Meta-analysis of gestational duration and spontaneous preterm birth identifies new maternal risk loci


1Research Unit of Clinical Medicine, Medical Research Center Oulu, University of Oulu, and Department of Children and Adolescents, Oulu University Hospital, Oulu, Finland
2Research Unit of Population Health, Faculty of Medicine, University of Oulu, Oulu, Finland
3Contributors of FinnGen are listed in supplemental material
4Division of Human Genetics, Cincinnati Children’s Hospital Medical Center, Center for Prevention of Preterm Birth, Perinatal Institute and March of Dimes Prematurity Research Center Ohio Collaborative, Cincinnati Children’s Hospital Medical Center, Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, Ohio, United States of America
5Department of Epidemiology Research, Statens Serum Institut, Copenhagen, Denmark
6Department of Obstetrics and Gynaecology, Sahlgrenska Academy, Institute of Clinical Science, University of Gothenburg, Gothenburg, Sweden
7Department of Genetics and Bioinformatics, Health Data and Digitalization, Norwegian Institute of Public Health, Oslo, Norway
8Institute for Molecular Medicine Finland (FIMM), Helsinki Institute of Life Science, University of Helsinki, Helsinki, Finland
9Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA.
10Psychiatric & Neurodevelopmental Genetics Unit, Department of Psychiatry, Analytic and Translational Genetics Unit, Department of Medicine, and the Department of Neurology, Massachusetts General Hospital, Boston, MA, USA.
11Center for Child, Adolescent, and Maternal Health Research, Faculty of Medicine and Health Technology, University of Tampere, Tampere, Finland
12Department of Obstetrics and Gynecology, Tampere University Hospital, Tampere, Finland
13Burroughs Wellcome Fund, Research Triangle Park, Durham, NC, USA.
14Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland
*Corresponding author
**equal contribution

Abstract

Background Preterm birth (<37 weeks of gestation) is a major cause of neonatal death and morbidity. Up to 40% of the variation in timing of birth results from genetic factors, mostly due to the maternal genome.

Methods We conducted a genome-wide meta-analysis of gestational duration and spontaneous preterm birth in 68,732 and 98,371 European mothers, respectively.

Results We detected 19 associated loci of which seven were novel. The loci mapped to several biologically plausible genes, including HAND2 whose expression was previously shown to decrease during gestation, associated with gestational duration, and GC encoding Vitamin D-binding protein, associated with preterm birth. Downstream in silico-analysis suggested regulatory roles as underlying mechanisms for the associated loci. LD score regression found birth weight measures as the most strongly correlated traits, highlighting the unique nature of spontaneous preterm birth phenotype. Tissue expression and colocalization analysis revealed reproductive tissues and immune cell types as the most relevant sites of action.

Conclusion We report novel genetic risk loci that associate with preterm birth or gestational duration, and reproduce findings from previous genome-wide association studies. Altogether, our findings provide new insight into the genetic background of preterm birth. Better characterization of the causal genetic mechanisms will be important to public health as it could suggest new strategies to treat and prevent preterm birth.
Funding:
The FinnGen project is funded by two grants from Business Finland (HUS 4685/31/2016 and UH 4386/31/2016) and the following industry partners: AbbVie Inc., AstraZeneca UK Ltd, Biogen MA Inc., Bristol Myers Squibb (and Celgene Corporation & Celgene International II Sàrl), Genentech Inc., Merck Sharp & Dohme Corp, Pfizer Inc., GlaxoSmithKline Intellectual Property Development Ltd., Sanofi US Services Inc., Maze Therapeutics Inc., Janssen Biotech Inc, Novartis AG, and Boehringer Ingelheim. Research regarding spontaneous preterm birth performed at the University of Oulu was financed by the Jane and Aatos Erkko Foundation (MH, MR), Competitive State Research Financing of the Expert Responsibility Area of Oulu University Hospital (MR), Sigrid Jusélius Foundation (MH), Foundation for Pediatric Research (MR), Emil Aaltonen Foundation (AP), and Alma and K.A. Snellman Foundation (AP). BF received support from an Oak Foundation fellowship and a Novo Nordisk Foundation grant (12955). GZ is supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number R01HD101669, the Burroughs Wellcome Fund (Grant 10172896), the March of Dimes Prematurity Research Center Ohio Collaborative and the Bill & Melinda Gates Foundation. The Norwegian Mother and Child Cohort Study was supported by the Norwegian Ministry of Health and the Ministry of Education and Research, by the National Institute of Environmental Health Sciences (contract no. N01-ES-75558), the National Institute of Neurological Disorders and Stroke (U01 NS 047537-01 and U01 NS 047537-06A1), the Norwegian Research Council/FUGE (151918/S10, 183220/S10 and FRI-MEDBIO 249779). BJ was funded by the Swedish Research Council (2015-02559) and by from the Norwegian Research Council, a grant from the Jane and Dan Olsson Foundations, a grant (ALFGBG-426411) from the Swedish government to researchers in the public health service. The Danish National Birth Cohort was established with a significant grant from the Danish National Research Foundation. Additional support was obtained from the Danish Regional Committees, the Pharmacy Foundation, the Egmont Foundation, the March of Dimes Birth Defects Foundation, the Health Foundation and other minor grants. The DNBC Biobank has been supported by the Novo Nordisk Foundation and the Lundbeck Foundation.

Acknowledgements:

We want to acknowledge the participants and investigators of the FinnGen study. The following biobanks are acknowledged for delivering biobank samples to FinnGen: Auria Biobank (www.auria.fi/biopankki), THL Biobank (www.thl.fi/biobank), Helsinki Biobank
(www.helsinginbiopankki.fi), Biobank Borealis of Northern Finland (https://www.ppshp.fi/Tutkimus-ja-opetus/Biopankki/Pages/Biobank-Borealis-briefly-in-English.aspx), Finnish Clinical Biobank Tampere (www.tays.fi/en-US/Research_and_development/Finnish_Clinical_Biobank_Tampere), Biobank of Eastern Finland (www.ita-suomenbiopankki.fi/en), Central Finland Biobank (www.ksshp.fi/fi-FI/Potilaalle/Biopankki), Finnish Red Cross Blood Service Biobank (www.veripalvelu.fi/verenluovutus/biopankkitoiminta), and Terveystalo Biobank (www.terveystalo.com/fi/Yritystietoa/Terveystalo-Biopankki/Biopankki/). All Finnish Biobanks are members of BBMRI.fi infrastructure (www.bbmri.fi). Finnish Biobank Cooperative–FINBB (https://finbb.fi/) is the coordinator of BBMRI-ERIC operations in Finland. Finnish biobank data can be accessed through Fingenious® services (https://site.fingenious.fi/en/), managed by FINBB. We would like to thank the research participants and employees of 23andMe, Inc, for making this work possible. CSC–IT Center for Science, Finland, is acknowledged for computational resources. This study includes data from the Norwegian Mother, Father and Child Cohort Study (MoBa) conducted by the Norwegian Institute of Public Health and from the Danish National Birth Cohort (DNBC), and we would like to thank the research participants of the Norwegian MoBa study and the DNBC. We thank Maarit Haarala (University of Oulu, Oulu, Finland) and Riitta Vikeväinen (Oulu University Hospital, Oulu, Finland) for technical assistance.

Data availability statement

Meta-analysis summary statistics of the top 10,000 SNPs will be deposited in an appropriate data repository. Summary statistics of the Northern/Central Finnish cohort are available upon reasonable request. The full GWAS summary statistics for the 23andMe discovery data set will be made available through 23andMe to qualified researchers under an agreement with 23andMe that protects the privacy of the 23andMe participants. Please visit https://research.23andme.com/collaborate/#dataset-access/ for more information and to apply to access the data. For information regarding Finngen data access, please visit https://www.finngen.fi.
Proper timing of birth is crucial for the survival and long-term health of newborn infants. Preterm birth, defined as birth that occurs prior to 37 completed weeks of gestation, is the most common cause of neonatal death and a prevalent cause of death among children under 5 years. Moreover, preterm birth is the underlying cause of several long-term morbidities including neurodevelopmental problems, cerebral palsy, learning difficulties, and sensory loss. Globally, preterm birth affects approximately 11% of births, equal to 15 million pregnancies, each year. In Scandinavian countries and Finland, the annual incidence of preterm birth is approximately 5–6%.

While intrauterine growth restriction and preeclampsia are major causes of medically indicated preterm birth, approximately 70% preterm births occur after spontaneous onset of labor. There are just a few ways to predict the risk or to prevent the occurrence of spontaneous preterm birth (SPTB). Genetic variants in maternal and fetal genomes have been recognized as factors that contribute to the risk of SPTB and to variation in gestational duration. Family studies suggest that approximately 30%–40% of the variation in timing of birth is explained by genetic factors, with contributions from the maternal genome most important. Recent genome-wide association studies (GWAS) have identified some robust associations. Variants in genes including WNT4, EBF1, AGTR2, and KCNAB1 were associated with timing of birth in mothers, and a fetal GWAS meta-analysis discovered a locus near genes that encode pro-inflammatory cytokines associated with gestational duration.

In the present study, our aim was to strengthen knowledge of the genetic background of SPTB by identifying and replicating associations of genetic loci in relation to timing of spontaneous singleton birth. To that end, we conducted a case-control meta-analysis of SPTB and a quantitative meta-analysis of gestational duration in 98,371 and 68,732 European mothers, respectively.
Results

Overview of genome-wide meta-analysis

The genome-wide meta-analysis of SPTB (n=98,371) and gestational duration (n=68,732) with maternal data from the FinnGen study, 23andMe, Inc., and the cohort from Northern and Central Finland detected 19 independent loci with at least one variant associated at $p<5\times10^{-8}$ (Figure 1a, Tables 1-2). Variants associated with gestational duration were enriched in several categories, whereas variants associated with SPTB were mostly annotated as intronic or intergenic (Figure 1b). We considered an associated locus to be novel if there were no genome-wide significant associations with gestational duration or SPTB for any of the variants within a ±1 Mb range around the meta-analysis lead variant in the GWAS Catalog$^{11}$ (Table S2) or in a recent maternal meta-analysis of the timing of parturition$^{10}$. We detected five novel loci associated with gestational duration, and two novel risk loci for SPTB. The results of the meta-analysis with a strict SPTB definition in the FinnGen GWAS are shown in Figure S1. The effect estimates of the associated loci were similar with the main analysis (Figure S2).

LDSC-based genomic inflation factor$^{12}$ indicated minimal confounding effects in the meta-analysis of gestational duration ($\lambda_{GC}=1.077$, intercept=1.025) or SPTB ($\lambda_{GC}=1.038$, intercept=1.007) (Figure S3). The meta-analysis test statistics were homogenous among populations, and the effect estimates of the risk loci for SPTB were similar across individual cohorts (Tables 1-2, Figure S4). According to LDSC heritability estimates, the current results explain approximately 17.5% of the variation in gestational duration and 6% in SPTB on a liability scale (Table S3). We further used LDSC to evaluate shared genetic architecture between the meta-analysis outcomes and 773 other complex traits (Figure 2, Table S4). The analysis detected significant correlations between birth weight–related measures and both gestational duration and SPTB. As expected, longer duration of gestation was associated with higher birth weight, whereas preterm birth was linked to lower birth-weight measures. In addition, specific measures of physical fitness, alertness, and lack of depression correlated with a longer duration of pregnancy or term birth (Figure 2, Table S4).

MAGMA gene set enrichment analysis based on the full distribution $p$-values indicated involvement of gonad development and steroid hormone biosynthesis in gestational duration, whereas kinetochore-microtubule and neuron differentiation were the top pathways in SPTB.
MAGMA tissue expression analysis across GTEx v8 did not yield significant results but ranked several reproductive tissues, including uterus and ovary, among the most relevant tissue types for both gestational duration- and SPTB-associated genes (Figure S5). When visualized in a gene-expression heatmap across the GTEx v8 tissues, some of the genes, including *EBF1*, *GNAQ*, *HAND2*, *ZBTB38*, and *COL27A1*, clustered in a profile of higher expression in tissues including blood vessel, uterus, and fallopian tube (Figure S5). Gene set analysis based on a list of genes corresponding to all significant loci in the current study identified enrichment of multiple pathways, with GO terms referring to regulation of morphogenesis and development of various organs and tissues among multiple top pathways (Table S6).

**Replication and joint analysis** We used data from the Nordic data sets to test for replication of the associated loci and to perform joint analysis (Table S7). Loci near *WNT4*, *EEFSEC*, *EBF1*, and *AGTR2* were not included, since the same replication data was used in the study that discovered these associations. While all effects among the genome-wide significant meta-analysis loci and the replication population were in the same direction, the strongest associations in the replication population were detected for *ZBTB38*, *HAND2*, *TET3*, and *KCNAB1*. Joint analysis of the replication data and the meta-analysis variants with suggestive significance ($p<1e^{-6}$ to $5e^{-8}$) detected *DNAH2* and *RAP2C* as additional loci associated with gestational duration. In addition, the current meta-analysis detected many of the associations in the preprint reporting maternal meta-analysis of the timing of parturition, including nine loci for gestational duration and two for SPTB.

**Characterization of association signals**
To gain insight into the nature of the associated loci, we explored previous associations with other complex traits in the literature and with data from FinnGenR7 and the IEU openGWAS project, and performed colocalization analysis with expression quantitative trait loci (eQTLs) to evaluate if the associated variants affect their target genes by regulating gene expression. In the FinnGen data, we screened the meta-analysis lead variants for associations with all $>3000$ phenotypes in freeze 7 (Figure S8), whereas data from the IEU openGWAS project was queried in a PheWAS for all associated variants within the associated meta-analysis loci (Table S8).
The lead variant rs1991431 in ZBTB38 with a replicable association was associated with hyperplasia of prostate (BHP) in the FinnGen (Figure S8), and other associated variants in the locus were linked to various complex traits including cell counts of lymphocytes and monocytes, and ZBTB38 mRNA expression on the IEU openGWA S data. The same alleles of several meta-analysis variants (e.g., T allele of variant rs9846396 associated with longer gestational duration; Z-score = 6.36, \( p = 2.04\times10^{-10} \)) were also associated with taller height, higher body mass measures, and increased risk of prostate cancer (Table S8). We detected colocalization for variants associated with gestational duration and ZBTB38 expression in tissues including monocytes, T cells, and B cells, and alleles associated with longer gestational duration were linked to higher ZBTB38 expression (Figure 3, Table S9).

Like variants in ZBTB38, the polymorphisms in the WNT3A locus were associated with height and body mass indices, and the encoded protein was implicated in cell fate and patterning during embryogenesis\(^\text{14,15}\). Concordantly, variants associated with gestational duration colocalized with WNT3A expression in the placenta (Figure 3, Table S9).

We detected a replicable association of variants near HAND2. HAND2 plays a role in heart development with previous associations with traits including atrial fibrillation and platelet count\(^\text{11}\). HAND2 is expressed in the human uterine tissue, where it is upregulated by the progesterone receptor, and involved in immune tolerance of the decidua by regulating a distinct set of genes including interleukin \(\text{15}^{\text{16,17}}\).

\(TET3\) and \(KCNAB1\) showed strong replication (Table S7). Variants in both genes were associated with birth weight of offspring (Table S8), and alleles associated with longer gestational duration in the current meta-analysis were linked to higher birth-weight measures. We observed a similar positive correlation between gestational duration and birth weight at the genome-wide level in the LDSC analysis (Figure 2). Variants associated with gestational duration colocalized with \(KCNA1B\) expression two reproductive tissues and blood vessel (Table S9). Alleles linked to longer gestational duration were associated with increased \(KCNAB1\) expression in all tissues.

The case-control meta-analysis of SPTB detected two novel associated loci: \(GC\) and \(LINC02824\), which encode vitamin D-binding protein and a long noncoding RNA, respectively. \(GC\) is involved in vitamin D transport and storage, and circulating vitamin D
levels have been linked to preterm birth and other pregnancy- and reproductive health–related outcomes in observational studies\textsuperscript{18,19}. 
Discussion

The current genome-wide meta-analysis of SPTB (n=98,371) and gestational duration (n=68,732) identified several associated loci. We detected loci that had no previous associations with gestational duration or SPTB, and our findings further reinforce the associations of genes from previous GWASs of mothers who gave birth preterm. The associated loci with strong replication in the current analysis, including ZBTB38, HAND2, TET3, and KCNAB1, also showed association in the recent meta-analysis of the timing of parturition[10], establishing these genes as strong candidates for further molecular biological studies of SPTB. The inferred functions of the assigned candidate genes were consistent with a role in the timing of birth.

Association of ZBTB38, zinc finger and BTB domain containing 38, with benign hyperplasia of prostate (BHP) is a compelling finding given that both gestational duration and BHP are affected by changes in estrogen and androgen levels[20,21]. ZBTB38 was further associated with cell counts of various immune cells, and our results suggest that increased ZBTB38 expression in these cell types may play a role in regulating length of pregnancy. Alleles associated with longer gestational duration showed association with increased height, body mass, and risk of prostate cancer. It remains to be determined whether ZBTB38 confers its effect on birth timing through pregnancy-specific mechanisms or by contributing to more general immune pathways that influence gestation. Our findings for ZBTB38 associations are in keeping with reported correlations among maternal height, gestational duration, and fetal growth, and further comply with detected associations between maternal birth weight–elevating alleles and longer gestational duration and between maternal gestation-prolonging alleles and the risk of prostate carcinoma[8,22,23].

An association near HAND2 showed strong replication. HAND2 encodes heart and neural crest derivatives expressed 2, a transcription factor best known for its roles in cardiac morphogenesis and limb development. Decreasing expression of HAND2 in the decidua during pregnancy may contribute to regulation of gestational duration[24]. The expression of HAND2 in the human uterine tissue, and its gradually decreasing expression in the decidua during pregnancy[24], makes it an interesting candidate gene and a potential biomarker for SPTB.
Variants in Tet methylcytosine dioxygenase 3 (TET3) and potassium voltage-gated channel subfamily A regulatory beta subunit 1 (KCNAB1) were previously associated with birth weight of offspring, and alleles associated with longer gestational duration correlated with birth-weight measures, complying with known correlations of the length of gestation and fetal growth. It is possible that the associations between these loci and gestational duration explain the effect of the mentioned loci on birth weight. Our results further suggest that KCNAB1 expression contributes to the timing of birth. Of note, TET3 was suggested to play a role in embryo implantation. Both TET3 and KCNAB1 represent interesting targets for further study to determine their specific roles related to the regulation of gestational duration.

Interestingly, COL27A1 was associated with phenotypes including embryonic growth retardation, abnormal placenta morphology, and abnormal placenta vasculature in data in the knock-out mice as investigated via IMPC (https://www.mousephenotype.org/). COL27A1 is most abundantly expressed in the endometrium, and the gene encodes collagen type XXVII alpha 1 chain, which is a fibrillar, developmentally regulated protein. Further, the meta-analysis lead variant of the COL27A1 locus is near miR-455, which has roles in cartilage development, adipogenesis, and preeclampsia, and may protect endometrial cells against oxidative stress.

The novel meta-analysis loci associated with gestational duration, such as variants in GNAQ and KCNN2, comprised further intriguing candidates. GNAQ (protein subunit alpha q) plays a role in survival of immune cells and was identified as part of a transcriptomic signature related to human labor in the choriodedecidua. KCNN2 (potassium calcium-activated channel subfamily N member 2) has no obvious connection to pregnancy-related regulation, but KCNN3, another molecule in the KCNN family of potassium channel genes, plays a role in uterine function. The role of these risk loci and corresponding causal genes remains to be verified.

The case-control meta-analysis of SPTB detected an association in GC, encoding GC vitamin D binding protein, as of special interest because of its known involvement in vitamin D transport and storage. Protein encoded by GC is the primary carrier of vitamin D that binds to the vitamin and its plasma metabolites and transports them to their target tissues. Previous studies have suggested links among plasma vitamin D levels and preterm birth or other pregnancy-related outcomes including pre-eclampsia, polycystic ovary syndrome, and endometriosis. Vitamin D deficiency was found associated with many adverse outcomes.
including those related to pregnancy, whereas increased levels of the protein product of \textit{GC} showed association with a reduced risk of certain immune-mediated diseases\textsuperscript{19,31,32}. The precise role of \textit{GC} in the context of human pregnancy and SPTB remains to be determined.

Altogether, our results highlight the unique nature of SPTB. At the genome-wide level, birth weight measures were the only traits that showed significant correlations with both gestational duration and SPTB in a comprehensive set of complex phenotypes. However, the Bonferroni correction deployed for the 773 tests is likely overly conservative since the traits include closely related phenotypes. Gene set enrichment for gestational duration and pathway analysis for gestational duration and SPTB highlighted involvement of gonad development and steroid hormone biosynthetic processes and GO terms referring to regulation of morphogenesis among multiple top pathways. Gene set enrichment analysis of SPTB implied kinetochore microtubule as the top pathway. Proper function of the kinetochore-microtubule pathway is essential for preserving genomic integrity and prevention of birth defects\textsuperscript{33}. Tissue analysis pinpointed several reproductive tissues, including uterus and ovary, among the most relevant tissue types for both SPTB- and gestational duration-associated genes.

Multiple variants in the candidate loci were individually associated with birth weight indices. Associated genes had primary roles in steroid hormone–regulating processes and tissue and organ morphogenesis. Reproductive tissues of the mother were among the principal locations where the associated genes were expressed. Our results suggest that many of the associated variants contribute to pregnancy outcomes by regulating expression of their target genes, mainly but not exclusively in reproductive tissues and immune cell types. Hence, our results indicate that those tissue and cell types are the most relevant when considering the regulatory events related to pregnancy and preterm birth, and should be the primary targets in future molecular biological studies of SPTB and gestational duration.

The current analysis was restricted to individuals of predominantly European descent. Future studies should include ancestrally diverse populations to better understand the genetic architecture of the timing of birth and to ensure the broad applicability of results from genetic studies\textsuperscript{34}. 
In conclusion, the current meta-analysis detected multiple loci that were associated with gestational duration or SPTB and produced intriguing candidates for further studies. Our results highlight the intricate nature of spontaneous birth as a trait and emphasize the importance of reproductive and immune tissues and cell types. The new genetic discoveries prime further research including large-scale complex investigations and individual regulatory pathway analyses utilizing labor-inducing tissues and cells. Studies may eventually reveal signaling pathways that activate spontaneous preterm birth and contribute towards effective prevention of SPTB.
Methods

Study cohorts, phenotype descriptions, and ethics information
We conducted a meta-analysis of SPTB (8,542 cases and 89,829 controls) and a quantitative meta-analysis of gestational duration (n=68,733) with data from mothers of European ancestry. The data originated from FinnGen, Northern/Central Finland, and 23andMe project.

The FinnGen research project (launched 2017) combines genome information with health care data from national registries. The project aims to collect data from 500,000 Finnish participants. Preterm and term birth were defined as births before and after 37 weeks of gestation. FinnGen preterm endpoint in Preparatory Phase Data Freeze 6 comprised individuals with World Health Organization International Classification of Diseases, Eight, Ninth, and Tenth Revision (ICD-8, ICD-9, and ICD-10) codes O60, 644, and 63497, respectively. We excluded births with ICD-9 code 644 (“early or threatened labor”) if they occurred after 37 weeks of gestation according to birth register data, and individuals with multiple gestation or birth, preeclampsia/eclampsia, and polyhydroamnios. Controls were people with spontaneous term birth. The GWAS of SPTB comprised 4,925 cases and 49,105 controls, and the GWAS of gestational duration comprised 24,391 mothers for whom gestational duration was available.

We additionally performed GWAS with a “strict” definition of SPTB, in which we only included cases with births indicated as spontaneous and preterm in the endpoint data.

FinnGen participants provided informed consent under the Finnish Biobank Act. Older cohorts with study-specific consents were transferred to the Finnish biobanks after approval by Fimea, the National Supervisory Authority for Welfare and Health. Recruitment protocols followed the biobank protocols approved by Fimea. The Coordinating Ethics Committee of the Hospital District of Helsinki and Uusimaa (HUS) approved the FinnGen study protocol (Nr HUS/990/2017).

The FinnGen study is approved by the Finnish Institute for Health and Welfare (permit numbers THL/2031/6.02.00/2017, THL/1101/5.05.00/2017, THL/341/6.02.00/2018, THL/2222/6.02.00/2018, THL/283/6.02.00/2019, THL/1721/5.05.00/2019, THL/1524/5.05.00/2020, and THL/2364/14.02/2020), the Digital and Population Data Service Agency (permit numbers VRK43431/2017-3, VRK/6909/2018-3, and VRK/4415/2019-3), the Social Insurance Institution (permit numbers KELA 58/522/2017, KELA 131/522/2018,

The study subjects from Northern and Central Finland were sampled in Oulu and Tampere University Hospital districts. Informed consent was obtained from the mother. SPTB was defined as birth prior to 36 wk + 1 d of gestation. Term birth was defined as birth at 38–41 wk (38 wk + 0 d to 41 wk + 6 d) of gestation. We excluded births with multiple gestation, preeclampsia, polyhydramnios, intrauterine growth restriction, placental abruption, anomalies of the fetus, clinical chorioamnionitis or acute septic infection in the mother, alcohol or narcotic use, and accidents. Term births were from families without previous preterm births. The analysis comprised 286 cases and 488 controls. Ethical approval was received from the participating centers (Oulu University Hospital 79/2003, 14/2010, and 73/2013).

Summary statistics of the 23andMe research program were obtained by request. Women included in the analyses provided informed consent and answered surveys online in accordance with a human subjects protocol approved by Ethical & Independent Review Services (http://www.eandireview.com). The summary data comprised unrelated mothers of European ancestry with self-reported gestational duration for their first singleton live birth. Individuals reporting a medical indication for preterm delivery were excluded. Preterm birth was defined as birth before 37 weeks of gestation, and control samples were people with term delivery. The meta-analysis included 43,568 individuals (3331 cases and 40,237 controls).

Data used in the replication and joint analysis originated from European women in the Nordic data sets including FIN cohort (N=888; Finland), MoBa (N=1,834; Norway), and DNBC (N=5,921; Denmark), for which the summary statistics were obtained via collaboration\textsuperscript{8,35,36}.
To protect individual level data, we show distribution of gestational duration in the Finnish studies as categories (Table S1). Other cohorts were described previously\textsuperscript{8,35,36}.

**DNA sample preparation, genotyping, imputation, and quality control**

Various methods were used to extract DNA from the FinnGen samples. Genotyping was done with Illumina and Affymetrix arrays (Illumina Inc, San Diego, CA and Thermo Fisher Scientific, Santa Clara, CA). Sample quality control (QC) entailed excluding individuals of uncertain sex, non-Finnish ancestry, high missingness (>5%), and excess heterozygosity (±4SD). For genotype QC, variants with missingness >2%, minor allele count (MAC)<3, and deviation from Hardy–Weinberg equilibrium (HWE) \(p<1\times10^{-6}\) were excluded. Imputation was conducted against a Finnish population–specific SISuv3 reference with Beagle4.1\textsuperscript{37}. Variants with imputation info (INFO)<0.7 were excluded (https://github.com/FINNGEN/finngen-documentation).

DNA from Northern/Central Finnish study was extracted with UltraClean Blood DNA Isolation Kit (MO BIO Laboratories, Inc., Carlsbad, CA), Puregene Blood Core Kit (Qiagen, Hilden, Germany), or prepIT-L2P kit (DNA Genotek, Ontario, Canada). Genotyping was performed with the Infinium HumanCoreExome BeadChip (Illumina, San Diego, CA) by the Technology Centre, Institute for Molecular Medicine Finland (FIMM), University of Helsinki. Variants with minor allele frequency (MAF)<1%, HWE \(p<1\times10^{-4}\), or genotyping rate <90%, and samples with >10% missingness, were excluded. Prephasing was conducted with SHAPEIT2\textsuperscript{38}, and imputation with IMPUTE2\textsuperscript{39}, against the 1000 Genomes Project (1KGP) v3 reference panel\textsuperscript{40}. Variants with INFO<0.7 were excluded.

DNA of the 23andMe samples was extracted from saliva samples, followed by genotyping with custom Illumina platforms by the National Genetics Institute (NGI). Samples with <97% European ancestry, and variants with HWE \(p<1\times10^{-20}\), call rate <95%, or allele frequency discrepancy with 1KGP Europeans, were excluded. Imputation was done with Minimac2\textsuperscript{42}, using the 1KGP phase1\textsuperscript{40}. 
GWAS and meta-analysis

FinnGen GWAS was conducted with Scalable and Accurate Implementation of GEneralized mixed model (SAIGE). Gestational duration were inverse normalized. GWAS covariates were age, sex, genotyping batch, and ten leading principal components. MAC was set to five.

We used SNPTESTv2 in GWAS of Northern/Central Finnish cohort. A frequentist case–control association test was implemented for SPTB, and a quantitative trait test for gestational duration was carried out with a linear model. Covariates were three multidimensional scaling dimensions, defined with Plink1.9. We used SNPTEST defaults to achieve mean centering and scaling of gestational duration and to apply quantile normalization. Post-GWAS QC entailed excluding variants with MAF<1% and SNPTEST info<0.7.

In the 23andMe data, preterm birth was analyzed with logistic regression, and linear regression was applied in the GWAS of gestational duration. Covariates were maternal age and the top five principal components.

We used METAL to conduct a fixed-effect inverse variance–weighted meta-analysis of SPTB and a sample size–weighted $p$-value–based meta-analysis of gestational duration. Sample size–based meta-analysis allows combining results when β-coefficients and standard errors from individual studies are in different units. Genomic coordinates of the meta-analysis cohorts were aligned into the GRCh38 coordinates. Genomic inflation factor was calculated with linkage disequilibrium score regression (LDSC). We report associations based on at least two individual meta-analysis cohorts, and excluded remaining rare variants with MAF<0.01%.

Characterization of association signals

We defined associated loci as genomic regions within a ±1 Mb window around the lead variant. The locus was defined as novel if there were no previous genome-wide significant associations for SPTB or gestational duration in the ±1 Mb window in the National Human Genome Research Institute–European Bioinformatics Institute (NHGRI-EBI) GWAS Catalog or in the preprint of meta-analysis of the timing of parturition.

We used LDSC to estimate SNP-based heritability and to test for genetic correlation between gestational duration or SPTB with a comprehensive set of phenotypes downloaded from the
Integrative Epidemiology Unit (IEU) OpenGWAS Project. We used FUMA GWAS (Functional Mapping and Annotation of Genome-Wide Association Studies) to aid functional annotation of the GWAS results. FUMA was used to prioritize genes for enrichment testing and assessment and visualization of tissue-specific expression among GTExv8 tissues. FUMA implements MAGMA (Multi-marker Analysis of GenoMic Annotation) in gene-based analyses and gene-set enrichment analyses for GWAS summary data with curated gene sets and GO terms from Molecular Signature Database, MSigDB. In addition, we tested lists of gestational duration- and STPB-associated candidate genes for averaged gene expression across GTExv8 tissues (hierarchical clustering) and for enrichment against various gene sets with hypergeometric tests in FUMA’s GENE2FUNC process.

To gain insight into the associated loci, we checked previous associations in the FinnGen R7 data and performed phenome-wide association study (PheWAS) within 1 Mb window around the meta-analysis index variants by querying GWAS data in the IEU OpenGWAS Project, which includes approximately 40,000 studies.

We performed colocalization analysis with HyprColoc to assess if associated variants were also quantitative trait loci (QTLs) that affect mRNA expression. Colocalization was tested for variants within a 1Mb window around the meta-analysis lead variant. We estimated betas from the meta-analysis Z-scores. We used expression QTLs (eQTLs) from the eQTLCatalogue, which contains uniformly processed cis-eQTLs from most of the available public studies. From the GTEx data in the eQTLCatalogue, we included eQTLs based on GTEx v8 and LCLs from an earlier release. We report colocalization results with a posterior probability (PP)>0.6 and meta-analysis \( p<5\times10^{-8} \). We did not test eQTLs with FDR>0.05.

**Replication and joint analysis**

We tested the lead variants within each locus for association in the replication data from the data set of Nordic birth studies. For most of the tested variants, replication data from two out of three cohorts (FIN and MoBa, N=2,722) was available. Joint test of the replication data and meta-analysis variants with \( p<1\times10^{-6} \) was done to see if additional loci reached genome-wide significance. We did not test replicated loci originally reported in Zhang et al. since the same replication data was used.
Figure 1. Meta-analysis of gestational duration and SPTB. A) Loci with genome-wide significant associations ($p < 5e-08$) are highlighted in the Manhattan plots. Chromosomal positions are shown at the x-axis, and the y-axis shows association $p$ values at the -$\log_{10}$ scale. The meta-analysis detected 14 loci associated with gestational duration and four loci associated with SPTB. Peaks highlighted in pink represent novel loci, and peaks highlighted in green show known loci. B) Annotations as number of SNPs per functional consequences on genes.
Figure 2. Genetic correlations between A) gestational duration or B) SPTB and other complex traits. Genetic correlation between gestational duration or SPTB and a comprehensive set of 773 complex traits was analyzed with LD score regression. Top 10 correlated traits, followed by their respective p values, are shown.
Figure 3. Colocalization analysis of meta-analysis associations with expression quantitative trait (eQTL) data. The variants are colored according to their LD ($r^2$) with the lead SNP, based on pairwise LD in European population of the 1000 Genomes Project Phase 3. A–C) ZBTB38 variants were implicated in gestational duration-linked gene regulation in various immune regulatory cell types. D) Variants in WNT3A locus colocalized with WNT3A expression in placenta.
Table 1. Loci associated with gestational duration in meta-analysis of 68,732 women of European ancestry. Loci highlighted in bold had no previous associations ($p < 5\text{e}^{-8}$) with gestational duration or preterm birth.

<table>
<thead>
<tr>
<th>Chr:Pos</th>
<th>Rsid</th>
<th>A1</th>
<th>Freq</th>
<th>Z-score</th>
<th>p-value</th>
<th>HetPVal</th>
<th>Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>5:158460958</td>
<td>rs6881817</td>
<td>T</td>
<td>0.244</td>
<td>-11.41</td>
<td>3.74E-30</td>
<td>0.726</td>
<td>EBF1</td>
</tr>
<tr>
<td>1:22141722</td>
<td>rs3820282</td>
<td>T</td>
<td>0.149</td>
<td>9.189</td>
<td>3.95E-20</td>
<td>0.075</td>
<td>WNT4</td>
</tr>
<tr>
<td>23:116053119</td>
<td>rs5950512</td>
<td>A</td>
<td>0.401</td>
<td>-8.791</td>
<td>1.48E-18</td>
<td>0.512</td>
<td>AGTR2</td>
</tr>
<tr>
<td>3:128159974</td>
<td>rs2999049</td>
<td>T</td>
<td>0.720</td>
<td>-7.972</td>
<td>1.56E-15</td>
<td>0.258</td>
<td>EEFSEC</td>
</tr>
<tr>
<td>3:123365694</td>
<td>rs10934646</td>
<td>A</td>
<td>0.634</td>
<td>-7.193</td>
<td>6.35E-13</td>
<td>0.121</td>
<td>ADYS</td>
</tr>
<tr>
<td>1:228015567</td>
<td>rs708119</td>
<td>C</td>
<td>0.670</td>
<td>6.514</td>
<td>7.32E-11</td>
<td>0.207</td>
<td>WNT3A</td>
</tr>
<tr>
<td>3:141414608</td>
<td>rs1991431</td>
<td>A</td>
<td>0.441</td>
<td>6.402</td>
<td>1.54E-10</td>
<td>0.583</td>
<td>ZBTB38</td>
</tr>
<tr>
<td>6:49592080</td>
<td>rs10948514</td>
<td>T</td>
<td>0.274</td>
<td>6.267</td>
<td>3.68E-10</td>
<td>0.999</td>
<td>RHAG</td>
</tr>
<tr>
<td>2:74009288</td>
<td>rs71848031</td>
<td>D</td>
<td>0.057</td>
<td>-6.061</td>
<td>1.36E-09</td>
<td>0.913</td>
<td>TET3</td>
</tr>
<tr>
<td>3:156137629</td>
<td>rs4679760</td>
<td>C</td>
<td>0.426</td>
<td>-5.992</td>
<td>2.07E-09</td>
<td>0.360</td>
<td>KCNAB1</td>
</tr>
<tr>
<td>4:173813320</td>
<td>rs7697038</td>
<td>A</td>
<td>0.323</td>
<td>5.780</td>
<td>7.45E-09</td>
<td>0.355</td>
<td>HAND2</td>
</tr>
<tr>
<td>5:114208039</td>
<td>rs13175113</td>
<td>T</td>
<td>0.486</td>
<td>5.668</td>
<td>1.45E-08</td>
<td>0.174</td>
<td>KCNN2</td>
</tr>
<tr>
<td>7:51940143</td>
<td>rs15114398</td>
<td>T</td>
<td>0.996</td>
<td>5.655</td>
<td>1.56E-08</td>
<td>0.712</td>
<td>COBL</td>
</tr>
<tr>
<td>9:77936015</td>
<td>rs11145617</td>
<td>A</td>
<td>0.210</td>
<td>-5.586</td>
<td>2.32E-08</td>
<td>0.584</td>
<td>GNAQ</td>
</tr>
<tr>
<td>9:114160401</td>
<td>rs2808791</td>
<td>T</td>
<td>0.474</td>
<td>5.467</td>
<td>4.58E-08</td>
<td>0.429</td>
<td>COL27A1</td>
</tr>
</tbody>
</table>

Chr, chromosome; Pos, position (Grch38); A1, effect allele; HetPval, p value for heterogeneity. Novel loci are highlighted in bold; reported genes are the nearest genes. For replicated loci, candidate genes are based on previous studies.
Table 2. Loci associated with SPTB in meta-analysis of 98,372 women of European ancestry. Loci highlighted in bold had no previous associations \((p < 5\text{e}-8)\) with gestational duration or preterm birth.

<table>
<thead>
<tr>
<th>Chr:Pos</th>
<th>Rsid</th>
<th>A1</th>
<th>Freq</th>
<th>OR</th>
<th>CI95%</th>
<th>p-value</th>
<th>HetPVal</th>
<th>Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>5:158467739</td>
<td>rs12520982</td>
<td>T</td>
<td>0.714</td>
<td>0.818</td>
<td>0.78-0.86</td>
<td>1.71E-13</td>
<td>0.385</td>
<td>EBF1</td>
</tr>
<tr>
<td>3:128348271</td>
<td>rs1553758423</td>
<td>D</td>
<td>0.707</td>
<td>1.210</td>
<td>1.14-1.28</td>
<td>2.32E-10</td>
<td>0.630</td>
<td>EEFSEC</td>
</tr>
<tr>
<td>4:71763252</td>
<td>rs2276461</td>
<td>A</td>
<td>0.052</td>
<td>1.290</td>
<td>1.18-1.41</td>
<td>9.84E-09</td>
<td>0.626</td>
<td>GC</td>
</tr>
<tr>
<td>12:126705821</td>
<td>rs192808132</td>
<td>A</td>
<td>0.993</td>
<td>0.581</td>
<td>0.48-0.70</td>
<td>1.99E-08</td>
<td>0.462</td>
<td>LINC02824</td>
</tr>
</tbody>
</table>

Chr, chromosome; Pos, position (Grch38); A1, effect allele; OR, odds ratio; CI95%, 95% confidence interval; HetPval, \(p\) value for heterogeneity. Novel loci are highlighted in bold; reported genes are the nearest genes. For replicated loci, candidate genes are based on previous studies.


