Integrated Protein Network Analysis of Whole Exome Sequencing of Severe Preeclampsia

Jessica Schuster, Ph.D. [1,2], George A. Tollefson B.S. [1], Valeria Zarate B.S.[1], Anthony Agudelo B.S. [1], Joan Stabila B.S. [1], Ashok Ragavendran Ph.D. [3,4], James Padbury M.D. [1,2,5], Alper Uzun Ph.D. [1,2,4,5*]

1) Pediatrics, Women and Infants Hospital, Providence, RI; 2) Pediatrics, Warren Alpert Medical School, Brown University, Providence, RI; 3) Center for Computation and Visualization, Brown University, Providence, RI; 4) Computational Biology of Human Disease, Brown University, Providence, RI. 5) Center for Computational Molecular Biology, Brown University, Providence, RI.

Short title: Whole Exome Sequencing of Severe Preeclampsia

*** Corresponding author:**

Associate Professor of Pediatrics Alper Uzun, PhD

email: alper_uzun@brown.edu

Abstract

Preeclampsia is a hypertensive disorder of pregnancy, which complicates up to 15 % of US deliveries. It is an idiopathic disorder with complex disease genetics associated with several different phenotypes. We sought to determine if the genetic architecture of preeclampsia can be described by clusters of patients with variants in genes in shared protein interaction networks. We performed a case-control study using whole exome sequencing on early onset preeclamptic mothers with severe features and control mothers with uncomplicated pregnancies. The study was conducted at Women & Infants Hospital of Rhode Island (WIH). A total of 143 patients were enrolled, 61 women with early onset preeclampsia with severe features based on ACOG criteria, and 82 control women at term, matched for race and ethnicity. The main outcomes are variants associated with severe preeclampsia and demonstration of the genetic architecture of preeclampsia. A network analysis and visualization tool, Proteinarium, was used to confirm there are clusters of patients with shared gene networks associated with severe preeclampsia. The majority of the sequenced patients appear in two significant clusters. We identified one case dominant and one control dominant cluster. Thirteen genes were unique to the case dominated cluster. Among these genes, LAMB2, PTK2, RAC1, QSOX1, FN1, and VCAM1 have known associations with the pathogenic mechanisms of preeclampsia. Using the exome-wide sequence variants, combined with these 13 identified network genes, we generated a polygenetic risk score for severe preeclampsia with an AUC of 0.57. Using bioinformatic analysis, we were able to identify subsets of patients with shared protein interaction networks, thus confirming our hypothesis about the genetic architecture of preeclampsia. The unique genes identified

in the cluster associated with severe preeclampsia were able to increase the predictive power of the polygenic risk score.

Introduction

Preeclampsia is a hypertensive disorder of pregnancy. It is associated with a higher risk of hypertension and cardiovascular disease later in life. Women who had preeclampsia have a twofold increased risk of death from cardiovascular diseases [1, 2]. There is evidence that preeclampsia originates in part from genetic causes that include contributions from the maternal, paternal and fetal genome [3-7]. The role of genetics in preeclampsia is supported by family-based observations [8, 9] with more than 100 studies showing a 2- to 5-fold increased risk among family members of affected women [10-15]. The heritability of preeclampsia is up to 52% [8, 16]. The recurrence risk for preeclampsia in the daughters of either eclamptic or preeclamptic patients is 20-40% [17, 18]. However, there is no current consensus among the published results in regards to associated genes and the pathogenesis of disease.

Genetic risk for most complex diseases involves the interaction of multiple genes in discrete networks and pathways [19]. Although complex diseases show increased recurrence risk in families, they do not follow a simple Mendelian pattern of inheritance [20]. Computational methods have been used to analyze the network of genes that are linked to a variety of disorders like autism and to find biological subnetworks due to the genetic heterogeneity of the disease [21]. There are several studies employing computational methods to identify important genes associated with hypertension. Ran et al analyzed protein-protein interaction (PPI) network topology and molecular connectivity between protein pathways to identify associations with hypertension [22]. Researchers developed a machine-learning algorithm to predict novel hypertension associated genes [23].

We hypothesize that the genetic architecture of complex diseases like preeclampsia is described by clusters of patients with variants in genes in shared protein interaction networks. We sought to test this hypothesis using whole exome sequencing in carefully selected patients with severe preeclampsia**.** We compared variants identified in women with early onset, idiopathic preeclampsia with term controls without personal or family history of pregnancy related hypertensive disorders. We built and implemented *Proteinarium,* a multi-sample, PPI tool, to identify clusters of patients with shared PPI networks.

Methods

Study population

Women & Infants Hospital of Rhode Island (WIH) is the only provider of high-risk perinatal services in Rhode Island, northeastern Connecticut and southeastern Massachusetts. We used this population-based service to enroll preeclamptic mothers with early onset, severe features, based on ACOG criteria, as well as term mothers with no history of preeclampsia [24].

This case/control study was approved by the Institutional Review Board of WIH (Project ID: WIH 16-0031). Between the years 2016-2020, we reviewed the records of all early-onset preeclamptic mothers with severe features delivering < 34 weeks. Following informed consent, we asked explicit questions about preeclampsia in mother, grandmother, first order relatives and also paternal relatives. Clinical history, with an emphasis on additional risk factors including medical illnesses and drug use was recorded. Hypertensive disorders include a broad range of different phenotypes. Again, in order to leverage the likelihood of genetic discovery associated with preeclampsia, we

excluded mothers with personal or family history of other hypertensive disorders. Controls were mothers who delivered ≥ 37 weeks' gestation for whom the formal genetic interview revealed no history of preterm birth or pregnancy related hypertensive disorders on either the maternal or paternal side of the pedigree. A total of 143 patients were enrolled, 61 women with early onset preeclampsia with severe features, and 82 control women at term, matched for race and ethnicity.

Whole Exome Sequencing

Residual maternal whole blood was obtained from each mother and stored at -80°C. Samples were sent to an outside facility for whole exome sequencing that was blind to disease status. The library was sequenced on an Illumina HiSeq 4000 using 150 bp paired-end protocols.

Sequence Data

For variant discovery we used the Gene Analysis Tool Kit (GATK) V4 to analyze the sequence reads [25]. Haplotype caller was applied for variant detection [26]. Variants were flagged as low quality and filtered using established metrics: if three or more variants were detected within 10bp; if four or more alignments mapped to different locations equally well; if coverage was less than ten reads; if quality score < 30; if low quality for a particular sequence depth (variant confidence/unfiltered depth < 1.5); and if strand bias was observed (Phred-scaled p-values using Fisher's Exact Test > 200).

Genotype Testing

To identify variants that were differentially abundant between cases and controls, we used a Markov Chain Monte Carlo (MCMC) Fisher Exact Test to compare the frequency of the homozygous reference, homozygous alternative, and the heterozygous genotypes between cases and controls. Eigenstrat detected no significant population stratification [27].

Variant Annotation

We applied a strict filter-based annotation using ANNOVAR [28]. We identified deleterious variants with Polyphen 2 HDIV, SIFT and CADD [29-32]. We used the following thresholds: Polyphen 2 HDIV prediction if a change is damaging (>=0.957), a SIFT score (<0.05), a CADD score >15, and minor allele frequency (MAF) <0.05 from the 1000 Genome Project [32].

Network Analysis

We hypothesized that the genetic architecture underlying complex disorders is best explained by subsets of patients with variants in shared networks and pathways sufficient to express the phenotype. To analyze our whole exome sequencing data, we implemented *Proteinarium,* our multi-sample PPI analysis and visualization tool [1]. We determined the genetic similarity between the clusters identified using separation testing. Combining genome wide variants with the unique genes identified by this network analysis, we generated a polygenic risk score prediction model. These analyses are explained in detail in the Supplementary Methods.

Results

The clinical characteristics and the race/ethnicity distribution of the patients are shown in Table 1. As can be seen from Table 1, gestational age at delivery, systolic blood pressure, frequency of proteinuria, impaired liver function, thrombocytopenia, cerebral visual symptoms and fetal growth retardation were all significantly different between the groups, which was expected by our definition of severe preeclampsia.

High quality sequence data with a Phred score $\geq$30 from well-balanced pools with over 19,000,000 reads/patient, 40X average depth of coverage, with more than 80% of sequence reads with at least 20X coverage were observed. We identified a total of 528,630 variants including 187,915 exonic variants. The work flow for the univariate analysis is shown in Figure 1. After application of the initial filters for coverage and variant pathogenicity, there were 8,867 predicted deleterious variants (available at Online_Supplemental_Table 1). Among these, 21 variants were nominally associated with preeclampsia by genotype testing. All were non-synonymous, exonic variants (Table 2). Nonetheless, none of these variants met genome-wide significance after correction for multiple comparison testing.

Proteinarium was used to identify clusters of patients with shared networks associated with severe preeclampsia and the resulting dendrogram is shown in Figure 2A. Out of the 143 patients sequenced, 129 patients were assigned to two statistically significant clusters. ($p < 0.0001$). The inset in Figure 2A shows the number of cases and controls in each cluster. Cluster A had significantly more cases than controls, containing 47 of the 61 case subjects. The layered network for the case-dominated Cluster A is shown in Figure 2B. There are 13 genes which are unique to Cluster A highlighted in red in the layered network

graph. Most have defined functional roles or implications for preeclampsia, Table 3. Cluster B had significantly more controls than cases, including 61 of the 82 subjects. The layered network for Cluster B is shown in Figure 2C. The unique genes from the layered network graph of Cluster B, shown in blue, are listed in Supplemental Table 2. When we compared the sequence data of the samples not assigned to clusters with those that were assigned, we did not find significant differences in the average depth of coverage. Likewise, there were no significant differences in clinical/phenotypic characteristics when comparing the subjects in the significant clusters with the subjects that were not in these clusters (data not shown).

The comparison of the unique genes from the case and the control dominated clusters revealed a positive separation score, confirming that the layered PPI networks of these two patient subgroups exist in distinct areas of the interactome. We ran GO term analysis using DAVID software on all genes of the network from Cluster A and from Cluster B, Supplementary Table 3 [33, 34]. We found significantly enriched biological processes, molecular functions and cellular components based on Bonferroni corrected p-value for the case and control dominated networks. Prominent among the biological processes and molecular functions associated with preeclampsia were antigen processing and presentation, cellular movement (axon guidance and microtubules) and T cell receptor signaling.

We previously reported the Database for Preeclampsia (dbPEC) which archives a curation-based collection of genes associated with preeclampsia and their association with clinical features and concurrent conditions [35]. We compared the genes from our univariate analysis and the genes from both case and control dominated layered networks to those in

the database. We found two overlapping genes from the univariate gene list (TTN and CCL14) that were included in dbPEC. We also found three overlapping genes from the layered network of Cluster A (FN1, KIF2A, VCAM1). By over representation analysis, Cluster A is significantly enriched for genes previously shown to be associated with preeclampsia in dbPEC ($p < 0.0033$).

Aggregating information from an array of risk alleles and or genes, also known as a polygenic risk score (PRS), is a means to predict an individual's phenotype or risk of disease based on their genomic profile.[36] Gene-based models trained on the training set and tested on the test set with an LFDR threshold of 0.1 achieved the highest AUC for the ROC of 0.524. We hypothesized that our PPI network analysis would provide increased predictive power, and thus we also tested these models including the 13 unique genes identified in the preeclampsia dominant cluster. This resulted in an increase in the AUC to 0.57 with a 95% confidence interval between 0.383 and 0.732.

Discussion

Preeclampsia is a life-threatening, multi-system hypertensive disorder of pregnancy, which complicates up to 15 % of US deliveries [8, 36-38]. The incidence is increasing[38]. It is recognized as a leading cause of maternal and fetal morbidity and mortality worldwide [36]. Preeclampsia is characterized by varying degrees of maternal symptoms including elevated blood pressure, proteinuria and fetal growth restriction [39]. Many clinicians believe that preeclampsia, severe preeclampsia, and early vs late preeclampsia are different disorders [40-42]. Previously, using bioinformatic methods, we showed that there are discrete gene sets associated with these different phenotypes of preeclampsia [35].

We performed whole exome sequencing on women with idiopathic early-onset preeclampsia with severe features and singleton births <34 weeks' gestation and compared them to term controls with no family history of preeclampsia. We developed *Proteinarium,* a multi-sample, PPI analysis and visualization tool, to identify clusters of patients with shared protein-protein interaction networks [43]. Using seed genes from each patient, *Proteinarium* built individual networks based on the STRING database. The similarities between individual PPI networks were evaluated using a distance metric for clustering the samples. We identified a single, significant cluster with a predominance of cases with early-onset, severe features of preeclampsia. We also identified a single control-dominated cluster. The separation test of the unique genes from case and control dominated clusters confirmed that the two subnetworks forming clusters A and B exist in the different regions of the interactome. These results support our hypothesis that the genetic architecture of complex diseases is characterized by clusters of patients that have variants in shared gene networks and provide insights into the genetics of severe preeclampsia.

Several of the unique genes from the case dominated network have very plausible mechanistic connections to preeclampsia. Laminin β2 (LAMB2) is a glomerular basement membrane (GBM) component, required for proper functioning of the glomerular filtration barrier. It has a role in proteinuria [44] and serum laminin levels in preeclamptic patients are significantly higher than those in normal pregnancy [45]. Hypoxia-induced upregulation of Quiescin Sulfhydryl Oxidase 1 (QSOX1) and an elevation in intracellular H_2O_2 leads to increased apoptosis in the placentae of pregnancies complicated by preeclampsia [46]. QSOX1 protein is found in circulating

extracellular vesicles of both preeclampsia and healthy pregnant women [47]. Fibronectin 1 (FN1) might promote the development of preeclampsia by modulating differentiation of human extravillous trophoblasts, as well as formation of focal adhesions [48-50]. Vascular Cell Adhesion Molecule 1 (VCAM1) is involved in cellular adhesion and serum concentrations of sVCAM-1 are significantly elevated in both mild and severe preeclampsia [51]. Invasion of maternal decidua and uterine spiral arteries by extravillous trophoblasts is required for establishment of normal placenta. Human trophoblast migration requires Rac Family Small GTPase 1 (RAC1) and Cell Division Cycle 42 (CDC42) [52]. Lower levels were found in preeclampsia samples than in normal term pregnancy samples, and decline significantly in severe preeclampsia [53]. Protein tyrosine kinase 2 (PTK2) (focal adhesion kinase) is differentially expressed in preeclampsia and reported as among the promising biomarkers for preeclampsia [54]. In the case-dominated subnetwork we observed Kinesin Family Member 2A (KIF2A) which is upregulated in the preeclamptic placenta [4]. Up-regulated genes in the preeclampsia placenta have been shown to be associated with the regulation of diverse cellular processes, including matrix degradation, trophoblast cell invasion, migration and proliferation [4].

There have been several sequencing efforts including whole genome, whole exome and targeted sequencing on an array of preeclampsia phenotypes from diverse populations [37, 55-64]. There is no consensus among the published results in regards to associated genes and variants. Since preeclampsia is a complex, polygenic disease, the lack of a consensus among these univariate comparisons might be expected in these early-stage studies. Among the 20 genes identified in our univariate analysis, only Titin (TTN) was identified in prior studies [55, 59]. Protein-altering mutations in TTN have been

identified in patients with cardiomyopathy and women with preeclampsia are more likely to carry TTN mutations associated with idiopathic cardiomyopathy and peripartum cardiomyopathy [59]. Additioinally, we found 2 genes, Major Histocompatibility Complex, Class II, DQ Alpha 1 (HLA-DQA1) and Inositol 1,4,5-Trisphosphate Receptor Type 1 (ITPR1) that were reported in previous studies of preeclampsia [56, 62]. None of these overlapping genes were among the unique genes identified in the shared layered networks. Likewise, no overlapping variants or genes were found in a recent genome-wide association meta-analysis study investigating genetic predispositions associated with preeclampsia [37].

Our analysis allowed us to identify clusters of patients with shared PPI networks associated with severe preeclampsia. Within the significant clusters, there were unique imputed genes (RAC1, KIF5B, PTK2, KIF5A, FN1, QSOX1, ARF4, VCAM1, CDC42, KIF2A) that were not among the top 60 seed genes selected by genotype testing. Nonetheless, our approach allowed us to identify these influential genes in the mechanism(s) underlying preeclampsia that would not otherwise have been identified by whole genome univariate variant analysis.

We also examined the unique proteins of the network of the control dominated cluster. Proteins in this network are associated with the ubiquitination process. They may serve a role that confer resilience against preeclampsia [65, 66]. Although there are studies showing a relationship with hypertension - ubiquitination process and pregnancy, this still needs further investigation [66].

Whole exome sequencing, combined with a novel, multi-sample network analysis, and careful phenotyping contributed to our discovery despite the relatively modest size of

our study. Concepts developed from network theory suggest that related diseases involve proteins in similar neighborhoods of the interactome [67]. Based on these concepts, we hypothesized that the genetic architecture of preeclampsia is described by subgroups of patients with variants in shared genes in specific networks and pathways. We identified a significant subgroup of cases with shared PPI networks associated with severe preeclampsia. We believe that the careful phenotyping resulted in the high percentage of subjects being successfully assigned to significant clusters and the ability to observe distinct separation between the case and control dominated clusters.

We used the Identified genes and their associated variants to generate a polygenic risk score. Of greatest importance, the unique genes in the case dominant cluster enhanced the predicted power of our polygenic risk score. These results compare favorably with results of others employing a similar approach [68]. The recent meta-analysis of GWAS studies generated a polygenic risk score from different genomic elements with an odds ratio of 1.25 in prediction of preeclampsia [37].

Strengths and Limitations

While we were not expecting each patient to appear in a significant cluster and our study included only a modest sample size, we identified a significant subgroup of patients with shared PPI networks associated with severe preeclampsia. In order to leverage the likelihood of genetics discovery, we focused exclusively on women with severe, early-onset preeclampsia. Our analysis was restricted to evaluation of genetic variants in the maternal genome only. Future studies including fetal and/or paternal data will enhance the likelihood of genetic discovery.

Conclusion

Using our unique network analysis, we were able to identify subsets of patients with shared networks, thus confirming our hypothesis about the genetic architecture of preeclampsia. Strict phenotyping of both cases and controls improved the likelihood of identifying these otherwise difficult to find genetic associations. Our network analysis identified genes which were imputed from the interactome and these imputed genes provide insights for severe preeclampsia that may otherwise have not been identified. As such, these are important candidates to include in meta-analyses of genetic associations with preeclampsia. Inclusion of the unique genes identified in cluster associated with severe preeclampsia increased the predictive power of the polygenic risk score. These results provide promise to further our understanding the mechanism underlying complex diseases like preeclampsia.

Acknowledgements

We thank the Kilguss Research Core at Women & Infants Hospital and The Center for Computation and Visualization (CCV) at Brown University.

Sources of Funding

This work was supported by grants from the National Institutes of Health *5P20GM109035-05* and *5P20GM121298-05.*

Disclosures

None

References

1. Bokslag A, van Weissenbruch M, Mol BW, and de Groot CJ, Preeclampsia; short and long-term consequences for mother and neonate. Early Hum Dev, 2016. **102**: p. 47-50. doi: 10.1016/j.earlhumdev.2016.09.007
2. Neerukonda S, Shariati F, Hart T, Stewart M, Elkayam U, and Qamruddin S, Cardiovascular effects of preeclampsia. Curr Opin Cardiol, 2020. **35**(4): p. 357-359. doi: 10.1097/HCO.0000000000000756
3. Cnattingius S, Reilly M, Pawitan Y, and Lichtenstein P, Maternal and fetal genetic factors account for most of familial aggregation of preeclampsia: a population-based Swedish cohort study. Am J Med Genet A, 2004. **130A**(4): p. 365-71. doi: 10.1002/ajmg.a.30257
4. Kobayashi H, The Impact of Maternal-Fetal Genetic Conflict Situations on the Pathogenesis of Preeclampsia. Biochem Genet, 2015. **53**(9-10): p. 223-34. doi: 10.1007/s10528-015-9684-y
5. Nilsson E, Salonen Ros H, Cnattingius S, and Lichtenstein P, The importance of genetic and environmental effects for pre-eclampsia and gestational hypertension: a family study. BJOG, 2004. **111**(3): p. 200-6. doi: 10.1111/j.1471-0528.2004.00042x.x
6. Than NG, Romero R, Tarca AL, Kekesi KA, Xu Y, Xu Z, Juhasz K, Bhatti G, Leavitt RJ, Gelencser Z, et al., Integrated Systems Biology Approach Identifies Novel Maternal and Placental Pathways of Preeclampsia. Front Immunol, 2018. **9**: p. 1661. doi: 10.3389/fimmu.2018.01661
7. Zusterzeel PL, te Morsche R, Raijmakers MT, Roes EM, Peters WH, and Steegers EA, Paternal contribution to the risk for pre-eclampsia. J Med Genet, 2002. **39**(1): p. 44-5. doi: 10.1136/jmg.39.1.44
8. Chappell S and Morgan L, Searching for genetic clues to the causes of pre-eclampsia. Clin Sci (Lond), 2006. **110**(4): p. 443-58. doi: 10.1042/CS20050323
9. Nejatizadeh A, Stobdan T, Malhotra N, and Pasha MA, The genetic aspects of pre-eclampsia: achievements and limitations. Biochem Genet, 2008. **46**(7-8): p. 451-79. doi: 10.1007/s10528-008-9163-9
10. Arngrimsson R, Bjornsson S, Geirsson RT, Bjornsson H, Walker JJ, and Snaedal G, Genetic and familial predisposition to eclampsia and pre-eclampsia in a defined population. Br J Obstet Gynaecol, 1990. **97**(9): p. 762-9. doi:
11. Chesley LC, Annitto JE, and Cosgrove RA, The familial factor in toxemia of pregnancy. Obstet Gynecol, 1968. **32**(3): p. 303-11. doi:
12. Cincotta RB and Brennecke SP, Family history of pre-eclampsia as a predictor for pre-eclampsia in primigravidas. Int J Gynaecol Obstet, 1998. **60**(1): p. 23-7. doi:
13. Mutze S, Rudnik-Schoneborn S, Zerres K, and Rath W, Genes and the preeclampsia syndrome. J Perinat Med, 2008. **36**(1): p. 38-58. doi: 10.1515/JPM.2008.004
14. Sutherland A, Cooper DW, Howie PW, Liston WA, and MacGillivray I, The indicence of severe pre-eclampsia amongst mothers and mothers-in-law of pre-eclamptics and controls. Br J Obstet Gynaecol, 1981. **88**(8): p. 785-91. doi: 10.1111/j.1471-0528.1981.tb01304.x
15. Ward K, Genetic factors in common obstetric disorders. Clin Obstet Gynecol, 2008. **51**(1): p. 74-83. doi: 10.1097/GRF.0b013e3181616545

16. Salonen Ros H, Lichtenstein P, Lipworth L, and Cnattingius S, Genetic effects on the liability of developing pre-eclampsia and gestational hypertension. Am J Med Genet, 2000. **91**(4): p. 256-60. doi:
17. Genc MR and Schantz-Dunn J, The role of gene-environment interaction in predicting adverse pregnancy outcome. Best Pract Res Clin Obstet Gynaecol, 2007. **21**(3): p. 491-504. doi: 10.1016/j.bpobgyn.2007.01.009
18. Serrano NC, Immunology and genetic of preeclampsia. Clin Dev Immunol, 2006. **13**(2-4): p. 197-201. doi: 10.1080/17402520600876903
19. Loscalzo J, Kohane I, and Barabasi AL, Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. Mol Syst Biol, 2007. **3**: p. 124. doi: 10.1038/msb4100163
20. Smith GD and Ebrahim S, 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol, 2003. **32**(1): p. 1-22. doi: 10.1093/ije/dyg070
21. Wall DP, Esteban FJ, Deluca TF, Huyck M, Monaghan T, Velez de Mendizabal N, Goni J, and Kohane IS, Comparative analysis of neurological disorders focuses genome-wide search for autism genes. Genomics, 2009. **93**(2): p. 120-9. doi: 10.1016/j.ygeno.2008.09.015
22. Ran J, Li H, Fu J, Liu L, Xing Y, Li X, Shen H, Chen Y, Jiang X, Li Y, et al., Construction and analysis of the protein-protein interaction network related to essential hypertension. BMC Syst Biol, 2013. **7**: p. 32. doi: 10.1186/1752-0509-7-32
23. Li YH, Zhang GG, and Wang N, Systematic Characterization and Prediction of Human Hypertension Genes. Hypertension, 2017. **69**(2): p. 349-355. doi: 10.1161/HYPERTENSIONAHA.116.08573
24. *American College of Obstetricians and Gynecologists (ACOG)*. [cited October 8, 2020; Available from: www.acog.org.
25. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al., From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics, 2013. **43**: p. 11 10 1-11 10 33. doi: 10.1002/0471250953.bi1110s43
26. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, Kling DE, Gauthier LD, Levy-Moonshine A, Roazen D, et al., Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv, 2018: p. 201178. doi: 10.1101/201178
27. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, and Reich D, Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet, 2006. **38**(8): p. 904-9. doi: 10.1038/ng1847
28. Wang K, Li M, and Hakonarson H, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res, 2010. **38**(16): p. e164. doi: 10.1093/nar/gkq603
29. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, and Sunyaev SR, A method and server for predicting damaging missense mutations. Nat Methods, 2010. **7**(4): p. 248-9. doi: 10.1038/nmeth0410-248
30. Ng PC and Henikoff S, SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res, 2003. **31**(13): p. 3812-4. doi: 10.1093/nar/gkg509

31. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, and Shendure J, A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet, 2014. **46**(3): p. 310-5. doi: 10.1038/ng.2892
32. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, et al., A global reference for human genetic variation. Nature, 2015. **526**(7571): p. 68-74. doi: 10.1038/nature15393
33. Huang da W, Sherman BT, and Lempicki RA, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc, 2009. **4**(1): p. 44-57. doi: 10.1038/nprot.2008.211
34. Huang da W, Sherman BT, and Lempicki RA, Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res, 2009. **37**(1): p. 1-13. doi: 10.1093/nar/gkn923
35. Triche EW, Uzun A, DeWan AT, Kurihara I, Liu J, Occhiogrosso R, Shen B, Parker J, and Padbury JF, Bioinformatic approach to the genetics of preeclampsia. Obstet Gynecol, 2014. **123**(6): p. 1155-61. doi: 10.1097/AOG.0000000000000293
36. Valenzuela FJ, Perez-Sepulveda A, Torres MJ, Correa P, Repetto GM, and Illanes SE, Pathogenesis of preeclampsia: the genetic component. J Pregnancy, 2012. **2012**: p. 632732. doi: 10.1155/2012/632732
37. Steinthorsdottir V, McGinnis R, Williams NO, Stefansdottir L, Thorleifsson G, Shooter S, Fadista J, Sigurdsson JK, Auro KM, Berezina G, et al., Genetic predisposition to hypertension is associated with preeclampsia in European and Central Asian women. Nat Commun, 2020. **11**(1): p. 5976. doi: 10.1038/s41467-020-19733-6
38. Bornstein EE, Y.; Chervenak,F.A.;Grünebaum, A., Concerning trends in maternal risk factors in the United States: 1989–2018. EClinicalMedicine, 2020. **29-30**. doi:
39. Jebbink J, Wolters A, Fernando F, Afink G, van der Post J, and Ris-Stalpers C, Molecular genetics of preeclampsia and HELLP syndrome - a review. Biochim Biophys Acta, 2012. **1822**(12): p. 1960-9. doi: 10.1016/j.bbadis.2012.08.004
40. Carreiras M, Montagnani S, and Layrisse Z, Preeclampsia: a multifactorial disease resulting from the interaction of the feto-maternal HLA genotype and HCMV infection. Am J Reprod Immunol, 2002. **48**(3): p. 176-83. doi: 10.1034/j.1600-0897.2002.01076.x
41. Raymond D and Peterson E, A critical review of early-onset and late-onset preeclampsia. Obstet Gynecol Surv, 2011. **66**(8): p. 497-506. doi: 10.1097/OGX.0b013e3182331028
42. American College of O, Gynecologists, and Task Force on Hypertension in P, Hypertension in pregnancy. Report of the American College of Obstetricians and Gynecologists' Task Force on Hypertension in Pregnancy. Obstet Gynecol, 2013. **122**(5): p. 1122-31. doi: 10.1097/01.AOG.0000437382.03963.88
43. Armanious D, Schuster J, Tollefson GA, Agudelo A, DeWan AT, Istrail S, Padbury J, and Uzun A, Proteinarium: Multi-sample protein-protein interaction analysis and visualization tool. Genomics, 2020. doi: 10.1016/j.ygeno.2020.07.028
44. Zhang A and Huang S, Progress in pathogenesis of proteinuria. Int J Nephrol, 2012. **2012**: p. 314251. doi: 10.1155/2012/314251
45. Furuhashi N, Kimura H, Nagae H, Yajima A, Kimura C, and Saito T, Serum laminin levels in normal pregnancy and preeclampsia. Gynecol Obstet Invest, 1993. **36**(3): p. 172-5. doi: 10.1159/000292620
46. Li J, Tong C, Xu P, Wang L, Han TL, Wen L, Luo X, Tan B, Zhu F, Gui S, et al., QSOX1 regulates trophoblastic apoptosis in preeclampsia through hydrogen peroxide

production. J Matern Fetal Neonatal Med, 2019. **32**(22): p. 3708-3715. doi: 10.1080/14767058.2018.1471459
47. Tan KH, Tan SS, Sze SK, Lee WK, Ng MJ, and Lim SK, Plasma biomarker discovery in preeclampsia using a novel differential isolation technology for circulating extracellular vesicles. Am J Obstet Gynecol, 2014. **211**(4): p. 380 e1-13. doi: 10.1016/j.ajog.2014.03.038
48. Brubaker DB, Ross MG, and Marinoff D, The function of elevated plasma fibronectin in preeclampsia. Am J Obstet Gynecol, 1992. **166**(2): p. 526-31. doi: 10.1016/0002-9378(92)91663-u
49. Zhao M, Li L, Yang X, Cui J, and Li H, FN1, FOS, and ITGA5 induce preeclampsia: Abnormal expression and methylation. Hypertens Pregnancy, 2017. **36**(4): p. 302-309. doi: 10.1080/10641955.2017.1385795
50. Auer J, Camoin L, Guillonneau F, Rigourd V, Chelbi ST, Leduc M, Laparre J, Mignot TM, and Vaiman D, Serum profile in preeclampsia and intra-uterine growth restriction revealed by iTRAQ technology. J Proteomics, 2010. **73**(5): p. 1004-17. doi: 10.1016/j.jprot.2009.12.014
51. Kim SY, Ryu HM, Yang JH, Kim MY, Ahn HK, Lim HJ, Shin JS, Woo HJ, Park SY, Kim YM, et al., Maternal serum levels of VCAM-1, ICAM-1 and E-selectin in preeclampsia. J Korean Med Sci, 2004. **19**(5): p. 688-92. doi: 10.3346/jkms.2004.19.5.688
52. Grewal S, Carver JG, Ridley AJ, and Mardon HJ, Implantation of the human embryo requires Rac1-dependent endometrial stromal cell migration. Proc Natl Acad Sci U S A, 2008. **105**(42): p. 16189-94. doi: 10.1073/pnas.0806219105
53. Fan M, Xu Y, Hong F, Gao X, Xin G, Hong H, Dong L, and Zhao X, Rac1/beta-Catenin Signalling Pathway Contributes to Trophoblast Cell Invasion by Targeting Snail and MMP9. Cell Physiol Biochem, 2016. **38**(4): p. 1319-32. doi: 10.1159/000443076
54. Sado T, Naruse K, Noguchi T, Haruta S, Yoshida S, Tanase Y, Kitanaka T, Oi H, and Kobayashi H, Inflammatory pattern recognition receptors and their ligands: factors contributing to the pathogenesis of preeclampsia. Inflamm Res, 2011. **60**(6): p. 509-20. doi: 10.1007/s00011-011-0319-4
55. Zhang L, Cao Z, Feng F, Xu YN, Li L, and Gao H, A maternal GOT1 novel variant associated with early-onset severe preeclampsia identified by whole-exome sequencing. BMC Med Genet, 2020. **21**(1): p. 49. doi: 10.1186/s12881-020-0989-2
56. Hansen AT, Bernth Jensen JM, Hvas AM, and Christiansen M, The genetic component of preeclampsia: A whole-exome sequencing study. PLoS One, 2018. **13**(5): p. e0197217. doi: 10.1371/journal.pone.0197217
57. Melton PE, Johnson MP, Gokhale-Agashe D, Rea AJ, Ariff A, Cadby G, Peralta JM, McNab TJ, Allcock RJ, Abraham LJ, et al., Whole-exome sequencing in multiplex preeclampsia families identifies novel candidate susceptibility genes. J Hypertens, 2019. **37**(5): p. 997-1011. doi: 10.1097/HJH.0000000000002023
58. Kaartokallio T, Wang J, Heinonen S, Kajantie E, Kivinen K, Pouta A, Gerdhem P, Jiao H, Kere J, and Laivuori H, Exome sequencing in pooled DNA samples to identify maternal pre-eclampsia risk variants. Sci Rep, 2016. **6**: p. 29085. doi: 10.1038/srep29085
59. Gammill HS, Chettier R, Brewer A, Roberts JM, Shree R, Tsigas E, and Ward K, Cardiomyopathy and Preeclampsia. Circulation, 2018. **138**(21): p. 2359-2366. doi: 10.1161/CIRCULATIONAHA.117.031527

60. Glotov AS, Kazakov SV, Vashukova ES, Pakin VS, Danilova MM, Nasykhova YA, Masharsky AE, Mozgovaya EV, Eremeeva DR, Zainullina MS, et al., Targeted sequencing analysis of ACVR2A gene identifies novel risk variants associated with preeclampsia. J Matern Fetal Neonatal Med, 2019. **32**(17): p. 2790-2796. doi: 10.1080/14767058.2018.1449204
61. Soellner L, Kopp KM, Mutze S, Meyer R, Begemann M, Rudnik S, Rath W, Eggermann T, and Zerres K, NLRP genes and their role in preeclampsia and multi-locus imprinting disorders. J Perinat Med, 2018. **46**(2): p. 169-173. doi: 10.1515/jpm-2016-0405
62. Emmery J, Hachmon R, Pyo CW, Nelson WC, Geraghty DE, Andersen AM, Melbye M, and Hviid TV, Maternal and fetal human leukocyte antigen class Ia and II alleles in severe preeclampsia and eclampsia. Genes Immun, 2016. **17**(4): p. 251-60. doi: 10.1038/gene.2016.20
63. Johnson MP, Brennecke SP, East CE, Goring HH, Kent JW, Jr., Dyer TD, Said JM, Roten LT, Iversen AC, Abraham LJ, et al., Genome-wide association scan identifies a risk locus for preeclampsia on 2q14, near the inhibin, beta B gene. PLoS One, 2012. **7**(3): p. e33666. doi: 10.1371/journal.pone.0033666
64. Thomsen LC, McCarthy NS, Melton PE, Cadby G, Austgulen R, Nygard OK, Johnson MP, Brennecke S, Moses EK, Bjorge L, et al., The antihypertensive MTHFR gene polymorphism rs17367504-G is a possible novel protective locus for preeclampsia. J Hypertens, 2017. **35**(1): p. 132-139. doi: 10.1097/HJH.0000000000001131
65. Berryman K, Buhimschi CS, Zhao G, Axe M, Locke M, and Buhimschi IA, Proteasome Levels and Activity in Pregnancies Complicated by Severe Preeclampsia and Hemolysis, Elevated Liver Enzymes, and Thrombocytopenia (HELLP) Syndrome. Hypertension, 2019. **73**(6): p. 1308-1318. doi: 10.1161/HYPERTENSIONAHA.118.12437
66. Fredrickson EK and Gardner RG, Selective destruction of abnormal proteins by ubiquitin-mediated protein quality control degradation. Semin Cell Dev Biol, 2012. **23**(5): p. 530-7. doi: 10.1016/j.semcdb.2011.12.006
67. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, and Barabasi AL, Disease networks. Uncovering disease-disease relationships through the incomplete interactome. Science, 2015. **347**(6224): p. 1257601. doi: 10.1126/science.1257601
68. Fabbri C, Kasper S, Kautzky A, Zohar J, Souery D, Montgomery S, Albani D, Forloni G, Ferentinos P, Rujescu D, et al., A polygenic predictor of treatment-resistant depression using whole exome sequencing and genome-wide genotyping. Transl Psychiatry, 2020. **10**(1): p. 50. doi: 10.1038/s41398-020-0738-5

Figure Legends

Figure 1. Figure shows the univariate work flow for analysis of the whole exome sequencing results.

Figure 2. A) Dendrogram shows statistically significant ($p < 0.05$) clusters of patients. Case dominated cluster (Cluster A) and control dominated cluster (Cluster B) are presented by dashed lines. Cases are represented in red and controls are represented in blue color. **B**) Layered network graphs for the case dominated cluster A are presented. Unique genes of cluster A are in red color. **C)** Layered network graphs for the control dominated cluster B are presented. Unique genes of cluster B are in blue color.

Supplemental Table 1. Unique genes from control dominated cluster (Cluster B). *Genes alphabetically ordered.

Supplemental Table 2. Significantly enriched biological processes, molecular functions and cellular components based on Bonferroni corrected p-value for case and control dominated networks.

Table 1. Clinical characteristics of patients. Mean + SD.

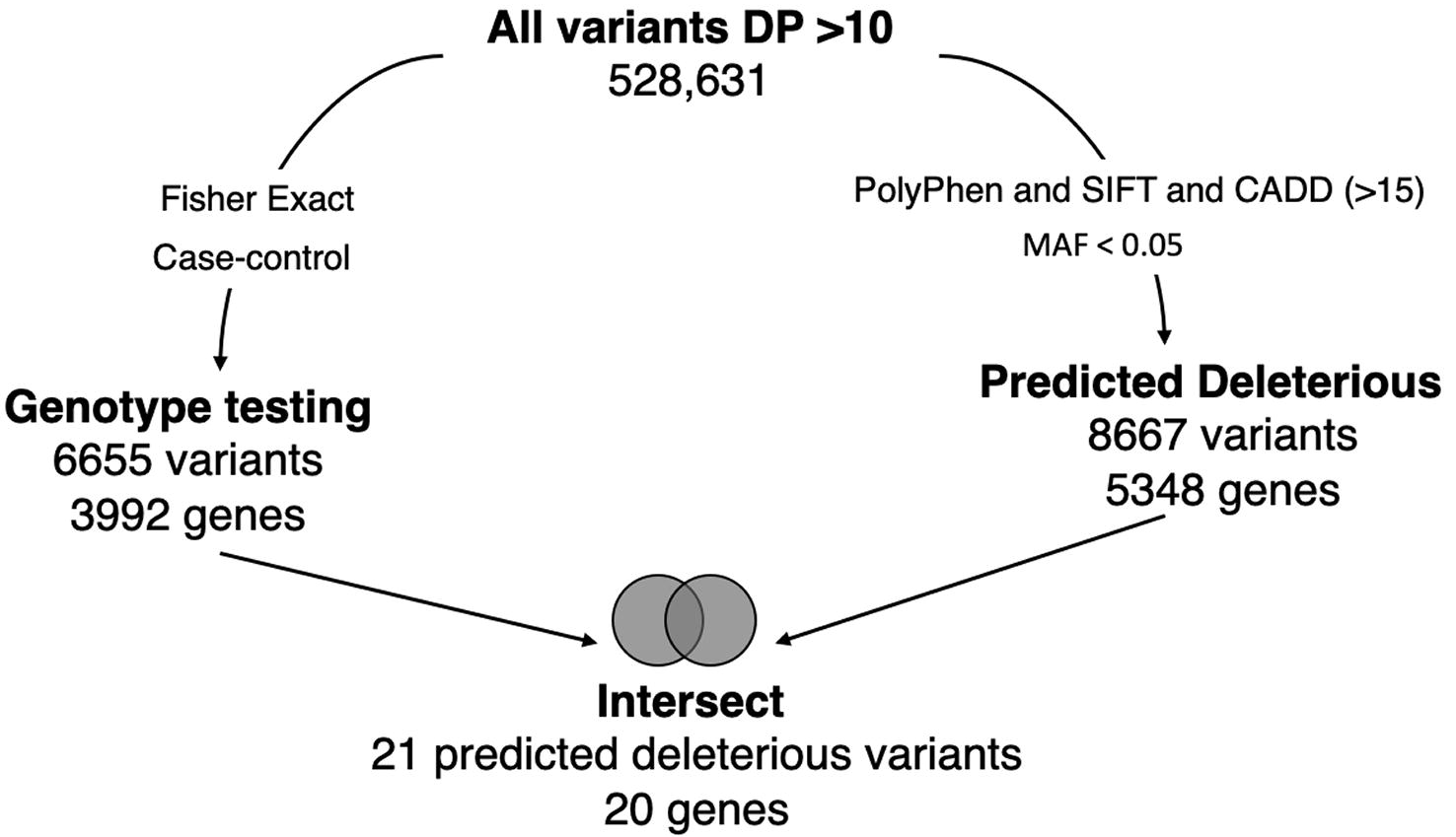
Categories	case (n=61)	control (n=82)
Gestational age of delivery and life style		
Age (mean)	29.1 ± 5.0	29.4 ± 5.3
Grava (mean)	2.1 ± 1.2	2.5 ± 1.6
Job strenuous (%)	26.2%	28.0%
Obesity (%)	31.1%	23.1%
Race/Ethnicity		
African_American (%)	9.8%	4.8%
Asian (%)	3.2%	3.6%
Caucasian (%)	55.7%	56.1%
Hispanic (%)	22.9%	28.0%
Native_American (%)	1.6%	1.2%
Other_Racial_ID (%)	6.5%	6.1%
Abnormal laboratory values		
Systolic_bp (mean, mmHg)	170.8 ± 14.4	117.6 ± 9.6
Proteinuria (%)	65.5%	0.00%
Impaired_liver_function (%)	55.7%	2.4%
Thrombocytopenia (%)	14.7%	0.0%
Cerebral_visual_symptoms (%)	55.7%	0.0%
Fetal Gorth Restriction (FGR) (%)	29.5%	2.4%
Preterm delivery		
Preterm_delivery_before_34_weeks_for_sPEC (%)	55.7%	0.0%
Preterm_delivery_before_37_weeks (%)	60.6%	3.6%

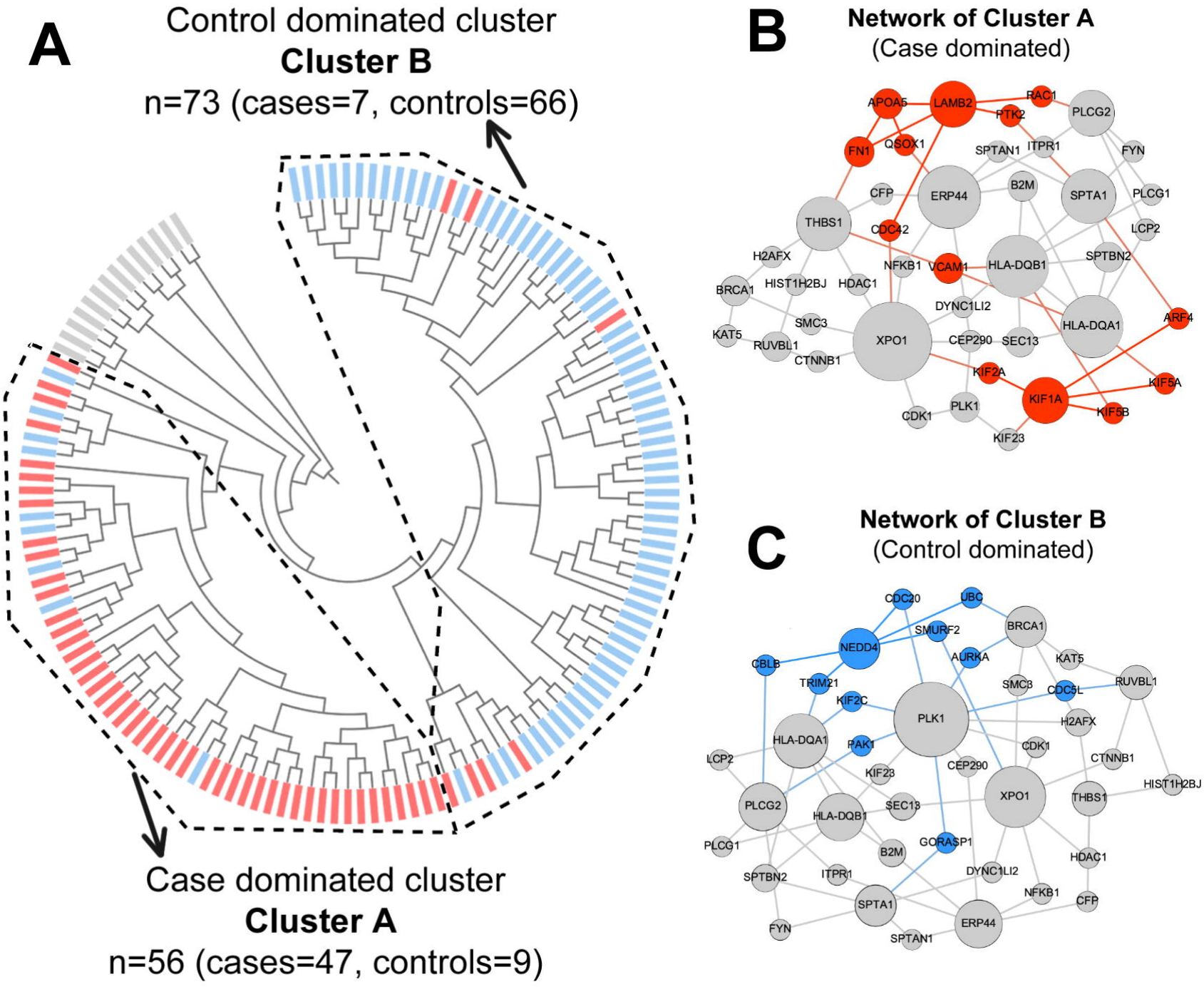
Table 2. Pathogenic, nominally significant (based on genotype testing, p <0.05) gene variants identified by univariate analysis. Genomic positions are based on Human Feb. 2009 (GRCh37/hg19) Assembly.

Chr	Pos	Gene	HGNC ID	SNP	p value	Polyphen 2_HDIV	SIFT	CADD
1	97770920	DPYD	3012	rs1801160	0.032	0.998	0	23.5
1	104117921	AMY2B	478	rs140978983	0.035	1	0	26.1
1	109446750	GPSM2	29501	rs61754640	0.022	0.994	0.02	19.3
1	226125385	LEFTY2	3122	rs2295418	0.022	1	0	16.6
2	69177269	GKN2	24588	rs62133344	0.036	1	0	18.5
2	70504399	PCYOX1	20588	rs34041544	0.030	1	0.01	26.4
2	179486345	TTN	12403	rs114331773	0.017	1	0	15.7
2	179666982	TTN	12403	rs35683768	0.022	0.999	0	15.7
6	76024704	FILIP1	21015	rs62415695	0.009	1	0.01	15.4
6	84904604	CEP162	21107	rs17790493	0.024	1	0	15.9
7	103130222	RELN	9957	rs73714410	0.034	0.972	0.02	27.9
12	124221796	ATP6V0A2	18481	rs74922060	0.010	1	0.03	23.0
13	113750905	MCF2L	14576	rs140657264	0.024	0.999	0	26.6
16	29825022	PRRT2	30500	rs76335820	0.043	0.995	0.02	18.4
17	34311387	CCL14	10612	rs16971802	0.047	0.974	0.02	16.2
17	37321347	ARL5C	31111	rs9912267	0.028	1	0	18.6
18	28604374	DSC3	3037	rs35630063	0.021	1	0	21.1
19	56249615	NLRP9	22941	rs80009430	0.012	1	0	16.0
20	3641868	GFRA4	13821	rs146579049	0.017	1	0	18.3
20	36954724	BPI	1095	rs5743523	0.008	0.998	0.02	15.5
22	31494813	SMTN	11126	rs80055673	0.011	1	0.03	18.7

Table 3. Unique genes from case dominated cluster (Cluster A). *Genes alphabetically ordered.

Gene Name	Gene*	HGNC id	Cluster	Imputed
Apolipoprotein A5	APOA5	17288	A	No
ADP ribosylation factor 4	ARF4	655	A	Yes
Cell division cycle 42	CDC42	1736	A	Yes
Fibronectin 1	FN1	3778	A	Yes
Kinesin family member 1A	KIF1A	888	A	No
Kinesin family member 2A	KIF2A	6318	A	Yes
Kinesin family member 5A	KIF5A	6323	A	Yes
Kinesin family member 5B	KIF5B	6324	A	Yes
Laminin subunit beta 2	LAMB2	6487	A	No
Protein tyrosine kinase 2	PTK2	9611	A	Yes
Quiescin sulfhydryl oxidase 1	QSOX1	9756	A	Yes
Rac family small gtpase 1	RAC1	9801	A	Yes
Vascular cell adhesion molecule 1	VCAM1	12663	A	Yes


All variants DP >10
528,631
Fisher Exact
Case-control
PolyPhen and SIFT and CADD (>15)
MAF < 0.05
Genotype testing
6655 variants
3992 genes
Predicted Deleterious
8667 variants
5348 genes
Intersect
21 predicted deleterious variants
20 genes


A
Control dominated cluster
Cluster B
n=73 (cases=7, controls=66)
Case dominated cluster
Cluster A
n=56 (cases=47, controls=9)
B
Network of Cluster A
(Case dominated)
APOA5
LAMB2
RAC1
PTK2
PLCG2
FN1
QSOX1
SPTAN1
ITPR1
FYN
CFP
ERP44
B2M
SPTA1
PLCG1
THBS1
CDC42
LCP2
H2AFX
NFKB1
VCAM1
HLA-DQB1
SPTBN2
BRCA1
HIST1H2BJ
HDAC1
DYNC1LI2
ARF4
KAT5
SMC3
XPO1
HLA-DQA1
RUVBL1
CEP290
SEC13
CTNNB1
KIF2A
KIF5A
KIF1A
CDK1
PLK1
KIF5B
KIF23
C
Network of Cluster B
(Control dominated)
CDC20
UBC
SMURF2
BRCA1
NEDD4
AURKA
KAT5
CBLB
TRIM21
SMC3
CDC5L
RUVBL1
KIF2C
PLK1
H2AFX
HLA-DQA1
PAK1
CDK1
LCP2
CTNNB1
CEP290
KIF23
HIST1H2BJ
XPO1
THBS1
PLCG2
SEC13
HLA-DQB1
PLCG1
GORASP1
B2M
HDAC1
SPTBN2
ITPR1
DYNC1LI2
NFKB1
CFP
SPTA1
FYN
ERP44
SPTAN1