Ultra-low coverage genome-wide association study – insights into gestational age using 17,844 embryo samples with preimplantation genetic testing

Shumin Li[1,#], Bin Yan[1,#], Thomas K.T. Li[6,#], Jianliang Lu[1], Yifan Gu[4,5], Yueqiu Tan[4,5], Fei Gong[4,5], Tak-Wah Lam[1], Pingyuan Xie[2,3,*], Yuexuan Wang[1,7,*], Ge Lin[3,4,5,*], Ruibang Luo[1,*]

[1] Department of Computer Science, The University of Hong Kong, Hong Kong, China
[2] Hunan Normal University School of Medicine, Changsha, 410013, Hunan, China
[3] National Engineering and Research Center of Human Stem Cell, Changsha, Hunan, China
[4] NHC Key Laboratory of Human Stem Cell and Reproductive Engineering, School of Basic Medical Science, Institute of Reproductive and Stem Cell Engineering, Central South University, Changsha 410008, Hunan, China
[5] Clinical Research Center for Reproduction and Genetics in Hunan Province, Reproductive and Genetic Hospital of CITIC-Xiangya, Changsha 410013, Hunan, China
[6] Department of Obstetrics & Gynecology, Queen Mary Hospital, The University of Hong Kong, Hong Kong, China
[7] College of Computer Science and Technology, Zhejiang University, Hangzhou, China

[#] These authors contributed equally to this work
[*] To whom correspondence should be addressed: Ruibang Luo, Email: rbluo@cs.hku.hk, Ge Lin, Email: linggf@hotmail.com; Yuexuan Wang, Email: amywang@hku.hk, Pingyuan Xie, Email：plainxie192@126.com

Abstract

Background: Very low coverage (0.1 to 1x) whole genome sequencing (WGS) has become a promising and affordable approach to discover genomic variants of human populations for Genome-Wide Association Study (GWAS). To support genetic screening using Preimplantation Genetic Testing (PGT) in a large population, the sequencing coverage goes below 0.1x to an ultra-low level. However, its feasibility and effectiveness for GWAS remains undetermined.

Methods: We devised a pipeline to process ultra-low coverage WGS data and benchmarked the accuracy of genotype imputation at the combination of different coverages below 0.1x and sample sizes from 2,000 to 16,000, using 17,844 embryo PGT with approximately 0.04x average coverage and the standard Chinese sample HG005 with known genotypes. We then applied the imputed genotypes of 1,744 transferred embryos who have gestational ages and complete follow-up records to GWAS.

Results: The accuracy of genotype imputation under ultra-low coverage can be improved by increasing the sample size and applying a set of filters. From 1,744 born embryos, we identified 11 genomic risk loci associated with gestational ages and 166 genes mapped to these loci according to positional, expression quantitative trait locus and chromatin interaction strategies. Among these mapped genes, *CRHBP, ICAM1* and *OXTR* were more frequently reported as preterm birth related. By joint analysis of gene expression data from previous studies, we constructed interrelationships of mainly

CRHBP, ICAM1, PLAGL1, DNMT1, CNTLN, DKK1 and *EGR2* with preterm birth, infant disease and breast cancer.

Conclusions: This study not only demonstrates that ultra-low coverage WGS could achieve relatively high accuracy of adequate genotype imputation and is capable of GWAS, but also provides insights into uncovering genetic associations of gestational age trait existed in the fetal embryo samples from Chinese or Eastern Asian populations.

Keywords: Ultra-low coverage Whole Genome Sequencing, Imputation, Single Nucleotide Polymorphisms, Genome-Wide Association Study, Gestational Age, Preterm Birth

Background

Detection and characterization of genetic variants associated with traits and diseases are fundamental to the study of human genetics. Genome-Wide Association Study (GWAS) is an approach widely used in genetic research that aims to decode the associations of specific genetic variations with particular diseases or traits in sample populations. In the past decade, GWAS has facilitated discovery of over one hundred thousand variants associated with complex traits in human (1). Whole-genome sequencing (WGS) has emerged as a dominant technology in GWAS because it enables one to generate a comprehensive view of the genomic variation landscape for not only a specific trait but also for common diseases. Thus, WGS-based approaches hold a significant advantage over genome-wide genotyping arrays or exome sequencing in the analysis of complete genetic variations. However, with a fixed budget, the high cost of sequencing many DNA samples is a limitation for GWAS (2-4). Recently, to reduce the cost of sequencing, a number of low (0.5-1x) or extremely low-coverage (0.1-0.5x) WGS has been carried out as an alternative method of genotyping (2, 5, 6). It is, however, unclear whether ultra-low coverage WGS (ulcWGS) below 0.1x data can capture enriched genetic variations across the entire allele frequency (AF) spectrum. When considering a balance between number of samples sequenced and sequencing read coverage, effective genotype imputation could provide more authentic DNA variants that would be helpful for genetic research.

Genotype imputation can be used to infer missing genotypes and to increase the accuracy of detecting genetic variants, such as single nucleotide polymorphisms (SNPs).

96 In general, performance of genotype imputation is largely affected by sample size,
97 sequencing coverage, analysis methods, and other parameters (7). A main challenge to
98 use very low coverage WGS is how to achieve an adequately accurate imputation for
99 downstream analyses. Previous attempts have shown the efficiency of low coverage
100 WGS, for example, a high r^2 of imputation accuracy observed by using 10 low coverage
101 WGS (~0.5x) as compared to known genotypes (6). Pasaniuc et al. reported that the
102 GWAS signals obtained from using 909 whole-exome sequencing (~0.24x) are
103 comparable to using genotyping array (2). Gilly et al. found that more true association
104 signals were identified by WGS (~1.0x) than the traditional array-based study (5).
105 Using ulcWGS (0.06x-0.1x) with 141,431 samples from a Chinese genomic study, the
106 accuracy of imputed genotypes reached 0.71 (8). Even though the distribution of
107 genetic background from large number of samples is expected to compensate for the
108 low sequencing coverages, it has never been determined how many samples are needed
109 to achieve a relatively high accuracy. More importantly, lack of comparative data with
110 coverages less than 0.05x results in the limited application of ulcWGS to GWAS.

111 Gestational age is an important complex trait associated with biological processes
112 and human disease. Biologically, gestational duration plays a vital role in both mental
113 and physical health of children at an age of five-year-old (9). Gestational age shorter
114 than 37 weeks is categorized as Preterm Birth (PTB). Previous studies found the
115 contribution of both the maternal and fetal genomes to variation of gestational ages (10-
116 12). However, they focused on European and African samples by involving few
117 samples from Chinese or Eastern Asian ancestries. Overall, biological mechanisms

118 underlying variation of gestational durations remain unclear, primarily because
119 insufficient maternal or fetal genotypes with widespread gestational ages have been
120 collected (13). Recently, Preimplantation Genetic Testing (PGT) with trophectoderm
121 biopsy for embryo aneuploidy screening has become a common practice in *in vitro*
122 fertilization (14, 15), and poses as an expectant source of genotypes for GWAS.
123 However, if the average sequencing coverage of PGT is even lower than the lowest
124 levels that have been reported in GWAS so far, it is necessary to examine whether such
125 PGT datasets are appropriately applied to GWAS.

126 In this study, we devised a pipeline for analyzing and applying the ulcWGS of
127 17,844 embryo samples for GWAS. Our result shows that a large sample size is
128 effective to increase the accuracy of genotype imputation even at an ultra-low coverage.
129 Furthermore, using the imputed genotypes of 1,744 embryos that were successfully
130 transferred and born with a widespread of gestational ages, we demonstrate the power
131 of using ulcWGS in GWAS and provides insights into understanding genetic
132 association of gestational age in embryos acquired from Chinese and East Asians.
133 Refreshing the lowest coverage used in GWAS, our finding also provides a foundation
134 for exploring the utilization of an even lower coverage for dissecting genotype-
135 phenotype associations.
136

137 # Methods

138 ## Samples and sequencing coverage

139 The whole PGT dataset of 17,844 embryos was from the Clinical Research Center for
140 Reproduction and Genetics in Hunan Province, Reproductive and Genetic Hospital of

141 China International Trust Investment Corporation - Xiangya. The protocol of embryo
142 culture and biopsy was published in a related study (16). Three WGA kits were applied
143 to the biopsied TE cells by following the manufactures' guides, including REPLI-g
144 Mini Kit (called MDA), WGA4 GenomePlex Single Cell Whole Genome
145 Amplification Kit (called dop-PCR) and Rubicon Genomics PicoPlex Single Cell
146 Whole Genome Amplification Kit (called PicoPlex). A 1-2 μg of the WGA product was
147 subjected to library construction and sequencing on the four platforms, including BGI-
148 Seq 500, Illumina MiSeq, Ion Proton, Ion Torrent (Additional file 1: Table S2).
149

150 **Study design**

151 We developed a three-step pipeline to carry out genotype imputation using ulcWGS
152 data and to perform GWAS of the detected SNPs (Fig. 1). Firstly, the raw reads of the
153 17,844 embryo samples were aligned to the hs37d5 reference genome. Sequencing
154 coverage of these embryo samples displays a distribution with an average coverage
155 0.04x (Additional file 3: Figure S1). To our best knowledge, it is below the coverage of
156 any dataset used previously for genotype imputation. After removing potential PCR
157 duplicates, the aligned reads were used to call population SNPs. Secondly, we
158 conducted genotype imputation on each individual sample at the called population
159 SNPs and assessed the genotyping accuracy based on the standard Chinese sample
160 HG005 with known genotypes from GIAB (Genome in a Bottle, NIST)(17). Last, we
161 applied the imputed genotypes from 1,744 born embryos with complete follow-up

162 records to GWAS and explored biological associations between the genetic variants
163 detected in the born embryos and their gestational ages.
164

165 **Sequencing read processing and alignment**

166 In the first step of Fig. 1, the raw reads of the PGT samples were delivered in two
167 types, fastq or BAM. For BAM data, we used bedtools (18) to extract the raw reads into
168 single-end fastq files. The raw reads of each sample were then aligned to the hs37d5
169 reference genome using BWA (19). BWA-mem was applied to samples sequenced by
170 the Ion Torrent with longer reads and "bwa aln" was used for the rest of the samples
171 with shorter reads. Samtools rmdup (20) was used to remove the potential PCR
172 duplicates.
173

174 **Population SNP calling**

175 In the population SNP calling stage, we modified the method of Liu et al. (8). The first
176 stage is to use log-likelihood estimation for AF estimation, and the second stage is to
177 use log-likelihood ratio test for determining allelic types. More details are described in
178 Additional file 2: Supplementary Methods.

179 SNP calls of the raw population were filtered following the rules, 1) calls that
180 overlapped with the 35-kmer problematic alignment regions in hs37d5 were removed
181 (21); 2) calls that overlapped with regions with ENCODE mappability uniqueness
182 score unequal to 1 were removed (22), tool "bigWigToBedGraph" used afterward to
183 convert the bigwig into bed format) (23).
184

185 **Genotype imputation**

186 We used STITCH (24) version 1.5.7 for genotype imputation. The ancestral haplotypes
187 number *k* was set as 20, the assumed number of generations *nGen* was set to 2000, and
188 the reads were binned into windows with *gridWindowSize* 10000. The *diploid* mode in
189 STITCH was used. Although using a reference panel is optional in STITCH, we used
190 the IMPUTE2 1000 genome haplotypes phase 1 reference panel as it improves the
191 accuracy of imputation when the sample size is small. With sample size larger than
192 10,000, the improvement was not significant. All parameters were optimized by
193 maximizing the r^2 of the estimated AFs between imputation and population SNP
194 calling in a randomly chosen 5 Mbp genomic region (chr3:180-185Mbp). When
195 applying STITCH to the whole genome, we divided the genome into 5 Mbp windows
196 with a 500 Kbp overlap between two windows.

197 For benchmarking, seqtk was used to subsample the HG005 Illumina WGS raw
198 reads to 0.01x-0.1x (Paired-end 250bp, 300-fold). We aligned the reads to hs37d5 by
199 using the same pipeline as used in the embryo samples. Because a computer takes a few
200 years to impute whole genome with tens of thousands of samples, we worked on only
201 chromosome 1, the longest one in human. The genotype imputation was benchmarked
202 according to 80 combinations of the 10-scale coverages of HG005 with 8 sizes of our
203 samples from 2,000 to 16,000, respectively.

204 All bi-allelic SNPs with MAF ≥ 0.01 found in chromosome 1 with the population
205 SNP calling from the 17,844 samples are included for genotype imputation. We used
206 the 167,814 SNPs both in our SNP callset and HG005 known genotypes for

207 benchmarking. The imputed genotypes were compared to the truth released by GIAB
208 for estimating the imputation accuracy. Then, genotype imputation was applied to all
209 17,844 embryo datasets at the 31,622,332 bi-allelic sites with MAF ≥ 0.01 found in
210 population SNP calling. The entire imputation process spent 19 days and used 15
211 machines with 16 cores (two 8-core Intel Xeon Silver 4108 CPU). Two filters "INFO
212 score ≥ 0.4 and HWE *p*-value >1e-6" were applied to select the imputed genotypes.
213

214 **Genome-wide association study (GWAS)**

215 To conduct GWAS, we used score statistics (25) that is implemented in ANGSD (26).
216 Variants satisfying four conditions were selected as inputs, including 1) known in
217 dbSNP150, 2) MAF ≥ 0.01, 3) INFO score ≥ 0.4, and 4) HWE *p*-value > 1e-6. To
218 remove biases, we specified 16 covariates for ANGSD, 8 most significant principal
219 components calculated from the inputs of PCA (Additional file 2: Supplementary
220 Methods), and 8 clinical records including maternal age, maternal BMI, fetal sex, either
221 parent with single-gene disease, either parent with chromosome abnormality, multiple
222 pregnancy, preeclampsia and gestational diabetes of mellitus. Except the default
223 parameters, we set *minHigh* to 15 (requiring at least 15 high credible genotypes from
224 the input) instead of 10 to achieve better accuracy with a large sample size. ANGSD
225 did not create an output beta-coefficient, so we followed the ideas of Skotte et al. (25)
226 by incorporating the genotype probabilities and all 16 covariates into a linear regression
227 model, with the gestational age as a response variable. The effect size was calculated
228 by the coefficient of genotypes.

229 **Independent SNPs and genomic risk loci**

230 The significant SNPs from our GWAS were mapped to genomic risk loci using FUMA
231 pipeline (27) and the LD information of 1000G EAS variants (28). We first defined
232 "independent significant SNP", a SNP that meets genome-wide significance level
233 (p value $\leq 4.515e-8$) and is independent of other significant SNPs (with LD $r^2 <$
234 0.6). FUMA also generated a set of lead SNPs with low LD and with other ($r^2 < 0.1$)
235 from the independent significant SNPs. The genomic risk loci were identified by
236 starting from these lead SNPs and through iteratively merging related genomic regions
237 to them according to FUMA's rules.

238 Also, the FUMA pipeline sorted out a set of candidate SNPs from our inputs
239 that meets one of two conditions, 1) the independent significant SNPs and 2) SNPs that
240 are linked to the independent significant SNPs (with LD $r^2 \geq 0.6$). For condition 2,
241 the SNPs can be from our imputed genotypes if p value below 0.05 or from the reference
242 panel of 1000G EAS. ANNOVAR was used to annotate the candidate SNPs (29).

243

244 **Functional annotation of the mapped genes**

245 DAVID online tool (30) was used to analyze the enrichment of Gene Ontology (GO)
246 biological processes and KEGG pathways for the coding genes mapped to the risk
247 loci.

248

249 **Gene mapping**

250 We used three gene-mapping strategies provided by FUMA (27), including positional,
251 expression quantitative trait locus (eQTL) and chromatin interaction. For positional

252 mapping, ANNOVAR annotations were used. The candidate SNPs were mapped to the
253 nearest genes within a maximum 10 Kbp distance. For eQTL mapping, expression data
254 of all tissue types in GTEv6, GTEv7, and GTEv8 (31) were used. We required False
255 Discovery Rate (FDR) < 0.05 and p value <0.001 for a valid eQTL mapping. All
256 chromatin interaction data in FUMA were used (32-35). The promoter was set to
257 upstream 2000 bp to downstream 500 bp of transcriptional starting sites. We required
258 FDR < 1e-6 for a valid chromatin interaction mapping.
259

260 **Analysis of genome-wide mRNA expression data**

261 We first extracted the genome-wide microarray and RNA-seq data of human mRNA
262 expression from GEO/NCBI database. The mRNA data includes three subsets in
263 maternal PTB, infant PTB and breast cancer (Additional file 1: Table S8). Based on the
264 normalized expression data provided by the database, we analyzed differentially
265 expressed genes (DEGs) between different conditions, including 1) PTB vs. normal
266 term, 2) BPD or sepsis vs. infant without BPD or sepsis, and 3) breast cancer vs. control
267 samples. For microarray platform-based data, we used the *limma* package in R
268 programming language and conducted empirical Bayes moderated t-test. DEGs were
269 detected with a fold change above 1.5 and p value below 0.05. For RNA-seq data with
270 raw counts, we utilized the Edge-R method to identify DEGs. The DEGs are listed in
271 Additional file 1: Table S9.
272

Results

Benchmarking genotype imputation using the ultra-low coverage sequencing data of 17,844 embryos and HG005

We estimated the accuracy of the imputed genotypes from the SNPs of chromosome 1 both called in our 17,844 samples and the known genotypes in HG005. The genotype imputation was benchmarked according to 80 combinations of the 10-scale coverages of HG005 with 8 sizes of our samples. A monotonic increase in accuracy with sequencing coverage was observed (Fig. 2a), consistent with previous studies (2, 24). The accuracy for sample size of 2,000 stayed at around 0.48 under all sequencing coverages. But for a larger sample size of 16,000, its accuracy increased from 0.48 at 0.01x to 0.66 at 0.1x. This result suggests that at ultra-low coverages, increase in sample size could obtain higher accuracy (Additional file 1: Table S1). In general, a lower coverage with a larger sample size results in better performance than a higher coverage with a smaller sample size. For example, the genotype accuracy at 0.05x with 14,000 samples versus 0.1x with 4,000 embryos was 0.61 versus 0.55. A larger sample number is therefore more efficient in optimizing genotype imputation than increasing sequence coverage. It is also noticed that at the two lowest coverages in our experiments, the contribution of increasing sample size was not significant and the accuracy plateaued at 0.52 (0.01x) and 0.55 (0.02x). Using the same datasets, we evaluated allele accuracy that relaxed zygosity correctness from genotype accuracy. The corresponding accuracies were much better (increased to 0.7 and higher) while maintaining the same trend with increasing sample size and coverage (Fig. 2b). Therefore, when the

295 genotypes are incorrectly imputed for some SNPs, the non-reference allele could be
296 correctly detected.

297 We next examined several important quality metrics that are widely used to filter
298 falsely imputed genotypes. The INFO score of IMPUTE2 style (36) denotes the
299 certainty of an imputed genotype and has been accepted as a quality metric of
300 imputation. For a combination of 0.04x coverage with 16,000 samples, we
301 benchmarked ten different INFO score thresholds from 0.1 to 1.0 and detected
302 corresponding SNPs. As increase in the INFO scores, we observed a consistent increase
303 in genotype accuracy from ~0.60 to 0.99, but a rapid decrease in number of SNPs that
304 meet these thresholds (Fig. 3a). Thus, INFO score could act as an effective metric to
305 evaluate the accuracies at ultra-low coverage, but its thresholds should be meticulously
306 chosen in order to retain sufficient SNPs. Effect of MAF scores on genotype accuracy
307 of HG005 were then tested as a potential metric. We divided the genotyping results at
308 different sequencing coverages (sample size fixed to 16,000) into bins of MAF ranges
309 (0.01 MAF a bin) and calculated the genotype accuracy each bin. The genotype
310 accuracy increased rapidly from MAF 0 to 0.05 and reached a turning point at 0.05.
311 After this point, the accuracy became slow increasing (Fig. 3b). However, even for the
312 most common SNPs (MAF 0.4~0.5), the accuracy was converged at ~70%. The
313 accuracy of the two lowest coverages 0.01x and 0.02x fluctuated especially at low MAF
314 cutoffs. Because such fluctuation was not observed during INFO testing, MAF might
315 not be a reliable metric to change genotype accuracy at ultra-low coverage. In
316 subsequent analyses, we followed the common practice to use SNPs with MAF ≥ 0.01

317 for GWAS. Finally, we combined HWE *p*-values with INFO scores as a filter but
318 without losing too many SNPs. HWE *p*-values could evaluate the probability of the
319 imputed genotype at a certain SNP that is significantly different from the expectations
320 by Hardy-Weinberg Equilibrium. We summarized the genotype accuracy of different
321 combinations of INFO scores and HWE *p*-value cutoffs in Table 1. When INFO scores
322 were set above 0.4, the accuracies of genotype and allele were 70.0% and 83.4%,
323 respectively, with 48,176 SNPs left. The "INFO score ≥ 0.4 and HWE *p*-value >1e-6"
324 resulted in an increased accuracy 71.5%, with 28773 SNPs left. Thus, our GWAS
325 utilized this setting, "INFO score ≥ 0.4 and HWE *p*-value >1e-6" as filtering criteria.
326 To summarize our benchmarking results for future study on ulcWGS, we built a
327 regression model (Formula 1) to calculate the expected genotype accuracy using
328 sequencing coverage and sample size as inputs.

$$acc = 2.227 * c + 8.937e^{-6} * s + 0.494 \ \ (c \geq 0.01, s \geq 4000) \qquad (1)$$

329 where *acc* is the expected genotype accuracy, *c* denotes the sequencing coverage, *s*
330 denotes the sample size. The model has a r^2 of 0.874 (Additional file 3: Figure S2).
331

332 **GWAS of gestational ages using 1,744 born embryos**

333 With the solid foundation laid out in the previous section, we have obtained sufficient
334 good quality SNPs for GWAS. Among the 17,844 sequenced embryos, 1,744 were
335 transferred and gave birth to a baby. The gestational age of all 1,744 born embryos are
336 well documented and thus were chosen for biological associated study. We revised the
337 population SNP calling method used in Liu et al. (8). A total of 151,793,444 SNPs were
338 detected and of 141,718,305 are bi-allelic. The MAF spectrum bi-allelic novel and

339 known variants in dbSNPv150 (37), 1000G (28) and gnomAD (38) is shown in
340 Additional file 3: Figure S3a. The transitions/transversions ratio for "all bi-allelic SNPs"
341 and "bi-allelic SNPs known in dbSNPv150" were 3.05 and 3.58, respectively. Pearson
342 correlation coefficient of the non-reference AFs between the 301 Chinese samples in
343 1,000G (so-called 1000G CHN (28)) and the corresponding SNPs obtained in our
344 dataset was 0.986 (Additional file 3: Figure S3b). This result supports a strong
345 correlation between the two datasets, and high confidence of the known variants used
346 in our analysis. We also performed genotype imputation at bi-allelic population SNPs
347 with MAF ≥ 0.01. Pearson correlation coefficient was 0.985, showing a high
348 consistency between the estimated AFs in population SNP calling and in imputation.

349 Three different whole genome amplification (WGA) methods and four different
350 sequencing platforms were used in the PGT dataset (Additional file 1: Table S2). It is
351 not uncommon that large number of samples may use multiple sequencing platforms
352 and WGA. Removing these unrelated covariates from GWAS as much as possible is
353 essential especially when the sequencing coverage is ultra-low. Such covariates should
354 be detected and disregarded in GWAS. We applied Principal Component Analysis (PCA)
355 to the imputed genotypes of SNPs with MAF ≥ 0.05 among all 17,844 embryo samples
356 (Additional file 2: Supplementary Methods, Additional file 3: Figure S4a). The first and
357 second principal components distinguish the differences of sequencing platforms
358 (Additional file 3: Figure S4b) and of WGA methods (Additional file 3: Figure S4c).
359 Therefore, we used the top eight principal components and eight other clinical records
360 as covariates in GWAS. PCA was also applied to the GWAS samples and the top

361 principal components were included in the subsequent analyses as covariates for
362 removing the biases.

363 We used the state-of-the-art one-stage GWAS strategy (39) to analyze the
364 1,107,198 imputed SNPs in the 1,744 transferred and born embryo samples with
365 complete follow-up records. The distribution of gestational ages shown in Additional
366 file 3: Figure S5 include 162 preterm deliveries (gestational age < 37 weeks), 42 early
367 preterm deliveries (gestational age <34 weeks), and 8 very early preterm deliveries
368 (gestational age <28 weeks). The gestational ages were standardized by z-score and
369 incorporated as a quantitative trait.

370 A total 1,107, 198 SNPs with imputed genotypes were selected for GWAS that
371 are in accord with, 1) MAF ≥ 0.01, 2) known in dbSNPv150, and 3) passed the filter
372 "INFO score ≥ 0.4 and HWE *p*-value >1e-6". The Q-Q plot shows a large deviation of
373 the observed p values from the null hypothesis (Additional file 3: Figure S6). The
374 linkage disequilibrium score regression (LDSC) software package (40) with 1000G
375 EAS reference was used to estimate $\lambda_{GC} = 0.992$, mean $\chi^2 = 1.012$. The LD score
376 regression intercept was 0.952, standard error = 0.021, indicating that the population
377 stratification and other factors were well-controlled. We identified 40 significant SNPs
378 satisfying Bonferroni-corrected significant levels of 4.515e-8. The Manhattan plot
379 shows the distribution of the detected SNPs cross all chromosomes (Fig. 4a, Additional
380 file 1: Table S3).

381 We used FUMA (27) pipeline for GWAS downstream analysis. First, FUMA
382 generated a set of candidate SNPs and the 11 independent SNPs (Additional file 1:

383 Table S4). By annotation, most of the SNPs are located and enriched in intergenic and
384 intronic regions (Fig. 4b), which is similar to the previous study (41). Specifically, there
385 are 11 SNPs located in exons (0.8% of total), and 4 of them are non-synonymous SNPs
386 (Additional file 1: Table S5). Additionally, we observed the distribution of regulatory
387 elements and chromatin states with the candidate SNPs (Additional file 3: Figure S7).
388 According to RegulomeDB scores assigned to each candidate SNP, 1.09% SNPs were
389 classified as likely to affect regulator binding (score 2a and 2b) and 0.21% as likely to
390 affect regulator binding and linked to expression of a gene target (score 1d and 1f).
391 These proportions of SNPs hold a relatively high likelihood to affect the regulatory
392 elements along noncoding regions. Second, we identified 11 leading SNPs from their
393 corresponding genomic risk loci by FUMA. Fig. 4e shows the DNA length, number of
394 SNPs and mapped genes of these risk loci. The zoom in locus plot of the 11 risk loci
395 are shown in Additional file 3: Figure S8. Of 4 risk loci were reported to be associated
396 with refractive astigmatism, adolescent idiopathic scoliosis, glomerular filtration rate,
397 among others, indicating a possible connection of these diseases or traits with PTB
398 (Additional file 1: Table S6).

399 By integrating strategies positional, eQTL and chromatin interaction mappings,
400 we identified a total of 166 genes mapped to the 11 risk loci, including 48 and 19 genes
401 from two and three strategies, respectively (Fig. 4c, Additional file 1: Table S7). There
402 were 24 genes detected within or less than 10 Kbp from the candidate SNPs and 7 of
403 them were shown in Fig. 4a based on the location of the genomic risk loci. Importantly,
404 *CNTLN* was reported as a PTB related gene (42), and *PIN1* involves inhibition of breast

405 cancer (43). A Circos plot shows the graphic distribution of the mapped genes via eQTL
406 and chromatin interaction, and their links with the genomic risk loci (Fig. 4d). The
407 breakdown of each chromosome is shown in Additional file 3: Figure S9. Even though
408 not within any risk loci, *CRHBP* was linked through chromatin interaction mapping to
409 two loci, chr5:75101342-75164623 and chr5:78251511-78271282. Enrichment
410 analysis of DEGs in 30 tissue types in GTEx v8 (44) exhibits significant overexpression
411 of the mapped genes in both ovary and uterus (Additional file 3: Figure S10a). The
412 gene-set enrichment analysis also indicates their association with the immune system,
413 breast cancer and transcriptional regulation (Additional file 3: Figure S10b).
414

415 **Association of the 166 mapped genes from GWAS with preterm birth,**
416 **infant disease and breast cancer**

417 GWAS of gestational age related PTB has been implicated in biological functions that
418 include immune response, inflammatory response, and coagulation factors (11, 45-47).
419 We compared our 166 mapped genes with reported PTB markers by collection of 8
420 published resources, here classed to 3 PTB sets, including dbPTB from Sheikh et al.
421 (48) and Uzun et al (49), PTB-merged from 5 data resources (10), and PNAS-identified
422 DEGs of PTB in 2019 (10). We found that *CRHBP, EMR1, ICAM1, MBL2, OXTR,* and
423 *THBS4* have been reported in at least 2 PTB sets. Specifically, *ICAM1* was present in
424 all 3 sets and 6 data resources, and both *CRHBP* and *OXTR* in 5 resources (Additional
425 file 1: Table S7). In addition, there were 8 genes overlapped with 1 PTB set (Fig. 5a).
426 This result pinpoints a relationship of the detected risk loci with PTB, and possible roles
427 of the overlapped genes in PTB. There are totally 1,930 genes reported as PTB-related;

428 however, only 50 were frequently recognized by at least 5 data resources, hereafter
429 referred to as PTB marker genes (Additional file 1: Table S7). The 50-PTB marker set
430 was significantly enriched with inflammatory and immune response related processes
431 or pathways (Fig. 5b). Similarly, those PTB genes that overlapped with 3 or 4 resources
432 mainly participate in the same biological functions. The PTB-related genes listed at Fig.
433 5a involve immune response (*EMR1, ICAM1, PTPRZ1* and *MBL2*), inflammatory
434 response (*CRHBP, ICAM1, PTPRZ1* and *MBL2*), coagulation (*MBL2*), apoptosis
435 (*PLAGL1*) and cell adhesion (*ICAM1* and *THBS4*), emphasizing their associations with
436 PTB.

437 To determine the relationship between the mapped genes and PTB, we first
438 analyzed the DEGs between maternal PTB and normal term birth based on six
439 published datasets of genome-wide gene expression (Additional file 1: Tables S8 and
440 S9). Significant overexpression of *CRHBP* and *OXTR* were identified in two datasets,
441 while overexpressed (*ICAM1, EGR2,* and *PLAGL1*) and underexpressed (*THBS4*)
442 genes were present in one dataset (Table 2). Another *DKK1* expressed higher levels
443 above 2.0-fold in two datasets, suggesting its importance in PTB. Then, we analyzed
444 four infant PTB datasets of gene expression, and found differentially increased
445 expression of *ICAM1, CRHBP, DKK1, EGR2,* and *PLAGL1,* consistent with maternal
446 PTB. In contrast, significantly underexpression of *THBS4, PIN1* and *GCNT4* was
447 commonly detected by maternal and infant subgroups (Table 2). It is also noticed that
448 *DNMT1* shows increased expression in both fetal and maternal groups (Table 2). DNA
449 methylation was suggested to involve generation of early PTB (10); however, the role

450 of *DNMT1* as a DNA methyltransferase in PTB has not been determined yet. This
451 analysis provides evidence that these mapped DEGs above including *DNMT1* are
452 associated with both maternal and infant PTB. Relatively, *OXTR* seems to be related to
453 only maternal factor. A heatmap of mRNA expression shows clustering of 6 expression
454 profile cross PTB and normal term conditions (Fig. 5c). Two clusters display distinct
455 expression patterns of these PTB genes.

456 It is well known that bronchopulmonary dysplasia (BPD) is the most common
457 respiratory disorder among children born preterm (50, 51). The pathogenesis of BPD
458 involves multiple prenatal and postnatal mechanisms affecting the development of
459 immature lung. Also, neonatal sepsis is associated with severe morbidity and mortality
460 during the neonatal stage. The incidence of late-onset sepsis increases with increase in
461 survival rate of preterm and low weight babies (52). Thus, we examined the possible
462 relationship of the 166 genes with infant BPD and sepsis by analyzing gene expression
463 data derived from samples of preterm infants (Additional file 1: Table S8). We first
464 identified differentially increased expression of *CRHBP, ICAM1* and *EGR2* under PTB
465 of maternal & infant and BPD conditions, but differentially decreased expression of
466 *DKK1* (Table 2). Under sepsis condition, differentially overexpressed *CNTLN, ICAM1,*
467 and *PLAGL1* in PTB were consistently observed. By contrast, *GCNT4* expression was
468 always significantly decreased under maternal and infant subpopulations. The gene
469 clustering shows diversity of gene expression across different samples with infant PTB;
470 however, these up-regulated genes were grouped together (Fig. 5d), supporting the idea
471 that BPD or sepsis induces the change in these PTB genes.

472 Breast cancer is one of the most frequently diagnosed malignancies observed
473 during pregnancy. It often presents characteristics of high malignancy and are hormone
474 receptor negative like Estrogen receptor (ER)-, HER2+ or triple negative breast cancer
475 (TNBC). We collected gene expression data mainly presenting three subtypes of breast
476 cancer, TNBC, HER2+ and ER or Progesterone receptor (PR) (Additional file 1: Table
477 S8). By analysis of DEGs, we detected the expression of PTB related genes *DKK1,*
478 *ICAM1, DKK1, EGR2, PLAGL1, GCNT4* and *THBS4* in all three subtypes. Increased
479 *ICAM1* and decreased *PLAGL1* were consistently identified in these datasets (Table 2).
480 In fact, *ICAM1* has been reported as TNBC markers (53) and acts as prognostic
481 molecule of breast cancer (54). Both *ICAM1* and *DKK1* could increase expression in
482 TNBC cells (55). However, other PTB related genes do not display similar changes in
483 expression under the cancer subtypes. For example, *OXTR* was detected by only one
484 dataset of ER+/-, while *CRHBP, EMR1, PIN1* and *MBL2* were not found among any of
485 the subtypes. Conversely, underexpression of *PLAGL1* and *GCNT4* were found in all
486 three types of breast cancer. In addition, overexpression of *DNMT1* was observed in
487 TNBC and HER2+ subtypes, that validates its oncogenic roles in breast cancer and drug
488 target of TNBC (56, 57).

489 To identify further interactions between the selected PTB genes from Table 2 and
490 the top 50 PTB markers, we calculated Pearson correlation coefficient by comparing
491 their gene expression of maternal and infant PTB, BPD and sepsis subsets, respectively.
492 We built the corresponding co-expression networks of the PTB genes (Additional file
493 3: Figure S11). Clearly, the co-expressed genes involve immune and inflammatory

494 responses, coagulation, as well as apoptosis, angiogenesis, among others. We then
495 constructed the networks cross different subpopulations. As shown in Fig. 6a, the PTB
496 genes could involve both maternal and infant PTB processes, especially *ICAM1*,
497 *PLAGL1, EGR2* and *CRHBP* that link to TLR4, a known preterm marker associated
498 with immune and inflammatory processes (58). Similarly, co-expression interactions of
499 these genes with the top PTB markers were observed under maternal PTB and infant
500 BPD (Fig. 6b). Both *OXTR* and *PIN1* only display gene correlations under maternal
501 PTB, whereas *MBL2* only correlates under infant PTB. This analysis indicates that the
502 predicted PTB-related genes including *ICAM1*, *CRHBP*, *PLAGL1, EGR2, CNTLN,* and
503 *DKK1*, play an important role in preforming biological activities associated with PTB,
504 infant disease, possibly breast cancer, due to the gestational age-induced.
505

506 ## Discussion

507 Low or very low coverage sequencing data have been increasingly used in discovery of
508 genetic variation and GWAS. However, if the coverages are further decreased to
509 extremely ultra-low levels, how many samples are required to obtain a relatively high
510 accuracy of genotype, and whether these samples can be appropriately applied to
511 GWAS. To address this challenge, we used 17,844 PGT datasets of embryo samples
512 with an average of 0.04x coverage and achieved genotype imputation performance
513 comparable to those using low coverage samples. We first demonstrate that increase in
514 number of ulcWGS samples is more efficient than changing sequencing coverages and
515 indicates the ability of such ultra-low coverage PGT samples to obtain adequate
516 accuracy of phenotypes. Furthermore, we found that INFO score and HWE *p*-value are

517 effective to act as filters to improve the accuracy at ultra-low coverages while
518 meanwhile keep enough SNPs for downstream analyses. To our best knowledge, the
519 samples used in this study hold the lowest average WGS coverage for GWAS so far,
520 and our study provides a framework for guiding other researchers who work on
521 ulcWGS data.

522 Gestational age is a multi-factor phenotype involved in maternal and fetal
523 biological activities (59). To investigate effects of the fetal genome on gestational age,
524 we used the imputed genotypes of 1,744 born embryo samples, a cohort of samples who
525 successfully gave birth after *in vitro* fertilization. From the 166 genes mapped to the 11
526 genomic risk loci, we identified a set of the PTB-related genes that were previously
527 reported (12, 59). *CRHBP, ICAM1,* and *OXTR* are primary representative genes
528 showing evidence of genetic association with gestational age among our samples.
529 *CRHBP* is an important gene in maternal and fetal gestation that could regulate the
530 pregnancy length by increasing/decreasing the concentration of CRH (60, 61). *ICAM1*
531 involves disease induced PTB during pregnancy (62-64). Oxytocin signaling is
532 mediated by Oxytocin receptor (*OXTR*), which is related to gestational age (65). We
533 validated these predicted PTB genes by analysis of DEGs from maternal and infant
534 PTB subpopulations.

535 PTB is neonatal birth occurring before 37 weeks of gestation age and is a leading
536 cause of infant morbidity and mortality. Understanding of genetic and molecular
537 mechanisms of PTB and its association with gestation duration is currently insufficient.
538 Recently, Zhang et al. reported replicable loci in six genes (*EBF1, EEFSEC, AGTR2,*

539 *WNT4, ADCY5* and *RAP2C*) associated with gestational duration (12). This study
540 identified 3 genes (*EBF1, EEFSEC* and *AGTR2*) strongly associated with PTB in a
541 European ancestry cohort of 43,568 women. However, we did not detect such 6 genes
542 from the reported PTB markers and only identified them as DEGs from few published
543 datasets of PTB. This could be due to the heterogeneous sources of samples used in
544 different PTB studies or be explained in part by the genetic complexity of incomplete
545 gestation induced PTB complications. Here, we collected almost 2,000 PTB-related
546 genes from previous PTB studies. The differences in study design, source and subtype
547 of samples, and statistical methods would be important factors that could account for
548 the diversity of PTB variants and genes among the various studies. Considering the
549 possible association of PTB with other traits, we compared differential gene programs
550 on multiple phenotypes between mother and infant reported in previous studies, and
551 demonstrated a high risk of *CRHBP, ICAM1, DNMT1, CNTLN, PLAGL1, DKK1* and
552 *EGR2* with PTB among our sample cohorts of Chinese or Eastern Asian ancestry. To
553 support this finding, we reconstructed co-expressive networks linking the PTB-related
554 genes in GWAS with the reported PTB markers. Indeed, the correlated PTB genes are
555 mainly involved in immune & inflammation related processes and signaling pathways,
556 as well as coagulation factors. Thus, these findings have biological implications for
557 dissecting genetic associations of gestational factors with disease or traits in different
558 human ancestries.

559 Pregnancy is associated with an increased risk of developing breast cancer (66).
560 Although several reports indirectly described relationship between breast cancer and

561 PTB (67), evidence is still lacking. To determine whether a correlation exists between
562 PTB women and breast cancer pathogenesis, we compared differential expression of
563 the PTB related genes in PTB samples and three subtypes of hormone receptor-related
564 breast cancer samples. Although several DEGs were found in both PTB and cancer
565 subtypes, their expression patterns are not consistent. It perhaps is because of different
566 subtype of samples derived from heterogeneous populations, as well as different
567 phenotypes involved in the analyses. Nevertheless, our data provides additional
568 evidence that PTB might be related with breast cancer hormone-related subtypes. A
569 graphic overview of our results is summarized in Fig. 6c. We proposed interplays of
570 gestational age with PTB, fetal disease and breast cancer. The representative PTB genes
571 *CRHBP, ICAM1, THBS4, DNMT1, CNTLN, PLAGL1, DKK1* and *EGR2* are likely
572 associated with these phenotypes by targeting immune and inflammatory response,
573 coagulation, and cell adhesion.
574

575 ## Conclusions

576 This study benchmarked the ability of ulcWGS to be used for genotype imputation. As
577 the first study using a large cohort of human embryo samples with ulcWGS, we
578 demonstrate its power and effectiveness in GWAS. We detected 40 significant SNPs
579 and 11 genomic risk loci that contains independent significant SNPs and are associated
580 with gestational age. From 166 genes mapped to the risk loci. We establish
581 interrelationships between the mapped genes and maternal or infant diseases and
582 provide insights into understanding the genetic associations of gestational ages. Our

583 findings should expand current GWAS related to gestational duration and preterm trait
584 by including Chinese and East Asian samples and would therefore be helpful to future
585 research.
586

587 **Supplementary Information**

588

589 **Additional file 1: Table S1.** Genotype imputation performance at different ultra-low
590 coverages and sample sizes. **Table S2**. WGA methods and sequencing platforms used
591 in the PGT experiments. **Table S3**. The summary of 40 significant SNPs satisfying
592 Bonferroni-corrected significant levels of 4.526e-8 from 1,744 embryo PGT samples.
593 **Table S4**. The list of candidate SNPs. **Table S5**. Nonsynonymous SNPs that are
594 independent significant SNPs or in LD (r2>0.6) with one of the independent significant
595 SNPs. **Table S6**. The reported genomic risk loci from GWAS catalog that were detected
596 in our dataset. **Table S7**. A list of 166 mapped genes by positional, eQTL, and chromatin
597 interaction mappings. **Table S8**. Data sources of genome-wide mRNA expression in
598 Preterm birth, infant disease and breast cancer. **Table S9**. A list of the 166 mapped genes
599 that were also identified by differentially expressed genes (DEGs) derived from
600 analyzing the genome-wide gene expression datasets listed.
601 **Additional file 2: Supplemental Methods**
602 **Additional file 3: Figure S1.** Sequencing coverage statistics of the 17,844 embryos.
603 **Figure S2.** Visualizing the linear regression model for genotype accuracy prediction.
604 **Figure S3.** Quality control of SNP calling. Figure S4. Figure S4. Principal component

605 analysis of all 17,844 embryo samples. **Figure S5.** Distribution of gestational age in the
606 1,744 born embryo samples. **Figure S6.** Quantile-Quantile plot of the 1,107,198 studied
607 SNPs. **Figure S7.** Pie charts of th**e** candidate SNPs. **Figure S8.** Zoom in locus plots of
608 the 11 genomic risk loci. **Figure S9.** Circos plots of chromatin interactions and eQTL
609 mapping. **Figure S10.** Gene set analysis of the 166 mapped genes. **Figure 11.** Co-
610 expression network of PTB-related genes in maternal and infant subtypes.
611

612 **Abbreviations**

613 WGS: whole genome sequencing; NGS: next-generation sequencing; GWAS:
614 genome-wide association study; bp: base pair; SNP: single nucleotide polymorphism;
615 PTB: preterm birth; PGT: preimplantation genetic testing; MAF: minor allele
616 frequency; LD: linkage disequilibrium; HWE: Hardy-Weinberg equilibrium; WGA:
617 whole genome amplification; PCA: principal component analysis; eQTL: expression
618 quantitative trait locus; FDR: false discovery rate; GO: gene ontology; BPD:
619 bronchopulmonary dysplasia; ER: Estrogen receptor; PR: Progesterone receptor;
620 HER2: human epidermal growth factor receptor 2; TNBC: Triple-negative breast
621 cancer; DEG: differentially expressed gene; AF: allele frequency. GIAB: Genome in a
622 Bottle; BWA: Burrows-Wheeler Aligner. GEO: Gene Expression Omnibus: NCBI:
623 National Center for Biotechnology Information.
624

625 # Declarations

626 ## Ethics approval and consent to participate

627 This study has been reviewed and approved by the Institutional Review Board (IRB) of
628 China International Trust Investment Corporation - Xianngya (IRB Reference No. LL-
629 SC-2020-004). The study used data for clinical PGT samples and their complete follow-
630 up records collected at the Reproductive & Genetic Hospital of CITIC-XIANGYA
631 between 2013 and 2019 for retrospective analysis. All patients provided written
632 informed consent for PGT and medical research. Following the regulations of the
633 Human Genetic Resources Administration of China, all genetic materials involved in
634 this study have been reviewed and approved by the Ministry of Science and Technology
635 of the People´s Republic of China (Approval No. [2022] GH1831).

636

637 ## Consent for publication

638 Not applicable

639

640 ## Availability of data and materials

641 The datasets supporting the conclusions of this article are included within the article
642 and its additional files. The significant SNPs and candidate SNPs are listed in
643 Additional file 1: Tables S3 and S4, respectively. The processed or raw counts of mRNA
644 expression datasets shown in Additional file 1: Table S8 was downloaded from Gene
645 Expression Omnibus (GEO)/National Center for Biotechnology Information (NCBI).

646

Funding

The study was supported by the Early Career Schema (27204518) of the Hong Kong Research Grants Council for RL, partially by General Research Funding of Hong Kong Research Grants Council (17113721) for RL and (17117918) for BY, by the University Grants Committees Fund from the University of Hong Kong for RL, partially by National Key Research and Developmental Program of China (2018YFC1004900) for YT, by the Innovation and Technology Fund (ITF/331/17FP) of Innovation and Technology Commission of the Hong Kong SAR government for TL.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SL, RL and BY developed the computational pipeline and the algorithms. RL, GL, YW, TKL, and TL designed the experiments. PX, YG, FG and YT contributed to the clinical data collection. SL, BY, JL, YG and FG performed data processing and analysis. RL, GL, SL, BY, PX and YW participated in drafting the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable

669

670 ## Reference

671 1. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10
672 Years of GWAS Discovery: Biology, Function, and Translation. American journal of
673 human genetics. 2017;101(1):5-22.
674 2. Pasaniuc B, Rohland N, McLaren PJ, Garimella K, Zaitlen N, Li H, et al. Extremely
675 low-coverage sequencing and imputation increases power for genome-wide association
676 studies. 2012;44(6):631.
677 3. Wang Z, Chatterjee N. Increasing mapping precision of genome-wide association
678 studies: to genotype and impute, sequence, or both? Genome Biology. 2017;18(1):118.
679 4. Quick C, Anugu P, Musani S, Weiss ST, Burchard EG, White MJ, et al. Sequencing
680 and imputation in GWAS: Cost-effective strategies to increase power and genomic
681 coverage across diverse populations. 2020;44(6):537-49.
682 5. Gilly A, Southam L, Suveges D, Kuchenbaecker K, Moore R, Melloni GEM, et al.
683 Very low-depth whole-genome sequencing in complex trait association studies.
684 Bioinformatics. 2019;35(15):2555-61.
685 6. Homburger JR, Neben CL, Mishne G, Zhou AY, Kathiresan S, Khera AV. Low
686 coverage whole genome sequencing enables accurate assessment of common variants
687 and calculation of genome-wide polygenic scores. Genome medicine. 2019;11(1):74.
688 7. Marchini J, Howie B. Genotype imputation for genome-wide association studies.
689 Nature reviews Genetics. 2010;11(7):499-511.
690 8. Liu S, Huang S, Chen F, Zhao L, Yuan Y, Francis SS, et al. Genomic Analyses from
691 Non-invasive Prenatal Testing Reveal Genetic Associations, Patterns of Viral Infections,
692 and Chinese Population History. Cell. 2018;175(2):347-59.e14.
693 9. Cronin FM, Segurado R, McAuliffe FM, Kelleher CC, Tremblay RE. Gestational
694 Age at Birth and 'Body-Mind' Health at 5 Years of Age: A Population Based Cohort
695 Study. PloS one. 2016;11(3):e0151222.
696 10. Knijnenburg TA, Vockley JG, Chambwe N, Gibbs DL, Humphries C, Huddleston
697 KC, et al. Genomic and molecular characterization of preterm birth. Proceedings of the
698 National Academy of Sciences of the United States of America. 2019;116(12):5819-27.
699 11. Liu X, Helenius D, Skotte L, Beaumont RN, Wielscher M, Geller F, et al. Variants
700 in the fetal genome near pro-inflammatory cytokine genes on 2q13 associate with
701 gestational duration. Nature communications. 2019;10(1):3927.
702 12. Zhang G, Feenstra B, Bacelis J, Liu X, Muglia LM, Juodakis J, et al. Genetic
703 Associations with Gestational Duration and Spontaneous Preterm Birth. The New
704 England journal of medicine. 2017;377(12):1156-67.
705 13. Wadon M, Modi N, Wong HS, Thapar A, O'Donovan MC. Recent advances in the
706 genetics of preterm birth. Annals of human genetics. 2019.
707 14. Brezina PR, Kutteh WH, Bailey AP, Ke RW. Preimplantation genetic screening
708 (PGS) is an excellent tool, but not perfect: a guide to counseling patients considering
709 PGS. Fertil Steril. 2016;105(1):49-50.
710 15. Huang L, Bogale B, Tang Y, Lu S, Xie XS, Racowsky C. Noninvasive
711 preimplantation genetic testing for aneuploidy in spent medium may be more reliable

712 than trophectoderm biopsy. Proc Natl Acad Sci U S A. 2019;116(28):14105-12.
713 16. Tan Y, Yin X, Zhang S, Jiang H, Tan K, Li J, et al. Clinical outcome of
714 preimplantation genetic diagnosis and screening using next generation sequencing.
715 GigaScience. 2014;3(1):30.
716 17. Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, et al.
717 Integrating human sequence data sets provides a resource of benchmark SNP and indel
718 genotype calls. Nature Biotechnology. 2014;32(3):246-51.
719 18. Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis.
720 Current protocols in bioinformatics. 2014;47:11.2.1-34.
721 19. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler
722 transform. Bioinformatics. 2009;25(14):1754-60.
723 20. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
724 Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
725 21. Li H. FermiKit: assembly-based variant calling for Illumina resequencing data.
726 Bioinformatics. 2015;31(22):3694-6.
727 22. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The
728 human genome browser at UCSC. Genome Res. 2002;12(6):996-1006.
729 23. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed:
730 enabling browsing of large distributed datasets. Bioinformatics. 2010;26(17):2204-7.
731 24. Davies RW, Flint J, Myers S, Mott R. Rapid genotype imputation from sequence
732 without reference panels. Nature genetics. 2016;48(8):965-9.
733 25. Skotte L, Korneliussen TS, Albrechtsen A. Association testing for next-generation
734 sequencing data using score statistics. Genet Epidemiol. 2012;36(5):430-7.
735 26. Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: Analysis of Next Generation
736 Sequencing Data. BMC Bioinformatics. 2014;15:356.
737 27. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and
738 annotation of genetic associations with FUMA. Nature communications.
739 2017;8(1):1826.
740 28. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A
741 global reference for human genetic variation. Nature. 2015;526(7571):68-74.
742 29. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic
743 variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.
744 30. Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, et al. DAVID: a web
745 server for functional enrichment analysis and functional annotation of gene lists
746 (2021 update). Nucleic Acids Res. 2022.
747 31. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45(6):580-5.
748 32. Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, et al. A Compendium of
749 Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. Cell
750 reports. 2016;17(8):2042-59.
751 33. Giusti-Rodriguez PM, Sullivan PF. Using three-dimensional regulatory chromatin
752 interactions from adult and fetal cortex to interpret genetic results for psychiatric
753 disorders and cognitive traits. 2019:406330.
754 34. Akbarian S, Liu C, Knowles JA, Vaccarino FM, Farnham PJ, Crawford GE, et al.
755 The PsychENCODE project. Nature neuroscience. 2015;18(12):1707-12.

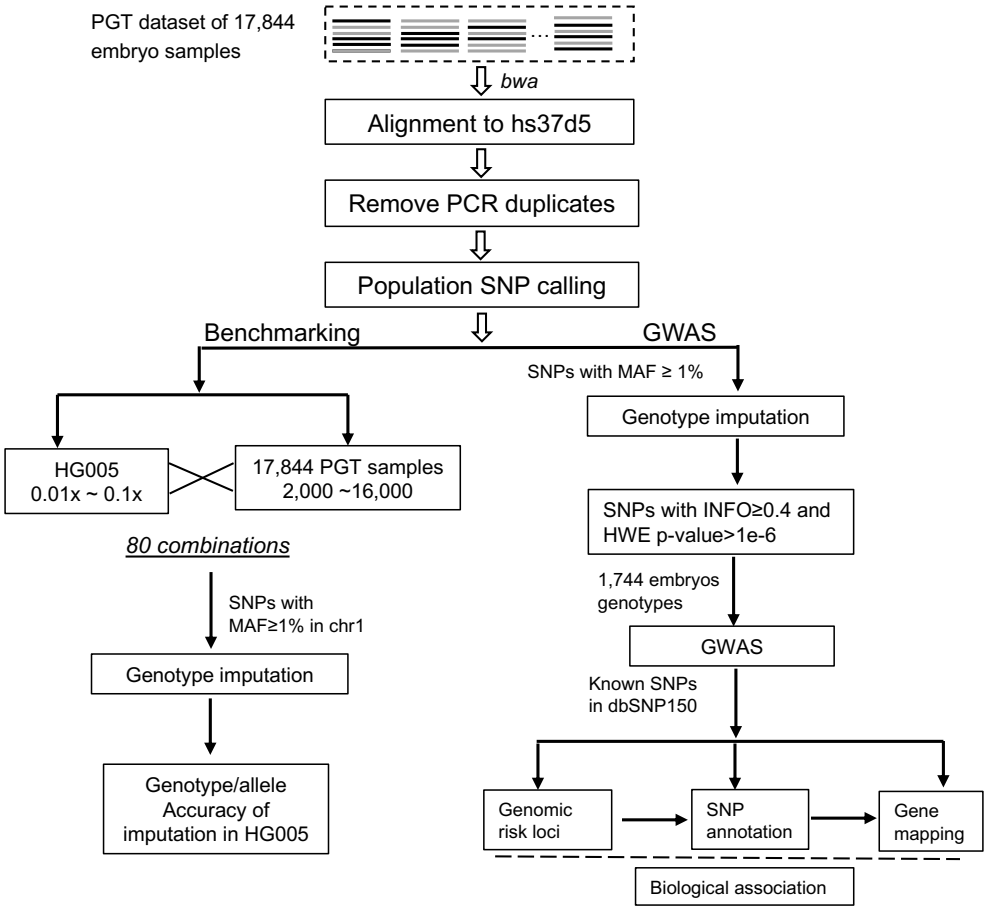
35. Noguchi S, Arakawa T, Fukuda S, Furuno M, Hasegawa A, Hori F, et al. FANTOM5 CAGE profiles of human and mouse samples. Scientific data. 2017;4:170112.
36. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5(6):e1000529.
37. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308-11.
38. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-43.
39. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet. 2006;38(2):209-13.
40. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Patterson N, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015;47(3):291-5.
41. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337(6099):1190-5.
42. Pasaniuc B, Rohland N, McLaren PJ, Garimella K, Zaitlen N, Li H, et al. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nature Genetics. 2012;44:631.
43. Wulf G, Ryo A, Liou Y-C, Lu KP. The prolyl isomerase Pin1 in breast development and cancer. Breast Cancer Res. 2003;5(2):76-82.
44. Consortium GT. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science (New York, NY). 2020;369(6509):1318-30.
45. Vora B, Wang A, Kosti I, Huang H, Paranjpe I, Woodruff TJ, et al. Meta-Analysis of Maternal and Fetal Transcriptomic Data Elucidates the Role of Adaptive and Innate Immunity in Preterm Birth. Front Immunol. 2018;9:993-.
46. Strauss JF, 3rd, Romero R, Gomez-Lopez N, Haymond-Thornburg H, Modi BP, Teves ME, et al. Spontaneous preterm birth: advances toward the discovery of genetic predisposition. American journal of obstetrics and gynecology. 2018;218(3):294-314.e2.
47. Velez DR, Fortunato SJ, Thorsen P, Lombardi SJ, Williams SM, Menon R. Preterm birth in Caucasians is associated with coagulation and inflammation pathway gene variants. PloS one. 2008;3(9):e3283-e.
48. Sheikh IA, Ahmad E, Jamal MS, Rehan M, Assidi M, Tayubi IA, et al. Spontaneous preterm birth and single nucleotide gene polymorphisms: a recent update. BMC Genomics. 2016;17(Suppl 9):759.
49. Uzun A, Sharma S, Padbury J. A bioinformatics approach to preterm birth. American journal of reproductive immunology (New York, NY : 1989). 2012;67(4):273-7.
50. Siffel C, Kistler KD, Lewis JFM, Sarda SP. Global incidence of bronchopulmonary

dysplasia among extremely preterm infants: a systematic literature review. The journal of maternal-fetal & neonatal medicine : the official journal of the European Association of Perinatal Medicine, the Federation of Asia and Oceania Perinatal Societies, the International Society of Perinatal Obstet. 2019:1-11.
51. Cai Y, Ma F, Qu L, Liu B, Xiong H, Ma Y, et al. Weighted Gene Co-expression Network Analysis of Key Biomarkers Associated With Bronchopulmonary Dysplasia. Front Genet. 2020;11:539292.
52. Villamor-Martinez E, Lubach GA, Rahim OM, Degraeuwe P, Zimmermann LJ, Kramer BW, et al. Association of Histological and Clinical Chorioamnionitis With Neonatal Sepsis Among Preterm Infants: A Systematic Review, Meta-Analysis, and Meta-Regression. Front Immunol. 2020;11:972.
53. Rosette C, Roth RB, Oeth P, Braun A, Kammerer S, Ekblom J, et al. Role of ICAM1 in invasion of human breast cancer cells. Carcinogenesis. 2005;26(5):943-50.
54. Schroder C, Witzel I, Muller V, Krenkel S, Wirtz RM, Janicke F, et al. Prognostic value of intercellular adhesion molecule (ICAM)-1 expression in breast cancer. J Cancer Res Clin Oncol. 2011;137(8):1193-201.
55. Xu WH, Liu ZB, Yang C, Qin W, Shao ZM. Expression of dickkopf-1 and beta-catenin related to the prognosis of breast cancer patients with triple negative phenotype. PloS one. 2012;7(5):e37624.
56. Shin E, Lee Y, Koo JS. Differential expression of the epigenetic methylation-related protein DNMT1 by breast cancer molecular subtype and stromal histology. J Transl Med. 2016;14:87.
57. Wong KK. DNMT1: A key drug target in triple-negative breast cancer. Semin Cancer Biol. 2020.
58. Robertson SA, Hutchinson MR, Rice KC, Chin PY, Moldenhauer LM, Stark MJ, et al. Targeting Toll-like receptor-4 to tackle preterm birth and fetal inflammatory injury. Clin Transl Immunology. 2020;9(4):e1121.
59. Zhang G, Srivastava A, Bacelis J, Juodakis J, Jacobsson B, Muglia LJ. Genetic studies of gestational duration and preterm birth. Best Pract Res Clin Obstet Gynaecol. 2018;52:33-47.
60. Petraglia F, Imperatore A, Challis JR. Neuroendocrine mechanisms in pregnancy and parturition. Endocr Rev. 2010;31(6):783-816.
61. Majzoub JA, McGregor JA, Lockwood CJ, Smith R, Taggart MS, Schulkin J. A central theory of preterm and term labor: putative role for corticotropin-releasing hormone. American journal of obstetrics and gynecology. 1999;180(1 Pt 3):S232-41.
62. Díaz-Pérez FI, Hiden U, Gauster M, Lang I, Konya V, Heinemann A, et al. Post-transcriptional down regulation of ICAM-1 in feto-placental endothelium in GDM. Cell adhesion & migration. 2016;10(1-2):18-27.
63. Chen X, Scholl TO. Maternal biomarkers of endothelial dysfunction and preterm delivery. PloS one. 2014;9(1):e85716.
64. Labarrere CA, Bammerlin E, Hardin JW, Dicarlo HL. Intercellular adhesion molecule-1 expression in massive chronic intervillositis: implications for the invasion of maternal cells into fetal tissues. Placenta. 2014;35(5):311-7.
65. Kim SC, Lee JE, Kang SS, Yang HS, Kim SS, An BS. The regulation of oxytocin

844 and oxytocin receptor in human placenta according to gestational age. Journal of
845 molecular endocrinology. 2017;59(3):235-43.
846 66. Keyser EA, Staat BC, Fausett MB, Shields AD. Pregnancy-associated breast cancer.
847 Rev Obstet Gynecol. 2012;5(2):94-9.
848 67. Froehlich K, Schmidt A, Heger JI, Al-Kawlani B, Aberl CA, Jeschke U, et al.
849 Breast cancer, placenta and pregnancy. Eur J Cancer. 2019;115:68-78.
850

851

852 # Figures



853

854 **Fig. 1. An overview of the analysis and benchmarking pipeline for ultra-low**
855 **coverage WGS.**

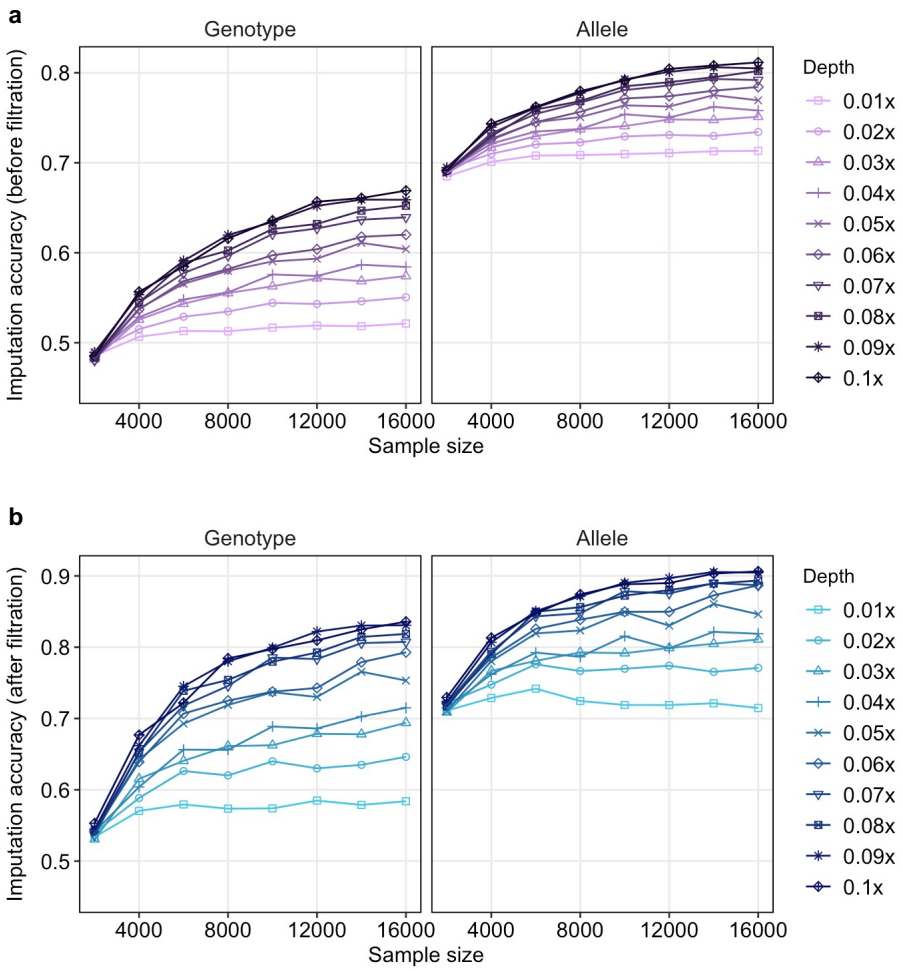


Fig. 2. Imputation accuracy at different coverages and sample sizes. The accuracies of imputed genotype (a) or allele (b) were obtained by comparing with the known genotypes in HG005. After using filter "INFO score ≥ 0.4 and HWE *p*-value > 1e-6", the accuracies of imputed genotype (c) or allele (d).

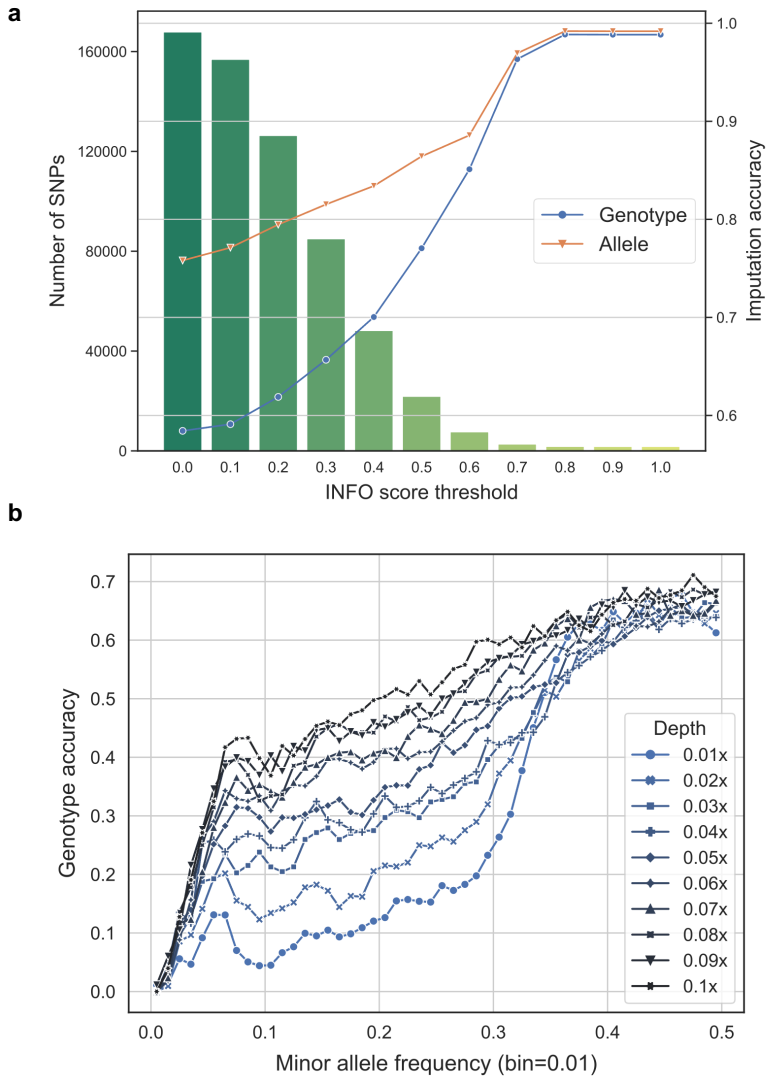


Fig. 3. Performance of different imputation result filters. The accuracies of our samples were calculated against the known genotypes in HG005. a, effect of INFO score filtering cutoffs on genotype and allele accuracies. The imputation was conducted by using 0.04x sequencing coverage of HG005 and with 16000 embryo samples. b, effect of MAF cutoffs on genotype accuracy at multiple sequencing coverages. The imputation was conducted through different sequencing coverages of HG005 with 16000 embryo samples.

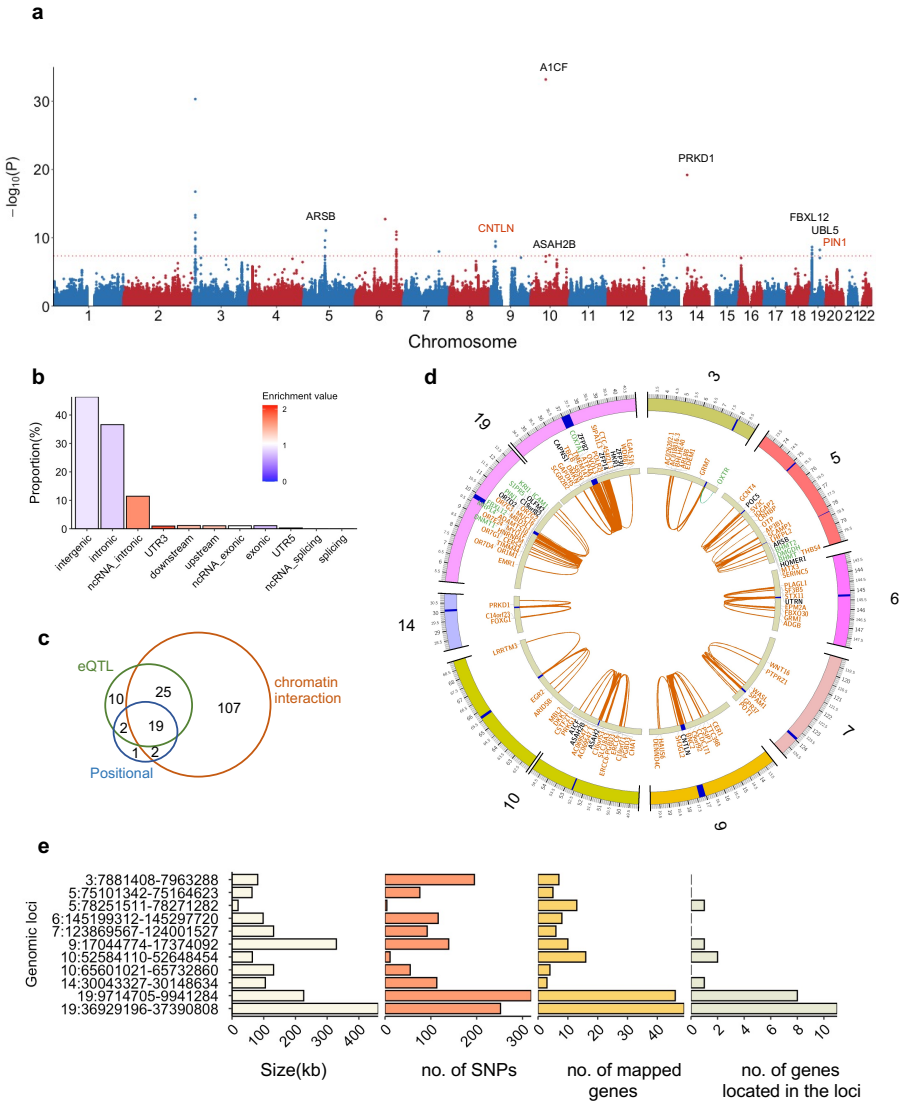


Fig. 4. SNP-based genome-wide association on gestational age. a, Manhattan plot of the SNPs in GWAS. The red dash line represents the genome-wide significance level 4.515e-8. The SNP "rs946934582" with p value of 2.764e-144 is beyond the scale, thus hereby listed alone. The genes shown are linking with the candidate SNPs and position of the corresponding genomic risk loci. b, functional annotation and enrichment test result of the candidate SNPs in FUMA. c, a Venn diagram of the 166 genes that could be mapped to the 11 genomic risk loci by positional, eQTL and chromatin interaction strategies. d, a Circos plot of the chromatin interactions and eQTL mapping in the 11 genomic risk loci from eight chromosomes. The outer ring is chromosomes, the regions in blue denote genomic risk loci. The middle ring represents the mapped genes. The color of the gene symbols shows how they were mapped, eQTL in green, chromatin interaction in orange, and both eQTL and chromatin interaction in black. The inner ring shows the linking edges, eQTL in green, and chromatin interaction in orange. e, a summary of the 11 genomic risk loci.

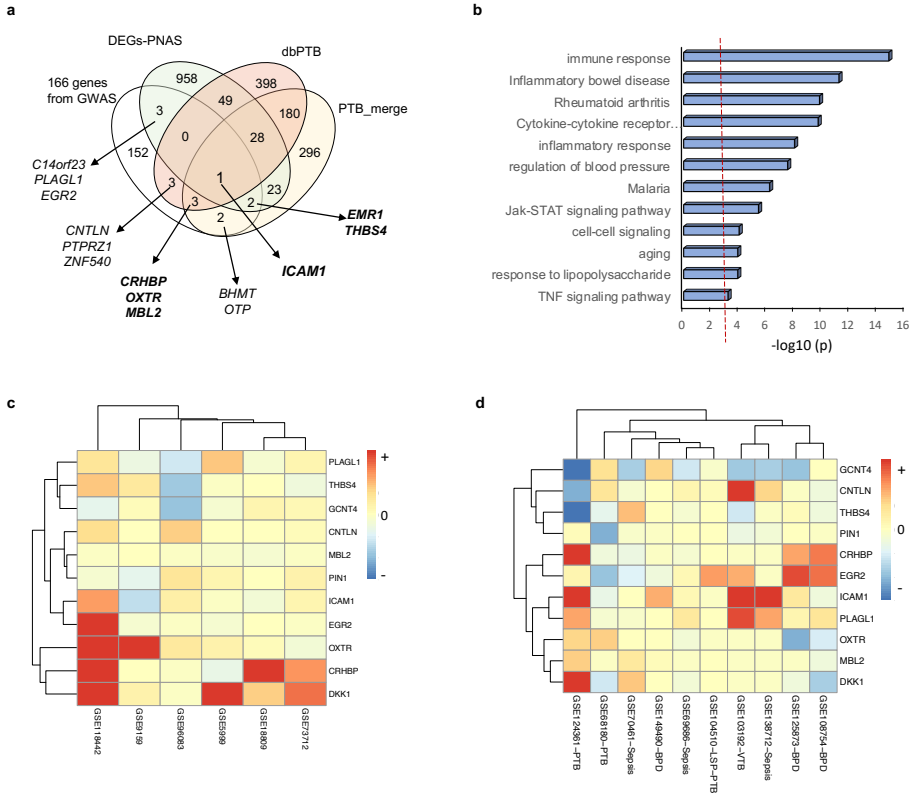


Fig. 5. Comparison of the 166 genes mapped to 11 genomic risk loci with PTB and infant disease. a, the 166 mapped genes were compared with 3 sets of reported PTB genes including dbPTB from Sheikh et al. (48) and Uzun et al (49), PTB-merged from 5 resources (10), and PNAS-identified DEGs in 2019 (10) (see Additional file 1: Table S8). b, a bar graph showing significantly enriched GO biological processes and KEGG pathways based on the 50 PTB marker genes (see Additional file 1: Table S7). c-d, heatmaps showing expression profiling and clusters of the PTB genes predicted from GWAS under maternal PTB (c) and PTB infant with BPD and sepsis (d). The gene expression data was extracted from the listed GSE accession numbers of NCBI/GEO.

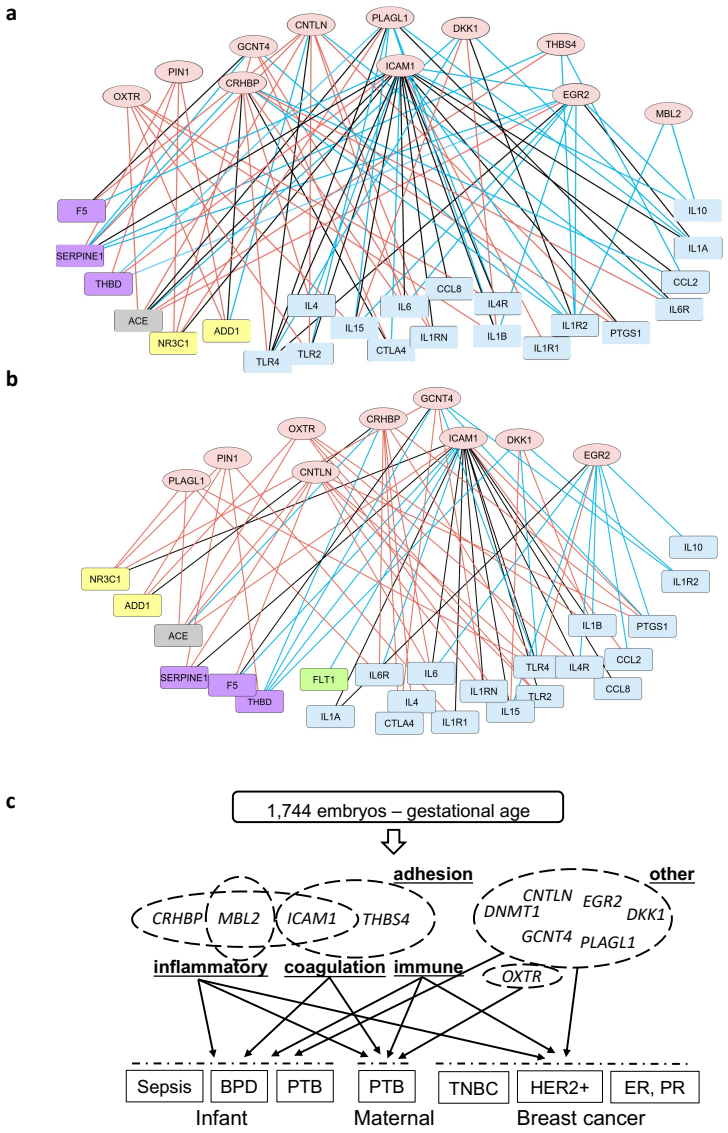


Fig. 6. Associated analysis of the PTB-related genes in maternal and infant subtypes. a, a gene co-expression network merging maternal and infant PTB subtypes. b, a gene co-expression network merging maternal PTB and BPD of preterm infant. In A and B, oval nodes represent PTB related genes predicted from GWAS using the 1,744 born embryo samples, rectangle nodes refer to the 50 top PTB markers summarized from previously studies (Additional file 1: Table S7), involving immune or inflammation (blue), apoptosis (yellow), angiogenesis (green), coagulation (purple) and other biological processes (grey). Co-expressive edges (Pearson correlation p value < 0.01) linking nodes represent maternal (red), infant (blue) and both maternal and infant (black). c, a graphic summary to illustrate gestational age's association with PTB, infant disease and breast cancer.

Tables

Table 1. Genotype imputation performance with different filtering criteria

Filtering criteria	SNP number	Genotype accuracy	Allele accuracy
None	167814	0.584	0.758
Known SNP in 1000G reference panel	141161	0.578	0.751
INFO score ≥ 0.1	156835	0.591	0.771
INFO score ≥ 0.2	126331	0.618	0.794
INFO score ≥ 0.3	84932	0.656	0.815
INFO score ≥ 0.4	48176	0.7	0.834
MAF ≥ 0.05	162788	0.597	0.777
HWE *p*-value > 1e-9	84237	0.595	0.695
HWE *p*-value > 1e-6	77419	0.596	0.692
HWE *p*-value > 1e-3	65561	0.599	0.687
HWE *p*-value > 1e-6 and INFO score ≥ 0.1	68505	0.610	0.713
HWE *p*-value > 1e-6 and INFO score ≥ 0.2	55972	0.646	0.751
HWE *p*-value > 1e-6 and INFO score ≥ 0.3	42202	0.681	0.786
HWE *p*-value > 1e-6 and INFO score ≥ 0.4	28773	0.715	0.818

Totally 16,000 samples with an average of 0.04x sequencing coverage were used for imputation.

Table 2. Main mapped genes differentially expressed in PTB, infant disease and breast cancer

Genes mapped to risk loci	Overlapped No. with the reported PTB datasets	No. of DEGs detected in gene expression data						
		Maternal PTB	Infant PTB	Infant BPD	Infant sepsis	Breast cancer ER, PR	Breast cancer TNBC	Breast cancer HER2+
ICAM1	6	1(up)	2(up)	1(up)	1(up)	1(up),2(down)	2(up)	1(up)
CRHBP	5	2(up)	1(up)	1(up)				
OXTR	5	2(up)				1(up)	1(down)	
THBS4	3	1(down)	1(down)		1(up)	1(up)	2(down)	1(up),1(down)
EGR2	2	1(up)	1(up)	1(up)		1(up)	2(down)	2(down)
CNTLN	1	1(up)	1(up),1(down)		1(up)	1(down)		1(down)
MBL2	4		1(up)					
EMR1	2			1(up)				
PLAGL1	1	1(up)	2(up)		1(up)	2(down)	2(down)	2(down)
DKK1		2(up)	1(up)	1(down)		2(down)	3(up),1(down)	1(up),1(down)
GCNT4		1(down)	2(down)	1(down)	2(down)	1(up),2(down)	1(down)	1(down)
DNMT1		1(up)	1(up)				1(up)	1(up)
PIN1		1(down)	1(down)					

PTB related genes identified in our GWAS were compared with DEGs identified in published gene expression datasets (Additional file 1: Table S9). “up” and “down” indicate differentially overexpressed and underexpressed gene, respectively.