

Variants in the fetal genome near pro-inflammatory cytokine genes on 2q13 are associated with gestational duration

Xueping Liu^{1,a}, Dorte Helenius^{2,3,a}, Line Skotte^{1,a}, Robin N. Beaumont⁴, Matthias Wielscher⁵, Frank Geller¹, Julius Juodakis⁶, Anubha Mahajan⁷, Jonathan P. Bradfield^{8,9}, Frederick T.J. Lin¹⁰, Suzanne Voegelzang^{11,12,13}, Mariona Bustamante^{14,15,16}, Tarunveer S. Ahluwalia¹⁷, Niina Pitkänen¹⁸, Carol A. Wang¹⁹, Jonas Bacelis²⁰, Maria C. Borges^{21,22}, Ge Zhang^{23,24,25}, Bruce A. Bedell²⁶, Robert M. Rossi^{25,27}, Kristin Skogstrand^{2,28}, Shouneng Peng^{29,30}, Early Growth Genetics (EGG) Consortium^b, iPSYCH-BROAD Working Group^c, Wesley K. Thompson^{2,3}, Vivek Appadurai^{2,3}, Debbie A. Lawlor^{21,22,31}, Ilkka Kalliala^{32,33}, Christine Power³⁴, Mark I. McCarthy^{7,35,36}, Heather A. Boyd¹, Mary L. Marazita^{37,38}, Hakon Hakonarson^{8,39,40}, M. Geoffrey Hayes^{10,41,42}, Denise M. Scholtens⁴³, Fernando Rivadeneira^{11,12,44}, Vincent W.V. Jaddoe^{11,12,13}, Rebecca K. Vinding¹⁷, Hans Bisgaard¹⁷, Bridget A. Knight⁴⁵, Katja Pakkala^{18,46}, Olli Raitakari^{18,47}, Øyvind Helgeland^{48,49,50}, Stefan Johansson^{48,51}, Pål R. Njølstad^{48,49}, João Fadista¹, Andrew Schork^{2,3}, Ron Nudel^{2,3}, Daniel E. Miller⁵², Xiaoting Chen⁵², Matthew T. Weirauch^{27,52,53}, Preben Bo Mortensen^{2,54,55}, Anders D. Børghlum^{2,55,56}, Merete Nordentoft^{2,57,58}, Ole Mors^{2,59}, Ke Hao^{29,30}, Kelli K. Ryckman^{26,60}, David M. Hougaard^{2,28}, Leah C. Kottyan^{27,52}, Craig E. Pennell¹⁹, Leo-Pekka Lyytikäinen^{61,62}, Klaus Bønnelykke¹⁷, Martine Vrijheid^{14,15,16}, Janine F. Felix^{11,12,13}, William L. Lowe, Jr.¹⁰, Struan F.A. Grant^{8,39,40}, Elina Hyppönen^{34,63,64}, Bo Jacobsson^{6,65}, Marjo-Riitta Jarvelin^{66,67}, Louis J. Muglia^{23,24,25,27}, Jeffrey C. Murray²⁶, Rachel M. Freathy^{4,68}, Thomas M. Werge^{2,3,58}, Mads Melbye^{1,69}, Alfonso Buil^{2,3,d}, Bjarke Feenstra^{1,d}

^a These authors contributed equally to this work.

^b A full list of the members of the consortium is provided in the Supplementary Material.

^c A list of collaborators in the working group is provided in the Supplementary Material.

^d These authors jointly directed this work.

Correspondence should be addressed to B.F. (fee@ssi.dk).

1. Department of Epidemiology Research, Statens Serum Institut, Copenhagen, Denmark.
2. iPSYCH, The Lundbeck Foundation Initiative for Integrative Psychiatric Research, Aarhus, Denmark.
3. Institute of Biological Psychiatry, Mental Health Center Sct. Hans, Mental Health Services Copenhagen, Roskilde, Denmark.
4. Institute of Biomedical and Clinical Science, University of Exeter Medical School, University of Exeter, Royal Devon and Exeter Hospital, Barrack Road, Exeter, EX2 5DW, UK.

5. Department of Epidemiology and Biostatistics, MRC-PHE Centre for Environment and Health, School of Public Health, Imperial College London, London, UK.
6. Department of Obstetrics and Gynecology, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden.
7. Wellcome Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK.
8. Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA.
9. Quantinuum Research, LLC, San Diego, CA, USA.
10. Division of Endocrinology, Metabolism and Molecular Medicine, Department of Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA.
11. The Generation R Study Group, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands.
12. Department of Epidemiology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands.
13. Department of Pediatrics, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands.
14. ISGlobal, Barcelona Institute for Global Health, Barcelona, Spain.
15. Universitat Pompeu Fabra (UPF), Barcelona, Spain.
16. CIBER Epidemiología y Salud Pública (CIBERESP), Spain.
17. COPSAC, Copenhagen Prospective Studies on Asthma in Childhood, Herlev and Gentofte Hospital, University of Copenhagen, Copenhagen, Denmark.
18. Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku, 20520 Turku, Finland.
19. School of Medicine and Public Health, Faculty of Medicine and Health, The University of Newcastle, Newcastle, New South Wales, Australia.
20. Department of Obstetrics and Gynecology, Sahlgrenska University Hospital, Gothenburg, Sweden.
21. MRC Integrative Epidemiology Unit at the University of Bristol, Bristol, UK.
22. Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK.
23. Division of Human Genetics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA.

24. Center for Prevention of Preterm Birth, Perinatal Institute, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA.
25. March of Dimes Prematurity Research Center Ohio Collaborative, Cincinnati, OH, USA.
26. Department of Pediatrics, University of Iowa, Iowa City, IA, USA.
27. Department of Pediatrics, University of Cincinnati College of Medicine. Cincinnati, OH, USA.
28. Statens Serum Institut, Center for Neonatal Screening, Department for Congenital Disorders, Copenhagen, Denmark.
29. Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, 1425 Madison Avenue, New York, NY 10029, USA.
30. Icahn Institute of Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, 1425 Madison Avenue, New York, NY 10029, USA.
31. NIHR Bristol Biomedical Research Centre, Bristol, UK.
32. Department of Surgery and Cancer, IRDB, Faculty of Medicine, Imperial College, London W12 0NN, UK.
33. Department of Obstetrics and Gynaecology, University of Helsinki and Helsinki University Hospital, Haartmaninkatu 2, 00029 HUS, Finland.
34. Population, Policy and Practice, Great Ormond Street Institute for Child Health, University College London, London, UK.
35. Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford OX3 7LJ, UK.
36. NIHR Oxford Biomedical Research Centre, Churchill Hospital, Oxford, OX3 7LJ, UK.
37. Center for Craniofacial and Dental Genetics, Department of Oral Biology School of Dental Medicine University of Pittsburgh, Pittsburgh, PA, USA.
38. Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA.
39. Division of Human Genetics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA.
40. Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA.
41. Department of Anthropology, Northwestern University, Evanston, IL, USA.
42. Center for Genetic Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA.

43. Division of Biostatistics, Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA.
44. Department of Internal Medicine, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands.
45. NIHR Exeter Clinical Research Facility, University of Exeter Medical School, University of Exeter, Royal Devon and Exeter Hospital, Barrack Road, Exeter, EX2 5DW, UK.
46. Paavo Nurmi Centre, Sports & Exercise Medicine Unit, Department of Health and Physical Activity, University of Turku, 20520 Turku, Finland.
47. Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital, 20521 Turku, Finland.
48. K. G. Jebsen Center for Diabetes Research, Department of Clinical Science, University of Bergen, Bergen, Norway.
49. Department of Pediatrics, Haukeland University Hospital, Bergen, Norway.
50. Norwegian Institute of Public Health, Division of Health data and Digitalization, Department of Genetic Research and Bioinformatics, Oslo, Norway.
51. Center for Medical Genetics and Molecular Medicine, Haukeland University Hospital, Bergen, Norway.
52. Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA.
53. Divisions of Biomedical Informatics and Developmental Biology, Cincinnati Children's Hospital, Cincinnati, Ohio, USA.
54. National Centre for Register-Based Research, Aarhus University, Aarhus, Denmark.
55. iSEQ, Centre for Integrative Sequencing, Aarhus University, Aarhus, Denmark.
56. Department of Biomedicine - Human Genetics, Aarhus University, Aarhus, Denmark.
57. Mental Health Center Copenhagen, Mental Health Services in the Capital Region of Denmark, Copenhagen, Denmark.
58. Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.
59. Psychosis Research Unit, Aarhus University Hospital, Risskov, Denmark.
60. Department of Epidemiology, University of Iowa, Iowa City, IA, USA.

61. Department of Clinical Chemistry, Fimlab Laboratories, Tampere 33520, Finland.
62. Department of Clinical Chemistry, Finnish Cardiovascular Research Center - Tampere, Faculty of Medicine and Life Sciences, University of Tampere, Tampere 33014, Finland.
63. Australian Centre for Precision Health, University of South Australia Cancer Research Institute, Adelaide, Australia.
64. South Australian Health and Medical Research Institute, Adelaide, Australia.
65. Department of Genes and Environment, Division of Epidemiology, Norwegian Institute of Public Health, Oslo, Norway.
66. Institute of Health Sciences, University of Oulu, Oulu, Finland.
67. Department of Epidemiology and Biostatistics, School of Public Health, Medical Research Council-Health Protection Agency Centre for Environment and Health, Faculty of Medicine, Imperial College London, UK.
68. Medical Research Council Integrative Epidemiology Unit at the University of Bristol, Oakfield House, Oakfield Grove Bristol, BS8 2BN, UK.
69. Department of Medicine, Stanford University School of Medicine, Stanford, California, USA.

Abstract

The duration of pregnancy is influenced by fetal and maternal genetic and non-genetic factors. We conducted a fetal genome-wide association meta-analysis of gestational duration, and early preterm, preterm, and postterm birth in 84,689 infants. One locus on chromosome 2q13 was associated with gestational duration; the association was replicated in 9,291 additional infants (combined $P = 3.96 \times 10^{-14}$). Analysis of 15,536 mother-child pairs showed that the association was driven by fetal rather than maternal genotype. Functional experiments showed that the lead SNP, rs7594852, alters the binding of the HIC1 transcriptional repressor. Genes at the locus include several interleukin 1 family members with roles in pro-inflammatory pathways that are central to the process of parturition. Further understanding of the underlying mechanisms will be of great public health importance, since giving birth either before or after the window of term gestation is associated with increased morbidity and mortality.

Keywords

Gestational duration, preterm birth, postterm birth, genome-wide association study, pro-inflammatory cytokines, interleukin 1

Introduction

Pregnancy in eutherian mammals is characterized by tightly regulated physiological processes to ensure normal fetal development and delivery after a narrowly defined period of gestation^{1,2}. A conundrum first posed by Sir Peter Medawar more than 60 years ago is how the semi-allogeneic fetus is protected from attack by the mother's immune system³. Compared to many other mammals, humans have a highly invasive placentation process with direct contact with the maternal circulation^{1,4}, and the immunological paradox of pregnancy continues to be an important research topic. It is well-established that successful gestation depends on numerous mechanisms, some of which involve inflammatory pathways⁵. After conception, an inflammatory phase ensures implantation of the blastocyst in the uterine wall⁶. This is followed by a long anti-inflammatory phase in which the maternal adaptive immune response is dampened to allow the development, growth and maturation of the fetus. Eventually, a second inflammatory phase results in gradual ripening of gestational tissues, followed by parturition⁶. Many other pathways are dynamically regulated over the course of a pregnancy and are required for the successful completion of pregnancy and timely parturition⁷.

Correct timing of parturition is critical for the health of the newborn. Preterm birth, defined as birth before 37 completed weeks of gestation, is not only a major cause of perinatal mortality and morbidity⁸, but is also associated with long-term adverse health outcomes including cerebral palsy⁹, diabetes¹⁰, increased blood pressure¹¹, and various psychiatric disorders¹². Postterm birth, defined as birth after a gestation of 42 completed weeks (hereafter weeks) or more is associated with increased risks of fetal and neonatal mortality

and morbidity plus increased maternal morbidity¹³. Each of these outcomes affects approximately 5% to 10% of all births in high income countries^{14,15} and preterm birth rates are considerably higher in some low and middle income countries¹⁶.

Although timing of parturition is influenced by many non-genetic risk factors, including parity, maternal stress, smoking, urogenital infection, educational attainment and socioeconomic status, there is compelling evidence for a substantial genetic impact^{17,18}. For example, twin and family studies have estimated that the heritability of gestational duration ranges from 25% to 40%¹⁹. Several studies have shown that the duration of pregnancy has both heritable maternal and fetal components^{20,21}. Estimates from a Swedish family study analysing 244,000 births indicated that fetal genetic factors explained about 10% of the variation in gestational duration, whereas maternal factors accounted for about 20%²¹.

Little is known about specific fetal and maternal genetic contributions to gestational duration. Most genome-wide association studies (GWAS) of birth timing have been limited in sample size and have not identified robustly associated genetic loci^{22–26}. Recently, however, a GWAS based on samples from 43,568 women of European ancestry identified *maternal* genetic variants at six loci associated with gestational duration at $P < 5 \times 10^{-8}$ with replication in three independent data sets²⁷. Three of these loci were also associated with preterm birth as a dichotomous outcome.

In the current study, our goal was to identify *fetal* genetic variants associated with timing of parturition. We conducted a GWAS meta-analysis of gestational duration as a quantitative trait and early preterm (<34 weeks), preterm (<37 weeks), and postterm (≥ 42 weeks) birth as dichotomous outcomes, in 84,689 infants from cohorts included in the Early Growth Genetics (EGG) consortium, the Initiative for Integrative Psychiatric Research (iPSYCH), and the Genomic and Proteomic Network for Preterm Birth Research (GPN), with replication analyses in 9,291 infants from additional cohorts. Since a child inherits half of its genetic material from its mother, their genotypes at a given locus are highly correlated, and it may not be clear whether a genetic association reflects the effect of the child's own genotype on the timing of their delivery, or an effect of their mother's genotype on the timing of parturition. For 15,536 of the infants, maternal genetic data were also available, allowing us to address whether identified genetic effects were of fetal or maternal origin.

Results

Discovery stage

Characteristics of the 20 studies included in the discovery stage are presented in **Supplementary Table 1**. The discovery data set included information on 84,689 infants, 4,775 of whom were born preterm (<37 weeks), with 1,139 of these considered early preterm infants (<34 weeks). A further 60,148 infants were born ≥ 39 weeks and <42 weeks of gestation and were used as term controls. Finally, 7,933 infants were born postterm (≥ 42 weeks). Our study design is illustrated in **Supplementary Fig. 1**. After imputation using reference data from the Haplotype Reference Consortium release 1.1²⁸ or the integrated phase III release of the 1000

Genomes Project²⁹, each contributing study performed GWAS analyses for at least one of the four study traits, assuming an additive genetic model (see Methods for details). Final meta-analysis results were obtained for >7.5 million SNPs for each of gestational duration, early preterm birth, preterm birth, and postterm birth, with genomic inflation factors <1.05. Quantile-quantile and Manhattan plots for the four phenotypes are shown in **Supplementary Fig. 2**. In the discovery GWAS meta-analysis, one locus (on chromosome 2q13) was associated with gestational duration and postterm birth at genome-wide significance ($P < 5 \times 10^{-8}$) (**Supplementary Fig. 2A and 2B**), and two loci (on chromosomes 1p33 and 3q28) were significantly associated with early preterm birth (**Supplementary Fig. 2C**). No locus reached genome-wide significance for preterm birth (**Supplementary Fig. 2D**). We selected one lead SNP for each of the three loci reaching genome-wide significance for analysis in the replication stage.

A locus harboring pro-inflammatory cytokine genes

At the 2q13 locus, rs7594852 was the SNP most significantly associated with gestational duration ($P = 1.88 \times 10^{-12}$, **Figure 1A**) and was selected as the lead SNP for replication stage analysis. For postterm birth, we also selected rs7594852 ($P = 4.64 \times 10^{-8}$, **Figure 1B**) as the lead SNP for the replication analyses from a set of highly correlated SNPs ($r^2 > 0.98$) with very similar P values. The association of rs7594852 with gestational duration was replicated, with a P value of 3.69×10^{-3} in the replication sample and an overall P value of 3.96×10^{-14} in the combined discovery and replication analysis (**Table 1**). In the combined analysis, each additional fetal rs7594852-C allele was associated with an additional 0.37 days (95% confidence interval (CI) = 0.22–0.51) of gestational duration. For postterm birth the statistical power to replicate the association was modest at 40% (**Supplementary Table 2**) and the SNP did not reach nominal significance in the replication stage analysis, although the direction of the effect was consistent with the discovery stage (**Table 1**). rs7594852 is intronic in *CKAP2L* and is located in a linkage disequilibrium (LD) block that encompasses *IL1A*, *IL1B*, and several other genes encoding proteins in the interleukin 1 cytokine family (**Figures 1A and 1B**). In an additional analysis conditioning on rs7594852, we found no evidence for multiple independent signals at the locus (**Supplementary Fig. 3**). **Figure 2** shows a forest plot of association results for rs7594852 across all studies. The estimated variance in gestational duration explained by rs7594852 was 0.066%. The variance explained by all common SNPs (minor allele frequency, MAF>1%) was 7.3% (SE=0.7%).

No association was seen at the 2q13 locus in case-control analyses of early preterm birth or preterm birth (**Supplementary Fig. 4**), suggesting that the locus primarily influences gestational duration in the later stages of pregnancy. To further investigate this question, we binned the 51,357 births from the largest contributing study (iPSYCH) in 10 groups of equal size by gestational duration. We then estimated the frequency of the rs7594852-C allele in each group and in the whole sample. In the overall meta-analysis, each additional fetal rs7594852-C allele was associated with increased gestational duration (**Table 1**). The frequencies of the rs7594852-C allele in the two groups with shortest gestational durations were only slightly lower than the frequency in the whole sample (**Supplementary Fig. 5**). The lowest allele frequency (0.514) was seen in the third group, representing a mean gestational duration of 276 days. The allele frequency then gradually

increased in the next groups with the highest frequency (0.551) observed for the group representing the longest gestational duration (mean of 296 days) (**Supplementary Fig. 5**).

To explore possible functional mechanisms underlying the association signal at the 2q13 locus, we annotated the 283 variants within 1Mb of the lead SNP that were associated with gestational duration with $P < 1 \times 10^{-4}$ (**Supplementary Table 3**). Among these, 6 variants were categorized as exonic (2 synonymous and 4 missense variants, see **Supplementary Table 4**). Additionally, 190 of the 283 variants have been reported by the GTEx³⁰ and GEUVADIS³¹ consortia as cis-eQTLs ($P < 1 \times 10^{-4}$) for several nearby genes, including *IL1A*, *IL1B*, *IL36B*, and *IL1RN*, in different tissues. Of note, the lead variant rs7594852 C-allele is associated with decreased expression of *IL1A* in skin ($P = 7.36 \times 10^{-6}$) and decreased expression of *IL1B* in lymphoblastoid cell lines ($P = 4.41 \times 10^{-6}$). Among the 283 top variants at the 2q13 locus, 104 are located in likely enhancer regions of (mainly) cytokine genes (see Methods). GWAS Catalog³² annotation revealed that the variant rs10167914 ($P = 2.66 \times 10^{-7}$ for gestational duration, $r^2 = 0.64$ with rs7594852) has been reported to be associated with endometriosis³³ ($P < 5 \times 10^{-8}$), with the risk allele for endometriosis corresponding to increased gestational duration in our data (**Supplementary Table 3**).

Exome sequencing data were available for 11,455 subjects from the iPSYCH study. To investigate whether exonic variants could explain the observed association at the 2q13 locus, we tested all exonic variants in a 1 MB region around the lead SNP rs7594852 for association with gestational duration (see Methods). Among 841 exonic variants tested, 14 were associated with gestational duration at $P < 0.01$, with the lowest $P = 0.0005$. These variants were, however, either not in LD with rs7594852 ($r^2 < 0.01$) or did not remain associated after conditioning on rs7594852 genotype (**Supplementary Table 5**). Thus, we found no exonic variants likely to explain the observed association.

Early preterm birth associations

Our early preterm birth meta-analysis revealed two genome-wide significant loci in the discovery stage. At the first locus on chromosome 3q28, rs112912841 yielded the lowest P value (OR = 1.64, 95% CI = 1.38–1.94, $P = 9.85 \times 10^{-9}$). This SNP is intronic in *LPP* (**Supplementary Fig. 6A**). The most significant SNP at the second locus on chromosome 1p33, rs1877720 (OR = 1.64, 95% CI = 1.37–1.96, $P = 4.33 \times 10^{-8}$), is intronic in *SPATA6* (**Supplementary Fig. 6B**). There was no evidence for multiple independent signals at these loci when the lead SNP was included as a co-variate in conditional regression analyses (**Supplementary Fig. 7**). We conducted replication analyses for the two lead SNPs (rs112912841 and rs1877720) in an independent Finnish case-control study (107 infants born early preterm and 865 born at term), but statistical power was low (**Supplementary Table 2**) and we could not confirm the associations (**Supplementary Fig. 8 and Supplementary Table 6**). We also conducted family-based association tests of the lead SNPs at these two loci in a set of 276 early preterm birth trios of European ancestry from Iowa, but could not find support for the signals in that data set either (**Supplementary Table 7**).

Fetal or maternal genetic effects?

Gestational duration is a complex outcome influenced by both the maternal and fetal genomes. To disentangle fetal and maternal genetic effects, we first compared our fetal association results with those from a recent maternal GWAS of gestational duration²⁷. In our fetal analysis, each C allele of the lead variant, rs7594852, was associated with an additional 0.37 days (95% CI = 0.22–0.51) of gestational duration (estimate based on 51,357 infants from the iPSYCH study, see Methods). In the published maternal GWAS ($n = 43,568$), the direction of effect was the same: each maternal rs7594852-C allele was associated with an additional 0.22 days (95% CI = -0.01–0.45) of gestational duration, though the confidence intervals were wide and included the null value. The fact that the maternal effect estimate was approximately half the size of the fetal effect estimate (0.22 vs. 0.37 days) is as expected if the association is of fetal origin. Next, we performed joint maternal-fetal genetic association analyses of the lead variant, rs7594852, at the 2q13 locus in 15,536 mother-child pairs that met the inclusion criteria from the discovery stage (see Methods for details). Here we found that conditioning on maternal genotype did not attenuate the fetal genetic association while the maternal genetic association conditional on fetal genotype was non-significant (**Supplementary Table 8**). Taken together, these results indicate that the association signal for gestational duration at the 2q13 locus represents a fetal genetic effect. Conversely, we examined the lead variants at 4 of the 6 loci, which were reported to be significant in the maternal GWAS and were available in our meta-analysis (the remaining 2 were not autosomal)²⁷. The direction of effects in the fetal GWAS was consistent with the published maternal GWAS results, but fetal effect size estimates were smaller (**Supplementary Table 9**).

Functional analyses

Exome sequencing data and variant annotation of the 2q13 locus did not identify any exonic variants likely to explain the association with gestational duration, suggesting that the underlying mechanism might instead involve altered gene regulatory mechanisms. To investigate potential mechanisms, we conducted a series of functional experiments and analyses. We first prioritized all SNPs in the LD block containing the association signal based on their overlap with functional genomics datasets from cell types relevant to gestation (see Methods). The highest ranked variant was the discovery lead variant rs7594852, which overlaps with 17 different datasets, with no other variant intersecting more than three (**Supplementary Table 10**). We then used the Cis-BP database³⁴ to identify transcription factors that might bind differentially to this variant. These analyses revealed that the rs7594852-C allele might alter the binding of the hypermethylated in cancer 1 (HIC1) protein (**Figure 3A**). The HIC1 protein is a C₂H₂ zinc finger transcriptional repressor with a consensus DNA binding sequence containing a core “GGCA” motif³⁵. Protein binding microarray enrichment scores (E-scores) indicate strong binding of HIC1 to the cytosine “C” allele (E-score = 0.48), and only moderate binding to the alternative thymidine (“T”) allele (E-score = 0.32). We confirmed the presence of multiple histone marks overlapping rs7494852 in various fetal cell and tissue types (chorion, amniocytes, trophoblasts)³⁶, indicating that the chromatin in this locus is likely accessible and active in these fetal cells (**Figure 3B**). In particular, the fetal side of the placenta displays a strong H3K4me3 modification signal, a histone mark often found in active regulatory regions. Using an electrophoretic mobility shift assay, we detected enhanced binding of HIC1 to the

rs7594852- C allele (**Figure 3C**), as predicted. Densitometric analysis for HIC1 band intensity demonstrated a statistically significant difference in average intensity between the rs7594852-C and rs7594852-T alleles respectively (204.0 ± 60.0 vs 76.6 ± 42.3 , $P = 0.013$; $n = 4$ per group). Additional experiments are needed to investigate which of the genes at the 2q13 locus have altered transcriptional levels in relevant cell types due to the enhanced binding of HIC1 to the rs7594852-C allele.

Next, we examined associations of the lead SNP, rs7594852 with gene expression using RNA sequencing data from 102 human placentas³⁷. Here we found that among 118 genes/transcripts with transcription start sites within 500 kb of rs7594852, there were five nominally significant ($P < 0.05$) cis-eQTL associations (**Supplementary Table 11**). Three of these genes (*IL1A*, *IL36G*, *IL36RN*) encode proteins in the interleukin 1 cytokine family. The rs7594852-C allele, which in our data was associated with increased gestational duration, corresponded to decreased expression of the cytokine-encoding genes *IL1A* and *IL36G*. Conversely, the rs7594852-C allele corresponded to increased expression of *IL36RN*, which encodes an antagonist to the interleukin-36-receptor.

Finally, to evaluate a possible general effect of the 2q13 locus on inflammatory markers we tested for associations between the lead SNP rs7594852 and levels of inflammatory markers in peripheral blood from newborns, using data from the iPSYCH study. None of the cytokines encoded by genes at the 2q13 locus had been assayed, but measurements of the biomarkers BDNF, CRP, EPO, IgA, IL8, IL-18, MCP1, S100B, TARC, and VEGFA were available for 8,138 iPSYCH samples. However, we found no significant associations between rs7594852 genotype and levels of these biomarkers (results not shown). We also used the GWAS Catalog³² to identify 39 SNPs known from previous studies to be associated with cytokine levels and examined associations for these SNPs with gestational duration in our discovery stage meta-analysis. These SNPs did not show more evidence of association with gestational duration than expected by chance (**Supplementary Fig. 9**).

Discussion

In this genome-wide meta-analysis including a total of 93,980 infants in the discovery and replication stages, we identified a fetal locus on chromosome 2q13 that was robustly associated with gestational duration. The lead SNP at the locus, rs7594852, was also associated with postterm birth as a dichotomous outcome in the discovery stage, but not with either preterm birth or early preterm birth. An analysis of allele frequency in different strata of the gestational age distribution confirmed that genetic variation at the locus influences timing of parturition in later stages of pregnancy.

Gestational duration is a complex phenotype influenced by both fetal and maternal genetic contributions, as well as environmental factors. Since mothers and children share half of their genetic material, we investigated the possibility that the 2q13 association signal could represent a maternal, rather than a fetal, effect. This was not the case: our analysis of more than 15,000 mother-child pairs showed that the association had a fetal origin independent of the maternal genotype. Furthermore, the lead SNP, rs7594852, was not significantly associated

with gestational duration in a maternal GWAS including more than 40,000 women²⁷. More generally, we found that fetal genetic variants explained 7.3% of the variance in gestational duration, which is broadly in line with estimates of 11% and 13.1% from large family studies in populations of Scandinavian origin^{20,21}.

While the lead SNP at the new 2q13 locus is intronic in *CKAP2L*, which encodes a mitotic spindle protein, the locus also harbors a number of genes encoding proteins in the interleukin 1 family of pro-inflammatory cytokines. It is well-established that IL-1 signalling plays a central role in the process leading to parturition in healthy term pregnancies³⁸. However, infections or trauma can also induce increased secretion of IL-1 and other pro-inflammatory cytokines, provoking preterm birth³⁸. In our data, genetic variation at the 2q13 locus was only associated with gestational duration in later stages of pregnancy. We hypothesize that this locus is involved in genetic regulation of a pro-inflammatory cytokine signalling mechanism by which the mature fetus communicates to the mother that it is ready to be born. In a first step towards understanding the molecular mechanisms underlying the genetic association, we found that the rs7594852-C allele creates a strong binding site for the transcriptional repressor HIC1. It is conceivable that the variant thereby delays signalling from the fetus to the mother, thus prolonging gestation, possibly until other (redundant) mechanisms stimulating parturition take effect. We had limited statistical power to directly evaluate effects of the variant allele on expression levels of genes at the locus, but the results of eQTL analyses in 102 placental samples collected after birth were compatible with the hypothesis that the rs7594852-C allele may lead to prolonged gestational duration through decreased gene expression of one or more members of the interleukin 1 cytokine family genes. eQTL results from the GTEx and GEUVADIS data bases for skin tissue and lymphoblastoid cell lines, respectively, support this suggestion. However, to fully investigate this hypothesis, larger sample sizes and further functional follow-up experiments are needed, including fine-mapping of the locus and characterization of pro-inflammatory cytokine signalling shortly before parturition e.g., through non-invasive techniques allowing quantification of cell-free fetal RNA^{39,40} and measurement of cytokine levels.

The fact that the 2q13 locus was not associated with preterm birth or early preterm birth in our case-control analyses underlines that genetic triggers of parturition probably change as gestation advances. Genetic analysis of preterm birth is further complicated by the possible influence of a wide range of environmental factors. Population-based heritability analyses of gestational duration demonstrate that inclusion of early preterm births results in decreased estimates of the fetal genetic effect²⁰ and previous GWAS efforts have not identified robust fetal genetic associations with preterm birth^{22–24}. To refine outcome definitions, we applied a range of exclusion criteria, but although our analyses were based on almost 5,000 infants born preterm, with 1,139 of these considered early preterm births, no loci were robustly associated with preterm birth. Statistical power calculations suggested that our preterm birth analyses were well-powered to detect associations with common (MAF>0.1) fetal genetic variants with odds ratios >1.25 (**Supplementary Fig. 10**). Fetal genetic contributions to preterm birth may therefore involve smaller effect sizes or less frequent variants. Studies with larger sample sizes would be needed to address this question.

Our study had some limitations, including the restriction to infants of European descent. Further studies are therefore warranted to delineate fetal and maternal genetic contributions to gestational duration in populations of non-European ancestries. A second limitation of our study was the differences in gestational

duration ascertainment from study to study. In many of the older cohorts that recruited women who were pregnant prior to routine use of ultrasound scan dating, gestational duration estimates were based largely on maternal-reported last menstrual period at the time of pregnancy, whereas estimates for infants born more recently were predominantly based on first-trimester ultrasound screening (**Supplementary Table 1**). Also, the degree to which children were excluded based on maternal conditions or pregnancy complications differed among cohorts. However, while these sources of heterogeneity may have caused some underestimation of effect sizes at genuinely associated loci, it should not have resulted in increased false-positive rates. Our extensive exclusion criteria aimed to focus on ‘natural’ gestational duration rather than specific causes such as preterm birth due to pregnancy complications, assisted delivery or congenital anomalies. Although such exclusions can result in selection bias⁴¹, the overall consistency between studies (**Table 1, Figure 2**), despite their varying ability to completely apply all of our pre-specified exclusion criteria, provides some reassurance that any such bias may not be large. One might also speculate as to whether spurious signals could arise from the case groups of various diseases that were included in the discovery stage analyses. However, we consider this unlikely since the association analyses were stratified by disease group and since we did not observe heterogeneity of effect estimates between studies of various design (including population-based cohorts) for the 2q13 lead variant (**Table 1, Figure 2**).

In conclusion, parturition is a complex physiological process involving multiple redundant mechanisms influenced by maternal and fetal factors². An enhanced understanding of these mechanisms is of great public health importance, since giving birth either before or after the window of term gestation is associated with increased morbidity and mortality^{8,13}. Our study identified the first robustly associated fetal genetic locus for gestational duration. The effect was observed in later stages of pregnancy and our results raise the hypothesis that variants at the associated locus influence the regulation of pro-inflammatory cytokines in the IL-1 family. Our findings provide a foundation for further functional studies that are required to refine our understanding of the biology of the timing of parturition.

URLs

1000 Genomes Project, <http://www.1000genomes.org/>; ANNOVAR, <http://annovar.openbioinformatics.org/>; Cis-BP, <http://cisbp.cibr.utoronto.ca/>; dbGaP, <https://www.ncbi.nlm.nih.gov/gap>; EGG consortium, <http://egg-consortium.org/>; Ensembl Variant Effect Predictor, <https://www.ensembl.org/vep>; GeneHancer, <http://www.genecards.org/>; GEUVADIS data browser, <http://www.ebi.ac.uk/Tools/geuvadis-das/>; GWAS Catalog, <http://www.genome.gov/gwastudies/>; Haplotype Reference Consortium, <http://www.haplotype-reference-consortium.org/>; iPSYCH: <http://ipsych.au.dk/about-ipsych/>; LD Score regression, <https://github.com/bulik/ldsc>; METAL, <http://www.sph.umich.edu/csg/abecasis/metal/>; NCBI Genotype-Tissue Expression (GTEx) eQTL database and browser, <http://www.ncbi.nlm.nih.gov/projects/gap/eql/index.cgi>; PLINK: <https://www.cog-genomics.org/plink2>; RVTESTS, <https://github.com/zhanxw/rvtests>; SNPTTEST, https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html

References

1. Moffett, A. & Loke, C. Immunology of placentation in eutherian mammals. *Nature Reviews Immunology* **6**, 584–594 (2006).
2. Smith, R. Parturition. *N. Engl. J. Med.* **356**, 271–283 (2007).
3. Medawar, P. B. Some immunological and endocrinological problems raised by the evolution of viviparity in vertebrates. *Symp Soc Exp Biol* **7**, 320–338 (1953).
4. Gundling, W. E. & Wildman, D. E. A review of inter- and intraspecific variation in the eutherian placenta. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **370**, 20140072 (2015).
5. Racicot, K., Kwon, J. Y., Aldo, P., Silasi, M. & Mor, G. Understanding the complexity of the immune system during pregnancy. *American Journal of Reproductive Immunology* **72**, 107–116 (2014).
6. Chavan, A. R., Griffith, O. W. & Wagner, G. P. The inflammation paradox in the evolution of mammalian pregnancy: turning a foe into a friend. *Curr. Opin. Genet. Dev.* **47**, 24–32 (2017).
7. Romero, R., Dey, S. K. & Fisher, S. J. Preterm labor: One syndrome, many causes. *Science* **345**, 760–765 (2014).
8. Blencowe, H. *et al.* Born Too Soon: The global epidemiology of 15 million preterm births. *Reprod. Health* **10**, (2013).
9. Moster, D., Lie, R. T. & Markestad, T. Long-term medical and social consequences of preterm birth. *N. Engl. J. Med.* **359**, 262–73 (2008).
10. Li, S. *et al.* Preterm birth and risk of type 1 and type 2 diabetes: Systematic review and meta-analysis. *Obes. Rev.* **15**, 804–811 (2014).
11. Parkinson, J. R. C., Hyde, M. J., Gale, C., Santhakumaran, S. & Modi, N. Preterm birth and the metabolic syndrome in adult life: a systematic review and meta-analysis. *Pediatrics* **131**, e1240-63 (2013).
12. Nosarti, C. *et al.* Preterm birth and psychiatric disorders in young adult life. *Arch. Gen. Psychiatry* **69**, E1-8 (2012).
13. Olesen, A. W., Westergaard, J. G. & Olsen, J. Perinatal and maternal complications related to postterm delivery: A national register-based study, 1978-1993. *Am. J. Obstet. Gynecol.* **189**, 222–227 (2003).
14. European Perinatal Health Report 2010. (2018).
15. Martin, J. A., Hamilton, B. E., Osterman, M. J., Curtin, S. C. & Matthews, T. J. Births: final data for 2013. *Natl. Vital Stat. Reports From Centers Dis. Control Prev. Natl. Cent. Heal. Stat. Natl. Vital Stat. Syst.* **64**, 1–65 (2015).
16. Blencowe, H. *et al.* National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. *Lancet (London, England)* **379**, 2162–2172 (2012).

17. Bezold, K. Y., Karjalainen, M. K., Hallman, M., Teramo, K. & Muglia, L. J. The genomics of preterm birth: from animal models to human studies. *Genome Med.* **5**, 34 (2013).
18. York, T. P., Eaves, L. J., Neale, M. C. & Strauss, J. F. The contribution of genetic and environmental factors to the duration of pregnancy. *Am. J. Obstet. Gynecol.* **210**, 398–405 (2014).
19. Clausson, B., Lichtenstein, P. & Cnattingius, S. Genetic influence on birthweight and gestational length determined by studies in offspring of twins. *BJOG.* **107**, 375–381 (2000).
20. Lunde, A., Melve, K. K., Gjessing, H. K., Skjaerven, R. & Irgens, L. M. Genetic and environmental influences on birth weight, birth length, head circumference, and gestational age by use of population-based parent-offspring data. *Am. J. Epidemiol.* **165**, 734–741 (2007).
21. York, T. P. *et al.* Fetal and maternal genes' influence on gestational age in a quantitative genetic analysis of 244,000 Swedish births. *Am. J. Epidemiol.* **178**, 543–550 (2013).
22. Zhang, H. *et al.* A genome-wide association study of early spontaneous preterm delivery. *Genet. Epidemiol.* **39**, 217–226 (2015).
23. Rappoport, N. *et al.* A genome-wide association study identifies only two ancestry specific variants associated with spontaneous preterm birth. *Sci. Rep.* **8**, 226 (2018).
24. Bacelis, J. *et al.* Literature-Informed Analysis of a Genome-Wide Association Study of Gestational Age in Norwegian Women and Children Suggests Involvement of Inflammatory Pathways. *PLoS One* **11**, e0160335 (2016).
25. Uzun, A., Dewan, A. T., Istrail, S. & Padbury, J. F. Pathway-based genetic analysis of preterm birth. *Genomics* **101**, 163–170 (2013).
26. Schierding, W. *et al.* GWAS on prolonged gestation (post-term birth): analysis of successive Finnish birth cohorts. *J. Med. Genet.* **55**, 55–63 (2018).
27. Zhang, G. *et al.* Genetic Associations with Gestational Duration and Spontaneous Preterm Birth. *N. Engl. J. Med.* NEJMoa1612665 (2017). doi:10.1056/NEJMoa1612665
28. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
29. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
30. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
31. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
32. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
33. Sapkota, Y. *et al.* Meta-analysis identifies five novel loci associated with endometriosis highlighting key genes involved in hormone metabolism. *Nat. Commun.* **8**, 15539 (2017).

34. Weirauch, M. T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).
35. Pinte, S. *et al.* The tumor suppressor gene HIC1 (hypermethylated in cancer 1) is a sequence-specific transcriptional repressor: Definition of its consensus binding sequence and analysis of its DNA binding and repressive properties. *J. Biol. Chem.* **279**, 38313–38324 (2004).
36. Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* **28**, 1045–1048 (2010).
37. Peng, S. *et al.* Expression quantitative trait loci (eQTLs) in human placentas suggest developmental origins of complex diseases. *Hum. Mol. Genet.* **26**, 3432–3441 (2017).
38. Nadeau-Vallée, M. *et al.* A critical role of interleukin-1 in preterm labor. *Cytokine Growth Factor Rev.* **28**, 37–51 (2016).
39. Koh, W. *et al.* Noninvasive in vivo monitoring of tissue-specific global gene expression in humans. *Proc.Natl.Acad.Sci.U.S.A* **111**, 7361–7366 (2014).
40. Ngo, T. T. M. *et al.* A Noninvasive Molecular Clock for Fetal Development Predicts Gestational Age and Preterm Delivery. *bioRxiv* 212910 (2017). doi:10.1101/212910
41. Munafò, M. R., Tilling, K., Taylor, A. E., Evans, D. M. & Smith, G. D. Collider scope: When selection bias can substantially influence observed associations. *Int. J. Epidemiol.* **47**, 226–235 (2018).
42. Pedersen, C. B. *et al.* The iPSYCH2012 case-cohort sample: new directions for unravelling genetic and environmental architectures of severe mental disorders. *Mol. Psychiatry* **23**, 6–14 (2018).
43. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am.J.Hum.Genet.* **81**, 559–575 (2007).
44. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat.Rev.Genet.* **11**, 499–511 (2010).
45. Zhan, X., Hu, Y., Li, B., Abecasis, G. R. & Liu, D. J. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics* **32**, 1423–1426 (2016).
46. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet.Epidemiol.* **34**, 816–834 (2010).
47. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
48. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
49. Higgins, J. P. & Thompson, S. G. Quantifying heterogeneity in a meta-analysis. *Stat.Med.* **21**, 1539–1558 (2002).
50. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. <http://www.R-project.org/> (2018).

51. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
52. Fishilevich, S. *et al.* GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database J. Biol. Databases Curation* **2017**, (2017).
53. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
54. Matteini, A. M. *et al.* Novel gene variants predict serum levels of the cytokines IL-18 and IL-1ra in older adults. *Cytokine* **65**, 10–16 (2014).
55. Voruganti, V. S. *et al.* Genome-wide association replicates the association of Duffy antigen receptor for chemokines (DARC) polymorphisms with serum monocyte chemoattractant protein-1 (MCP-1) levels in Hispanic children. *Cytokine* **60**, 634–638 (2012).
56. Ahola-Olli, A. V *et al.* Genome-wide Association Study Identifies 27 Loci Influencing Concentrations of Circulating Cytokines and Growth Factors. *Am. J. Hum. Genet.* **100**, 40–50 (2017).
57. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
58. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
59. Liu, T. *et al.* Cistrome: An integrative platform for transcriptional regulation studies. *Genome Biol.* **12**, (2011).
60. Griffon, A. *et al.* Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. *Nucleic Acids Res.* **43**, (2015).
61. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22**, 1790–1797 (2012).
62. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

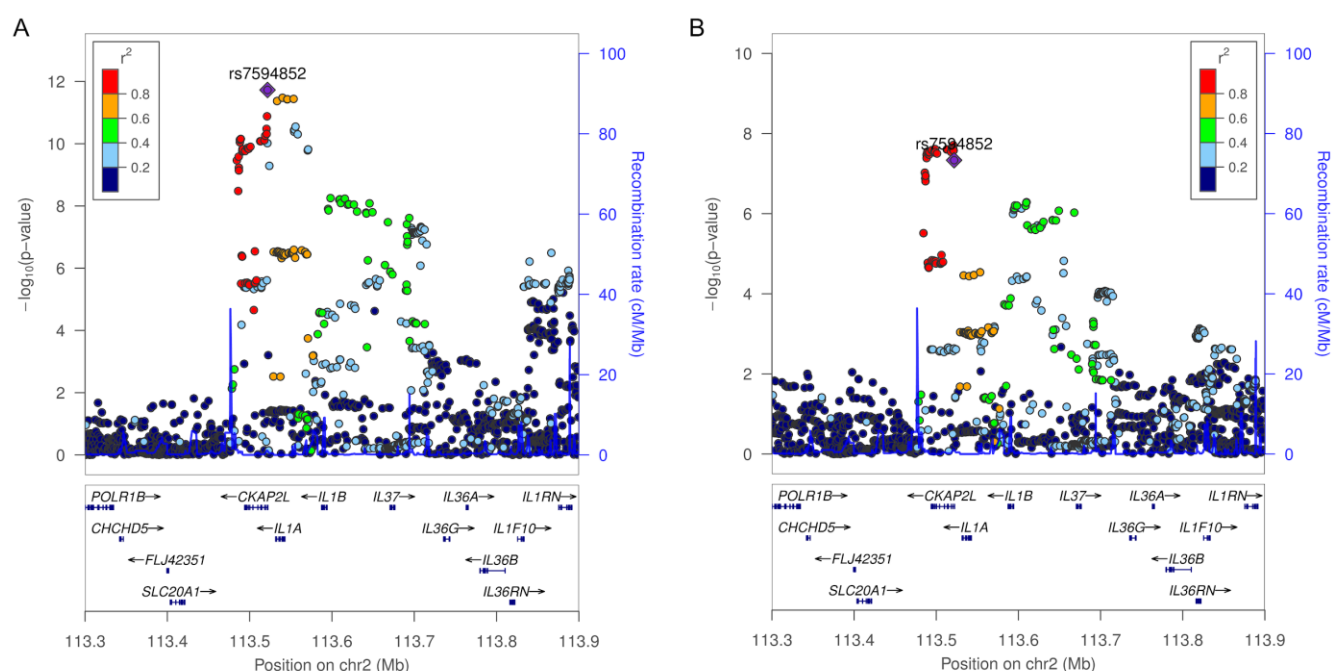


Figure 1. Discovery stage results at the 2q13 locus. Regional association plots for (A) gestational duration, and (B) postterm birth. SNP position is shown on the x-axis and association ($-\log_{10} P$ value) with gestational duration and postterm birth, respectively, on the left y-axis. The lead SNP, rs7594852, from the gestational duration analysis is represented by a purple diamond, and the other SNPs are colored to reflect their LD with the lead SNP (based on pairwise r^2 values from the DNBC). In the postterm birth analysis, rs7607470 had a slightly lower P value, but this SNP is highly correlated with rs7594852 ($r^2 > 0.99$), and the latter SNP was selected for the replication stage analyses of both gestational duration and postterm birth. Estimated recombination rates are from HapMap (right y-axis).

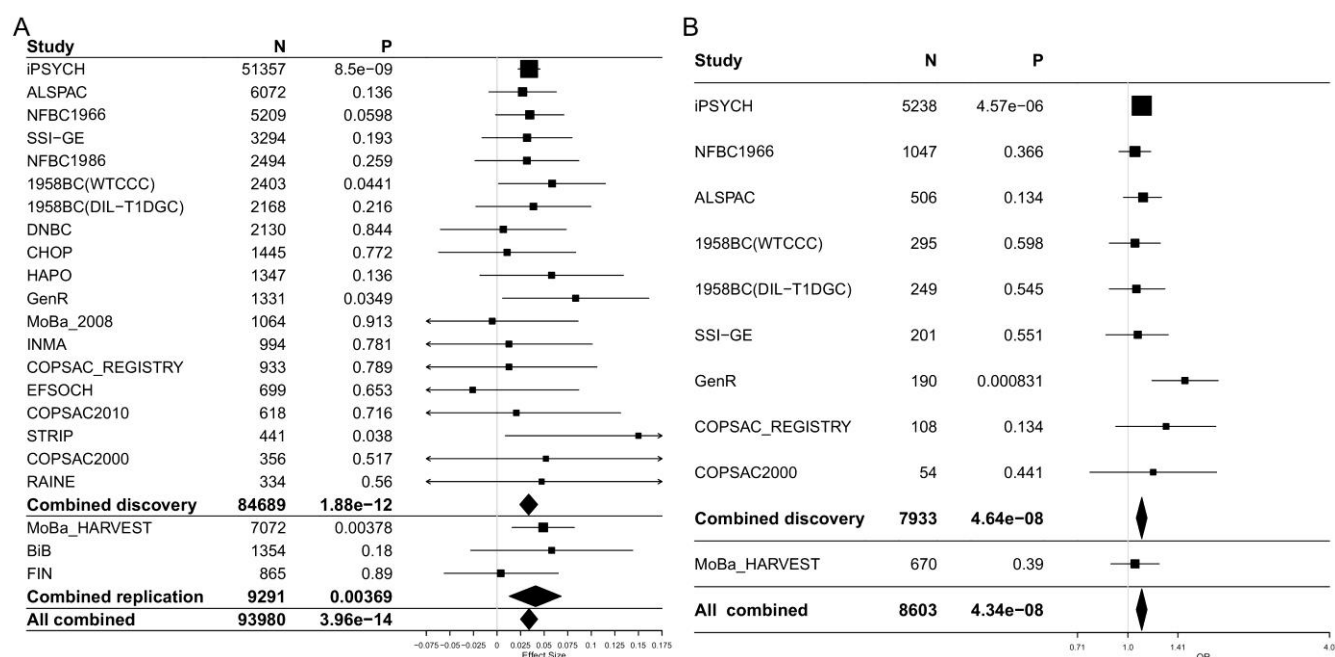


Figure 2. Forest plots showing association results for rs7594852 in the discovery and replication stage cohorts. (A), Gestational duration effect estimates with 95% CIs, and (B) postterm birth ORs with 95% CIs.

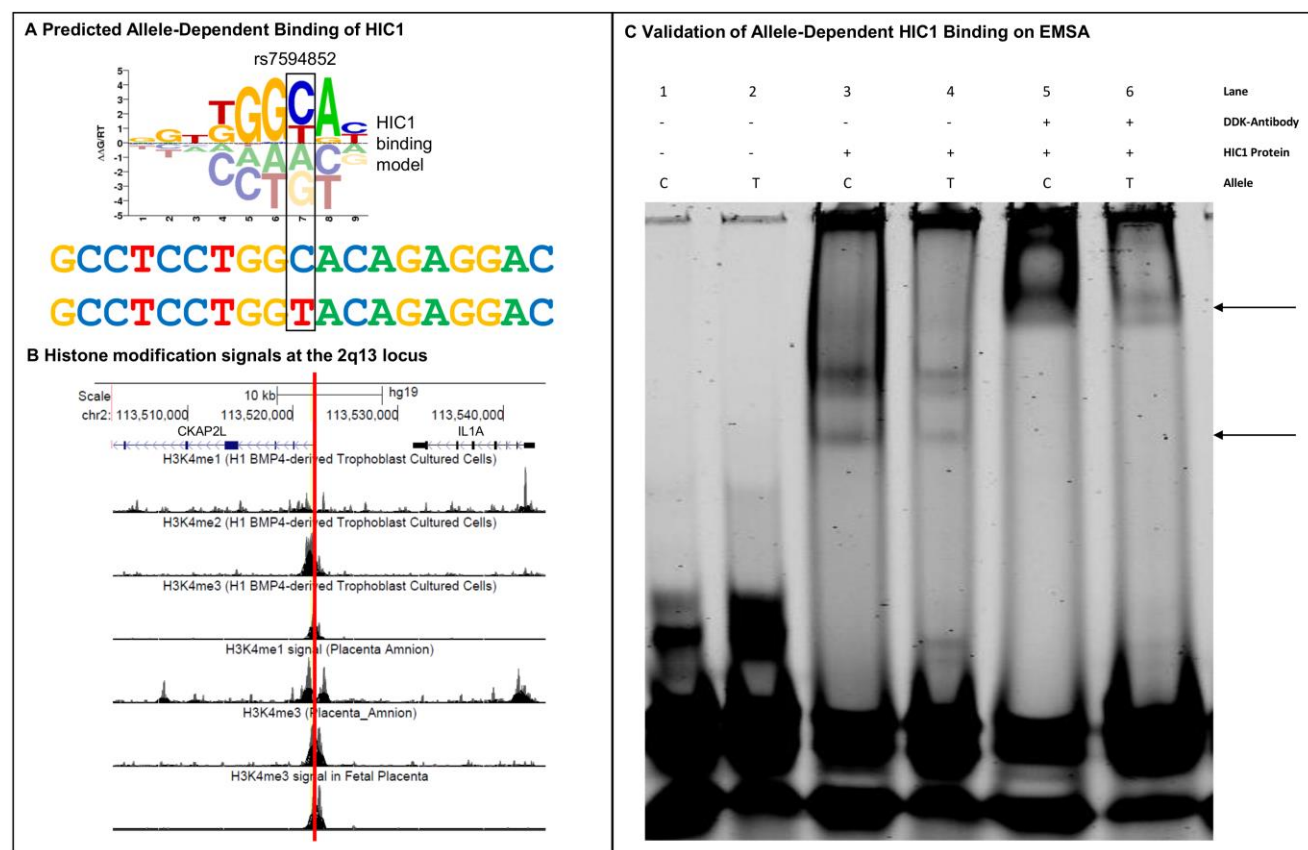


Figure 3. HIC1 Binding at the 2q13 locus. (A) The rs7594852-C allele creates a stronger binding site for the hypermethylated in cancer 1 (HIC1) protein. The sequence logo of the HIC1 binding motif shows the DNA binding preferences of HIC1. Tall nucleotides above the dashed line indicate DNA bases that are preferred by HIC1, 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 ($\Delta\Delta G$) in units of gas constant (R) and temperature (T). The DNA sequence flanking the rs7594852-C allele is shown directly below, with the alternative rs7594852-T allele shown at the bottom. The rs7594852-T allele changes the HIC1 binding site sequence from C (most favored) to T (less favored). (B) UC Santa Cruz Genome Browser screenshot depicting the rs7594852 locus. The purple (*CKAP2L*) and black (*IL1A*) graphics at the top indicate the locations of exons (columns), untranslated regions (rectangles), and introns (horizontal lines), with arrows indicating the direction of transcription. The red vertical line indicates the position of rs7594852, which overlaps strong signals obtained from chromatin immunoprecipitation sequencing (ChIP-seq) experiments (indicating histone modification by mono-, di-, or trimethylation of histone H3 on lysine 4; H3K4) in trophoblast cultured cells, placental amnion, or fetal placenta. (C) Experimental validation of allele-dependent binding of human purified recombinant HIC1 protein with a c-Myc/DDK tag to rs7594852 and flanking sequence via electrophoretic mobility shift assay (EMSA). Arrows indicate allele-dependent binding of HIC1 (bottom arrow) and a “supershift” of the protein-DNA complex induced by the binding of the anti-DDK antibody to the complex (top arrow). The presence of multiple bands in lanes 3 and 4 is likely due to the presence of multiple HIC1 isoforms.

Table 1. Discovery, replication and combined results for the lead SNP rs7594852 at the 2q13 locus.

Phenotype	Sample sets	rs7594852 effect allele (C) frequency		Number		Beta/OR*	P	I ²	P _{het}
						(95% CI)		(95% CI)	
		cases	controls	cases	controls				
Gestational duration	Combined discovery		0.53		84689	0.034 (0.024–0.043)	1.88 × 10 ⁻¹²	0 (0.0–13.4)	0.97
	MoBa_HARV EST		0.54		7072	0.049 (0.016–0.082)	3.78 × 10 ⁻³		
	BiB		0.51		1354	0.058 (-0.028–0.144)	0.18		
	FIN		0.58		865	0.004 (-0.056–0.065)	0.89		
	Combined replication				9291	0.041 (0.013–0.068)	3.69 × 10 ⁻³	0 (0.0–97.4)	0.41
	All combined				93980	0.034 (0.025–0.043)	3.96 × 10⁻¹⁴	0 (0.0–99.6)	0.64
Postterm birth	Combined discovery	0.55	0.53	7933	60148	1.1 (1.06–1.14)	4.64 × 10 ⁻⁸	7.3 (0.0–88.1)	0.37
	MoBa_HARV EST	0.55	0.54	670	5626	1.05 (0.89–1.24)	0.39		
	All combined			8603	65774	1.1 (1.06–1.14)	4.34 × 10⁻⁸	0 (0.0–99.7)	0.58

* For the quantitative trait of quantile transformed gestational duration, the column reports beta estimates. For the dichotomous trait postterm birth, odds ratio (OR) estimates are given; CI, confidence interval; I², heterogeneity estimate (proportion of variance that is due to between study differences); P_{het}, P value from the *Cochran Q* test of between study heterogeneity.

Methods

Discovery stage cohorts

Analyses were performed among participants of studies in the Early Growth Genetics (EGG) consortium, the Initiative for Integrative Psychiatric Research (iPSYCH), and the Genomic and Proteomic Network for Preterm Birth Research (GPN). The iPSYCH sample ($n = 51,357$) included patient groups of six mental disorders: autism, ADHD, schizophrenia, bipolar disorder, depression and anorexia⁴². Participating studies from the EGG consortium included the Avon Longitudinal Study of Parents and Children (ALSPAC, $n = 6,072$) study, the Children's Hospital of Philadelphia (CHOP, $n = 1,445$) cohort, three sub-samples from the Copenhagen Prospective Studies on Asthma in Childhood (COPSAC_REGISTRY, $n = 933$; COPSAC2000, $n = 356$; COPSAC2010, $n = 618$), a sub-sample from the Danish National Birth Cohort (DNBC, $n = 2,130$), the Exeter Family Study of Children's Health (EFSOCH, $n = 699$) study, the GENERATION R (GenR, $n = 1,331$) study, the Hyperglycemia and Adverse Pregnancy Outcome (HAPO, $n = 1,347$) study, the Infancia y Medio Ambiente (INMA, $n = 994$) study, a sub-sample of the Norwegian Mother and Child cohort study (MoBa_2008, $n = 1064$), two Northern Finland Birth Cohort studies (NFBC1966, $n = 5,209$, and NFBC1986, $n = 2,494$), the Western Australian Pregnancy Cohort Study (Raine Study, $n = 334$), a sub-sample of Statens Serum Institut's genetic epidemiology (SSI-GE, $n = 3,294$) studies, the Special Turku coronary Risk factor Intervention Project (STRIP, $n = 441$), the Diabetes and Inflammation Laboratory (1958BC (DIL-T1DGC), $n = 2,168$) cohort, and the Wellcome Trust Case Control Consortium 1958 British Birth Cohort (1958BC (WTCCC), $n = 2,403$). Study protocols within the EGG consortium were approved at each study centre by the local ethics committee and written informed consent had been obtained from all participants and/or their parent(s) or legal guardians. Regarding the iPSYCH and SSI-GE cohorts, parents are informed in writing about the neonatal screening and that the samples can be used for research, pending approval from relevant authorities⁴². The GPN study genotype and phenotype data were downloaded from dbGaP (<https://www.ncbi.nlm.nih.gov/gap>, accession number: phs000714.v1.p1) and 190 early preterm cases and 274 term infants were included in our early preterm and preterm birth meta-analysis. Study descriptions, relevant sample sizes and basic characteristics of samples in the discovery stage are presented in **Supplementary Table 1**.

Replication stage cohorts

Three population based cohorts were used for replication stage analyses of the lead SNPs for gestational duration and early preterm birth, namely an additional sub-sample of the Norwegian Mother and Child cohort study (MoBa_HARVEST, $n = 7,072$), the Born in Bradford study (BiB, $n = 1,354$), and a Finnish cohort from Helsinki (FIN, $n = 865$). We also included a set of mother-father-child trios from Iowa ($n = 276$ trios) for family-based association testing of the lead SNPs for early preterm birth. The study characteristics of these four studies are described in **Supplementary Table 1**.

Exclusion criteria for cases and controls

We excluded pregnancies based on the following criteria: 1) stillbirths; 2) twins or any multiple births; 3) ancestry outliers using principal component analysis; 4) outliers in birth weight or birth length (gestational duration possibly wrong); 5) Caesarian section, if due to pregnancy complications; Caesarian sections due to complications during labor were not excluded. Caesarian sections were allowed for cases in the postterm birth analysis; 6) Physician initiated births (induced births were allowed for cases in the postterm birth analysis); 7) placental abruption, placenta previa, pre-eclampsia/eclampsia, hydramnios, placental insufficiency, cervical insufficiency, isoimmunization, gestational diabetes, cervical cerclage; 8) pre-existing medical conditions in the mother, such as diabetes, hypertension, autoimmune diseases (including systemic lupus erythematosus, rheumatoid arthritis and sclerodermia), immuno-compromised patients; and 9) known congenital anomalies. Further, the study sample was restricted to individuals of European ancestry, in most cohorts by principal component analysis. Some cohorts were not able to perform exclusions according to all criteria, but applied as many criteria as possible (see **Supplementary Table 1** for details).

Data cleaning and imputation

Genotyping in each of the contributing studies was conducted using various high-density SNP arrays (see **Supplementary Table 1** for details). Data cleaning was done locally for each study, with sample level exclusion criteria based on high genotype missing rate, high autosomal heterozygosity rate, discrepancy between reported sex and the sex inferred from genotyping, and sample heterogeneity, as well as SNP-level exclusion criteria based on call rate, Hardy-Weinberg disequilibrium, duplicate discordance, Mendelian inconsistencies, and low minor allele frequency. Imputation was performed based on reference data from the Haplotype Reference Consortium (HRC) release 1.1²⁸ for most studies. The iPSYCH sample was imputed based on the integrated phase III release of the 1000 Genomes Project²⁹. Study-specific details on data cleaning filters and imputation are given in **Supplementary Table 1**. SNP positions were based on National Center for Biotechnology Information (NCBI) build 37 (hg19) and alleles were labelled on the positive strand of the reference genome.

GWAS analysis

We analysed four traits for association with fetal genotypes, including three binary case-control traits (early preterm birth, preterm birth, and postterm birth) and one quantitative trait (gestational duration). Early preterm birth cases were defined as infants born before gestational week 34+0 (i.e. <238 days of gestation); preterm birth cases were infants born before gestational week 37+0 (i.e. <259 days of gestation, including the early cases); postterm birth cases were infants born at or after gestational week 42+0 (i.e. ≥ 294 days of gestation). The controls used in the analyses of these three traits were defined as infants born at or after gestational week 39+0 and before gestational week 42+0 (i.e. ≥ 273 days of gestation, and < 294 days of gestation). Case groups in each study contributing to the discovery stage analyses were required to contain at least 50 individuals.

For the dichotomous traits early preterm birth, preterm birth, and postterm birth, the genome-wide association analyses within discovery studies was done by logistic regression using imputed allelic dosage data under an additive genetic model. For the quantitative trait gestational duration, we applied a rank-based inverse normal transformation. More specifically, in each cohort gestational duration in days (in some cohorts converted from weeks; see **Supplementary Table 1**) was regressed on infant sex and the resulting residuals were quantile transformed to a standard normal distribution before being tested for association with fetal SNP genotypes. The DNBC and MoBa_2008 samples represent case-control studies of preterm birth, which means that the distribution of gestational duration is bimodal for these studies. In these two cohorts, we transformed gestational duration to be on the same scale as the population-based cohorts (See **Supplementary Methods** and **Supplementary Fig. 11** for details).

Some of the analyzed cohorts represent case-control studies of various diseases. For these, the association analyses of the four outcomes of interest were done in strata defined by disease group. Thus, the iPSYCH study was divided into six patient groups (autism ($n = 7,147$), ADHD ($n = 8,606$), schizophrenia ($n = 1,101$), bipolar disorder ($n = 864$), depression ($n = 13,836$), and anorexia ($n = 1,924$)) and a population control group ($n = 17,879$), which were analyzed separately and combined by fixed-effects meta-analysis. Similarly, the SSI-GE sample was split into six patient groups (atrial septal defects ($n = 368$), febrile seizures ($n = 1,350$), hydrocephalus ($n = 289$), hypospadias ($n = 301$), opioid dependence ($n = 685$), and postpartum depression ($n = 301$)) (see **Supplementary Table 1** for details), which were analyzed separately and combined by fixed-effects meta-analysis. Genome-wide association analyses in each cohort/sub-sample was conducted using PLINK⁴³, SNPTEST⁴⁴ or RVTESTS⁴⁵.

We obtained effect size estimates of the lead variant for gestational duration in the unit of days based on the iPSYCH study. Using a linear model, we regressed gestational duration in days on SNP allele dosage within each iPSYCH disease group, adjusting for infant sex. A combined estimate was obtained by fixed-effects inverse-variance meta-analysis. The iPSYCH study was also used to obtain frequency estimates for the lead variant for gestational duration within samples grouped by gestational duration. In this case, iPSYCH disease status was omitted from the model, since sample sizes would be too small if analyses were stratified by gestational duration groups as well as iPSYCH disease groups.

Meta-analysis

Prior to meta-analysis, SNPs with a minor allele frequency (MAF) < 0.01 and poorly imputed SNPs ($r^2_{\text{hat}} < 0.3$ from MACH⁴⁶ or $\text{info} < 0.4$ from SNPTEST⁴⁴) were excluded. Furthermore, SNPs available in less than 50% of the discovery cohorts for each trait were excluded. To adjust for inflation in test statistics generated in each cohort, genomic control⁴⁷ was applied once to each individual study (see **Supplementary Table 12** for λ values in each study). The sub-samples within iPSYCH and SSI-GE were meta-analyzed separately first and estimates were then adjusted by genomic control again. Finally, we combined results from all discovery cohorts using fixed-effects inverse variance-weighted meta-analysis as implemented in METAL⁴⁸. Final meta-analysis results were obtained for 7,646,297 SNPs for gestational duration with a genomic inflation factor (λ) of 1.049, 7,588,467

SNPs for early preterm birth ($\lambda=1.005$), 7,545,601 SNPs for preterm birth ($\lambda=1.013$), and 7,583,965 SNPs for postterm birth ($\lambda=1.026$). Heterogeneity between studies was estimated using the I^2 statistic⁴⁹. Combined analysis of the discovery and replication stage data was also conducted by fixed-effects inverse variance-weighted meta-analysis. We considered SNPs with $P < 0.05$ in the replication stage and $P < 5 \times 10^{-8}$ in the combined analysis to indicate robust evidence of association.

Power analysis

We assessed the statistical power of our study design by computer simulations in R⁵⁰. For gestational duration, we simulated a quantitative trait influenced by an additive genetic effect and allowed the effect size and the effect allele frequency to vary. For early preterm, preterm, and postterm birth, we simulated disease state from a logistic regression model allowing the odds ratio for a log-additive genetic effect and the frequency of the effect allele to vary. For each combination of effect size and effect allele frequency, we simulated 5000 data sets using the relevant sample size (e.g., for preterm birth: 4,775 cases and 60,148 controls in the discovery stage). We then conducted association tests on the simulated data sets and calculated power as the proportion of tests with a P value lower than the relevant significance level ($P < 5 \times 10^{-8}$ for the discovery stage and $P < 0.05$ for the replication stage).

Bioinformatics analysis

To investigate the functional characteristics of our findings, we annotated all variants with $P < 1 \times 10^{-4}$ at the 2q13 locus using ANNOVAR⁵¹ (accessed 1 June, 2017), a tool that retrieves variant and region-specific functional annotations from several databases. We retrieved eQTL information for these variants from the GTEx V6³⁰, and GEUVADIS³¹ project databases. We also queried GeneHancer⁵², a database of human enhancers and their inferred target genes, which has integrated four different enhancer datasets, including the Encyclopedia of DNA Elements (ENCODE), the Ensembl regulatory build, the functional annotation of the mammalian genome (FANTOM) project and the VISTA Enhancer Browser. Gene-enhancer scores (>5) were included in the annotation of the variants. We further downloaded all reported variants in the National Human Genome Research (NHGRI) GWAS Catalog³² (accessed 24 Nov, 2017) associated with a trait or disease at $P < 5 \times 10^{-8}$, and searched for SNPs in LD ($r^2 > 0.2$) with the lead SNP at 2q13 locus. Further annotation of these variants was performed with the Ensembl Variant Effect Predictor⁵³.

To assess possible enrichment of cytokine-related variants in the association results for gestational duration, we did a quantile-quantile plot of observed versus expected $-\log_{10} P$ values of SNPs known to be associated with cytokine levels (**Supplementary Fig. 9**). The cytokine related SNPs were restricted to cytokine GWAS publications^{54–56}, in which the association had been reported in the GWAS Catalog with $P < 5 \times 10^{-8}$.

Exome analysis

Exome sequencing data were available for a subset of samples in the iPSYCH study and analysis was restricted to the overlap between iPSYCH exome samples and the part of the iPSYCH cohort that were included in the GWAS. In total, $n = 11,455$ individuals, sampled from either schizophrenia ($n = 1,036$), bipolar ($n = 816$), ADHD

($n = 2,353$), autism ($n = 2,664$), affective disorder ($n = 1$) or controls ($n = 4,585$) were analyzed. For these samples, variants within a 1 MB region (113MB - 114MB) containing the 2q13 association signal were extracted and combined with the genotype data for the lead variant rs7594852. For these variants, association analysis was performed with gestational duration, transformed as described above. We adjusted the regression model for sex and the first three principal components obtained from the genotyping data. Due to small sample size in the strata of inclusion diagnosis, we did not perform analyses within strata of inclusion diagnosis but instead performed adjustment for four indicator variables denoting whether the individual has schizophrenia, ADHD, bipolar disorder or autism. In addition, association analysis conditioned on rs7594852 was performed by adding rs7594852 dosage as a covariate.

Biomarker analysis

Measurement of the biomarkers BDNF, CRP, EPO, IgA, IL8, IL-18, MCP1, S100B, TARC, and VEGFA was conducted based on infant dried blood spot samples obtained a few days after birth during routine neonatal screening. We tested each measured analyte for association with the lead SNP rs7594852 for gestational duration. We first fitted a linear model with age as a predictor and then normalized and log-transformed the residuals. The log-transformed residuals were tested for association with rs7594852 dosage while adjusting for infant sex, 6 principal components and iPSYCH disorders.

eQTL analysis in placental samples

eQTL analyses were conducted based on existing RNA sequencing data in placental samples from the Rhode Island Child Health Study. Placenta tissues were from singleton, term pregnancies without pregnancy complications. Details on sample collection, DNA and RNA extraction, SNP array genotyping and RNA sequencing have been described previously³⁷. In the eQTL analyses of the current study, the sample was restricted to 102 infants of European ancestry. Only genes/transcripts with transcription start sites within 500 kb of the lead SNP rs7594852 for gestational duration, with a total read count > 50 across all samples, and with >1 counts per million (cpm) in at least two samples, were considered.

Computational prediction of gene regulatory mechanisms

In order to prioritize genetic variants for experimental validation, we ranked all variants at the 2q13 locus with $r^2 > 0.8$ to the lead SNP, rs7594852, by their likelihood of being functional based on the strength of the supporting functional genomic data (e.g., ChIP-seq peaks for transcription factors or histone marks, open chromatin as measured by DNase-seq, see **Supplementary Table 10** for details). We used a wide range of functional genomic data in our analysis obtained from sources such as the UCSC Genome Browser⁵⁷, Roadmap Epigenomics⁵⁸, Cistrome⁵⁹, and ReMap-ChIP⁶⁰. By restricting our analysis to studies performed in relevant cell lines (placenta, chorion, amnion, trophoblasts, neutrophils, and macrophages), we prioritized those variants likely to have regulatory function in these cells. Variants were ranked based upon the total number of datasets they overlap, which is a similar strategic scheme to that employed by RegulomeDB⁶¹.

Electrophoretic mobility shift assays (EMSAs)

EMSAs were performed to determine whether the rs7594852 polymorphism at the 2q13 locus differentially affected HIC1 binding. Recombinant human HIC1 purified protein (ORIGENE #TP322752) was obtained from ORIGENE (expressed in HEK293 using TrueORF clone, RC222752) with a c-Myc/DDK tag. Double-stranded IRDye700 5' end-labelled 39 bp oligonucleotides, identical except for the nucleotide at rs7594852 (either the "C" or "T" allele), were obtained from IDT. The oligo sequence of the common "C" allele is:

5' -IRDye700/ GCCAGACCCCGCCTCTGGCACAGAGGACCACGCCCCGGC-3'.

The alternative "T" allele oligo sequence is:

5' -IRDye700/ GCCAGACCCCGCCTCTGGTACAGAGGACCACGCCCCGGC -3'.

The DNA binding reaction buffer contained 1x binding buffer, 1x DTT/Tw20, 1µg poly(dI-dC), 0.05% NP-40 (LI-COR EMSA buffer kit), and 1mM zinc acetate. Binding reactions contained 435 ng of purified HIC1 protein. 50fmol fluorescent oligo DNAs were then added to the appropriate protein/binding mix and incubated for 20 min at room temperature. For supershift assays, 1 ug per lane of mouse anti-DDK (FLAG) monoclonal antibody (ORIGENE #TA50011-100) was incubated with the binding buffer for 20 min prior to addition of and incubation with oligo DNA. 1x orange loading dye (LI-COR kit) was added to samples, which were then resolved on (pre-cast, pre-run at 100V for 60min) 6% TBE gels (Novex,13 ThermoFisher) in 0.5x TBE buffer for 120 min at 80V (4C). Fluorescent bands were then imaged using a LI-COR chemiluminescent imaging system. EMSA experiments display representative panels of 2-3 replicates. Densitometric analysis for HIC1 band intensity was performed using a Licor Odyssey scanner.

Estimating fetal and maternal genetic effects

For the 2q13 locus, we analyzed up to 15,536 mother-child pairs from seven studies with both fetal and maternal genotypes available (ALSPAC, BiB, DNBC, EFSOCH, FIN, MoBa_2008, and MoBa_HARVEST). We used linear regression to test the association between quantile transformed gestational duration (same transformation as in the main analysis) and fetal genotype conditional on maternal genotype and vice versa. In the same complete mother-child pairs (i.e. where genotype data were available for both mother and child), we estimated unconditional effects of fetal and maternal genotype, respectively. We combined the results from the individual studies using fixed-effects meta-analysis.

Fraction of variance explained

To estimate the fraction of variance in gestational duration explained by the lead variant rs7594852 at the 2q13 locus we fitted a linear regression model of quantile transformed gestational duration in the iPSYCH cohort (n = 51,357) where the variant was genome-wide significant. The model was corrected for iPSYCH disease group and the fraction of variance explained by rs7594852 genotype dosage was extracted. The

estimate of variance explained by all common autosomal SNPs (MAF > 1%) was calculated based on the discovery stage meta-analysis results using LD Score regression⁶².

Acknowledgements

For study-specific acknowledgements, please see **Supplementary Material**.

Funding

B.F. received support from an Oak Foundation fellowship, a Novo Nordisk Foundation grant (12955), and a Bill and Melinda Gates Foundation subward (137097); X.L. received support from the Nordic Center of Excellence in Health-Related e-Sciences; D.H., V.A., A.S., R.N., T.W. and A.B.D. received funding from the Lundbeck Foundation (R102-A9118, R155-2014-1724, R248-2017-2003); L.S. reports funding from a Carlsberg Foundation postdoctoral fellowship (CF15-0899); R.M.F. is a Sir Henry Dale Fellow (Wellcome and Royal Society grant: WT104150); R.N.B. is funded by Wellcome and Royal Society (grant: WT104150); M.C.B. and D.A.L. work in a unit that receives UK MRC funding (MC_UU_00011/6); M.C.B. is supported by the MRC Skills Development Fellowship MR/P014054/1; D.A.L.'s contribution to this work was funded by grants from the US NIH and European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement (Grant number 669545; DevelopObese) and European Union's Horizon 2020 research and innovation programme under grant agreement 733206 (LIFECYCLE); D.A.L. is also an NIHR senior investigator (NF-SI-0611-10,196); F.R. received partial funding from the Netherlands Organization for Health Research and Development (VIDI 016.136.367); J.F.F. and V.W.V.J. received partial funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 733206 (LIFECYCLE) and 633595 (DynaHEALTH); V.W.V.J. received partial funding from the Netherlands Organization for Health Research and Development (VIDI 016.136.361) and the European Research Council (ERC Consolidator Grant, ERC-2014-CoG-648916); I.K. reports funding from Sigrid Juselius Foundation and Biomendicum Foundation postdoctoral fellowship; L.J.M. was supported by the March of Dimes Prematurity Research Center Ohio Collaborative, NICHD HD 091527, and the Bill and Melinda Gates Foundation (OPP1113966); M.T.W. was supported by NIH R01 NS099068, NIH R01 GM055479-19A1, a Lupus Research Alliance "Novel Approaches" award, CCRF Endowed Scholar, a CCHMC CpG Pilot study award, and a CCHMC Trustee award; K.K.R. received support from March of Dimes (21-FY13-19); C.P. was supported by the NIHR Great Ormond Street Hospital Biomedical Research Centre; M.I.M. is a Wellcome Senior Investigator and a NIHR Senior Investigator. His work is supported by Wellcome (090532, 093831, 203141, 106130) and by the NIH (U01DK105535). The views expressed in this article are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health. For study-specific funding, please see **Supplementary Material**.

Disclosures

D.A.L received support from Roche Diagnostics and Medtronic for biomarker research unrelated to the work presented in this paper. J.C.M. is a part time employee of the Bill and Melinda Gates Foundation.

Author contributions

Statistical analysis:

X.L., D.H., L.S., R.N.B., M.W., F.G., J.J., A.M., J.P.B., F.T.J.L., S.V., T.S.A., N.P., C.A.W., M.C.B., G.Z., I.K., M.G.H., D.M.S., J.F., R.N., L.-P.L., J.F.F., E.H., R.M.F., A.B., B.F.

Study design:

D.A.L., C.P., M.I.M., H.A.B., M.L.M., H.H., V.W.V.J., R.K.V., H.B., K.P., O.R., C.E.P., K.B., J.F.F., S.F.A.G., E.H., T.M.W., M.M., A.B., B.F.

Sample collection:

D.A.L., C.P., H.A.B., H.H., V.W.V.J., R.K.V., H.B., B.A.K., K.P., O.R., D.M.H., C.E.P., K.B., M.V., J.F.F., W.L.L., S.F.A.G., E.H., M.-R.J., J.C.M., M.M., B.F.

Genotyping:

X.L., F.G., J.J., A.M., M.B., J.B., B.A.B., W.K.T., V.A., D.A.L., M.I.M., M.L.M., H.H., M.G.H., F.R., H.B., K.P., O.R., Ø.H., S.J., P.R.N., A.S., P.B.M., A.D.B., M.N., O.M., K.K.R., D.M.H., C.E.P., K.B., W.L.L., S.F.A.G., B.J., M.-R.J., L.J.M., J.C.M., R.M.F., T.M.W., A.B.

Functional follow-up experiments:

R.M.R., K.S., S.P., D.E.M., X.C., M.T.W., K.H., D.M.H., L.C.K., L.J.M.

Writing and overall study direction:

X.L., D.H., L.S., J.C.M., L.J.M., R.M.F., T.M.W., M.M., A.B., B.F.

All authors reviewed and edited the manuscript.