scientific reports

Check for updates

OPEN

# Protein interaction networks define the genetic architecture of preterm birth

Alper Uzun[1,2,4], Jessica S. Schuster[1,4], Joan Stabila[1], Valeria Zarate[1], George A. Tollefson[1], Anthony Agudelo[1], Prachi Kothiyal[3], Wendy S. W. Wong[3] & James Padbury[1,2]✉

**The likely genetic architecture of complex diseases is that subgroups of patients share variants in genes in specific networks sufficient to express a shared phenotype. We combined high throughput sequencing with advanced bioinformatic approaches to identify such subgroups of patients with variants in shared networks. We performed targeted sequencing of patients with 2 or 3 generations of preterm birth on genes, gene sets and haplotype blocks that were highly associated with preterm birth. We analyzed the data using a multi-sample, protein–protein interaction (PPI) tool to identify significant clusters of patients associated with preterm birth. We identified shared protein interaction networks among preterm cases in two statistically significant clusters, $p < 0.001$. We also found two small control-dominated clusters. We replicated these data on an independent, large birth cohort. Separation testing showed significant similarity scores between the clusters from the two independent cohorts of patients. Canonical pathway analysis of the unique genes defining these clusters demonstrated enrichment in inflammatory signaling pathways, the glucocorticoid receptor, the insulin receptor, EGF and B-cell signaling, These results support a genetic architecture defined by subgroups of patients that *share* variants in genes in specific networks and pathways which are sufficient to give rise to the disease phenotype.**

Genome-wide association studies (GWAS) are a contemporary approach to the investigation of complex diseases that have made possible discovery of insights not previously recognized[1–3]. However, GWAS have failed to demonstrate the "missing heritability" in many common diseases[4–8]. The computational approaches underlying GWAS reflect the "common disease-common variant hypothesis," that complex disease architecture is due to additive genetic effects of variants in individual genes. However, the genetics of complex diseases suggests that is unlikely. The more likely genetic architecture is that subgroups of patients share variants in genes in specific networks and pathways which are sufficient to give rise to a shared phenotype. It is also likely that variants in genes in different networks and pathways express similar phenotypes and define different subgroups of patients.

Preterm birth is an important, complex genetic disorder affecting up to 10% of pregnant women[11]. We built the Database for Preterm Birth, containing a validated collection of genes with an a priori connection to preterm birth[9]. This was the result of a semantic data mining and curation of published literature and publicly available, high-throughput databases. We used this resource to analyze a large genome wide association study to identify the biological networks and pathways associated with preterm birth[10]. In this report, we present the results of targeted sequence of the genes, flanking sites and haplotype blocks identified by gene set enrichment of that GWAS[10]. Further, in order to leverage the likelihood of genetic discovery, we exploited an "extreme phenotype" of preterm birth by concentrating our enrollment on patients with a family history of preterm birth. We compared variants identified in women with 2–3 generations of preterm birth with term controls without history of preterm birth. We then used *Proteinarium,* a multi-sample, protein–protein interaction analysis (PPI) tool, to identify clusters of patients with shared PPI networks associated with preterm birth[11].

[1]Pediatrics, Women & Infants Hospital, Providence, RI, USA. [2]Center for Computational Molecular Biology, Brown Medical School, Brown University, Providence, RI, USA. [3]INOVA Translational Medicine Institute, INOVA Health System, Falls Church, VA, USA. [4]These authors contributed equally: Alper Uzun and Jessica S. Schuster. ✉email: jpadbury9@gmail.com

## Material and methods

**Patient identification and enrollment.** Large epidemiological studies drawn from population based analyses support a predominantly maternal origin for the genetic contribution(s) to risk of preterm birth, with little contribution by paternal or fetal genetic factors[12–14]. We therefore concentrated our efforts on identification of maternal genetic variants. Women & Infants Hospital of Rhode Island is the sole provider of high-risk perinatal services in Rhode Island, northeastern Connecticut and southeastern Massachusetts. We used this *population-based* service to enroll patients with a prior history of preterm birth. The study was approved by the Institutional Review Board of the Women & Infants Hospital of RI, 08-0117. All methods were performed in accordance with the relevant guidelines and regulations. An informatically driven retrieval from our electronic medical record gave us a daily report on all preterm births. A clinical research assistant, formally trained in genetic interviews, reviewed the records of all patients delivering ≥ 24 weeks and < 34 weeks. Following informed consent, women underwent an interview focused on family history of preterm birth. We asked explicit questions about preterm birth in mother, grandmother, her first order relatives and also paternal relatives. Careful clinical history with an emphasis on additional risk factors for prematurity including medical illnesses, drug use, psychiatric disorders and employment history was recorded on all patients. We excluded patients delivered prematurely for considerations related to preeclampsia, drug use, diabetes or multiple gestation. Controls were patients who delivered ≥ 37 weeks gestation in whom the same, formal genetic history revealed no history of preterm birth on either maternal or paternal side of the pedigree. All of the patients' identifying data was coded and redacted for the purposes of data analysis. 190 patients were enrolled for targeted sequencing. Samples were taken from 122 women with multiple generations of preterm birth, and 68 race, ethnicity matched control women at term. Residual maternal whole blood was obtained for extraction of genomic DNA. The samples were stored continuously at − 80 °C until processing.

**Sample preparation.** We targeted the 329 genes and 132 haplotype blocks for sequencing that are highly associated with preterm birth[10]. Genomic DNA from maternal whole blood was extracted using QIAamp DSP DNA blood mini kit from Qiagen following the manufacture's protocol. Samples were quantified using Qubit technology (Life Technologies, Carlsbad, CA, USA) and sequencing libraries were constructed from 2 µg each of case/control DNA. Library preparation was performed using Illumina TruSeq DNA LT Sample prep Kit (Illumina, San Diego, CA, USA), with enzymatic fragmentation using dsDNA Fragmentase (NEB), followed by indexing and clean-up. DNA capture was performed using custom capture probes from SeqCap EZ choice kit (Roche NimbleGen).

**Targeted sequencing.** The library was sequenced on an Illumina HiSeq 2500 using 100 bp paired-end protocols. Following sequencing, the multiplex indices were used to bin the samples for each patient and QC sequence data was recorded. High quality sequence data from well-balanced pools was observed. There was an average of 25 million reads from each patient, with an average Q30 of 91%. Reads were then mapped to the to the human reference sequence (Hg19) with BWA[15], sorted and indexed with SAMtools[16].

**Sequence data, variant calling and genotype testing.** Variants were flagged as low quality and filtered using established metrics: if three or more variants were detected within 10 bp; 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. For variant discovery we used the Gene Analysis Tool Kit (GATK) version 3.2 to analyze the sequence reads[17]. Duplicate reads were marked and removed using Picard Tools version 1.77. Haplotype caller was applied for variant detection[18]. Twenty-five base pairs upstream and downstream of each exon were included in the design of capture probes and in variant detection. Variants were annotated using ANNOVAR for pathogenicity prediction scores. We used Eigenstrat to control for population stratification during genotype testing of differential abundance of variants in cases and controls[19]. To investigate the frequency of potentially relevant single variants, we extracted variants with the following filters: coverage ≥ 10×, a Polyphen 2 HDIV prediction if a change is damaging (≥ 0.957), a SIFT score (< 0.05), a CADD score > 10, and minor allele frequency (MAF) < 0.05 from the Exome Aggregation Consortium (ExAC)[20], and significant difference by genotype testing[22].

**Network analysis.** In order to identify patients with shared networks and pathways associated with preterm birth, we used *Proteinarium*[11]. This is a tool for analysis of protein–protein interactions that uses the String interactome database, Dijkstra's algorithm and the Jaccard index to build a network similarity matrix of protein–protein interactions (PPI) between samples[11]. The top 30 genes, based on the most significant variants (ranked by Eigenstrat genotype p value) for each patient, were used as the seed genes for input into *Proteinarium*. We selected *Proteinarium's* user defined output minimum path length of 2, which includes 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. The output of *Proteinarium* is a UPGMA generated dendrogram that shows clusters of patients with shared PPI networks, gene lists forming the networks and the group assignment for each patient. Statistical significance for each branch under the dendrogram is calculated by Fisher exact test comparing the probability of observing a cluster of that size relative to the total number of samples and group assignment.

| | Cases N = 128 (%) | Controls N = 68 (%) | *p*-value |
|---|---|---|---|
| Maternal age | 27 ± 6 | 27 ± 7 | N.S |
| Gravida | 3 ± 2 | 2 ± 1 | N.S |
| Gestational age, week | 31 ± 3 | 40 ± 1 | 0.001 |
| Birth weight, g | 1724 ± 607 | 3451 ± 371 | 0.001 |
| African American | 15 (12) | 8 (13) | N.S |
| Asian | 4 (3) | 3 (5) | N.S |
| Caucasian (non-Hispanic) | 75 (59) | 37 (60) | N.S |
| Hispanic or Latino | 30 (23) | 14 (23) | N.S |
| Native American | 1 (1) | 0 (0) | N.S |
| Other | 3 (2) | 0 (0) | N.S |

**Table 1.** Clinical characteristics of patients. Mean ± SD.

| Number of patients | Case definition |
|---|---|
| 15 | 3 generations of PTB |
| 57 | 2 generations of PTB |
| 12 | Generational skips |
| 6 | Intra-generational PTB |
| 32 | Same mother with multiple PTB |
| 68 | No personal or family history PTB |

**Table 2.** Family history of preterm birth among enrolled patients.

**Network separation testing.** Computational methods have been developed to identify disease-disease similarity by comparison of individual networks from the protein–protein interactome[21]. This network-based approach compares the shortest distances between proteins *within* each disease or network to the shortest distances *between* the disease networks. This approach has been applied to other complex disease phenotypes[22]. We computed the separation between our clusters using the seed genes identified for each of the patients. Using the union of the seed genes for patients within each of the clusters, we identified genes unique to each cluster to use as input for the separation analysis[21]. A negative score indicates an overlap between networks within the interactome. The more negative the score the greater the similarity/overlap between two networks.

**Phenotypic analysis.** In order to identify significant differences in the clinical, phenotypic characteristics between the clusters identified by *Proteinarium,* we used a Bayesian generalized linear model implemented via the Arm package in R and the optimal model was determined using stepwise model selection with the MASS package (https://CRAN.R-project.org/).

## Results

The clinical characteristics of the patients and their distribution by race/ethnicity are shown in Table 1. The only significant difference between the groups was in gestational age and birth weight. As described in the Methods, in order to leverage genetic discovery, the patients were carefully phenotyped with respect to history of preterm birth. The distribution of family history of preterm birth among the enrolled patients is shown in Table 2. Of the enrolled patients, 84 had a multi-generational history of preterm birth, 6 had an intra-generational history of preterm birth, 32 were first generation with multiple preterm births and 68 control patients had no family history of preterm births.

We identified a total of almost 140,000 variants, the bulk of which were in intronic regions captured from the haplotype blocks previously identified[10]. We restricted our subsequent analyses to variants in regions with greater than tenfold coverage which resulted in 39,472 variants. There were also almost 7000 exonic variants and several splice variants. After application of the initial filters for coverage and variant pathogenicity, there were a total of 264 variants (Supplemental Table 1). Of these, there were 9 variants that were nominally associated with preterm birth. All were non-synonymous, exonic variants (Table 3).

A single SNV in the AOAH gene was more abundant in the cases, whereas the remaining eight variants were only present in the controls. Nonetheless, none met significance after correction for multiple comparison testing. None of the splice variants passed genotype testing for differential abundance between preterm cases and controls.

For replication of these univariate data, we compared our results to a cohort of patients recruited at the INOVA Translational Medicine Institute, Falls Church VA [23,24]. They enrolled 816 families who underwent 60X whole genome sequencing. From these families, there were 60 cases and 321 controls that met our strict phenotypic criteria (singleton pregnancy, less than 34 weeks gestation, no history of preeclampsia or drug use).

| Chr | Gene | pos | Exonic function | Polyphen2_HDIV_score | CADD_phred | SIFT_score | ExAC_ALL |
|---|---|---|---|---|---|---|---|
| 1 | *COL16A1* | 32,164,127 | nSNV | 1 | 14 | 0 | 0.002 |
| 7 | *AOAH* | 36,656,035 | n SNV | 1 | 21 | 0.01 | 0.1172 |
| 9 | *QRFP* | 133,769,023 | n SNV | 0.999 | 18 | 0 | 0.0043 |
| 10 | *SORBS1* | 97,144,031 | n SNV | 1 | 17 | 0.01 | 0.0007 |
| 11 | *ATM* | 108,098,576 | n SNV | 0.98 | 18 | 0 | 0.0074 |
| 12 | *TBX5* | 114,837,349 | n SNV | 1 | 23 | 0 | 0.0034 |
| 16 | *CHST4* | 71,571,658 | n SNV | 0.999 | 17 | 0.02 | 0.0033 |
| 19 | *MYH14* | 50,747,534 | n SNV | 1 | 15 | 0 | 0.0029 |
| 20 | *TCFL5* | 61,485,507 | n SNV | 1 | 16 | 0 | 0.0028 |

**Table 3.** Nominally significant genes from univariate analysis. Nonsynonymous SNV (nSNV).

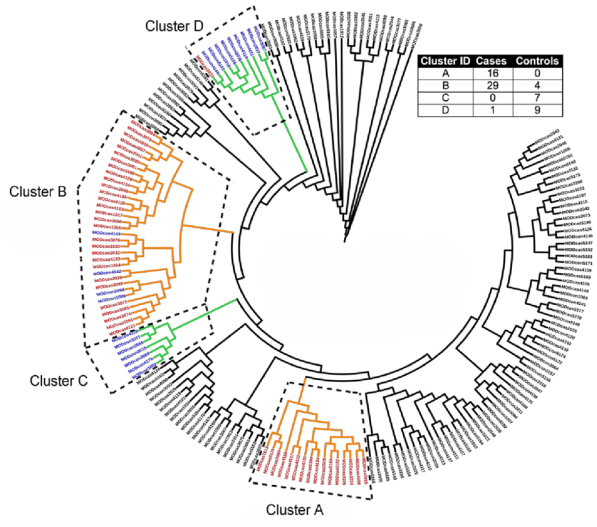


**Figure 1.** Dendrogram showing significant clusters of patients (colored). Inset: distribution of cases and controls in each of the clusters.

INOVA provided the variant data for the genes and haplotype block intervals described above. With similar functional filters, among the 264 variants identified from our cohort, 165 (63%) were also identified in the INOVA sequence data.

Among our cases with a family history of preterm birth, we found that each patient had an average of 163 variants that passed the coverage and Eigenstrat genotype testing above. In order to identify clusters of patients with shared networks associated with preterm birth, the top 30 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 Fig. 1.

For ease of visualization the dendrogram has been circularized and the significant clusters have been highlighted in colors. The inset in Fig. 1 shows the distribution of cases and controls that belonged to each cluster. There were four significant clusters identified at a Fisher exact test with $p < 0.001$. Out of the 190 patients sequenced, a total of 66 subjects were assigned to statistically significant clusters. The two largest significant clusters (A and B) had significantly more cases than controls, encompassing 45 of the 122 cases. There were also two small control-dominated clusters. The layered networks for the case-dominated clusters A and B are shown in Fig. 2. The unique genes associated with these clusters are highlighted in light blue. There were 9 unique genes in cluster A and likewise 6 genes unique to cluster B (Table 4). All of the genes from the layered network graph of the two case dominated clusters and group membership of each gene are shown in Supplemental Table 2.

For replication of this network analysis, a similar approach using the filters described above to identify variants in individual patients was applied to the INOVA cohort. The top 30 seed genes for each subject were used for input into *Proteinarium*. There were four significant case-dominated clusters identified encompassing 40 of the 60 cases. The layered networks for these four clusters are shown in Fig. 3. The unique genes associated with these clusters are highlighted in light blue. The gene lists for the layered networks for the case dominated clusters from the preterm birth cohort and the replications cohort and group membership is shown in Supplemental Table 2.

We used separation testing to compare the networks identified by *Proteinarium*. The two case dominated clusters from our preterm birth cohort (clusters A and B) showed overlap with each other within the interactome.

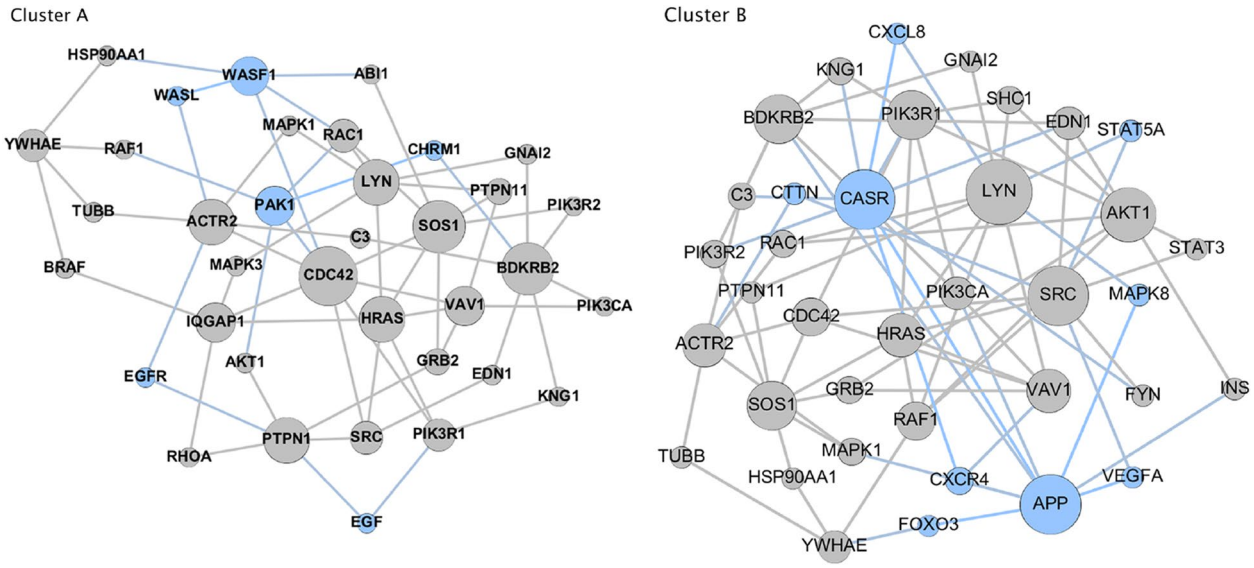


**Figure 2.** Layered network graphs for the case dominated clusters A & B from Fig. 1. The unique genes associated with each cluster are highlighted in light blue.

| Genes* | Cluster |
|---|---|
| *APP* | A |
| *CASR* | A |
| *CTTN* | A |
| *CXCL8 (IL-8)* | A |
| *CXCR4* | A |
| *FOXO3* | A |
| *MAPK8* | A |
| *STAT5A* | A |
| *FOXO3* | A |
| *CHRM1* | B |
| *EGF* | B |
| *EGFR* | B |
| *PAK1* | B |
| *WASF 1* | B |
| *WASL* | B |

**Table 4.** Unique genes from case dominated preterm birth clusters shown in Fig. 2: Cluster A and Cluser *Alphabeticall.

This was reflected by negative separation values (separation = −0.285) when compared to the control dominated clusters that showed higher separation values (ranging from −0.037 to 0.11).

We also compared the separation of the case dominated clusters from our preterm birth cohort to the case dominated clusters in the INOVA replication cohort. The results are shown in Table 5. Case clusters B' and D' from the replication cohort showed the greatest similarity when compared to our case clusters B and A, respectively, indicating significant overlap within the interactome. In order to identify the pathways modulated by the genes defining the networks within significant clusters of cases, we carried out comparative pathway analysis for the genes defining each cluster using Ingenuity Pathway Analysis. For this analysis we included the genes unique to each of the layered network graphs of the two case dominated clusters (clusters A and B) in the preterm birth cohort and the two case dominated clusters in the replication cohort that had the smallest separation with those clusters (clusters B' and D'). To avoid false positive inferences, we used stringent criteria with $p$ values ranging from $10^6$ to $10^8$. The pathway similarity is illustrated in the heatmap shown in Fig. 4. Canonical pathway analysis of the unique genes defining these clusters demonstrated enrichment in numerous signaling pathways. This includes signaling by IL6, IL-7, IL-15, IL-8, JAK/Cytokine signaling, toll-like receptors, the glucocorticoid receptor, the insulin receptor, EGF and B-cell signaling.

We also sought to determine if there were underlying phenotypic differences between the patients in the different clusters we identified using *Proteinarium*. The clinical phenotype of the case patients in Cluster A, were

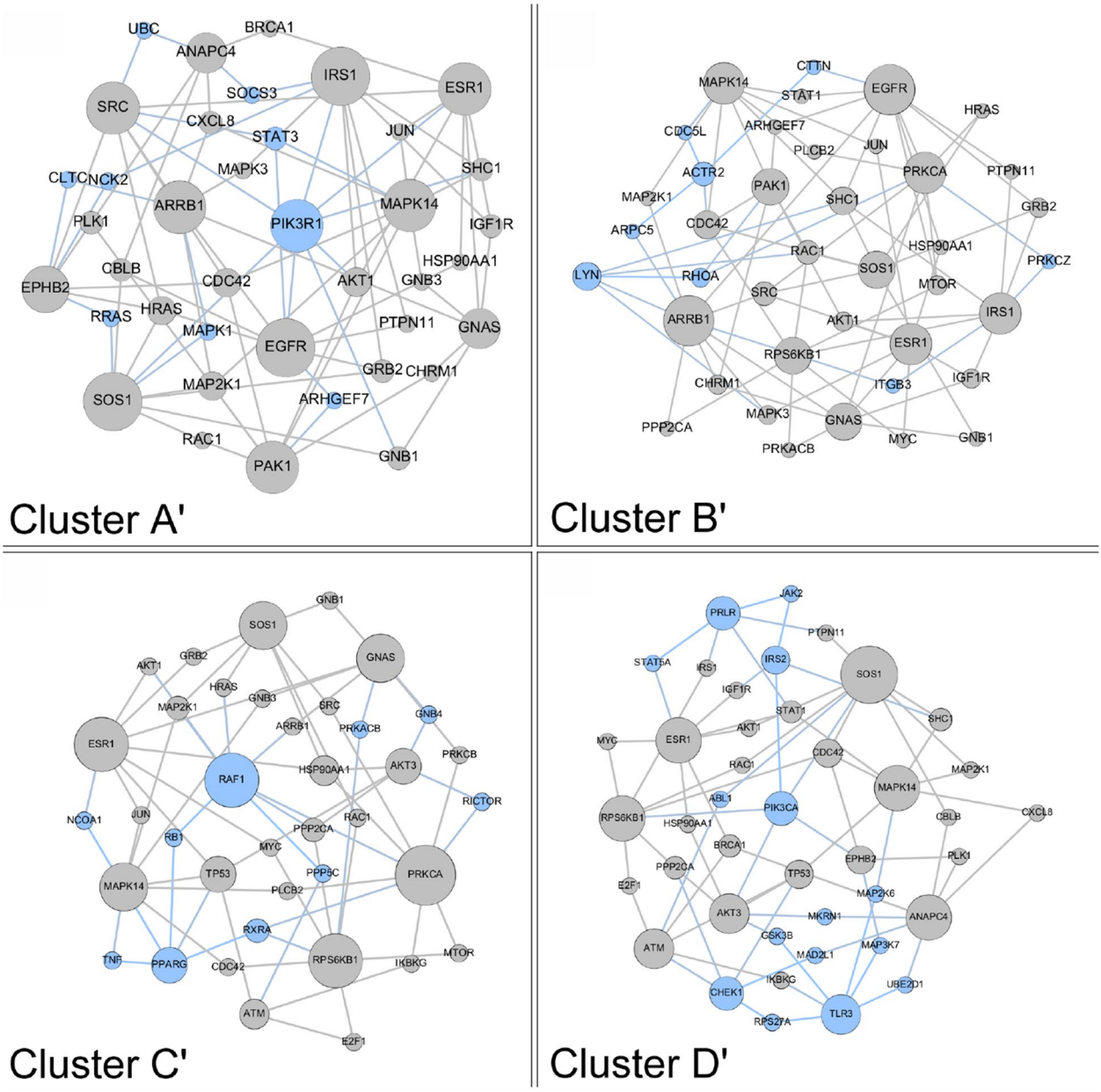


**Figure 3.** Layered network graphs for the INOVA replication cohort showing significant clusters A', B', C', D'. Unique genes to each cluster are shown in light blue.

| | Replication cohort | | | | |
|---|---|---|---|---|---|
| | | A' | B' | C' | D' |
| Preterm birth cohort | A | −0.224 | 0.010 | −0.193 | **−0.330** |
| | B | 0.114 | **−0.491** | −0.123 | −0.239 |

**Table 5.** Separation scores for comparison of case dominated clusters from the preterm birth cohort and the replication cohort. The values in bold are significant.

compared to the characteristics of the remaining patients not included in Cluster A or Cluster B. An analogous analysis was performed for Cluster B. In addition, the characteristics of patients in Clusters A and B were compared to each other. The distribution of clinical characteristics for these comparisons is shown in Table 6. Comparing the cases from Cluster A to the subjects in neither of the two clusters, there was a significant difference in the distribution of maternal racial background and in the generational history of preterm birth ($p < 0.05$). For Cluster B compared to the subjects in neither of the two clusters, there was a significant difference in the


**Figure 4.** Comparative network analysis from Ingenuity Pathway Analysis. Comparison of case dominated clusters A, D,' B, B'. All pathways significant $p < 10^{6}$ to $10^{8}$.

proportion with chorioamnionitis ($p < 0.05$). Comparison of the patients in Cluster A and Cluster B, showed a significant difference in income and the distribution of maternal racial background ($p < 0.05$). Nonetheless, the majority of these differences in clinical characteristics were very modest.

## Discussion

We performed targeted sequencing of gene sets and haplotype blocks that are highly associated with preterm birth on carefully phenotyped patients. We enrolled women with idiopathic, singleton births < 34 weeks gestation, the majority of whom had multiple generations of preterm birth. We compared them to term controls with no family history of preterm birth. We used *Proteinarium,* a multi-sample, protein–protein interaction tool, to identify clusters of patients with shared protein–protein interaction networks associated with preterm birth[11]. Using seed genes from each patient, *Proteinarium* mapped the input genes onto the STRING PPI interactome to build individual networks. The similarities between all subjects' PPI Networks were used as the distance metric for clustering samples. We identified two significant clusters with a predominance of preterm birth patients encompassing 45 out of the 122 women with a multi-generation history of preterm birth. We also found two small control-dominated clusters. For replication, we compared our data to a large birth cohort collected at INOVA Health [24]. Sequence data analyzed from INOVA's 60 cases and 321 controls identified 40 preterm cases in four significant clusters. Separation analysis of the layered PPI networks of the significant clusters from our preterm birth cohort and the replication cohort demonstrated overlap of these clusters within the interactome. Canonical pathway analysis of the unique genes defining these clusters demonstrated enrichment in inflammatory signaling by *IL-6, IL-7, IL-15, IL-8,* JAK/Cytokine signaling, toll-like receptors, the glucocorticoid receptor, the insulin receptor, EGF and B-cell signaling, $p \sim 10^{-6}$ to $10^{-8}$.

The individual, unique genes from the two case-dominated preterm birth clusters are shown in Table 4. Several of the genes unique to Cluster A have been association with inflammation and immune responses. *CXCR4* is among the gene sets and pathways upregulated in whole blood from women with spontaneous preterm birth when compared to patients delivering at term[25]. Moreover, *CXCR4* is located in genomic regions with large ROH that we have shown to be in greater abundance in women delivering preterm than full term[26]. *CXCR4* has also been evolutionarily linked to preterm birth[27]. *CXCL8* (IL-8) is a monocyte macrophage chemoattractant. It has been widely studied in labor and shown to be expressed in multiple gestational tissues including myometrium, cervix and decidua[28]. *CXCL8* is upregulated in chorio-decidua samples collected from preterm labor patients when compared to patients at term and in labor[28]. This is consistent with the increase in inflammatory cells in the decidua during labor. The gene *FOXO3* codes for the forkhead box 03 transcription factor that regulates

| Covariate | Phenotype and clinical characteristics | Cluster A | Cluster B | Remaining patients |
|---|---|---|---|---|
| Generational status* | 3 Generations of PTB | 0.000 | 0.091 | 0.085 |
| | 2 Generations of PTB | 0.375 | 0.485 | 0.246 |
| | Generational skips | 0.000 | 0.061 | 0.070 |
| | Intragenerational | 0.125 | 0.000 | 0.028 |
| | Multiple PTB | 0.500 | 0.242 | 0.113 |
| | No personal or family history of PTB | 0.000 | 0.121 | 0.451 |
| | NA | 0.000 | 0.000 | 0.007 |
| Maternal racial background*,^ | Caucasian | 0.125 | 0.697 | 0.585 |
| | African American | 0.500 | 0.061 | 0.106 |
| | Hispanic or Latino | 0.313 | 0.152 | 0.254 |
| | Asian | 0.000 | 0.091 | 0.035 |
| | Native American | 0.000 | 0.000 | 0.007 |
| | Other | 0.063 | 0.000 | 0.014 |
| | Do not know | 0.000 | 0.000 | 0.000 |
| Income^ | $0–$19,999 | 0.313 | 0.091 | 0.211 |
| | $20,000–$29,999 | 0.000 | 0.212 | 0.092 |
| | $30,000–$49,999 | 0.000 | 0.091 | 0.070 |
| | $50,000+ | 0.000 | 0.364 | 0.254 |
| | Other | 0.688 | 0.242 | 0.373 |
| Previous preterm | No | 0.375 | 0.333 | 0.662 |
| | Yes | 0.625 | 0.667 | 0.338 |
| Chorioamnionitis+ | No | 0.813 | 0.848 | 0.965 |
| | Yes | 0.188 | 0.152 | 0.035 |

**Table 6.** Proportion of patient clinical characteristics within clusters. *$p < 0.05$, Cluster A vs remaining patients. +$p < 0.05$ Cluster B vs remaining patients. ^$p < 0.05$ Cluster A vs Cluster B.

inflammation in non-gestational tissues. It has also been shown to be expressed in human myometrial tissue[29]. Further, higher *FOXO3* gene and protein expression have been demonstrated in myometrium from women in labor compared to non-laboring samples. In isolated cells, *FOXO3* silencing was associated with a significant decrease in IL-1 induced I-L6 and IL-8 expression and cyclooxygenase 2 production. Overexpression of *FOXO3*, increased cytokine expression, prostaglandin production and *MMP9* expression are all observed in myometrial cells following administration of IL-1B. Thus, *FOXO3* may be implicated in the pathways regulating labor and it may be a potential target for prevention of preterm birth[29]. The *STAT5A* gene is present in a network of cell-mediated immune responses, cellular movement and hematologic system development that was identified in a genome-wide association study looking at single nucleotide polymorphisms (SNP) in peripheral blood of patients who are in preterm labor compared to full term labor[30]. Most importantly, serial sampling and transcriptional profiling of circulating immune cells during pregnancy has been carried out to identify patterns of gene signatures across pregnancy that are associated with gestational age and thus define a "gestational clock" [31]. These studies revealed an important role for IL-2 dependent *STAT5A* signaling in modulating T-cell function during pregnancy. Baseline gene expression was compared with expression following activation with receptor-specific ligands including interferon and a cocktail of interleukins. The endogenous *STAT5A* signal in naïve cells and the phospho-*STAT5A* response to interleukins and neutrophils were among the strongest features correlated to immunological adaptations to pregnancy and association with gestational age[31]. The Comparative Toxicogenomics Database (CTD) contains more than 5,000 curated and inferred gene–disease associations (including preterm birth) extracted from the published literature by formal curation[32]. *STAT5A* is also among the CTD gene disease phenotypic associations with preterm birth. MAPK8 is also among the CTD gene disease associations associated with preterm birth[32]. Lastly, several genes in Cluster A are involved in cellular migration and invasion, which may be important in the onset of labor. *CTTN* is the cortactin gene. It functions as a key regulator of actin cytoskeleton and has roles in actin-based cellular processes including cell migration and invasion. Polymorphisms in the *VEGFA* gene in a discrete haplotype block have been shown to be associated with preterm birth[30]. This may play an important role in angiogenesis during placentation. Nonetheless, the latter study was of modest size and the findings just reached nominal significance.

In support of the network analysis, many of the genes are involved in uterine contractility, signal transduction and cell–cell signaling. Moreover, there is substantial literature based evidence that 4 out of the 6 unique genes in Cluster B are involved in uterine contractility. Signaling via *EGF* and the EGF receptor in human amnion cells regulates their proliferation and increases calcium mobilization and PGE2 production which may also have significant effects on uterine contractility[33]. *PAK1*, one of the genes identified in the Database for Preterm Birth[9], encodes a member of the serine threonine P21 activating kinases. *PAK1* is only present in pregnant myometrial tissue. PAKs have shown to regulate uterine contractility[34]. We have also previously reported that *PAK1* is located in a genomic region with runs of homozygosity (ROH) which are significantly more abundant

in mothers delivering preterm than term[26]. Activation of the *CHRM1* receptor by acetylcholine increases uterine contractility[35]. Of functional significance, *CHRM1* has been shown to be down-regulated in preterm human myometrium compared to patients at term and not-in-labor[36]. *WASF1* is an A-kinase anchoring protein. It has been implicated in preterm labor by the Ontario Birth Study where it was shown to be differentially expressed in patients undergoing preterm labor[37]. It was further shown to be responsive to glucocorticoids in another study of peripheral blood mononuclear cells from patients delivering preterm[37].

Previous investigations have been undertaken to identify social, environmental and genetic associations with preterm birth[14,38–41]. This has included case controlled studies of single nucleotide polymorphisms (SNP) in the protein coding regions, regulatory and intronic sequences of specific genes have been described. Because of the prominent inferences of inflammatory reactivity and alterations in uterine contractility, candidate have largely been investigated[14,38–53]. Modest associations in some SNPs and alterations in the expression of genes regulating inflammatory mechanisms have been identified. Nonetheless, the results do not explain the cause of prematurity without evidence of inflammation or infection. Additionally, while treatments directed at infection and inflammation in animal models have been successful, they have not demonstrated benefit in the treatment of preterm birth or prolongation of gestation in humans[49].

More recently, investigations using high throughput and multi-omic techniques have been undertaken. Sakabe et al. compared transcriptome and regulatory maps of decidua-derived stromal cells to a genome-wide association study of gestational duration[54]. Using a combination of techniques, including ATAC-seq, Hi-C and ChIP-seq, they identified the chromatin landscapes in decidua-derived stromal cells. These were then compared to the heritability of annotations in the GWAS of pregnancy-related traits. They showed the heritability of gestational-duration was enriched for functional annotations in decidual stromal cells[54]. Volozanoka recently carried out targeted sequencing of genes shown to be related to cervical insufficiency following a systematic literature analysis similar to that underlying this study[9,55]. They identified 12 genes that were normally linked to cervical insufficiency. However, this was a modest study involving only 21 patients and there were no overlaps with our gene set. Zhou et al. recently analyzed publically available gene expression data sets derived from maternal blood in the second and third trimesters of women with spontaneous preterm birth and term birth[56]. Expression of a single gene, *EBF1*, was associated with preterm birth[57]. In a large genome-wide association study of over 1300 cases of spontaneous preterm birth in comparison to 12,000 ancestry-matched controls, they identified only two intergenic loci associated with preterm birth. The authors concluded that the genetic contributions to preterm birth are unlikely due to single common genetic variants but could be explained by interactions of multiple variants or environmental influences[57]. Meta-analysis of maternal and fetal transcriptomic data from genomic databases for studies related to preterm birth identified genes differentially expressed in spontaneous preterm birth relative to term. Ontogeny analysis demonstrated that maternal changes were enriched in immune-related pathways, upregulation of innate immunity and downregulation of adaptive immunity[58]. By comparison, analysis of the transcription profile in cord blood showed a downregulation of innate immune findings. Nonetheless, the results demonstrate a significant relationship of immune functions in the pathogenesis of preterm birth. Multi-omic analysis of preterm birth from the parent study of our INOVA replication samples has been reported[24]. It was a large study integrating whole genome sequencing, RNA sequencing and DNA methylation data for 270 preterm births and 520 control families. In their univariate analysis there were no variants that reached genome-wide significance. However, there were groups of genes associated with preterm birth in subsets of patients that were identified in secondary analyses[24]. Integration of the three data sources (WGS, RNAseq, DNAm) identified a set of 72 candidate biomarker genes for very early preterm birth (VEPTB) and genes associated with PTB. *RAB31* and *RBPJ* were identified by all three data types in preterm birth patients. Additionally, pathways associated with VEPTB included the EGFR and prolactin signaling pathways, inflammation- and immunity-related pathways, chemokine signaling, IFN-γ signaling, and Notch1 signaling[24].

Our study has many strengths which contributed to successful identification of genetic variants in genes in networks associated with preterm birth. First, we used a very carefully phenotyped cohort of women with a strong family history of preterm birth. Our control cases were as carefully ascertained to have no family history of preterm birth. Second, we carried out targeted sequencing on genes with a demonstrated role in preterm birth. Third, we employed a novel analysis of protein–protein interactions to identify clusters of patients with shared PPI networks associated with preterm birth. Our study also has limitations and areas that deserve consideration. A power calculation was not carried out. This was an opportunistic, discovery sample. The significant clusters of preterm birth cases included unique genes that were both in our targeted sequencing as well as imputed via network analysis. Even though not included in our original sequencing, the fact that the unique imputed genes had a strong association with preterm birth, uterine contractility and immune responses is noteworthy. This was a modest sample size. We identified significant clusters of patients with networks and pathways associated with preterm birth but those findings were restricted to 45 out of the 122 cases. We did not anticipate being able to assign each of the cases to a significant cluster. We believe that the targeted sequencing contributed to our successful discovery even though all cases were not assigned to significant clusters. Nonetheless, the fact that we were able to identify significant case dominated clusters in our preterm birth cohort at all and that we were able to demonstrate similarity to clusters of patients with shared networks in the replication cohort lends validity to our hypothesis on the genetic architecture of this complex disease and the genetic leverage provided by the family history of preterm birth. The fact that this was not a whole exome study likely contributed to the number of cases we were able to assign to significant clusters. Futures studies employing similar techniques but with whole exome sequencing are likely to expand the number of case clusters that we will be able to identify using this approach. It is beyond the scope of this report to thoroughly discuss the control dominated clusters. Nonetheless, several elements deserve mention. We interpret the networks and pathways that are shared between control patients to represent protective genes or genes that confer resiliency against preterm birth. It is notable that we were able to identify significant clusters in the controls from our preterm birth cohort but none were

identified in the replication cohort. We attribute this to the careful phenotyping that was used to enroll patients in the primary study.

In summary, we used a novel multi-sample, protein–protein interaction tool, to identify clusters of patients with shared protein–protein interaction networks associated with preterm birth. We showed similarity between these networks and results from an independent replication cohort. Our results provide insights into the genetics of PTB and support a genetic architecture defined by subgroups of patients that *share* variants in genes in specific networks and pathways which are sufficient to give rise to the disease phenotype.

Received: 13 June 2020; Accepted: 10 February 2021
Published online: 10 January 2022

## References

1. Cordell, H. J. Detecting gene-gene interactions that underlie human diseases. *Nat. Rev. Genet.* **10**, 392–404. https://doi.org/10.1038/nrg2579 (2009).
2. Moore, J. H. Detecting, characterizing, and interpreting nonlinear gene-gene interactions using multifactor dimensionality reduction. *Adv. Genet.* **72**, 101–116. https://doi.org/10.1016/B978-0-12-380862-2.00005-9 (2010).
3. Wang, K., Li, M. & Hakonarson, H. Analysing biological pathways in genome-wide association studies. *Nat. Rev. Genet.* **11**, 843–854. https://doi.org/10.1038/nrg2884 (2010).
4. Eichler, E. E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* **11**, 446–450. https://doi.org/10.1038/nrg2809 (2010).
5. Gibson, G. Hints of hidden heritability in GWAS. *Nat. Genet.* **42**, 558–560. https://doi.org/10.1038/ng0710-558 (2010).
6. Maher, B. Personal genomes: The case of the missing heritability. *Nature* **456**, 18–21. https://doi.org/10.1038/456018a (2008).
7. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753. https://doi.org/10.1038/nature08494 (2009).
8. McClellan, J. & King, M. C. Genetic heterogeneity in human disease. *Cell* **141**, 210–217. https://doi.org/10.1016/j.cell.2010.03.032 (2010).
9. Uzun, A. *et al.* dbPTB: A database for preterm birth. *Database J. Biol. Databases Curation*. https://doi.org/10.1093/database/bar069 (2012).
10. Uzun, A., Dewan, A. T., Istrail, S. & Padbury, J. F. Pathway-based genetic analysis of preterm birth. *Genomics* **101**, 163–170. https://doi.org/10.1016/j.ygeno.2012.12.005 (2013).
11. Armanious, D. *et al.* Proteinarium: Multi-sample protein-protein interaction analysis and visualization tool. bioRxiv:589085. https://doi.org/10.1101/589085 (2019).
12. Boyd, H. A. *et al.* Maternal contributions to preterm delivery. *Am. J. Epidemiol.* **170**, 1358–1364. https://doi.org/10.1093/aje/kwp324 (2009).
13. Svensson, A. C. *et al.* Maternal effects for preterm birth: A genetic epidemiologic study of 630,000 families. *Am. J. Epidemiol.* **170**, 1365–1372. https://doi.org/10.1093/aje/kwp328 (2009).
14. Weinberg, C. R. & Shi, M. The genetics of preterm birth: Using what we know to design better association studies. *Am. J. Epidemiol.* **170**, 1373–1381. https://doi.org/10.1093/aje/kwp325 (2009).
15. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324 (2009).
16. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352 (2009).
17. McKenna, A. *et al.* The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303. https://doi.org/10.1101/gr.107524.110 (2010).
18. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinform.* **43**, 111011–111033. https://doi.org/10.1002/0471250953.bi1110s43 (2013).
19. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909. https://doi.org/10.1038/ng1847 (2006).
20. Walsh, R. *et al.* Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. *Genet. Med.* **19**, 192–203. https://doi.org/10.1038/gim.2016.90 (2017).
21. Menche, J. *et al.* Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science* **347**, 1257601. https://doi.org/10.1126/science.1257601 (2015).
22. Qiao, D. *et al.* Whole exome sequencing analysis in severe chronic obstructive pulmonary disease. *Hum. Mol. Genet.* **27**, 3801–3812. https://doi.org/10.1093/hmg/ddy269 (2018).
23. Goldmann, J. M. *et al.* Parent-of-origin-specific signatures of de novo mutations. *Nat. Genet.* **48**, 935–939. https://doi.org/10.1038/ng.3597 (2016).
24. Knijnenburg, T. A. *et al.* Genomic and molecular characterization of preterm birth. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 5819–5827. https://doi.org/10.1073/pnas.1716314116 (2019).
25. Heng, Y. J. *et al.* Maternal whole blood gene expression at 18 and 28 weeks of gestation associated with spontaneous preterm birth in asymptomatic women. *PLoS ONE* **11**, e0155191. https://doi.org/10.1371/journal.pone.0155191 (2016).
26. Uzun, A. *et al.* Structural and genomic variation in preterm birth. *Pediatr. Res.* **80**, 829–836. https://doi.org/10.1038/pr.2016.152 (2016).
27. Lynch, V. J., Leclerc, R. D., May, G. & Wagner, G. P. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nat. Genet.* **43**, 1154–1159. https://doi.org/10.1038/ng.917 (2011).
28. Hamilton, S. A., Tower, C. L. & Jones, R. L. Identification of chemokines associated with the recruitment of decidual leukocytes in human labour: Potential novel targets for preterm labour. *PLoS ONE* **8**, e56946. https://doi.org/10.1371/journal.pone.0056946 (2013).
29. Lim, R., Barker, G. & Lappas, M. A novel role for FOXO3 in human labor: Increased expression in laboring myometrium, and regulation of proinflammatory and prolabor mediators in pregnant human myometrial cells. *Biol. Reprod.* **88**, 156. https://doi.org/10.1095/biolreprod.113.108126 (2013).
30. Romero, R. *et al.* Identification of fetal and maternal single nucleotide polymorphisms in candidate genes that predispose to spontaneous preterm labor with intact membranes. *Am. J. Obstet. Gynecol.* **202**(431), e431-434. https://doi.org/10.1016/j.ajog.2010.03.026 (2010).
31. Aghaeepour, N. *et al.* An immune clock of human pregnancy. *Sci. Immunol.* https://doi.org/10.1126/sciimmunol.aan2946 (2017).
32. Davis, A. P. *et al.* The comparative toxicogenomics database: Update 2019. *Nucleic Acids Res.* **47**, D948–D954. https://doi.org/10.1093/nar/gky868 (2019).

33. Tahara, M. *et al.* Expression of messenger ribonucleic acid for epidermal growth factor (EGF), transforming growth factor-alpha (TGF alpha), and EGF receptor in human amnion cells: possible role of TGF alpha in prostaglandin E2 synthesis and cell proliferation. *J. Clin. Endocrinol. Metab.* **80**, 138–146. https://doi.org/10.1210/jcem.80.1.7829602 (1995).
34. Moore, F. *et al.* Up-regulation of p21- and RhoA-activated protein kinases in human pregnant myometrium. *Biochem. Biophys. Res. Commun.* **269**, 322–326. https://doi.org/10.1006/bbrc.2000.2290 (2000).
35. Lopez Bernal, A. The regulation of uterine relaxation. *Semin. Cell Dev. Biol.* **18**, 340–347. https://doi.org/10.1016/j.semcdb.2007.05.002 (2007).
36. Charpigny, G. *et al.* A functional genomic study to identify differential gene expression in the preterm and term human myometrium. *Biol. Reprod.* **68**, 2289–2296. https://doi.org/10.1095/biolreprod.102.013763 (2003).
37. Menon, R., Fortunato, S. J., Thorsen, P. & Williams, S. Genetic associations in preterm birth: A primer of marker selection, study design, and data analysis. *J. Soc. Gynecol. Investig.* **13**, 531–541. https://doi.org/10.1016/j.jsgi.2006.09.006 (2006).
38. Paquette, A. G. *et al.* Comparative analysis of gene expression in maternal peripheral blood and monocytes during spontaneous preterm labor. *Am. J. Obstet. Gynecol.* **218**, 345e341–345e330. https://doi.org/10.1016/j.ajog.2017.12.234 (2018).
39. Pennell, C. E. *et al.* Genetic epidemiologic studies of preterm birth: guidelines for research. *Am. J. Obstet. Gynecol.* **196**, 107–118. https://doi.org/10.1016/j.ajog.2006.03.109 (2007).
40. Plunkett, J. & Muglia, L. J. Genetic contributions to preterm birth: Implications from epidemiological and genetic association studies. *Ann. Med.* **40**, 167–195. https://doi.org/10.1080/07853890701806181 (2008).
41. Romero, R. *et al.* The use of high-dimensional biology (genomics, transcriptomics, proteomics, and metabolomics) to understand the preterm parturition syndrome. *BJOG Int. J. Obstet. Gynaecol.* **113**(Suppl 3), 118–135. https://doi.org/10.1111/j.1471-0528.2006.01150.x (2006).
42. Bezold, K. Y., Karjalainen, M. K., Hallman, M., Teramo, K. & Muglia, L. J. The genomics of preterm birth: from animal models to human studies. *Genome Med.* **5**, 34. https://doi.org/10.1186/gm438 (2013).
43. Clausson, B., Lichtenstein, P. & Cnattingius, S. Genetic influence on birthweight and gestational length determined by studies in offspring of twins. *BJOG Int. J. Obstet. Gynaecol.* **107**, 375–381. https://doi.org/10.1111/j.1471-0528.2000.tb13234.x (2000).
44. Fujimoto, T. *et al.* A single nucleotide polymorphism in the matrix metalloproteinase-1 (MMP-1) promoter influences amnion cell MMP-1 expression and risk for preterm premature rupture of the fetal membranes. *J. Biol. Chem.* **277**, 6296–6302. https://doi.org/10.1074/jbc.M107865200 (2002).
45. Genc, M. R., Gerber, S., Nesin, M. & Witkin, S. S. Polymorphism in the interleukin-1 gene complex and spontaneous preterm delivery. *Am. J. Obstet. Gynecol.* **187**, 157–163. https://doi.org/10.1067/mob.2002.122407 (2002).
46. Kalish, R. B., Vardhana, S., Gupta, M., Perni, S. C. & Witkin, S. S. Interleukin-4 and -10 gene polymorphisms and spontaneous preterm birth in multifetal gestations. *Am. J. Obstet. Gynecol.* **190**, 702–706. https://doi.org/10.1016/j.ajog.2003.09.066 (2004).
47. Landau, R. *et al.* beta2-Adrenergic receptor genotype and preterm delivery. *Am. J. Obstet. Gynecol.* **187**, 1294–1298. https://doi.org/10.1067/mob.2002.128524 (2002).
48. Lorenz, E., Hallman, M., Marttila, R., Haataja, R. & Schwartz, D. A. Association between the Asp299Gly polymorphisms in the Toll-like receptor 4 and premature births in the Finnish population. *Pediatr. Res.* **52**, 373–376. https://doi.org/10.1203/00006450-200209000-00011 (2002).
49. Nadeau-Vallee, M. *et al.* Novel noncompetitive IL-1 receptor-biased ligand prevents infection- and inflammation-induced preterm birth. *J. Immunol.* **195**, 3402–3415. https://doi.org/10.4049/jimmunol.1500758 (2015).
50. Ozkur, M. *et al.* Association of the Gln27Glu polymorphism of the beta-2-adrenergic receptor with preterm labor. *Int. J. Gynaecol. Obstet.* **77**, 209–215. https://doi.org/10.1016/s0020-7292(02)00035-8 (2002).
51. Papazoglou, D., Galazios, G., Koukourakis, M. I., Kontomanolis, E. N. & Maltezos, E. Association of -634G/C and 936C/T polymorphisms of the vascular endothelial growth factor with spontaneous preterm delivery. *Acta Obstet. Gynecol. Scand.* **83**, 461–465. https://doi.org/10.1111/j.0001-6349.2004.00403.x (2004).
52. Simhan, H. N., Krohn, M. A., Roberts, J. M., Zeevi, A. & Caritis, S. N. Interleukin-6 promoter -174 polymorphism and spontaneous preterm birth. *Am. J. Obstet. Gynecol.* **189**, 915–918. https://doi.org/10.1067/s0002-9378(03)00843-3 (2003).
53. Witkin, S. S. *et al.* Polymorphism in intron 2 of the fetal interleukin-1 receptor antagonist genotype influences midtrimester amniotic fluid concentrations of interleukin-1beta and interleukin-1 receptor antagonist and pregnancy outcome. *Am. J. Obstet. Gynecol.* **189**, 1413–1417. https://doi.org/10.1067/s0002-9378(03)00630-6 (2003).
54. Sakabe, N. *et al.* Transcriptome and regulatory maps of decidua-derived stromal cells inform gene discovery in preterm birth. BioRxiv: 2020.2004.2006.017079. https://doi.org/10.1101/2020.04.06.017079 (2020).
55. Volozonoka, L. *et al.* Genetic landscape of preterm birth due to cervical insufficiency: Comprehensive gene analysis and patient next-generation sequencing data interpretation. *PLoS ONE* **15**, e0230771. https://doi.org/10.1371/journal.pone.0230771 (2020).
56. Zhou, G. *et al.* EBF1 gene mRNA levels in maternal blood and spontaneous preterm birth. *Reprod. Sci.* **27**, 316–324. https://doi.org/10.1007/s43032-019-00027-2 (2020).
57. Rappoport, N. *et al.* A genome-wide association study identifies only two ancestry specific variants associated with spontaneous preterm birth. *Sci. Rep.* **8**, 226. https://doi.org/10.1038/s41598-017-18246-5 (2018).
58. Vora, B. *et al.* Meta-analysis of maternal and fetal transcriptomic data elucidates the role of adaptive and innate immunity in preterm birth. *Front. Immunol.* **9**, 993. https://doi.org/10.3389/fimmu.2018.00993 (2018).


## Acknowledgements

This work was supported by a grant from National Foundation March of Dimes Prematurity Initiative # 21-FY08-563, National Institutes of Health grants P30GM103410, P20GM121298, P30GM114750, P20GM109035 and the Kilguss Research Core at Women & Infants Hospital. The INOVA study was supported by the Inova Health System, a nonprofit healthcare system in Northern Virginia.

## Author contributions

A.U.: Conceptualization, Data Analysis, Methodology, Writing-Editing; J.S.S.: Conceptualization, Data Analysis, Methodology, Writing-Editing; J.S.: Experimental Analysis, Editing; V.Z.: Experimental analysis, Editing; G.A.T.: Data Analysis, Validation, Editing; A.A.: Data Analysis, Validation, Editing; P.K.: Collaboration, Data Sharing, Editing; W.S.W.W.: Collaboration, Data Sharing, Editing; J.P.: Conceptualization, Project Administration, Data Analysis, Writing-Review & Editing, Funding Acquisition.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-03427-0.

**Correspondence** and requests for materials should be addressed to J.P.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.