (Chang *et al.*, 2015) binary format: BED, BIM and FAM files. Case and control datasets were merged into a single PLINK dataset for subsequent quality control and association analyses.

The BED file stores genotype information as a binary matrix of samples by SNPs, where genotypes are encoded in binary format instead of human readable text.

The BIM file contains SNP metadata. Each row corresponds to one variant and includes the following columns: chromosome number, SNP identifier (rsID), genetic distance, base-pair position and the two alleles (allele 1 and allele 2).

The FAM file contains sample information for each individual in the dataset, including family ID, individual ID, paternal ID, maternal ID, sex and phenotype. Each row corresponds to one sample.

## The workflow of the study

Data quality control was performed as an initial step to ensure high-quality genotype data.

Based on evidence from previous studies identified through a literature review, a set of 43 candidate SNPs associated with PTB was selected and evaluated for potential association with the phenotype in the study dataset.

While candidate variant analysis focuses on a predefined set of variants selected based on prior evidence, it is limited to known loci and cannot identify novel variants associated with the phenotype. Therefore, association analysis was applied as a broader, hypothesis-free approach.

Association analysis is a statistical method used to test the relationship between genetic variants and a phenotype by systematically evaluating each variant across the genome, allowing the identification of novel associations without prior assumptions.

To consider the effects of clinical and maternal factors, six models with different combinations of these variables were used in all analyses:

- Model 1: neonatal sex and maternal age
- Model 2: birth weight category and exact birth-weight percentile for GA and sex
- Model 3: GD or preeclampsia, SPB and PPROM and latency period
- Model 4: maternal age and number of pregnancies
- Model 5: previous pregnancy complications and maternal ancestry
- Model 6: GA, birth weight and neonatal sex

These models were applied in logistic regression analysis using PLINK v1.9 (Shaun Purcell, Christopher Chang, 2026) and in association analysis using SNPTEST v2.5.4 (Marchini and Howie, 2010) on non-imputed genotype data. In addition, genotype imputation was performed to infer missing variants and the SNPTEST v2.5.4 analysis was repeated on the imputed dataset to assess associations with increased genomic coverage.

The workflow of the study is presented in Figure 3.

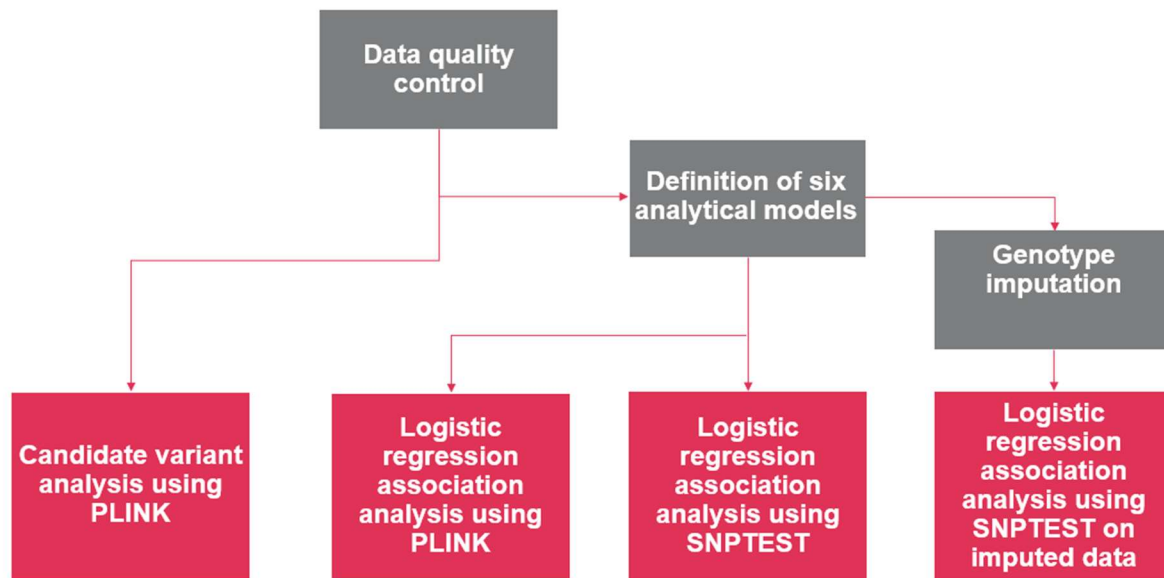


**Figure 3.** The workflow of the study

All methods are described in detail in the following sections.

### Data quality control

Genotype data quality control was performed using PLINK v2.0. The initial dataset included 54 individuals (21 females and 33 males; 29 cases and 25 controls) comprising 221,408 SNPs. Standard quality control filters were applied to remove low quality samples and variants. Individuals with genotype missingness greater than 2% were excluded, as well as SNPs with missing genotype rates exceeding 2%. Variants with a minor allele frequency (MAF) below 0.01 were also removed and SNPs deviating from Hardy-Weinberg equilibrium ( $p\text{-value} < 1 \times 10^{-4}$ ) were excluded. After applying these filters, no samples were removed due to missing genotype data and 1,344 SNPs were excluded due to deviation from Hardy-Weinberg equilibrium. The final dataset used for subsequent analyses consisted of 54 individuals and 220,064 SNPs.

Quality control filtering was performed using the following command:

```

plink2 --bfile genotype_data \
--geno 0.02 \
--mind 0.02 \
--maf 0.01 \
--hwe 1e-4 \
--make-bed \
--out samplesnpqc
  
```

## Candidate variant analysis using PLINK

Candidate variant analysis is a hypothesis driven approach that examines genetic variants previously implicated in the biological mechanisms underlying PTB. Based on evidence from previous studies identified through a literature review, a set of 43 candidate SNPs associated with PTB was selected and evaluated for potential association with the studied phenotype. The list of candidate variants is provided in Appendix 1.

The presence of candidate variants in the dataset was verified by matching them against the variant list in the BIM file using the *grep* command:

```
grep -Ff snp.txt samplesnpqc.bim
```

Only one candidate variant rs17713682 was found.

This variant was extracted using PLINK v2.0:

```
plink2 \  
--bfile samplesnpqc \  
--snp rs17713682 \  
--make-bed \  
--out rs17713682_only
```

Association between the candidate SNP and the phenotype was evaluated using logistic regression implemented in PLINK v2.0. The binary phenotype (case/control status) was obtained from the phenotype column of the FAM file. Below is the command to run logistic regression:

```
plink2 \  
--bfile rs17713682_only \  
--glm \  
--out rs17713682_chi2
```

Due to the relatively small sample size, Fisher's exact test was additionally performed to compare allele frequencies between cases and controls. This analysis was conducted using PLINK v1.9.

```
plink \  
--bfile rs17713682_only \  
--fisher \  

```

```
--out rs17713682_fisher
```

## Logistic regression association analysis using PLINK

Logistic regression is a statistical method commonly applied in genetic association studies to test whether genetic variants are associated with a binary outcome variable, namely case-control status, while accounting for one or more explanatory variables such as phenotype markers. The method estimates the probability of the outcome using a logit link function and reports the effect of predictors as odds ratios (OR) together with confidence intervals (CI) and p-values.

In this study, logistic regression association analysis was performed using PLINK v1.9.

A covariate file for the association analysis was prepared according to the input requirements of PLINK v1.9. The file contained sample identifiers together with phenotype and covariate information corresponding to the genotype data. The first two columns included the Family ID (FID) and Individual ID (IID), which must match exactly the identifiers in the genotype dataset. The remaining columns contained the phenotype variable and covariates used in the regression analysis. All variables were encoded in numeric format, missing values were represented using a numeric missing-value code (-9). The file is space delimited, each row corresponds to one individual included in the analysis. Covariate file content can be found in Appendix 2.

Each SNP was tested independently under an additive genetic model, where the genotype was coded according to the number of effect alleles carried by an individual (0, 1, or 2). An additive genetic model is a framework used in genetic association studies in which the effect of a genetic variant increases linearly with the number of risk alleles carried by an individual. Under this model, genotypes are typically coded as 0, 1, or 2, corresponding to the number of copies of the risk allele, each additional allele is assumed to contribute an equal increment to the trait or disease risk. This approach is widely applied in GWAS, because it provides a simple and biologically plausible approximation of genetic effects while maintaining statistical power compared to more complex dominance models.

The phenotype variable was provided through the phenotype file using the *--pheno* and *--pheno-name* parameters.

To evaluate the potential influence of neonatal and maternal phenotype markers on the association results, logistic regression analyses were performed using six different covariate models. Covariates were included in the regression model using the *--covar* and *--covar-name* options in PLINK v1.9.

For each model, logistic regression was performed using the *--logistic* function. Empirical significance levels were estimated using permutation testing, with one million permutations (*mperm=1000000*). The *hide-covar* option was used to suppress covariate-specific results in the main output file. For each SNP, the analysis produced allele frequencies, OR, corresponding 95% CI and p-values.

The resulting association statistics were exported to output files and subsequently used to identify genetic variants showing potential association with the studied phenotype under different covariate models.

Below is example of command line for running logistic regression for model 1:

```
plink \
--bfile samplesnpqc \
--pheno covariate.txt \
--pheno-name pheno \
--logistic hide-covar mperm=1000000 \
--covar covariate.txt \
--covar-name pheno_sex mothers_age \
--ci 0.95 \
--out logistic_01_sex_maternal_age
```

Logistic regression results were filtered to retain only variants tested under the additive genetic model with p-values below the genome-wide significance threshold ( $p\text{-value} < 5 \times 10^{-8}$ ). The header line was preserved, and the resulting variants were sorted by p-value in ascending order. The filtered results were written to a separate output file for further analysis.

```
{
head -n 1 logistic_01_sex_maternal_age.assoc.logistic
awk 'NR>1 && $5=="ADD" && $12 < 5e-8'
logistic_01_sex_maternal_age.assoc.logistic | sort -k12,12gr
} > significant_logistic_01_sex_maternal_age.txt
```

## Logistic regression association analysis using SNPTEST

Logistic regression association analysis was performed using the SNPTEST v2.5.4 software on non-imputed genotype data. SNPTEST v2.5.4 provides a flexible and robust framework for regression based genetic association testing, even when directly genotyped data are used. SNPTEST v2.5.4 allows consistent implementation of statistical models, including logistic regression with covariate adjustment, which is essential for controlling potential confounding factors. In addition, using SNPTEST v2.5.4 for both non-imputed and imputed datasets ensure methodological consistency across analyses, facilitating direct comparison of results obtained from different analytical approaches.

Prior to association analysis with SNPTEST genotype data were prepared from the quality-controlled dataset with PLINK v2.0. Rare variants with MAF less than one percent are present in only a small number of individuals, especially in studies with limited sample size. This leads to low statistical power, making it difficult to detect true associations. For each autosomal chromosome variants with  $MAF < 0.01$  were excluded and the data were exported in Variant Call Format (VCF). These chromosome-specific files were used as genotype scaffolds for subsequent analyses.

```
for CHR in {1..22}; do
  plink2 --bfile samplesnpqc \
    --maf 0.01 \
    --chr ${CHR} \
    --allow-no-sex \
    --export vcf \
    --out scaffold.chr${CHR}
done
```

The chromosome-specific genotype scaffold files were pre-phased separately for each autosomal chromosome using EAGLE v2.4.1 (Loh *et al.*, 2016). Pre-phasing is the process of estimating haplotypes by determining which alleles occur together on the same chromosome for each individual. Pre-phasing was performed with a reference genetic map based on the GRCh38 genome build, which reflects local recombination rates across the genome. These genetic maps were obtained from the BEAGLE resource (Browning *et al.*, 2021) and used as input in EAGLE v2.4.1 to improve haplotype estimation accuracy.

```

for CHR in {1..22}; do
    eagle --vcf scaffold.chr${CHR}.vcf \
    --geneticMapFile ../GRCh38.map/genetic_map_hg38_withX.txt \
    --numThreads=2 \
    --Kpbwt=20000 \
    --outPrefix prephased_chr${CHR}
done

```

Pre-phased chromosome-specific VCF files were converted to Oxford-format genotype files (GEN and SAMPLE) using PLINK v2.0 for subsequent analysis with SNPTEST v2.5.4. Genotype probabilities were retained during conversion, and sample identifiers were formatted to match the SNPTEST v2.5.4 input requirements.

```

for CHR in {1..22}; do
    plink2 --vcf prephased_chr${CHR}.vcf.gz \
    dosage=GP-force \
    --recode oxford \
    --out prephased_chr${CHR} \
    --double-id
done

```

A sample file required for association analysis was generated from the phenotype file and formatted according to the requirements of SNPTEST v2.5.4. The file contains sample identifiers and phenotype information corresponding to the genotype data. The first header line includes column names, while the second header line specified the data type for each column (D: categorical variables, B: binary phenotypes, C: continuous variables). Each subsequent row represents one individual, including identifiers ID\_1 and ID\_2, missing status, sex and phenotype or covariate variables. The file is space-delimited and maintains the same sample order as the genotype data files. The content of sample file can be found in Appendix 3.

Association analysis was performed using a frequentist score test under an additive genetic model implemented in SNPTEST v2.5.4. The score test evaluates the null hypothesis of no genetic effect in a computationally efficient manner, making it suitable for genome-wide analyses. The additive genetic model assumes a linear relationship between the number of minor alleles, coded as 0, 1, or 2 and the probability of the outcome. Case-control status was defined as the phenotype variable and specified using the *--pheno* parameter. To account for potential confounding effects, selected

covariates were included in the regression model. The resulting output files contain summary statistics for each tested SNP, including effect estimates, standard errors and p-values.

Different combinations of covariates were tested in separate association models to evaluate the potential influence of confounding factors on the association results. For example, neonatal sex and maternal age were included as covariates in the first model.

```
for CHR in {1..22}; do
  snptest -data prephased_chr${CHR}.gen sample.txt \
  -o chr${CHR}_01_sex_maternal_age.snptest \
  -frequentist 1 \
  -lower_sample_limit 30 \
  -method score \
  -pheno pheno \
  -cov_names pheno_sex mothers_age
done
```

The SNPTEST v2.5.4 output files were post-processed using an R script <https://github.com/natalija-osmianskaja/master-thesis/blob/main/snptest.extract.R> to extract variants passing predefined filtering criteria. For each chromosome, result files were filtered using a significance threshold of  $p\text{-value} < 5 \times 10^{-8}$ , and the resulting variants were sorted by p-value to prioritize the most significant associations. The processed results were saved as chromosome-specific files for further evaluation.

```
for CHR in {1..22}; do
  Rscript ../../r_scripts/snptest.extract.R \
  chr${CHR}_01_sex_maternal_age.snptest 0.4 5e-8 \
  | { read header; echo "$header"; sort -k42,42g; } \
  > chr${CHR}_01_sex_maternal_age.txt
done

(
head -1 chr1_01_sex_maternal_age.txt
awk 'FNR>1' chr*_01_sex_maternal_age.txt | LC_ALL=C sort -g -
k8,8
) > significant_snptest1_01_sex_maternal_age.txt
```

## Logistic regression association analysis using SNPTEST on imputed data

When SNPTEST is applied to non-imputed genotype data, the analysis is restricted to the set of variants that were directly measured by the genotyping array. Genotyping arrays capture only a predefined subset of common variants across the genome, so large portions of genetic variation remain unobserved. This leads to limited genomic coverage, meaning that many regions and variants within them are simply not tested.

Genotype imputation is a statistical method used in genetic studies to predict missing or untyped genetic variants based on observed genotype data and known patterns of genetic variation from a reference population. Imputation uses reference panel, such as the 1000 Genomes Project (The 1000 Genomes Project Consortium *et al.*, 2015), which contains dense genotype information from many individuals. By comparing the study data with these reference haplotypes, statistical algorithms infer the most likely genotypes at variants that were not directly genotyped.

The result of imputation is typically genotype probabilities or dosage values for millions of additional variants. This process increases genomic coverage, improves the ability to detect genetic associations, and allows results to be compared across different studies, even when different genotyping platforms are used. However, when imputed data are analyzed in small cohorts, effect estimates may be unstable, and random variation may appear as significant associations, which increases the risk of false positive results.

The first step of imputation involves downloading the reference panel data from the 1000 Genomes Project repository (1000 Genomes Project High-coverage phased genotype data, 2020).

After downloading, chromosome identifiers must be recoded, as the reference data use the format “CHR1”, whereas the genotyped dataset uses numeric chromosome labels such as “1”. This harmonization step is required to ensure compatibility between datasets for downstream imputation analysis. This can be done by running BCFTOOLS (Danecek *et al.*, 2021):

```
for CHR in {1..22}; do
  bcftools annotate \
    --rename-chrs <(printf "chr%s\t%s\n" "$CHR" "$CHR") \
    -Oz -o ../hg38_recoded/CCDG_chr${CHR}_as_${CHR}.vcf.gz \
    CCDG_14151_B01_GRM_WGS_2020-08-05_chr${CHR}.filtered.shapeit2-
    duohmm-phased.vcf.gz
done
```

The genotyped data must be scaffolded and pre-phased using EAGLE, as described in the previous section. After that, the resulting VCF files are processed using BCFTOOLS to normalize variant representation. VCF files may contain positions with more than two alleles (multiallelic variants), while imputation pipeline requires biallelic variants only. Multiallelic variants are split into biallelic records and saved in compressed VCF format for each chromosome, ensuring consistency and compatibility with imputation analysis with following command:

```
for CHR in {1..22}; do
  bcftools norm -m -both -O z \
  -o cleaned.prephased_chr${CHR}.vcf.gz \
  prephased_chr${CHR}.vcf.gz
done
```

Before imputation the data should be indexed to enable fast and efficient access to specific genomic regions:

```
for CHR in {1..22}; do
  tabix -f -p vcf cleaned.prephased_chr${CHR}.vcf.gz
done
```

Imputation is performed using BEAGLE v5.5 software:

```
for CHR in {1..22}; do
  java -Xmx3641m -jar beagle.17Dec24.224.jar \
  gt=cleaned.prephased_chr${CHR}.vcf.gz \
  ref=../hg38_recoded/CCDG_chr${CHR}_as_${CHR}.vcf.gz \
  map=../GRCh38.map/plink.chr${CHR}.GRCh38.map \
  out=chrchr${CHR}_imputed nthreads=4 ne=20000 gp=true
done
```

Although multiallelic variants may be removed prior to imputation, it is still necessary to perform post-imputation filtering using BCFTOOLS. During imputation with BEAGLE, multiallelic variants can be reintroduced from the reference panel, and additional variant types such as indels may also appear in the output. Therefore, filtering to retain only biallelic SNPs ensures that the dataset meets the requirements of downstream association analysis tools, such as SNPTTEST v2.5.4, which typically expect biallelic variants only. Furthermore, indexing the filtered VCF files

using TABIX is required to enable efficient access to genomic regions during subsequent analyses. Overall, this ensures consistency, compatibility and data quality in the final imputed dataset.

```
for CHR in {1..22}; do
  bcftools view -m2 -M2 -v snps chrchr${CHR}_imputed.vcf.gz -Oz
-o chr${CHR}_imputed.biallelic.vcf.gz
  tabix -p vcf chr${CHR}_imputed.biallelic.vcf.gz
done
```

Afterwards, the data should be converted using PLINK v2.0 to Oxford-format genotype files (GEN and SAMPLE) for subsequent analysis with SNPTEST v2.5.4:

```
for CHR in {1..22}; do
  plink2 --vcf chr${CHR}_imputed.biallelic.vcf.gz dosage=DS --
recode oxford --out chr${CHR}_imputed.biallelic.impute --double-
id
done
```

Finally, SNPTEST v2.5.4 can be run using the same commands and models as described in the previous section, allowing direct comparison between results obtained from genotyped and imputed datasets.

## Manhattan Plot Visualization

Manhattan plots of association results were generated in R using the GGLOT2 library. Association output files were imported and processed to retain relevant columns, including SNP identifier, chromosome, base-pair position (BP) and p-value, while entries with missing or invalid values were removed. P-values were transformed to  $-\log_{10}(P)$  for visualization. Genomic positions were converted into cumulative coordinates across chromosomes to allow continuous representation along the x axis. SNPs were plotted using alternating colors for consecutive chromosomes to improve visual distinction between chromosomal regions.

Genome-wide significance ( $p\text{-value} < 5 \times 10^{-8}$ ) and suggestive significance ( $p\text{-value} < 5 \times 10^{-5}$ ) thresholds were indicated by horizontal reference lines. The most significant SNPs exceeding these thresholds were identified and optionally annotated on the plot. Separate Manhattan plots were generated for each model, and outputs were saved as high-resolution PNG files suitable for publication.

## QQ Plot Visualization

Quantile-quantile (QQ) plots of logistic regression results were generated in R using GGPLOT2. Association files were processed to extract p-values, with invalid or missing entries removed prior to analysis.

Expected p-values under the null hypothesis of no association were calculated based on the uniform distribution and transformed to  $-\log_{10}(P)$ . Observed p-values were similarly transformed and plotted against expected values to assess deviations from the null distribution.

A diagonal reference line representing the expected distribution under the null hypothesis was included to facilitate visual comparison. Deviations from this line were used to evaluate inflation or deflation of test statistics.

The genomic inflation factor ( $\lambda$ ) is expected to be close to 1 under a well-calibrated model; values below 1 indicate overly conservative test statistics (deflation), whereas elevated values indicate inflation and an excess of small p-values.

The genomic inflation factor ( $\lambda$ ) was calculated for each model to quantify systematic deviation from the null expectation and was annotated on the corresponding plots. Plots were exported as high-resolution PNG files (300 dpi) for publication.

The source code used for logistic regression Manhattan and QQ plot generation is provided in [https://github.com/natalija-osmianskaja/master-thesis/blob/main/logistic\\_regression\\_plots.R](https://github.com/natalija-osmianskaja/master-thesis/blob/main/logistic_regression_plots.R)

The source code used for generating Manhattan and QQ for SNPTTEST data can be found at:

[https://github.com/natalija-osmianskaja/master-thesis/blob/main/snptest\\_plots.R](https://github.com/natalija-osmianskaja/master-thesis/blob/main/snptest_plots.R)

[https://github.com/natalija-osmianskaja/master-thesis/blob/main/snptest\\_imputed\\_plots.R](https://github.com/natalija-osmianskaja/master-thesis/blob/main/snptest_imputed_plots.R)

## Identification and prioritization of overlapping SNPs across association models

To enable comparison of association signals across different models, a Python-based workflow was developed to identify overlapping SNPs and prioritize variants based on statistical significance and consistency across models. This approach was applied uniformly to results obtained from logistic regression analyses with PLINK v1.9, SNPTTEST v2.5.4 output performed on non-imputed and imputed datasets.

For each model-specific result file, relevant columns containing SNP identifiers, genomic positions (base-pair coordinates, GRCh38), and p-values were extracted. Column names were standardized across input files to ensure compatibility between outputs generated by different tools. Data cleaning procedures included removal of missing or non-numeric values and harmonization of SNP identifiers to avoid duplication due to formatting inconsistencies.

All processed datasets were merged into a single combined table, and SNPs were grouped by their unique identifier and genomic position. For each SNP, the set of models in which it was detected was recorded, along with the total number of models and the minimum p-value observed across all analyses. This enabled identification of variants consistently detected across multiple analytical frameworks.

To prioritize candidate variants, SNPs were ranked based on two criteria: the number of models in which the SNP was identified, reflecting reproducibility across analyses, and the lowest observed p-value, reflecting statistical strength of association. Variants detected in a greater number of models and exhibiting lower p-values were assigned higher priority for downstream interpretation.

The workflow generated both a comprehensive merged dataset of all SNPs and a summarized intersection table containing SNP-level metrics, including model representation and best p-value. These outputs were used to guide selection of variants for further functional annotation and biological interpretation.

The content of the script can be found in [https://github.com/natalija-osmianskaja/master-thesis/blob/main/snp\\_intersections.py](https://github.com/natalija-osmianskaja/master-thesis/blob/main/snp_intersections.py)

Example of running the script for logistic regression models:

```
python snp_intersections.py \
  --analysis-name logistic \
  --base-dir data \
  --out-dir results \
  --files \
    1:significant_logistic_01_sex_maternal_age.txt \
    2:significant_logistic_02_weight_category_exact_ga.txt \
    3:significant_logistic_03_gd_pre_spb_latency.txt \
    4:significant_logistic_04_mothers_age_nr_preg.txt \
    5:significant_logistic_05_prev_preg_mother_ancestry.txt \
    6:significant_logistic_06_w_ga_weight_sex.txt \
  --snp-cols snp \
```

```
--position-col bp \  
--pvalue-col p
```

Example of running the script for SNPTTEST non-imputed data:

```
python snp_intersections.py \  
  --analysis-name snptest \  
  --base-dir data \  
  --out-dir results \  
  --files \  
    1:significant_snptest1_01_sex_maternal_age.txt \  
    2:significant_snptest1_02_weight_category_exact_ga.txt \  
    3:significant_snptest1_03_gd_pre_spb_latency.txt \  
    4:significant_snptest1_04_mothers_age_nr_preg.txt \  
    5:significant_snptest1_05_prev_preg_mother_ancestry.txt \  
  --snp-cols alternate_ids rsid \  
  --position-col position \  
  --pvalue-col frequentist_add_pvalue
```

Example of running the script for SNPTTEST imputed data:

```
python snp_intersections.py \  
  --analysis-name snptest \  
  --base-dir data \  
  --out-dir results \  
  --files \  
    1:significant_snptest2_01_sex_maternal_age.txt \  
    2:significant_snptest2_02_weight_category_exact_ga.txt \  
    3:significant_snptest2_03_gd_pre_spb_latency.txt \  
    4:significant_snptest2_04_mothers_age_nr_preg.txt \  
    5:significant_snptest2_05_prev_preg_mother_ancestry.txt \  
  --snp-cols alternate_ids rsid \  
  --position-col position \  
  --pvalue-col frequentist_add_pvalue
```

## Annotation

In this study, following genotype imputation, variant identifiers were represented in chromosome-position-allele format (e.g., 17:12513046:G:A) rather than reference SNP (rs17713682) requiring additional querying of dbSNP (National Library of Medicine; National Institutes of Health; U.S. Department of Health and Human Services, 2026) to retrieve corresponding reference SNP

identifiers where available. The dbSNP resource is a comprehensive public repository of genetic variation (Sherry, 2001). It provides rsIDs, genomic coordinates, allele information, population frequency data and functional annotations, facilitating consistent identification and comparison of variants across studies. In the context of this analysis, dbSNP was used to map imputed variants lacking rsIDs to their corresponding reference identifiers, when such mappings were available.

Functional annotation of genetic variants was carried out using GeneBe (Variant annotation and interpretation tool, 2024), which applies an automated framework for the assignment of ACMG variant pathogenicity criteria (Stawiński and Płoski, 2024).

Gene annotation and functional interpretation also were performed using the GeneCards (Stelzer *et al.*, 2016) database. GeneCards is an integrative platform that provides comprehensive information on human genes, including gene function, expression profiles, associated pathways and links to other genomic resources. In this study, genes located in proximity to prioritized SNPs were identified based on genomic coordinates, and corresponding gene entries were queried in GeneCards. The database was used to retrieve information on gene function, biological relevance, and potential involvement in disease-related pathways, supporting the interpretation of genetic association results.

### Declaration on the use of artificial intelligence

ChatGPT (OpenAI, 2026) was used solely for grammar correction and language polishing. No content generation or analytical interpretation was relied upon without critical evaluation by the author.

# RESULTS

## Study cohort

SNP genotype data for the case group (generated using microarray genotyping) and SNP variant data for the control group (derived from whole-genome sequencing) were provided by the Population Genomics Laboratory, Institute for Translational Health Research, Faculty of Medicine, Vilnius University.

Genotype quality control resulted in a final dataset comprising 54 individuals, including 29 cases consisting of 12 females and 17 males, and 25 controls consisting of 10 females and 15 males and 220,064 SNPs. No individuals were excluded based on genotype missingness. A total of 1344 SNPs were removed due to deviation from Hardy-Weinberg equilibrium, while additional variants were filtered out based on missingness and minor allele frequency thresholds.

Following data quality control, the resulting dataset was used for downstream genetic analyses, including candidate variant analysis and genome-wide association analyses performed using logistic regression with PLINK v1.9 and SNPTEST v2.5.4 on both non-imputed and imputed genotype data.

For each association analysis, six covariate models were applied consistently. Association results were summarized using tabulated outputs of significant or top-ranked SNPs, and the genome-wide distribution of association signals was visualized using Manhattan plots and QQ plots.

## Candidate variant analysis results

Among the 43 candidate SNPs identified from previous studies (Appendix 1), only the rs17713682 variant was present in our study dataset after quality control filtering. The SNP rs17713682 was identified as having an offspring-driven genetic effect (OD), indicating that variation in the fetal genome contributes to its association with gestational duration, with a moderate positive effect size suggesting a role in prolonging pregnancy (Solé-Navais *et al.*, 2023). According to GeneBe (Variant annotation and interpretation tool, 2024), rs17713682 is located in a non-coding region, specifically annotated as an intergenic variant not predicted to directly affect protein sequence.

This SNP is located on chromosome 17 at position 12513046 bp (GRCh38) and was analyzed for its association with the studied phenotype using logistic regression and Fisher's exact test.

Logistic regression analysis performed with PLINK v2.0 indicated that the T allele was associated with reduced odds of the PTB (OR = 0.31), suggesting a protective association among carriers. However, the association did not reach statistical significance (p-value = 0.069), and the confidence interval included 1, indicating uncertainty in the estimated effect.

The result of candidate rs17713682 logistic regression analysis is presented in Table 1.

**Table 1.** Association of candidate SNP rs17713682 with the studied phenotype.

| <b>SNP<br/>(chr:position,<br/>GRCh38)</b> | <b>Gene</b> | <b>Effect<br/>allele</b> | <b>Reference<br/>allele</b> | <b>OR<br/>(95% CI)</b> | <b>P-<br/>value</b> | <b>N</b> |
|-------------------------------------------|-------------|--------------------------|-----------------------------|------------------------|---------------------|----------|
| rs17713682<br>(chr17:12513046)            | -           | T                        | G                           | 0.31<br>(0.09-1.09)    | 0.069               | 54       |

*\* Position (GRCh38) - genomic position based on the GRCh38 reference genome; Gene - nearest annotated gene (- indicates intergenic region); Effect allele - allele associated with the estimated effect; Reference allele - comparison allele; OR - odds ratio; CI - confidence interval; P-value - significance level of association; N - number of individuals included in the analysis.*

Fisher's exact test with PLINK v1.9 produced similar results (OR = 0.38, p-value = 0.102), supporting the absence of a statistically significant association (Table 2).

**Table 2.** Association analysis of candidate SNP rs17713682 using Fisher's exact test.

| <b>SNP<br/>(chr:position,<br/>GRCh38)</b> | <b>Effect<br/>allele</b> | <b>Reference<br/>allele</b> | <b>Freq<br/>cases<br/>(F_A)</b> | <b>Freq<br/>controls<br/>(F_U)</b> | <b>OR</b> | <b>P-value</b> |
|-------------------------------------------|--------------------------|-----------------------------|---------------------------------|------------------------------------|-----------|----------------|
| rs17713682<br>(chr17:12513046)            | T                        | G                           | 0.086                           | 0.200                              | 0.38      | 0.102          |

*\* Position (GRCh38) - genomic position based on the GRCh38 reference genome; Effect allele - allele associated with the estimated effect; Reference allele - comparison allele; F\_A - allele frequency in cases; F\_U - allele frequency in controls; OR - odds ratio for the effect allele; P-value - statistical significance of association.*

Overall, the candidate variant analysis did not provide statistically significant evidence for an association between rs17713682 and PTB. The absence of a significant association may reflect the limited sample size and restricted number of candidate variants available in the dataset. Further studies with larger cohorts and broader variant coverage are required to clarify the potential role of this variant.

### Logistic regression association analysis results using PLINK

Logistic regression analyses were performed using PLINK 1.9 for six covariate models to evaluate the potential influence of clinical and maternal factors on genetic associations. No variants reached the conventional genome-wide significance threshold ( $p < 5 \times 10^{-8}$ ) in any model. However, several SNPs showed lower p-values ( $p < 5 \times 10^{-5}$ ) primarily in model 1 (neonatal sex and maternal age), model 3 (clinical complications and latency) and model 4 (maternal age and number of pregnancies).

The detected association signals were not fully consistent across models, nevertheless, several variants, including rs3129111, rs11683045, rs10065905, rs202197010, rs9666203 and rs10250442, were observed in all three models. Additional variants, including rs200527317, rs66964741, rs62455706, rs2370186 and rs11237479, were identified in at least two models.

Most identified variants showed odds ratios below 1, however, wide confidence intervals and occasional extreme effect sizes ( $OR > 50$ ) indicate instability of regression models. These findings are likely related to the limited sample size and sparse genotype counts within the dataset. Overall, these findings should be interpreted as exploratory and require validation in larger cohorts.

The top-ranking variants identified in each model are summarized in Tables 3-6. Model 1 produced the largest number of variants with low p-values, whereas models 3 and 4 identified fewer but partially overlapping signals. No variants meeting the reporting thresholds were identified in models 2, 5 or 6.

**Table 3.** Logistic regression association analysis using PLINK for model 1, including neonatal sex and maternal age as covariates

| <b>SNP<br/>(chr:position, GRCh38)</b> | <b>Effect allele</b> | <b>OR (95% CI)</b>        | <b>P-value</b>          |
|---------------------------------------|----------------------|---------------------------|-------------------------|
| rs200527317<br>(chr13:48027628)       | T                    | 0.00851 (0.000872-0.0831) | $0.9682 \times 10^{-5}$ |
| rs3129111<br>(chr6:29099995)          | T                    | 0.00841 (0.000847-0.0835) | $0.9332 \times 10^{-5}$ |
| rs66964741<br>(chr17:78869989)        | G                    | 0.01539 (0.00208-0.114)   | $0.8459 \times 10^{-5}$ |
| rs11683045<br>(chr2:237188003)        | G                    | 0.00781 (0.000766-0.0796) | $0.8343 \times 10^{-5}$ |
| rs10065905<br>(chr5:153326566)        | G                    | 0.00473 (0.000359-0.0622) | $1.206 \times 10^{-5}$  |
| rs62455706<br>(chr7:65188030)         | A                    | 0.01378 (0.00178-0.1068)  | $0.8617 \times 10^{-5}$ |
| rs9666203<br>(chr11:67868856)         | A                    | 0.02093 (0.00332-0.132)   | $0.9398 \times 10^{-5}$ |
| rs10250442<br>(chr7:106981726)        | T                    | 0.00932 (0.00100-0.0866)  | $1.137 \times 10^{-5}$  |
| rs62110898<br>(chr18:76401830)        | T                    | 0.02868 (0.00530-0.155)   | $1.162 \times 10^{-5}$  |
| rs28415028<br>(chr6:86533364)         | A                    | 0.00644 (0.000606-0.0684) | $1.045 \times 10^{-5}$  |

|                               |   |                          |                        |
|-------------------------------|---|--------------------------|------------------------|
| rs7439888<br>(chr4:89094314)  | T | 0.01511 (0.00227-0.1008) | $1.185 \times 10^{-5}$ |
| rs2370186<br>(chr2:195066289) | C | 0.02615 (0.00498-0.1373) | $1.022 \times 10^{-5}$ |
| rs80238130<br>(chr19:4511271) | T | 0.02819 (0.00550-0.1447) | $1.171 \times 10^{-5}$ |
| rs202197010<br>(chr5:8089059) | C | 0.01772 (0.00285-0.1104) | $1.315 \times 10^{-5}$ |

*\* Position (GRCh38) - genomic position based on the GRCh38 reference genome; Effect allele - allele associated with the estimated effect; OR (95% CI) - odds ratio and its 95% confidence interval for the effect; P-value - significance level of association*

**Table 4.** Logistic regression association analysis using PLINK for model 3, including GD or preeclampsia, SPB and PPRM and latency period as covariates

| <b>SNP<br/>(chr:position, GRCh38)</b> | <b>Effect allele</b> | <b>OR (95% CI)</b>   | <b>P-value</b>        |
|---------------------------------------|----------------------|----------------------|-----------------------|
| rs3129111<br>(chr6:29099995)          | T                    | 0.017 (0.0024-0.123) | $5.15 \times 10^{-5}$ |
| rs11683045<br>(chr2:237188003)        | G                    | 0.012 (0.0013-0.108) | $8.12 \times 10^{-5}$ |

**Table 5.** Logistic regression association analysis using PLINK for model 4, including maternal age and number of pregnancies as covariates

| <b>SNP<br/>(chr:position, RCh38)</b> | <b>Effect allele</b> | <b>OR (95% CI)</b> | <b>P-value</b> |
|--------------------------------------|----------------------|--------------------|----------------|
|--------------------------------------|----------------------|--------------------|----------------|

|                                |   |                        |                       |
|--------------------------------|---|------------------------|-----------------------|
| rs3129111<br>(chr6:29099995)   | T | 0.018 (0.0029-0.115)   | $2.10 \times 10^{-5}$ |
| rs11683045<br>(chr2:237188003) | G | 0.029 (0.0054-0.153)   | $3.22 \times 10^{-5}$ |
| rs202197010<br>(chr5:8089059)  | C | 0.0055 (0.00048-0.064) | $3.22 \times 10^{-5}$ |
| rs10065905<br>(chr5:153326566) | G | 0.0068 (0.00065-0.072) | $3.38 \times 10^{-5}$ |
| rs9666203<br>(chr11:67868856)  | A | 0.022 (0.0035-0.136)   | $4.11 \times 10^{-5}$ |
| rs10250442<br>(chr7:106981726) | T | 0.010 (0.0011-0.092)   | $4.66 \times 10^{-5}$ |
| rs11237479<br>(chr11:78392632) | T | 55.76 (7.99-389.1)     | $4.98 \times 10^{-5}$ |

Manhattan plots generated from logistic regression association analyses across six covariate models (Figure 4) showed the absence of genome-wide significant associations. Variant with lower p-value ( $p < 5 \times 10^{-5}$ ) were primarily observed in model 1, 3 and 4, while the remaining models showed minimal evidence of association. However, the overall lack of robust replication, together with wide confidence intervals and extreme effect size estimates, suggests that many of the observed associations may reflect false-positive findings arising from limited sample size and sparse genotype counts. These results indicate limited statistical power and highlight the strong influence of covariate specification on the detected association signals.

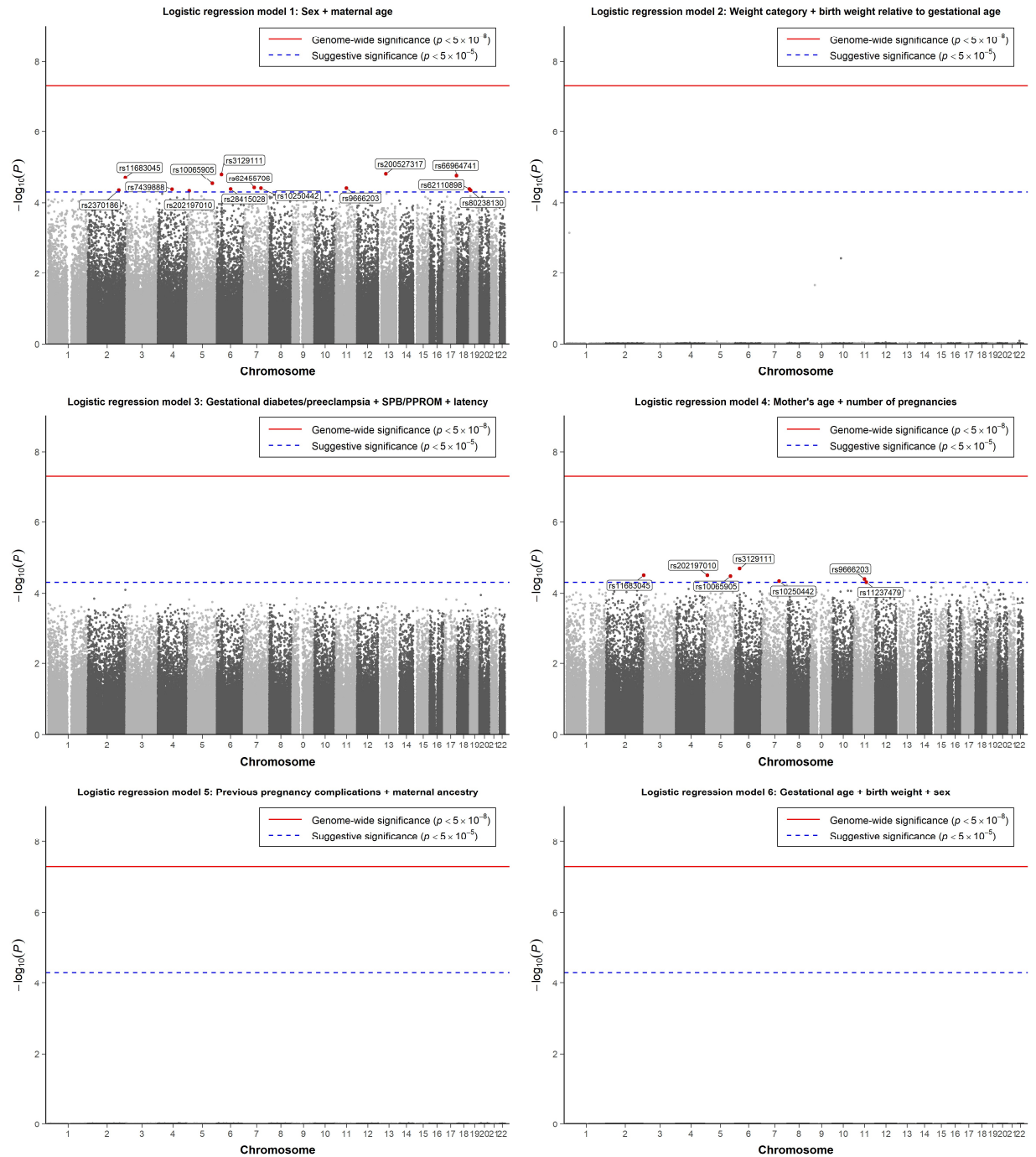


**Figure 4.** Manhattan plots from logistic regression association analysis using PLINK across six covariate models

QQ plots (Figure 5) were used to compare the distribution of observed association test statistics relative to those expected under the null hypothesis of no association. QQ plots provide a diagnostic tool for detecting systematic bias, inflation or deflation. The QQ plots revealed

substantial deviations from the expected null distribution across models, indicating poor calibration of test statistics. The genomic inflation factor ( $\lambda$ ) is expected to be close to 1 under a well-calibrated model. Values below 1 indicate overly conservative test statistics (deflation), whereas elevated values indicate inflation and an excess of small p-values.

Models 1, 3 and 4 demonstrated inflation of association statistics ( $\lambda \approx 1.7 - 1.9$ ), indicating an excess of low p-values. Given the limited sample size and sparse genotype distribution, this inflation is more likely attributable to model instability, sparse-data effects, or residual confounding than to true genetic associations. In contrast, models 2, 5 and 6 showed pronounced deflation ( $\lambda \approx 0$ ), indicating overly conservative association estimates and reduced statistical power. The considerable variability in genomic inflation factors across models indicates strong sensitivity of the association results to covariate specification. Overall, the QQ plot analysis suggests that the observed association signals are likely to be influenced by methodological limitations, including small sample size, sparse genotype counts and model instability, rather than robust underlying genetic effects.

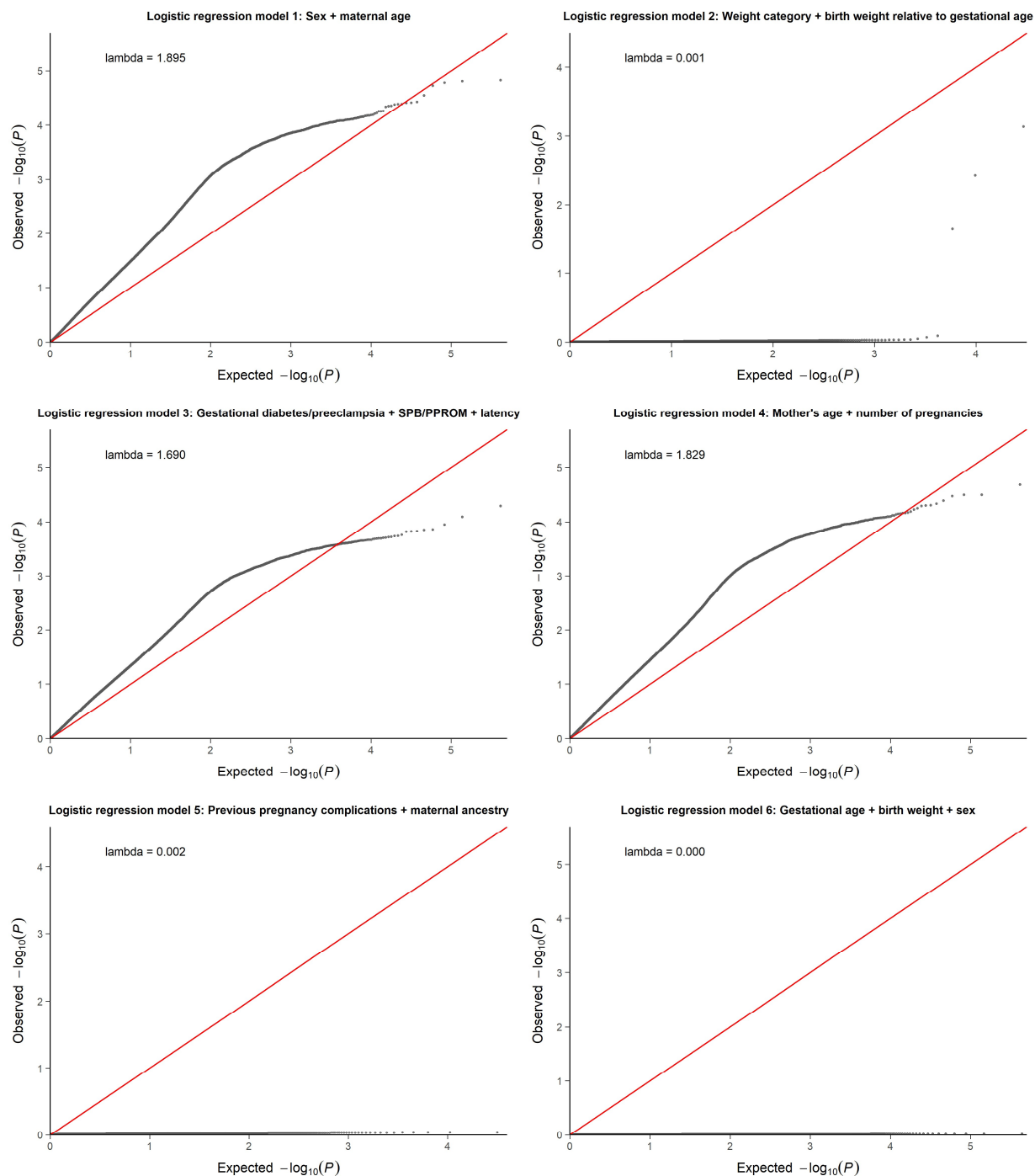


**Figure 5.** QQ plots from logistic regression association analysis using PLINK across six covariate models

While no genome-wide significant associations were identified, several loci demonstrated recurrent association signals across multiple models and may warrant further investigation in larger cohorts or meta-analysis approaches to validate their potential role in the studied phenotype. To facilitate downstream analysis, SNPs were prioritized based on statistical significance ( $p < 5 \times 10^{-5}$ ) and reproducibility across multiple models, with variants identified in more than one model considered high-priority candidate loci.

**Table 6.** SNP prioritization for further analysis based on statistical significance and consistency across logistic regression association analysis using PLINK

| SNP<br>(chr:position, RCh38)   | Models<br>identified | Best p-value          | Priority level | Notes                                              |
|--------------------------------|----------------------|-----------------------|----------------|----------------------------------------------------|
| rs3129111<br>(chr6:29099995)   | 1, 3, 4              | $1.55 \times 10^{-5}$ | High           | Strongest and most consistent signal across models |
| rs11683045<br>(chr2:237188003) | 1, 3, 4              | $1.89 \times 10^{-5}$ | High           | Reproduced in three models                         |
| rs10065905<br>(chr5:153326566) | 1, 3, 4              | $2.86 \times 10^{-5}$ | High           | Reproduced in three models                         |
| rs202197010<br>(chr5:8089059)  | 1, 3, 4              | $3.22 \times 10^{-5}$ | High           | Reproduced in three models                         |
| rs9666203<br>(chr11:67868856)  | 1, 3, 4              | $3.89 \times 10^{-5}$ | High           | Reproduced in three models                         |
| rs10250442<br>(chr7:106981726) | 1, 3, 4              | $3.93 \times 10^{-5}$ | High           | Reproduced in three models                         |
| rs200527317<br>(chr1:48027628) | 1, 3                 | $1.49 \times 10^{-5}$ | Medium         | Strong p-value but limited to two models           |

|                               |      |                       |        |                          |
|-------------------------------|------|-----------------------|--------|--------------------------|
| rs66964741<br>(chr1:78869989) | 1, 3 | $1.65 \times 10^{-5}$ | Medium | Reproduced in two models |
| rs62455706<br>(chr1:65188030) | 1, 3 | $3.76 \times 10^{-5}$ | Medium | Reproduced in two models |
| rs2370186<br>(chr2:195066289) | 1, 3 | $4.44 \times 10^{-5}$ | Medium | Reproduced in two models |
| rs11237479<br>(chr3:78392632) | 3, 4 | $4.98 \times 10^{-5}$ | Medium | Reproduced in two models |

Functional annotation of highly prioritized SNPs was performed using the (Variant annotation and interpretation tool, 2024) , (National Library of Medicine; National Institutes of Health; U.S. Department of Health and Human Services, 2026) and GeneCards (Stelzer *et al.*, 2016) (Table 7).

**Table 7.** Functional annotation of prioritized SNPs identified in the logistic regression association analysis using PLINK

| <b>SNP<br/>(chr:position,<br/>RCh38)</b> | <b>Gene</b>            | <b>Location</b> | <b>Functional consequence</b>                                                                                                                                                                                                                                         |
|------------------------------------------|------------------------|-----------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| rs3129111<br>(chr6:29099995)             | <i>OR2J1</i>           | intronic        | The variant is located within an intronic region of the <i>OR2J1</i> gene.<br><br>The <i>OR2J1</i> gene encodes an olfactory receptor and is primarily involved in sensory perception pathways. No clear association with pregnancy-related traits has been reported. |
| rs11683045<br>(chr2:237188003)           | <i>ENSG00000224132</i> | intronic        | The variant is located within a long non-coding RNA locus, with potential regulatory rather than protein-altering effects.                                                                                                                                            |

|                                 |                        |            |                                                                                                                                                                                                                  |
|---------------------------------|------------------------|------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                 |                        |            | <i>ENSG00000224132</i> is an RNA gene and is affiliated with the lncRNA class.                                                                                                                                   |
| rs10065905<br>(chr5:153326566)  | <i>ENSG00000310234</i> | intronic   | The variant is located within a long non-coding RNA locus, with potential regulatory rather than protein-altering effects.<br><br><i>ENSG00000310234</i> is an RNA gene and is affiliated with the lncRNA class. |
| rs202197010<br>(chr5:8089059)   | <i>ENSG00000308020</i> | intronic   | The variant is located within a long non-coding RNA locus, with potential regulatory rather than protein-altering effects.<br><br><i>ENSG00000308020</i> is an RNA gene and is affiliated with the lncRNA class. |
| rs9666203<br>(chr11:67868856)   | <i>LOC112268076</i>    | intronic   | The variant is located within an intronic region of <i>LOC112268076</i> gene.<br><br><i>LOC112268076</i> (Translation Initiation Factor IF-2-Like) is a protein coding gene.                                     |
| rs10250442<br>(chr7:106981726)  | -                      | intergenic | This variant is located in a non-coding region and does not directly alter the protein sequence.                                                                                                                 |
| rs200527317<br>(chr13:48027628) | <i>LOC124903172</i>    | intronic   | The variant is located within an intronic region of <i>LOC124903172</i> gene.<br><br><i>LOC124903172</i> is an RNA gene and is affiliated with the ncRNA class.                                                  |
| rs66964741<br>(chr17:78869989)  | <i>TIMP2</i>           | intronic   | The variant is located within an intronic region of <i>TIMP2</i> gene.<br><br><i>TIMP2</i> gene is a member of the TIMP gene family. The proteins encoded by this gene family                                    |

|                                |                     |          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|--------------------------------|---------------------|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                |                     |          | are natural inhibitors of the matrix metalloproteinases, a group of peptidases involved in degradation of the extracellular matrix. In addition to an inhibitory role against metalloproteinases, the encoded protein has a unique role among TIMP family members in its ability to directly suppress the proliferation of endothelial cells. As a result, the encoded protein may be critical to the maintenance of tissue homeostasis by suppressing the proliferation of quiescent tissues in response to angiogenic factors, and by inhibiting protease activity in tissues undergoing remodelling of the extracellular matrix. |
| rs62455706<br>(chr7:65188030)  | <i>INTS4P1</i>      | intronic | The variant is located within an intronic region of <i>INTS4P1</i> gene.<br><i>INTS4P1</i> (Integrator Complex Subunit 4 Pseudogene 1) is a pseudogene.                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| rs2370186<br>(chr2:195066289)  | <i>LOC105376755</i> | intronic | The variant is located within an intronic region of <i>LOC105376755</i> gene.<br><i>LOC105376755</i> is an RNA gene and is affiliated with the ncRNA class.                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| rs11237479<br>(chr11:78392632) | <i>GAB2</i>         | intronic | The variant r is located within an intronic region of the <i>GAB2</i> gene.<br><i>GAB2</i> is a member of the GRB2-associated binding protein (GAB) gene family. These proteins contain pleckstrin homology (PH) domain and bind SHP2 tyrosine phosphatase and GRB2 adapter protein. They act as adapters for                                                                                                                                                                                                                                                                                                                       |

|  |  |  |                                                                                                                                                                                                                                                                                                         |
|--|--|--|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|  |  |  | <p>transmitting various signals in response to stimuli through cytokine and growth factor receptors, and T- and B-cell antigen receptors. The protein encoded by this gene is the principal activator of phosphatidylinositol-3 kinase in response to activation of the high affinity IgE receptor.</p> |
|--|--|--|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

Functional annotation of prioritized SNPs identified in the logistic regression association analysis using PLINK (Table 7) showed that most variants were located within intronic regions or non-coding RNA loci, whereas only one variant was classified as intergenic. No variants were found within protein-coding exons, suggesting that the identified loci are unlikely to directly alter protein structure and may instead influence regulatory mechanisms affecting gene expression.

Several SNPs were mapped to genes with limited or unclear functional relevance to the studied phenotype, including olfactory receptor genes (e.g., *OR2J1*) and multiple long non-coding RNA (lncRNA) loci. These non-coding regions may have regulatory roles. However, their biological significance in this context remains uncertain. Additionally, several variants were located within poorly characterized or predicted genomic regions (e.g., pseudogenes), further limiting biological interpretation.

Among the annotated genes, *TIMP2* may represent a biologically relevant candidate due to its involvement in extracellular matrix remodeling and tissue homeostasis. Another variant was located within *GAB2*, a signaling adaptor gene implicated in cellular response pathways. Despite these potentially relevant genes, the overall pattern of predominantly intronic and non-coding annotations, combined with limited functional evidence, suggests that the identified associations should be interpreted with caution and may reflect indirect regulatory effects or false-positive findings rather than clear causal mechanisms. Further functional studies and replication analyses in larger cohorts are required to clarify the biological relevance of these loci.

## Logistic regression association analysis results using SNPTEST

SNPTEST v2.5.4 association analysis was performed using the non-imputed genotype dataset and the same six covariate models described in the Methods section and previously applied in the logistic regression analysis using PLINK. The aim was to evaluate whether genetic associations with the studied phenotype remained detectable after adjustment for different maternal and neonatal factors.

Models 1-5 produced a large number of variants exceeded the conventional genome-wide significance threshold of  $p\text{-value} < 5 \times 10^{-8}$  and effect size estimates were consistently large across all models. In particular, the unusually high number of significant variants, diffuse distribution of association signals across the genome, and consistently large effect size estimates strongly suggest inflation of association statistics, model instability, sparse-data effects and potential overfitting.

Model 1 (neonatal sex and maternal age) identified 626 variants, variants exceeding the genome-wide significant threshold. The most significant variants are: rs75516371 ( $p\text{-value } 7.49 \times 10^{-11}$ ), rs479547 ( $p\text{-value } 9.98 \times 10^{-11}$ ) and rs201101435 ( $p\text{-value } 1.17 \times 10^{-10}$ ). Several variants demonstrated extremely large effect size estimates  $|\beta| > 2$ , further indicating unstable model estimates.

Model 2 (birth weight category and exact birth-weight percentile for GA and sex) identified 396 significant variants. The strongest associations were observed for rs2633428 ( $p\text{-value } 7.27 \times 10^{-10}$ ), rs546293974 ( $p\text{-value } 7.59 \times 10^{-10}$ ) and rs6047156 ( $p\text{-value } 9.38 \times 10^{-10}$ ).

Model 3 (GD or preeclampsia, SPB and PPRM and latency period) identified 351 significant variants. The strongest associations were observed for rs62079648 ( $p\text{-value } 6.89 \times 10^{-10}$ ), rs508075 ( $p\text{-value } 8.56 \times 10^{-10}$ ) and rs6875908 ( $p\text{-value } 9.62 \times 10^{-10}$ ).

Model 4 (maternal age and number of pregnancies) identified 559 variants exceeding genome-wide significant threshold. The strongest signals were observed for rs75516371 ( $p\text{-value } 1.69 \times 10^{-10}$ ), rs78315775 ( $p\text{-value } 2.30 \times 10^{-10}$ ) and rs71494354 ( $p\text{-value } = 2.32 \times 10^{-10}$ ).

Model 5 (previous pregnancy complications and maternal ancestry) identified 221 genome-wide significant variants. The strongest associations were observed for rs546293974 ( $p\text{-value } 1.80 \times 10^{-9}$ ), rs10083699 ( $p\text{-value } 2.01 \times 10^{-9}$ ) and rs2325829 ( $p\text{-value } 2.08 \times 10^{-9}$ ).

Model 6 (GA, birth weight and neonatal sex) did not identify any genome-wide significant variants.

The Manhattan plots (Figure 6) illustrate the genome-wide distribution of association signals across all six models. Models 1-5 show a substantial excess of highly significant variants, with numerous SNPs exceeding the genome-wide significance threshold ( $p < 5 \times 10^{-8}$ ) across multiple chromosomes. The association signals are widely dispersed rather than confined to specific genomic regions, indicating a lack of distinct, well-defined loci and instead suggesting a diffuse pattern of significance.

Across these models, the density of significant SNPs appears elevated across nearly all chromosomes, with no clear clustering into singular peaks typically expected in GWAS of complex traits. This pattern further supports the presence of inflated test statistics rather than true biological signals. The overall similarity of Manhattan plots across models 1-5 indicates that the inclusion of different covariates had limited impact on the overall distribution of association signals.

In contrast, model 6 showed no genome-wide significant associations, with all variants remaining below the significance threshold. This sharp deviation from the other models suggests that the inclusion of gestational age, birth weight and neonatal sex may have absorbed a substantial portion of the variation or alternatively reflects model instability.

Overall, the Manhattan plots reinforce the interpretation that the observed widespread significant associations in models 1-5 are unlikely to represent true genetic effects, but rather reflect systematic inflation, potentially driven by small sample size, sparse data structure and overfitting.

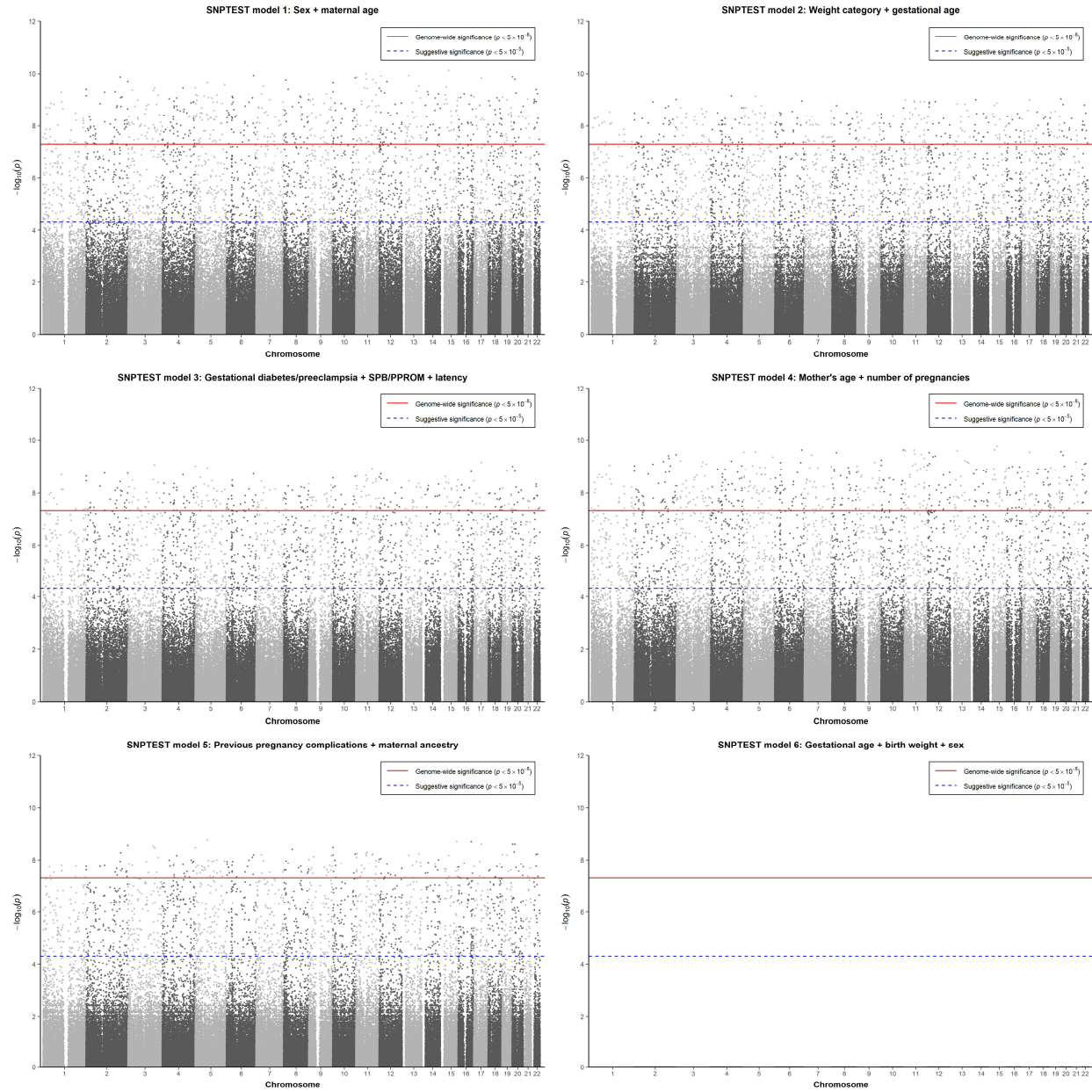


**Figure 6.** Manhattan plots from logistic regression association analysis results using SNPTEST using non-imputed data across six covariate models

QQ plots (Figure 7) demonstrate a substantial deviation of the observed test statistics from the expected null distribution across models 1-5. In these models the observed  $-\log_{10}(p)$  values deviated early from the null expectation and showed pronounced upward curvature, indicating an excess of small p-values and inflation of association statistics.

s systematically exceed the expected values, with early departure from the null line and strong upward curvature, indicating an excess of small p-values. This pattern is consistent with substantial inflation of association statistics rather than true polygenic signal.

The genomic inflation factor ( $\lambda$ ) further supports this interpretation, with elevated values detected in model 1 ( $\lambda = 2.25$ ), model 2 ( $\lambda = 1.99$ ), model 3 ( $\lambda = 2.02$ ), model 4 ( $\lambda = 2.18$ ) and model 5 ( $\lambda = 1.79$ ). These values substantially exceed the expected  $\lambda \approx 1$  under the null hypothesis, confirming poor calibration of the test statistics. The similar QQ plot patterns observed across these models suggest that the inflation is systematic and was not substantially reduced by different covariate adjustments.

In contrast, model 6 showed no deviation from the expected distribution, with  $\lambda = 0.00$  and all points lying along the null line. This indicates an absence of detectable association signal, consistent with the lack of genome-wide significant variants observed in the Manhattan plot.

Overall, the QQ plot analysis reinforces the conclusion that models 1-5 were strongly affected by test statistic inflation, most likely driven by small sample size, sparse genotype counts and model instability. Conversely, model 6 appeared substantially deflated, indicating reduced statistical power for detection of association signals

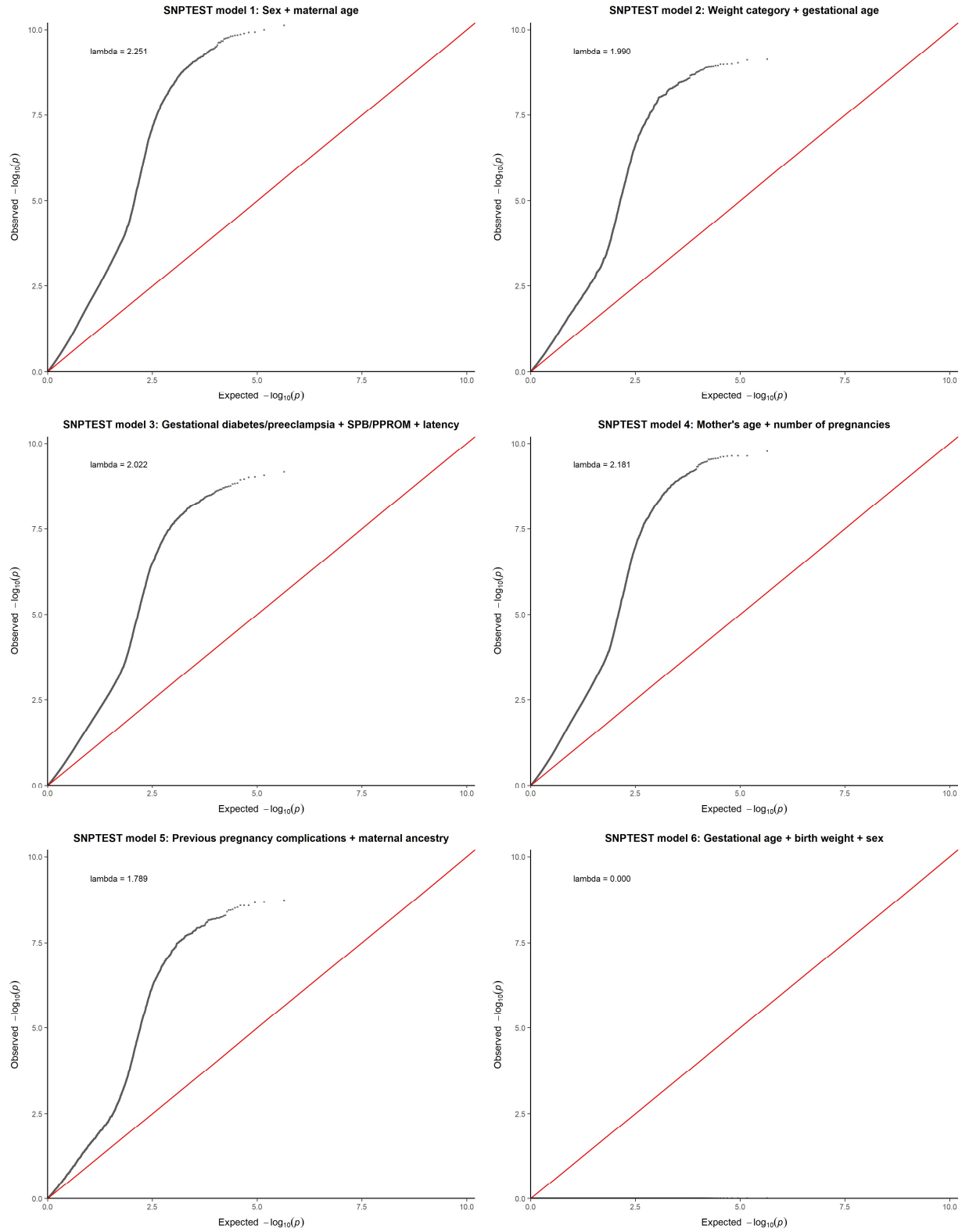


**Figure 7.** QQ plots from SNPTEST v2.5.4 analyses on non-imputed data across six covariate models

185 variants were repeatedly detected across models 1-5, including rs75516371, rs479547, rs201101435, rs113066651, rs6047156, rs11608006, rs71494354 and rs62079648. Variants were prioritized according to best p-value. Top 20 SNPs are presented in Table 8.

**Table 8.** Top 20 SNPs prioritized for further analysis based on statistical significance and consistency across logistic regression association analysis with SNPTEST using non-imputed data

| <b>SNP<br/>(chr:position,<br/>GRCh38)</b> | <b>Models<br/>identified</b> | <b>Best p-value</b>    | <b>Priority level</b> | <b>Notes</b>                                      |
|-------------------------------------------|------------------------------|------------------------|-----------------------|---------------------------------------------------|
| rs75516371<br>(chr15:47290631)            | 1, 2, 3, 4, 5                | $7.49 \times 10^{-11}$ | High                  | Strongest signal.<br>Reproduced in five<br>models |
| rs479547<br>(chr11:57663572)              | 1, 2, 3, 4, 5                | $9.98 \times 10^{-11}$ | High                  | Reproduced in five<br>models                      |
| rs201101435<br>(chr13:38364572)           | 1, 2, 3, 4, 5                | $1.17 \times 10^{-10}$ | High                  | Reproduced in five<br>models                      |
| rs113066651<br>(chr6:156478010)           | 1, 2, 3, 4, 5                | $1.18 \times 10^{-10}$ | High                  | Reproduced in five<br>models                      |
| rs71224202<br>(chr12:9427023)             | 1, 2, 3, 4, 5                | $1.28 \times 10^{-10}$ | High                  | Reproduced in five<br>models                      |
| rs6047156<br>(chr20:2164342)              | 1, 2, 3, 4, 5                | $1.38 \times 10^{-10}$ | High                  | Reproduced in five<br>models                      |
| rs201966219<br>(chr2:196947436)           | 1, 2, 3, 4, 5                | $1.42 \times 10^{-10}$ | High                  | Reproduced in five<br>models                      |
| rs11608006<br>(chr11:7813823)             | 1, 2, 3, 4, 5                | $1.46 \times 10^{-10}$ | High                  | Reproduced in five<br>models                      |

|                                 |               |                        |      |                           |
|---------------------------------|---------------|------------------------|------|---------------------------|
| rs71494354<br>(chr10:130572534) | 1, 2, 3, 4, 5 | $1.52 \times 10^{-10}$ | High | Reproduced in five models |
| rs62079648<br>(chr17:43323879)  | 1, 2, 3, 4, 5 | $1.55 \times 10^{-10}$ | High | Reproduced in five models |
| rs55930305<br>(chr11:61340568)  | 1, 2, 3, 4, 5 | $1.67 \times 10^{-10}$ | High | Reproduced in five models |
| rs1233740<br>(chr20:15529488)   | 1, 2, 3, 4, 5 | $1.72 \times 10^{-10}$ | High | Reproduced in five models |
| rs4753943<br>(chr11:99137867)   | 1, 2, 3, 4, 5 | $1.77 \times 10^{-10}$ | High | Reproduced in five models |
| rs623937<br>(chr8:12670736)     | 1, 2, 3, 4, 5 | $1.84 \times 10^{-10}$ | High | Reproduced in five models |
| rs201381023<br>(chr2:241236984) | 1, 2, 3, 4, 5 | $2.10 \times 10^{-10}$ | High | Reproduced in five models |
| rs268050<br>(chr12:44953925)    | 1, 2, 3, 4, 5 | $2.10 \times 10^{-10}$ | High | Reproduced in five models |
| rs78315775<br>(chr15:32485928)  | 1, 2, 3, 4, 5 | $2.30 \times 10^{-10}$ | High | Reproduced in five models |
| rs546293974<br>(chr5:69556655)  | 1, 2, 3, 4, 5 | $2.31 \times 10^{-10}$ | High | Reproduced in five models |
| rs4881072<br>(chr10:3062296)    | 1, 2, 3, 4, 5 | $2.31 \times 10^{-10}$ | High | Reproduced in five models |
| rs2266788<br>(chr11:116789970)  | 1, 2, 3, 4, 5 | $2.35 \times 10^{-10}$ | High | Reproduced in five models |

Functional annotation of highly prioritized SNPs was performed using the GeneBe (Variant annotation and interpretation tool, 2024) , dbSNP (National Library of Medicine; National Institutes of Health; U.S. Department of Health and Human Services, 2026) and GeneCards (Stelzer *et al.*, 2016) (Table 9).

**Table 9.** Functional annotation of prioritized SNPs identified in logistic regression association analysis with SNPTEST using non-imputed data

| SNP<br>(chr:position,<br>GRCh38) | Gene          | Location   | Functional consequence                                                                                                                                                                                                                                                                                                                                                   |
|----------------------------------|---------------|------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| rs75516371<br>(chr15:47290631)   | <i>SEMA6D</i> | intronic   | The variant is located within an intronic region of <i>SEMA6D</i> gene.<br><br><i>SEMA6D</i> encodes a transmembrane semaphorin involved in axon guidance, cell signaling, and developmental processes. Although primarily studied in the context of neural development, semaphorins have also been implicated in processes such as cell migration and tissue remodeling |
| rs479547<br>(chr11:57663572)     | -             | intergenic | This variant is located in a non-coding region and does not directly alter the protein sequence.                                                                                                                                                                                                                                                                         |
| rs201101435<br>(chr13:38364572)  | -             | intergenic | This variant is located in a non-coding region and does not directly alter the protein sequence.                                                                                                                                                                                                                                                                         |
| rs113066651<br>(chr6:156478010)  | -             | intergenic | This variant is located in a non-coding region and does not directly alter the protein sequence.                                                                                                                                                                                                                                                                         |

|                                 |                        |            |                                                                                                                                                                                                                                                                                                                                                                                                                                       |
|---------------------------------|------------------------|------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| rs71224202<br>(chr12:9427023)   | <i>DDX12P</i>          | intronic   | <p>The variant is located within an intronic region of <i>DDX12P</i> pseudogene.</p> <p><i>DDX12P</i> pseudogene predicted to enable DNA helicase and involved in establishment of sister chromatid cohesion.</p>                                                                                                                                                                                                                     |
| rs6047156<br>(chr20:2164342)    | <i>STK35</i>           | intronic   | <p>The variant is located within an intronic region of <i>STK35</i> gene.</p> <p><i>STK35</i> (Serine/Threonine Kinase 35) is a protein coding gene. The protein encoded by this gene is a kinase that is predominantly found in the nucleus. However, it can interact with PDLIM1/CLP-36 in the cytoplasm and localize to actin stress fibers. The encoded protein may be a regulator of actin stress fibers in non-muscle cells</p> |
| rs201966219<br>(chr2:196947436) | -                      | intergenic | <p>This variant is located in a non-coding region and does not directly alter the protein sequence.</p>                                                                                                                                                                                                                                                                                                                               |
| rs11608006<br>(chr11:7813823)   | <i>ENSG00000271758</i> | intronic   | <p>The variant is located within an intronic region of <i>ENSG00000271758</i> gene.</p> <p><i>ENSG00000271758</i> is an RNA gene and is affiliated with the lncRNA class.</p>                                                                                                                                                                                                                                                         |
| rs71494354<br>(chr10:130572534) | -                      | intergenic | <p>This variant is located in a non-coding region and does not directly alter the protein sequence.</p>                                                                                                                                                                                                                                                                                                                               |

|                                |                |            |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
|--------------------------------|----------------|------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| rs62079648<br>(chr17:43323879) | -              | intergenic | This variant is located in a non-coding region and does not directly alter the protein sequence.                                                                                                                                                                                                                                                                                                                                                                               |
| rs55930305<br>(chr11:61340568) | <i>TKFC</i>    | intronic   | <p>The variant is located within an intronic region of <i>TKFC</i> gene.</p> <p><i>TKFC</i> (Triokinase And FMN Cyclase) is a protein coding gene. The product of this gene phosphorylates dihydroxyacetone and catalyzes the formation of riboflavin 4',5'-phosphate (aka cyclin FMN) from FAD. Diseases associated with TKFC include Triokinase And Fmn Cyclase Deficiency Syndrome and Sengers Syndrome</p>                                                                 |
| rs1233740<br>(chr20:15529488)  | <i>MACROD2</i> | intronic   | <p>The variant is located within an intronic region of <i>MACROD2</i> gene.</p> <p>The protein encoded by <i>MACROD2</i> (Mono-ADP Ribosylhydrolase 2) gene is a deacetylase involved in removing ADP-ribose from mono-ADP-ribosylated proteins. The encoded protein has been shown to translocate from the nucleus to the cytoplasm upon DNA damage. Diseases associated with <i>MACROD2</i> include Hypogonadotropic Hypogonadism 21 with or without anosmia and autism.</p> |
| rs4753943<br>(chr11:99137867)  | <i>CNTN5</i>   | intronic   | The variant is located within an intronic region of <i>CNTN5</i> gene.                                                                                                                                                                                                                                                                                                                                                                                                         |

|                                 |                               |            |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
|---------------------------------|-------------------------------|------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                 |                               |            | <p><i>CNTN5</i> (Contactin 5) is a protein coding gene. The protein encoded by this gene is a member of the immunoglobulin superfamily, and contactin family, which mediate cell surface interactions during nervous system development. This protein is a glycosylphosphatidylinositol (GPI)-anchored neuronal membrane protein that functions as a cell adhesion molecule. It may play a role in the formation of axon connections in the developing nervous system. Diseases associated with <i>CNTN5</i> include Coffin-Siris Syndrome 6 and Chromosome 3Pter-P25 Deletion Syndrome</p> |
| rs623937<br>(chr8:12670736)     | -                             | intergenic | <p>This variant is located in a non-coding region and does not directly alter the protein sequence.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| rs201381023<br>(chr2:241236984) | <i>HDLBP</i> ,<br><i>ANO7</i> | intronic   | <p>The variant is located within an intronic region of <i>HDLBP</i> and <i>ANO7</i> genes.</p> <p>The protein encoded by <i>HDLBP</i> (High Density Lipoprotein Binding Protein) gene binds high density lipoprotein (HDL) and may function to regulate excess cholesterol levels in cells. The encoded protein also binds RNA and can induce heterochromatin formation. Diseases associated with <i>HDLBP</i> include Chromosome 2Q37 Deletion Syndrome and Alacrima, Achalasia, And Impaired Intellectual Development Syndrome. Among its related</p>                                     |

|                                |                        |                     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|--------------------------------|------------------------|---------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                |                        |                     | <p>pathways are Plasma lipoprotein assembly, remodeling, and clearance and Transport of inorganic cations/anions and amino acids/oligopeptides.</p> <p><i>ANO7</i> (Anoctamin 7) is a prostate-specific gene, which encodes a cytoplasmic protein, as well as a polytopic membrane protein which may serve as a target in prostate cancer diagnosis and immunotherapy. Diseases associated with <i>ANO7</i> include Miyoshi Muscular Dystrophy 3 and Immunodeficiency 15B. Among its related pathways are Infectious disease and SARS-CoV-2 Infection. Gene Ontology (GO) annotations related to this gene include intracellularly calcium-gated chloride channel activity and phospholipid scramblase activity</p> |
| rs268050<br>(chr12:44953925)   | -                      | intergenic          | This variant is located in a non-coding region and does not directly alter the protein sequence.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
| rs78315775<br>(chr15:32485928) | <i>ENSG00000307111</i> | intronic            | <p>The variant is located within an intronic region of <i>ENSG00000307111</i> gene.</p> <p><i>ENSG00000271758</i> is an RNA gene and is affiliated with the lncRNA class.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| rs546293974<br>(chr5:69556655) | <i>OCLN</i>            | 3 prime UTR variant | The variant causes a 3 prime UTR change involving the alteration of a non-conserved nucleotide.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |

|                                |                |                     |                                                                                                                                                                                                                                                                                                                                                                                                      |
|--------------------------------|----------------|---------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                |                |                     | <i>OCN</i> (Occludin) gene encodes an integral membrane protein that is required for cytokine-induced regulation of the tight junction paracellular permeability barrier. Mutations in this gene are thought to be a cause of band-like calcification with simplified gyration and polymicrogyria (BLC-PMG), an autosomal recessive neurologic disorder that is also known as pseudo-TORCH syndrome. |
| rs4881072<br>(chr10:3062296)   | <i>PFKP-DT</i> | intronic            | The variant is located within an intronic region of <i>PFKP-DT</i> gene.<br><br><i>PFKP-DT</i> is an RNA gene and is affiliated with the lncRNA class.                                                                                                                                                                                                                                               |
| rs2266788<br>(chr11:116789970) | <i>APOA5</i>   | 3 prime UTR variant | The variant causes a 3 prime UTR change involving the alteration of a non-conserved nucleotide.<br><br><i>APOA5</i> (Apolipoprotein A5) is a protein coding gene. Apolipoprotein plays an important role in regulating the plasma triglyceride levels, a major risk factor for coronary artery disease.                                                                                              |

Functional annotation of top 20 prioritized SNPs identified in the SNPTEST v2.5.4 association analysis using non-imputed data revealed that majority of variants were located within non-coding genomic regions, including intronic and intergenic regions. Several intergenic variants (e.g., rs479547, rs201101435, rs113066651, rs201966219, rs71494354, rs62079648, rs623937 and rs268050) may potentially influence regulatory elements such as enhancers or silencers rather than directly altering protein-coding sequences.

Multiple variants were located within intronic regions of genes with diverse biological functions, including *SEMA6D*, *STK35*, *TKFC*, *MACROD2*, *CNTN5*, *HDLBP/ANO7* and several lncRNA or pseudogene loci. These genes are involved in processes such as cell signaling, neural development, cytoskeletal regulation, lipid metabolism, RNA processing and membrane transport. Two variants, rs546293974 in *OCN* and rs2266788 in *APOA5*, were identified as 3' UTR variants, suggesting possible effects on post-transcriptional regulation and gene expression.

Overall, the annotated variants were predominantly associated with regulatory or non-coding genomic functions, and no clear enrichment of genes directly linked to pregnancy-related traits or reproductive pathways was observed.

### Logistic regression association analysis results using SNPTEST on imputed data

SNPTEST v2.5.4 association analysis was subsequently performed using the imputed genotype dataset using the same six covariate models applied in previous analyses. The aim was to evaluate whether genotype imputation improved detection of association signals after adjustment for different maternal and neonatal covariates.

Across all six models a large number of variants exceeded the conventional genome-wide significance threshold ( $p\text{-value} < 5 \times 10^{-8}$ ). Compared with the non-imputed analysis, the number of significant variants was markedly increased, and the statistical significance of top signals was substantially stronger with p-values reaching as low as  $\sim 10^{-11}$ . Effect size estimates remained consistently large across all models, with many variants exhibiting  $|\beta| > 2$  and in some cases substantially larger. Such effect sizes are highly unusual in GWAS of complex traits and likely reflect effect size inflation, model instability and potential overfitting or imputation-related artefacts.

Model 1 (neonatal sex and maternal age) identified 2,743 genome-wide significant variants with the most significant variants including rs115819657 ( $p\text{-value } 1.45 \times 10^{-11}$ ), rs28798180 ( $p\text{-value } 1.72 \times 10^{-11}$ ) and rs191335330 ( $p\text{-value } 1.74 \times 10^{-11}$ ). These variants showed extremely large effect sizes ( $|\beta| > 20$  in several cases).

Model 2 (birth weight category and exact birth-weight percentile for GA and sex) identified 1,398 genome-wide significant variants. The strongest associations were observed for rs118030376 ( $p\text{-value } 1.45 \times 10^{-11}$ ).

value  $1.26 \times 10^{-10}$ ), rs67494416 (p-value  $2.33 \times 10^{-10}$ ) and rs71224202 (p-value  $2.97 \times 10^{-10}$ ) with consistently large effect estimates ( $\beta \approx 2-26$ ).

Model 3 (GD or preeclampsia, SPB and PPRM and latency period) identified 1,246 genome-wide significant variants, with top signals including rs10083699 (p-value  $2.01 \times 10^{-9}$ ), rs6434642 (p-value  $2.26 \times 10^{-9}$ ) and rs4273241 (p-value  $2.38 \times 10^{-9}$ ).

Model 4 (maternal age and number of pregnancies) identified 2,323 genome-wide significant variants, with strong clustering of associations observed particularly on chromosomes 13 and 11, where multiple neighboring SNPs shared similar p-values and effect sizes.

Model 5 (previous pregnancy complications and maternal ancestry) identified 596 genome-wide significant variants, with recurrent signals observed across chromosomes 9, 13 and 15, consistent with patterns observed in other models.

Model 6 (gestational age, birth weight and neonatal sex) did not identify any genome-wide significant variants, consistent with the non-imputed analysis where no variants reached genome-wide significance.

The Manhattan plots generated for the imputed dataset (Figure 8) show a markedly different pattern compared to the non-imputed analysis. Across models 1-5 association signals are still distributed genome-wide, however, the number of variants exceeding the genome-wide significance threshold (p-value  $< 5 \times 10^{-8}$ ) is substantially reduced. While some SNPs surpass the threshold, they are fewer in number and less densely distributed than in the non-imputed data, indicating a reduction in extreme association signals after imputation.

The overall distribution of signals across chromosomes remains relatively diffuse with no strong, well-defined peaks corresponding to distinct genomic loci. Similar to the non-imputed results, signals are not concentrated within specific regions but rather scattered across the genome. However, the visual intensity of associations is lower, suggesting that imputation has partially mitigated the effect size inflation observed previously.

The Manhattan plots for models 1-5 remain broadly similar to each other, indicating that different covariate adjustments continue to have limited influence on the overall structure of association

signals. This consistency suggests that the underlying signal pattern is not strongly driven by the included covariates.

In contrast, model 6 again shows a complete absence of genome-wide significant variants, with all SNPs remaining below the significance threshold. This consistent lack of significant findings supports the interpretation that inclusion of gestational age, birth weight and neonatal sex accounts for a substantial portion of variability or alternatively reflects reduced statistical power and model instability.

Overall, the Manhattan plots generated on imputed dataset suggest a partial attenuation of the inflated signals observed in the non-imputed data. Although some significant associations persist, the reduced number and magnitude of these signals indicate improved calibration, while the absence of clear locus-specific peaks continues to argue against robust, biologically meaningful genetic associations in this dataset.

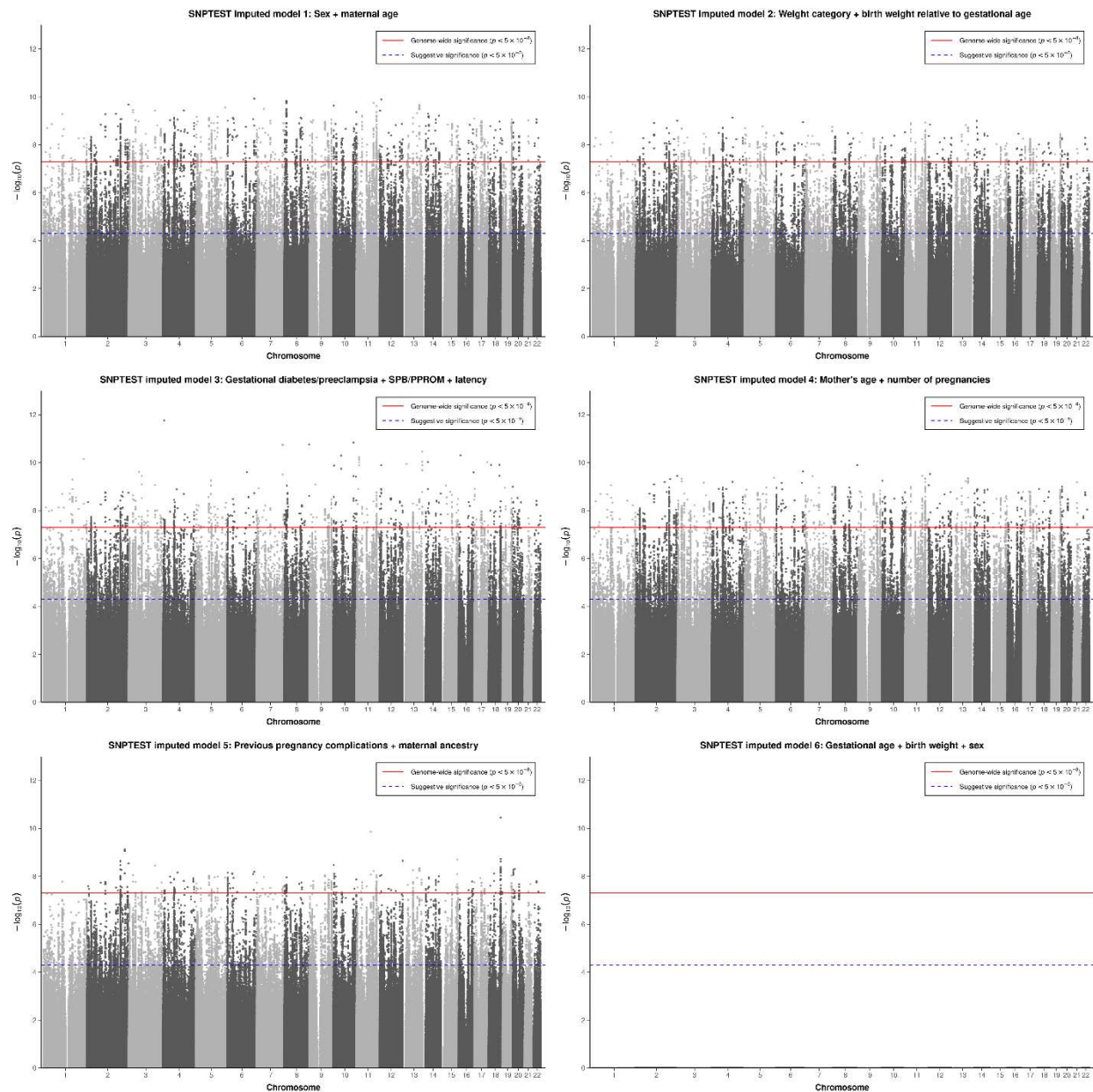


**Figure 8.** Manhattan plots from logistic regression association analysis results using SNPTTEST using imputed data across six covariate models

QQ plots for the imputed dataset (Figure 9) show a clear deviation of observed test statistics from the expected null distribution across models 1-5, although the magnitude of this deviation is reduced compared to the non-imputed analysis. In these models, the observed  $-\log_{10}(p)$  values exceed the expected values early, with noticeable upward curvature, indicating an excess of small

p-values. However, the departure from the null line is less extreme, suggesting partial attenuation of test statistic inflation after imputation.

The genomic inflation factor ( $\lambda$ ) remains elevated but is consistently lower than in the non-imputed data, with values of  $\lambda = 2.18$  (model 1),  $\lambda = 1.98$  (model 2),  $\lambda = 1.96$  (model 3),  $\lambda = 2.11$  (model 4) and  $\lambda = 1.76$  (model 5). Although these values still substantially exceed the expected  $\lambda \approx 1$  under the null hypothesis, they indicate a modest improvement in calibration. The similarity in QQ plot shapes across models 1-5 suggests that the inflation is systematic and not substantially influenced by different covariate adjustments.

In contrast, model 6 shows no deviation from the expected distribution, with  $\lambda = 0.00$  and all points lying along the null line. This indicates a complete absence of detectable association signal, consistent with the corresponding Manhattan plot and suggesting either over-adjustment or lack of statistical power in this model.

Overall, the QQ plots for the imputed data indicate that, while imputation reduces the extent of inflation observed in the non-imputed analysis, substantial deviation from the null distribution persists in models 1-5. This supports the interpretation that the observed excess of significant associations is still largely driven by systematic inflation rather than true underlying genetic effects.

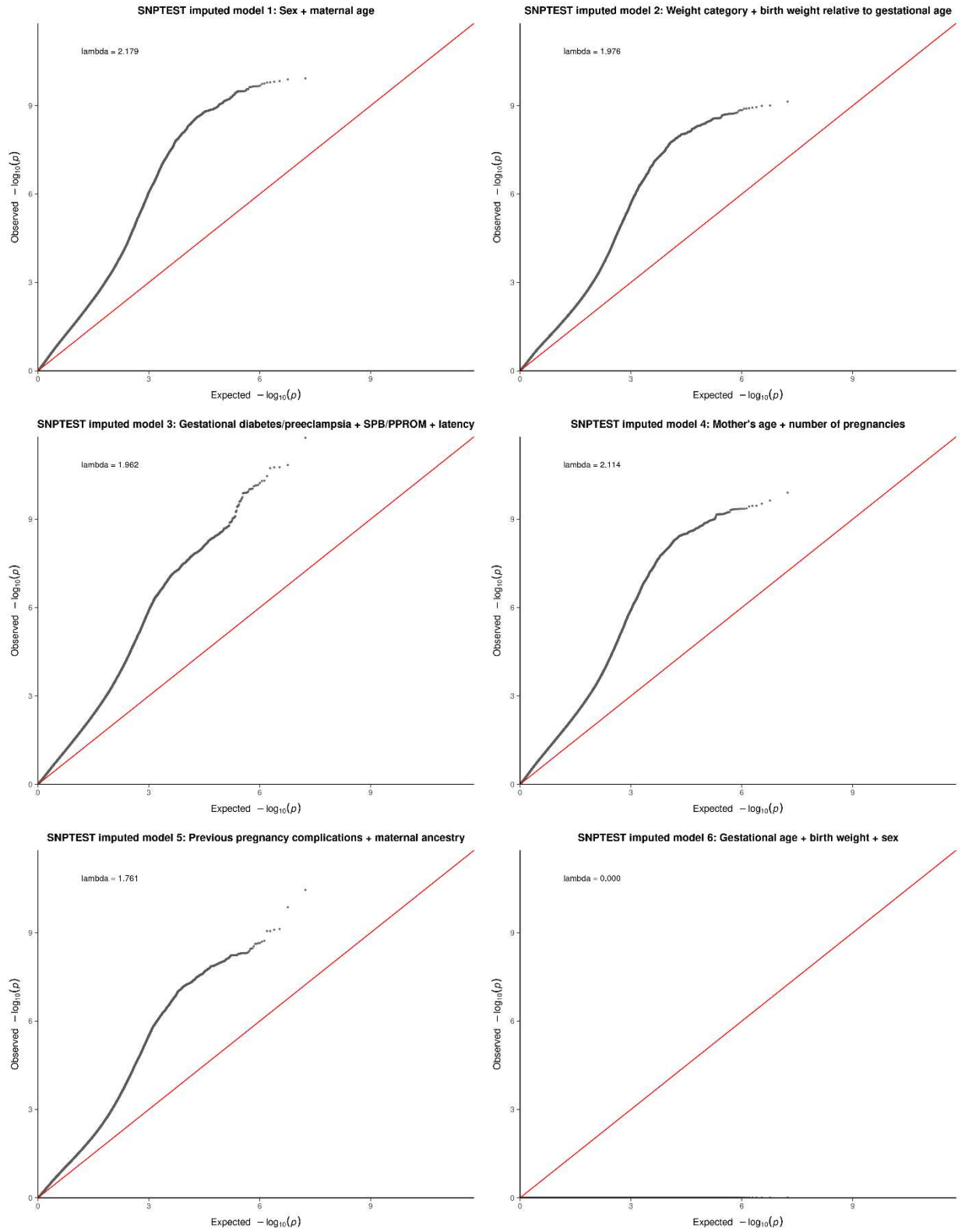


**Figure 9.** QQ plots from logistic regression association analysis results using SNPTEST using imputed data across six covariate models

409 variants were repeatedly detected across models 1-5. Several SNPs formed clustered signals within specific genomic regions, showing similar levels of statistical significance and consistent detection across models. This pattern is indicative of linkage disequilibrium, where nearby variants are inherited together and reflect a single underlying association signal rather than independent effects. Top 20 SNPs were prioritized for further analysis based on statistical significance and consistency across SNPTEST models using imputed data (Table 10).

**Table 10.** Top 20 SNPs prioritized for further analysis based on statistical significance and consistency across logistic regression association analysis with SNPTEST using imputed data

| <b>SNP (chr:position, GRCh38)</b> | <b>Models identified</b> | <b>Best P-value</b>    | <b>Priority level</b> | <b>Notes</b>                                   |
|-----------------------------------|--------------------------|------------------------|-----------------------|------------------------------------------------|
| rs67494416<br>(chr6:156478010)    | 1, 2, 3, 4, 5            | $1.18 \times 10^{-10}$ | High                  | Strongest signal.<br>Reproduced in five models |
| rs71224202<br>(chr12:9427023)     | 1, 2, 3, 4, 5            | $1.28 \times 10^{-10}$ | High                  | Reproduced in five models                      |
| rs622140<br>(chr8:12670279)       | 1, 2, 3, 4, 5            | $1.47 \times 10^{-10}$ | High                  | Reproduced in five models                      |
| rs631894<br>(chr8:12670599)       | 1, 2, 3, 4, 5            | $1.55 \times 10^{-10}$ | High                  | Reproduced in five models                      |
| rs623934<br>(chr8:12670728)       | 1, 2, 3, 4, 5            | $1.62 \times 10^{-10}$ | High                  | Reproduced in five models                      |
| rs619909<br>(chr8:12670189)       | 1, 2, 3, 4, 5            | $1.65 \times 10^{-10}$ | High                  | Reproduced in five models                      |
| rs4753943<br>(chr11:99137867)     | 1, 2, 3, 4, 5            | $1.77 \times 10^{-10}$ | High                  | Reproduced in five models                      |

|                                 |               |                        |      |                           |
|---------------------------------|---------------|------------------------|------|---------------------------|
| rs623937<br>(chr8:12670736)     | 1, 2, 3, 4, 5 | $1.84 \times 10^{-10}$ | High | Reproduced in five models |
| rs201381023<br>(chr2:241236984) | 1, 2, 3, 4, 5 | $2.10 \times 10^{-10}$ | High | Reproduced in five models |
| rs7984995<br>(chr13:96461783)   | 1, 2, 3, 4, 5 | $2.17 \times 10^{-10}$ | High | Reproduced in five models |
| rs4771951<br>(chr13:96461467)   | 1, 2, 3, 4, 5 | $2.18 \times 10^{-10}$ | High | Reproduced in five models |
| rs1327915<br>(chr13:96462337)   | 1, 2, 3, 4, 5 | $2.18 \times 10^{-10}$ | High | Reproduced in five models |
| rs4142574<br>(chr13:96463189)   | 1, 2, 3, 4, 5 | $2.21 \times 10^{-10}$ | High | Reproduced in five models |
| rs1887600<br>(chr13:96468788)   | 1, 2, 3, 4, 5 | $2.23 \times 10^{-10}$ | High | Reproduced in five models |
| rs4881072<br>(chr10:3062296)    | 1, 2, 3, 4, 5 | $2.31 \times 10^{-10}$ | High | Reproduced in five models |
| rs2266788<br>(chr11:116789970)  | 1, 2, 3, 4, 5 | $2.35 \times 10^{-10}$ | High | Reproduced in five models |
| rs4771253<br>(chr13:96470318)   | 1, 2, 3, 4, 5 | $2.37 \times 10^{-10}$ | High | Reproduced in five models |
| rs7319073<br>(chr13:96454485)   | 1, 2, 3, 4, 5 | $2.73 \times 10^{-10}$ | High | Reproduced in five models |
| rs1328060<br>(chr13:96454644)   | 1, 2, 3, 4, 5 | $2.73 \times 10^{-10}$ | High | Reproduced in five models |

|                              |               |                        |      |                           |
|------------------------------|---------------|------------------------|------|---------------------------|
| rs395407<br>(chr5:170262600) | 1, 2, 3, 4, 5 | $2.73 \times 10^{-10}$ | High | Reproduced in five models |
|------------------------------|---------------|------------------------|------|---------------------------|

Functional annotation of highly prioritized SNPs was performed using the GeneBe, dbSNP and GeneCards (Table 11).

**Table 11.** Functional annotation of prioritized SNPs identified in logistic regression association analysis with SNPTTEST using imputed data

| SNP<br>(chr:position,<br>GRCh38) | Gene                                   | Location   | Functional consequence                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
|----------------------------------|----------------------------------------|------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| rs67494416<br>(chr6:156478010)   | -                                      | intergenic | This variant is located in a non-coding region and does not directly alter the protein sequence.                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| rs71224202<br>(chr12:9427023)    | <i>DDX12P</i> ,<br><i>LOC124902873</i> | intronic   | <p>The variant is located within an intronic region of <i>DDX12P</i> and <i>LOC124902873</i> genes.</p> <p><i>DDX12P</i> (DEAD/H-Box Helicase 12, Pseudogene) is a pseudogene predicted to enable DNA helicase activity and to be involved in establishment of sister chromatid cohesion. Diseases associated with <i>DDX12P</i> include Warsaw Breakage Syndrome. Gene Ontology (GO) annotations related to this gene include RNA binding and DNA helicase activity.</p> <p><i>LOC124902873</i> is an RNA gene and is affiliated with the ncRNA class.</p> |

|                               |              |            |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
|-------------------------------|--------------|------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| rs622140<br>(chr8:12670279)   | -            | intergenic | This variant is located in a non-coding region and does not directly alter the protein sequence.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| rs631894<br>(chr8:12670599)   | -            | intergenic | This variant is located in a non-coding region and does not directly alter the protein sequence.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| rs623934<br>(chr8:12670728)   | -            | intergenic | This variant is located in a non-coding region and does not directly alter the protein sequence.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| rs619909<br>(chr8:12670189)   | -            | intergenic | This variant is located in a non-coding region and does not directly alter the protein sequence.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| rs4753943<br>(chr11:99137867) | <i>CNTN5</i> | intronic   | <p>The variant is located within an intronic region of <i>CNTN5</i> gene.</p> <p><i>CNTN5</i> gene encodes Contactin 5, a member of the immunoglobulin superfamily and Contactin family, which mediates cell surface interactions during nervous system development. This protein is a glycosylphosphatidylinositol (GPI)-anchored neuronal membrane protein that functions as a cell adhesion molecule. It may play a role in the formation of axon connections in the developing nervous system.</p> <p>Diseases associated with <i>CNTN5</i> include Coffin-Siris Syndrome 6 and Chromosome 3Pter-P25 Deletion Syndrome. Among its</p> |

|                                 |                               |            |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
|---------------------------------|-------------------------------|------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                 |                               |            | related pathways are metabolism of proteins and post-translational modification: synthesis of GPI-anchored proteins.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| rs623937<br>(chr8:12670736)     | -                             | intergenic | This variant is located in a non-coding region and does not directly alter the protein sequence.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
| rs201381023<br>(chr2:241236984) | <i>ANO7</i> ,<br><i>HDLBP</i> | intronic   | <p>The variant is located within an intronic region of <i>ANO7</i> and <i>HDLBP</i> genes.</p> <p><i>ANO7</i> is a prostate-specific gene which encodes a cytoplasmic protein, as well as a polytopic membrane protein which may serve as a target in prostate cancer diagnosis and immunotherapy. Diseases associated with <i>ANO7</i> include Miyoshi Muscular Dystrophy 3 and Immunodeficiency 15B. Among its related pathways are Infectious disease and SARS-CoV-2 Infection.</p> <p>The protein encoded by <i>HDLBP</i> gene binds high density lipoprotein (HDL) and may function to regulate excess cholesterol levels in cells. The encoded protein also binds RNA and can induce heterochromatin formation.</p> <p>Diseases associated with <i>HDLBP</i> include Chromosome 2Q37 Deletion Syndrome and Alacrima, Achalasia, And Impaired Intellectual Development Syndrome. Among its related pathways are Plasma lipoprotein assembly, remodeling, and clearance and Transport of inorganic</p> |

|                               |               |          |                                                                                                                                                                                                                                                                                                                                                                                                                                                |
|-------------------------------|---------------|----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                               |               |          | cations/anions and amino acids/oligopeptides.                                                                                                                                                                                                                                                                                                                                                                                                  |
| rs7984995<br>(chr13:96461783) | <i>HS6ST3</i> | intronic | <p>The variant is located within an intronic region of <i>HS6ST3</i> gene.</p> <p>Heparan sulfate (HS) sulfotransferases, such as <i>HS6ST3</i>, modify HS to generate structures required for interactions between HS and a variety of proteins. These interactions are implicated in proliferation and differentiation, adhesion, migration, inflammation, blood coagulation, and other diverse processes (Habuchi <i>et al.</i>, 2000).</p> |
| rs4771951<br>(chr13:96461467) | <i>HS6ST3</i> | intronic | <p>The variant is located within an intronic region of <i>HS6ST3</i> gene.</p> <p>Heparan sulfate (HS) sulfotransferases, such as <i>HS6ST3</i>, modify HS to generate structures required for interactions between HS and a variety of proteins. These interactions are implicated in proliferation and differentiation, adhesion, migration, inflammation, blood coagulation, and other diverse processes.</p>                               |
| rs1327915<br>(chr13:96462337) | <i>HS6ST3</i> | intronic | <p>The variant is located within an intronic region of <i>HS6ST3</i> gene.</p> <p>Heparan sulfate (HS) sulfotransferases, such as <i>HS6ST3</i>, modify HS to generate structures required for interactions between HS and a variety of proteins. These</p>                                                                                                                                                                                    |

|                               |                |          |                                                                                                                                                                                                                                                                                                                                                                                                                  |
|-------------------------------|----------------|----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                               |                |          | interactions are implicated in proliferation and differentiation, adhesion, migration, inflammation, blood coagulation, and other diverse processes.                                                                                                                                                                                                                                                             |
| rs4142574<br>(chr13:96463189) | <i>HS6ST3</i>  | intronic | <p>The variant is located within an intronic region of <i>HS6ST3</i> gene.</p> <p>Heparan sulfate (HS) sulfotransferases, such as <i>HS6ST3</i>, modify HS to generate structures required for interactions between HS and a variety of proteins. These interactions are implicated in proliferation and differentiation, adhesion, migration, inflammation, blood coagulation, and other diverse processes.</p> |
| rs1887600<br>(chr13:96468788) | <i>HS6ST3</i>  | intronic | <p>The variant is located within an intronic region of <i>HS6ST3</i> gene.</p> <p>Heparan sulfate (HS) sulfotransferases, such as <i>HS6ST3</i>, modify HS to generate structures required for interactions between HS and a variety of proteins. These interactions are implicated in proliferation and differentiation, adhesion, migration, inflammation, blood coagulation, and other diverse processes.</p> |
| rs4881072<br>(chr10:3062296)  | <i>PFKP-DT</i> | intronic | <p>The variant is located within an intronic region of <i>PFKP-DT</i> gene.</p> <p><i>PFKP-DT</i> is an RNA gene and is affiliated with the lncRNA class.</p>                                                                                                                                                                                                                                                    |

|                                |               |                        |                                                                                                                                                                                                                                                                                                                                                                                                                  |
|--------------------------------|---------------|------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| rs2266788<br>(chr11:116789970) | <i>APOA5</i>  | 3 prime UTR<br>variant | <p>The variant causes a 3 prime UTR change involving the alteration of a non-conserved nucleotide.</p> <p><i>APOA5</i> (Apolipoprotein A5) is a protein coding gene. Apolipoprotein plays an important role in regulating the plasma triglyceride levels, a major risk factor for coronary artery disease.</p>                                                                                                   |
| rs4771253<br>(chr13:96470318)  | <i>HS6ST3</i> | intronic               | <p>The variant is located within an intronic region of <i>HS6ST3</i> gene.</p> <p>Heparan sulfate (HS) sulfotransferases, such as <i>HS6ST3</i>, modify HS to generate structures required for interactions between HS and a variety of proteins. These interactions are implicated in proliferation and differentiation, adhesion, migration, inflammation, blood coagulation, and other diverse processes.</p> |
| rs7319073<br>(chr13:96454485)  | <i>HS6ST3</i> | intronic               | <p>The variant is located within an intronic region of <i>HS6ST3</i> gene.</p> <p>Heparan sulfate (HS) sulfotransferases, such as <i>HS6ST3</i>, modify HS to generate structures required for interactions between HS and a variety of proteins. These interactions are implicated in proliferation and differentiation, adhesion, migration, inflammation, blood coagulation, and other diverse processes.</p> |

|                               |               |          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
|-------------------------------|---------------|----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| rs1328060<br>(chr13:96454644) | <i>HS6ST3</i> | intronic | <p>The variant is located within an intronic region of <i>HS6ST3</i> gene.</p> <p>Heparan sulfate (HS) sulfotransferases, such as <i>HS6ST3</i>, modify HS to generate structures required for interactions between HS and a variety of proteins. These interactions are implicated in proliferation and differentiation, adhesion, migration, inflammation, blood coagulation, and other diverse processes.</p>                                                                         |
| rs395407<br>(chr5:170262600)  | <i>LCP2</i>   | intronic | <p>The variant is located within an intronic region of <i>LCP2</i> gene.</p> <p><i>LCP2</i> gene encodes an adapter protein that acts as a substrate of the T cell antigen receptor (TCR)-activated protein tyrosine kinase pathway. The encoded protein is associated with growth factor receptor bound protein 2 and is thought to play a role TCR-mediated intracellular signal transduction. A similar protein in mice plays a role in normal T-cell development and activation.</p> |

Functional annotation of prioritized SNPs identified from the SNPTEST v2.5.4 analysis on imputed data (Table 11) shows that the majority of variants are located in non-coding regions, including intergenic regions and introns. No variants were identified within protein-coding exons, indicating that the detected associations are unlikely to directly alter protein structure and are more consistent with regulatory effects on gene expression.

A substantial proportion of SNPs are intergenic, suggesting potential involvement in distal regulatory elements such as enhancers or silencers. Among intronic variants, several variants map

to genes with diverse biological roles. Notably, variants were identified in *CNTN5*, involved in cell adhesion and neural development, and in a region overlapping *ANO7* and *HDLBP*, which are associated with membrane function and lipid metabolism, respectively.

A prominent feature of the results is a cluster of multiple SNPs within the *HS6ST3* locus on chromosome 13, all showing similar annotations. This pattern suggests a single underlying signal likely driven by linkage disequilibrium, rather than independent associations. Additional variants were located in pseudogenes and non-coding RNA loci, including *DDX12P* and *PFKP-DT*, that may have regulatory roles but remain poorly characterized.

Overall, the predominance of intronic and intergenic variants, together with clustering patterns and limited direct functional relevance of many annotated genes, indicates that these findings should be interpreted cautiously. In the context of observed statistical inflation, the results are consistent with indirect regulatory signals, linkage disequilibrium effects, or potential false-positive associations, rather than clear causal genetic mechanisms.

## DISCUSSION

The aim of this study was to investigate genomic variation in preterm and term-born neonates and assess potential associations between genetic variants, PTB and neonatal complications, as well as to explore their possible functional relevance.

After genotype quality control, the final dataset included 54 individuals and 220,064 SNPs.

The candidate variant analysis was substantially limited because only one of the 43 previously reported candidate SNPs, rs17713682, remained after quality control filtering. This considerably restricted the ability to evaluate previously published associations. Although rs17713682 showed a protective direction of effect, with lower odds of the phenotype among carriers of the T allele in both logistic regression and Fisher's exact test, the association did not reach statistical significance. The observed effect direction may still be biologically relevant. However, the wide confidence interval and limited statistical power prevent firm conclusions. Therefore, the candidate variant analysis did not provide convincing supporting an association between rs17713682 and studied phenotype in this cohort.

Association analyses were performed using logistic regression under six covariate models with PLINK v1.9 and SNPTEST v2.5.4, applied consistently to both non-imputed and imputed datasets. Overall, the findings should be interpreted as exploratory due to the limited sample size, sparse genotype distribution and evidence of test statistic inflation observed in several models.

No SNPs reached the conventional genome-wide significance threshold in the logistic regression association analysis performed with PLINK 1.9. This finding is consistent with the limited statistical power expected in a dataset containing only 54 individuals. Several SNPs showed suggestive associations, particularly in models 1, 3 and 4, and some variants were repeatedly detected across these models. The most recurring signals included rs3129111, rs11683045, rs10065905, rs202197010, rs9666203 and rs10250442. Their recurrence across different covariate adjustments suggests that these variants may represent potentially interesting loci for further investigation. However, the observed effect estimates were frequently very large, confidence intervals were broad, and several odds ratios were extreme. Such patterns indicate unstable model estimation, likely caused by sparse genotype counts, low minor allele frequencies and model overfitting.

The QQ plots from the logistic regression performed with PLINK 1.9 further support a cautious interpretation of the results. Models 1, 3 and 4 demonstrated inflation of association statistics, whereas models 2, 5 and 6 showed substantial deflation. This variability indicates that the association results were highly sensitive to covariate selection. In a well-calibrated GWAS, genomic inflation factors are expected to be close to 1. The observed deviations in this study from this expectation suggest that the detected signals are unlikely to represent robust genome-wide associations. Instead, they may reflect a combination of small sample size, sparse data structure, residual confounding and instability introduced by including multiple covariates in a limited dataset.

The SNPTEST v2.5.4 analysis of non-imputed data produced a markedly different pattern. Models 1-5 identified hundreds of genome-wide significant variants, whereas model 6 identified none. At first glance, this could suggest stronger detection power compared with logistic regression. However, the widespread distribution of significant SNPs across many chromosomes, the absence of clearly defined locus-specific peaks, and the high genomic inflation factors indicate substantial test statistic inflation. The large effect size estimates further support this interpretation. In GWAS of complex traits, especially pregnancy-related outcomes, very large genetic effects are uncommon. These findings strongly suggest that the majority of significant associations identified by SNPTEST were likely driven by statistical artefacts rather than genuine biological effects. SNPTEST may be particularly sensitive to sparse genotype distributions and unstable likelihood estimation in very small cohorts, which can contribute to inflated association statistics.

The imputed SNPTEST v2.5.4 analysis further increased the number of genome-wide significant variants, with thousands of significant SNPs identified in some models. Imputation can improve genomic coverage and detect associations with variants that are not directly genotyped. Extremely large effect estimates and dense clusters of significant variants were observed, particularly on chromosomes 8 and 13. These clustered signals may reflect linkage disequilibrium, but the overall pattern remains concerning inflation and potential imputation-related artefacts. The reliability of imputation may also have been affected by the small sample size and possible differences between the study population and the reference panel used for imputation. As in the non-imputed SNPTEST analysis, model 6 did not identify genome-wide significant variants, suggesting that the inclusion

of gestational age, birth weight and neonatal sex either absorbed phenotype-related variation or resulted in an overly conservative and underpowered model.

Across all analyses, one important observation was that the results depended strongly on both the statistical method and the covariate model used. Logistic regression with PLINK produced no genome-wide significant associations, while logistic regression with SNPTEST identified many significant variants in models 1-5. This inconsistency suggests that the results are not methodologically robust. In a reliable association study, the strongest signals would be expected to show consistency across statistical approaches, stable effect estimates, and acceptable calibration in QQ plots. In this study, although some variants were recurrent across models, the broader evidence indicates that most significant SNPTEST findings should be interpreted with caution.

Functional annotation revealed that most prioritized variants were located in non-coding regions, including intronic, intergenic, pseudogene and long non-coding RNA regions. This is not unexpected, as many GWAS signals are located outside protein-coding regions and may influence gene regulation rather than protein structure. Some annotated genes, such as *SEMA6D*, *STK35*, *GAB2*, *MACROD2*, *CNTN5*, *DDB1* and *HDLBP* have roles in cellular signaling, cytoskeletal regulation, DNA repair, metabolism or development. These processes may be broadly relevant to placental development, inflammation and fetal growth. However, no direct mechanistic relationship between these variants and the studied phenotype can be established. Functional interpretation of these findings should be considered hypothesis-generating rather than confirmatory.

The primary limitation of this study is the small sample size. With only 29 cases and 25 controls, the study had limited statistical power to detect modest genetic effects and was vulnerable to unstable effect estimates and false-positive findings. Given the large number of variants tested relative to the cohort size, the multiple testing burden remained substantial despite the use of genome-wide significance thresholds. Recruitment of PTB cohorts is also inherently challenging because PTB affects a highly sensitive and clinically vulnerable population. Ethical considerations, clinical burden, and parental consent constraints often limit recruitment and reduce participation rates, making it difficult to assemble large, well-characterized cohorts comparable to those used in large-scale GWAS consortia.

Beyond methodological limitations, PTB itself is a highly complex and multifactorial condition. Its etiology involves a combination of genetic, environmental, maternal and neonatal factors, including inflammation, infection, stress and obstetric complications. Genetic variation represents only one component of this multifaceted risk architecture. Therefore, the absence of strong genetic associations in this study does not exclude a biological basis but may reflect the relatively modest contribution of common genetic variants compared to environmental and clinical influences.

Furthermore, it is plausible that the underlying biological mechanisms of PTB are driven more by regulatory processes than by direct genetic variation in coding regions. Many prioritized variants in this study were located in non-coding regions, supporting the possibility that regulatory pathways may contribute to disease risk. Epigenetic modifications, such as DNA methylation, histone modification and gene expression regulation, may be particularly relevant in the context of pregnancy and fetal development. These mechanisms are dynamic and responsive to environmental exposures, and therefore epigenetic analyses may provide additional insights into PTB susceptibility that are not captured by SNP-based approaches.

Finally, the choice of genotyping strategy may also have influenced the results. Although SNP arrays combined with imputation provide cost-effective genome-wide coverage, they rely heavily on reference panels and may fail to capture rare or population-specific variants. WGS offers a more complete assessment of genetic variation, including rare variants, structural variants, and non-coding regions, without reliance on imputation. Therefore, it is possible that a study design based entirely on WGS for both cases and controls would provide more accurate and consistent variant detection and may improve the ability to identify true associations.

In summary, the limitations of this study arise from a combination of practical constraints, methodological factors, and the inherent complexity of PTB. These considerations underscore the need for larger, collaborative studies, integrative multi-omics approaches, and careful methodological design to better understand the genetic and biological basis of this condition.

Despite these limitations, this study provides useful exploratory insights. It demonstrates the importance of comparing multiple statistical approaches, evaluating QQ plots and genomic inflation factors, and interpreting genome-wide significant results in the context of sample size and model calibration. The prioritization of variants based on both statistical significance and recurrence across models provides a useful framework for selecting candidates for future

validation. However, none of the identified variants should currently be considered as confirmed associations to PTB.

Future studies should include a substantially larger and independently collected cohort, preferably with harmonized genotyping or sequencing platforms for cases and controls. Replication in an external dataset is essential before any biological conclusions are made. Additional analyses, such as principal component adjustment for population structure, stricter imputation quality filtering, sensitivity analyses with reduced covariate models, and permutation testing, would help evaluate the robustness of association signals. Functional follow-up of recurrent non-coding variants may also be informative, particularly if replicated associations are found near genes involved in placental development, immune regulation, or fetal growth.

In conclusion, this study did not identify robust genome-wide significant associations consistently supported across analytical methods and diagnostic evaluations. The candidate variant rs17713682 showed a protective direction of effect but did not achieve statistical significance. Logistic regression performed with PLINK identified several suggestive and recurrent SNPs, while SNPTTEST v2.5.4 analyses produced many genome-wide significant signals that were likely affected by test statistic inflation. Overall, the findings therefore should be regarded as exploratory and interpreted cautiously. Larger, well-powered studies with independent replication are required to clarify whether any of the prioritized variants contribute to the genetic architecture of the studied phenotype.

## CONCLUSIONS

1. No statistically significant genome-wide associations with preterm birth (PTB) were identified in the case-control GWAS analysis, indicating that the available dataset was insufficient to detect robust genetic effects associated with the studied phenotype.
2. Annotation and functional evaluation of genomic variation identified several suggestive and recurrent variants, including rs17713682, which demonstrated a non-significant protective trend. However, the observed associations were not stable and are likely influenced by limited sample size, model instability and inflation of test statistics; therefore, the findings should be considered exploratory and insufficient to confirm biologically relevant mechanisms underlying PTB.
3. Analyses assessing associations between genetic variants and neonatal complications did not provide reliable evidence of significant genetic effects. Although SNPTEST v2.5.4 identified numerous statistically significant associations, substantial inflation of test statistics suggests that these findings are unlikely to represent true biological signals. Larger, well-designed studies are required to reliably determine the genetic contribution to neonatal complications and PTB. Overall, the results should be considered exploratory and larger, well-designed studies are required to reliably determine the genetic contribution to the studied phenotype.

## **RECOMMENDATION**

Future studies should prioritize increasing sample size through multi-center collaboration to improve statistical power and reduce the risk of false-positive findings. Ensuring methodological consistency, particularly by using the same genotyping or sequencing platform for both cases and controls, would help minimize technical bias.

Given the predominance of non-coding variants and the complex nature of the studied phenotype, integrative approaches should be considered in future research. These may include whole-genome sequencing for comprehensive variant detection and incorporation of epigenetic and transcriptomic data to better capture regulatory mechanisms. Replication in independent cohorts and functional validation of prioritized variants are essential steps before drawing biological conclusions or considering clinical relevance.

## **ACKNOWLEDGEMENTS**

I would like to express my sincere gratitude to my supervisor, Alina Urnikytė, for her guidance, continuous support and insightful discussions throughout this project. Her expertise was essential in shaping the research strategy, interpreting complex results and maintaining a clear scientific direction.

I am also very grateful to Svetlana Dauengauer-Kirlienė for her significant contribution in preparing and providing the covariate data, as well as for developing biologically meaningful models that greatly strengthened the analytical framework of this study.

I would also like to thank all individuals and institutions involved in data generation and support, whose contributions made this research possible.

## REFERENCES

1. 1000 Genomes Project High-coverage phased genotype data (2020) “1000 Genomes Project High-coverage phased genotype data.” 1000 Genomes Project. Available at: [https://42basepairs.com/browse/s3/1000genomes/1000G\\_2504\\_high\\_coverage/working/20201028\\_3202\\_phased](https://42basepairs.com/browse/s3/1000genomes/1000G_2504_high_coverage/working/20201028_3202_phased).
2. Andzane, D. *et al.* (2025) “Preterm birth in Latvia: Two-decade national trends, structure, and risk factors,” *Acta Obstetrica et Gynecologica Scandinavica*, 104(10), pp. 1929–1938. Available at: <https://doi.org/10.1111/aogs.70015>.
3. Anum, E.A. *et al.* (2009) “Genetic Contributions to Disparities in Preterm Birth,” *Pediatric Research*, 65(1), pp. 1–9. Available at: <https://doi.org/10.1203/PDR.0b013e31818912e7>.
4. Araújo, B.F.D. *et al.* (2012) “Analysis of neonatal morbidity and mortality in late-preterm newborn infants,” *Jornal de Pediatria*, pp. 259–66. Available at: <https://doi.org/10.2223/JPED.2196>.
5. Bacelis, J. *et al.* (2016) “Literature-Informed Analysis of a Genome-Wide Association Study of Gestational Age in Norwegian Women and Children Suggests Involvement of Inflammatory Pathways,” *PLOS ONE*. Edited by Z. Xu, 11(8), p. e0160335. Available at: <https://doi.org/10.1371/journal.pone.0160335>.
6. Bhattacharjee, E. *et al.* (2023) “Genetic variants associated with spontaneous preterm birth in women from India: a prospective cohort study,” *The Lancet Regional Health - Southeast Asia*, 14, p. 100190. Available at: <https://doi.org/10.1016/j.lansea.2023.100190>.
7. Boyd, H.A. *et al.* (2009) “Maternal Contributions to Preterm Delivery,” *American Journal of Epidemiology*, 170(11), pp. 1358–1364. Available at: <https://doi.org/10.1093/aje/kwp324>.
8. Browning, B.L. *et al.* (2021) “Fast two-stage phasing of large-scale sequence data,” *The American Journal of Human Genetics*, 108(10), pp. 1880–1890. Available at: <https://doi.org/10.1016/j.ajhg.2021.08.005>.
9. Brubaker, D. *et al.* (2016) “Finding lost genes in GWAS via integrative—omics analysis reveals novel sub-networks associated with preterm birth,” *Human Molecular Genetics*, p. ddw325. Available at: <https://doi.org/10.1093/hmg/ddw325>.
10. Chang, C.C. *et al.* (2015) “Second-generation PLINK: rising to the challenge of larger and richer datasets,” *Gigascience*, 4(1), pp. s13742-015-0047–8. Available at: <https://doi.org/10.1186/s13742-015-0047-8>.
11. Cook-Deegan, R. and McGuire, A.L. (2017) “Moving beyond Bermuda: sharing data to build a medical information commons,” *Genome Research*, 27(6), pp. 897–901. Available at: <https://doi.org/10.1101/gr.216911.116>.

12. Cunningham, M., Kortsalioudaki, C. and Heath, P. (2013) “Genitourinary pathogens and preterm birth:,” *Current Opinion in Infectious Diseases*, 26(3), pp. 219–230. Available at: <https://doi.org/10.1097/QCO.0b013e328360dc31>.
13. Danecek, P. *et al.* (2021) “Twelve years of SAMtools and BCFtools,” *GigaScience*, 10(2), p. giab008. Available at: <https://doi.org/10.1093/gigascience/giab008>.
14. Denis, M. *et al.* (2014) “Placental Genome and Maternal-Placental Genetic Interactions: A Genome-Wide and Candidate Gene Association Study of Placental Abruption,” *PLoS ONE*. Edited by Z. Chen, 9(12), p. e116346. Available at: <https://doi.org/10.1371/journal.pone.0116346>.
15. Desai, M. *et al.* (2007) “Epidemiology and burden of malaria in pregnancy,” *The Lancet Infectious Diseases*, 7(2), pp. 93–104. Available at: [https://doi.org/10.1016/S1473-3099\(07\)70021-X](https://doi.org/10.1016/S1473-3099(07)70021-X).
16. Fernández De Gamarra-Oca, L. *et al.* (2021) “Long-Term Neurodevelopmental Outcomes after Moderate and Late Preterm Birth: A Systematic Review,” *The Journal of Pediatrics*, 237, pp. 168-176.e11. Available at: <https://doi.org/10.1016/j.jpeds.2021.06.004>.
17. Galinsky, R. *et al.* (2013) “The Consequences of Chorioamnionitis: Preterm Birth and Effects on Development,” *Journal of Pregnancy*, 2013, pp. 1–11. Available at: <https://doi.org/10.1155/2013/412831>.
18. Goldenberg, R.L. *et al.* (2008) “Epidemiology and causes of preterm birth,” *The Lancet*, 371(9606), pp. 75–84. Available at: [https://doi.org/10.1016/S0140-6736\(08\)60074-4](https://doi.org/10.1016/S0140-6736(08)60074-4).
19. 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. Available at: <https://doi.org/10.1097/NMC.0000000000000656>.
20. Gudmundsson, J. *et al.* (2009) “Genome-wide association and replication studies identify four variants associated with prostate cancer susceptibility,” *Nature Genetics*, 41(10), pp. 1122–1126. Available at: <https://doi.org/10.1038/ng.448>.
21. Haahr, T. *et al.* (2016) “Treatment of bacterial vaginosis in pregnancy in order to reduce the risk of spontaneous preterm delivery – a clinical recommendation,” *Acta Obstetrica et Gynecologica Scandinavica*, 95(8), pp. 850–860. Available at: <https://doi.org/10.1111/aogs.12933>.
22. Haas, J.S. *et al.* (2005) “Pregpregnancy Health Status and the Risk of Preterm Delivery,” *Archives of Pediatrics & Adolescent Medicine*, 159(1), p. 58. Available at: <https://doi.org/10.1001/archpedi.159.1.58>.
23. Habuchi, H. *et al.* (2000) “The Occurrence of Three Isoforms of Heparan Sulfate 6-O-Sulfotransferase Having Different Specificities for Hexuronic Acid Adjacent to the Targeted N-Sulfoglucosamine,” *Journal of Biological Chemistry*, 275(4), pp. 2859–2868. Available at: <https://doi.org/10.1074/jbc.275.4.2859>.

24. Hasegawa, K. *et al.* (2023) “Genome-wide association study of preterm birth and gestational age in a Japanese population,” *Human Genome Variation*, 10(1), p. 19. Available at: <https://doi.org/10.1038/s41439-023-00246-9>.
25. Hassed, S.J. *et al.* (2012) “RBPJ Mutations Identified in Two Families Affected by Adams-Oliver Syndrome,” *The American Journal of Human Genetics*, 91(2), pp. 391–395. Available at: <https://doi.org/10.1016/j.ajhg.2012.07.005>.
26. Huang, Q. -t. *et al.* (2015) “Chronic hepatitis C virus infection is associated with increased risk of preterm birth: a meta-analysis of observational studies,” *Journal of Viral Hepatitis*, 22(12), pp. 1033–1042. Available at: <https://doi.org/10.1111/jvh.12430>.
27. Hur, Y.M. *et al.* (2023) “A genome-wide and candidate gene association study of preterm birth in Korean pregnant women,” *PLOS ONE*. Edited by N. Mahdieh, 18(11), p. e0294948. Available at: <https://doi.org/10.1371/journal.pone.0294948>.
28. Institute of Hygiene (2026) “Newborns born in Lithuanian obstetric hospitals by sex, birth weight, and gestational age.” Available at: [https://sveikstat.hi.lt/infographic.aspx?ing\\_uid=1080](https://sveikstat.hi.lt/infographic.aspx?ing_uid=1080).
29. Jakuskiene, R. *et al.* (2011) “Neonatal outcomes of very preterm infants admitted to a tertiary center in Lithuania between the years 2003 and 2005,” *European Journal of Pediatrics*, 170(10), pp. 1293–1303. Available at: <https://doi.org/10.1007/s00431-011-1431-8>.
30. Juvinao-Quintero, D.L. *et al.* (2023) “Genetic association study of Preterm birth and Gestational age in a population-based case-control study in Peru.” *Genetic and Genomic Medicine*. Available at: <https://doi.org/10.1101/2023.11.22.23298891>.
31. Knijnenburg, T.A. *et al.* (2019) “Genomic and molecular characterization of preterm birth,” *Proceedings of the National Academy of Sciences*, 116(12), pp. 5819–5827. Available at: <https://doi.org/10.1073/pnas.1716314116>.
32. Kondracki, A.J. and Hofferth, S.L. (2019) “A gestational vulnerability window for smoking exposure and the increased risk of preterm birth: how timing and intensity of maternal smoking matter,” *Reproductive Health*, 16(1), p. 43. Available at: <https://doi.org/10.1186/s12978-019-0705-x>.
33. Kozuki, N. *et al.* (2013) “The associations of parity and maternal age with small-for-gestational-age, preterm, and neonatal and infant mortality: a meta-analysis,” *BMC Public Health*, 13(S3), p. S2. Available at: <https://doi.org/10.1186/1471-2458-13-S3-S2>.
34. Liu, B. *et al.* (2020) “Maternal cigarette smoking before and during pregnancy and the risk of preterm birth: A dose–response analysis of 25 million mother–infant pairs,” *PLOS Medicine*. Edited by S.J. Stock, 17(8), p. e1003158. Available at: <https://doi.org/10.1371/journal.pmed.1003158>.

35. Liu, X. *et al.* (2019) “Variants in the fetal genome near pro-inflammatory cytokine genes on 2q13 associate with gestational duration,” *Nature Communications*, 10(1), p. 3927. Available at: <https://doi.org/10.1038/s41467-019-11881-8>.
36. Loh, P.-R. *et al.* (2016) “Reference-based phasing using the Haplotype Reference Consortium panel,” *Nature Genetics*, 48(11), pp. 1443–1448. Available at: <https://doi.org/10.1038/ng.3679>.
37. Marchi, J. *et al.* (2015) “Risks associated with obesity in pregnancy, for the mother and baby: a systematic review of reviews,” *Obesity Reviews*, 16(8), pp. 621–638. Available at: <https://doi.org/10.1111/obr.12288>.
38. Marchini, J. and Howie, B. (2010) “Genotype imputation for genome-wide association studies,” *Nature Reviews Genetics*, 11(7), pp. 499–511. Available at: <https://doi.org/10.1038/nrg2796>.
39. McDonald, S.D. *et al.* (2010) “Overweight and obesity in mothers and risk of preterm birth and low birth weight infants: systematic review and meta-analyses,” *BMJ*, 341(jul20 1), pp. c3428–c3428. Available at: <https://doi.org/10.1136/bmj.c3428>.
40. Mladenicić, T. *et al.* (2024) “Protective Effect of EBF Transcription Factor 1 (EBF1) Polymorphism in Sporadic and Familial Spontaneous Preterm Birth: Insights from a Case-Control Study,” *International Journal of Molecular Sciences*, 25(20), p. 11192. Available at: <https://doi.org/10.3390/ijms252011192>.
41. National Library of Medicine; National Institutes of Health; U.S. Department of Health and Human Services (2026) “dbSNP.” Available at: <https://www.ncbi.nlm.nih.gov/snp/>.
42. Ohuma, E.O. *et al.* (2023) “National, regional, and global estimates of preterm birth in 2020, with trends from 2010: a systematic analysis,” *The Lancet*, 402(10409), pp. 1261–1271. Available at: [https://doi.org/10.1016/S0140-6736\(23\)00878-4](https://doi.org/10.1016/S0140-6736(23)00878-4).
43. Olson-Chen, C., Balaram, K. and Hackney, D.N. (2018) “Chlamydia trachomatis and Adverse Pregnancy Outcomes: Meta-analysis of Patients With and Without Infection,” *Maternal and Child Health Journal*, 22(6), pp. 812–821. Available at: <https://doi.org/10.1007/s10995-018-2451-z>.
44. OpenAI (2026) “ChatGPT.” Available at: <https://chat.openai.com>.
45. Pasanen, A. *et al.* (2022) “Meta-analysis of gestational duration and spontaneous preterm birth identifies new maternal risk loci.” *Genetic and Genomic Medicine*. Available at: <https://doi.org/10.1101/2022.10.31.22281753>.
46. Paternoster, L. *et al.* (2011) “Genome-Wide Population-Based Association Study of Extremely Overweight Young Adults – The GOYA Study,” *PLoS ONE*. Edited by P. Awadalla, 6(9), p. e24303. Available at: <https://doi.org/10.1371/journal.pone.0024303>.

47. Qin, J. *et al.* (2014) “Reported Estimates of Adverse Pregnancy Outcomes among Women with and without Syphilis: A Systematic Review and Meta-Analysis,” *PLoS ONE*. Edited by C. Rubens, 9(7), p. e102203. Available at: <https://doi.org/10.1371/journal.pone.0102203>.
48. Rappoport, N. *et al.* (2018) “A genome-wide association study identifies only two ancestry specific variants associated with spontaneous preterm birth,” *Scientific Reports*, 8(1), p. 226. Available at: <https://doi.org/10.1038/s41598-017-18246-5>.
49. Risnes, K. *et al.* (2021) “Mortality Among Young Adults Born Preterm and Early Term in 4 Nordic Nations,” *JAMA Network Open*, 4(1), p. e2032779. Available at: <https://doi.org/10.1001/jamanetworkopen.2020.32779>.
50. Shaun Purcell, Christopher Chang (2026) “PLINK.” Available at: [www.cog-genomics.org/plink/2.0/](http://www.cog-genomics.org/plink/2.0/).
51. Sherry, S.T. (2001) “dbSNP: the NCBI database of genetic variation,” *Nucleic Acids Research*, 29(1), pp. 308–311. Available at: <https://doi.org/10.1093/nar/29.1.308>.
52. Solé-Navais, P. *et al.* (2023) “Genetic effects on the timing of parturition and links to fetal birth weight,” *Nature Genetics*, 55(4), pp. 559–567. Available at: <https://doi.org/10.1038/s41588-023-01343-9>.
53. Soneji, S. and Beltrán-Sánchez, H. (2019) “Association of Maternal Cigarette Smoking and Smoking Cessation With Preterm Birth,” *JAMA Network Open*, 2(4), p. e192514. Available at: <https://doi.org/10.1001/jamanetworkopen.2019.2514>.
54. Stawiński, P. and Płoski, R. (2024) “Genebe.net: Implementation and validation of an automatic ACMG variant pathogenicity criteria assignment,” *Clinical Genetics*, 106(2), pp. 119–126. Available at: <https://doi.org/10.1111/cge.14516>.
55. Stelzer, G. *et al.* (2016) “The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses,” *Current Protocols in Bioinformatics*, 54(1). Available at: <https://doi.org/10.1002/cpbi.5>.
56. 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. 1365–1372. Available at: <https://doi.org/10.1093/aje/kwp328>.
57. Teune, M.J. *et al.* (2011) “A systematic review of severe morbidity in infants born late preterm,” *American Journal of Obstetrics and Gynecology*, 205(4), p. 374.e1-374.e9. Available at: <https://doi.org/10.1016/j.ajog.2011.07.015>.
58. The 1000 Genomes Project Consortium *et al.* (2015) “A global reference for human genetic variation,” *Nature*, 526(7571), pp. 68–74. Available at: <https://doi.org/10.1038/nature15393>.

59. Torloni, M.R. *et al.* (2009) “Maternal BMI and preterm birth: A systematic review of the literature with meta-analysis,” *The Journal of Maternal-Fetal & Neonatal Medicine*, 22(11), pp. 957–970. Available at: <https://doi.org/10.3109/14767050903042561>.
60. Uffelmann, E. *et al.* (2021) “Genome-wide association studies,” *Nature Reviews Methods Primers*, 1(1), p. 59. Available at: <https://doi.org/10.1038/s43586-021-00056-9>.
61. Urniyte, 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. Available at: <https://doi.org/10.3390/genes13040569>.
62. Variant annotation and interpretation tool (2024) “GeneBe. Variant annotation and interpretation tool.” Available at: <https://genebe.net/>.
63. Vogel, J.P. *et al.* (2018) “The global epidemiology of preterm birth,” *Best Practice & Research Clinical Obstetrics & Gynaecology*, 52, pp. 3–12. Available at: <https://doi.org/10.1016/j.bpobgyn.2018.04.003>.
64. Winn, V.D. *et al.* (2007) “Gene Expression Profiling of the Human Maternal-Fetal Interface Reveals Dramatic Changes between Midgestation and Term,” *Endocrinology*, 148(3), pp. 1059–1079. Available at: <https://doi.org/10.1210/en.2006-0683>.
65. Xiao, P.-L. *et al.* (2015) “Association between maternal HIV infection and low birth weight and prematurity: a meta-analysis of cohort studies,” *BMC Pregnancy and Childbirth*, 15(1), p. 246. Available at: <https://doi.org/10.1186/s12884-015-0684-z>.
66. York, T.P. *et al.* (2009) “Estimating Fetal and Maternal Genetic Contributions to Premature Birth From Multiparous Pregnancy Histories of Twins Using MCMC and Maximum-Likelihood Approaches,” *Twin Research and Human Genetics*, 12(4), pp. 333–342. Available at: <https://doi.org/10.1375/twin.12.4.333>.
67. Zhang, G. *et al.* (2017) “Genetic Associations with Gestational Duration and Spontaneous Preterm Birth,” *New England Journal of Medicine*, 377(12), pp. 1156–1167. Available at: <https://doi.org/10.1056/NEJMoal612665>.

# SUMMARY

## Genome-Wide Association Study of Preterm Birth Using Genotyping Data

Natalija Ošmianskaja

Systems Biology Master Program

Vilnius University, Faculty of Medicine

Preterm birth is a complex pregnancy-related condition with significant health consequences and a partially understood genetic basis. This study aimed to investigate genetic variants associated with preterm birth using genome-wide association approaches. The research was motivated by the need to better understand the genetic contribution to this phenotype, particularly in underrepresented populations.

Genotyping data were analyzed using multiple statistical approaches, including candidate variant analysis, logistic regression (PLINK), and association testing with SNPTEST, both with and without imputed data. Quality control procedures were applied, and results were visualized using Manhattan and QQ plots. Additional steps included identification of overlapping variants across models and functional annotation.

The analysis identified several suggestive genetic variants; however, no robust genome-wide significant associations were consistently supported across methods. Some signals detected, particularly in SNPTEST analyses, were likely influenced by statistical inflation and limited sample size.

In conclusion, the findings should be considered exploratory and do not provide strong evidence for specific genetic determinants of preterm birth. Larger, well-powered studies with independent replication are required to clarify the genetic architecture of this complex phenotype.

## SUMMARY IN LITHUANIAN

Priešlaikinio gimdymo viso genomo asociacijos tyrimas naudojant genotipavimo duomenis

Natalija Ošmianskaja

Sistemų biologijos magistrantūros studijų programa

Vilniaus universitetas, Medicinos fakultetas

Priešlaikinis gimdymas yra sudėtinga su nėštumu susijusi būklė, turinti reikšmingų sveikatos pasekmių ir iš dalies neaiškų genetinį pagrindą. Šio tyrimo tikslas buvo ištirti genetinius variantus, susijusius su priešlaikiniu gimdymu, taikant viso genomo asociacijos tyrimo metodus. Tyrimas buvo motyvuotas poreikiu geriau suprasti genetinį šio fenotipo pagrindą, ypač nepakankamai ištirtose populiacijose.

Genotipavimo duomenys buvo analizuojami taikant kelis statistinius metodus, įskaitant kandidatinių variantų analizę, logistinės regresijos analizę (PLINK) ir asociacijų analizę naudojant SNPTEST, tiek su imputuotais duomenimis, tiek be jų. Buvo atliktos duomenų kokybės kontrolės procedūros, o rezultatai vizualizuoti naudojant Manhattan ir QQ grafikus. Papildomai buvo identifikuojami tarp modelių sutampantys variantai ir atliekama funkcinė anotacija.

Analizės metu nustatyti keli galimi genetiniai variantai, tačiau nė vienas genomo mastu reikšmingas ryšys nebuvo nuosekliai patvirtintas skirtingais metodais. Kai kurie signalai, ypač gauti naudojant SNPTEST, greičiausiai buvo paveikti statistinės infliacijos ir riboto imties dydžio.

Apibendrinant, gauti rezultatai laikytini preliminariais ir nepateikia tvirtų įrodymų apie konkrečius genetinius priešlaikinio gimdymo veiksnius. Norint geriau suprasti šio sudėtingo fenotipo genetinę architektūrą, būtini didesni, pakankamos galios tyrimai su nepriklausomu rezultatų patvirtinimu.

## APPENDICES

### Appendix 1. Candidate variant list investigated in the study (GRCh38)

| Variant     | Gene             | Chromosome | Start (GRCh38) | End (GRCh38) |
|-------------|------------------|------------|----------------|--------------|
| rs10167914  | –                | 2          | 112805784      | 112805784    |
| rs10934853  | <i>EEFSEC</i>    | 3          | 128319530      | 128319530    |
| rs112912841 | <i>LPP</i>       | 3          | 188269895      | 188269895    |
| rs113828443 | <i>RNU6-410P</i> | 4          | 55029115       | 55029115     |
| rs11708067  | <i>ADCY5</i>     | 3          | 123346931      | 123346931    |
| rs12037376  | <i>WNT4</i>      | 1          | 22135618       | 22135618     |
| rs143530128 | <i>RN7SKP155</i> | 8          | 123591544      | 123591544    |
| rs17713682  | –                | 17         | 12513046       | 12513046     |
| rs1877720   | <i>SPATA6</i>    | 1          | 48358735       | 48358735     |
| rs200745338 | <i>EEFSEC</i>    | 3          | 128150615      | 128150617    |
| rs201450565 | <i>EEFSEC</i>    | 3          | 128339773      | 128339773    |
| rs2387280   | –                | 16         | 4289683        | 4289683      |
| rs2659685   | <i>EEFSEC</i>    | 3          | 128403553      | 128403553    |
| rs28654158  | <i>ADCY5</i>     | 3          | 123393445      | 123393445    |
| rs2946169   | <i>LINC02227</i> | 5          | 158491951      | 158491951    |
| rs2955117   | <i>EEFSEC</i>    | 3          | 128162770      | 128162770    |
| rs2963463   | <i>LINC02227</i> | 5          | 158468041      | 158468041    |
| rs2999052   | <i>EEFSEC</i>    | 3          | 128173194      | 128173194    |
| rs3129768   | <i>HLA-DQA1</i>  | 6          | 32627306       | 32627306     |
| rs34555419  | <i>TET3</i>      | 2          | 74008799       | 74008799     |
| rs3820282   | <i>WNT4</i>      | 1          | 22141722       | 22141722     |
| rs4383453   | <i>ADCY5</i>     | 3          | 123349512      | 123349512    |
| rs56318008  | <i>WNT4</i>      | 1          | 22143914       | 22143914     |
| rs6090040   | <i>TCEA2</i>     | 20         | 64060707       | 64060707     |
| rs62270785  | <i>MRPS22</i>    | 3          | 139285491      | 139285491    |
| rs625036    | <i>HIVEP3</i>    | 1          | 41805108       | 41805108     |
| rs6472846   | <i>GDAP1</i>     | 8          | 74402911       | 74402911     |
| rs6780427   | <i>LEKR1</i>     | 3          | 156979308      | 156979308    |
| rs6831441   | <i>LINC02269</i> | 4          | 173812145      | 173812145    |
| rs72898946  | <i>FAF1</i>      | 1          | 50493590       | 50493590     |
| rs7594852   | <i>CKAP2L</i>    | 2          | 112764177      | 112764177    |
| rs7650602   | <i>ZBTB38</i>    | 3          | 141428572      | 141428572    |
| rs7729301   | <i>LINC02227</i> | 5          | 158459945      | 158459945    |
| rs9823520   | –                | 3          | 14252332       | 14252332     |
| rs9861425   | <i>ADCY5</i>     | 3          | 123354036      | 123354036    |

## Appendix 2. Covariate file covariate.txt, used for logistic regression with PLINK

FID IID pheno sex w\_ga weight pheno\_sex weight\_category exact\_ga  
gd\_preeclampsia spb\_pprom latency\_period\_hrs latency\_period\_cat  
mog mothers\_age nr\_preg prev\_preg\_complications mother\_ancestry  
mother\_family\_hist

NANE-01-3 NANE-01-3 2 1 34 3160 1 3 98.59 0 0 27 1 0 31 2 0 0 0  
NANE-011-3 NANE-011-3 2 1 31 1750 1 0 49.15 0 0 384 1 0 30 2 0 0  
0  
NANE-012-3 NANE-012-3 2 2 32 1885 0 0 51.6 0 0 3 0 1 36 2 0 1 0  
NANE-013-3 NANE-013-3 2 1 32 2200 1 0 79.7 1 0 2 0 0 30 1 0 0 1  
NANE-014-3 NANE-014-3 2 1 33 1980 1 0 28.07 0 0 183 1 0 37 3 1 0  
1  
NANE-019-3 NANE-019-3 2 2 32 1600 0 0 18 0 1 27 1 1 36 7 1 0 0  
NANE-02-3 NANE-02-3 2 2 27 840 0 0 14.7 0 0 0 0 1 37 6 1 0 0  
NANE-021-3 NANE-021-3 2 2 32 1880 0 0 51 0 0 0 0 0 31 3 0 0 0  
NANE-024-3 NANE-024-3 2 1 31 1600 1 0 28.17 0 0 0 0 1 26 1 0 1 0  
NANE-025-3 NANE-025-3 2 2 24 900 0 2 95.4 0 1 75 1 1 38 1 0 0 1  
NANE-030-3 NANE-030-3 2 2 33 2060 0 0 47.9 0 1 7 0 0 34 2 0 0 0  
NANE-031-3 NANE-031-3 2 1 33 2410 1 0 77.76 0 0 30 1 0 40 2 1 1  
1  
NANE-04-3 NANE-04-3 2 1 24 750 1 0 98.59 0 1 212 1 1 37 1 0 0 0  
NANE-046-3 NANE-046-3 2 2 29 820 0 1 1 1 0 0 0 1 36 2 0 0 1  
NANE-05-3 NANE-05-3 2 2 24 600 0 0 33.2 0 1 1 0 1 26 2 0 0 0  
NANE-06-3 NANE-06-3 2 2 34 1920 0 0 13.27 0 0 13 0 0 30 1 0 0 0  
NANE-015-3 NANE-015-3 2 1 32 1990 1 0 54.94 0 1 79 1 0 32 4 1 1  
1  
NANE-026-3 NANE-026-3 2 1 31 1230 1 1 2.46 0 1 74 1 0 26 2 1 0 0  
NANE-027-3 NANE-027-3 2 2 33 2260 0 0 70.9 1 1 0 0 0 20 1 0 0 0  
NANE-03-3 NANE-03-3 2 2 31 1250 0 1 5.6 1 0 0 0 1 28 2 0 0 0  
NANE-034-3 NANE-034-3 2 1 32 1990 1 0 54.94 0 1 0 0 0 32 2 1 0 1  
NANE-036-3 NANE-036-3 2 1 31 1470 1 0 14.36 0 0 0 0 0 30 1 0 0 0

NANE-039-3 NANE-039-3 2 1 26 1000 1 0 59.44 0 1 139 1 1 40 2 1 1  
 0  
 NANE-040-3 NANE-040-3 2 1 24 780 1 0 73.02 0 1 28 1 1 34 2 0 0 0  
 NANE-041-3 NANE-041-3 2 2 33 1530 0 1 3.9 1 1 1 0 0 39 2 1 0 1  
 NANE-043-3 NANE-043-3 2 1 31 1790 1 0 55.03 0 0 52 1 0 29 1 0 1  
 1  
 NANE-045-3 NANE-045-3 2 1 29 1300 1 0 32.46 0 1 0 0 1 24 2 1 1 0  
 NANE-047-3 NANE-047-3 2 1 24 560 1 0 13.27 0 0 0 0 1 39 5 1 4 0  
 NANE-07-3 NANE-07-3 2 1 23 640 1 0 31.59 0 0 64 1 1 30 1 0 1 0  
 S3420Nr18 ANE-001-3 1 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9  
 -9 -9  
 S3420Nr36 ANE-009-3 1 1 40 3680 1 0 16.49 1 1 14 0 0 31 1 0 0 0  
 S3420Nr39 ANE-010-3 1 1 40 3250 1 0 9.2 0 0 0 0 0 37 2 0 0 0  
 S3420Nr42 ANE-011-3 1 1 40 3600 1 0 15.14 0 0 3 0 0 36 3 0 0 0  
 S3420Nr45 ANE-013-3 1 1 40 3200 1 0 13.6 0 0 1 0 0 31 2 0 0 0  
 S3420Nr48 ANE-014-3 1 2 41 3300 0 0 22.1 1 0 4 0 0 26 1 0 0 0  
 S3420Nr51 ANE-015-3 1 1 40 3600 1 0 43.6 0 0 7 0 0 32 2 0 0 0  
 S3420Nr54 ANE-017-3 1 2 38 2840 0 0 17.8 0 0 4 0 0 32 1 0 0 0  
 S3420Nr57 ANE-019-3 1 1 41 3800 1 0 51.6 0 0 10 0 0 32 1 0 0 0  
 S3420Nr60 ANE-021-3 1 1 40 3500 1 0 34.5 1 1 14 0 0 27 2 0 0 0  
 S3420Nr63 ANE-022-3 1 1 39 3830 1 0 76.6 1 0 1 0 0 29 3 1 0 0  
 S3420Nr66 ANE-023-3 1 1 40 3460 1 0 31.1 1 1 13 0 0 21 1 0 0 0  
 S3420Nr69 ANE-025-3 1 1 40 3800 1 0 62.3 1 0 5 0 0 27 1 0 0 0  
 S3420Nr72 ANE-027-3 1 2 38 3090 0 0 39 0 0 5 0 0 32 2 0 0 0  
 S3420Nr75 ANE-028-3 1 2 41 3860 0 0 72.3 1 0 7 0 0 30 2 1 0 0  
 S3420Nr20 ANE-002-2 1 2 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -  
 9 -9  
 S3420Nr78 ANE-029-3 1 1 40 3790 1 0 61.4 0 0 2 0 0 34 1 0 0 0  
 S3420Nr81 ANE-030-3 1 2 39 3620 0 0 72.5 0 0 2 0 0 32 3 1 0 0  
 S3420Nr84 ANE-031-3 1 2 39 3550 0 0 66.6 0 1 15 0 0 30 1 0 0 0  
 S3420Nr87 ANE-032-3 1 2 38 3520 0 0 78.3 0 0 9 0 0 31 1 0 0 0

|           |           |   |    |    |      |    |    |       |    |    |    |    |    |    |    |    |    |    |
|-----------|-----------|---|----|----|------|----|----|-------|----|----|----|----|----|----|----|----|----|----|
| S3420Nr90 | ANE-033-3 | 1 | 1  | 40 | 3840 | 1  | 0  | 65.9  | 0  | 0  | 2  | 0  | 0  | 28 | 2  | 0  | 0  | 0  |
| S3420Nr24 | ANE-004-3 | 1 | 1  | 40 | 4160 | 1  | 0  | 24.63 | 0  | 0  | 3  | 0  | 0  | 34 | 1  | 0  | 0  | 0  |
| S3420Nr27 | ANE-005-3 | 1 | -9 | -9 | -9   | -9 | -9 | -9    | -9 | -9 | -9 | -9 | -9 | -9 | -9 | -9 | -9 | -9 |
| S3420Nr30 | ANE-006-3 | 1 | 1  | 40 | 3700 | 1  | 0  | 16.83 | 1  | 1  | 10 | 0  | 0  | 29 | 1  | 0  | 0  | 0  |
| S3420Nr33 | ANE-008-3 | 1 | 1  | 40 | 3900 | 1  | 0  | 20.22 | 0  | 1  | 8  | 0  | 0  | 26 | 2  | 0  | 0  | 0  |

### Appendix 3. Sample file sample.txt, used for SNPTEST association analysis

```
ID_1 ID_2 pheno sex w_ga weight pheno_sex weight_category
exact_ga gd_preeclampsia spb_pprom latency_period_hrs
latency_period_cat mog mothers_age nr_preg
prev_preg_complications mother_ancestry mother_family_hist
0 0 B C C C C C C C C C C C C C C C
NANE-01-3 NANE-01-3 1 1 34 3160 1 3 98.59 0 0 27 1 0 31 2 0 0 0
NANE-011-3 NANE-011-3 1 1 31 1750 1 0 49.15 0 0 384 1 0 30 2 0 0
0
NANE-012-3 NANE-012-3 1 2 32 1885 0 0 51.6 0 0 3 0 1 36 2 0 1 0
NANE-013-3 NANE-013-3 1 1 32 2200 1 0 79.7 1 0 2 0 0 30 1 0 0 1
NANE-014-3 NANE-014-3 1 1 33 1980 1 0 28.07 0 0 183 1 0 37 3 1 0
1
NANE-019-3 NANE-019-3 1 2 32 1600 0 0 18 0 1 27 1 1 36 7 1 0 0
NANE-02-3 NANE-02-3 1 2 27 840 0 0 14.7 0 0 0 0 1 37 6 1 0 0
NANE-021-3 NANE-021-3 1 2 32 1880 0 0 51 0 0 0 0 0 31 3 0 0 0
NANE-024-3 NANE-024-3 1 1 31 1600 1 0 28.17 0 0 0 0 1 26 1 0 1 0
NANE-025-3 NANE-025-3 1 2 24 900 0 2 95.4 0 1 75 1 1 38 1 0 0 1
NANE-030-3 NANE-030-3 1 2 33 2060 0 0 47.9 0 1 7 0 0 34 2 0 0 0
NANE-031-3 NANE-031-3 1 1 33 2410 1 0 77.76 0 0 30 1 0 40 2 1 1
1
NANE-04-3 NANE-04-3 1 1 24 750 1 0 98.59 0 1 212 1 1 37 1 0 0 0
NANE-046-3 NANE-046-3 1 2 29 820 0 1 1 1 0 0 0 1 36 2 0 0 1
NANE-05-3 NANE-05-3 1 2 24 600 0 0 33.2 0 1 1 0 1 26 2 0 0 0
NANE-06-3 NANE-06-3 1 2 34 1920 0 0 13.27 0 0 13 0 0 30 1 0 0 0
NANE-015-3 NANE-015-3 1 1 32 1990 1 0 54.94 0 1 79 1 0 32 4 1 1
1
NANE-026-3 NANE-026-3 1 1 31 1230 1 1 2.46 0 1 74 1 0 26 2 1 0 0
NANE-027-3 NANE-027-3 1 2 33 2260 0 0 70.9 1 1 0 0 0 20 1 0 0 0
NANE-03-3 NANE-03-3 1 2 31 1250 0 1 5.6 1 0 0 0 1 28 2 0 0 0
NANE-034-3 NANE-034-3 1 1 32 1990 1 0 54.94 0 1 0 0 0 32 2 1 0 1
NANE-036-3 NANE-036-3 1 1 31 1470 1 0 14.36 0 0 0 0 0 30 1 0 0 0
```

NANE-039-3 NANE-039-3 1 1 26 1000 1 0 59.44 0 1 139 1 1 40 2 1 1  
 0  
 NANE-040-3 NANE-040-3 1 1 24 780 1 0 73.02 0 1 28 1 1 34 2 0 0 0  
 NANE-041-3 NANE-041-3 1 2 33 1530 0 1 3.9 1 1 1 0 0 39 2 1 0 1  
 NANE-043-3 NANE-043-3 1 1 31 1790 1 0 55.03 0 0 52 1 0 29 1 0 1  
 1  
 NANE-045-3 NANE-045-3 1 1 29 1300 1 0 32.46 0 1 0 0 1 24 2 1 1 0  
 NANE-047-3 NANE-047-3 1 1 24 560 1 0 13.27 0 0 0 0 1 39 5 1 4 0  
 NANE-07-3 NANE-07-3 1 1 23 640 1 0 31.59 0 0 64 1 1 30 1 0 1 0  
 ANE-001-3 ANE-001-3 0 NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA  
 NA NA  
 ANE-009-3 ANE-009-3 0 1 40 3680 1 0 16.49 1 1 14 0 0 31 1 0 0 0  
 ANE-010-3 ANE-010-3 0 1 40 3250 1 0 9.2 0 0 0 0 0 37 2 0 0 0  
 ANE-011-3 ANE-011-3 0 1 40 3600 1 0 15.14 0 0 3 0 0 36 3 0 0 0  
 ANE-013-3 ANE-013-3 0 1 40 3200 1 0 13.6 0 0 1 0 0 31 2 0 0 0  
 ANE-014-3 ANE-014-3 0 2 41 3300 0 0 22.1 1 0 4 0 0 26 1 0 0 0  
 ANE-015-3 ANE-015-3 0 1 40 3600 1 0 43.6 0 0 7 0 0 32 2 0 0 0  
 ANE-017-3 ANE-017-3 0 2 38 2840 0 0 17.8 0 0 4 0 0 32 1 0 0 0  
 ANE-019-3 ANE-019-3 0 1 41 3800 1 0 51.6 0 0 10 0 0 32 1 0 0 0  
 ANE-021-3 ANE-021-3 0 1 40 3500 1 0 34.5 1 1 14 0 0 27 2 0 0 0  
 ANE-022-3 ANE-022-3 0 1 39 3830 1 0 76.6 1 0 1 0 0 29 3 1 0 0  
 ANE-023-3 ANE-023-3 0 1 40 3460 1 0 31.1 1 1 13 0 0 21 1 0 0 0  
 ANE-025-3 ANE-025-3 0 1 40 3800 1 0 62.3 1 0 5 0 0 27 1 0 0 0  
 ANE-027-3 ANE-027-3 0 2 38 3090 0 0 39 0 0 5 0 0 32 2 0 0 0  
 ANE-028-3 ANE-028-3 0 2 41 3860 0 0 72.3 1 0 7 0 0 30 2 1 0 0  
 ANE-002-2 ANE-002-2 0 2 NA NA NA NA NA NA NA NA NA NA NA NA NA NA  
 NA NA  
 ANE-029-3 ANE-029-3 0 1 40 3790 1 0 61.4 0 0 2 0 0 34 1 0 0 0  
 ANE-030-3 ANE-030-3 0 2 39 3620 0 0 72.5 0 0 2 0 0 32 3 1 0 0  
 ANE-031-3 ANE-031-3 0 2 39 3550 0 0 66.6 0 1 15 0 0 30 1 0 0 0  
 ANE-032-3 ANE-032-3 0 2 38 3520 0 0 78.3 0 0 9 0 0 31 1 0 0 0

|           |           |   |    |    |      |    |    |       |    |    |    |    |    |    |    |    |    |    |
|-----------|-----------|---|----|----|------|----|----|-------|----|----|----|----|----|----|----|----|----|----|
| ANE-033-3 | ANE-033-3 | 0 | 1  | 40 | 3840 | 1  | 0  | 65.9  | 0  | 0  | 2  | 0  | 0  | 28 | 2  | 0  | 0  | 0  |
| ANE-004-3 | ANE-004-3 | 0 | 1  | 40 | 4160 | 1  | 0  | 24.63 | 0  | 0  | 3  | 0  | 0  | 34 | 1  | 0  | 0  | 0  |
| ANE-005-3 | ANE-005-3 | 0 | NA | NA | NA   | NA | NA | NA    | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| ANE-006-3 | ANE-006-3 | 0 | 1  | 40 | 3700 | 1  | 0  | 16.83 | 1  | 1  | 10 | 0  | 0  | 29 | 1  | 0  | 0  | 0  |
| ANE-008-3 | ANE-008-3 | 0 | 1  | 40 | 3900 | 1  | 0  | 20.22 | 0  | 1  | 8  | 0  | 0  | 26 | 2  | 0  | 0  | 0  |