Integrated Protein Network Analysis of Whole Exome Sequencing Study 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.

Funding: NIH-NIGMS, 5P20GM109035-05

* Corresponding author E-mail: alper_uzun@brown.edu

Abstract:

To identify clusters of patients with shared networks of genes associated with early onset severe preeclampsia from whole exome sequencing data through novel bioinformatic analysis.

We performed a case-control study using whole exome sequencing (WES) on early onset preeclamptic mothers with severe features delivering < 34 weeks and mothers who delivered ≥ 37 weeks. Genotype testing identified variants that were differentially abundant between cases and controls. A Protein-Protein interaction (PPI) analysis and visualization tool, *Proteinarium,* was implemented to identify clusters of patients with shared networks associated with severe preeclampsia.

A total of 61 early onset preeclamptic women with severe features and 82 race and ethnicity matched control women at term were sequenced. We identified 8,867 predicted deleterious variants. 21 of these variants were nominally associated with preeclampsia by genotype testing. Using Proteinarium129 out of the 143 sequenced patients were assigned to statistically significant clusters, Cluster A and B ($p < 0.0001$). Case dominated Cluster A contained 47 of the 61 case subjects. There were 13 unique genes in the PPI network of Cluster A compared to control dominated Cluster B. Amongst these unique genes, LAMB2, PTK2, RAC1, QSOX1, FN1, and VCAM1 have known associations with the pathogenic mechanisms of preeclampsia.

Our network analysis identified genes that were unique to a large cluster of patients with shared networks that provide insights for severe preeclampsia. We also identified genes imputed from the interactome that may otherwise have not been identified by conventional analysis. Strict phenotyping of both cases and controls improved the

likelihood of identifying these otherwise difficult to find genetic associations. These uniquely identified genes and their associated variants are potential candidates for developing polygenic risk scores for severe preeclampsia. These results support our hypothesis on the genetic architecture of complex diseases and are generalizable to other phenotypes.

Introduction

Complex diseases are the result of the interaction of variants in multiple genes with environmental and lifestyle factors. Genetic risk for most complex diseases involves the interaction of multiple genes in discrete networks and pathways [1]. Although complex diseases show increased recurrence risk in families, they do not follow a simple Mendelian pattern of inheritance [2]. Rather, the likely genetic architecture of complex diseases is that subgroups of patients share variants in genes in specific networks and pathways sufficient to express a shared phenotype. It is also probable that alterations in genes in different networks define different clusters of patients with a similar phenotype. Completion of the Human Genome Project has enhanced our approach to complex diseases. Genomic technologies, genome-wide association studies, high-throughput sequencing and bioinformatics methods have revealed insights to the pathogenesis of complex diseases. This includes enhanced understanding of Alzheimer's disease, autism, asthma, Parkinson's disease, multiple sclerosis, and ovarian cancer. For example, computational methods have been used to analyze the network of genes that are linked to autism and also used to find biological subnetworks due to the genetic heterogeneity of the disease [3]. In ovarian cancer, overlapping differentially expressed genes have been identified in different groups of drug-resistant ovarian cancer cells and bioinformatics methods have been applied to identify hub genes to determine potentially effective treatment options [4].

Preeclampsia is a life-threatening, multi-system hypertensive disorder of pregnancy, which complicates 5 to 7% of US deliveries [5, 6]. It is recognized as a leading cause of maternal and fetal morbidity and mortality worldwide [6]. Preeclampsia is characterized by varying degrees of maternal symptoms including elevated blood pressure, proteinuria and

fetal growth retardation [7]. It is a complex disease associated with several different phenotypes. Many clinicians agree that preeclampsia and severe preeclampsia, or early and late preeclampsia are different disorders [8-10]. Previously, using bioinformatics methods, we showed that there are discrete gene sets associated with severe preeclampsia [11]. To date however, other than these observations, there is limited evidence demonstrating whether severe preeclampsia and other hypertensive disorders of pregnancy represent different genetic etiologies.

The evidence that preeclampsia originates in part from genetic causes is based on family and epidemiological studies [12, 13]. Preeclampsia has contributions from the maternal, paternal and fetal genome [14-17]. The classical approach to genetics is twin studies where it has been shown that the heritability of preeclampsia is up to 52% [5, 12, 18]. The recurrence risk for preeclampsia in the daughters of either eclamptic or preeclamptic patients is 20-40% [19, 20]. A significant role for genetics in the development of preeclampsia is also supported by family based studies [5, 21]. More than 100 family studies in different populations have reported a 2- to 5-fold increased risk of preeclampsia among family members of affected women [22-27].

We sought to test our hypothesis on the genetic basis for preeclampsia using whole exome sequencing in carefully selected patients with severe preeclampsia. In the present study we compared variants identified by whole-exome sequencing in early onset preeclamptic mothers with severe features and term controls without personal or family history of pregnancy related hypertensive disorders. We then used *Proteinarium,* a multi-sample, protein-protein interaction analysis (PPI) tool, to identify clusters of patients with shared PPI networks associated with preeclampsia.

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 [28]. We retrieved clinical data from all enrolled subjects from their electronic medical records.

This case/control study was approved by the Institutional Review Board of WIH (Project ID: WIH 16-0031). 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 for all patients. In addition, employment history and strenuous effort on the job were recorded. We excluded mothers with personal or family history of other hypertensive disorders. In our control cohort, we enrolled 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 for whole exome sequencing, 61 early onset preeclamptic women with severe features, and 82 race and ethnicity matched control women at term.

Whole Exome Sequencing

Residual maternal whole blood was obtained from each subject for extraction of genomic DNA. EDTA stabilized whole blood was stored continuously at -80°C until processing. These residual whole blood samples were sent to Beijing Genome Institute (BGI) at the Children's Hospital of Philadelphia for whole exome sequencing. The sequencing facility was blinded to the preeclampsia disease status. The library was sequenced on an Illumina HiSeq 4000 using 150 bp paired-end protocols. QC sequence data was recorded.

Sequence Data

For variant discovery we used the Gene Analysis Tool Kit (GATK) version 4 to analyze the sequence reads [29]. Haplotype caller was applied for variant detection [30]. 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). A variant identified by any one of these filters was labeled "low quality" and not considered for further analysis.

Genotype Testing

In order 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 alleles, the homozygosis alternative alleles, and the heterozygous genotypes between cases and controls.

Variant Annotation

We applied a strict filter-based annotation using ANNOVAR. We identified deleterious variants with Polyphen 2 HDIV, SIFT and CADD [31-34]. 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 [34].

Network Analysis

We hypothesize 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. For that purpose, we analyzed our whole exome sequencing data using *Proteinarium,* a multisample PPI analysis and visualization tool [35]. The top 60 genes, corresponding to the most significant, differentially abundant variants between cases and controls for each patient (ranked by genotype testing p value) were used as the seed genes for input into *Proteinarium*. *Proteinarium* was implemented with the minimum path length parameter set to 2, in order to include only those pathways in which seed proteins are connected directly to each other and/or via a single intermediary protein. We refer to these intermediary connecting proteins as imputed proteins. *Proteinarium* clusters the subjects based on similarities of their PPI networks using the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) algorithm [36], outputting a dendrogram visualization of the clustering. Statistical significance for each branch under the dendrogram is calculated by Fisher exact test comparing the abundance of cases and controls in each cluster relative to the total number of samples and their group assignment.

Network Separation Testing

Measuring the genetic similarity between diseases by comparing their protein-protein interaction networks from the interactome is possible. There are computational approaches like separation-based methods which compare the shortest distances between network proteins *within* each disease or network to the shortest distances *between* the disease networks [37]. A positive separation score indicates that there is a physical separation between networks within the interactome. The greater the score, the more dissimilar the networks. We computed the separation score between the networks of significant clusters identified by Proteinarium [35].

Comparative Phenotypic Analysis

We performed a series of comparative phenotypic analyses between the significant clusters of patients identified by Proteinarium using univariate and multivariate approaches. Before performing phenotypic analyses comparing the clinical characteristics of the clustered patients, we removed all the control patients from the case dominated cluster, and all the case patients from the control dominated cluster. We performed univariate and multivariate analysis of the clinical characteristics of the remaining patients in each of the significant clusters using the R stats v3.4.3 package.

Results

Clinical Characteristics:

The clinical characteristics of the patients with early onset preeclampsia with severe features and term control patients are shown in Table 1. The distribution of

race/ethnicity is also shown in Table 1. In order to leverage the likelihood of genetic discovery, both the cases and controls were carefully phenotyped with respect to early onset preeclampsia with severe features, clinical characteristics and family history. As can be seen from the table, gestational age at delivery, the 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. There were no other significant univariate or multivariate phenotypic differences.

Sequence Results:

High quality sequence data with a Phred score $\geq$30 from well-balanced pools with over 19,000,000 reads/patient were observed. The sequence showed high quality 40X average depth of coverage with more than 80% at least 20X coverage. We identified a total of 528,630 variants. There were 187,915 exonic variants. The work flow for the univariate analysis is shown in Figure 1. After application of the initial filters for coverage (DP > 10) and variant pathogenicity (SIFT and PolyPhen 2 HDIV and CADD), there were a total of 8,867 predicted deleterious variants (available at www.dbpec.brown.edu). Amongst these, 21 variants were also 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.

In order identify clusters of patients with shared networks associated with severe preeclampsia, the top 60 genes based on the most significant variants (ranked by genotype p value) for each patient were used as the seed genes for input into *Proteinarium*. The resulting dendrogram is shown in Figure 2. Out of the 143 patients sequenced, 129 of these patients

(90 % of the mothers) were assigned to statistically significant clusters. The two significant clusters have been highlighted in red and blue on the dendrogram ($p < 0.0001$). The inset in Figure 2 shows the number of cases and controls that are in each cluster. Cluster A had significantly more cases than controls, containing 47 of the total 61 case subjects. The layered network for the case-dominated Cluster A is shown in Figure 3. There are 13 genes which are unique to Cluster A and they are 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 total 82 control subjects. The layered network for the control-dominated Cluster B is shown in Figure 3. The unique genes from the layered network graph of Cluster B, shown in blue, are listed in Supplemental Table 1.

We used separation testing to compare the case and control dominated networks identified by *Proteinarium [37]*. The comparison of the unique genes from case dominated cluster and the control dominated cluster 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 case dominated Cluster A and control dominated Cluster B, Table 4 [38, 39]. We found significantly enriched biological processes, molecular functions and cellular components based on Bonferroni corrected p-value for case and control dominated networks. Prominent among the biological processes and molecular functions were antigen processing and presentation, cellular movement (axon guidance and microtubules) and T cell receptor signaling.

We previously published the Database for Preeclampsia (dbPEC) [11]. dbPEC consists of the genes associated with preeclampsia, the clinical features, and concurrent conditions.

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

Discussion

We identified clusters of patients with shared protein-protein interaction networks associated with early onset severe preeclampsia. These results provide insights into the genetics of severe preeclampsia and support our hypothesis that the genetic architecture of complex diseases is characterized by clusters of patients that have variants in shared gene networks. To generate these results, first we performed whole exome sequencing. For our case cohort, we enrolled women with idiopathic early onset preeclampsia with severe features and singleton births <34 weeks' gestation. We compared them to term controls with no family history of preeclampsia. Then, we used *Proteinarium,* a multi-sample, PPI analysis and visualization tool, to identify clusters of patients with shared protein-protein interaction networks [35]. Using seed genes from each patient, *Proteinarium* mapped the input genes onto the PPI interactome based on STRING database to build individual networks. The similarities between all subjects' PPI networks were based on distance metrics and were then used for clustering samples. We identified a single, significant cluster with a predominance of patients with early onset, severe features of

preeclampsia encompassing 47 out of the 61 women. We also identified a single control-dominated cluster with 66 out of 82 control patients.

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. We reviewed the association of the unique genes from the case dominated network with preeclampsia and note that several of these genes 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 [40]. In addition, it was shown previously that serum laminin levels in preeclamptic patients are significantly higher than those in normal pregnancy [41]. It has been shown that 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 [42]. QSOX1 protein is also found in circulating extracellular vesicles of both preeclampsia and healthy pregnant women [43]. It has been reported that Fibronectin 1 (FN1) might promote the development of preeclampsia by modulating differentiation of human extravillous trophoblasts, as well as formation of focal adhesions [44-46]. Vascular Cell Adhesion Molecule 1 (VCAM1) is involved in cellular adhesion. It has been reported that serum concentrations of sVCAM-1 are significantly higher in both mild and severe preeclampsia than in normal pregnancy [47]. Although Thrombospondin 1 (THBS1) is not a unique gene in the case dominated network, it is notable that THBS1, which is also associated with focal adhesions, is an intermediate protein in this network where it has edges to both VCAM1 and FN1 [48]. Invasion of maternal decidua and uterine spiral arteries by extravillous trophoblasts (EVT) is required for establishment of

normal placenta and adequate blood supply toward the fetus. Human trophoblast migration requires Rac Family Small GTPase 1 (RAC1) and Cell Division Cycle 42 (CDC42) [49]. Lower levels were found in preeclampsia samples than in normal term pregnancy samples, and these levels significantly declined in severe preeclampsia samples compared with mild preeclampsia samples [50]. Protein tyrosine kinase 2 (PTK2) which is also called focal adhesion kinase is differentially expressed in preeclampsia. Investigations were carried out to evaluate the features of inflammatory response, placental dysfunction and PTK2 was reported as among the promising biomarkers for preeclampsia [51]. In the case-dominated subnetwork we observed Kinesin Family Member 2A (KIF2A) which has been reported in the literature to be upregulated in the preeclamptic placenta [17]. Up-regulated genes in the preeclampsia placenta haven been shown to be associated with the regulation of diverse cellular processes, including matrix degradation, trophoblast cell invasion, migration and proliferation [17].

Although our study included only a modest sample size, we identified a significant subgroup of patients with shared PPI networks associated with severe preeclampsia. We were not expecting each patient to appear in a significant cluster. However, we believe that our careful phenotyping resulted in the high percentage of patients being successfully assigned to significant clusters (77% of the severe preeclamptic mothers). We also believe that careful phenotyping allowed us to observe distinct separation between case and control dominated clusters in the dendrogram. We believe that the whole exome sequencing, combined with this novel multi sample network analysis, combined with very carefully chosen phenotype of preeclampsia patients contributed to our discovery despite the relatively modest size.

There have been several sequencing efforts including whole genome, whole exome and targeted sequencing on an array of preeclampsia phenotypes from diverse populations [52-61]. There is no consensus amongst the published results in regards to associated genes and variants. Since preeclampsia is a polygenic disease, lack of a consensus among these studies focused solely on univariate etiologies might be expected in these early stage studies. Nonetheless, we compared our significant gene list from the univariate analysis to the results of these previously published studies. Among the 20 genes identified in our univariate analysis only Titin (TTN) was identified in two prior studies [52, 56]. We also compared the genes from our network based approach to the results of these other studies. 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 [53, 59]. None of these overlapping genes were among the unique genes identified in the shared layered networks.

Our protein-protein interaction analysis allowed us to identify clusters of patients with shared PPI networks associated with preeclampsia. Within the significant clusters there were unique genes that were imputed during network analysis (RAC1, KIF5B, PTK2, KIF5A, FN1, QSOX1, ARF4, VCAM1, CDC42, KIF2A). In other words, they were not amongst the top 60 seed genes selected by differential abundance between cases and controls. Nonetheless, this 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.

While our primary aim was to identify associations in the case dominated cluster, we also examined the network of control dominated cluster where we were able to identify

additional unique proteins. Proteins in this network were associated with the ubiquitination process. They may be thought to serve as protective proteins that confer resilience against preeclampsia [62, 63]. Although there are studies showing the relationship with hypertension - ubiquitination process or pregnancy, this still needs further investigation [63].

Conclusion

The results of our study provide several promises of future use to further understanding the mechanism underlying preeclampsia. Identified genes and their associated variants, particularly the unique genes in the case dominant cluster, are candidates for generating polygenic risk scores for severe preeclampsia. 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. Strict phenotyping of both cases and controls improved the likelihood of identifying these otherwise difficult to find genetic associations.

Acknowledgements

This work was supported by a grant from National Institutes of Health grant *5P20GM109035-05* and the Kilguss Research Core at Women & Infants Hospital.

References

1. Loscalzo, J., I. Kohane, and A.L. Barabasi, *Human disease classification in the postgenomic era: a complex systems approach to human pathobiology.* Mol Syst Biol, 2007. **3**: p. 124.
2. Smith, G.D. and S. Ebrahim, *'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?* Int J Epidemiol, 2003. **32**(1): p. 1-22.
3. Wall, D.P., F.J. Esteban, T.F. Deluca, M. Huyck, T. Monaghan, N. Velez de Mendizabal, et al., *Comparative analysis of neurological disorders focuses genome-wide search for autism genes.* Genomics, 2009. **93**(2): p. 120-9.
4. Yuan, D., H. Zhou, H. Sun, R. Tian, M. Xia, L. Sun, et al., *Identification of key genes for guiding chemotherapeutic management in ovarian cancer using translational bioinformatics.* Oncol Lett, 2020. **20**(2): p. 1345-1359.
5. Chappell, S. and L. Morgan, *Searching for genetic clues to the causes of pre-eclampsia.* Clin Sci (Lond), 2006. **110**(4): p. 443-58.
6. Valenzuela, F.J., A. Perez-Sepulveda, M.J. Torres, P. Correa, G.M. Repetto, and S.E. Illanes, *Pathogenesis of preeclampsia: the genetic component.* J Pregnancy, 2012. **2012**: p. 632732.
7. Jebbink, J., A. Wolters, F. Fernando, G. Afink, J. van der Post, and C. Ris-Stalpers, *Molecular genetics of preeclampsia and HELLP syndrome - a review.* Biochim Biophys Acta, 2012. **1822**(12): p. 1960-9.
8. Carreiras, M., S. Montagnani, and Z. Layrisse, *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.

9. Raymond, D. and E. Peterson, *A critical review of early-onset and late-onset preeclampsia.* Obstet Gynecol Surv, 2011. **66**(8): p. 497-506.

10. American College of, O., Gynecologists, and P. Task Force on Hypertension in, *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.

11. Triche, E.W., A. Uzun, A.T. DeWan, I. Kurihara, J. Liu, R. Occhiogrosso, et al., *Bioinformatic approach to the genetics of preeclampsia.* Obstet Gynecol, 2014. **123**(6): p. 1155-61.

12. Madejczyk M, K.G., BRĘBOROWICZ GH, *Etiopathology of preeclampsia.* ARCHIVES of PERINATAL MEDICINE, 2009. **15**(3): p. 8.

13. Nilsson, E., H. Salonen Ros, S. Cnattingius, and P. Lichtenstein, *The importance of genetic and environmental effects for pre-eclampsia and gestational hypertension: a family study.* BJOG, 2004. **111**(3): p. 200-6.

14. Cnattingius, S., M. Reilly, Y. Pawitan, and P. Lichtenstein, *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.

15. Zusterzeel, P.L., R. te Morsche, M.T. Raijmakers, E.M. Roes, W.H. Peters, and E.A. Steegers, *Paternal contribution to the risk for pre-eclampsia.* J Med Genet, 2002. **39**(1): p. 44-5.

16. Than, N.G., R. Romero, A.L. Tarca, K.A. Kekesi, Y. Xu, Z. Xu, et al., *Integrated Systems Biology Approach Identifies Novel Maternal and Placental Pathways of Preeclampsia.* Front Immunol, 2018. **9**: p. 1661.

17. Kobayashi, H., *The Impact of Maternal-Fetal Genetic Conflict Situations on the Pathogenesis of Preeclampsia.* Biochem Genet, 2015. **53**(9-10): p. 223-34.

18. Salonen Ros, H., P. Lichtenstein, L. Lipworth, and S. Cnattingius, *Genetic effects on the liability of developing pre-eclampsia and gestational hypertension.* Am J Med Genet, 2000. **91**(4): p. 256-60.

19. Genc, M.R. and J. Schantz-Dunn, *The role of gene-environment interaction in predicting adverse pregnancy outcome.* Best Pract Res Clin Obstet Gynaecol, 2007. **21**(3): p. 491-504.

20. Serrano, N.C., *Immunology and genetic of preeclampsia.* Clin Dev Immunol, 2006. **13**(2-4): p. 197-201.

21. Nejatizadeh, A., T. Stobdan, N. Malhotra, and M.A. Pasha, *The genetic aspects of pre-eclampsia: achievements and limitations.* Biochem Genet, 2008. **46**(7-8): p. 451-79.

22. Arngrimsson, R., S. Bjornsson, R.T. Geirsson, H. Bjornsson, J.J. Walker, and G. Snaedal, *Genetic and familial predisposition to eclampsia and pre-eclampsia in a defined population.* Br J Obstet Gynaecol, 1990. **97**(9): p. 762-9.

23. Chesley, L.C., J.E. Annitto, and R.A. Cosgrove, *The familial factor in toxemia of pregnancy.* Obstet Gynecol, 1968. **32**(3): p. 303-11.

24. Cincotta, R.B. and S.P. Brennecke, *Family history of pre-eclampsia as a predictor for pre-eclampsia in primigravidas.* Int J Gynaecol Obstet, 1998. **60**(1): p. 23-7.

25. Mutze, S., S. Rudnik-Schoneborn, K. Zerres, and W. Rath, *Genes and the preeclampsia syndrome.* J Perinat Med, 2008. **36**(1): p. 38-58.

26. Sutherland, A., D.W. Cooper, P.W. Howie, W.A. Liston, and I. MacGillivray, *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.

27. Ward, K., *Genetic factors in common obstetric disorders.* Clin Obstet Gynecol, 2008. **51**(1): p. 74-83.

28. *American College of Obstetricians and Gynecologists (ACOG)*. Available at: www.acog.org. Retrieved October 8, 2020.

29. Van der Auwera, G.A., M.O. Carneiro, C. Hartl, R. Poplin, G. Del Angel, A. Levy-Moonshine, 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.

30. Poplin, R., V. Ruano-Rubio, M.A. DePristo, T.J. Fennell, M.O. Carneiro, G.A. Van der Auwera, et al., *Scaling accurate genetic variant discovery to tens of thousands of samples.* bioRxiv, 2018: p. 201178.

31. Adzhubei, I.A., S. Schmidt, L. Peshkin, V.E. Ramensky, A. Gerasimova, P. Bork, et al., *A method and server for predicting damaging missense mutations.* Nat Methods, 2010. **7**(4): p. 248-9.

32. Ng, P.C. and S. Henikoff, *SIFT: Predicting amino acid changes that affect protein function.* Nucleic Acids Res, 2003. **31**(13): p. 3812-4.

33. Kircher, M., D.M. Witten, P. Jain, B.J. O'Roak, G.M. Cooper, and J. Shendure, *A general framework for estimating the relative pathogenicity of human genetic variants.* Nat Genet, 2014. **46**(3): p. 310-5.

34. Genomes Project, C., A. Auton, L.D. Brooks, R.M. Durbin, E.P. Garrison, H.M. Kang, et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.

35. Armanious, D., J. Schuster, G.A. Tollefson, A. Agudelo, A.T. DeWan, S. Istrail, et al., *Proteinarium: Multi-sample protein-protein interaction analysis and visualization tool.* Genomics, 2020.

36. Michener, C.D., Sokal, R.R, *A quantitative approach to a problem of classification.* Evolution, 1957. **11**: p. 490–499.

37. Menche, J., A. Sharma, M. Kitsak, S.D. Ghiassian, M. Vidal, J. Loscalzo, et al., *Disease networks. Uncovering disease-disease relationships through the incomplete interactome.* Science, 2015. **347**(6224): p. 1257601.

38. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nat Protoc, 2009. **4**(1): p. 44-57.

39. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic Acids Res, 2009. **37**(1): p. 1-13.

40. Zhang, A. and S. Huang, *Progress in pathogenesis of proteinuria.* Int J Nephrol, 2012. **2012**: p. 314251.

41. Furuhashi, N., H. Kimura, H. Nagae, A. Yajima, C. Kimura, and T. Saito, *Serum laminin levels in normal pregnancy and preeclampsia.* Gynecol Obstet Invest, 1993. **36**(3): p. 172-5.

42. Li, J., C. Tong, P. Xu, L. Wang, T.L. Han, L. Wen, et al., *QSOX1 regulates trophoblastic apoptosis in preeclampsia through hydrogen peroxide production.* J Matern Fetal Neonatal Med, 2019. **32**(22): p. 3708-3715.

43. Tan, K.H., S.S. Tan, S.K. Sze, W.K. Lee, M.J. Ng, and S.K. Lim, *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.

44. Brubaker, D.B., M.G. Ross, and D. Marinoff, *The function of elevated plasma fibronectin in preeclampsia.* Am J Obstet Gynecol, 1992. **166**(2): p. 526-31.

45. Zhao, M., L. Li, X. Yang, J. Cui, and H. Li, *FN1, FOS, and ITGA5 induce preeclampsia: Abnormal expression and methylation.* Hypertens Pregnancy, 2017. **36**(4): p. 302-309.

46. Auer, J., L. Camoin, F. Guillonneau, V. Rigourd, S.T. Chelbi, M. Leduc, et al., *Serum profile in preeclampsia and intra-uterine growth restriction revealed by iTRAQ technology.* J Proteomics, 2010. **73**(5): p. 1004-17.

47. Kim, S.Y., H.M. Ryu, J.H. Yang, M.Y. Kim, H.K. Ahn, H.J. Lim, 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.

48. Murphy-Ullrich, J.E. and M. Hook, *Thrombospondin modulates focal adhesions in endothelial cells.* J Cell Biol, 1989. **109**(3): p. 1309-19.

49. Grewal, S., J.G. Carver, A.J. Ridley, and H.J. Mardon, *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.

50. Fan, M., Y. Xu, F. Hong, X. Gao, G. Xin, H. Hong, et al., *Rac1/beta-Catenin Signalling Pathway Contributes to Trophoblast Cell Invasion by Targeting Snail and MMP9.* Cell Physiol Biochem, 2016. **38**(4): p. 1319-32.

51. Sado, T., K. Naruse, T. Noguchi, S. Haruta, S. Yoshida, Y. Tanase, et al., *Inflammatory pattern recognition receptors and their ligands: factors contributing to the pathogenesis of preeclampsia.* Inflamm Res, 2011. **60**(6): p. 509-20.

52. Zhang, L., Z. Cao, F. Feng, Y.N. Xu, L. Li, and H. Gao, *A maternal GOT1 novel variant associated with early-onset severe preeclampsia identified by whole-exome sequencing.* BMC Med Genet, 2020. **21**(1): p. 49.

53. Hansen, A.T., J.M. Bernth Jensen, A.M. Hvas, and M. Christiansen, *The genetic component of preeclampsia: A whole-exome sequencing study.* PLoS One, 2018. **13**(5): p. e0197217.

54. Melton, P.E., M.P. Johnson, D. Gokhale-Agashe, A.J. Rea, A. Ariff, G. Cadby, et al., *Whole-exome sequencing in multiplex preeclampsia families identifies novel candidate susceptibility genes.* J Hypertens, 2019. **37**(5): p. 997-1011.

55. Kaartokallio, T., J. Wang, S. Heinonen, E. Kajantie, K. Kivinen, A. Pouta, et al., *Exome sequencing in pooled DNA samples to identify maternal pre-eclampsia risk variants.* Sci Rep, 2016. **6**: p. 29085.

56. Gammill, H.S., R. Chettier, A. Brewer, J.M. Roberts, R. Shree, E. Tsigas, et al., *Cardiomyopathy and Preeclampsia.* Circulation, 2018. **138**(21): p. 2359-2366.

57. Glotov, A.S., S.V. Kazakov, E.S. Vashukova, V.S. Pakin, M.M. Danilova, Y.A. Nasykhova, 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.

58. Soellner, L., K.M. Kopp, S. Mutze, R. Meyer, M. Begemann, S. Rudnik, et al., *NLRP genes and their role in preeclampsia and multi-locus imprinting disorders.* J Perinat Med, 2018. **46**(2): p. 169-173.

59. Emmery, J., R. Hachmon, C.W. Pyo, W.C. Nelson, D.E. Geraghty, A.M. Andersen, et al., *Maternal and fetal human leukocyte antigen class Ia and II alleles in severe preeclampsia and eclampsia.* Genes Immun, 2016. **17**(4): p. 251-60.

60. Johnson, M.P., S.P. Brennecke, C.E. East, H.H. Goring, J.W. Kent, Jr., T.D. Dyer, 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.

61. Thomsen, L.C., N.S. McCarthy, P.E. Melton, G. Cadby, R. Austgulen, O.K. Nygard, 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.

62. Berryman, K., C.S. Buhimschi, G. Zhao, M. Axe, M. Locke, and I.A. Buhimschi, *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.

63. Fredrickson, E.K. and R.G. Gardner, *Selective destruction of abnormal proteins by ubiquitin-mediated protein quality control degradation.* Semin Cell Dev Biol, 2012. **23**(5): p. 530-7.

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

Categories	**case (n=61)**	**control (n=82)**	**p-value**
Age (mean)	29.14 ± 5.05	29.45 ± 5.34	7.294E-01
Grava (mean)	2.19 ± 1.27	2.54 ± 1.62	1.177E-01
Job_strenuous (%)	26.23%	28.05%	8.102E-01
Obesity (%)	31.15%	23.17%	2.958E-01
African_American (%)	9.84%	4.88%	2.761E-01
Asian (%)	3.28%	3.66%	9.028E-01
Caucasian (%)	55.74%	56.10%	9.661E-01
Hispanic (%)	22.95%	28.05%	4.906E-01
Native_American (%)	1.64%	1.22%	3.213E-01
Other_Racial_ID (%)	6.56%	6.10%	9.121E-01
Systolic_bp (mean)	170.81 ± 14.47	117.67 ± 9.63	3.502E-44
Proteinuria (%)	65.57%	0.00%	1.606E-15
Impaired_liver_function (%)	55.74%	2.47%	1.825E-11
Thrombocytopenia (%)	14.75%	0.00%	2.055E-03
Cerebral_visual_symptoms (%)	55.74%	0.00%	3.263E-12
FGR (%)	29.51%	2.44%	3.590E-05
Preterm_delivery_before_34_weeks_for_sPEC (%)	55.74%	0.00%	3.263E-12
Preterm_delivery_before_37_weeks (%)	60.66%	3.66%	2.856E-13

Table 2. Nominally significant genes from univariate analysis. Genomic positions are based on Human Feb. 2009 (GRCh37/hg19) Assembly.

Chr	Pos	Gene	HGNC ID	SNP	Polyphen2_HDIV	SIFT	CADD_phred
1	97770920	DPYD	3012	rs1801160	0.998	0	23.5
1	104117921	AMY2B	478	rs140978983	1	0	26.1
1	109446750	GPSM2	29501	rs61754640	0.994	0.02	19.3
1	226125385	LEFTY2	3122	rs2295418	1	0	16.6
2	69177269	GKN2	24588	rs62133344	1	0	18.5
2	70504399	PCYOX1	20588	rs34041544	1	0.01	26.4
2	179486345	TTN	12403	rs114331773	1	0	15.7
2	179666982	TTN	12403	rs35683768	0.999	0	15.7
6	76024704	FILIP1	21015	rs62415695	1	0.01	15.4
6	84904604	CEP162	21107	rs17790493	1	0	15.9
7	103130222	RELN	9957	rs73714410	0.972	0.02	27.9
12	124221796	ATP6V0A2	18481	rs74922060	1	0.03	23.0
13	113750905	MCF2L	14576	rs140657264	0.999	0	26.6
16	29825022	PRRT2	30500	rs76335820	0.995	0.02	18.4
17	34311387	CCL14	10612	rs16971802	0.974	0.02	16.2
17	37321347	ARL5C	31111	rs9912267	1	0	18.6
18	28604374	DSC3	3037	rs35630063	1	0	21.1
19	56249615	NLRP9	22941	rs80009430	1	0	16.0
20	3641868	GFRA4	13821	rs146579049	1	0	18.3
20	36954724	BPI	1095	rs5743523	0.998	0.02	15.5
22	31494813	SMTN	11126	rs80055673	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

Table 4. Significantly enriched biological processes, molecular functions and cellular components based on Bonferroni corrected p-value for case and control dominated networks. GO terms that are associated with case dominated networks were represented with "Cluster A" and with control dominated networks were represented with "Cluster B".

GO term ID	Definition	p value
Case dominated cluster (A)		
Biological Processes		
GO:0019886	Antigen processing and presentation of exogenous peptide antigen via MHC class II	2.09E-06
GO:0007411	Axon guidance	3.75E-06
GO:0007018	Microtubule-based movement	1.31E-03
GO:0050852	T cell receptor signaling pathway	1.32E-03
GO:0038096	Fc-gamma receptor signaling pathway involved in phagocytosis	1.19E-02
GO:0031295	T cell costimulation	3.20E-02
Molecular Function		
GO:0003777	Microtubule motor activity	6.23E-06
GO:0008017	Microtubule binding	2.36E-02
GO:0019899	Enzyme binding	2.36E-02
GO:0005088	Ras guanyl-nucleotide exchange factor activity	2.84E-02
Cellular Components		
GO:0005829	Cytosol	2.18E-05
GO:0016020	Membrane	6.02E-04
GO:0005871	Kinesin complex	1.33E-03
GO:0008091	Spectrin	3.66E-02
GO:0012507	ER to Golgi transport vesicle membrane	4.59E-02
Control dominated cluster (B)		
Biological Processes		
GO:0050852	T cell receptor signaling pathway	1.29E-06
GO:0019886	Antigen processing and presentation of exogenous peptide antigen via MHC class II	5.36E-05
GO:0042787	Protein ubiquitination involved in ubiquitin-dependent protein catabolic process	1.09E-03

GO:0007062	Sister chromatid cohesion	3.08E-03
GO:0007067	Mitotic nuclear division	1.73E-02
GO:0031145	Anaphase-promoting complex-dependent catabolic process	2.54E-02
Molecular Function		
GO:0005515	Protein binding	6.13E-04
Cellular Components		
GO:0005829	Cytosol	1.59E-08
GO:0005654	Nucleoplasm	2.84E-04
GO:0005813	Centrosome	6.55E-03
GO:0008091	Spectrin	3.12E-02
GO:0012507	ER to Golgi transport vesicle membrane	3.73E-02
GO:0005634	Nucleus	4.27E-02
GO:0043234	Protein complex	4.75E-02

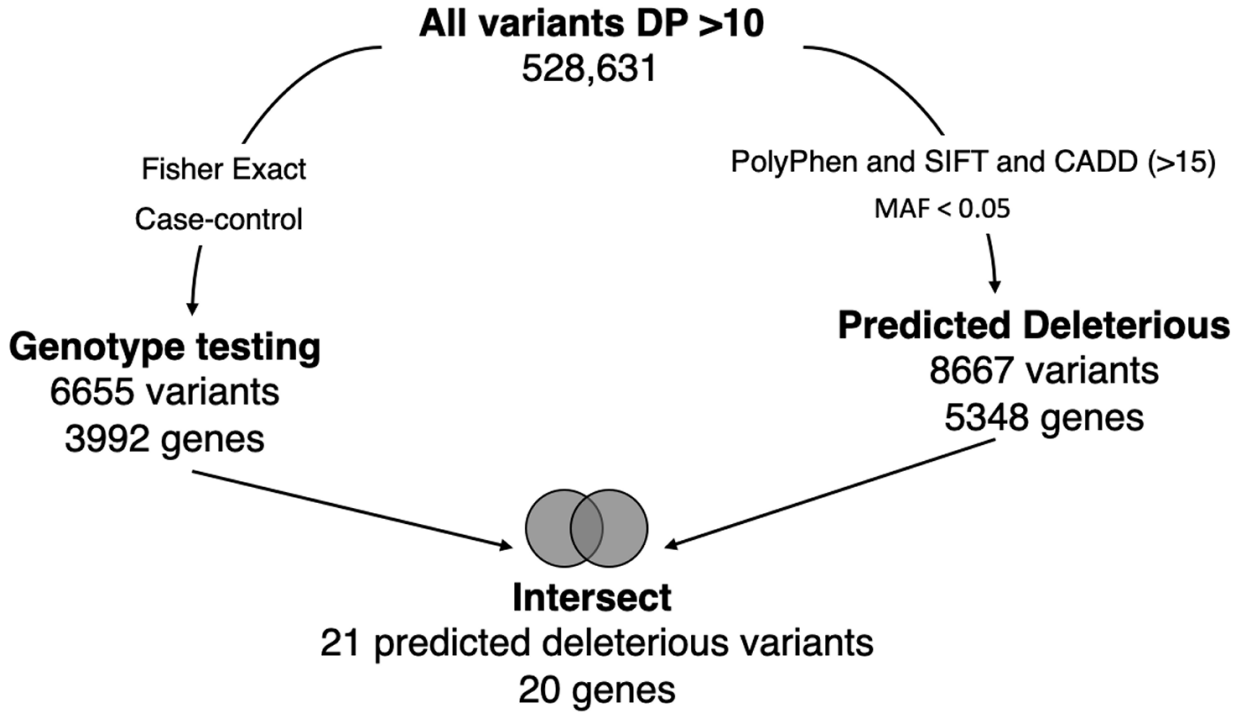

Figure 1. Figure shows the univariate work flow for whole exome sequencing.

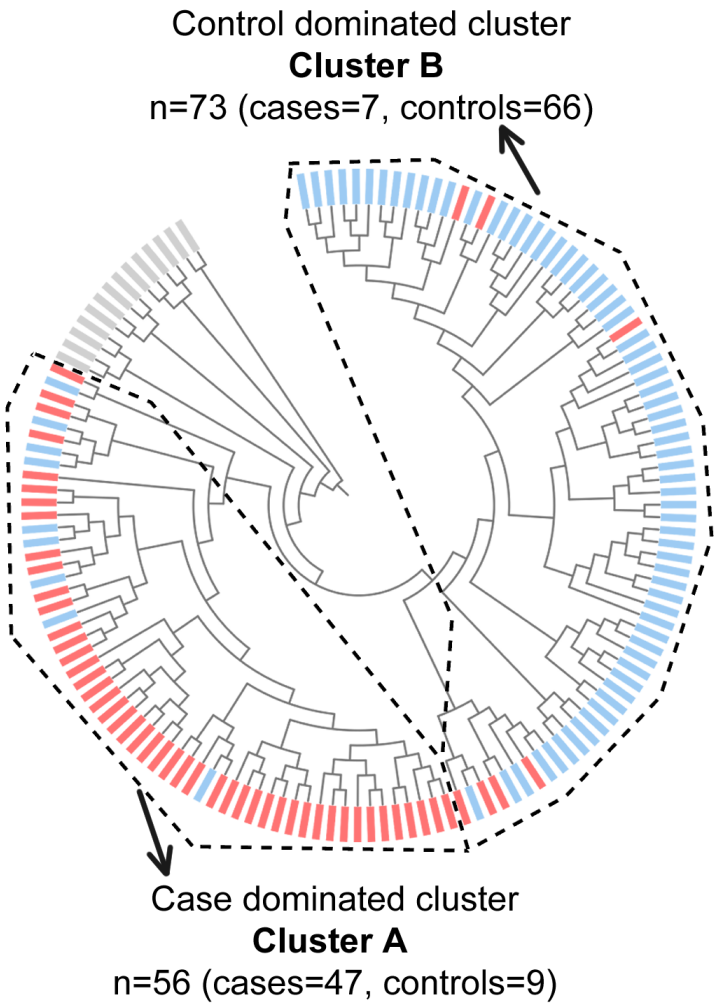

Figure 2. Dendrogram is showing significant clusters of patients (colored). Case dominated cluster (Cluster A) and control dominated cluster (Cluster B) is presented in dashed lines. Cases are represented in red and controls were represented in blue color.

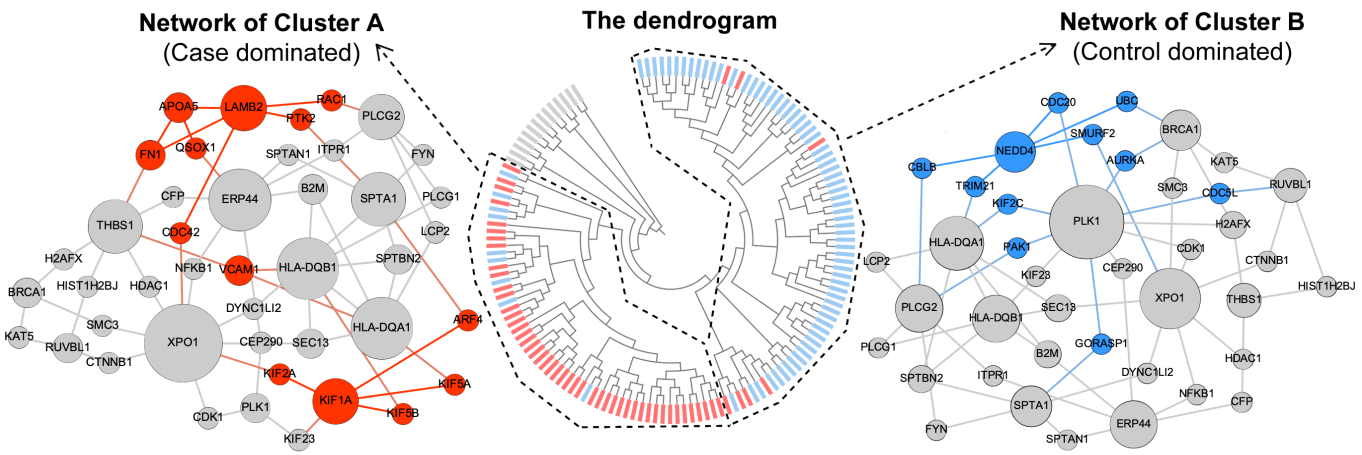

Figure 3. Layered network graphs for the case dominated cluster A and control dominated cluster B represented. The unique genes associated with each cluster are highlighted in their related cluster colors. Unique genes of cluster A are in red and unique genes of cluster B are in blue color.

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

Gene Name	Gene*	HGNC id	Cluster	Imputed
Aurora kinase A	AURKA	11393	B	Yes
Cbl proto-oncogene B	CBLB	1542	B	Yes
Cell division cycle 20	CDC20	1723	B	Yes
Cell division cycle 5 like	CDC5L	1743	B	Yes
Golgi reassembly stacking protein 1	GORASP1	16769	B	Yes
Kinesin family member 2C	KIF2C	6393	B	Yes
NEDD4 E3 ubiquitin protein ligase	NEDD4	7727	B	No
P21 (RAC1) activated kinase 1	PAK1	8590	B	Yes
SMAD specific E3 ubiquitin protein ligase 2	SMURF2	16809	B	Yes
Tripartite motif containing 21	TRIM21	11312	B	Yes
Ubiquitin C	UBC	12468	B	Yes