GENOMIC AND MOLECULAR CHARACTERIZATION OF PRETERM BIRTH - SUPPORTING INFORMATION

Theo A. Knijnenburg[1*], Joseph G. Vockley[2*], Nyasha Chambwe[1*], David Gibbs[1*], Crystal Humphries[1], Kathi Huddleston[2], Lisa Klein[2], Prachi Kothiyal[2], Ryan Tasseff[1], Varsha Dhankani[1], Dale Bodian[2], Wendy S. W. Wong[2], Gustavo Glusman[1], Denise E. Mauldin[1], Michael Miller[1], Joseph Slagel[1], Summer Elasady[1], Jared C. Roach[1], Roger Kramer[1], Kalle Leinonen[1], Jasper Linthorst[1], Raj Baveja[3], Robin Baker[3], Benjamin D. Solomon[2], Greg Eley[2], Ram Iyer[2], Larry Maxwell[2], Brady Bernard[1], Ilya Shmulevich[1], Leroy Hood[1#] and John E. Niederhuber[2,4,5#]

[1] Institute for Systems Biology, 401 Terry Ave. North, Seattle, WA, 98109.

[2] Inova Translational Medicine Institute, Inova Health System and Inova Fairfax Medical Center, 3300 Gallows Road, Falls Church, VA, 22042.

[3] Fairfax Neonatal Associates, Inova Children's Hospital, 3300 Gallows Road, Falls Church, VA, 22042.

[4] Adjunct Professor, Johns Hopkins University School of Medicine, Baltimore, MD 21287.

[5] Professor, Department of Public Health Sciences, University of Virginia School of Medicine

*equal contribution

[#]corresponding authors (lhood@systemsbiology.org, john.niederhuber@inova.org)

www.pnas.org/cgi/doi/10.1073/pnas.1716314116

TABLE OF CONTENTS

METHODS

COHORT AND CLINICAL DATA

STUDY POPULATION

The study population reported in this analysis consisted of 791 family trios (father, mother and newborn). All families were accessioned at the time of delivery at the Inova Fairfax Medical Center, Falls Church, VA, from 2011-2013. Participating families were enrolled in the Inova Translational Medicine Institute (ITMI) research protocol “Molecular Study of Preterm Birth”. Study participants provided written informed consent for use of their clinical and genomic data for research purposes. The study was approved by the Institutional Review Board of the Inova Health System and the Western Institutional Review Board (Protocol #1124761). All patient data and information was anonymized and de-identified prior to analysis.

Clinical information concerning pregnancy, delivery, and newborn status were obtained through (i) self-reported data using a study-specific case report form (CRF) administered by interview and (ii) from electronic medical records (EMR). Data from the CRFs and medical records of the mothers, fathers and newborns were combined into 167 clinical elements pertaining to the mother, 63 to the father and 76 clinical elements related to the newborn. The study cohort is representative of the population giving birth at Inova Fairfax Medical Center and of the diversity of the population of Northern Virginia in terms of race, ethnicity and socioeconomic status.

ADMIXTURE ANALYSIS

In order to assign each mother and father to a population, we employed WGS data to estimate admixture coefficients using the glu-genetics ‘struct.admix’ method (https://github.com/bioinformed/glu-genetics/blob/master/doc/modules/struct/admix.rst) based on data from the 1000 Genomes Project for reference population definitions(1). The maternal cohort was composed of a range of ancestries including African (AFA, n=43), Asian (EAS, n=75), European (EUR, n=377), Ad Mixed American (AMR, n=127), and mixed (n=169). Specifically, for each mother we determined the four admixture coefficients (EUR, AFA, EAS and AMR). If the admixture coefficient for one population was greater than 0.8, the mother was labelled as being a part of that population, otherwise the label was ‘Mixed’.

CLINICAL DATA TRANSFORMATION

A series of transformation and aggregation steps were necessary to prepare and utilize the clinical CRF and EMR data in subsequent analyses. All clinical data were transformed to be compliant with The Health Insurance Portability and Accountability Act of 1996 (HIPAA).

PTB RELATED CLINICAL PHENOTYPES

We defined nine PTB related phenotypes, which allowed us to divide the samples into cases and controls for each of these phenotypes. See **Figure 1B** and **S1**. The nine PTB related phenotypes were gathered from the EMR and CRF data.

Three of the phenotypes were based on the gestational age as follows:

1. Very early preterm if gestation was less than 28 weeks (VEPTB)
2. Early preterm if gestation was less than 34 weeks (EPTB)
3. Preterm if gestation was less than 37 weeks (PTB)

As a control set, we considered all full term births, i.e. gestation was 37 or more weeks and less than 42 weeks. Additional phenotypes include PROM (premature rupture of membrane), Pre-eclampsia and Idiopathic PTB. Idiopathic PTB was defined as delivery without evidence of infection, no history of cervical incompetence or insufficiency, no evidence of uterine structural abnormalities, no evidence of placentation abnormalities, and absence of any maternal disease(s) known to cause premature delivery. Finally, we created PTB related phenotypes involving issues with pregnancy related maternal organs, i.e. Placenta-related (mostly placental previa/abruption), Uterine-related (mostly uterine anomalies and endometriosis) and Cervix-related (those requiring placement of cervical cerclage/incompetent shortened cervix).

SEQUENCING DATA AND GENOMIC TESTS

SAMPLE COLLECTION, PREPARATION AND SEQUENCING

BD VacutainerR K2-EDTA tubes were used to collect whole blood samples from mother and father. Newborn samples were collected at the time of first heel stick to obtain blood spots for the Commonwealth of Virginia Commonwealth newborn screening program or when diagnostic laboratory studies, such as blood chemistries or complete blood count, were clinically indicated. Specimens were immediately transported to ITMI's specimen processing laboratory for initial processing and for biobanking under liquid nitrogen vapor. DNA samples were prepared using a QiaSymphonyR automated DNA extractor, QIAGEN Inc., Valentia, CA. Batches of samples were sent to Complete Genomics Inc., Mountain View, CA. for whole genome sequencing, assembly and variant calling.

Genome sequences were assembled using Complete Genomics Assembly Pipeline (CGA) versions 2.0.0-2.0.4 and aligned to the GRCh37 human reference genome. The data were stored in the Amazon Web Services' storage cloud and included raw sequence reads, read mappings, coverage, variant calls with annotations and summary statistics. Coverage statistics were calculated using weight-sum

sequence coverage depth, breadth and uniformity. Sex concordance and consistency with Mendelian genetics were determined. The variant data were annotated using available databases, including Online Mendelian Inheritance in Man (OMIM), dbSNP, Catalogue of Somatic Mutations in Cancer (COSMIC) and Polyphen2, and stored in Variant Call Format (VCF) 4.1. On average, each genome had >40x coverage over the >70% of the genome and >80% of the exome. Variants from the masterVar files from all genomes were merged into a single VCF 4.1 file using mkvcf (beta) from the CGA™ tools suite, version 1.6.0. All analyses were based on human genome version hg19/GRCh37.

GENOMIC COHORT

Whole genome sequences were successfully collected on 2434 individuals. The genomes consisted of 791 mothers, 791 fathers and 839 newborns. In total there were 784 complete family trios, i.e. the genomes from the mother, father and newborn were successfully sequenced. There were 63 families with twins (in most cases both newborns were sequenced) and one family with triplets. For analysis, only one newborn was (randomly) picked from each multiple birth familiy. See **Dataset S1**.

PREPROCESSING OF GENOMIC DATA FOR VARIANT ANALYSIS

The called genome variants in the original merged VCF file, 196,136,835 in total, were filtered to remove low-quality, multiallelic, commonly mutated or misannotated, and extremely heterozygous sites. Commonly mutated or misannotated segments (CMS) are those in which we observed an enrichment of rare (or even private) detrimental annotated variants. Such enrichment is suggestive of some genomic and/or technical artifact, such as an annotated exon that is not actually transcribed. Extremely heterozygous (exHet) sites are those observed as heterozygous in the vast majority of the genomes. This deviation from Hardy-Weinberg equilibrium cannot be biologically true. Explanations for these artifacts include ‘genomic compressions’ (present in the reference sequence at lower copy number than in real genomes), low-quality calls (annotated by CGA as VQLOW), calls annotated as Mendelian inheritance errors (MIEs), and hemizygous calls were considered as missing data. Additionally, we filtered variants such that only those with call-rate greater than 0.90 and minor-allele frequency (MAF) greater than 0.01 remained. After filtering, the total number of variants remaining was 6,987,906. See **Dataset S1**.

BATCH CORRECTION

Given the extended period of time in which the sequencing took place, we expected some level of batch effects due to technical changes and variability within the sequencing pipeline. We defined two ‘batch variables’, one based on the Complete Genomics Assembly Pipeline version, which was either 2.0.0, 2.0.1, 2,0.2, 2.0.3 or 2.0.4, and the other based on the shipping date of the sample (a continuous valued variable of the format yyyymmdd).

We performed simple pairwise statistical tests, where we compared the genotype of each variant (a categorical variable indicating homozygous reference allele, heterozygous alternative or homozygous alternative) with 1) the pipeline version using a (categorical-categorical) Chi-square test was applied, and 2) the shipping dates using a (categorical-numerical) non-metric Kruskal-Wallis test. The tests were performed for the maternal, paternal and neonatal genomes separately. Significant results were called at an FDR<10%. Any variant significantly associated with batch in at least one of these six tests (2 batch variables x the three sample groups: mothers, fathers, newborns) was discarded for further analysis. In total 1,223,174 (17.5% of 6,987,906) variants were filtered out based on these statistical tests.

GENOMIC STATISTICAL TESTS

EIGENSTRAT

We performed the EIGENSTRAT(2) statistical genomic association test for each of the nine phenotypes and each of the three sample groups; mothers, fathers and newborns, separately. Using EIGENSTRAT's smartpca function, we generated ten principal components and used these to adjust the phenotype and the genotype. P-values were computed using the chi-squared distribution with one degree of freedom. To avoid inflated p-values due to small sample sizes, we did not test variants that had less than ten samples with non-reference (homozygous or heterozygous) allele calls in both the cases and controls. (For variants, where our cohort had more non-reference than reference alleles, we switched reference and non-reference.)

We observed that phenotypes with a smaller number of cases had more associated variants, indicating that in those scenarios, the statistical tests might be overly optimistic (**Figure S4**).

FBAT

We performed the Family Based Association Test (FBAT)(3) statistical genomic association test for each of the nine phenotypes separately. FBAT was run for the 784 complete family trios under the additive genetic model. To avoid inflated p-values due to small sample sizes, we did not test variants with a minor allele frequency in our cohort smaller than 5% (MAF>5%).

MAPPING VARIANTS TO GENES

The variant data were annotated using Oncotator (v1.9.9.0) on the hg19 genome build with annotation data sources (version: oncotator_v1_ds_April052016) (4). Variant classification was based on Gencode v19 canonical transcript annotations. We report a subset of these annotations: variant class, mutation consequence, predicted pathogenicity and allele frequencies from reference population datasets, 1000

Genomes (1000gp3 20130502), NHLBI GO Exome Sequencing Project (ESP) and the Exome Aggregation Consortium (ExAC) (version 0.3.1) (**Dataset S2**).

DISTRIBUTION OF ASSOCIATED VARIANTS ACROSS THE GENOME

Variants associated with PTB-related phenotypes were primarily single nucleotide polymorphisms (SNPs) and mostly in intronic or intergenic regions (**Figure S11**, **Dataset S21**). Variants associated with VEPTB identified using FBAT were in intronic or intergenic regions while some RNA-associated variants were identified by EIGENSTRAT in the maternal and newborn genomes (**Figure S11**, **Dataset S21**)

RNA-SEQ AND DNA METHYLATION DATA AND DIFFERENTIAL EXPRESSION TESTS

MRNA AND RNA-SEQUENCING: RNA EXTRACTION, LIBRARY PREPARATION, SEQUENCING AND DATA PREPROCESSING

ALPHA AND BETA MRNA REDUCTION

Globin mRNA is substantially depleted from total RNA samples using the GlobinClear-Human Kit (Life Technologies # AM1980), as described by the vendor. Briefly, 1.25 µg of total RNA isolated from whole blood is combined with biotinylated capture oligonucleotides complementary to globin mRNAs and the mixture is incubated at 50°C for 15 minutes to allow for duplex formation. Streptavidin magnetic beads are added to each specimen, and the resulting mixture is incubated for an additional 30 minutes at 50°C to allow binding of the biotin moieties by Streptavidin. These complexes, comprising Streptavidin magnetic beads bound to biotinylated oligonucleotides that are specifically hybridized to globin mRNAs and are then captured using a magnet. The globin-depleted supernatant is transferred to a new container and further purified using RNA binding beads. The final globin mRNA-depleted RNA samples are quantitated by spectrophotometry using a NanoDrop ND-8000 spectrophotometer.

RNA-SEQ USING THE ILLUMINA MRNA STRANDED KIT

RNA samples were converted into cDNA libraries using the Illumina TruSeq Stranded mRNA sample preparation kit (Illumina # RS-122-2103). Briefly, total RNA samples are concentration normalized, and polyadenylated RNA is purified using oligo-dT attached to magnetic beads. Purified mRNA is fragmented using heat in the presence of divalent cations. Fragmented RNA is converted into double-stranded cDNA, with dUTP utilized in place of dTTP in the second strand master mix. A single 'A' base is added to the cDNA and forked adaptors that include index, or barcode, sequences are attached via ligation. The resulting molecules are amplified via polymerase chain reaction (PCR). Final libraries are quantified, normalized and pooled. Pooled libraries are bound to the surface of a flow cell and each

bound template molecule is clonally amplified up to 1000-fold to create individual clusters. Four fluorescently labeled nucleotides are then flowed over the surface of the flow cell and incorporated into each nucleic acid chain. Fluorescence is measured for each cluster during each cycle to identify the base that was added to each cluster. The dye is then enzymatically removed to allow incorporation of the next nucleotide during the next cycle.

MIRNA SAMPLE PREPARATION AND SEQUENCING

Total RNA samples were converted into indexed cDNA sequencing libraries using Illumina's TruSeq Small RNA sample preparation kit (Illumina # RS-200-0012). Starting with 1000 ng of total RNA, a single stranded adenylated DNA adapter was added to the 3' hydroxyl group using T4 RNA Ligase 2 deletion mutant. The T4 RNA ligase 2 deletion mutant prevents ligation of the adapter to the 5' end due to the absence of ATP. A 5' adapter was then added to the 5' phosphate using T4 RNA Ligase in the presence of ATP. Following adapter ligation, single stranded cDNA was created by a reverse transcription reaction. The cDNA was then PCR amplified using a common sequencing primer and an indexed primer that is unique to each sample. The cDNA libraries were analyzed for quality and fragment size ranges using the Agilent 2200 Tapestation (D1000 Screentape, Agilent # 5067-5582). The libraries were then size-selected, retaining fragments of between 125 - 160 bps, using BluePippin (3% cassettes, Sage Science # BDF3010), resulting in a mean library size of approximately 135 bps. The final libraries were then quantitated by qPCR (KAPA Library Quant Kit, KAPA Biosystems # KK4824), and normalized to 2 nM in preparation for sequencing.

RNA-SEQ DATA PRE-PROCESSING

A total of 684 samples were sequenced on the Illumina HiSeq2000 at Expression Analysis. The samples were aligned to the human reference genome (GRCh37) using BWA and raw counts of reads overlapping genes were calculated(5). Initially, 24,000 genes were detected and those with a read count of less than two in five or more samples were excluded from further analysis. This brought the total number of genes for analysis to 20,240 genes.

MIRNA-SEQ DATA PRE-PROCESSING

Small RNAs were sequenced for 766 samples on the Illumina HiSeq2000 at Expression Analysis. Fastq reads were clipped to prevent adapter contamination using Expression Analysis FASTQ processing utilities ea-utils; specifically fastq-mcf module was used for adapter trimming (6, 7). Trimmed reads were aligned to the human reference genome (GRCh37). Alignments that correspond to known miRNA loci are quantified using mirBase miRNA gene definition and reads were quantified for 234 microRNAs across these samples.

MIRNA TARGET IDENTIFICATION

We retrieved targets of miRNAs from publicly available databases using the multiMiR package in R(8), which allows simultaneous retrieval from multiple miRNA target databases. Targets of miRNAs are divided between high confidence miRNA targets that have been experimentally validated, and those that are predicted using a variety of computational approaches. We retrieved the set of validated gene targets from mirecords(9), mirtarbase(10) and tarbase(11) for the miRNAs measured in this study.

MIRNA EXPRESSION AND CORRELATION WITH TARGETS

Examination of miRNAs did not identify differentially expressed miRNAs for any of the clinical phenotypes; however, further analysis did identify significant correlation between miRNAs and predicted target genes that were differentially expressed in certain PTB phenotypes (**Dataset S22**).

DNA METHYLATION PROFILING BY ARRAY: LIBRARY PREPARATION, DATA PREPROCESSING AND GENE MAPPING

METHYLATION DATA PROCESSING

A total of 784 samples were analyzed on the Illumina Human Methylation 450K platform at Expression Analysis using standard protocols. Illumina's GenomeStudio software was used to read and process bead information. Probes were rejected if more than 2 samples had p-values > 0.05, indicating poor probe performance. Intra-sample normalization was performed using BMIQ(12). The high quality methylation probes found in Naeem *et al.*(13) were used to avoid the negative impacts of genomic features. Beta values were logit transformed to M-values, making the data approximately normal.

MAPPING PROBES TO GENES

The Illumina 450K Methylation probes were mapped to nearest genes using the FDb.InfiniumMethylation.hg19 Bioconductor annotation package (version 2.1.999).

MOLECULAR DATA ANALYSIS PIPELINE

The mRNA, miRNA-seq and DNA methylation data sets are broadly referred to as molecular data types. We applied a unified data QC and batch correction pipeline described below.

OUTLIER DETECTION

Outliers were determined using multivariate robust distance(14, 15), computed using minimum covariance determinant (MCD) as implemented in the robustbase R-package [https://cran.r-project.org/web/packages/robustbase/citation.html]. Robust estimates of location and scatter were made over a given number of repetitions, each time randomly sampling a set of features (methylation

probes, genes, or miRNAs). If the absolute value of the difference of robust median distance and robust deviation of distance was greater than the X^2 statistic (q=0.999, df=features, as used in (15)), then the individual sample warrants additional attention. This computation was computationally intensive, so a smaller number of features (genes or probes) were sampled. The number of features was selected by comparing different sizes over multiple runs. Using 31 features and 100 replicates gave reproducible results in terms of robust distance estimates.

After applying outlier detection, two sample outliers were detected in mRNA-seq and DNA methylation data and they were removed from subsequent analysis. No extreme outliers were detected in the small RNA-seq data; therefore, all samples were used for downstream analysis.

FILTERING BASED ON BLOOD DRAW DATE

For the majority of mothers (71%), whole blood samples were taken one day after birth. For 93% of the mothers, blood was drawn within the first four days after birth. Molecular data derived from samples taken more than four days after birth were not used in further analysis.

NORMALIZATION

We processed the Expression Analysis (EA) summarized raw count data by carrying out upper quartile normalization(16) as implemented in the edgeR Bioconductor package(17). DNA methylation data was normalized using BMIQ(12) and logit transformed from beta values to M-values.

BATCH EFFECT CORRECTION

Batch effects correlating with processing date were detected in the molecular data. mRNA-seq and miRNA-seq count data was transformed to continuous values using the Limma-Voom transformation(18). The Limma Batch effect removal method was used to correct for batches correlating with batch ID in the context of covariates of ancestry from ADMIXTURE(19) and maternal age.

DIFFERENTIAL EXPRESSION AND METHYLATION ANALYSIS

We developed a unified differential expression and methylation workflow to identify significant differences in the levels of expression or methylation using a generalized linear model framework implemented using the Limma package in R(20). Age, admixture (as calculated based on the genomic data), and blood draw dates were used as covariates in the analysis. In order to improve the robustness of the results, a 95% adjusted p-value interval (the 2.5% and 97.5% quantiles) was constructed by bootstrapping (sampling with replacement) the data using bootstrap 1000 samplings.

INTEGRATIVE GENOMIC AND MOLECULAR ANALYSIS

GENE LISTS FROM LITERATURE

We mined existing databases and the literature to identify published PTB and other pregnancy associated genes. **Dataset S4** presents an overview of the gene lists considered. Our main source for literature identified PTB genes was the Database for Preterm Birth (dbPTB) (accessed September 2014) that lists 640 genes associated with PTB(21). Other sources of PTB related genes included Capece *et al.*(22) who published a list for genes associated with preterm prelabor rupture of membranes (PPROM) (Capece_PROM 33 genes) and spontaneous preterm birth (sPTB) (Capece_sPTB 102 genes). Two doctoral dissertations by McElroy (McElroy_PTB)(23) and Plunkett (Plunkett_PTB)(24) summarized candidate gene studies in PTB and those gene lists were included in our analysis. We included a set of genes that are involved in human birth timing, and reported to evolve at an accelerated rate in the human lineage (Plunkett_HBT)(25). Finally we included a list of genes implicated in Pre-eclampsia from the Pre-eclampsia SNP Resource (PESNPdb)(26). We found that some of these data sources were largely redundant with respect to each other, and therefore used the four most encompassing gene lists (dbPTB, Plunkett_HBT, McElroy_PTB and PESNPdb) in subsequent enrichment analyses.

OVERLAP ANALYSIS

To determine whether the overlap between any two gene lists was statistically significant, we performed the hypergeometric test. The gene universe was determined by intersecting the full list of genes between the two platforms under consideration. Genes commonly represented were eligible for overlap analysis. Any comparisons at the genomic level defined the gene universe as the set of all protein coding genes (gencode annotation hg19 version 19).

E/M QTL ANALYSIS

For both the eQTL and mQTL analysis, we examined genes and methylation probes that were listed or overlapped those listed in the VEPTB candidate set of 72 genes. Correlation was examined with variants located within a window of 1Mb from the gene or probe. To perform this analysis, we used MatrixEQTL software which employs a linear regression model to detect eQTLs. We also factored in the mother’s age, admixture ratio, and blood draw date as covariates to remove possible confounders(27).

VARIANT COLLAPSING AND ENRICHMENT

Variant enrichment analysis was done to gauge the overlap between variants and genomic locations. This was done using all variants with an FDR<10%. To remove redundant variants, we used the Linkage-disequilibrium based SNP pruning software(28). The pruned variants were examined for overlap with genomic locations using bedtools(29) and compared against the background variants, the ~6 million variants that were tested for association. A hypergeometric test was performed on the variant location to determine whether the enrichment was statistically significant.

VEPTB CANDIDATE GENES

We defined a set of VEPTB candidate genes based on either strong association of variants in particular genes with the phenotype, or multiple lines of evidence at the transcriptional and epigenetic level. Specifically, if a variant part of a given gene had a p-value smaller than 10^{-8}, that gene was included in the candidate set. Alternatively, a gene was included in the candidate set if it was present in at least two of the following sets: differentially expressed (bootstrap FDR q-value < 10% and |log2(FC)| > 0.5), differentially methylated (bootstrap FDR q-value < 10%), weakly associated variants (FDR < 10%). The final VEPTB candidate gene list includes 72 genes (see **Dataset S12**).

PATHWAY ANALYSIS

The VEPTB candidate set of 72 genes was used as input to the ConsensusPathDB-human web service (CPDB)(30) Release 30 (09.01.2015). NetPath, Wikipathways, Reactome, and KEGG pathway databases were utilized, and hypergeometric tests were performed to look for over-representation within pathways. Pathways had to contain at least 2 genes to be tested. An FDR adjustment was made to the p-values to correct for multiple testing. Gene backgrounds are supplied by CPDB, which includes all genes that are implicated in at least one pathway; a total of 10,773 genes.

SUPPLEMENTARY NOTE 1 - VALIDATION ON INDEPENDENT COHORT

We employed an independent cohort of more than 1300 family trios to validate the statistically significant variants of the genome-wide tests in the original cohort.

VALIDATION DATASET

The validation cohort is part of ITMI's Longitudinal Childhood Genome Study (also known as the First 1,000 Days of Life and Beyond study). For this study, families were recruited prenatally, from 2012-2016. Follow up occurred at delivery at the Inova Fairfax Medical Center, VA. Participating families, with written informed consent, were enrolled by the Inova Translational Medicine Institute (ITMI) staff under the research protocol "Genomic Correlations to Childhood Health Outcomes" approved by the

Institutional Review Board of the Inova Health System (#15-1804) and the Western Institutional Review Board (#20120204). All patient data and information de-identified prior to analysis.

Although the validation cohort is similar to the original cohort in the sense that it is also comprised of sequenced family trios, some of which involved PTB, there are two main differences: First, the validation cohort was not 'enriched' for PTB, EPTB and VEPTB cases. This means that the case groups (PTB, EPTB or VEPTB) are much smaller than the control group (FTB). For example, of the 1330 sequenced newborns in the validation cohort, only 8 (0.6%) were born before 28 weeks of gestation (VEPTB), and 115 (8.6%) were before 37 weeks of gestation (PTB). See **Dataset S6** for an overview of the samples in the validation cohort. Second, the samples in the validation cohort were sequenced using the Illumina technology, whereas sequencing for the original cohort was done using Complete Genomics. See below for details.

SAMPLE COLLECTION, PREPARATION AND SEQUENCING FOR THE VALIDATION COHORT

Genome sequences were assembled using Illumina HiSeq and aligned to the GRCh37 human reference genome. The data were stored in the Amazon Web Services' storage cloud and included read mappings, coverage, variant calls with annotations and summary statistics. Further details on cohort demographics and sequencing information is described in (31). Coverage statistics were calculated using weight-sum sequence coverage depth, breadth and uniformity. Sex concordance and consistency with Mendelian genetics were determined. The variant data were annotated using available databases, including dbSNP, dbNSFP, snpEff, and ensembl, and stored in Variant Call Format (VCF) 4.1. On average, each genome had >40x coverage over the >70% of the genome and >80% of the exome. Variants from the masterVar files from all genomes were normalized using vtnorm and uploaded to Cloudera Impala version 2.5.0-cdh5.7.1.

STATISTICAL TESTS ON THE VALIDATION DATASET

Using EIGENSTRAT and FBAT we uncovered 1291 distinct variants in the original dataset at an FDR of 10% for the PTB, EPTB and VEPTB phenotypes for fathers and mothers (**Dataset S2**). We set out to validate these variants in the validation cohort. For this, we extracted the variant calls for the 1291 variants from the individual VCFs. No-calls, low quality calls and multi-allelic calls were annotated as missing values. For families with twins or triplets only one infant was selected (at random) to be part of the dataset. Then we applied three filters in the following order: First, we removed all variants with a MAF smaller than 5%, leading to the removal of 403 variants. Second, we removed all variants with excessive heterozygosity, i.e. 50% more heterozygous calls than could be expected based on the Hardy–Weinberg equilibrium, leading to removal of zero variants. Third, we removed nearby and correlated variants using PLINK VIF pruning(28) with standard parameters, leading to removal of 533 variants. In total, 355 variants were obtained after filtering and were used for statistical testing.

We applied two statistical tests: First, we applied the Cochran-Armitage chi-square test for trends in proportions in homozygous reference, heterozygous alternative and homozygous alternative alleles across the case and controls groups. This is the same test that underlies EIGENSTRAT, however we did not perform the ancestry correction using principal components analysis, which is performed in EIGENSTRAT before application of the Cochran-Armitage chi-square test. This is because we extracted only the 1291 variants for validation, whereas all (or many more) variants would be needed for the ancestry correction. Second, we performed a hypergeometric test. Here, we grouped heterozygous alternative and homozygous alternative alleles together, leading to two groups; samples that are homozygous reference and samples that have at least one alternative allele.

We tested 12 different scenarios, i.e. the three different phenotypes; PTB, EPTB and VEPTB (as cases with the FTB samples as controls) for four different sampling groups: the father, mothers, newborn and their union, i.e. all samples. The latter category is labeled as 'All' in **Dataset S7** and **Dataset S8**.

RESULTS

For the validation cohort, we identified 51 unique variants that are in or around 10 unique genes at a FDR threshold of 10% and a p-value threshold of 0.01 across the 12 different testing scenarios using the Cochran-Armitage chi-square test (**Dataset S7**, **Figure S3A**). The large majority of validated hits were found in the 'All' group, where we jointly analyzed the samples for the fathers, mothers and newborns. In the separate groups, we found fewer variants that met the statistical threshold, which may due to the small sample sizes of the case groups (PTB, EPTB and VEPTB). Noteworthy, for VEPTB we did find a significant exonic variant in *IL28RA* for all of the groups. *IL28RA* (Interleukin 28 Receptor, Alpha, aka *IFNLR1*, Interferon Lambda Receptor 1) belongs to the class II cytokine receptor family, and forms a receptor complex with interleukin 10 receptor, beta (*IL10RB*). Importantly, heterozygous and homozygous alternative alleles for *IL28RA* were found in the VEPTB group for multiple ancestries (based on self-reported race) (**Figure S3C,D**).

The hypergeometric test showed more conservative results, but *IL28RA* (*IFNLR1*) was still identified as a significant hit (**Figure S3B**, **Dataset S7, S8**).

ENRICHMENT OF PTB ASSOCIATED VARIANTS IN VEPTB GENES IN THE VALIDATION COHORT

We employed the same validation cohort to identify PTB associated variants in or around the 72 VEPTB related genes. Specifically, we studied 43,036 variants in or around (within 2kB) these genes. We applied the previously mentioned set of filters (MAF, excessive heterozygosity and pruning nearby, correlated variants). This resulted in 21,292 variants that were tested using the Cochran-Armitage chi-square test and the hypergeometric test. We compared these results with an identical statistical testing procedure applied to a set of 72 randomly selected genes. This set is defined in **Supplementary Note**

4 and found in **Dataset S12**. In the random set, we started with 42,800 variants that were filtered down to 20,023.

Once again, we observed that the Cochran-Armitage chi-square test showed more optimistic p-values than the hypergeometric test that turned out to be more conservative (**Figure S3E,F**). Yet, both test show an enrichment of statistical hits (low p-values) for the VEPTB genes compared to the background set. Using a Bonferroni corrected p-value of 0.05, we observed very few significant events for the father, mother and newborn cohorts, and almost all significant events in the 'All' group (**Dataset S8**).

DISCUSSION

Cohort size and ancestry as well as study design are important factors when studying a complex and heterogeneous phenotype. For example, in our study we had to discard 15% of the variants because of detected batch effects (**Supplementary Appendix - Methods - Sequencing data and genomic tests - Batch correction**). This could easily have resulted in missing previously identified genes. Additionally, in our validation cohort we were able to reproduce only a small number of variants, likely due to a different study design and sequencing technology. Specifically, the validation cohort was not 'enriched' for PTB, EPTB and VEPTB cases. This means that the case groups (PTB, EPTB or VEPTB) are much smaller than the control group (FTB). Also, the samples in the validation cohort were sequenced using the Illumina technology, whereas sequencing for the original cohort was done using Complete Genomics. In the PTB literature, it is not uncommon for reported biomarkers to fail to reproduce(23, 32, 33) and genes identified in this study showed only a marginal overlap with published biomarkers for PTB (**Supplementary Note 3**). From these observations we conclude that for identification of biomarkers for complex diseases it may be important to look beyond single associations between genomic loci and the phenotype.

SUPPLEMENTARY NOTE 2 - VEPTB PATHWAY GENE ANALYSIS

ENRICHMENT TESTS

We hypothesized that genes in the VEPTB pathways will be enriched for genetic variants, and differentially methylated or expressed genes associated with preterm birth and other clinical phenotypes. We carried out an extensive post-hoc pathway analysis in order to test this hypothesis.

DATASET

Based on integrative pathway analysis of the 72 VEPTB genes, we selected the top 27 pathways implicated in VEPTB by our integrative analysis (**Dataset S14**, criteria for inclusion FDR q-value < 5%). These pathways are summarized in **Figure S7,** along the themes of NOTCH signaling pathways,

immune/growth factor pathways and sugar metabolism pathways. We created a VEPTB candidate pathway gene set across these significant pathways excluding the 72 VEPTB genes used to infer pathway enrichment. This resulted in a total of 1,324 genes; we called this set the 'VEPTB pathway genes'.

METHODS

We tested for enrichment of variants in genes that were associated with clinical phenotypes using a hypergeometric test. The background gene set for this test was the set of all annotated genes excluding pseudogenes (ensemble gtf Homo_sapiens.GRCh38.93.gtf). Similarly, we tested the enrichment of differentially expressed and methylated genes in VEPTB pathway genes using a hypergeometric test. We considered all genes measured on the expression and methylation platforms as the background gene list to test for enrichment.

RESULTS

At the genomic level, we found statistically significant enrichment of variants that mapped to VEPTB pathway genes and were associated with preterm birth in parental genomes (**Dataset S23**). Similarly, we found statistically significant enrichment of genes that are differentially expressed or methylated in the more extreme preterm categories (EPTB and VEPTB) and in candidate PTB pathways compared to all other measured genes (**Dataset S23**).

In summary, in-depth analysis of genes in VEPTB candidate pathways supports our findings that these pathways are disrupted in preterm birth.

MACHINE LEARNING

In order to further investigate the associated pathways, we applied Random Forests(34) classification to predict status based on the VEPTB pathway genes. Specifically, the target (or class labels) for prediction is a binary vector across all samples (mothers), where 1 indicates VEPTB (the cases) and 0 indicates FTB (the controls)

Six sets of genes were used for prediction.

1.) The 72 VEPTB genes
2.) The 1,324 VEPTB pathway genes, i.e. the union of genes from the 72 VEPTB associated pathways merged excluding the 72 VEPTB genes.
3.) VEPTB pathway genes per pathway
4.) 72 randomly selected genes.
5.) 1,324 randomly selected genes

6.) Randomly selected genes, where the set size is identical to each of the pathways (excluding VEPTB genes).

The data types used are:

1.) DNA methylation data (Methylation probes were mapped to the closest gene as described in methylation data processing.)
2.) RNA-Seq gene expression data
3.) A joint data matrix of methylation and RNA-Seq

RF analysis was performed for each combination of gene set and data type using a 10-fold cross validation, where we roughly balanced the number of VEPTB and FTB samples in the test folds (see below). The random draws were repeated 100 times. 'Random' genes were drawn from the top 25% of most varying genes for the methylation and RNA-Seq data. In each case, we recorded the area under the ROC curve (AUC) as a measure of prediction performance. Perfect predictions have an AUC of 1, while random predictions have an AUC of 0.5.

The pseudo-algorithm is as follows:

For one of the 18 combinations of gene set, and data type:

1.) Split the VEPTB samples into 10 parts
    a.) For each split, make that split a test set, and the other 9 parts a training set. The result is 10 training and 10 testing sets.
2.) Split the FTB samples into 10 parts
    a.) Each part will be used to create one training and one test set. For each of the 10 parts, split into 10 sub-parts, making one sub-part a test set, the other 9 sub-parts a training set. The result is 10 training and 10 testing sets. This selection procedure produces a roughly balanced number of VEPTB and FTB samples in the test folds.
3.) For Vi in VEPTB training sets and Fj in FTB training sets:
    a.) Train the Random Forest on parts (Vi-train, Fj-train)
    b.) Test the Random Forest on (Vi-test,Fj-test) of the test sets
4.) When done with all 100 combinations, compute AUC.

We employed the R package 'RandomForest'(35). Random Forests were trained using 200 trees (ntree = 200). All other parameters were set to their default values.

A summary of the AUC is found in **Dataset S15**. Overall, the mean AUC was the highest for the joint data, followed by the methylation data, and finally the RNA-Seq data (**Figure 3D**). In each case, the

VEPTB genes and the VEPTB pathway genes have a comparable AUC, which is higher than genes selected at random. With the mean AUC for each associated pathway, pathways can be ranked on prediction accuracy. Compared to random sets, the candidate pathways improved prediction. We observed that the AUC varied across data types (**Figure S8**). For example, the Notch1 signaling pathway had a much lower prediction error using the methylation data compared to the RNA-Seq data, which might indicate what biological systems are most pertinent in this context.

SUPPLEMENTARY NOTE 3 ASSESSMENT OF REPRODUCIBILITY OF SPONTANEOUS PTB BIOMARKERS

Next we determined the extent that previously reported PTB biomarkers from literature correlated with the phenotype in this cohort using two approaches.

APPROACH 1

Conde-Agudelo *et al.*(36) summarized biomarkers studies in the following categories: inflammation-related, placental proteins/hormone related, angiogenesis-related, coagulation-related and proteomics-based biomarkers (**Dataset S24**). We expanded protein descriptions in these categories to include all gene identifiers associated with a given hormone or protein complex.

We mapped study data at the genomic and molecular level to genes and compare the validity of these biomarkers in this cohort. For genomic variants, we considered only those non-intergenic variants that were significantly associated with the phenotype (FDR < 10%). For gene expression and DNA methylation, we considered the differential methylation or expression status of a given biomarker gene as long as it was measured by the assay.

49/54 proposed biomarkers were measured for expression and 34/54 proposed biomarkers were measured for methylation. 8 biomarkers were differentially methylated or expressed in this cohort (**Figure S12**). Conde-Agudelo *et al.*(36) outlined a set of the most predictive biomarkers that included desmoplakin isoform-1 (*DSP*), statifin (*SFN*), thrombospondin-1 precursor (*THBS1*), matrix metalloproteinase-8 (*MMP8*) and prolactin (*PRL*). In our study, we did not find evidence of differential expression or methylation of these highly predictive biomarker set. Similarly, this most predictive biomarker set according to Conde-Agudelo et al. does not overlap with the 72 gene VEPTB candidate set although we found significant differential expression and methylation in *IL10* and differential methylation in *ALPL* from the VEPTB candidate gene set (**Figure S12**).

APPROACH 2

Saade *et al.* (37) identified 2 serum proteins, insulin-like growth factor–binding protein 4 (IGFBP4) and sex hormone–binding globulin (SHBG), as predictors of spontaneous preterm delivery. By analyzing the gene expression data in our cohort we assessed whether these genes show differential expression with respect to the PTB phenotypes. This analysis showed that IGFBP4 and SHBG were not differentially expressed. However, IGFBP2 a paralog of IGFBP4 was significantly lower expressed in VEPTBs compared to FTBs. Although IGFBP2 and IGFBP4 showed significant anti-correlation in our cohort (results not shown), IGFBP4 itself was not differentially expressed between VEPTB and FTB.

SUPPLEMENTARY NOTE 4 - RARE VARIANT ANALYSIS

Within the candidate VEPTB gene set, we examined all variants with a MAF < 1% in the 44 VEPTB and the 521 PTB mothers. Variants were annotated with CADD scores from the Combined Annotation-Dependent Depletion method(38). The CADD score measures the deleteriousness of single nucleotide variants, and as such indicates the degree of detrimentality. We performed statistical tests using T-tests and SKAT(39) on the set of 72 VEPTB genes as well as a set of 72 random genes for comparison. These analyses are outlined in detail below.

SELECTING 72 'RANDOM' GENES

A method for sampling genes was developed, producing a random set of 72 genes for comparison to the candidate set. The method aims to produce a set of genes that matches the variance profile of a given set of genes, in this case the candidate genes. The variance of the RNA-seq and methylation data is considered simultaneously. To accomplish this, the variance for each gene was computed using the RNA-seq and DNA methylation data over the samples, giving two vectors of variance, one for each data type. The two variance vectors are binned into 100 bins, producing a matrix of bins, where each axis of the matrix belongs to a single data type. The matrix of bins is populated with the candidate genes, the number of genes in each bin providing a sampling size for each bin. Then, the rest of the genes, available for the random set, are assigned to a matrix bin. The candidate sample sizes are used to uniformly sample a number of random genes from each bin. Many bins have zero candidate genes, and are therefore skipped. The generated set then has the same variance profile for both expression and methylation data as the candidate set. Additional features, such as gene length and GC percentage were not considered in selecting the random genes. The 72 random genes are tabulated in **Dataset S12**.

T-TESTS

Three comparisons were made for each gene using T-tests. Specifically, for each gene and each mother (sample) we either recorded: 1) the number of rare variants in the gene (**Dataset S16**), 2) the

mean CADD score of the rare variants (**Dataset S17**), or 3) the maximum CADD score (max CADD scores) of the rare variants (**Dataset S18**). The T-test was used to compare the group of the 44 VEPTB mothers with the group of the 521 PTB mothers. Additionally, the T-tests were performed within each of the five ancestry groups, within gene components (exon, intron, etc.), and both within ancestry groups and gene componets.

Overall, the 44 VEPTB mothers had an average of 928 rare variants across the 72 candidate genes, whereas the 521 FTB mothers had an average of 834 rare variants.

The T-tests comparing the mean CADD scores did not show any genes with FDR < 5% in any comparison (**Dataset S17**). When using the max CADD score per-gene-per-mother, *NAV3* in the Asian group had an FDR of 0.001 with CADD scores higher in the VEPTB group. When comparing max CADD scores by ancestry and gene region, *SBF2* in the ncRNA intron region was significant in the African group with an FDR of 0.003. Also *NAV3* in the Asian group was again significant in the intronic region with an FDR of 0.004. *NAV3* is part of the glucocorticoid receptor pathway which plays a role in regulating the expression of inflammatory genes.

When comparing the counts of rare variants between the two groups, *SIRPA* in the Asian group had higher counts with FDR 5.1×10^{-5}, and specifically higher counts in the introns of SIRPA with FDR 4.4×10^{-4} (**Dataset S16**). These results can possibly be connected to the eQTL results reported earlier in this manuscript.

We observed that sub-setting by ancestry and gene region leads to groups with a very small number of rare variants, especially for the VEPTB group, potentially leading to statistics, which may not be robust.

We observed that the 72 random genes often showed mildly significant results as well (**Figure S9**). This included *CUX1*, which has been observed as differentially expressed in preeclampsia(40). This observation underscores the difficulty in generating a truly random, unrelated set of genes in the context of complex, heterogeneous disease.

Inspection of the locations of the rare variants inside of the 72 candidate genes (inside of intron, exons, UTRs, etc.) did not provide a consistent picture. Analysis of the rare variants in the 72 VEPTB candidate genes led to some statistically significant and biologically interesting findings. These results encourage a comprehensive investigation of the role of these variants in (very early) preterm birth as well as related phenotypes.

RARE VARIANT ANALYSIS WITH SKAT

In a complementary approach, we applied a widely used rare variant statistical test, i.e. the sequence kernel association test (SKAT)(39) to test for association of rare variants with VEPTB. Specifically, we tested all rare variants (MAF < 1%) annotated to VEPTB candidate and random genes for association with the VEPTB phenotype. Using the default implementation of SKAT, we found two genes

(*ADCYAP1* and *CUEDC1*) with statistically significant associations with VEPTB using a FDR threshold of 10%. *ADCYAP1* is in the random set while *CUEDC1* is a VEPTB candidate gene (**Dataset S25**). Next we tested variants annotated as non-synonymous deleterious variants. Here we found two associations with VEPTB: *GRIA3* and again *CUEDC1* (also in **Dataset S25**).

The CUE Domain Containing 1 gene (*CUEDC1*) plays a role in protein trafficking and degradation(41). *CUEDC1* has previously been shown to be hypermethylated in preeclampsia patients in the third trimester relative to controls(42). In our study we also found suggestive association of *CUEDC1* with VEPTB as *CUEDC1* was differentially expressed (**Dataset S11**) and differentially methylated (**Dataset S10**) in VEPTB mothers compared to FTB. *ADCYAP1* encodes pituitary adenylate cyclase-activating polypeptide (PACAP) that is mediator of neuroendocrine stress response. Ressler *et al.* showed that genetic variation in *ADCYAP1* underlies post-traumatic stress disorder (PTSD)(43), while Shaw *et al.* demonstrated that active PTSD is significantly associated with spontaneous preterm birth(44). It is conceivable that rare variants in *ADCYAP1* can disrupt PACAP function, undermine stress response and contribute to a preterm birth outcome. *GRIA3* encodes a glutamate receptor that has been implicated in X-linked mental retardation(45-47). To the best of our knowledge, these three genes have not been previously implicated in preterm birth. However, their function and associations shown in other studies suggest that they could play potential roles in preterm birth, and extensive validation of these results will need to be carried out in an independent cohort with larger numbers of VEPTB samples.

SUPPLEMENTARY NOTE 5 - CO-MORBIDITY ANALYSIS

The WGS data enabled us to explore the relationship between the prevalence of PTB with genetic risks of other diseases through a comorbidity analysis using published GWAS markers. We obtained a copy of the GWAS catalog on March 9th, 2017 (48). We selected from the GWAS catalog entries with stated chromosomal coordinates (chromosome and position), a single rsid value in the 'strongest SNP-risk allele' column, an explicitly stated risk allele (A, C, G or T), a numerical and positive value in the 'OR or beta' column and a stated, unitless confidence interval range in the '95% CI (text)' column. This yielded a set of 7645 SNPs, which we remapped to GRCh37 coordinates by rsid using dbSNP version 149.

We extracted from the merged (multi-sample) VCF file all rows corresponding to 4068 remapped GWAS catalog entries; the remaining SNPs were not observed in the data set. We retained for further analysis only those SNPs with a risk allele frequency stated in the GWAS catalog. For each SNP and phenotype with multiple statements in the GWAS catalog, we preferred the statement with the lowest p-value and retained the information on location, risk allele and odds ratio. Then, for each phenotype we group linked SNPs by maximal clustering with a 15 kb distance cutoff. Within each linkage group, we ignored SNPs for which the risk allele is consistent with both the reference and the alternate allele (ambiguous, 402 SNPs) or is inconsistent with both the reference and the alternate allele (3 SNPs), and ignored batch-associated SNPs (see the section on ‘Batch correction’ earlier in this document). We

prioritized the remaining SNPs in each linkage group by decreasing effect size and selected the highest effect size variant as representative SNP for the linkage group.

Finally, we computed for each individual in the data set a relative disease risk score for each phenotype by summing the log odds ratios for each observed risk allele of the representative SNP for each linkage group; for individuals homozygous for the risk allele, the contribution of such SNPs was doubled relative to heterozygous individuals.

We analyzed the genetic risk profile scores of the 56 disease phenotypes with at least 10 contributing SNPs. The cohort can be broken down into the following subsets:

1. Gender (M-Mother, F-Father, NB-Newborn)
2. Admix categories (AFA,AMR,ASA,EUR,MIX)
3. PTB term categories (V-VEPTB, E-EPTB, P-PTB, F-FTB)

Then, the continuous risk profile scores of each phenotype were analyzed by a two-way ANOVA model using

1. The PTB term categories as a factor
2. The Admix categories as a factor
3. The interaction effect between both

All analyses were done separately for M, F, and NB. P-values, Q-values and F-statistics were computed for each of the two factors and the interaction effect. We observed that variation in genetic risk profile scores is largely explained by the Admix factor, i.e. the risk profiles varied substantially across the different populations/ancestries, resulting in many very low p-values (**Dataset S19**). However, genetic risk profile scores did not significantly vary across PTB categories. Specifically, only one of the 56 phenotypes showed a significant p-value (defined as an uncorrected p-value of 10^{-3} or less; given the multiple testing and relatively large sample size, p-values of 10^{-3} or higher should not be considered as significant). The single significant association was between the genetic risk of prostate cancer based on the maternal genomes and the PTB status. Particularly, EPTB mothers had a lower risk than PTB (late preterm) and FTB mothers, but also than VEPTB mothers (results not shown). This non-consistent result together with the association between women and prostate cancer let us to discard this observation. There were also no significant interaction effects. **Figure S10** displays the genetic risk profile for one of the phenotypes, extreme obesity, across the PTB and Admix factors.

We performed two follow-up analyses aimed at ameliorating the small sample size of the cohort. First, we used a random forest regression model to remove the variation in the genetic risk profile scores that are explained by Admix coefficients, and then reran the analysis on the residuals. For this analysis, fathers and mothers were grouped together to enlarge statistical power. Newborns were ignored. Second, fathers and mothers were grouped together, but the analysis was done within each Admix group. The resulting p-values for each Admix group were combined by a p-value combination scheme

for dependent p-values (49). The statistical tests were run for FTB vs. PTB (where PTB includes EPTB and VEPTB) using the Wilcoxon rank sum test for equal medians, and across the term categories (VEPTB, EPTB, PTB and FTB) using the Kruskal-Wallis nonparametric one-way ANOVA test. None of the results gave statistically significant hits (**Dataset S20**). In conclusion, the WGS data enabled us to explore the relationship between the prevalence of PTB with genetic risks of other diseases through a comorbidity analysis using published GWAS markers for 56 common diseases. We observed that variation in genetic risk varied substantially with ancestry, but not significantly across PTB categories.

SUPPLEMENTARY FIGURE LEGENDS

FIGURE S1 | OVERVIEW OF THE BINARY CLINICAL PHENOTYPES IN STUDY COHORT

Heatmap visualization, where each column represents a family in the study ordered from left to right by increasing gestational age. The binary clinical phenotypes are depicted on the rows with cases in black, controls in white and non-used samples in grey.

FIGURE S2 | MANHATTAN PLOTS OF GENOMIC ASSOCIATIONS ACROSS THE NINE CLINICAL PHENOTYPES

(A) Genome wide significance values ($-\log_{10}$ p-values) for all variants tested for association with preterm birth (PTB), early preterm birth (EPTB) and very early preterm birth (VEPTB). Association tests were performed using EIGENSTRAT on the paternal, maternal and neonatal genomes separately and using FBAT on the family trios. The green horizontal line represents the global p-value threshold of 10^{-8}. Stacked points represent variants within close proximity of one another. **(B)** Similar to (A), but for phenotypes PROM, Pre-eclampsia and Idiopathic PTB. **(C)** Similar to (A), but for phenotypes Cervix-related, Uterine-related and Placenta-Related

FIGURE S3 | OVERVIEW OF THE RESULTS OF THE STATISTICAL TESTS ON THE VALIDATION COHORT

(A,B) QQ-plot comparing the obtained p-values (y-axis) from the genomic association tests compared with the randomly expected p-values (x-axis, generated from a uniform distribution) for all 355 variants across the twelve different testing scenarios (three phenotypes; PTB, EPTB, VEPTB and four groups; 'All', mothers, fathers, newborns) using the Cochran-Armitage chi-square test **(A)** and the hypergeometric test **(B)**. The most significant events are annotated with gene, chromosome, position, rsid, phenotype, test and ratio of alleles across the cases and controls. **(C)** Heatmap of the distribution of samples grouped by preterm birth categories across self-reported race for all samples in the validation cohort. Note that the self-reported race for the Newborns is always 'Unknown'. **(D)** Heatmap

of the distribution of samples with homozygous and heterozygous alternative alleles in the significant variant in *IL28RA* (aka *IFNLR1*) broken down by controls (FTB) and cases (VEPTB) and self-reported race. **(E,F)** QQ-plot comparing the obtained p-values (y-axis) from the genomic association tests compared with the randomly expected p-values (x-axis, generated from a uniform distribution) for all variants in/around the 72 VEPTB genes (brown dots) and the random set of 72 genes (green dots) across the twelve different testing scenarios (three phenotypes; PTB, EPTB, VEPTB and four groups; 'All', mothers, fathers, newborns) using the Cochran-Armitage chi-square test **(E)** and the hypergeometric test **(F)**.

FIGURE S4 | QUANTILE-QUANTILE PLOTS FOR GENOMIC ASSOCIATION TESTS

QQ-plots comparing the obtained p-values (y-axis) from the genomic association tests compared with the randomly expected p-values (x-axis, generated from a uniform distribution). A QQ-plot is depicted for each combination of a test (rows) and clinical phenotype (columns). Magenta lines depict the statistical significance threshold of 10^{-8} and the green lines demarcate the threshold of FDR < 10%.

FIGURE S5 | QUANTILE-QUANTILE PLOTS FOR GENOMIC ASSOCIATION TESTS IN SINGLE-GESTATION COHORT

QQ-plots comparing the obtained p-values (y-axis) from the genomic association tests compared with the randomly expected p-values (x-axis, generated from a uniform distribution). The red line represent the QQ metrics for the single-gestation cohort; the grey lines are the QQ metrics for the complete cohort. A QQ-plot is depicted for each combination of a test (rows) and clinical phenotype (columns). Magenta lines depict the statistical significance threshold of 10^{-8} and the green lines demarcate the threshold of FDR < 10%.

FIGURE S6 | EQTL ASSOCIATION FOR *SIRPA*

The boxplot shows the difference in expression levels of *SIRPA* per genotype of the associated variant at chr20:1,825,838. This marker had a significant eQTL association with SIRPA (p-value 2.08×10^{-4}, q-value 2.79×10^{-2}). The marker was also found significantly associated with VEPTB based on the EIGENSTRAT test (p-value 3.44×10^{-6}, q-value 5.28×10^{-2}).

FIGURE S7 | NETWORK DIAGRAM OF PATHWAYS ENRICHED FOR VEPTB GENES

Gene set overlap of significantly enriched pathways (FDR < 5%) led to three clusters representing Immune/Growth factor pathways, NOTCH signaling pathways and sugar metabolism pathways. Edges highlight the overlap of genes with Jaccard indices greater than 0.1; the widths of the lines represent the strength of the overlap. For example, STAT5B is part of six pathways and PRKCD and VAV2 are

present in five. Several genes are found in both growth factor and immune pathways, including STAT family members that respond to both cytokines and growth factors - see also **Dataset S14** and **Figure 4C**.

FIGURE S8 | SUMMARY OF VEPTB CLASS PREDICTION ERROR BY PATHWAY

The difference in prediction error between the VEPTB and random class based on a random forest classifier for DNA methylation and mRNA expression of genes in VEPTB candidate pathways. Each row shows one of the 27 associated pathways. The x-axis shows the mean VEPTB class prediction error minus the mean VEPTB class prediction error using randomly selected genes. Means were computed over cross validation sets. Zero indicates that the mean prediction was not different between genes in the pathway, and randomly selected genes. A negative value implies that pathway genes produced a better prediction with a lower error rate.

FIGURE S9 | T-TESTS FOR MAXIMUM CADD SCORES OF RARE VARIANTS BETWEEN FTB AND VEPTB GROUPS

Barplots showing T-statistics comparing the rare variant maximum CADD scores per gene per mother between FTB and VEPTB mothers for the 72 VEPTB candidate genes and random genes. Positive T statistics indicates the VEPTB group having larger maximum CADD scores on average.

FIGURE S10 | GENETIC RISK SCORES FOR EXTREME OBESITY ACROSS THE PTB COHORT

Top-left: Distribution bar graph of the genetic risk scores for early onset extreme obesity across all samples (mothers, fathers, newborns) in the cohort. **Bottom-left**: Bar plot showing $-\log_{10}$ p-values (left), $-\log_{10}$ Q-values (center) and F-statistics (right) for the PTB, Admix and interactions factors. **Right**: Distribution of genetic risk scores stratified by PTB term categories (V-VEPTB, E-EPTB, P-PTB, F-FTB) and gender (M-Mother, F-Father, NB-Newborn) (top), by Admix categories (AFA,AMR,ASA,EUR,MIX) and gender (center), and by all three factors (PTB, Admix and Gender) (bottom).

FIGURE S11 | DISTRIBUTION OF ASSOCIATED VARIANTS ACROSS THE GENOME

(A) Heatmap summarizing the number of variants by genomic test and clinical phenotype. ‘F’,‘M’, ‘NB’ represent the variants associated with a clinical phenotype for the EIGENSTRAT test for each of the paternal, maternal and neonatal genomes. FAM represents results from the FBAT test. Summary barplots representing the distribution of variants by **(B)** Variant type - SNP single nucleotide polymorphism, DNP, TNP, ONP, INS, DEL. **(C)** Variant Classification - annotated functional classes include: 5’Flank, 5’UTR, missense mutation, silent, splice site, intron, 3’UTR, IGR (intergenic), RNA (encompassing variants annotated to functional RNA transcripts) and lincRNA. The bars indicate the

proportions of variants, which were significantly associated with the PTB-related clinical phenotypes for each test. Asterisks (*) indicate a statistically significant difference in the distribution of variants across the test and genome conditions for a given phenotype (Fisher's exact test $p < 0.05$).

FIGURE S12 | REPRODUCIBILITY ANALYSIS OF PTB BIOMARKERS

Venn diagram summarizing biomarkers reproduced in this cohort via association tests in WGS and differential methylation and expression analysis for VEPTB. Representative violin plots showing differences in methylation and expression levels between full term birth (FTB) and very early preterm birth (VEPTB).

SUPPLEMENTARY DATASET DESCRIPTIONS

DATASET S1 | OVERVIEW OF GENOMIC DATA AND FILTERING STRATEGY

This table presents the total number of WGS samples (family trios) and the total number of variants considered, filtered and analyzed in this study.

DATASET S2 | GENOMIC ASSOCIATION RESULTS ACROSS PHENOTYPES

Overview of all variants, which were statistically associated with a clinical phenotype at the FDR threshold of 10%. Variants that meet the p-value threshold of 10^{-8} are colored. There is a separate table for each of the clinical phenotypes for which at least one significant event was found.

DATASET S3 | OVERLAP OF GENES WITH ASSOCIATED VARIANTS IN MULTIPLE PHENOTYPES

The first column contains all genes that had a variant found to be significant ($p<10^{-8}$) in at least one of the nine clinical phenotypes. Shaded black boxed to the right indicate the clinical phenotypes for which the genes were significant.

DATASET S4 | OVERVIEW OF LITERATURE SOURCES FOR GENES IMPLICATED IN PRETERM BIRTH AND RELATED PHENOTYPES

An overview of the seven gene lists derived from literature including a short description, source and number of genes. For each of the seven lists, a table with the corresponding genes is available.

DATASET S5 | SUMMARY OF OVERLAP ANALYSIS BETWEEN GENOMIC, MOLECULAR AND LITERATURE IDENTIFIED GENE SETS

Statistical enrichment (using the hypergeometric test) between gene sets derived from the genomic associations tests, differential gene expression analysis, differential DNA methylation analysis and PTB gene sets from literature. There are separate tables for each of the nine clinical phenotypes.

DATASET S6 | OVERVIEW OF VALIDATION DATASET

Overview of the distribution of the samples in the validation cohort broken down by preterm birth categories based on gestational age.

DATASET S7 | STATISTICALLY SIGNIFICANT VARIANTS REPRODUCED IN VALIDATION COHORT

Overview of the statistically significant variants (uncovered in the original cohort) in the validation cohort for each of the three PTB phenotypes (PTB, EPTB and VEPTB) and four groups (All, Fathers, Mother, Newborn). The number in the cells indicate the number of statistically significant variants at FDR<10% and p-value<0.01 using the Cochran-Armitage chi-square test. This number is broken down into the genes, where these variants were found with the number of genes associated to a gene between parentheses.

DATASET S8 | STATISTICALLY SIGNIFICANT VARIANTS IN RANDOM AND VEPTB GENES

Overview of the statistically significant variants in or around the 72 VEPTB genes and a random set of 72 genes in the validation cohort for each of the three PTB phenotypes (PTB, EPTB and VEPTB) and four groups (All, Fathers, Mother, Newborn). The number in the cells indicates the number of statistically significant variants and genes at a Bonferroni corrected p-value of 0.05 using the Cochran-Armitage chi-square test and the hypergeometric test.

DATASET S9 | GENOMIC ASSOCIATION RESULTS ACROSS PHENOTYPES FOR THE SINGLE-GESTATION COHORT

Overview of all variants, which were statistically associated with a clinical phenotype at the FDR threshold of 10% in the single-gestation cohort. Variants that meet the p-value threshold of 10^{-8} are colored. There is a separate table for each of the clinical phenotypes for which at least one significant event was found.

DATASET S10 | DIFFERENTIAL METHYLATION RESULTS ACROSS PHENOTYPES

Overview of all DNA methylation probes, which were differentially methylated between cases and controls for a clinical phenotype at the FDR threshold < 10%. The associated gene for each probe is

also given. There is a separate table for each of the clinical phenotypes for which at least one significant event was found.

DATASET S11 | DIFFERENTIAL GENE EXPRESSION RESULTS ACROSS PHENOTYPES

Overview of all genes, which were differentially expressed between cases and controls for a clinical phenotype at the FDR threshold < 10% and a log fold-change > 0.5. There is a separate table for each of the clinical phenotypes for which at least one significant event was found.

DATASET S12 | VEPTB GENE LISTS

VEPTB Candidate Genes lists 72 genes that showed significant associations with VEPTB at an FDR of 10% in at least two of the three data types (WGS, gene expression and DNA methylation data) or genome-wide significance (p-value < 10^{-8}) at the genomic level. **Random Genes** an equal sized list of genes that match the variance profile of the VEPTB candidate genes used for background testing. See **Supplementary Note 4** on rare variant analysis for details.

DATASET S13 | E/M QTL ANALYSIS WITHIN THE VEPTB CANDIDATE GENE LIST

Overview of eQTLs and mQTLS within the VEPTB candidate gene list. The first column denotes the filter used to determine the set of variant QTL pairs examined. Variants: the number of unique variants examined (with a MAF>0.01). Quantitative Traits: The number of DEGs or DMPs observed. Pairs: the unique number of DEG-SNP or DMP-SNP pairs observed. eQTL: expression Quantitative trait loci; mQTL: methylation quantitative trait loci. DMP: Differentially methylated Probe DEG: Differentially expressed gene.

DATASET S14 | VEPTB CANDIDATE GENES SIGNIFICANTLY ENRICHED PATHWAYS

The full set of enriched pathways with FDR < 5% using the 72 VEPTB candidate genes. Results were derived using the ConsensusPathDB service. NetPath, Wikipathways, Reactome, and KEGG pathway databases were utilized, and hypergeometric tests were performed to look for over-representation within pathways. Pathways had to contain at least 2 genes to be tested. An FDR adjustment was made to the p-values to correct for multiple testing. Gene backgrounds are supplied by CPDB, which includes all genes that participate in at least one pathway; a total of 10,773 genes.

DATASET S15 | PREDICTION OF VEPTB USING RANDOM FOREST CLASSIFIER

Distribution of mean AUC in VEPTB class prediction by random forest. VEPTB group indicates the class prediction AUC, TEST the overall prediction error (both classes, FTB and VEPTB), Cand is the candidate genes, Pathway all genes within the candidate pathways, minus the candidate genes

themselves, Rand Cand is a random set of genes the same size as the candidate set (100 draws performed) and Rand Path is a random set of genes that is the same size as the Pathway set (100 draws performed).

DATASET S16 | T-TESTS ON THE COUNTS OF RARE VARIANTS FOR VEPTB CANDIDATE GENES

Statistics from the T-tests comparing the counts of rare variants between the FTB and VEPTB mothers for the 72 VEPTB candidate genes. The excel document contains four sheets for tests by gene, by gene and ancestry, by gene and part of gene, and by gene, ancestry, and part-of-gene.

DATASET S17 | T-TESTS ON THE MAXIMUM CADD SCORES OF RARE VARIANTS FOR VEPTB CANDIDATE GENES

Statistics from the T-tests comparing the maximum CADD scores of the rare variants between the FTB and VEPTB mothers for the 72 VEPTB candidate genes. The excel document contains four sheets for tests by gene, by gene and ancestry, by gene and part of gene, and by gene, ancestry, and part-of-gene.

DATASET S18 | T-TESTS ON THE ALL CADD SCORES OF RARE VARIANTS FOR VEPTB CANDIDATE GENES

Statistics from the T-tests comparing all CADD scores of the rare variants between the FTB and VEP15TB mothers for the 72 VEPTB candidate genes. The excel document contains four sheets for tests by gene, by gene and ancestry, by gene and part of gene, and by gene, ancestry, and part-of-gene.

DATASET S19 | RESULTS OF TWO-WAY ANOVA CO-MORBIDITY ANALYSIS

This table lists the $-\log_{10}$ p-values (rounded to the nearest integer) for the two-way ANOVA analysis to explain the genetic risk scores for 56 diseases (rows). In the model, we included two factors: 1) PTB term categories and 2) Admix categories. For these two factors and for the interaction effect (INT) the p-values are listed. Note that the cohort was stratified by gender (M-Mother, F-Father, NB-Newborn).

DATASET S20 | ASSOCIATION BETWEEN GENETIC RISK SCORES AND PTB STATUS

This table lists the $-\log_{10}$ p-values (rounded to the nearest integer) for the association between PTB status and the genetic risk scores for 56 diseases (rows). There is one sheet ('PTB vs. FTB') where the samples were split into FTB and PTB (where PTB includes EPTB and VEPTB), and testing was performed using the Wilcoxon rank sum test for equal medians. The second sheet

('VEPTB,EPTB,PTB,FTB') contains results obtained by an analysis across the term categories (VEPTB, EPTB, PTB and FTB) using the Kruskal-Wallis nonparametric one-way ANOVA test. The second column ('All M+F samples (compensated for ADMIX)') is based on a statistical test on the residual genetic risk score after the variation explained by Admix coefficients was removed using a Random Forest regression model. The subsequent 5 columns display results of these tests using the (uncompensated) genetic risk scores applied to the samples of a particular Admix group. The final column lists the combined P-value of the AFA, AMR, ASA and EUR results.

DATASET S21 | DISTRIBUTION OF PHENOTYPE ASSOCIATED VARIANTS ACROSS FUNCTIONAL GENOMIC REGIONS

This table presents the variant type and class of moderately significant variants (FDR<10%) for each of the phenotypes and genomic association tests based on oncotator variant classification: SNP single nucleotide polymorphism, DNP di-nucleotide polymorphism, TNP tri-nucleotide polymorphism, MNP multi-nucleotide polymorphism, ONP other nucleotide polymorphism, INS insertion, DEL deletion. The p-values were based on Fisher's exact test.

DATASET S22 | CORRELATION OF DIFFERENTIALLY EXPRESSED GENES WITH MICRORNA EXPRESSION

Overview of all miRNA-gene pairs that were significantly correlated with each other at the FDR threshold < 10%. Besides a summary table, there are separate tables for each of the clinical phenotypes for which at least one significant event was found.

DATASET S23 | DETAILS OF VEPTB CANDIDATE PATHWAY ENRICHMENT

The odds ratios of the hypergeometric test for enrichment of significant differentially expressed, differentially methylated and genomic association in genes attributed to the phenotypes on the rows and enriched in VEPTB candidate pathways.

DATASET S24 | PUBLISHED PTB BIOMARKERS

List of the biomarkers described in Conde *et al.*.

DATASET S25 | SKAT RARE VARIANT ASSOCIATION TESTS FOR VEPTB CANDIDATE AND RANDOM GENES

Results for Sequencing Kernel Association Tests (SKAT) for VEPTB. We separately tested all variants (first sheet in Excel document) and the deleterious variants (second sheet in Excel document). We

report the SKAT Q statistic, p-value and Benjamini-Hochberg corrected FDR values. The number of markers represents the number of rare variant positions for the given gene used in the test.

REFERENCES

1. Genomes Project C*, et al.* (2015) A global reference for human genetic variation. *Nature* 526(7571):68-74.
2. Price AL*, et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* 38(8):904-909.
3. Horvath S, Xu X, & Laird NM (2001) The family based association test method: strategies for studying general genotype--phenotype associations. *European journal of human genetics : EJHG* 9(4):301-306.
4. Ramos AH*, et al.* (2015) Oncotator: cancer variant annotation tool. *Human mutation* 36(4):E2423-2429.
5. Li H & Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754-1760.
6. Aronesty E (2011) Command-line tools for processing biological sequencing data.
7. Aronesty E (2013) Comparison of sequencing utility programs. *The Open Bioinformatics Journal* 7:1-8.
8. Ru Y*, et al.* (2014) The multiMiR R package and database: integration of microRNA-target interactions along with their disease and drug associations. *Nucleic acids research* 42(17):e133.
9. Xiao F*, et al.* (2009) miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res* 37(Database issue):D105-110.
10. Hsu SD*, et al.* (2014) miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res* 42(Database issue):D78-85.
11. Papadopoulos GL, Reczko M, Simossis VA, Sethupathy P, & Hatzigeorgiou AG (2009) The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res* 37(Database issue):D155-158.
12. Teschendorff AE*, et al.* (2013) A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* 29(2):189-196.

13. Naeem H*, et al.* (2014) Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the HumanMethylation450 array. *BMC genomics* 15:51.

14. Rousseeuw PJ & Hubert M (2011) Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1(1):73-79.

15. Hubert M & Debruyne M (2010) Minimum covariance determinant. *Wiley interdisciplinary reviews: Computational statistics* 2(1):36-43.

16. Bullard JH, Purdom E, Hansen KD, & Dudoit S (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11:94.

17. Robinson MD, McCarthy DJ, & Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26(1):139-140.

18. Law CW, Chen Y, Shi W, & Smyth GK (2014) voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 15(2):R29.

19. Alexander DH, Novembre J, & Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19(9):1655-1664.

20. Ritchie ME*, et al.* (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* 43(7):e47.

21. Uzun A*, et al.* (2012) dbPTB: a database for preterm birth. *Database : the journal of biological databases and curation* 2012:bar069.

22. Capece A, Vasieva O, Meher S, Alfirevic Z, & Alfirevic A (2014) Pathway analysis of genetic factors associated with spontaneous preterm birth and pre-labor preterm rupture of membranes. *PloS one* 9(9):e108578.

23. McElroy JJ (2013) Genetics of spontaneous idiopathic preterm birth: exploration of maternal and fetal genomes. (Vanderbilt University).

24. Plunkett J (2010) Genetic Influences on Preterm Birth.

25. Plunkett J*, et al.* (2011) An evolutionary genomic approach to identify genes involved in human birth timing. *PLoS genetics* 7(4):e1001365.

26. Tuteja G, Cheng E, Papadakis H, & Bejerano G (2012) PESNPdb: a comprehensive database of SNPs studied in association with pre-eclampsia. *Placenta* 33(12):1055-1057.

27. Shabalin AA (2012) Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* 28(10):1353-1358.

28. Purcell S*, et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* 81(3):559-575.

29. Maurano MT*, et al.* (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337(6099):1190-1195.

30. Kamburov A, Stelzl U, Lehrach H, & Herwig R (2013) The ConsensusPathDB interaction database: 2013 update. *Nucleic acids research* 41(Database issue):D793-800.

31. Bodian DL*, et al.* (2016) Utility of whole-genome sequencing for detection of newborn screening disorders in a population cohort of 1,696 neonates. *Genetics in medicine : official journal of the American College of Medical Genetics* 18(3):221-230.

32. Plunkett J & Muglia LJ (2008) Genetic contributions to preterm birth: implications from epidemiological and genetic association studies. *Annals of medicine* 40(3):167-195.

33. Crider KS, Whitehead N, & Buus RM (2005) Genetic variation associated with preterm birth: a HuGE review. *Genetics in medicine : official journal of the American College of Medical Genetics* 7(9):593-604.

34. Breiman L (2001) Random forests. *Machine learning* 45(1):5-32.

35. Liaw A & Wiener M (2002) Classification and regression by randomForest. *R news* 2(3):18-22.

36. Conde-Agudelo A, Papageorghiou AT, Kennedy SH, & Villar J (2011) Novel biomarkers for the prediction of the spontaneous preterm birth phenotype: a systematic review and meta-analysis. *BJOG : an international journal of obstetrics and gynaecology* 118(9):1042-1054.

37. Saade GR*, et al.* (2016) Development and validation of a spontaneous preterm delivery predictor in asymptomatic women. *American journal of obstetrics and gynecology* 214(5):633. e631-633. e624.

38. Kircher M*, et al.* (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics* 46(3):310-315.

39. Wu MC*, et al.* (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics* 89(1):82-93.

40. Sitras V*, et al.* (2009) Differential placental gene expression in severe preeclampsia. *Placenta* 30(5):424-433.

41. Colland F*, et al.* (2004) Functional proteomics mapping of a human signaling pathway. *Genome research* 14(7):1324-1332.

42. Yan Y*, et al.* (2013) Screening for preeclampsia pathogenesis related genes. *European review for medical and pharmacological sciences* 17(22):3083-3094.

43. Ressler KJ*, et al.* (2011) Post-traumatic stress disorder is associated with PACAP and the PAC1 receptor. *Nature* 470(7335):492-497.

44. Shaw JG*, et al.* (2014) Posttraumatic stress disorder and risk of spontaneous preterm birth. *Obstetrics & Gynecology* 124(6):1111-1119.

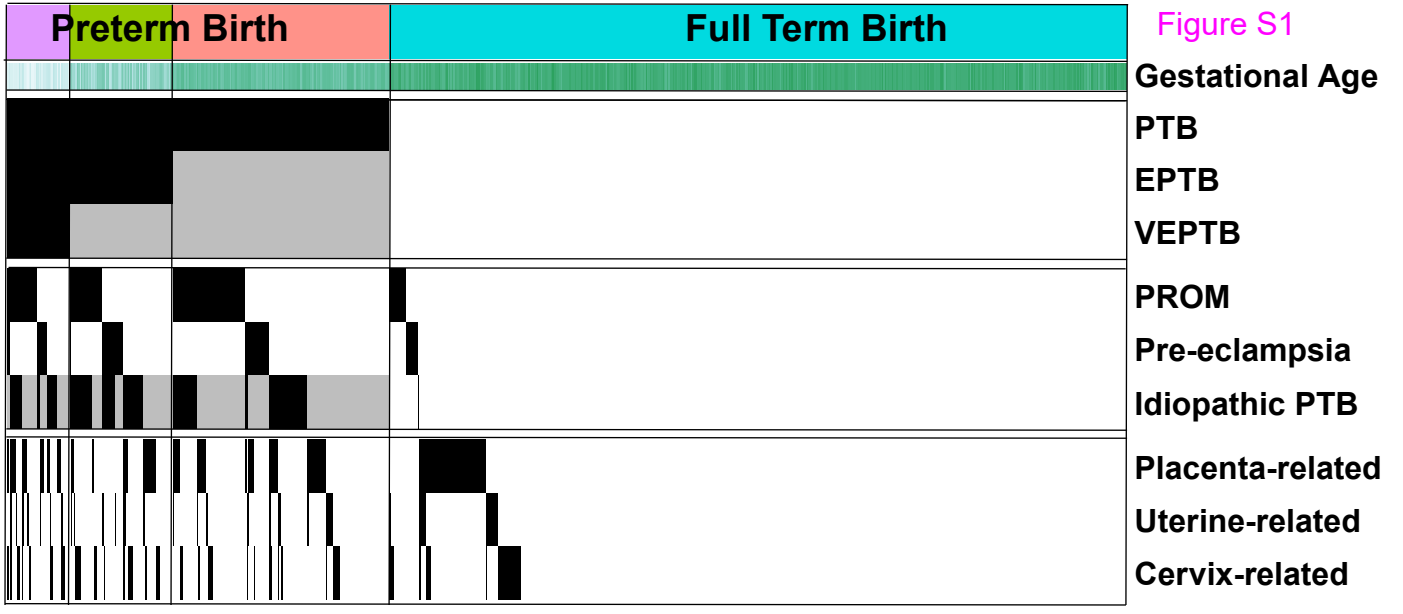
45. Wu Y*, et al.* (2007) Mutations in ionotropic AMPA receptor 3 alter channel properties and are associated with moderate cognitive impairment in humans. *Proceedings of the National Academy of Sciences* 104(46):18163-18168.

46. Bonnet C*, et al.* (2009) Aberrant GRIA3 transcripts with multi‐exon duplications in a family with X‐linked mental retardation. *American Journal of Medical Genetics Part A* 149(6):1280-1289.

47. Philips AK*, et al.* (2014) X-exome sequencing in Finnish families with Intellectual Disability-four novel mutations and two novel syndromic phenotypes. *Orphanet J Rare Dis* 9:49.

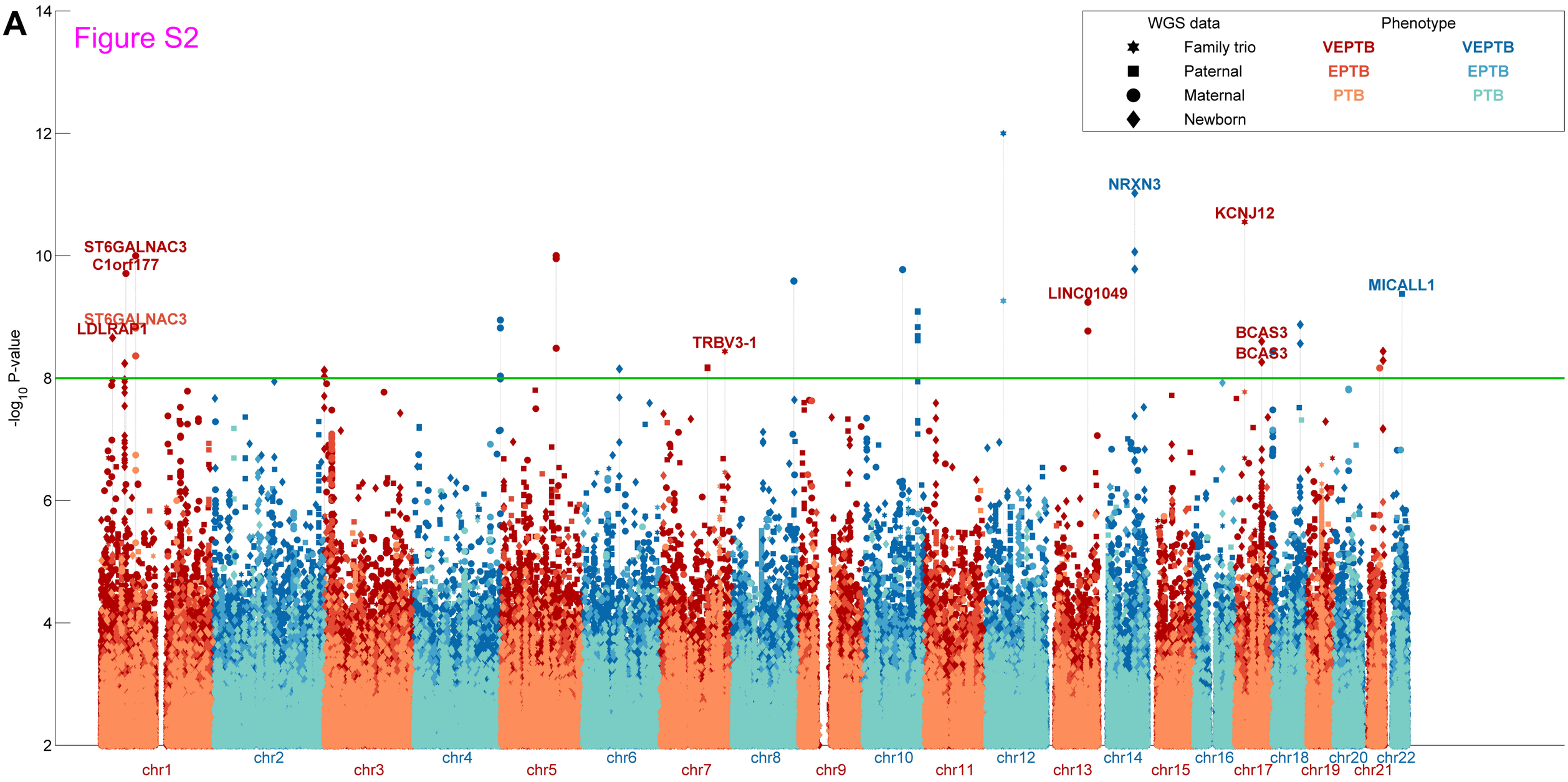
48. Welter D*, et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research* 42(Database issue):D1001-1006.

49. Poole W, Gibbs DL, Shmulevich I, Bernard B, & Knijnenburg TA (2016) Combining dependent P-values with an empirical adaptation of Brown's method. *Bioinformatics* 32(17):i430-i436.

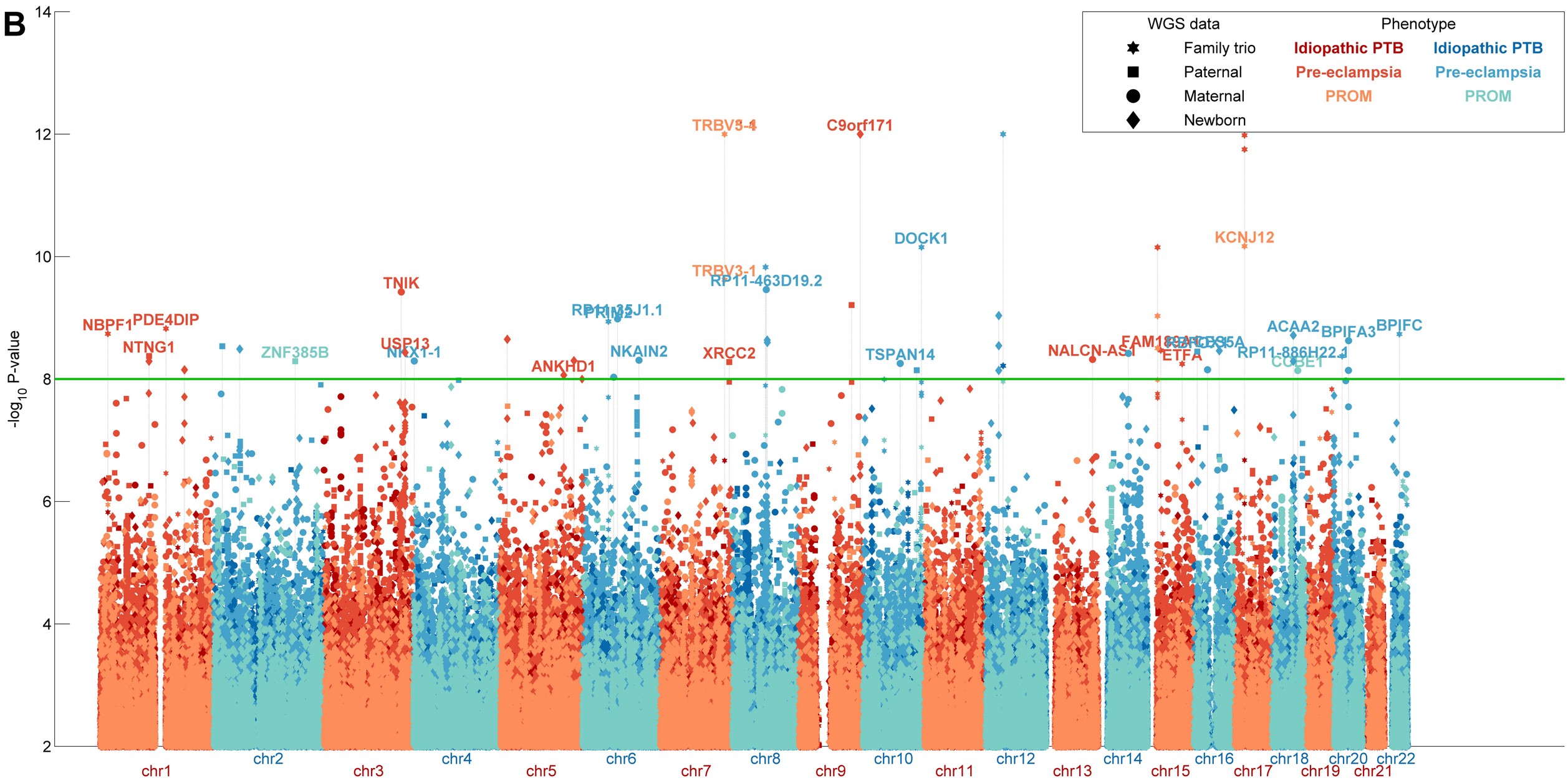

Preterm Birth
Full Term Birth
Figure S1
Gestational Age
PTB
EPTB
VEPTB
PROM
Pre-eclampsia
Idiopathic PTB
Placenta-related
Uterine-related
Cervix-related

A

Figure S2



B



C

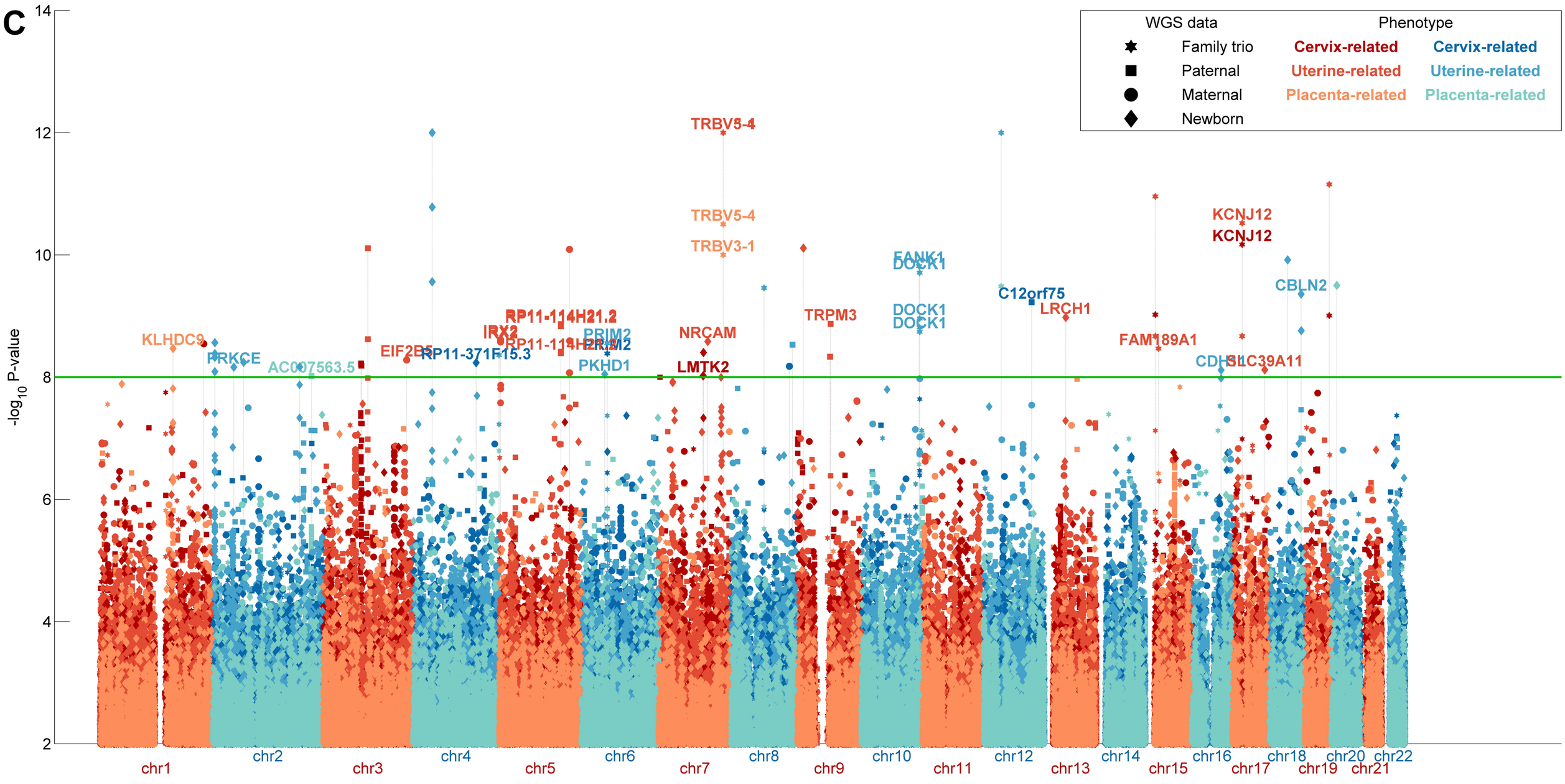
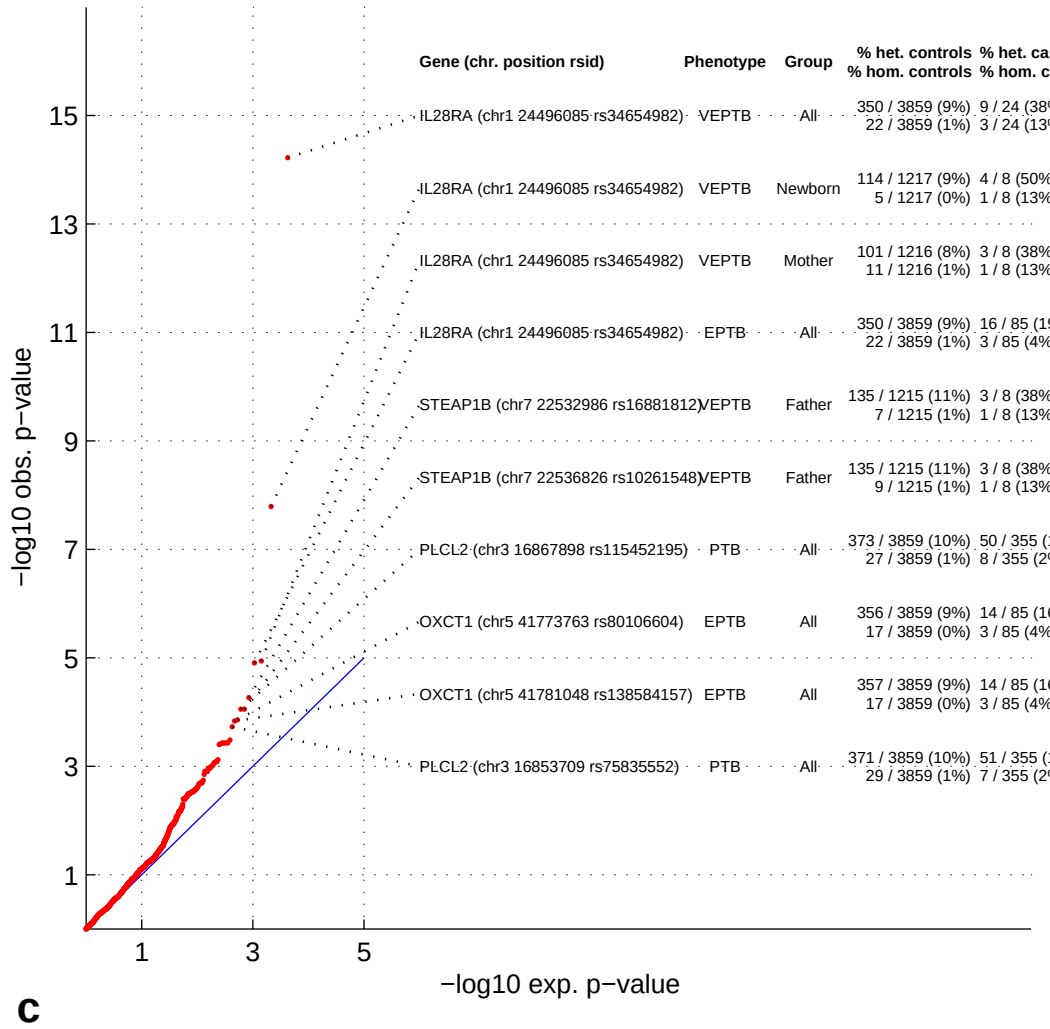


Figure S3

a

Cochran−Armitage chi−square test

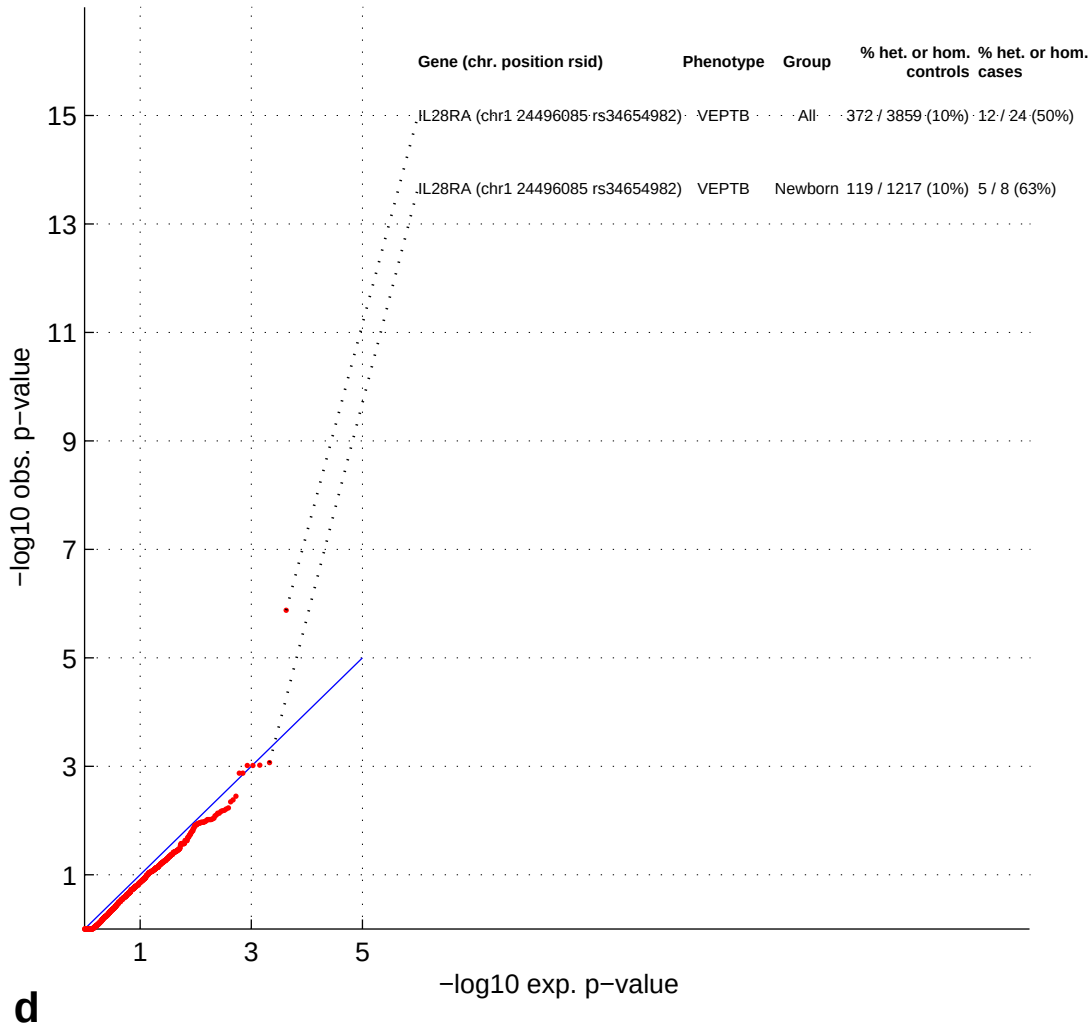
Gene (chr. position rsid)	Phenotype	Group	% het. controls % hom. controls	% het. cases % hom. cases
IL28RA (chr1 24496085 rs34654982)	VEPTB	All	350 / 3859 (9%) 22 / 3859 (1%)	9 / 24 (38%) 3 / 24 (13%)
IL28RA (chr1 24496085 rs34654982)	VEPTB	Newborn	114 / 1217 (9%) 5 / 1217 (0%)	4 / 8 (50%) 1 / 8 (13%)
IL28RA (chr1 24496085 rs34654982)	VEPTB	Mother	101 / 1216 (8%) 11 / 1216 (1%)	3 / 8 (38%) 1 / 8 (13%)
IL28RA (chr1 24496085 rs34654982)	EPTB	All	350 / 3859 (9%) 22 / 3859 (1%)	16 / 85 (19%) 3 / 85 (4%)
STEAP1B (chr7 22532986 rs16881812)	VEPTB	Father	135 / 1215 (11%) 7 / 1215 (1%)	3 / 8 (38%) 1 / 8 (13%)
STEAP1B (chr7 22536826 rs10261548)	VEPTB	Father	135 / 1215 (11%) 9 / 1215 (1%)	3 / 8 (38%) 1 / 8 (13%)
PLCL2 (chr3 16867898 rs115452195)	PTB	All	373 / 3859 (10%) 27 / 3859 (1%)	50 / 355 (14%) 8 / 355 (2%)
OXCT1 (chr5 41773763 rs80106604)	EPTB	All	356 / 3859 (9%) 17 / 3859 (0%)	14 / 85 (16%) 3 / 85 (4%)
OXCT1 (chr5 41781048 rs138584157)	EPTB	All	357 / 3859 (9%) 17 / 3859 (0%)	14 / 85 (16%) 3 / 85 (4%)
PLCL2 (chr3 16853709 rs75835552)	PTB	All	371 / 3859 (10%) 29 / 3859 (1%)	51 / 355 (14%) 7 / 355 (2%)



b

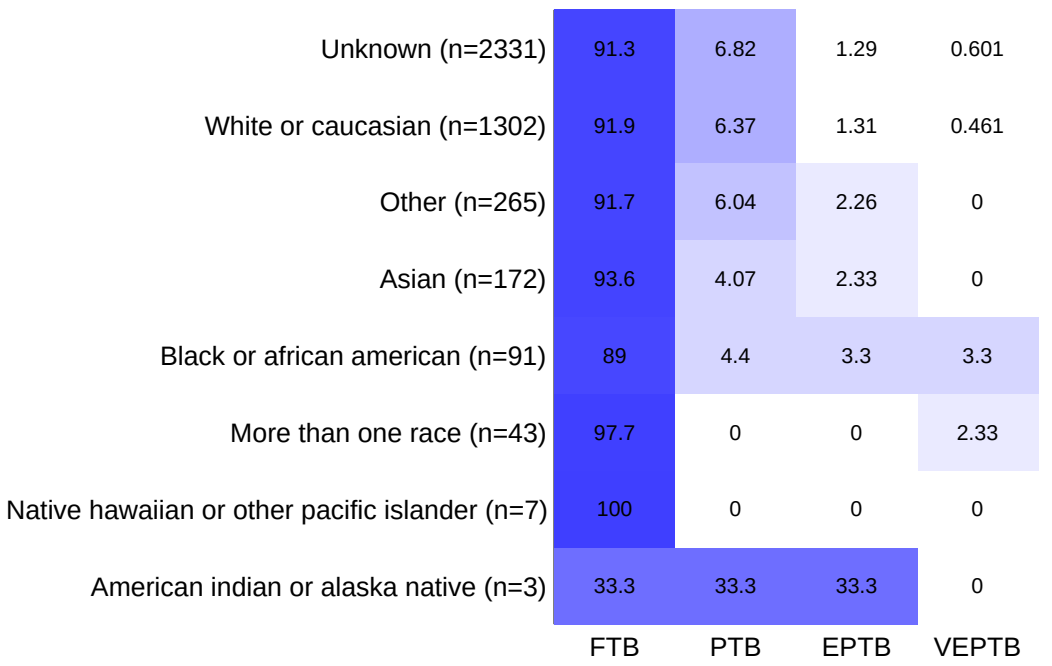
Hypergeometric test

Gene (chr. position rsid)	Phenotype	Group	% het. or hom. controls	% het. or hom. cases
IL28RA (chr1 24496085 rs34654982)	VEPTB	All	372 / 3859 (10%)	12 / 24 (50%)
IL28RA (chr1 24496085 rs34654982)	VEPTB	Newborn	119 / 1217 (10%)	5 / 8 (63%)



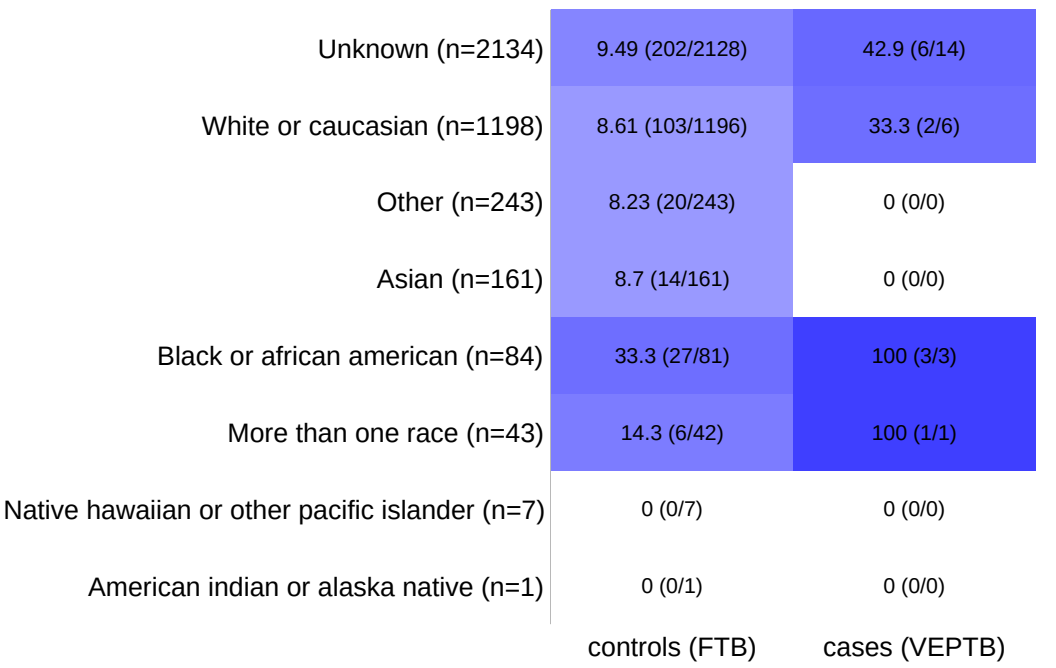
c

% term category across self reported race



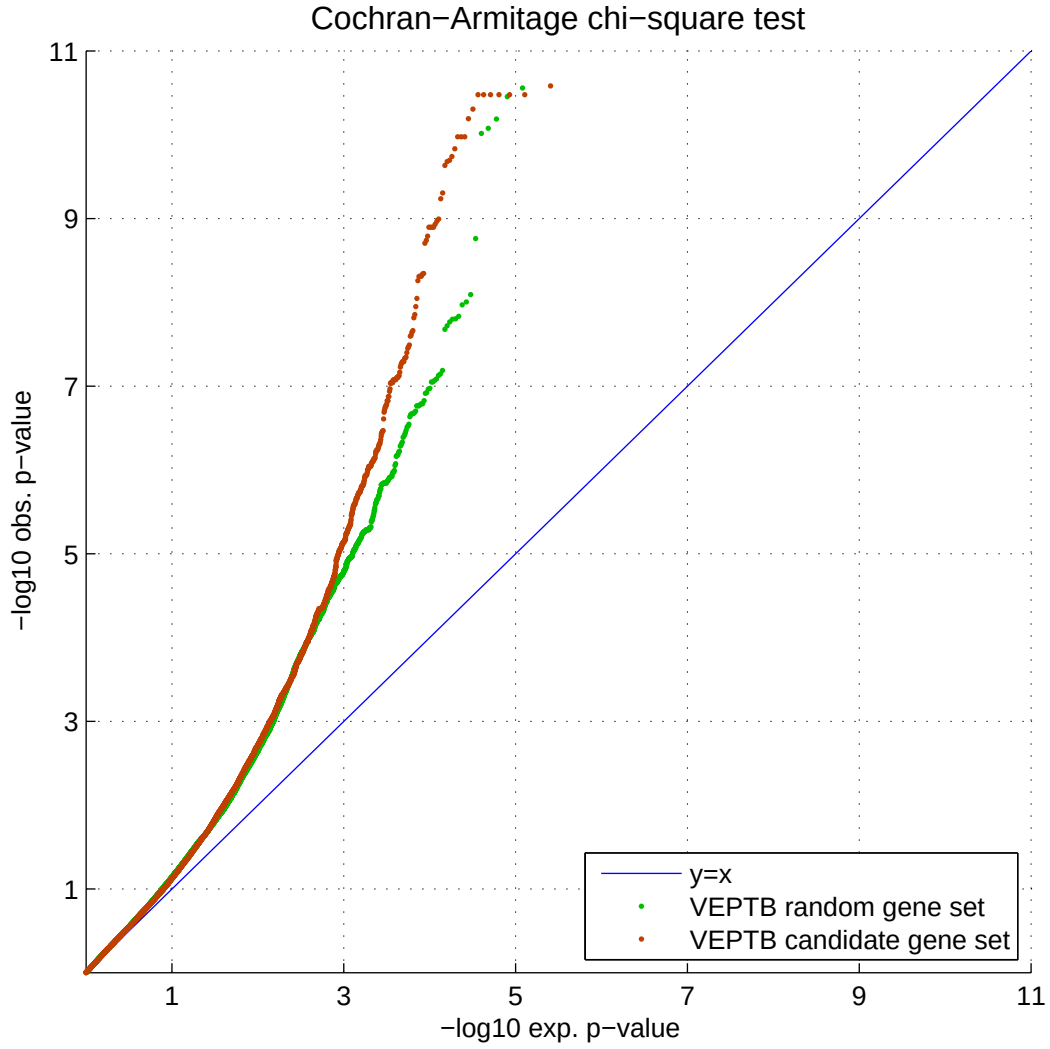
d

% hom./het. alternative allelle in IL28RA



e

Cochran−Armitage chi−square test



f

Hypergeometric test

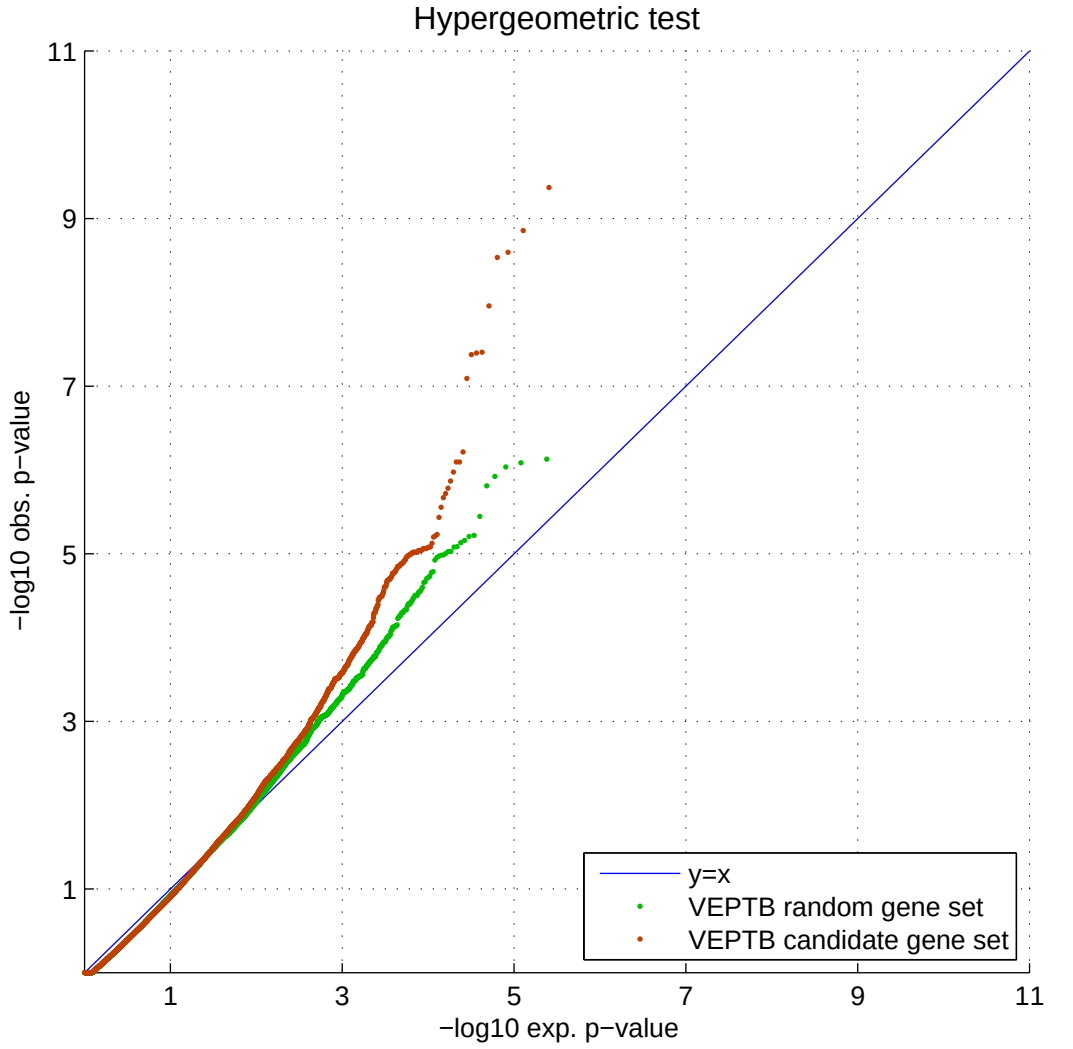


Figure S4

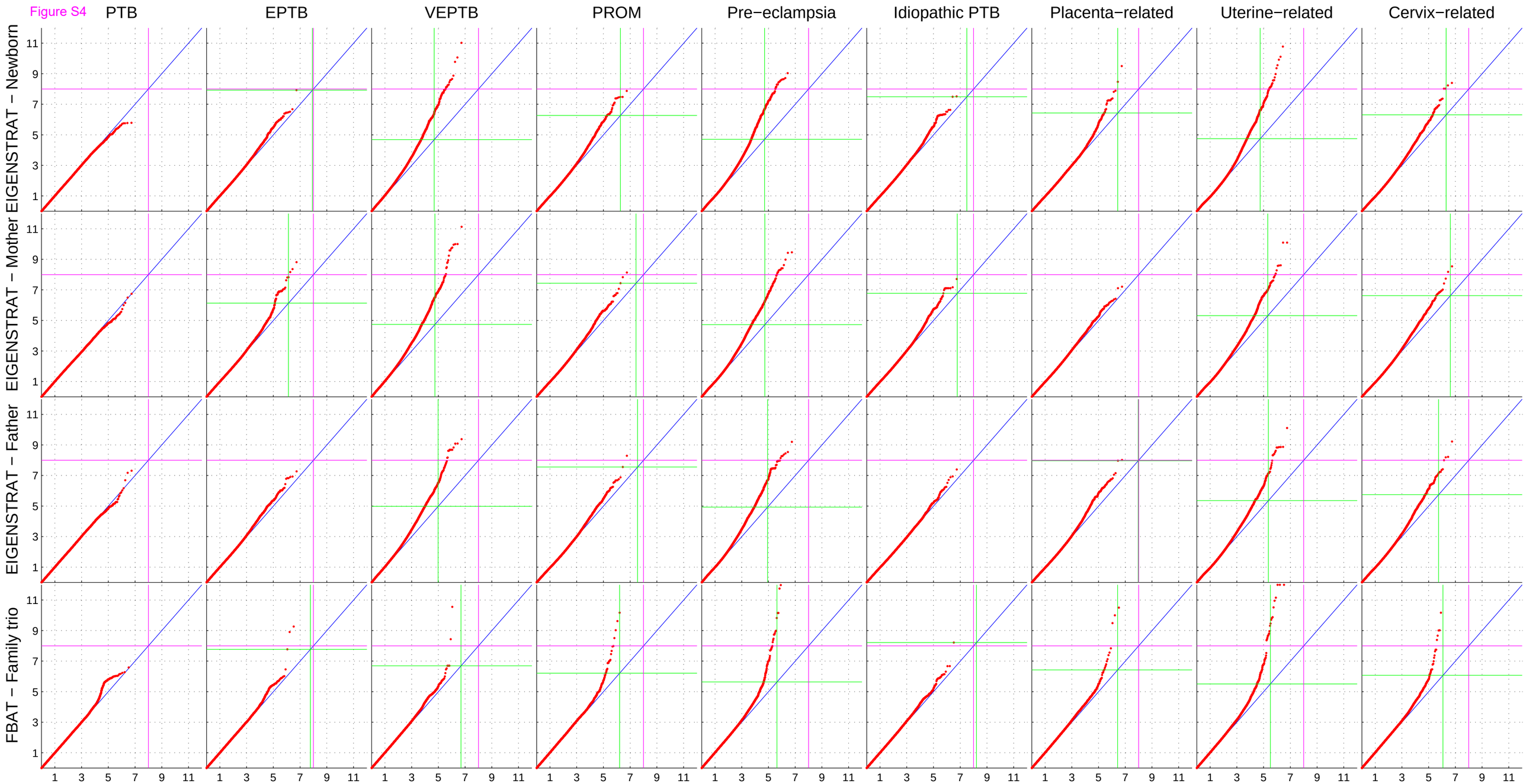

PTB
EPTB
VEPTB
PROM
Pre−eclampsia
Idiopathic PTB
Placenta−related
Uterine−related
Cervix−related
EIGENSTRAT − Newborn
EIGENSTRAT − Mother
EIGENSTRAT − Father
FBAT − Family trio
1
3
5
7
9
11

Figure S5

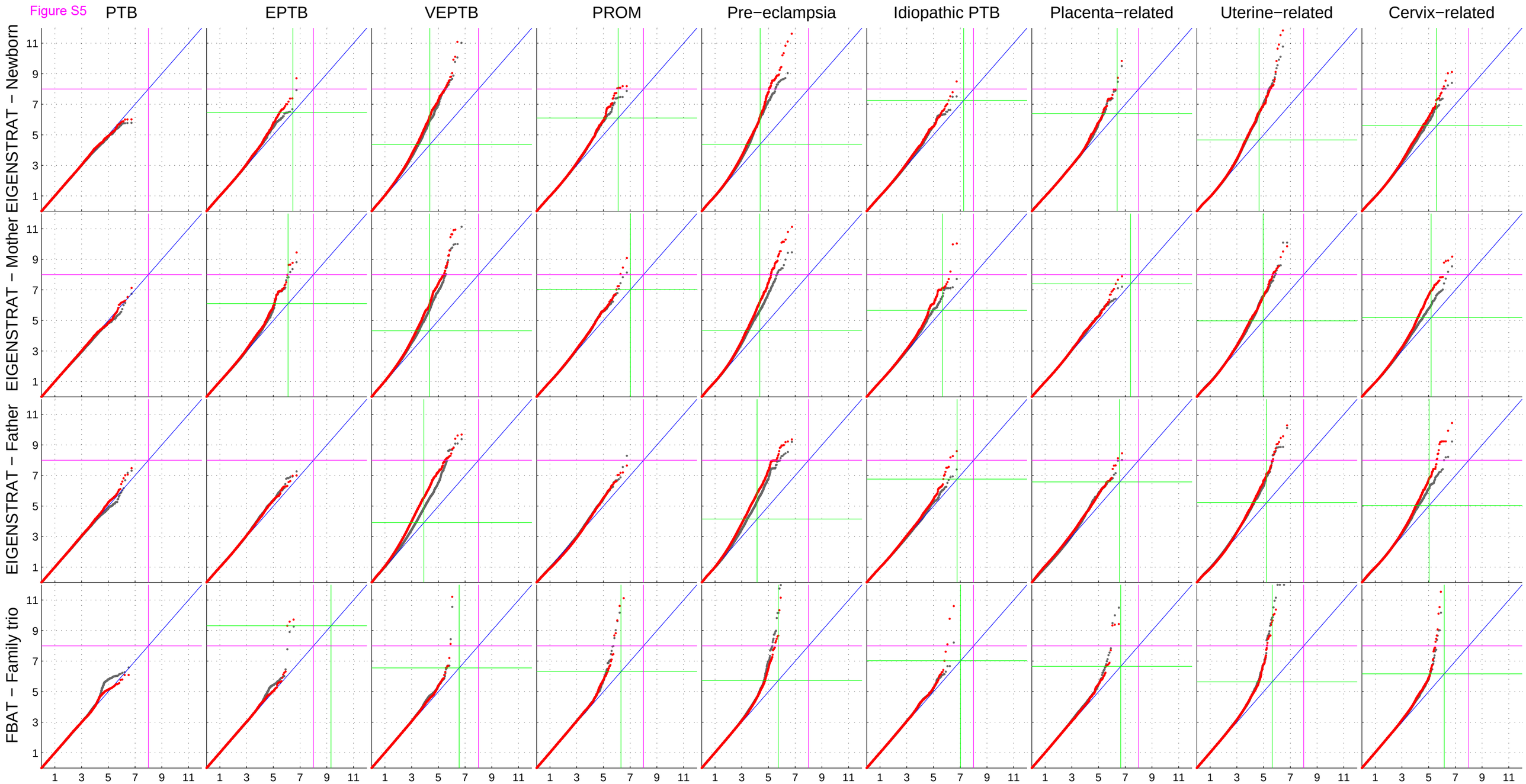

PTB
EPTB
VEPTB
PROM
Pre−eclampsia
Idiopathic PTB
Placenta−related
Uterine−related
Cervix−related
EIGENSTRAT − Newborn
EIGENSTRAT − Mother
EIGENSTRAT − Father
FBAT − Family trio
1
3
5
7
9
11

Figure S6

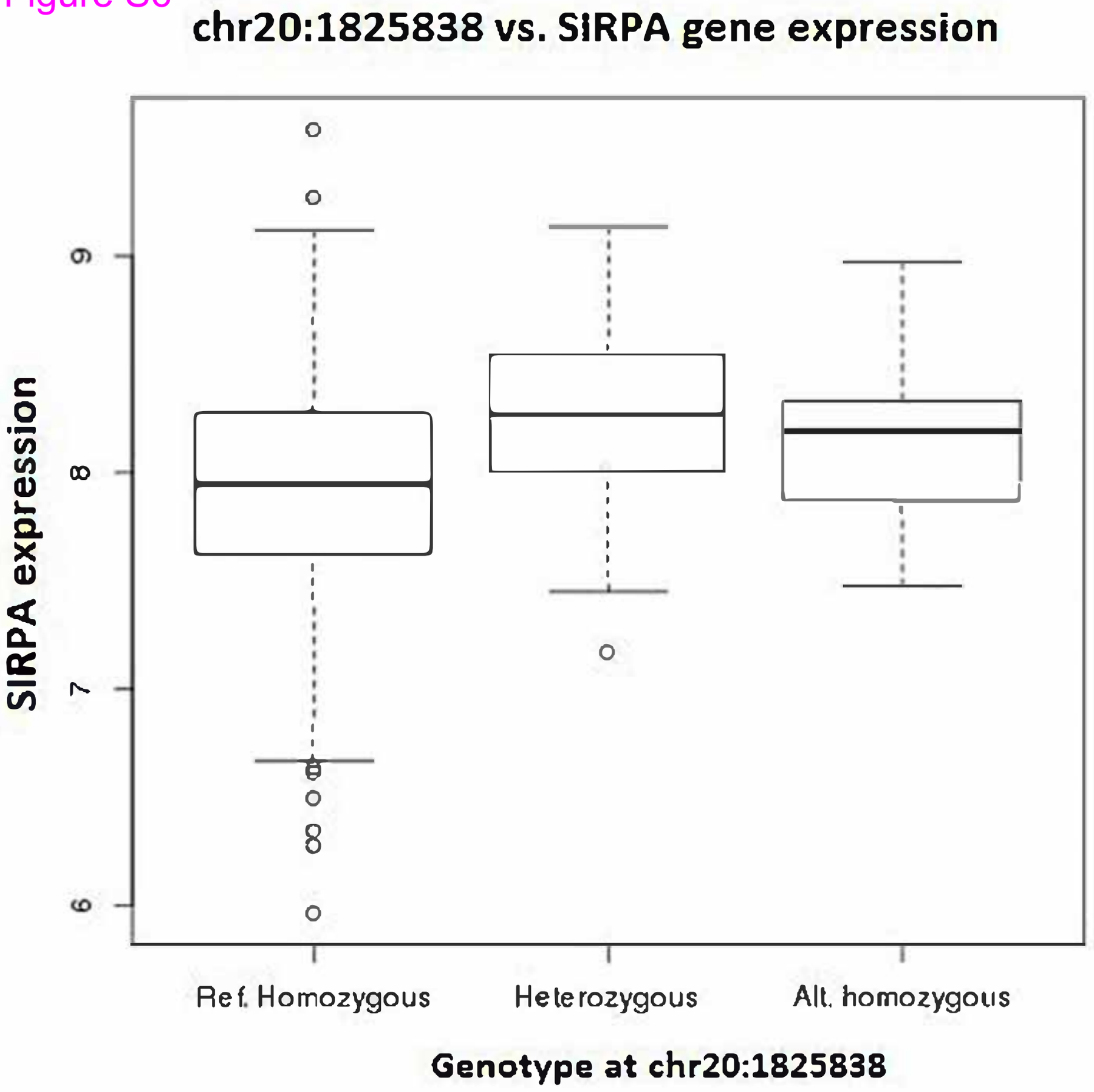
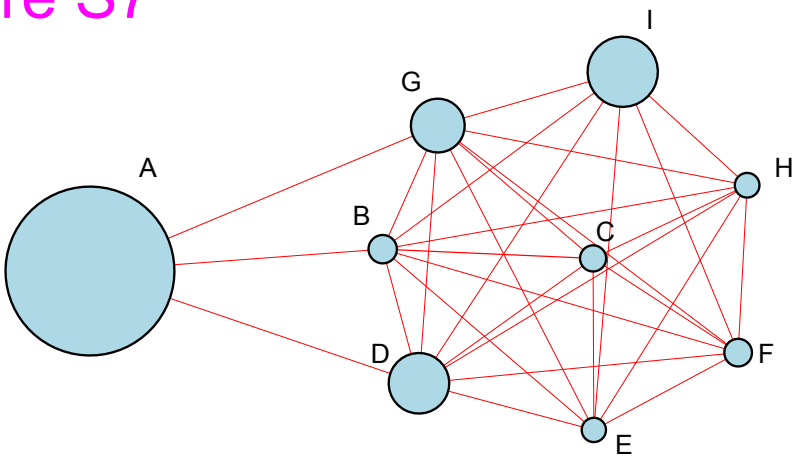
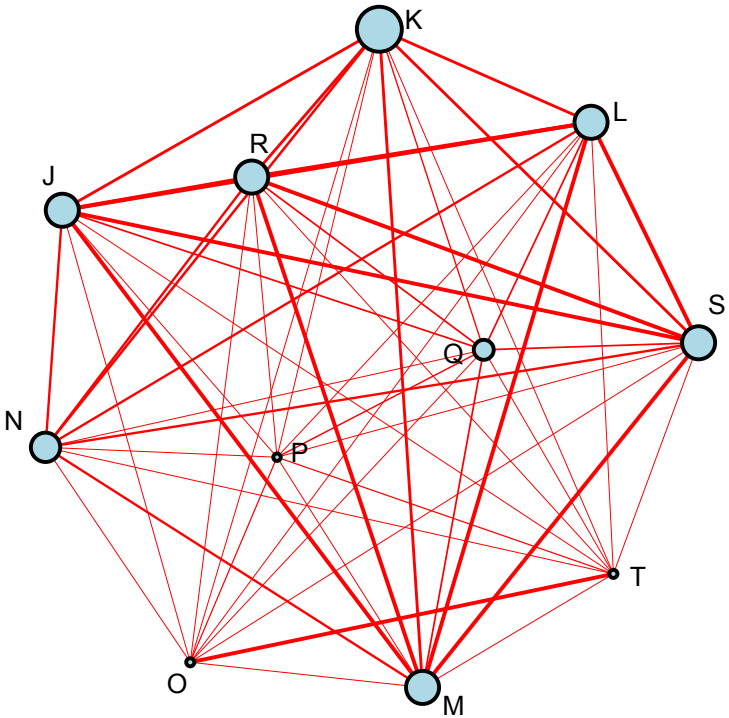

chr20:1825838 vs. SIRPA gene expression
SIRPA expression
9
8
7
6
Ref. Homozygous
Heterozygous
Alt. homozygous
Genotype at chr20:1825838

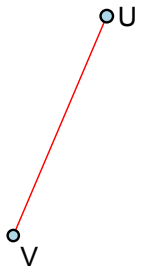
Figure S7



A. EGFR1
B. Prolactin Signaling
C. Prolactin
D. EGF-EGFR Signaling
E. Oncostatin M Signaling
F. IL6
G. BDNF signaling
H. AGE-RAGE
I. Chemokine signaling



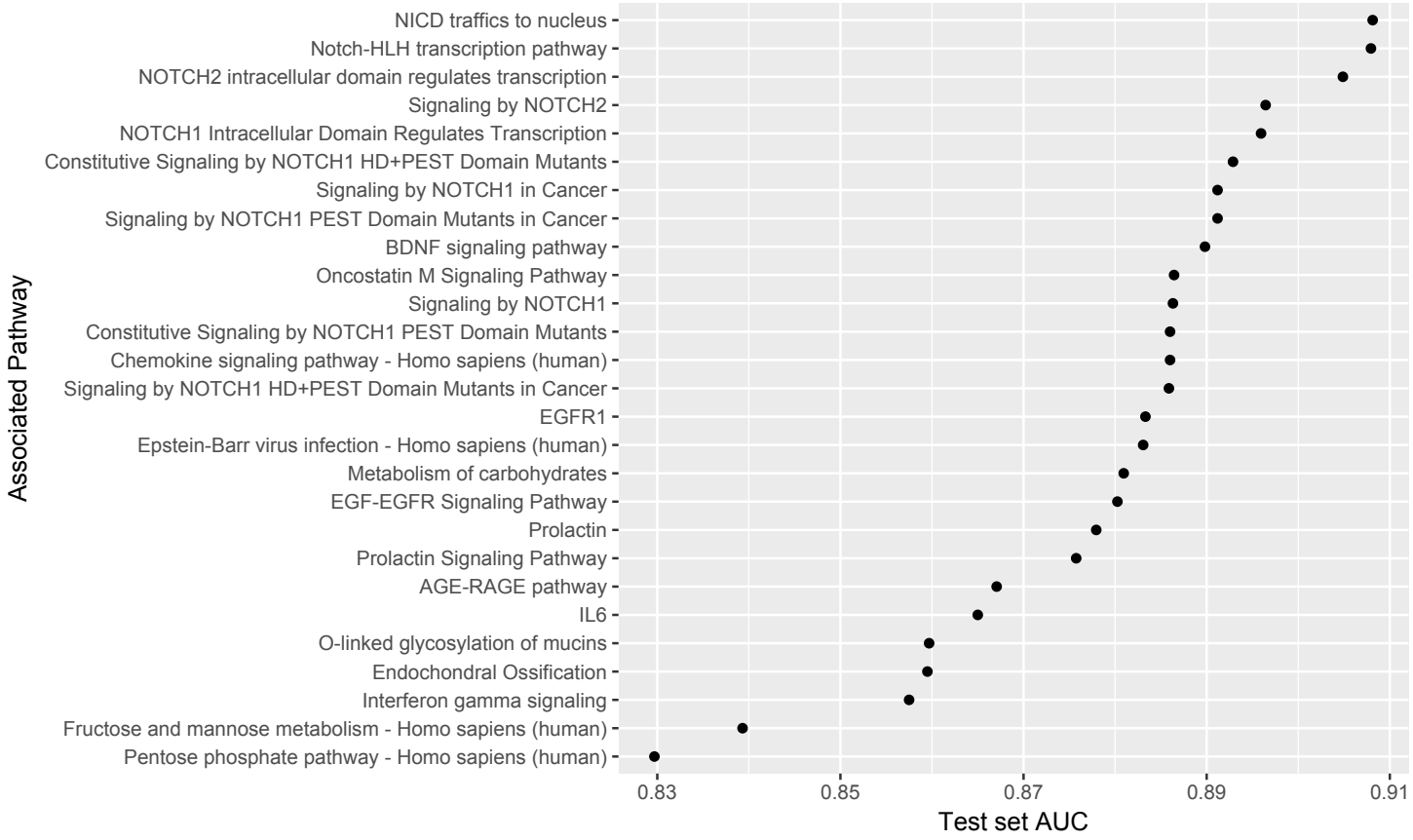
J. Signaling by NOTCH1 PEST Domain Mutants in Cancer
K. Signaling by NOTCH1
L. Signaling by NOTCH1 HD+PEST Domain Mutants in Cancer
M. Constitutive Signaling by NOTCH1 HD+PEST Domain Mutants
N. NOTCH1 Intracellular Domain Regulates Transcription
O. Notch-HLH transcription
P. NOTCH2 intracellular domain regulates transcription
Q. Signaling by NOTCH2
R. Signaling by NOTCH1 in Cancer
S. Constitutive Signaling by NOTCH1 PEST Domain Mutants
T. NICD traffics to nucleus



U. Fructose and mannose metabolism
V. Pentose phosphate pathway

Figure S8

Prediction using Methylation data



Prediction using RNA-seq data

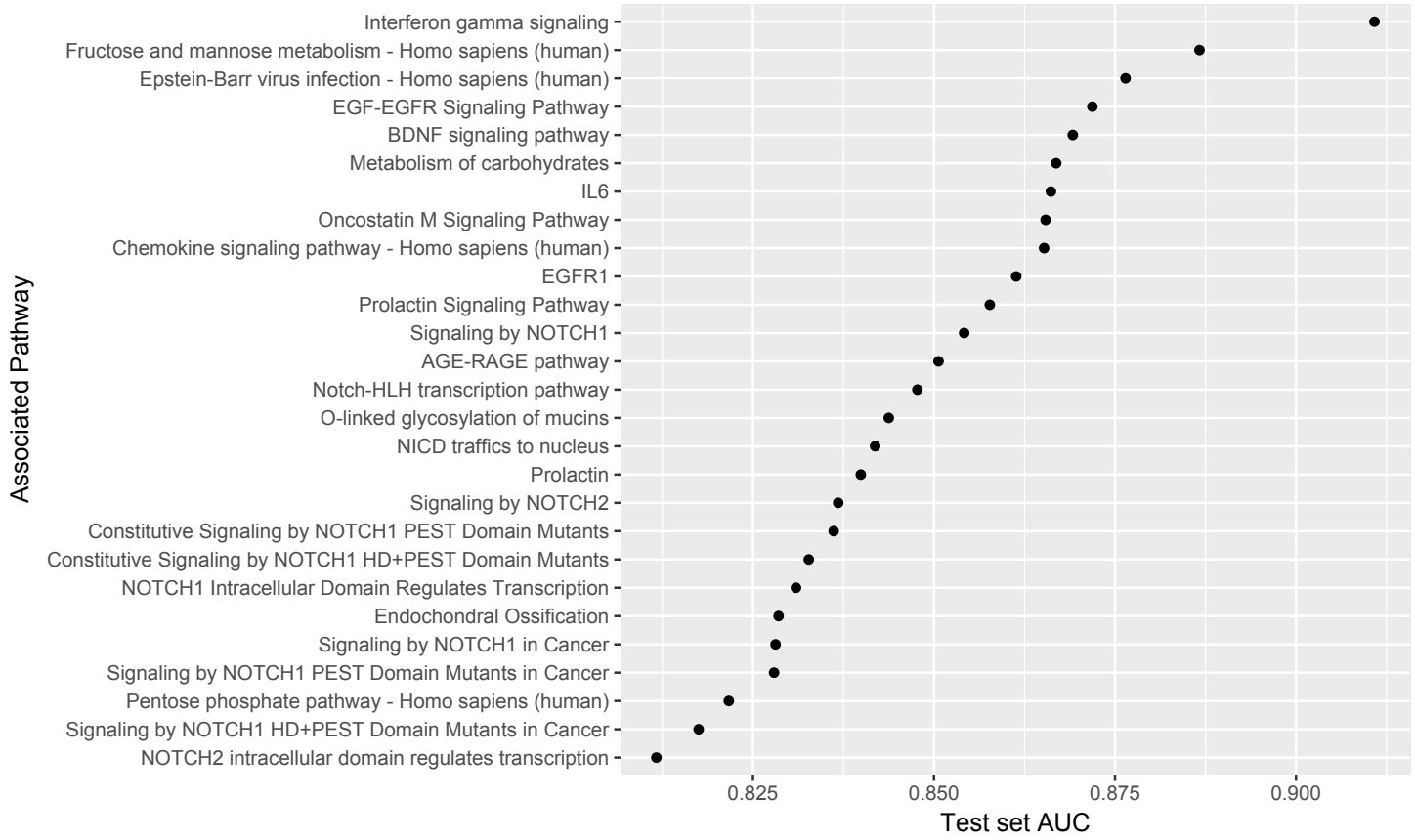


Figure S9

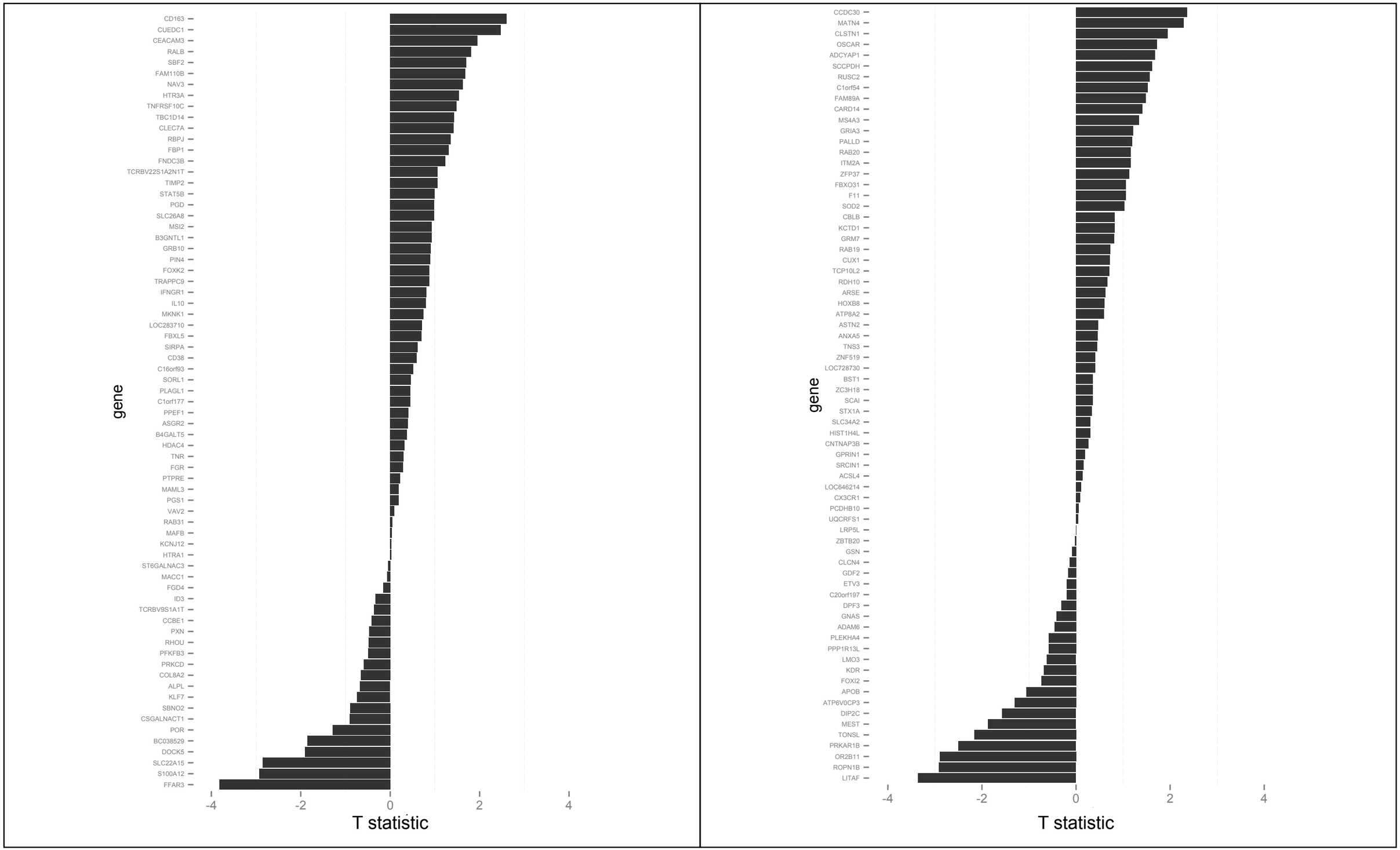
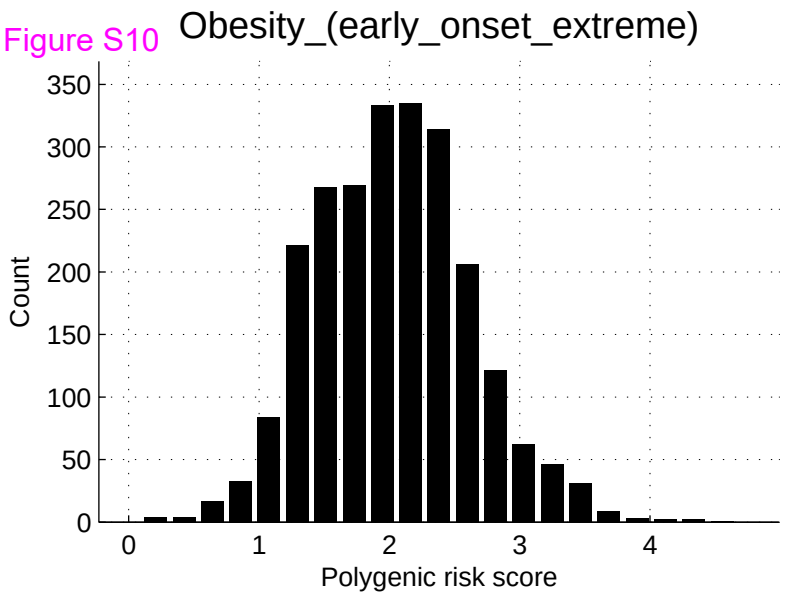
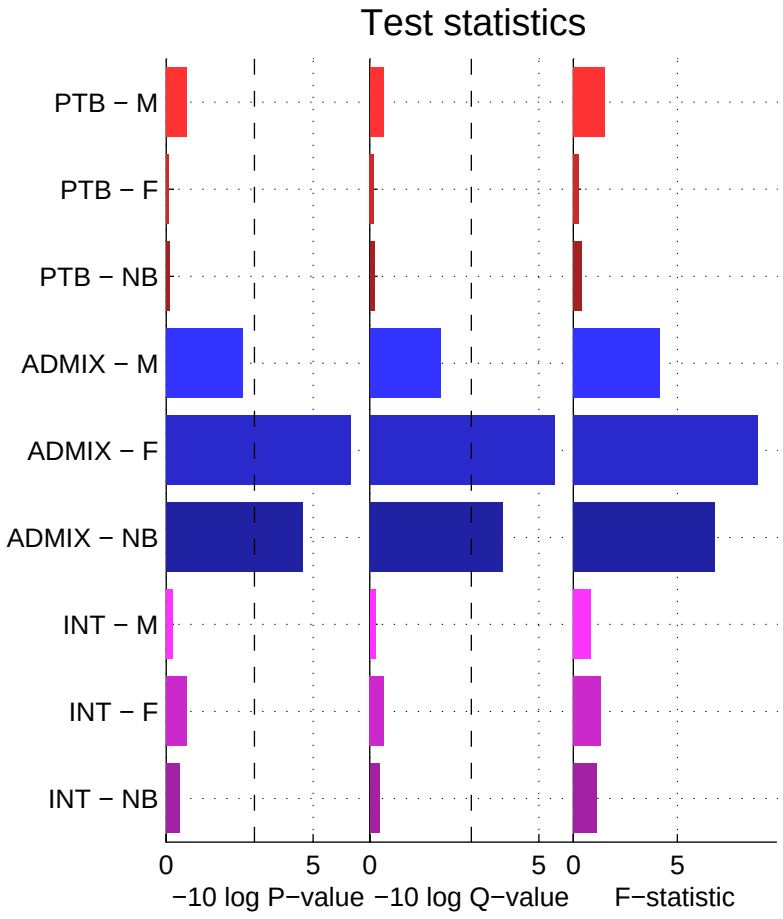
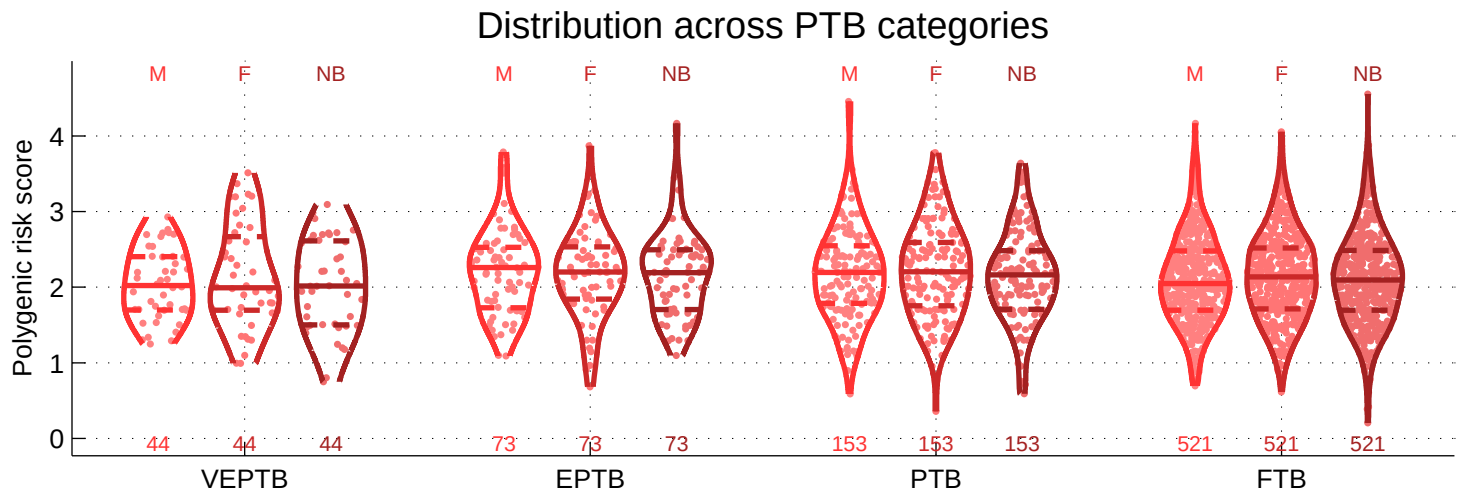
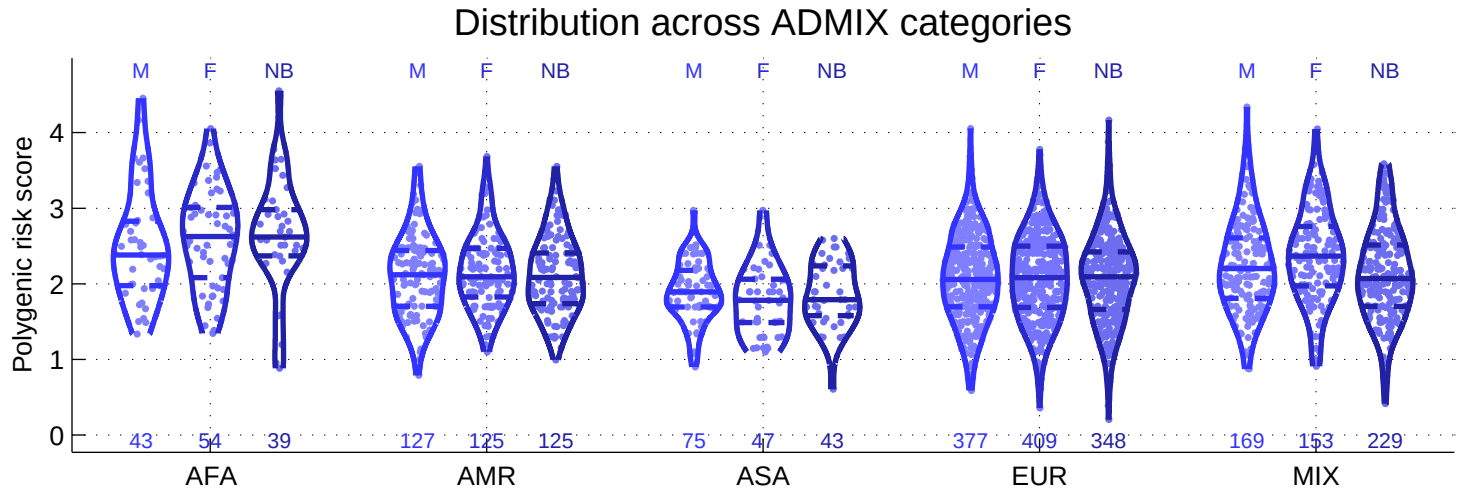

Candidate Genes
CD163
CUEDC1
CEACAM3
RALB
SBF2
FAM110B
NAV3
HTR3A
TNFRSF10C
TBC1D14
CLEC7A
RBPJ
FBP1
FNDC3B
TCRBV22S1A2N1T
TIMP2
STAT5B
PGD
SLC26A8
MSI2
B3GNTL1
GRB10
PIN4
FOXK2
TRAPPC9
IFNGR1
IL10
MKNK1
LOC283710
FBXL5
SIRPA
CD38
C16orf93
SORL1
PLAGL1
C1orf177
PPEF1
ASGR2
B4GALT5
HDAC4
TNR
FGR
PTPRE
MAML3
PGS1
VAV2
RAB31
MAFB
KCNJ12
HTRA1
ST6GALNAC3
MACC1
FGD4
ID3
TCRBV9S1A1T
CCBE1
PXN
RHOU
PFKFB3
PRKCD
COL8A2
ALPL
KLF7
SBNO2
CSGALNACT1
POR
BC038529
DOCK5
SLC22A15
S100A12
FFAR3
gene
-4
-2
0
2
4
T statistic
Random Genes
CCDC30
MATN4
CLSTN1
OSCAR
ADCYAP1
SCCPDH
RUSC2
C1orf54
FAM89A
CARD14
MS4A3
GRIA3
PALLD
RAB20
ITM2A
ZFP37
FBXO31
F11
SOD2
CBLB
KCTD1
GRM7
RAB19
CUX1
TCP10L2
RDH10
ARSE
HOXB8
ATP8A2
ASTN2
ANXA5
TNS3
ZNF519
LOC728730
BST1
ZC3H18
SCAI
STX1A
SLC34A2
HIST1H4L
CNTNAP3B
GPRIN1
SRCIN1
ACSL4
LOC646214
CX3CR1
PCDHB10
UQCRFS1
LRP5L
ZBTB20
GSN
CLCN4
GDF2
ETV3
C20orf197
DPF3
GNAS
ADAM6
PLEKHA4
PPP1R13L
LMO3
KDR
FOXI2
APOB
ATP6V0CP3
DIP2C
MEST
TONSL
PRKAR1B
OR2B11
ROPN1B
LITAF
gene
-4
-2
0
2
4
T statistic

Figure S10


Obesity_(early_onset_extreme)
Count
Polygenic risk score


Test statistics
PTB − M
PTB − F
PTB − NB
ADMIX − M
ADMIX − F
ADMIX − NB
INT − M
INT − F
INT − NB
−10 log P−value
−10 log Q−value
F−statistic


Distribution across PTB categories
Polygenic risk score
VEPTB
EPTB
PTB
FTB


Distribution across ADMIX categories
Polygenic risk score
AFA
AMR
ASA
EUR
MIX

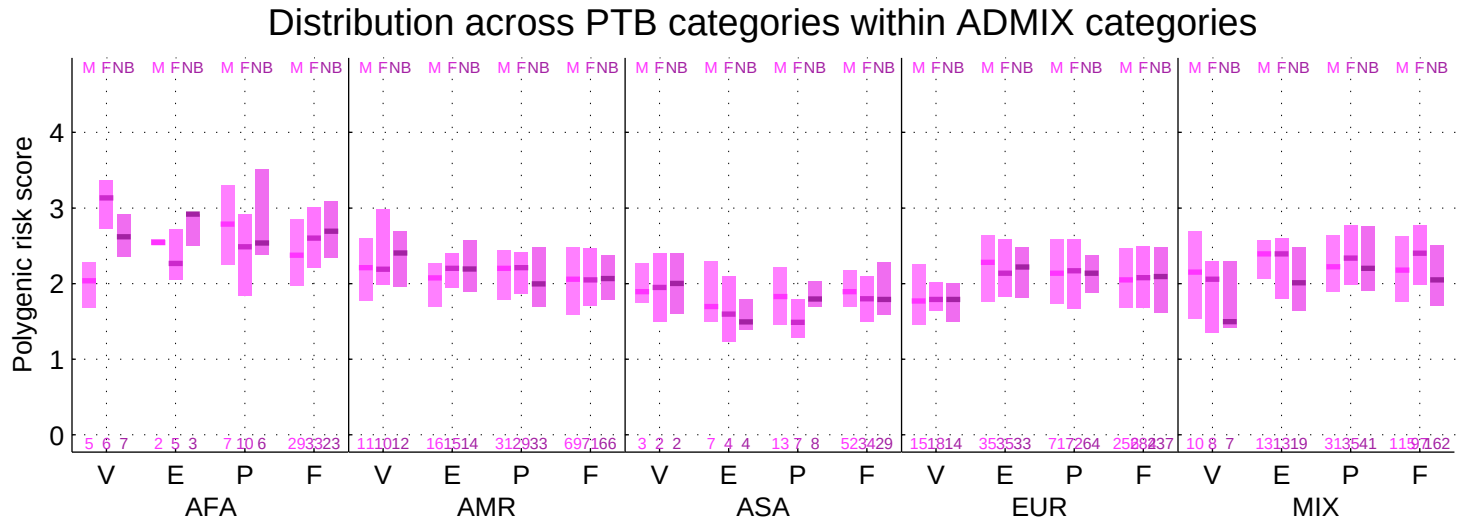

Distribution across PTB categories within ADMIX categories
Polygenic risk score
AFA
AMR
ASA
EUR
MIX

Figure S11

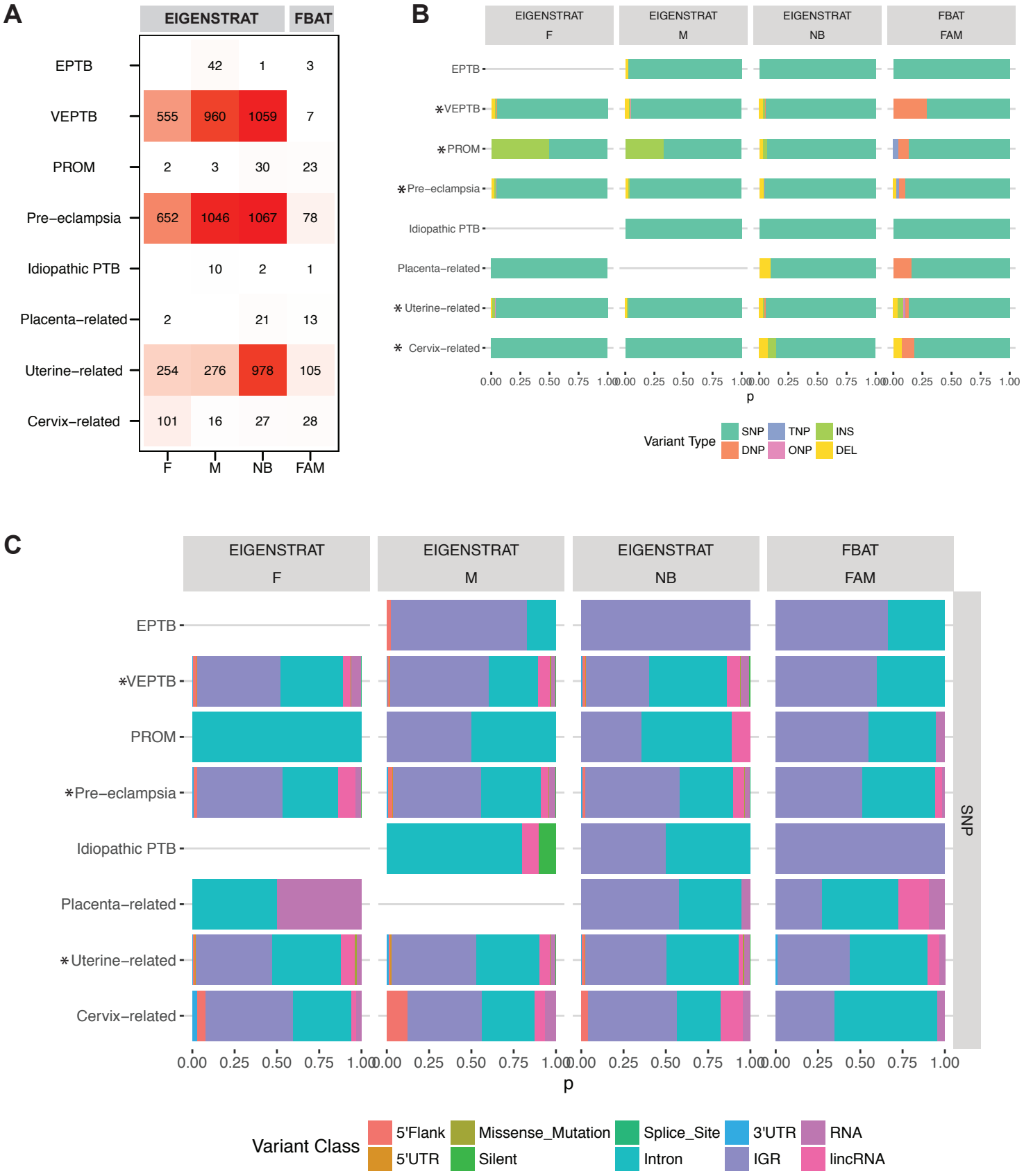
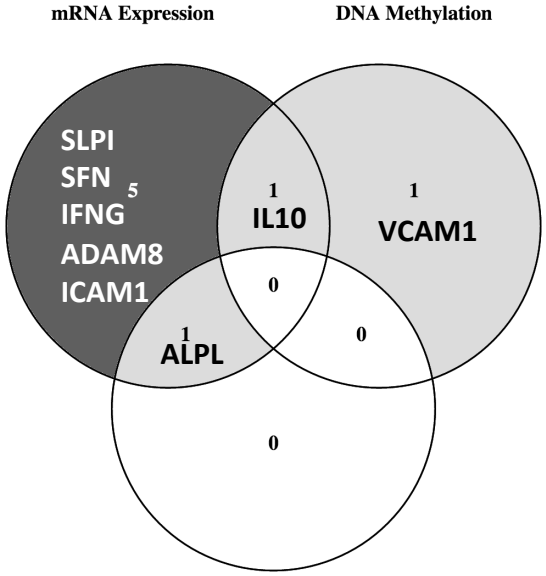
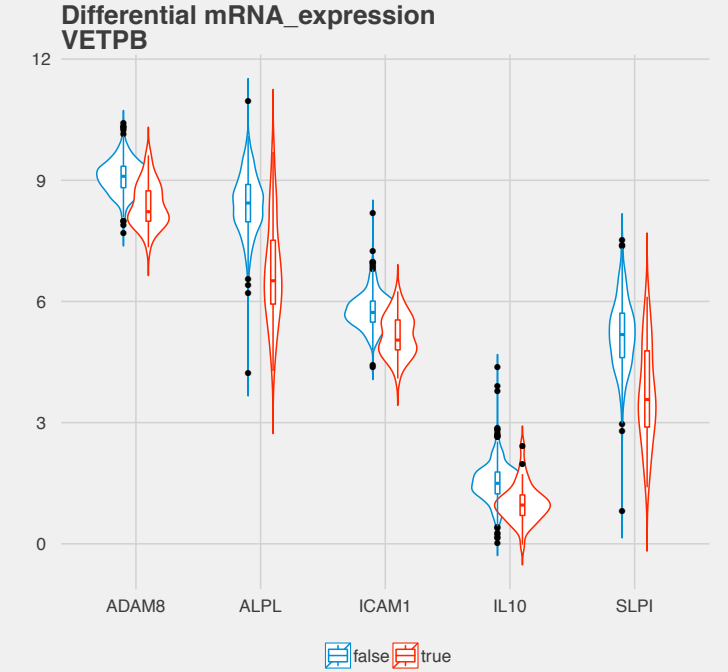
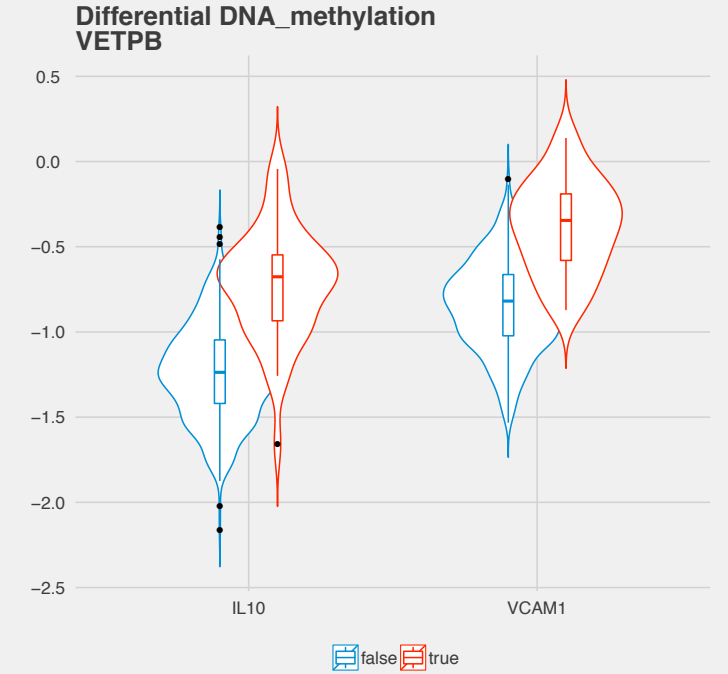

A
EIGENSTRAT
FBAT
EPTB
VEPTB
PROM
Pre−eclampsia
Idiopathic PTB
Placenta−related
Uterine−related
Cervix−related
42 1 3
555 960 1059 7
2 3 30 23
652 1046 1067 78
10 2 1
2 21 13
254 276 978 105
101 16 27 28
F M NB FAM
B
EIGENSTRAT F
EIGENSTRAT M
EIGENSTRAT NB
FBAT FAM
EPTB
∗VEPTB
∗PROM
∗Pre−eclampsia
Idiopathic PTB
Placenta−related
∗Uterine−related
∗ Cervix−related
0.00 0.25 0.50 0.75 1.00
p
Variant Type
SNP
TNP
INS
DNP
ONP
DEL
C
EIGENSTRAT F
EIGENSTRAT M
EIGENSTRAT NB
FBAT FAM
EPTB
∗VEPTB
PROM
∗Pre−eclampsia
Idiopathic PTB
Placenta−related
∗Uterine−related
Cervix−related
SNP
0.00 0.25 0.50 0.75 1.00
p
Variant Class
5'Flank
Missense_Mutation
Splice_Site
3'UTR
RNA
5'UTR
Silent
Intron
IGR
lincRNA

Figure S12


mRNA Expression
DNA Methylation
SLPI
SFN
IFNG
ADAM8
ICAM1
5
1
IL10
1
VCAM1
0
1
ALPL
0
0
DNA Sequence


Differential mRNA_expression
VETPB
12
9
6
3
0
ADAM8
ALPL
ICAM1
IL10
SLPI
false
true


Differential DNA_methylation
VETPB
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
IL10
VCAM1
false
true