Whole genome sequencing and analysis of 4,053 individuals in trios and mother-infant duos from the Born in Guangzhou Cohort Study

Xiu Qiu (xiu.qiu@bigcs.org)

Division of Birth Cohort Study, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, 510623, China.

Shujia Huang

Guangzhou Women and Children's Medical Center

Mingxi Huang

Guangzhou Women and Children's Medical Center

Siyang Liu

Sun Yat-sen University https://orcid.org/0000-0001-6780-9419

Chengrui Wang

Guangzhou Women and Children's Medical Center

Jianrong He

Guangzhou Women and Children's Medical Center

Yashu Kuang

Guangzhou Women and Children's Medical Center

Jinhua Lu

Guangzhou Women and Children's Medical Center

Yuqin Gu

Sun Yat-sen University

Xiaoyan Xia

Guangzhou Women and Children's Medical Center

Shanshan Lin

Guangzhou Women and Children's Medical Center

Huimin Xia

Guangzhou Women and Children's Medical Center

Biological Sciences - Article

Keywords:

Posted Date: June 10th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1732885/v1

Whole genome sequencing and analysis of 4,053 individuals in trios and mother-infant duos from the Born in Guangzhou Cohort Study

Shujia Huang[1*], Mingxi Huang[1*], Siyang Liu[2*], Chengrui Wang[1], Jianrong He[1,3,4], Yashu Kuang[1,3], Jinhua Lu[1,3], Yuqin Gu[2], Xiaoyan Xia[1,7], Shanshan Lin[1,7], Huimin Xia[3,5,6#], Xiu Qiu[1,3,7#]

Affiliations:

[1] Division of Birth Cohort Study, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, 510623, China.

[2] School of Public Health (Shenzhen), Sun Yat-sen University, Shenzhen, Guangdong 510006, China.

[3] Provincial Clinical Research Center for Child Health, Guangzhou, 510623, China.

[4] Department of Obstetrics and Gynecology, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, 510623, China

[5] Provincial Key Laboratory of Research in Structure Birth Defect Disease, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, 510623, China.

[6] Department of Pediatric Surgery, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, 510623, China

[7] Department of Women's Health, Provincial Key Clinical Specialty of Woman and Child Health, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, 510623, China

* These authors contributed equally to this work: Shujia Huang, Mingxi Huang, Siyang Liu

# These authors supervised this work: Xiu Qiu and Huimin Xia. Correspondence should be addressed to xiu.qiu@bigcs.org and huimin.xia@bigcs.org

ABSTRACT

Large-scale birth cohorts that recruit trios/duos families are essential resource for determining the genetic and environmental contribution to maternal-infant's health. While large-scale genomic studies of unrelated individuals were carried out around the world, genomic study based on birth cohorts is scarce, especially in China and Asia. Here, we present the Phase I genomic study of the Born in Guangzhou Cohort Study (BIGCS), which consists of the analysis of the whole genome sequencing data of 4,053 Chinese participants in trios or mother-infant duos living in South China. We identify 18.3 million novel genetic variants and construct a reference panel that enables more accurate genotype imputation for individuals of Chinese ancestry. We identify a new ancestral component specific to southern Chinese and for the first time, have dissected the genetic relationship of 10 Chinese dialects. We reveal seven novel East-Asian specific genetic associations with total bile acid, gestational weight gain and lipid metabolism level in the maternal peripheral or fetal cord blood and have recognized the distinction of genetic effect on the same trait between adults and infants. Using inter-generational mendelian randomization, we have dissected five mechanisms related to maternal and fetal genetic effect of seven adult phenotypes on six fetal growth-related pregnancy outcomes. Our findings fill the gap of the missing diversity in human genetics and demonstrate the great value of genomic study of birth cohort in advancing the worldwide genetic knowledge and maternal-infant health.

INTRODUCTION

Advances in sequencing technology and analytical approaches have enabled the possibility to identify genes and variants that are causal for both common and rare traits and diseases. However, the underrepresentation of non-European individuals in the current genomic studies has limited the applicability of the results to the worldwide population[1]. In addition, the large-scale international and national genomic projects initiated and published nowadays have a focus on unrelated adult individuals[2–5], and thus are not able to answer fundamental questions such as the genetic basis of traits involved early in life. Furthermore, we have limited knowledge on how intrauterine and early-life environments interact with genetics and altogether contribute to the trait variation and disease liability.

Birth cohort that recruits families in trios and duos in a longitudinal effort provides an effective and systematic strategy to investigate the effects of early life exposure on the near- and long-term health status of the progenitors and offsprings[6,7]. It represents a foundational resource for human genetic study despite more difficulties are needed to overcome to construct and maintain such a cohort compared to the cross-sectional study that recruits randomly chosen individuals at a single point in time. To date, a few European birth cohort studies reported insights into the genetic basis of particular traits and developed protocols to dissect the causal link between intrauterine or early-life exposures on birth or long-term health outcomes[8–11]. Nonetheless, all of the studies rely on array instead of whole-genome sequencing technology and thus variants that are difficult to impute such as the rare and individual genetic variants are hardly investigated. In China, several birth cohort studies have been launched in recent years such as the China-Anhui birth cohort study[12], the Jiaxing Birth Cohort[13], the Shanghai Birth Cohort Study (SBC)[14] and the China National Birth Cohort (CNBC)[15]. However, those projects are mainly focusing on traditional epidemiology analytical strategy with limited genetic data. Such a design hinders the possibility to causally dissect the interaction between multifactorial risk factors and the health outcome. Considering the difference of lifestyle, genetic background and population characteristics between Chinese and

European population and the necessity of an in-depth understanding of the genetic and environmental factors that may contribute to early growth and adult health, large-scale genomic study on the birth cohorts in the worldwide are urgently needed.

Launched in South China, the Born in Guangzhou Cohort Study (BIGCS) represents the largest prospective and longitudinal birth cohort in China and Asia[16]. By the end of 2021, the study has recruited and deeply phenotyped more than 50,000 trio or duo families whose health status as well as physical and biological measurements were tracked from the prenatal period to adult life. Here, we report the BIGCS Phase I genomic study, which consists of analyses of the whole genome sequencing (WGS) of 4,053 participants in trios or duos (332 trios, 1,406 duos and 245 unrelated individuals) from China using a low-coverage whole-genome sequencing design (lc-WGS, ~6.63x on average)[17]. We are aiming to provide high-quality variation dataset and haplotype reference panel utilizing the family information to improve the global and regional genetic knowledge. We also use the data generated in the study to explore the population genetic characteristic of the cohort, which includes participants of previously under-studied southern Chinese ancestry and analyze the genetic relationship between 10 Chinese dialects for the first time. Furthermore, we carry out genome-wide association studies (GWAS) of 18 adult and infant quantitative traits to identify genes that impact the level of adult or infant glucolipid metabolism and to compare the genetic effect on the same trait between adults and infants. Finally, we conduct inter-generational mendelian randomization to dissect different mechanisms underlying the observational associations between maternal phenotypes and fetal growth measurements.

RESULTS

Whole-genome sequencing of 332 trios, 1,406 duos and 245 unrelated individuals in BIGCS

A total of 4,053 individuals from the BIGCS program at Guangzhou Women and Children Medical Center (GWCMC) were included in the study, covering 13 ethnicities (Han Chinese and 12 minorities) from 30 of the 34 administrative divisions of China (**Fig. 1a**; Fig. S1; Table S1). The participants consist of 332 parent-offspring trios (father-mother-offspring), 1,406 parent-offspring duos (14 father-offspring and 1,392 mother-offspring) and 245 unrelated single individuals (70 infants, 150 adult females and 25 adult males) (**Fig. 1b**). Whole blood from the parents and the cord blood of the infants were sequenced using 100bp paired-end (PE) reads to an average sequencing depth of 6.63x (Fig. S2a,b), covering ~98.56%±0.64% of the non-N sequences in human genome (Table S2).

After a series of quality control and computational processing (Extended Data Fig.1) (**see Methods**), we identified a final call set of 56,230,613 bi-allelic variants, which consists of 51,052,456 SNPs and 5,178,157 small insertion-deletion polymorphisms (Indels) (Extended Data Table 1). The overall Ts/Tv ratio is 2.09 and the heterozygous-to-homozygous (Het/Hom) proportion is 1.48, both are consistent with the statistical expectation[18]. We divided the variants into six categories based on the non-reference allele frequency (AF) of the BIGCS population (**Fig. 1c**). Among these, 22.28 million (39.62%) are singleton (Allele count (AC) = 1) variants; 11.92 million (21.19%) are doubletons (AC=2); 9.54 million (16.96%) are very rare variants (AC>2 and AF ≤ 0.1%); 5.23 million (9.30%) are rare variants (0.1%< AF ≤1%); 1.79 million (3.19%) are low frequency variants (1%< AF ≤5%) and 5.48 million (9.74%) are common variants (AF > 5%) (Extended Data Table 1). Approximately 32.56% of the variants (18.3 million) that we identified

were not reported in NCBI dbSNP (Build 154)[19], 93.4% of which were classified as singletons or doubletons (AC≤2). The length of the Indels distributes between -20bp and 20bp (Fig. S3) and demonstrates a 3 base-pair (bp) enrichment characteristics as expected in coding regions due to strong negative selection on the frameshift Indels[20] (**Fig. 1d**). There were 653,289 (1.16%) protein-coding variants predicted by ensembl VEP[21] (v95), 159,674 deleterious variants by SIFT[22] and 143,184 probably/possibly damaging variants by PolyPhen-2[23] (Extended Data Table 1). The number of variants per individual demonstrates a geographical and ethnic pattern. Individuals from East, Southwest and South China have noticeably more outlier of variant than the others (Fig. S4a), likely due to more frequent gene flow in ancient China[24]. Ethnic minorities, Tujia, Mongols, Miao, Mulao and Yi, have a higher level of average individual variant number than the average, whereas the She, Buyi and Manchu individuals showed an overall decreased amount of variants (Fig. S4b). In order to evaluate the accuracy of the detected variant, we genotyped 240 individuals on Illumina GSA-24 (v1.0) BeadChip (**see Methods**). Among the 450k variants in the autosomal chromosomes detected in both of the WGS and SNP arrays data, we achieved 0.99 genotype concordance rate for variants with minor allele frequency (MAF) > 0.05. For low-frequency variants with MAF between 0.01 and 0.05, the concordance rate dropped slightly to 0.98 (Fig. S5a). These statistics indicates the high quality of our final call set. As genotype refinement via the BEAGLE tools[25] dramatically reduced the discordance rate (Fig. S5b), it emphasizes the importance of utilizing the family relatedness information in a birth cohort to improve variant quality.

To investigate the novel genetic characteristics and information of the Chinese population, we compared BIGCS dataset with four Chinese genetic resources: CMDB[26] (v1.0; 0.06x; 141,431genomes), ChinaMAP[5] (v1.0; 40.8x; 10,588 genomes), gnomAD EAS[27] (v3.0; 1,567 genomes) and 1KGP3[2] CHN (the combination of CHB, CHS and CDX; 301 genomes). There are 197,403,927 non-redundant variants (175,440,792 SNPs and 21,963,135 5s, respectively) in the combination of all these five datasets. A total of 19,736,896 BIGCS specific variants (10% of the overall variants) were identified, which mainly consists of rare variants (MAF < 1%, 99.95**%**) and 67.65% were singletons (**Fig. 1e**). Nonetheless, these BIGCS specific variants still include 6,278 low-frequency variants (3,232 SNPs and 3,046 Indels, MAF = 1%~5%) and 3,868 common variants (2,477 SNPs and 1,391 Indels, MAF > 5%).

Finally, utilizing the trios and duos design, we have constructed and released a haplotype reference panel that consists of 2,245 unrelated parental individuals and 43,055,086 high quality SNPs plus 4,184,387 Indels on the 22 autosomes and the X chromosome (**see Methods**). To compare the BIGCS panel with the most widely used reference panels for genotype imputation of individuals with Chinese ancestry, we used an independent high-coverage WGS data set of 50 Chinese individuals (40X coverage, 11,174,603 biallelic variants consisting of 9,816,793 SNPs and 1,357,810 Indels). We mimicked a typical imputation process using the 0.93 million SNP sites present in the Affymetrix Genome-Wide Human SNP Array 6.0, and evaluated the imputation accuracy (R^2) between the genotype dosage derived from the imputation and that from the high coverage sequencing using sites not present on the array (**see Methods**). The highest R^2 is achieved by the BIGCS panel in the comparison with the 1KGP3 reference panel (n=2,504), the haplotype reference consortium panel (HRC, n=32,470)[28] and the GenomeAsia100K Project reference panel (GAsP, n=1,654)[29] (**Fig. 1f**). The mean R^2 values of BIGCS panel were 0.918 compared to the 1KGP3 (R^2 =0.743), HRC panel (R^2 =0.837) and GAsP panel (R^2 =0.909) for rare variants with MAF ≤ 1%, which have a high proportion of population-specific variants. For low frequency

(1%≤MAF<5%) and common variants (MAF≥5%), the mean R^2 values of BIGCS reference panel are also the best among the 1KGP3 (low frequency: 0.923 versus 0.890, common: 0.965 versus 0.957), HRC (low frequency: 0.923 versus 0.787, common: 0.965 versus 0.934) and GAsP (low frequency: 0.923 versus 0.892, common: 0.965 versus 0.961). The higher imputation accuracy of BIGCS panel unravels the lack of Chinese population references in the globe and emphasizes the importance of family design in national population genome sequencing projects. To accelerate the broad utility of the BIGCS reference panel, we have released the data on the BIGCS imputation server (http://gdbig.bigcs.com.cn/).

Population structure analysis reveals a new genetic component specific to southern Chinese and the genetic relationship between ten Chinese dialects

The involvement of a high proportion of southern Chinese in the BIGCS study is different from what was investigated in 1000 genome project (1KGP) and other genomic studies[2,5,30]. We evaluated the genetic diversity of the BIGCS cohort using standard population genetic methodologies. First, we performed PCA and pairwise *F*st analysis in the BIGCS samples and the 1KGP Phase III population, which was treated as an outgroup to help distinguish the ethnic ancestry of BIGCS participants. As expected, we find that the first two principal components (PC1 and PC2) reveal the continental differentiation of the worldwide populations, and BIGCS individuals are most closely related to the East Asian populations (EAS) (Fig. S6). The observed genetic relationship is also reflected by the pairwise *F*st (*F*st ≤ 0.01 with all the five EAS groups in 1KGP) (Extended Data Fig. 2a). The genetic differentiation is very small among sub-populations in BIGCS characterized by the China geography (*F*st ≤ 0.004) (Extended Data Fig. 2b).

Secondly, we conducted PCA using all the unrelated parental individuals of BIGCS (N=2,245). Different from a few previous studies that investigated the Chinese population structure using genetic and geographical data, BIGCS participants also reported the dialects they speak, providing the first opportunity to link the Chinese genetic structure to the Chinese dialects. In total, our study has covered 10 Chinese dialects including 8 major dialects (Mandarin, Cantonese, Min, Hakka, Gan, Xiang, Wu and Hui) and 2 minority dialects (Zhuang and Li) according to the *Language Atlas of China*[31] (**Fig. 2a**). We find that the genetic relationship not only indicates the geographic distance (Extended Data Fig. 3, Fig. S7) but also reflects the language history (**Fig. 2b-c**). Specifically, Mandarin has a long history in China and gradually becomes the official language of China in 1932 [32]. Many Han Chinese living in the northern and southwestern China are native speakers of Mandarin and have been undergoing migration throughout the Chinese modern history[31] (**Fig. 2a**). As a result, Mandarin speakers demonstrate a greater level of genetic diversity compared to other dialect groups (**Fig. 2c**). Cantonese, Hakka and Min are the three main local dialects spoken by residences in Guangdong province located in the Southeast China and are clearly distinguishable from each other (**Fig. 2c**). Cantonese demonstrates notable genetic distance from Min, which is also a dialect in Fujian and Taiwan provinces in eastern China. The Hakka dialect falls in the middle of Cantonese and Min in the PC space (**Fig. 2b**). Xiang is spoken in Hunan province in central China and forms a unique cluster (**Fig. 2c**). Despite the small sample size, we can observe the genetic difference among the Gan, Hui, Wu, Li and Zhuang dialect groups (**Fig. 2c**).

Moreover, we investigated the ancestral components for the BIGCS and 1KGP3 population using ADMIXTURE (**Fig. 2d**). We inferred the optimal number of ancestral components to be K = 11, which shows the lowest cross-validation error, and visualized the structure for all the ethnic groups. Compared to the East Asian populations included in the 1KGP3, the BIGCS displays a specific genetic component in South China (marked in purple in **Fig. 2d**; Extended Data Fig. 4 and Table S3). To gain more insights on where the new ancestral component occurred, we investigated the new ancestral component with dialects information and found that the new component was substantially enriched in Cantonese speakers. Among the 20 participants that are almost pure for the new ancestral component (>98% of the genome according to ADMIXTURE), 16 of whom are Cantonese speakers (Table S4). These results suggest that the potential new ancestry in BIGCS is likely belong to the native Cantonese. This observation reflects that the southern Han Chinese (CHS) recruited in the 1KGP missed important ancestral components from South China.

Finally, as linkage disequilibrium (LD) decay is a useful measure in population genetic studies, which is affected by the recombination rate and the number of generations[33], we computed and compared the pattern of LD decay of the BIGCS population with other ancestral groups in 1KGP3 using PopLDdecay[34]. We observe faster LD decay in BIGCS population compared to the EAS group (Extended Data Fig. 5a), indicating the BIGCS population may have greater numbers of generations than the East Asian populations involved in 1KGP3. We further compared the pattern of LD decay for the BIGCS with the five 1KGP EAS populations, including JPT, CHS, CHB, CDX and KHV, respectively. We find that the LD of BIGCS population decays faster than CHS, CHB or JPT and slower than CDX and KHV (Extended Data Fig. 5b). This suggests that East Asians with a southern ancestry such as BIGCS, CDX, KHV may have an older population history than the East Asians resided in the northern areas, consistent with observations from recent studies based on ancient DNA[35].

GWAS of multiple adult and infant traits unravels novel East-Asian specific genetic association and distinct maternal and infant genetic effect on the same trait

There have been investigations using genome-wide association analyses (GWAS) for a few maternal and children's health related traits, but almost all were conducted in European populations[9–11,36]. More importantly, it is still unknown whether and to what extent the genetic effect of the same trait may differ between adults and infants. The scale of the BIGCS WGS data of individuals with comprehensive Chinese ancestries should empower the discovery of novel genes or pathways which may have an effect on the health related traits for East Asian mothers and infants. Therefore, we performed GWAS of 12 adult (including fathers and mothers) and 6 infant's quantitative traits (Extended Data Table 2). These traits correspond to the glycolipid metabolism and anthropometric measurements in BIGCS, respectively. Based on the power calculation, we focused on the individuals with Han ancestry and the variants with MAF > 1% using the linear mixed model in SAIGE[37] (**see Methods**). The statistic inflation (GC lambda) was negligible ($0.98 < \lambda_{GC} < 1.01$) for all the analyzed traits (Extended Data Table 2).

In total, we identified 11 and 4 independent loci reaching genome-wide significance ($P < 5\times10^{-8}$) for the 12 adult and 6 infant traits, respectively (**Fig. 3**; Extended Data Table 2; Table S5). LocusZoom[38] plots for all the significant loci were showed in Fig. S9. Replication of six of the signals that have replication data in a GWAS analysis that involved around 30K Chinese individuals sequenced by noninvasive prenatal testing (NIPT) indicates 100% replication rate (**see**

Methods). Among these signals, eight were located within ±50 kb from previously reported variants for the same trait in the studies across various population ancestry in GWAS catalog[39] (Table S5). Of the rest of the seven novel genetic loci, three belongs to maternal trait while four belongs to the infant traits (**Fig. 3**).

The novel locus associated with the maternal gestational weight gain (GWG, defined as the difference between the self-reported pre-pregnancy weight and the last recorded weight before delivery) rate (kg/week) is a 4 bases deletion located in the intron of gene *TTC28* in chromosome 22 (rs3840091, beta=0.33, P= 7.26×10^{-10}, **Fig. 3b**; **Fig. 4a**). We found that *TTC28* is mainly expressed in ovary according to the Human Protein Atlas (HPA)[40] (Fig. S10a) and several variants found in *TTC28* were associated with ovarian cyst, ovarian carcinoma, breast cancer and obesity in European populations[41,42]. This suggests that *TTC28* may have a crucial role for fertility function. The 4bp rs3840091 deletion variant in *TTC28*, which accelerated the weight gain significantly during gestation is common in BIGCS (AF=12.18%) and East Asian population (AF=13.04%) but is less common in European (AF=4.15%) or African (AF=2.17%) population according to the gnomAD database[27]. This may explain why the genetic effect on GWG was not identified before[39].

Notably, we also identify a novel missense variant associated with total bile acid (TBA) in the coding region of *SLC10A1* (rs2296651, c.800C>T/p.Ser267Phe, P=8.36×10^{-40}) (**Fig. 3b;** Table S5). This variants was previously reported as a pathogenic variant that caused sodium taurocholate co-transporting polypeptide (NCTP) deficiency[43]. The NCTP helps transport bile acids from the portal blood back into hepatocytes in enterohepatic recirculation[44] and is mainly expressed in liver according to the HPA (Fig. S10b). One recent case study associated this variant with cholestatic jaundice in early infants[45] and another case study found that a pediatric patient with NCTP deficiency gave birth to two female monozygotic twins with cholestatic liver disease due to raised serum BA[46]. We found that the allele frequency of rs2296651 was significantly higher in Southern China (9.96%) compared with that in Northern China (2.46%), and it's not found in non-East Asian populations (**Fig. 4b, c**). We infer that this observation may be explained by the fact of enduring infection by hepatitis B virus (HBV) among East Asian populations especially whom lived in South China, as the p.Ser267Phe variant were also reported to be associated with resistance to chronic HBV infection[47–49]. Our study suggests, for the first time, using a large-scale WGS data and molecular phenotypes from an East Asian population, that there is a clear association between p.Ser267Phe and high level of TBA (**Fig. 4d**). Although the p.Ser267Phe variant might prevent HBV from cellular entry, our study revealed that it still paid the cost by resulting in cholestasis of pregnancy, which harms the fetus and increases perinatal morbidity and mortality. Further study is needed to investigate how this information may contribute to the treatment of cholestasis during pregnancy to improve adult and infant health.

When investigating the four variants associated with the three infant traits, we notice a difference in the genetic effect between the adults and infants for the same trait (**Fig. 3b;** Table S5). For example, given the similar sample size, the chr12 locus associated with the low-density lipoprotein (LDL) level (rs137994041, *SOAT2*, beta=0.34, P=1.27×10^{-9}) in the infant cord blood did not show significant genetic association with LDL level in the maternal peripheral blood (beta=-0.002, P=0. 96). The rs137994041 is located in the intronic region of *SOAT2* and turns out to be a novel GWAS signal of LDL level in infant cord blood first revealed in this study. By investigating the RNA expression from HPA, we notice that *SOAT2* is highly expressed in small intestine and

duodenum (Fig. S10c). It suggests that there may exist some important genes such as *SOAT2* specifically affecting the lipid metabolism specifically for the infants but not the adults. On the contrary, there are also two loci, including the *APOE* locus associated with the total cholesterol (TC) level (rs72654473, beta=-0.48, P=3.77×10^{-16}) and the *CELSR2* locus associated with the LDL level (rs611917, beta=-0.59, P=2.66×10^{-14}) in the maternal peripheral blood, that do not show significant genetic effect on the same lipid trait in the infant cord blood (P>0.05). Except for those three loci, the rest of the seven loci demonstrated similar genetic effect on the same lipid trait in both the maternal peripheral and the infant's cord blood (**Fig. 3**; Table S5). To facilitate meta-analysis of the above and future traits, we have provided an online interface for conducting meta-analysis in our project website (http://gdbig.bigcs.com.cn/gwas_meta_analysis/jobs.html).

Inter-generational mendelian randomization dissects causal effects of maternal phenotype on fetal growth

According to the observational association analysis, fetal growth measurements such as birth weight, birth length and gestational duration are associated with a range of health outcomes, and are correlated with maternal phenotypes such as maternal height, weight, blood pressure (BP) and glucose level (**Table 1**). It's essential to understand the mechanisms underlying the relationship, that is, the consequence of socioeconomic status, shared genetics or intrauterine effects[11,50,51].-We utilized the trio and duo design in our study to investigate five mechanisms underlying the observed association between the maternal phenotype and the fetal growth-related pregnancy outcomes. Those mechanisms include (1) the mixed effect of maternal and/or fetal genetic effect by the maternal transmitted alleles (h1); (2) the maternal intrauterine effect mediated by the maternal non-transmitted alleles (h2); (3) the fetal genetic effect by the paternal transmitted alleles (h3); (4) the maternal effect (a linear combination of the effects of h1, h2 and h3) and (5) the fetal genetic effect (a linear combination of the effects of h1, h2 and h3) of a maternal phenotype on a pregnancy outcome (**see Methods**). For this purpose, we conducted an inter-generational mendelian randomization (MR) analysis by testing the haplotype genetic scores of seven maternal quantitative measurements (including maternal height, pre-pregnancy BMI, BP, fasting plasma glucose (FPG) level, TBA, triglyceride (TG) and TC) against three widely investigated pregnancy outcomes (including birth weight, birth length and gestational duration at birth) and three quantitative pregnancy outcomes (including the high-density lipoprotein (HDL), LDL and TC concentration in the cord blood) (**see Methods**).

The majority of the genetic scores constructed from the maternal genotypes (h1+h2), the maternal transmitted alleles (h1) or the maternal non-transmitted alleles (h2) explained more than 5% variance of the corresponding maternal phenotypes respectively (except for TBA and h1, h2 in TC), assuring the instrumental strength of the genetic scores in the MR analysis (Extended Data Table 3). In the MR analysis, the underlying mechanism of the observed phenotypic associations were dissected (**Table 1**; Table S6). For birth weight, while maternal height, pre-pregnancy BMI and FPG have demonstrated the most significant phenotypic associations, we have only observed causal effects of maternal height, FPG and BP on the birth weight outcome. The ratio estimates demonstrate a maternal effect of approximately 14.82g greater birth weight per unit (1.68 cm) increase in maternal height and a maternal effect of around 15.7g greater birth weight per unit (0.08 mmol/L) increase in maternal FPG. Although BMI demonstrated the strongest phenotypic association with birth weight, we didn't identify significant associations in any of the five

mechanisms investigated in our study, consistent with a previous study[52] and different from another study investigating the effects of obesity-related maternal traits on birth weight[53].

For birth length, maternal height and BMI have demonstrated the strongest phenotypic associations. However, only the genetic score of paternal transmitted alleles of height demonstrated a significant effect (P = 0.0017). Interestingly, despite no significant phenotypic association was identified, the ratio estimates of TBA indicate a significant negative maternal effect of 0.1 cm and an intrauterine effect of 0.08 cm shorter birth length per unit (0.49 umol/L and 0.32 umol/L) increase in maternal TBA. Despite no phenotypic associations were observed, maternal BP demonstrated significant negative fetal genetic effect, which is mainly from maternal transmitted alleles on both birth weight (P=0.01) and birth length (P=0.03), corresponding to 21.0g smaller birth weight and 0.076 cm shorter birth length per 0.5 mmHg increase in maternal BP.

We didn't find any significant associations between all the genetic scores investigated in our study and the gestational duration, although the associations between the genetic scores of maternal height, BP and FPG and gestational duration were previously reported in European population[52]. This may be due to the relatively small sample size in our study and/or the difference of genetic background from European population. For the three quantitative outcomes in the fetal cord blood, maternal BP has been found to impact the LDL in the cord blood through a negative maternal effect and a positive fetal genetic effect. Maternal TC affects both the LDL and TG in the cord blood via a positive fetal genetic effect (Table S6).

DISCUSSION

The study presented here reflects wealth of knowledge that can be obtained from whole genome sequencing and analysis of related individuals from a large-scale birth cohort. When properly utilizing the population and family relatedness information, we are able to reliably detect SNPs and Indels with high validation rate and construct a haplotype reference panel for more accurate genotype imputation, better than the existing and most widely used reference panels, despite the relative low sequencing coverage (~6.63x) (**Fig. 1**). This supports the prediction that a design of low and intermediate sequencing of large-scale individuals provides a better population reference compared with that of high depth sequencing of a few individuals[17]. In total, we have identified 15.78 and 2.53 million novel SNPs and Indels that were not present in the dbSNP (Build 154). Compared to 197 million variants that were discovered in the Chinese population from all the four WGS dataset published nowadays[2,5,26,27,30], our call set contributes to 19 million new variants that were not previously identified in any of those Chinese studies (10% of the total 197 million).

Compared to the few Chinese population genetic study published nowadays[5,26,33], the unique composition of BIGCS, namely the recruitment of southern Chinese combined with the collection of the speaking dialect information of the participants offer the opportunity to generate new insights into the Chinese local and global human shared ancestry. We have identified a novel southern Chinese specific genetic component (**Fig. 2**), which belongs to the Cantonese speakers and were not previously reported before to our knowledge. In addition, we have revealed the genetic relationship of 10 Chinese dialects including 8 major dialects - Mandarin, Cantonese, Min, Hakka, Gan, Xiang, Wu and Hui and 2 minor ones-Zhuang and Li. Those discoveries renewed our knowledge of the genetic diversity of the Chinese population, not only from a geographical but also a cultural perspective.

There is very limited knowledge regarding whether the genetic effects may change for the same trait in the human's developmental and aging program. In addition, studies in the genotype-phenotype associations have long been biased against the European population[1]. In the genome-wide association study of 12 maternal and 6 infants' quantitative traits that are either anthropometric measurements or involved in glycolipid metabolism (**Fig. 3**), we have identified novel loci (*SLC10A1*, *TTC28*, *SOAT2 and APOE*) associated with maternal total bile acid (TBA) and gestational weight gain (GWG) during pregnancy, as well as the LDL and the HDL level in the cord blood of infants, respectively. Notably, the sentinel missense variant (rs2296651) in the *SLC10A1* locus that impacts the plasma bile acid level in the Chinese population is almost absent in other populations. Interestingly, distinct genetic effects on the same trait are observed between the mothers and the infants. For example, the *SOAT2* locus that is significantly associated with the LDL in the infant cord blood is not associated with the LDL level in the maternal plasma. As another example, the *CELSR2* locus associated with LDL and the *APOE* locus associated with the total cholesterol (TC) in the maternal plasma did not significantly impact the LDL and the TC level in the infant cord blood. Those discoveries provide clear evidence that genetic effect on lipid metabolism changes as human grows and develops.

In the inter-generational mendelian randomization analysis, we have utilized the genetic data of the deep-phenotyped birth cohort to dissect the five mechanisms underlying the observed associations between maternal phenotypes and birth outcomes. We identify several notable causal relationships: (1) birth weight is positively affected by maternal height and FPG through a maternal causal effect but there is no causal relation between maternal BMI and birth weight; (2) birth length is positively affected by the paternal transmitted alleles of adult height and negatively affected by the maternal and intrauterine effect of TBA; (3) maternal BP negatively influences birth weight and length mainly through a fetal effect; (4) maternal BP and TC demonstrated maternal or fetal genetic effects on the LDL and TG in the cord blood, respectively. As intrauterine effect corresponds to modifiable maternal phenotypes, lowering abnormally high maternal total bile acid may result in greater birth length. In the future, the causal effects of more comprehensive maternal phenotypes should be examined in the birth cohort to identify maternal predisposition that may implicate clinical interventions so as to prevent adverse birth outcomes.

Different from microarray genotyping, whole genome sequencing is a rapidly evolving and relatively new technology for population genetic and genome-wide association analysis. For several gene association signals identified in this study, we have replicated it in another independent cohort. Nonetheless, due to limited genetic study of the corresponding traits in the pregnancy and child period, replication is not available for all of the signals. The current sample size of the study (N=4,053 including 332 trios, 1,406 duos and 245 unrelated individuals) limits our power to detect low-frequency and rare variants associated with the important traits collected in the birth cohort. In our inter-generational mendelian randomization analysis, we didn't identify the potential causal relation between height, BP and FPG and gestational duration as reported by one previous study[52]. We cannot conclude at the moment on the reasons for this difference. Possibilities can be due to lack of power in our study, different genetic and physiological features between populations and artifacts induced by meta-analysis in the previous study. Future efforts are required to develop efficient methods to compensate the time-consuming clinical trials to validate and accommodate the difference of the discoveries.

The method adopted and developed as well as the novel genetic loci and causal correlation identified in this study provide a proof-of-concept methodological framework for future genetic study of human population. Along with the continuous development and the release of genomic data in the BGICS, which has recruited and deeply phenotyped more than 50,000 maternal-infant pairs in China, future effects will be prioritized to address the abovementioned limitations, update the variation dataset, the reference panel, the genotype-phenotype associations and dissect the contribution and interplay between environmental and genetic factors in a systematic framework.

MATERIALS AND METHODS

Cohort description

BIGCS is a largest-scale prospective cohort of mothers and their children from the prenatal period to adult life in the city of Guangzhou, China[16], launched in 2012. The BIGCS cohort recruits pregnant women attending their first routine antenatal examinations (usually around the sixteenth gestational week) and their husbands and offspring at the Guangzhou Women and Children Medical Center (GWCMC). Eligible participants were identified and asked for consent to take part in BIGCS by trained personnel. All participants in BIGCS have provided informed consent.

BIGCS has recruited more than 50,000 of trios/duos families by the end of 2021. During the pregnancy period, epidemiological information, clinical assessment and biological samples are obtained at recruitment, 24-28 and 35-38 gestational weeks, respectively. Phenotypic and clinical information such as antenatal screening test results (e.g. Oral Glucose Tolerance Test (OGTT), coagulation test, and maternal biochemistry), ultrasound examinations, and obstetric complications and prescribed medication are obtained from the electronic medical records (EMR). After delivery, maternal information including living habits, anthropometric measurements and physical and mental health-related status are collected at multiple time-points. Information about the neonates, including the details of the delivery, birth characteristics (e.g. birth weight, Apgar scoring) and perinatal outcomes are also obtained in EMR. Children are followed up by doctors in the cohort clinics at ~6 weeks, 6 months, 1, 3, 6 and around 9 years. All cohort children are tracked till 18 years old. More details of BIGCS were summarized online at http://www.bigcs.com.cn/menu_f4.html and https://clinicaltrials.gov/ct2/show/NCT02526901.

From all the BIGCS participants, with the aim to understand the genetic composition of the cohort, the genetic contribution to pregnancy specific phenotypes and to use the genetic data to strengthen causal inference in observational research, we selected 4,215 individuals for whole genome sequencing, including 353 trios, 1,493 duos (8 father-offspring and 1,485 mother-offspring) and 170 unrelated single individuals (53 infants, 100 adult females and 17 adult males) and conducted bioinformatics and statistical analyses. This study was approved by the Ethics Committee of GWCMC and the Human Genetic Resources Administration of China (HGRAC).

Whole-genome sequencing of the 4,215 participants

Paired-end (PE) 100 bp whole-genome sequencing with a mean insert size of 214 bp was performed on the BGISEQ-500 platform with Magnetic Beads Blood Genomic DNA Extraction Kit (Shenzhen, China) and the average sequencing depth was ~6.63x (Fig. S2a, b). The duplication rate is low which is 1.7% on average (Fig. S2c). The adaptor sequences and poor quality bases from raw sequencing data were filtered by SOAPnuke (v1.5.6)[54]. A read was removed if the read sequence matched the adaptor sequences with less than 2 mismatches or if the proportion of low-quality bases (base quality < 12) was greater than 50% or if it contained more than 10% N bases.

SNP array genotyping of 240 participants

We selected 240 adult female participants out of the total 4,215 participants for array genotyping using the Illunima GSA-24 (v1.0) BeadChip SNP array (https://www.illumina.com/products/by-type/microarray-kits/infinium-global-screening.html).

to generate a gold-standard SNP dataset and these genotyped data were used to be the benchmark in quality control (QC) processes of variants calling in this study.

High-coverage whole-genome sequencing of 50 participants

We performed 140bp PE sequencing for 50 extra healthy Chinese participants of BIGCS cohort who were not included in the participants mentioned above to a high-coverage (40x on average) using Illumina HiSeq X10 platform. We aligned the clean reads to the same GRCh38/hg38 reference genome by BWA-MEM (v0.7.17)[55] and applied the GATK(v4.1.8.1) best practice joint calling protocol to detect and genotype variants for these participants. After VQSR and removal of the multi-allelic variants, we obtained 11,174,603 high-quality genotyped biallelic variants, including 9,816,793 SNPs and 1,357,810 Indels and the results were used to benchmark the genotype imputation accuracy.

BIGCS variant calling pipeline

In order to make the variant calling for thousands of samples easier and more efficient, we developed a handy variant calling pipeline for BIGCS, named as ilus (see data and code availability), which was based on the best practice of GATK[18] multi-sample joint calling framework and several useful QC processes and a series of data statistic functions (Extended Data Fig. 1).

Before performing ilus, we removed four samples without phenotype information and five samples had high sequencing error (>0.01) (Fig. S2d). We used the human reference genome assembly 38 (GRCh38/hg38) in our study and constructed an ilus pipeline to conduct alignment and variant calling for the 4,209 samples. The pipeline uses BWA-MEM (v0.7.17)[55] to align sequencing reads, VerifyBamID2 (v1.0.6) to detect high contaminated samples[56], samtools[57] to sort and merge alignment reads from different sequencing lanes or libraries of the same sample, GATK (v4.1.8.1) MarkDuplicates to identify and mark duplicates reads, GATK BaseRecalibrator to recalibrate the quality of bases, GATK HaplotypeCaller in GVCF mode and GATK GenotypeGVCFs to jointly call variants, GATK VQSR with truth-sensitivity-filter-level of 99.0 for both the SNPs and Indels, and in-house scripts written in Python and R for data statistics.

Contaminated and sex ambiguous samples

We assessed the level of DNA contamination by VerifyBamID2 (v1.0.6) in ilus for each sample and removed 50 samples with contamination level $\alpha > 0.01$ (Fig. S2e). We re-inferred the sex for each sample by computing the rate of average sequencing depth between the chromosome X and autosomes (X/A ratio in short). The expected relative X/A ratio is 0.5 for male and 1.0 for female, respectively. We used 0.75 as a threshold of X/A ratio and infer a sample as male (<0.75) or as female (≥0.75) (Fig. S2f). We found the inferred sex contradicts the self-reported sex in 21 samples. We confirmed the correctness of the inferred sex for 7 of them by reviewing the follow-up records, and removed the rest of the 14 samples because of the ambiguous sex judgement.

Genotype calling and family relatedness

After the filtering of contaminated and sex ambiguous samples, we performed variant calling and genotyping for the remaining 4,142 samples. For the initial screening before any filtration processes, we identified 83,877,374 candidate variants with a transition-to-transversion ratio (Ts/Tv) of 1.93 for single nucleotide polymorphisms (SNPs).

We used KDG (with default parameters) which allows parentage to be assigned from low-depth sequencing data to identify the family relatedness for all the samples[58]. We find that the predicted relationship of 68 samples contradicts the self-reported family (trios or duos) relationship. We confirmed the correctness of relationship for 14 of those samples by double-checking the follow-up records, while the rest 54 samples (~1%) were removed from our study.

We filtered out the samples that had extraordinarily low or high number of variants (N < mean-3SD or N > mean + 3SD) as well as the samples that had variant low call rate (less than 90% of the total variants). Therefore, we additionally removed 28 and 7 samples due to abnormal number of variants or low call rate, respectively. Finally, we obtained 4,053 samples, consisting of 332 parent-offspring trios (father-mother-offspring), 1,406 parent-offspring duos (14 father-offspring and 1,392 mother-offspring) and 245 unrelated single individuals (70 infants, 150 adult females and 25 adult males) (**Fig. 1b**). The 4,053 participants cover 13 ethnicities (Han Chinese and 12 minorities) from 30 of 34 administrative divisions of China (**Fig. 1a;** Fig. S1; Table S1).

We then performed the variant calling and genotyping for those samples again and obtained the variation call set consisting of 82,778,859 candidate variants, including 74,001,658 SNVs (Ts/Tv=1.94) and 10,799,865 Indels. Multiple types variants were calculated multi-times at the same chromosomal position.

Variant QC, annotation and validation

Variants QC step1: Recalculating InbreedingCoeff and ExcessHet by adding pedigree-related annotations and variant quality score recalibration (VQSR). We initially called the variants without pedigree information by performing the HaplotypeCaller and GenotypeGVCFs modules of GATK (v4.1.7.0). Since many of the samples in BIGCS are trios or duos, we recalculated the inbreeding coefficient (InbreedingCoeff) and the rate of excessive heterozygosity (ExcessHet) for each variant by adjusting pedigree-rated information by using VariantAnnotator module of GATK. Then the VQSR of GATK were performed to filter out low quality variants. Know variants for VQSR were downloaded from GATK resource bundle and GATK best practice parameters was adopted. The --truth-sensitivity-filter-level was set to be 99.0 for both of SNPs and Indels.

Variants QC step2: Calculating Genotype posteriors and filtration before LD-based refinement. In order to further improve the accuracy of variant calling and to filter less reliable genotypes, we performed GATK CalculateGenotypePosteriors to derive genotype posterior probability and improve the genotype likelihoods (PLs) for each sample with family pedigree prior information of BIGCS and the allele frequency prior information from 1KGP3 population (1000G.phase3.integrated.sites_only.no_MATCHED_REV.hg38.vcf.gz). All the biallelic variants marked as PASS were selected and the rest were removed. Furthermore, we filtered out variants or masked the genotype as missing when matching any one of the following criteria:
(1) ExcessHet ≥ 60 in the INFO column calculated by GATK.
(2) Mendelian violation rate > 5%.
(3) Sequencing depth less than 4x in each of the samples.
(4) Allele fraction < 0.2 for all non-reference alleles of heterozygous genotypes in all of the samples.

(5) Missing genotype call rate > 10%.
(6) Variants in low complexity regions (LCR).

After those filtrations, we obtained 58,837,923 biallelic variants, including 52,964,853 biallelic SNPs (Ts/Tv=2.08) and 5,873,070 biallelic Indels.

Variants QC step3: LD-based genotype refinement by BEAGLE. To further improve the genotyping accuracy, we used BEAGLE software version 4.0, which was the version of BEAGLE that can integrate pedigree structure and genotype likelihoods as input to perform LD-based genotype refinement[59]. Before the phasing step, we set individual genotypes with GQ < 20 in the sample FORMAT field calculated by GATK as missing by VCFtools[60] (version 0.1.16). For X chromosome, we assume the variants in the non-pseudo-autosomal regions (non-PARs) include only homozygous genotype for the males. Therefore, we converted all the heterozygous variants in non-PAR regions of chromosome X as missing in males and then phased all samples together. In order to accelerate the phasing process, we ran BEAGLE in parallel for autosomes and X chromosome by splitting each chromosome into 10kb chunks (with about 300 variants per chunk on average). We used BCFtools software to perform splitting and merging the variants in VCF format[57]. After removing low quality variants with dosage $R^2 \leq 0.3$ estimated by BEAGLE and merging back the variants on Y chromosome, the final variants call set includes 4,053 samples and 56,230,613 biallelic variants, contained 51,052,456 SNPs (Ts/Tv=2.09) and 5,178,157 Indels. Variants were then annotated by Variant Effect Predictor[21] (VEP) (v95 GRCh38) with the default parameters.

SNPs validation. We compared the SNP genotypes of final call set versus the call set derived from the abovementioned DNA array of 240 samples. The genotype concordance rates in varying non-reference allele frequency bins were summarized in Figure S4a and were generally greater than 0.97.

BIGCS reference panel construction

We constructed the BIGCS haplotype reference panel for chromosomes 1-22 and X using BEAGLE 5.0. For X chromosome, the heterozygous genotypes had been set to missing in non-PAR regions for male samples before phasing as described above, and the chromosome were divided into PARs and non-PAR regions for all samples (male and female) and phased. We extracted all of the 2,245 unrelated samples from the above phasing result, which consist of 43,055,086 high quality SNPs plus 4,184,387 Indels to generate the BIGCS reference panel. The phased data were converted to m3vcf format for each chromosomes, which is a classic reference format in Minmac3 (v2.0.1)[61]. To evaluate the performance of BIGCS panel, we used the variants detected from high-coverage WGS data (~40x on average described above) from extra 50 healthy Chinese individuals who are unrelated to any individuals in the BIGCS panel. We extracted the 0.93 million SNPs present in the Affymetrix Genome-Wide Human SNP Array 6.0 from these 50 samples, and performed genotype imputation for the rest variants using Minmac3. Imputation accuracies were estimated using Pearson's R^2 between the genotype dosage from high-coverage WGS and those imputed by the BIGCS reference panel.

PCA, ADMIXTURE and F_{st} analysis

The population genetic structure was analyzed via PCA and ADMIXTURE[62] by merging a dataset of autosomal biallelic SNPs of BIGCS unrelated parental samples and a full 1KGP3 dataset. We used PLINK2 (v2.00a3LM) [63] to select the SNPs with MAF ≥ 5%, genotype missing rate < 5% and HWE *P* value > 1.0×10^{-6}. Moreover, we performed LD prune using "--indep-pairwise 50 5 0.5" for SNPs in PLINK2, yielding 1,215,873 biallelic sites for PCA among BIGCS unrelated samples and a full 1KGP3. PCA was carried out using smartpca command from the EIGENSOFT program[64]. We computed the top 20 PCs with options "numoutevec: 20" and found that PC1 and PC2 reflected the main genetic differentiation with worldwide populations (Fig. S6). We applied ADMITXTURE analysis from K=2 to K=15 based on the same SNPs used in PCA with the default parameters. The K value with smallest cross-validation error rate was chosen as the best model, which was K=11 in our study (Fig. S8).

We also performed PCA in the full BIGCS unrelated parental samples (N=2,245) base on the same filtration criteria for SNPs described above and found the genetic evidences linked to the dialects of China.

We next computed the pairwise *F*st estimates between the populations using VCFtools[60] (v0.1.16) based on the shared autosomal common biallelic SNPs (MAF≥5% and genotype missing call rate < 10%) of the above BIGCS unrelated parental samples and 1KGP3 (2,504 individuals) and of the BIGCS unrelated parental samples alone. We built up pairwise *F*st values matrix for the populations with window size and window step set to 20,000 and 5,000. Extended Data Fig.2 was generated by performing hierarchical clustering using the complete-linkage method implemented in the *hclust* function of *pheatmap* package in R.

Genome-wide association analysis

We applied SAIGE[37] to carry out linear regression model for genotype-phenotype association tests using the default parameters. We only performed GWAS for Han Chinese adults and infants. We analyzed18 relevant quantitative traits for adults and infants in our study, including 3 parental traits (Height, Weight and BMI), 9 maternal traits during pregnancy and 6 traits of infants. Before performing GWAS, we filtered out individuals and variants which matched any one of the following criteria for each trait:

For individuals:
(1) Samples having rare disease;
(2) Samples having relatedness within the selected samples;
(3) Samples whose trait were missing.

For variants (including SNPs and Indels):
(1) The variants had MAF < 0.01;
(2) The variants had HWE *p* value < 1.0×10^{-6};
(3) The variants had genotype missing call rate > 0.01 (since LD-based refinement was applied there was no missing call).

After performing the above filtrations, the sample size and variants count for GWAS of each trait was presented in Extended Data Table S2 and the covariates used for each trait are described below:

For the GWAS of height, pre-pregnancy weight and BMI of parents which were measured at recruitment, we used the sex, age and the first 10 PCs from the PCA as the covariates. For the GWAS of 9 quantitative traits for mother, we used the age, pre-pregnancy BMI, gravidity, parity and the first 10 PCs from PCA as the covariates. For the GWAS of fetal cord blood and maternal lipid metabolism, including HDL, LDL, TC and TG, we additionally included gestational duration of fetus as a covariate. For the GWAS of infant traits, we used the infants' sex and the first 10 PCs as covariates. For the birth weight and birth length, we additionally included gestational duration of the fetus as a covariate. We note that the birth weight and birth length phenotype in infants were related to gestational duration. Since we are mostly interested in the genetic effects on infants' phenotype in this study, we used the gestational duration and fetus sex as covariates.

For all the GWAS, we used 5×10^{-8} as the genome-wide significant *P* value threshold, and the lead SNP of independent locus was defined as the variant with the smallest *P* value and not in linkage disequilibrium (LD $r^2 < 0.1$) with any other variants within a window of 1.0 Mb. LocusZoom was applied to visualized the loci (Fig. S9).

Replication of GWAS significant loci

For replication, we compared all the significant variants to an independent study that had the same traits of our study in about 30,000 Chinese pregnancy women with NIPT genome data with an average sequencing depth ~0.1x for each sample (accession number GVM000325, manuscript in submission).

Haplotype-based polygenic risk score (PRS) and association analysis with fetal growth

Data preparation. Before constructing genotype and haplotype-based PRS, we need to select suitable proxy loci which had been reported by GWAS as a base dataset. We downloaded the GWAS SNPs of maternal height, BMI, FPG and BP in European ancestry (https://doi.org/10.1371/journal.pmed.1003305.s003), which had been selected and reported by a previous study[52]. There are 2,130 SNPs for height, 628 SNPs for BMI, 22 SNPs for FPG and 831 SNPs for BP in this dataset. We converted the genome coordinates of these SNPs from human genome GRCh37 (hg19) to human genome GRCh38 assembly using the CrossMap software[65]. After removing the variants failing the conversion and the variants that were not in our variant dataset, we obtained 2007, 603, 19 and 759 SNPs for maternal height, BMI, FPG and BP, respectively.

We also used the GWAS SNPs of maternal height, BMI, FPG, total bile acid (TBA) and total cholesterol (TC) from ~30,000 Chinese pregnancy women with NIPT data (accession number GVM000325, manuscript in submission). We selected the SNPs for each of these traits above based on the conditional & joint COJO analysis by the GCTA software[66] with the parameters "--maf 0.01 --cojo-wind 2000 –cojo-slct". After that, we have obtained 33, 17, 13, 8, 47 and 34 SNPs for maternal height, BMI, FPG, TBA, TG and TC in the raw dataset, respectively.

Since PRS is specific to a certain population, we used the Beta, SE, *P* value and allele frequency from the NIPT GWAS for all the overlap loci for maternal height, BMI, FPG, TBA, TG and that were present in both the European and the NIPT Chinese GWAS and for the non-overlap loci. For BP which Chinese data is not available, we used the values from the European data. After removing the loci that were not in the BIGCS variant dataset, we finally obtained 2033, 617, 30,

759, 7, 45 and 28 SNPs for maternal height, BMI, FPG, BP, TBA, TG and TC, respectively (Table S7-S13).

PRS calculation. We constructed the genotype and haplotype-based PRS for the corresponding phenotypes using the following equation:

$$PRS_i = \frac{\sum_j^n b_j G_{ij}}{n} \quad \text{(I)}$$

Where, i denotes the ith individual, b_j is the estimated effect size for the jth SNP reported by GWAS studies and n is the total number of SNPs that were used to compute PRS for the ith individual. There were two ways to compute G_{ij}:

- For maternal and fetal genotype, G_{ij} is the genotype dosage of SNP.
- For investigating the genetic effects of parental transmitted alleles and maternal non-transmitted allele underlying the maternal observed traits, G_{ij} is computed by the following equation:

$$G_{ij} = \begin{cases} \dfrac{h_1 D_c}{\max(1, G_c)}, & (for\ materna\ transmitted\ allele) \\ \dfrac{h_2 D_m}{\max(1, G_m)}, & (for\ maternal\ non-transmitted\ allele) \\ \dfrac{h_3 D_c}{\max(1, G_c)}, & (for\ paternal\ transmitted\ allele) \end{cases} \quad \text{(II)}$$

Where, G_{ij} represents the haplotype dosage value of maternal transmitted allele (h_1), maternal non-transmitted allele (h_2) and paternal transmitted allele (h_3) of the ith individual in the jth SNP, respectively. The haplotype allele of h1, h2 and h3 were determined using the phased genotype and pedigree data and the value is either 0 or 1. $G_c = h_1 + h_3$ represents the infant's genotype, $G_m = h_1 + h_2$ indicates the maternal genotype; D_c and D_m are the maternal and infant genotype dosage, respectively**.**

Associations between maternal genotype PRS and maternal phenotypes. We built a linear regression model to test and verify the explained fraction of the maternal phenotypic variances by the contributions of maternal genotype-based PRS, the maternal transmitted (h1) and maternal non-transmitted (h2) haplotype-based PRS, which were used as genetic instrumental variables for the corresponding maternal phenotypes. Here is the regression equation:

$$Y \sim X_{prs}\beta_{prs} + \boldsymbol{X}_{cov}\boldsymbol{\beta}_{cov} + e \quad \text{(III)}$$

Where Y is the maternal phenotype value and X_{prs} are the PRS value. For the association analysis of all the maternal traits, we used maternal age and the top 10 PCs as the covariates. The summary statistics of the regression were presented in Extended Data Table 3.

Inter-generational mendelian randomization model to explain the causality between the maternal phenotype and fetal growth. In order to investigate the genetic and intrauterine effect of mother and fetus on fetal growth relevant birth outcomes such as gestational duration, birth weight, birth length and biochemical measurements from the cord blood, we constructed a mendelian randomization (MR) model using the three haplotype-based PRS for the corresponding maternal phenotypes based on the h1, h2 and h3, respectively, as the genetic instrumental variable, and the fetal growth values as outcome phenotypes. In this model, the association of PRS score of maternal non-transmitted haplotype (h2) with a birth outcome was a signal of maternal intrauterine effect, while the association with the PRS score of paternal transmitted haplotype (h3) was a direct fetal effect on a birth outcome.

However, considering the h1 haplotype was transmitted from maternal genome to fetal genome, it contains both of the maternal and fetal effect on the birth outcome. Under the assumption of additivity between maternal and fetal effect and zero parent of origin effect, the effect of h1 should be the summation of the effect of h2 and h3. Thus, we can model the joint maternal effect and the fetal genetic effect by the linear combinations of a regression coefficients of the three haplotype-based PRS ($\boldsymbol{\beta}_{h1}$, $\boldsymbol{\beta}_{h2}$ and $\boldsymbol{\beta}_{h3}$), which was first defined in Zhang's study[52]. We used the same joint maternal effect and fetal genetic effect on a fetal growth defined by the following equations:

- Maternal effect: $\boldsymbol{\beta}_M = \frac{1}{2}(\boldsymbol{\beta}_{h1} + \boldsymbol{\beta}_{h2} - \boldsymbol{\beta}_{h3})$ is the joint effect of maternal haplotypes and excludes the possible fetal effect of the paternal transmitted haplotype;
- Fetal genetic effect: $\boldsymbol{\beta}_F = \frac{1}{2}(\boldsymbol{\beta}_{h1} + \boldsymbol{\beta}_{h3} - \boldsymbol{\beta}_{h2})$ is the joint effect of maternal and paternal transmitted haplotypes and exclude the possible maternal effect of the maternal non-transmitted haplotype, i.e. the intrauterine effect.

The $\boldsymbol{\beta}_{h1}$, $\boldsymbol{\beta}_{h2}$ and $\boldsymbol{\beta}_{h3}$ are the coefficients of the association analysis between the haplotype-based PRS scores and birth outcome using the following linear regression model:

$$Y \sim X_{h1}\beta_{h1} + X_{h2}\beta_{h2} + X_{h3}\beta_{h3} + \boldsymbol{X}_{cov}\boldsymbol{\beta}_{cov} + e \quad \text{(IV)}$$

Where, Y was the value of birth outcome. X_{h1}, X_{h2} and X_{h3} represented the haplotype-based PRS score of h1, h2 and h3, respectively. X_{cov} represented a list of the covariates depending on the traits of MR analysis.

The computation of the PRS has been described above in the "PRS calculation" section and equation (I). The covariates used for each trait in our study are described below:

- For maternal height and prepregnant BMI PRS as exposure, and birth weight, birth length, HDL/LDL/TG of cord blood as outcome, the list of covariates includes maternal age, gestation age and first 10 PCs of PCA results.
- For maternal height and prepregnant BMI PRS as exposure, and gestational duration (GAW) as outcome, the list of covariates includes maternal age and first 10 PCs of PCA results.

- For maternal BP, FPG, TBA, TG or TC PRS as exposure, and birth weight, birth length, HDL/LDL/TG of cord blood as outcome, the list of covariates includes maternal age, maternal height, maternal prepregnant BMI, gestation age and first 10 PCs of PCA results.
- For maternal BP, FPG, TBA, TG or TC PRS as exposure, and gestational duration as outcome, the list of covariates includes maternal age, maternal height, maternal prepregnant BMI, and first 10 PCs of PCA results.

We perform a linear hypothesis testing of the linear combination of the regression coefficients ($\boldsymbol{\beta}$). The *P* values of the testing result demonstrates the statistical significance of the maternal and fetal genetic effect towards the fetal growth (**Table 1** and Table S6).

Data Availability

Raw sequencing data have been deposited to the Genome Sequence Archive for Human (https://ngdc.cncb.ac.cn/gsa/) at the BIG Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, under the accession number (in application). In compliance with the regulations of the Ministry of Science and Technology of the People's Republic of China, the raw sequencing data contain information unique to an individual and thus require controlled access. To facilitate the use of the BIGCS information, we have established the genome database server GDBIG (http://gdbig.bigcs.com.cn/) with user-friendly website interface. Researchers can register and gain the assess to use data including allele frequencies of all variants, BIGCS reference panel and GWAS summary statistics online. Details on how to use the server can be referred to a companion paper of this study. Researchers who are interested to collaborate with the BIGCS group, are welcome to contact Xiu Qiu (xiu.qiu@bigcs.org) and data.bigcs@bigcs.org.

Code Availability

The ilus (the variant calling pipeline generator) and GDBIGtools are all available in the Github repository using the following links:
ilus: https://github.com/ShujiaHuang/ilus
GDBIGtools: https://github.com/BIGCS-Lab/GDBIGtools
Script for distinguish of the parental haplotype alleles from infant genotype and calculation of the genotype/haplotype-based PRS:
https://github.com/ShujiaHuang/genotools/blob/master/scripts/mr.py
Other software and databases used in this study are publicly available, and the URLs are listed below:
SOAPnuke (v1.5.6): https://github.com/BGI-flexlab/SOAPnuke
BWA-MEM (v0.7.17): https://github.com/lh3/bwa
verifyBamID2 (v1.0.6): https://github.com/Griffan/VerifyBamID
GATK (v4.1.8.1): https://github.com/broadgsa/gatk/
SAMtools (v1.9): http://samtools.github.io/
bedtools (v2.27.1-65-gc2af1e7-dirty): https://github.com/arq5x/bedtools2/
Variant Effect Predictor (release 95): https://github.com/Ensembl/ensembl-vep
Beagle (v4.0): https://faculty.washington.edu/browning/beagle/beagle.r1399.jar
Minimac3 (v 2.0.1): http://genome.sph.umich.edu/wiki/Minimac3
CrossMap (version 0.2.2): http://crossmap.sourceforge.net/

dbSNP Build 154: http://www.ncbi.nlm.nih.gov/SNP/
GATK bundle (hg38): https://console.cloud.google.com/storage/browser/genomics-public-data/resources/broad/hg38/v0
Human genome reference (GRCh38/hg38):ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/seqs_for_alignment_pipelines.ucsc_ids/GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.gz
The low complexity regions of GRCh38: https://github.com/lh3/varcmp/blob/master/scripts/LCR-hs38.bed.gz
The GWAS Catalog: https://www.ebi.ac.uk/gwas/
The Human Protein Atlas: https://www.proteinatlas.org/
The public GWAS SNPs used in constructing genotype-based PRS and haplotype-based PRS: https://doi.org/10.1371/journal.pmed.1003305.s003
We used Python (version 3.7.6) and R (version 4.1.1) extensively to analyze data and create plots. The Venn and ADMIXTURE plots were created by using a Python library: https://github.com/ShujiaHuang/geneview
Figure S10 was created by using: https://www.proteinatlas.org/
Figure 4c was created by using: https://popgen.uchicago.edu/ggv/

References

1. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26–31 (2019).
2. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
3. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
4. Maretty, L. *et al.* Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. *Nature* **548**, 87–91 (2017).
5. Cao, Y. *et al.* The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. *Cell Res.* **30**, 717–731 (2020).
6. Canova, C. & Cantarutti, A. Population-based birth cohort studies in epidemiology. *International Journal of Environmental Research and Public Health* vol. 17 1–6 (2020).
7. Harville, E. W., Kruse, A. N. & Zhao, Q. The Impact of Early-Life Exposures on Women's Reproductive Health in Adulthood. *Curr. Epidemiol. Reports 2021 84* **8**, 175–189 (2021).
8. Middeldorp, C. M., Felix, J. F., Mahajan, A. & McCarthy, M. I. The early growth genetics (Egg) and early genetics and lifecourse epidemiology (eagle) consortia: Design, results and future prospects. *Eur. J. Epidemiol.* **34**, 279–300 (2019).
9. Zhang, G. *et al.* Genetic associations with gestational duration and spontaneous preterm birth. *N. Engl. J. Med.* **377**, 1156–1167 (2017).
10. Liu, X. *et al.* Variants in the fetal genome near pro-inflammatory cytokine genes on 2q13 associate with gestational duration. *Nat. Commun.* **10**, 1–13 (2019).
11. Horikoshi, M. *et al.* Genome-wide associations for birth weight and correlations with adult disease. *Nature* **538**, 248–252 (2016).
12. Tao, F. B. *et al.* Cohort profile: The china-anhui birth cohort study. *Int. J. Epidemiol.* **42**, 709–721 (2013).
13. Zheng, J. S. *et al.* Cohort Profile: The Jiaxing Birth Cohort in China. *Int. J. Epidemiol.* **46**, 1382-1382g (2017).
14. Lin, J. *et al.* Cohort Profile: The Shanghai Sleep Birth Cohort Study. *Paediatr. Perinat. Epidemiol.* **35**, 257–268 (2021).
15. Hu, Z. B. *et al.* Profile of China National Birth Cohort. *Chinese J. Epidemiol.* **42**, 569–574 (2021).
16. Qiu, X. *et al.* The Born in Guangzhou Cohort Study (BIGCS). *Eur. J. Epidemiol.* **32**, 337–346 (2017).
17. Li, Y., Sidore, C., Kang, H. M., Boehnke, M. & Abecasis, G. R. Low-coverage sequencing: Implications for design of complex trait association studies. *Genome Res.* **21**, 940–951 (2011).
18. DePristo, M. a *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
19. Sherry, S. T. *et al.* DbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
20. Besenbacher, S. *et al.* Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. *Nat. Commun.* **6**, 1–9 (2015).
21. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).

22. Ng, P. C. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
23. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **Chapter 7**, Unit7.20 (2013).
24. Shen, Z., Li, P., Sun, H. & Pang, L. Geographical patterns of Chinese ethnic minority population composition and ethnic diversity. *Chinese Geogr. Sci.* **21**, 454–464 (2011).
25. Browning, S. R. & Browning, B. L. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
26. Liu, S. *et al.* Genomic Analyses from Non-invasive Prenatal Testing Reveal Genetic Associations, Patterns of Viral Infections, and Chinese Population History. *Cell* **175**, 347-359.e14 (2018).
27. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
28. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
29. Jeffrey D. Wall, Eric W. Stawiski, Aakrosh Ratan, Hie Lim Kim, Changhoon Kim, Ravi Gupta, Kushal Suryamohan, Elena S. Gusareva, Rikky Wenang Purbojati, Tushar Bhangale, Vadim Stepanov, Vladimir Kharkov, Markus S. Schröder, Vedam Ramprasad, Jennifer Tom, S, S. C. S. & A. S. P. The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature* **576**, 106–111 (2019).
30. Zhang, P. *et al.* NyuWa Genome resource: A deep whole-genome sequencing-based variation profile and reference panel for the Chinese population. *Cell Rep.* **37**, 110017 (2021).
31. Baker, H. D. R. Language atlas of China. *Bull. Sch. Orient. African Stud.* **56**, 398–399 (1993).
32. Barnes, D. The language of instruction in Chinese communities. *Int. Rev. Educ.* **24**, 371–374 (1978).
33. Stumpf, M. P. H. & Goldstein, D. B. Demography, Recombination Hotspot Intensity, and the Block Structure of Linkage Disequilibrium. *Curr. Biol.* **13**, 1–8 (2003).
34. Zhang, C., Dong, S. S., Xu, J. Y., He, W. M. & Yang, T. L. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* **35**, 1786–1788 (2019).
35. Wang, T. *et al.* Human population history at the crossroads of East and Southeast Asia since 11,000 years ago. *Cell* **184**, 3829-3841.e21 (2021).
36. Beaumont, R. N. *et al.* Genome-wide association study of offspring birth weight in 86 577 women identifies five novel loci and highlights maternal genetic effects that are independent of fetal genetics. *Hum. Mol. Genet.* **27**, 742–756 (2018).
37. Zhou, W. *et al.* Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. *Nat. Genet.* **52**, 634–639 (2020).
38. Pruim, R. J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010).
39. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).

40. Thul, P. J. & Lindskog, C. The human protein atlas: A spatial map of the human proteome. *Protein Sci.* **27**, 233–244 (2018).
41. Sakaue, S. *et al.* A cross-population atlas of genetic associations for 220 human phenotypes. *Nat. Genet.* **53**, 1415–1424 (2021).
42. Phelan, C. M. *et al.* Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. *Nat. Genet.* **49**, 680–691 (2017).
43. Pan, W. *et al.* Genetic polymorphisms in Na+-taurocholate co-transporting polypeptide (NTCP) and ileal apical sodium-dependent bile acid transporter (ASBT) and ethnic comparisons of functional variants of NTCP among Asian populations. *Xenobiotica* **41**, 501–510 (2011).
44. Claro Da Silva, T., Polli, J. E. & Swaan, P. W. The solute carrier family 10 (SLC10): Beyond bile acid transport. *Molecular Aspects of Medicine* vol. 34 252–269 (2013).
45. Li, H. *et al.* Clinical and genetic analysis of a pediatric patient with sodium taurocholate cotransporting polypeptide deficiency. *Chinese J. Contemp. Pediatr.* **20**, 279–284 (2018).
46. Tan, H. J., Deng, M., Qiu, J. W., Wu, J. F. & Song, Y. Z. Monozygotic Twins Suffering From Sodium Taurocholate Cotransporting Polypeptide Deficiency: A Case Report. *Front. Pediatr.* **6**, (2018).
47. Peng, L. *et al.* The p.Ser267Phe variant in SLC10A1 is associated with resistance to chronic hepatitis B. *Hepatology* **61**, 1251–1260 (2015).
48. Watashi, K. *et al.* Cyclosporin A and its analogs inhibit hepatitis B virus entry into cultured hepatocytes through targeting a membrane transporter, sodium taurocholate cotransporting polypeptide (NTCP). *Hepatology* **59**, 1726–1737 (2014).
49. Liu, C. *et al.* The p.Ser267Phe variant of sodium taurocholate cotransporting polypeptide (NTCP) supports HBV infection with a low efficiency. *Virology* **522**, 168–176 (2018).
50. Barker, D. The developmental origins of chronic adult disease. *Acta Pædiatrica* **93**, 26–33 (2007).
51. Calkins, K. & Devaskar, S. U. Fetal Origins of Adult Disease. *Curr. Probl. Pediatr. Adolesc. Health Care* **41**, 158–176 (2011).
52. Chen, J. *et al.* Dissecting maternal and fetal genetic effects underlying the associations between maternal phenotypes, birth outcomes, and adult phenotypes: A mendelian-randomization and haplotype-based genetic score analysis in 10,734 mother–infant pairs. *PLoS Med.* **17**, 1–28 (2020).
53. Tyrrell, J. *et al.* Genetic evidence for causal relationships between maternal obesity-related traits and birth weight. *JAMA - J. Am. Med. Assoc.* **315**, 1129–1140 (2016).
54. Chen, Y. *et al.* SOAPnuke: A MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *Gigascience* **7**, 1–6 (2018).
55. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–60 (2009).
56. Zhang, F. *et al.* Ancestry-agnostic estimation of DNA sample contamination from sequence reads. *Genome Res.* **30**, 185–194 (2020).
57. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–93 (2011).
58. Dodds, K. G. *et al.* Exclusion and genomic relatedness methods for assignment of parentage using genotyping-by-sequencing data. *G3 Genes, Genomes, Genet.* **9**, 3239–

3247 (2019).
59. Browning, B. L. & Yu, Z. Simultaneous Genotype Calling and Haplotype Phasing Improves Genotype Accuracy and Reduces False-Positive Associations for Genome-wide Association Studies. *Am. J. Hum. Genet.* **85**, 847–861 (2009).
60. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
61. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
62. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
63. Chang, C. C. *et al.* Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* **4**, 1–16 (2015).
64. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet. 2006 388* **38**, 904–909 (2006).
65. Zhao, H. *et al.* CrossMap: A versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006–1007 (2014).
66. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

Acknowledgements

This study was supported by the National Natural Science Foundation of China (81673181, 82173525, 31900487, 82003471), the Ministry of Science and Technology of the People's Republic of China (2016YFC1000205, 2016YFC1000304-1), the Department of Science and Technology of Guangdong Province (2020A1515110859, 2020B1111170001, 2019B030316014, 2019B020227001, 2019B030301004) and the Guangzhou Municipal Science and Technology Bureau Basic Research Foundation (202102010254). We are grateful to all the participants in this project, and the whole BIGCS team. We thank Dr. Ge Zhang from the Division of Human Genetics, Cincinnati Children's Hospital Medical Center for useful discussions on the methodology of mendelian randomization analysis in mother-infant pairs. Particularly, we would like to thank the professional technical support service provide by Liang Wei, Shunping Liu and Wenzhi Lai from Guangzhou Aixunpan technology Co., Ltd on setting up the GDBIG website. We thank the Tianhe-2 Supercomputer Center in Guangzhou for support of computational and storage resources.

Author Contributions

Conceptualization, X. Qiu, S. Huang and S. Liu; Sample collection & Data curation, Y. Kuang, J. Lu, X. Xia; Investigation, X. Qiu, S. Huang, M. Huang and C. Wang; Methodology, S. Huang and S. Liu; Formal analysis, S. Huang, M. Huang and C. Wang; Visualization, S. Huang, M. Huang and C. Wang; Software, S. Huang, M. Huang and C. Wang; Validation, S. Liu, M. Huang, S. Huang, J. Lu, X. Xia, Y. Kuang and Y. Gu; Writing-original draft, S. Huang and S. Liu; Writing-review & editing, X. Qiu, S. Liu, S. Huang, J. He, and S. Lin; Project administration, X. Qiu and S. Huang; Supervision, X. Qiu and H. Xia; Resource, X. Qiu and H. Xia.

Competing interest

The authors declare no competing interest.

Figures and Tables

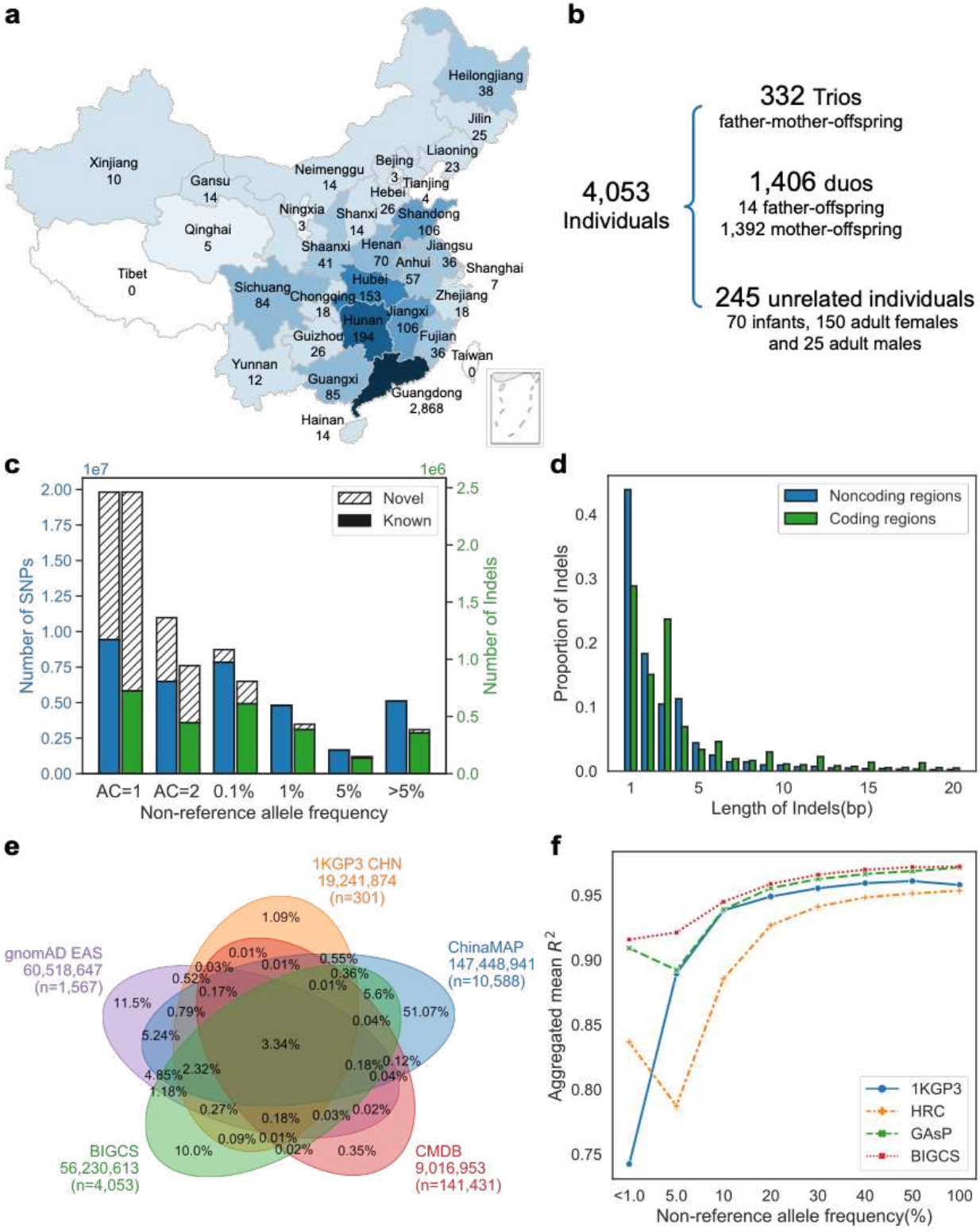


Fig. 1 | The geography distribution of samples and patterns of genetic variants of BIGCS. (a) Geographical distribution and statistics of 4,053 individuals in the BIGCS resource. Samples were

assigned to provinces based on their identity records that reflect their birth place. (b) The composition of samples in the study. (c) The number and allele frequency spectrum of known and novel variants: SNPs (left axis, in blue) and Indels (right axis, in green). The rectangle with slash represents novel variants, while the solid one represents known variants. (d) The length distribution and proportion of Indels in noncoding (blue) and coding regions (green). (e) The Venn plot of the variants (SNPs and Indels) identified in BIGCS compared with those in CMDB, ChinaMAP, gnomAD EAS and 1KGP3 CHN data resources. Each colored oval represents one resource, alongside with the name, variants number and sample size of the resource. The denominator of the percentage number in the oval is the number of total variants in the combined five datasets (N=197,403,927). (f) Imputation aggregated mean R^2 between the imputed genotypes by each of the 1KGP3, HRC, GAsP and BIGCS reference panels and the true genotypes in the 50 high-coverage WGS samples in different non-reference allele frequency bins defined by the corresponding reference panel. Each colored line represents one reference panel.

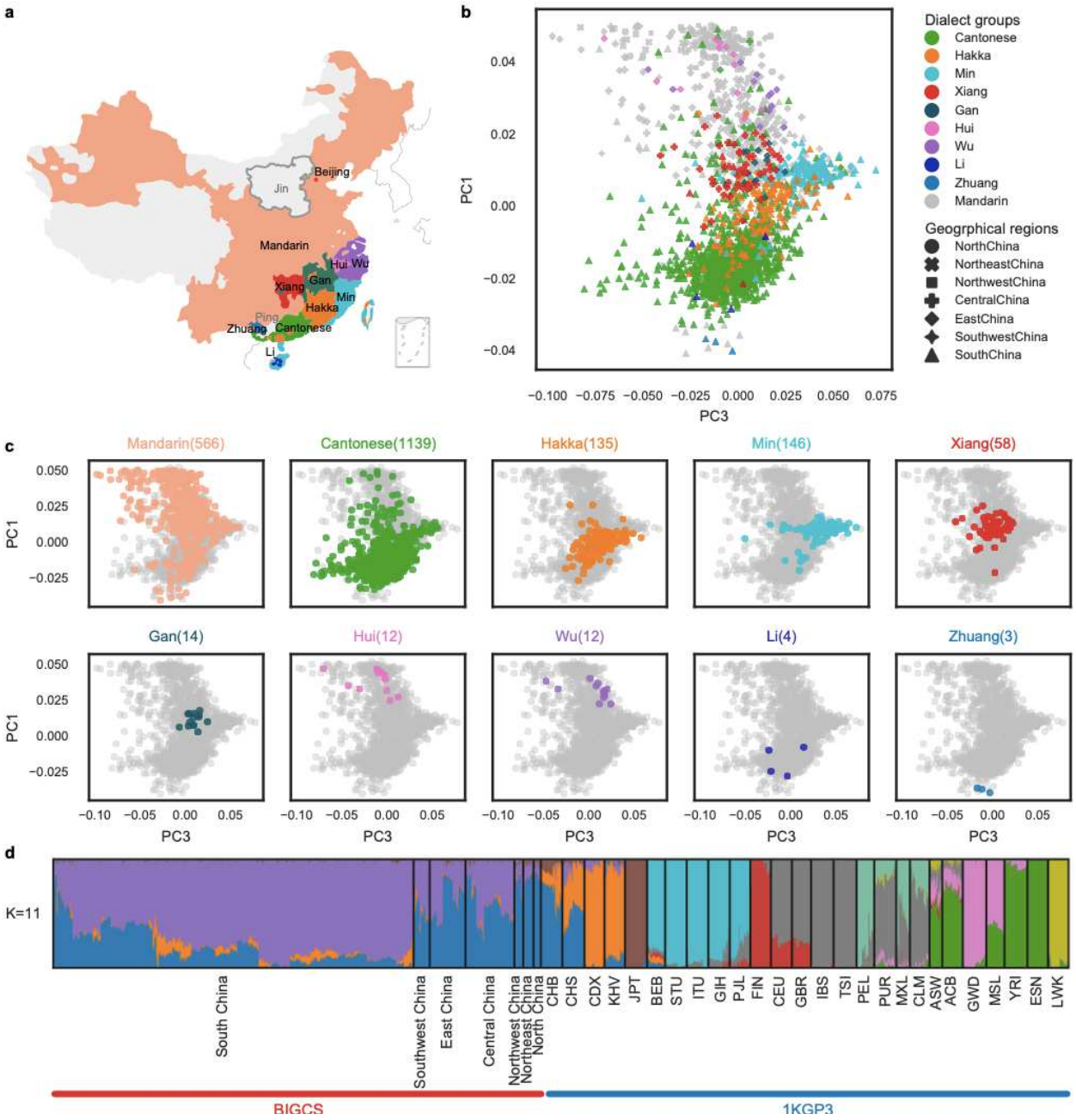


Fig. 2 | PCA and ADMIXTURE analysis of the BIGCS populations. (a) A map of China of the eight major Chinese dialect groups and the two minority dialects, Li and Zhuang Chinese in BIGCS study according to the "*Language Atlas of China*". (b) PCA of the BIGCS colored by the dialects they speak and shaped by the geographical regions that they were born. Each point represents one participant and is placed according to their PC values. (c) The same PCA colored by a specific dialect and the rest individuals are included but marked as gray in each subplot. Title is the name of dialects with sample size in parenthesis on each subplot. (d) ADMIXTURE analysis of all the 2,245 unrelated BIGCS samples and the full 2504 1KGP3 samples for K=11. Each small bar represents one individual, and the proportion of color represents the proportion of specific

ancestral components. BIGCS samples have been divided into seven geographical groups (North, Northeast, Northwest, Central, East, Southwest and South) of China according to their birth place.

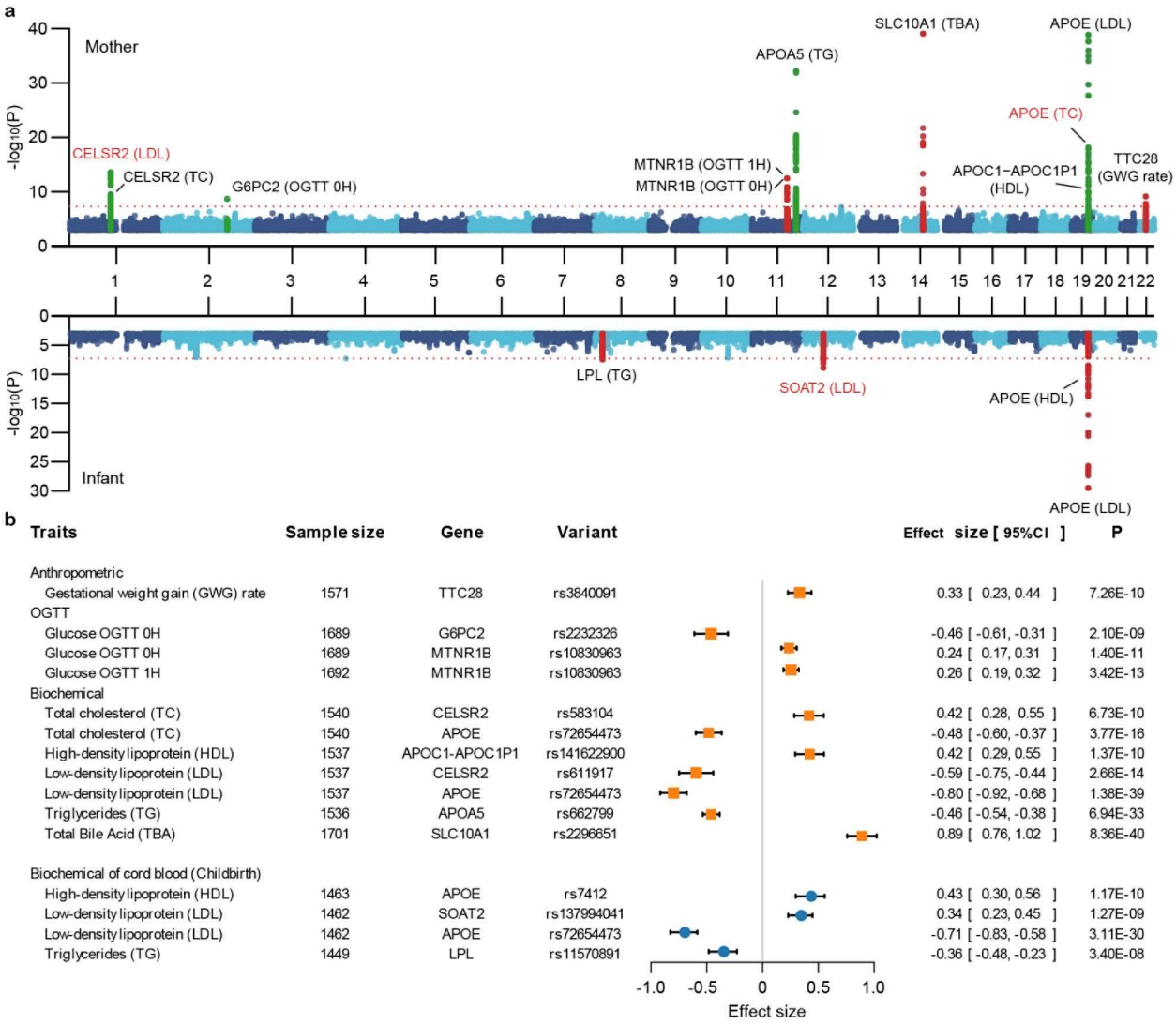


Fig. 3 | Overview of the 15 significant GWAS findings of 12 maternal quantitative traits (upside) and the 6 infant traits (downside). (a) Manhattan plot of all the variants with *P* value less than 10^{-3}. The x-axis is the chromosomal position and y-axis represents the minus logarithmic transformation of the *P* value from the GWAS regression model. The red dash horizontal line represents the *P* value of genome-wide significance level at $P=5.0\times10^{-8}$. Each significant locus is annotated by the gene ID with the corresponding trait in parenthesis. The vertical dots in red and green indicate novel and known GWAS loci for specific traits in the plot, respectively. The three gene symbols in red indicate distinct maternal and infant genetic effect on the same trait. (b) The forest plot for the 15 independent loci reaching genome-wide significance level of $P < 5\times10^{-8}$. Orange boxplots represented effect size (β value) of variant-traits association of mothers while the blue ones represented the effect size of corresponding traits in the cord blood of the infants. Data statistics

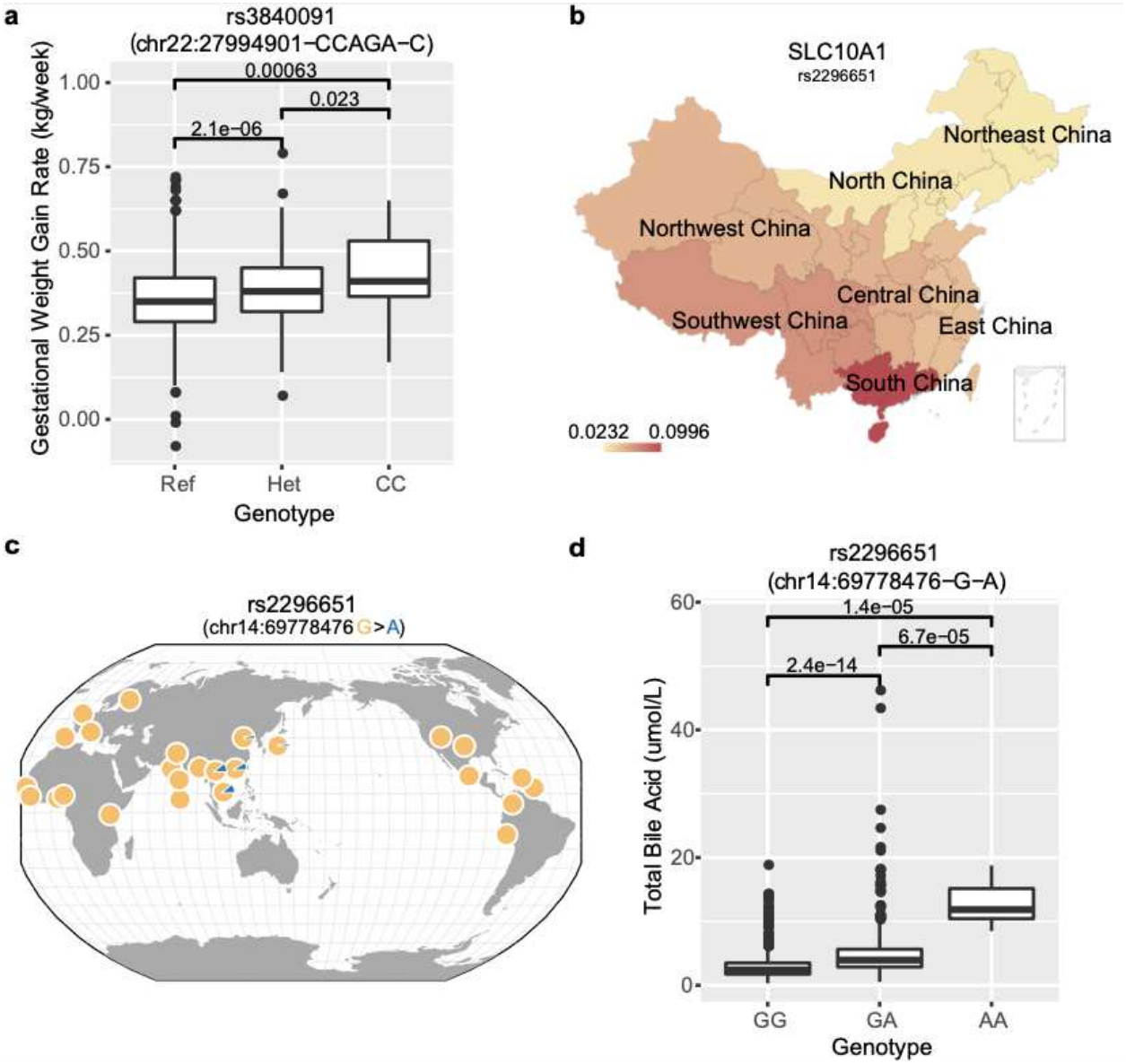


Fig. 4 | Two genetic associations with total bile acid and gestational weight gain first identified in the study. Comparison of gestational weight gain (GWG) rate (a) and total bile acid (d) according to the three genotypes of the lead SNP rs3840091 and rs2296651. P value of linear regression is indicated between any two of the genotype categories. The non-reference allele frequencies of rs2296651 (p.Ser267Phe) in China and in the world are visualized using data from BIGCS by PyEchart (b) and from 1KPG data by Chicago allele frequency website (c), respectively.

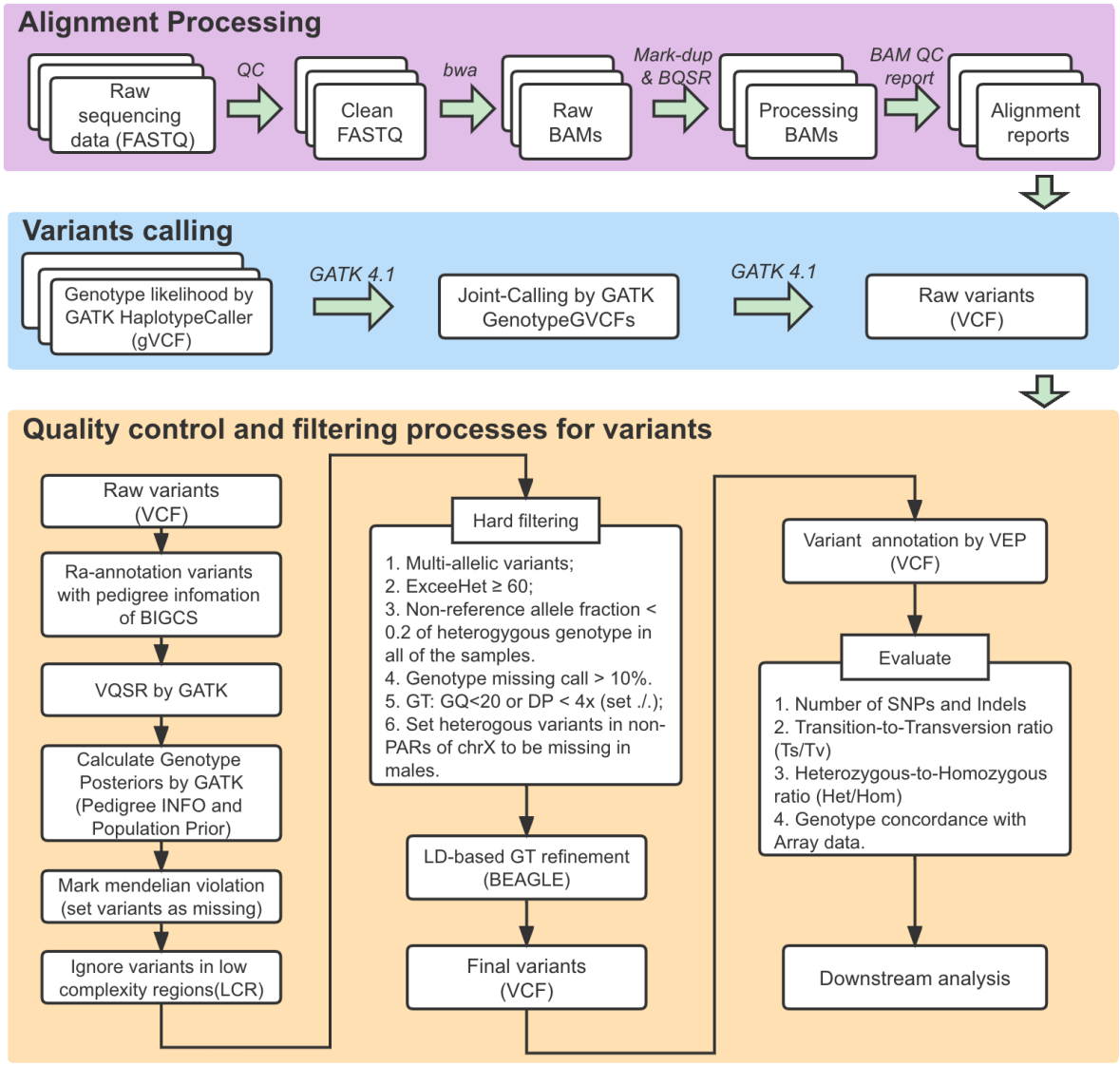
Table 1 | Intergenerational mendelian randomization analysis for the maternal phenotypes and the fetal growth measurements.

Maternal Traits	Birth Weight (g, n=1,655)			Birth Length (cm, n=1,654)			Gestational Duration (week, n=1,655)		
	Beta	SE	*P* value	Beta	SE	*P* value	Beta	SE	*P* value
Height (cm)									
[1] Maternal height	14.65	1.76	1.4E-16*	0.045	0.008	2.6E-09*	0.017	0.007	0.010*
[2] Maternal transmitted allele (h1)	28.53	8.53	8.5E-04*	0.045	0.037	2.2E-01	0.038	0.031	0.228
[2] Maternal non-transmitted allele (h2)	26.34	8.54	2.1E-03*	0.058	0.037	1.1E-01	0.030	0.031	0.336
[2] Paternal transmitted allele (h3)	23.47	8.55	6.1E-03*	0.114	0.036	1.7E-03*	-0.033	0.031	0.297
[3] Maternal effect	14.82	7.39	4.5E-02*	-0.009	0.032	7.7E-01	0.051	0.027	0.063
[3] Fetal genetic effect	11.63	7.33	1.13E-01	0.046	0.031	1.4E-01	-0.012	0.027	0.647
BMI (kg/m^2)									
Maternal BMI	29.30	3.08	5.7E-21*	0.085	0.013	2.1E-10*	0.015	0.012	0.197
Maternal transmitted allele (h1)	11.63	8.54	1.7E-01	0.037	0.036	3.1E-01	-0.009	0.031	0.767
Maternal non-transmitted allele (h2)	13.04	8.56	1.3E-01	0.003	0.037	9.3E-01	0.034	0.031	0.285
Paternal transmitted allele (h3)	7.41	8.56	3.9E-01	0.012	0.037	7.5E-01	-0.030	0.031	0.333
Maternal effect	8.44	7.39	2.5E-01	0.014	0.032	6.6E-01	0.027	0.027	0.311
Fetal genetic effect	2.92	7.47	7.0E-01	0.023	0.032	4.7E-01	-0.037	0.027	0.174
[4] **BP (mmHg)**									
Maternal BP	-0.08	0.32	8.1E-01	0.000	0.001	9.7E-01	-0.004	0.001	0.001*
Maternal transmitted allele (h1)	-20.97	8.14	1.0E-02*	-0.076	0.036	3.3E-02*	-0.004	0.031	0.898
Maternal non-transmitted allele (h2)	11.66	8.14	1.5E-01	0.052	0.036	1.4E-01	-0.021	0.031	0.494
Paternal transmitted allele (h3)	0.38	8.21	9.6E-01	-0.056	0.036	1.2E-01	-0.002	0.031	0.956
Maternal effect	-4.49	6.95	5.2E-01	0.017	0.030	5.9E-01	-0.012	0.027	0.664
Fetal genetic effect	-16.62	7.15	2.0E-02*	-0.095	0.031	2.4E-03*	0.008	0.027	0.778
FPG (mmol/L)									
Maternal FPG	7.96	3.12	1.1E-02*	0.005	0.013	7.2E-01	0.012	0.011	0.307
Maternal transmitted allele (h1)	5.96	8.17	4.7E-01	0.026	0.036	4.7E-01	-0.015	0.031	0.643
Maternal non-transmitted allele (h2)	10.32	8.18	2.1E-01	-0.048	0.036	1.8E-01	0.055	0.031	0.077
Paternal transmitted allele (h3)	-14.32	8.15	7.9E-02	-0.012	0.036	7.4E-01	-0.034	0.031	0.275

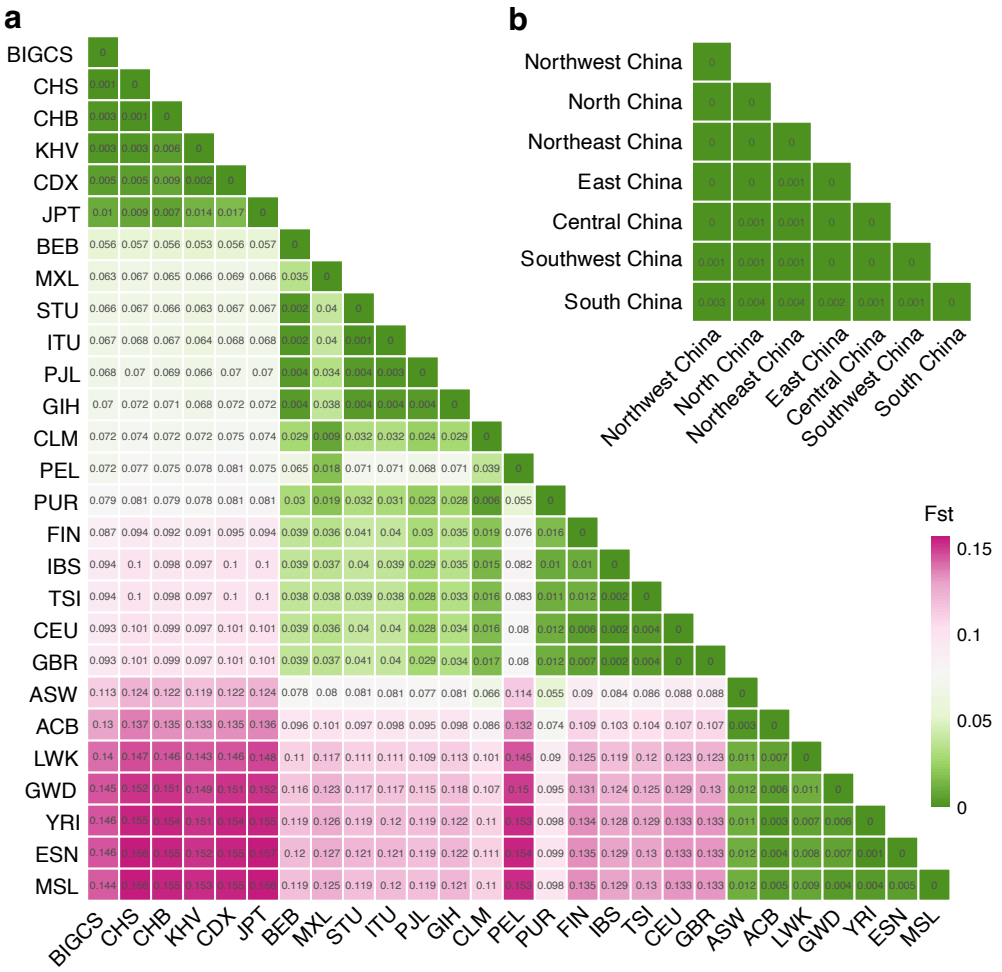
Maternal effect	15.75	7.16	2.8E-02*	-0.003	0.031	9.3E-01	0.037	0.027	0.178
Fetal genetic effect	-8.99	6.97	2.0E-01	0.032	0.031	2.9E-01	-0.051	0.027	0.055
TBA (umol/L)									
Maternal TBA	3.53	2.47	1.5E-01	0.006	0.011	5.6E-01	0.002	0.009	0.804
Maternal transmitted allele (h1)	-6.89	8.15	4.0E-01	-0.061	0.036	8.6E-02	-0.035	0.031	0.268
Maternal non-transmitted allele (h2)	-9.90	8.18	2.3E-01	-0.081	0.036	2.3E-02*	0.018	0.031	0.560
Paternal transmitted allele (h3)	7.22	8.14	3.8E-01	0.046	0.036	2.0E-01	-0.012	0.031	0.695
Maternal effect	-12.26	7.14	8.6E-02	-0.096	0.031	2.1E-03*	-0.001	0.027	0.963
Fetal genetic effect	5.55	7.21	4.4E-01	0.036	0.031	2.5E-01	-0.034	0.028	0.223

(1) The association analysis between the maternal phenotypes and birth outcomes.
(2) The effects of haplotype-based PRS for maternal transmitted, maternal non-transmitted and paternal transmitted alleles.
(3) The maternal effect and fetal genetic effect, which was modeled by linear combination of the effects of parental transmitted (h1 and h3) and maternal non-transmitted (h2) haplotypes.
(4) BP was calculated as the mean value of the systolic blood pressure and the diastolic blood pressure.
**P* value less than 0.05.

Extended data figures

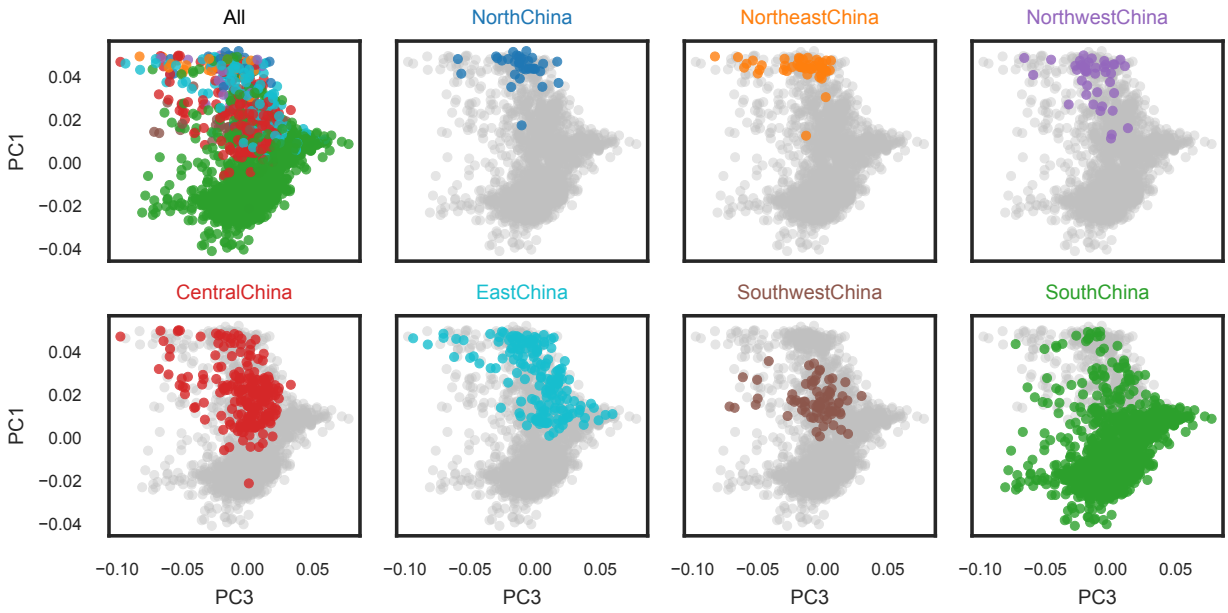


Extended data Fig.1 | Sequencing data alignment, variants calling, filtering and genotype refinement processes.
Related to Figure 1.



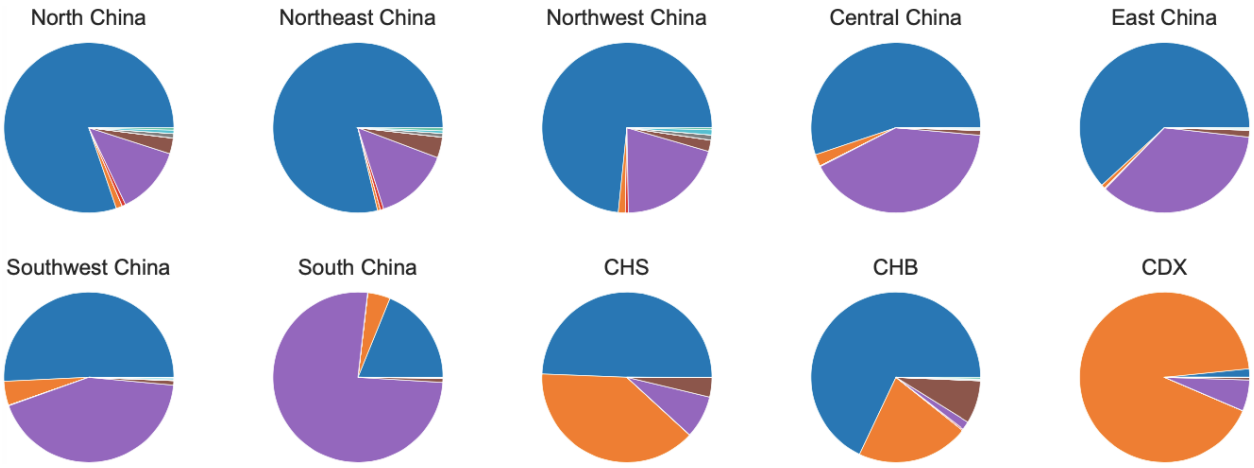
Extended data Fig.2 | Pairwise *F*st analysis.

(a) Pairwise *F*st analysis for BIGCS and the 26 populations in the 1KGP3 samples. (b) Pairwise *F*st analysis of BIGCS subpopulations defined by the seven geographical divisions in China. Related to Figure 2.



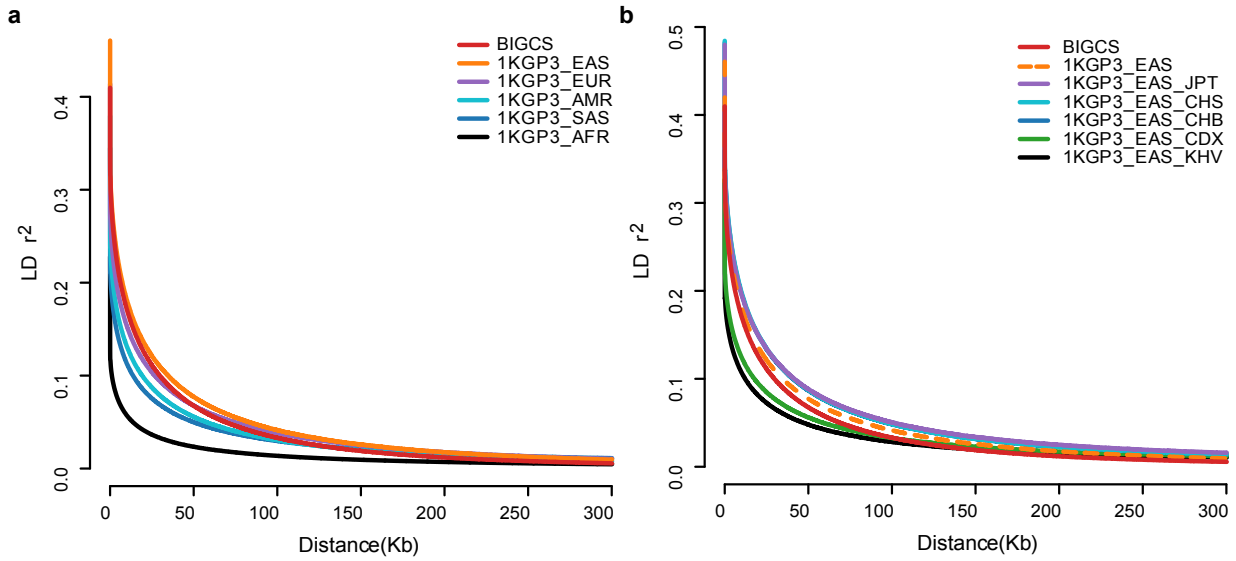
Extended data Fig.3 | Visualization of the individual geographical origin in the PC spaces

Colors represents different geographical region and the rest of the individuals were marked as gray in each subplot. Related to Figure 2.



Extended data Fig.4 | Geographic distribution of the ancestral components of ADMIXTURE analysis in Fig.2

Each pie chart represents the average ancestral proportions of a population in BIGCS and three Chinese population groups (CHB, CHS and CDX) in 1KGP3. Related to Figure 2.



Extended Fig.5 | The pattern of LD decay in BIGCS populations compared with other ancestral groups in 1KGP3.

(a) LD decay comparison between the BIGCS and the five continental populations in 1KGP3. (b) LD decay comparison between the BIGCS and the EAS populations in 1KGP3. Each colored line represents one population group and the BIGCS represents the LD decay value computed from all the unrelated participants. Related to Figure 2.

Extended data Table 1 | The summary statistics of variants discovered in the 4,053 individuals of BIGCS.

Variants type	AC=1	AC=2	MAF≤0.1%	0.1%<MAF≤1%	1%<MAF≤5%	MAF>5%	Total
Total variants (SNPs and Indels)	22,276,474	11,919,262	9,648,220	5,279,786	1,907,131	5,199,740	56,230,613
Known variants	10,155,543	6,931,714	8,551,334	5,218,588	1,892,992	5,172,409	37,922,580
Novel variants*	12,120,931	4,987,548	1,096,886	61,198	14,139	27,331	18,308,033
SNPs	19,814,479	10,975,400	8,829,910	4,839,893	1,749,629	4,843,145	51,052,456
Ts/Tv (Autosomal)	1.99	2.09	2.21	2.16	2.15	2.18	2.09
Het/Hom (Autosomal)	-	-	-	-	-	-	1.46
Known SNPs	9,432,687	6,487,985	7,930,364	4,827,755	1,749,356	4,842,499	35,270,646
Novel SNPs*	10,381,792	4,487,415	899,546	12,138	273	646	15,781,810
Indels (Insertions and Deletions)	2,461,995	943,862	818,310	439,893	157,502	356,595	5,178,157
Known Indels	722,856	443,729	620,970	390,833	143,636	329,910	2,651,934
Novel Indels*	1,739,139	500,133	197,340	49,060	13,866	26,685	2,526,223
Insertions	700,178	270,876	237,912	137,660	55,299	141,068	1,542,993
Known Insertion	173,805	110,741	162,461	111,025	46,470	128,634	733,136
Novel Insertion*	526,373	160,135	75,451	26,635	8,829	12,434	809,857
Deletions	1,761,817	672,986	580,398	302,233	102,203	215,527	3,635,164
Known Deletion	549,051	332,988	458,509	279,808	97,166	201,276	1,918,798
Novel Deletion*	1,212,766	339,998	121,889	22,425	5,037	14,251	1,716,366
Variants Location							
Exon(protein-coding region)	284,818	154,686	111,704	54,002	15,357	32,722	653,289
Intron	11,831,432	6,342,069	5,108,542	2,785,876	990,029	2,659,269	29,717,217
Splice-site	45,168	23,608	17,272	8,522	2,773	6,816	104,159
UTR	374,429	210,279	166,708	90,089	29,620	70,796	941,921
Upstream	980,243	513,867	419,617	230,654	85,249	233,302	2,462,932
Downsteam	801,760	423,750	347,360	191,070	70,611	193,169	2,027,720
Intergenic	6,595,366	3,475,822	2,847,084	1,573,558	587,315	1,650,000	16,729,145
Annotation function by VEP							
Synonymous	103,377	52,301	39,370	20,442	6,837	16,385	238,712
Missense	156,572	93,223	66,250	31,203	8,023	15,492	370,763
Stoplost	510	201	147	87	23	63	1,031
Stopgain	4,826	2,559	1,365	579	117	190	9,636
Startlost	646	310	221	96	27	60	1,360
frameshift insertion	3,238	1,109	592	234	66	121	5,360
frameshift deletion	8,191	2,391	1,328	567	107	155	12,739
inframeshift insertion	1,226	448	356	116	34	67	2,247
inframeshift deletion	4,679	1,676	1,514	593	101	148	8,711
Deleterious variants							
SIFT:deleterious	70,601	42,263	28,307	12,393	2,619	3,491	159,674
Polyphen2:probably damaging	38,391	22,914	14,439	6,050	1,086	1,163	84,043
Polyphen2:possibly damaging	25,712	15,540	10,694	4,788	991	1,416	59,141
ClinVar:Pathogenic	448	260	198	59	8	18	991
ClinVar:Likely pathogenic	130	98	50	13	2	3	296

*Not present in dbSNP build 154.

Extended data Table 2 | Summary of the GWAS discoveries for the 12 adult traits and 6 infant traits.

Group	Triats (Phenotype)	Sample size	No. of variants	Lambda value	GWAS significant loci (*P* value ≤5e-8)		
					Total	Known	Novel
Adult/Mother	**Anthropometric**						
	Height (prepregnancy)	2196	7,143,653	1.005	0	0	0
	Weight (prepregnancy)	2166	7,140,326	0.999	0	0	0
	BMI (prepregnancy)	2172	7,140,326	1.001	0	0	0
	Gestational weight gain (GWG) rate	1571	7,065,536	0.994	1	0	1
	OGTT glucose (Collected in 24-28 weeks of pregnancy)						
	OGTT 0H (Fasting)	1689	7,076,809	1	2	2	0
	OGTT 1H	1692	7,077,011	0.997	1	1	0
	OGTT 2H	1690	7,076,710	0.922	0	0	0
	Biochemical (Collected in 16-20 weeks of pregnancy)						
	Total cholesterol (TC)	1540	7,063,930	0.997	2	2	0
	High-density lipoprotein (HDL)	1537	7,063,589	0.999	1	1	0
	Low-density lipoprotein (LDL)	1537	7,063,422	0.995	2	2	0
	Triglycerides (TG)	1536	7,063,063	0.994	1	1	0
	Total Bile Acid (TBA)	1701	7,078,159	0.984	1	0	1
	Fetal growth (Collected at birth)						
	Birth length	1606	7,071,240	0.975	0	0	0
	Birth weight	1619	7,071,977	0.999	0	0	0
	Total				11	9	2
Infant	**Biochemical of cord blood (Collected at birth)**						
	Total cholesterol (TC)	1464	7,048,352	1.001	0	0	0
	High-density lipoprotein (HDL)	1463	7,048,037	0.999	1	1	0
	Low-density lipoprotein (LDL)	1462	7,048,193	1	2	1	1
	Triglycerides (TG)	1449	7,049,658	0.897	1	1	0
	Fetal growth (Collected at birth)						
	Birth length	1733	7,075,756	0.911	0	0	0
	Birth weight	1746	7,080,318	0.89	0	0	0
	Total				4	3	1

Extended data Table 3 | Linear correlation and regression of the maternal phenotypes on the maternal PRS[#]

Trait		Sample size	Beta	SE	CI2.5%	CI97.5%	*P* value	R^2
Height (cm)	Maternal genotype (h1+h2)	1655	1.68	0.11	1.46	1.89	6.92E-51	0.163
	Maternal transmitted allele (h1)	1655	1.14	0.11	0.92	1.36	1.27E-23	0.097
	Maternal nontransmitted allele (h2)	1655	1.25	0.11	1.04	1.47	1.45E-28	0.109
BMI (kg/m^2)	Maternal genotype (h1+h2)	1655	0.49	0.07	0.36	0.62	9.06E-14	0.085
	Maternal transmitted allele (h1)	1655	0.28	0.07	0.15	0.41	2.41E-05	0.063
	Maternal nontransmitted allele (h2)	1655	0.42	0.07	0.29	0.55	2.61E-10	0.076
BP (mmHg)	Maternal genotype (h1+h2)	1517	0.68	0.18	0.33	1.03	1.32E-04	0.073
	Maternal transmitted allele (h1)	1517	0.49	0.18	0.14	0.84	5.74E-03	0.069
	Maternal nontransmitted allele (h2)	1517	0.48	0.18	0.13	0.83	7.74E-03	0.069
FPG (mmol/L)	Maternal genotype (h1+h2)	1584	0.08	0.01	0.06	0.10	6.70E-14	0.076
	Maternal transmitted allele (h1)	1584	0.05	0.01	0.03	0.07	7.11E-07	0.057
	Maternal nontransmitted allele (h2)	1584	0.06	0.01	0.04	0.08	3.20E-08	0.061
TBA (umol/L)	Maternal genotype (h1+h2)	1617	0.49	0.07	0.35	0.63	2.22E-11	0.025
	Maternal transmitted allele (h1)	1617	0.32	0.07	0.18	0.47	1.01E-05	0.009
	Maternal nontransmitted allele (h2)	1617	0.32	0.07	0.17	0.46	1.73E-05	0.009
TG (mmol/L)	Maternal genotype (h1+h2)	1457	0.20	0.02	0.17	0.24	4.69E-30	0.124
	Maternal transmitted allele (h1)	1457	0.13	0.02	0.10	0.17	2.51E-13	0.077
	Maternal nontransmitted allele (h2)	1457	0.17	0.02	0.13	0.20	2.08E-20	0.097
TC (mmol/L)	Maternal genotype (h1+h2)	1457	0.21	0.02	0.17	0.26	5.65E-18	0.058
	Maternal transmitted allele (h1)	1457	0.13	0.02	0.09	0.18	8.18E-08	0.028
	Maternal nontransmitted allele (h2)	1457	0.18	0.02	0.13	0.23	2.08E-12	0.041

[#]R^2 indicates instrumental strength in the mendelian randomization analysis. Beta indicates the changes of the trait per unit changes of the PRS.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- BIGCSSupplementaryinformation.pdf
- TableS1.xlsx
- TableS2.xlsx
- TableS3.xlsx
- TableS4.xlsx
- TableS5.xlsx
- TableS6.xlsx
- TableS7.xlsx
- TableS8.xlsx
- TableS9.xlsx
- TableS10.xlsx
- TableS11.xlsx
- TableS12.xlsx
- TableS13.xlsx
- ExtendeddataTable1.docx
- ExtendeddataTable2.docx
- ExtendeddataTable3.docx
- ExtendedDataFig.1.pdf
- ExtendedDataFig.2.pdf
- ExtendedDataFig.3.pdf
- ExtendedDataFig.4.pdf
- ExtendedDataFig.5.pdf