ORIGINAL ARTICLE

Genetic Associations with Gestational Duration and Spontaneous Preterm Birth

G. Zhang, B. Feenstra, J. Bacelis, X. Liu, L.M. Muglia, J. Juodakis, D.E. Miller, N. Litterman, P.-P. Jiang, L. Russell, D.A. Hinds, Y. Hu, M.T. Weirauch, X. Chen, A.R. Chavan, G.P. Wagner, M. Pavličev, M.C. Nnamani, J. Maziarz, M.K. Karjalainen, M. Rämet, V. Sengpiel, F. Geller, H.A. Boyd, A. Palotie, A. Momany, B. Bedell, K.K. Ryckman, J.M. Huusko, C.R. Forney, L.C. Kottyan, M. Hallman, K. Teramo, E.A. Nohr, G. Davey Smith, M. Melbye, B. Jacobsson, and L.J. Muglia

ABSTRACT

BACKGROUND

Despite evidence that genetic factors contribute to the duration of gestation and the risk of preterm birth, robust associations with genetic variants have not been identified. We used large data sets that included the gestational duration to determine possible genetic associations.

METHODS

We performed a genomewide association study in a discovery set of samples obtained from 43,568 women of European ancestry using gestational duration as a continuous trait and term or preterm (<37 weeks) birth as a dichotomous outcome. We used samples from three Nordic data sets (involving a total of 8643 women) to test for replication of genomic loci that had significant genomewide association ($P<5.0\times10^{-8}$) or an association with suggestive significance ($P<1.0\times10^{-6}$) in the discovery set.

RESULTS

In the discovery and replication data sets, four loci (*EBF1*, *EEFSEC*, *AGTR2*, and *WNT4*) were significantly associated with gestational duration. Functional analysis showed that an implicated variant in *WNT4* alters the binding of the estrogen receptor. The association between variants in *ADCY5* and *RAP2C* and gestational duration had suggestive significance in the discovery set and significant evidence of association in the replication sets; these variants also showed genomewide significance in a joint analysis. Common variants in *EBF1*, *EEFSEC*, and *AGTR2* showed association with preterm birth with genomewide significance. An analysis of mother–infant dyads suggested that these variants act at the level of the maternal genome.

CONCLUSIONS

In this genomewide association study, we found that variants at the *EBF1*, *EEFSEC*, *AGTR2*, *WNT4*, *ADCY5*, and *RAP2C* loci were associated with gestational duration and variants at the *EBF1*, *EEFSEC*, and *AGTR2* loci with preterm birth. Previously established roles of these genes in uterine development, maternal nutrition, and vascular control support their mechanistic involvement. (Funded by the March of Dimes and others.)

The authors' full names, academic degrees, and affiliations are listed in the Appendix. Address reprint requests to Dr. Muglia at Cincinnati Children's Hospital Medical Center, MLC 7009, 3333 Burnet Ave., Cincinnati, OH 45229-3026, or at louis.muglia@cchmc.org.

Drs. Jacobsson and Muglia contributed equally to this article.

This article was published on September 6, 2017, at NEJM.org.

This is the *New England Journal of Medicine* version of record, which includes all *Journal* editing and enhancements. The Author Final Manuscript, which is the author's version after external peer review and before publication in the *Journal*, is available under a CC BY license at PMC5561422.

N Engl J Med 2017;377:1156-67.
DOI: 10.1056/NEJMoa1612665

Preterm birth (defined as birth before 37 completed weeks of gestation) affects 9.6% of pregnancies in the United States[1] and more than 15 million pregnancies worldwide each year. It is the leading cause of death in neonates and children under the age of 5 years.[2,3] The majority of preterm births arise by spontaneous, idiopathic onset of uterine contractions or rupture of fetal membranes.[4] A substantial body of evidence has shown genetic influence in the duration of gestation and the risk of preterm birth.[5] For example, twin and family studies suggest that 30 to 40% of the variation in birth timing and in the risk of preterm birth arises from genetic factors that largely but not exclusively reside in the maternal genome.[6-10]

Preterm birth, and gestational duration in general, is a complicated phenotype that is affected by both maternal and fetal genomes. The definition of preterm birth as a dichotomous trait on the basis of a somewhat arbitrary cutoff of 37 weeks of gestation, rather than time of birth for a specified level of fetal maturity or as a continuous trait, limits the interpretation of data and reduces the statistical power to detect association.[11] Therefore, we tested for genetic associations with gestational duration (a quantitative trait) as well as preterm birth (a dichotomous trait). To date, individual genomewide association studies (see Glossary) of spontaneous preterm birth have included approximately 1000 case mothers or infants with control groups of similar size. No replicated loci with genomewide significance have been reported.[12-14] To overcome sample-size limitations, we conducted genomewide discovery in a large cohort of women of European ancestry and tested associations that were identified in the discovery set for replication in three Nordic data sets.

Glossary

1000 Genomes Project: An international collaboration to produce an extensive public catalogue of human genetic variation. Phase 1 of the project described genomes of 1092 samples from 14 populations. The haplotypes that are inferred from the 1000 Genomes Project data (reference haplotypes) can be used to impute genotypes at sites not on the original SNP array.

Genomewide association study (GWAS): An approach used in genetics research to look for associations between many (typically hundreds of thousands) specific genetic variations (most commonly, single-nucleotide polymorphisms [SNPs]) and particular diseases or traits.

Genomewide complex trait analysis (GCTA): A computational tool that was originally designed to estimate the proportion of phenotypic variance explained by genomewide or chromosomewide SNPs for complex traits, has many other functions to analyze the genetic architecture of complex traits.

GTEx: The Genotype–Tissue Expression project collects and analyzes multiple human tissues from donors to assess correlations between genotype and tissue-specific levels of gene expression. If the expression levels of genes are treated as quantitative traits, variations in gene expression that are highly correlated with genetic variation can be identified as expression quantitative trait loci.

GWAS Catalog: A quality controlled, manually curated, literature-derived collection of all published genomewide association studies assaying at least 100,000 SNPs and all SNP-trait associations with P values of less than 1.0×10^{-5}. The GWAS Catalog is provided jointly by the National Human Genome Research Institute and the European Bioinformatics Institute.

Locus: The specific chromosomal location of a gene or other DNA sequence of interest.

Phenotype variance: The proportion of interindividual difference (variance) in a phenotypic trait usually combines the genotype variance with the environmental variance.

RNA sequencing: RNA sequencing uses next-generation sequencing to detect and quantify RNA in a biologic sample to measure transcript abundance.

Single-nucleotide polymorphism (SNP): A single-nucleotide variation in a genetic sequence; a common form of variation in the human genome.

Transcript: An RNA sequence resulting from transcription of a DNA sequence (often a gene).

METHODS

DISCOVERY DATA SET

We performed a two-stage genomewide association study to discover and replicate genetic loci associated with gestational duration and the risk of preterm birth. Women in the discovery data set were participants in the research program of 23andMe, a personal genomics and biotechnology company. All the women provided written informed consent and answered surveys online according to a human-subjects protocol approved by Ethical and Independent Review Services (www.eandireview.com). Unrelated women of European ancestry who self-reported the gestational duration of their first live singleton birth were included in the analysis. Women with a medical indication for their preterm delivery were excluded; those who did not specify a medical indication were included to optimize sample size. Preterm-birth status was determined on the basis of dichotomization of gestational duration (preterm, <37 weeks; term, ≥37 weeks).

DNA extraction and genotyping were performed by the National Genetics Institute. We restricted analyses to 43,568 women with more than 97% European ancestry, as determined by means of an analysis of local ancestry.[15] Genotype

data were imputed against the reference haplotypes of phase 1 of the 1000 Genomes Project.[16]

We used linear regression to test single-marker genetic associations with gestational duration and logistic regression to test such associations with preterm birth on the basis of imputed allelic dosage (i.e., the expected allele count, as reported by the imputation program). We included as covariates the maternal age and the top five principal components to account for residual population structure (i.e., the difference in allele frequencies between subpopulations).

We clustered single-nucleotide polymorphisms (SNPs) into association regions (or loci) by first identifying SNPs with an association of $P<1.0\times10^{-4}$ and then grouping these SNPs into a region if they were adjacent to each other (<250 kb). The SNP with the smallest P value within each region was chosen as the index SNP. Regions with suggestive significance ($P<1.0\times10^{-6}$) were tested in the replication stage.

TESTS OF REPLICATION

To test for replication, we used data from 8643 mothers and 4090 infants collected from three Nordic birth studies,[17] in which samples from preterm births were enriched and samples from births that were post-term or close to the preterm–term boundary (37 to 38 weeks of gestation) were excluded (Table S1 and Fig. S2 in the Supplementary Appendix, available with the full text of this article at NEJM.org). Genotyping on samples from these studies was conducted with the use of various SNP arrays, as described previously.[17] Participants of non-European ancestry were identified and excluded with the use of principal components analysis. We performed genomewide imputation using the reference haplotypes extracted from phase 1 of the 1000 Genomes Project.[16]

We tested single-marker genetic association in each replication data set using methods similar to those used in the discovery stage. We used the fixed-effects inverse-variance method to calculate the replication P values after adjustment for the genomic inflation factor (which is used to quantify the excess false positive rate) in a combined analysis of the three Nordic data sets. Among the SNPs that had an association with genomewide suggestive significance ($P<1.0\times10^{-6}$) in the discovery stage or SNPs in close linkage disequilibrium ($r^2>0.80$), those that showed significant association (and in the same direction) in tests of replication were regarded as statistical evidence of replication of a putative locus. The significance level of each locus was corrected by the effective number of independent SNPs[18] that were tested in the locus and the total number of loci that were tested in the replication data sets (Table S6 in the Supplementary Appendix). An association was considered to be replicated if the P value of the most strongly associated SNP was less than the threshold of significance and had a combined discovery and replication P value of less than 5.0×10^{-8}.

We also performed association tests on samples obtained from the 4090 infants and joint maternal–fetal genetic association analysis in 3184 mother–infant pairs from the three replication sets that met the inclusion criteria (see the section describing the replication methods in the Supplementary Appendix) to evaluate whether the observed associations were driven by variants in the maternal genome or by variants in the fetal genome.

STATISTICAL ANALYSIS

We used the GWAS Catalog,[19] a database of all published genomewide association studies that is produced by a collaboration between the National Human Genome Research Institute and the European Bioinformatics Institute, to check whether the SNPs that were associated with gestational duration or preterm birth have been associated with other traits previously. In addition, we used the GTEx[20] database, which stores information on genotype and tissue-specific gene-expression levels, to determine whether any of the implicated SNPs could influence tissue-specific gene expression. We examined whether multiple independent variants at a given locus influenced birth timing by means of an approximate conditional and joint analysis.[21] We estimated the fraction of phenotype variance that was explained by all common SNPs[22] by means of genomewide complex trait analysis (GCTA)[23] or sets of significant SNPs using a genetic-score approach.[24] A detailed description of these analyses is provided in the section describing other statistical and bioinformatics analyses in the Supplementary Appendix.

RESULTS

STUDY DATA SETS

In the discovery data set, of the 43,568 women who had been identified through 23andMe, 37,803 (86.8%) had delivered at term (37 to 42 weeks),

3331 (7.6%) before term (<37 weeks), and 2434 (5.6%) after term (>42 weeks) (Tables S1 and S2 and Fig. S1 in the Supplementary Appendix). Maternal age was strongly associated with gestational duration ($P=2.3\times10^{-41}$) (Table S3 in the Supplementary Appendix). In the three Nordic birth studies,[17] the sex of the infant and maternal height were associated with gestational duration (Table S4 in the Supplementary Appendix).

DISCOVERY-STAGE FINDINGS IN MOTHERS

Single-marker association tests were performed across 15,635,593 SNPs (see the Methods section in the Supplementary Appendix). The summary statistical outcomes of the top 10,000 SNPs have been deposited in the GeneStation repository (www.genestation.org/analysis/gwas/Zhang_2017/discovery), and summary statistics of the complete data set are available on request from 23andMe. The corresponding author holds a copy of the summary statistical outcomes of the full set of SNPs. Test results were adjusted for genomic inflation factors (Fig. S3 in the Supplementary Appendix).

With respect to gestational duration, 12 loci were identified with an association of $P<1.0\times10^{-6}$, 4 of which had an association of $P<5.0\times10^{-8}$ (Fig. 1A, and Tables S5 and S6 in the Supplementary Appendix). With respect to preterm birth,

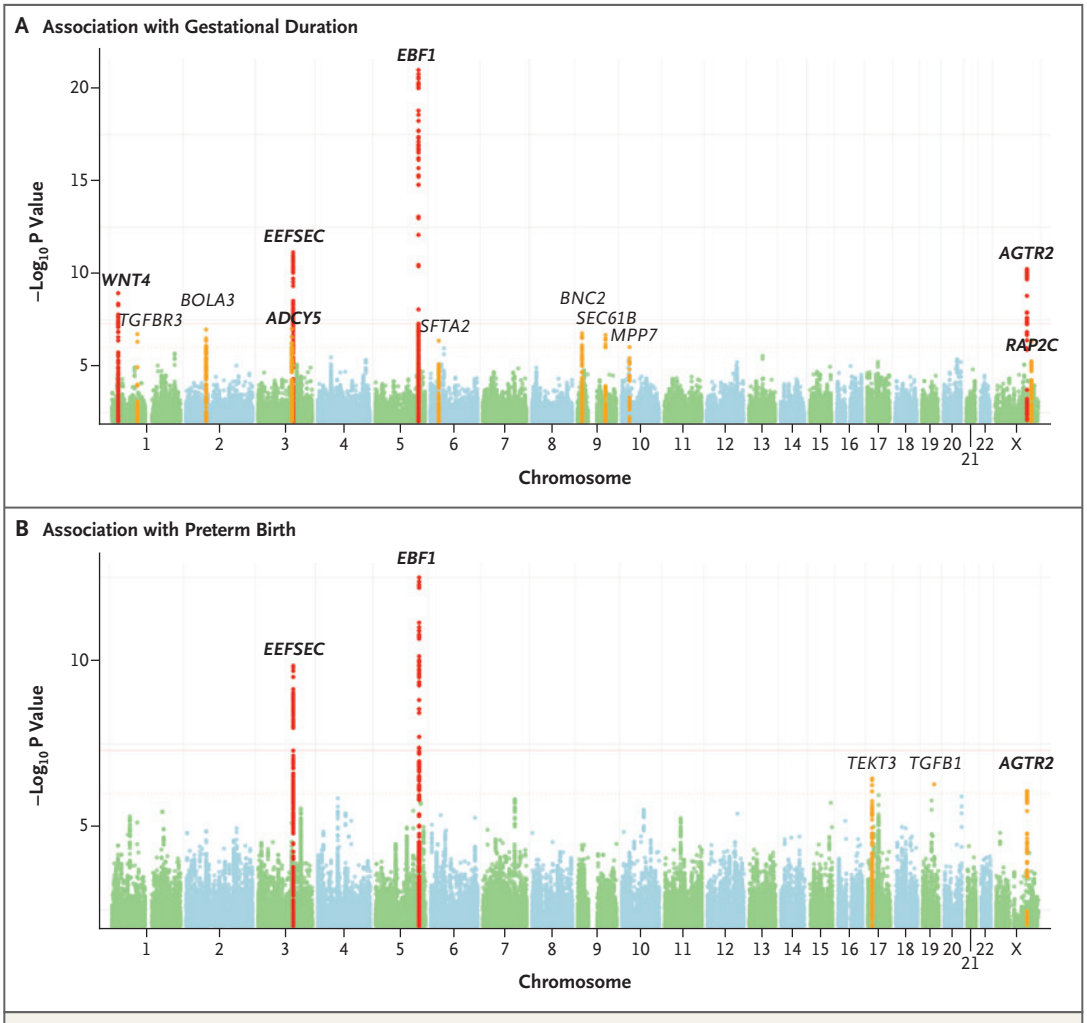

Figure 1. Results of the Discovery-Stage Genomewide Association Study.

Panel A shows the 12 loci that were associated with gestational duration with "suggestive significance" ($P<1.0\times10^{-6}$, in orange), 4 of which were associated with "genomewide significance" ($P<5.0\times10^{-8}$, in red). Panel B shows the 5 loci that had an association with preterm birth with suggestive significance (in orange), 2 of which had an association of genomewide significance (in red). The top 3 loci that were associated with gestational duration (*EBF1*, *EEFSEC*, and *AGTR2*) were also associated with preterm birth. The names of the 6 replicated loci are highlighted in bold.

Table 1. Replicated Loci Associated with Gestational Duration or Preterm Birth.*

Gene and SNP	Chromosome	Alleles (A/B)†	Discovery Stage				Replication Stage			Joint Analysis
			Frequency	Effect Size‡	P Value§	r^2¶	Frequency	Effect Size‡	P Value§	P Value§
Association with gestational duration										
EBF1	5									
rs2963463		C/T	0.272	−1.29	1.0×10^{-21}	—	0.264	−1.11	0.002	7.7×10^{-24}
rs2946171		T/G	0.219	−1.24	1.1×10^{-17}	0.71	0.206	−1.46	1.4×10^{-4}	8.1×10^{-21}
EEFSEC	3									
rs2955117		G/A	0.286	0.91	7.2×10^{-12}	—	0.279	1.33	1.6×10^{-4}	9.5×10^{-15}
rs200745338		D/I	0.237	0.99	1.5×10^{-11}	0.72	0.232	1.91	7.6×10^{-7}	7.5×10^{-16}
AGTR2	X									
rs201226733		I/D	0.422	−0.82	5.7×10^{-11}	—	0.420	−1.67	9.2×10^{-8}	7.2×10^{-16}
rs5950491		C/A	0.423	−0.83	6.8×10^{-11}	0.92	0.425	−1.75	4.7×10^{-8}	6.6×10^{-16}
WNT4	1									
rs56318008		C/T	0.139	1.05	1.2×10^{-9}	—	0.153	2.27	1.8×10^{-7}	3.4×10^{-14}
rs12037376		G/A	0.145	1.00	4.5×10^{-9}	0.91	0.157	2.41	2.1×10^{-8}	5.6×10^{-14}
ADCY5	3									
rs4383453		G/A	0.200	−0.81	9.6×10^{-8}	—	0.197	−0.59	0.15	3.7×10^{-8}
rs9861425		A/C	0.453	−0.60	6.1×10^{-7}	0.34	0.470	−1.38	9.5×10^{-6}	4.2×10^{-10}
RAP2C	X									
rs200879388		I/D	0.351	−0.66	4.5×10^{-7}	—	0.364	−1.10	9.2×10^{-4}	3.4×10^{-9}
Association with preterm birth										
EBF1	5									
rs2963463		C/T	0.272	1.23	3.2×10^{-13}	—	0.265	1.13	0.002	4.5×10^{-15}
rs2946169		C/T	0.217	1.22	1.1×10^{-10}	0.68	0.207	1.16	5.5×10^{-4}	2.2×10^{-13}
EEFSEC	3									
rs201450565		D/I	0.233	0.81	1.4×10^{-10}	—	0.135	0.82	0.002	1.9×10^{-12}
rs200745338		D/I	0.237	0.83	9.0×10^{-9}	0.24	0.232	0.80	3.5×10^{-7}	3.3×10^{-14}

AGTR2	X									
rs201386833		D/I	0.410	1.15	8.5×10^{-7}	—	0.41	1.18	2.3×10^{-6}	1.0×10^{-11}
rs5950506		G/A	0.420	1.14	1.2×10^{-6}	0.92	0.418	1.18	1.6×10^{-6}	1.1×10^{-11}

* For each locus, the single-nucleotide polymorphism (SNP) with the most significant association with either gestational duration or preterm birth in the discovery stage (index SNP) and the SNP with the most significant association in the replication stage are shown. Only SNPs with a P value of less than 1.0×10^{-6} in the discovery stage and their close proxies ($r^2>0.80$) were tested for replication. For each region, the protein-encoding gene closest to the index SNP is shown. More details regarding all the discovery and replication loci are provided in Tables S5, S8, and S9 in the Supplementary Appendix.

† Alleles were determined on the basis of the positive strand of the reference genome. Allele B is used as the reference allele for frequency and effect.

‡ For gestational duration, the effect is the regression coefficient, which shows the estimated changes in the number of gestational days per B allele. A positive value indicates longer gestational duration. For preterm birth, the effect is the estimated odds ratio of the reference B allele. Odds ratios of more than 1 indicate an increased risk of preterm birth (i.e., a reduced gestational duration).

§ In the discovery stage, P values were adjusted by means of genomic inflation factors. In the replication stage, P values were calculated from the inflation-adjusted effect sizes and standard error of the three Nordic studies with the use of a fixed-effects meta-analysis. In the joint analysis, P values were calculated from the 23andMe and combined Nordic studies with the use of the inverse variance method.

¶ The r^2 statistic (squared correlation coefficient) between the index SNP and the SNP with the most significant association in the replication stage was estimated from haplotype data obtained from samples in phase 1 of the 1000 Genomes Project.

5 loci were identified with an association of $P<1.0\times10^{-6}$, 2 of which had an association of $P<5.0\times10^{-8}$ (Fig. 1B). Of these 17 loci, the top 3 that were associated with gestational duration (*EBF1*, *EEFSEC*, and *AGTR2*) were also associated with preterm birth, so altogether, 14 independent loci were selected for replication. In similar association tests in a subgroup of discovery participants who explicitly checked "spontaneous delivery" in the questionnaire, the results were similar to those obtained from the full discovery data sets (Table S7 in the Supplementary Appendix).

TESTS OF REPLICATION

Among the candidate 14 loci, 6 were significantly replicated: *EBF1*, *EEFSEC*, and *AGTR2*, which were associated with both gestational duration and preterm birth, and *WNT4*, *ADCY5*, and *RAP2C*, which had significant association with gestational duration but not with preterm birth (Table 1, and Tables S8 and S9 in the Supplementary Appendix). In addition, showing marginal significance ($P<0.05$) were associations between the *BOLA3* locus with gestational duration and associations between the *TEKT3* and *TGFB1* loci with preterm birth.

ANNOTATION OF IMPLICATED SNPS

Within these replicated loci, there are SNPs that have been reported as being associated with other complex traits (Tables S6 and S10 in the Supplementary Appendix).[19] Three SNPs (rs10934853, rs2999052, and rs2687729) at the *EEFSEC* locus were associated with both gestational duration and preterm birth (Tables S8 and S9 in the Supplementary Appendix). The alleles that were associated with a longer duration of gestation have also been associated with an increased risk of prostate cancer (rs10934853-A),[25] a reduced risk of hypospadias (rs2999052-C),[26] and a later age of menarche (rs2687729-G).[27,28] At the *WNT4* locus, five significant SNPs in close linkage disequilibrium that were associated with gestational duration had been previously found to be associated with endometriosis,[29] ovarian cancer,[30] and bone mineral density.[31] The *WNT4* alleles that we observed to be associated with increased gestational duration have been previously identified as risk alleles for endometriosis, ovarian cancer, and low bone mineral density (Table S10 in the Supplementary Appendix). According to the GTEx database, some SNPs at four replicated loci (*EBF1*, *EEFSEC*,

WNT4, and *ADCY5*) can influence the messenger RNA expression level of nearby genes[20] (Tables S11 and S12 in the Supplementary Appendix).

SNPs at the *ADCY5* locus have been associated with birth weight[32] and blood glucose traits.[33] More recently, a meta-analysis has revealed SNPs at the *ADCY5*, *WNT4*, and *EBF1* loci that are associated with birth weight.[34] The SNPs that were implicated in this meta-analysis, at the *ADCY5* and *WNT4* loci, appear to influence birth weight through the fetal genome. None of these SNPs were in close linkage disequilibrium with the SNPs showing significant association with gestational duration, whereas the SNP at the *EBF1* locus (rs7729301) seems to influence birth weight through the maternal effect, and the allele (G) that was associated with reduced birth weight was associated with a shorter gestational duration (Tables S8 and S10 in the Supplementary Appendix).

MATERNAL OR FETAL GENETIC EFFECT?

We then tested for associations between the implicated SNPs and gestational duration and preterm birth in the infant samples (Tables S13 and S14 in the Supplementary Appendix). We observed the same associations as those in the maternal samples, and in the same direction, but with smaller effect sizes (Table S15 in the Supplementary Appendix). The effect sizes that were estimated from infant samples were highly correlated with the effect sizes that were estimated from the maternal samples ($r^2=0.95$) and were approximately half the size of those effects, a finding that supports the theory that the effect observed in infants is due to sharing of one maternal allele by descent (Fig. S4 in the Supplementary Appendix). In addition, joint association analysis in mother–infant pairs showed significant associations exclusively with maternal genotypes but not with fetal genotypes, which also indicates a maternal origin of the observed genetic associations (Table S16 in the Supplementary Appendix). The estimated phenotypic variances that were explained by all common SNPs in mothers (minor allele frequency, >0.01) were approximately 17% for gestational duration and 23% for preterm birth, as transformed to the heritability of the underlying disease liability (Table S18 in the Supplementary Appendix). Findings for detection of allelic heterogeneity, dominance effects, percentage of the variance explained, and gene set enrichment and pathway analyses are provided in the section describing other statistical and bioinformatics analyses in the Supplementary Appendix.

FUNCTIONAL EVIDENCE IMPLICATING *WNT4*

The *WNT4* locus implicates the endometrium as a determinant of preterm birth.[35] The function of *WNT4* is critical to decidualization of the endometrium and subsequent implantation and establishment of pregnancy.[36] Using RNA sequencing, we confirmed a substantial induction of *WNT4* expression with decidualization, since we detected no transcripts per million before decidualization in vitro and 29.5 transcripts per million after decidualization.

We used the catalogue of inferred sequence-binding preferences[37] to predict transcription factors, the binding of which would be altered by the implicated variants at the *WNT4* locus. We predicted that the thymidine (T) allele of rs3820282 ($r^2=0.94$ with the index SNP rs56318008) in the first intron of *WNT4* would alter the binding of estrogen receptor 1 (ESR1). This allele creates a near-perfect half-site for ESR1 (Fig. 2A). The derived enrichment score of the protein-binding microarrays, which evaluates the allele-specific binding strength, was 0.46 (indicating strong binding) for this allele, whereas that of the alternative allele, a cytosine (C) residue, was 0.09 (indicating no binding). This finding is consistent with the observation that ESR1 is capable of binding to this locus in a cellular context.[38] We confirmed the presence of H3K4me3 marks and an open chromatin domain overlapping rs3820282 in an immortalized endometrial stromal-cell line, which showed that the chromatin over this locus was probably accessible and active in these cells (Fig. 2A). (H3K4me3 is a histone modification that is found in active promoter regions, and an open chromatin domain indicates that the region is accessible to transcription factors.) Using an electrophoretic mobility shift assay, we detected enhanced binding of ESR1 to the T allele of rs3820282 (Fig. 2B). Collectively, these data indicate that the association between gestational duration and the *WNT4* locus is driven by the modulation of the binding of ESR1 through rs3820282.

DISCUSSION

We identified and replicated six maternal genomic loci that were robustly associated with gestational

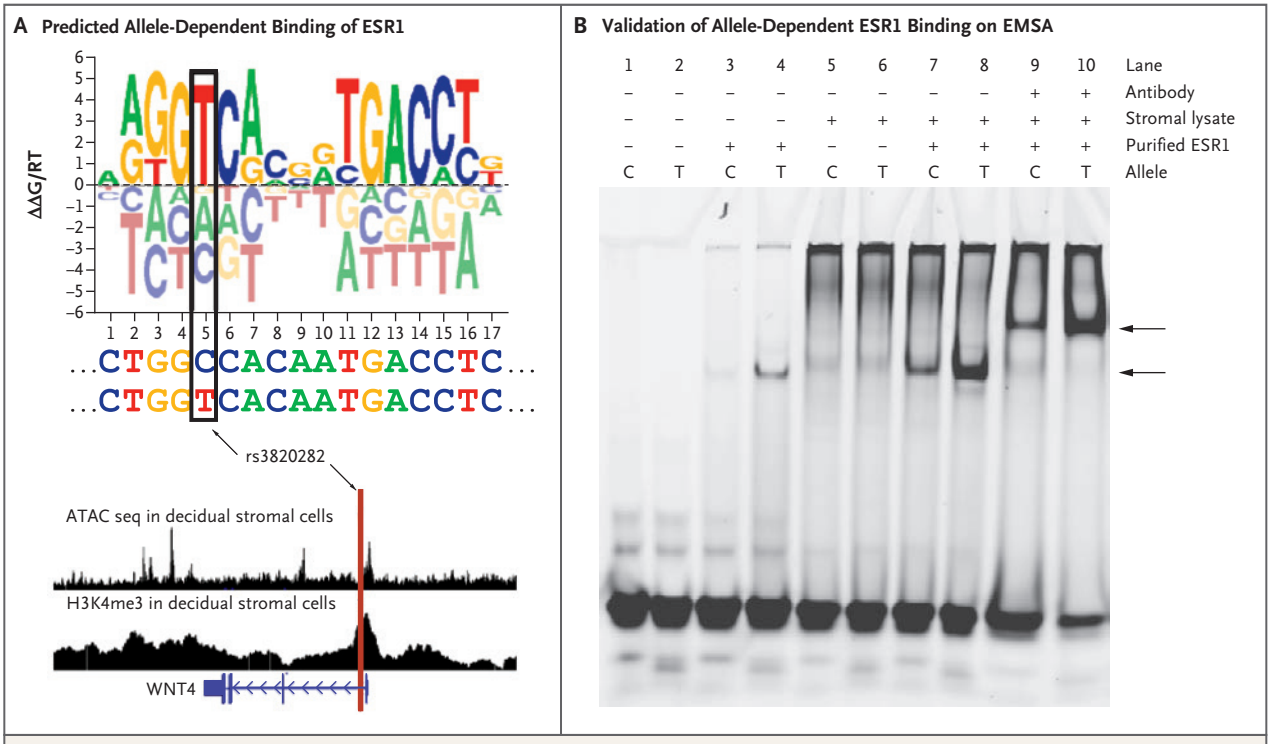

Figure 2. ESR1 Binding at the *WNT4* Locus.

Panel A shows how the rs3820282 T allele creates a strong binding site for estrogen receptor 1 (ESR1). The sequence logo of the ESR1 binding motif shows the DNA-binding preferences of ESR1. Tall nucleotides above the dashed line indicate DNA bases that are preferred by ESR1, whereas bases below the dashed line are disfavored. The y axis indicates the relative free energies of binding for each nucleotide at each position. The height of each nucleotide can be interpreted as the free energy difference from the average (ΔΔG) in units of gas constant (R) and temperature (T). The sequence located in the *WNT4* promoter is shown directly below the x axis, with the T allele for rs3820282 at the bottom. The T allele changes the sequence from C (most disfavored) to T (most preferred). In the UCSC Genome Browser screen shot depicted below the graph, the variant rs3820282 overlaps strong signals obtained from the assay for transposase-accessible chromatin with high-throughput sequencing (ATAC seq, a technique used in molecular biology to study chromatin accessibility) and H3K4me3 signals (a chemical mark present on histone H3 that is indicative of an active promoter) in decidual stromal cells at the *WNT4* locus. The red vertical line indicates the position of rs3820282. The purple graphic below the screen shot indicates the locations of the WNT4 exons (columns), untranslated regions (rectangles), and introns (horizontal lines), with the arrows indicating the direction of transcription. Panel B shows the experimental validation of allele-dependent binding of ESR1 to rs3820282 on electrophoretic mobility shift assay (EMSA), with the arrows indicating allele-dependent binding of ESR1 (bottom arrow) and a "supershift" of the protein-DNA complex induced by the binding of the ESR1 antibody to the complex (top arrow).

duration and that contained genes in which the established functions are consistent with a role in the timing of birth. Three of these loci were also associated with preterm birth with genome-wide significance.

EBF1, which encodes early B-cell factor 1, is essential for normal B-cell development,[39] and genomewide association studies have implicated it in the control of blood pressure,[40,41] carotid-artery intima–media thickness,[42] hypospadias,[26] and metabolic risk.[43] It remains to be determined whether *EBF1* confers its effect on birth timing through pregnancy-specific mechanisms or by contributing to more general cardiovascular or metabolic traits that influence gestation. In addition, the association between this locus and gestational duration may explain the effect of this locus on birth weight, as reported by Horikoshi et al.[34]

EEFSEC, which encodes selenocysteine tRNA-specific eukaryotic elongation factor, participates in the incorporation of selenocysteine into selenoproteins. Selenoproteins serve critical cellular homeostatic functions in maintaining redox status and antioxidant defenses, as well as modulating inflammatory responses.[44] These physiologic functions have been linked to the parturition pro-

cess and preterm birth.[45-47] Moreover, the SNPs we identified in *EEFSEC* are in high linkage disequilibrium with SNPs that have been associated with the risk of prostate cancer,[25] the risk of hypospadias,[26] and the age at menarche.[27] The identification of the selenocysteine pathway suggests a potential benefit for further evaluation of the role of maternal selenium micronutrient status on prematurity risk, as suggested by one study that showed an association between a reduced selenium concentration and preterm birth.[48] In addition, Malawi, the country with the highest global risk of preterm birth,[49] has a high prevalence of selenium deficiency.[50]

It has been suggested that *AGTR2*, which encodes angiotensin II receptor type 2, plays a role in modulating uteroplacental circulation and harbors variants that may contribute to the risk of preeclampsia.[51,52] It is unlikely that the association that we identified with *AGTR2* indicates a risk of preeclampsia rather than of spontaneous preterm birth, because women with preeclampsia as a reason for their delivery were excluded from the Nordic studies and women were excluded from the 23andMe discovery data set if medical indications for delivery were reported.

WNT4, which encodes wingless-type MMTV integration site family member 4, was strongly replicated in the Nordic populations. *WNT4* mutations have been found in women with müllerian duct abnormalities, primary amenorrhea, and hyperandrogenism,[53] and common variants in *WNT4*, which are in high linkage disequilibrium with our index SNPs, are associated with endometriosis,[29] ovarian cancer,[30] and bone mineral density.[31] Our analysis indicates that the T allele of the putative causative variant rs3820282 in the Nordic populations is associated with an increased gestational duration and is protective for preterm birth. The rs3820282 variant is located in an active chromatin domain in the first intron of *WNT4*, and the T allele generates a strong ESR1-binding site and as such probably alters the estrogen-based regulation of *WNT4* or adjacent genes. The role of augmented estrogen signaling as the functional consequence of the polymorphism is further supported by the association of the same region with endometriosis and ovarian cancer, both of which are hormone-responsive disorders. Finally, the population prevalence of endometriosis among women of Asian, European, or African ancestry corresponds to the frequencies of the T allele of rs3820282 (0.49, 0.14, and 0.01, respectively).[54,55]

ADCY5, which encodes adenylyl cyclase type 5, and *RAP2C*, which encodes a member of the RAS oncogene family, had associations of nearly genomewide significance in the discovery stage and were successfully replicated (Table 1). SNPs at the *ADCY5* locus have been reported to be associated with birth weight[32] and type 2 diabetes[33]; however, none were in linkage disequilibrium with the SNPs showing significant association with gestational duration. The SNP rs2747022 in the *RAP2C* region (in gene *FRMD7*) was previously reported to be associated with spontaneous preterm delivery in Danish and Norwegian studies.[56] (The samples used in this study overlap with our replication samples.)

Our study had limitations regarding the characteristics of the discovery data set, in which data regarding gestational duration in the 23andMe samples were self-reported. One study has shown that approximately 90% of the gestational durations that were reported by mothers agreed with the associated medical records.[57] Also, we could not distinguish spontaneous births from medically indicated births among women who carried their pregnancies to term. In addition, all the participants in our study were of European ancestry, in both the discovery and replication data sets. Thus, whether the same loci are involved in birth timing among women of other ancestries is uncertain.

Our study shows the utility of combining large samples that have self-reported phenotyping with more modestly sized but precisely phenotyped replication studies to reveal maternal loci associated with gestational duration and preterm birth. With this foundation, we anticipate that larger studies of samples with maternal and fetal genotyping associated with data regarding gestational duration will further refine our understanding of human pregnancy and the risk of adverse pregnancy outcomes.

Supported by grants (22-FY15-003, to Dr. Muglia; and 21-FY16-121) from the March of Dimes, grants from the National Institutes of Health, grants (to Drs. Jiang, Hu, Hinds, and Litterman and Ms. Russell) from the Cincinnati Children's Hospital Medical Center, a grant (to Dr. Muglia) from the Fifth Third Foundation, a grant (OPP1113966, to Dr. Muglia) from the Bill and Melinda Gates Foundation, a grant from the Jane and Aatos Erkko Foundation, grants (FUGE 183220/S10 and FRIMEDKLI-05 ES236011) from the Norwegian Research Council, a grant from

the Jane and Dan Olsson Foundations, a grant (ALFGBG-507701) from the Swedish government to researchers in the public health service, and a grant (FP7/2007-2013) from the European Community's Seventh Framework Program (grant agreement HEALTH-F4-2007-201413). Dr. Wagner is supported by a grant (54860) from the John Templeton Foundation; Dr. Feenstra, by an Oak Foundation fellowship; Dr. Liu, by the Nordic Center of Excellence in Health-Related e-Sciences; the Norwegian Mother and Child Cohort Study, 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 (UO1 NS 047537-01 and UO1 NS 047537-06A1), by the Norwegian Research Council/FUGE (151918/S10 and FRI-MEDBIO 249779), and by the Swedish Research Council (2015-02559); Dr. Palotie, by the Academy of Finland; Dr. Ryckman, by the March of Dimes, Bill and Melinda Gates Foundation, and National Institutes of Health; Dr. Chavan and Ms. Maziarz, by the John Templeton Foundation; and Dr. Boyd, by the National Institutes of Health Genes, Environment, and Health Initiative and the Danish Council for Independent Research, Medical Sciences. Support from the Functional Genomics Core at Cincinnati Children's Hospital Medical Center was made possible through a grant (P30 AR070549) from the National Institute of Arthritis and Musculoskeletal and Skin Diseases.

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

We thank the participants in the Finnish birth cohort, the Mother Child Cohort of Norway, and the Danish National Birth Cohort (a part of the Danish National Biobank resource, which is supported by the Novo Nordisk Foundation); dbGAP for depositing and hosting the phenotype and genotype data; the research participants and employees of 23andMe; Hugh Taylor of Yale Medical School for providing biopsy samples of endometrial stromal fibroblasts for RNA sequencing; and Gil More, also of Yale Medical School, for providing immortalized human endometrial stromal fibroblasts for chromatin immunoprecipitation sequencing.

APPENDIX

The authors' full names and academic degrees are as follows: Ge Zhang, M.D., Ph.D., Bjarke Feenstra, Ph.D., Jonas Bacelis, B.S., Xueping Liu, Ph.D., Lisa M. Muglia, Ph.D., Julius Juodakis, B.S., Daniel E. Miller, B.S., Nadia Litterman, Ph.D., Pan-Pan Jiang, Ph.D., Laura Russell, M.S., David A. Hinds, Ph.D., Youna Hu, Ph.D., Matthew T. Weirauch, Ph.D., Xiaoting Chen, Ph.D., Arun R. Chavan, M.Sci., Günter P. Wagner, Ph.D., Mihaela Pavličev, Ph.D., Mauris C. Nnamani, Ph.D., Jamie Maziarz, M.Sc., Minna K. Karjalainen, Ph.D., Mika Rämet, M.D., Ph.D., Verena Sengpiel, M.D., Ph.D., Frank Geller, M.Sc., Heather A. Boyd, Ph.D., Aarno Palotie, M.D., Ph.D., Allison Momany, B.S., Bruce Bedell, M.A., Kelli K. Ryckman, Ph.D., Johanna M. Huusko, Ph.D., Carmy R. Forney, B.S., Leah C. Kottyan, Ph.D., Mikko Hallman, M.D., Ph.D., Kari Teramo, M.D., Ph.D., Ellen A. Nohr, Ph.D., George Davey Smith, D.Sc., Mads Melbye, M.D., D.M.Sc., Bo Jacobsson, M.D., Ph.D., and Louis J. Muglia, M.D., Ph.D.

The authors' affiliations are as follows: the Division of Human Genetics (G.Z., L.J.M.), Center for Autoimmune Genomics and Etiology (M.T.W., D.E.M., X.C., C.R.F., L.C.K.) and the Divisions of Biomedical Informatics and Developmental Biology (M.T.W.), Cincinnati Children's Hospital Medical Center, and the Center for Prevention of Preterm Birth, Perinatal Institute, Cincinnati Children's Hospital Medical Center and March of Dimes Prematurity Research Center Ohio Collaborative (G.Z., L.M.M., M.P., J.M.H., L.J.M.), Cincinnati; the Department of Epidemiology Research, Statens Serum Institut (B.F., X.L., F.G., H.A.B., M.M.), and the Department of Clinical Medicine, University of Copenhagen (M.M.), Copenhagen, and the Research Unit of Gynecology and Obstetrics, Institute of Clinical Research, University of Southern Denmark, Odense (E.A.N.) — all in Denmark; the Department of Obstetrics and Gynecology, Sahlgrenska University Hospital Östra (J.B., V.S.), the Department of Obstetrics and Gynecology, Institute of Clinical Sciences (J.J.), and the Department of Obstetrics and Gynecology (B.J.), Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden; 23andMe, Mountain View (N.L., P.-P.J., L.R., D.A.H., Y.H.), and the Department of Medicine, Stanford University School of Medicine, Stanford (M.M.) — both in California; the Department of Ecology and Evolutionary Biology, Yale University (A.R.C., G.P.W., M.C.N., J.M.), and the Department of Obstetrics, Gynecology, and Reproductive Sciences, Yale Medical School (G.P.W.), New Haven, and the Yale Systems Biology Institute, West Haven (A.R.C., G.P.W., M.C.N., J.M.) — all in Connecticut; the Department of Obstetrics and Gynecology, Wayne State University, Detroit (G.P.W.); the PEDEGO Research Unit and Medical Research Center Oulu, University of Oulu, and the Department of Children and Adolescents, Oulu University Hospital, Oulu (M.K.K., M.R., J.M.H., M.H.), and the Institute for Molecular Medicine Finland, University of Helsinki (A.P.), and Obstetrics and Gynecology, University of Helsinki and Helsinki University Hospital (K.T.), Helsinki — all in Finland; the Analytic and Translational Genetics Unit, Department of Medicine, the Psychiatric and Neurodevelopmental Genetics Unit, Department of Psychiatry, and the Department of Neurology, Massachusetts General Hospital, Boston (A.P.), and the Program in Medical and Population Genetics and the Stanley Center for Psychiatric Research, Broad Institute of the Massachusetts Institute of Technology and Harvard, Cambridge (A.P.) — both in Massachusetts; the Departments of Pediatrics (A.M., B.B.) and Epidemiology (K.K.R.), College of Public Health, and the Department of Pediatrics (K.K.R.), Carver College of Medicine, University of Iowa, Iowa City; the Medical Research Council Integrative Epidemiology Unit at the University of Bristol, School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom (G.D.S.); and the Department of Genetics and Bioinformatics, Area of Health Data and Digitalization, Norwegian Institute of Public Health, Oslo (B.J.).

REFERENCES

1. Martin JA, Hamilton BE, Osterman MJ. Births in the United States, 2014. NCHS Data Brief 2015;216:1-8.
2. Yoshida S, Martines J, Lawn JE, et al. Setting research priorities to improve global newborn health and prevent stillbirths by 2025. J Glob Health 2016;6(1):010508.
3. Liu L, Oza S, Hogan D, et al. Global, regional, and national causes of child mortality in 2000-13, with projections to inform post-2015 priorities: an updated systematic analysis. Lancet 2015;385:430-40.
4. Butler AS, Behrman RE. Preterm birth: causes, consequences, and prevention. Washington, DC: National Academies Press, 2007.
5. Bezold KY, Karjalainen MK, Hallman M, Teramo K, Muglia LJ. The genomics of preterm birth: from animal models to human studies. Genome Med 2013;5:34.
6. Clausson B, Lichtenstein P, Cnattingius S. Genetic influence on birthweight and gestational length determined by studies in offspring of twins. BJOG 2000;107:375-81.
7. York TP, Eaves LJ, Lichtenstein P, et al. Fetal and maternal genes' influence on gestational age in a quantitative genetic

analysis of 244,000 Swedish births. Am J Epidemiol 2013;178:543-50.
8. Plunkett J, Feitosa MF, Trusgnich M, et al. Mother's genome or maternally-inherited genes acting in the fetus influence gestational age in familial preterm birth. Hum Hered 2009;68:209-19.
9. Kistka ZA, DeFranco EA, Ligthart L, et al. Heritability of parturition timing: an extended twin design analysis. Am J Obstet Gynecol 2008;199(1):43.e1-5.
10. Boyd HA, Poulsen G, Wohlfahrt J, Murray JC, Feenstra B, Melbye M. Maternal contributions to preterm delivery. Am J Epidemiol 2009;170:1358-64.
11. Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ 2006;332:1080.
12. Monangi NK, Brockway HM, House M, Zhang G, Muglia LJ. The genetics of preterm birth: progress and promise. Semin Perinatol 2015;39:574-83.
13. Wu W, Clark EAS, Manuck TA, Esplin MS, Varner MW, Jorde LB. A genome-wide association study of spontaneous preterm birth in a European population. F1000Research2013;2:255.
14. Zhang H, Baldwin DA, Bukowski RK, et al. A genome-wide association study of early spontaneous preterm delivery. Genet Epidemiol 2015;39:217-26.
15. Durand EY, Do CB, Mountain JL, Macpherson JM. Ancestry composition: a novel, efficient pipeline for ancestry deconvolution. bioRxiv. October 18, 2014.
16. Abecasis GR, Auton A, Brooks LD, et al. An integrated map of genetic variation from 1,092 human genomes. Nature 2012;491:56-65.
17. Zhang G, Bacelis J, Lengyel C, et al. Assessing the causal relationship of maternal height on birth size and gestational age at birth: a Mendelian randomization analysis. PLoS Med 2015;12(8):e1001865.
18. Gao X, Starmer J, Martin ER. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. Genet Epidemiol 2008;32:361-9.
19. Welter D, MacArthur J, Morales J, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 2014;42:D1001-D1006.
20. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet 2013;45:580-5.
21. Yang J, Ferreira T, Morris AP, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet 2012;44:369-75.
22. Yang J, Benyamin B, McEvoy BP, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet 2010;42:565-9.
23. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 2011;88:76-82.
24. Zhang G, Karns R, Sun G, et al. Extent of height variability explained by known height-associated genetic variants in an isolated population of the Adriatic coast of Croatia. PLoS One 2011;6(12):e29475.
25. Gudmundsson J, Sulem P, Gudbjartsson DF, et al. Genome-wide association and replication studies identify four variants associated with prostate cancer susceptibility. Nat Genet 2009;41:1122-6.
26. Geller F, Feenstra B, Carstensen L, et al. Genome-wide association analyses identify variants in developmental genes associated with hypospadias. Nat Genet 2014;46:957-63.
27. Perry JR, Day F, Elks CE, et al. Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. Nature 2014;514:92-7.
28. Elks CE, Perry JR, Sulem P, et al. Thirty new loci for age at menarche identified by a meta-analysis of genome-wide association studies. Nat Genet 2010;42:1077-85.
29. Albertsen HM, Chettier R, Farrington P, Ward K. Genome-wide association study link novel loci to endometriosis. PLoS One 2013;8(3):e58257.
30. Kuchenbaecker KB, Ramus SJ, Tyrer J, et al. Identification of six new susceptibility loci for invasive epithelial ovarian cancer. Nat Genet 2015;47:164-71.
31. Kemp JP, Medina-Gomez C, Estrada K, et al. Phenotypic dissection of bone mineral density reveals skeletal site specificity and facilitates the identification of novel loci in the genetic regulation of bone mass attainment. PLoS Genet 2014;10(6):e1004423.
32. Freathy RM, Mook-Kanamori DO, Sovio U, et al. Variants in ADCY5 and near CCNL1 are associated with fetal growth and birth weight. Nat Genet 2010;42:430-5.
33. Dupuis J, Langenberg C, Prokopenko I, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet 2010;42:105-16.
34. Horikoshi M, Beaumont RN, Day FR, et al. Genome-wide associations for birth weight and correlations with adult disease. Nature 2016;538:248-52.
35. Sonderegger S, Pollheimer J, Knöfler M. Wnt signalling in implantation, decidualisation and placental differentiation — review. Placenta 2010;31:839-47.
36. Li Q, Kannan A, Das A, et al. WNT4 acts downstream of BMP2 and functions via β-catenin signaling pathway to regulate human endometrial stromal cell differentiation. Endocrinology 2013;154:446-57.
37. Weirauch MT, Yang A, Albu M, et al. Determination and inference of eukaryotic transcription factor sequence specificity. Cell 2014;158:1431-43.
38. Griffon A, Barbier Q, Dalino J, van Helden J, Spicuglia S, Ballester B. Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. Nucleic Acids Res 2015;43(4):e27.
39. György I, Boller S, Nechanitzky R, et al. Transcription factor Ebf1 regulates differentiation stage-specific signaling, proliferation, and survival of B cells. Genes Dev 2012;26:668-82.
40. Ehret GB, Munroe PB, Rice KM, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature 2011;478:103-9.
41. Wain LV, Verwoert GC, O'Reilly PF, et al. Genome-wide association study identifies six new loci influencing pulse pressure and mean arterial pressure. Nat Genet 2011;43:1005-11.
42. Xie G, Myint PK, Voora D, et al. Genome-wide association study on progression of carotid artery intima media thickness over 10 years in a Chinese cohort. Atherosclerosis 2015;243:30-7.
43. Singh A, Babyak MA, Nolan DK, et al. Gene by stress genome-wide interaction analysis and path analysis identify EBF1 as a cardiovascular and metabolic risk gene. Eur J Hum Genet 2015;23:854-62.
44. Labunskyy VM, Hatfield DL, Gladyshev VN. Selenoproteins: molecular pathways and physiological roles. Physiol Rev 2014;94:739-77.
45. Burnum KE, Hirota Y, Baker ES, et al. Uterine deletion of Trp53 compromises antioxidant responses in the mouse decidua. Endocrinology 2012;153:4568-79.
46. Goldenberg RL, Culhane JF, Iams JD, Romero R. Epidemiology and causes of preterm birth. Lancet 2008;371:75-84.
47. Muglia LJ, Katz M. The enigma of spontaneous preterm birth. N Engl J Med 2010;362:529-35.
48. Rayman MP, Wijnen H, Vader H, Kooistra L, Pop V. Maternal selenium status during early gestation and risk for preterm birth. CMAJ 2011;183:549-55.
49. Born too soon: the global action report on preterm birth. Geneva: World Health Organization, 2012.
50. Hurst R, Siyame EW, Young SD, et al. Soil-type influences human selenium status and underlies widespread selenium deficiency risks in Malawi. Sci Rep 2013;3:1425.
51. Akbar SA, Khawaja NP, Brown PR, Tayyeb R, Bamfo J, Nicolaides KH. Angiotensin II type 1 and 2 receptors gene polymorphisms in pre-eclampsia and normal pregnancy in three different populations. Acta Obstet Gynecol Scand 2009;88:606-11.
52. Zhou A, Dekker GA, Lumbers ER, et al. The association of AGTR2 polymor-

phisms with preeclampsia and uterine artery bilateral notching is modulated by maternal BMI. Placenta 2013;34:75-81.
53. Philibert P, Biason-Lauber A, Gueorguieva I, et al. Molecular analysis of WNT4 gene in four adolescent girls with mullerian duct abnormality and hyperandrogenism (atypical Mayer-Rokitansky-Küster-Hauser syndrome). Fertil Steril 2011;95:2683-6.
54. Missmer SA, Hankinson SE, Spiegelman D, Barbieri RL, Marshall LM, Hunter DJ. Incidence of laparoscopically confirmed endometriosis by demographic, anthropometric, and lifestyle factors. Am J Epidemiol 2004;160:784-96.
55. Mangtani P, Booth M. Epidemiology of endometriosis. J Epidemiol Community Health 1993;47:84-8.
56. Myking S, Boyd HA, Myhre R, 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 2013; 8(4):e61781.
57. Little RE. Birthweight and gestational age: mothers' estimates compared with state and hospital records. Am J Public Health 1986;76:1350-1.