Transcriptome and regulatory maps of decidua-derived stromal cells inform gene discovery in preterm birth – Supplementary Information

Supplementary File 1 – Excel spreadsheet showing statistics for the data generated

Supplementary File 2 – Gene Ontology terms of differentially expressed genes

Supplementary File 3 – This document

Supplementary File 4 – Full fine-mapping results of high PIP SNPs

Supplementary File 5 – TORUS estimates

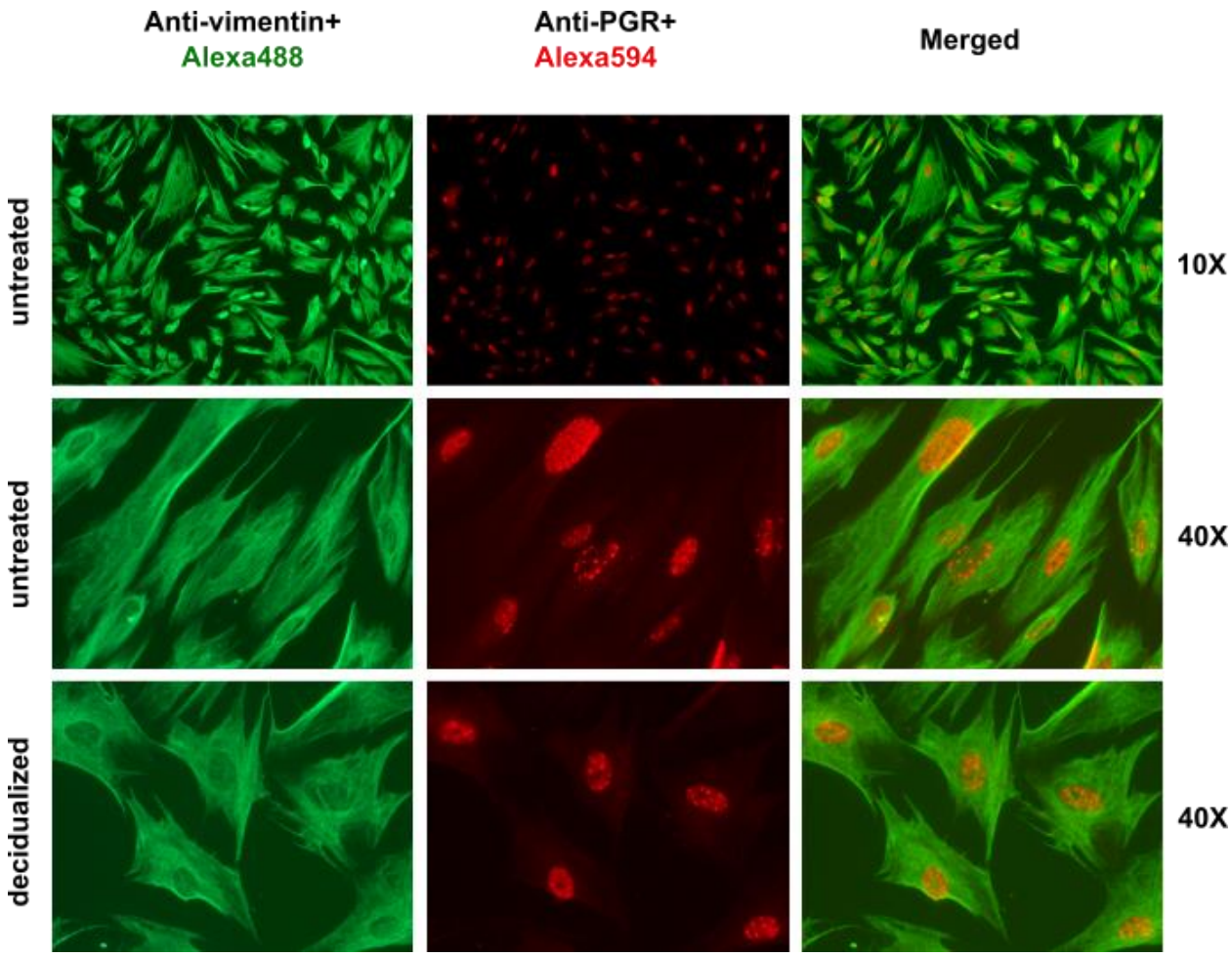

Figure S1. Immunofluorescence characterization of cultured decidua-derived mesenchymal stromal cells (MSCs) collected from human placental membrane. Cells were fixed with 4% paraformaldehyde for 15 minutes and permeabilized with 0.5% Triton X-100 for 15 minutes. Samples were incubated in 10% normal goat serum solution containing 1/100 dilution of the primary antibody anti-vimentin (Biolegend,677801) and 1/50 dilution of the anti-progesterone receptor (PGR; Abcam, ab62621) for 3 hours. Conjugated secondary antibodies (1/1000 dilution) were goat anti-mouse Alexa Fluor®488- (Thermofisher, A-11001) and goat anti-rabbit Alexa 594 (Thermofisher, A-11037). Images are representative of three cell lines control (untreated) or after 48 h treatment with 0.5mM 8-Br-cAMP+ 1uM MPA (decidualized). Images have been taken on widefield microscope using same expositions and any post imaging treatment are the same in all the images.

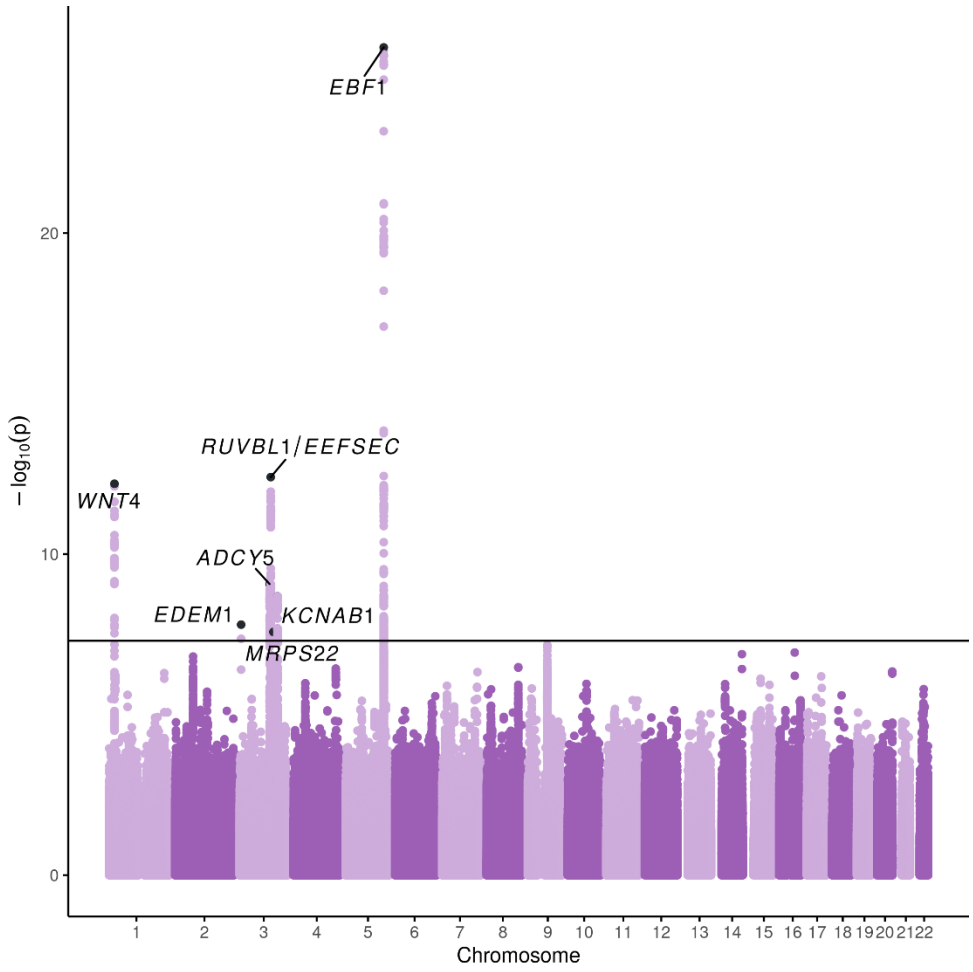

Figure S2– Manhattan plot of a GWAS for gestational duration GWAS. The horizontal black line denotes the threshold for genome-wide significance ($p < 5x10^{-8}$). For the 6 independent genome-wide significant loci, the most significant p-value is highlighted in black, and labeled with the nearest gene(s). See Figure S2 for the QQ plot of these data, Table S1 for a description of the lead SNP, and Table S5 for a description of the study populations.

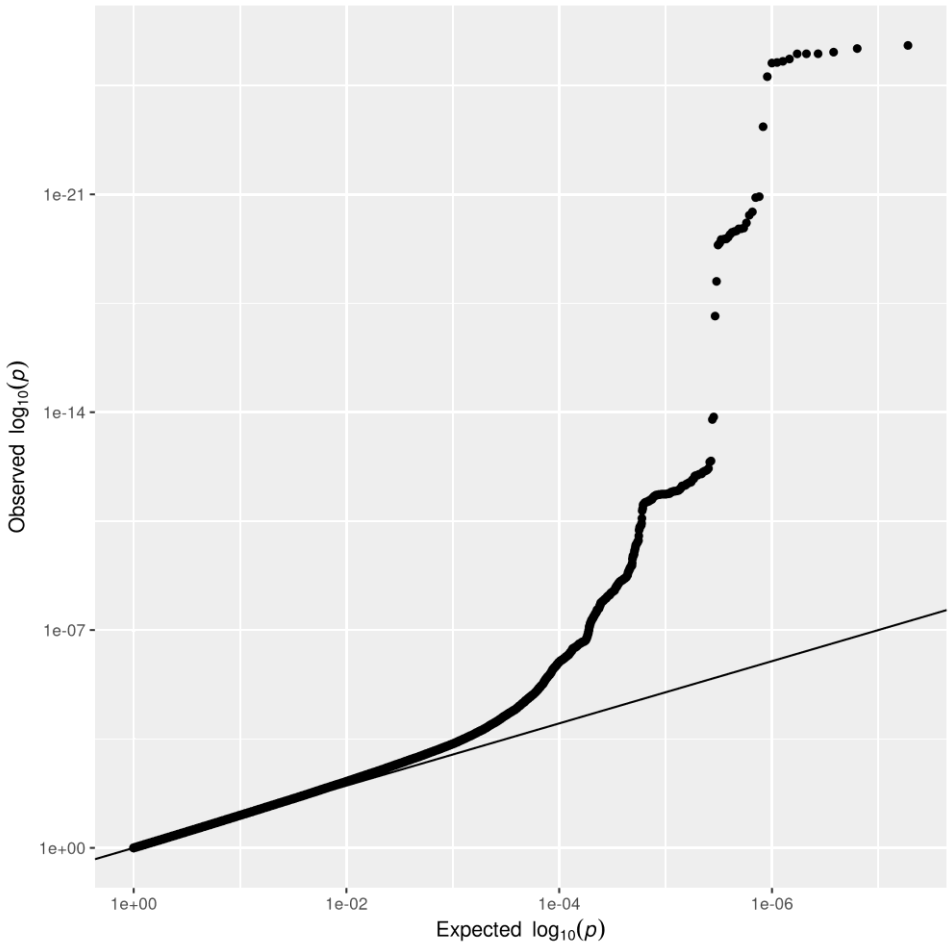

Figure S3 - QQ plot of p-values from the GWAS of gestational duration.

Table S1 – Lead SNPs at each genome-wide significant locus. AF: allele frequency. 1KG_EUR: European ancestry from 1000 Genome Project.

SNP	p-value	locus	ID	Beta(S.E)	GWAS_AF	1KG_EUR_AF
chr1:22468215:C:T	6E−13	chr1:21736588-23086883	rs3820282	-0.9508(0.132)	0.149	0.1421
chr3:123085359:T:C	9E−10	chr3:121974097-123517768	rs6794886	0.5766(0.094)	0.522	0.555
chr3:127869598:C:A	4E−13	chr3:126214943-128194861	rs144609957	0.769(0.106)	0.274	0.265
chr3:139004333:A:G	3E−08	chr3:137371083-139954597	rs62270785	2.152(0.387)	0.0155	0.0149
chr3:155855501:A:AT	4E−10	chr3:154714218-156008700	rs66960245	-0.6232(0.099)	0.459	0.518
chr5:157888115:T:C	2E−26	chr5:156628700-158825698	rs6881996	1.141(0.107)	0.741	0.732

Table S2 – Estimates of the enrichment parameters of TORUS. The parameters (alpha1) represent log-OR enrichment of GWAS causal variants in a given annotation. The first row shows the intercept term in TORUS logistic regression model.

Model Parameter	Model Estimate	95% CI Low	95% CI High	Parameter p-value
Intercept	−12.54	−12.57	−12.52	0
Untreated-H3K4me3	0.292	−1.207	1.791	0.351
Decidualized-H3K27ac	0.179	−0.469	0.827	0.294
Decidualized-H3K4me1	3.15	2.206	4.093	3E−11
Decidualized-HiC	1.35	−0.13	2.83	0.0369

Table S3 - All fine-mapped SNPs with PIP > 0.01 at the HAND2 locus (chr4:174264132-176570716)

SNP	ID	annotation	PIP	p-value	pcHi-C genes
chr4:174726131 (T/G)	rs5014764	H3K4me1, H3K27ac	0.0399	7E−07	
chr4:174728566 (T/G)	rs13121843	H3K4me1, H3K27ac, H3K4me3, ATAC, pcHi-C	0.0505	8E−07	HAND2
chr4:174728703 (C/T)	rs13141656	H3K4me1, H3K27ac, H3K4me3, ATAC, pcHi-C	0.381	4E−07	HAND2
chr4:174729014 (G/A)	rs7663453	H3K4me1, H3K27ac, H3K4me3, pcHi-C	0.329	4E−07	HAND2
chr4:174729270 (A/G)	rs7689307	H3K4me1, H3K27ac, H3K4me3, pcHi-C	0.0743	5E−07	HAND2
chr4:174729550 (C/T)	rs12512745	H3K4me1, H3K27ac, pcHi-C	0.0576	5E−07	HAND2
chr4:174741209 (C/T)	rs13140184	H3K4me1, H3K27ac	0.0506	6E−07	

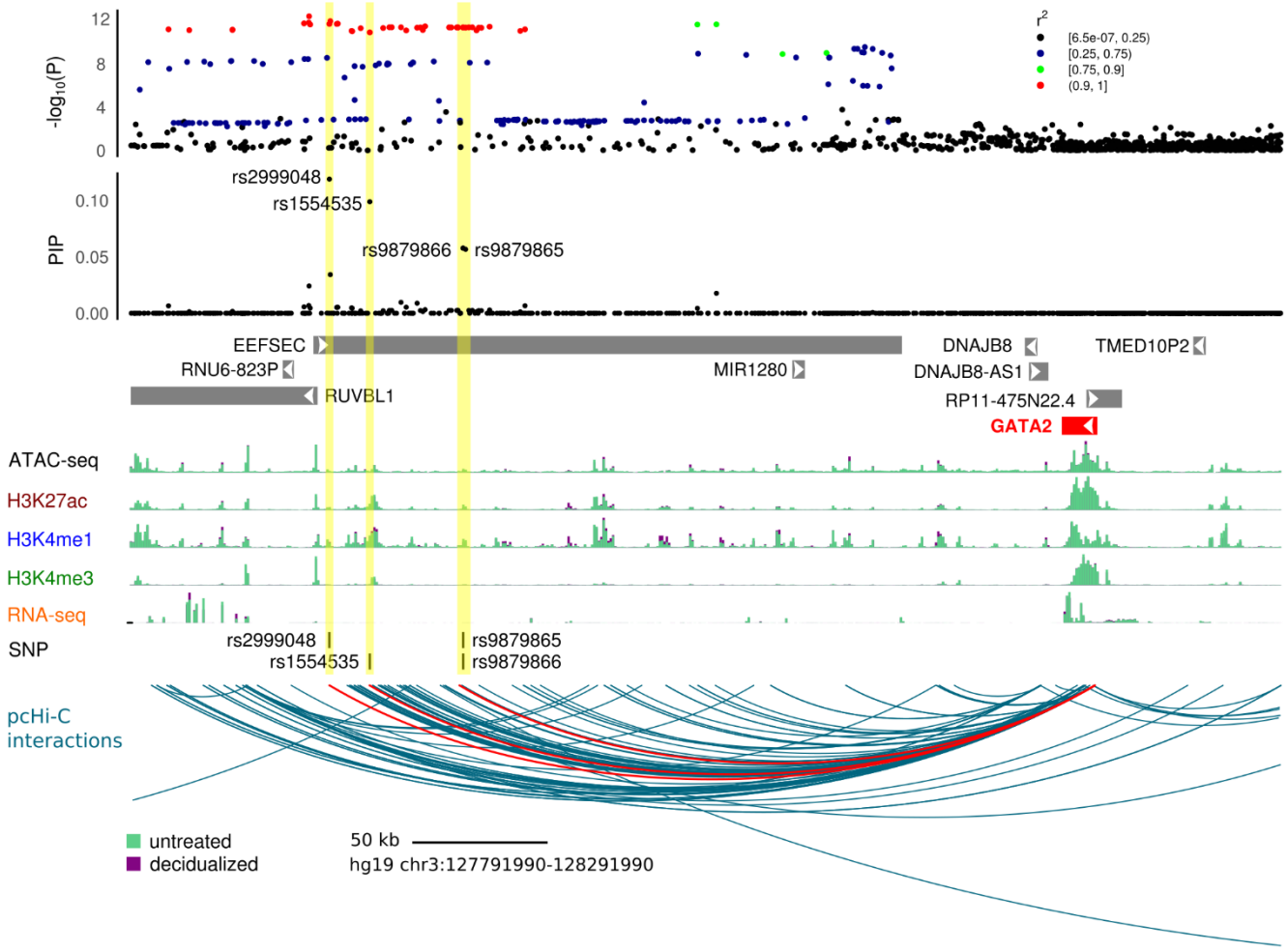

Figure S4 - Fine-mapping a GWAS locus of gestational duration: likely causal variants near *GATA2* and their functional annotations. The upper panel shows the significance of the SNPs in the GWAS and the middle panel shows fine-mapping results (PIPs) in the region. The vertical yellow bar highlights the four SNPs with high PIPs. These SNPs are located in a region annotated with ATAC-seq, H3K27ac, H3K4me1 and H3K4me3 peaks (lower panel). The sequences containing the four SNPs all interact with the *GATA2* promoter (red arcs). rs2999048 is spanned by an H3K4me1 peak in 3/129 tissues of the Epigenome Roadmap data set whereas rs1554535 is not spanned by enhancer marks in any tissue. rs9879865 and rs9879866 are spanned by H3K27ac or H3K4me1 peaks in 24 and 26 tissues, respectively.

Table S4 - All SNPs with PIP > 0.01 at the *GATA2* locus (chr3:126214943-128194861)

SNP	ID	annotation	PIP	p-value	pcHi-C genes
chr3:127822320 (C/T)	rs7641133	H3K4me1	0.0346	7E−12	
chr3:127869598 (C/A)	rs144609957		0.0242	4E−13	
chr3:127878416 (G/A)	rs2999048	H3K4me1, H3K27ac, pcHi-C	0.118	2E−12	GATA2
chr3:127878817 (T/C)	rs2999049	pcHi-C	0.0343	1E−12	GATA2
chr3:127889287 (A/G)	rs3122173	H3K4me1, pcHi-C	0.181	5E−12	GATA2
chr3:127895226 (A/G)	rs2687729	H3K4me1, pcHi-C	0.0493	5E−12	GATA2
chr3:127895986 (G/A)	rs1554535	H3K4me1, pcHi-C, H3K27ac	0.0989	1E−11	GATA2
chr3:127898501 (A/C)	rs2811476	H3K4me1, ATAC, H3K4me3, H3K27ac, pcHi-C	0.0649	6E−12	GATA2
chr3:127936527 (T/C)	rs9879865	H3K4me1, ATAC, H3K27ac, pcHi-C	0.0578	4E−12	GATA2
chr3:127936532 (T/C)	rs9879866	H3K4me1, ATAC, H3K27ac, pcHi-C	0.0578	4E−12	GATA2
chr3:127937645 (G/A)	rs9847576	H3K4me1, H3K27ac	0.0567	4E−12	
chr3:128046643 (G/A)	rs4857841	pcHi-C	0.0177	2E−12	GATA2

GWAS

The GWAS results used in this study were an extension of previously published results[1]. As in the previous study, we included summary results from 23andMe for a GWAS of gestational duration in 42,121 mothers of European ancestry who reported gestational duration of their first live singleton birth. Meta-analyzed those data with results of GWASs of gestational during in 14,263 European mothers from six additional studies. The sample size of each study is shown in Table S5. The description of these data sets and the association test procedures are provided below.

Table S5 - Sample sizes of the data sets included in our GWAS of gestational duration

Data sets	Male	Female	Total
23andMe (U.S.)	**21779**	**20342**	**42121**
Six European data sets	**7252**	**7011**	**14263**
ALSPAC	*3820*	*3783*	*7603*
DNBC	*1001*	*911*	*1912*
MoBa	*891*	*913*	*1804*
FIN	*699*	*623*	*1322*
HAPO	*610*	*593*	*1203*
GPN	*231*	*188*	*419*

Study descriptions

Study 1: 23andMe GWA summary results of gestational duration from 23andMe (www.23andme.com, Sunnyvale, CA, USA) as described in Zhang et al[1].

Study 2: The Avon Longitudinal Study of Parents and Children (ALSPAC) is a prospective birth cohort study. 14,541 pregnant women resident in the former county of Avon (situated around the city of Bristol in the South West of England) with expected dates of delivery 1st April 1991 to 31st December 1992 were recruited[2,3]. The children arising from these women, and their partners were followed up intensively over nearly three decades. Genotype data of the mothers

and children were generated using the Illumina HumanHap550 quad (children) and Illumina human660W quad (mothers). This resulted in a dataset of 17,842 participants (either mothers or offspring), each with 465,740 SNPs genotyped. From this data set, 7,603 mothers who passed genotype QC and inclusion/exclusion criteria were included in the analysis. Informed consent for the use of data collected via questionnaires and clinics was obtained from participants following the recommendations of the ALSPAC Ethics and Law Committee at the time. The study website contains details of all the data that is available through a fully searchable data dictionary and variable search tool (http://www.bristol.ac.uk/alspac/researchers/our-data/). Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. Consent for biological samples has been collected in accordance with the Human Tissue Act (2004).

Study 3: The Danish National Birth Cohort (DNBC) followed over 100,000 pregnancies between 1996 and 2003 with extensive epidemiologic data on health outcomes in both mother and child[4]. The current study used the data downloaded from the Database of Genotypes and Phenotypes (dbGaP) (phs000103.v1.p1), which contains data from a genome-wide case/control study using approximately 1,000 preterm mother-child pairs (gestational age between 22–37 weeks) from the DNBC, along with 1,000 control pairs in which the child was born at ~40 weeks gestation. Gestational duration in this dataset was determined by a consensus algorithm combining all available information from multiple sources: self-reported date of last menstrual period, self-reported delivery date, and gestational age at birth registered in the Medical Birth Register and the National Patient Register. We identified 1,912 mothers and used them in the GWA analysis. The study protocol was approved by the Danish Scientific Ethical Committee and the Danish Data Protection Agency.

Study 4: The Mother Child dataset of Norway (MoBa) is a nationwide Norwegian pregnancy study administered by the Norwegian Institute of Public Health. The study includes more than 114,000 children, 95,000 mothers and 75,000 fathers recruited from 1999 through 2008[5]. Gestational age was estimated by ultrasound at gestational weeks 17–19. In the few cases without ultrasound dating, gestational age was estimated using the date of the last menstrual period. Singleton live-born spontaneous pregnancies with mothers in the age group 20–34 years were selected. Random sampling was done from two gestational age ranges 154-258 days (cases)

and 273–286 days (controls). Pregnancies involving pre-existing medical conditions, pregnancies with complications as well as pregnancies conceived by in vitro fertilization, were excluded from the study. In total, blood samples from 3,120 mothers and children were genotyped[6]. 1,804 mothers that passed QC and inclusion/exclusion criteria were included in the analysis. All parents gave informed, written consent. The study was approved by The Regional Committee for Medical Research Ethics in South-Eastern, Norway.

Study 5: The Finnish dataset (FIN) was collected for a genetic study of spontaneous preterm birth[7]. Briefly, whole blood samples were collected from more than 1,600 mother/child pairs from the Helsinki (southern Finland) University Hospitals between 2004 and 2014. All the studied samples are of Finnish descent. Crown-rump length at the first ultrasound screening between 10+ and 13 weeks was used to determine the gestational age. 2,962 blood samples from mothers and children were genotyped. After genotype quality control (QC) procedure and applying the phenotype-based inclusion/exclusion criteria, 1,322 mothers were selected and used in the analysis. The study was approved by the Ethics Committee of Oulu University Hospital and that of Helsinki University Central Hospital. Written informed consent was given by all participants.

Study 6: The Hyperglycemia and Adverse Pregnancy Outcome (HAPO) Study is a multicenter, international study in which high quality phenotypic data related to fetal growth and maternal glucose metabolism has been collected from 25,000 pregnant women of varied racial and socio-demographic backgrounds using standardized protocols that were uniform across centers. For the current study, we utilized phenotype and genotype data of European descent downloaded from dbGaP (phs000096.v2.p1). Gestational duration in this dataset was determined by last menstrual period or ultrasound estimation from 6-24 weeks. After genotype quality control (QC) and application of the phenotype-based inclusion/exclusion criteria, we identified 1,203 mothers and used them in the GWA analysis.

Study 7: The Genomic and Proteomic Network for Preterm Birth Research (GPN) Study is a multicenter observational genome-wide association study (GWAS) designed to determine the genetic predisposition to idiopathic preterm birth. Phenotype data and genotype data from 743 spontaneous preterm births (20 to less than 34 weeks gestation), and 752 controls (39 to less than

42 weeks gestation) of diverse ethnic background (White, Hispanics, African Americans, and Others) were collected. For this current study, we identified 419 mothers of European descent from the data downloaded from dbGaP (phs000714.v1.p1).

Data selection

For the GWAS, we included only singleton pregnancies with spontaneous live birth deliveries with or without premature rupture of membranes (PROM). C-sections after spontaneous onset of labor were retained. Medically indicated induced deliveries or C-sections were excluded. Pregnancies with known gestational or fetal complications (e.g. placental abnormities, chorioamnionitis, preeclampsia, and congenital anomalies) and pregnancies involving pre-existing medical conditions (i.e. hypertension or diabetes) or maternal risk exposure (e.g. drug use during pregnancy) known be associated with preterm birth were also excluded.

Genotyping

Genotyping of these data sets was conducted on DNA extracted from blood using the various SNP arrays. Specifically: the ALSPAC genotype data were generated using the Illumina HumanHap550 quad (children) and Illumina human660W quad (mothers). The cleaned genotype calls of 465,740 SNPs of 17,842 subjects were obtained from ALSPAC. The DNBC samples were genotyped using Human660W-Quad bead arrays from Illumina. The raw genotype intensity data (.idat) files were obtained from dbGaP and genotype calls were performed using the CRLMM algorithm[8,9]. The samples from the MoBa dataset were genotyped using the Illumina Human660W-Quadv1_A bead chip (Illumina Inc.) and the genotype calls were determined using the CRLMM algorithm. For the FIN dataset, genotyping was conducted using Affymetrix 6.0 (Affymetrix, California, United States) and various other Illumina arrays (Illumina, California, United States). For the Affymetrix SNP Array 6.0, genotype calls were determined using the CRLMM algorithm in chips that passed the vendor-suggested QC (Contrast QC > 0.4). For the Illumina chips, the genotype calling was conducted using Illumina's genotyping module v1.94 in GenomeStudio v2011.1. The HAPO samples of European descent were genotyped using Human610-Quad array. We obtained the raw genotype intensity data (.idat) files and performed genotype calls using the CRLMM algorithm. The processed genotype calls in plink format of the GPN data set were obtained from dbGaP (phs000714.v1.p1). Data from participants of apparent

duplications with others (IBD>0.8), with sex discrepancies (between known sex and genetically inferred sex), and children with high Mendelian errors (>10%) were removed.

Genotyping QC

We performed similar genotype QC across all the dataset. We first performed sample-level QC based on call rate, overall heterogeneity and sex discrepancies. We checked the pedigree relationship based on IBD analysis. Genetic ancestry was assessed by principal components analysis (PCA) anchored by 1000 Genomes reference samples. Individuals with non-European ancestry were excluded. We then performed marker level QC: SNPs with low call rate (<98%), low minor allele frequency (<0.01) or significant deviation from Hardy-Weinberg Equilibrium ($P < 5\times10^{-6}$) were excluded.

Imputation

We conducted genome wide imputation following a standard two-step imputation procedure: the genotype data was first pre-phased together using the Shapeit2 software[10] and then the estimated haplotypes were used to impute non-genotyped SNPs using the reference haplotypes extracted from the Phase III 1000 Genomes Project[11] using Minimac4[12] (FIN and MoBa).

Association analysis

Single-marker genetic association tests were conducted in individual data sets separately, using regression methods and imputed genotype data. Fetal sex was included as covariate. The test results from the 7 data sets were then combined by fixed effect meta-analysis using the inverse-variance method.

Public data used in this study

Table S6 – Transcription factor ChIP-seq

PGR	GSE94038	SRR2051516	GSM1703567
NR2F2	GSE52008	SRR1023110	GSM1257396
FOSL2	GSE94038	SRR2051517	GSM1703568
FOXO1	GSE94037	SRR2051519	GSM1703607
input-NR2F2	GSE52008	SRR1023111	GSM1257397
input-FOXO1+PolII+PGR+FOSL2	GSE94038	SRR2051518	GSM1703569
GATA2	GSE108409	SRR6410464	GSM2897813
input-GATA2	GSE108409	SRR6410465	GSM2897814

References

1 Zhang, G. *et al.* Genetic Associations with Gestational Duration and Spontaneous Preterm Birth. *N Engl J Med* **377**, 1156-1167, doi:10.1056/NEJMoa1612665 (2017).

2 Boyd, A. *et al.* Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. *International journal of epidemiology* **42**, 111-127, doi:10.1093/ije/dys064 (2013).

3 Fraser, A. *et al.* Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *International journal of epidemiology* **42**, 97-110, doi:10.1093/ije/dys066 (2013).

4 Olsen, J. *et al.* The Danish National Birth Cohort--its background, structure and aim. *Scandinavian journal of public health* **29**, 300-307 (2001).

5 Magnus, P. *et al.* Cohort Profile Update: The Norwegian Mother and Child Cohort Study (MoBa). *International journal of epidemiology* **45**, 382-388, doi:10.1093/ije/dyw029 (2016).

6 Myking, S. *et al.* X-chromosomal maternal and fetal SNPs and the risk of spontaneous preterm delivery in a Danish/Norwegian genome-wide association study. *PLoS One* **8**, e61781, doi:10.1371/journal.pone.0061781 (2013).

7 Plunkett, J. *et al.* An evolutionary genomic approach to identify genes involved in human birth timing. *PLoS genetics* **7**, e1001365, doi:10.1371/journal.pgen.1001365 (2011).

8 Carvalho, B., Bengtsson, H., Speed, T. P. & Irizarry, R. A. Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. *Biostatistics* **8**, 485-499, doi:10.1093/biostatistics/kxl042 (2007).

9 Scharpf, R. B., Irizarry, R. A., Ritchie, M. E., Carvalho, B. & Ruczinski, I. Using the R Package crlmm for Genotyping and Copy Number Estimation. *Journal of statistical software* **40**, 1-32 (2011).

10 Delaneau, O., Marchini, J. & Zagury, J. F. A linear complexity phasing method for thousands of genomes. *Nature methods* **9**, 179-181, doi:10.1038/nmeth.1785 (2012).

11 Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:10.1038/nature11632 (2012).

12 Fuchsberger, C., Abecasis, G. R. & Hinds, D. A. minimac2: faster genotype imputation. *Bioinformatics* **31**, 782-784, doi:10.1093/bioinformatics/btu704 (2015).