PLOS | ONE

RESEARCH ARTICLE

Targeted Sequencing and Meta-Analysis of Preterm Birth

Alper Uzun[1,2], Jessica Schuster[1], Bethany McGonnigal[1], Christoph Schorl[3], Andrew Dewan[4], James Padbury[1,2,5]*

1 Department of Pediatrics, Women & Infants Hospital of Rhode Island, Providence, Rhode Island, United States of America, **2** Brown Alpert Medical School, Providence, Rhode Island, United States of America, **3** Molecular Biology, Cell Biology & Biochemistry, Brown University, Providence, Rhode Island, United States of America, **4** Department of Epidemiology and Public Health, Yale University, New Haven, Connecticut, United States of America, **5** Center for Computational Molecular Biology, Brown University, Providence, Rhode Island, United States of America

* jpadbury@wihri.org

OPEN ACCESS

Citation: Uzun A, Schuster J, McGonnigal B, Schorl C, Dewan A, Padbury J (2016) Targeted Sequencing and Meta-Analysis of Preterm Birth. PLoS ONE 11 (5): e0155021. doi:10.1371/journal.pone.0155021

Editor: Cristina Cereda, Center of Genomic & Post Genomics, ITALY

Received: November 26, 2015

Accepted: April 22, 2016

Published: May 10, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information file.

Funding: This work was supported by grants from the National Foundation March of Dimes #21-FY14-154, the Rhode Island Foundation #20133978, and the National Institutes of Health #P30 GM114750-01, #P20 RR18728 and #P30 GM103410. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Understanding the genetic contribution(s) to the risk of preterm birth may lead to the development of interventions for treatment, prediction and prevention. Twin studies suggest heritability of preterm birth is 36–40%. Large epidemiological analyses support a primary maternal origin for recurrence of preterm birth, with little effect of paternal or fetal genetic factors. We exploited an "extreme phenotype" of preterm birth to leverage the likelihood of genetic discovery. We compared variants identified by targeted sequencing of women with 2–3 generations of preterm birth with term controls without history of preterm birth. We used a meta-genomic, bi-clustering algorithm to identify gene sets coordinately associated with preterm birth. We identified 33 genes including 217 variants from 5 modules that were significantly different between cases and controls. The most frequently identified and connected genes in the exome library were IGF1, ATM and IQGAP2. Likewise, SOS1, RAF1 and AKT3 were most frequent in the haplotype library. Additionally, SERPINB8, AZU1 and WASF3 showed significant differences in abundance of variants in the univariate comparison of cases and controls. The biological processes impacted by these gene sets included: cell motility, migration and locomotion; response to glucocorticoid stimulus; signal transduction; metabolic regulation and control of apoptosis.

Introduction

Despite significant advances in the care of pregnant mothers and infants, preterm birth remains a leading cause of newborn morbidity, mortality and hospitalization in the first year of life in the United States [1]. In developed countries 70% of infant mortality is secondary to preterm birth (birth before 37 completed weeks of gestation). The rate of preterm birth varies in different societies and in different ethnic groups from 3.8% in Eastern Asia to rates reaching close to 17% in disadvantaged African American groups [2, 3]. Neonatal morbidity and mortality after preterm birth are inversely related to gestational length. Survivors are at increased risk

of cerebral palsy, intellectual disabilities, respiratory problems and other long term conditions [4]. Moreover, despite numerous attempts at intervention, the incidence of prematurity has shown minimal improvement over the last two decades [2]. The risk factors associated with prematurity are many including: adverse sociodemographic status, race/ethnicity, infection, stress, trauma and prior history of a premature birth [4–10]. The leading etiology is idiopathic. A large number of clinical/epidemiologic studies have examined the individual and collective contribution of each of these factors. A family history of preterm birth and inter-pregnancy interval of <18 months also increase the risk of prematurity [9].

A precise estimate of the contribution(s) of genetic factors to preterm birth has been difficult to achieve [11–17]. Twin studies suggest heritability is 36–40%, however differences in gestational age used and other details cloud the precision of those estimates [18, 19]. A history of a previous preterm birth is one of the best predictors of a subsequent preterm delivery. Mothers who were preterm themselves or who have a first order relative with preterm birth have an increased risk for delivering prematurely. These observations support the importance of genetic factors in preterm birth [13, 20, 21]. Large epidemiological studies drawn from population based analyses in Sweden and Denmark support a maternal origin for the genetic contribution(s) to risk of preterm birth, with little contribution by paternal or fetal genetic factors [17, 22–24].

Attempts to identify the genetic contributions to the risk of preterm birth have been pursued widely [13–17, 25, 26]. Studies have focused on genomic and proteomic approaches to the mechanism(s) of preterm labor. Polymorphic changes in the protein coding regions, regulatory and intronic sequences of specific genes have been described. In most of these studies, candidate genes or proteins involved in inflammatory reactivity or uterine contractility have been investigated [13–18, 25–37]. The results suggest that alteration in the expression of these proteins interacts with infection and/or other environmental influences associated with preterm birth. The results however, do not provide insight into the causes of prematurity in the absence of early inflammation or infection. Moreover, while interventions directed at infection or inflammation have been successful in experimental models they have largely been unsuccessful in treatment or prevention of preterm birth in humans [38]. Thus, there is abundant information that demonstrates important genetic contribution(s) to the risk of preterm birth and further suggests that preterm birth is a complex, polygenic disorder that entails activation and/or suppression of a host of genes [4]. In addition, linkage analyses have been limited because large pedigrees with a family history of preterm birth are not widely available, however one such study has been reported [39]. In spite of the data suggesting an association between genetics and PTB, there is a gap in our knowledge of the precise genetic contributions and whether they are discrete or multifactorial.

We have developed an alternative approach to identify a more manageable set of genes for preterm birth which incorporates some elements of the discovery in genome wide investigations. We previously used a bioinformatics approach for mining published literature and screening publicly available high-throughput databases to develop a validated collection of genes with *a priori* connection to preterm birth [40]. We used gene set enrichment analysis (GSEA) of this refined gene set to analyze a large genome wide association study to identify the contribution(s) of individual biological pathways to the genetic architecture of preterm birth [40, 41]. We identified important genes and networks associated with preterm birth. In order to identify the variants underlying these associations, we targeted the exons, flanking sequence and splice sites of the 329 genes and 132 haplotype blocks that we showed were associated with preterm birth [41]. We were as interested in variants that were associated with increased risk for preterm birth as we were with variants that were associated with reduced risk. We exploited an "extreme phenotype" of preterm birth to leverage the likelihood of genetic discovery by

concentrating our enrollment on patients with a prior 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 used a meta-analytic, bi-clustering algorithm to identify gene sets coordinately associated with preterm birth. We identified 33 genes including 217 variants from 5 modules significantly different between cases and controls. The biological processes impacted by these gene sets included: cell motility, migration and locomotion; response to glucocorticoid stimulus; signal transduction; metabolic regulation and control of apoptosis.

Results

Library Design and Univariate Sequence Analysis

Sequencing was carried out on 32 women with 2 or 3 generations of preterm birth and 16 controls. We targeted the exons, flanking sequence and splice sites of the 329 genes and 132 haplotype blocks that we had previously shown were highly associated with preterm birth [41]. We identified over 13,000 variants in the targeted exome library and 11,000 variants in the haplotype block library [41]. Using the univariate analysis strategy discussed in the Methods, we identified 205 and 168 variants that were significantly different in abundance between cases and controls from the exome and haplotype block libraries at $p<0.05$, respectively. These variants and their associated genes are shown in S1 and S2 Tables. Fig 1 shows a Manhattan plot for these combined results with a threshold at -logP 1.3.

Meta-analysis

The genes containing variants that showed significant differences between cases and controls were examined for their association with networks and biological pathways using GSEA. We analyzed genes whose variants differed from controls with a p-value <0.1. We ran GSEA independently for each of the 48 patients. The significant gene sets from the GSEA of each patient were then compared by adapting a newly described meta-analytic approach known as iterative

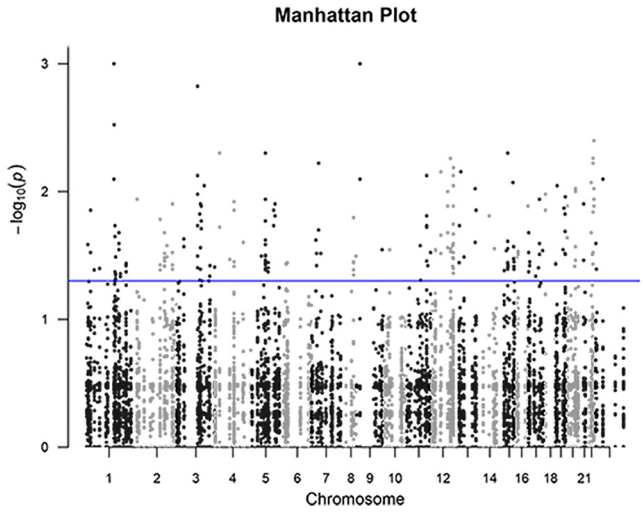

Fig 1. Manhattan Plot of Significant Variants. The 13,000 variants from the targeted exome library and 11,000 variants from the haplotype block library were compared for difference in abundance in the cases versus the controls. The figure shows a Manhattan plot of all variants across 22 autosomes with the vertical axis being the -logP value from the statistical test for association, with the threshold line (-logP 1.3) indicating p-value of 0.05. There were 205 and 168 variants that significantly differed in abundance in cases versus controls from the exome and haplotype block libraries respectively.

doi:10.1371/journal.pone.0155021.g001

binary bi-clustering (iBBiG) [42]. The iBBiG algorithm identifies "modules" of gene sets and patient subsets from binary data [42]. Our analytical pipeline is illustrated in Fig 2.

For each module we analyzed the patient subsets by comparing the number of cases versus the number of controls. These results are summarized in Table 1, which lists the module

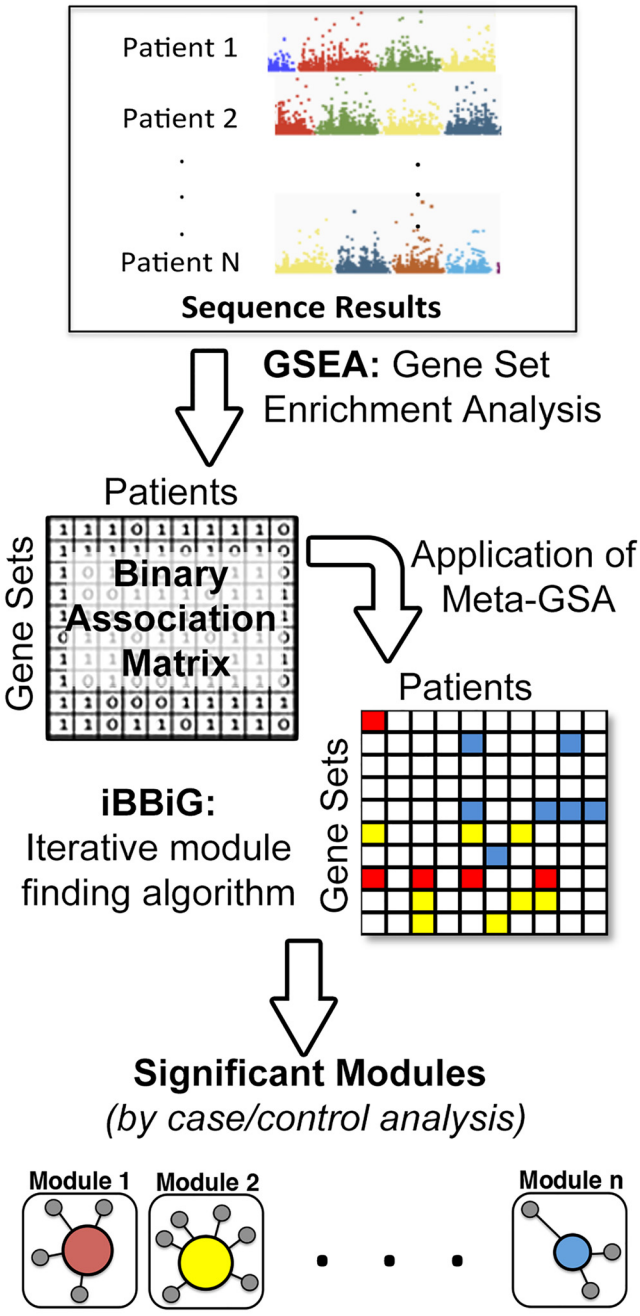

Fig 2. Meta-analysis and analytical pipeline: The genes harboring variants in each patient were analyzed by gene set enrichment using the MSig database C2 collection of gene sets [43]. The significant gene sets for each patient were combined into a binary association matrix. The iBBiG algorithm extracts modules of gene sets and patient subsets from the data matrix. The modules are represented by different colors. Fisher's exact test was used to identify modules with significant differences in the number of cases and controls.

doi:10.1371/journal.pone.0155021.g002

Table 1. The number of case and control patients and Fisher's Exact p-value for each of the significant modules from the two targeted sequencing libraries.

Libraries	Modules	Cases	Controls	p-value
Exome Library	M4	2	9	0.0002
	M8	9	0	0.0200
Haplotype Library	H2	9	0	0.0200
	H8	1	4	0.0300
	H9	0	5	0.0025

doi:10.1371/journal.pone.0155021.t001

number, the numbers of cases, the numbers of controls, and the p-value (Fisher's exact test). This analysis of the exome library identified 2 modules, for which there were significant differences in number of cases and controls in the patient subsets. For the haplotype library 3 significant modules were identified. Fig 3A and 3B shows a network of all of the modules and the patients assigned to each module from the exome and haplotype libraries. For the exome library, Module 8 (blue) contained significantly more cases than controls and Module 4 (red) contained significantly more controls than cases. For the haplotype library, Module 2 (light blue) contained significantly more cases than controls and Module 8 (green) and Module 9

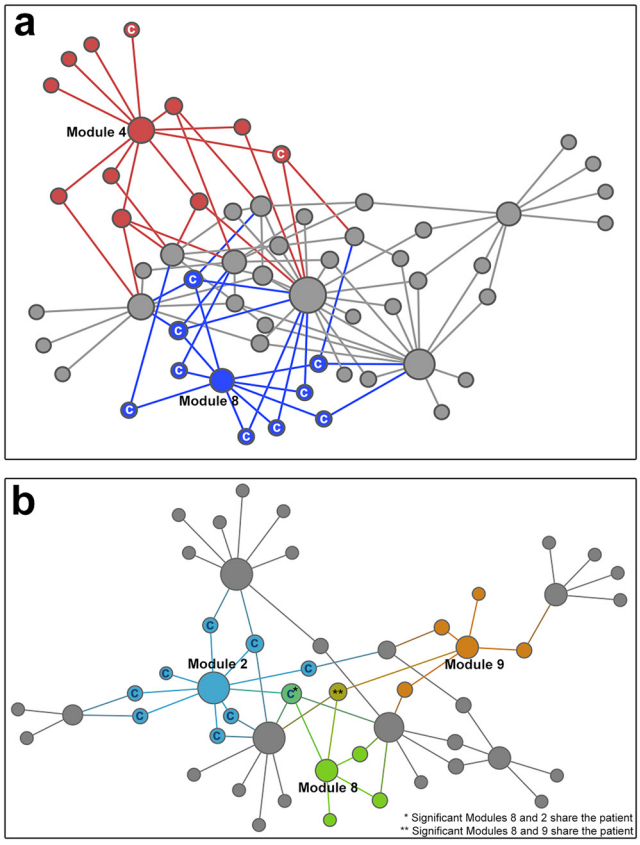

Fig 3. Network analysis of cases and controls. Cases are labeled by letter "c" in the significant modules. **(A)** Network of all modules of the exome library and patients, with the patients from significant Modules 4 and 8 highlighted in red and blue, respectively. **(3B)** Network of all modules of the haplotype block library and patients, with the patients from significant Modules 2, 8 and 9 highlighted in light blue, green and orange, respectively.

doi:10.1371/journal.pone.0155021.g003

(orange) contained significantly more controls than cases. Not surprisingly, patient subsets overlapped between modules. However for the exome library, the significant modules were represented by discrete patients. In the haplotype library two patients were shared among the 3 significant modules.

In addition to patient subsets, each module contained gene sets. We extracted the genes from each gene set. We found significant overlap when analyzing these modules. Several gene sets were included in more than one module and there were multiple genes within each gene set that were shared among modules. This overlap is displayed for the exome library and the haplotype block library in Fig 4A and 4B. The significant modules and associated genes are displayed as Insets of Fig 4A and 4B. Inset **a1** shows the 6 genes in Module 8 and inset **a2** shows the 8 genes in Module 4. Inset **b1** shows the 17 genes in Module 2, **b2** shows the 5 genes in Module 8 and **b3** shows the 6 genes in Module 9. Table 2 lists these individual genes, their genomic location, and the five significant modules to which they belong.

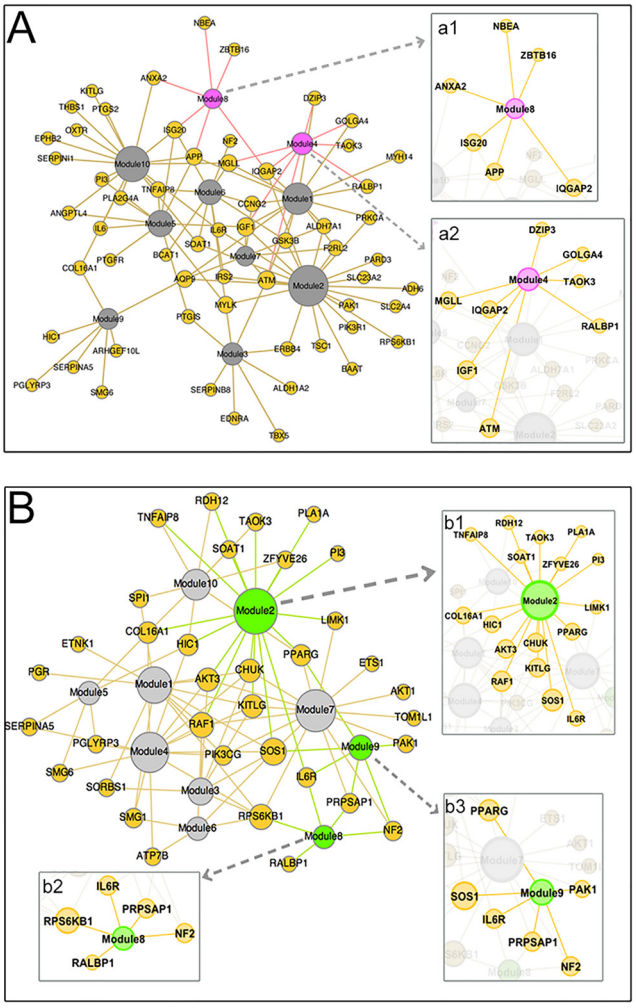

Fig 4. Network of modules and their gene sets. (A) Network output showing all 10 modules from the exome library and the genes contained in each module. The two significant modules are displayed as insets. Inset **a1** and **a2** display the genes of E8 and E4 respectively. **(B)** Network output showing all 10 modules from the haplotype block library and the genes contained in each module. Insets **b1, b2 and b3** show the genes of H2, H9 and H3 respectively.

doi:10.1371/journal.pone.0155021.g004

Table 2. Gene names, IDs and chromosome numbers identified in the significant modules from both targeted sequencing libraries.

Gene	HGNC ID	Chr	Modules
DZIP3	30938	3	E4
GOLGA4	4427	3	E4
TAOK3	18133	12	E4, H2
RALBP1	9841	18	E4, H8
IGF1	5464	12	E4
ATM	795	11	E4
MGLL	17038	3	E4
IQGAP2	6111	5	E4,E8
NBEA	7648	13	E8
ZBTB16	12930	11	E8
APP	620	21	E8
ISG20	6130	15	E8
ANXA2	537	15	E8
TNFAIP8	17260	5	H2
RDH12	19977	14	H2
SOAT1	11177	1	H2
PLA1A	17661	3	H2
PI3	8947	20	H2
LIMK1	6613	7	H2
PPARG	9236	3	H2, H9
ZFYVE26	20761	14	H2
SOS1	11187	2	H2, H9
KITLG	6343	12	H2
CHUK	1974	10	H2
RAF1	9829	3	H2
AKT3	393	1	H2
HIC1	4909	17	H2
COL16A1	2193	1	H2
IL6R	6019	1	H2, H8, H9
PRPSAP1	9466	17	H8, H9
NF2	7773	22	H8, H9
RPS6KB1	10436	17	H8
PAK1	8590	11	H9

E4 = Exome Library Module 4, **E8 =** Exome Library, Module 8,
H2 = Haplotype Library, Module 2, **H8 =** Haplotype Library, Module 8, **H9 =** Haplotype Library, Module 9

doi:10.1371/journal.pone.0155021.t002

The most highly connected genes in the exome library, IGF1, ATM and IQGAP2, were identified in 4–5 modules, Fig 4A. Similarly, from the haplotype block library, SOS1, RAF1 and AKT3 were identified in 5–6 modules, Fig 4B.

We recognize that some of the individual variants we found during the initial univariate testing might be important and not identified in the meta-analysis. Table 3 shows the variants in SERPINB8, AZU1 and WASF3 that were significantly different in cases and controls ($p < 0.05$) with predicted deleterious effects according to PolyPhen 2 HDIV.

Table 3. Significant variants from the Exome library with annotations.

Gene	HGNC id	Chr	Function	dbSNP 138	Polyphen 2HDIV	p-value
SERPINB8	8952	18	exonic	rs3826616	0.998	0.036
AZU1	913	19	exonic	rs28626600	0.995	0.037
WASF3	12734	13	exonic	rs17084492	0.968	0.036

doi:10.1371/journal.pone.0155021.t003

Gene Ontology Analysis

We performed gene ontology analysis to identify the biological processes for the genes belonging to the significant modules [44]. A total of 80 groups of biological processes were identified which segregated into 9 individual and overlapping clusters, S3 Table. Fig 5 shows these individual and shared ontology groups. The most abundant association was with mechanisms regulating programmed cell death. Likewise, control of cell motility, migration and cell cycle regulation were associated with several of the most highly connected genes in Module 4 of the exome library. Metabolic processes, phosphate and lipid metabolism, protein phosphorylation and various forms of signal transduction were common biological functions attributed to the other most highly connected genes from the exome library. Similarly, the results from the gene ontology analysis of the haplotype library showed a high degree of association with cellular metabolism, signal transduction and nucleic acid metabolism. Regulation of immune cell system development, responses to glucocorticoid signaling, signal transduction pathways in

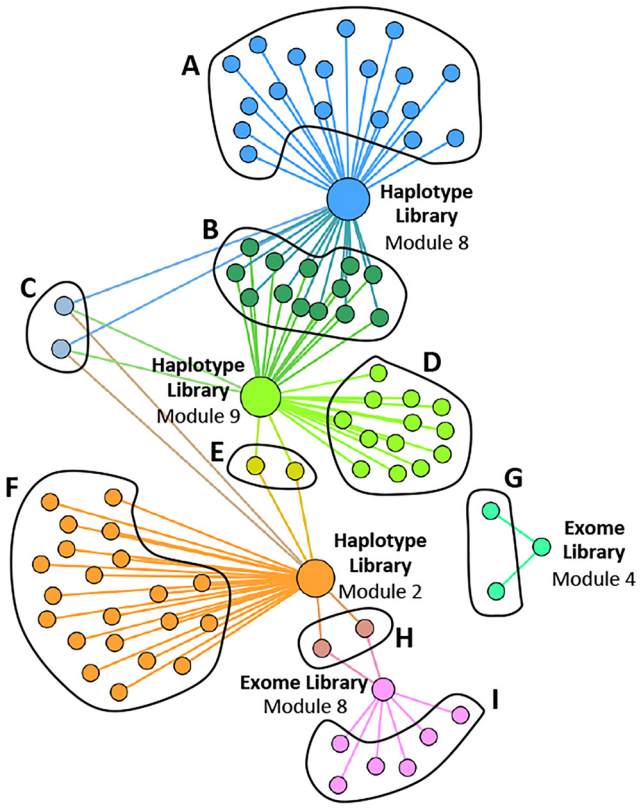

Fig 5. Ontology groups. Diagram showing the clusters of terms from the Gene Ontology analysis for biological processes related to preterm birth. Gene Ontology Database terms for biological processes shown in clusters A thru I are detailed in S3 Table.

doi:10.1371/journal.pone.0155021.g005

immune regulatory cells and regulation of smooth muscle cell proliferation were associated with the genes identified in the haplotype library.

Discussion

We are striving to identify the genetic basis for preterm birth. Our prior pathway analysis supports our strategy to look for variants in shared networks or pathways of genes that contribute to risk or resilience [41]. In order to increase the likelihood of discovery, we leveraged genetic risk by concentrating our enrollment on patients with a prior 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 performed targeted sequencing of the haplotype blocks surrounding previously identified significant variants from a GWAS for preterm birth [41]. We also performed targeted sequencing of the exons and flanking sequences of the genes nearest these variants. We then filtered the resulting 25,000 variants for those which were significantly different between preterm birth cases and matched controls. We adapted a new meta-analysis to identify modules of gene sets that were increased in abundance in cases or the controls. Several gene sets were included in more than one module and there were multiple genes within each gene set that were shared among modules. The most frequently identified and connected genes in the exome library were IGF1, ATM and IQGAP2. Likewise, SOS1, RAF1 and AKT3 were most frequent in the haplotype library. Additionally, SERPINB8, AZU1 and WASF3 showed significant differences in abundance of variants in the univariate comparison of cases and controls. IGFI has been previously implicated in preterm birth in association with variants in coagulation and inflammation pathway genes [45]. IGF1 has also been implicated through reduced expression in the placenta from preterm birth compared to controls [46]. The IGF-I receptor has also been associated with preterm birth through a linkage analysis in the Finish population [39]. Interestingly, none of the remaining genes were previously implicated in preterm births. Their inclusion in this discovery sent was solely through our prior imputation [41]. The high degree of overlap of the same gene in different gene sets and the same genes in different modules is consistent with the well-recognized redundancy of genes and their networks in nature [47].

Whole exome sequencing (WES) has been undertaken to identify the genetic architecture of complex diseases [48–50]. While successful at identifying large numbers of variants, specificity is limited. A recent WES project from the National Heart, Lung and Blood Institute (NHLBI) identified almost 500,000 nucleotide variants which were rare [50]. Remarkably, individual patients were predicted to have up to 300 putatively deleterious variants but actual phenotype genotype correlations were not available. It was only after careful review of the literature and prioritization of the identified variants that targets for resequencing were identified [51, 52]. Our strategy is the inverse of the approach just described. We carried out the biological reductionism first through a robust literature curation and aggregation of genes from public databases [40]. We then used gene set enrichment with this biologically validated gene set to analyze a large genome wide association study of preterm birth [41]. We identified a modest number of genes and significant haplotype blocks. In this report we describe the results of targeted deep sequencing of these genes and significant haplotype blocks in women with a multigenerational history of preterm birth and compared the findings to patients delivering at term with no family history of preterm birth.

Our results are consistent with our *a priori* hypothesis that preterm birth would not be associated with single gene variant(s) but rather with variants in networks of genes. Additionally, we anticipated that we would find networks with gene variants in some but not all of the cases and controls. This is consistent with the notion that a minimal but sufficient disruption in

several pathways is sufficient to lead to a clinical disease or phenotype but that different networks or modules can result in similar clinical or phenotypic outcomes [40, 53]. This approach is powerful at identifying subsets of patients with networks of genes that are associated with clinical disease phenotypes.

We compared cases with a multigenerational history of preterm birth to patients delivering at term with no family history of preterm birth. This approach is consistent with recommendations on design of studies to define rare variants and "missing heritability" which include careful phenotyping of cases, carefully-matched controls, use of *prior* data on genes or variants to identify targets and/or assess results [54, 55]. Our data are from a modest size cohort of patients. Nonetheless, our targeted strategy allowed us to find significant associations which both enhanced and reduced the risk of preterm birth. Other investigators have used combinations of targeted sequencing and/or targeted patient enrollment to enhance discovery of rare variants using modest patient size cohorts and have reported similar success [55–60]. As shown here, prior genetic analysis and prior filtering of both patients and gene targets improves the likelihood of identifying otherwise difficult-to-find rare variants [55–60]. Replication in another cohort of patients, comparison with genes associated evolutionarily with preterm birth and the addition of phylogenomic analyses are needed to validate and add veracity to these candidate genes [61].

These results illustrate an effective way to use large data sets and layered approaches employing pathway analysis, gene set enrichment and meta-genomic analysis, to identify the genes in networks and pathways associated with complex disease. We discovered modules of genes for which the variants in these genes taken together might prevent or result in preterm birth for a specified subset of patients. This meta-genomic approach is also suitable for meta-analysis of sequence/variant results from independent projects to identify gene networks in additional subsets of patients/phenotypes. These results are generalizable to other disorders.

Methods

Patient Identification and Enrollment

Women & Infants Hospital of Rhode Island 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 patients with a prior history of preterm birth. An informatically driven retrieval from our electronic medical record gave us a daily report on all preterm births. A clinical research assistant trained in genetic interviews reviewed the records of all patients delivering < 34 weeks. Controls were patients who delivered ≥ 37 weeks gestation in whom a careful, formal genetic history revealed no history of preterm birth on either maternal or paternal side of the pedigree. Following informed consent, women underwent a careful interview. There were explicit questions in the formal questionnaire of preterm birth in mother, grandmother, her first order relatives and also paternal relatives. Informed consent was obtained from all participants. 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. The study was approved by the Institutional Review Board, 08–0117. 48 samples were selected for targeted sequencing. Samples were taken from 23 women with 2 generations of preterm birth, 9 women with 3 generations of preterm birth and 16 control women at term. The clinical characteristics of the patients are shown in Table 4. The only significant difference was the older gestational age among the controls. All of the patients' identifying data was coded and redacted for the purposes of data analysis. Residual maternal whole blood was obtained for extraction of genomic DNA. The samples were stored continuously at -80°C until processing.

Table 4. Clinical characteristics of maternal patients (mean ± SD).

Study Group	Maternal Age	Gravida	Gestational Age	Race
Preterm	25.0 ± 5.3	2.8 ± 2.3	31.9 ± 2.3	A8; AS1; H8; W18;NA1
Controls	24.8 ± 4.5	1.8 ± 1	40.0 ± 0.7*	A3; AS1; H3; W9

* P<0.05A;

A African-American; AS Asian; H Hispanic; W White; NA Native American

doi:10.1371/journal.pone.0155021.t004

Sample preparation

We targeted 329 genes and 132 haplotype blocks that are highly associated with preterm birth for sequencing [41]. Genomic DNA from whole blood was extracted using DNA kit 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 **μ**g 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 ds DNA Fragmentase (NEB), followed by indexing and clean-up. DNA capture was performed using custom capture probes from SeqCap EZ choice kit (Roche NimbleGen) Post-capture quality control and targeted sequencing were performed at the Brown University Genomics Core.

Targeted sequencing

The library was sequenced on our Illumina *HiSeq* 2500 using 100 base pair paired-end protocols. Initial cluster counts of ~300,000 were obtained. 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 were an average of 22,000,000 reads from each patient, with an average of 99% perfect index reads and a Q30 of 91%. The mean Phred score for each patient was 36. These data were then aligned to the human reference sequence (Hg19). Reads were mapped to the to the human reference sequence (Hg19) with BWA [62] sorted and indexed with SAMtools [63].

Sequence data, variant calling and zygosity testing

Variants were flagged as low quality and filtered using the established metrics: if three or more variants detected within 10bp; if four or more alignments map to different locations equally well; if coverage of less than five reads; if quality score < 30; if low quality for a particular sequence depth (variant confidence/unfiltered depth < 1.5); and if strand bias (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. Following the filters described in Methods, we implemented GATK's *Haplotype Caller* [64, 65]. Duplicated reads marked and removed using Picard Tools version 1.77. Haplotype caller was applied for variant detection on 329 gene set library and the Haplotype blocks library. 100 base pairs upstream and downstream of the each gene were included in the variant detection.

Annotations

Variants were annotated using ANNOVAR for functional prediction scores Polyphen 2 HDIV [prediction if a change is damaging ($>=0.957$), possibly damaging ($0.453<=$ Polyphen

2 HDIV < = 0.956) or benign < = 0.452], PhyloP [prediction of a conserved (>0.95) or non-conserved (<0.95) site], and chromosome position [66].

Univariate Analysis

In order to focus on putatively relevant variants, we eliminated the bulk of the identified variants for immediate investigation because they were equally common in cases and controls. Using a Markov Chain Monte Carlo (MCMC) Fisher Exact Test, we created a 2 x 3 contingency table for zygosity testing to compare the frequency of homozygosity for the reference allele, heterozygosity or homozygosity for the minor allele. The results are shown in Table 3.

Meta-Analysis using GSEA and iBBiG

For each patient (total of 48), we built a gene list for each patient for GSEA. In order to ensure that we analyzed enough genes in each patient to conduct gene set enrichment, we relaxed the significance threshold on genes from p <.05 to p <.1. We ran GSEA on each patient's gene list independently [43]. We used pre-ranked GSEA analysis against a collection of curated gene sets (C2) from MSigDB [43]. The resultant gene sets were considered significant if they obtained a nominal p-value below 0.05. Next, we transformed the significant gene sets into a binary matrix where the rows of the matrix were the gene sets and the columns were the individual patients. The binary matrix was used as input into the iterative binary bi-clustering algorithm (iBBiG). This algorithm was used to identify groups of gene sets that are coordinately associated with subsets of patients and their phenotypes (preterm, term) across the GSEA results. Fisher exact test was used to compare the abundance of cases and controls in each module output from iBBiG.

Gene Ontology

We sought to provide a representation of the biological processes encompassed by these gene sets using the Gene Ontology (GO) Database. The Gene Ontology Database describes genes in terms of their associated biological processes, molecular functions and cellular components. Using GOstat, the genes shown in the clusters were tested for their statistical association with GO terms [67]. The program identifies Gene Ontology terms for which genes in the list were overrepresented. For each GO term, a p-value was calculated indicating the probability that the observed counts could have resulted from randomly distributing the associated GO terms between our genes and modules. GOstat corrects for multiple comparisons by employing a false discovery rate, $p < .05$.

Supporting Information

S1 Table. Genes and variants from zygosity testing (Exome library).
(DOCX)

S2 Table. Genes and variants from zygosity testing (Haplotype library).
(DOCX)

S3 Table. Gene ontology terms from each significant module. Genes from each module were used in GO to describe biological functions.16 Each biological function and module number is shown for the nine clusters. Modules are shown as large solid figures. Connections ("edges") between biological functions in different clusters are shown.
(DOCX)

Acknowledgments

This work was supported by grants from the National Foundation March of Dimes #21-FY14-154, the Rhode Island Foundation #20133978, and the National Institutes of Health #P30 GM114750-01, #P20 RR18728 and #P30 GM103410.

Author Contributions

Conceived and designed the experiments: JP AU BM CS. Performed the experiments: JP AU BM CS. Analyzed the data: JP AU JS AD. Contributed reagents/materials/analysis tools: CS. Wrote the paper: JP AU JS AD.

References

1. Ventura SJ, Hamilton BE, Mathews TJ, Centers for Disease C, Prevention. Pregnancy and childbirth among females aged 10–19 years—United States, 2007–2010. Morbidity and mortality weekly report Surveillance summaries. 2013; 62 Suppl 3:71–6. Epub 2013/11/23. PMID: 24264493.
2. Adams MM, Elam-Evans LD, Wilson HG, Gilbertz DA. Rates of and factors associated with recurrence of preterm delivery. Jama. 2000; 283(12):1591–6. Epub 2000/03/29. PMID: 10735396.
3. Ekwo E, Moawad A. The risk for recurrence of premature births to African-American and white women. Journal of the Association for Academic Minority Physicians: the official publication of the Association for Academic Minority Physicians. 1998; 9(1):16–21. Epub 1998/05/20. PMID: 9585671.
4. Muglia LJ, Katz M. The enigma of spontaneous preterm birth. N Engl J Med. 2010; 362(6):529–35. Epub 2010/02/12. 362/6/529 [pii] doi: 10.1056/NEJMra0904308 PMID: 20147718.
5. Allen MC, Alexander GR, Tompkins ME, Hulsey TC. Racial differences in temporal changes in newborn viability and survival by gestational age. Paediatric and perinatal epidemiology. 2000; 14(2):152–8. Epub 2000/05/03. PMID: 10791659.
6. Copper RL, Goldenberg RL, Das A, Elder N, Swain M, Norman G, et al. The preterm prediction study: maternal stress is associated with spontaneous preterm birth at less than thirty-five weeks' gestation. National Institute of Child Health and Human Development Maternal-Fetal Medicine Units Network. American journal of obstetrics and gynecology. 1996; 175(5):1286–92. Epub 1996/11/01. PMID: 8942502.
7. Iams JD, Goldenberg RL, Mercer BM, Moawad A, Thom E, Meis PJ, et al. The Preterm Prediction Study: recurrence risk of spontaneous preterm birth. National Institute of Child Health and Human Development Maternal-Fetal Medicine Units Network. American journal of obstetrics and gynecology. 1998; 178(5):1035–40. Epub 1998/06/03. PMID: 9609580.
8. Mercer BM, Goldenberg RL, Das A, Moawad AH, Iams JD, Meis PJ, et al. The preterm prediction study: a clinical risk assessment system. American journal of obstetrics and gynecology. 1996; 174(6):1885–93; discussion 93–5. Epub 1996/06/01. PMID: 8678155.
9. Mercer BM, Goldenberg RL, Meis PJ, Moawad AH, Shellhaas C, Das A, et al. The Preterm Prediction Study: prediction of preterm premature rupture of membranes through clinical findings and ancillary testing. The National Institute of Child Health and Human Development Maternal-Fetal Medicine Units Network. American journal of obstetrics and gynecology. 2000; 183(3):738–45. Epub 2000/09/19. PMID: 10992202.
10. Rolett A, Kiely JL. Maternal sociodemographic characteristics as risk factors for preterm birth in twins versus singletons. Paediatric and perinatal epidemiology. 2000; 14(3):211–8. Epub 2000/08/19. PMID: 10949212.
11. Adams KM, Eschenbach DA. The genetic contribution towards preterm delivery. Semin Fetal Neonatal Med. 2004; 9(6):445–52. Epub 2005/02/05. S1084275604000302 [pii] doi: 10.1016/j.siny.2004.04.001 PMID: 15691782.
12. Crider KS, Whitehead N, Buus RM. Genetic variation associated with preterm birth: a HuGE review. Genet Med. 2005; 7(9):593–604. Epub 2005/11/23. 00125817-200511000-00001 [pii]. PMID: 16301860.
13. Menon R, Fortunato SJ, Thorsen P, Williams S. Genetic associations in preterm birth: a primer of marker selection, study design, and data analysis. J Soc Gynecol Investig. 2006; 13(8):531–41. Epub 2006/11/08. S1071-5576(06)01499-7 [pii] doi: 10.1016/j.jsgi.2006.09.006 PMID: 17088082.
14. Pennell CE, Jacobsson B, Williams SM, Buus RM, Muglia LJ, Dolan SM, et al. Genetic epidemiologic studies of preterm birth: guidelines for research. American journal of obstetrics and gynecology. 2007;

196(2):107–18. Epub 2007/02/20. S0002-9378(06)02475-6 [pii] doi: 10.1016/j.ajog.2006.03.109 PMID: 17306646.

15. Plunkett J, Muglia LJ. Genetic contributions to preterm birth: implications from epidemiological and genetic association studies. Ann Med. 2008; 40(3):167–95. Epub 2008/04/03. 791840773 [pii] doi: 10.1080/07853890701806181 PMID: 18382883.

16. Romero R, Espinoza J, Gotsch F, Kusanovic JP, Friel LA, Erez O, et al. The use of high-dimensional biology (genomics, transcriptomics, proteomics, and metabolomics) to understand the preterm parturition syndrome. BJOG: an international journal of obstetrics and gynaecology. 2006; 113 Suppl 3:118–35. Epub 2007/01/09. BJO1150 [pii] doi: 10.1111/j.1471-0528.2006.01150.x PMID: 17206980.

17. Weinberg CR, Shi M. The Genetics of Preterm Birth: Using What We Know to Design Better Association Studies. American journal of epidemiology. 2009; 170(11):1373–81. doi: 10.1093/aje/kwp325 PMID: 19854804

18. Clausson B, Lichtenstein P, Cnattingius S. Genetic influence on birthweight and gestational length determined by studies in offspring of twins. BJOG: an international journal of obstetrics and gynaecology. 2000; 107(3):375–81. Epub 2000/03/31. PMID: 10740335.

19. Treloar SA, Macones GA, Mitchell LE, Martin NG. Genetic influences on premature parturition in an Australian twin sample. Twin Res. 2000; 3(2):80–2. Epub 2000/08/06. PMID: 10918619.

20. Johnstone F, Inglis L. Familial trends in low birth weight. Br Med J. 1974; 3(5932):659–61. Epub 1974/09/14. PMID: 4425792; PubMed Central PMCID: PMC1611690.

21. Ward K, Argyle V, Meade M, Nelson L. The heritability of preterm delivery. Obstetrics and gynecology. 2005; 106(6):1235–9. Epub 2005/12/02. 106/6/1235 [pii] doi: 10.1097/01.AOG.0000189091.35982.85 PMID: 16319246.

22. Boyd HA, Poulsen G, Wohlfahrt J, Murray JC, Feenstra B, Melbye M. Maternal Contributions to Preterm Delivery. American journal of epidemiology. 2009; 170(11):1358–64. doi: 10.1093/aje/kwp324 PMID: 19854807

23. Little J. Invited Commentary: Maternal Effects in Preterm Birth—Effects of Maternal Genotype, Mitochondrial DNA, Imprinting, or Environment? American journal of epidemiology. 2009; 170(11):1382–5. doi: 10.1093/aje/kwp326 PMID: 19854805

24. Svensson AC, Sandin S, Cnattingius S, Reilly M, Pawitan Y, Hultman CM, et al. Maternal effects for preterm birth: a genetic epidemiologic study of 630,000 families. American journal of epidemiology. 2009; 170(11):1365–72. Epub 2009/10/27. kwp328 [pii] doi: 10.1093/aje/kwp328 PMID: 19854802.

25. Aidoo M, McElroy PD, Kolczak MS, Terlouw DJ, ter Kuile FO, Nahlen B, et al. Tumor necrosis factor-alpha promoter variant 2 (TNF2) is associated with pre-term delivery, infant mortality, and malaria morbidity in western Kenya: Asembo Bay Cohort Project IX. Genet Epidemiol. 2001; 21(3):201–11. Epub 2001/10/23. doi: 10.1002/gepi.1029 [pii] 10.1002/gepi.1029. PMID: 11668577.

26. Roberts AK, Monzon-Bordonaba F, Van Deerlin PG, Holder J, Macones GA, Morgan MA, et al. Association of polymorphism within the promoter of the tumor necrosis factor alpha gene with increased risk of preterm premature rupture of the fetal membranes. American journal of obstetrics and gynecology. 1999; 180(5):1297–302. Epub 1999/05/18. S0002937899003014 [pii]. PMID: 10329893.

27. Bezold KY, Karjalainen MK, Hallman M, Teramo K, Muglia LJ. The genomics of preterm birth: from animal models to human studies. Genome medicine. 2013; 5(4):34. Epub 2013/05/16. doi: 10.1186/gm438 PMID: 23673148; PubMed Central PMCID: PMC3707062.

28. Fujimoto T, Parry S, Urbanek M, Sammel M, Macones G, Kuivaniemi H, 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. 2002; 277(8):6296–302. Epub 2001/12/14. doi: 10.1074/jbc.M107865200 M107865200 [pii]. PMID: 11741975.

29. Genc MR, Gerber S, Nesin M, Witkin SS. Polymorphism in the interleukin-1 gene complex and spontaneous preterm delivery. American journal of obstetrics and gynecology. 2002; 187(1):157–63. Epub 2002/07/13. S0002937802000972 [pii]. PMID: 12114904.

30. Gibson G. Hints of hidden heritability in GWAS. Nature genetics. 2010; 42(7):558–60. Epub 2010/06/29. ng0710-558 [pii] doi: 10.1038/ng0710-558 PMID: 20581876.

31. Kalish RB, Vardhana S, Gupta M, Perni SC, Witkin SS. Interleukin-4 and -10 gene polymorphisms and spontaneous preterm birth in multifetal gestations. American journal of obstetrics and gynecology. 2004; 190(3):702–6. Epub 2004/03/26. doi: 10.1016/j.ajog.2003.09.066 S0002937803019343 [pii]. PMID: 15042002.

32. Landau R, Xie HG, Dishy V, Stein CM, Wood AJ, Emala CW, et al. beta2-Adrenergic receptor genotype and preterm delivery. American journal of obstetrics and gynecology. 2002; 187(5):1294–8. Epub 2002/11/20. S000293780200412X [pii]. PMID: 12439523.

33. Lorenz E, Hallman M, Marttila R, Haataja R, Schwartz DA. Association between the Asp299Gly polymorphisms in the Toll-like receptor 4 and premature births in the Finnish population. Pediatr Res. 2002; 52(3):373–6. Epub 2002/08/24. PMID: 12193670.

34. Ozkur M, Dogulu F, Ozkur A, Gokmen B, Inaloz SS, Aynacioglu AS. Association of the Gln27Glu polymorphism of the beta-2-adrenergic receptor with preterm labor. International journal of gynaecology and obstetrics: the official organ of the International Federation of Gynaecology and Obstetrics. 2002; 77(3):209–15. Epub 2002/06/18. S0020729202000358 [pii]. PMID: 12065131.

35. Papazoglou D, Galazios G, Koukourakis MI, Kontomanolis EN, Maltezos E. Association of -634G/C and 936C/T polymorphisms of the vascular endothelial growth factor with spontaneous preterm delivery. Acta Obstet Gynecol Scand. 2004; 83(5):461–5. Epub 2004/04/03. doi: 10.1111/j.0001-6349.2004.00403.x AOG403 [pii]. PMID: 15059159.

36. Simhan HN, Krohn MA, Roberts JM, Zeevi A, Caritis SN. Interleukin-6 promoter -174 polymorphism and spontaneous preterm birth. American journal of obstetrics and gynecology. 2003; 189(4):915–8. Epub 2003/10/31. S0002937803008433 [pii]. PMID: 14586325.

37. Witkin SS, Vardhana S, Yih M, Doh K, Bongiovanni AM, Gerber S. 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. American journal of obstetrics and gynecology. 2003; 189(5):1413–7. Epub 2003/11/25. S0002937803006306 [pii]. PMID: 14634579.

38. Nadeau-Vallee M, Quiniou C, Palacios J, Hou X, Erfani A, Madaan A, et al. Novel Noncompetitive IL-1 Receptor-Biased Ligand Prevents Infection- and Inflammation-Induced Preterm Birth. Journal of immunology. 2015; 195(7):3402–15. Epub 2015/08/26. doi: 10.4049/jimmunol.1500758 PMID: 26304990.

39. Haataja R, Karjalainen MK, Luukkonen A, Teramo K, Puttonen H, Ojaniemi M, et al. Mapping a new spontaneous preterm birth susceptibility gene, IGF1R, using linkage, haplotype sharing, and association analysis. PLoS Genet. 2011; 7(2):e1001293. Epub 2011/02/10. doi: 10.1371/journal.pgen.1001293 PMID: 21304894; PubMed Central PMCID: PMC3033387.

40. Uzun A, Laliberte A, Parker J, Andrew C, Winterrowd E, Sharma S, et al. dbPTB: a database for preterm birth. Database: the journal of biological databases and curation. 2012; 2012:bar069. Epub 2012/02/11. doi: 10.1093/database/bar069 PMID: 22323062; PubMed Central PMCID: PMC3275764.

41. Uzun A, Dewan AT, Istrail S, Padbury JF. Pathway-based genetic analysis of preterm birth. Genomics. 2013; 101(3):163–70. Epub 2013/01/10. doi: 10.1016/j.ygeno.2012.12.005 PMID: 23298525; PubMed Central PMCID: PMC3570639.

42. Gusenleitner D, Howe EA, Bentink S, Quackenbush J, Culhane AC. iBBiG: iterative binary bi-clustering of gene sets. Bioinformatics. 2012; 28(19):2484–92. Epub 2012/07/14. doi: 10.1093/bioinformatics/bts438 PMID: 22789589; PubMed Central PMCID: PMC3463116.

43. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America. 2005; 102(43):15545–50. Epub 2005/10/04. doi: 10.1073/pnas.0506580102 PMID: 16199517; PubMed Central PMCID: PMC1239896.

44. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics. 2000; 25(1):25–9. Epub 2000/05/10. doi: 10.1038/75556 PMID: 10802651; PubMed Central PMCID: PMC3037419.

45. Velez DR, Fortunato SJ, Thorsen P, Lombardi SJ, Williams SM, Menon R. Preterm birth in Caucasians is associated with coagulation and inflammation pathway gene variants. PloS one. 2008; 3(9):e3283. Epub 2008/09/27. doi: 10.1371/journal.pone.0003283 PMID: 18818748; PubMed Central PMCID: PMC2553267.

46. Demendi C, Borzsonyi B, Nagy ZB, Rigo J Jr., Pajor A, Joo JG. Gene expression patterns of insulin-like growth factor 1, 2 (IGF-1, IGF-2) and insulin-like growth factor binding protein 3 (IGFBP-3) in human placenta from preterm deliveries: influence of additional factors. European journal of obstetrics, gynecology, and reproductive biology. 2012; 160(1):40–4. Epub 2011/11/11. doi: 10.1016/j.ejogrb.2011.10.005 PMID: 22071113.

47. Gustafsson M, Nestor CE, Zhang H, Barabasi AL, Baranzini S, Brunak S, et al. Modules, networks and systems medicine for understanding disease and aiding diagnosis. Genome medicine. 2014; 6(10):82. Epub 2014/12/05. doi: 10.1186/s13073-014-0082-6 PMID: 25473422; PubMed Central PMCID: PMC4254417.

48. Litchfield K, Summersgill B, Yost S, Sultana R, Labreche K, Dudakia D, et al. Whole-exome sequencing reveals the mutational spectrum of testicular germ cell tumours. Nature communications. 2015; 6:5973. Epub 2015/01/23. doi: 10.1038/ncomms6973 PMID: 25609015; PubMed Central PMCID: PMC4338546.

49. Ng SB, Nickerson DA, Bamshad MJ, Shendure J. Massively parallel sequencing and rare disease. Human molecular genetics. 2010; 19(R2):R119–24. Epub 2010/09/18. doi: 10.1093/hmg/ddq390 PMID: 20846941; PubMed Central PMCID: PMC2953741.

50. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. 2012; 337 (6090):64–9. Epub 2012/05/19. doi: 10.1126/science.1219240 PMID: 22604720; PubMed Central PMCID: PMC3708544.

51. Dorschner MO, Amendola LM, Turner EH, Robertson PD, Shirts BH, Gallego CJ, et al. Actionable, pathogenic incidental findings in 1,000 participants' exomes. American journal of human genetics. 2013; 93(4):631–40. Epub 2013/09/24. doi: 10.1016/j.ajhg.2013.08.006 PMID: 24055113; PubMed Central PMCID: PMC3791261.

52. Gordon AS, Tabor HK, Johnson AD, Snively BM, Assimes TL, Auer PL, et al. Quantifying rare, deleterious variation in 12 human cytochrome P450 drug-metabolism genes in a large-scale exome dataset. Human molecular genetics. 2014; 23(8):1957–63. Epub 2013/11/28. doi: 10.1093/hmg/ddt588 PMID: 24282029; PubMed Central PMCID: PMC3959810.

53. Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nature reviews Genetics. 2011; 12(1):56–68. Epub 2010/12/18. doi: 10.1038/nrg2918 PMID: 21164525; PubMed Central PMCID: PMC3140052.

54. MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, et al. Guidelines for investigating causality of sequence variants in human disease. Nature. 2014; 508(7497):469–76. Epub 2014/04/25. doi: 10.1038/nature13127 PMID: 24759409; PubMed Central PMCID: PMC4180223.

55. Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, et al. Searching for missing heritability: designing rare variant association studies. Proceedings of the National Academy of Sciences of the United States of America. 2014; 111(4):E455–64. Epub 2014/01/21. doi: 10.1073/pnas.1322563111 PMID: 24443550; PubMed Central PMCID: PMC3910587.

56. Gonzaga-Jauregui C, Lupski JR, Gibbs RA. Human genome sequencing in health and disease. Annual review of medicine. 2012; 63:35–61. Epub 2012/01/18. doi: 10.1146/annurev-med-051010-162644 PMID: 22248320; PubMed Central PMCID: PMC3656720.

57. Gracia-Aznarez FJ, Fernandez V, Pita G, Peterlongo P, Dominguez O, de la Hoya M, et al. Whole exome sequencing suggests much of non-BRCA1/BRCA2 familial breast cancer is due to moderate and low penetrance susceptibility alleles. PloS one. 2013; 8(2):e55681. Epub 2013/02/15. doi: 10.1371/journal.pone.0055681 PMID: 23409019; PubMed Central PMCID: PMC3568132.

58. He Z, O'Roak BJ, Smith JD, Wang G, Hooker S, Santos-Cortez RL, et al. Rare-variant extensions of the transmission disequilibrium test: application to autism exome sequence data. American journal of human genetics. 2014; 94(1):33–46. Epub 2013/12/24. doi: 10.1016/j.ajhg.2013.11.021 PMID: 24360806; PubMed Central PMCID: PMC3882934.

59. Nelson MR, Wegmann D, Ehm MG, Kessner D, Jean P St, Verzilli C, et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science. 2012; 337(6090):100–4. Epub 2012/05/19. doi: 10.1126/science.1217876 PMID: 22604722; PubMed Central PMCID: PMC4319976.

60. Steinberg KM, Yu B, Koboldt DC, Mardis ER, Pamphlett R. Exome sequencing of case-unaffected-parents trios reveals recessive and de novo genetic variants in sporadic ALS. Scientific reports. 2015; 5:9124. Epub 2015/03/17. doi: 10.1038/srep09124 PMID: 25773295; PubMed Central PMCID: PMC4360641.

61. Romano JD, Tharp WG, Sarkar IN. Adapting simultaneous analysis phylogenomic techniques to study complex disease gene relationships. Journal of biomedical informatics. 2015; 54:10–38. Epub 2015/01/17. doi: 10.1016/j.jbi.2015.01.002 PMID: 25592479.

62. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25(14):1754–60. Epub 2009/05/20. doi: 10.1093/bioinformatics/btp324 PMID: 19451168; PubMed Central PMCID: PMC2705234.

63. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25(16):2078–9. Epub 2009/06/10. doi: 10.1093/bioinformatics/btp352 PMID: 19505943; PubMed Central PMCID: PMC2723002.

64. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research. 2010; 20(9):1297–303. Epub 2010/07/21. doi: 10.1101/gr.107524.110 PMID: 20644199; PubMed Central PMCID: PMC2928508.

65. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Current protocols in bioinformatics / editoral board, Baxevanis Andreas D [et al]. 2013; 11(1110):11 0 1–0 33. Epub

2014/11/29. doi: 10.1002/0471250953.bi1110s43 PMID: 25431634; PubMed Central PMCID: PMC4243306.

66. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research. 2010; 38(16):e164. Epub 2010/07/06. doi: 10.1093/nar/gkq603 PMID: 20601685; PubMed Central PMCID: PMC2938201.

67. Gene Ontology C. Gene Ontology Consortium: going forward. Nucleic acids research. 2015; 43(Database issue):D1049–56. Epub 2014/11/28. doi: 10.1093/nar/gku1179 PMID: 25428369; PubMed Central PMCID: PMC4383973.