

1 **Genetic association study of Preterm birth and Gestational age in a**  
2 **population-based case-control study in Peru**

3  
4 *Genetics of PTB and GA in pregnant women in Peru*

5  
6 Diana L. Juvinao-Quintero<sup>1\*</sup>, Sixto E. Sanchez<sup>2,3</sup>, Tsegaselassie Workalemahu<sup>4</sup>, Nelida Pinto<sup>3</sup>,  
7 Liming Liang<sup>1,5</sup>, Michelle A. Williams<sup>1</sup> and Bizu Gelaye<sup>1,6</sup>

8  
9 Affiliations:

10  
11 <sup>1</sup> Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA.

12 <sup>2</sup> Facultad de Medicina Humana, Instituto de Investigación, Universidad de San Martín de  
13 Porres, Lima, Peru.

14 <sup>3</sup> Asociación Civil PROESA, Lima, Peru.

15 <sup>4</sup> Department of Obstetrics and Gynecology, University of Utah Health, Salt Lake City, UT, USA.

16 <sup>5</sup> Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA.

17 <sup>6</sup> The Chester M. Pierce, M.D. Division of Global Psychiatry, Massachusetts General Hospital,  
18 and Harvard Medical School, Boston, MA, USA.

19  
20 \*Corresponding Author:

21 Email: [djq@hsph.harvard.edu](mailto:djq@hsph.harvard.edu) (DLJQ)

## 22 ABSTRACT

23 Preterm birth (PTB) is an adverse pregnancy outcome affecting ~15 million pregnancies  
24 worldwide. Genetic studies have identified several candidate loci for PTB, but results remain  
25 inconclusive and limited to European populations. Thus, we conducted a genome-wide  
26 association study (GWAS) of PTB and gestational age at delivery (GA) among 2,212 Peruvian  
27 women. PTB cases delivered  $\geq 20$  weeks' but  $< 37$  weeks' gestation, while controls delivered at  
28 term ( $\geq 37$  weeks but  $< 42$  weeks). After imputation (TOPMED) and quality control, we assessed  
29 the association of ~6 million SNPs with PTB and GA using multivariable regression models  
30 adjusted for maternal age and the first two genetic principal components. *In silico* functional  
31 analysis (FUMA-GWAS) was conducted among top signals detected with an arbitrary  $P <$   
32  $1.0 \times 10^{-5}$  in each GWAS. We sought to replicate genetic associations with PTB and GA identified  
33 in Europeans, and we developed a genetic risk score for GA based on European markers.  
34 Mean GA was  $30 \pm 4$  weeks in PTB cases (N=933) and  $39 \pm 1$  in the controls (N=1,279). PTB  
35 cases were slightly older and had higher C-sections and vaginal bleeding than controls. No  
36 association was identified at genome-wide level. Top suggestive ( $P < 1.0 \times 10^{-5}$ ) signals were  
37 seen at rs13151645 (*LINC01182*) for PTB, and at rs72824565 (*CTNNA2*) for GA. Top PTB  
38 variants were enriched for biological pathways associated with polyketide, progesterone, steroid  
39 hormones, and glycosyl metabolism. Top GA variants were enriched in intronic regions and  
40 cancer pathways, and these genes were upregulated in the brain and subcutaneous adipose  
41 tissue. In combination with non-genetic risk factors, top SNPs explained 14% and 15% of the  
42 phenotypic variance of PTB and GA in our sample, but these results need to be interpreted with  
43 caution. Variants in *WNT4* associated with GA in Europeans were replicated in our study. The  
44 genetic risk score based in European markers, was associated with a 2-day longer GA  
45 ( $R^2=0.003$ ,  $P=0.002$ ) per standard deviation increase in the score in our sample. This genetic  
46 association study identified various signals suggestively associated with PTB and GA in a non-  
47 European population; they were linked to relevant biological pathways related to the metabolism  
48 of progesterone, prostanoid, and steroid hormones, and genes associated with GA were  
49 significantly upregulated in relevant tissues for the pathophysiology of PTB based on the *in-*  
50 *silico* functional analysis. None of these top variants overlapped with signals previously  
51 identified for PTB or GA in Europeans.

52

53 **Keywords:** preterm birth, gestational age at delivery, GWAS, genetic risk score, pregnancy,  
54 maternal genotype.

55

56

57

## 58 INTRODUCTION

59 Preterm birth (PTB)—the premature onset of labor before 37 weeks of gestation [1-3]—affects  
60 10% (~ 15 million) of newborns globally [1, 4] with the majority of them (80%) occurring in the  
61 Sub-Saharan and South Asia region [1]. PTB is one of the leading causes of neonatal morbidity  
62 and mortality, and the leading cause of infant mortality before 5 years [5]. Preterm infants who  
63 survive are at higher risk of short-term and long-term disability; they are more prone to suffer  
64 from respiratory diseases [6, 7], necrotizing enterocolitis [8], neurodevelopmental impairment [9,  
65 10], hypertension [11, 12], and glucose intolerance [13]. The pathogenesis of PTB remains  
66 largely unknown but is recognized as a multifactorial syndrome [14], influenced by multiple  
67 inflammatory-driven processes like microbial-induced inflammation, decidual hemorrhage,  
68 vascular disease, and disruption of the maternal-fetal immune tolerance, among others [2, 14].  
69 Common PTB risk factors include maternal sociodemographic (ethnicity, low socioeconomic  
70 status, education, marital status, working conditions, and age) and behavioral characteristics  
71 (psychosocial distress, depression, substance abuse, smoking) [2, 15]; pregnancy history  
72 (previous PTB, preeclampsia, gestational diabetes), present pregnancy status (multigravida,  
73 intrauterine infection, uteroplacental ischemia or hemorrhage, uterine overdistention, cervical  
74 length, other comorbidities), and biological factors (nutritional status, BMI, inflammatory  
75 markers, fetal fibronectin) and genetic markers [2].

76

77 Twin studies have demonstrated the genetic contribution to PTB, with a reported heritability of  
78 ~27% to 36% relative to maternal genetic effects [15], while evidence from genome-wide  
79 association studies (GWAS) have revealed several single nucleotide polymorphisms or SNPs  
80 with small effects on PTB and gestational duration [15-17]. Some of the challenges in studying  
81 the genetics of PTB are their reliance on both, the maternal and fetal genetic effects [18], the  
82 heterogeneity in the clinical definition of PTB, measurement error in the assessment of  
83 gestational age [18], and overall differences in the genetic risk observed between populations

84 [15]. These challenges have limited the possibility of replicating signals across studies. GWAS  
85 in European ancestry individuals have identified six loci associated with gestational duration  
86 (*EBF1*, *EEFSEC*, *AGTR2*, *WNT4*, *ADCY5*, and *RAP2C*), four of which were also associated  
87 with PTB (*EBF1*, *EEFSEC*, and *AGTR2*) based on maternal genetic effects. These loci were  
88 linked to relevant biological pathways of PTB, including uterine development, maternal nutrition,  
89 and vascular control [16]. Using the fetal genotype, one variant in the *SLIT2* gene was  
90 associated with PTB in a Finnish population [19]. The *SLIT2* gene and its receptor *ROBO1* were  
91 upregulated in the placenta of PTB infants and participated in the regulation of genes involved in  
92 inflammation, decidualization, and fetal growth [19]. A recent GWAS using the maternal  
93 genotype from UK samples identified several SNPs and gene transcripts associated with  
94 spontaneous PTB, like the SNP rs14675645 (*ASTN1*), and transcripts from the microRNA-142  
95 and *PPARG1-FOXP3* gene associated with PPRM [20]. Overall, these markers were linked to  
96 inflammation and immune response pathways [20]. In a trans-ethnic GWAS using the fetal  
97 genotype [15], two intergenic SNPs were identified in association with extreme PTB cases (< 30  
98 weeks of gestation) in populations from African (chr 1, rs17591250) and American (chr 8,  
99 rs1979081) ancestries, which differed from SNPs previously identified in Europeans, and could  
100 not be replicated in additional cohorts [15]. Lastly, GWAS investigating gene-environment  
101 interactions in the risk of PTB have identified SNPs in *PTPRD* and *COL24A1* where the  
102 genotype interacts with maternal lifetime stress and pre-pregnancy BMI, respectively, in  
103 increasing the risk of spontaneous PTB in African Americans [21, 22].

104

105 So far, genetic evidence points towards a population-specific genetic risk for PTB and  
106 gestational age [15, 17, 18], which may be reflective of the interaction between the genotype  
107 and specific environmental risk factors. However, most genetic studies have focused on  
108 European ancestry populations, with a limited representation of admixture populations like those  
109 in the Americas [15], limiting our ability to identify the mechanisms associated with PTB in

110 minority populations at increased risk. Thus, we conducted the present study to identify  
111 maternal genetic variants associated with PTB and gestational age at delivery (GA) in a case-  
112 control study from Lima, Peru. We hypothesize that unique markers associated with PTB and  
113 GA will be identified, reflecting distinct mechanistic pathways to PTB in this admixture  
114 population.

115

## 116 RESULTS

### 117 *Population Characteristics*

118 As shown in Table 1 women were, on average, 28 years of age, with most of them being  
119 multiparous, and had normal pre-pregnancy BMI, and low educational attainment. Only a small  
120 proportion of these mothers reported smoking tobacco or using illicit drugs during pregnancy (<  
121 2%), while 17% of them reported using alcohol. Compared to controls, PTB cases were older,  
122 more likely to deliver through C-section, and had poorer self-reported health in pregnancy. PTB  
123 cases were also more likely to be preeclamptic, report vaginal bleeding during pregnancy, and  
124 deliver infants with lower birthweight compared to control mothers. Additional characteristics of  
125 the study participants are provided in S1 Table.

126

127 **Table 1. Sociodemographic characteristics of study participants in Lima, Peru (N= 2,212).**

	Overall Sample (N= 2,212) n (%)	Preterm- Births (N= 933) n (%)	Full-term births (N= 1,279) n (%)	P*
Maternal age (years), Mean (SD)	28 (7)	29 (7)	28 (7)	0.01
Maternal age by groups				
18-19	211 (10)	74 (8)	137 (11)	0.03
20-29	1,095 (50)	462 (50)	633 (50)	
30-34	444 (20)	181 (20)	263 (21)	
≥ 35	438 (20)	207 (22)	231 (18)	
Gestational age (weeks), Mean (SD)	35 (5)	30 (4)	39 (1)	<0.001
PPROM, yes (%)	515 (24)	379 (42)	136 (11)	<0.001
Primiparous, yes (%)	680 (31)	273 (29)	407 (32)	0.20

Mode of delivery				
Natural	822 (37)	287 (31)	535 (42)	<0.001
C-section	1385 (63)	644 (69)	741 (58)	
Maternal Education				
≤ high school	1,557 (71)	634 (69)	923 (73)	0.036
> high school	636 (29)	290 (31)	346 (27)	
Marital status				
Married or living with partner	1,896 (86)	786 (85)	1,110 (87)	0.11
Not married or living alone	304 (14)	141 (15)	163 (13)	
Employed during pregnancy				
No	999 (45)	414 (45)	585 (46)	0.60
Yes	1,198 (55)	510 (55)	688 (54)	
Pre-pregnancy BMI (kg/m <sup>2</sup> ), Mean (SD)	25 (5)	25 (5)	25 (5)	0.04
Pre-pregnancy BMI (categories)				
Underweight (<18.5)	53 (2.5)	29 (3.2)	24 (1.9)	0.12
Normal weight (18.5–24.9)	1,196 (56)	512 (57)	684 (55)	
Overweight (25 – 29.9)	614 (29)	241 (27)	373 (30)	
Obese (≥ 30)	275 (13)	118 (13)	157 (13)	
Planned pregnancy, yes (%)	733 (33)	291 (31)	442 (35)	0.08
Health perception, Fair/poor (%)	1,230 (56)	603 (65)	627 (50)	<0.001
Alcohol use, yes (%)	370 (17)	168 (18)	202 (16)	0.20
Tobacco smoking, yes (%)	25 (1)	12 (1)	13 (1)	0.60
Drug abuse, yes (%)	4 (0.2)	0 (0)	4 (0.3)	0.14
Preeclampsia, yes (%)	180 (8.2)	90 (9.7)	90 (7.1)	0.03
Vaginal bleeding, yes (%)	432 (20)	273 (29)	159 (12)	<0.001
Spontaneous Abortions, yes (%)	616 (28)	277 (30)	339 (27)	0.10
Infant birthweight (g), Mean (SD)	2,664 (1,169)	1,656 (670)	3,340 (867)	<0.001
Child sex				
Female	993 (46)	397 (43)	596 (47)	0.08
Male	1,189 (54)	519 (57)	670 (53)	

128 SD: standard deviation. PROM, Preterm premature rupture of membranes. \*Wilcoxon rank sum test;  
 129 Pearson's Chi-squared test; Fisher's exact test.

130

### 131 **Ancestry Assessment**

132 The PC analysis using combined genetic data from PAGE and the 1000 Genomes allowed us to  
 133 determine the ancestry of PAGE participants, who generally clustered close to PEL samples  
 134 (Peruvians from Lima, Peru), and a few others were dispersed among samples from other  
 135 American regions (CLM Colombians, MXL Mexican American) (Fig 1). Using cleaned, directly  
 136 genotyped genetic data from PAGE alone, we computed genetic PCs to use as covariates in  
 137 genetic association analyses. We extracted the first 2 PCs explaining 56.5% and 9.65% of the

138 total variation in the sample. Some dispersion was observed in the distribution of PAGE  
139 samples across these two genetic PCs.

140

141 **Fig 1. Genetic ancestry analysis showing the clustering of PAGE samples relative to**  
142 **global populations from different ancestries included in the 1000 Genomes project.**

143

#### 144 ***GWAS of PTB and GA***

145 We provided full GWAS results to allow access to non-European-ancestry summary statistics  
146 and facilitate future trans-ethnic GWAS meta-analyses of PTB and GA. As shown in Fig 2, we  
147 did not identify genetic associations with PTB or GA at genome-wide significance in our dataset.  
148 This result was somehow expected due to the appropriately QC'd but underpowered genetic  
149 dataset. When examining QQ plots, there was no indication of bias due to genomic inflation in  
150 the GWAS based on  $\lambda = 1.0$  (Fig 2). We reported markers with suggestive association with PTB  
151 and GA at  $P < 1.0 \times 10^{-5}$ . In total, 8 SNPs were identified suggestively associated with PTB, most  
152 of them related with a lower risk of PTB (i.e., longer gestational age at delivery) per additional  
153 risk allele, except for two markers mapping to the *KDM4C* and *SOX5* loci. The strongest signal  
154 was observed in rs13151645 mapping to the non-coding RNA *LINC01182*. This SNP was  
155 associated with a 26% (95%CI= 0.65, 0.84) lower risk of PTB per additional risk allele, which  
156 was consistent with the association observed between rs13151645 and GA ( $\beta = 0.6$  weeks,  
157 95%CI= 0.30, 0.91) (Table 2). Other SNPs in strong correlation ( $r^2 > 0.6$ ) with rs13151645 but  
158 with larger GWAS  $P$ -values, were seen in the nearby region (Fig 3), suggesting that *LINC01182*  
159 could be a good candidate locus for PTB study. Additionally, we observed a similar direction of  
160 association with GA at top signals detected in the GWAS of PTB, even though  $P$ -values were  
161 larger (i.e., less statistically significant).

162



163 **Fig 2. Quantile and Manhattan plots summarizing results of the GWAS of Preterm Birth**  
 164 **(A, B) and Gestational age at delivery (C, D) conducted among women in the PAGE study**  
 165 **(N=2,212).**

166

167 **Table 2.** Top genetic associations detected in the GWAS of preterm birth (PTB) among women  
 168 in the PAGE study (N=2,212). Suggestive associations were identified with  $P < 1.0 \times 10^{-5}$ .  
 169

Gene and SNP	Ch r	Alleles (A1/A0 )	Fre q	$r^2$	GWAS of PTB		Look up in GWAS of GA	
					OR (95% CI)	P	$\beta$ (95% CI)	P
<i>MIR3681HG</i>								
rs13401913	2	T/C	0.06	---	0.52 (0.39,0.70)	$8.5 \times 10^{-6}$	1.32 (0.69,1.95)	$4.2 \times 10^{-5}$
rs4849108	2	G/A	0.28	< 0.1	0.74 (0.64,0.84)	$1.1 \times 10^{-5}$	0.56 (0.24,0.88)	$5.8 \times 10^{-4}$
<i>LINC01182</i>								
<b>rs13151645</b>	<b>4</b>	<b>T/T</b>	<b>0.65</b>	---	0.74 (0.65,0.84)	<b><math>3.9 \times 10^{-6}</math></b>	0.60 (0.30,0.91)	$1.0 \times 10^{-4}$
rs2058581	4	T/T	0.64	0.88	0.74 (0.65,0.85)	$6.6 \times 10^{-6}$	0.58 (0.28,0.88)	$1.8 \times 10^{-4}$
<i>KDM4C</i>								
rs148983777	9	C/G	0.07	---	1.69 (1.34,2.13)	$9.1 \times 10^{-6}$	-1.21 (-1.76,- 0.65)	$1.9 \times 10^{-5}$
<i>RP11-45P17.5: AKR1C3</i>								
rs7098485	10	G/G	0.51	---	0.76 (0.67,0.86)	$5.4 \times 10^{-6}$	0.56 (0.27,0.85)	$9.9 \times 10^{-5}$
rs2050359	10	A/A	0.51	0.97	0.75 (0.67,0.85)	$9.1 \times 10^{-6}$	0.58 (0.29,0.87)	$1.8 \times 10^{-4}$
<i>SOX5</i>								
rs17408900	12	C/T	0.17	---	1.44 (1.23,1.68)	$6.3 \times 10^{-6}$	-0.70 (-1.08,- 0.33)	$2.6 \times 10^{-4}$

170 In bold, SNP identified with the smallest P-value of association in the GWAS of PTB. Genome-wide  
 171 significant signals if  $P < 5.0 \times 10^{-8}$ . Associations were adjusted for maternal age (years) and the first two  
 172 genetic PCs. SNP coordinates are based on the GRCh37hg19 genome build. A0: reference allele, A1:  
 173 effect allele, Freq: frequency of the risk allele,  $r^2$ : linkage disequilibrium (LD) showing the correlation  
 174 between SNPs within the same genetic region. Pairwise LD correlations were calculated using the 1,000  
 175 genomes (phase 3) and the mixed American population as reference. Effect size is interpreted as the  
 176 odds of PTB per additional risk allele. In the look-up of top signals of PTB in the GWAS of gestational age  
 177 at delivery, the effect size is the estimated change in the number of gestational weeks, per additional risk  
 178 alleles.  
 179



180 In the GWAS of GA (Table 3), 20 SNPs mapping to 12 genes across the genome were  
 181 borderline associated ( $P < 1.0 \times 10^{-5}$ ). For most of these markers, the risk allele investigated was  
 182 associated with shorter GA at delivery, as it was evident for associations in the *DNAJC6*,  
 183 *CTNNA2* and *GUCY1A2* loci. Positive associations with GA (i.e., longer gestational length) were  
 184 identified in the *ALPK1:RTEL1P1*, *ARNT2*, and *MACROD2* genes. The marker most strongly  
 185 associated with GA was detected in the intronic variant rs72824565 mapping within the  
 186 *CTNNA2* gene ( $\beta = -3.48$  weeks, 95%CI= -4.78, -2.19). Few other signals in correlation with  
 187 rs72824565 were seen in the region of *CTNNA2* (Fig 3). Top signals were nominally associated  
 188 with shorter gestation, directionally consistent with their effect sizes in PTB GWAS.

189

190 **Table 3.** Top genetic associations detected in the GWAS of gestational age at delivery (GA,  
 191 continuous) among women in the PAGE study (N=2,212). Suggestive associations were  
 192 identified at  $P < 1.0 \times 10^{-5}$ .

193

Gene and SNP	Chr	Alleles (A1/A0)	Freq	$r^2$	GWAS of GA		Look up in GWAS of PTB	
					$\beta$ (95% CI)	P	OR (95% CI)	P
<i>OR2L13</i> chr1:247983582	1	GTCC/G	0.02		-2.90 (-4.02,-1.77)	$4.7 \times 10^{-7}$	2.54 (1.57,4.11)	$1.5 \times 10^{-4}$
<i>DNAJC6</i> chr1:65376701	1	G/GT	0.16	< 0.1	-0.91 (-1.31,-0.52)	$6.9 \times 10^{-6}$	1.34 (1.13,1.58)	$5.5 \times 10^{-4}$
<i>CTNNA2</i> <b>rs72824565</b>	<b>2</b>	<b>G/T</b>	<b>0.01</b>	---	-3.48 (-4.78,-2.19)	<b><math>1.4 \times 10^{-7}</math></b>	2.55 (1.45,4.49)	$1.2 \times 10^{-3}$
rs72824567	2	G/A	0.01	1.0	-3.36 (-4.63,-2.09)	$2.5 \times 10^{-7}$	2.51 (1.44,4.37)	$1.2 \times 10^{-3}$
rs112058880	2	G/A	0.01	0.8	-3.36 (-4.63,-2.09)	$2.5 \times 10^{-7}$	2.51 (1.44,4.37)	$1.2 \times 10^{-3}$
rs115416830	2	A/G	0.02	< 0.1	-3.11 (-4.36,-1.86)	$1.1 \times 10^{-6}$	2.13 (1.25,3.64)	$5.3 \times 10^{-3}$
<i>MTHFD2P1</i> rs28440168	3	T/C	0.02		-2.72 (-3.90,-1.54)	$6.2 \times 10^{-6}$	2.24 (1.36,3.68)	$1.5 \times 10^{-3}$
<i>ALPK1:RTEL1P1</i> rs4834267	4	C/T	0.61	---	0.67 (0.38,0.96)	$7.4 \times 10^{-6}$	0.78 (0.69,0.89)	$1.0 \times 10^{-4}$
rs4833403	4	C/T	0.68	< 0.1	0.70 (0.39,1.00)	$9.8 \times 10^{-6}$	0.75 (0.66,0.85)	$1.6 \times 10^{-5}$
<i>SPOCK1</i> rs73294123	5	G/C	0.07		-1.38 (-1.92,-0.83)	$7.7 \times 10^{-7}$	1.48 (1.18,1.86)	$6.3 \times 10^{-4}$
<i>TPK1</i> rs717885	7	A/C	0.02		-2.57 (-3.66,-1.48)	$3.9 \times 10^{-6}$	2.06 (1.30,3.27)	$2.0 \times 10^{-3}$
<i>GLUD1</i> chr10:87071607	10	C/CAG	0.19		-0.83 (-1.19,-0.46)	$8.3 \times 10^{-6}$	1.33 (1.14,1.55)	$2.2 \times 10^{-4}$
<i>GUCY1A2</i>								

rs6591172	11	T/C	0.03	---	-2.34 (-3.23,-1.45)	$2.7 \times 10^{-7}$	2.01 (1.38,2.91)	$2.6 \times 10^{-4}$
rs6591171	11	A/G	0.03	1.0	-2.34 (-3.23,-1.45)	$2.7 \times 10^{-7}$	2.01 (1.38,2.91)	$2.6 \times 10^{-4}$
rs12290638	11	G/T	0.03	0.9	-2.30 (-3.19,-1.42)	$3.6 \times 10^{-7}$	2.00 (1.38,2.91)	$2.5 \times 10^{-4}$
rs12271785	11	A/G	0.02	< 0.1	-2.54 (-3.53,-1.55)	$4.9 \times 10^{-7}$	2.21 (1.46,3.36)	$2.0 \times 10^{-4}$
rs7102076	11	G/A	0.02	< 0.1	-2.55 (-3.59,-1.51)	$1.6 \times 10^{-6}$	2.17 (1.40,3.36)	$5.6 \times 10^{-4}$
<i>DCLK1</i>								
rs56135847	13	A/G	0.33		-0.70 (-0.99,-0.40)	$3.4 \times 10^{-6}$	1.24 (1.10,1.40)	$6.3 \times 10^{-4}$
<i>ARNT2</i>								
rs17225178	15	A/T	0.02		2.38 (1.38,3.37)	$3.2 \times 10^{-6}$	0.33 (0.20,0.54)	$1.3 \times 10^{-5}$
<i>MACROD2</i>								
chr20:15143602	20	A/AT	0.04		1.79 (1.06,2.53)	$1.8 \times 10^{-6}$	0.55 (0.40,0.77)	$5.1 \times 10^{-4}$

194 In bold, SNP identified with the smallest P-value of association with GA. Genome-wide significant signals  
 195 if  $P < 5.0 \times 10^{-8}$ . Associations were adjusted for maternal age (years) and the first two genetic PCs. A0:  
 196 reference allele, A1: effect allele, Freq: frequency of the risk allele,  $r^2$ : linkage disequilibrium (LD) showing  
 197 the correlation between SNPs within the same genetic region. Pairwise LD correlations were calculated  
 198 using the 1,000 genomes (phase 3) and the mixed American population as reference. Effect size  
 199 corresponds to the change in gestational weeks per additional risk allele. In the look-up of signals of GA  
 200 in the GWAS of PTB, the effect size is interpreted as the odds of PTB per additional risk allele.  
 201

202 **Fig 3. Regional plots showing the genetic context of top signals detected with suggestive**  
 203 **association with Preterm Birth (rs13151645 in chr4) and Gestational age at delivery**  
 204 **(rs72824565 in chr2).**

205  
 206 Top independent signals identified in the GWAS in PAGE were looked up in a large meta-  
 207 analysis of GWAS of PTB and GA conducted in Europeans [17]. We observed no replication of  
 208 our signals in the GWAS in Europeans using a nominally significant GWAS  $P < 0.05$ . For top  
 209 signals in PAGE identified in the GWAS in Europeans (7 and 16 top SNPs for PTB and GA,  
 210 respectively), only a few of them (3/7 PTB and 1/16 GA SNPs) showed same direction of  
 211 association between studies.

212  
 213 Using the Nagelkerke's pseudo- $R^2$  and the adjusted- $R^2$  obtained from multivariable logistic and  
 214 linear regressions, respectively, we assessed the percent of the total variance in PTB and GA  
 215 explained by top SNPs detected with the smallest  $P$ -value in each GWAS. Combining the effect  
 216 of the top 8 PTB-associated SNPs, they explained 7.2% of the variance in the risk of PTB in

217 PAGE samples. This value was slightly larger than the variance explained by non-genetic risk  
218 factors (i.e., maternal age, gravidity, pre-pregnancy BMI, vaginal bleeding, maternal education,  
219 child sex, and two genetic PCs) (adjusted- $R^2= 6.9\%$ ) (S2 Table). Likewise, combining the effect  
220 of the top 20 GA-associated SNPs, they explained 10% of the GA variance, almost double the  
221 variance captured by non-genetic risk factors (adjusted- $R^2= 5.5\%$ ). Adding the effect of non-  
222 genetic and genetic risk factors, we captured 14% and 15.3% of the total variance in PTB and  
223 GA in our sample, respectively.

224

### 225 ***Replication of European Markers and Genetic Score***

226 We extracted summary statistics for 15 unique SNPs previously identified in association with  
227 PTB and/or GA in a meta-analysis of GWAS conducted among Europeans [16]. Out of the 11  
228 GA SNPs reported by Zhang *et al.*, seven were observed in our dataset, but only three of them  
229 had the same direction of association with GA as in the reference study. True replication was  
230 only seen for two GA SNPs mapping to the *WNT4* gene ( $P=1.0\times 10^{-3}$ ) (Table 4). Similarly, out of  
231 the 6 SNPs previously reported in association with PTB, only two were identified in our dataset,  
232 both with the same direction of association with PTB as in the reference study, but with  
233 replication  $P > 0.03$  (significant if  $P < 0.05/2$  SNPs) (Table 4). We searched for proxies for SNPs  
234 in the reference study missing in our GWAS, but the two proxies (LD  $R^2 > 0.7$ ) in rs5950498 and  
235 rs5991030 identified for target SNPs rs5950506 (PTB) and rs5950491 (GAD) in Chr X, were  
236 also not included in our GWAS. Additionally, for each participant in PAGE, we constructed a  
237 standardized GRS for GA based on five independent SNPs previously identified by Zhang *et al.*  
238 (S3 Fig). We showed that an SD increase in the GRS was associated with an average 0.29-  
239 week (~ 2 days) (95%CI= 0.08, 0.49 weeks) increase in GA (S3 Table). The GRS was able to  
240 capture 0.3% of the variance in GA in our sample, and in combination with other non-genetic

241 risk factors (maternal age, parity, pre-pregnancy BMI, vaginal bleeding, maternal education,  
242 child sex, PC1, and PC2), it explained up to 6.0% of the variance in GA in our dataset.

243

244

245

246

247 **Table 4.** Look-up of risk loci for preterm birth and gestational duration identified in Europeans  
248 (Zhang *et al.* 2017), in the GWAS conducted in PAGE. SNP associations were replicated if a  
249 similar direction of association was observed between the reference and the observed GWAS  
250 SNP at a nominal GWAS *P*-value < 0.05.

251

Reference GWAS in Europeans (Zhang et al. 2017)						GWAS in PAGE		
Gene	SNP	Alleles (A0/A1)	Chr: Position	$\beta$	P	A1	$\beta$	P
<i>Association with GA</i>								
<i>EBF1</i>	rs2963463	C/T	5:157895049	-0.16	$7.7 \times 10^{-24}$	T	0.01	$9.6 \times 10^{-01}$
	rs2946171	T/G	5:157921940	-0.21	$8.1 \times 10^{-21}$	G	-0.02	$9.3 \times 10^{-01}$
<i>EEFSEC</i>	rs2955117	G/A	3:127881613	0.19	$9.5 \times 10^{-15}$	A	-0.04	$8.1 \times 10^{-01}$
	rs200745338	D/I	3:127869457	0.27	$7.5 \times 10^{-16}$	-	-	-
<i>AGTR2</i>	rs201226733	I/D	23:115164770	-0.24	$7.2 \times 10^{-16}$	-	-	-
	rs5950491	C/A	23:115129714	-0.25	$6.6 \times 10^{-16}$	-	-	-
<b><i>WNT4</i></b>	<b>rs56318008</b>	<b>C/T</b>	<b>1:22470407</b>	0.32	<b><math>3.4 \times 10^{-14}</math></b>	<b>T</b>	<b>0.48</b>	<b><math>1.1 \times 10^{-03}</math></b>
	<b>rs12037376</b>	<b>G/A</b>	<b>1:22462111</b>	0.34	<b><math>5.6 \times 10^{-14}</math></b>	<b>A</b>	<b>0.48</b>	<b><math>1.1 \times 10^{-03}</math></b>
<i>ADCY5</i>	rs4383453	G/A	3:123068359	-0.08	$3.7 \times 10^{-08}$	A	0.20	$3.6 \times 10^{-01}$
	rs9861425	A/C	3:123072883	-0.20	$4.2 \times 10^{-10}$	A	0.20	$2.4 \times 10^{-01}$
<i>RAP2C</i>	rs200879388	I/D	23:13130057	-0.16	$3.4 \times 10^{-09}$			
<i>Association with PTB</i>								
<i>EBF1</i>	rs2963463	C/T	5:157895049	1.13	$4.5 \times 10^{-15}$	T	1.00	$9.8 \times 10^{-01}$
	rs2946169	C/T	5:157918959	1.16	$2.2 \times 10^{-13}$	T	1.01	$9.0 \times 10^{-01}$
<i>EEFSEC</i>	rs201450565	D/I	3:128058610	0.82	$1.9 \times 10^{-12}$	-	-	-
	rs200745338	D/I	3:127869457	0.80	$3.3 \times 10^{-14}$	-	-	-
<i>AGTR2</i>	rs201386833	D/I	23:115164281	1.18	$1.0 \times 10^{-11}$	-	-	-
	rs5950506	G/A	23:115175748	1.18	$1.1 \times 10^{-11}$	-	-	-

252 In bold, associations of SNPs in *WNT4* with GA that were replicated in the GWAS in PAGE at nominal *P*  
253 < 0.05. The effect estimate for GA is interpreted as the change in gestational weeks, per additional risk  
254 allele; positive values indicate longer gestational age at delivery associated with the SNP. For PTB, effect  
255 estimates are interpreted as the odds of PTB, per additional risk allele; values larger than 1 indicate  
256 increased odds of PTB associated with the SNP. SNPs in Zhang et al. without data in the GWAS in  
257 PAGE were filtered out during genetic QC.

258

### 259 ***In-silico Functional Analysis***

260 Based on the FUMA-GWAS analysis, five SNPs were identified as independent significant  
261 signals for PTB at  $P < 1.0 \times 10^{-5}$ ; these signals were enriched in intronic and in non-coding RNAs  
262 within intronic regions (Fisher  $P < 0.01$ ) (S1 Fig). Using the full distribution of GWAS  $P$ -values ( $P$   
263  $< 0.05$ ), the MAGMA gene-set analysis identified a positive association of input PTB SNPs with  
264 GO terms associated with neuropeptide hormone activity, positive regulation of glucocorticoid  
265 secretion, inositol 1,3,4,5 tetrakisphosphate binding, and corticosterone secretion (adjusted- $P <$   
266  $2.0 \times 10^{-2}$ ). Similarly, a tissue expression analysis identified a positive association between SNPs  
267 for PTB and genes expressed in adipose tissue, the heart, mammary breast cells, and the  
268 adrenal gland ( $P < 0.05$ ). An eQTL look-up using BIOS QTL identified transcripts in *AKR1C3*,  
269 *AKR1CL1*, *AKR1C4*, among others, associated with the top PTB SNP rs7098485. Using  
270 prioritized genes (i.e., top 8 PTB SNPs at GWAS  $P < 1.0 \times 10^{-5}$ ), differential gene expression in  
271 the above-mentioned tissues was identified for the *AKR1C3* gene associated with the SNP  
272 rs7098485. A gene-set enrichment analysis unveiled different biological processes related to the  
273 metabolism of polyketides, progesterone, glycoside, quinone, and steroids, in association with  
274 genes in *AKR1C3*, *AKR1C4*, *AKR1CL1*, all linked to the PTB SNP rs7098485.

275

276 *In-silico* interrogation of signals detected in the GWAS of GA revealed 15 independent  
277 significant SNPs (LD  $< 0.1$ ,  $P < 1.0 \times 10^{-5}$ ) and 229 GWAS-tagged SNPs in LD  $> 0.6$  with the  
278 independent SNPs, mapping to 41 genes. These SNPs were enriched in intronic, intergenic,  
279 and non-coding RNAs within exonic and intronic regions (S2 Fig). Gene set analysis revealed a  
280 positive association of GA SNPs with GO terms related to nucleotide sugar transmembrane  
281 transporters activity and the positive regulation of catecholamines secretion. Two GA SNPs in  
282 rs4834267 and rs17225178 were identified as eQTLs of transcripts in the *ALPK1* and *ARNT2*

283 genes, respectively, and data from the GWAScatalog revealed that rs4834267, rs73294123,  
284 and rs6591172, were previously reported in association with gut microbiota [23], hepatitis B [24],  
285 and genetic interactions in Alzheimer's disease [25], respectively. Tissue-specific enrichment  
286 analysis showed that genes associated with GA SNPs were enriched for DEGs up-regulated in  
287 adipose tissue and the brain (S2 Fig), while gene-set enrichment analysis revealed enrichment  
288 of GA SNPs for cancer modules and cancer gene neighborhoods (S4 Table).

289

## 290 **DISCUSSION**

291 This may be the largest study investigating the maternal genetic factors associated with PTB  
292 and GA in South American samples. Our study included data from the PAGE study which were  
293 2,212 women, 933 of whom were PTB cases. None of the ~ 6 million SNPs interrogated  
294 surpassed significance in our analysis; however, the top signals identified with  $P < 1.0 \times 10^{-5}$   
295 were associated with a lower risk of PTB ( $n = 6/8$  top SNPs), and with shorter GA ( $n = 16/20$  top  
296 SNPs) in the specific GWAS analyses. Two SNPs in *WNT4* associated with GA in Europeans in  
297 a previous study were replicated in our sample ( $P < 0.001$ ), and a GRS for GA constructed  
298 using five European SNPs, was associated with a 2-day longer GA in our study. The GRS  
299 explained 0.3% of the variance in GA, and this value was lower than the variance explained by  
300 the top 20 SNPs detected in our GWAS of GA (adjusted- $R^2 = 10\%$ ), suggesting that PTB genetic  
301 risk factors from European-ancestry populations may not be generalizable to non-European,  
302 admixed American populations. Functional annotation of top markers identified in the GWAS  
303 provided insights into their possible biological role in PTB risk and GA duration based on  
304 maternal genetic effects.

305

306 Our strongest signal for the GWAS of PTB (rs13151645), maps to the long intergenic non-  
307 coding RNA (lincRNAs) 1182 (*LINC01182*), which was suggestively associated with a 26%

308 lower risk of PTB and explained 1.2% of the variance in PTB in our sample. No previous  
309 evidence, to our knowledge, supports the role of rs13151645 or the *LINC01182* gene in PTB  
310 risk. However, several SNPs in the *LINC01182* gene in our data had a strong correlation (LD >  
311 0.5) with our index SNP rs13151645, suggesting that this region may be of biological  
312 importance for PTB. For instance, Zhou *et al.* [26] identified five lincRNAs associated with  
313 circadian rhythm genes or “clock genes” that were downregulated in the placentas of women  
314 with spontaneous PTB versus controls. Similarly, these lincRNAs were linked to pathways  
315 relevant to PTB (immunity, inflammation, oxidative stress, apoptosis, etc.) [26]. Replication of  
316 our top signal at *LINC01182* in a larger dataset is necessary to confirm its potential mechanistic  
317 role in PTB risk.

318

319 Another variant of interest observed with a nominal association was rs7098485 near the Aldo-  
320 Keto Reductase Family 1 Member C3 (*AKR1C3*). Other SNPs in high LD with rs7098485 were  
321 observed in this locus (n = 11) in our data. *AKR1C3* is a protein-coding gene member of the  
322 superfamily of aldo/keto reductases, which are enzymes that catalyze the conversion of  
323 aldehydes and ketones to their specific alcohols using NADH/NADPH as cofactors [27]. This  
324 enzyme is involved in the metabolism of sex hormones (estrogen, androgens, and  
325 progesterone) and prostaglandins [28], which are acidic lipids that promote inflammatory  
326 responses [29]. Prostaglandins are ubiquitously produced by different mammalian cells [29, 30],  
327 and during gestation, they are known to promote labor by stimulating smooth muscle  
328 contractility in the uterus and cervical ripening [2, 14, 30, 31]. Thus, the role of *AKR1C3* on PTB  
329 may be through alterations in the function of this enzyme that result in a reduction in the  
330 biochemical availability of prostaglandins. This is consistent with our observation of *AKR1C3*  
331 SNPs having effect sizes suggestive of protective effects on PTB. This concept may be  
332 additionally supported by the previous identification of an eQTL associated with our SNP  
333 rs7098485, where the T risk allele (same risk allele for PTB) was associated with lower



334 expression of the transcript ENSG00000196139 in *AKR1C3* (FDR < 0.001). To confirm our  
335 hypothesis of a potential functional role of rs7098485 on *AKR1C3* in PTB risk, a formal  
336 functional analysis is required. The expression of *AKR1C3* has also been associated with better  
337 survival in patients with endometrial cancer [32, 33], and with other diseases like breast cancer  
338 [28] and asthma [34].

339  
340 Our *in-silico* functional analysis using top PTB-associated SNPs supported the role of these  
341 markers in the metabolism of progesterone, glycoside, and prostanoid (i.e., prostaglandin),  
342 primarily attributed to *AKR1C3* function. Another relevant pathway for PTB identified in our gene  
343 set analysis was linked to the regulation of steroid secretion (glucocorticoids and  
344 corticosterone); these are compounds commonly used as an antenatal treatment to improve  
345 pregnancy outcomes in women at risk of preterm delivery [35]. Our risk factor analysis  
346 demonstrated that the top 8 SNPs explained a slightly larger proportion of the variance in PTB  
347 compared to non-genetic risk factors (7.2% vs 6.9%), indicating the potential of genetic markers  
348 in adding value to the risk prediction of PTB already achieved by common risk factors. Of note,  
349 results of this genetic variance analysis should be interpreted with caution, as variance in GA  
350 and PTB was calculated in the same sample used to conduct the GWAS; thus, results may be  
351 overestimated. Furthermore, the SNPs included in the calculation of the phenotypic variance  
352 were only suggestively associated with the traits in our GWAS. None of the markers previously  
353 detected for PTB in Europeans were replicated in our sample ( $P > 0.03$ ), and calculating a GRS  
354 was not possible as only one out of six SNPs previously reported in Europeans (rs2963463 in  
355 *EBF1*), surpassed QC for the GRS in our dataset.

356  
357 The intronic SNP rs72824565 mapping to the Catenin Alpha 2 (*CTNNA2*) gene was the most  
358 significant signal in the GWAS of GA. This variant was associated with a 3-week shorter  
359 gestational duration, and this association was consistent across the three other GWAS SNPs

360 observed in high correlation with rs72824565 in *CTNNA2*. The influence of this protein-coding  
361 gene on lower GA may be through fetal neurodevelopmental impairments. Specifically,  
362 *CTNNA2* acts as a link between the cadherin adhesion receptors and the cytoskeleton in  
363 regulating cell-cell adhesion and differentiation in the nervous system [36]. It also participates in  
364 cortical neuronal migration and neurite growth [36]. Bi-allelic loss-of-function mutations in  
365 *CTNNA2* have been associated with pachygyria syndrome, a neurodevelopmental brain defect  
366 associated with impairments in neuronal migration, resulting in lower gyri complexity in the  
367 cerebral cortex [36]. This syndrome is accompanied by motor and cognitive delays in affected  
368 patients [36]. Our findings at *CTNNA2* are consistent with the pathophysiology of PTB, as  
369 infants born preterm are more likely to suffer from neurodevelopmental disorders and long-term  
370 cognitive and motor disabilities [2, 14]. Other variants in *CTNNA2* have been previously  
371 associated with educational attainment [37], externalizing behaviors (attention deficit  
372 hyperactivity disorder, substance abuse, antisocial behavior) [38], and acute myeloid leukemia  
373 [39], among others. Another SNP detected with suggestive inverse association with GA in our  
374 study was located in chr1:65376701, mapping to the *DNAJC6* gene, a member of the heat  
375 shock family of proteins (HSP). SNPs mapping to this gene, and similar members of this family  
376 of proteins, have been vinculated with higher risk of sPTB [40]. This multi-omic study conducted  
377 in Europeans identified SNPs and transcripts of HSPs associated with higher sPTB,  
378 hypothesizing that activation of these proteins mediated signalling may promote labor [40],  
379 either through their role in activating immune reponses as a consequence of infection, or  
380 through their role in maturation and activation of nuclear hormone receptors, like glucocorticoid,  
381 androgen, estrogen and progesterone receptors [40].

382

383 For additional top SNPs identified in the GWAS of GA, the *in-silico* functional analysis  
384 suggested enrichment of their mapping genes for cancer modules in relation to SNPs in *ARNT2*,  
385 *DCLK1*, and *SPOCK1*. Furthermore, top GA-SNPs were associated with the positive regulation

386 of catecholamines secretion in the gene-set analysis (*CHRNA2, GDNF, CXCL12, KCNB1*).  
387 Catecholamines are stress hormones that have been previously associated with an increased  
388 risk of spontaneous PTB [41]. GA-associated SNPs also showed upregulation of differentially  
389 expressed genes in adipose tissue, brain caudate, and brain putamen regions in the basal  
390 ganglia, supporting the concept that GA variants may be associated with PTB risk through their  
391 influence on fetal neurodevelopmental traits. Our risk factor analysis suggested that top SNPs  
392 associated with GA explained almost double the variance in the trait captured by common non-  
393 genetic risk factors (10% vs 5.5%), and this variance was close to the one reported by Zhang *et*  
394 *al.* in their GWAS (10% for us vs 17% in their study).

395

396 We replicated two associations in *WNT4* previously identified in Europeans, and the GRS using  
397 five European SNPs was associated with a 2-day increase in GA in our sample; the GRS  
398 captured 0.3% of the variance in GA in PAGE participants, which was less than the variance  
399 explained by our top SNPs of the GWAS. The replication of markers in *WNT4* across  
400 populations is worth noting, as it indicates that this locus may represent a common mechanistic  
401 pathway for PTB, which causal protective effect on the disease still needs to be validated.  
402 Biologically, *WNT4* is related to the decidualization of the endometrium for subsequent  
403 implantation and establishment of pregnancy [16].

404

405 Our study had several limitations. First, we did not differentiate PTB cases into two clinical  
406 presentations as PPRM or spontaneous PTB, limiting our ability to identify markers more  
407 targeted to the pathophysiological mechanisms of each subtype of PTB. The main reason why  
408 we did not do this stratified analysis was due to sample size constraints, as having smaller  
409 comparison groups would have been disadvantageous for the GWAS knowing the very large  
410 sample sizes required for this type of analysis (~100k). In addition, we did not consider in our  
411 analysis the fetal genetic effects, knowing that GA may be influenced by both, maternal and

412 fetal genetic effects [2]. However, previous studies looking at both genotypes coincided with the  
413 major influence that the maternal genotype has on GA and PTB determination versus the fetal  
414 genotype [16]. Lastly, our study may have been limited by the population studied, and by its  
415 specific socioeconomic and environmental context, with findings that may not be generalizable  
416 to other populations, as has been demonstrated in other trans-ethnic GWAS of PTB [15].  
417 Despite these limitations, our study has several strengths. First, this is the largest GWAS of PTB  
418 and GA conducted among South American individuals, which expands the literature on the  
419 genetics of pregnancy outcomes in low-and-middle income populations from non-European  
420 ancestry. By making our GWAS results available, we facilitate their inclusion in future meta-  
421 analyses of GWAS that are more tailored to South American and other Latino and Hispanic  
422 populations; thus, promoting the discovery of genetic-based treatments that are more useful to  
423 these minority populations. As another strength, we used the most up-to-date reference panel  
424 for the genetic imputation to include the largest number of SNPs to be interrogated in the  
425 GWAS. Even though our results were null, we conducted different post hoc analyses to evaluate  
426 the biological meaning of the top signals identified in each GWAS. Lastly, we performed a  
427 replication analysis of signals previously identified in Europeans, and we developed a GRS for  
428 GA in PAGE samples based on European markers.

429  
430 Further validation of our findings in larger datasets of similar ancestral backgrounds, may help in  
431 elucidating the mechanisms associated with the pathophysiology of PTB in the context of  
432 minority and vulnerable populations, and to identify genetic markers that serve in the early  
433 prediction of PTB in women at higher risk.

434

435 **CONCLUSIONS**

436 Our study identified genetic markers suggestively associated with PTB and GA among Peruvian  
437 women. These markers were independent of those previously reported in Europeans, and they  
438 were linked to pathways related to the metabolism of steroids, progesterone, and prostanoids  
439 (*AKR1C3*), fetal neurodevelopment (*CTNNA2*), response to stressors (*DNAJC6*), and  
440 catecholamine secretion (*CXCL12*). Prioritized genes for GA were upregulated in metabolically  
441 relevant tissues such as adipose tissue and the brain, supporting some of our findings in the  
442 pathway analysis. Replication of suggestive markers in larger studies of the same population  
443 background is warranted to validate their role in PTB and GA. The discovery of new markers for  
444 PTB and the validation of existing ones will aid in developing genetic scores that permit an  
445 accurate risk stratification of women, especially in high-risk non-European ancestry populations.  
446

## 447 **METHODS**

### 448 ***Study Population and Analytic Sample***

449 This study was conducted among participants of the Placental Abruption Genetic Epidemiology  
450 study (PAGE), a case-control study from Lima, Peru, designed to investigate the genetic and  
451 environmental determinants of placenta abruption and other adverse pregnancy outcomes like  
452 preterm birth. Study procedures have been described elsewhere [42, 43]. Briefly, women were  
453 recruited from seven participating hospitals in Lima between March 2013 and December 2015.  
454 Preterm women were identified by daily monitoring of admission logbooks from antepartum  
455 wards, the emergency room (intensive care units), delivery wards, and surgery rooms [42].  
456 Eligible women for this study were older than 18 years of age, had singleton pregnancies, had  
457 sufficient phenotypic and medical information to determine PTB status, and provided a saliva  
458 sample at delivery to extract genomic data (described below). Women were excluded if preterm  
459 deliveries were medically indicated. The total number of participants remained 933 PTB cases  
460 and 1,279 controls. Eligible women were invited to participate in a 30-minute in-person interview

461 during their hospital stay, where research staff explained the study objectives and obtained  
462 written informed consent from study participants. All the study protocols were approved by the  
463 IRB of participating hospitals, and the Swedish Medical Center, Seattle, WA, where the study  
464 was administratively based.

465

#### 466 ***Data collection***

467 Trained research staff conducted in-person interviews to gather information on the participants'  
468 sociodemographic, lifestyle and anthropometric characteristics using structured questionnaires.  
469 Sociodemographic factors collected included maternal age, education level (higher vs lower  
470 than high school), marital status (married living with a partner vs not married living alone),  
471 employment status, health perception during pregnancy (good to excellent vs fair to poor), and  
472 infant sex. Lifestyle characteristics included smoking or alcohol use during pregnancy, and drug  
473 abuse. Pre-pregnancy body mass index (BMI) was measured continuous and categorically  
474 (underweight if BMI < 18 kg/m<sup>2</sup>, normal weight if BMI 18-24.9 kg/m<sup>2</sup>, overweight if BMI 25-30  
475 kg/m<sup>2</sup>, and obese if BMI 30-49 kg/m<sup>2</sup>). Maternal obstetric variables were abstracted from  
476 medical records and included gestational age at delivery (weeks), mode of delivery (natural vs  
477 C-section), planned pregnancy, diagnosis of preterm premature rupture of fetal membranes  
478 (PPROM), gravidity (primiparous vs multiparous), preeclampsia, vaginal bleeding, spontaneous  
479 abortions, and infant birthweight (g).

480

#### 481 ***Gestational Age at Delivery and Preterm Birth***

482 Gestational age at delivery (GA) was obtained from the maternal self-reported date of the last  
483 menstrual period (LMP) during the interview. When values were missing for LMP, information on  
484 GA was obtained from an early pregnancy ultrasound or fundal height examinations performed  
485 ≤ 20 weeks of gestation according to medical records. We defined preterm births (PTB)

486 following the American College of Obstetrics and Gynecologists criteria as deliveries occurring <  
487 37 weeks of gestation [44]. PTB cases can be further classified into three pathophysiological  
488 groups: spontaneous PTB (sPTB), preterm with premature rupture of fetal membranes  
489 (PPROM), or medically induced PTB [45]. Women with sPTB have a medical diagnosis of  
490 spontaneous labor with intact fetal membranes prior to completing 37 weeks of gestation.  
491 Women with PPRM have a physician diagnosis of premature rupture of fetal membranes  
492 before labor and prior to completing 37 weeks of gestation. Of the 933 PTB cases in the study,  
493 527 (56%) were sPTB and 379 (41%) were PPRM. A high proportion of C-sections was seen  
494 in both clinical subgroups (70% in sPTB and 68% in PPRM). Controls were selected at  
495 random from women delivering at or after 37 weeks and before 42 weeks across the  
496 participating hospitals. By study protocol, controls were captured within 48 hours of the  
497 identification of a PTB case. To preserve a larger sample size for genetic analyses and increase  
498 power [46], we run the GWAS using the general PTB definition and including both clinical  
499 subgroups.

500

### 501 ***DNA Extraction and Genotyping***

502 At delivery, maternal saliva samples were collected, plated, and stored at room temperature  
503 using the Oragene™ saliva cells kit (OGR500, DNA Genotek, Ottawa, Canada) [47, 48].  
504 Genomic DNA was extracted using the Qiagen DNAeasy™ system and following manufacturer  
505 protocols (Qiagen, Valencia, CA) [47]. Direct genotyping of ~ 300,000 genome-wide genetic  
506 variants was performed at the Genomics Shared Resource, Roswell Park Cancer Institute,  
507 Buffalo, NY, using the Illumina HumanCore-24 BeadChip array (Illumina Inc., San Diego, CA)  
508 [47, 48].

509



## 510 **Data Quality Control and Imputation**

511 Prior to performing genetic imputation, we conducted different quality control (QC) steps on the  
512 directly genotyped genetic variants or SNPs. Firstly, we filtered out SNPs based on minor allele  
513 frequency (MAF)  $< 0.01$  ( $n=39,439$ ), missing genotyping rate  $> 5\%$  ( $n=19,700$ ), and deviation  
514 from the Hardy-Weinberg equilibrium (HWE  $P < 1.0 \times 10^{-5}$ ,  $n=223$ ) in the control group. After  
515 quality control, 249,299 directly genotyped SNPs remained in the dataset. Similarly, samples  
516 were filtered out if they had genotyping failure rate  $> 0.1$  ( $n=38$ ), were duplicates or related  
517 considering an identity by descent (IBD) value  $> 0.9$  ( $n=20$ ), had excess  
518 heterozygosity/homozygosity rate (outside the range of  $\pm 3$  standard deviations from the mean  
519 heterozygosity;  $n=29$ ), and had discordance between the reported sex and sex predicted using  
520 genetic information ( $F \geq 0.8$ ;  $n=10$ ). We did not exclude one sample identified with ancestry  
521 divergence (outside the following range for the first two genetic principal components (PCs):  
522 PC1  $> -0.01$  and  $< 0.01$ , and PC2  $> -0.02$  and  $< 0.02$ ). This sample was retained knowing the  
523 ethnical homogeneity of our dataset (most are mestizos - or mixed race/ethnicity), and because  
524 we included the first two genetic PCs as covariates to account for potential population  
525 stratification in downstream analyses (see below). A total of 2,212 samples (933 cases and  
526 1,279 controls) were retained for further analyses. We then performed genetic imputation to  
527 infer the genotype of SNPs not included in the array and to maximize our chances of identifying  
528 true genetic associations with PTB and/or GA. First, we used the Michigan server  
529 (<https://imputationserver.sph.umich.edu>) [49] to conduct a pre-phasing step and check for  
530 inconsistencies between our input genotype, and data from the 1000 Genomes phase 3  
531 (version 5) Admixed American (AMR) population used as the reference panel. Cleaned pre-  
532 phased genotypes were then phased using EAGLE (version 2.4) [50] to infer haplotypes and  
533 improve imputation accuracy [47]. Imputation of phased genotypes was conducted using the  
534 TOPMed imputation server (<https://imputation.biodatacatalyst.nih.gov/>) and reference

535 panel (version r2). Imputed genotype data was then QCed to exclude SNPs that were non-  
536 biallelic SNPs, duplicated, had quality imputation score (INFO) < 0.8, HWE  $P < 1.0 \times 10^{-5}$ , MAF <  
537 0.01, and genotyping call rate < 99%, leaving in total 6,047,004 high-quality SNPs for the  
538 analysis. SNPs were annotated to their rsID (dbSNP142, GRCh37/hg19) using data from the  
539 Haplotype Reference Consortium (HRC) reference panel (HRC.r1-1) [51], or using the  
540 chromosome and position if the rsID was not available.

#### 541 ***Ancestry Analysis***

542 Using directly genotyped SNPs that passed quality control, we performed a combined genetic  
543 PC analysis between samples from PAGE and the 1000 Genomes Project (phase 3, version  
544 2013.05.02) to determine the genetic ancestry of participants. First, we generated PCs for SNPs  
545 identified in common between the 1000 Genomes and PAGE datasets. Then, we retrieved  
546 genetic variation captured by the first three PCs obtained from the combined dataset analysis  
547 and used scatterplots to visually represent the clustering of samples based on their genetic  
548 relatedness across pairs of PCs (i.e., PC1 vs PC2 and PC1 vs PC3). We also generated PCs  
549 for the PAGE dataset alone and retrieved the first two PCs, explaining in total 66.1% of the total  
550 genetic variation in the sample, to adjust for population stratification in subsequent analyses.

551

#### 552 ***Statistical Analysis***

553 We described characteristics of study participants using the mean and standard deviations (SD)  
554 for continuous variables, and proportions and percentages for categorical variables.

555 Comparisons between groups (PTB cases and controls) were done using t-tests (Wilcoxon rank  
556 sum test if non-parametric) and chi-squared tests for continuous and categorical variables,  
557 respectively. Using an additive genetic model (i.e., assuming a linear increase in PTB risk/GA  
558 per additional risk allele), we fitted multivariable logistic and linear regression models to test the  
559 association of each SNP as the exposure against PTB and GA as the outcome, respectively,

560 using PLINK (version 1.90) [52]. Genome-wide associations at each SNP were adjusted for  
561 maternal age, and the first two genetic PCs to account for population stratification. We used the  
562 genomic inflation factor or  $\lambda$  to identify population stratification or residual confounding in the  
563 GWAS if  $\lambda > 1.0$ . Adjusted regression coefficients (ORs/Betas), 95% confidence intervals, and  
564 the genomic control uncorrected  $P$ -values (~ to genomic controlled corrected  $P$ -values) were  
565 reported for each SNP. To determine if there was consistency in the genetic effect between  
566 traits, we looked up association estimates for top signals ( $P < 1.0 \times 10^{-5}$ ) identified in the GWAS  
567 of PTB, in the GWAS of GA, and vice versa. Similarly, we looked up our top suggestive signals  
568 in the largest and most recent European GWAS meta-analysis conducted to date for GA and  
569 PTB [17] (<http://egg-consortium.org/Gestational-duration-2023.html>). We applied correction for  
570 multiple testing (6,047,004 tests performed) using the Bonferroni method and regarded GWAS  
571 significant associations at  $P < 5.0 \times 10^{-8}$ , or nominally significant associations at  $P < 1.0 \times 10^{-5}$ .  
572 Results of the GWAS were summarized using Manhattan and Quantile plots (Q-Q plots). We  
573 used stepwise adjusted regressions to obtain the percent of the total variance in the traits  
574 explained by each one of the top SNPs ( $P < 1.0 \times 10^{-5}$ ) identified in the GWAS of PTB and GA,  
575 and by all of them in combination. We reported the adjusted- $R^2$  obtained from the linear  
576 regressions between the SNPs (exposures) and GA (outcome), or the Nagelkerke's pseudo- $R^2$   
577 obtained from the logistic regressions between the SNPs and PTB. Similarly, we estimated the  
578 percent of the total variance in the traits (GA and PTB) explained by the non-genetic risk factors  
579 of maternal age, gravidity, pre-pregnancy BMI, presence of vaginal bleeding, maternal  
580 education, child sex, and the first two genetic PCs. We then compared the proportion of the  
581 variance in the traits explained by the genetic, the non-genetic risk factors, and by the  
582 combination of both factors.  
583

## 584 **Replication of European Markers and Polygenic Score**

585 We retrieved summary statistics for genome-wide significant associations obtained in a recent  
586 meta-analysis of GWAS of PTB and GA conducted by Zhang *et al.* in European cohorts  
587 (N=52,211, spontaneous PTB cases with GA < 37 weeks = 5,896 and term controls = 46,315)  
588 [16]. Across the discovery and replication stages of this metanalyses, 11 SNPs were identified  
589 in association with PTB (in 6 genes), 6 with GA (in 3 genes), and 2 SNPs were identified in  
590 common between both traits [16]. We looked up associations for these 15 unique SNPs in the  
591 GWAS conducted in PAGE and considered evidence of replication if a similar direction of  
592 association was identified between studies with replication  $P < 0.05/\text{number of SNPs}$   
593 interrogated. Since effect estimates were expressed in days for SNPs associated with GA in  
594 Zhang *et al.*, we divided them by 7 to match the interpretation of GA used in our study (in  
595 weeks). For SNPs in Zhang *et al.* [16] not present in our GWAS, we searched for proxy SNPs  
596 with an LD  $R^2 > 0.7$  with the target SNP using LDlink (<https://ldlink.nih.gov/>) [53] and the 1000  
597 genomes as the reference panel (AMR and EUR populations). Using individual-level genotype  
598 data in PAGE and summary statistics from Zhang *et al.*, we developed a genetic risk score  
599 (GRS) for each participant in PAGE using PLINK (version 1.90). The score for each participant  
600 represents the sum of the total number of risk alleles present in their genotype, weighted by the  
601 effect of each SNP in the score as obtained from GWAS summary data. SNPs from the  
602 reference study were clumped to include only independent SNPs (LD < 0.1) in the score,  
603 resulting in five independent GA-associated SNPs but only one independent PTB-associated  
604 SNP; thus, we only constructed the GRS for GA using PAGE samples. The score was  
605 standardized to represent a weekly change in GA per SD increase in the GRS. The variance  
606 ( $R^2$ ) in GA explained by the standardized score was obtained using the adjusted- $R^2$  from a full  
607 model (with the score), minus the adjusted- $R^2$  from a basic model including only covariates  
608 (maternal age, gestational age, PC1, and PC2).

609

### 610 ***In-silico Functional Analysis***

611 We used summary data from GWAS conducted among PAGE participants to perform an *in-*  
612 *silico* functional analysis using tools available in the FUMA-GWAS (Functional Mapping and  
613 Annotation of GWAS) web tool (<https://fuma.ctglab.nl/>) [54]. This analysis was restricted to  
614 associations with  $P < 0.05$  in the GWAS. For gene annotation purposes, we specified as the  
615 reference panel the 1000 Genomes AMR population, selected an LD threshold  $< 0.6$  to identify  
616 independent SNPs, and chose a  $P < 1.0 \times 10^{-5}$  as the minimum  $P$ -value to detect lead SNPs.  
617 Other settings were left as default. FUMA-GWAS uses two applications, the SNP2GENE  
618 (version 1.3.5d) and the GENE2FUNC (version 1.3.5), to provide complete SNP annotation and  
619 biological interpretation of GWAS data, respectively. Using the SNP2GENE function, we  
620 reported information on the lead SNP in a region, the number of independent SNPs ( $r^2 < 0.1$ )  
621 and their functional consequences (intronic, exonic, non-coding RNA, etc.), the nearest gene,  
622 and output from the MAGMA analysis. For this latter, gene set and tissue-specificity analyses  
623 were performed using curated data obtained from the MsigDB (version 7.0) and GTEx (v7)  
624 databases, respectively. When relevant, we also reported phenotypic data obtained from the  
625 GWAScatalog and expression Quantitative Trait Loci (eQTL) retrieved from the BIOS QTL [55],  
626 regarding lead SNPs in our dataset. Using genes prioritized in the SNP2GENE analysis, we run  
627 the GENE2FUNC function to test the enrichment of candidate genes for curated gene sets  
628 obtained from MsigDB, WikiPathways, and reported genes in the GWAScatalog. Enrichment  
629 was also tested for tissue-specific gene expression using data from GTEx (v8) based on pre-  
630 calculated differentially expressed gene sets (DEGs). In both cases, enrichment analysis was  
631 performed using hypergeometric tests, with correction for multiple testing using the Bonferroni  
632 method. Further information on the protocol implemented in the FUMA-GWAS analysis can be  
633 found on their website (<https://fuma.ctglab.nl/tutorial>).

634

## 635 **ACKNOWLEDGEMENTS**

636 The authors are grateful to the participants of the collaborating hospitals for their involvement in  
637 the study. We also want to recognize the important contribution of Ms. Elena Sanchez and the  
638 research staff of the Asociacion Civil Proyectos en Salud (PROESA), Peru, for their expert  
639 technical assistance with this project.

640

## 641 **AUTHOR CONTRIBUTIONS**

642 DLJQ and BG contributed to the conception and design of the study. LL and TW provided their  
643 expert feedback on the methodology of the study and the interpretation of results. DLJQ  
644 conducted the analyses and wrote the paper in collaboration with BG. SES contributed with  
645 access to the data. SES, LL, TW, and MAW provided their expert clinical, epidemiological, and  
646 statistical input on the manuscript. All authors read the paper and agreed with the version  
647 submitted for publication.

648

## 649 **FUNDING DISCLOSURE**

650 This work was supported by the National Institutes of Health (R01-HD059827; 1R21HD102822).

651

## 652 **DATA AVAILABILITY**

653 Complete summary data of the GWAS of GA and PTB would be available from the GWAS  
654 catalog EMBL-EBI repository. Study accession numbers will be released upon approval of the  
655 manuscript for publication.

656

## 657 **CONFLICT OF INTERESTS**

658 The authors have no conflict of interest to disclose in relation to this work.

659

## 660 REFERENCES

- 661 1. Chawanpaiboon S, Vogel JP, Moller A-B, Lumbiganon P, Petzold M, Hogan D, et al.  
662 Global, regional, and national estimates of levels of preterm birth in 2014: a systematic  
663 review and modelling analysis. *The Lancet Global Health*. 2019;7(1):e37-e46.
- 664 2. Goldenberg RL, Culhane JF, Iams JD, Romero R. Epidemiology and causes of preterm  
665 birth. *Lancet*. 2008;371(9606):75-84.
- 666 3. De Costa A, Moller A-B, Blencowe H, Johansson EW, Hussain-Alkhateeb L, Ohuma EO, et  
667 al. Study protocol for WHO and UNICEF estimates of global, regional, and national preterm  
668 birth rates for 2010 to 2019. *PLOS ONE*. 2021;16(10):e0258751.
- 669 4. Calvert C, Brockway M, Zoega H, Miller JE, Been JV, Amegah AK, et al. Changes in  
670 preterm birth and stillbirth during COVID-19 lockdowns in 26 countries. *Nature Human  
671 Behaviour*. 2023;7(4):529-44.
- 672 5. Alamneh TS, Teshale AB, Worku MG, Tessema ZT, Yeshaw Y, Tesema GA, et al. Preterm  
673 birth and its associated factors among reproductive aged women in sub-Saharan Africa:  
674 evidence from the recent demographic and health surveys of sub-Saharan African countries.  
675 *BMC Pregnancy and Childbirth*. 2021;21(1):770.
- 676 6. Landry JS, Menzies D. Occurrence and severity of bronchopulmonary dysplasia and  
677 respiratory distress syndrome after a preterm birth. *Paediatr Child Health*. 2011;16(7):399-  
678 403.
- 679 7. Vento M, Moro M, Escrig R, Arruza L, Villar G, Izquierdo I, et al. Preterm resuscitation with  
680 low oxygen causes less oxidative stress, inflammation, and chronic lung disease.  
681 *Pediatrics*. 2009;124(3):e439-49.
- 682 8. Deshpande G, Rao S, Patole S, Bulsara M. Updated meta-analysis of probiotics for  
683 preventing necrotizing enterocolitis in preterm neonates. *Pediatrics*. 2010;125(5):921-30.
- 684 9. Adams-Chapman I, Heyne RJ, DeMauro SB, Duncan AF, Hintz SR, Pappas A, et al.  
685 Neurodevelopmental Impairment Among Extremely Preterm Infants in the Neonatal  
686 Research Network. *Pediatrics*. 2018;141(5).
- 687 10. Arpino C, Compagnone E, Montanaro ML, Cacciatore D, De Luca A, Cerulli A, et al.  
688 Preterm birth and neurodevelopmental outcome: a review. *Childs Nerv Syst*.  
689 2010;26(9):1139-49.
- 690 11. Norman M. Preterm Birth—An Emerging Risk Factor for Adult Hypertension? *Seminars in  
691 Perinatology*. 2010;34(3):183-7.
- 692 12. de Jong F, Monuteaux MC, van Elburg RM, Gillman MW, Belfort MB. Systematic review  
693 and meta-analysis of preterm birth and later systolic blood pressure. *Hypertension*.  
694 2012;59(2):226-34.
- 695 13. Finken MJ, Keijzer-Veen MG, Dekker FW, Frölich M, Hille ET, Romijn JA, et al. Preterm  
696 birth and later insulin resistance: effects of birth weight and postnatal growth in a population  
697 based longitudinal study from birth into adult life. *Diabetologia*. 2006;49(3):478-85.
- 698 14. Romero R, Dey SK, Fisher SJ. Preterm labor: one syndrome, many causes. *Science*.  
699 2014;345(6198):760-5.



- 700 15. Rappoport N, Toung J, Hadley D, Wong RJ, Fujioka K, Reuter J, et al. A genome-wide  
701 association study identifies only two ancestry specific variants associated with spontaneous  
702 preterm birth. *Sci Rep*. 2018;8(1):226.
- 703 16. Zhang G, Feenstra B, Bacelis J, Liu X, Muglia LM, Juodakis J, et al. Genetic Associations  
704 with Gestational Duration and Spontaneous Preterm Birth. *N Engl J Med*.  
705 2017;377(12):1156-67.
- 706 17. Solé-Navais P, Flatley C, Steinhorsdottir V, Vaudel M, Juodakis J, Chen J, et al. Genetic  
707 effects on the timing of parturition and links to fetal birth weight. *Nature Genetics*.  
708 2023;55(4):559-67.
- 709 18. Zhang G, Srivastava A, Bacelis J, Juodakis J, Jacobsson B, Muglia LJ. Genetic studies of  
710 gestational duration and preterm birth. *Best Pract Res Clin Obstet Gynaecol*. 2018;52:33-  
711 47.
- 712 19. Tiensuu H, Haapalainen AM, Karjalainen MK, Pasanen A, Huusko JM, Marttila R, et al.  
713 Risk of spontaneous preterm birth and fetal growth associates with fetal SLIT2. *PLoS*  
714 *Genet*. 2019;15(6):e1008107.
- 715 20. Gupta JK, Care A, Goodfellow L, Alfirevic Z, Müller-Myhsok B, Alfirevic A. Genome and  
716 transcriptome profiling of spontaneous preterm birth phenotypes. *Scientific Reports*.  
717 2022;12(1):1003.
- 718 21. Hong X, Surkan PJ, Zhang B, Keiser A, Ji Y, Ji H, et al. Genome-wide association study  
719 identifies a novel maternal gene  $\times$  stress interaction associated with spontaneous  
720 preterm birth. *Pediatr Res*. 2021;89(6):1549-56.
- 721 22. Hong X, Hao K, Ji H, Peng S, Sherwood B, Di Narzo A, et al. Genome-wide approach  
722 identifies a novel gene-maternal pre-pregnancy BMI interaction on preterm birth. *Nat*  
723 *Commun*. 2017;8:15608.
- 724 23. Ishida S, Kato K, Tanaka M, Odamaki T, Kubo R, Mitsuyama E, et al. Genome-wide  
725 association studies and heritability analysis reveal the involvement of host genetics in the  
726 Japanese gut microbiota. *Commun Biol*. 2020;3(1):686.
- 727 24. Tian C, Hromatka BS, Kiefer AK, Eriksson N, Noble SM, Tung JY, et al. Genome-wide  
728 association and HLA region fine-mapping studies identify susceptibility loci for multiple  
729 common infections. *Nat Commun*. 2017;8(1):599.
- 730 25. Wang H, Yang J, Schneider JA, De Jager PL, Bennett DA, Zhang HY. Genome-wide  
731 interaction analysis of pathological hallmarks in Alzheimer's disease. *Neurobiol Aging*.  
732 2020;93:61-8.
- 733 26. Zhou G, Fichorova RN, Holzman C, Chen B, Chang C, Kasten EP, et al. Placental circadian  
734 lincRNAs and spontaneous preterm birth. *Frontiers in Genetics*. 2023;13.
- 735 27. GeneCards, The human gene database: Weizmann Institute of Science; [Available from:  
736 <https://www.genecards.org/cgi-bin/carddisp.pl?gene=AKR1C3&keywords=AKR1C3>].
- 737 28. Byrns MC, Duan L, Lee SH, Blair IA, Penning TM. Aldo-keto reductase 1C3 expression in  
738 MCF-7 cells reveals roles in steroid hormone and prostaglandin metabolism that may  
739 explain its over-expression in breast cancer. *J Steroid Biochem Mol Biol*. 2010;118(3):177-  
740 87.
- 741 29. Ricciotti E, FitzGerald GA. Prostaglandins and inflammation. *Arterioscler Thromb Vasc Biol*.  
742 2011;31(5):986-1000.

- 743 30. O'Brien WF. The role of prostaglandins in labor and delivery. *Clin Perinatol.*  
744 1995;22(4):973-84.
- 745 31. Romero R, Yeo L, Chaemsaitong P, Chaiworapongsa T, Hassan SS. Progesterone to  
746 prevent spontaneous preterm birth. *Semin Fetal Neonatal Med.* 2014;19(1):15-26.
- 747 32. Hojnik M, Šuster NK, Š S, Grazio SF, Verdenik I, Rižner TL. 783 AKR1C3 – a potential  
748 prognostic biomarker for patients with endometrial carcinomas. *International Journal of*  
749 *Gynecologic Cancer.* 2021;31(Suppl 3):A122.
- 750 33. Hojnik M, Kenda Šuster N, Smrkolj Š, Frković Grazio S, Verdenik I, Rižner TL. AKR1C3 Is  
751 Associated with Better Survival of Patients with Endometrial Carcinomas. *J Clin Med.*  
752 2020;9(12).
- 753 34. Jin Y. Activities of aldo-keto reductase 1 enzymes on two inhaled corticosteroids:  
754 implications for the pharmacological effects of inhaled corticosteroids. *Chem Biol Interact.*  
755 2011;191(1-3):234-8.
- 756 35. Antenatal corticosteroids revisited: repeat courses. *NIH Consens Statement.* 2000;17(2):1-  
757 18.
- 758 36. Schaffer AE, Breuss MW, Caglayan AO, Al-Sanaa N, Al-Abdulwahed HY, Kaymakçalan H,  
759 et al. Biallelic loss of human CTNNA2, encoding  $\alpha$ N-catenin, leads to ARP2/3 complex  
760 overactivity and disordered cortical neuronal migration. *Nat Genet.* 2018;50(8):1093-101.
- 761 37. Lee JJ, Wedow R, Okbay A, Kong E, Maghzian O, Zacher M, et al. Gene discovery and  
762 polygenic prediction from a genome-wide association study of educational attainment in 1.1  
763 million individuals. *Nat Genet.* 2018;50(8):1112-21.
- 764 38. Karlsson Linnér R, Mallard TT, Barr PB, Sanchez-Roige S, Madole JW, Driver MN, et al.  
765 Multivariate analysis of 1.5 million people identifies genetic associations with traits related  
766 to self-regulation and addiction. *Nat Neurosci.* 2021;24(10):1367-76.
- 767 39. Lv H, Zhang M, Shang Z, Li J, Zhang S, Lian D, et al. Genome-wide haplotype association  
768 study identify the FGFR2 gene as a risk gene for acute myeloid leukemia. *Oncotarget.*  
769 2017;8(5):7891-9.
- 770 40. Huusko JM, Tiensuu H, Haapalainen AM, Pasanen A, Tissarinen P, Karjalainen MK, et al.  
771 Integrative genetic, genomic and transcriptomic analysis of heat shock protein and nuclear  
772 hormone receptor gene associations with spontaneous preterm birth. *Scientific Reports.*  
773 2021;11(1):17115.
- 774 41. Holzman C, Senagore P, Tian Y, Bullen B, Devos E, Leece C, et al. Maternal  
775 catecholamine levels in midpregnancy and risk of preterm delivery. *Am J Epidemiol.*  
776 2009;170(8):1014-24.
- 777 42. Chahal HS, Gelaye B, Mostofsky E, Sanchez SE, Mittleman MA, Maclure M, et al. Physical  
778 Exertion Immediately Prior to Placental Abruption: A Case-Crossover Study. *Am J*  
779 *Epidemiol.* 2018;187(10):2073-9.
- 780 43. Workalemahu T, Enquobahrie DA, Gelaye B, Sanchez SE, Garcia PJ, Tekola-Ayele F, et  
781 al. Genetic variations and risk of placental abruption: A genome-wide association study and  
782 meta-analysis of genome-wide association studies. *Placenta.* 2018;66:8-16.
- 783 44. Preterm labor. *International Journal of Gynecology & Obstetrics.* 1995;50(3):303-13.
- 784 45. Gelaye B, Kirschbaum C, Zhong QY, Sanchez SE, Rondon MB, Koenen KC, et al. Chronic  
785 HPA activity in mothers with preterm delivery: A pilot nested case-control study. *J Neonatal*  
786 *Perinatal Med.* 2020;13(3):313-21.

- 787 46. Mead EC, Wang CA, Phung J, Fu JYX, Williams SM, Merialdi M, et al. The Role of  
788 Genetics in Preterm Birth. *Reproductive Sciences*. 2023.
- 789 47. Workalemahu T, Enquobahrie DA, Gelaye B, Tadesse MG, Sanchez SE, Tekola-Ayele F,  
790 et al. Maternal-fetal genetic interactions, imprinting, and risk of placental abruption. *J*  
791 *Matern Fetal Neonatal Med*. 2022;35(18):3473-82.
- 792 48. Workalemahu T, Enquobahrie DA, Gelaye B, Thornton TA, Tekola-Ayele F, Sanchez SE, et  
793 al. Abruptio placentae risk and genetic variations in mitochondrial biogenesis and oxidative  
794 phosphorylation: replication of a candidate gene association study. *American journal of*  
795 *obstetrics and gynecology*. 2018;219(6):617.e1-e17.
- 796 49. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation  
797 genotype imputation service and methods. *Nat Genet*. 2016;48(10):1284-7.
- 798 50. Loh P-R, Palamara PF, Price AL. Fast and accurate long-range phasing in a UK Biobank  
799 cohort. *Nature Genetics*. 2016;48(7):811-6.
- 800 51. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference  
801 panel of 64,976 haplotypes for genotype imputation. *Nat Genet*. 2016;48(10):1279-83.
- 802 52. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation  
803 PLINK: rising to the challenge of larger and richer datasets. *GigaScience*. 2015;4(1).
- 804 53. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific  
805 haplotype structure and linking correlated alleles of possible functional variants.  
806 *Bioinformatics*. 2015;31(21):3555-7.
- 807 54. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and  
808 annotation of genetic associations with FUMA. *Nature Communications*. 2017;8(1):1826.
- 809 55. Bonder MJ, Luijk R, Zhernakova DV, Moed M, Deelen P, Vermaat M, et al. Disease  
810 variants alter transcription factor levels and methylation of their binding sites. *Nat Genet*.  
811 2017;49(1):131-8.

812

813

#### 814 **SUPPORTING INFORMATION**

815 **S1 Table. Pregnancy history and maternal mental health characteristics of study**  
816 **participants (N=2,212).**

817 **S2 Table. Variance in preterm birth explained by top SNP with the smallest P-value**  
818 **detected in the GWAS in PAGE, and by known non-genetic risk factors.**

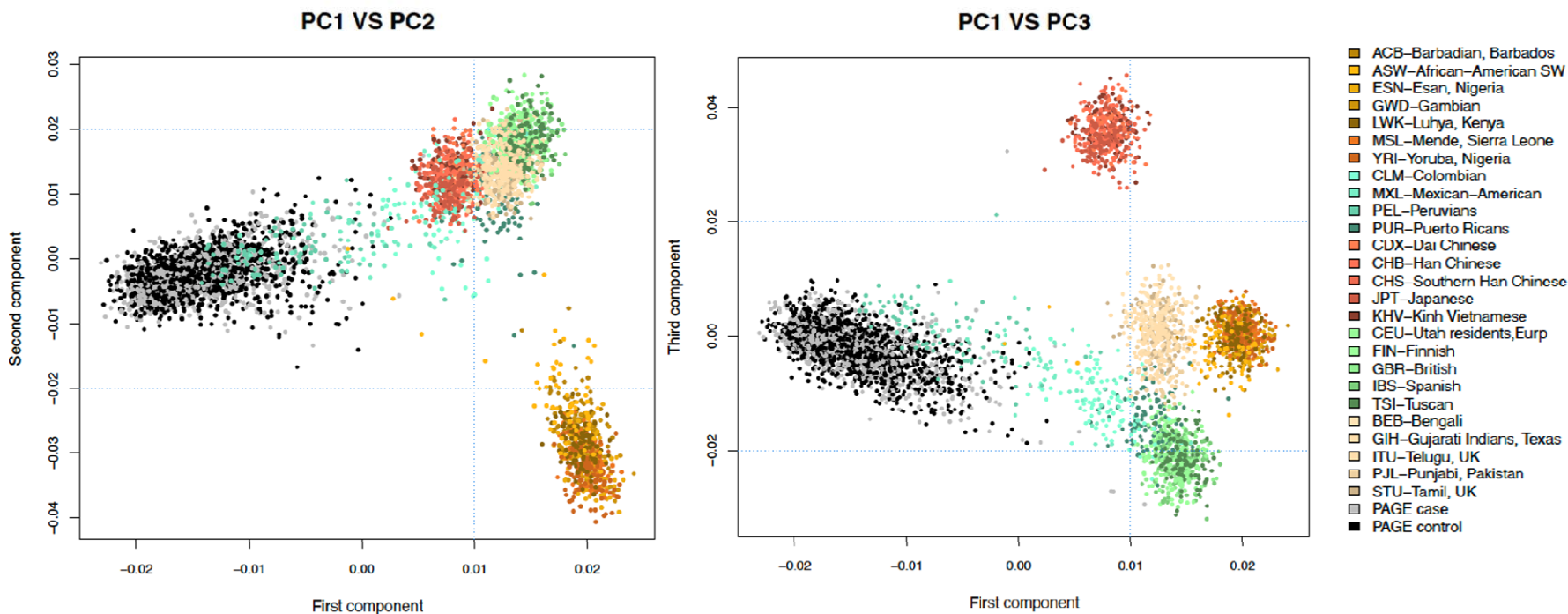
819 **S3 Table. Variance in gestational duration (GAD) explained by top SNP with the smallest**  
820 **P-value detected in the GWAS in PAGE, and by known non-genetic risk factors.**

821 **S4 Table. Enrichment analysis showing pathways associated with genes annotated to**  
822 **top SNPs detected in the GWAS of preterm birth and gestational duration in PAGE.**

823 **S1 Fig. Functional analysis of top-five independent (linkage disequilibrium < 0.1) genetic**  
824 **variants nominally associated with preterm birth (PTB) at  $P < 1.0 \times 10^{-5}$ .**

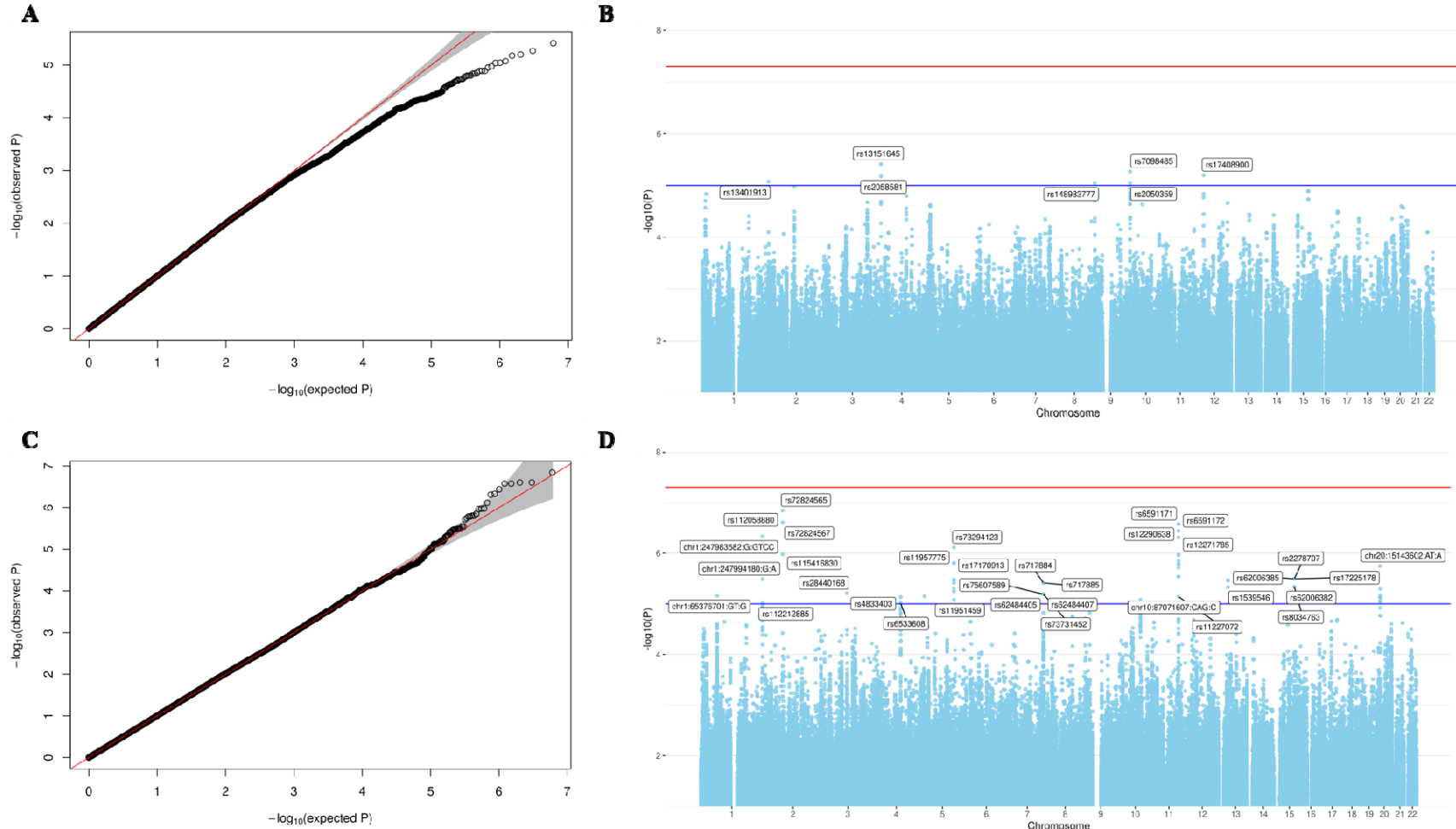
825 **S2 Fig. Functional analysis of top-12 independent (linkage disequilibrium < 0.1) genetic**  
826 **variants nominally associated with gestational duration (GAD) at  $P < 1.0 \times 10^{-5}$ .**

827 **S3 Fig. Distribution of the polygenic risk score for gestational duration among study**  
828 **participants (N=2,212).**



829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842

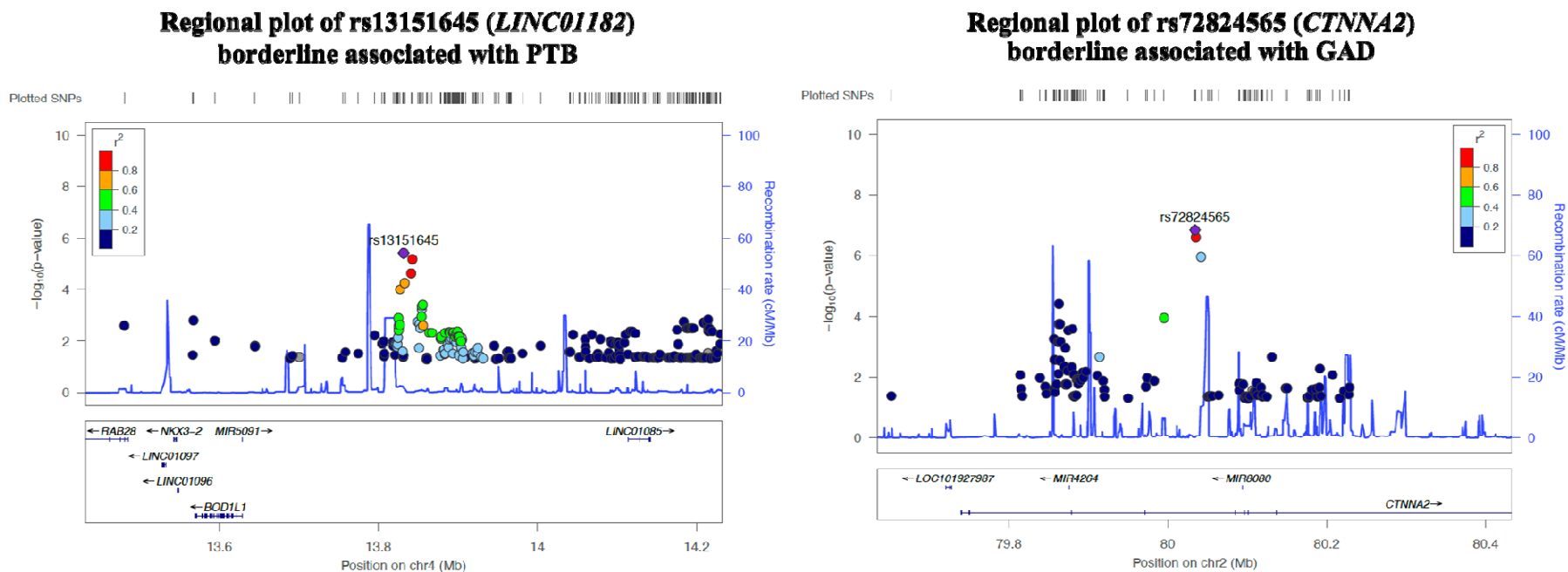
**Fig 1. Genetic ancestry analysis showing the clustering of PAGE samples relative to global populations from different ancestries included in the 1000 Genomes project.** More genetically related populations clustered more closely with each other than genetically distant populations. Most of the PAGE participants (black and grey dots) aligned close to the 1000 Genomes PEL samples (Peruvians from Lima, Peru), and few others were dispersed among samples from other populations in the Americas (CLM Colombians, MXL Mexican Americans). PC1-PC3: first three principal components. Subpopulations: **ACB** African Caribbean in Barbados, **ASW** African Ancestry in Southwest US, **BEB** Bengali in Bangladesh, **CDX** Chinese Dai in Xishuangbanna, **CEU** Utah residents with Northern and Western European ancestry, **CHB** Han Chinese in Beijing, **CHS** Southern Han Chinese, **CLM** Colombian in Medellin, **ESN** Esan in Nigeria, **FIN** Finnish in Finland, **GBR** British in England and Scotland, **GIH** Gujarati Indian in Houston, **GWD** Gambian in Western Division, **IBS** Iberian populations in Spain, **ITU** Indian Telugu in the UK, **JPT** Japanese in Tokyo, **KHV** Kinh in Ho Chi Minh City, **LWK** Luhya in Webuye, **MSL** Mende in Sierra Leone, **MXL** Mexican Ancestry in Los Angeles, **PEL** Peruvian in Lima, **P.JL** Punjabi in Lahore, **PUR** Puerto Rican in Puerto Rico, **STU** Sri Lankan Tamil in the UK, **TSI** Toscani in Italy, **YRI** Yoruba in Ibadan.



843

844 **Fig 2. Quantile and Manhattan plots summarizing results of the GWAS of Preterm Birth (A, B) and Gestational age at**  
 845 **delivery (C, D) conducted among women in the PAGE study (N=2,212).** In the Manhattan plots (B, D), the horizontal blue line  
 846 corresponds to the suggestive genome-wide association threshold at  $P < 1.0 \times 10^{-5}$ ; the horizontal red line is the Bonferroni significant  
 847 threshold at  $P < 5.0 \times 10^{-8}$ . Associations were adjusted for maternal age and the first two genetic PCs. Top SNPs were annotated with  
 848 their rsID using the Haplotype Reference Consortium panel (GRCh37/hg19), or the SNP coordinates reported in TOPMED  
 849 (GRCh38/hg38).





850

851 **Fig 3. Regional plots showing the genetic context of top signals detected with suggestive association with Preterm Birth**  
 852 **(rs13151645 in chr4) and Gestational age at delivery (rs72824565 in chr2).** The colored legend represents the correlation (LD)  
 853 of nearby SNPs with our sentinel SNP shown by the purple diamond. SNPs more correlated with the SNP of interest are shown as red  
 854 dots. Pairwise SNP correlations were calculated using the 1,000 genomes (phase 3) and the admixed American population as  
 855 reference. At the bottom of the plot, showing representative genes in the region and their transcriptional orientation.  
 856

857