SCIENTIFIC REPORTS

OPEN

# A genome-wide association study identifies only two ancestry specific variants associated with spontaneous preterm birth

Received: 19 September 2017
Accepted: 7 December 2017
Published online: 09 January 2018

Nadav Rappoport[1,2], Jonathan Toung[3], Dexter Hadley[1,2], Ronald J. Wong[3], Kazumichi Fujioka[3], Jason Reuter[4], Charles W. Abbott[4], Sam Oh[5], Donglei Hu[5], Celeste Eng[5], Scott Huntsman[5], Dale L. Bodian[6], John E. Niederhuber[6,7], Xiumei Hong[7], Ge Zhang[8], Weronika Sikora-Wohfeld[3], Christopher R. Gignoux[4], Hui Wang[3], John Oehlert[3], Laura L. Jelliffe-Pawlowski[9], Jeffrey B. Gould[3], Gary L. Darmstadt[3], Xiaobin Wang[7], Carlos D. Bustamante[4], Michael P. Snyder[4], Elad Ziv[5], Nikolaos A. Patsopoulos[10,11,12], Louis J. Muglia[8], Esteban Burchard[5], Gary M. Shaw[3], Hugh M. O'Brodovich[3], David K. Stevenson[3], Atul J. Butte[1,2,5] & Marina Sirota[1,2,5]

**Preterm birth (PTB), or the delivery prior to 37 weeks of gestation, is a significant cause of infant morbidity and mortality. Although twin studies estimate that maternal genetic contributions account for approximately 30% of the incidence of PTB, and other studies reported fetal gene polymorphism association, to date no consistent associations have been identified. In this study, we performed the largest reported genome-wide association study analysis on 1,349 cases of PTB and 12,595 ancestry-matched controls from the focusing on genomic fetal signals. We tested over 2 million single nucleotide polymorphisms (SNPs) for associations with PTB across five subpopulations: African (AFR), the Americas (AMR), European, South Asian, and East Asian. We identified only two intergenic loci associated with PTB at a genome-wide level of significance: rs17591250 (P = 4.55E-09) on chromosome 1 in the AFR population and rs1979081 (P = 3.72E-08) on chromosome 8 in the AMR group. We have queried several existing replication cohorts and found no support of these associations. We conclude that the fetal genetic contribution to PTB is unlikely due to single common genetic variant, but could be explained by interactions of multiple common variants, or of rare variants affected by environmental influences, all not detectable using a GWAS alone.**

Preterm birth (PTB), the delivery of an infant prior to 37 weeks of gestation, is a significant determinant of infant morbidity and mortality. Globally, in 2010 an estimated 11% of births, totaling 15 million were delivered preterm[1,2]. In the US, PTB occurs in nearly 10% of all births. Infants born preterm are at risk for a variety of long-term adverse outcomes, such as respiratory illness, blindness, and cerebral palsy, with associated

[1]Institute for Computational Health Sciences, University of California, San Francisco, 94143, CA, USA. [2]Department of Pediatrics, University of California, San Francisco, San Francisco, CA, USA. [3]Department of Pediatrics, Stanford University School of Medicine, Stanford, CA, USA. [4]Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA. [5]Institute for Human Genetics, University of California, San Francisco, San Francisco, CA, USA. [6]Inova Translational Medicine Institute, Inova Health System, Falls Church, VA, USA. [7]Department of Population, Family and Reproductive Health, Center on the Early Life Origins of Disease, Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD, USA. [8]Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. [9]Department of Biostatistics, University of California, San Francisco, CA, USA. [10]Systems Biology and Computer Science Program, Ann Romney Center of Neurological Diseases, Department of Neurology, Division of Genetics, Brigham & Women's Hospital, Boston, MA, USA. [11]Harvard Medical School, Boston, MA, USA. [12]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. Nadav Rappoport and Jonathan Toung contributed equally to this work. Correspondence and requests for materials should be addressed to A.J.B. (email: atul.butte@ucsf.edu) or M.S. (email: marina.sirota@ucsf.edu)

complications resulting in nearly one million deaths each year. PTB is the top cause of death in children under-five years of age[3]. Although survival for most children born premature has improved considerably, they remain at increased risk for a variety of severe neurodevelopmental, gastrointestinal, and respiratory complications, many of which extend well beyond the neonatal period and contribute to lifelong challenges for individuals and their families as well as to burdensome economic costs to society.

While approximately 20% to 30% of PTBs are medically indicated or performed to minimize potential complications from delivery, the vast majority of PTBs are spontaneous, caused by either preterm labor or preterm premature rupture of membranes (PPROM)[4]. The exact mechanism of spontaneous PTB is unknown, though a multitude of social[5,6], environmental[7,8], and maternal factors, such as a shortened interpregnancy interval[9–11], young maternal age[1,2,12–15], extremes of body mass index[16–18], and increased stress levels, have been implicated. Various observations, such as the tendency for recurrent PTBs[19–22] and the increased risk of preterm delivery for women who themselves were born preterm[23,24], suggest a heritable component to the risk of PTB[25]. Twin studies estimate that maternal genetic variants account for approximately 27% to 36% of the incidence of PTB[26,27]. A recent study estimated that 11% of variation is due to fetal genetic effects[28]. A follow up study which focused on quantifying the extent that heritable factors and environmental exposures predict the timing of birth and explain differences between racial groups, found a significant effect in individuals with European ancestry, but not African Americans[29]. Furthermore, PTB rates vary among races and ethnicities, with elevated frequencies observed in African Americans when compared to European Americans[19,22,30,31], and environmental and socio-economic factors alone do not readily account for these differences[32]. Currently, it is unclear how the genetics of PTB differ across racial and ethnic groups.

Despite the evidence for contribution of genetics to the risk of PTB, to date very few reproducible associations have been identified through the genome-wide screens conducted for PTB. The majority of approaches adopted thus far have focused on candidate genes in pathways previously associated with PTB, such as immunity and inflammation[33–37]. The study of the genetics of PTB is further complicated by the likely complex interactions of genes and the environment within these highly interactive causal pathways. Linkage and rare-variant analyses as well as a few genome-wide association studies (GWAS) have been performed to study the genetic factors contributing to the risk of PTB[38], however, replication efforts have been limited. The challenge in carrying out such analyses is further complicated by possible effects arising from maternal and fetal genetic influences.

Several genome-wide linkage-based strategies have been used to identify maternal and fetal genetic variants on chromosomes 15 and X that contribute to PTB in Finnish multiplex and nuclear families[39,40]. Bream *et al*. used candidate gene linkage analyses to identify a set of mutations potentially linked to fetal-mediated PTB and another set of variants linked to maternal-mediated PTB based on 257 families with a family history of PTB[41]. A study by Chittoor *et al*. examined the susceptibility of Mexican Americans to PTB using a linkage analysis across 1,439 individuals and found a significant linkage signal for a region on chromosome 18 containing 52 potential candidate genes[42]. Whole exome sequencing of a limited set of maternal genomes from European families revealed an over-representation of rare variants in the complement/coagulation factor cascade[43]. There have been several large GWAS that have identified variants associated with low birth weight[44–46], but only a limited number have focused on spontaneous PTB. Evolutionary approaches have been applied to identify genes involved in human birth timing[47], and a recent study of mother-infant pairs found two single nucleotide polymorphisms (SNPs) associated with early spontaneous PTB using data collected by the Genomic and Proteomic Network (GPN) for Preterm Birth Research. Neither of these SNP associations, however, could be replicated in independent cohorts[48]. A very recent large scale study identified several genomic regions in European mothers who have delivered preterm[49]. The aforementioned work as well as other studies to date have focused on maternal effects, and not on fetal genetic influences. In the current study, we focused on fetal associations in several populations. Finally, there are several databases summarizing up-to-date findings of the genetics of PTB; however, high-value genetic candidates for diagnostic and therapeutic targets for further study have not been identified[50,51].

In addition to the evidence for the role of genetic factors in PTB, there are pronounced and persistent racial and ethnic disparities observed in the rates of PTB, with significantly higher rates observed in African Americans[52]. The vast majority of genetic studies to date have been performed in European populations, resulting in poor generalizability of these findings to minority populations[53]. Although the factors that underlie this disparity remain elusive, they likely involve complex interactions between genetics, neighborhood-level environmental exposures, and infection and inflammation[54]. A recent study by Hong *et al*. showed that maternal COL24A1 variants have a significant genome-wide interaction with maternal pre-pregnancy overweight/obesity on PTB risk in a population of African Americans[55]. Nonetheless, even after decades of basic science research and public health initiatives, this racial disparity remains poorly understood and relatively unchanged, and specific fetal genetic markers are yet to be found.

In this study, we performed a large ancestry-informed GWAS of over 1,300 infants that were born preterm. To our knowledge, this is one of the largest reported cross-ethnic GWAS performed for this phenotype. Leveraging publicly available data, we were able to employ a large cohort of over 12,000 individuals to serve as a control for our analyses. Using this approach, after extensive filtering and quality control, we identified only two intergenic variants that are statistically significantly associated with PTB and show some evidence in several independent cohorts.

## Results

We carried out a genome-wide association analysis comparing 1,349 extreme preterm infants delivered between 25 and 30 weeks of gestation and documented clinically as spontaneous PTB in the years 2005 to 2008 (Figure S1), with 12,595 ancestry-matched controls. The workflow for this analysis is shown in Fig. 1.

The 1,349 preterm cases were defined as spontaneous PTB deliveries based on birth certificate records. These cases were previously studied in an effort to identify genetic markers for PTB-associated chronic lung disease[56].

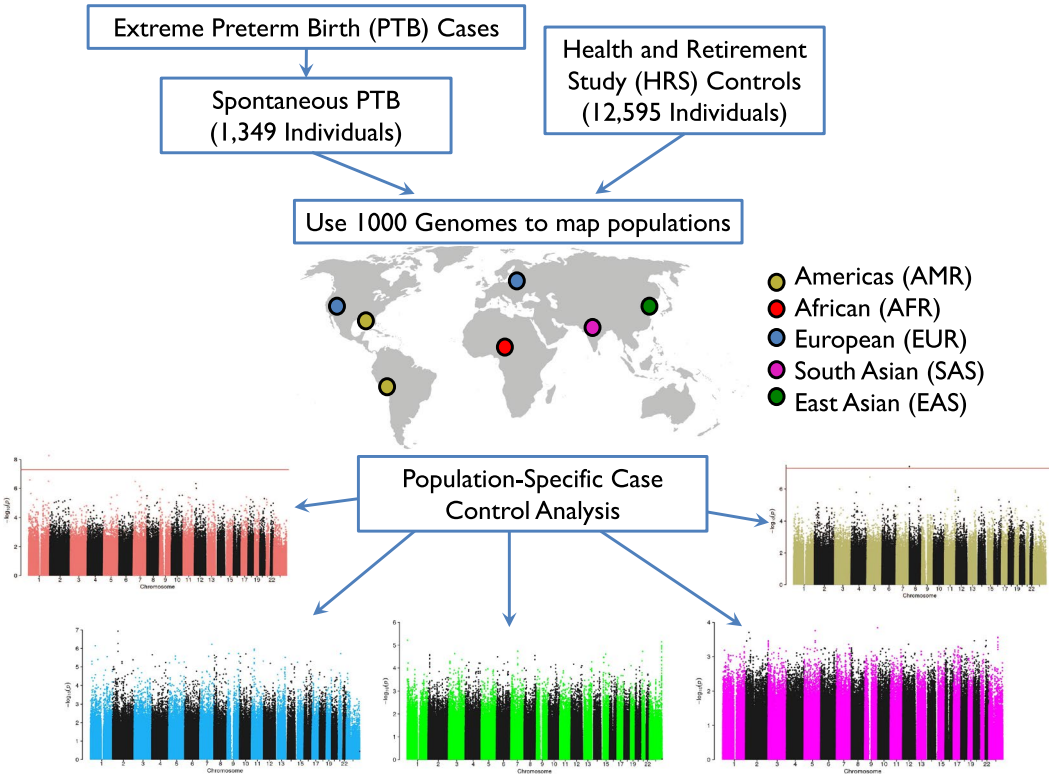


**Figure 1.** Analysis Workflow. We started with a set of PTB cases and chose to focus on those that were born preterm spontaneously (1,349 individuals). We also assembled a large group of control individuals who have a very low chance of being born preterm (12,595 individuals). We used the 1,000 Genomes dataset in order to map both cases and controls to five populations: Americas (AMR), African (AFR), European (EUR), South Asian (SAS), and East Asian (EAS). Finally, we carried out case control analysis in each of the five populations. World map image adapted from Wikipedia: https://upload.wikimedia.org/wikipedia/commons/1/17/BlankMap-World-noborders.png.

The ancestry-matched control population was obtained from the Health and Retirement Study (HRS)[57], the vast majority of whom are older than 55 years of age at the time of recruitment, which started in 1992 (Table S1 and Figure S3). In the early 1990s, major advances in neonatal care, led by the introduction of surfactant therapy[58], greatly increased the likelihood of survival for infants born before 30 weeks[59–62]. Since all of the controls were retirees born well before 1990, we assumed that they were very unlikely to have been born extremely preterm (before 30 weeks), and, therefore, can serve as an appropriate control population for our genome analyses.

To account for differences in population substructures between the cases and controls, we used 1,815 individuals from Phase 3 of the 1,000 Genomes Project as an anchor point to match cases and controls (Fig. 2, Figure S4). Cases and controls were classified into one of the five population cohorts based on their relative ancestral distance to individuals in the 1,000 Genomes Project[63] (See Methods): African (AFR), the Americas (AMR), East Asian (EAS), European (EUR), and South Asian (SAS) using principle component analysis (Table 1).

We tested 2,015,750 SNPs for their association to the spontaneous PTB phenotype in each of the five population cohorts separately, with sex and the first ten principal components of genetic ancestry as covariates in the model. The genomic inflation factor ($\lambda$)[64] was 1.02, 1.05, 1.04, 1.02, and 1.13 for AFR, AMR, EAS, EUR, and SAS subpopulations, respectively, suggesting little evidence for residual population stratification. The Q-Q plots for these two populations show a well-calibrated distribution of p-values with the majority of values falling on the diagonal (Figure S5).

Overall, we identified two loci associated with spontaneous PTB at a genome-wide level of significance (Table 2), both are intergenic loci - one variant in the AFR population and one in the AMR population (Fig. 3). The analyses in the other three (i.e., EAS, EUR, and SAS) populations did not result in any significant genome-wide associations (Figure S6).

The most significant variant in the AFR population corresponded to rs17591250 (OR = 2.814, P = 4.55E-09), an intergenic SNP on chromosome 1. This variant is located between the pseudogene YWHAQP9 (distance ~100 Kb) and RP11-136B18.1 gene (distance ~ 45 Kb). The minor allele frequency of this variant in our AFR case-control cohort of 0.072 is similar to the one on gnomAD[65] of 0.07804 for the AFR population and slightly higher than the one reported in dbSNP (0.041)[66]. The second significant variant was found in the AMR population and corresponds to rs1979081 (OR = 0.5656, P = 3.72E-08) (Table 2). This variant is also intergenic, located

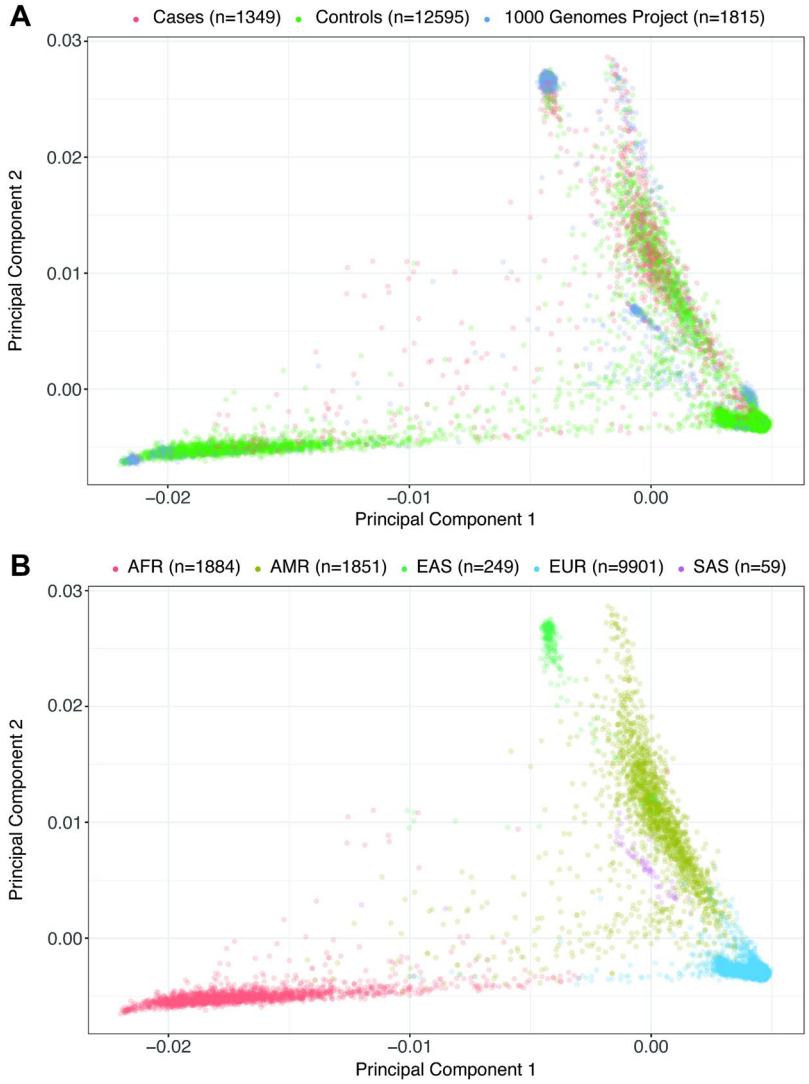


**Figure 2.** Principal components of genetic ancestry across 15,734 individuals. (**A**) This is a PCA plot showing the first and second principal components of genetic ancestry colored by dataset for the three datasets in our analyses. The PTB cases are shown in red, HRS controls shown in green, and 1,000 Genomes Project shown in blue. (**B**) This is a PCA plot similar to the one above, however colored by the five populations that our cases and controls were stratified into.

on chromosome 8, 864 bp upstream to FAM87A (Family With Sequence Similarity 87 Member A) gene and ~22 Kb from FBXO25 (F-Box Protein 25) gene.

To further test whether these results are robust, we imputed the region of each variant after removing the significant variants, and tested whether the association was recovered (see Methods). Both variants were imputed with high certainty (>0.99). In the AFR population, rs146565706, a variant 1931bp from rs17591250, also had a significant association to PTB (P = 3.91E-09). We also identified rs17591327 (560 bp away) with the same P-value as rs17591250 (Figure S7). In the AMR imputation analysis, the variant rs1979081 was not fully recovered, but the region around the variant has several other loci with elevated significance (Figure S7) pointing to the potential association of the region with PTB.

We have queried several replication datasets to see whether our findings are replicated in independent cohorts. We recognize that none of the existing replication datasets that we used were well matched to our discovery cohort based on phenotypic and demographic parameters. The replication results are summarized in Table S2. We analyzed 307 PTB cases and 2,826 controls from the Latino population from the Genes-environments and Admixture in Latino Americans babies (GALA II) study[67] and found that variant rs17591250 on chromosome 1 was modestly significant (OR = 0.7, P = 0.02775;, however, we observed a flip in the odds ratio (OR) for this population. This could be due to the difference in minor allele frequencies for this variant in different ethnic groups (AFR = 0.04 vs. EUR = 0.20 and AMR = 0.19 dbSNP build ID: 123/150[66]). We did not observe significant association for rs1979081 in this cohort.

| Population | Sex | PTB Cases | HRS Controls |
|---|---|---|---|
| EUR (n = 9,890) | Female | 121 | 5,575 |
| | Male | 139 | 4,055 |
| | Total | 260 | 9,630 |
| AFR (n = 1,874) | Female | 103 | 1,078 |
| | Male | 87 | 606 |
| | Total | 190 | 1,684 |
| AMR (n = 1,847) | Female | 337 | 672 |
| | Male | 408 | 430 |
| | Total | 745 | 1,102 |
| EAS (n = 249) | Female | 72 | 77 |
| | Male | 59 | 41 |
| | Total | 131 | 118 |
| SAS (n = 59) | Female | 7 | 17 |
| | Male | 16 | 19 |
| | Total | 23 | 36 |

**Table 1.** Number of spontaneous PTB cases and HRS controls in the presented GWAS.

Another validation cohort was obtained from Inova Translational Medicine Institute (ITMI, Inova Health System, Falls Church, VA). This cohort consists of 800 trios (mother, father, and infant) on whom whole genome sequencing was performed. In this cohort, cases were defined as early preterm ($\leq$34 weeks of gestation) and controls as full-term ($\geq$37 weeks of gestation). The analysis of the infants' genomes was carried out separately in the AFR and the EUR subpopulations (see Methods). We found rs17591250 was nominally significant in the EUR population (OR = 0.563, P = 0.04842), however we again observed a flip in the OR, which could be explained by the difference in allele frequencies in the different populations. In the ITMI cohort, we also examined variants within 1Kb of the associations and were able to identify regional support for the association (Table S3). We did not observe significant association for rs1979081 in this cohort, but we saw some regional support. The fact that the p-values are close to 0.05 and effect directions are different in the validation sample indicate a high uncertainty of the estimation.

In addition, we examined maternal genomic influences in these regions using several additional existing cohorts. The FIN cohort is composed of over 800 European mother/child pairs. We queried this cohort for variants within 1Kb of the associations (see Methods). In the mothers' cohort, 6 variants within the 1Kb window around rs1979081 had an association P-value $<$0.05, and OR ranged between 1.7515 and 1.8143 and in the babies' cohort a single variant was statistically associated (OR = 1.3966, P = 0.04373) (Table S4). While we observe some statistical association in this region, the evidence is weak, which might point to the fetal vs. maternal genetic effects on PTB. We did not observe any significant association for the region surrounding rs17591250 in this cohort.

Finally, we carried out an association analysis using the Boston Birth Cohort (BBC) as an independent cohort. This cohort consists of 698 African-American mothers who had preterm deliveries and 1,035 mothers of term controls (See Methods). We were not able to validate our findings in this cohort, which might point to the fetal vs. maternal effects on these genetic associations.

## Discussion

In this report, we present the largest reported genome-wide study of early spontaneous PTB with over 1,300 cases of early spontaneous PTB compared with 12,000 ancestry-matched control individuals obtained from a separate independent study. We identified only two loci that were significantly associated with PTB in the AFR and AMR cohorts.

The most significant AFR variant discovered was rs17591250, an intergenic variant on chromosome 1 between the pseudogene YWHAQP9 and RP11-136B18.1 gene. It is ~300 Kb away from the Zona Pellucida Glycoprotein 4 (ZP4) gene. ZP4 is an extracellular matrix protein that surrounds the oocyte and early embryo. The most significant AMR variant found was rs1979081. The variant is also intergenic on chromosome 8 between FAM87A and FBXO25. Two other genes are ~130 Kb farther: ZNF596 (Zinc Finger Protein 596) and TDRP (Testis Development Related Protein). TDRP is known to be highly expressed in the placenta[68,69]. FBXO25 is mostly expressed in the testis. FAM87A and ZNF596 were found to be highly expressed in the brain and testis and FAM87A was also highly expressed in the pancreas[68,69]. This SNP is a known eQTL in different tissues for the lncRNA RP11-91J19.4 gene (Consortium [68]), which is highly expressed in the brain. The two significant variants found are in intergenic regions, and we cannot yet ascribe any direct effect from these SNPs.

There are several limitations of this study that need to be noted. While the case and control populations were genotyped on the same platforms and the cases and controls were matched based on their ancestry using a clustering approach, the study design was not optimal. We have carried out extensive quality control to eliminate any potential batch effects, but ideally the cases and controls should emerge from the same population at risk and be genotyped in a single pass to avoid potential bias and batch effects. Also, once the cases were stratified by the five ancestral populations, the number of cases in each ancestry population was relatively small. Due to the small sample size for the SAS and EAS cohorts in particular, and the extreme unbalanced case-control populations for the

| CHR | SNP | BP | Minor allele | OR | P-value | POP | Genes (distance in bp) | Case genotype counts | Control genotype counts | Validated |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | rs17591250 | 238386465 | G | 2.814 | 4.55E-09 | AFR | RP11-136B18.1 (45329) | GG:2,GA:52 | GG:3,GA:204 | G,T |
| | | | | | | | RP11-136B18.1 (45329) | AA:127 | AA:1451 | |
| 8 | rs1979081 | 334288 | A | 0.566 | 3.72E-08 | AMR | FAM87 (864) | AA:19,AG:147 | AA:38,AG:352 | M,I |
| | | | | | | | FBXO25 (22520) | GG:535,00:35 | GG:665,00:29 | |

**Table 2.** SNPs with genome-wide significant associations with spontaneous PTB. OR and p-values for the other discovery and validation cohorts are given in Table S2. OR - odds-ratio. POP - The population in which the association was identified. Missing genotypes are labeled as '00'. M, I, G, T P-value $<0.05$ in external cohort FIN-mothers, FIN-infants, GALA II, ITMI respectively.

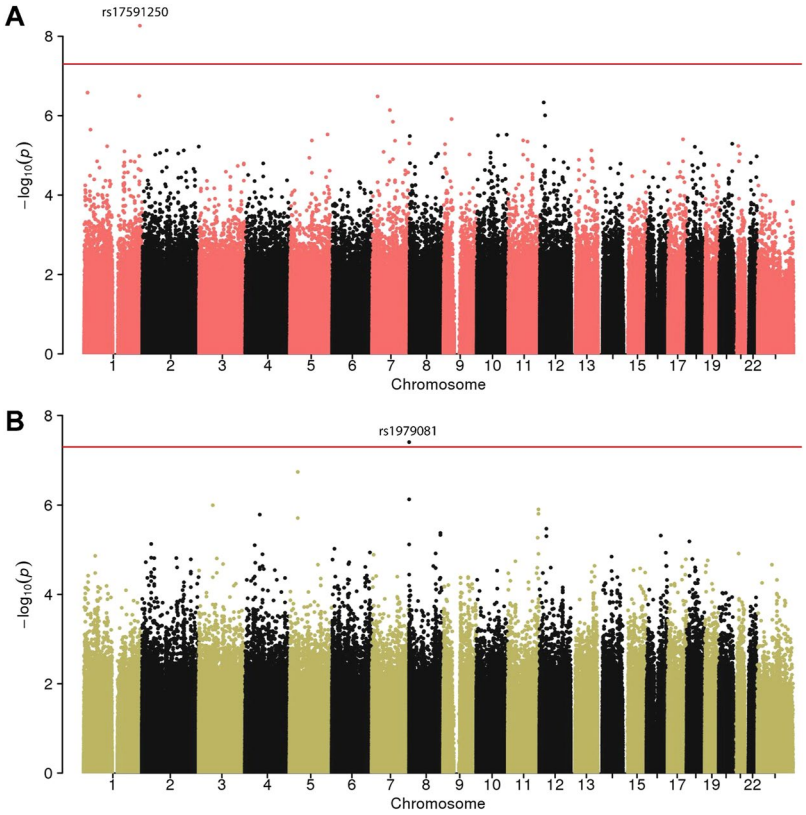


**Figure 3.** Manhattan Plots in the African (AFR) and Americas (AMR) populations. R (**A**) and AMR (**B**) populations. The chromosomes are on the x-axis and the -log (P-value) is shown on the y-axis. Genome-wide significant variants are shown above the dotted line.

EUR, we were not able to identify any variants that passed the genome-wide significance threshold for an association with PTB. We used strict filtering tests and thresholds to remove variants with low-call scores, missing or differentially missing in cases and controls; and therefore, we might not have been able to identify some relevant associations. We assumed that the HRS was a reliable cohort for use as a control set for the PTB phenotype due to the low survival of preterm babies prior to 1990. Other phenotypic factors such as age[70,71], BMI[16], and others need to be considered in future analyses. Finally, we define PTB here as a binary trait, while another approach would be to model gestational age at delivery as a continuous phenotype.

The limitations of our validation cohorts should also be recognized. None of the cohorts ideally mimic our discovery population, which is focused on early spontaneous PTBs. The validation cohorts are not as strictly phenotyped, are sampled from diverse ethnic backgrounds and several of them were not collected to study preterm birth specifically. The validation cohorts are also limited by sample size. Finally, while several of the cohorts contain a limited number of fetal samples, some of them focus on maternal effects, and albeit sharing half of the DNA with the fetus, these can only be used as supporting evidence and cannot be interpreted as validation. Much more rigorous validation of our findings with better matched cohorts is needed.

Using genotype approaches (i.e., GWAS), we were only able to identify markers (e.g., tagSNPs) associated with PTB and not necessarily any causal variants. Fine mapping gene sequencing or use of a whole genome sequencing approaches are needed to exactly locate causal mutations that might be responsible for the phenotype. With numerous sequencing and genotyping efforts underway, we hope to leverage some of these resources for further

validation and analyses. Finally, we have focused our work here on identifying potential genetic risk variants associated with PTB; however, we recognize that this is only a part of the complete picture. In the current analysis, environmental factors were not captured even though previous work had shown that these factors might have an effect[36]. Future studies should be designed to allow us to explore pertinent environmental exposures and examine gene-environment interactions in the context of the phenotype of interest.

While the sample size of this study is considerable, it does seem like the study is still underpowered to identify robust fetal genome-wide associations and additional studies are needed. Our results highlight only two significant variants, though of relatively small effect, throughout the genome suggesting that there are more elusive sources of heritability to consider for PTB. First, the GWAS we performed is best powered to find common variants of small effect in large populations[72]. However, human genomes are rich in structural diversity outside of SNPs that are individually rare, but collectively common in the human population[73,74] and result in large genomic spans of deletions, duplications, and inversions that impact genes with large effects. Second, differences in genetic architecture reflect the complex, often opposing effects of selection, population history, migration and mutation rates[75], and the cumulative genetic effects of these influences are also likely to be common in complex phenotypes[76,77]. Third, and for PTB in particular, currently unknown gene-environment (GxE) interactions most likely contribute to the phenotype. These and other shortcomings[75] can be addressed with more sophisticated models of disease genetics of PTB that depend on deeper phenotyping and sequencing of subjects to identify the interaction of environmental triggers and rare variants that likely contribute to the phenotype.

In conclusion, we have presented an ancestry-informed GWAS of PTB including over 1,300 early spontaneous PTB cases and over 12,000 ancestry-matched controls across five populations. We leveraged publicly-available genotypic data to identify a large control group that is highly likely to be depleted for the phenotype of PTB, and we also mapped cases and controls based on their ancestry lineage to carry out population-specific case-control analyses. The analytical approach we present here can be extended to study other phenotypes of interest. Finally, we identified two different variants associated with PTB in two different populations. Neither of the associations we identified has previously been linked to PTB to our knowledge, and the biological significance of these associations has yet to be determined. While additional validation is needed, we anticipate that there might be other variants in these genomic regions that may contribute to a wide range of varying capacities for expression among the general population that singly or in combination might contribute to individual differences in response to environmental stressors and the risk of PTB.

## Methods

**Case and control cohorts.** The 1,349 cases used in this study were obtained from the California Perinatal Quality Care Collaborative (CPQCC) database during the 2005 to 2008 calendar years. Inclusion criteria included gestational age 25 to 29 and 6/7 weeks (Figure S1), birthweight less than 1500 g (Figure S2), and spontaneous delivery defined by PPROM, premature labor, or tocolysis as recorded on birth certificates. These cases were previously analyzed as part of a GWAS on bronchopulmonary dysplasia and additional inclusion criteria are detailed in this prior study[56].

The 12,595 controls used in this analysis were obtained from the University of Michigan HRS (http://hrsonline.isr.umich.edu/gwas), a longitudinal survey of Americans who ranged from 50 to 80 + years old at the time of data collection (Table S1; Figure S3). Data were downloaded from dbGap (accession pht002614.v1.p1) on January 2016.

**Genotyping.** For the cases, genomic DNA was extracted from bloodspots[78] and genotyped using the HumanOmni2.5-4 v1 BeadChip (Illumina, San Diego, CA), as previously described[56]. For the controls, saliva was collected and genotyped by the NIH Center for Inherited Disease Research using the same Human Omni2.5-4 v1 BeadChip methodology, as previously described.

We merged raw intensities data (*.idat) files of cases and controls and carried out joint calling, to reduce batch effects. Clustering and genotyping were performed using GenomeStudio software 2011.1 (Illumina, 2011). Quality control procedures were performed, including filtering by call rate, mismatch between observed and reported gender and possible gender abnormalities as described in Wang *et al.*[56], and tri-allelic and ambiguous (AT/CG) SNPs were removed. Overall, a total of 2,015,750 SNPs were used in our downstream analyses.

**Ancestry matching of cases and controls.** To minimize potential confounding factors introduced by population stratification, we matched cases and controls to defined population cohorts as provided by the 1,000 Genomes Project. We performed PCA on our cases, controls, and 1,815 individuals from the 1,000 Genomes Project using a subset of 102,008 common (minor allele frequency greater than 5%) autosomal SNPs that are in Hardy-Weinberg equilibrium. These SNPs are also relatively independent, with pairwise linkage disequilibrium values (R2) less than 0.20. Next, utilizing population labels provided for individuals in the 1,000 Genomes Project and the top 10 principal components, we inferred population labels for our cases and controls using the k-nearest neighbor algorithm[79] (Fig. 2 and Figure S4).

**Association analysis.** We analyzed potential associations between 2,015,750 SNPs and spontaneous PTB status using logistic regression in PLINK v1.90b3.42[80]. We stratified our analyses by population (inferred by principal components of genetic ancestry) and used the top ten principal components of genetic ancestry and sex as covariates. Variants were filtered out in case of missing calls >5% either in cases or controls, or P-value of differential missingness between cases and controls <0.001, or Hardy-Weinberg equilibrium <0.001 in cases and controls separately or jointly, or MAF <0.001. Outlier samples were filtered out if standard deviation using first 10 first principal components was larger than 6. This was performed iteratively re-computing PCA and removing

outliers until no more outliers were detected using EIGENSTRAT software[81]. Sample outliers with identity by descent >0.2 were removed as well. We confirmed the results visually for the lack of population structure by plotting cases, control and matching 1,000 Genome Project samples by the top two principle components (Figure S8)

**Manual inspection of variant intensities.** In order to ensure the quality of the genotyping data, following joint calling of cases and controls using GenomeStudio software, we manually inspected the variant calling and clustering for the two variants that we identified as significantly associated in the study (Figure S9).

**Manhattan plots.** Manhattan plots were generated in R version 3.3.1 using qqman library version 0.1.2[82]. Significance line was plotted at the level of genome-wide significant threshold of 5E-8. Zoomed-in manhattan plots were generated using the LocusZoom web tool[83].

**Imputation.** For each one of the two significant variants found, we extracted the genotyping calls from ± 1mbp excluding the variant of interest. Prephasing was done with EAGLE2 v2.3.1[84] and imputation using IMPUTE2[85] and 1000 G V3 as a reference.

**Replication cohorts.** Validation of associations has been carried out in three independent cohorts: (i) The Study of African Americans, Asthma, Genes and Environments (SAGE II)[67] (ii) Genes-environments and Admixture in Latino Americans babies (GALA II) study[67] (iii) Inova Translational Medicine Institute's (ITMI) PTB study cohort.

*SAGE II cohort.* The Study of African Americans, Asthma, Genes, and Environments (SAGE II) is a gene-environment interaction study of asthma in African-American children in the USA initiated in 2006. Over 1,500 Individuals aged 8–21 years were collected from the San Francisco Bay during 2006–2013. Available genotyping data cover 114 PTB cases (labor prior to week 36) and two sets of controls: 996 controls born non-preterm (after 36 weeks of gestation) and a subset of 225 individuals who were born full-term (≥40 weeks of gestation). DNA was extracted from whole blood samples and genotyped with the Affymetrix Axiom LAT1 array[86].

*GALA II cohort.* The GALA II is a case-control study using protocols and questionnaires as described in Nishimura *et al.*[67]. Subjects were recruited from five urban study centers across the mainland US and Puerto Rico (n = 7,683). Genome-wide SNP genotypes are from the Axiom™ LAT1 array (Affymetrix, >800,000 SNPs). Latino children were recruited from the San Francisco Bay Area, Chicago, Houston, New York City, and Puerto Rico. Although the original dataset was collected to study asthma and drug response in these children, several questions related to birth timing were part of the original questionnaire, including whether the child was born early, on time or late and how many weeks early or late. Using the questionnaire data, we were able to identify PTB case and control populations in order to investigate the genetic determinants that are responsible for the observed phenotype of PTB. The 307 cases here are individuals who reported being born between 25 and 36 weeks of gestation. We used two set controls: the first one includes 2,826 individuals who were born non-preterm and the second is a subset of 309 individuals who were born full-term (≥40 weeks of gestation).

*ITMI cohort.* A validation cohort was obtained from Inova Translational Medicine Institute (ITMI), Inova Health System, Falls Church, Virginia. This cohort consisting of 800 trios (mother, father, and infant) on which whole genome sequencing was performed. In this cohort, cases were defined as early preterm (≤34 weeks gestation) and controls as full-term (≥37 weeks gestation). The analysis of the infants' genomes was carried out separately in the AFR (31 early preterm cases and 53 controls) and the EUR (82 early preterm cases and 249 controls) sub-populations.

**Study participants and whole genome sequencing**: Trios comprised of neonates and both parents were enrolled in Inova Translational Medicine Institute's "Molecular Study of Pre-term Birth"[87]. Gestational ages of the infants ranged from 22 to 41 weeks. Whole genome sequencing of peripheral blood samples collected from each participant was performed at Complete Genomics (Mountain View, CA), and sequencing data were processed as described[87]. The genomic data were used to assign ancestry using principal components analysis with the 1,000 Genomes Project data as reference as described above.

**Variant filtering**: Filtering required genotypes to be fully called with GQ ≥42, allele balance >0.225, read depth ≥9, and call rate ≥80%, with variants annotated as VQLOW excluded. Vt[88] was used for multi-allelic variant decomposition, block substitution decomposition, and variant normalization. The resulting genotype calls were converted to bed format with PLINK v1.90b2a.

**Association analyses**: For each ancestry group, association tests were performed on early preterm (≤34 weeks gestation) vs. full-term (≥37 weeks gestation) infants from singleton births. The association testing was performed with a logistic model using gestational age, gender, and the eigenvectors from the ancestry-based PCA as covariates as described above. The AFR group was comprised of 84 neonates, 31 preterm cases and 53 controls. The EUR group had 331 infants, 82 cases and 249 controls. The association analyses and genotype frequencies were computed with PLINK v1.90b2a. We have also examined variants within 1Kb of the associations.

**Additional maternal cohorts.** In addition to replication cohorts, we looked at maternal genetic influences in the regions we found in two additional cohorts (i) The Boston Birth Cohort (BBC); (ii) The Finnish cohort (FIN).

*BBC cohort.* The GWAS of PTB in the Boston Birth Cohort (BBC) is funded by NICHD (R01 HD041702, Principal Investigator: Wang, Xiaobin). This study includes 698 mothers with PTB (ranging from 230/7 to 366/7 weeks of gestation) and 1035 mothers with term births (370/7 to 430/7 weeks). All of these mothers

are African Americans. The preterm and term mothers were matched on maternal age (±5 years), parity and years of delivery (in frequency). Maternal DNA was extracted from white blood cells and genotyped using the Illumina HumanOmni2.5 array. SNP data filtering followed the NIH GENEVA consortium published protocol. The recruitment and characteristics of the BBC has been described previously[89,90]. Eligible mother-infant pairs were recruited 1 to 3 days post-delivery. Mothers who delivered singleton livebirths were included in the study. Pregnancies that were a result of *in vitro* fertilization or multiple gestations, fetal chromosomal abnormalities or major birth defects were excluded. Analyses used the imputed SNPs, with genotype dosage as the predictor. The analysis was done using SNPTEST2 (-Frequentist function), with adjustment of batch, the first three principle components from PCA, infant's gender and maternal age and parity.

*FIN cohort.* The Finnish cohort (FIN) cohort includes genotypes of 817 babies (253 cases) and 888 mothers (334 cases) and gestational age on labor[49]. Mother/child pair samples were collected from the Helsinki (southern Finland) University Hospitals between 2004 and 2014. Regression to gestational week on labor was performed for babies and mothers separately using imputed allelic dosage data assuming additive allelic effects. Maternal age and infant gender were included as covariates. We queried this cohort for variants within a 1Kb of our associations.

**Data Sharing and Availability.** The complete summary statistics and results files are available through ImmPort (http://www.immport.org/) SDY1205, https://doi.org/10.21430/M37N6PJEQT. Results of the analysis and interactive visualization is available as an R Shiny App at http://comphealth.ucsf.edu/ptb_gwas.

**Ethical statement.** The controls from this study came from a publicly available dataset called the Health and Retirement Study[57]. The cases were obtained from California blood spots and are restricted to data sharing. California has determined that all requests for the use of California Biobank Program biospecimens for research studies will need to seek an exemption from NIH or other granting or funder requirements regarding the uploading of study results into an external bank or repository (including into the NIH dbGaP or other bank or repository). This applies to any uploading of genomic data and/or sharing of these biospecimens or individual data derived from these biospecimens. Such activities have been determined to violate the statutory scheme of the California Health and Safety Code Sections 124980(j), 124991(b), (g), (h) and 103850(a) and (d), which protect the confidential nature of biospecimens and individual data derived from biospecimens.

## References


1. Hediger, M. L., Scholl, T. O., Schall, J. I. & Krueger, P. M. Young maternal age and preterm labor. *Ann Epidemiol* **7**, 400–6 (1997).
2. da Silva, A. A. *et al.* Young maternal age and preterm birth. *Paediatr Perinat Epidemiol* **17**, 332–9 (2003).
3. You, D. *et al.* Global, regional, and national levels and trends in under-5 mortality between 1990 and 2015, with scenario-based projections to 2030: a systematic analysis by the un inter-agency group for child mortality estimation. *Lancet* **386**, 2275–86 (2015).
4. Green, N. S. *et al.* Research agenda for preterm birth: recommendations from the march of dimes. *Am J Obstet Gynecol* **193**, 626–35 (2005).
5. Morgen, C. S., Bjork, C., Andersen, P. K., Mortensen, L. H. & Nybo Andersen, A. M. Socioeconomic position and the risk of preterm birth–a study within the danish national birth cohort. *Int J Epidemiol* **37**, 1109–20 (2008).
6. Reagan, P. B. & Salsberry, P. J. Race and ethnic differences in determinants of preterm birth in the usa: broadening the social context. *Soc Sci Med* **60**, 2217–28 (2005).
7. Stieb, D. M., Chen, L., Eshoul, M. & Judek, S. Ambient air pollution, birth weight and preterm birth: a systematic review and meta-analysis. *Environ Res* **117**, 100–11 (2012).
8. Windham, G. C., Hopkins, B., Fenster, L. & Swan, S. H. Prenatal active or passive tobacco smoke exposure and the risk of preterm delivery or low birth weight. *Epidemiology* **11**, 427–33 (2000).
9. Basso, O., Olsen, J., Knudsen, L. B. & Christensen, K. Low birth weight and preterm birth after short interpregnancy intervals. *Am J Obstet Gynecol* **178**, 259–63 (1998).
10. DeFranco, E. A., Stamilio, D. M., Boslaugh, S. E., Gross, G. A. & Muglia, L. J. A short interpregnancy interval is a risk factor for preterm birth and its recurrence. *Am J Obstet Gynecol* **197**, 264e1–6 (2007).
11. Zhu, B. P. Effect of interpregnancy interval on birth outcomes: findings from three recent us studies. *Int J Gynaecol Obstet* **89** (Suppl 1), S25–33 (2005).
12. Astolfi, P. & Zonta, L. A. Risks of preterm delivery and association with maternal age, birth order, and fetal gender. *Hum Reprod* **14**, 2891–4 (1999).
13. Hsieh, T. T. *et al.* Advanced maternal age and adverse perinatal outcomes in an asian population. *Eur J Obstet Gynecol Reprod Biol* **148**, 21–6 (2010).
14. Seoud, M. A. *et al.* Impact of advanced maternal age on pregnancy outcome. *Am J Perinatol* **19**, 1–8 (2002).
15. Stewart, C. P. *et al.* Preterm delivery but not intrauterine growth retardation is associated with young maternal age among primiparae in rural nepal. *Matern Child Nutr* **3**, 174–85 (2007).
16. Hendler, I. *et al.* The preterm prediction study: association between maternal body mass index and spontaneous and indicated preterm birth. *Am J Obstet Gynecol* **192**, 882–6 (2005).
17. Hickey, C. A., Cliver, S. P., McNeal, S. F. & Goldenberg, R. L. Low pregravid body mass index as a risk factor for preterm birth: variation by ethnic group. *Obstet Gynecol* **89**, 206–12 (1997).
18. Schieve, L. A. *et al.* Prepregnancy body mass index and pregnancy weight gain: associations with preterm delivery. the nmihs collaborative study group. *Obstet Gynecol* **96**, 194–200 (2000).
19. Adams, M. M., Elam-Evans, L. D., Wilson, H. G. & Gilbertz, D. A. Rates of and factors associated with recurrence of preterm delivery. *JAMA* **283**, 1591–6 (2000).
20. Ananth, C. V., Getahun, D., Peltier, M. R., Salihu, H. M. & Vintzileos, A. M. Recurrence of spontaneous versus medically indicated preterm birth. *Am J Obstet Gynecol* **195**, 643–50 (2006).
21. Basso, O., Olsen, J. & Christensen, K. Study of environmental, social, and paternal factors in preterm delivery using sibs and half sibs. a population-based study in denmark. *J Epidemiol Community Health* **53**, 20–3 (1999).
22. Kistka, Z. A. *et al.* Racial disparity in the frequency of recurrence of preterm birth. *Am J Obstet Gynecol* **196**, 131e1–6 (2007).
23. Bhattacharya, S. *et al.* Inherited predisposition to spontaneous preterm delivery. *Obstet Gynecol* **115**, 1125–33 (2010).

24. Porter, T. F., Fraser, A. M., Hunter, C. Y., Ward, R. H. & Varner, M. W. The risk of preterm birth across generations. *Obstet Gynecol* **90**, 63–7 (1997).
25. Plunkett, J. & Muglia, L. J. Genetic contributions to preterm birth: implications from epidemiological and genetic association studies. *Ann Med* **40**, 167–95 (2008).
26. Clausson, B., Lichtenstein, P. & Cnattingius, S. Genetic influence on birthweight and gestational length determined by studies in offspring of twins. *BJOG* **107**, 375–81 (2000).
27. Treloar, S. A., Macones, G. A., Mitchell, L. E. & Martin, N. G. Genetic influences on premature parturition in an australian twin sample. *Twin Res* **3**, 80–2 (2000).
28. Lunde, A., Melve, K. K., Gjessing, H. K., Skjærven, R. & Irgens, L. M. Genetic and environmental influences on birth weight, birth length, head circumference, and gestational age by use of population-based parent-offspring data. *American journal of epidemiology* **165**, 734–741 (2007).
29. York, T. P., Strauss III, J. F., Neale, M. C. & Eaves, L. J. Racial differences in genetic and environmental risk to preterm birth. *PloS one* **5**, e12391 (2010).
30. Anum, E. A., Springel, E. H., Shriver, M. D. & Strauss, J. F. Genetic contributions to disparities in preterm birth. *Pediatr Res* **65**, 1–9 (2009).
31. Palomar, L., DeFranco, E. A., Lee, K. A., Allsworth, J. E. & Muglia, L. J. Paternal race is a risk factor for preterm birth. *Am J Obstet Gynecol* **197**, 152e1–7 (2007).
32. Goldenberg, R. L. *et al.* Medical, psychosocial, and behavioral risk factors do not explain the increased risk for low birth weight among black women. *Am J Obstet Gynecol* **175**, 1317–24 (1996).
33. Annells, M. F. *et al.* Interleukins-1, -4, -6, -10, tumor necrosis factor, transforming growth factor-beta, fas, and mannose-binding protein c gene polymorphisms in australian women: Risk of preterm birth. *Am J Obstet Gynecol* **191**, 2056–67 (2004).
34. Engel, S. A. *et al.* Risk of spontaneous preterm birth is associated with common proinflammatory cytokine polymorphisms. *Epidemiology* **16**, 469–77 (2005).
35. Kalish, R. B., Vardhana, S., Gupta, M., Perni, S. C. & Witkin, S. S. Interleukin-4 and -10 gene polymorphisms and spontaneous preterm birth in multifetal gestations. *Am J Obstet Gynecol* **190**, 702–6 (2004).
36. Macones, G. A. *et al.* A polymorphism in the promoter region of tnf and bacterial vaginosis: preliminary evidence of gene-environment interaction in the etiology of spontaneous preterm birth. *Am J Obstet Gynecol* **190**, 1504–8 (2004). discussion 3A.
37. Roberts, A. K. *et al.* Association of polymorphism within the promoter of the tumor necrosis factor alpha gene with increased risk of preterm premature rupture of the fetal membranes. *Am J Obstet Gynecol* **180**, 1297–302 (1999).
38. Monangi, N. K., Brockway, H. M., House, M., Zhang, G. &Muglia, L. J. The genetics of preterm birth: Progress and promise. *Semin Perinatol* (2015).
39. Haataja, R. *et al.* Mapping a new spontaneous preterm birth susceptibility gene, igf1r, using linkage, haplotype sharing, and association analysis. *PLoS Genet* **7**, e1001293 (2011).
40. Karjalainen, M. K. *et al.* A potential novel spontaneous preterm birth gene, ar, identified by linkage and association analysis of x chromosomal markers. *PLoS One* **7**, e51378 (2012).
41. Bream, E. N. *et al.* Candidate gene linkage approach to identify dna variants that predispose to preterm birth. *Pediatr Res* **73**, 135–41 (2013).
42. Chittoor, G. *et al.* Localization of a major susceptibility locus influencing preterm birth. *Mol Hum Reprod* **19**, 687–96 (2013).
43. McElroy, J. J. *et al.* Maternal coding variants in complement receptor 1 and spontaneous idiopathic preterm birth. *Hum Genet* **132**, 935–42 (2013).
44. Freathy, R. M. *et al.* Variants in adcy5 and near ccnl1 are associated with fetal growth and birth weight. *Nat Genet* **42**, 430–5 (2010).
45. Mook-Kanamori, D. O. *et al.* Variants near ccnl1/lekr1 and in adcy5 and fetal growth characteristics in different trimesters. *J Clin Endocrinol Metab* **96**, E810–5 (2011).
46. Urbanek, M. *et al.* The chromosome 3q25 genomic region is associated with measures of adiposity in newborns in a multi-ethnic genome-wide association study. *Hum Mol Genet* **22**, 3583–96 (2013).
47. Plunkett, J. *et al.* An evolutionary genomic approach to identify genes involved in human birth timing. *PLoS Genet* **7**, e1001365 (2011).
48. Zhang, H. *et al.* A genome-wide association study of early spontaneous preterm delivery. *Genet Epidemiol* **39**, 217–26 (2015).
49. Zhang, G. *et al.* Genetic associations with gestational duration and spontaneous preterm birth. *New England Journal of Medicine* (2017).
50. Dolan, S. M. *et al.* Synopsis of preterm birth genetic association studies: the preterm birth genetics knowledge base (ptbgene). *Public Health Genomics* **13**, 514–23 (2010).
51. Uzun, A. *et al.* dbptb: a database for preterm birth. *Database (Oxford)* **201**2, bar069 (2012).
52. Culhane, J. F. & Goldenberg, R. L. Racial disparities in preterm birth. *Semin Perinatol* **35**, 234–9 (2011).
53. Bustamante, C. D., Burchard, E. G. & De la Vega, F. M. Genomics for the world. *Nature* **475**, 163–5 (2011).
54. Menon, R., Dunlop, A. L., Kramer, M. R., Fortunato, S. J. & Hogue, C. J. An overview of racial disparities in preterm birth rates: caused by infection or inflammatory response? *Acta Obstet Gynecol Scand* **90**, 1325–31 (2011).
55. Hong, X. *et al.* Genome-wide approach identifies a novel gene-maternal pre-pregnancy bmi interaction on preterm birth. *Nature Communications* **8** (2017).
56. Wang, H. *et al.* A genome-wide association study (gwas) for bronchopulmonary dysplasia. *Pediatrics* **132**, 290–7 (2013).
57. Sonnega, A. *et al.* Cohort profile: the health and retirement study (hrs). *Int J Epidemiol* **43**, 576–85 (2014).
58. Jobe, A. H. Pulmonary surfactant therapy. *N Engl J Med* **328**, 861–8 (1993).
59. Enhorning, G. *et al.* Prevention of neonatal respiratory distress syndrome by tracheal instillation of surfactant: a randomized clinical trial. *Pediatrics* **76**, 145–53 (1985).
60. Hoekstra, R. E. *et al.* Improved neonatal survival following multiple doses of bovine surfactant in very premature neonates at risk for respiratory distress syndrome. *Pediatrics* **88**, 10–8 (1991).
61. Hoekstra, R. E., Ferrara, T. B. & Payne, N. R. Effects of surfactant therapy on outcome of extremely premature infants. *Eur J Pediatr* **153**, S12–6 (1994).
62. Horbar, J. D., Wright, E. C. & Onstad, L. Decreasing mortality associated with the introduction of surfactant therapy: an observational study of neonates weighing 601 to 1300 grams at birth. the members of the national institute of child health and human development neonatal research network. *Pediatrics* **92**, 191–6 (1993).
63. Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
64. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
65. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–91 (2016).
66. Sherry, S. T. *et al.* dbsnp: the ncbi database of genetic variation. *Nucleic Acids Res* **29**, 308–11 (2001. https://www.ncbi.nlm.nih.gov/pubmed/11125122
67. Nishimura, K. K. *et al.* Early-life air pollution and asthma risk in minority children. the gala ii and sage ii studies. *Am J Respir Crit Care Med* **188**, 309–18 (2013).
68. Consortium, G. T. The genotype-tissue expression (gtex) project. *Nat Genet* **45**, 580–5 (2013).
69. Uhlen, M. *et al.* Proteomics. tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
70. Goldenberg, R. L., Culhane, J. F., Iams, J. D. & Romero, R. Epidemiology and causes of preterm birth. *Lancet* **371**, 75–84 (2008).

71. Martius, J. A., Steck, T., Oehler, M. K. & Wulf, K. H. Risk factors associated with preterm (!37 + 0 weeks) and early preterm birth (!32 + 0 weeks): univariate and multivariate analysis of 106 345 singleton births from the 1994 statewide perinatal survey of bavaria. *Eur J Obstet Gynecol Reprod Biol* **80**, 183–9 (1998).
72. Bush, W. S. & Moore, J. H. Chapter 11: Genome-wide association studies. *PLoS Comput Biol* **8**, e1002822 (2012).
73. Itsara, A. *et al.* Population analysis of large copy number variants and hotspots of human genetic disease. *Am J Hum Genet* **84**, 148–61 (2009).
74. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–8 (2004).
75. Eichler, E. E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* **11**, 446–50 (2010).
76. Hill, W. G., Goddard, M. E. & Visscher, P. M. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet* **4**, e1000008 (2008).
77. Wright, S. Evolution in mendelian populations. *Genetics* **16**, 97–159 (1931).
78. St Julien, K. R. *et al.* High quality genome-wide genotyping from archived dried blood spots without dna amplification. *PLoS One* **8**, e64710 (2013).
79. Hadley, D. *et al.* The impact of the metabotropic glutamate receptor and other gene family interaction networks on autism. *Nat Commun* **5**, 4074 (2014).
80. Chang, C. C. *et al.* Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
81. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904–9 (2006).
82. Turner, S. D. qqman: an r package for visualizing gwas results using qq and manhattan plots. *bioRxiv* 005165 (2014).
83. Pruim, R. J. *et al.* Locuszoom: regional visualization of genome-wide association scan results. *Bioinforma.* **26**, 2336–7 (2010).
84. Loh, P. R. *et al.* Reference-based phasing using the haplotype reference consortium panel. *Nat Genet* **48**, 1443–1448 (2016).
85. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
86. White, M. J. *et al.* Novel genetic risk factors for asthma in african american children: Precision medicine and the sage ii study. *Immunogenetics* **68**, 391–400 (2016).
87. Bodian, D. L. *et al.* Germline variation in cancer-susceptibility genes in a healthy, ancestrally diverse cohort: implications for individual genome sequencing. *PLoS One* **9**, e94554 (2014).
88. Tan, A., Abecasis, G. R. & Kang, H. M. Unified representation of genetic variants. *Bioinformatics* **31**, 2202–4 (2015).
89. Wang, G. *et al.* Preterm birth and random plasma insulin levels at birth and in early childhood. *JAMA* **311**, 587–96 (2014).
90. Wang, X. *et al.* Maternal cigarette smoking, metabolic gene polymorphism, and infant birth weight. *JAMA* **287**, 195–202 (2002).


## Acknowledgements

This study was supported in part by March of Dimes Prematurity Research Center at Stanford, the Stanford Child Health Research Institute, and NIH (NHLBI) Grant 1RC2HL101748-01. Marina Sirota is supported by K01LM012381. The ITMI Cohort was supported by Inova Health System, and a generous gift of the Odeen family to ITMI. The GALA II study was supported in part by the National Heart Lung and Blood Institute R01 HL117004 and R01 HL104608, K12 HL119997, U01 OD019769, the National Institute of Health and Environmental Health Sciences R21ES24844, the Tobacco-Related Disease Research Program under Award Number 24RT-0025, the RWJF Amos Medical Faculty Development Award, the Sandler Foundation and the American Asthma Foundation to EGB, and the National Institute on Minority Health and Health Disparities under Award Number P60 MD006902, U54 MD009523 and R25 MD006832. The Boston Birth Cohort is supported in part by the March of Dimes PERI grants (20-FY02-56, #21-FY07-605, PI: Xiaobin Wang), and NIH grants (R21ES011666, R21HD066471, 2R01HD041702, PI: Xiaobin Wang). The content is solely the responsibility of the authors and does not necessarily represent the official views of the California Department of Public Health or the National Institutes of Health. Nikolaos A. Patsopoulos was support by a Career Independence Award from the National Multiple Sclerosis Society (TA 3056-A-2). We would like to thank David Hinds, Idit Kosti, Jieming Chen and Jeremy Pierce for useful discussion and insightful comments on the analysis.

## Author Contributions

N.R., J.T., G.M.S., H.M.O., D.K.S., A.J.B. and M.S. had desgined the study. R.W., K.F. H.W. and J.O. had curated the data. N.R., J.T., D.H., W.S., C.G., C.B., M.S., E.Z., N.P., G.S., H.O., D.S., A.B. and M.S. had developed the methodology. J.R., C.A., S.O., D.H., C.E., S.H., D.B., J.N., X.H., G.Z., X.W., L.M. and E.B. had provided data for validation. R.W., K.F., H.W., J.O., L.J., J.G., G.D. G.S., H.O. and D.S. had generated the data. N.R. and J.T. analysed the results. All authors contributed to writing and reviewed the manuscript.

## Additional Information

**Supplementary information** accompanies this paper at https://doi.org/10.1038/s41598-017-18246-5.

**Competing Interests:** The authors declare that they have no competing interests.

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