

# **DEVELOPMENT OF A NETWORK-BASED MEASURE OF GENETIC RISK AND ITS APPLICATION IN PRETERM BIRTH AND RETINOPATHY OF PREMATURITY**

By

Ryan Swan

A DISSERTATION

Presented to the Department of Medical Informatics and Clinical Epidemiology  
And the Oregon Health & Science University

School of Medicine

In partial fulfillment of  
The requirements of the degree of

Doctor of Philosophy

AUGUST 2019

School of Medicine  
Oregon Health & Science University

**Certificate of Approval**

This is to certify that the PhD Dissertation of

**Ryan M. Swan**

*“Development of a network-based measure of genetic risk and its applications in preterm birth and retinopathy of prematurity”*

Has been approved

---

Dissertation Advisor – Michael F. Chiang, M.D.

---

Committee Member – Kemal Sonmez, Ph.D.

---

Committee Member – Jayashree Kalpathy-Cramer, Ph.D.

---

Committee Member – Shannon McWeeney, Ph.D.

---

Committee Member – Michael Mooney, Ph.D.

DEVELOPMENT OF A NETWORK-BASED MEASURE OF GENETIC RISK AND ITS APPLICATION IN PRETERM BIRTH AND RETINOPATHY OF PREMATUREITY	1
ABSTRACT	7
ACKNOWLEDGEMENTS	8
CHAPTER 1: INTRODUCTION AND BACKGROUND	9
1.1 INTRODUCTION	9
1.2 STATEMENT OF AIMS	12
1.3 ORGANIZATION OF THIS DISSERTATION	13
1.4 RETINOPATHY OF PREMATUREITY	15
1.4.1 PRESENTATION OF ROP	15
1.4.2 GENETIC CONTRIBUTIONS TO DISEASE	16
1.5 SPONTANEOUS PRETERM BIRTH	17
1.5.1 RISK MEASURES OF PRETERM BIRTH	18
1.5.2 GENETIC RISK OF PRETERM BIRTH	19
1.5.3 OUTLOOK FOR GENETIC STUDY	21
1.6 POLYGENIC RISK SCORING	21
1.7 GENE SET ANALYSIS	23
1.8 PROTEIN-PROTEIN INTERACTION DATABASES	25
1.9 NETWORK-BASED MODELING	26
CHAPTER 2: THE GENETICS OF RETINOPATHY OF PREMATUREITY	30
ABSTRACT	30
2.1 INTRODUCTION	32
2.2 METHODS	35
2.3 CANDIDATE GENES IN ROP	35
2.3.1 VEGF AND ASSOCIATED RECEPTORS	35
2.3.2 FEVR, NORRIE DISEASE AND THE WNT PATHWAY	44
2.3.3 IGF-1	46
2.3.4 eNOS	46
2.3.5 INFLAMMATORY MEDIATORS	47
2.3.6 BRAIN-DERIVED NEUROTROPHIC FACTOR	48
2.3.7 RENIN-ANGIOTENSIN SYSTEM	48
2.3.8 ANGIOPOIETINS	49

2.3.9 ERYTHROPOIETIN	49
2.3.10 HYPOXIA INDUCIBLE FACTOR	50
2.3.11 HEME-OXYGENASE-1	51
2.4 OTHER CANDIDATE FACTORS	52
2.5 DISCUSSION	54
2.5.1 SUMMARY OF PREVIOUS STUDIES	54
2.5.2 LIMITATIONS OF PREVIOUS STUDIES	55
2.5.3 FUTURE DIRECTIONS OF STUDYING ROP GENETICS	56
2.5.4 EXPECTED BENEFITS OF GENETIC STUDIES OF ROP	58
2.5.5 CONCLUSIONS	59
2.6 Author Contributions	60
CHAPTER 3: A NETWORK MODEL OF POLYGENIC RISK	61
3.1 ABSTRACT	61
3.2 INTRODUCTION	62
3.3 METHODS	67
3.3.1 GWAS DATA	67
Health and Retirement Study	67
GIANT Consortium Data	67
Danish National Birth Cohort	68
Imputation	69
Reference Genome	69
Phenotypic Data	69
3.3.2 NETWORK DATA	70
Ensembl	70
STRING	70
Assessing Genomic Coverage	71
Mapping From GWAS To STRING	71
3.3.3 POLYGENIC RISK MODEL	72
Early Network Context Integration	72
Late Network Context Integration	73
Calculation of PRS from SNP Values	74
3.4 Results	75
3.4.1 IMPUTATION ANALYSIS	75



3.4.2 EDA OF STRING NETWORK	76
3.4.3 SNP MAPPING MODEL	79
SNP Mapping Process	79
Investigation of Network Coverage	79
Traditional PRS Formulation	83
3.4.4 EARLY NETWORK CONTEXT METHODS	83
Inclusion of All Gene Regions	83
Connectivity Gated Score	85
Network Connectivity Analysis	85
Connectivity Score Performance	88
Candidate Gene Score	89
Simulation of Height Candidate List	89
Candidate Gene Score Construction	90
Permutation Testing of PRS Fit	92
3.4.5 LATE NETWORK CONTEXT METHODS	93
Modularity Assessment of HRS Data	94
Permutation Testing	94
Module Enrichment	95
Insulated Heat Diffusion Method	99
HotNet2 Community Assessment	101
3.4.6 PRETERM BIRTH PRS ASSESSMENT	102
GWAS Evaluation	102
Traditional PRS Formulation	103
Early Context Model	104
Inclusion of All Gene Regions	104
Connectivity Gated Model	104
Candidate Gated Model	104
ROP Candidate Gene Model	105
Permutation Analysis	107
Late Context Model	108
Module Analysis	108
HotNet2 Analysis	111
3.4.7 PRETERM BIRTH AND ROP OVERLAP ASSESSMENT	111

Performance of ROP Candidate Gene Score vs Preterm Birth Score	114
Enrichment of Preterm Birth for ROP Genes	115
3.5 DISCUSSION	117
3.6 CONCLUSIONS	122
CHAPTER 4: DISCUSSION AND CONCLUSIONS	123
APPENDIX	133
HRS GWAS Evaluation	133
Data Acquisition	133
Initial EDA	133
Data Cleaning	134
Pre-imputation Association	138
Imputation	139
Secondary Data QA/QC	139
DNBC GWAS EVALUATION	140
TOP CONNECTIVITY GENES	145
FULL SETS OF ENRICHED GENE ONTOLOGY TERMS	146
FULL LIST OF GENES EXCLUSIVE TO ROP CANDIDATE GENE SCORE	148
CANDIDATE GENE SCORES WITH FIXED AND UNFIXED THRESHOLDS	150
ROP ENRICHMENT FOR PTB GENES	151
ADDITIONAL PRS PLOTS	152
GIANT PRS Plots	152
DNBC PRS Plots	155
COMPUTATIONAL ASSETS	158
BIBLIOGRAPHY	159

## ABSTRACT

Despite incredible advances in the recruitment and phenotyping of patients for genomewide association studies (GWAS) in the past decade, the ability to ascertain the causes of complex disease remains a significant challenge. Recent research implies that instead of single variants of large effect, many variants of extremely small effect represent the majority of signal associated with genetic disease. To approximate these broad effects, enormous sample sizes are required. However, such recruitment is often impossible in rare diseases. To this end, we introduce a modified model of polygenic risk score (PRS) formulation incorporating protein-protein interaction network topology. This method allows the investigation of the degree to which network effects can augment existing risk scores and provides a complementing framework within which to assess the composition of existing polygenic risk measures. Secondly, we assess the ability to discern the degree to which coexisting conditions are due to similar genetic causes using the framework described above. We assess the degree to which genetic effects are detectable in a small population enriched for prematurity. We evaluate the overlap between signal for preterm birth with that of and retinopathy of prematurity (ROP), a coincident condition believed to have independent genetic causes. We believe this work is an important step toward increasing the predictive power and interpretability of genetic risk score methods, and that the evolution of such score will help inform and direct research in genetic disease to the benefit of patients and clinicians.

## ACKNOWLEDGEMENTS

While a dissertation is a famously solitary endeavor, it represents the effort of a multitude of people without whom its completion would be impossible.

My first thanks is to my committee. To Michael Chiang for his support throughout the dissertation process, and for taking a chance on me when I was a masters student trying to find a home at OHSU. My thanks to Jayashree Kalpathy-Cramer for her guidance and constant vigilance towards statistical integrity. My thanks to Kemal Sonmez for his help getting me up and running with cluster computing, and for believing in my programming ability on my first phone call with a faculty member many years ago. And thanks to Shannon McWeeney for her enormous investments of time and effort in her attempts to get me to open my mind and think like something approximating a real live scientist.

I would also like to thank Michael Mooney for his mentorship both as a scientist and a teacher, and for his infinite patience helping me learn how to think about networks and genetics.

I would like to thank the DMICE fellows, especially Aurora Blucher, Eric Leung, and Aaron Coyner, for being there to talk through the many issues that come up in research and in life.

I thank my parents Joel and Teresa for the many, many phone calls they insist they were happy to answer when my belief in my endeavors or myself faltered, not only in the production of this dissertation but throughout my life.

And my thanks to Brianna Barrett for not just being there for me but for living the experience with me, and somehow believing in me until the end. You're the best B.

# CHAPTER 1: INTRODUCTION AND BACKGROUND

## 1.1 INTRODUCTION

Retinopathy of prematurity (ROP) is a disorder of the retinal vasculature affecting premature infants. Estimated to lead to blindness in more than 50,000 infants annually, it is a leading cause of lifelong blindness both in the United States and worldwide and represents an enormous social and economic burden. While treatments for ROP are advancing, the ability of care facilities to better treat the complications of preterm birth has led to an increasing incidence of ROP over time.

Surgical interventions for ROP exist, but long-term outcomes for patients often do not approach full visual acuity. Early treatments focused on cryotherapy, inducing scarring of the retina in order to halt disease progress. These surgical interventions have since been unseated by laser ablation therapies, which serve the same function but result in more predictable outcomes with reduced tissue damage to patients. Drug interventions such as bevacizumab have also proven effective, but concerns about nervous system exposure to anti-angiogenic compounds remains a concern. Due to these shortcomings, new treatments and therapies are high in demand.

ROP has long been hypothesized to have a genetic basis of disease. Identification of such a causal link would be useful for scheduling and prioritization of high-risk patients. Despite numerous single-gene studies, a strong genetic predictor of disease has to date not been found. While studies focusing on single genes or small subsets of genes have been

successful at identifying genetic variants with significant effect, these studies have also been troubled by challenges with replicability and interpretation.

An additional challenge in genetic studies are cases where the definition of a disease phenotype is not easily specified. While certain disease phenotypes such as Huntington's disease, Down syndrome, and cystic fibrosis carry definitive phenotypes with discrete genetic causes, many if not the vast array of complex disease phenotypes like autism, diabetes, and cardiovascular disease exist with definitions that contain several modes of progression with many potentially participating biological systems.

ROP is a rare and complex disease with several axes of severity and differing manifestation between populations, and as such, phenotyping presents a significant challenge. Approximation of disease severity has been shown to vary a great deal, even when limited to expert clinicians specializing in ROP diagnosis.<sup>(1)</sup> Due to these concerns, larger studies of individuals are required, but the rarity of the disease makes recruitment of large study cohorts to facilitate large-scale genetic testing a difficult task.

Additional complication is introduced due to ROP's coincident occurrence with preterm birth. Preterm birth is implicated in a host of infant disease phenotypes including chronic lung disease, periventricular leukomalacia, and death. Preterm birth has also been hypothesized to have a genetic component of disease, and this genetic component may become a surrogate trait for ROP as well. As such, methods separating such potential risks would be of particular use.

One method by which the causes of complex genetic disease has been investigated is polygenic risk scoring (PRS). PRS involves exhaustive calculation of all documented genetic

variant effects from genome-wide association study (GWAS) summary statistics, which are then summed to produce an aggregated genetic scoring metric. As the cost of large-scale genomic studies has decreased, PRS has increased in adoption, finding applications in a variety of traits ranging from cardiovascular disease, to schizophrenia, to educational attainment.(2–4)

Despite the popularity of such measures, PRS has yet to achieve clinical utility. (5) Historically, such measures are only able to capture a fraction of known heritable variability. PRS also do not provide context regarding the biological cause of disease, instead serving as a broad summary measure calculated for each individual. While scores are able to identify individuals at heightened risk for negative outcomes, often in the extremes of the PRS distribution, the degree of specificity and sensitivity of such scores has so far failed to reach a level that would allow prescription of lifestyle changes or medications which may carry a competing burden on patients.

In order to address these concerns and provide an analysis of the genetic determinants of ROP, we evaluate a modification of PRS incorporating network topology measures from protein-protein interaction databases and currently available annotation resources. We interrogate the ability of such additional contextual information to augment the power of PRS and to elucidate the factors that contribute to the predictive ability of PRS.

## 1.2 STATEMENT OF AIMS

**Specific Aim 1. Develop network-based polygenic risk model approaches with early and late incorporation of network context from the Michigan Health and Retirement Study (HRS) genome-wide association data. Evaluate these scores against traditional polygenic risk scores to determine if there is additional information provided by the network approach.**

We formulate two novel models of polygenic risk score construction using network information derived from protein-protein interaction database resources. These models are evaluated with respect to traditional PRS. Their contributions to phenotype prediction and their ability to elucidate the sources of heritable variability are evaluated. We propose guidelines for the use of PRS in future studies and the role that network context can play in constructing new scores going forward.

**Specific Aim 2. Assess genetic factors contributing to preterm birth utilizing both the traditional and network approaches from Aim 1 in the GENEVA study (Genome-Wide Association Studies of Prematurity And Its Complications). Assess the impact of inclusion of ROP candidate genes in early network approach as well as contribution of ROP candidates to late network PRS to determine the role of related trait genes in prematurity disease network.**

Using a dataset created for the investigation of the causes of preterm birth, we evaluate the performance of the described network PRS model on datasets of limited size. We incorporate previous studies of ROP genetics to curate a list of candidate genes, and using network-modified PRS incorporating additional context, we evaluate a novel method of quantifying the amount of overlap between preterm birth and ROP.



### 1.3 ORGANIZATION OF THIS DISSERTATION

To provide background, in the remainder of this section we will consider a detailed treatment of the phenotypes of ROP (Section 1.4) and preterm birth (Section 1.5) with special emphasis on existing research into each condition's genetic basis for disease.

We will then discuss methods of analyzing genomic data, beginning with PRS (Section 1.6) and then continuing on to a short treatment of methods of Gene-Set Analysis (GSA) (Section 1.7) and protein-protein interaction data (Section 1.8) which form the basis of network-topology-based approaches (Section 1.9).

Following these introductory reviews, in Chapter 2 we will proceed to describe the support for a genetic basis for ROP rooted in previous studies investigating specific biochemical and protein components theorized to contribute to the development of advanced ROP. These investigations form the underpinnings for a candidate-gene approach which will be presented in Chapter 3 elucidating the degree to which ROP and preterm birth exist simultaneously as heritable predispositions toward disease.

In Chapter 3 we address Aim 1 by presenting network-based methods of constructing and evaluating PRS. We first consider the degree of general support that the inclusion of network context represents, and whether that additional information that can be exploited by PRS methods. We evaluate both an early network context approach using network information to shape the construction of PRS, which attempts to increase the predictive power of PRS over previously proposed methods, and a late network context approach taking network context into account after the construction of PRS, which seeks to provide biological context to existing formulations of PRS.

We will also address Aim 2 by applying these methods to a dataset including individuals enriched for preterm birth. We will attempt to discern the degree to which ROP candidate genes identified in Chapter 2 function as drivers of preterm birth, and we will introduce a novel approach to assess the ability of the developed network PRS methods to detect and describe this overlap of effect.

Chapter 4 will serve as a recapitulation of findings, as well as a discussion of new directions and a broad look at the implications of the dissertation work presented herein.

## **1.4 RETINOPATHY OF PREMATURITY**

Retinopathy of prematurity (ROP) is a disorder of the retinal vasculature affecting premature infants. Estimated to lead to blindness in more than 50,000 infants annually, it is a leading cause of lifelong blindness both in the United States and worldwide and represents an enormous social and economic burden.(6–10) While treatments for ROP are advancing, the ability of care facilities to better treat the complications of preterm birth has led to an increasing incidence of ROP over time.(6,11–13)

ROP has been suggested as a model for study not only because of its extreme burden posed to patients, but also due to features of the disease that make it uniquely suited to genetic studies of high impact. The course of disease completes in under a year in the first year of life, making it more tractable to follow the full course of disease in a short span of time. In addition, the manifestation of ROP involves dysfunction in angiogenic systems that have broad impact on diseases such as tumor progression in cancer, but offers a cleaner model of progression that can be tracked more quickly.

### **1.4.1 PRESENTATION OF ROP**

ROP is a complex phenotype described by several axes of disease. The most longstanding measures of ROP are those of disease progression (stage) and the physical extent of the retinal area affected by the disease (zone and clock-face hours). Later revisions of the standard for diagnosis added metrics describing venous tortuosity and dilation (Plus/Pre-Plus disease) and also added a designation of a faster progressing subtype of disease referred to as aggressive posterior ROP (AP-ROP).(14–17) Threshold disease is defined as ROP with a 50% or

greater likelihood of progressing to retinal detachment, and is described by any of the following conditions:

*Zone I ROP, any stage, with plus disease*

*Zone I ROP, stage 3, no plus disease*

*Zone II ROP, stage 2 or 3, with plus disease*

However, despite these guidelines, diagnosis of ROP ultimately relies on clinician judgment. The degree of training and expertise necessary for the diagnosis of ROP is considerable, and only a small subset of clinicians with substantial experience are able to effectively identify high-risk cases. Even among clinicians who are experts in ROP diagnosis, clinicians have been shown to differ in diagnosis of individual features of ROP, as well as overall disease severity.(1,16) While the ranking of disease severity cases by expert clinicians appears robust, the definition of a case that requires treatment may be substantially different. These differences have led to demand for quantitative measures of disease risk to supplement clinician diagnoses and serve as a quantitative baseline around which training can be structured.

#### **1.4.2 GENETIC CONTRIBUTIONS TO DISEASE**

Initially described as retrolental fibroplasia, early investigations into ROP were focused on the role played by oxygen exposure in premature infants as a contributor to development of disease.(18) These studies were successful in preventing progression in many infants by attenuating and monitoring oxygen levels shortly after birth.(19) But despite these advances, a subset of patients exists in which ROP develops despite these measures.(20)

Epidemiological studies into ROP progression suggest that genetic differences in predisposition to ROP progression exist between patients.(21–23) In a study of 409 premature infants under 1,251g of weight, the incidence in white infants was found to be roughly twice that of black infants.(23) Genetic heritability estimated from monozygotic and dizygotic twin studies was 0.70 and 0.73 respectively.(24,25)

Model organism studies have also provided supporting evidence that ROP progression involves a genetic component. Differences have been described between strains of rats in both the avascular area of disease and expression of RNA related to angiogenic factors.(26–28)

Individual inquiries have been made into many disparate genetic systems, primarily focused on a small subset of genes related to angiogenic function. Special attention has been made to describe the action of vascular endothelial growth factor (VEGF) and associated signaling proteins on progression.(29–37) Other studies have focused on the Wnt signaling pathway, which is known to be involved in Familial Exudative Vitreoretinopathy (FEVR), a disease with similar symptoms to ROP.(38–61) Insulin-like growth factor 1 (IGF-1) has also been proposed as a contributing factor to ROP progression.(60,62–71)

## **1.5 SPONTANEOUS PRETERM BIRTH**

Worldwide, preterm birth has risen from the second largest direct cause of death in children under five years of age after pneumonia in 2012, and is now the leading cause of death in children under five worldwide.(72,73) An estimated 14.9 million infants are born under 28 weeks of gestational age each year,(74) though evidence indicates that these rates have a slightly declining trend as a percentage of all births due to increased capability for intervention

worldwide.(75) However, as the raw number of births worldwide rises each year, it is likely that the number of preterm births will also increase.

### **1.5.1 RISK MEASURES OF PRETERM BIRTH**

Many of the current predictive measures for PTB focus on physical maternal traits known to be associated with preterm birth. Cervicovaginal fFN, fetal breathing movements, and transvaginal sonographic cervical length measurements are all physical attributes shown to be associated to varying degrees with preterm labor.(76–79) Higher BMI also seems to provide a protective effect.(80)

In addition to these morphological attributes, several other conditions are related to spontaneous preterm birth that may indicate a degree of genetic involvement with substantial error contributed by environmental factors. A woman with a sister who has given preterm birth is at 80% higher risk of delivering preterm herself.(81) Mothers who experience one early spontaneous preterm birth tend to repeat in later pregnancies.(82) A woman who has had a preterm birth is also more likely to have a grandmother who was born preterm than a woman who has not given birth preterm.(83)

Differences between racial backgrounds and preterm birth rates also exist. African-American and Afro-Caribbean individuals report a preterm birth rate of 16-18% compared to 5-9% for Caucasian mothers.(84) African-American mothers are also more than three times more likely to have a very early preterm birth than Caucasian mothers.(85)

Despite these observations, there are significant confounding effects from environment in each case, and it is difficult to discern whether environmental or genetic causes

predominate. More targeted studies performed with the intent of separating these effects by focusing on wholly genetic determinants have met with mixed results.(86)

### **1.5.2 GENETIC RISK OF PRETERM BIRTH**

Cohort studies of the genetic contribution to gestational age have estimated the narrow sense heritability of preterm birth at 13.3% and the broad-sense heritability at 24.5%.(87) More recent estimates are that as much as 40% of preterm birth outcomes are due to maternal genetic factors.(88)

Several studies have identified biomarkers of varying strength that indicate increased risk of preterm birth. Inflammatory agents in particular appear to have a relationship to PTB incidence. Serum metalloproteinase-9 levels display a rise in the hours before labor,(89) and levels of fetal fibronectin at 24 weeks gestation is also associated with increased risk of preterm labor.(90)

Candidate gene investigations have focused on inflammatory factors with putative function in PTB. A specific variant of TNF-alpha was found to interact with bacterial vaginosis to produce an elevated risk of spontaneous preterm birth.(91) IL6 was also investigated as a possible contributor and was found to impart a greater risk of PTB in black women.(92) Maternal smoking has also been implicated to interact with specific metabolic gene polymorphisms.(93)

Several genome-wide association studies have been conducted with the aim of identifying single nucleotide polymorphism (SNP) events associated with preterm birth outcomes. While these studies were able to detect weak associations between SNPs and preterm birth, they have generally been unable to detect single variants of strong effect,

though this is likely in part due to small sample size and the assumptions associated with the proposed model of disease.(94–98)

The most successful of these GWAS studies used a group of 979 cases and 985 controls and was able to identify a handful of SNPs with significance levels below  $1.4E-10$ , approaching genome-wide significance.(97) However the results of the replication cohort were unclear, and no gene achieved an OR of greater than 1.4, indicating that it is unlikely there are genes of large effect predisposing mothers to preterm birth.

Additional studies seeking to extend genome-wide association studies and incorporate additional data modalities using network and pathway models have met with some success. A pathway analysis of data acquired from a literature review of candidate gene studies was able to identify evidence of an autoimmune/hormonal regulation component contributing to preterm birth risk.(99) Combination of protein-protein interaction network data with tissue specific gene expression data has been used to identify candidate genes that may have functional importance.(100)

Studies investigating an epigenetic component of disease have also attempted to link environmental and genetic disease risk. Studies of cord blood in preterm and full term infants show different DNA methylation profiles.(101,102) Preliminary studies have also shown evidence that microRNA expression is associated with preterm birth.(103,104)

Taken together, these studies provide evidence that PTB is potentially based in genetic mechanisms of disease, though no single genetic factor has emerged as a strong candidate for involvement in the majority of PTB cases.



### **1.5.3 OUTLOOK FOR GENETIC STUDY**

Significant challenges exist when evaluating genetic contributions to preterm birth risk. There are large effects from confounding due to possible contribution from maternal and neonate genetic systems. This is further complicated by extensive documented environmental effects, and the fact that many of these environmental factors correlate strongly with socioeconomic background and educational background.(105,106)

Despite these challenges, there are indications that detection of genetic effects should be tractable with current data sets. While no single strong genetic determinant of preterm birth risk has been found to date, it seems likely that a sizeable genetic component is determined by genetic and epigenetic factors. The aforementioned early successes of integrating existing GWAS and candidate gene data using network and pathway approaches is also encouraging, showing that it is possible to find additional context using such methods. As price for genetic assays decreases, increased investigation into genetic and epigenetic factors over time is also likely to create new datasets that can provide additional context, making methods of integrating different modalities important. The study of preterm birth in particular is also likely to benefit from models that are able to quantify coincident genetic effects, as many other infant disorders are strongly correlated with gestational age.

### **1.6 POLYGENIC RISK SCORING**

Despite the widespread use of GWAS techniques, it has thus far proved difficult to find common causative genetic variants associated with large portions of genetic risk, especially in complex traits. This problem, known as the missing heritability problem, poses a major

challenge to large-scale genetic studies that may assay thousands or millions of individual single-nucleotide polymorphisms.(107) This problem has forced development of new techniques meant to interrogate larger subsets of variants in combination in order to capture a larger percentage of the genetic variance associated with a given phenotype in hopes of providing a more comprehensive analysis of complex traits.

Polygenic risk scores (PRS) are one proposed method of addressing the missing heritability problem. Originally developed for studies of complex traits like schizophrenia and bipolar disorder, the canonical PRS estimates the odds ratios associated with each SNP in a GWAS study using a logistic or linear regression model. A significance threshold is value is set, below which all SNPs with a more extreme significance are included. These SNPs are then summed either on their binary presence or as a weighted quantity using the calculated odds ratios. The resulting score can be evaluated for significance between case and control groups, or used as a variable for regression in the case of quantitative traits.(108) While many PRS computation methods exist, the predominant method of calculating scores involves weighted sums due to ease of interpretation and model parsimony.(109)

Investigation of PRS in schizophrenia found that scores calculated in a male discovery group were correlated with presence of schizophrenia in a target female group. Score alleles were also found to be significantly more abundant in target cases than controls, a trend which increased when more score alleles were added. Specificity was also evaluated against a number of other diseases with varying heritability and similarity to schizophrenia, and scores were found to reflect these differences.(108)

Further characterization of PRS performance demonstrated that ideal power could be achieved in cases where 95% of genes surveyed had no role in phenotype outcome, heritability was high, and roughly equally sized discovery and target groups were used. It was also demonstrated that while useful power is achievable for association testing between phenotypic groups, it is much more difficult to achieve the specificity and sensitivity necessary to create predictive measures without prohibitively high sample sizes.(110)

With the adoption of calculation tools into the popular genomic analysis package PLINK, PRS analysis has been applied to a variety of complex trait analyses including schizophrenia, BMI, height, and bipolar disorder.(108,111–113) The PRSice software package, dedicated specifically to PRS calculation, has also achieved substantial adoption and can be used to calculate scores and additional contextual information.(114)

Despite their relative popularity, the small amount of contextual information provided by PRS analysis limits their ability to elucidate the underlying mechanisms of disease. In addition, the global scope of an analysis makes identification of related subgroups of genetic elements difficult. These limitations provide opportunities for new methods with the ability to produce more nuanced results that describe the mechanisms of these interactions.

## **1.7 GENE SET ANALYSIS**

In the evaluation of GWAS data it is often advantageous to consider genomic regions as representative of larger entities with known or hypothesized function. Gene Set Analysis (GSA) methods incorporate context derived from annotation sources in order to provide additional

context to genomic data by grouping genomic regions as genes or other biologically significant elements and then jointly considering these elements as groups.

In general, gene sets exist as sets of prespecified gene entities. GSA data sources can be derived from various existing resources. Gene sets can be taken from experimentally validated data organized into biological pathways as in the case of Pathway Commons.<sup>(115)</sup> Gene sets may also be derived from network curated sets as in HPRD, or from hierarchically grouped ontological information as in the case of the Gene Ontology Database.<sup>(116,117)</sup> In other cases, gene sets may represent specific disease biomarkers known to be associated with a given condition.

Each of these approaches carries unique considerations. One example of a difficulty introduced by GSA is that while many gene sets may be catalogued, they are not necessarily mutually exclusive, and the quantification of such overlap can lead to difficulties in interpretation of results.

Also, unlike network analysis, gene sets must be curated before association testing. Due to this consideration, they are limited in their ability to detect undocumented novel interactions that may be of interest to investigators.

GSA methods for GWAS analysis can be grouped into two-step and one-step methods. In two-step methods, gene-level statistics are first calculated and then those gene-level statistics are aggregated at the level of all genes.<sup>(118)</sup> In the one-step case, all SNPs in a gene set are used to calculate a summary statistic without consideration of individual gene-level effects.

In one sense, PRS can be considered a one-step GSA method, incorporating all SNPs in a single gene-set that encompasses the full genome without regard for gene-level groupings. In the methods provided in this dissertation, we make use of two-step methods in the grouping of SNPs, but also introduce pure network context and pruning methods that go beyond the traditional application of GSA.

The insulated heat diffusion treatment of late context make use of a two-step approach using minimum p-values, but this method should be compatible with most two-step approaches that provide gene-level summary statistics.

In addition, though the PRS methods presented in this dissertation make use of GSA techniques, we present an application of these techniques specific to PRS construction with significant consideration for the needs of that particular use case.

## **1.8 PROTEIN-PROTEIN INTERACTION DATABASES**

Protein-Protein Interaction (PPI) data attempts to capture the relationship between proteins interacting in a biological system. These interactions can represent a wide array of phenomena, including coupling, degradation, or post-translational modification, with evidence derived from a wide variety of sources including immunoprecipitation assays, western-blot analysis, and genetic experimental evidence.

A variety of database resources has developed over time attempting to catalogue and consolidate the large amount of disparate PPI evidence. These resources run the gamut from strict experimental evidence to predicted interactions based on algorithmic inference, while

applying a variety of unique methods to prioritize and grade the degree of confidence placed in these interactions.

Databases may also be narrowly focused on one organism, as in the case of the Human Protein Reference Database,(119) which contains experimentally derived human-specific interaction data, or they may be extremely broad as in the case of GeneMANIA, which contains algorithmically predicted relationships across several model organisms in addition to humans.(120)

In this study we focus on the Search Tool for Retrieval of Interacting Genes/Proteins (STRING).(121) The STRING database provides both experimentally validated and algorithmically predicted protein interactions with a focus on providing a confidence score representative of the strength of evidence supporting that interaction.

Recent studies into the distribution of association signals from genetic variants have indicated that signal is scattered across the genome, with a large number of SNPs possessing a miniscule association potentially as the result of far-acting trans effects which are difficult to detect using traditional GWAS methods.(122,123) With this in mind, selection of a database resource with maximal coverage is a strong concern. STRING manages to provide that level of coverage while also providing a parameter for exclusion of spurious interactions, making it a well suited resource for investigations of polygenic risk.

## **1.9 NETWORK-BASED MODELING**

While polygenic risk scores provide a means of testing for the presence of a genetic effect, they provide little context for the underlying mechanism of disease. Network models provide

means of representing high-dimensional or complex data and allow for a variety of additional measures to be computed describing higher-order interactions.

Network-based approaches have been used as a flexible system for modeling a variety of biological processes from experimental data. Cell function has been modeled from protein-protein interaction, coexpression and transcriptome data.(124–126) Metabolic changes resulting from mutation have been modeled from differences in genetics between tumor subtypes.(127) Complex behavioral differences in animal strains have been mapped to networks of genetic data.(128) Networks serve as a flexible contextual tool for the analysis of natural systems.

Weighted networks model nodes as features and edges as weights denoting the strength of various interactions between those features, with a subset of highly connected nodes denoted as hubs and description of membership in various highly connected areas of the network known as modules. This convention has proven to have a great deal of flexibility in general modeling, and has proven applicability to various biological systems including interactions between genetic variants, cellular components, and expressed proteins.(129–131)

Many biological systems have been observed to adhere to a scale-free topology where the number of out edges from one node to others decreases exponentially at each degree.(132) This conceptualization is useful as it indicates a relatively small subset of network hubs exist with many connections to other nodes. Analysis of which nodes connect with one another, and their relative overlap with various hubs, allows for added context to be inferred about the role of these nodes assuming that membership in a module indicates a similar functional role.

A number of statistics have been developed in order to describe differences between networks.<sup>(131)</sup> Modules in genetic networks where nodes may represent individual alleles or genomic regions can be investigated for enrichment of many different functional relationships, including genes with similar ontology or pathway membership. These summary measures can be used to infer cellular or functional groups that may be disrupted by a given condition.

Overall network structure may also be compared between different samples. One measure of differences in network structure is to capture topological overlap of members of each module between groups. More advanced network operations may also be employed in order to find eigenvectors indicating centrality or importance of specific nodes between samples.

The treatment of networks considered in this dissertation is primarily interested in methods of community detection applied to large unweighted graph constructs. We make particular use of two methods of dividing the protein-protein interaction graph into smaller units in the late network context model of analysis.

The first of these methods is the Louvain algorithm described by Blondel et. al.<sup>(133)</sup> This method operates by first assigning every node to its own community, then iteratively considering the effect of adding each node from its community to the community of one of its neighbors. After each assignment, the gain in modularity is assessed and the node is placed into the community that maximizes that node's contribution to modularity.

The second specific method made use of in the late network context approach is HotNet2, an insulated heat diffusion approach for finding subnetworks enriched for high signal within biological networks.<sup>(134)</sup> HotNet2 operates by using a random walk with random restart



process. Nodes are first assigned heat values. Each time step, these heat values are then allowed to spread to neighboring nodes via a random walker in a proportion governed by the number of neighbors. Nodes then retain a fraction of their total heat  $\beta$  which can vary depending on graph topology or the specific analysis. A two-stage statistical test using a null derived from Monte-Carlo simulation is then performed to assess whether the number of detected subnetworks is greater than that expected by chance.

## CHAPTER 2: THE GENETICS OF RETINOPATHY OF PREMATURITY

### ABSTRACT

**TOPIC:** Retinopathy of prematurity (ROP) is a proliferative retinal vascular disease in premature infants, and is a major cause of childhood blindness worldwide. In addition to known clinical risk factors such as low birth weight and gestational age, there is a growing body of evidence supporting a genetic basis for ROP.

**CLINICAL RELEVANCE:** While comorbidities and environmental factors have been identified as contributing to ROP outcomes in premature infants, most notably gestational age and oxygen, respectively, a subset of infants progresses to severe disease despite an absence of these factors. The contribution of genetic factors may explain these differences and allow better early detection and treatment of premature infants at risk of severe ROP.

**METHODS:** To comprehensively review genetic factors that potentially contribute to the development and severity of ROP, we conducted a literature search focusing on the genetic basis for ROP. Terms related to other heritable retinal vascular diseases like “familial exudative vitreoretinopathy”, as well as to genes implicated in animal models of ROP, were also used to capture research in diseases with similar pathogenesis to ROP in humans with known genetic components.

**RESULTS:** Contributions across several genetic domains are described including vascular endothelial growth factor, the Wnt signaling pathway, insulin-like growth factor 1, inflammatory mediators, and brain-derived neurotrophic factor.

**CONCLUSIONS:** Most candidate gene studies of ROP have limitations such as inability to replicate results, conflicting results from various studies, small sample size, and differences in clinical characterization. Additional difficulty arises in separating the contribution of genetic factors like Wnt signaling to ROP and prematurity. Although studies have implicated involvement of multiple signaling pathways in ROP, the genetics of ROP have not been clearly elucidated. Next-generation sequencing and genome-wide association studies have potential to expand future understanding of underlying genetic risk factors and pathophysiology of ROP.

## 2.1 INTRODUCTION

Retinopathy of prematurity (ROP) is a retinal vascular disorder affecting premature low birth weight infants, and is a major cause of childhood blindness in the United States and internationally. Beyond the clinical impact, infancy-acquired visual loss from ROP represents an enormous social and economic burden.(7–10) Furthermore, as the incidence of premature births worldwide increases and as medical technology becomes better able to treat the complications of premature birth, the number of infants at risk for ROP is increasing rapidly.(6,11–13)

Oxygen plays a central role in ROP.(135–139) Oxygen environment and a key transcription factor that oxygen regulates (e.g. Hypoxia inducible factor [HIF]) are thought to modulate ROP. In terms of ROP pathogenesis, a two-phase hypothesis has been proposed and has become widely accepted.(140,141) In phase 1, there is delayed physiologic retinal vascular development and vasoattenuation, which is aggravated by hyperoxia and loss of nutrients and growth factors. In phase 2, vasoproliferation occurs at the junction of vascularized and avascular retina. Mouse oxygen-induced retinopathy (OIR) model (exposure to 75% oxygen for 5 days followed by room air), a widely used animal model of ROP, best represents the two-phase hypothesis.(142,143) During the vasoproliferative phase, the avascular retina releases pro-angiogenic growth factors such as vascular endothelial growth factor (VEGF), which are induced by hypoxia and may cause aberrant vessel growth and neovascularization. Oxygen fluctuations with intermittent hypoxia is also implicated in development of ROP in clinical studies(144–146) and OIR animal model studies especially in rats (e.g. cycling between 50 and 10% oxygen).(147,148) Growing neovascular vessels lead to fibrovascular membranes that may

pull on the retina, causing tractional retinal detachment and eventual blindness. The phenotype of ROP is classified based on location, extent, and severity of these pathologic changes.(149) Some infants show a rapidly progressing, severe form of ROP, known as aggressive posterior ROP (AP-ROP).(15,149–152)

Early investigations into ROP risk factors focused primarily on prematurity itself, as well as environmental factors including oxygen exposure after birth.(136,137) Various studies focusing on oxygen exposure have proven its importance as a primary predictor of ROP outcomes.(135–137) However, some high-risk infants with extremely low birth weight (BW) and gestational age (GA) do not develop ROP, whereas some low-risk infants do develop severe ROP. In these infants at phenotypic extremes, a study showed that known clinical risk factors were not significantly associated with development of ROP.(20) In addition, it is not understood why certain infants are predisposed to AP-ROP with very high likelihood of blindness. This heterogeneity of ROP risk suggests that other factors, such as genetics may be involved in creating a predisposition to ROP. Before specific genetic variations were investigated in ROP, epidemiologic studies suggested racial and ethnic differences in ROP incidence.(21–23) The Cryotherapy for ROP (CRYO-ROP) study of 4,099 premature infants found 7.4% of white infants reached threshold disease, while only 3.2% of black infants achieved a similar level of disease.(23) Also, twin and sibling studies have supported the involvement of a genetic component of disease. Two studies of monozygotic and dizygotic twins found that the heritability of ROP was 0.70 and 0.73, respectively.(24,25) Evidence of genetic effects is also supported by data from the oxygen-induced retinopathy (OIR) phenotype in rodent models, in which studies of different rat strains have found differences in the retinal avascular area and

VEGF expression between strains.(26–28) Investigations into this genetic component in humans and animal models have implicated the involvement of multiple genes, but have not discovered a genetic component of large effect. It is likely that knowledge of such a genetic component could be used to identify possible targets to improve outcomes of screening and treatment.

Many signaling molecules and related pathways have been suspected in the pathogenesis of ROP due to known biochemical and clinical associations: VEGF, insulin-like growth factor-1 (IGF-1), erythropoietin (EPO), and inflammatory mediators. In addition to ROP, the growth of abnormal, leaky blood vessels is a common pathologic component of other blinding neovascular eye diseases, such as diabetic retinopathy (DR) and neovascular age-related macular degeneration (AMD), both of which have strong evidence of a genetic predisposition to disease.(153–155) Moreover, because ROP progresses more rapidly and presents with relatively homogeneous clinical characteristics, the correlation of genotype and phenotype is easier than with a chronic disease such as DR or AMD.(141) Thus, the study of ROP genetics may give us important insights into the pathophysiology of other more prevalent adult and pediatric neovascular retinal diseases.

This review summarizes current research into genetic factors contributing to ROP risk in both human and animal models and recommends future directions for research into the underlying genetics of pathways that contribute to disease.

## 2.2 METHODS

Pubmed was queried from January 1980 to June 2017. The following search terms were used: retinopathy of prematurity AND genetics, retinopathy of prematurity AND gene, retinopathy of prematurity AND single nucleotide polymorphism (SNP), retinopathy of prematurity AND variant, and retinopathy of prematurity AND polymorphism. Criteria for inclusion included the relevance, clinical importance, level of statistical evidence provided, and scientific importance of articles to the subject of this paper. Articles cited in the reference lists of other articles were reviewed and included when considered appropriate. All articles with English abstracts were reviewed.

## 2.3 CANDIDATE GENES IN ROP

### 2.3.1 VEGF AND ASSOCIATED RECEPTORS

VEGF plays a crucial role in ROP. Increased VEGF in avascular retina stimulates pathological retinal neovascularization, which may result in blinding complications like tractional retinal detachment. Moreover, VEGF is a proven therapeutic target, as intravitreal anti-VEGF therapy has shown efficacy in promoting regression of severe ROP.(156) There have been many genetic studies on associations between the VEGF gene and incidence or severity of ROP.

**Table 1** summarizes results of SNP studies in human VEGF gene (*VEGFA*). rs2010963 (also known as -634G>C and +405 G>C) is the most extensively studied SNP. In a British study of

188 preterm infants on rs2010963 in 2004, the G allele was found to have higher frequency among infants with ROP.(29) This result was supported by a 2015 study in 102 preterm infants from Egyptian hospitals showing that G allele was significantly higher in infants with ROP.(30) However, one study in Hungary reported the opposite results – higher frequency of C allele in severe ROP – and 5 other studies found no significant association between rs2010963 and ROP. In addition, rs833061 (-460C>T) and *VEGFA* +13553C>T have been reported to be associated with ROP. However, replication has not been attempted for +13553C>T and the association of rs833061 and ROP has not been replicated in 3 other studies. *VEGFA* haplotypes have also been reported to be associated with ROP. A study performed in an Italian population of 342 infants focused on the distribution of polymorphisms in a handful of genes implicated in ROP showed evidence that *VEGFA* haplotype (TCCT) decreases risk of ROP.(31)

VEGF promotes angiogenesis and hyper-permeability by binding to the VEGF receptor 2 (VEGFR-2) on vascular endothelium, whereas VEGFR-1 acts as a decoy receptor.(157) However, studies on VEGFR-1 (*FLT1*) and -2 (*KDR*) genes found no associations with ROP (**Table 2**).



**Table 1. Studies Investigating the Association Between VEGFA Genes and Retinopathy of Prematurity (ROP)**

Table lists investigated polymorphism and presence of statistical significance. Where noted in the original study, information is provided in parentheses regarding the birth weight (BW) and gestational age (GA) of patients.

Brackets denote range of patient values and  $\pm$  denotes one standard deviation of range of each variable.

Polymorphism	Study country	Subjects	Results	Reference
rs2010963 (-634G>C, +405G>C)	United Kingdom	91 treatment-requiring ROP (BW 779g [440-1185g], GA 25 wk [23-32 wk]), 97 non-treatment-requiring preterm infants (BW 920 g [448-2302g], GA 26 wk $\pm$ 2.9 wk)	Higher frequency G allele among infants with threshold ROP	(29)
	Hungary	115 treatment-requiring ROP (BW 1160g $\pm$ 270g, GA 28.5 wk $\pm$ 2.0 wk), 86 mild or no ROP (BW 1200g $\pm$ 270g, GA 29.2 wk $\pm$ 2.9 wk)	Higher frequency C allele among treated infants	(34)
	Turkey	42 treatment-requiring ROP (BW 1097.5g $\pm$ 264.3g, GA 28.2 wk $\pm$ 2.4 wk), 50 regressed ROP (BW 1253.0g $\pm$ 212.2g, GA 29.7 wk $\pm$ 2.0 wk), 31 no ROP (BW 1345.6g $\pm$ 225.9g)	No significant association	(32)
	United States	61 stage 4/5 ROP (BW 882g [600-1300g], GA 26 wk [23-30 wk]), 61 normal controls (BW 2430-3960g, GA 34-40 wk)	No significant association	(158)
	Japan	127 ROP (944g [3778-2168g], GA 27 wk [22-33 wk]), 77 no ROP (BW 1596g [692-2400g], GA 32 wk [22-33 wk])	No significant association	(33)
	Egypt	62 ROP (BW 1400g [1000-2110g], GA 32 wk [28-34 wk]), 40 no ROP (BW 1640g [1009-2800g], GA 33 wk [29-35 wk])	High frequency of G allele in ROP	(30)
	Poland	60 treatment-requiring ROP (BW 900g $\pm$ 225g, GA 26.7 wk $\pm$ 2.3 wk), 20 regressed ROP (BW 1029g $\pm$ 231g, GA 27.5 wk $\pm$ 1.6 wk), 101 no ROP (BW 1153g $\pm$ 225g, GA 29.2 wk $\pm$ 2.05 wk)	No significant association	(35)
	Iran	15 treatment-requiring ROP (BW 879g $\pm$ 81g, GA 27 wk $\pm$ 13 wk), 30 regressed ROP (BW 884g $\pm$ 63g, GA 27 wk $\pm$ 12 wk), 66 no ROP (BW 980g $\pm$ 81g, GA 27 wk $\pm$ 10 wk)	No significant association	(37)
rs1547651	Caucasian	43 ROP, 299 no ROP (all subjects GA $\leq$ 28 weeks)	No significant association	(31)
rs3025039 (+936C>T)	Caucasian	43 ROP, 299 no ROP (all subjects GA $\leq$ 28 weeks)	No significant association	(31)
	Iran	15 treatment-requiring ROP (BW 879g $\pm$ 81g, GA 27 wk $\pm$ 13 wk), 30 regressed ROP (BW 884g $\pm$ 63g, GA 27 wk $\pm$ 12 wk), 66 no ROP (BW	No significant association	(37)

		980g $\pm$ 81g, GA 27 wk $\pm$ 10 wk)		
	United States	33 stage 4/5 ROP, 49 normal controls	No significant association	(36)
	Egypt	62 ROP (BW 1400g [1000-2110g], GA 32 wk [28-34 wk]), 40 no ROP (BW 1640g [1009-2800g], GA 33 wk [29-35 wk])	No significant association	(30)
rs833058	Italy	43 ROP, 299 no ROP (all subjects GA $\leq$ 28 weeks)	No significant association	(31)
rs833061 (-460C>T)	Italy	43 ROP, 299 no ROP (all subjects GA $\leq$ 28 weeks)	No significant association	(31)
	Hungary	115 treatment-requiring ROP (BW 1160g $\pm$ 270g, GA 28.5 wk $\pm$ 2.0 wk), 86 mild or no ROP (BW 1200g $\pm$ 270g, GA 29.2 wk $\pm$ 2.9 wk)	High frequency of 460TT/405CC haplotype in treatment-requiring ROP	(34)
	Turkey	42 treatment-requiring ROP (BW 1097.5g $\pm$ 270g, GA 28.2 wk $\pm$ 2.4 wk), 50 regressed ROP (BW 1253.0g $\pm$ 212.2g, GA 29.7 wk $\pm$ 2.0 wk), 31 no ROP (BW 1345.6g $\pm$ 225.9g)	No significant association	(32)
	United States	61 stage 4/ 5 ROP (BW 882g [600-1300g], GA 26 wk [23-30 wk]), 61 normal controls (BW 2430-3960g, GA 34-40 wk)	No significant association	(158)
+13553C>T	Japanese	127 ROP (BW 944g [378-2168g], GA 27 wk [22-33 wk]), 77 no ROP (BW 1596g [692-2400g], GA 32 wk [22-33 wk])	A significant association between the TT genotype and non-severe ROP for gestational age	(33)
+702C>T	United States	33 stage 4/5 ROP, 49 normal controls	No significant association	(36)
+1612G>A	United States	33 stage 4/5 ROP, 49 normal controls	No significant association	(36)
-2578C>A	United States	ROP (BW 2430-3960g, GA 34-40 wk), no ROP (BW 600-1300g, GA 23-30 wk) (number of patients not reported)	No significant association	(159)
	Hungary	90 treatment-requiring ROP (BW 1160g $\pm$ 300g, GA 28.5 wk $\pm$ 2.4 wk), 110 mild (stage 1 or 2) or no ROP (BW 1200g $\pm$ 280g, GA 28.5 wk $\pm$ 2.4 wk)	No significant association	(160)

**Table 2. Summary of candidate gene studies of retinopathy of prematurity other than *VEGFA***

Gene	Variant	Study country	Subjects	Results	Reference
ACE	rs1799752	Italy	299 ROP, 43 no ROP	No significant association	(31)
	rs4291				
	287-bp insertion in intron 16	Kuwait	74 ROP (53 regressed, 21 stage 4/5 ROP), 107 no ROP	The incidence of the II genotype was higher in ROP cases, while the incidence of the DD genotype was significantly higher in advanced stage ROP cases compared to spontaneously regressing ROP cases. (I, insertion, D, deletion)	(161)
AGT	rs699	Italy	43 ROP, 299 no ROP	No significant association	(31)
AGTR1	rs5186	Italy	43 ROP, 299 no ROP	No significant association	(31)
	rs427832	United States	102 ROP, 228 no ROP	Significant association with ROP at $p < 0.01$ level of significance	(162)
ANGPT2	-35G>C	United States	Not specified	No significant association	(159)
		Hungary	90 treatment-requiring ROP, 110 no or mild (stage 1 or 2) ROP	No significant association	(160)
BDNF	rs7934165 rs2049046	United States	126 treatment-requiring ROP, 467 stage 1/2 ROP	Two intronic SNPs found to be associated with difference between mild and threshold ROP	(163)
	rs7934165	United States	140 treatment-requiring ROP, 1257 no or mild (stage 1 or 2) ROP	Meta-analysis of two studies provided evidence of association of variant with severe ROP	(163)
CETP	rs289747	United States	102 ROP, 228 no ROP	Significant association with ROP at $p < 0.01$ level	(162)
CFH	rs52985	United States	102 ROP, 228 no ROP	Increased protection against ROP as number of T alleles increased ( $p = 0.01$ )	(162)
	rs800292	United States	102 ROP, 228 no ROP	Increased protection against ROP as number of T alleles increased ( $p = 0.01$ )	(162)
EPAS1	rs1867785	United States	102 ROP, 228 no ROP	Significantly higher incidence of A allele in ROP	(162)

<i>GP1BA</i>	rs2243093	United States	102 ROP, 228 no ROP	Significant association with ROP at $p < 0.01$ level	(162)
<i>LRP5</i>	rs143924910 (c.3656G>A), c.4148A>C, rs141407040 (c.4619C>T)	Japan	53 advanced ROP	Direct sequencing of coding regions of LRP5 revealed 3 nonsynonymous DNA variants in 3 patients.	(60)
	3-bp insertion in exon 1	Japan	17 advanced ROP, 51 no ROP	Single patient with advanced ROP shown to have 3 bp insertion in exon 1 CTG repeat area not observed in 28 unaffected patients.	(47)
<i>NOS3</i>	rs2070744 (-786T>C)	Italy	43 ROP, 299 no ROP	No significant association	(31)
		Hungary	105 treatment-requiring ROP, 127 stage 1 or 2 ROP	No significant association	(164)
		United States	15 ROP, 131 no ROP	significantly higher frequency of C allele in ROP	(165)
		United States	19 stage 4/5 ROP, 34 normal	significantly higher frequency of C allele in ROP	(70)
	rs1799983 (894G>T)	United States	14 stage 4/5 ROP, 32 normal	No significant association	(70)
		United States	15 ROP, 131 no ROP	significantly higher frequency of T allele in ROP	(165)
		Italy	43 ROP, 299 no ROP	No significant association	(31)
	27-bp VNTR in intron 4 (b/a)	United States	15 stage 4/5, 32 normal controls	No significant association	(70)
		Hungary	105 treatment-requiring ROP, 127 stage 1 or 2 ROP	The aa genotype presented an independent risk factor for ROP requiring treatment.	(164)
	rs61722009	Italy	43 ROP, 299 no ROP	No significant association	(31)
<i>FLT1</i>	c.+6724(TG) 13-23 dinucleotide repeat	Japan	127 ROP, 77 no ROP	No significant association	(33)
<i>FZD4</i>	c.97 C>T; c.502 C>T (double missense mutation)	United States	93 ROP, 98 normal controls	Seven of 93 (7.5%) patients with ROP showed c.97 C>T; c.502 C>T double missense mutation.	(61)

	rs80358282 (c.205C>T), rs184709254 (c.380G>A), c.631T>C	Japan	53 advanced ROP	Direct sequencing of coding regions of <i>FZD4</i> revealed 3 nonsynonymous DNA variants in 4 patients.	(60)
	c.766A>G	Unspecified	10 sporadic FEVR cases 20 advanced ROP cases	PCR amplification of a large DNA fragment revealed one severe ROP case with c.766A>G. Significance not investigated.	(42)
	c.1109C>G, c.609G>T	Canada	71 severe ROP, 33 mild or no ROP	Direct sequencing of coding regions of <i>FZD4</i> revealed 2 nonsynonymous DNA variants in 2 patients.	(43)
<i>HMOX1</i>	rs3074372	Italy	43 ROP, 299 no ROP	No significant association	(31)
<i>IGF1R</i>	c.3174G>A	United States	52 stage 4/5 ROP, 33 normal controls	No significant association	(166)
		Hungary	108 treatment-requiring ROP, 120 stage 1 or 2 ROP, 164 normal controls	No significant association	(69)
<i>IHH</i>	rs3099	United States	102 ROP, 228 no ROP	Significant association with ROP at p < 0.01 level	(162)
<i>IL10</i>	-1082G>A	Germany	31 stage 1 or 2 ROP, 13 stage 3 ROP, 29 no ROP	No significant association	(71)
<i>IL1B</i>	+3953C>T	Germany	31 stage 1 or 2 ROP, 13 stage 3 ROP, 29 no ROP	No significant association	(71)
<i>KDR</i>	32G>A	Turkey	42 treatment-requiring ROP, 50 regressed ROP, 31 normal controls	No significant association	(32)
	g.+4422(AC)11-14 dinucleotide repeat	Japan	127 ROP, 77 no ROP	No significant association	(33)
<i>NDP</i>	Sequencing of all 3 exons and UTRs	United States	54 severe ROP, 36 mild or no ROP, 22 normal controls, 31 normal parents	Six of 54 (11%) infants with severe ROP had polymorphisms in the NDP.	(53)
	Direct sequencing of coding regions and noncoding exon 1	Japan	53 advanced ROP	No meaningful sequence changes	(60)

	237A>G	Japan	17 advanced ROP, 51 no ROP	Single patient with AP-ROP found to have heterozygous substitution not observed in 51 unaffected cases	(47)
	14 bp deletion in exon 1	Australia	31 ROP (Stage 2 or greater), 90 no ROP	Two twins with stage 3 regressed ROP and one unrelated patient with regressed stage 2 ROP displayed 14 bp deletion in CT repeat region. Also observed in a control patient. No statistical analysis.	(48)
	5 bp deletion in exon 1 26C>G 71 bp deletion in exon 1	UK	31 ROP stage 3 or more, 16 regressed ROP, 2 no ROP	One patient had 5 bp deletion and C>G transversion at +26, one patient had 71 bp deletion in same exon 1 region. No statistical analysis.	(51)
	12 bp insertion in exon 1 14 bp deletion in exon 1	Japan	100 advanced ROP (stage 4/5), 6 regressed stage 3 ROP, 130 no ROP	Two advanced ROP patients found to have disruptions in exon 1 of ND gene. No statistical analysis.	(50)
	597C>A 110C>G	Kuwait	95 ROP, 115 no ROP	Significant association was found between ROP and 597C>A polymorphism. No significance found between 110C>G polymorphism and ROP.	(52)
	121C>T R121W L108P	United States	16 ROP, 50 normal controls	One patient with a heterozygous base substitution, one pair of twins with novel R121W mutation, and one pair of twins with L108P missense mutation observed. No statistical analysis.	(49)
<i>TBX5</i>	rs1895602	United States	102 ROP, 228 no ROP	Significant association with ROP at p < 0.01 level	(162)
<i>TGFB1</i>	-509C>T	United Kingdom	91 treatment-requiring ROP, 97 stage 1/2 or no ROP	No significant association	(29)
<i>TLR4</i>	rs4986790 (c.896A>G)	Germany	31 stage 1 or 2 ROP, 13 stage 3 ROP, 29 no ROP	No significant association	(71)
<i>TNF</i>	-308G>A	United Kingdom	91 treatment-requiring ROP and 97 stage 1/2 or no ROP	No significant association	(29)
		Germany	31 stage 1 or 2 ROP, 13 stage 3 ROP, 29 no ROP	No significant association	(71)

<i>TSPAN12</i>	Direct sequencing of coding regions of <i>TSPAN12</i>	Japan	53 advanced ROP	No meaningful sequence changes	(60)
----------------	---	-------	-----------------	--------------------------------	------

### 2.3.2 FEVR, NORRIE DISEASE AND THE WNT PATHWAY

Familial Exudative Vitreoretinopathy (FEVR) and Norrie disease are developmental diseases of the retina with known genetic causes with similar pathology to ROP. FEVR and Norrie disease are both hereditary disorders occurring in full-term infants, characterized by failure of peripheral retinal vascularization leading to retinal detachment.(40,167) While Norrie disease progresses quickly in early childhood and is accompanied by additional pathologies like deafness and irregular mental development, FEVR may not progress to retinal detachment until patients reach adulthood and is restricted to abnormalities in ocular development.(168) FEVR is known to be caused by mutations in *FZD4*, *LRP5*, *TSPAN12*, *NDP*, etc.(41,45,46,55) and Norrie disease is caused by mutations in *NDP* gene.(167) These genes encode proteins which are components of the Wnt/beta-catenin signaling pathway – a group of signal transduction pathways that play roles in cell survival, proliferation, and migration throughout the body.

The canonical (beta-catenin dependent) Wnt pathway has known roles in a variety of diseases with angiogenic properties including DR and AMD.(38,39) Frizzled-4 and low-density-lipoprotein receptor related protein 5 (LRP-5) are receptors for Wnt ligands, and tetraspanin-12 is an auxiliary membrane protein. Norrin, a product of *NDP* gene, binds to the Frizzled-4, LRP-5, and tetraspanin-12 receptor complex and activates signals on endothelial cells. Mutations of these genes have been investigated in ROP (**Table 2**).

Mutations in the *FZD4* gene were found in up to 7.5% of patients with severe ROP (**Table 2**).(42,43,60,61) A 2015 study of 421 patients displaying various vitreoretinopathies found a significant association between the *FZD4* double missense mutation [P33S(;)P168S] and both ROP and FEVR.(61) A study of 53 Japanese patients with advanced ROP was performed



using direct sequencing of *FZD4*, *TSPAN12*, *NDP*, and *LRP5*. Investigators identified six nonsynonymous DNA variants in the coding regions of *FZD4* and *LRP5*, but detected no changes in *NDP* or *TSPAN12*, demonstrating involvement of Wnt with ROP.(60)

Mutations in the *NDP* gene have also been found in ROP patients with variable frequencies (**Table 2**). (49,50,52) SNP studies in Kuwaiti populations have supported evidence of a link between *NDP* and ROP,(52) while other studies have implied that mutations in the regulatory region of *NDP* are also a contributor to the development of ROP.(51) The relationship between SNPs residing in the UTR of *NDP* and progression of ROP to advanced disease has also been investigated. The Kuwaiti study by Haider found that 83% of patients with severe disease possessed *NDP* 597C>A polymorphisms in their UTR, while none of those whose disease resolved spontaneously possessed this polymorphism.(52)

Taken together, these findings intriguingly suggest involvement for the Wnt pathway and associated genes in ROP development, and serve as strong candidates for further sequencing research. It should be noted that it may be difficult or nearly impossible to differentiate ROP from FEVR in premature infants (which has recently been proposed as a new classification, ROPER [ROP vs. FEVR]) due to the clinical similarity of the two conditions.(169) In future studies, in-depth analysis of clinical features, retinal imaging with fluorescein angiography, genetic and phenotypic analysis of relatives, and functional analysis of genetic variants may be helpful for better understanding of genetics in ROP as well as FEVR.

### 2.3.3 IGF-1

Insulin-like growth factor 1 (IGF-1), a growth hormone promoting somatic growth and maturation, has also been proposed as a contributing factor to ROP progression.(62) IGF-1-deficient mice showed a decrease in vascular development(51) and lower birth weight(170) than those of controls. In human babies, low IGF-1 levels were also associated with low birth weight,(171) and persistent low serum IGF-1 levels were associated with severity of ROP.(62,172) Based on these findings, IGF-1 replacement therapy has recently been investigated.(173) A phase 2 trial of administering a complex of recombinant human IGF-1 and IGFBP-3 to prevent ROP was undertaken, but the study did not meet its primary endpoint of reducing severity of ROP.(174)

Investigations of specific polymorphisms of *IGF-1* gene have been unsuccessful finding a significant association. A study linked a c.3174G>A polymorphism in the IGF-1 receptor gene (*IGF1R*) to low levels of plasma IGF-1.(68) A 2006 study of 392 infants in Hungary was unable to detect a difference in the prevalence of the *IGF1R* c.3174G>A among severe ROP, mild ROP and full-term groups (**Table 2**).(69) A 2007 study in an American population was also unable to find a link between advanced ROP and *IGF1R* c.3174G>A polymorphism (**Table 2**).(166)

### 2.3.4 eNOS

Endothelial nitric oxide synthase (eNOS) is one of the constitutive enzymes that synthesize NO, which is known to play a regulatory role in retinal and choroidal blood flow.(175,176) In an eNOS-deficient mouse OIR model, neovascularization and vaso-obliteration were both reduced.(177) Moreover, eNOS gene polymorphisms have shown reduced NO

levels.(178) Thus, the association between ROP and eNOS gene (*NOS3*) polymorphisms have been investigated. A literature search showed that 3 SNPs (rs2070744, rs1799983 and rs61722009) and one variable number tandem repeat (VNTR), 27-bp VNTR in intron 4, had been observed in ROP patients (**Table 2**). Although some studies reported positive associations between rs2070744, rs1799983, or the 27-bp VNTR and ROP, others found contradictory results (**Table 2**).

### 2.3.5 INFLAMMATORY MEDIATORS

Growing evidence suggests that perinatal inflammation and infection may increase the risk for ROP by direct proangiogenic effects and/or modifying known risk factors.(71) Studies have reported higher plasma levels of inflammatory cytokines including IL-6, IL-8, and TNF(179) and higher vitreous levels of inflammatory cytokines including IL-6, IL-7, IL-10, IL-15, etc. in eyes with advanced ROP.(180)

Dammann et al investigated 4 SNPs of inflammation-associated genes (IL1B, TNF, IL10, TLR4) in preterm patients, but none showed significant association, although there were trends towards higher stage of ROP with the presence of *TNF* and *IL1B* SNPs (**Table 2**).(71) *TNF* - 308G>A polymorphism also showed no significant associations with ROP (**Table 2**).

A recent study has also shown an angiogenic role for mast cells and associated factors including mast cell tryptase and monocyte chemotactic protein-1, making them a potential target for ROP research.(181)

### 2.3.6 BRAIN-DERIVED NEUROTROPHIC FACTOR

Brain-derived neurotrophic factor (BDNF), a neuronal trophic factor in brain and retina, may promote survival of several types of retinal neurons.(182–185) Although the exact role of BDNF in retinal angiogenesis is unknown, reduced BDNF levels have been demonstrated in patients with severe ROP, suggesting a possible role of BDNF in development of severe ROP.(186–188) In an animal model study, the retinal level of BDNF was lower in the OIR mouse model compared to that in normal controls.(188)

In a large-scale candidate gene study, which analyzed 1614 Tag SNPs of the 145 candidate genes in 817 infants in the discovery cohort and 543 in the US replication cohort, it was found that two SNPs (rs7934165 and rs2049046) in the intronic region of *BDNF* were associated with severe ROP. Although these results were not independently confirmed in the replication cohort, the association with rs7934165 did increase in significance with severe ROP in their meta-analysis of the combined data. Interestingly, reduced serum BDNF in the severe ROP group was also found in the same discovery cohort.(189) Further studies on the functional effects of intronic variants of *BDNF* and replication studies in different populations are warranted.

### 2.3.7 RENIN-ANGIOTENSIN SYSTEM

The Renin Angiotensin system (RAS) has been linked to retinal vascular development and pathological angiogenesis. Blockade of RAS with inhibitors of angiotensin-converting enzyme (ACE) and angiotensin receptor blockers ameliorated OIR, suggesting that inhibiting RAS may be beneficial in ROP.(190) A SNP study of *ACE* gene showed association with DR.(189)

However, results from genetic studies on RAS component genes in ROP are inconclusive (**Table 2**). A study in Italy showed no associations between ROP and SNPs of ACE gene (*ACE*), angiotensinogen gene (*AGT*) and angiotensinogen type 1 receptor gene (*AGTR1*). In a study of 181 premature Kuwaiti infants on 287-bp insertion(I)/deletion(D) in intron 16, the frequency of II genotype was higher in ROP patients compared to normal controls, but the frequency of DD genotype was higher in advanced ROP patients compared to regressed ROP.(161) A candidate gene study of 228 infants with ROP and 102 controls found a SNP in the *AGTR1* gene to be associated with ROP, though this association was not significant after Bonferroni correction.(162)

### **2.3.8 ANGIOPOIETINS**

Angiopoietin(Ang)-1 and -2 are growth factors that are essential for retinal vascular development. Ang-1 binds tyrosine kinase receptor Tie2 and promotes vascular maturation and stabilization.(191) In an OIR model, intravitreal Ang-1 promoted normal vascular regeneration while inhibiting pathological angiogenesis and vascular leakage.(192) In contrast, Ang-2, a competitive antagonist of Ang-1/Tie-2, promotes neovascularization in animal models.(193,194) Vitreous levels of Ang-1 and Ang-2 in eyes of stage 4 ROP were higher than those of control eyes.(195) However, in two studies of Ang-2 gene promoter polymorphism (*ANGPT2* -35G>C), no association was found with ROP (**Table 2**).

### **2.3.9 ERYTHROPOIETIN**

Erythropoietin (EPO), a hormone known to stimulate red blood cell formation in bone marrow, and EPO receptors are expressed in retina, and their expression is regulated by oxygen

status.(196,197) Mouse models of ROP have shown that vascular stability is affected by EPO levels, with exogenous restoration of EPO leading to a reduction in blood vessel dropout during the first phase of ROP.(195) Conversely, elevated levels of EPO during the second stage of ROP exacerbated vasoproliferation, and the vitreous level of EPO is elevated in eyes with stage 4 ROP. Increased erythropoietin receptor signaling has also been shown to influence severe OIR models of disease through VEGFR2-mediated angiogenesis, making it an important target for clinical research in human patients.(198,199)

While a variant of EPO was investigated in a candidate-gene study by Mohamed et. al., significance for this variant was not reported in the study results.(162)

### **2.3.10 HYPOXIA INDUCIBLE FACTOR**

HIF-1 plays a central role in oxygen homeostasis.(200) According to the oxygen environment, HIF-1 regulates transcription of genes such as VEGF, VEGFR1, PDGF, SDF-1 and Ang2, which have been suggested to play important roles in retinal angiogenesis.(193) In a study of Hif1 $\alpha$  knockout mice in an OIR model of disease, disruption of HIF-1 was shown to lead to decreased VEGF abundance, indicating a possible role in neovascularization.(201) Additionally, organ system pharmacology studies in mouse models have indicated that stabilization of HIF-1 may be important for protection against oxygen toxicity in premature infants.(202)

Likewise, homologous recombination models in mice studying HIF-1 $\alpha$ -like factor (HLG) and HIF2 $\alpha$  found decreasing expression of these genes led to decreased EPO expression and resistance to hyperoxia treatments meant to induce ROP.(203) HIF1 $\alpha$  was also shown to

upregulate annexin A2 expression in OIR mice during hypoxia, supporting a role in OIR models.(204)

HIF2 $\alpha$ 's closest human analogue, known as Endothelial PAS Domain Protein 1 (EPAS1), serves as the main regulator of EPO induction and has also been shown to have a connection to ROP.(205) A candidate gene study of 153 genes in 347 infants under 32 weeks gestational age found an association between EPAS1 with development of severe ROP.(162)

### **2.3.11 HEME-OXYGENASE-1**

Heme oxygenase-1 plays important roles in inflammatory responses, oxidative stress, iron-metabolism, and vascular physiology. However, in a candidate gene study, rs3074372 in *HMOX1* showed no significant association with ROP (**Table 2**).

## 2.4 OTHER CANDIDATE FACTORS

In addition to the above described factors and pathways, a number of other potential targets and mechanisms have been identified that lack genetic studies in patients with ROP. The 'a' disintegrin and metalloproteinase (ADAM) family of proteases are involved in the degradation of extracellular matrix components as well as interactions mediated by integrin.(206) Several subtypes of ADAM family are implicated in the pathogenesis of ROP. ADAM17 knockout mice showed less neovascularization in OIR models without affecting normal vascular development.(207) Moreover, ADAM 8, 9, and 10 was found to play a role in development of plus disease in OIR mouse models. Adam8<sup>-/-</sup> and Adam9<sup>-/-</sup> mice and mice lacking ADAM10 in endothelial cells showed less severe tortuosity and dilation mimicking less plus disease in ROP.(208) Further evaluations in humans including genetic analysis are warranted.

In conjunction with ADAM17, studies have also considered the family of tissue inhibitor of metalloproteinases (TIMP) family of proteins. The TIMP-3 protein specifically is a known physiological ADAM17 inhibitor.(209) Mouse model investigations into the application of this protein as a potential treatment showed that TIMP-3 application was linked to decreased neovascular tuft formation.(208)

In addition to these studies, large candidate gene studies of ROP have been successful identifying targets with undiscovered connections to ROP. The previously mentioned study by Mohamed et al. implicated genes with function in embryonic development (*IHH*), transcription (*TBX5*), and protein localization (*GP1BA*, *CETP*) (**Table 2**).(162) The same study also found an



association between ROP and complement factor H (*CFH*), known to be associated with development of AMD.(154)

## 2.5 DISCUSSION

### 2.5.1 SUMMARY OF PREVIOUS STUDIES

Most genetic studies in ROP have used the candidate gene approach and focused on genes related to angiogenesis, inflammation, and retinal (neuro)development. Among them, *VEGFA* polymorphisms and FEVR-related genes have been most extensively studied in different populations. However, no *VEGFA* polymorphisms have been proven to be associated with ROP, because most positive studies have not been replicated in other populations (**Table 1**). Variants of Wnt pathway genes, which are known to cause FEVR or Norrie disease, have been also found in ROP patients, suggesting possible associations of these variants in at least a small proportion of severe ROP patients (**Table 2**). However, these results also have limitations in that we may not confidently distinguish between premature infants with severe ROP and FEVR-related genetic variants and prematurely-born infants with FEVR, as Hartnett et al. pointed out.(152) In addition the polygenic nature of many diseases makes identification of causative variants difficult in small sample sizes focused on a small number of variants.(107) Recently, results of a large-scale candidate gene study using Tag SNPs of the 145 candidate genes in a multiracial cohort were reported.(189) Although no SNPs were significantly associated with the presence versus absence of ROP in this study, one SNP of *BDNF* gene was significantly associated with severe ROP in their meta-analysis combining the discovery and replication cohorts, which warrants further genetic and biological studies.

### 2.5.2 LIMITATIONS OF PREVIOUS STUDIES

It is difficult to draw meaningful conclusions from most of the candidate gene studies reviewed here due to the following limitations: (1) the sample sizes of most individual studies were small; (2) no replication study has been performed for many variants; (3) there are conflicting results among studies of the same variants; (4) most studies were conducted using only one or a few clinical sites; (5) ocular phenotype was not standardized; (6) confounding variables were not reported or standardized; (7) meta-analysis is not possible for most variants due to different study protocols between studies; (8) there are variabilities in neonatal care such as oxygen treatment protocol(135), incidence of (severe) ROP, and diagnosis and management of ROP between physicians, study hospitals, study countries and study periods.(6,210–213) Differences in neonatal care may affect survival rate, systemic morbidities of prematurity, incidence of ROP and severity of ROP, making it difficult to find exact roles of genetic variants. Moreover, there are unexplained differences in outcome of premature birth such as mortality. Also, differences in diagnosis and management of ROP may cause bias in phenotypic categorization of subjects, which is a huge problem in genetic studies. It should be noted that genetic risk factors for stage 1-3 ROP and stage 4 or 5 ROP could be different, as different biochemical processes may be involved and management protocols and treatment outcomes of study centers are also important factors for stage 4 or 5 ROP.

Most importantly, candidate gene studies have inherent limitations of not being able to find novel genetic factors. Other approaches to detect novel variants or genes associated with ROP are necessary.

### 2.5.3 FUTURE DIRECTIONS OF STUDYING ROP GENETICS

It is very challenging to study the genetics of multifactorial diseases such as ROP. To overcome the current limitations mentioned above and to study the contribution of genetics efficiently, it is necessary to improve the methodology for studying the genetics of ROP. It is essential that investigators leverage new methods that interrogate genetic factors agnostically and at high sample sizes, in order to maximize study power and facilitate simultaneous investigation of many, rather than single, genetic elements. Genome-wide Association Studies (GWAS) test for association across hundreds of thousands of SNPs simultaneously using array-based technology. GWAS can be helpful to find genes or pathways associated with ROP. In other ophthalmological diseases such as AMD(154,214–216), DR(217,218), glaucoma(219–221) and myopia(222–224), GWAS has been successful in finding susceptibility loci. However, a large-scale GWAS has not been conducted in ROP. Massively parallel sequencing, also called next-generation sequencing (NGS), enables sequencing of specific regions, whole exome, or whole genome in a short period of time at high depth and affordable cost. Whole exome sequencing or targeted exome sequencing can be helpful for finding novel variants with possible functional consequences. Exome genotyping arrays may also provide a method of interrogating for SNPs involved in ROP.

In addition to these genetic evaluations, integration of sequence data with data regarding post-transcriptional and post-translational modification, including transcriptomics, metabolomics, and proteomics, will be important to identify biomarkers that may be useful for early detection, diagnosis, and prediction of treatment response. Studies of epigenetics in DR have also shown promise, with epigenetic changes associated with processes of microvascular

complications(225), mitochondrial dysfunction(226), microRNA expression(227), and capillary cell apoptosis.(228,229) These findings suggest that interrogation of epigenetic factors may be an important method of discovering new treatments in ROP.

Second, large-scale multi-center collaboration of the type offered by consortium studies can help provide structure to such studies. Consortium approaches facilitate recruitment of larger cohorts and make available more sophisticated computational approaches allowing investigators to control for more complicated confounding effects. Previous large international consortium attempts at examining the role of genetics in multifactorial disease have met with success(154,219,230–232), and two consortium studies investigating the genetic causes of ROP are currently ongoing at centers in North America.(233,234)

Third, standardization of ocular phenotypes and confounding factors is crucial. For this, ocular and systemic factors should be acquired systematically, and known risk factors including GA and BW should be assessed in a standardized fashion and strictly controlled for. Additionally, the importance of environmental effects should be noted, as differences between study populations and sites has the ability to have a profound effect on phenotype. Heterogeneity of study subjects in race, ethnicity, and physical covariates, as well as differences between treatment sites and attending clinicians can affect study outcomes. This is especially important to distinguish genetic variants associated with ROP from those associated with prematurity itself. Also, objective phenotyping such as image-based diagnosis should be considered. Compared to clinical ophthalmoscopic diagnosis, consensus image-based diagnosis may enable reduction of intra- and inter-grader discrepancy in ROP diagnosis.

It is also important to note that additional basic research studies using representative animal models such as mouse or rat OIR model are required to test hypotheses. While animal models face many limitations including differences in biology, most notably their use of full-term rather than premature animals, these models' ability to control for phenotypic, environmental, and genetic stratification factors distinguishes them as a valuable method of testing hypotheses and adding insight to human observational studies.

#### **2.5.4 EXPECTED BENEFITS OF GENETIC STUDIES OF ROP**

Finding genetic variants affecting ROP will be useful in at least three ways. First, genetic risk factors may be incorporated into risk modelling to predict development and progression of ROP. A refined risk analysis system with clinical and genetic risk factors may help clinicians to identify both high- and low-risk patients. Second, identifying specific genes or biological pathways that contribute to the pathogenesis of ROP may be helpful for development of new therapeutics. In AMD, genetic studies have revealed the importance of complement pathway in the pathogenesis of AMD, which has led to development of new investigational agents under clinical trials such as lampalizumab, an inhibitor of complement factor D. Third, studying ROP genetics can also contribute to the understanding of pathophysiologies of other ocular vascular diseases such as AMD or DR and other angiogenesis-related diseases like cancer.<sup>15</sup> Fourth, a better understanding of the genetics of retinopathy of prematurity may lead to better understanding of the pathophysiologic mechanisms of common neonatal diseases of prematurity such as chronic lung disease.

### **2.5.5 CONCLUSIONS**

Evidence suggests a genetic contribution to ROP, including epidemiologic studies, twin studies and risk analysis studies. To date, a number of candidate gene studies have been performed. However, it is still unclear which genes or variants are significantly and strongly associated with development and progression of ROP. Large-scale studies using NGS and GWAS with standardized phenotyping have potential to expand understanding of genetic contributions and pathophysiology of ROP.

## 2.6 Author Contributions

This chapter originally published as ‘The genetics of retinopathy of prematurity: a model for neovascular retinal disease’, by Ryan Swan, Sang Jin Kim, MD, PhD,<sup>2,3</sup> J. Peter Campbell, MD, MPH,<sup>2</sup> R. V. Paul Chan, MD,<sup>4,5</sup> Kemal Sonmez, PhD,<sup>6</sup> Kent D. Taylor, PhD,<sup>7</sup> Xiaohui Li, MD,<sup>7</sup> Yii-Der Ida Chen, PhD,<sup>7</sup> Jerome I. Rotter, MD,<sup>7</sup> Charles Simmons, MD,<sup>8</sup> Michael F. Chiang, MD<sup>1,2</sup>, and the Imaging and Informatics in ROP Research Consortium in Ophthalmology Retina on Mar 8, 2018. DOI 10.1016/j.oret.2018.01.016.



## CHAPTER 3: A NETWORK MODEL OF POLYGENIC RISK

### 3.1 ABSTRACT

**TOPIC:** Polygenic risk scoring (PRS) is a proven method of genetic disease prediction but faces difficulties due to sample size requirements and difficulty of interpretation. We investigate the utility of incorporating network context into the construction and interpretation of PRS.

**CLINICAL RELEVANCE:** While PRS has found success in large cohorts with annotation of common phenotypes, rare disease represents a case in which recruitment of large sample sizes is difficult or impossible. In this paper we consider the specific case of preterm birth (PTB) and retinopathy of prematurity (ROP), two coincident conditions which would benefit from enhanced risk detection and contextual information regarding the degree to which genetic causes of these conditions are the same.

**METHODS:** To investigate the degree to which network context can supplement the information provided by traditional PRS, we propose novel methods of integrating network context into PRS construction before and after score construction, as well as evaluation of an existing method (HotNet2). These new methods are first developed and evaluated using data from the Michigan Health and Retirement Study (HRS), and then validated using a small study of PTB using data from the Danish National Birth Cohort (DNBC). The degree to which ROP candidate study targets overlap with those for PTB is also considered.

## 3.2 INTRODUCTION

Recent advances in genetic testing have led to a boom in the number of studies completed investigating various traits. As genetic assay development becomes more robust, cost of testing has decreased exponentially. Despite the advent of whole-genome sequencing techniques, due to low cost and adequate genomic coverage the majority of studies at scale are still performed using microarrays targeted to a subset of genomic loci in the genome at increased likelihood of single nucleotide polymorphisms (SNPs).

Although studies have become larger in sample size, allowing increased statistical power to detect differences between case and control groups, studies have been unable to uncover single SNPs of large effect in most cases. This problem has been referred to as the case of missing heritability, as studies have failed to capture a majority of the theorized heritable variability in complex phenotypes.(107)

One theorized reason for this missing heritability is that instead of single SNPs conferring large proportions of the heritable variability, it may be that many or all SNPs contribute a miniscule amount of effect.(122,123) Polygenic Risk Scores (PRS) are one proposed solution to these difficulties. Instead of focusing on single variants or a small subset of curated SNPs, PRS instead create a measure of the aggregate of genetic signal represented by many SNPs in a study.

PRS have quickly become a popular tool for investigation of genetic disease, finding application in traits such as height, obesity, cancer, and risk of cardiovascular disease. While single SNPs may account for small percentages of the total heritable variability, PRS have been

able to capture a significant percent and in some cases the majority of theorized heritable variability.

Despite these early successes, criticism of PRS exists. While PRS is able to capture a large proportion of heritable variability, scores offer no biological context for interpretation and exist in most cases as a simple additive sum of individual SNP effect. This limits the ability of researchers to translate PRS results into hypotheses that may inform future research, and also limits the ability to develop more effective methods of constructing scores that leverage such context to achieve increased predictive power.

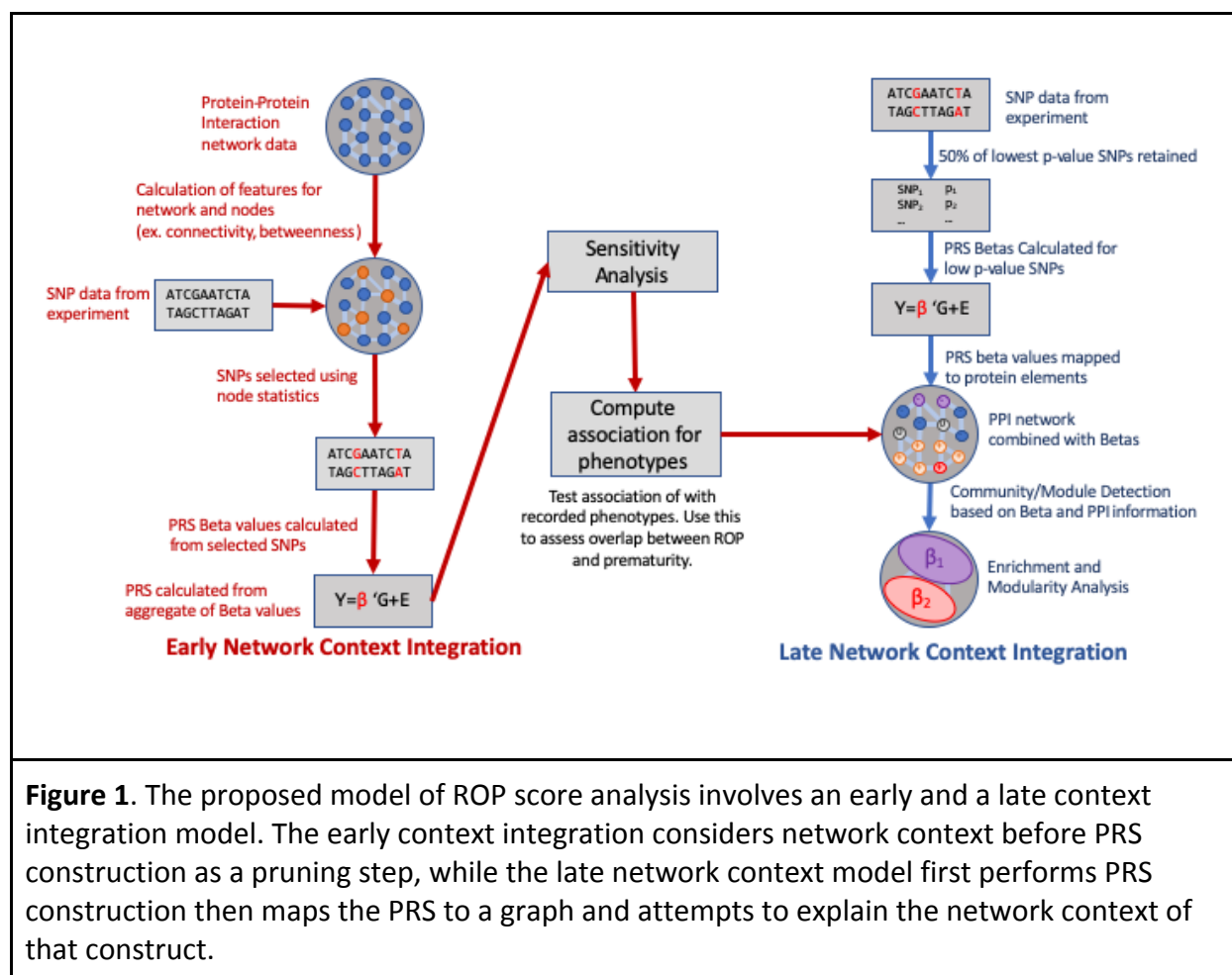
Additionally, PRS to date have required enormous cohorts to achieve meaningful predictive power. While common phenotypes like biometric data and the most commonly occurring diseases documented in population-scale studies are well suited to such approaches, many complex phenotypes occur only in a limited subset of the population, which will never be able to achieve massive recruitment. As such, studies have so far been limited in the phenotypes PRS are able to interrogate.

An added challenge arises when considering genetic conditions which may have coincident presentation. PRS exist solely as a summation of individual SNP effects, and this decreases the ability to compare between interlinked disorders or distinguish the degree to which such signals can be separated from one another.

In light of these challenges, we attempt in this aim to evaluate the utility of incorporating additional biological features into the construction and presentation of polygenic risk scores. Biological network data incorporating protein-protein interaction (PPI) information represents one avenue by which to increase the information content of such scores, and these

network-based approaches have shown promise in genetic studies of cancer phenotypes as well as other complex disease traits.(130,134)

We propose two complementary methods of incorporating network context into PRS. The first is an early network context integration approach, which leverages network features in the construction of a PRS. We evaluate the role of connectivity in the distribution of genetic risk, and give evidence that such network context provides additional information to PRS construction. (Figure 1)



**Figure 1.** The proposed model of ROP score analysis involves an early and a late context integration model. The early context integration considers network context before PRS construction as a pruning step, while the late network context model first performs PRS construction then maps the PRS to a graph and attempts to explain the network context of that construct.

Two early network context integration methods are discussed. The first relies solely on PPI data, leveraging connectivity as the primary thresholding method in order to assess

whether focus on high connectivity nodes increases the ability to predict disease outcome. The second takes into account additional information in the form of pre-existing candidate gene data resulting from biological studies, then augments that information using PPI information relevant to the provided gene products in order to test if focus on clinically validated targets and their partners outperforms an indiscriminate model of SNP inclusion.

We also propose a late network context approach, which seeks to provide additional context information after PRS construction. We first propose a method leveraging community detection after network thresholding derived from PRS model fitting to evaluate whether network submodules constructed from genes participating in the final PRS are enriched for specific gene ontology function. A second proposed method of late network context applies an existing insulated heat diffusion approach to detect small modules significantly enriched for genetic association signal.

We also investigate enrichment of these resulting scores for biologically relevant gene sets, attempting to provide a representation of which biological features serve as drivers in a fully specified PRS.

As phenotypes for consideration of these methods we consider first the case of height data. This hallmark trait is well investigated in many studies and as such serves as a benchmark to compare our method with other existing methods of PRS construction. In order to validate on a use case representing a dataset of limited size, we consider the case of preterm birth. We in turn consider the retinopathy of prematurity phenotype as a trait occurring coincident with preterm birth, and assess the degree to which the genetic causes of these disorders can be separated using a network augmented candidate gene approach.

Using this framework we introduce a novel method by which to expand the information included in PRS construction and evaluate its contribution to the predictive utility of such scores. We also provide a novel framework by which to evaluate PRS after construction and inform the development of new scoring metrics using network context which may increase the reach and applicability of such scores in the detection and interrogation of complex disease.

## **3.3 METHODS**

### **3.3.1 GWAS DATA**

#### **Health and Retirement Study**

The University of Michigan Health and Retirement Study (HRS) is a biennial longitudinal health survey of a cohort of adults age 50 and older in the United States with a focus on monitoring health and demographic features in participants over an extended time period. The HRS is sponsored by the Institute for Social Research at the University of Michigan and the United States Social Security Administration.(235)

In addition to biometric and disease related phenotypes, the HRS also includes genomic data for participants made available through the database of Genotypes and Phenotypes (dbGaP) which is operated as a service of the National Center for Biotechnology Information. This genetic data represents 12,507 participants with phenotypes including height.

Height data represents a well studied and easily measured example of a complex physical trait. Previous studies have assessed the narrow sense heritability of height at 80%, and the contribution of common variants to phenotypic variation in height are estimated at roughly 50% of this heritability. (236,237)

For these reasons, height data has proven popular as a phenotype in studies of polygenic risk methods, making it a well documented and easily comparable trait of study.

#### **GIANT Consortium Data**

The Genetic Investigation of Anthropometric Traits (GIANT) consortium is an international collaboration investigating the role of genetics in variability in human body size

and shape. GIANT represents a large-scale collaboration spanning several iterations over a ten year timespan. GWAS meta-analysis summary statistics provided by GIANT incorporate hundreds of thousands of individuals, and the large sample size of these data make them well positioned for the construction of polygenic risk scores.

Two meta-analyses included in the GIANT consortium's studies are used for the basis of this experiment. A large meta-analysis of 253,288 individuals across 79 GWAS by Wood et. al. serves as a means of constructing a polygenic risk score for the interrogation of GWAS data from HRS.(238)

In order to simulate a candidate gene study, a GIANT meta-analysis by Yang, et. al. incorporating 133,154 individuals across 38 studies is used.(239) This simulated candidate gene list is then expanded using network context, and the resulting gene set is used to filter the variants provided by the Wood study previously described.

#### **Danish National Birth Cohort**

The Danish National Birth Cohort (DNBC) is a large cohort of pregnant women and their infants with the intended purpose of providing data regarding pregnancy and childhood as well as the role that exposure in early life to environmental stimuli has on later disease.(240) The genetic homogeneity of this population makes it a useful study group for the evaluation of genetic scoring.

A subset of the DNBC representing 1000 preterm mother child pairs with spontaneous onset of labor prior to 37 weeks of gestation was made publicly available as part of a study included in the Gene Environment Association Studies initiative (GENEVA).(241) Infants from



this cohort form the basis of our investigation into preterm birth, and also serves as a trial case investigating the ability of network-based PRS methods to detect signal in small sample sizes.

### **Imputation**

DNBC and HRS data was imputed to reference build hg37 using the IMPUTE2 utility with use of SHAPEIT2 for pre-phasing of genotype data.(242,243) PLINK and GTOOL were used for variant file conversions and manipulations.(244,245)

In order to assure concordance between imputation methods, an evaluation of differences between the supplied BEAGLE (246) and IMPUTE2 methods was performed.

### **Reference Genome**

All data was updated to hg38 reference genome coordinates in order to coincide with gene regions represented in Ensembl Homo Sapiens gene database build 92. The LiftOver utility maintained by UCSC Genome Browser was used to perform conversion across reference builds for DNBC and HRS data.(247)

GIANT data was updated to hg38 coordinates in Python using SNP coordinates made available through the Single Nucleotide Polymorphism Database (dbSNP).(248)

### **Phenotypic Data**

In order to assess replicability of PRS generation, phenotype data for height from the HRS was transformed to age-adjusted sex-standardized values using a linear model in the R software package. Resulting PRS values were compared to those provided by the HRS.

Preterm birth phenotypes are binary representing a case/control study design with cases represented as infants born before 37 weeks gestation and infants born after 40 weeks gestation considered controls.

### **3.3.2 NETWORK DATA**

#### **Ensembl**

The Ensembl database is a project of the European Bioinformatics Institute and the Wellcome Trust Sanger Institute attempting to provide annotated genomic information in a centralized comprehensive resource.(249) Ensembl provides 153 annotated assemblies including those for homo sapiens, with notable information including gene and regulatory region genomic position information. For mapping of SNP positions to gene products, gene regions from the Ensembl Genes 92 build using hg38 was used.

#### **STRING**

The Search Tool for Retrieval of Interacting Genes/Proteins (STRING) is a protein-protein interaction (PPI) database maintained by the Swiss Institute of Bioinformatics, CPR-NNF Center for Protein Research, and the European Molecular Biology Laboratory representing experimental evidence of interaction and predicted interaction between 24.6 million protein entities across 5,090 organisms including homo sapiens.(121)

STRING attempts to provide maximal coverage of all documented interactions. Due to the highly connected nature of biological systems as well as the dispersion of genetic signal throughout the genome in complex traits, STRING is a well positioned database for genomewide studies as it provides maximal genomic coverage.(118)

STRING also provides a reasoned scoring system ranking the evidence of these interactions. This system allows a simple parameter by which to adjust the density of the PPI graph to prevent excessive sparsity or over-connectedness. In order to incorporate a large number of protein interactions while assuring strong evidence of those interactions, a high confidence segment of the STRING database was specified (score > 700).

### **Assessing Genomic Coverage**

In order to increase the coverage of gene regions and thereby account for upstream and downstream regulatory features, gene regions were extended to incorporate additional area. An assessment of the GIANT data set indicated maximal STRING network coverage at a window of 80,000 base pairs. **(Figure 6a)** A second analysis was performed to assess genomic coverage, and it was found that a larger window of 200,000 bp was necessary for each mapped SNP to be represented at least once, however this resulted in significant multimapping of SNPs between genes with the average SNP participating in approximately 9.5 proximal gene regions. An extension of 80,000 bp was sufficient to capture 75% of represented SNPs while participating in an average of 4.6 gene regions. **(Figure 6b)**

In the interest of maximizing graph coverage while minimizing overlap between gene regions, a window size of 80,000 bp was selected for mapping of SNPs to the Ensembl reference.

### **Mapping From GWAS To STRING**

Using summary association statistics produced by PLINK, SNP coordinates were assigned to all gene regions with a boundary falling within 80,000 bp. After assignment to Ensembl gene

regions, these were mapped to analogous Ensembl protein stable IDs using linking tables acquired through the Ensembl BioMart utility.

The STRING database was loaded into Python using the iGraph network analysis tool.(250) SNPs were then assigned to STRING entities by matching Ensembl protein stable IDs. The resulting graph incorporates 19,586 protein entities with 360,341 edges.

### **3.3.3 POLYGENIC RISK MODEL**

Two alternative methods of network PRS construction were used to incorporate network context into PRS. The first method, which we will refer to as early context, incorporates network information before SNPs are assigned to the STRING graph. In the early context method the organization of network nodes and edges guides pruning of the graph, and SNPs assigned to the retained nodes are used to construct the final PRS for evaluation.

#### **Early Network Context Integration**

Two alternate methods of performing early context integration are described. The first method of early network context integration, which we will refer to as the early network connectivity model, uses measures of centrality in order to select the most highly connected nodes represented by the STRING database. Evaluation of the GIANT data found that when looking at minimum and maximum values assigned to gene regions, highly connected nodes displayed a trend toward higher median values for beta and p-values. **(Figure 2)** In order to assess the broad implications of this trend, the subset of nodes with degree greater than the median value in the graph are selected, and SNPs associated with these nodes are then used in further score construction

A second method of early network context integration makes use of curated candidate gene lists specific to the phenotype of interest. Curated candidate genes from clinician review or previous genetic studies are matched with their STRING entity counterparts. The coverage of the graph is then extended to direct neighbors of these candidate genes. SNPs associated with the resulting subgraph are then isolated and used in further score construction.

### **Late Network Context Integration**

A second method of network context integration is proposed in order to add information to interpretation after the construction of polygenic risk scores. In this method, PRS construction is first performed using all available SNPs or a subset of SNPs as described in the early context methods. The resulting p-value and beta values are then assigned to a STRING graph object.

The resulting graph object is then assessed for modularity and clustering of signal which has been incorporated into the PRS. Two alternate methods of deriving context from this information are proposed, as well as an integrated approach for combining the output of these two methods.

The first method of performing late network context analysis of PRS is proposed using the modularity of the nodes represented in the final set of beta values. In this proposed method SNPs are first pruned by a permissive significance threshold (generally  $\sim 1e-4$ ), and these SNPs are assigned to the STRING graph object. The resulting graph is pruned to include only those nodes that are matched with at least one SNP object.

This pruned graph is then analyzed using a community modularity maximization algorithm.<sup>(133)</sup> The resulting communities are used to create a maximum possible community

beta value, representing the sum of all beta values within the community. The resulting community beta values are evaluated against a null distribution created through permutation in order to assess if they are significantly enriched for effect. Plots of hierarchical gene ontology terms are created from those available through the AmiGO2 ontology annotation database.(251)

A second method incorporates all SNP association data as well as all graph nodes in the STRING graph object. In this method, SNPs resulting from the GWAS are loaded onto the STRING graph. The minimum SNP p-value assigned to each node is first isolated. This value is transformed to its  $-\log_{10}$  equivalent heat value.

The list of nodes and heats is then analyzed by the HotNet2 insulated heat diffusion algorithm,(134) which outputs a list of subnetworks significantly enriched for signal. These subnetworks can then be analyzed for enrichment for biological context with the goal of identifying novel targets or interactions.

#### **Calculation of PRS from SNP Values**

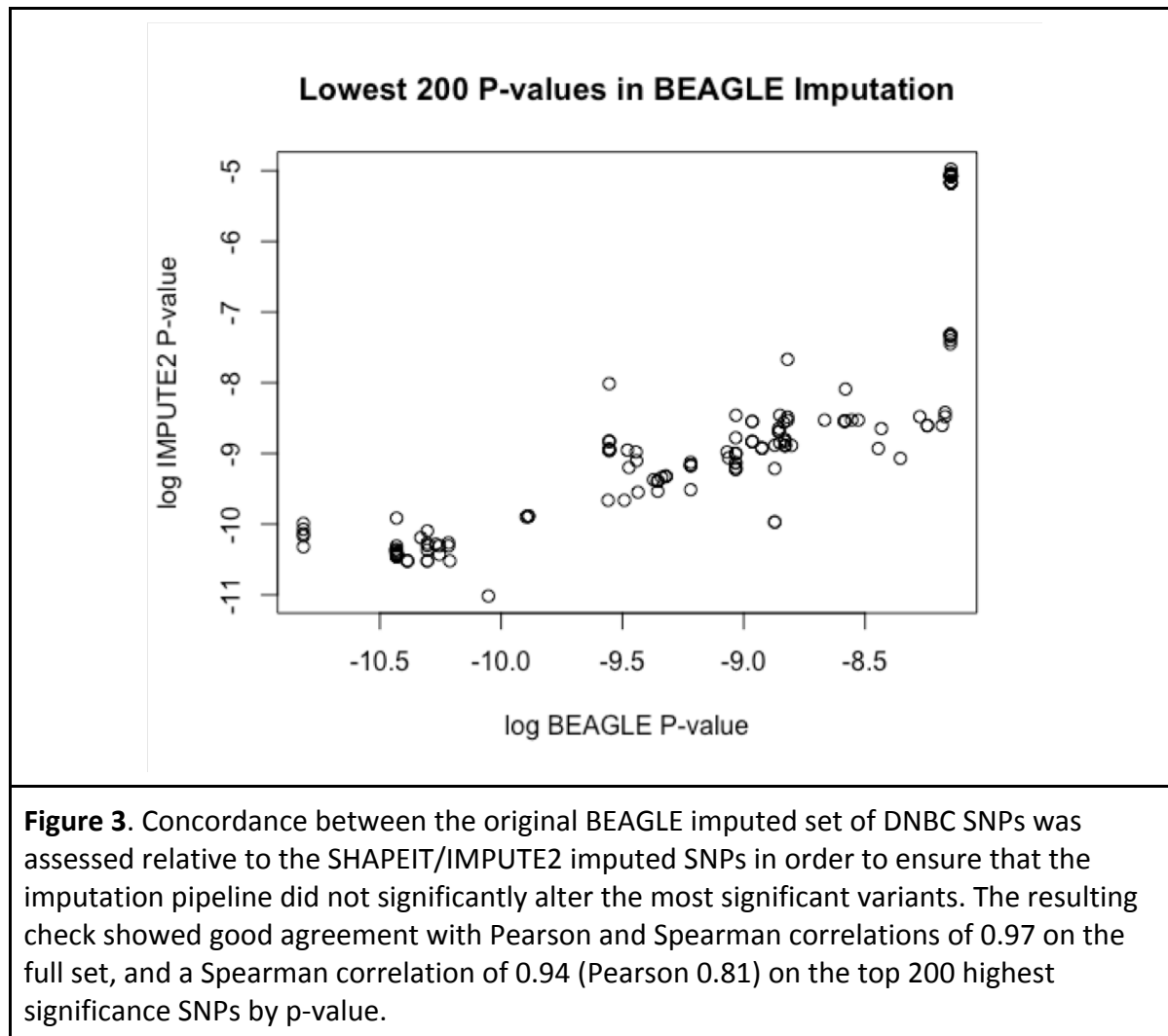
Output of early context methods are reduced to unique SNP IDs. The resulting list of SNP IDs is isolated from the full SNP set using PLINK. Filtered or unfiltered output is then used as input to PRSice, which performs clumping, threshold selection, and scoring of the test group assuming an additive model of disease.

Statistical analysis is performed in R and Python.

## 3.4 Results

### 3.4.1 IMPUTATION ANALYSIS

In order to harmonize data from the HRS and DNBC studies, data was re-imputed from base data using the IMPUTE2 platform. DNBC data was reannotated with hg37 coordinates using the LiftOver utility from UCSC. Haplotype estimation was performed in SHAPEIT2 using an effective population size of 15000. Imputation was then performed using IMPUTE2 in 5 Mb chunks to the 1000 Genomes Phase 3 reference. The resulting imputed reference was filtered to include only SNPs meeting a quality score threshold of 0.8.



After imputation, concordance was assessed against the originally provided BEAGLE imputed values. The resulting SNPs showed a Spearman correlation of 0.97 for the full set of SNPs, while the 200 most significant SNPs were correlated with a Spearman correlation of 0.94.

As the only available 1000 Genomes hg38 reference is the product of using Lift Over to transfer the hg37 dataset to hg38 and thus no native imputation to hg38 exists, the imputed data sets were then reannotated with hg38 coordinates using LiftOver in order to achieve concordance with gene boundary data provided by the Ensembl database of human genes.

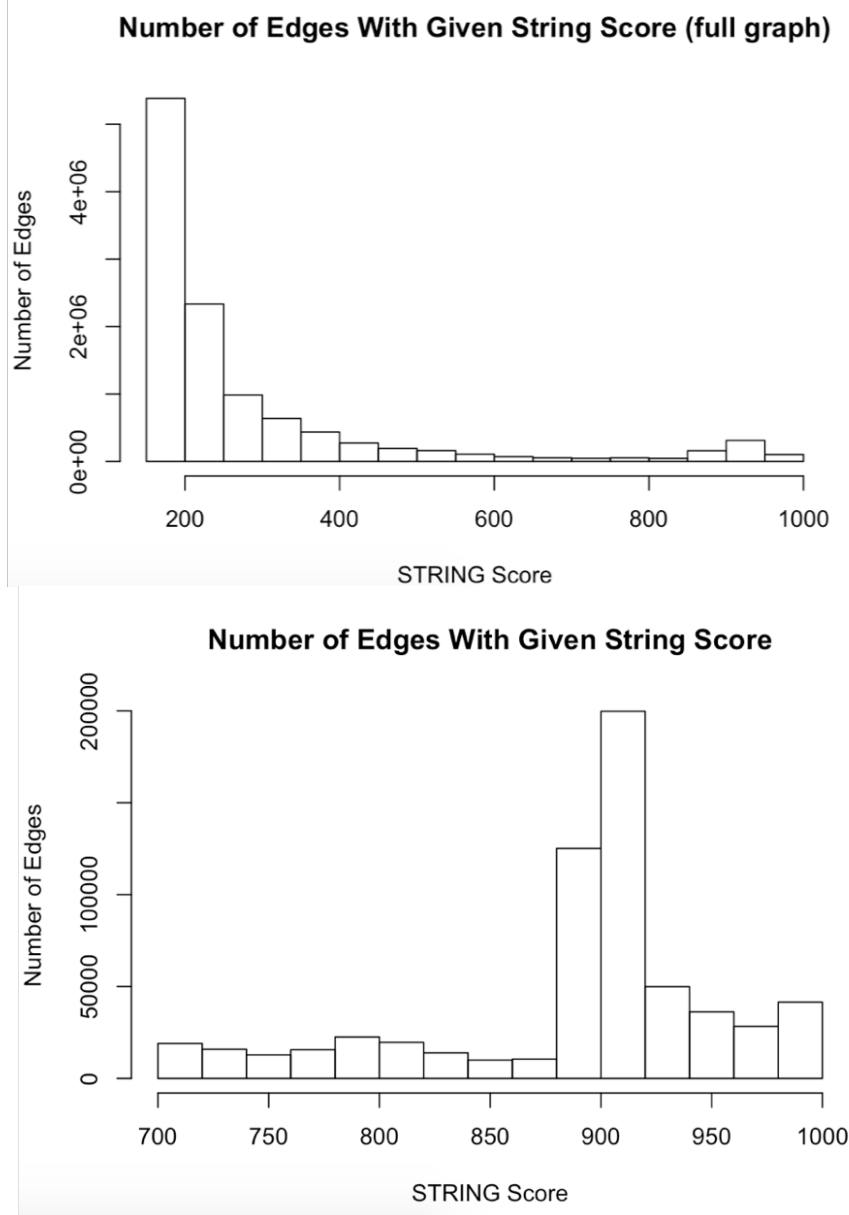
### **3.4.2 EDA OF STRING NETWORK**

Initial analysis of the STRING network showed that most edges were assigned low confidence scores for association. **(Figure 4)** The STRING network advises that edges above a score of 700 are considered high confidence, while edges between 400-700 represent medium confidence, and scores below 400 represent low confidence. The distribution of edges is bimodal, with a peak occurring at roughly a score of 900. **(Figure 4)** In order to include as many high confidence edges as possible without leading to an overconnected graph which would create problems for community detection algorithms, a threshold of 700 was applied to STRING for the formation of the template graph object.

The resulting graph represents 19,576 entities with 360,341 edges. Graph connectivity is roughly scale free, with a median node degree of seven and a maximum node degree of the graph at 1,249. The largest component of the graph represents 15,154 entities connected with 360,341 edges.



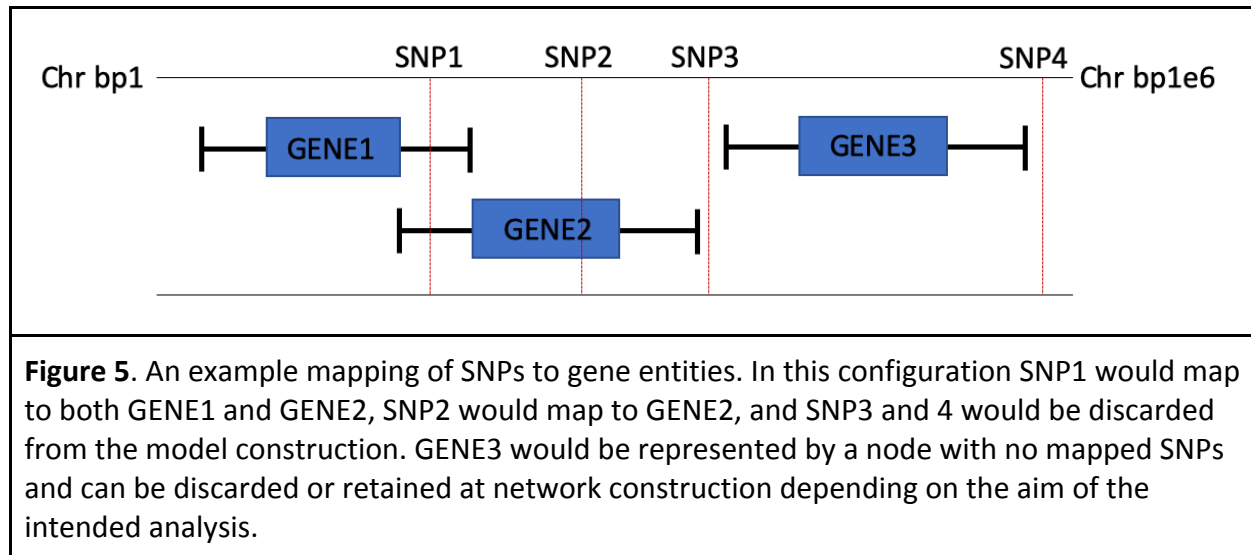
In order to map from ENSP identifiers in STRING to ENSG identifiers in the Ensembl genomic position data, a conversion table obtained from the Ensembl BioMart service was used. Of the 15,154 protein entities represented in the STRING graph, 14,381 matched at least one ENSG identifier in the Ensembl database. Of all STRING identifiers, 14,214 mapped to a HGNC identifier.



**Figure 4.** STRING edges were shown to be roughly bimodal by confidence of association (top). A second peak occurs at a confidence score of roughly 900 (zoomed detail, lower). Score are percent values multiplied by a factor of 100, thus a score of 700 represents a 0.70 relative proportion of strength of evidence for association compared to the full set of protein entities.

### 3.4.3 SNP MAPPING MODEL

#### SNP Mapping Process



SNPs were assigned to genes by beginning with coordinates from the Ensembl Homo Sapiens build 92 GRCh38 annotated as gene regions. Borders of gene regions were expanded to account for regulatory regions and maximize genomic coverage. SNP coordinates were assessed and assigned to all matching gene regions and written to a reference file for graph analysis.

(Figure 5)

#### Investigation of Network Coverage

In order to determine the best threshold to achieve maximal network coverage while also minimizing overlap between genes, a parameter sweep of gene window size was performed.

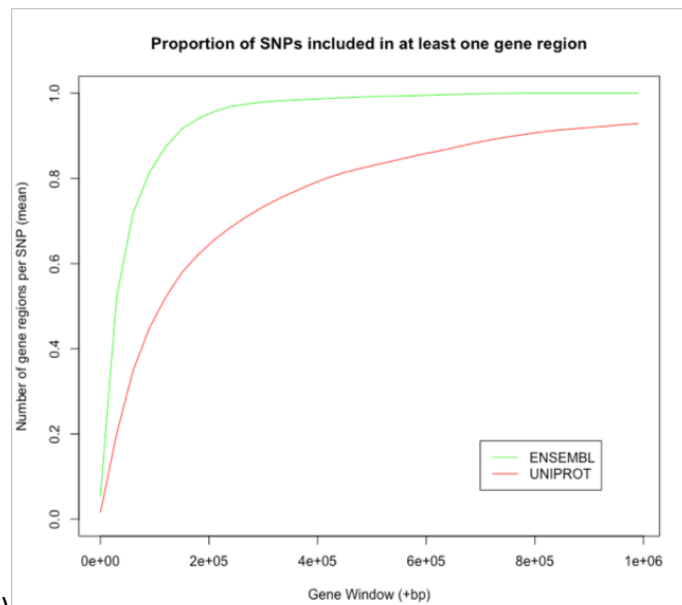
The first criteria investigated was the effect of window expansion on genomic coverage. Genomic coverage was found to increase rapidly until a window size of approximately 100 kb,

at which point roughly 80% of SNPs are included in at least one gene region. At 200 kb approximately 95% of SNPs are included in a gene region.

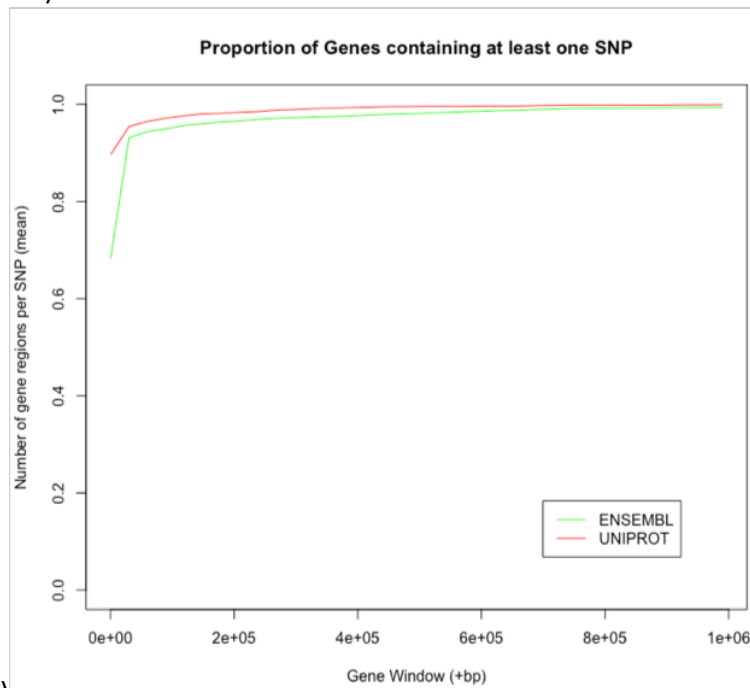
This analysis was balanced against an analysis of SNP participation in gene regions. In order to attempt to ensure that SNPs are as informative as possible, it is advantageous that SNPs participate in as few distinct gene regions as is possible while maintaining high overall genomic coverage. Gene region participation was found to increase roughly linearly with window size, indicating that the smallest gene region attainable would be advantageous to maximize this criteria. (**Figure 6**)

An additional analysis of network coverage with expanding window size was also evaluated. At 80 kb, roughly 95% of available genes are represented by at least one participating SNP.

In order to maximize these competing criteria, a window size of 80 kb was chosen in order to minimize cross gene participation while ensuring maximal network coverage.

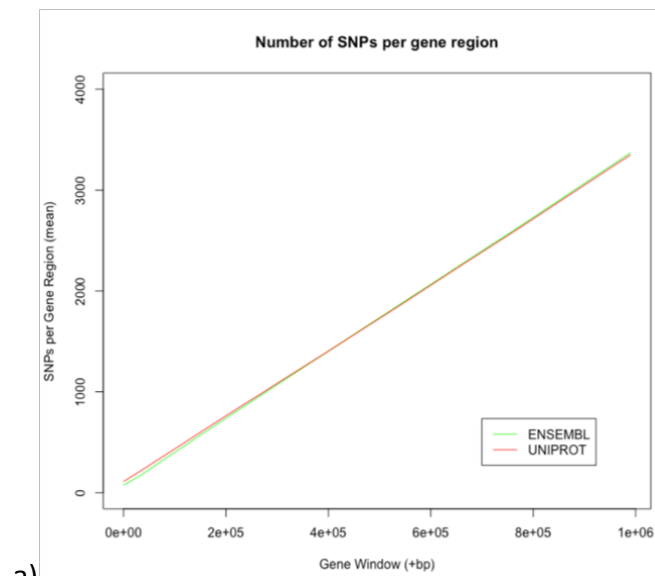


a)

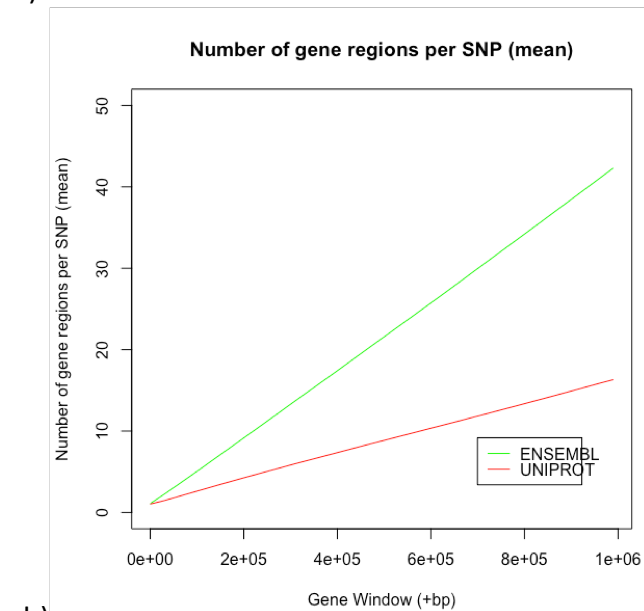


b)

**Figure 6.** SNP mapping to gene regions was investigated as a function of coverage and multimapping using both an Ensembl and Uniprot bridge onto a STRING network construct. The majority of SNPs could be found to map to a gene region with a gene region expansion of 200 kb (a) while most genes were found to contain at least one SNP at a window size of under 100 kb (b).



a)



b)

**Figure 7.** The number of SNPs per gene region scaled linearly with window size (a) as did the number of genes that SNPs were mapped (b).

### **Traditional PRS Formulation**

To serve as a control, a traditional PRS was constructed using the HRS dataset as a test set. As the basis for this study, GWAS summary statistics from the GIANT study representing height data of 253,288 individuals was used. After assignment to hg38 coordinates in order to match ENSG entity positions, a PRS was generated using the PRSice software package.

The resulting score reached a best-fit significance threshold for SNP inclusion at 0.0002 and obtained an  $R^2$  of 0.1773, representing a capture of roughly 18% of heritable variability in height. The model was validated using a test set of 9,915 individuals from the HRS using logistic regression, and the resulting fit of the model is highly significant ( $p < 1E-300$ ).

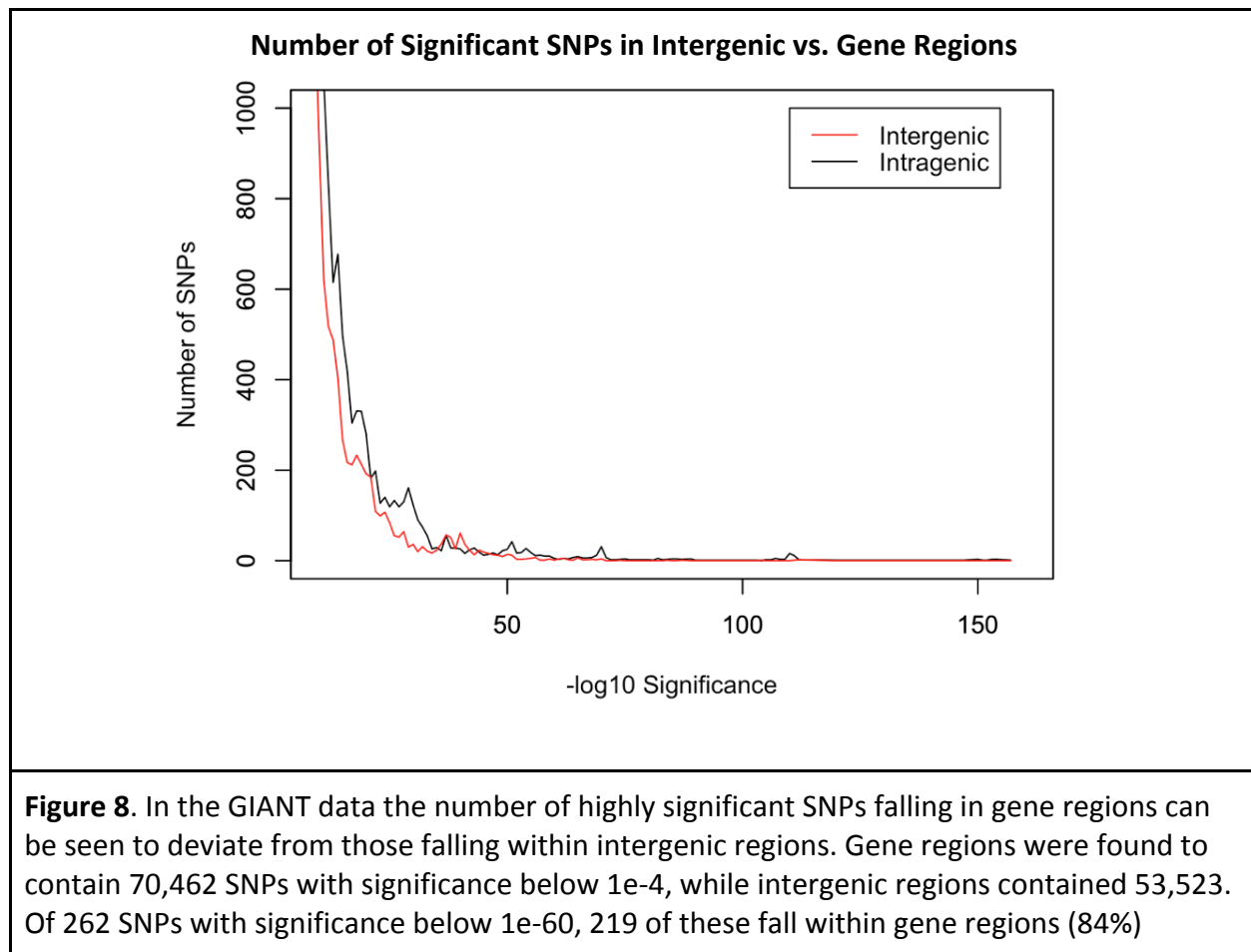
### **3.4.4 EARLY NETWORK CONTEXT METHODS**

In order to evaluate the utility of inclusion of network context during PRS construction, two different methods were employed: One attempting to utilize network connectivity to enhance study power, and a second method to assess the impact of integrating candidate gene lists gleaned from literature during the construction of the PRS.

#### **Inclusion of All Gene Regions**

In order to test the hypothesis of whether a score constructed from annotated gene regions would outperform a model including all SNPs, a score was created using all SNPs mapping to annotated gene regions. This model assumes that a basis of regions more densely populated with high significance SNPs may decrease noise contributed by low-significance SNPs from intergenic regions.

To evaluate the degree to which the assumption that gene regions are enriched for high-significance SNPs holds true, an evaluation of the number and quality of SNPs assigned to intergenic regions was performed. Gating of SNPs without gene region expansion resulted in inclusion of 1.47 million SNPs comprising 56% of the total set of 2.55 million SNPs in the GIANT meta-analysis summary statistics.



Evaluation of intragenic SNPs showed evidence of enrichment for highly significant SNPs with significance values less than 0.0001 when compared to the original set of SNPs. These



results indicate evidence that gene-regions do in fact contain a higher density of significant SNPs. **(Figure 8)**

The PRS built on all gene regions after expansion by 80 kb including 2,766 SNPs significant beyond the 0.00025 best-fit threshold achieved an  $R^2$  of 0.1639 and was highly significant with a p-value less than  $1E-300$ . While this score does not achieve predictive ability exceeding that of the score including all SNPs, the removal of 43% of SNPs results in a modest decrease in  $R^2$  of 7.6% indicating that intergenic regions may contribute only a small percentage of predictive SNPs. **(Figure 11)**

### **Connectivity Gated Score**

As genes code for biological products that interact with one another, it stands to reason that dysfunction in entities interacting with many other partners represent a likelihood of greater disruption in biological function. Operating under this assumption, it can be hypothesized that a score composed of high connectivity partners will outperform a score composed of all genes. In order to investigate whether this assumption is reflected in the STRING network mapping of protein interactions, we performed several analyses before arriving at a method for creating a connectivity-gated early network context score.

### **Network Connectivity Analysis**

In order to evaluate the role of connectivity on network nodes, an analysis of the effect size and significance of gene regions was performed. Three measures of connectivity were considered in order to perform this analysis: the degree centrality of nodes, representing the number of outgoing and incoming edges; the betweenness of nodes, representing the

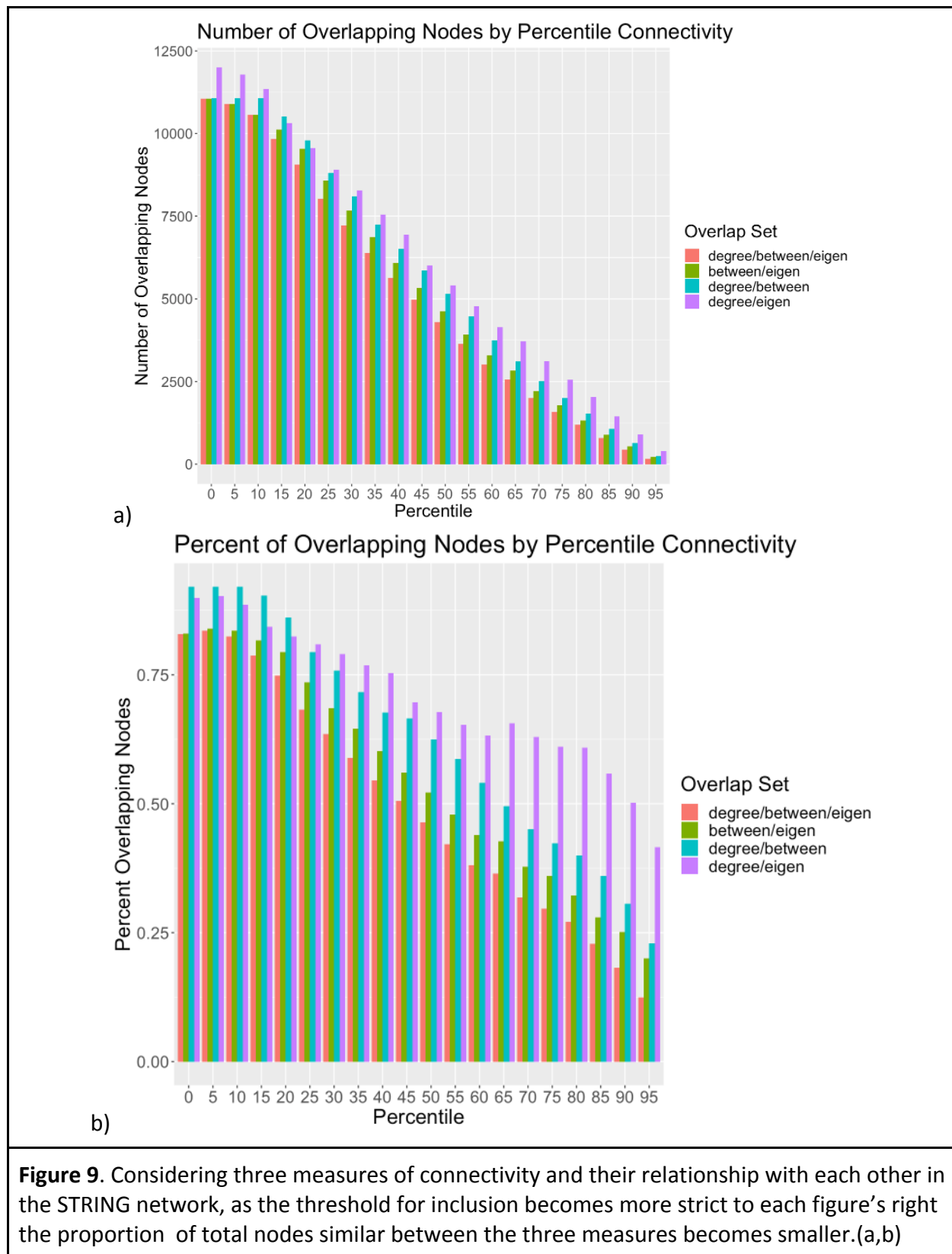
importance of specific edges as choke points in the graph; and the eigenvector centrality of nodes, a measure taking into account the degree of nodes as well as those nodes' neighbors, with the assumption that nodes interfacing with many high degree nodes have increased effect size.

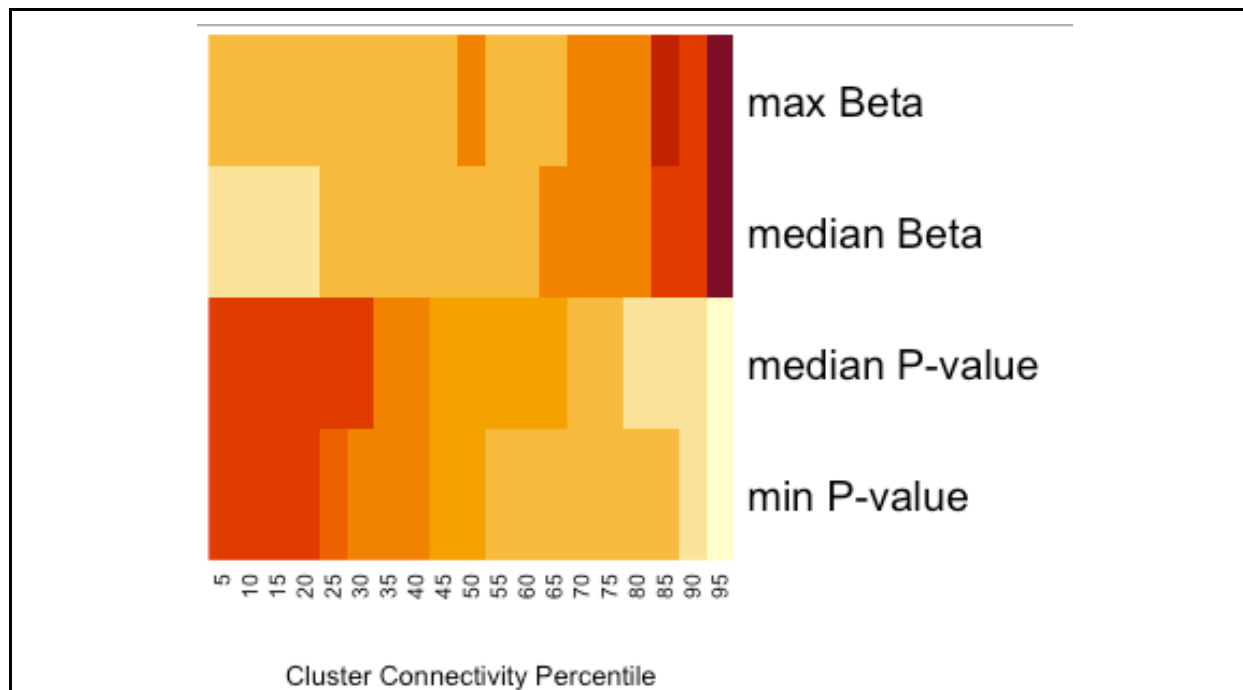
The overlap of these three connectivity measures was assessed in order to determine if a subset of nodes existed representing high centrality across all three measures. The degree of overlap between all three measures was found to be quite high. Degree and eigenvector centrality were the most similar of the three measures, while betweenness and eigenvector centrality were the most dissimilar. **(Figure 9)**

When considering the median p-value and effect size of SNPs assigned to gene regions, a slight but significant trend toward higher effect sizes and smaller significance values can be observed as the connectivity of nodes increases, especially toward the extremes of the distribution. **(Figure 10)**

When considering nodes falling in the top five percent of each measure, 12% of nodes fell in the top of all three measures. This subset of 267 nodes was compared to a random background representing a sample of 1,000 similarly sized nodes. The increases were found to be significant when considering the median p-value ( $p < 0.01$ ), the minimum p-value ( $p < 0.05$ ), and the median beta values ( $p < 0.05$ ). Maximum beta value was found not to be significant at a threshold of  $p < 0.05$ .

Inspection of these genes showed an enrichment for annotation regarding RNA polymerase function, histone function, and other cellular function proteins. **(Appendix Figure 9)**





**Figure 10.** The relationship between eigenvector centrality and degree centrality can be observed to be the most closely related, while betweenness and eigenvector centrality diverge the most. Focusing on the subset of nodes which fall into all three groups, a trend in median values toward more extreme effect sizes and p-values can be observed.

### Connectivity Score Performance

Taking into account the similarity of the connectivity measures assayed and the evidence of increasing effect with connectivity, a PRS encompassing the top 50% of nodes by degree centrality was constructed. Nodes having greater degree centrality values than the median of 7 were selected. SNPs falling in those gene regions were incorporated into a PRS.

The resulting PRS using a basis of 905,153 SNPs comprising 27% of all SNPs and 3000 score SNPs with a best-fit threshold of 0.0012 achieved an  $R^2$  of 0.1506. The score was again highly significant with a p-value < 1E-300. (**Figure 11**)

As with the all-gene-region score, the connectivity gated PRS fails to outperform the model incorporating all SNPs. However, a decrease to 36% of the total number of SNPs included in the full model results in only a 15% decrease in predictive ability when compared to the full model, again suggesting that a large percentage of signal may be included in connected regions.

### **Candidate Gene Score**

In order to evaluate the degree to which clinical evidence can be used to increase PRS predictive ability using automated techniques, we evaluated a method of forming scores using curated lists of disease targets from existing literature. This model assumes that curated lists of candidates resulting from focused biological hypotheses generated by subject experts describes a core set of genes with high likelihood of influencing disease progression. This method is in keeping with an omnigenic model of disease wherein core genes are influenced by many partner genes.<sup>(122,123)</sup>

### **Simulation of Height Candidate List**

As no list of candidate genes for height exists, we attempted to simulate a list of highly significant genes using summary statistics from a study performed by Lango Allen et. al. into the genetic causes of height as an earlier phase of the GIANT consortium project.<sup>(112)</sup>

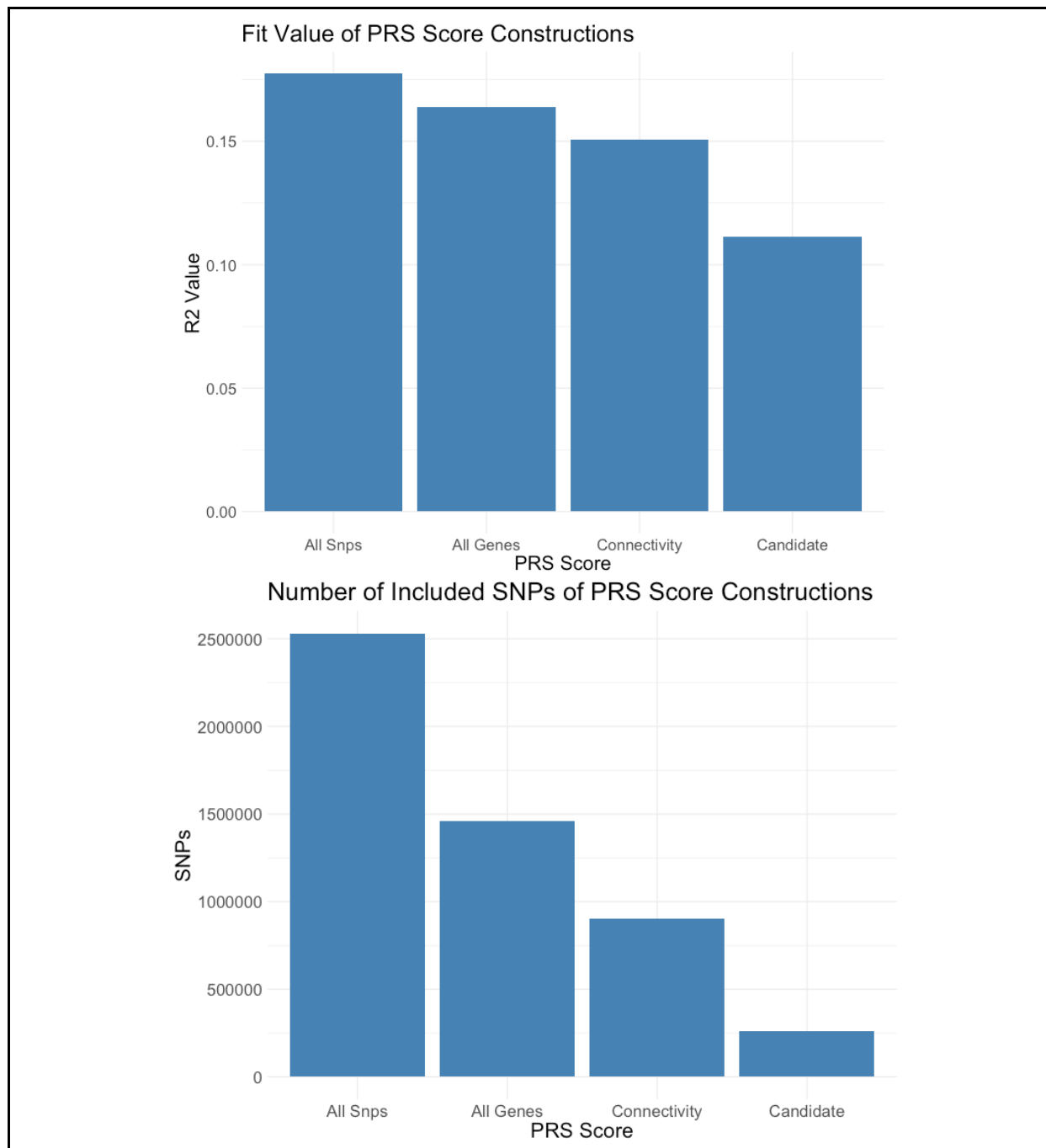
Using the ranked list of SNP p-values included as summary statistics, we selected a subset of SNPs comprising the 500 variants with the lowest p-value statistic. The protein entities including these SNPs were then identified by using the Ensembl database of gene regions.

### Candidate Gene Score Construction

This list of simulated candidate genes was selected on the STRING network graph. Each candidate gene's neighbors were then added to the subset of included genes. SNPs falling in these regions were then used as the basis of constructing a PRS.

The resulting subset of SNPs represents 261,338 SNPs, or 6% of the total included SNPs. The constructed score showed a fit  $R^2$  of 0.1113 and was significant with a p-value of 2.00E-256.

The candidate gene score represents a 94% reduction in the raw number of SNPs included, and this also represents a commensurate decrease in coverage. However, fit of the model decreases only 37%. In order to investigate whether the subset of candidate genes represents a meaningful subset containing a large amount of signal, the expanded set of candidate gene SNPs was subtracted from the full SNP set and PRS fit was recalculated. The resulting full SNP set with candidate gene SNPs removed resulted 2,270,498 SNPs (94% of total SNPs) and achieved an  $R^2$  of 0.1411. This may indicate that the specific SNPs involved in construction account for a greater percentage of total heritable variability in the test sample than a random sample of SNPs. **(Figure 11)**



**Figure 11.** GIANT constructed PRS measures were validated on the HRS participant genetic data. All scores were highly significant in their  $R^2$  measures. While the number of SNPs used in the candidate score is less than 5% that of the full set of SNPs, the HRS candidate score suffers only a 33% reduction in predictive ability.

### **Permutation Testing of PRS Fit**

In order to assess the degree to which performance differences in early network context PRS deviated from existing formulations, a background distribution was created by permuting random sets of SNPs selected from the full set at various sizes for comparison with the various PRS methods.

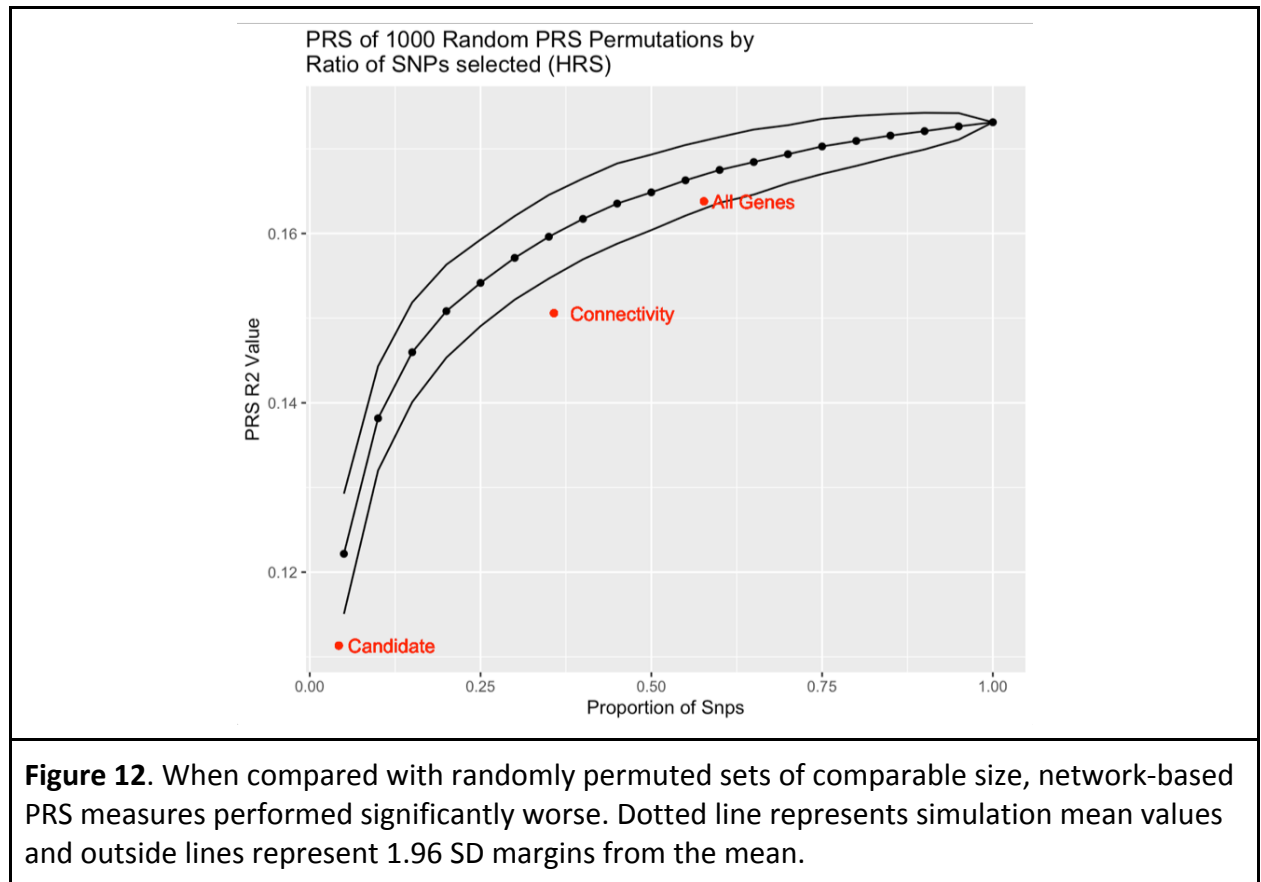
A set of 1,000 random permutations were performed at 5% increments of the full SNP population. Each method was then compared to a random set of the same size in order to assess its performance relative to random background.

When comparing to random background, the score built from all gene regions performed worse than the mean random set of SNPs of similar size, but was not significantly different from a random set of SNPs. Both the connectivity and candidate gene methods performed significantly worse than would be expected for a random set of selected SNPs.

### **(Figure 12)**

These results underscore the importance of full genomic coverage in prediction of complex phenotypes, and make clear the difficulties inherent with the incorporation of network information into score construction.





### 3.4.5 LATE NETWORK CONTEXT METHODS

As a complement to methods meant to increase the predictive value of PRS, we also evaluated methods of gleaning additional information from existing score constructions. While PRS have shown promise in identifying individuals at high risk for various conditions, they are also difficult to interpret.

We attempt to provide tools to provide additional interpretability to researchers when constructing and applying PRS, taking into account biological annotation that may allow increased ability to derive meaning from these scores.

## Modularity Assessment of HRS Data

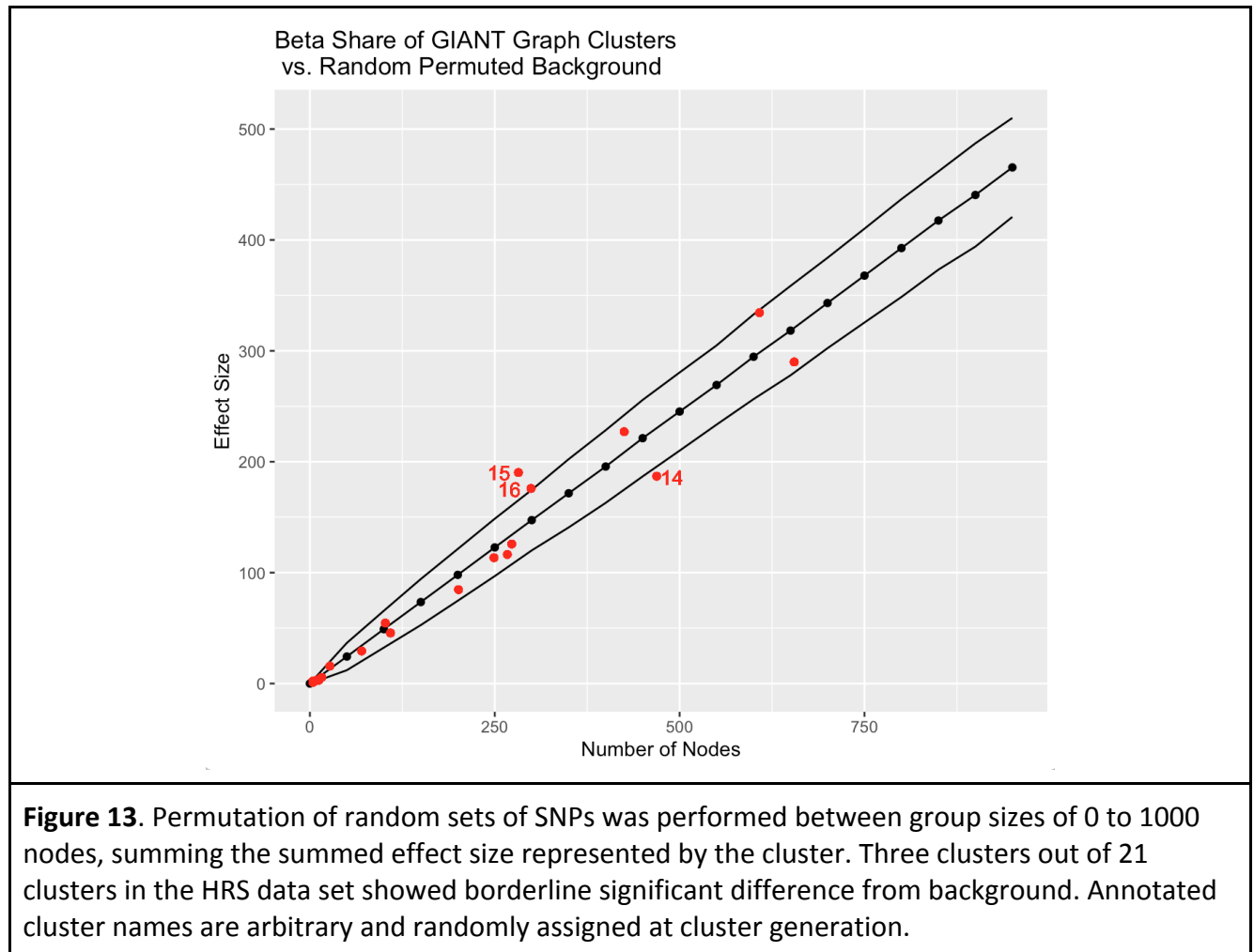
After construction of the height PRS from GIANT summary statistics in PRSice, the best-fit threshold for SNPs was identified to be  $p < 0.0001$ . After loading the STRING graph, genes only containing SNPs that failed to meet this threshold were subtracted from the set of nodes, leaving a largest connected component containing 4,513 nodes and 51,694 edges.

Community detection was performed on the reduced graph to determine if community structure could be derived implying regions of interacting genes. The community multilevel algorithm proposed by Blondel et. al. was applied to the pruned PPI graph using igraph in Python. The analysis resulted in 21 identified communities varying in size from 4 to 647 nodes.

## Permutation Testing

Absolute values of SNP effect sizes within clusters were summed as a proxy for the amount of potential signal each cluster contained, and these values were converted to a ratio of all effect size values included in the PRS. A null distribution was created by permuting 1,000 subsets of nodes in sizes increasing by 50 node increments between 50 and 1,000 nodes and arriving at a distribution of PRS beta share expected within each node size. Individual p-values are calculated by permuting 1,000 groups of nodes the same size as the cluster and performing a z-test.

When compared to background, three nodes are found to have borderline significance at  $p < 0.05$ , one node showing low enrichment for effect size and two significantly enriched for effect size. **(Figure 13)**



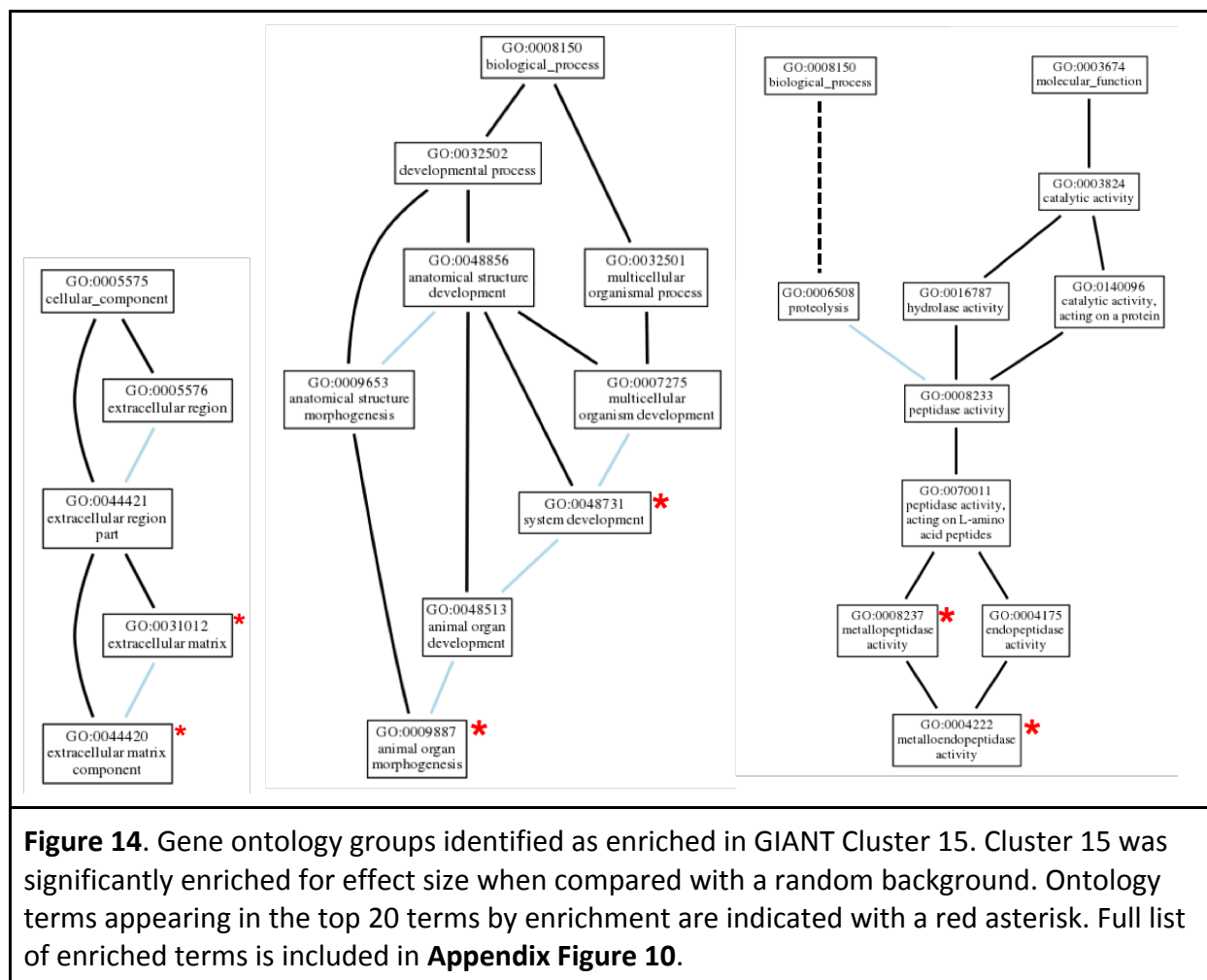
### Module Enrichment

A gene-set enrichment analysis (GSEA) was run using gene sets from the Molecular Signature Database (MSigDB). Clusters derived from the module detection process were individually analyzed via the GSEA process.

A Shiny utility in R using the FastGSEA package was created to investigate data by isolating clusters of nodes and investigating the gene-set enrichment of those clusters.(252–254)

Of the two modules with statistically significant enrichment for effect size, neither showed significant enrichment according to p-value adjusted for the number of gene sets investigated.

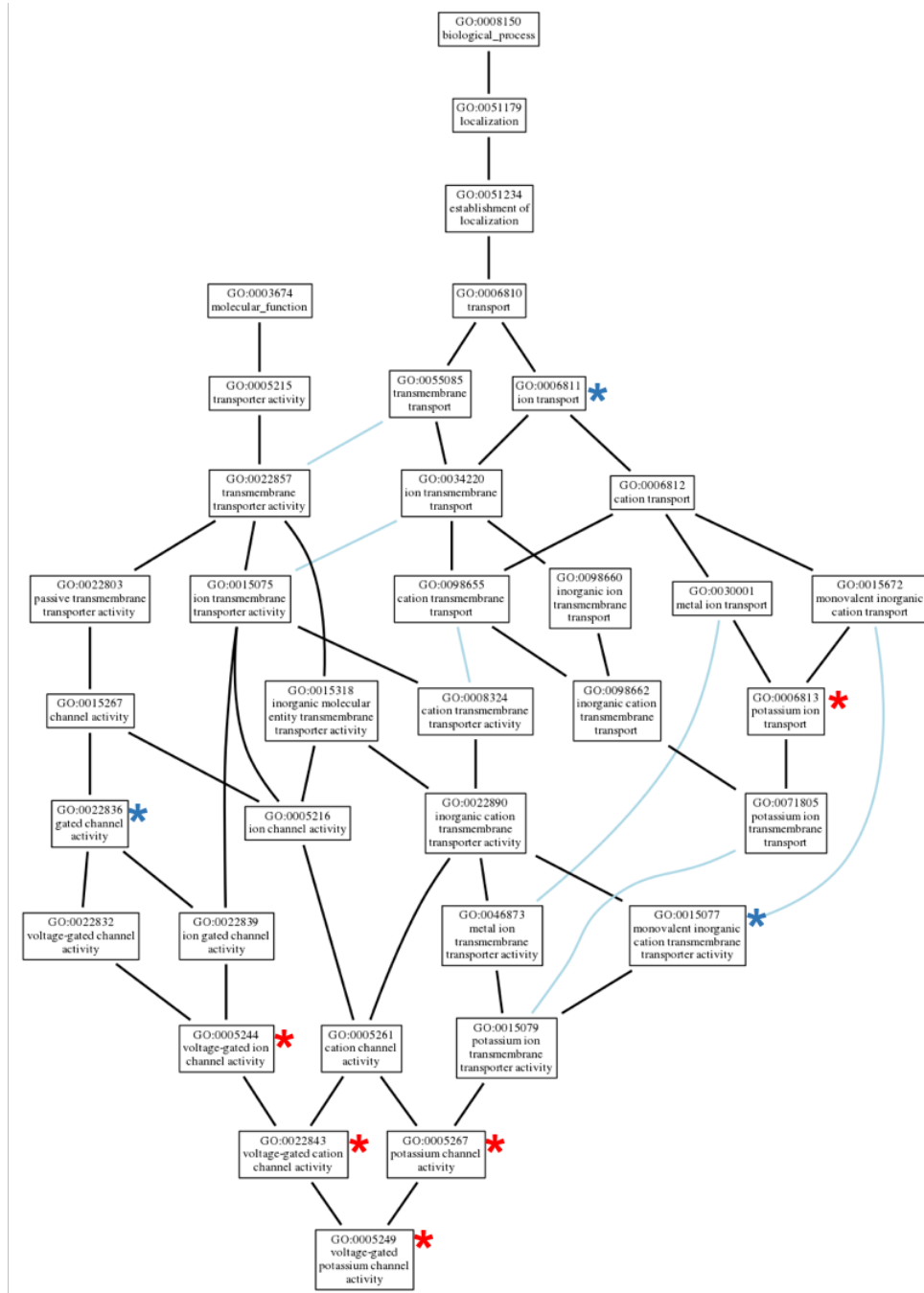
Cluster 15 showed the highest statistical significance for enrichment of effect size relative to cluster size with a  $p < 0.001$ . The cluster contains 276 nodes and accounts for 9% of all available effect size. The most highly enriched gene sets were those related to zinc and calcium ion binding and cellular matrix functions. Investigation of hierarchical models of gene ontology from AmiGO2 show enrichment for extracellular matrix functions, organ and skeletal development, and metalloproteinase activity. (**Figure 14**)



Cluster 16 was also enriched for share of effect size relative to cluster size with a  $p < 0.05$ . The cluster contains 299 nodes and accounts for 8% of total available effect size. The most highly enriched gene sets represented are those for ion transport activities. Hierarchical models from AmiGO2 show the strength of enrichment for potassium channel activities, with eight ontological categories in the same tree among the top 20 hits. **(Figure 15)**

Of all measured clusters, cluster 2 contains the greatest percentage of potential effect size overall at 16% of total potential effect size. Enrichments within this cluster include various chromatin-related processes involved in silencing, packaging, and expression. Interestingly, several systems relating to neural development also appear. Investigation of AmiGO2 hierarchical models show that these chromatin processes exist across many independent systems. The described neural development processes appear as upstream processes with involvement in forebrain development. **(Figure 16)**

Taken together these analyses show a moderate degree of ability to distinguish involvement of various ontological descriptions of biological context when considering communities found within constructed PRS. The enriched clusters show some evidence of enrichment for growth-related ontologies as well as ion-transport phenotypes that may be important for growth. This suggestion of new structure is interesting, but ultimately requires validation before it can be interpreted as a real relationship.

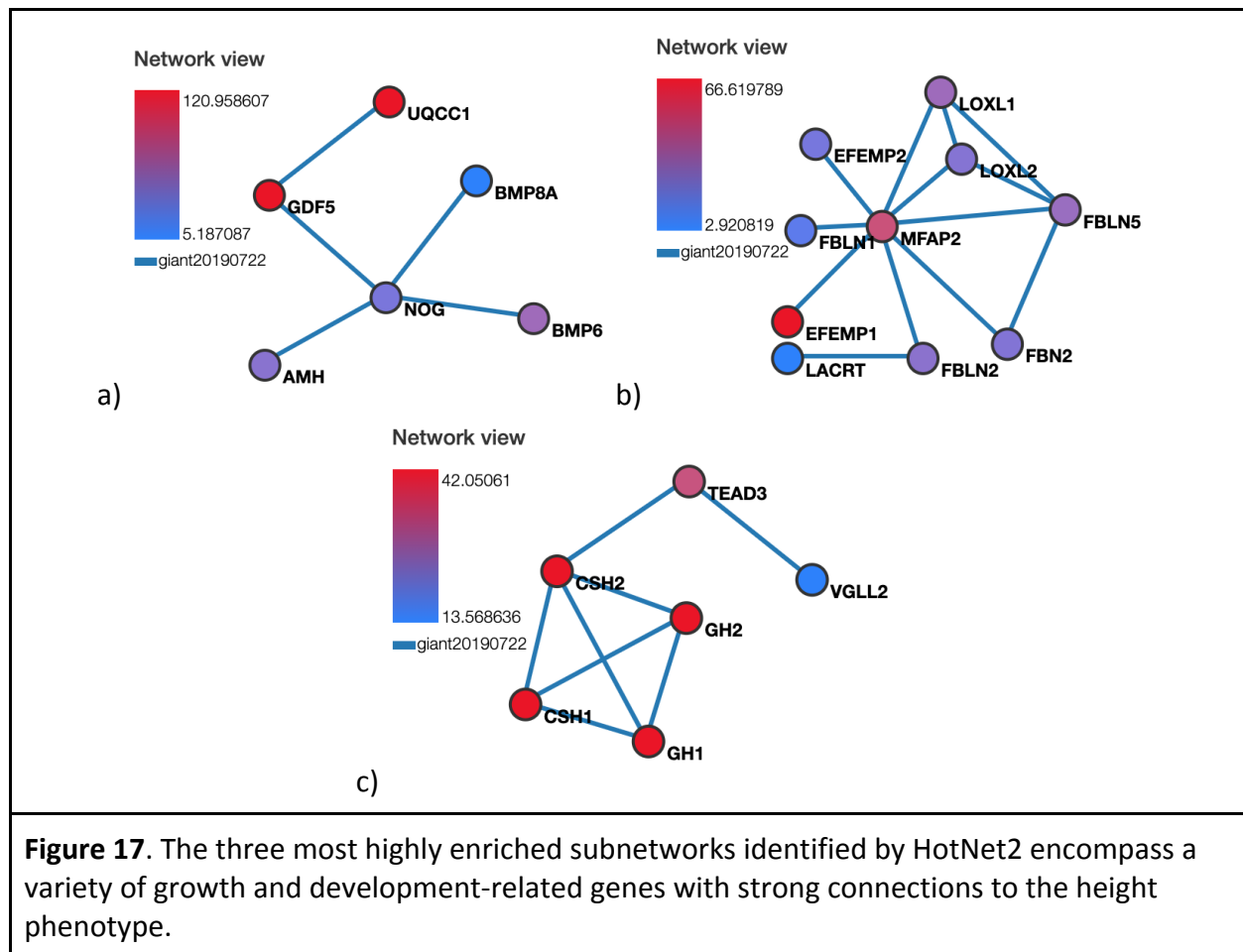


**Figure 15.** GIANT cluster 16 showed enrichment for terms relating to ion transport processes. Red asterisks indicate GO terms significantly enriched and appearing in the top 10 most enriched results. Blue asterisks represent GO terms enriched and appearing in the top 20 most highly enriched results. Full list of top terms presented in **Appendix Figure 11**.



taken and served as the per-node heat input into HotNet2. STRING network edges representing high confidence in protein-protein interaction (score > 700) were isolated and used as edge input to HotNet2.

The resulting analysis using an insulation value of 0.19 found evidence of enrichment of subnetworks of size 4 or greater ( $p=0.005$ ) and borderline significance of enrichment of hot subnetworks of 5 or more nodes ( $p=0.038$ ), but did not identify strong significance for subnetworks of increasing size.





## HotNet2 Community Assessment

Of networks HotNet2 identified as the highest heat, three compact subnetworks were contained above a threshold heat value of 200. The highest signal of these subnetworks contained GDF5, a known contributor to osteoarthritis and height(255). This network also contained BMP6, a protein known to be involved in growth plate function,(256) and the NOG gene, known to be involved in regulation of the BMP group of genes (two of which are also included in the hot subnet), which is involved in bone morphogenesis.(257) (**Figure 17a**)

The subnet showing the second highest amount of signal was a group of 10 genes with the highest contributor being EFEMP1, which has known developmental function.(258) This hot gene cluster is organized around MFAP2, a component of connective tissue microfibrils.(259) The major partners of MFAP2 in the enriched subcomponent are the LOXL family of genes, known to be involved in metastatic processes, and the FBLN family of genes, which have a theorized role in Marfan disease.(260) (**Figure 17b**)

The third highest subset is a group of six genes which contains GH1 and GH2, both growth hormones. The CSH family are highly signal enriched partners with known roles in infant development.(261) TEAD3 is another partner, and has been implicated in pituitary development.(262) (**Figure 17c**)

While the application of HotNet2 to genomic data is not a novel concept, this does demonstrate that subnetworks exist within the data representing compelling functional context. The knowledge of this signal enriches the available reach of PRS and its application, and can be considered an informative step of score construction.

### 3.4.6 PRETERM BIRTH PRS ASSESSMENT

In order to evaluate these methods on a small dataset with a phenotype of specific interest, we assess the ability of our network PRS method to capture signal in a small GWAS of preterm birth (1,793 individuals). We attempt to evaluate the amount of information gain possible in the DNBC preterm birth dataset using network models of PRS, both through increases in prediction and by providing additional context after score construction.

#### GWAS Evaluation

To create the dataset a GWAS was carried out on the DNBC data representing 592,839 SNPs captured using the Illumina Human660W-Quad BeadChip microarray. SNP annotation was converted from hg36 to hg37 using the LiftOver utility from UCSC genome browser utilities. Autosomal chromosomes were isolated from the dataset and used to perform association analysis. Patients identified during QC/QA as having major chromosomal dysfunction were removed.

Data was filtered before imputation using PLINK. SNPs with minor allele frequency greater than 0.05 were extracted. Variants with missing call rate greater than 0.1 were excluded and individuals with missing call rates greater than 0.01 were filtered from the dataset. Duplicate probe IDs were identified and filtered. (**Appendix Figure 7, Appendix Figure 6**)

Pre-imputation phasing of data was performed in SHAPEIT2 using an effective population size of 15,000 and a window parameter of 2 Mb. Imputation was performed in IMPUTE2 using the 1,000 Genomes Phase 3 reference panel. The resulting set of 25,920,975 imputed SNPs were filtered for a quality score greater than 0.8, representing high quality SNP

calls. The resulting data was thresholded for a call probability of 0.9 in PLINK as well as for missingness and minor allele frequency yielding a set of 5,405,340 imputed SNPs of high quality. Detailed GWAS QA/QC procedure descriptions are provided in the Appendix. (**Appendix Figure 8**)

The filtered data set was then lifted over from hg37 to hg38 in order to ensure compatibility with Ensembl gene coordinates in build 92.

Data was split into a training set of 1,344 patients (636 cases, 708 controls) and a test set of 449 patients (213 cases, 236 controls). Association was performed in the test set using PLINK assuming binary phenotype representing patients born under 37 weeks gestational age as case and patients born after 40 weeks gestational age as control.

#### **Traditional PRS Formulation**

A traditional PRS was constructed in PRSice as detailed with the HRS data set. The resulting score includes 153,222 SNPs after clumping. A best-fit threshold for SNP inclusion of 0.00055 was determined. Validation on the test set showed the score to be significant at  $p < 0.05$  threshold ( $p = 0.0295$ ) with a Nagelkerke  $R^2$  value of 0.014.

Additional fit analysis was performed using a pseudo- $R^2$  measure devised by Lee et. al. taking into account disease prevalence.(263) The Lee  $R^2$  measure uses an estimated prevalence of preterm birth of 10.8% derived from a 2017 study in a population of 1,911,757 North Carolina residents.(264) The Lee  $R^2$  value for the full SNP PRS is 0.012. (**Figure 18**)

## Early Context Model

### Inclusion of All Gene Regions

As with the HRS data, a PRS using all gene regions with SNP coverage was first created in PRSice as a baseline for the effect of network pruning. The subset represents 3.12 million SNPs or 57.78% of the total set. The full gene region model showed a Nagelkerke  $R^2$  of 0.0192 and a Lee  $R^2$  of 0.156. The model was borderline significant with a significance value of  $p=0.011$ .

**(Figure 18)**

### Connectivity Gated Model

A connectivity-gated model taking into account the top 50% of nodes by degree centrality was computed. The subset represents 1.96 million SNPs or 36.28% of all SNPs. The PRS resulted in a Nagelkerke  $R^2$  of 0.0101 and a Lee  $R^2$  of 0.0082. The model failed to achieve significance with a p-value of  $p=0.067$ . **(Figure 18)**

### Candidate Gated Model

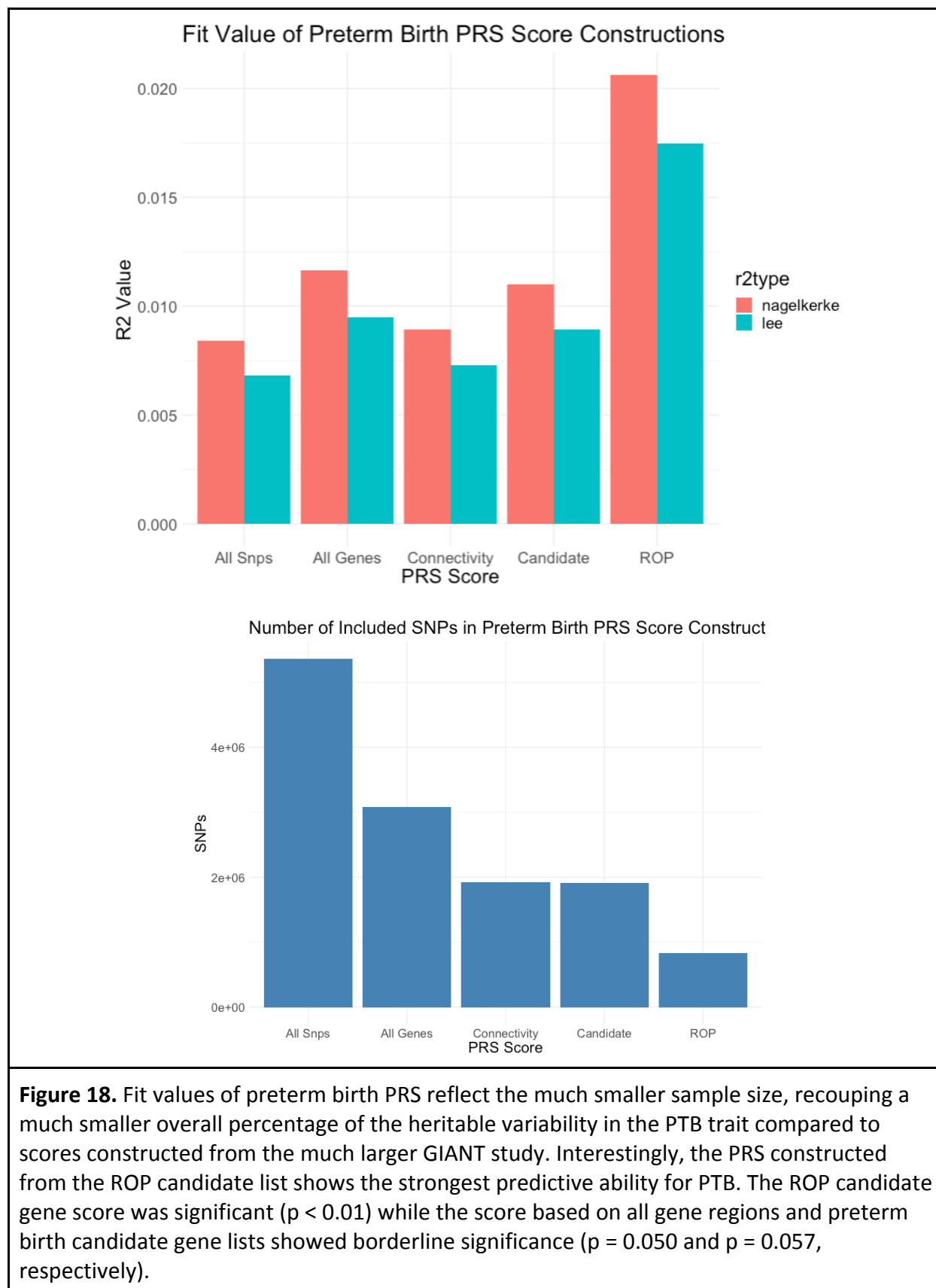
A candidate gene list containing 616 genes curated from the Database for Preterm Birth (dbPTB) was created. Genes represented in this set were selected from the graph and this selection was expanded to neighboring nodes. SNPs selected for the score represent 1.95 million SNPs or 36.02% of the total set of SNPs. The resulting subset of SNPs assigned to these genes was isolated and used to construct a PRS.

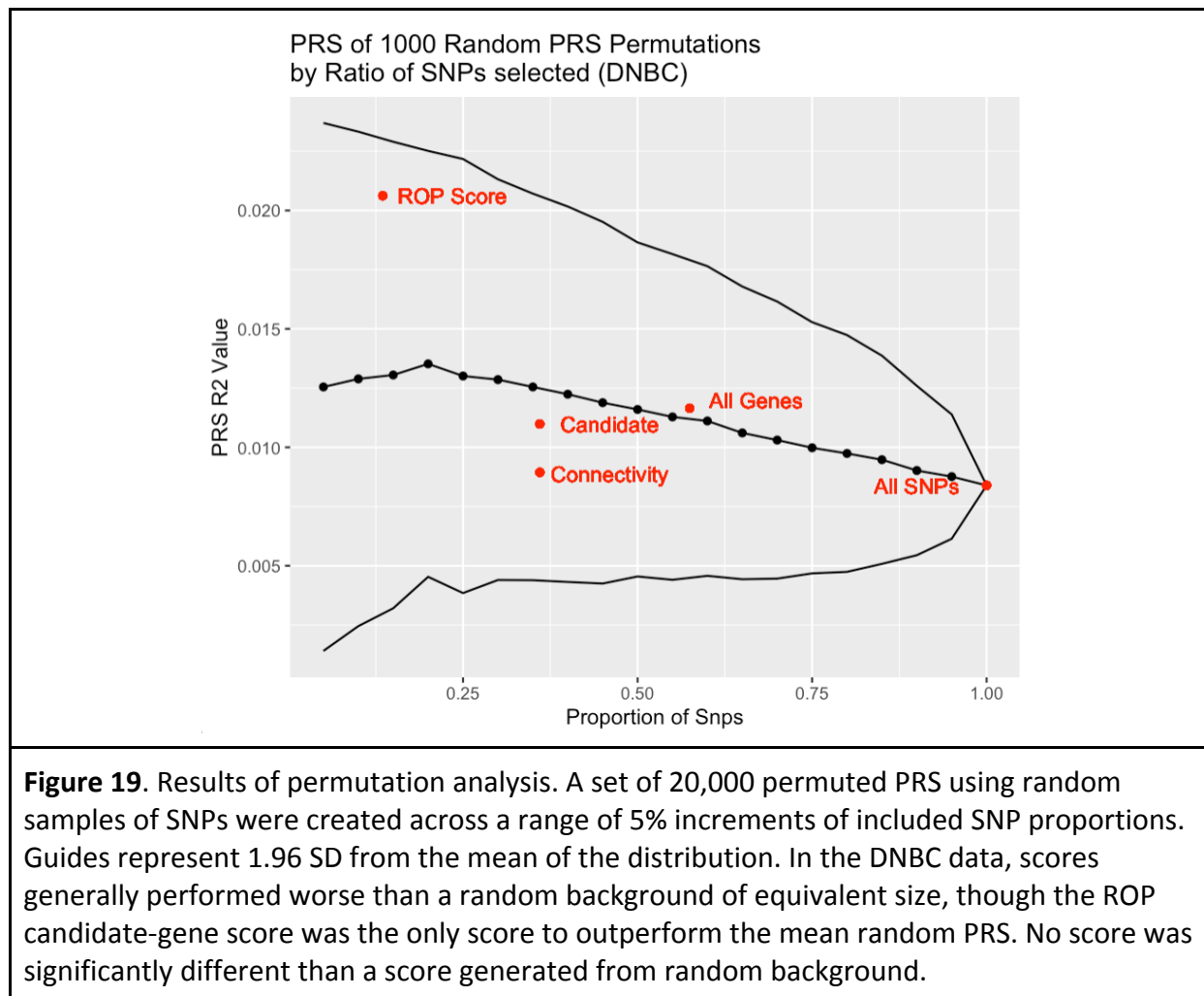
The resulting PRS had a Nagelkerke  $R^2$  of 0.0148 and a Lee  $R^2$  value of 0.0120. The model was borderline significant with a value of  $p=0.027$ . **(Figure 18)**

### ROP Candidate Gene Model

As a comparator to the PTB candidate-gene model, a second candidate-gene model was created using 125 genes with implied involvement in ROP based on the review conducted in Chapter 2. As with the PTB candidate score, these genes were selected on the graph, this selection was expanded to neighbors of these genes, and SNPs assigned to these genes were used in the construction of a score predicting PTB with the assumption that such a score's ability to predict the PTB phenotype would serve as an indicator of the strength of the interconnectedness of these two phenotypes.

The resulting PRS had a Nagelkerke  $R^2$  of 0.021 and a Lee  $R^2$  of 0.017 and was significant with a  $p < 0.01$ . **(Figure 18)** This surprising result seems to indicate that although the difference is small in absolute terms, the ROP candidate score outperforms all other PRS methods of PTB prediction evaluated. In section 3.4.7 we consider in greater detail the implications of this difference in predictive ability.





#### Permutation Analysis

In order to assess the degree to which the achieved  $R^2$  values were significant compared to a random subset of SNPs, a set of 1,000 permuted PRS using a random selection of SNPs from the full set was performed in 5% increments of SNP inclusion.

The results of permutation testing showed all scores were within 1.96 standard deviations of the mean of all permuted  $R^2$  values indicating that none were significantly different than a random subset of SNPs at a threshold of  $p < 0.05$ . The score incorporating all gene regions performed at a roughly equal level to a random subset of SNPs containing a 55%-

60% subset of random SNPs drawn from the full set without replacement. While we are able to see that the ROP score does not significantly outperform a random subset of the same size, we are able to assess that it performs better than all other scores with respect to random background. (**Figure 19**)

This finding again indicates that the role of network context in genetic heritability may be overshadowed by the need for broadly inclusive measures with full coverage of the genome.

### **Late Context Model**

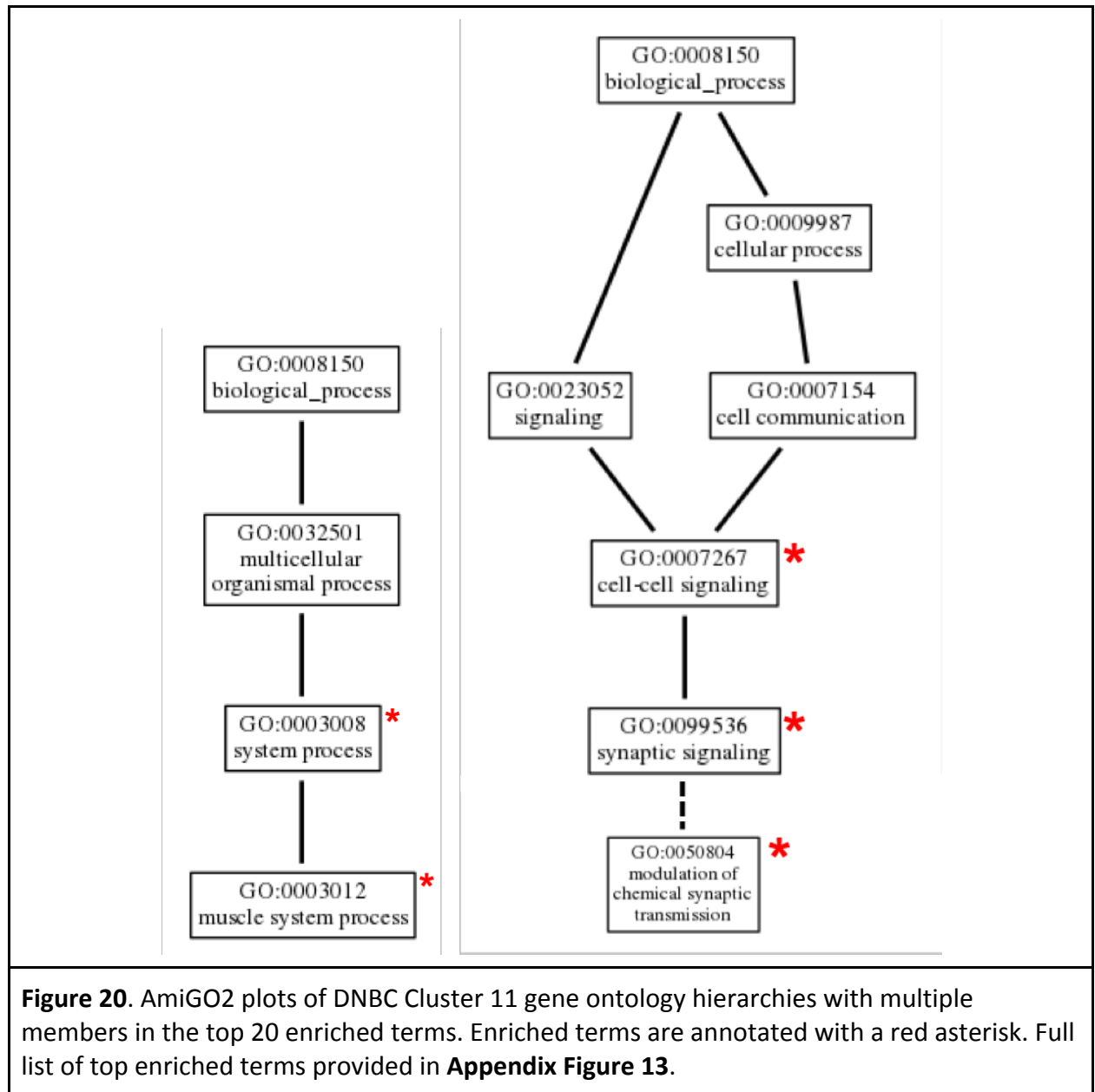
#### **Module Analysis**

As with the HRS data, a p-value threshold was applied to the DNBC graph. Due to the low significance values resulting from the small size of the study, the PRSice recommended best fit threshold of 0.0006 resulted in only 111 genes being admitted to the model. In order to achieve a connected largest component, this threshold was raised to 0.01. The resulting subset of genes represented 3,557 nodes, or 24.7% of the total nodes in the graph.

After community detection, 20 communities were distinguished, varying in size from 6 nodes to 492 nodes. Two clusters were significant at the  $p < 0.05$  level in their enrichment for effect size signal.

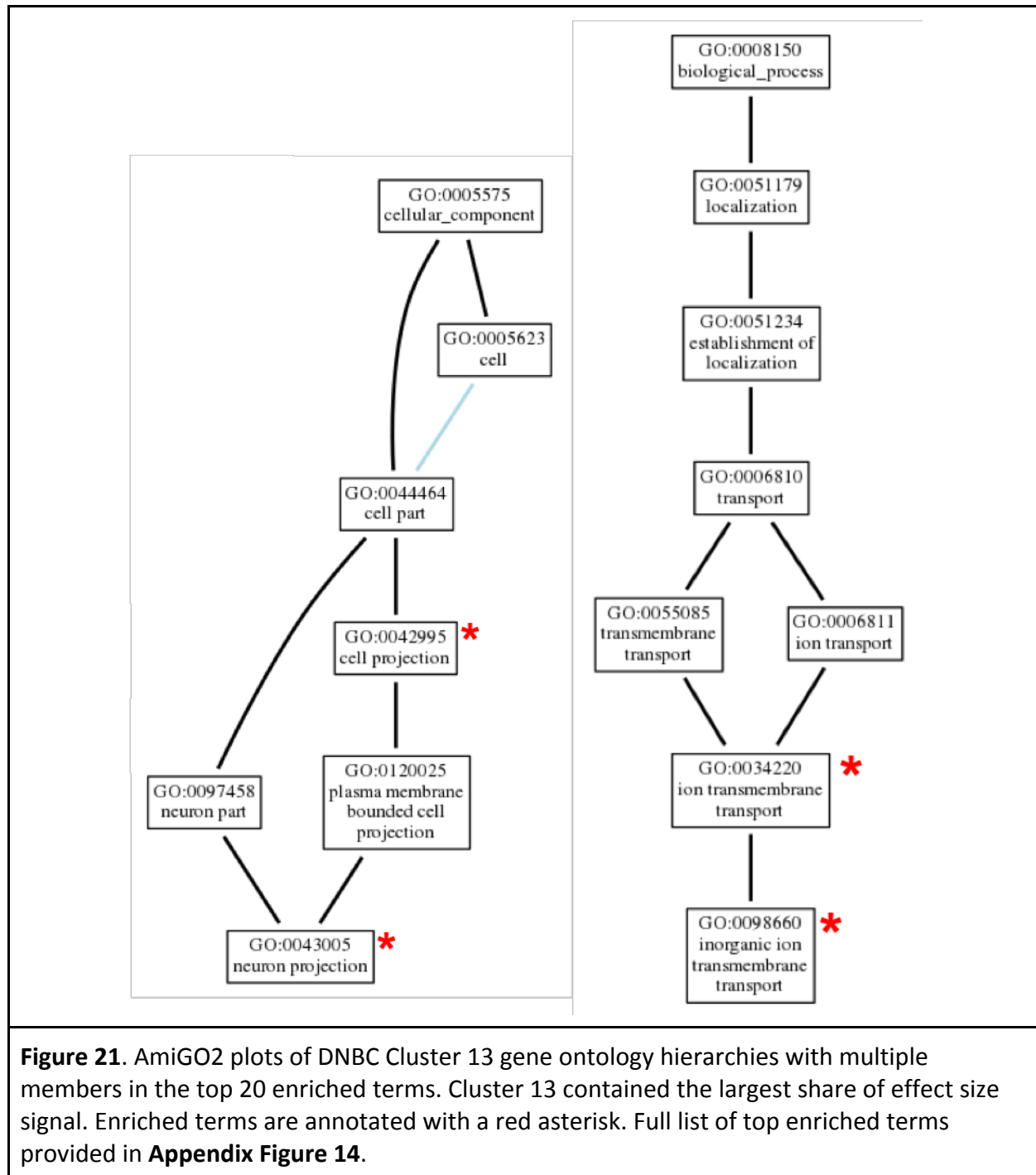
Cluster 11 includes 153 nodes accounting for 7.8% of total effect size, significant for high effect size relative to its cluster size ( $p=0.014$ ). The cluster is most enriched for cellular signalling and synaptic function, though no gene set achieves significance at an adjusted  $p < 0.05$ . Evaluation of AmiGO2 plots of nested gene ontology information show aggregation of signal in synaptic transmission and muscle system process ontologies. (**Figure 20**)





Cluster 13 includes 366 nodes accounting for 13.7% of total effect size. It does not achieve statistical significance for enrichment of effect size signal. No individual pathway reaches significance when considering adjusted p-values, but the top enriched gene sets represent cellular adhesion and catalytic activity. Investigation of AmiGO2 plots show that signal appears to be spread across several discrete systems. However, neuron projection and

ion transport show a degree of ontological enrichment, and the top performing group representing cellular adhesion may represent a clinically interesting target. (**Figure 21**)



Taken together these loose descriptive enrichments give a general picture of what causes may be contributing to preterm birth, though they ultimately represent a very small percentage of the total heritable variability that can be captured in a study of such small sample size. In addition, the degree to which information can be gleaned from a small sample study appears to be commensurately reduced when compared to the GIANT study, although communities enriched for effect size signal again appear to give more focused results when inspecting enrichment.

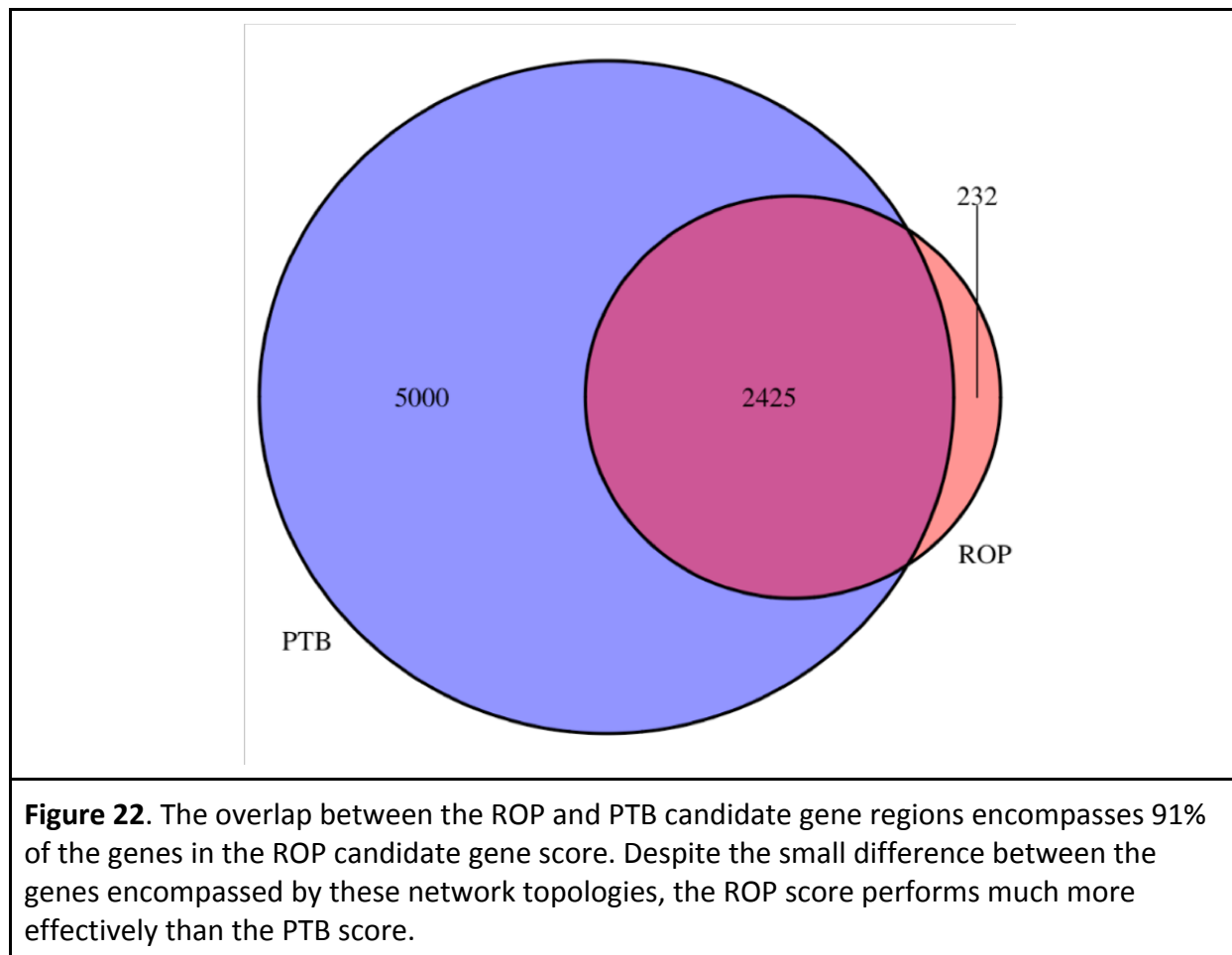
#### **HotNet2 Analysis**

HotNet analysis of the DNBC data found no subnetworks significantly enriched for signal. Due to low significance in the general GWAS, the top identified enriched subnetworks were extremely large, with the top enriched subnetwork containing 111 nodes and the smallest of the top 10 identified subnetworks was composed of 24 nodes.

#### **3.4.7 PRETERM BIRTH AND ROP OVERLAP ASSESSMENT**

In order to assess the degree to which ROP and PTB were genetically convolved, an assessment of the overlap between network topologies was considered. As the number of cases of ROP in the DNBC data set is quite small, the analysis makes use of candidate gene lists as a basis of inspecting network overlap by way of the PTB candidate gene score.

Of the 125 genes identified as candidates for ROP disease, 27 appear on the list of PTB genes curated by dbPtB. In order to assess better the degree to which these two gene sets are related, the selection of genes is expanded to include neighbors of both gene sets. The two gene sets are then inspected for the degree of overlap between the sets.



The expanded ROP cover set includes 2,657 genes (17% of the total graph) while the network cover for the PTB gene set includes 7,425 genes (49% of the total graph). The intersection of these two sets contains 2,425 genes (24% of all genes in the union of both candidate gene sets). (**Figure 22**)

In order to determine if the number of overlapping nodes was significantly different than would be expected by chance, a permutation analysis of the graph was performed with sets of randomly selected nodes of the same size as the two candidate gene sets and this

background was compared to the observed overlap. The analysis determined that the overlap between the two candidate gene groups was smaller than expected by chance ( $p < 0.01$ ).

One possibility for this result is that the clustering of genes on the graph is more compact than would be expected from a random set of SNPs is that curated gene sets may cluster in the graph, leading to the number of nodes included being smaller, and therefore the number of overlapping nodes would be expected to be smaller. In order to assess whether this significant result was due to statistically more compact regions than would be expected from a random sample due to curation of the gene sets reflecting a pattern of investigating the partners of known proteins, permutations of random gene sets the same size as the candidate gene lists were generated and the size of their cover was compared to the observed candidate gene lists. The result of this analysis found that the graph cover represented by the ROP candidate list was significantly smaller than would be expected from a random subset of SNPs ( $p < 0.01$ ), as is the cover of the dbPtB candidate graph ( $p = 0.01118$ ).

With the understanding that both subgraphs are significantly more compact than could be expected from a random graph, a further permuted analysis was undertaken to attempt to investigate whether the ratio of overlap between these two compact groups was significant, noting that most random gene sets would tend toward a larger size, making this analysis conservative compared to a random graph. The ratio of overlapping genes between the two candidate scores was shown to be higher than expected by chance by permutation testing ( $p < 0.01$ ) indicating a large degree of overlap between these two expanded candidate scores after taking into account the size of the sets and their compact size.

### Performance of ROP Candidate Gene Score vs Preterm Birth Score

To assess the degree to which the candidate gene list of ROP was predictive of PTB, a candidate gene PRS was constructed using the ROP gene list, and this score was evaluated with respect to the preterm birth trait. Surprisingly, the ROP gene-set score achieved an  $R^2$  of 0.02141 ( $p < 0.01$ ) outperformed all other PRS built to investigate PTB, including the score incorporating all captured SNPs.

As the PRSice model sets a dynamic threshold for inclusion of SNPs based on the best fit for a specific model, this threshold was investigated to determine if using a fixed p-value threshold would decrease the predictive value of the ROP score relative to the other PRS. The results of this analysis showed that a more lenient threshold significantly penalized the scores incorporating all SNPs and gene SNPs due to increased SNP inclusion while the PTB candidate gene score remained relatively stable. The ROP score suffered a drop in predictive ability and showed marginally worse performance than the candidate gene score at a threshold of  $p < 0.01$ , though it still outperformed the all-SNP and gene-SNP scores. (**Appendix Figure 16**)

Investigation of the SNP set used to construct the ROP score included 61,327 SNPs mapping to 232 genes not included in the PTB candidate gene PRS. Of these 61,327 SNPs, 36,196 assigned to 161 genes are not SNPs that appear in the PTB score due to multimapping. After filtering these SNPs for a significance value of 0.249 (the same value used for the ROP PRS best fit), 1,225 SNPs mapped to 72 genes remain.

This list of genes was analyzed according to the percentage of total effect size and the minimum significance value included in the gene region. (**Appendix Figure 15**) Top results included genes with known developmental function. USP6NL is the largest contributor of effect,

containing 198 SNPs responsible for 19% of the total effect in the ROP specific group. USP6NL is potentially interesting as it has previously been identified as having decreased methylation in studies of infant rat lung.(265) USP6NL has also been shown to play a role in developmental competence in oocytes.(266)

RBMS1 is the second highest contributor of effect, with 7% of total effect across 99 SNPs represented in the ROP specific gene group. RBMS1 is a previously known participant in estradiol release.(267) It has also been described to display differential methylation and RNA expression in human placenta.(268) FLRT2 is the fourth highest contributor of effect, with 4.7% of total effect across 53 SNPs represented in the ROP specific gene group. FLRT2 is a previously patented biomarker for preeclampsia.(269) FLRT2 has also been implied to have a role in placental development in mouse models.(270)

Adding biological context to a group of genes implied to be the most highly predictive subset of the data investigated strongly implies that this network subset has substantial involvement with the preterm birth phenotype. Further, the identification of this group of genes associated with ROP candidate genes but distinct from the group of curated dbPtB candidates suggests a strong interdependence between network topologies associated with ROP and those associated with preterm birth.

### **Enrichment of Preterm Birth for ROP Genes**

In order to further characterize the degree to which ROP and PTB candidate gene covers are related, the PTB candidate gene score was analyzed for enrichment for ROP signal using GSEA methods.

The PTB candidate gene PRS gene set was used as a ranked list with  $-\log_{10}(p)$  as the ranking statistic. When performing GSEA on the PTB score genes using the MSigDb gene ontology gene set of 1,736 gene sets, the ROP candidate gene set of 127 genes ranks 1,497 with a normalized enrichment score of 1.427 and a p-value of 0.1069 (adjusted p-value 0.5694).

The top enriched gene sets are those for the GO\_MEMBRANE\_PROTEIN\_COMPLEX ( $p=3.2e-4$ ) and GO\_REGULATION\_OF\_MEMBRANE\_POTENTIAL sets ( $p=7.7e-04$ ). Though these are the top performing gene sets, neither significance score survives adjustment (both  $p=0.19$ ).

**(Appendix Figure 17)**



### 3.5 DISCUSSION

The analysis and prediction of rare and complex diseases using GWAS faces significant challenges due to the inherent limitations of such conditions. Limited potential for recruitment is likely to remain the largest obstacle to overcome as researchers attempt to arrive at better methods to identify at-risk individuals for these severe disease phenotypes. While the attainment of large cohort sizes may be possible for diseases like preterm birth, which are estimated to affect more than 10% of the population, diseases with rare presentation like ROP will remain difficult to achieve recruitment sizes in the tens or hundreds of thousands, leading to an open question of the degree to which genetic studies will be able to predict the occurrence of such conditions.

Several limitations of the current study exist, and indeed the foremost among these is the small size of the study cohort for preterm birth. The Danish National Birth Cohort GENEVA data set represents one of the most well curated preterm birth GWAS publicly available to researchers. However, in order to adequately create predictive measures for PTB, it is possible that a cohort ten to one-hundred times the size of the DNBC is necessary.

Additionally, the importance of strong phenotyping and tracking of environmental effects is of the utmost importance. While the classification of preterm infants as those under the age of 37 weeks is a valid distinction, more extreme preterm infants may represent a group more likely to possess genetic variants of high risk. Studies of ROP have found that gestational age represents a significant risk factor in regards to disease progression, and as such extreme phenotypic groups may present one avenue by which to increase signal in genetic studies of

these disorders. (271) However, such techniques necessarily decrease sample sizes and the tradeoffs of such an approach would need to be considered carefully.

It is also worth considering the fact that only infant genetics are considered in this study. Effects relating to PTB originating from the mother's genetic code are only able to be detected from their existence in the infant genome. Further complicating this analysis is the fact that X-chromosome signal is not represented in this study, as it is removed in accordance with accepted methods of PRS construction in order to avoid complications due to the silencing of X-chromosomes in females. As PTB by definition affects one sex of parent more than another, it is of interest to include as much sex-specific signal as possible. With these limitations in mind, it may be prudent to replicate the results of this study in the DNBC cohort of mothers matched with the infant sample.

Difficulties persist in the integration of additional contextual information with traditional PRS construction. While some evidence exists that scores including gene regions and specific gene sets may be able to perform at similar levels to traditional PRS, the inability of any score to significantly outperform a random set of SNPs speaks to the importance of genomic coverage over a more focused approach. Evidence remains strong that many complex traits are driven by infinitesimal contributions from many if not the majority of SNPs in the genome.

The difficulties faced by network methods are underscored by the continued validation of an omnigenic perspective toward complex disease. Such a model places importance on the idea of numerous weak trans-acting genetic elements influencing disease through a subset of clinically relevant core genes, and these relationships are difficult to detect and categorize.(122,123) As such, network databases of PPI data may have meaningful omissions

which make the construction of network-based scores difficult. Recent efforts have been made for so-called TWAS studies in order to exhaustively interrogate transcriptomic events with the goal of supplementing such annotation, but these resources remain in a nascent state.(272)

There remain many ways by which a network-based PRS may be improved. For example, it may be possible to create a weighted measure of genomic regions, while retaining the majority of intergenic SNPs and boosting the signal of SNPs more likely to have high effect. Several preexisting gene scoring methods exist that may be useful to this end, including VEGAS and MAGMA.

It may also be possible to improve the performance of network-based PRS methods by incorporating more advanced methods of network-based pruning. While the early approaches in this paper relied on pruning based on connectivity-based manipulation of the graph, this excludes valuable context regarding the degree to which regions of the network may possess elevated significance or effect values. The strong performance of HotNet2 on the GIANT dataset indicates that such subnetworks exist and can be linked to biological functions, and methods that are computationally efficient and able to identify such regions should remain an important focus.

Candidate score treatment may also be able to be improved by the use of more fitting network modeling. In this investigation, expansion of nodes to neighboring nodes was used as a starting point. It is possible that other network methods of creating connected components may be more effective. For example, use of shortest path methods or random-walk network metrics may better represent candidate gene regions with larger genomic coverage by retaining more genomic information.

The degree to which linkage disequilibrium is considered in network methods of PRS is also likely to remain an important concern. In this analysis, LD was largely dealt with after the network pruning of the graph. It may prove more prudent to perform this step either before network construction or, as is the case in the PRSet method, as a joint step in the gene-scoring step. (273)

Late network context methods of interrogating PRS constructions may be the most interesting future path for development. In large sample sizes, there appears to be appreciable value in leveraging existing methods for the dissection of network data. The ability of HotNet2 to identify subnetworks of interest to the height phenotype is impressive, and similar methods could be leveraged in not only the interpretation of data but possibly the creation of scores.

While the application of the Louvain algorithm to detect communities is a somewhat simplistic approach, it does yield interesting summary information that may be of use to researchers who have a strong grasp of gene ontology and gene set analysis. With more complex clustering methods, it is possible that such investigations would achieve more informative results. For example, the current treatment of network pruning in the late context modularity method ignores removed paths between nodes. A possible future direction would be to impose a weighting scheme on the graph to preserve these connections instead of removing them outright, with the goal of preserving additional network structure.

The investigation of overlap in ROP and PTB is interesting for several reasons. While no score significantly outperformed the random background model, the strong performance of the ROP candidate gene score in predicting PTB relative to other score formulations was unexpected. The ability to compare dissimilar candidate gene sets and assess the difference in

information content between them is an interesting direction to consider and there is a suggestion that such investigations could help detect biological systems previously unexplored for disease effects.

It is necessary to temper interpretation of this result with several caveats. Again, the sample size of the DNBC data set is quite small by PRS standards, and the PRS evaluation underscores the small amount of genetically heritable variability that can be detected with such a population (roughly 2% of PTB outcomes). Additionally, as explained in detail in Chapter 2, the genetic effects of both ROP and PTB are unsettled. It is possible that the large increase in predictive power of the ROP candidate gene score is due to random variability in the dataset or is a function of overfit to the small test set. The inability of the HotNet2 evaluation to come to useful conclusions on the DNBC data set also supports this concern. It remains difficult to isolate systems that are strongly indicative of PTB pathology from the current data.

With those warnings in mind, these results cannot be interpreted as favorable to the idea that the genetic causes of ROP can easily be separated from those causing PTB. Based on the candidate-gene analysis, there is a suggestion that the areas of the PPI network which have been the most closely scrutinized for ROP signal are the same that contain some of the strongest predictors of PTB. Taking into account this information, it is possible that the cohort sizes for detection of ROP-specific signal may be prohibitively large given the relative rarity of the disease and the lack of well annotated cases in large consortium studies.

### 3.6 CONCLUSIONS

In this study we have evaluated two methods by which to integrate network context into the construction and analysis of PRS. This study is to our knowledge the first attempt to leverage PPI network topology to augment the construction of PRS. We present three novel strategies for performing pruning of SNPs prior to score construction, each testing unique hypotheses regarding the information content of each score. While methods of integrating network context into filtering and weighting steps of construction of PRS did not outperform a random background sample, there is some evidence that such measures may have applications in the comparison and construction of candidate gene studies, as evidenced by the ability of candidate scores to find interesting subnetworks with significant contribution to disease outcome. Integration of network context after score construction as a method of elucidating the biological motivators of such scores shows promise, with HotNet2 showing an ability to identify networks with previously described biological function. The late context modularity approach also shows some promise for the description of which biological systems provide the most potential discriminatory ability in a constructed PRS. A novel method used to separate two coincident phenotypes using the example of ROP and DNBC shows some promise as a means of quantifying the feasibility of distinguishing two simultaneously occurring conditions with theorized genetic causes and may have a role in the estimation of power calculations for such studies. Though concerns about sample size and phenotyping represent real obstacles to the development of predicting the onset of rare and complex disease conditions, the importance of addressing them remains just as pressing.

## CHAPTER 4: DISCUSSION AND CONCLUSIONS

Despite the recent success of genetic risk scoring in analyzing common disease and biometric phenotypes, it is clear that these methods face considerable challenges when applied to increasingly rare and complex conditions. There is rising awareness of the challenges of a described omnigenic model of disease, where instead of a subset of genetic elements contributing substantial disruption driving disease, every genetic marker in the genome contributes an infinitesimal fraction of signal by acting weakly through a subset of core genes.<sup>(122,123)</sup> In such a model, the ability to pinpoint the biological causes of a condition from genetic data becomes more difficult despite increasing resources for recruitment of individuals predisposed to such conditions.

In this manuscript we have considered in depth the complex traits of preterm birth and retinopathy of prematurity, two conditions which are interdependent and occur simultaneously but are theorized to have distinct pathology and root biological causes. The concerns specific to these conditions make improved methods of early detection an important focus of current research.

In an attempt to create a novel evidence-based list of candidate genes to serve as a tool for the investigation of ROP, we have conducted an extensive review of existing literature. In doing so we have documented strong clinician interest in angiogenic factors in the form of many focused single-variant studies. The success of treatments for ROP using anti-angiogenic agents like bevacizumab has strengthened the perceived role of VEGF and the gene products

with which it associates, although definitive evidence of genetic involvement has remained elusive.

Implication of other systems including those related to Norrie disease and FEVR, red blood cell production systems involving Erythropoietin, and broad inflammatory mediator systems, among other varied causes, none of which show evidence of a single strong signal, implies a complex disease etiology consistent with other investigated complex traits with varied causes like height and cardiovascular risk. It is also worth noting that many of the candidate genes identified from previous studies are the result of single-variant tests of questionable power, often taking into account only a small sample of individuals. Taking this information into account, we suggest the need for approaches leveraging increased power and propose the integration of additional context with genomic data in order to arrive at the root causes of this disease.

The literature on ROP also supports the view that its genetic causes are similar to those of PTB. IGF-I and its involvement in preterm birth as a factor correlated with birth weight serves as one example.(274,275) Similar support is found for the involvement of inflammatory agents like the interleukin family of proteins and broad cardiovascular factors like eNOS. However, the same factors have theorized roles in PTB as well.(276) The phenotype of extreme low birth weight in preterm infants is thus to some degree a chicken-and-egg situation, and one that exposes infants to a host of potential disease conditions including retinopathy of prematurity.

These findings point to the need for assessment of the degree of interdependence between similar but distinct disorders at a genetic level before recruitment in order to understand the sample sizes necessary for meaningful findings while maximizing the potential



for predictive power. In order to form a basis for answering such questions and investigate ways in which PRS may be improved, we consider network methods of supplementing both the construction and interpretation of polygenic risk scores.

Our initial attempt to discern whether additional signal can be recouped by the use of network methods yielded modest evidence that it was possible to identify network features with heightened influence. Intergenic regions can be shown to include fewer SNPs of strong significance, and highly connected gene regions can be shown to contribute higher amounts of effect size and more extreme average significance than other regions selected at random.

However, these gains seem to disappear when placed in the larger genome-wide context of a polygenic risk score. Ultimately, our investigation into the use of networks to increase the power of PRS signal underscores the importance of global network coverage over subset or focused approaches. We find that decreasing SNP coverage even in a focused manner underperforms relative to inclusion of all SNPs in most cases. Indeed, we find that such methods underperform random subsets of SNPs of similar size. Additional care is required not only to focus on variants with high contribution to disease, but also to retain the maximum amount of signal available in a complete dataset. Methods of weighting while retaining a large subset of SNPs may be a fruitful area of investigation going forward. In particular, dynamic weighting approaches which take into account complex collections of traits represented by machine learning methods like neural networks may represent one possible avenue of improvement.

While connectivity and full gene set methods approximated the level of fit imparted by traditional PRS methods, the methods used to create genewise scores employed in this

document represents only a small subset of available gene weighting techniques. Approaches like VEGAS and GSA-SNP include additional context in the formulation of gene scores beyond the region represented or minimum p-value within the region.(277,278) Incorporation of more sophisticated scoring measures while retaining intergenic SNPs may represent one possible pathway to improvement of genetic scoring metrics.

Gene set and network context methods have been another area by which researchers have sought to increase the power of risk assessment. Methods by Levine and Horvath leveraging weighted gene co-expression networks and minibatching were shown to provide advantages over traditional PRS while also providing information uncorrelated with traditional PRS formulations.(279) Additional unpublished work by Choi is also underway investigating methods of incorporating gene set analysis into PRS formulation.(273) With the successes demonstrated by these methods, it seems likely that network methods will have some part to play in the expansion of PRS methods over time.

We have also provided evidence that secondary analysis of polygenic risk scores may be one avenue by which additional information may be gleaned from genetic studies of complex traits. HotNet2 analysis of summary GWAS data associated with height was able to identify a selection of gene networks with existing support for their biological involvement in trait development despite not achieving strong significance across the full analysis. Though these methods were unable to generalize to the extremely small sample sizes of the preterm birth dataset, their success in the height data set suggests that there may be potential for future development of tools using a similar approach for deriving network context from PRS

information and their application in midsized datasets in the tens of thousands of participants range should be evaluated.

We also evaluate a novel approach for performing community detection in constructed PRS. Community-based enrichment analysis of PRS results was less clear in its ability to provide additional context to studies of complex traits, but was able to yield interesting results regarding enrichment for gene ontology terms potentially relevant to disease context. Unlike HotNet, the Louvain community detection algorithm is able to quickly digest large graphs and arrive at subnetworks reflecting graph structure. However, this speed comes at the cost of structural context reflecting the clustering of signal in specific areas of the graph, which HotNet2 provides. It is possible that more complex community detection approaches are necessary for the task of enrichment analysis.

Such methods also rely on gene set annotations which must be defined *a priori*, and as such a deep understanding of these collections is necessary in order to derive meaningful contextual information. Interactive tools for broad exploration of complex gene sets are compelling, but must be matched with domain experience in order to yield useful conclusions.

A novel approach using network-enriched candidate-gene PRS to compare competing phenotypes met with qualified success. Despite the caveat of extremely small sample size, an analysis of the differences between the network area occupied by ROP candidate gene lists and PTB candidate gene lists highlighted that the area exclusive to ROP contained significant signal for PTB. This suggests that the extraction of signal exclusive to ROP from PTB may involve recruitment of especially large cohorts relative to other complex trait studies. Future directions for method development may focus on the ability to leverage pilot studies of rare genetic

disorders to estimate their dissimilarity from other distinct disorders in order to supplement power calculations to determine if adequate recruitment is a tractable goal.

In the case of the study at hand, it is an important caveat to note the small sample size of the discovery cohort used for the analysis of PTB. Due to the large number of SNPs considered relative to the sample, it is likely that the resulting GWAS is enriched for less common variants with spurious high significance in the training set. This enrichment is likely to explain to a degree the decreasing performance of the PRS at predicting the test set as a larger proportion of SNPs is added to the score, increasing the noise captured by the model. In contrast, with adequate sample size (as with the GIANT dataset), and therefore precise estimates of effect size, it appears that a significant amount of information is lost as the number of SNPs included in the score is reduced. Likewise, the small size of the test set leads to the possibility of overfit, as all predictive SNPs may not be represented in the test data.

When considering the candidate gene score results it is also important to note that the correlation structure of SNPs is a meaningful concern. When we subtract candidate SNPs that had previously been identified as important in a disorder of interest, this forms an estimate of the degree to which those regions influence the overall score, allowing us to assess their impact. However, correlated SNPs within the same LD block may recapitulate the same signal. A refocused approach evaluating clumped tag SNPs before the use of PRSice may better reflect the degree of signal captured by these regions.

More nuanced general methods for LD clumping are also an active interest in the improvement of PRS development. Methods by Shing Wan Choi for PRSet focus on gene-set-based clumping methods in order to ensure coverage across all gene sets.<sup>(273)</sup> A similar

approach to network-based SNPs may involve clumping based on single genes or network neighborhoods in order to ensure the best clumped representation of the graph object.

Also related to the candidate gene PRS, It is a central assumption of the omnigenic model of disease that trans-acting genetic elements asserting effect on a subset of core genes of high importance to the phenotype.(123) It is possible to consider a study model in which curated candidate genes or high-performing GWAS candidates or rare variants identify core genes in an omnigenic model of disease, or one in which genes with many high significance partners may represent core genes. However, definitive identification of such relationships is reliant on underlying annotation, and trans-acting effects that to date are still notoriously difficult to discern in GWAS studies, especially those of limited size. In order to better capture such effects, large Transcriptome Wide Association (TWAS) studies may be necessary to better describe the network architecture that gives rise to trans effects as well as secondary and tertiary cis effects, which are also of interest. Several recent studies have been attempted to further describe these relationships, though the area is still actively developing.(272,280)

A variety of general concerns also impact the development of meaningful risk scores in small populations. As with gene set methods, the context derived from network methods is only as good as the underlying annotation. While network annotation of protein interactions is substantial and PPI resources continue to grow quickly year after year,(121) there is reason to believe that the overall number of protein interactions captured remains relatively low. Protein folding is a complex process. Proteins can be composed of many different subunit components, and experimental evaluation of such products can prove difficult to replicate by high-

throughput screening methods.(281) For these reasons, continued expansion of such resources is an important focus as time continues.

It is also important to acknowledge the role of phenotyping in genetic studies. While preterm birth is a well specified disorder with a clear threshold of birth prior to 37 weeks delineating its formal phenotype, coding such a disorder in binary terms likely leads to a loss of significant information. It is fair to theorize that extreme preterm patients may have different genetic contributors to disease than mild or borderline cases. Likewise, presentation of retinopathy of prematurity varies substantially between mild and severe cases, and such differences may represent the involvement of unique or varied biological systems. The recruitment of large cohorts with exhaustive phenotyping will be an imperative concern as investigations into complex disease continue. While growing resources like the UK Biobank partially address the problems of recruitment, phenotyping continues to lag behind due to the burden of expert involvement in phenotyping at scale. Going forward, it is possible that automated imaging and analysis systems may play a role in decreasing this phenotyping load on clinicians, though curation of information dense resources designed with these goals in mind from the beginning are necessary for such progress.(118)

While PRS methods are a promising tool for the detection and prediction of common complex traits, their utility in small datasets remains limited. The decrease in predictive power between the GIANT dataset, based on hundreds of thousands of individuals, to the DNBC data set of fewer than 2,000 individuals, shows the gains in performance that are possible as a result of modern consortium genotyping projects like the UK Biobank.

In order to expand the utility of scoring methods to the large array of rare diseases, it will be necessary to incorporate additional information into these metrics. Network context represents only one way in which we can incorporate supplementary information into such scores. Tissue-specific expression values may be one manner by which to increase the contextual content of PRS. As investigations of metabolomic and microbiomic involvement in disease continue to become more complex, it is likely that the resulting resources will also increase the reach of predictive methods.

Massive consortium projects represent an invaluable resource for the construction of risk scores, but the importance of adequate phenotyping for rare conditions remains. In order to continue to serve patients, these large consortiums must be diligent about collection of data to allow exhaustive phenotyping of rare diseases.

PRS studies must also be aware of the problems that population stratification poses to results. While most genetic studies are performed in Caucasian populations, black women and other non-white ethnic groups remain those at the most risk of PTB and by extension ROP, with potential increases in the incidence of these conditions in developing nations as healthcare improves to the point that extremely preterm infants are able to survive after birth. In order to address these issues, recruitment of more diverse study populations from a variety of backgrounds will continue to be an important concern.

Hand-in-hand with phenotyping, environmental effects are likely to play a significant role in the investigation of complex traits. Environmental effects have been theorized to play an important role in both ROP and preterm birth.(18,282) These complicating factors are likely to

serve as a barrier to the discovery of root genetic causes of disease unless well curated datasets are created to support such studies.

In ROP, a variety of new techniques are currently being developed using machine learning methods in order to perform automated analysis of infants shortly after birth and then project progression. (1,283,284) The difficulty in recruitment of ROP cohorts of the size necessary for genetic score construction likely remains intractable, and as machine learning methods continue to rapidly increase in capability and adoption, an imaging perspective on disease prevention may come to predominate ROP prediction over future years.

Despite these limitations, detection of at-risk individuals for many diseases remains an important goal, especially in cases where the disease is complex and outcomes have serious or fatal consequences. While imaging approaches for ROP show great promise, a host of other rare diseases exists and imaging-related solutions are not adequate for all disease types. In cases where phenotypes are primarily biochemical or subtle in nature, genetic testing will remain an important front line. Because of these considerations, the demand for minimally invasive genetic screenings that provide maximal information will persist in cases of complex disease.



# **APPENDIX**

## **HRS GWAS Evaluation**

### **Data Acquisition**

HRS SNP data was acquired through dbGaP with Institutional Review Board approval. Restricted phenotypic data was acquired after submission for protected access through HRS internal review process.

### **Initial EDA**

GWAS SNP files were provided in PLINK binary matrix format. Variant capture was performed using the Illumina HumanOmni2.5-8v1 platform, encompassing 2.5 million variants mapped to the GRCh37 (hg19) genome build.

Phenotypic data was consolidated from HRS response data from prior to the 2014 collection wave. As HRS members did not have height values in all years, the median height value in inches across all years was selected for each participant. Participant age at the time of collection of that height value and patient gender were also captured.

Height values were entered into a linear regression controlling for age and gender. The resulting residuals were scaled according to a normal distribution. The resulting phenotypic values were output to FAM format for analysis by PLINK. All conversions took place in the R software package.

## Data Cleaning

Patients were filtered for missing and unknown phenotypic status according to reference codes provided by the HRS. Patients with height values greater than 84 inches or less than 24 inches were also filtered.

Genomic data was filtered according to procedures described by Marees et. al.(285) using the PLINK software package. SNPs were first assessed for duplicate ID values, and any duplicate probes were removed. Participants were removed from the study if individuals were missing more than 2% of SNP calls. Individual SNPs missing more than 2% of calls were also excluded. Participants were analyzed for discrepancies in sex annotation and removed if any discrepancy was detected.

Autosomal SNPs were isolated and analyzed for minor allele frequency. SNPs with a  $MAF < 0.05$  were removed from the data set. Hardy-Weinberg equilibrium (HWE) was assessed and SNPs strongly deviating from HWE were removed ( $p < 1e-6$ ).

Heterozygosity of SNPs was assessed and individuals deviating by more than three standard deviations from the heterozygosity rate mean were removed.

Cryptic relatedness was assessed in individuals in order to identify sibs or offspring/parent pairs. Those with an assessed probability of relatedness greater than 20% were removed.

Participants were analyzed for population stratification using multi-dimensional scaling (MDS). A panel of data was downloaded from the 1000 Genomes Project representing 629 individuals of varying annotated ethnic backgrounds.(286) Variants missing greater than 2% of calls were removed from the set and patients missing greater than 2% of total calls were

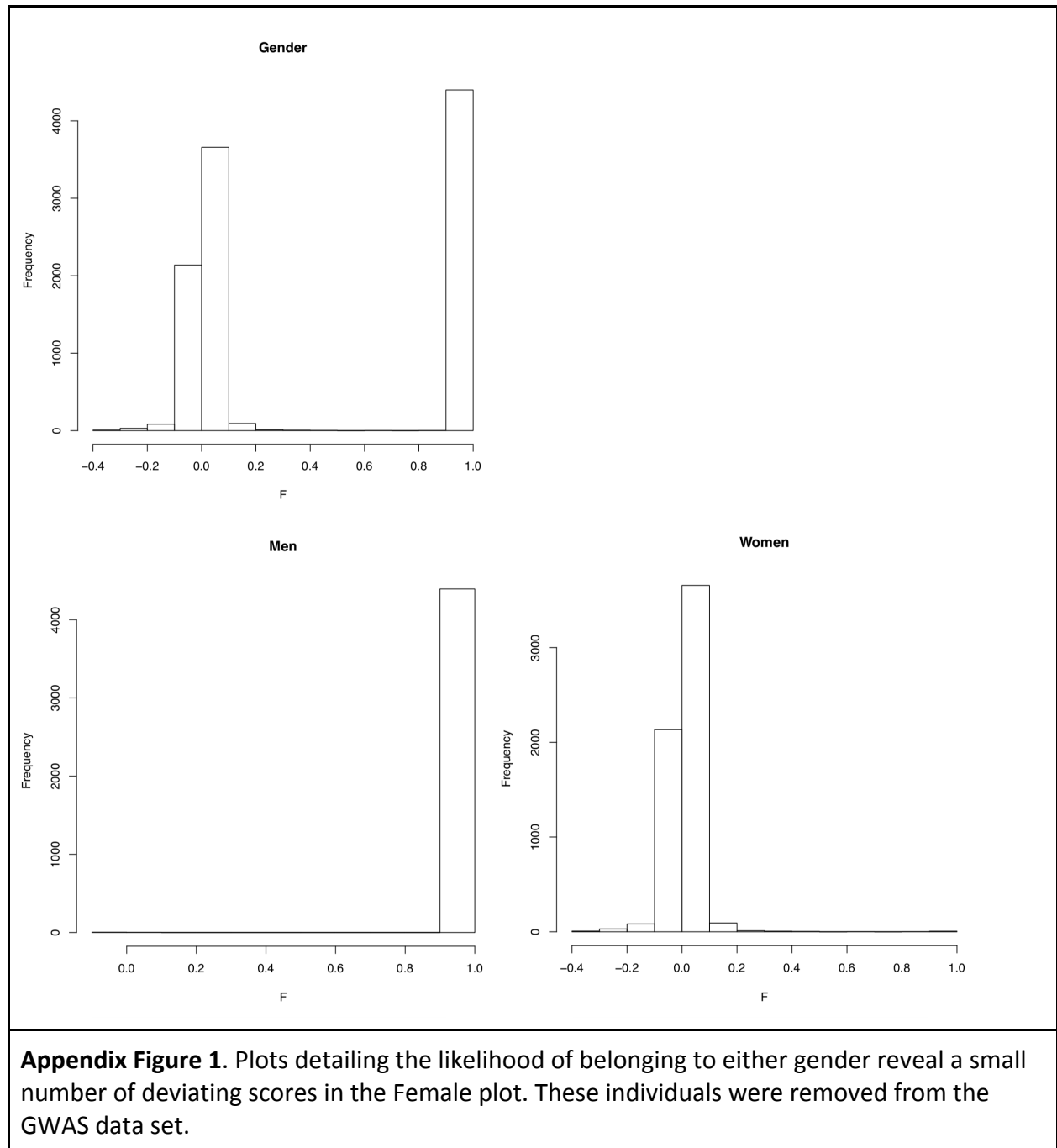
removed. As with the HRS data, SNPs with minor allele frequency greater than 0.05 were filtered from the total set.

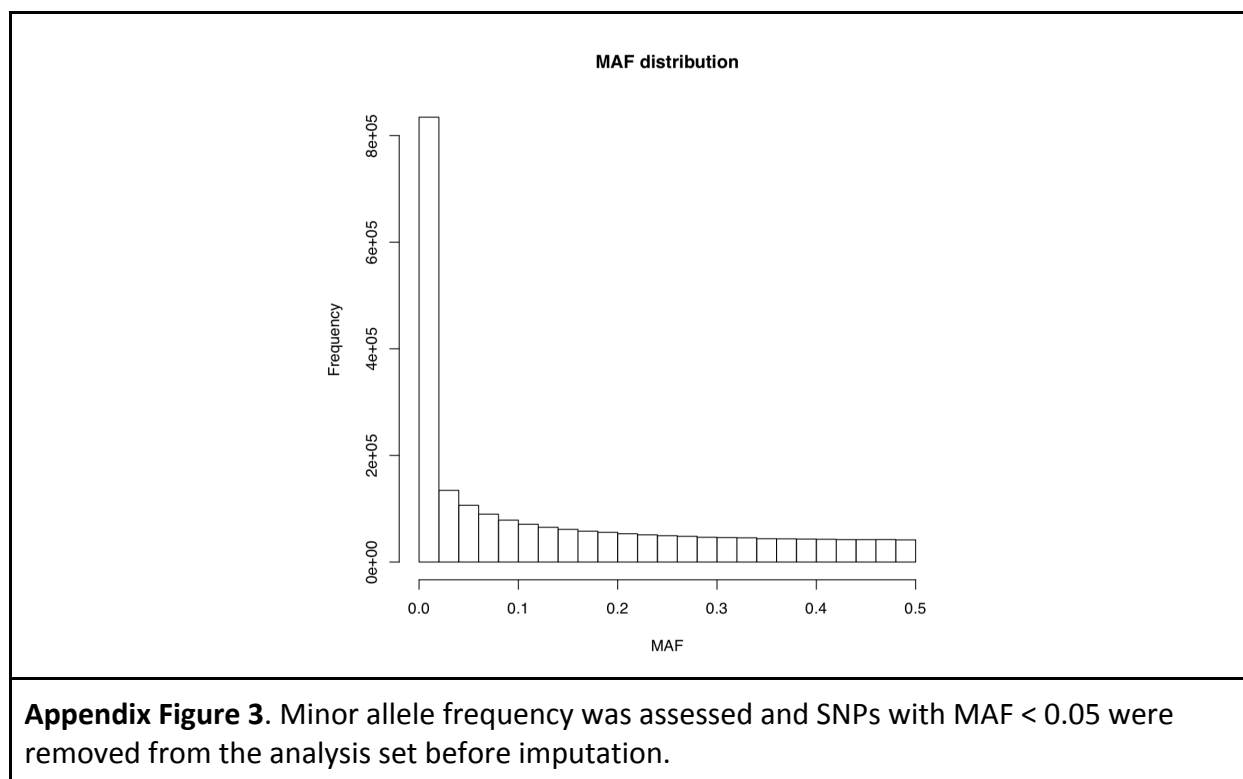
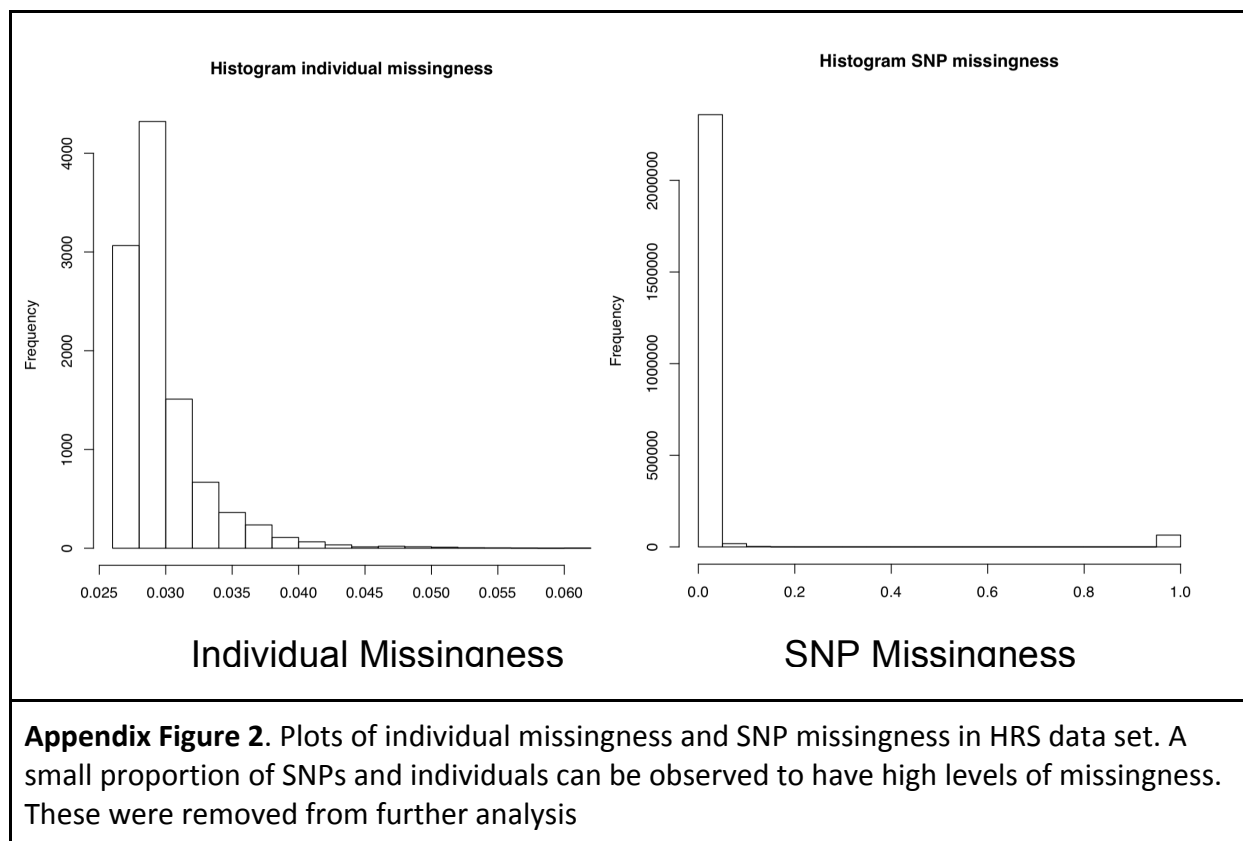
The intersection of SNPs contained in the HRS and 1000 Genomes data was taken and those variants were considered for stratification analysis. Strand positions were analyzed and flipped if in conflict with the reference build. SNPs remaining in conflict after flipping were removed.

Calculation of MDS coordinates were obtained from calculation of 10 dimensions using raw hamming distance. Cluster values were extracted for use in future analysis steps to correct for population stratification. Participants were analyzed in relation to the CEU group of 1000 Genomes participants, and those deviating from the main cluster were removed.

After QA/QC 9,746 HRS participants remained, with 1,198,351 SNPs carried forward for imputation.

## HRS GWAS QA/QC Figures

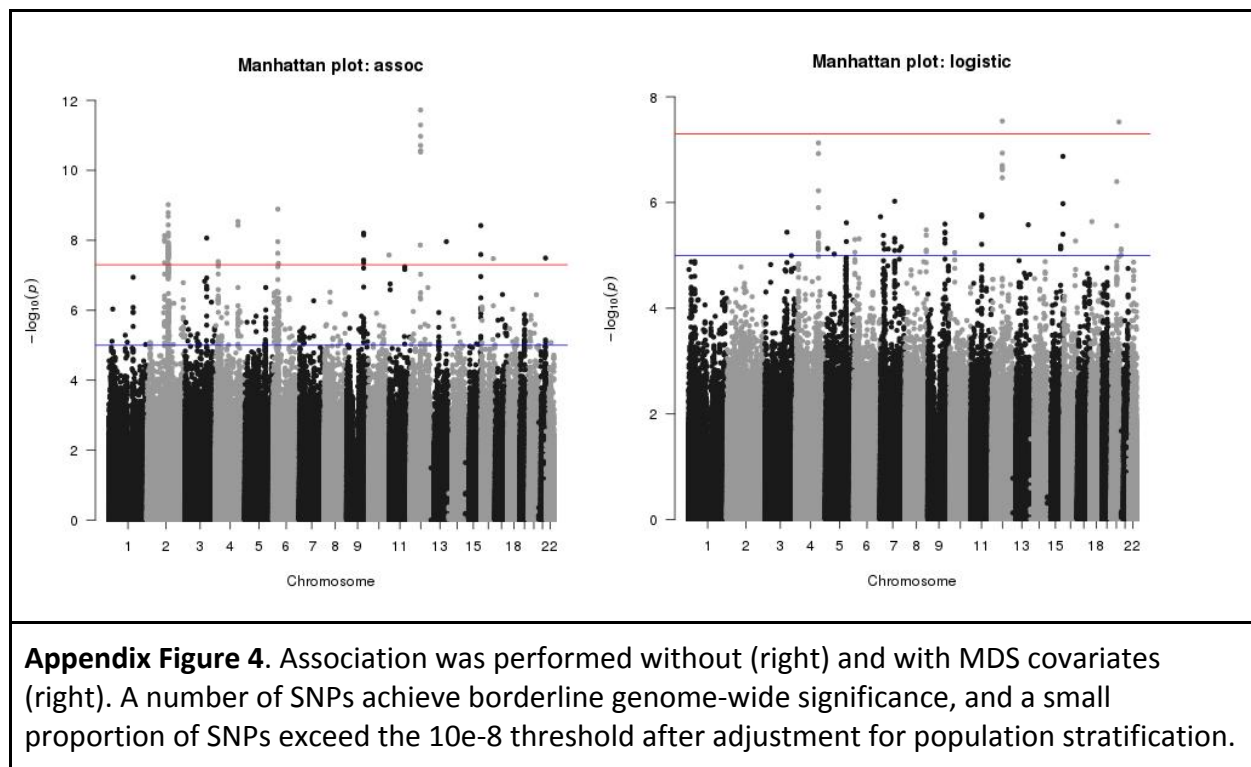




## Pre-imputation Association

For comparison to previously generated results and to assess data integrity prior to imputation, association analysis was performed on HRS individuals passing QA/QC. In order to prevent spurious association signals, the major histocompatibility complex representing the region from Chr6:2.5e6-3.5e6 was removed before testing.

Association test was performed in PLINK using ordinary least squares regression. Adjusted significance values were obtained by permutation in order to assess the impact of multiple testing on the data set. A second test was performed using linear regression with the covariates from the preceding MDS analysis. Manhattan plots and QQ-plots were created to assess the structure of the data.



## **Imputation**

Following QA/QC data was imputed to include SNPs that may have been missed by the platform as well as to assure compatibility with the GIANT study summary statistics by increasing genomic coverage.

Prephasing was performed by chromosome using SHAPEIT using an effective population size of 15000 individuals by chromosome using a reference from 1000 Genomes Phase 3 data.

Phased chromosomes were imputed in IMPUTE2 in 5 Mb chunks, avoiding centromeric regions. A 1000 Genomes Phase 3 reference panel with hg19 coordinates was used for imputation.

## **Secondary Data QA/QC**

Following imputation, SNPs with an info score of less than 0.8 assigned to them by IMPUTE2 were discarded, resulting in 7,485,299 SNPs. Filtering from the QA/QC step was repeated, resulted in a final set of 3,400,704 SNPs across 9,915 individuals.

## **DNBC GWAS EVALUATION**

Except where noted, DNBC GWAS evaluation and imputation followed the same procedure as was used for HRS data.

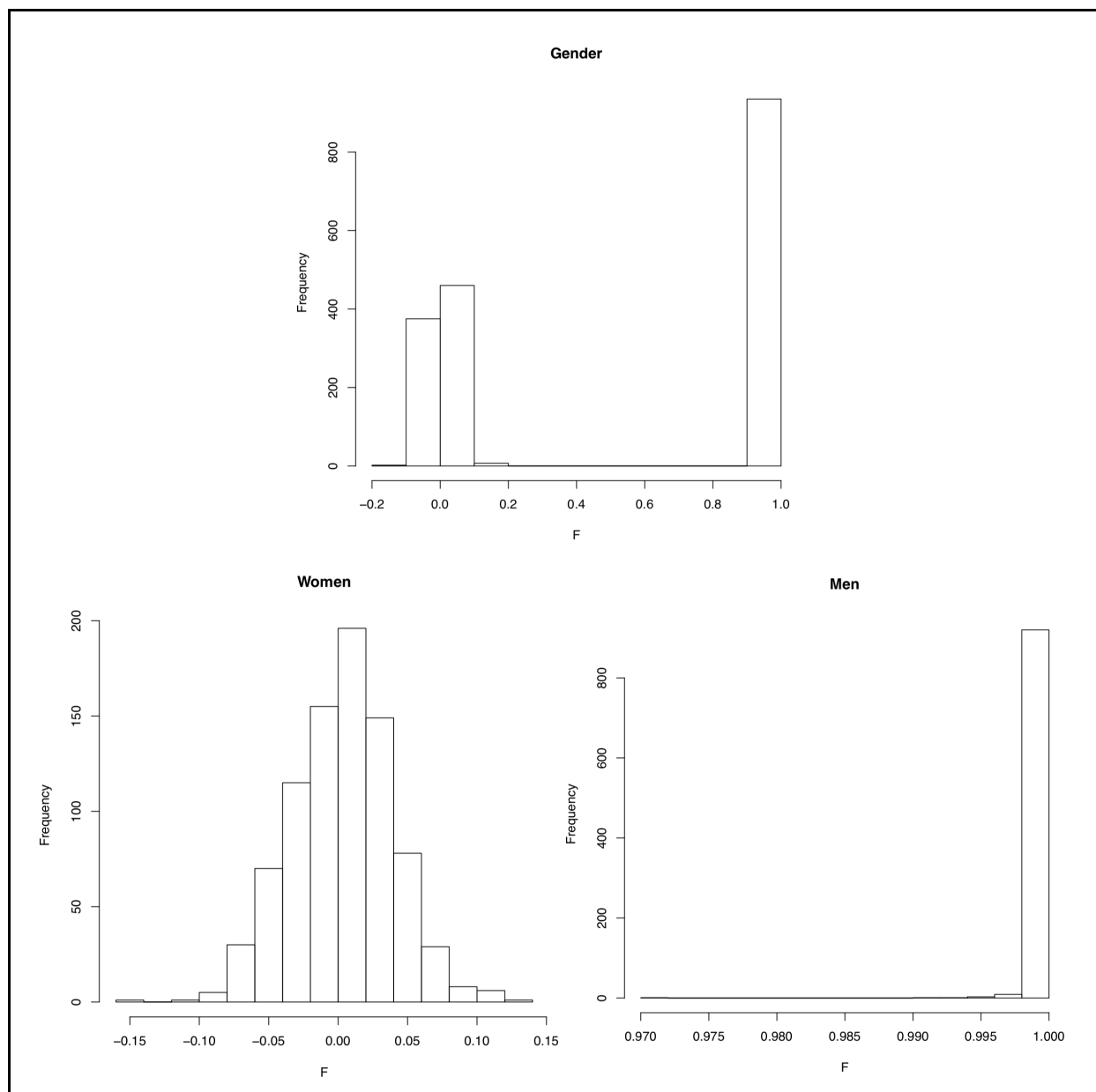
DNBC data was acquired via submission to dbGaP and data files were delivered electronically as PLINK-compatible matrices.

QA/QC was performed on DNBC data as stated. Association was performed using a Chi-squared test of association. A logistic regression using covariates derived from MDS was performed in substitution for linear regression due to the binary phenotype classification.

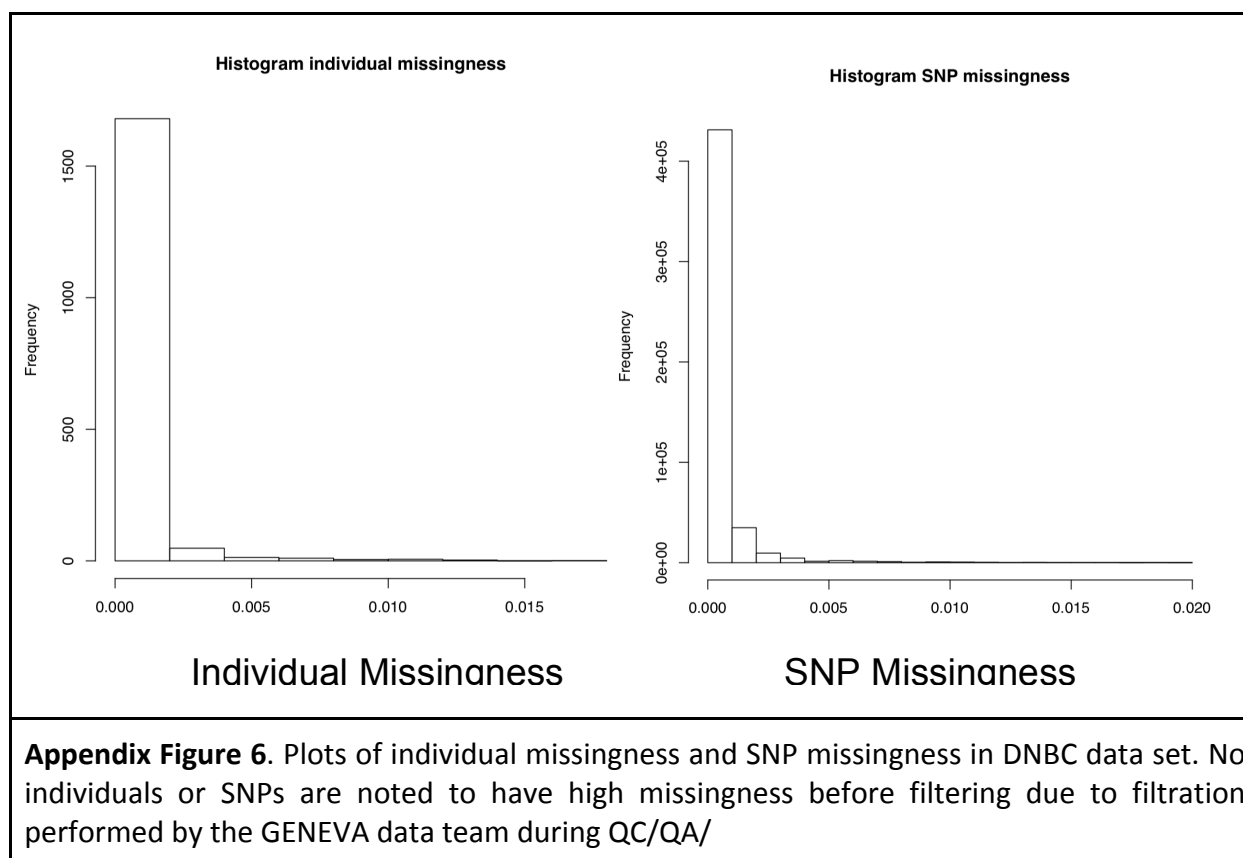
The DNBC dataset contains infant/mother pairs, and while analysis was focused on infants, imputation and phasing was performed with the full set of individuals in order to benefit from relatedness as recommended by SHAPEIT and IMPUTE2.

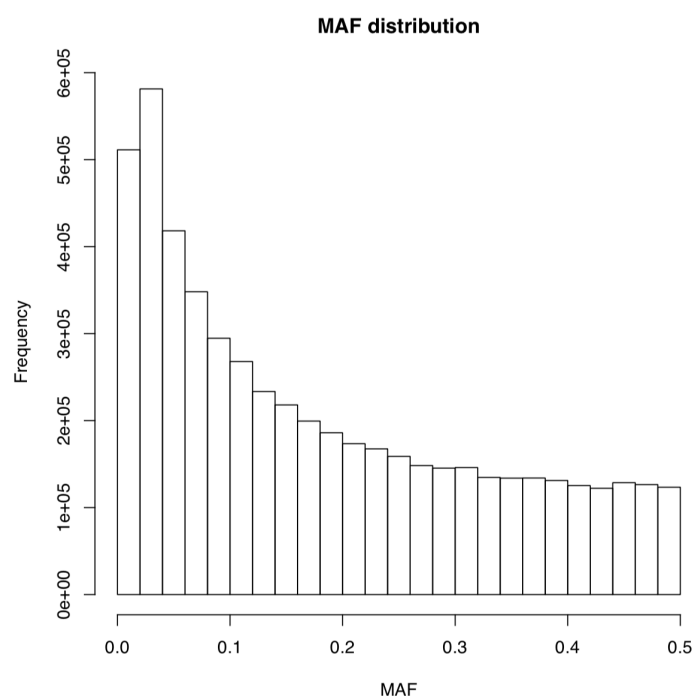
Following imputation secondary QA/QC was performed as stated above. Infants were selected from the full set prior to filtering and adults were removed. Association was performed as above, and adjusted p-values were derived prior to plotting Manhattan and QQ plots.



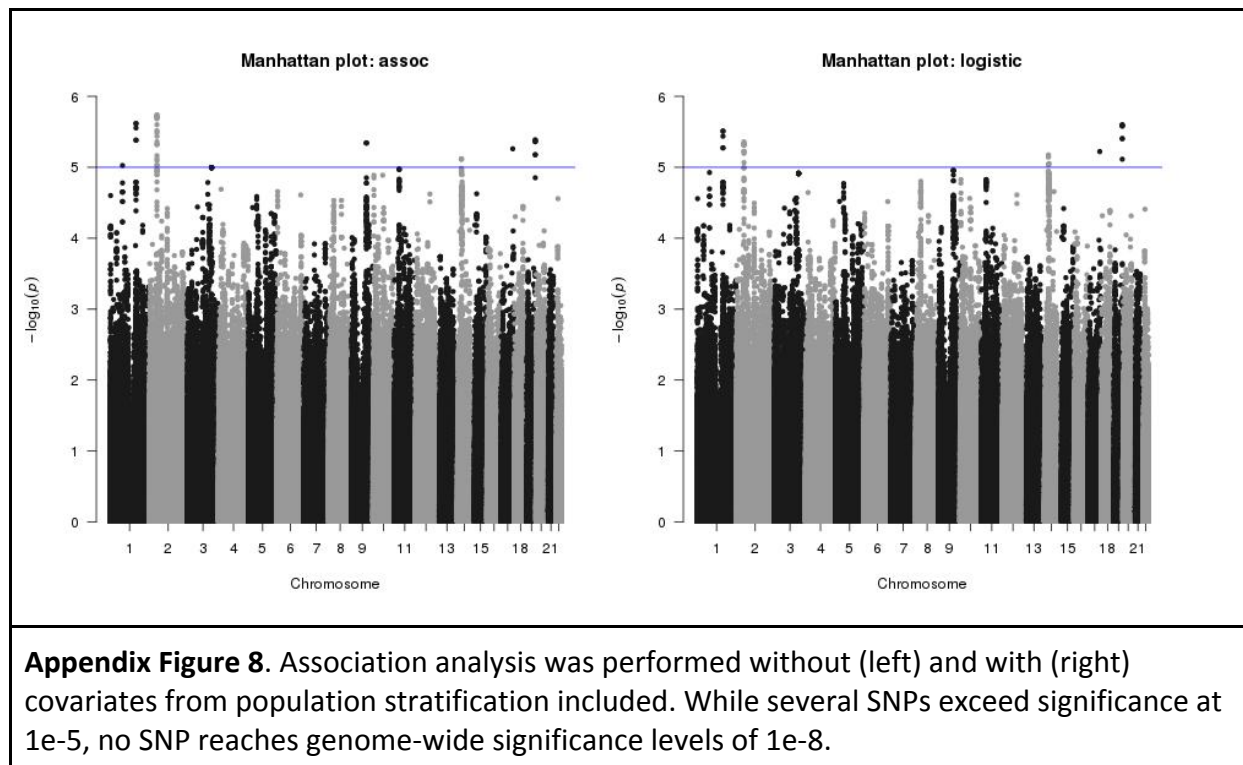


**Appendix Figure 5.** Gender checks of DNBC data showed no outliers for gender as annotated in data.





**Appendix Figure 7.** SNPs with MAF less than 0.05 were filtered from the data prior to imputation.



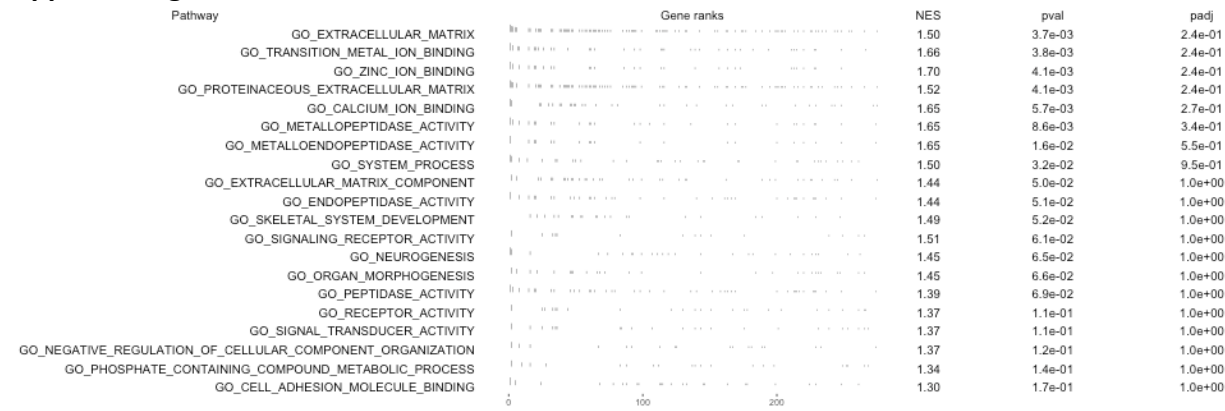
## TOP CONNECTIVITY GENES

AGTR1	CDC20	ELOC	HIST1H2BN	NFKB1	PPP2R5D	SKP1
AKT1	CDC23	EP300	HIST2H2AC	PCNA	PRPF19	SKP2
ARRB1	CDC42	ERCC2	HIST2H2BE	PLK1	PSMB8	SMURF1
ARRB2	CDK1	FBXW11	HIST2H3A	POLR2B	RACK1	SNU13
ATG7	CDK2	GSK3B	HRAS	POLR2C	RBBP4	SNW1
ATM	CDK4	GTF2F1	HSP90AA1	POLR2D	RBBP7	SOCS1
ATR	CDKN1A	H2AFV	HSPA8	POLR2E	RBX1	SOCS3
AURKA	CENPA	H2AFX	IL6	POLR2F	RELA	TNF
AURKB	CETN2	H2AFZ	JUN	POLR2G	RFC4	TOP2A
AVP	CFTR	HDAC1	KAT2B	POLR2I	RHOA	TP53
BIRC5	CHRM2	HDAC2	KAT5	POLR2J	RPA1	TRIM21
BRCA1	CREBBP	HDAC3	KEAP1	POLR2K	RPL11	UBA1
BTRC	CTNNB1	HIST1H2AC	MAD2L1	POLR2L	RPL23A	UBA52
CBL	CUL1	HIST1H2AD	MAPK14	PPIE	RPL5	UBB
CBLB	CUL3	HIST1H2AJ	MNAT1	PPP2CA	RPS14	UBC
CCNA2	CXCL8	HIST1H2BA	MRPS7	PPP2CB	RPS18	UBE2C
CCNB1	DVL2	HIST1H2BD	MYC	PPP2R1A	RPS27	UBE2D1
CCND1	EGF	HIST1H2BJ	NCBP1	PPP2R1B	RPS27A	UBE2E1
CCNH	EGFR	HIST1H2BK	NCBP2	PPP2R2A	RPS4Y1	UBE2I
CDC16	EIF4A3	HIST1H2BL	NCOR1	PPP2R5A	RPSA	UBE2N

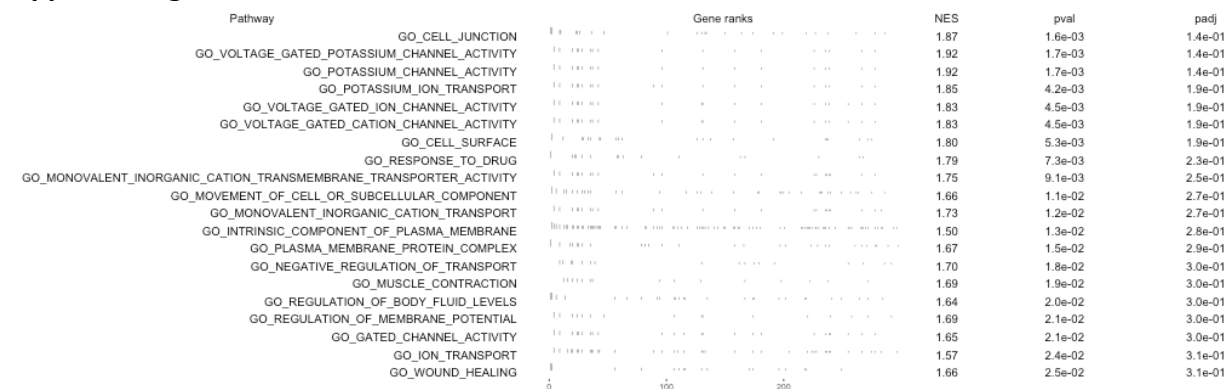
**Appendix Figure 9.** Of the 15,154 nodes contained in the full STRING network graph, 151 genes were in the top 5% for eigenvector centrality, degree centrality, and betweenness. Analysis of graph subsets found this highly connected group of proteins to be enriched for significance values relative to less connected genes. The presence of major histocompatibility complex proteins should be noted, as these will not be carried through for most PRS constructions, including our experiment.

## FULL SETS OF ENRICHED GENE ONTOLOGY TERMS

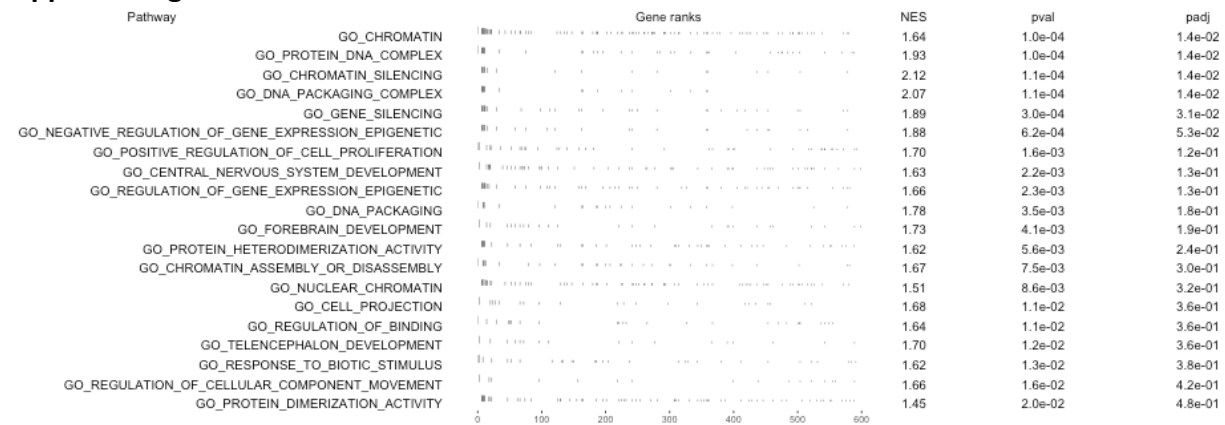
**Appendix Figure 10: GIANT Cluster 15**























**Appendix Figure 11: GIANT Cluster 16**



**Appendix Figure 12: GIANT Cluster 2**



### Appendix Figure 13: DNBC Cluster 11

Pathway	Gene ranks	NES	pval	padj
GO_CELL_CELL_SIGNALING		1.89	8.3e-03	7.1e-01
GO_SYNAPTIC_SIGNALING		1.80	1.6e-02	7.1e-01
GO_MUSCLE_SYSTEM_PROCESS		1.82	1.6e-02	7.1e-01
GO_REGULATION_OF_MEMBRANE_POTENTIAL		1.76	2.1e-02	7.2e-01
GO_REGULATION_OF_TRANSMEMBRANE_TRANSPORT		1.59	4.8e-02	8.2e-01
GO_RECEPTOR_BINDING		1.59	4.9e-02	8.2e-01
GO_POSTSYNAPTIC_MEMBRANE		1.55	5.9e-02	8.2e-01
GO_CHEMICAL_HOMEOSTASIS		1.51	7.5e-02	8.2e-01
GO_SYSTEM_PROCESS		1.51	7.6e-02	8.2e-01
GO_RESPONSE_TO_ENDOGENOUS_STIMULUS		1.49	8.0e-02	8.2e-01
GO_MOLECULAR_FUNCTION_REGULATOR		1.45	9.1e-02	8.2e-01
GO_SYNAPTIC_MEMBRANE		1.45	9.6e-02	8.2e-01
GO_HOMEOSTATIC_PROCESS		1.42	1.1e-01	8.2e-01
GO_REGULATION_OF_CATION_TRANSMEMBRANE_TRANSPORT		1.41	1.1e-01	8.2e-01
GO_REGULATION_OF_ION_TRANSPORT		1.39	1.2e-01	8.2e-01
GO_GATED_CHANNEL_ACTIVITY		1.39	1.2e-01	8.2e-01
GO_PLASMA_MEMBRANE_PROTEIN_COMPLEX		1.39	1.2e-01	8.2e-01
GO_MODULATION_OF_SYNAPTIC_TRANSMISSION		1.35	1.4e-01	8.2e-01
GO_POSTSYNAPSE		1.33	1.6e-01	8.2e-01
GO_ESTABLISHMENT_OF_LOCALIZATION_IN_CELL		1.32	1.6e-01	8.2e-01

### Appendix Figure 14: DNBC Cluster 13

Pathway	Gene ranks	NES	pval	padj
GO_NEGATIVE_REGULATION_OF_CELL_ADHESION		1.90	9.2e-03	9.9e-01
GO_SH3_DOMAIN_BINDING		1.84	1.3e-02	9.9e-01
GO_NEGATIVE_REGULATION_OF_CATALYTIC_ACTIVITY		1.83	1.5e-02	9.9e-01
GO_NEGATIVE_REGULATION_OF_MOLECULAR_FUNCTION		1.76	2.2e-02	9.9e-01
GO_CELL_PROJECTION		1.62	3.8e-02	9.9e-01
GO_INORGANIC_ION_TRANSMEMBRANE_TRANSPORT		1.59	4.9e-02	9.9e-01
GO_ANCHORING_JUNCTION		1.59	4.9e-02	9.9e-01
GO_NEURON_PART		1.53	6.5e-02	9.9e-01
GO_INTRACELLULAR_VESICLE		1.52	6.7e-02	9.9e-01
GO_NEURON_PROJECTION		1.48	8.3e-02	9.9e-01
GO_GOLGI_APPARATUS_PART		1.48	8.4e-02	9.9e-01
GO_LYTIC_VACUOLE		1.48	8.5e-02	9.9e-01
GO_VACUOLE		1.46	8.8e-02	9.9e-01
GO_NEGATIVE_REGULATION_OF_NERVOUS_SYSTEM_DEVELOPMENT		1.46	8.9e-02	9.9e-01
GO_ION_TRANSMEMBRANE_TRANSPORT		1.47	8.9e-02	9.9e-01
GO_HYDROLASE_ACTIVITY_ACTING_ON_ACID_ANHYDRIDES		1.45	9.4e-02	9.9e-01
GO_PROTEIN_DOMAIN_SPECIFIC_BINDING		1.46	9.5e-02	9.9e-01
GO_NEGATIVE_REGULATION_OF_NEURON_DIFFERENTIATION		1.46	9.6e-02	9.9e-01
GO_CELL_SUBSTRATE_JUNCTION		1.45	9.7e-02	9.9e-01
GO_VACUOLAR_PART		1.43	1.0e-01	9.9e-01

## FULL LIST OF GENES EXCLUSIVE TO ROP CANDIDATE GENE SCORE

Ensembl Gene Identifier	HGNC Symbol	#SNPs Included	Minimum p-value	Total Effect Size	%Total Effect Size
ENSP00000277575	USP6NL	198	5.31E-06	273.2346	0.188627206
ENSP00000294904	RBMS1	99	7.72E-05	102.6837	0.07088758
ENSP00000347546	BOC	74	0.001124	90.3197	0.062352106
ENSP00000332879	FLRT2	53	6.77E-06	68.6761	0.04741047
ENSP00000441802	PCDH7	78	0.0008061	59.9339	0.041375302
ENSP00000420381	CALU	43	0.003036	59.424	0.041023293
ENSP00000430333	SLIT3	49	0.005861	55.7825	0.038509388
ENSP00000342626	EYA1	52	6.50E-05	49.1472	0.033928716
ENSP00000401502	SIGLEC6	28	0.01164	43.518	0.030042604
ENSP00000377717	ST3GAL6	45	0.006287	36.6314	0.025288447
ENSP00000357986	PLEKHA1	27	0.001341	34.4275	0.023766987
ENSP00000285311	DKK2	37	0.002487	32.1802	0.022215566
ENSP00000328251	RIOX2	27	0.0002255	30.7427	0.021223189
ENSP00000381739	WNT8A	25	0.01178	30.5824	0.021112526
ENSP00000438468	HHAT	29	0.01041	30.0675	0.020757066
ENSP00000331040	OLIG2	24	0.004137	29.33	0.020247933
ENSP00000408464	DOCK8	18	0.001058	28.3645	0.019581401
ENSP00000205061	GLG1	23	0.01207	27.5792	0.01903927
ENSP00000332773	IGDCC3	16	0.004313	22.024	0.015204244
ENSP00000422591	SLIT2	17	0.00315	21.69	0.014973668
ENSP00000215909	LGALS1	17	0.01487	20.2522	0.013981084
ENSP00000305595	B3GNT2	14	0.005689	19.497	0.013459733
ENSP00000273221	IQSEC1	18	0.007923	18.4275	0.012721404
ENSP00000364028	ECE1	20	0.003988	18.335	0.012657547
ENSP00000405708	CCDC39	17	0.004709	18.1381	0.012521617
ENSP00000357196	PTPRK	13	0.01192	17.528	0.012100436
ENSP00000346829	WLS	24	0.01236	16.9501	0.011701483
ENSP00000339168	COLEC11	12	0.003041	15.48	0.010686601
ENSP00000352144	B4GALT4	11	0.002107	14.616	0.01009014
ENSP00000356016	CR1	13	0.007123	13.7488	0.009491469
ENSP00000339730	THBS4	12	0.01658	10.4269	0.007198199
ENSP00000282849	ADAMTS18	7	0.005817	9.804	0.006768181
ENSP00000344546	ARID1B	8	0.01495	9.1998	0.006351072
ENSP00000359305	TMED5	12	0.01678	9.1404	0.006310065
ENSP00000296575	HHIP	7	0.01603	7.9966	0.005520444
ENSP00000263388	NOTCH3	6	0.003058	6.8123	0.004702864



Ensembl Gene Identifier	HGNC Symbol	#SNPs Included	Minimum p-value	Total Effect Size	%Total Effect Size
ENSP00000443824	ZNRF3	9	0.002424	6.4606	0.004460068
ENSP00000307479	ARNT2	4	0.002501	5.1216	0.003535691
ENSP00000417038	PHACTR2	4	0.01179	4.934	0.003406181
ENSP00000218758	ACP5	6	0.01274	4.7969	0.003311535
ENSP00000362524	ANGPTL2	4	0.0128	4.5398	0.003134046
ENSP00000354458	C8A	5	0.01204	4.517	0.003118306
ENSP00000360281	C8B	5	0.01204	4.517	0.003118306
ENSP00000227495	ST3GAL4	5	0.0001987	4.3025	0.002970226
ENSP00000306459	B4GALT6	3	0.01711	3.674	0.002536342
ENSP00000412060	GREB1L	3	0.01975	3.62	0.002499063
ENSP00000220772	SFRP1	3	0.01972	3.599	0.002484566
ENSP00000322956	MFAP3	5	0.01014	3.5008	0.002416773
ENSP00000268124	POLG	3	0.001861	2.9099	0.002008846
ENSP00000357980	HTRA1	2	0.004095	2.841	0.001961281
ENSP00000436682	ARMS2	2	0.004095	2.841	0.001961281
ENSP00000265447	ANXA11	2	0.01043	2.774	0.001915028
ENSP00000356024	CR2	2	0.007123	2.598	0.001793526
ENSP00000351602	FUT4	2	0.01837	2.554	0.001763151
ENSP00000266066	SFRP5	2	0.01531	2.491	0.001719659
ENSP00000302936	DYNLRB2	2	0.02014	2.45	0.001691355
ENSP00000271588	HMCN1	2	0.01626	2.447	0.001689284
ENSP00000226284	IBSP	3	0.02016	2.4103	0.001663948
ENSP00000313875	CD46	2	0.02014	1.6145	0.001114568
ENSP00000255266	PDE6A	2	0.01729	1.5274	0.001054439
ENSP00000373347	SRGAP3	2	0.007039	1.4954	0.001032348
ENSP00000360519	RBP4	2	0.01026	1.4618	0.001009152
ENSP00000290894	SHF	1	0.0182	1.377	0.00095061
ENSP00000229922	CAP2	1	0.01607	1.303	0.000899525
ENSP00000294635	LRRCS3	1	0.01147	1.252	0.000864317
ENSP00000405176	TWIST2	1	0.01582	1.209	0.000834632
ENSP00000351605	FZD6	1	0.02239	0.8304	0.000573266
ENSP00000309714	SH3PXD2B	1	0.0144	0.8265	0.000570573
ENSP00000242728	BHLHE41	1	0.02117	0.8125	0.000560908
ENSP00000359221	RWDD3	1	0.02115	0.7811	0.000539232
ENSP00000419446	ADAM9	1	0.01393	0.7504	0.000518038
ENSP00000347198	SRGAP1	1	0.01658	0.7107	0.000490631

**Appendix Figure 15.** Genes appearing in the ROP candidate score but not included in the preterm birth score were found to have been previously identified as having gestation and development related functions. The ROP candidate gene score was the most predictive of all scores evaluated, outperforming even the score including all genomic SNPs.

## CANDIDATE GENE SCORES WITH FIXED AND UNFIXED THRESHOLDS

### PRS Results

Project	PRS	Nagelkerke R <sup>2</sup>	Lee R <sup>2</sup>	Sig	SNPs	Score SNPs	Threshold
GIANT	allSnps	0.177329		0	2,531,835	3375	0.00020
GIANT	allGenes	0.163862		0	1,462,671	2766	0.00025
GIANT	Connectivity	0.150591		0	905,153	3000	0.00025
GIANT	Candidate	0.111333		2.00E-256	261,338	3134	0.03400
DNBC	allSnps	0.00839188	0.006832	0.0958422	4,040,626	478	0.00985
DNBC	allGenes	0.0116527	0.009499	0.04992	3,315,137	243	0.0009
DNBC	Connectivity	0.00894214	0.007284	0.0854315	1,443,259	155	0.0009
DNBC	Candidate	0.010985	0.008942	0.0570404	1,438,785	2826	0.0231
DNBC	ROP	0.0206217	0.017491	0.00811833	629,022	1436	0.0111

### Fixed Threshold PRS Results

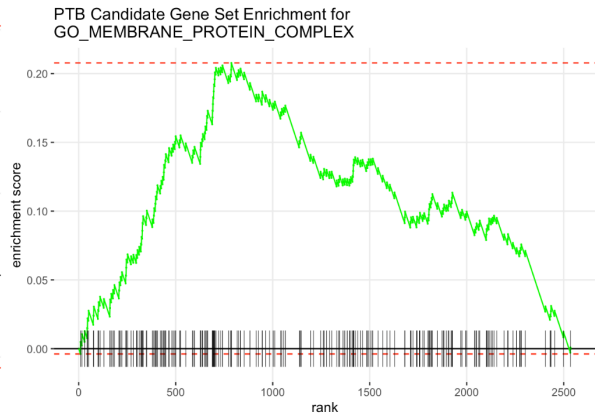
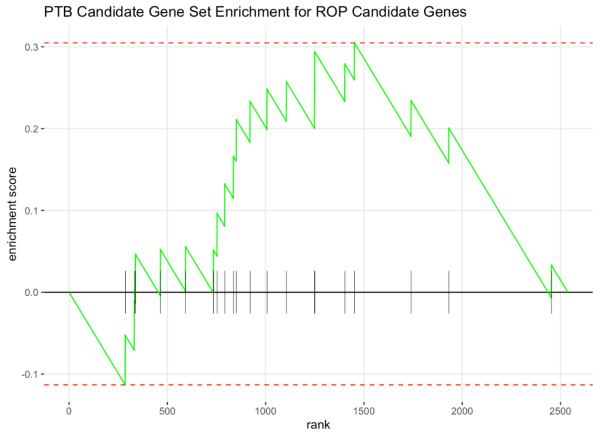
Project	PRS	Nagelkerke R <sup>2</sup>	Lee R <sup>2</sup>	Sig	SNPs	Score SNPs	Threshold
DNBC	allSnps	~1e-5		0.880597	4,040,626	3,381	0.01
DNBC	allGenes	~1e-6		0.984968	2,315,137	2,177	0.01
DNBC	Connectivity	~1e-5		0.885069	1,443,259	1366	0.01
DNBC	Candidate	0.010985		0.287942	1,438,785	1403	0.01
DNBC	ROP	0.0088		0.0882257	629,022	647	0.01

**Appendix Figure 16.** Comparison of score calculations with dynamic thresholding (above) and fixed thresholding (below) shows that enforcing a fixed threshold across all scores penalizes all values, but more severely penalizes network context scores using PTB context than those using context from ROP, which remains the most significant of the score set. Lee R<sup>2</sup> not calculated for fixed results.

ROP ENRICHMENT FOR PTB GENES

PTB Candidate Gene Top Enriched Gene Sets

Pathway	Gene ranks	NES	pval	padj
GO_PLASMA_MEMBRANE_PROTEIN_COMPLEX		2.23	3.2e-04	1.9e-01
GO_REGULATION_OF_MEMBRANE_POTENTIAL		2.19	7.7e-04	1.9e-01
GO_MULTICELLULAR_ORGANISMAL_SIGNALING		2.19	8.4e-04	1.9e-01
GO_MEMBRANE_PROTEIN_COMPLEX		2.01	9.2e-04	1.9e-01
GO_AUTOPHAGY		2.08	1.2e-03	1.9e-01
GO_CELL_CELL_CONTACT_ZONE		2.11	1.3e-03	1.9e-01
GO_MACROAUTOPHAGY		2.08	1.5e-03	1.9e-01
GO_REGULATION_OF_POSTSYNAPTIC_MEMBRANE_POTENTIAL		2.11	1.5e-03	1.9e-01
GO_SYNAPSE		1.98	1.5e-03	1.9e-01
GO_GATED_CHANNEL_ACTIVITY		2.07	1.9e-03	1.9e-01
GO_CELLULAR_RESPONSE_TO_EXTERNAL_STIMULUS		-1.98	5.8e-03	2.4e-01
GO_TRANSCRIPTION_ELONGATION_FROM_RNA_POLYMERASE_II_PROMOTER		-2.10	4.2e-03	2.2e-01
GO_REGULATION_OF_LEUKOCYTE_PROLIFERATION		-2.09	2.7e-03	2.0e-01
GO_CYTOKINE_RECEPTOR_BINDING		-2.17	2.7e-03	2.0e-01
GO_TRANSCRIPTION_FACTOR_COMPLEX		-2.03	2.4e-03	2.0e-01
GO_NUCLEAR_TRANSCRIPTION_FACTOR_COMPLEX		-2.35	1.9e-03	1.9e-01
GO_RESPONSE_TO_INTERFERON_GAMMA		-2.15	1.3e-03	1.9e-01
GO_CELLULAR_RESPONSE_TO_INTERFERON_GAMMA		-2.15	1.3e-03	1.9e-01
GO_POSITIVE_REGULATION_OF_LEUKOCYTE_PROLIFERATION		-2.72	4.9e-04	1.9e-01
GO_RNA_POLYMERASE_II_TRANSCRIPTION_FACTOR_COMPLEX		-2.66	4.5e-04	1.9e-01

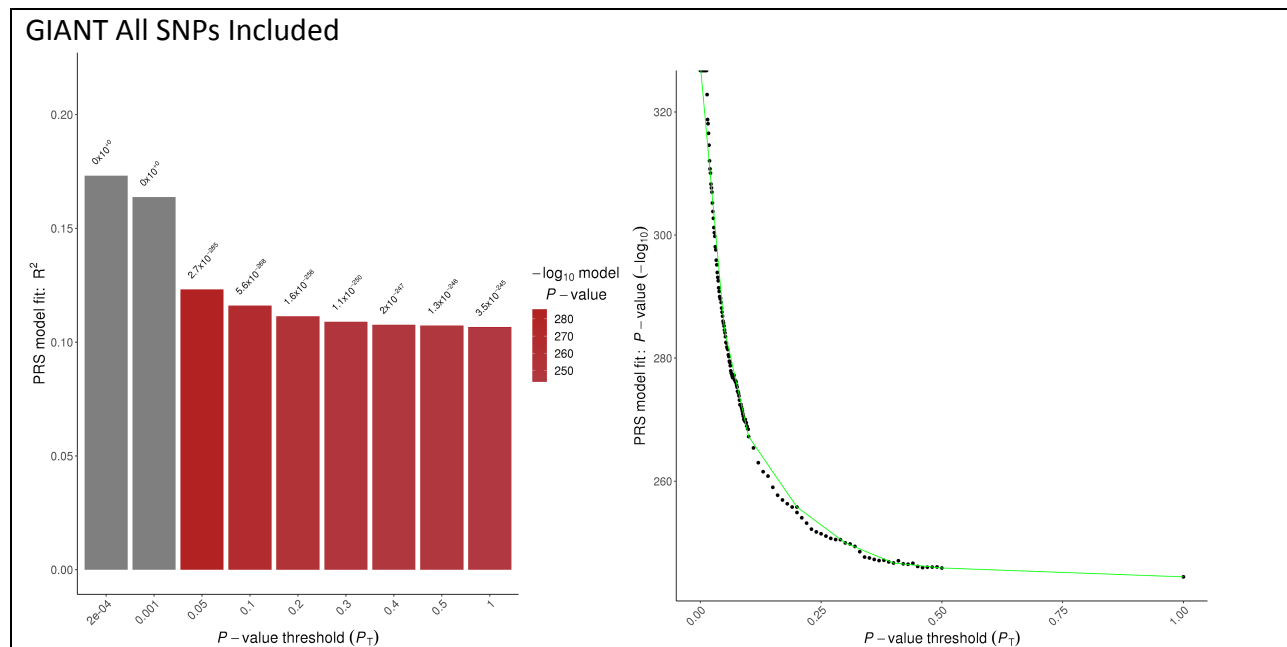


**Appendix Figure 17.** The PTB candidate gene PRS members show low enrichment for ROP compared to other curated pathway groups. The strongest observed enrichment was that for the GO\_MEMBRANE\_PROTEIN\_COMPLEX, though this is likely partly due to the large size of the second gene set.

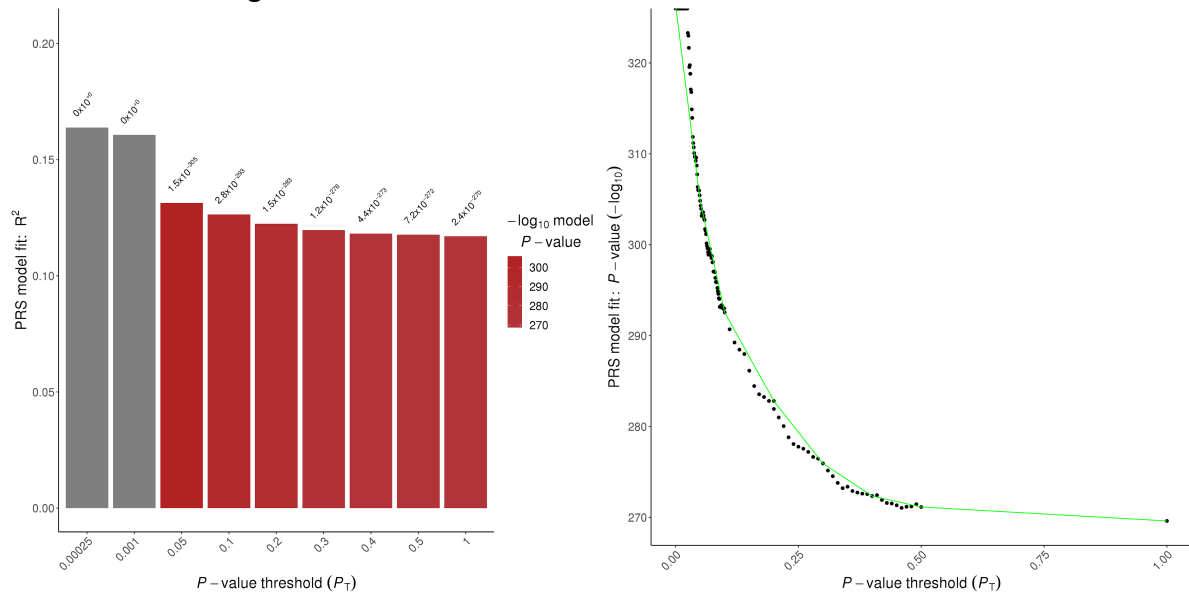
## ADDITIONAL PRS PLOTS

In addition to the provided plots, PRSice gives output summarizing the construction of polygenic risk scores. In the interest of clarity and replicability we provide these plots as an appendix.

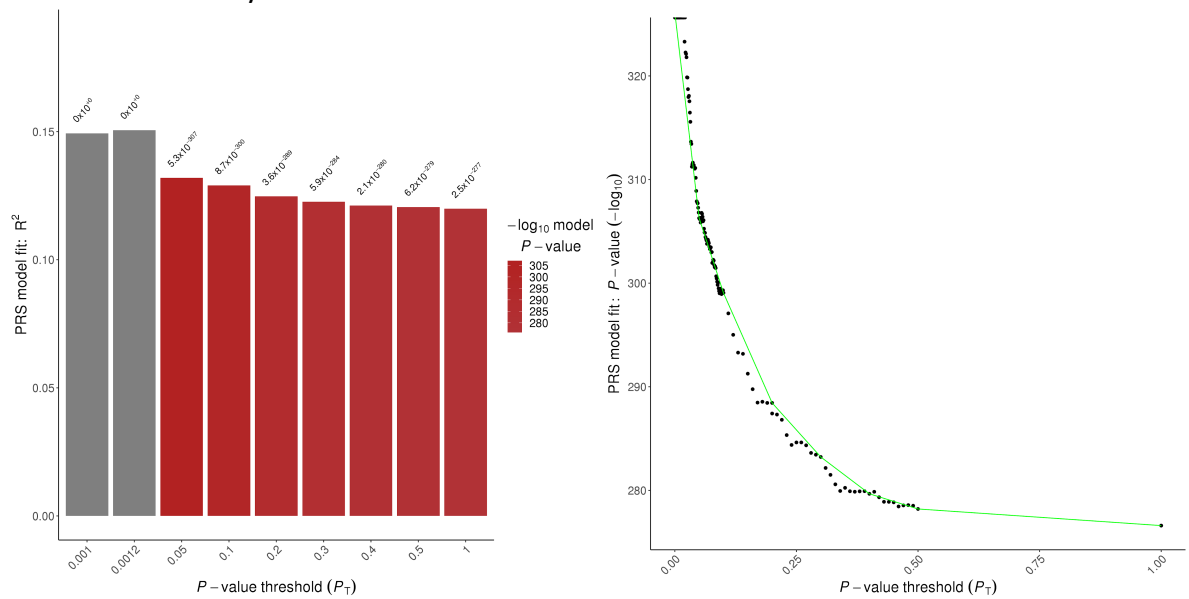
### GIANT PRS Plots



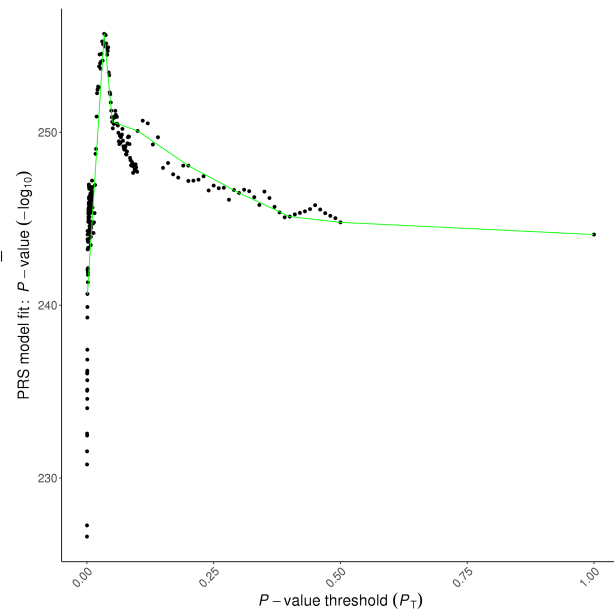
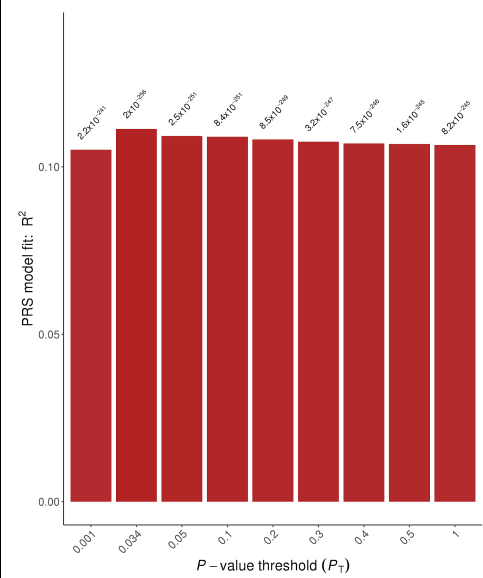
## GIANT All Gene Regions



## GIANT Connectivity Gated Score

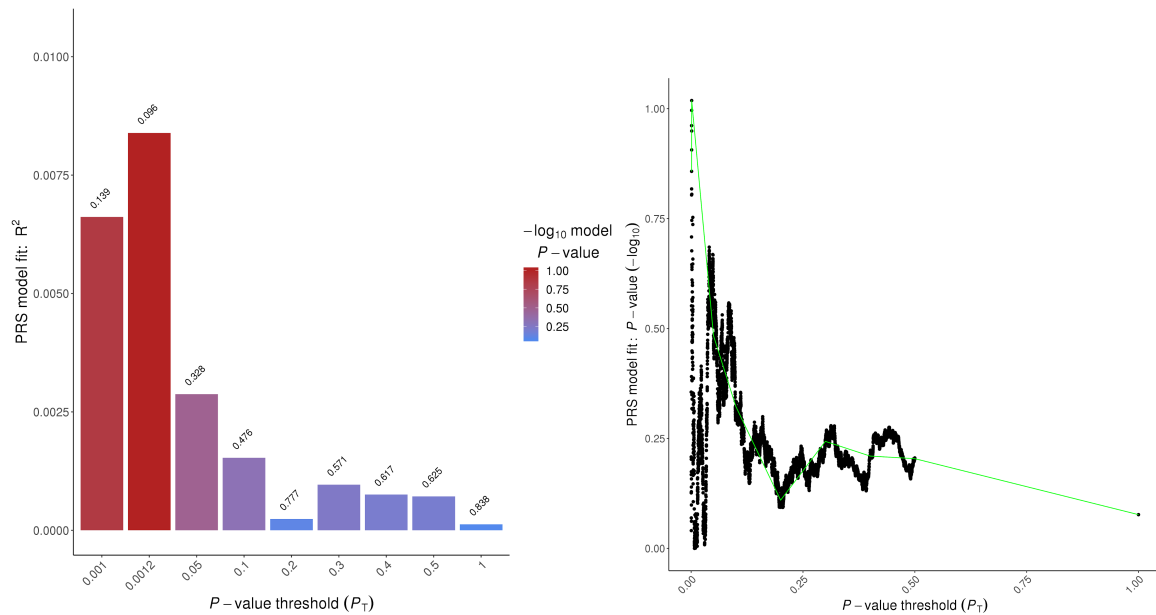


# GIANT Candidate Gated Score

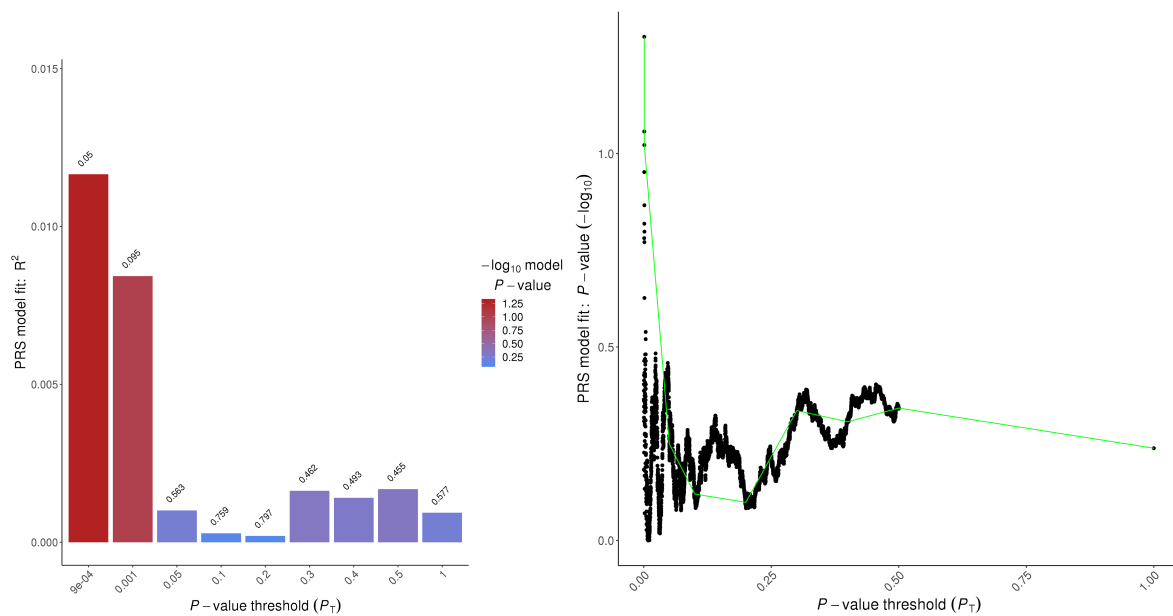


## DNBC PRS Plots

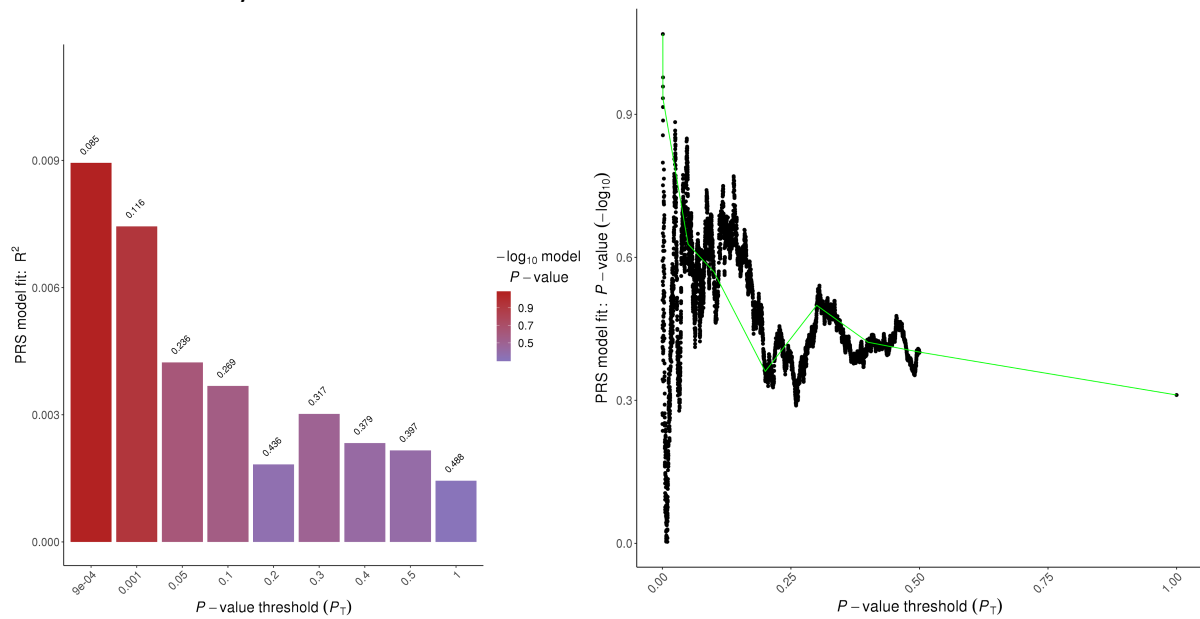
### DNBC All SNPs Included



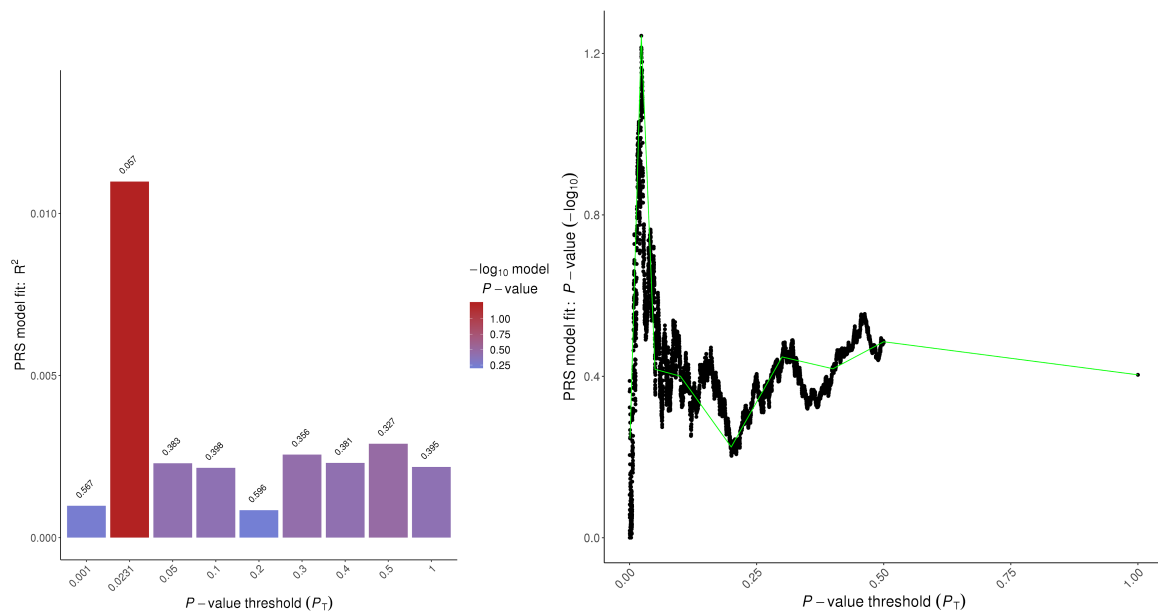
### DNBC All Genes Included



## DNBC Connectivity Gated Score

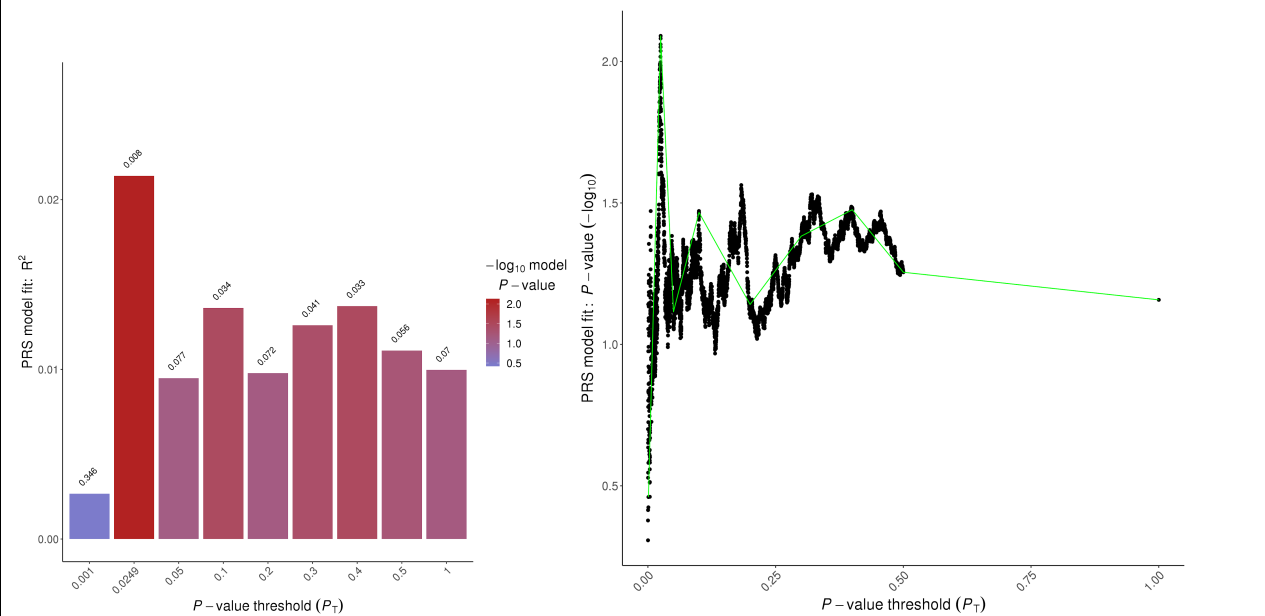


## DNBC PTB Candidate Gene Score





# DNBC ROP Gated Score



## COMPUTATIONAL ASSETS

Imputation, genomic filtering, and graph-based analysis were performed using the Exacloud High-Performance Cluster at Oregon Health & Science University, administered by the Advanced Computing Center. Exacloud runs the SLURM workload manager running on CentOS 7.

Genomic data manipulations were performed in PLINK, IMPUTE2, SHAPEIT, and PRSice. QQ-plots and Manhattan plots were produced using the qqman package for R. Network topology analysis and pipelining was performed in Python, using the igraph package.

Where appropriate, additional analysis was performed on a laptop computer using the R programming language. The fgsea package was used for GSEA analysis calculations, and Shiny was used for tool construction for late-context module analysis.

## BIBLIOGRAPHY

1. Kalpathy-Cramer J, Campbell JP, Erdogmus D, Tian P, Kedarisetti D, Moleta C, et al. Plus Disease in Retinopathy of Prematurity: Improving Diagnosis by Ranking Disease Severity and Using Quantitative Image Analysis. *Ophthalmology*. 2016 Nov;123(11):2345–51.
2. Lee JJ, Wedow R, Okbay A, Kong E, Maghzian O, Zacher M, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat Genet*. 2018 Jul 23;50(8):1112–21.
3. Khera AV, Chaffin M, Wade KH, Zahid S, Brancale J, Xia R, et al. Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood. *Cell*. 2019 Apr 18;177(3):587–96.e9.
4. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet*. 2018 Sep;50(9):1219–24.
5. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet*. 2018 Sep;19(9):581–90.
6. Gilbert C. Retinopathy of prematurity: a global perspective of the epidemics, population of babies at risk and implications for control. *Early Hum Dev*. 2008 Feb;84(2):77–82.
7. REESE, AB. A classification of retrolental fibroplasia. *Am J Ophthalmol*. 1953;36:1333–5.
8. SYMPOSIUM: retrolental fibroplasia (retinopathy of prematurity). *Am J Ophthalmol*. 1955 Aug;40(2):159–89.
9. Tasman W. Vitreoretinal changes in cicatricial retrolental fibroplasia. *Trans Am Ophthalmol Soc*. 1970;68:548–94.
10. Aiken JW, Vane JR. INTRARENAL PROSTAGLANDIN RELEASE ATTENUATES THE RENAL VASOCONSTRICTOR ACTIVITY OF ANGIOTENSIN. *J Pharmacol Exp Ther*. 1973 Mar 1;184(3):678–87.
11. Tysinger JW Jr, Weidenthal DT. Nasal heterotopia of the macula in retinopathy of prematurity. *Am J Ophthalmol*. 1977 Apr;83(4):499–500.
12. Tasman W. Retinal detachment in retrolental fibroplasia. *Albrecht Von Graefes Arch Klin Exp Ophthalmol*. 1975;195(2):129–39.
13. Harris GS. Retinopathy of prematurity and retinal detachment. *Can J Ophthalmol*. 1976 Jan;11(1):21–5.
14. Early Treatment For Retinopathy Of Prematurity Cooperative Group. Revised indications for the treatment of retinopathy of prematurity: results of the early treatment for retinopathy of prematurity randomized trial. *Arch Ophthalmol*. 2003 Dec;121(12):1684–94.

15. Palmer EA. Results of U.S. randomized clinical trial of cryotherapy for ROP (CRYO-ROP). *Doc Ophthalmol*. 1990 Mar;74(3):245–51.
16. Wallace DK, Quinn GE, Freedman SF, Chiang MF. Agreement among pediatric ophthalmologists in diagnosing plus and pre-plus disease in retinopathy of prematurity. *J AAPOS*. 2008 Aug;12(4):352–6.
17. Committee I, Others. ICROP Committee for classification of late stages of ROP: An international classification of retinopathy of prematurity: II The classification of retinal detachment. *Arch Ophthalmol*. 1987;105:906–12.
18. Patz A. The role of oxygen in retrolental fibroplasia. *Trans Am Ophthalmol Soc*. 1968;66:940–85.
19. Wright KW, Sami D, Thompson L, Ramanathan R, Joseph R, Farzavandi S. A physiologic reduced oxygen protocol decreases the incidence of threshold retinopathy of prematurity. *Trans Am Ophthalmol Soc*. 2006;104:78.
20. Port AD, Chan RVP, Ostmo S, Choi D, Chiang MF. Risk factors for retinopathy of prematurity: insights from outlier infants. *Graefes Arch Clin Exp Ophthalmol*. 2014 Oct;252(10):1669–77.
21. Ng YK, Fielder AR, Shaw DE, Levene MI. Epidemiology of retinopathy of prematurity. *Lancet*. 1988 Nov 26;2(8622):1235–8.
22. Tadesse M, Dhanireddy R, Mittal M, Higgins RD. Race, Candida sepsis, and retinopathy of prematurity. *Biol Neonate*. 2002;81(2):86–90.
23. Saunders RA, Donahue ML, Christmann LM, Pakalnis AV, Tung B, Hardy RJ, et al. Racial variation in retinopathy of prematurity. The Cryotherapy for Retinopathy of Prematurity Cooperative Group. *Arch Ophthalmol*. 1997 May;115(5):604–8.
24. Bizzarro MJ, Hussain N, Jonsson B, Feng R, Ment LR, Gruen JR, et al. Genetic susceptibility to retinopathy of prematurity. *Pediatrics*. 2006 Nov;118(5):1858–63.
25. Ortega-Molina JM, Anaya-Alaminos R, Uberos-Fernández J, Solans-Pérez de Larraya A, Chaves-Samaniego MJ, Salgado-Miranda A, et al. Genetic and Environmental Influences on Retinopathy of Prematurity. *Mediators Inflamm*. 2015 May 21;2015:764159.
26. van Wijngaarden P, Coster DJ, Brereton HM, Gibbins IL, Williams KA. Strain-dependent differences in oxygen-induced retinopathy in the inbred rat. *Invest Ophthalmol Vis Sci*. 2005 Apr;46(4):1445–52.
27. Floyd BNI, Leske DA, Wren SME, Mookadam M, Fautsch MP, Holmes JM. Differences between rat strains in models of retinopathy of prematurity. *Mol Vis*. 2005 Jul 19;11:524–30.
28. van Wijngaarden P, Brereton HM, Coster DJ, Williams KA. Genetic influences on susceptibility to oxygen-induced retinopathy. *Invest Ophthalmol Vis Sci*. 2007 Apr;48(4):1761–6.
29. Cooke RWI, Drury JA, Mountford R, Clark D. Genetic polymorphisms and retinopathy of prematurity. *Invest Ophthalmol Vis Sci*. 2004 Jun;45(6):1712–5.

30. Ali AA, Hussien NF, Samy RM, Hussein KA. Polymorphisms of Vascular Endothelial Growth Factor and Retinopathy of Prematurity. *J Pediatr Ophthalmol Strabismus*. 2015 Jul;52(4):245–53.
31. Poggi C, Giusti B, Gozzini E, Sereni A, Romagnuolo I, Kura A, et al. Genetic Contributions to the Development of Complications in Preterm Newborns. *PLoS One*. 2015 Jul 14;10(7):e0131741.
32. Kaya M, Çökakli M, Berk AT, Yaman A, Yesilirmak D, Kumral A, et al. Associations of VEGF/VEGF-receptor and HGF/c-Met promoter polymorphisms with progression/regression of retinopathy of prematurity. *Curr Eye Res*. 2013 Jan;38(1):137–42.
33. Kusuda T, Hikino S, Ohga S, Kinjo T, Ochiai M, Takahata Y, et al. Genetic variation of vascular endothelial growth factor pathway does not correlate with the severity of retinopathy of prematurity. *J Perinatol*. 2011 Apr;31(4):246–50.
34. Vannay Á, Dunai G, Bányász I, Szabó M, Vámos R, Treszl A, et al. Association of Genetic Polymorphisms of Vascular Endothelial Growth Factor and Risk for Proliferative Retinopathy of Prematurity [Internet]. Vol. 57, *Pediatric Research*. 2005. p. 396–8. Available from: <http://dx.doi.org/10.1203/01.pdr.0000153867.80238.e0>
35. Kwinta P, Bik-Multanowski M, Mitkowska Z, Tomasik T, Pietrzyk JJ. The clinical role of vascular endothelial growth factor (VEGF) system in the pathogenesis of retinopathy of prematurity. *Graefes Arch Clin Exp Ophthalmol*. 2008 Oct;246(10):1467–75.
36. Gismondi D, Ndoja L, Qu X, Shastri BS. Lack of association of VEGF gene 3'-UTR polymorphisms (C702T, C936T and G1612A) and the risk of developing advanced retinopathy of prematurity (ROP). *Graefes Arch Clin Exp Ophthalmol*. 2013 Jan;251(1):413–5.
37. Kalmeh ZA, Azarpira N, Mosallaei M, Hosseini H, Malekpour Z. Genetic polymorphisms of vascular endothelial growth factor and risk for retinopathy of prematurity in South of Iran. *Mol Biol Rep*. 2013 Jul;40(7):4613–8.
38. Kwan H, Pecinka V, Tsukamoto A, Parslow TG, Guzman R, Lin TP, et al. Transgenes expressing the Wnt-1 and int-2 proto-oncogenes cooperate during mammary carcinogenesis in doubly transgenic mice. *Mol Cell Biol*. 1992 Jan;12(1):147–54.
39. Jin T. The WNT signalling pathway and diabetes mellitus. *Diabetologia*. 2008 Oct;51(10):1771–80.
40. Criswick VG, Schepens CL. Familial exudative vitreoretinopathy. *Am J Ophthalmol*. 1969 Oct;68(4):578–94.
41. Robitaille J, MacDonald MLE, Kaykas A, Sheldahl LC, Zeisler J, Dubé M-P, et al. Mutant frizzled-4 disrupts retinal angiogenesis in familial exudative vitreoretinopathy. *Nat Genet*. 2002 Oct;32(2):326–30.
42. MacDonald MLE, Goldberg YP, Macfarlane J, Samuels ME, Trese MT, Shastri BS. Genetic variants of frizzled-4 gene in familial exudative vitreoretinopathy and advanced retinopathy of prematurity. *Clin Genet*. 2005 Apr;67(4):363–6.
43. Ells A, Guernsey DL, Wallace K, Zheng B, Vincer M, Allen A, et al. Severe retinopathy of prematurity associated with FZD4 mutations. *Ophthalmic Genet*. 2010 Mar;31(1):37–43.

44. Toomes C, Bottomley HM, Jackson RM, Towns KV, Scott S, Mackey DA, et al. Mutations in LRP5 or FZD4 underlie the common familial exudative vitreoretinopathy locus on chromosome 11q. *Am J Hum Genet.* 2004 Apr;74(4):721–30.
45. Xia C-H, Liu H, Cheung D, Wang M, Cheng C, Du X, et al. A model for familial exudative vitreoretinopathy caused by LRP5 mutations. *Hum Mol Genet.* 2008 Jun 1;17(11):1605–12.
46. Ye X, Wang Y, Cahill H, Yu M, Badea TC, Smallwood PM, et al. Norrin, frizzled-4, and Lrp5 signaling in endothelial cells controls a genetic program for retinal vascularization. *Cell.* 2009 Oct 16;139(2):285–98.
47. Hiraoka M, Takahashi H, Orimo H, Hiraoka M, Ogata T, Azuma N. Genetic screening of Wnt signaling factors in advanced retinopathy of prematurity. *Mol Vis.* 2010 Dec 5;16:2572–7.
48. Dickinson JL, Sale MM, Passmore A, FitzGerald LM, Wheatley CM, Burdon KP, et al. Mutations in the NDP gene: contribution to Norrie disease, familial exudative vitreoretinopathy and retinopathy of prematurity. *Clin Experiment Ophthalmol.* 2006 Sep;34(7):682–8.
49. Shastry BS, Pendergast SD, Hartzer MK, Liu X, Trese MT. Identification of missense mutations in the Norrie disease gene associated with advanced retinopathy of prematurity. *Arch Ophthalmol.* 1997 May;115(5):651–5.
50. Hiraoka M, Berinstein DM, Trese MT, Shastry BS. Insertion and deletion mutations in the dinucleotide repeat region of the Norrie disease gene in patients with advanced retinopathy of prematurity. *J Hum Genet.* 2001;46(4):178–81.
51. Talks SJ, Ebenezer N, Hykin P, Adams G, Yang F, Schulenberg E, et al. De novo mutations in the 5' regulatory region of the Norrie disease gene in retinopathy of prematurity. *J Med Genet.* 2001 Dec;38(12):E46.
52. Haider MZ, Devarajan LV, Al-Essa M, Kumar H. A C597→A Polymorphism in the Norrie Disease Gene Is Associated with Advanced Retinopathy of Prematurity in Premature Kuwaiti Infants. *J Biomed Sci.* 2002;9(4):365–70.
53. Hutcheson KA, Paluru PC, Bernstein SL, Koh J, Rappaport EF, Leach RA, et al. Norrie disease gene sequence variants in an ethnically diverse population with retinopathy of prematurity. *Mol Vis.* 2005 Jul 14;11:501–8.
54. Junge HJ, Yang S, Burton JB, Paes K, Shu X, French DM, et al. TSPAN12 regulates retinal vascular development by promoting Norrin- but not Wnt-induced FZD4/beta-catenin signaling. *Cell.* 2009 Oct 16;139(2):299–311.
55. Poulter JA, Ali M, Gilmour DF, Rice A, Kondo H, Hayashi K, et al. Mutations in TSPAN12 cause autosomal-dominant familial exudative vitreoretinopathy. *Am J Hum Genet.* 2010 Feb 12;86(2):248–53.
56. Ohlmann A, Tamm ER. Norrin: molecular and functional properties of an angiogenic and neuroprotective growth factor. *Prog Retin Eye Res.* 2012 May;31(3):243–57.
57. Katoh M. Therapeutics targeting angiogenesis: genetics and epigenetics, extracellular miRNAs and

- signaling networks (Review). *Int J Mol Med*. 2013 Oct;32(4):763–7.
58. van Wijk XMR, van Kuppevelt TH. Heparan sulfate in angiogenesis: a target for therapy. *Angiogenesis*. 2014 Jul;17(3):443–62.
  59. Griffith RM, Li H, Zhang N, Favazza TL, Fulton AB, Hansen RM, et al. Next-generation sequencing analysis of gene regulation in the rat model of retinopathy of prematurity. *Doc Ophthalmol*. 2013 Aug;127(1):13–31.
  60. Kondo H, Kusaka S, Yoshinaga A, Uchio E, Tawara A, Tahira T. Genetic variants of FZD4 and LRP5 genes in patients with advanced retinopathy of prematurity. *Mol Vis*. 2013 Feb 25;19:476–85.
  61. Dailey WA, Gryc W, Garg PG, Drenser KA. Frizzled-4 Variations Associated with Retinopathy and Intrauterine Growth Retardation: A Potential Marker for Prematurity and Retinopathy. *Ophthalmology*. 2015 Sep;122(9):1917–23.
  62. Hellstrom A, Perruzzi C, Ju M, Engstrom E, Hard AL, Liu JL, et al. Low IGF-I suppresses VEGF-survival signaling in retinal endothelial cells: direct correlation with clinical retinopathy of prematurity. *Proc Natl Acad Sci U S A*. 2001 May 8;98(10):5804–8.
  63. Lassarre C, Hardouin S, Daffos F, Forestier F, Franken F, Binoux M. Serum Insulin-Like Growth Factors and Insulin-Like Growth Factor Binding Proteins in the Human Fetus. Relationships with Growth in Normal Subjects and in Subjects with Intrauterine Growth Retardation. *Pediatr Res*. 1991 Mar 1;29(3):219–25.
  64. Reece EA, Wiznitzer A, Le E, Homko CJ, Behrman H, Spencer EM. The relation between human fetal growth and fetal blood levels of insulin-like growth factors I and II, their binding proteins, and receptors. *Obstet Gynecol*. 1994 Jul;84(1):88–95.
  65. Chen J, Smith LEH. Retinopathy of prematurity. *Angiogenesis*. 2007 Feb 27;10(2):133–40.
  66. Smith LE, Shen W, Perruzzi C, Soker S, Kinose F, Xu X, et al. Regulation of vascular endothelial growth factor-dependent retinal neovascularization by insulin-like growth factor-1 receptor. *Nat Med*. 1999 Dec;5(12):1390–5.
  67. Kondo T, Vicent D, Suzuma K, Yanagisawa M, King GL, Holzenberger M, et al. Knockout of insulin and IGF-1 receptors on vascular endothelial cells protects against retinal neovascularization. *J Clin Invest*. 2003 Jun;111(12):1835–42.
  68. Bonafè M, Barbieri M, Marchegiani F, Olivieri F, Ragno E, Giampieri C, et al. Polymorphic variants of insulin-like growth factor I (IGF-I) receptor and phosphoinositide 3-kinase genes affect IGF-I plasma levels and human longevity: cues for an evolutionarily conserved mechanism of life span control. *J Clin Endocrinol Metab*. 2003 Jul;88(7):3299–304.
  69. Balogh Á, Derzbach L, Vannay Á, Vászárhelyi B. Lack of association between insulin-like growth factor I receptor G+3174A polymorphism and retinopathy of prematurity. *Graefes Arch Clin Exp Ophthalmol*. 2006 Aug 1;244(8):1035–8.
  70. Shastri BS. Endothelial nitric oxide synthase gene promoter polymorphism (T-786C) may be associated with advanced retinopathy of prematurity. *Graefes Arch Clin Exp Ophthalmol*. 2013

Sep;251(9):2251–3.

71. Dammann O, Brinkhaus M-J, Bartels DB, Dördelmann M, Dressler F, Kerk J, et al. Immaturity, perinatal inflammation, and retinopathy of prematurity: a multi-hit hypothesis. *Early Hum Dev.* 2009 May;85(5):325–9.
72. Liu L, Johnson HL, Cousens S, Perin J, Scott S, Lawn JE, et al. Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000 [Internet]. Vol. 379, *The Lancet*. 2012. p. 2151–61. Available from: [http://dx.doi.org/10.1016/s0140-6736\(12\)60560-1](http://dx.doi.org/10.1016/s0140-6736(12)60560-1)
73. Liu L, Oza S, Hogan D, Chu Y, Perin J, Zhu J, et al. Global, regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the Sustainable Development Goals. *Lancet*. 2016 Dec 17;388(10063):3027–35.
74. Blencowe H, Cousens S, Oestergaard MZ, Chou D, Moller A-B, Narwal R, et al. National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. *Lancet*. 2012 Jun 9;379(9832):2162–72.
75. Osterman MJK, Kochanek KD, MacDorman MF, Strobino DM, Guyer B. Annual Summary of Vital Statistics: 2012–2013 [Internet]. Vol. 135, *PEDIATRICS*. 2015. p. 1115–25. Available from: <http://dx.doi.org/10.1542/peds.2015-0434>
76. Honest H, Bachmann LM, Sengupta R, Gupta JK, Kleijnen J, Khan KS. Accuracy of absence of fetal breathing movements in predicting preterm birth: a systematic review. *Ultrasound Obstet Gynecol*. 2004 Jul;24(1):94–100.
77. Sotiriadis A, Papatheodorou S, Kavvadias A, Makrydimas G. Transvaginal cervical length measurement for prediction of preterm birth in women with threatened preterm labor: a meta-analysis. *Ultrasound Obstet Gynecol*. 2010;35(1):54–64.
78. Honest H, Hyde CJ, Khan KS. Prediction of spontaneous preterm birth: no good test for predicting a spontaneous preterm birth. *Curr Opin Obstet Gynecol*. 2012 Dec;24(6):422–33.
79. Honest H, Bachmann LM, Coomarasamy A, Gupta JK, Kleijnen J, Khan KS. Accuracy of cervical transvaginal sonography in predicting preterm birth: a systematic review [Internet]. Vol. 22, *Ultrasound in Obstetrics and Gynecology*. 2003. p. 305–22. Available from: <http://dx.doi.org/10.1002/uog.202>
80. Hendler I, Goldenberg RL, Mercer BM, Iams JD, Meis PJ, Moawad AH, et al. The Preterm Prediction study: Association between maternal body mass index and spontaneous and indicated preterm birth [Internet]. Vol. 192, *American Journal of Obstetrics and Gynecology*. 2005. p. 882–6. Available from: <http://dx.doi.org/10.1016/j.ajog.2004.09.021>
81. Winkvist A. Familial patterns in birth characteristics: impact on individual and population risks [Internet]. Vol. 27, *International Journal of Epidemiology*. 1998. p. 248–54. Available from: <http://dx.doi.org/10.1093/ije/27.2.248>
82. Ananth CV, Getahun D, Peltier MR, Salihu HM, Vintzileos AM. Recurrence of spontaneous versus medically indicated preterm birth [Internet]. Vol. 195, *American Journal of Obstetrics and*



- Gynecology. 2006. p. 643–50. Available from: <http://dx.doi.org/10.1016/j.ajog.2006.05.022>
83. Porter T, Fraser A, Hunter C, Ward R, Varner M. The risk of preterm birth across generations [Internet]. Vol. 90, *Obstetrics & Gynecology*. 1997. p. 63–7. Available from: [http://dx.doi.org/10.1016/s0029-7844\(97\)00215-9](http://dx.doi.org/10.1016/s0029-7844(97)00215-9)
  84. Goldenberg RL, Cliver SP, Mulvihill FX, Hickey CA, Hoffman HJ, Klerman LV, et al. Medical, psychosocial, and behavioral risk factors do not explain the increased risk for low birth weight among black women [Internet]. Vol. 175, *American Journal of Obstetrics and Gynecology*. 1996. p. 1317–24. Available from: [http://dx.doi.org/10.1016/s0002-9378\(96\)70048-0](http://dx.doi.org/10.1016/s0002-9378(96)70048-0)
  85. Fiscella K. Race, Perinatal Outcome, and Amniotic Infection [Internet]. Vol. 51, *Obstetrical & Gynecological Survey*. 1996. p. 60–6. Available from: <http://dx.doi.org/10.1097/00006254-199601000-00022>
  86. Plunkett J, Muglia LJ. Genetic contributions to preterm birth: Implications from epidemiological and genetic association studies [Internet]. Vol. 40, *Annals of Medicine*. 2008. p. 167–79. Available from: <http://dx.doi.org/10.1080/07853890701806181>
  87. Wu W, Witherspoon DJ, Fraser A, Clark EAS, Rogers A, Stoddard GJ, et al. The heritability of gestational age in a two-million member cohort: implications for spontaneous preterm birth [Internet]. Vol. 134, *Human Genetics*. 2015. p. 803–8. Available from: <http://dx.doi.org/10.1007/s00439-015-1558-1>
  88. Svensson AC, Sandin S, Cnattingius S, Reilly M, Pawitan Y, Hultman CM, et al. Maternal Effects for Preterm Birth: A Genetic Epidemiologic Study of 630,000 Families [Internet]. Vol. 170, *American Journal of Epidemiology*. 2009. p. 1365–72. Available from: <http://dx.doi.org/10.1093/aje/kwp328>
  89. Tu FF, Goldenberg RL, Tamura T, Drews M, Zucker SJ, Voss HF. Prenatal plasma matrix metalloproteinase-9 levels to predict spontaneous preterm birth. *Obstet Gynecol*. 1998 Sep;92(3):446–9.
  90. Goldenberg RL, Mercer BM, Meis PJ, Copper RL, Das A, McNellis D. The preterm prediction study: Fetal fibronectin testing and spontaneous preterm birth. *Obstetrics & Gynecology*. 1996 May 1;87(5, Part 1):643–8.
  91. Macones GA, Parry S, Elkousy M, Clothier B, Ural SH, Strauss JF. A polymorphism in the promoter region of TNF and bacterial vaginosis: preliminary evidence of gene-environment interaction in the etiology of spontaneous preterm birth [Internet]. Vol. 190, *American Journal of Obstetrics and Gynecology*. 2004. p. 1504–8. Available from: <http://dx.doi.org/10.1016/j.ajog.2004.01.001>
  92. Engel SAM, Erichsen HC, Savitz DA, Thorp J, Chanock SJ, Olshan AF. Risk of spontaneous preterm birth is associated with common proinflammatory cytokine polymorphisms. *Epidemiology*. 2005 Jul;16(4):469–77.
  93. Wang X, Zuckerman B, Pearson C, Kaufman G, Chen C, Wang G, et al. Maternal Cigarette Smoking, Metabolic Gene Polymorphism, and Infant Birth Weight [Internet]. Vol. 57, *Obstetrical & Gynecological Survey*. 2002. p. 418–9. Available from: <http://dx.doi.org/10.1097/00006254-200207000-00005>

94. Plunkett J, Doniger S, Orabona G, Morgan T, Haataja R, Hallman M, et al. An Evolutionary Genomic Approach to Identify Genes Involved in Human Birth Timing [Internet]. Vol. 7, PLoS Genetics. 2011. p. e1001365. Available from: <http://dx.doi.org/10.1371/journal.pgen.1001365>
95. Olsen JOMMSF, Olsen J, Melbye M, Olsen SF. The Danish National Birth Cohort its background, structure and aim [Internet]. Vol. 29, Scandinavian Journal of Public Health. 2001. p. 300–7. Available from: <http://dx.doi.org/10.1080/140349401317115268>
96. Myking S, Boyd HA, Myhre R, Feenstra B, Jugessur A, Devold Pay AS, et al. X-chromosomal maternal and fetal SNPs and the risk of spontaneous preterm delivery in a Danish/Norwegian genome-wide association study. PLoS One. 2013 Apr 16;8(4):e61781.
97. Zhang H, Baldwin DA, Bukowski RK, Parry S, Xu Y, Song C, et al. A genome-wide association study of early spontaneous preterm delivery. Genet Epidemiol. 2015;39(3):217–26.
98. Dolan SM, Hollegaard MV, Merialdi M, Betran AP, Allen T, Abelow C, et al. Synopsis of preterm birth genetic association studies: the preterm birth genetics knowledge base (PTBGene). Public Health Genomics. 2010 May 20;13(7-8):514–23.
99. Capece A, Vasieva O, Meher S, Alfirevic Z, Alfirevic A. Pathway analysis of genetic factors associated with spontaneous preterm birth and pre-labor preterm rupture of membranes. PLoS One. 2014 Sep 29;9(9):e108578.
100. Brubaker D, Liu Y, Wang J, Tan H, Zhang G, Jacobsson B, et al. Finding lost genes in GWAS via integrative—omics analysis reveals novel sub-networks associated with preterm birth. Hum Mol Genet. 2016 Dec 1;25(23):5254–64.
101. Burris HH, Rifas-Shiman SL, Baccarelli A, Boeke CE, Kleinman K, Wen X, et al. Associations of LINE-1 (“Global”) DNA Methylation with Preterm Birth In a Prospective Cohort Study. In: JOURNAL OF DEVELOPMENTAL ORIGINS OF HEALTH AND DISEASE. CAMBRIDGE UNIV PRESS EDINBURGH BLDG, SHAFTESBURY RD, CB2 8RU CAMBRIDGE, ENGLAND; 2011. p. S62–S62.
102. Parets SE, Conneely KN, Kilaru V, Fortunato SJ, Syed TA, Saade G, et al. Fetal DNA Methylation Associates with Early Spontaneous Preterm Birth and Gestational Age. PLoS One. 2013 Jun 27;8(6):e67489.
103. Sanders AP, Burris HH, Just AC, Motta V, Svensson K, Mercado-Garcia A, et al. microRNA expression in the cervix during pregnancy is associated with length of gestation. Epigenetics. 2015;10(3):221–8.
104. Elovitz MA, Brown AG, Anton L, Gilstrap M, Heiser L, Bastek J. Distinct cervical microRNA profiles are present in women destined to have a preterm birth. Am J Obstet Gynecol. 2014 Mar;210(3):221.e1–11.
105. York TP, Strauss JF 3rd, Neale MC, Eaves LJ. Racial differences in genetic and environmental risk to preterm birth. PLoS One. 2010 Aug 25;5(8):e12391.
106. Ferguson KK, O’Neill MS, Meeker JD. Environmental contaminant exposures and preterm birth: a comprehensive review. J Toxicol Environ Health B Crit Rev. 2013;16(2):69–113.

107. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009 Oct 8;461(7265):747–53.
108. International Schizophrenia Consortium, Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*. 2009 Aug 6;460(7256):748–52.
109. Martin AR, Daly MJ, Robinson EB, Hyman SE, Neale BM. Predicting Polygenic Risk of Psychiatric Disorders. *Biol Psychiatry*. 2019 Jul 15;86(2):97–109.
110. Dudbridge F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet*. 2013 Mar;9(3):e1003348.
111. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet*. 2010 Nov;42(11):937–48.
112. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*. 2010 Oct 14;467(7317):832–8.
113. Group PGCBDW, Psychiatric GWAS Consortium Bipolar Disorder Working Group. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4 [Internet]. Vol. 43, *Nature Genetics*. 2011. p. 977–83. Available from: <http://dx.doi.org/10.1038/ng.943>
114. Euesden J, Lewis CM, O'Reilly PF. PRSice: Polygenic Risk Score software. *Bioinformatics*. 2015 May 1;31(9):1466–8.
115. Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, et al. Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res*. 2011 Jan;39(Database issue):D685–90.
116. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*. 2003 Oct;13(10):2363–71.
117. Gene Ontology Consortium. The Gene Ontology project in 2008. *Nucleic Acids Res*. 2008 Jan;36(Database issue):D440–4.
118. Mooney MA, Nigg JT, McWeeney SK, Wilmot B. Functional and genomic context in pathway analysis of GWAS data. *Trends Genet*. 2014 Sep;30(9):390–400.
119. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human Protein Reference Database--2009 update. *Nucleic Acids Res*. 2009 Jan;37(Database issue):D767–72.
120. Mostafavi et al S. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function | Kopernio [Internet]. [cited 2019 Jul 9]. Available from: <https://kopernio.com/viewer?doi=10.1186/gb-2008-9-s1-s4&route=6>

121. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 2019 Jan 8;47(D1):D607–13.
122. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell.* 2017 Jun 15;169(7):1177–86.
123. Liu X, Li YI, Pritchard JK. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell.* 2019 May 2;177(4):1022–34.e6.
124. Jeong H, Mason SP, Barabási AL, Oltvai ZN. Lethality and centrality in protein networks. *Nature.* 2001 May 3;411(6833):41–2.
125. Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. *Science.* 2003 Oct 10;302(5643):249–55.
126. Laub MT, McAdams HH, Feldblyum T, Fraser CM, Shapiro L. Global analysis of the genetic network controlling a bacterial cell cycle. *Science.* 2000 Dec 15;290(5499):2144–8.
127. Network TCGA, The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours [Internet]. Vol. 490, *Nature*. 2012. p. 61–70. Available from: <http://dx.doi.org/10.1038/nature11412>
128. Iancu OD, Oberbeck D, Darakjian P, Metten P, McWeeney S, Crabbe JC, et al. Selection for drinking in the dark alters brain gene coexpression networks. *Alcohol Clin Exp Res.* 2013 Aug;37(8):1295–303.
129. Leiserson MDM, Eldridge JV, Ramachandran S, Raphael BJ. Network analysis of GWAS data [Internet]. Vol. 23, *Current Opinion in Genetics & Development*. 2013. p. 602–10. Available from: <http://dx.doi.org/10.1016/j.gde.2013.09.003>
130. Horvath S. *Weighted Network Analysis: Applications in Genomics and Systems Biology*. Springer Science & Business Media; 2011. 421 p.
131. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol.* 2005 Aug 12;4:Article17.
132. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL. The large-scale organization of metabolic networks. *Nature.* 2000 Oct 5;407(6804):651–4.
133. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks [Internet]. *arXiv [physics.soc-ph]*. 2008. Available from: <http://arxiv.org/abs/0803.0476>
134. Leiserson MDM, Vandin F, Wu H-T, Dobson JR, Eldridge JV, Thomas JL, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes [Internet]. Vol. 47, *Nature Genetics*. 2015. p. 106–14. Available from: <http://dx.doi.org/10.1038/ng.3168>
135. Saugstad OD, Aune D. Optimal oxygenation of extremely low birth weight infants: a meta-analysis and systematic review of the oxygen saturation target studies. *Neonatology.* 2014;105(1):55–63.

136. Kinsey VE. Retrolental fibroplasia; cooperative study of retrolental fibroplasia and the use of oxygen. *AMA Arch Ophthalmol*. 1956 Oct;56(4):481–543.
137. Flynn JT, Bancalari E, Snyder ES, Goldberg RN, Feuer W, Cassady J, et al. A cohort study of transcutaneous oxygen tension and the incidence and severity of retinopathy of prematurity. *Trans Am Ophthalmol Soc*. 1991;89:77–92; discussion 92–5.
138. Hartnett ME, Lane RH. Effects of oxygen on the development and severity of retinopathy of prematurity. *J AAPOS*. 2013 Jun;17(3):229–34.
139. Heidary G, Vanderveen D, Smith LE. Retinopathy of prematurity: current concepts in molecular pathogenesis. *Semin Ophthalmol*. 2009 Mar;24(2):77–81.
140. Hartnett ME, Penn JS. Mechanisms and management of retinopathy of prematurity. *N Engl J Med*. 2012 Dec 27;367(26):2515–26.
141. Smith LEH. Through the eyes of a child: understanding retinopathy through ROP the Friedenwald lecture. *Invest Ophthalmol Vis Sci*. 2008 Dec;49(12):5177–82.
142. Smith LE, Wesolowski E, McLellan A, Kostyk SK, D’Amato R, Sullivan R, et al. Oxygen-induced retinopathy in the mouse. *Invest Ophthalmol Vis Sci*. 1994 Jan;35(1):101–11.
143. Connor KM, Krah NM, Dennison RJ, Aderman CM, Chen J, Guerin KI, et al. Quantification of oxygen-induced retinopathy in the mouse: a model of vessel loss, vessel regrowth and pathological angiogenesis. *Nat Protoc*. 2009 Oct 8;4(11):1565–73.
144. Cunningham S, Fleck BW, Elton RA, McIntosh N. Transcutaneous oxygen levels in retinopathy of prematurity. *Lancet*. 1995 Dec 2;346(8988):1464–5.
145. York JR, Landers S, Kirby RS, Arbogast PG, Penn JS. Arterial oxygen fluctuation and retinopathy of prematurity in very-low-birth-weight infants. *J Perinatol*. 2004 Feb;24(2):82–7.
146. Saito Y, Omoto T, Cho Y, Hatsukawa Y, Fujimura M, Takeuchi T. The progression of retinopathy of prematurity and fluctuation in blood gas tension. *Graefes Arch Clin Exp Ophthalmol*. 1993 Mar;231(3):151–6.
147. Penn JS, Henry MM, Tolman BL. Exposure to alternating hypoxia and hyperoxia causes severe proliferative retinopathy in the newborn rat. *Pediatr Res*. 1994 Dec;36(6):724–31.
148. Penn JS, Henry MM, Wall PT, Tolman BL. The range of PaO<sub>2</sub> variation determines the severity of oxygen-induced retinopathy in newborn rats. *Invest Ophthalmol Vis Sci*. 1995 Sep;36(10):2063–70.
149. Gole GA, Ells AL, Katz X, Holmström G, Fielder AR, Capone A, et al. The International Classification of Retinopathy of Prematurity Revisited : An International Committee for the Classification of Retinopathy of Prematurity. *Arch Ophthalmol*. 2005;123:991–9.
150. Zacharias L. Retrolental fibroplasia; a survey. *Am J Ophthalmol*. 1952 Oct;35(10):1426–54.
151. Schaffer DB, Palmer EA, Plotsky DF, Metz HS, Flynn JT, Tung B, et al. Prognostic Factors in the Natural Course of Retinopathy of Prematurity. *Ophthalmology*. 1993 Feb 1;100(2):230–7.

152. Hartnett ME. Features Associated With Surgical Outcome in Patients With Stages 4 and 5 Retinopathy of Prematurity Dear Editor: Reply. *Retina*. 2004 Aug;24(4):658.
153. Kuo JZ, Wong TY, Rotter JI. Challenges in elucidating the genetics of diabetic retinopathy. *JAMA Ophthalmol*. 2014 Jan;132(1):96–107.
154. Fritsche LG, Igl W, Bailey JNC, Grassmann F, Sengupta S, Bragg-Gresham JL, et al. A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nat Genet*. 2016 Feb;48(2):134–43.
155. Lim LS, Mitchell P, Seddon JM, Holz FG, Wong TY. Age-related macular degeneration. *Lancet*. 2012 May 5;379(9827):1728–38.
156. Mintz-Hittner HA, Kennedy KA, Chuang AZ, BEAT-ROP Cooperative Group. Efficacy of intravitreal bevacizumab for stage 3+ retinopathy of prematurity. *N Engl J Med*. 2011 Feb 17;364(7):603–15.
157. Miller JW, Le Couter J, Strauss EC, Ferrara N. Vascular endothelial growth factor a in intraocular vascular disease. *Ophthalmology*. 2013 Jan;120(1):106–14.
158. Shastry BS, Qu X. Lack of association of the VEGF gene promoter (–634 G→C and –460 C→T) polymorphism and the risk of advanced retinopathy of prematurity. *Graefes Arch Clin Exp Ophthalmol*. 2007 May 1;245(5):741–3.
159. Shastry BS. Lack of association of VEGF (–2578 C→A) and ANG 2 (–35 G→C) gene polymorphisms with the progression of retinopathy of prematurity. *Graefes Arch Clin Exp Ophthalmol*. 2009 Jun;247(6):859–60.
160. Bányász I, Bokodi G, Vannay A, Szebeni B, Treszl A, Vászárhelyi B, et al. Genetic polymorphisms of vascular endothelial growth factor and angiopoietin 2 in retinopathy of prematurity. *Curr Eye Res*. 2006 Jul;31(7-8):685–90.
161. Haider MZ, Devarajan LV, Al-Essa M, Kumar H. Angiotensin-converting enzyme gene insertion/deletion polymorphism in Kuwaiti children with retinopathy of prematurity. *Biol Neonate*. 2002 Aug;82(2):84–8.
162. Mohamed S, Schaa K, Cooper ME, Ahrens E, Alvarado A, Colaizy T, et al. Genetic contributions to the development of retinopathy of prematurity. *Pediatr Res*. 2009 Feb;65(2):193–7.
163. Hartnett ME, Cotten CM. Genomics in the neonatal nursery: Focus on ROP. *Semin Perinatol*. 2015 Dec;39(8):604–10.
164. Rusai K, Vannay A, Szebeni B, Borgulya G, Fekete A, Vászárhelyi B, et al. Endothelial nitric oxide synthase gene T-786C and 27-bp repeat gene polymorphisms in retinopathy of prematurity. *Mol Vis*. 2008 Feb 5;14:286–90.
165. Yanamandra K, Napper D, Pramanik A, Bocchini JA Jr, Dhanireddy R. Endothelial nitric oxide synthase genotypes in the etiology of retinopathy of prematurity in premature infants. *Ophthalmic Genet*. 2010 Dec;31(4):173–7.
166. Shastry BS. Assessment of the Contribution of Insulin-like Growth Factor I Receptor 3174 G→A

- Polymorphism to the Progression of Advanced Retinopathy of Prematurity. *Eur J Ophthalmol*. 2007 Nov 1;17(6):950–3.
167. Chen ZY, Battinelli EM, Fielder A, Bunday S, Sims K, Breakefield XO, et al. A mutation in the Norrie disease gene (NDP) associated with X-linked familial exudative vitreoretinopathy. *Nat Genet*. 1993 Oct;5(2):180–3.
  168. Shastri BS, Hiraoka M, Trese DC, Trese MT. Norrie disease and exudative vitreoretinopathy in families with affected female carriers. *Eur J Ophthalmol*. 1999 Jul;9(3):238–42.
  169. John VJ, McClintic JI, Hess DJ, Berrocal AM. Retinopathy of Prematurity Versus Familial Exudative Vitreoretinopathy: Report on Clinical and Angiographic Findings. *Ophthalmic Surg Lasers Imaging Retina*. 2016 Jan;47(1):14–9.
  170. Powell-Braxton L, Hollingshead P, Warburton C, Dowd M, Pitts-Meek S, Dalton D, et al. IGF-I is required for normal embryonic growth in mice. *Genes Dev*. 1993 Dec;7(12B):2609–17.
  171. Ashton IK, Zapf J, Einschenk I, MacKenzie IZ. Insulin-like growth factors (IGF) 1 and 2 in human foetal plasma and relationship to gestational age and foetal size during midpregnancy. *Acta Endocrinol*. 1985 Dec;110(4):558–63.
  172. Hellström A, Engström E, Hård A-L, Albertsson-Wikland K, Carlsson B, Niklasson A, et al. Postnatal serum insulin-like growth factor I deficiency is associated with retinopathy of prematurity and other complications of premature birth. *Pediatrics*. 2003 Nov;112(5):1016–20.
  173. Hellstrom A, Ley D, Hallberg B, Lofqvist C, Hansen-Pupp I, Ramenghi LA, et al. IGF-1 as a Drug for Preterm Infants: A Step-Wise Clinical Development. *Curr Pharm Des*. 2017;23(38):5964–70.
  174. IGF-1/IGFBP3 Prevention of Retinopathy of Prematurity - Full Text View - ClinicalTrials.gov [Internet]. [cited 2019 Jul 9]. Available from: <https://clinicaltrials.gov/ct2/show/NCT01096784>
  175. Schmetterer L, Polak K. Role of nitric oxide in the control of ocular blood flow. *Prog Retin Eye Res*. 2001 Nov;20(6):823–47.
  176. Ando A, Yang A, Mori K, Yamada H, Yamada E, Takahashi K, et al. Nitric oxide is proangiogenic in the retina and choroid. *J Cell Physiol*. 2002 Apr;191(1):116–24.
  177. Brooks SE, Gu X, Samuel S, Marcus DM, Bartoli M, Huang PL, et al. Reduced severity of oxygen-induced retinopathy in eNOS-deficient mice. *Invest Ophthalmol Vis Sci*. 2001 Jan;42(1):222–8.
  178. Tsukada T, Yokoyama K, Arai T, Takemoto F, Hara S, Yamada A, et al. Evidence of association of the eNOS gene polymorphism with plasma NO metabolite levels in humans. *Biochem Biophys Res Commun*. 1998 Apr 7;245(1):190–3.
  179. Silveira RC, Fortes Filho JB, Procianny RS. Assessment of the contribution of cytokine plasma levels to detect retinopathy of prematurity in very low birth weight infants. *Invest Ophthalmol Vis Sci*. 2011 Mar 10;52(3):1297–301.
  180. Sato T, Kusaka S, Shimojo H, Fujikado T. Vitreous levels of erythropoietin and vascular endothelial growth factor in eyes with retinopathy of prematurity. *Ophthalmology*. 2009 Sep;116(9):1599–603.

181. Matsuda K, Okamoto N, Kondo M, Arkwright PD, Karasawa K, Ishizaka S, et al. Mast cell hyperactivity underpins the development of oxygen-induced retinopathy. *J Clin Invest*. 2017 Nov 1;127(11):3987–4000.
182. Barde YA, Edgar D, Thoenen H. Purification of a new neurotrophic factor from mammalian brain. *EMBO J*. 1982 May 1;1(5):549–53.
183. Hyman C, Hofer M, Barde YA, Juhasz M, Yancopoulos GD, Squinto SP, et al. BDNF is a neurotrophic factor for dopaminergic neurons of the substantia nigra. *Nature*. 1991 Mar 21;350(6315):230–2.
184. Chen H, Weber AJ. BDNF enhances retinal ganglion cell survival in cats with optic nerve damage. *Invest Ophthalmol Vis Sci*. 2001 Apr;42(5):966–74.
185. Loeliger MM, Briscoe T, Rees SM. BDNF increases survival of retinal dopaminergic neurons after prenatal compromise. *Invest Ophthalmol Vis Sci*. 2008 Mar;49(3):1282–9.
186. Sood BG, Madan A, Saha S, Schendel D, Thorsen P, Skogstrand K, et al. Perinatal systemic inflammatory response syndrome and retinopathy of prematurity. *Pediatr Res*. 2010 Apr;67(4):394–400.
187. Rao R, Mashburn CB, Mao J, Wadhwa N, Smith GM, Desai NS. Brain-derived neurotrophic factor in infants <32 weeks gestational age: correlation with antenatal factors and postnatal outcomes. *Pediatr Res*. 2009 May;65(5):548–52.
188. Hellgren G, Willett K, Engstrom E, Thorsen P, Hougaard DM, Jacobsson B, et al. Proliferative retinopathy is associated with impaired increase in BDNF and RANTES expression levels after preterm birth. *Neonatology*. 2010 Nov 9;98(4):409–18.
189. Hartnett ME, Morrison MA, Smith S, Yanovitch TL, Young TL, Colaizy T, et al. Genetic variants associated with severe retinopathy of prematurity in extremely low birth weight infants. *Invest Ophthalmol Vis Sci*. 2014 Aug 12;55(10):6194–203.
190. Moravski CJ, Kelly DJ, Cooper ME, Gilbert RE, Bertram JF, Shahinfar S, et al. Retinal neovascularization is prevented by blockade of the renin-angiotensin system. *Hypertension*. 2000 Dec;36(6):1099–104.
191. Holash J, Wiegand SJ, Yancopoulos GD. New model of tumor angiogenesis: dynamic balance between vessel regression and growth mediated by angiopoietins and VEGF. *Oncogene*. 1999 Sep 20;18(38):5356–62.
192. Lee J, Kim KE, Choi D-K, Jang JY, Jung J-J, Kiyonari H, et al. Angiopoietin-1 guides directional angiogenesis through integrin  $\alpha\beta 5$  signaling for recovery of ischemic retinopathy. *Sci Transl Med*. 2013 Sep 18;5(203):203ra127.
193. Campochiaro PA. Molecular pathogenesis of retinal and choroidal vascular diseases. *Prog Retin Eye Res*. 2015 Nov;49:67–81.
194. Feng Y, vom Hagen F, Pfister F, Djokic S, Hoffmann S, Back W, et al. Impaired pericyte recruitment and abnormal retinal angiogenesis as a result of angiopoietin-2 overexpression. *Thromb Haemost*. 2007 Jan;97(1):99–108.



195. Sato T, Shima C, Kusaka S. Vitreous levels of angiopoietin-1 and angiopoietin-2 in eyes with retinopathy of prematurity. *Am J Ophthalmol*. 2011 Feb;151(2):353–7.e1.
196. Watanabe D, Suzuma K, Matsui S, Kurimoto M, Kiryu J, Kita M, et al. Erythropoietin as a retinal angiogenic factor in proliferative diabetic retinopathy. *N Engl J Med*. 2005 Aug 25;353(8):782–92.
197. Chen J, Connor KM, Aderman CM, Smith LEH. Erythropoietin deficiency decreases vascular stability in mice. *J Clin Invest*. 2008 Feb;118(2):526–33.
198. Yang Z, Wang H, Jiang Y, Hartnett ME. VEGFA activates erythropoietin receptor and enhances VEGFR2-mediated pathological angiogenesis. *Am J Pathol*. 2014 Apr;184(4):1230–9.
199. Yang N, Zhang W, He T, Xing Y. Exogenous erythropoietin aggravates retinal neovascularization in a murine model of proliferative retinopathy. *Turk J Med Sci*. 2017 Nov 13;47(5):1642–50.
200. Semenza GL. Oxygen sensing, homeostasis, and disease. *N Engl J Med*. 2011 Aug 11;365(6):537–47.
201. Lin M, Chen Y, Jin J, Hu Y, Zhou KK, Zhu M, et al. Ischaemia-induced retinal neovascularisation and diabetic retinopathy in mice with conditional knockout of hypoxia-inducible factor-1 in retinal Müller cells. *Diabetologia*. 2011 Jun;54(6):1554–66.
202. Hoppe G, Yoon S, Gopalan B, Savage AR, Brown R, Case K, et al. Comparative systems pharmacology of HIF stabilization in the prevention of retinopathy of prematurity. *Proc Natl Acad Sci U S A*. 2016 May 3;113(18):E2516–25.
203. Morita M, Ohneda O, Yamashita T, Takahashi S, Suzuki N, Nakajima O, et al. HLF/HIF-2 $\alpha$  is a key factor in retinopathy of prematurity in association with erythropoietin. *EMBO J*. 2003 Mar 3;22(5):1134–46.
204. Huang B, Deora AB, He K-L, Chen K, Sui G, Jacovina AT, et al. Hypoxia-inducible factor-1 drives annexin A2 system-mediated perivascular fibrin clearance in oxygen-induced retinopathy in mice. *Blood*. 2011 Sep 8;118(10):2918–29.
205. Rankin EB, Biju MP, Liu Q, Unger TL, Rha J, Johnson RS, et al. Hypoxia-inducible factor--2 (HIF-2) regulates hepatic erythropoietin in vivo. *J Clin Invest*. 2007;117(4):1068–77.
206. Wolfsberg TG, Primakoff P, Myles DG, White JM. ADAM, a novel family of membrane proteins containing A Disintegrin And Metalloprotease domain: multipotential functions in cell-cell and cell-matrix interactions. *J Cell Biol*. 1995 Oct;131(2):275–8.
207. Weskamp G, Mendelson K, Swendeman S, Le Gall S, Ma Y, Lyman S, et al. Pathological neovascularization is reduced by inactivation of ADAM17 in endothelial cells but not in pericytes. *Circ Res*. 2010 Mar 19;106(5):932–40.
208. Guaiquil VH, Hewing NJ, Chiang MF, Rosenblatt MI, Chan RVP, Blobel CP. A murine model for retinopathy of prematurity identifies endothelial cell proliferation as a potential mechanism for plus disease. *Invest Ophthalmol Vis Sci*. 2013 Aug 7;54(8):5294–302.
209. Mahmoodi M, Sahebjam S, Smookler D, Khokha R, Mort JS. Lack of tissue inhibitor of metalloproteinases-3 results in an enhanced inflammatory response in antigen-induced arthritis.

- Am J Pathol. 2005 Jun;166(6):1733–40.
210. Darlow BA, Lui K, Kusuda S, Reichman B, Håkansson S, Bassler D, et al. International variations and trends in the treatment for retinopathy of prematurity. *Br J Ophthalmol*. 2017 Oct;101(10):1399–404.
211. Mora JS, Waite C, Gilbert CE, Breidenstein B, Sloper JJ. A worldwide survey of retinopathy of prematurity screening. *Br J Ophthalmol*. 2018 Jan;102(1):9–13.
212. Moleta C, Campbell JP, Kalpathy-Cramer J, Chan RVP, Ostmo S, Jonas K, et al. Plus Disease in Retinopathy of Prematurity: Diagnostic Trends in 2016 Versus 2007. *Am J Ophthalmol*. 2017 Apr;176:70–6.
213. Manja V, Lakshminrusimha S, Cook DJ. Oxygen saturation target range for extremely preterm infants: a systematic review and meta-analysis. *JAMA Pediatr*. 2015 Apr;169(4):332–40.
214. Meng Q, Huang L, Sun Y, Bai Y, Wang B, Yu W, et al. Effect of high-density lipoprotein metabolic pathway gene variations and risk factors on neovascular age-related macular degeneration and polypoidal choroidal vasculopathy in China. *PLoS One*. 2015;10(12):e0143924.
215. Helgason H, Sulem P, Duvvari MR, Luo H, Thorleifsson G, Stefansson H, et al. A rare nonsynonymous sequence variant in C3 is associated with high risk of age-related macular degeneration. *Nat Genet*. 2013 Nov;45(11):1371–4.
216. Fritsche LG, Chen W, Schu M, Yaspan BL, Yu Y, Thorleifsson G, et al. Seven new loci associated with age-related macular degeneration. *Nat Genet*. 2013 Apr;45(4):433–9, 439e1–2.
217. Burdon KP, Fogarty RD, Shen W, Abhary S, Kaidonis G, Appukuttan B, et al. Genome-wide association study for sight-threatening diabetic retinopathy reveals association with genetic variation near the GRB2 gene. *Diabetologia*. 2015 Oct;58(10):2288–97.
218. Sheu WH-H, Kuo JZ, Lee I-T, Hung Y-J, Lee W-J, Tsai H-Y, et al. Genome-wide association study in a Chinese population with diabetic retinopathy. *Hum Mol Genet*. 2013 Aug 1;22(15):3165–73.
219. Khor CC, Do T, Jia H, Nakano M, George R, Abu-Amero K, et al. Genome-wide association study identifies five new susceptibility loci for primary angle closure glaucoma. *Nat Genet*. 2016 May;48(5):556–62.
220. Bailey JNC, Loomis SJ, Kang JH, Allingham RR, Gharahkhani P, Khor CC, et al. Genome-wide association analysis identifies TXNRD2, ATXN2 and FOXC1 as susceptibility loci for primary open-angle glaucoma. *Nat Genet*. 2016 Feb;48(2):189–94.
221. Vithana EN, Khor C-C, Qiao C, Nongpiur ME, George R, Chen L-J, et al. Genome-wide association analyses identify three new susceptibility loci for primary angle closure glaucoma. *Nat Genet*. 2012 Oct;44(10):1142–6.
222. Verhoeven VJM, Hysi PG, Wojciechowski R, Fan Q, Guggenheim JA, Höhn R, et al. Genome-wide meta-analyses of multiethnic cohorts identify multiple new susceptibility loci for refractive error and myopia. *Nat Genet*. 2013 Mar;45(3):314–8.

223. Solouki AM, Verhoeven VJM, van Duijn CM, Verkerk AJMH, Ikram MK, Hysi PG, et al. A genome-wide association study identifies a susceptibility locus for refractive errors and myopia at 15q14. *Nat Genet.* 2010 Oct;42(10):897–901.
224. Hysi PG, Young TL, Mackey DA, Andrew T, Fernández-Medarde A, Solouki AM, et al. A genome-wide association study for myopia and refractive error identifies a susceptibility locus at 15q25. *Nat Genet.* 2010 Oct;42(10):902–5.
225. Syreeni A, El-Osta A, Forsblom C, Sandholm N, Parkkonen M, Tarnow L, et al. Genetic examination of SETD7 and SUV39H1/H2 methyltransferases and the risk of diabetes complications in patients with type 1 diabetes. *Diabetes.* 2011 Nov;60(11):3073–80.
226. Zhong Q, Kowluru RA. Epigenetic changes in mitochondrial superoxide dismutase in the retina and the development of diabetic retinopathy. *Diabetes.* 2011 Apr;60(4):1304–13.
227. McArthur K, Feng B, Wu Y, Chen S, Chakrabarti S. MicroRNA-200b Regulates Vascular Endothelial Growth Factor–Mediated Alterations in Diabetic Retinopathy. *Diabetes.* 2011 Apr 1;60(4):1314–23.
228. Kowluru RA, Santos JM, Mishra M. Epigenetic modifications and diabetic retinopathy. *Biomed Res Int.* 2013 Oct 28;2013:635284.
229. Zhong Q, Kowluru RA. Regulation of matrix metalloproteinase-9 by epigenetic modifications and the development of diabetic retinopathy. *Diabetes.* 2013 Jul;62(7):2559–68.
230. Psaty BM, O'Donnell CJ, Gudnason V, Lunetta KL, Folsom AR, Rotter JJ, et al. Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium [Internet]. Vol. 2, Circulation: Cardiovascular Genetics. 2009. p. 73–80. Available from: <http://dx.doi.org/10.1161/circgenetics.108.829747>
231. Lu Y, Vitart V, Burdon KP, Khor CC, Bykhovskaya Y, Mirshahi A, et al. Genome-wide association analyses identify multiple loci associated with central corneal thickness and keratoconus. *Nat Genet.* 2013 Feb;45(2):155–63.
232. Jensen RA, Sim X, Smith AV, Li X, Jakobsdóttir J, Cheng C-Y, et al. Novel Genetic Loci Associated With Retinal Microvascular Diameter. *Circ Cardiovasc Genet.* 2016 Feb;9(1):45–54.
233. Genes Associated With Bronchopulmonary Dysplasia and Retinopathy of Prematurity - Full Text View - ClinicalTrials.gov [Internet]. [cited 2019 Jul 10]. Available from: <https://clinicaltrials.gov/ct2/show/NCT01780155>
234. i-ROP.com [Internet]. [cited 2019 Jul 10]. Available from: <http://i-rop.github.io/>
235. dbGaP | phs000428.v2.p2 | Health and Retirement Study (HRS) [Internet]. [cited 2019 Jul 9]. Available from: [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study\\_id=phs000428.v2.p2](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000428.v2.p2)
236. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet.* 2010 Jul;42(7):565–9.
237. Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, et al. Genome

- partitioning of genetic variation for complex traits using common SNPs. *Nat Genet.* 2011 Jun;43(6):519–25.
238. Wood et A. Defining the role of common variation in the genomic and biological architecture of adult human height | Kopernio [Internet]. [cited 2019 Jul 9]. Available from: <https://kopernio.com/viewer?doi=10.1038/ng.3097&token=WzY4ODgzMywiMTAuMTAzOC9uZy4zMdK3lI0.cWaNbTWvd25l9O4QW5kLc0lhRmE>
  239. Yang J, Loos et RJ. FTO genotype is associated with phenotypic variability of body mass index | Kopernio [Internet]. [cited 2019 Jul 9]. Available from: <https://kopernio.com/viewer?doi=10.1038/nature11401&token=WzY4ODgzMywiMTAuMTAzOC9uYXR1cmUxMTQwMSJd.Jb6qgA20E7ImUiKYaOhFzvo2hNQ>
  240. Olsen J, Melbye M, Olsen SF, Sørensen TI, Aaby P, Andersen AM, et al. The Danish National Birth Cohort--its background, structure and aim. *Scand J Public Health.* 2001 Dec;29(4):300–7.
  241. dbGaP | phs000103.v1.p1 | Genome-Wide Association Studies of Prematurity and Its Complications [Internet]. [cited 2019 Jul 9]. Available from: [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study\\_id=phs000103.v1.p1](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000103.v1.p1)
  242. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* 2009 Jun;5(6):e1000529.
  243. Delaneau O, The 1000 Genomes Project Consortium, Marchini J. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel [Internet]. Vol. 5, *Nature Communications*. 2014. Available from: <http://dx.doi.org/10.1038/ncomms4934>
  244. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience.* 2015 Feb 25;4(1):1–16.
  245. Freeman C, Marchini J. GTOOL. URL(accessed 08 23 13 ) [Internet]. 2007; Available from: <https://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html>
  246. Browning BL, Zhou Y, Browning SR. A one penny imputed genome from next generation reference panels [Internet]. Available from: <http://dx.doi.org/10.1101/357806>
  247. Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, et al. The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res.* 2019 Jan 8;47(D1):D853–8.
  248. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 2001 Jan 1;29(1):308–11.
  249. Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, et al. Ensembl 2019. *Nucleic Acids Res.* 2019 Jan 8;47(D1):D745–51.
  250. Csardi G, Nepusz T, Others. The igraph software package for complex network research. *InterJournal, Complex Systems.* 2006;1695(5):1–9.
  251. Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res.* 2015 Jan;43(Database issue):D1049–56.

252. Sergushichev AA. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation [Internet]. bioRxiv. 2016 [cited 2019 Jul 19]. p. 060012. Available from: <https://www.biorxiv.org/content/10.1101/060012v1>
253. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005 Oct 25;102(43):15545–50.
254. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst*. 2015 Dec 23;1(6):417–25.
255. Lettre G. The osteoarthritis and height GDF5 locus yields its secrets. *Nat Genet*. 2017 Jul 27;49(8):1165–6.
256. Perry MJ, McDougall KE, Hou S-C, Tobias JH. Impaired growth plate function in bmp-6 null mice. *Bone*. 2008 Jan;42(1):216–25.
257. Marcelino J, Sciortino CM, Romero MF, Ulatowski LM, Ballock RT, Economides AN, et al. Human disease-causing NOG missense mutations: effects on noggin secretion, dimer formation, and bone morphogenetic protein binding. *Proc Natl Acad Sci U S A*. 2001 Sep 25;98(20):11353–8.
258. Zhang Y, Marmorstein LY. Focus on molecules: fibulin-3 (EFEMP1). *Exp Eye Res*. 2010 Mar;90(3):374–5.
259. Faraco J, Bashir M, Rosenbloom J, Francke U. Characterization of the human gene for microfibril-associated glycoprotein (MFAP2), assignment to chromosome 1p36.1–p35, and linkage to D1S170. *Genomics*. 1995 Feb 10;25(3):630–7.
260. Collod G, Chu ML, Sasaki T, Coulon M, Timpl R, Renkart L, et al. Fibulin-2: genetic mapping and exclusion as a candidate gene in Marfan syndrome type 2. *Eur J Hum Genet*. 1996;4(5):292–5.
261. Männik J, Vaas P, Rull K, Teesalu P, Rebane T, Laan M. Differential expression profile of growth hormone/chorionic somatomammotropin genes in placenta of small- and large-for-gestational-age newborns. *J Clin Endocrinol Metab*. 2010 May;95(5):2433–42.
262. Lodge EJ, Russell JP, Patist AL, Francis-West P, Andoniadou CL. Expression Analysis of the Hippo Cascade Indicates a Role in Pituitary Stem Cell Development. *Front Physiol*. 2016 Mar 31;7:114.
263. Lee SH, Goddard ME, Wray NR, Visscher PM. A better coefficient of determination for genetic profile analysis. *Genet Epidemiol*. 2012 Apr;36(3):214–24.
264. Black KZ, Nichols HB, Eng E, Rowley DL. Prevalence of preterm, low birthweight, and small for gestational age delivery after breast cancer diagnosis: a population-based study. *Breast Cancer Res*. 2017 Jan 31;19(1):11.
265. Chen C-M, Liu Y-C, Chen Y-J, Chou H-C. Genome-Wide Analysis of DNA Methylation in Hyperoxia-Exposed Newborn Rat Lung. *Lung*. 2017 Oct;195(5):661–9.
266. O'Shea LC, Mehta J, Lonergan P, Hensey C, Fair T. Developmental competence in oocytes and cumulus cells: candidate genes and networks. *Syst Biol Reprod Med*. 2012 Apr;58(2):88–101.

267. Yin M, Lü M, Yao G, Tian H, Lian J, Liu L, et al. Transactivation of microRNA-383 by steroidogenic factor-1 promotes estradiol release from mouse ovarian granulosa cells by targeting RBMS1. *Mol Endocrinol*. 2012 Jul;26(7):1129–43.
268. Alvine T, Dhasarathy A, Bundy A, Bhattacharya A, Darland D, Hur J, et al. RBMS1 Methylation and mRNA Expression Are Differentially Regulated in Placenta Tissue from Obese Women (P11-131-19). *Curr Dev Nutr* [Internet]. 2019 Jun 1 [cited 2019 Jul 26];3(Supplement\_1). Available from: [https://academic.oup.com/cdn/article-abstract/3/Supplement\\_1/nzz048.P11-131-19/5517822](https://academic.oup.com/cdn/article-abstract/3/Supplement_1/nzz048.P11-131-19/5517822)
269. Conrad KP, Jeyabalan A, Founds SA, Hogge WA. Gene expression related to preeclampsia [Internet]. US Patent. 20110171650:A1, 2011 [cited 2019 Jul 26]. Available from: <https://patentimages.storage.googleapis.com/c1/2c/bf/0a9101e88a9364/US20110171650A1.pdf>
270. Tai-Nagara I, Yoshikawa Y, Numata N, Ando T, Okabe K, Sugiura Y, et al. Placental labyrinth formation in mice requires endothelial FLRT2/UNC5B signaling. *Development*. 2017 Jul 1;144(13):2392–401.
271. Kim SJ, Port AD, Swan R, Campbell JP, Chan RVP, Chiang MF. Retinopathy of prematurity: a review of risk factors and their clinical significance. *Surv Ophthalmol*. 2018 Sep;63(5):618–37.
272. Vösa U, Claringbould A, Westra H-J, Bonder MJ, Deelen P, Zeng B, et al. Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis [Internet]. *bioRxiv*. 2018 [cited 2019 Aug 14]. p. 447367. Available from: <https://www.biorxiv.org/content/10.1101/447367v1>
273. Choi SW. PRSet - PRSice-2 [Internet]. [cited 2019 Jul 24]. Available from: [http://www.prsice.info/prset\\_detail/](http://www.prsice.info/prset_detail/)
274. Rajaram S, Carlson SE, Koo WW, Rangachari A, Kelly DP. Insulin-like growth factor (IGF)-I and IGF-binding protein 3 during the first year in term and preterm infants. *Pediatr Res*. 1995 May;37(5):581–5.
275. Hikino S, Ihara K, Yamamoto J, Takahata Y, Nakayama H, Kinukawa N, et al. Physical growth and retinopathy in preterm infants: involvement of IGF-I and GH. *Pediatr Res*. 2001 Dec;50(6):732–6.
276. Hiden U, Glitzner E, Hartmann M, Desoye G. Insulin and the IGF system in the human placenta of normal and diabetic pregnancies. *J Anat*. 2009 Jul;215(1):60–8.
277. Liu JZ, McRae AF, Nyholt DR, Medland SE, Wray NR, Brown KM, et al. A versatile gene-based test for genome-wide association studies. *Am J Hum Genet*. 2010 Jul 9;87(1):139–45.
278. Nam D, Kim J, Kim S-Y, Kim S. GSA-SNP: a general approach for gene set analysis of polymorphisms. *Nucleic Acids Res*. 2010 Jul;38(Web Server issue):W749–54.
279. Levine ME, Langfelder P, Horvath S. A Weighted SNP Correlation Network Method for Estimating Polygenic Risk Scores. *Methods Mol Biol*. 2017;1613:277–90.
280. Zeng B, Lloyd-Jones LR, Montgomery GW, Metspalu A, Esko T, Franke L, et al. Comprehensive Multiple eQTL Detection and Its Application to GWAS Interpretation. *Genetics*. 2019 Jul;212(3):905–18.

281. Lalonde S, Ehrhardt DW, Loqué D, Chen J, Rhee SY, Frommer WB. Molecular and cellular approaches for the detection of protein-protein interactions: latest techniques and current limitations. *Plant J.* 2008 Feb;53(4):610–35.
282. Ferguson KK, Chin HB. Environmental chemicals and preterm birth: Biological mechanisms and the state of the science. *Curr Epidemiol Rep.* 2017 Mar;4(1):56–71.
283. Brown JM, Campbell JP, Beers A, Chang K, Ostmo S, Chan RVP, et al. Automated Diagnosis of Plus Disease in Retinopathy of Prematurity Using Deep Convolutional Neural Networks. *JAMA Ophthalmol.* 2018 Jul 1;136(7):803–10.
284. Coyner AS, Swan R, Campbell JP, Ostmo S, Brown JM, Kalpathy-Cramer J, et al. Automated Fundus Image Quality Assessment in Retinopathy of Prematurity Using Deep Convolutional Neural Networks. *Ophthalmol Retina.* 2019 May;3(5):444–50.
285. Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *Int J Methods Psychiatr Res.* 2018 Jun;27(2):e1608.
286. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. *Nature.* 2015 Oct 1;526(7571):68–74.