A Genome-Wide Association Study of spontaneous preterm birth in a European population [version 1; referees: 2 approved with reservations]

Wilfred Wu¹, Erin A S Clark², Tracy A Manuck², M Sean Esplin², Michael W Varner², Lynn B Jorde¹

¹Department of Human Genetics, University of Utah School of Medicine, Salt Lake City, UT, 84112, USA
²Department of Obstetrics and Gynecology, University of Utah School of Medicine, Salt Lake City, UT, 84112, USA

Abstract

Background: Preterm birth is defined as a birth prior to 37 completed weeks' gestation. It affects more than 10% of all births worldwide, and is the leading cause of neonatal mortality in non-anomalous newborns. Even if the preterm newborn survives, there is an increased risk of lifelong morbidity. Despite the magnitude of this public health problem, the etiology of spontaneous preterm birth is not well understood. Previous studies suggest that genetics is an important contributing factor. We therefore employed a genome-wide association approach to explore possible fetal genetic variants that may be associated with spontaneous preterm birth.

Methods: We obtained preterm birth phenotype and genotype data from the National Center for Biotechnology Information Genotypes and Phenotypes Database (study accession phs000103.v1.p1). This dataset contains participants collected by the Danish National Birth Cohort and includes 1000 preterm births and 1000 term births as controls. Whole genomes were genotyped on the Illumina Human660W-Quad_v1_A platform, which contains more than 500,000 markers. After data quality control, we performed genome-wide association studies for the 22 autosomal chromosomes.

Results: No single nucleotide polymorphism reached genome-wide significance after Bonferroni correction for multiple testing.

Conclusion: We found no evidence of genetic association with spontaneous preterm birth in this European population. Approaches that facilitate detection of both common and rare genetic variants, such as evaluation of high-risk pedigrees and genome sequencing, may be more successful in identifying genes associated with spontaneous preterm birth.
Corresponding author: Lynn B Jorde (lbj@genetics.utah.edu)


Copyright: © 2013 Wu W et al. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

Grant information: EASC is supported by National Institutes of Health, National Institute of Child Health and Human Development (K23HD061910).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: No competing interests were disclosed.

Introduction

Definition and significance of preterm birth

Preterm birth (PTB), defined as birth prior to 37 completed weeks’ gestation, is a major public health problem that affects more than 10% of all births worldwide. Globally, an estimated 15 million babies are born premature each year. Despite substantial public health efforts over the past several decades, the U.S. PTB rate remained at 11.72% in 2011.

PTB is the leading cause of neonatal mortality in non-anomalous newborns. It is also associated with a broad spectrum of lifelong morbidity in surviving preterm infants, including neuro-developmental delay, cerebral palsy, blindness, deafness, and chronic lung disease.

Definition of spontaneous preterm birth

Some PTBs are iatrogenic and can be attributed to obstetric intervention aimed at reducing maternal and/or fetal risk. The remaining PTBs are known as spontaneous PTB (SPTB) and are the focus of research efforts to identify genetic and environmental risk factors. Although SPTB is a pressing health issue, the incomplete understanding of its biology has inhibited development of effective prevention and treatment strategies.

Genetics of spontaneous preterm birth

The etiology of SPTB is complex and multifactorial, although genetic factors are important contributors. In addition, PTB prevalence varies among different population groups. African-American ancestry is consistently associated with an increased PTB risk, even after adjusting for epidemiologic risk factors, such as income, education, lack of prenatal care, and other socioeconomic factors.

A candidate gene approach has identified polymorphisms in genes that encode the progesterone receptor, tumor necrosis factor alpha, interleukins 4, 6, and 10, and mannose-binding lectin, that are mildly to modestly associated with SPTB. However, results have been inconsistent.

Our genome-wide approach for spontaneous preterm birth

Here, we employed an unbiased, genome-wide approach to search for possible candidate genes associated with SPTB. Genome-wide association studies (GWAS), or whole genome association studies, are a commonly used genetic approach to study a disease or a trait. GWAS compares thousands or even millions of common genetic markers, mainly single nucleotide polymorphisms (SNPs), across individuals with a disease or trait status. There are 1,688 publications identifying 11,299 SNPs that are significant for diseases or traits in 17 different categories. We obtained the SPTB phenotype and genotype data deposited in the National Center for Biotechnology Information (NCBI) Genotypes and Phenotypes Database (dbGaP), study accession phs000103.v1.p1, to perform a GWAS to explore genetic variants associated with SPTB.

Materials and methods

Data application and approval

We applied to and received approval from dbGaP for access to a dataset for SPTB phenotype and genotype (study accession phs000103.v1.p1). We followed the Data Use Certification Agreement we signed during the application.

To access the data, one must apply and agree to the dbGAP terms of usage. A detailed instructions and procedures for application can be obtained at https://dbgap.ncbi.nlm.nih.gov/.

Data content

This dataset contains participants collected by the Danish National Birth Cohort (DNBC). DNBC is a prospective cohort that enrolled more than 100,000 pregnant women in the first trimester, before most adverse outcomes occurred, and therefore is free from sampling or collection bias. In this dataset, there are approximately 1000 preterm births and 1000 term births as controls. These study subjects were collected from 1997 to 2003. All are singleton gestations. Each birth has records of mother-child pairs. With the exception of 24 children with one or two grandparents from other Nordic countries, all other children in the dataset had parents and all four grandparents born in Denmark.

The case (preterm) group contains births delivered before 37 gestational weeks. The control (term) group contains births delivered at approximately 40 weeks’ gestation. In both preterm and term groups, children born with any recognized congenital or genetic abnormality were excluded. Pregnancies with maternal conditions known to be associated with PTB, iatrogenic or spontaneous (placenta previa, placental abruption, hydramnios, isoimmunization, placental insufficiency, pre-eclampsia/eclampsia), were also excluded by DNBC.

The blood sample (buffy coat) was collected for each mother-child pair. Their whole genomes were genotyped on the Illumina Human660W-Quad_v1_A platform (Illumina, Inc., San Diego, California, USA), which contains more than 500,000 markers. Genotyping was performed by the Johns Hopkins University Center for Inherited Disease Research (Baltimore, Maryland, USA). Further data cleaning and harmonization were done at the GENEVA Coordinating Center at the University of Washington (Seattle, Washington, USA).

Quality control

In this study, we focused only on fetal genomes for further analysis. Individuals with missing genotypes greater than 3% were filtered out. In addition, individuals with a heterozygosity rate deviating more than 3 standard deviations were also excluded from further analysis. After per-individual quality control, we also performed per-SNP filtering. The SNPs that had missing genotypes greater than 3% were excluded. Those having significantly different genotype missing rate between the case and control groups were also eliminated. A conservative cut-off with p < 1x10^-4 was applied. We also excluded SNPs that significantly deviated from Hardy-Weinberg equilibrium in the control group – those with p < 1x10^-4 were filtered out.

Data analysis

After data quality control, we performed GWAS for the fetal genomes on 22 autosomal chromosomes using the PLINK software package v1.07 (http://pngu.mgh.harvard.edu/purcell/plink/). The
level of genome-wide significance was set at $9.18 \times 10^{-8}$, corresponding to Bonferroni correction for 544,675 multiple independent tests.

Quantile-quantile (QQ) plot was made using STATA Statistical Software, Release 12. Manhattan plots were generated using Haplovie™.

**Results**

**Data overview**

We started with a dataset comprised of 1,900 children. There were 31 children having a missing genotype rate greater than 3%. The average heterozygosity was 0.3238, with standard deviation of 0.0059. There were also 31 children having heterozygosity that deviated more than 3 standard deviations. In fact, there was considerable overlap when applying these two filtering criteria (Figure 1) - 26 individuals were identified by both exclusion criteria. During per-individual quality control, a total of 36 individuals were excluded, resulting in 1864 individuals. However, 66 of them had a missing phenotype, yielding a final total of 849 cases and 949 controls (Table 1).

Among the 560,768 markers, there were 1,933 SNPs that exceeded the missing rate threshold of 3%. There were 367 SNPs that had a significantly different missing rate between the case group and control group (P value $< 1 \times 10^{-5}$). Further, 885 SNPs significantly deviated from Hardy-Weinberg equilibrium in the control group (P value $< 1 \times 10^{-5}$). These three criteria also identified some overlapping SNPs; a total of 2,670 SNPs were excluded using all criteria. These quality control steps left 558,098 SNPs remaining in the dataset. Of these SNPs, 544,675 of them are located on 22 autosomes, and were included in the analysis (Table 1).

**Allelic test**

We carried out GWAS for these 1,798 individuals, of which 849 are SPTB cases and 949 are term controls, over 544,675 SNPs on 22 autosomal chromosomes. An allelic test was first carried out, and no SNPs reached genome-wide significance after Bonferroni correction for multiple testing (Manhattan plot (Figure 2) and QQ plot (Figure 3)).

**Other genetic models**

We then performed GWAS with different genetic models. We tested three classical Mendelian inheritance models here. The recessive model assumes that carrying two variant alleles is required to present a different phenotype; while in the dominant model, one variant allele is sufficient to present a different phenotype as carrying two variant alleles. The additive model assumes the heterozygotes present an intermediate phenotype between the two homozygotes and thus consider the three genotypes separately. The Manhattan plot for the additive model (Figure 4), dominant model (Figure 5), and recessive model (Figure 6) are shown. The dominant or recessive models refer to the action of the minor allele. Within these

---

**Figure 1. Per-individual quality control.** The X axis is the missing genotype rate for each individual. The Y axis is the heterozygosity rate for each individual. Each dot represents a person. The vertical dash line is the cut-off for per-individual missing rate: 3%. Individuals with missing rate greater than 3% are excluded. The two horizontal dot lines represent the mean heterozygosity ± 3 standard deviations. People with heterozygosity deviate above 3 standard deviations are also excluded. The two criteria overlap largely at the right lower part of the graph.
Figure 2. Manhattan plot for allelic test. The X axis is the position of each SNP grouped by different chromosomes, and presented with different colors. The Y axis is the P value for each test for each SNP, in $-\log_{10}$ scale. The horizontal blue line indicated the threshold of genome-wide significance after Bonferroni correction. The threshold is at 7.04, which corresponds to $-\log_{10}(0.05/544,675)$ for 544,675 independent tests. No SNP reached genome-wide significance.

Figure 3. QQ plot for allelic test. The X axis indicated the expected P value, in $-\log_{10}$ scale. The Y axis indicated the observed P value from allelic test, also in $-\log_{10}$ scale. The red diagonal line is the line of Y=X, where observed equals expected. The genome-wide significance threshold for $-\log_{10}(P)$ scale is at 7.04, which corresponds to $-\log_{10}(0.05/544,675)$ for 544,675 independent tests. No SNP reached genome-wide significance.
Table 1. Number of participants and markers in the dataset.

<table>
<thead>
<tr>
<th>I. Participants (Individuals)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Original dataset</td>
<td>1900</td>
</tr>
<tr>
<td>Per-individual quality control exclusion</td>
<td>36</td>
</tr>
<tr>
<td>Missing genotype rate &gt; 3%</td>
<td>31</td>
</tr>
<tr>
<td>Heterozygosity &gt; 3 sd</td>
<td>31</td>
</tr>
<tr>
<td>Missing phenotype</td>
<td>66</td>
</tr>
<tr>
<td>Remained for analysis</td>
<td>1,798</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>II. Markers (SNPs)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Original dataset</td>
<td>560,768</td>
</tr>
<tr>
<td>Per-marker quality control exclusion</td>
<td>2,670</td>
</tr>
<tr>
<td>Missing genotype rate &gt; 3%</td>
<td>1,933</td>
</tr>
<tr>
<td>Significantly deviate from Hardy-Weinberg equilibrium in control group</td>
<td>367</td>
</tr>
<tr>
<td>Not on autosomes</td>
<td>13,423</td>
</tr>
<tr>
<td>Remained for analysis</td>
<td>544,675</td>
</tr>
</tbody>
</table>

sd: standard deviations.

Discussion

Here we describe a negative result for associations between SPTB and genetic polymorphisms of 22 autosomal chromosomes in a homogeneous European population. Myking et al. reported a GWAS focusing on the X chromosome and they incorporated Danish cases and controls from DNBC, in addition to participants enrolled from the Norwegian Mother and Child Cohort Study (MoBa). Nevertheless, with a larger sample size (DNBC + MoBa), and fewer independent tests limited to the markers on X chromosomes, no SNP reached genome-wide significance after Bonferroni correction.

One way to decrease the probability of a negative result is to increase the sample size, either by recruiting more cases and controls directly, or by combining different studies and conducting a meta-analysis. Instead of sequencing thousands or millions of sporadic cases plus controls, another approach is to study SPTB using a family-based design – i.e., identify high-risk pedigrees in which a genetic mutation is more likely to be present in multiple individuals. Pedigree studies have the additional advantage of reduced phenotypic heterogeneity. Several loci associated with SPTB have been identified by using family-based linkage studies.

Figure 4. Manhattan plot for additive model. The X axis is the position of each SNP grouped by different chromosomes, and presented with different colors. The Y axis is the P value for each test for each SNP, in –log10 scale. The horizontal blue line indicated the threshold of genome-wide significance after Bonferroni correction. The threshold is at 7.04, which corresponds to –log10 (0.05/544,675) for 544,675 independent tests. No SNP reached genome-wide significance.
**Figure 5. Manhattan plot for dominant model.** This is the GWAS result assuming the dominant action of minor alleles. The Y axis is the P value for each test for each SNP, in –log10 scale. The horizontal blue line indicated the threshold of genome-wide significance after Bonferroni correction. The threshold is at 7.04, which corresponds to –log10 (0.05/544,675) for 544,675 independent tests. No SNP reached genome-wide significance.

**Figure 6. Manhattan plot for recessive model.** This is the GWAS result assuming the recessive action of minor alleles. The Y axis is the P value for each test for each SNP, in –log10 scale. The horizontal blue line indicated the threshold of genome-wide significance after Bonferroni correction. The threshold is at 7.04, which corresponds to –log10 (0.05/544,675) for 544,675 independent tests. No SNP reached genome-wide significance.
Another approach is to employ whole genome or whole exome sequencing. This will help to identify rare genetic variants with potentially larger effect sizes.

Another approach that may increase statistical power is to analyze the SPTB phenotype as a quantitative trait instead of a dichotomous one. The distribution of gestational age in the population is approximately normal. Therefore, to analyze SPTB as a quantitative trait (i.e., gestational age), samples should be drawn randomly from the population of newborns.

Conclusion

We found no evidence of genetic association with SPTB in Danish population using an unbiased genome-wide approach. A family-based design in a high risk pedigree, and whole genome or exome sequencing, may yield higher detection rates of both common and rare variants associated with SPTB.

Author contributions

WW conceived of and designed the study. EASC, TAM, MSE, MV, and LBJ provided input to the study design. WW and LBJ applied for data access and conducted the analysis. WW drafted the manuscript. EASC, TAM, MSE, MV, and LBJ provided critical review. All authors read and approved the final manuscript.

Competing interests

No competing interests were disclosed.

Grant information

EASC is supported by National Institutes of Health, National Institute of Child Health and Human Development (K23HD061910). We would like to acknowledge DNBC, dbGAP, and NCBI for depositing and hosting the phenotype and genotype data.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

We would like to acknowledge DNBC, dbGAP, and NCBI for depositing and hosting the phenotype and genotype data. We would like to thank Robert Weiss from Department of Human Genetics, University of Utah, for his detailed consultations for data analysis. We would also like to thank Isabella Johnsen and June Blackburn from University of Utah for their assistance during our application for dbGAP data access.

References


Page 8 of 12


45.  Stata Corp. LP: *STATA Statistical Software: Release 12*. College Station, TX, 2011.


Open Peer Review

Current Referee Status: ? ?

Version 1

Referee Report 24 November 2014

doi:10.5256/f1000research.2631.r6795

? David Olson1, Scott Williams2
1 Department of Obstetrics and Gynecology, University of Alberta, Edmonton, AB, Canada
2 Department of Genetics, Institute of Quantitative Biomedical Sciences, Geisel School of Medicine,
Dartmouth College, Hanover, NH, USA

From Scott Williams:

I have now read the paper and as it stands it is ok. I would however recommend several things that may
improve it:

1. State that the data generally suggests that most effects are maternal so that this may be a hard
task.

2. Note whether any of the candidate genes that have even nominal data fall in regions of statistical
significance. Although not proof positive such findings would support an association – albeit not
novel.

3. I always like to ask if the cases are in HWE as well as the controls as there are several papers I
could cite – including some of our own that indicate this is a good way to detect effects.

4. Lastly, I am concerned that their bonferonni correction is too conservative as the ~550,000 are not
all independent. We have paper in press on this point.

That said nothing the authors have done is incorrect – only what I would consider to be a too limited
exploration of the data. And since this type of study is by nature exploratory I would focus on lowering
type 2 errors as opposed to type 1 (e.g., bonferonni corrections).

From David Olson:

Preterm birth is a very complex and heterogeneous problem; there are many causes, and these can vary
according to gestational age. I would have expected the outcomes observed (no significant SNPs)
because the authors did not stratify their preterm birth subjects into homogeneous groups. They should
have separated single fetus pregnancies from multiple-fetus pregnancies, excluded those with other
pre-disposing problems such as pre-eclampsia, preterm premature rupture of membranes, incompetent
cervix, etc., and they should have stratified according to age: <28 weeks gestational age, 28-30, 30-34
and 34-37. I would suggest they re-analyze their data according to such stratification in an attempt to
create more homogeneous groups.
We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

**Competing Interests:** No competing interests were disclosed.

Mohamad Saad  
Division of Medical Genetics, University of Washington, Seattle, WA, USA

In their paper, “A Genome-Wide Association Study of spontaneous preterm birth in a European population”, the authors perform a GWAS on spontaneous preterm birth in a European population. The GWAS is based on 849 cases (preterm births) and 949 controls (term births), and use a dense genotyping panel that contains over 500,000 SNPs. The paper is nicely written and the analyses are well done. However, I have several issues that need to be taken care of. In its actual form, this paper lacks additional analyses. Below are my non-exhaustive comments:

1. In this study, no SNP reaches the genome-wide significance threshold. However, as we can see in the Manhattan plot (Figure 2), there are several suggestive signals, especially on chromosome 12, 15, and 3. The authors should investigate these signals and try to find the nearby genes.

2. They perform four models and show the corresponding Manhattan plots. First, they do not explain the reason why they perform these four models. Second, the allelic test should give similar results to the additive model test. The authors do not compare between these models, at least for the most significant signals. Otherwise, these analyses seem to be redundant.

3. Providing the Manhattan plots is far from enough. We do not have the ability to see the most significant SNP names. In addition, they should mention if there was a performed GWAS for preterm birth, and if so, they should compare their results with what can be found in the literature. Moreover, they mention that there is an evidence of a candidate gene. They should say how this gene was discovered, and show their SNP results in this gene or its region. In addition, they might consider performing the haplotypic test in this gene or in the region of the most significant SNPs. They should present these results in informative tables.

4. In the introduction, the authors say that there are genetic factors that contribute to SPTB. It is important to mention the heritability estimates or the familial recurrence risk estimate (if there is any) for SPTB.

5. The authors talk about negative results in their GWAS. They should discuss why they did not discover any genome-wide significant SNP (lack of power?). They also need to discuss the weaknesses of this study, and what could be done in the future.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.
**Competing Interests:** No competing interests were disclosed.