Genomic Analyses from Non-invasive Prenatal Testing Reveal Genetic Associations, Patterns of Viral Infections, and Chinese Population History

Siyang Liu[1,2*], Shujia Huang[1*], Fang Chen[1*], Lijian Zhao[1*], Yuying Yuan[1*], Stephen Starko Francis[3,4], Lin Fang[1], Zilong Li[5], Long Lin[5], Rong Liu[1], Yong Zhang[1], Huixin Xu[1], Shengkang Li[1], Yuwen Zhou[1,5], Robert W. Davies[6], Qiang Liu[1], Robin G. Walters[7], Kuang Lin[7], Jia Ju[1], Thorfinn Korneliussen[8], Melinda A. Yang[9], Qiaomei Fu[9], Jun Wang[1], Lijun Zhou[1], Anders Krogh[2], Hongyun Zhang[1], Wei Wang[1], Zhengming Chen[7], Zhiming Cai[10,11,12], Ye Yin[1], Huanming Yang[1,13], Mao Mao[1], Jay Shendure[14,15], Jian Wang[1,13#], Anders Albrechtsen[2#], Xin Jin[1,16,17#,] Rasmus Nielsen[8,18,19, 20#,] Xun Xu[1#]

1. BGI-Shenzhen, Shenzhen 518083, Guangdong, China

2. Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen 2200, Denmark

3. Division of Epidemiology, University of Nevada, Reno NV, USA

4. Department of Epidemiology and Biostatistics, University of California, San Francisco CA, USA

5. BGI Education Center, University of Chinese Academy of Sciences, Shenzhen 518083, Guangdong, China

6. Genetics and Genomic Biology and The Centre for Applied Genomics, Hospital for Sick Children, Toronto, Canada

7. Clinical Trial Service Unit and Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford, Oxford, England

8. Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Copenhagen 1350, Denmark

9. Key Laboratory of Vertebrate Evolution and Human Origins of Chinese Academy of Sciences, Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, Beijing 100044, China

10. Department of Urological Surgery, The First Affiliated Hospital of Shenzhen University (Shenzhen Second People's Hospital), Shenzhen 518035, Guangdong, China

11. Guangdong Key Laboratory of Systems Biology and Synthetic Biology for Urogenital Tumors, Shenzhen 518035, Guangdong, China

12. Shenzhen University Carson International Cancer Center, Shenzhen 518060, Guangdong, China

13. James D. Watson Institute of Genome Sciences, Hangzhou 310058, Zhejiang, China

14. Department of Genome Sciences, University of Washington, Seattle, WA, USA

15. Howard Hughes Medical Institute, Seattle WA, USA.

16. School of Medicine, South China University of Technology, Guangzhou 510006, Guangdong, China

17. School of Bioscience and Bioengineering, South China University of Technology, Guangzhou 510006, Guangdong, China

18. Department of Integrative Biology, University of California Berkeley, Berkeley, California 94720, USA

19. Department of Statistics, University of California Berkeley, Berkeley, California 942720, USA

20. Lead Contact

*. These authors contributed equally

#. Corresponding authors xuxun@genomics.cn, rasmus_nielsen@berkeley.edu, jinxin@genomics.cn, albrecht@binf.ku.dk, wangjian@genomics.cn

Summary

We analyze whole-genome sequencing data from 141,431 Chinese women generated for non-invasive pregnancy testing (NIPT). We use these data to characterize the population genetic structure of China, and to investigate genetic associations with maternal and infectious traits. We show that the present day distribution of alleles is a function of both ancient migration and very recent population movements. We reveal novel phenotype-genotype associations, including several replicated associations with height and BMI, an association between maternal age and a variant near *EMB*, and between twin pregnancy and *NRG1*. Finally, we identify a unique pattern of circulating viral DNA in plasma with high prevalence of hepatitis B and other clinically relevant maternal infections. A GWAS for viral infections identifies an exceptionally strong association between integrated herpesvirus 6 and *MOV10L1*, which affects piRNA processing and PIWI protein function. These findings demonstrate the great value and potential of accumulating NIPT data for worldwide medical and population genetic analyses.

Introduction

Sufficient large sample size is of fundamental importance in resolving biological questions in population and medical genetics. Given a fixed budget, sample size tends to play a more essential role compared to sequencing depth (Li et al., 2011). Previous studies have demonstrated that sequencing many individuals at a low depth generally provides a better representation of population genetic variation compared to sequencing a more limited number of individuals at a higher depth (Fumagalli, 2013). Furthermore, when using proper imputation techniques, even sequencing at an average depth of < 0.1x in a large enough cohort can be a cost-effective strategy for detecting genetic associations for complex traits (Pasaniuc et al., 2012).

Several large-scale national and international sequencing projects have been carried out in the past decade with sample sizes limited to tens of thousands (Auton et al., 2015; Francioli et al., 2014; Gudbjartsson et al., 2015; Maretty et al., 2017; Walter et al., 2015). Increasing the sample sizes of these studies is a major financial and logistical challenge. However, Non-invasive prenatal testing (NIPT) for fetal trisomy - by sequencing of maternal plasma cell-free DNA (cfDNA) (Zhang et al., 2015a) - has become the fastest adopted molecular test in history and provide an untapped resource for understanding population genetic variation and genetic associations. To date, over ten millions of NIPT tests have been carried out globally, among which 70% were conducted on Chinese women. These samples can be leveraged for population genetic investigations of population history, large-scale genetic association studies, and viral screening if the technical issues regarding the use of very large, very low depth (0.06x - 0.1x) samples

can be addressed.

Here, we analyze NIPT sequencing data of 141,431 pregnant women with informed consent. We demonstrate that allele frequencies can be estimated with high accuracy, allowing further population genetic analyses. We also show that efficient genotype imputation is feasible and can provide considerable mapping power. We use the data to carry out the hitherto largest analysis of population genetic variation in the Chinese population, perform a genome-wide association study (GWAS) on multiple traits in pregnant Chinese women, and survey the distribution of circulating viral DNA in the maternal plasma.

Results

Study participants and chromosomal coverage

The 141,431 participants were recruited from 31 out of the 34 administrative divisions in China (Figure S1A). Each individual was sequenced using 5-10 million single-end reads (35-49bp), corresponding to a sequencing depth of 0.06x to 0.1x per individual (Figure S1B). The reads were aligned to the hg19 reference using bwa[10] (see STAR Methods), with a resulting combined read depth distribution that is approximately Poisson and closely follows that expected from the ENCODE mappability track (Figures S1C and S1D). Based on read length and observed depth distribution we identified regions of the genome accessible to high-confidence mapping in the NIPT data, resulting in a total length of around 2.13 billion base pair accessible genome (75% of the non-N human

reference genome sequence). We also re-sequenced DNA derived from white blood cells of 40 participants to a mean depth of 15x. These data were used in the variant calling and genotype imputation evaluation.

Amount of genetic variations and accuracy of genotype imputation

Previous standard methods for allele frequency estimation and joint SNP calling, such as those implemented in GATK (DePristo et al., 2011) and Samtools (Li, 2011) did not scale up to sample sizes over a hundred thousand. Therefore, we developed a new method for fast maximum likelihood estimation of allele frequencies and joint SNP calling using a likelihood ratio test (STAR Methods). Using this method for initial screening, we identified 32.5 million bi-allelic candidate SNPs (Table S1). After recalibration using a gaussian mixture model, we identified a final call set of 9.04 million single nucleotide variants with a Transition/Transversion ratio of 2.1 for known variants and 2.4 for novel variants, respectively (Figure S2A), consistent with those obtained for 1000 Genomes Project (1KG) variants (Auton et al., 2015). 81.7% of the variants were called in the 1KG Han Chinese individuals (Auton et al., 2015), ~16% of the variants were in the remainder of 1KG or dbSNP database while 233,966 (2.6%) were novel variants **(**Figure 1A). 90% of the variants were found in the union of the gnomAD East Asian (Lek et al., 2016) and 1KG Han Chinese call sets (Figure 1B). Using experimental validation, we estimated an upper bound for the false positive rate (FPR) of SNP calling of 0.2% (Figure S2B-E). However, among novel variants the FPR was approx. 0.32. These SNPs comprised a small proportion of the total number of SNPs, but did include common variants, likely due to unresolved mapping issues. The squared correlation coefficient (R^2) of the

frequency of the non-reference allele, i.e. alternative allele estimated in our study and that computed in the 1KG Han Chinese was 0.98 (Figure 1C).

We subsequently imputed genotypes of 8.9 million known variable DNA sites with allele frequency greater than 0.01 using the 1KG Han Chinese as a reference panel (Flint et al., 2016) (STAR Methods). To estimate imputation accuracy, we compared the squared correlation coefficient (R^2) between the genotypes called in the medium coverage whole genome sequencing data of the 40 individuals (MC set, 15x) and the imputed genotypes in the low-coverage data (LC set). 2.13 million of the variants were well imputed with an info score greater than 0.4, and with a p-value from a chi-square test of Hardy-Weinberg equilibrium larger than 10^{-6}. The mean imputation accuracy of those variants was 0.89 while it was 0.71 for all variants combined (Figure 1D). The imputation accuracy was negatively correlated to the fraction of fetal DNA present in the plasma but the effect was not very pronounced (Figure S2F).

Population structure, recent population history and genetic adaptations

Even though the Chinese is the world's largest population that comprises 1.4 billion people, it is perhaps surprisingly understudied with respect to population genetic history. We applied information of the 141, 431 pregnant women regarding digital geographic location and self-reported ethnicity to study of the genetic variation in China at multiple time scales. Because of the uncertainty in genotype calling, we conduct all population genetic analyses using methods that either sample a single read per individual or use maximum likelihood estimates of allele frequencies without relying on genotype imputation.

A principal component analysis of all the 141, 431 participants suggested that the first three principal components reflected sequencing read length, latitudinal genetic differentiation and the sequencing error rate (Figure S3A-D). After removing participants with 49bp read length and with sequencing error rate greater than 0.00325, a principal component analysis of 45,387 self-reported Han Chinese from the 31 administrative divisions showed that the greatest differentiation of Han Chinese is along a latitudinal gradient (Figure SE-F), consistent with previous studies (Chen et al., 2009; Xu et al., 2009). In contrast, there is, perhaps surprisingly, very little differentiation from East to West. This observation may be explained by the fact that a large proportion of the western Han populations in China are recent immigrants organized by the central government starting from 1949 when the People's Republic of China was founded (Liang and White, 1996). While the Han Chinese were found to be relatively genetically homogenous, there was greater divergence among the minority ethnic groups for both latitude and longitude (Figure 2A-B). The most differentiated ethnic groups are the Turkic speaking Uyghur and Kazakhs, who reside in the Xinjiang province, and the Mongols residing primarily in Inner Mongolia. The Xibe, Tibetans, and Hui from central China, the Yi from southwestern China, and the Zhuang and Buyi minorities from southern China, also differ substantially from the Han Chinese that come from the same area. On the other hand, the Manchu from northeastern China were genetically closest to the Han Chinese in that area, consistent with historical accounts (Rhoads, 2000).

We further explored the patterns of allele sharing between Han Chinese and major global ethnic groups using private alleles defined from the 1KG populations and using

outgroup F3 statistics (Peter, 2016) (STAR Methods). In the northwest and central west, we observed private allele sharing with the 1KG European CEU panel both for individuals self-identified as Han Chinese and for individuals self-identified as belonging to a minority group. The strongest level of private allele sharing with the CEU was observed for people in the most northwest provinces of Xinjiang and Gansu (Figure 2C), likely reflecting the Turkic speaking ancestry in these minorities. When only the Han Chinese were included, the strongest level of allele sharing with Europeans was observed for people in the Qinghai, Gansu and Ningxia provinces (Figure 2D). These provinces are located in the Hexi corridor, the most important commercial hub on the Silk Road connecting China to the west since the establishment of the Han Dynasty (206 B.C.) (Yang et al., 2008). Thus, one potential explanation for the Western ancestry observed in these provinces is gene flow related to their location on the Silk Road. We also observed a pattern of increased allele sharing with the 1KG Indian ITU reference panels in southwestern populations from Xinjiang, Tibet, Yunnan, Guangxi and Hainan provinces (Figure 2E-F), consistent with their geographic proximity to the Indian subcontinent(Yang et al., 2017). Analyses based on the F3 statistic are mostly consistent for the CEU analysis, but for the ITU analysis, we also show high affinity between the Han Chinese in northern provinces and the ITU, likely due to the shared ancestry of the CEU and ITU populations. Furthermore, we applied the F3 statistic to learn patterns of allele sharing between the Chinese provincial populations and 1KGP neighbour populations including three Chinese populations, the Japanese and Vietnamese. We observe a pattern of allele sharing among the thirty-three administrative divisions

198 reflecting the geographical origin of the 1KGP populations (Figure S3G-K). Interestingly,
199 we found that the CHB, although annotated as the Han Chinese from Beijing, did not
200 have the closest affinity with Beijing individuals but tended to be closer to populations in
201 the coastal provinces: Shandong, Zhejiang, Jiangsu, Fujian, and Jiangxi(Figure S3G).
202 This likely reflects the recent multiethnic migration into Beijing consistent with the
203 demographic information available for our samples. We also investigated the inter-
204 provincial allele sharing between Han Chinese in the Chinese administrative divisions.
205 The difference in f3 statistic among provinces is very small, but all southern provinces
206 show more genetic affinity with other southern coastal provinces, while northern
207 provinces associate with northern coastal provinces(Results not shown). This observation
208 likely reflects a combination of internal migration events organized by the central
209 government since 1949 (Liang and White, 1996) and the country's oriented movement of
210 labour from the interior to the coastal areas since 1979 (Liang and Ma, 2004).

211 We inferred selection within Han Chinese populations using two approaches.
212 First, we identified variants with significant differentiation along each PC compared to a
213 null distribution expected under a model of genetic drift (STAR Methods). Second, we
214 conducted a scan of the rarer but more important pathogenic variants in the Clinvar
215 database (Landrum et al., 2014) by statistically comparing allele frequency differences of
216 those loci among North, Central and South Han Chinese against a null distribution
217 generated from the genome-wide data. In the PC scan, we identified six loci showing
218 genome-wide significance across latitude: *LILRA3*, *CR1*, *FADS2*, *DOCK9*, *ABCC11*, and
219 a cluster of *IGH* genes (Figure 3A). The *CR1*, *DOCK9*, and the *IGH* genes display a

higher allele frequency in the south while the *FADS2*, *ABCC11* and *LILRA3* genes display a higher derived allele frequency in the north (Figure 3B-G, Table S1). Three of these loci are known to be related to immune responses (the *IGH* genes, *LILRA3* and *CR1*). *DOCK9* is associated with bipolar disease and has been previously shown to under selection in East Asians (Suo et al., 2012). *FADS2* is a well-known target of selection associated with changes in diet to, or from, a diet with a high content of animal fat, and has previously been inferred to have been targeted by selection in Inuit (Fumagalli et al., 2015), South Asians (Kothapalli et al., 2016), Europeans (Buckley et al., 2017), and in Africa (Mathias et al., 2012). Our results suggest more recent selection has also been acting within China. The *ABCC11* locus is famously associated with earwax type and has previously been shown to be under selection in Asian, Native American, and European populations (Ohashi et al., 2011), and our results demonstrate that this locus is also under differential selection within China. We also investigated the geographical distribution of possibly pathogenic variants compiled from the ClinVar dataset (Landrum et al., 2014) as candidates for loci under selection. We calculated measures of allele frequency differentiation (Fisher's exact test between northern, central and southern Han Chinese, comparing against a frequency-matched dataset of 100,000 SNPs chosen at random) (STAR Methods). We identified and reported the nine out of the 42,058 possibly pathogenic variants in eight genes that display the most significant allele frequency difference among the three geographical regions (Fisher exact test with p-value $< 10^{-6}$, percentile p-value < 5e-3, Figure 3H-O, Table S2). Those SNPs include rs72554665, a polymorphic site in *G6PD*, a gene associated with resistance to malaria (Nkhoma et al.,

2009). This variant has a higher frequency in southern China, consistent with historically higher incidence rates of malaria in southern China than in northern and central China.

Phenotype-genotype associations of multiple complex traits

In the following, we demonstrate that NIPT data can be used effectively in GWAS. We first investigated associations with two common traits, height and body mass index (BMI) among 61.7K individuals with both phenotypes recorded (Figure S4A-B). We applied a score test (Skotte et al., 2012) to test the association between the traits and the genotype probabilities for each of the previously mentioned approx. 2 million imputed variants, incorporating covariates such as the first to fifth principal components (Figure S3A), maternal age, gestational age of the fetus, fetus sex, etc (STAR Methods). The genomic control factor lambda for height and BMI were 1.51 and 1.32 respectively(QQ-plots in Figure S4C-D). Due to the high polygenicity of the traits we also evaluated inflation of our test statistics using LD-score regression which did not show severe inflation (intercept 1.03, s.e. 0.03 and 1.10, s.e. 0.02, attenuation ratio 0.05, s.e. 0.04 and 0.24, s.e. 0.05 for height and BMI respectively), suggesting that confounding factors, such as population structure, were generally well controlled (Table S4). The estimated SNP heritability obtained from the LD score regression for height and BMI are 0.48 and 0.10, respectively. A comparison of the LD score regression statistics between Giant, UK Biobank and our study can be found in Table S4. We note that strong inflation was observed if covariates were not applied in the test model (genomic control factor lambda for height and BMI are 9.71 and 2.68).

In total, 48 and 13 loci reached genome-wide significance for association with height and BMI, respectively, at the classical $5x10^{-8}$ genome-wide significance level (Figure 4A-B, Table 1). Forty-one of the height loci were previously reported, although only thirty-six of them have previously been found in Asian populations (MacArthur et al., 2017). Seven height loci located in or around the genes *UBQLN2*, *MIR325HG*, *MAST2*, *STRBP/ZBTB26*, *C11orf24-LRP5*, *ARHGEF12* and *LINC00261* have not been previously reported (Table 2, Figure S5A-F). There was one new signal in the intronic region of DNA2, a locus first identified in GIANT (Wood et al., 2014) and another independent signal in the nearby gene MYPN was reported (conditional p-value = 4.8e-8). Three BMI-associated loci in the genes PLD5, TRPC6 and CBLN4 were also not previously reported (Table 2, Figure S5H-J). We attempted to replicate the novel and known associations for height and BMI in 32,000 genotyped Chinese participants from the China Kadoorie Biobank (CKB) cohort (Chen et al., 2011) and in the results of the GIANT consortia (Yengo et al, BioRxiv) and the UK Biobank (Ben Neale's website, STAR Methods). When regressing the effect size of the genome-wide significant variants in the three test sets including CKB, Giant and UK Biobank on the discovery set, i.e. the NIPT result, we observed a higher regression slopes for the test set of the same Chinese ancestry compared to the test sets of the European ancestry for height (CKB, slope 0.88; Giant, slope 0.614; UK Biobank, slope 0.495) and BMI (CKB, slope 0.709; Giant, slope 0.463; UK Biobank, slope 0.434), respectively(Figure S4E-J). When comparing the beta direction and the p-value of the lead or proxy SNPs between the discovery and test sets, in almost all cases the p-values and the effect sizes of the SNPs are similar(Table 2, Table

S5, Table S6). Only one known and one novel locus for height (FADS2 and LINC00261) and one locus for BMI (TRPC6) failed to be replicated at the significance level when correcting for multiple testing ($p < 0.05$ divided by the number of test loci) in all the CKB and the GIANT/UK Biobank analyses. However, the variant in FADS2 is nominally significant in the CKB cohort with a p-value of 0.002.

The height and BMI results provide a proof of concept for the use of NIPT data in GWAS and suggest that NIPT data can be used to investigate fertility and pregnancy related traits that otherwise would be prohibitively difficult to investigate in such large samples. To illustrate this, we investigated associations for two novel traits: (1) maternal age, which is expected to correlate with female fertility broadly defined, but may also be affected by several other factors, and (2) twin pregnancy. The maternal age distribution follows a bimodal distribution in the NIPT participants (Figure S4K). Similarly to height and BMI, we again do not observe any severe inflation when including covariates for both maternal age (lambda=1.29) and twin pregnancy (lambda=1.06) (QQ plot see Figure S4L-M). Strong inflation was also observed for these two traits if covariates were not applied in the test model for maternal age (lambda=1.88) and twin pregnancy (lambda=1.36). We find one significant association peak for maternal age located between the *HCN1* and *EMB* loci (rs16828019, p = 1.38E-11, Figures 4, S5K). This signal is located near a previously identified association peak for age at first birth, originally reported to be in the HCN1 gene (Barban et al., 2016). Our lead SNP is closer in location to *EMB* gene which encodes embigin, a transmembrane glycoprotein that is preferentially expressed in the early stages of embryogenesis and enhances integrin-

mediated cell-substratum adhesion in mice (Huang et al., 1993). It shows particularly high expression during early post implantation embryogenesis and is therefore a strong candidate gene for further studies of infertility in humans. In European countries, maternal age is associated with education attainment, which is a proxy for intelligence (Barban et al., 2016). We don't have records for education attainment in our study. Whether education has an impact on maternal age in the Chinese population requires further investigation.

Out of 137,646 individuals with ultrasound scans, 476 had more than one fetus (e.g., twins). The lead associated SNP for this trait was located in the gene *NRG1* (rs12056727, p=5.93E-9, odds ratio=1.99, Figures 4, S5L) and is a very strong eQTL for expression in the thyroid (effect size = 0.5, p-value 1.2e-19) in the GTEx database (Carithers and Moore, 2015). The tested allele T increases the twinning probability and the NRG1 expression in the thyroid. Furthermore, the SNP is associated with hyperthyroidism in the UK BioBank (Sudlow et al., 2015) (p=1.7e-7, odds ratio = 1.15). Thyroid function has previously been associated with fertility. The *NRG1* gene has mostly been investigated for its effects on behavior (schizophrenia in humans and response to stress and anxiety in rodents), but at least one study notes that matings between knockouts in mice have smaller litter size (Britto et al., 2004). However, the exact reason for this is unknown. More interestingly, twin pregnancies tend to be associated with lower levels of the thyroid-stimulating hormone (TSH) (Soldin, 2006), consistent with the SNP association with hyperthyroidism, which generally involves increased thyroid hormone levels and decreased TSH levels.

Circulating viral DNA in maternal plasma

Despite its importance for public health, few studies have been carried out on the distribution of viral DNA in blood plasma (the virome) at the population level. However, the sequence technologies used in NIPT studies provide an untapped resource for understanding viral epidemiology. We investigated the plasma virome of 138, 882 participants by querying reads that do not align to the human genome against the NCBI viral sequence database (Sayers et al., 2009) using BLAST (Altschul et al., 1990) (STAR Methods). The plasma virome, cleaned for phages and low prevalence virus, is represented in (Figure 5A), and all observed viruses, without removing phage and low abundance participants, can be found in Figure S6A-B. We examined sequencing coverage to individual viruses to determine potential misclassification or contamination, where prevalent viruses with relatively even coverage support true virus identification (Figure S7). Most viruses detected have even sequence coverage, while a few display localized peaks. To understand if these peaks are related to human homology, we aligned viral reference sequences to the human reference genome, hg19 and non-EBV decoys (Table S6). We found that only the localized peak of HCV corresponds to a human homologous sequence, yet these sequences are only found in a small number of our subjects (N=3). The peaks in coverage may represent highly conserved areas of viruses, misclassification of human sequences, uneven production of viral DNA, select viral integration events or introduced viral DNA from vaccine, further investigation and validation is required.

Interestingly, the blood virome in a recent study of Europeans (Moustafa et al., 2017) appeared to have a different viral distribution compared to the pregnant Chinese women participants analyzed in this study (Figure 5A). The use of differing sequencing approaches is a challenge to direct comparisons, but our participants were significantly enriched for hepatitis B virus (HBV) and Parvovirus B19 DNA and showed a lower prevalence of human herpesvirus 7 (HHV-7) DNA compared to Europeans (Moustafa et al., 2017). The prevalence of HBV DNA across Chinese populations estimated in our study is approximately 2.5% (with high mean abundance 25.6), which is less than the 9.8% prevalence reported in a Chinese 2014 population survey of HBV antibodies (IgM to HBV core antigen or surface antigen(HBsAg+)) (Yan et al., 2014). These differences are likely due to varying estimates derived from circulating HBV DNA versus antibody prevalence, the enrichment of our sample for relatively affluent younger women, and the late adoption of the HBV vaccine in China (since 1992)(Yan et al., 2014). We detected 3,421 participants for HBV DNA, yet 1,911 participants self-reported some type of HBV infection. As expected, we have the greatest sensitivity (78.7%) to detect reported active HBV infection (HBsAg+). Of 1032 individuals reporting latent infection (HBsAg-), we detected HBV DNA in 53 subjects, where the lower sensitivity (5.1%) is likely due to low levels of circulating HBV DNA during latent infection (Figure S6C). Interestingly, we detected HBV DNA in 2,959 individuals who did not report any HBV infection, suggesting an additional and potentially clinically important use of NIPT where circulating HBV DNA is associated with fetal transmission(Zou et al., 2012).

We detected many hits for human endogenous retrovirus K (HERV-K) (prevalence ~2.1%, mean abundance 0.22) which was previously shown to be active, capable of expression in humans, and associated with HIV infection (Zwolińska et al., 2013). All humans carry multiple copies of HERV-K in their genomes and the recovery of non-aligning HERV-K in a subset of subjects may be the result of insertionally polymorphic HERV-K (Wildschutte et al., 2016), or the result of co-option of HERV-K sequences by other exogenous viruses. Herpesvirus 6A/B (HHV-6A/B), the third most common viral group, has a prevalence of 0.8% (mean abundance 0.48). HHV-6A/B were grouped together due to high co-occurrence and potential for misclassification due to sequence homology. A distinct bimodal distribution clusters high abundance HHV-6A/B (Abundance>$10^{-0.5}$, Figure 5A), likely separating chromosomally integrated HHV-6A/B (ciHHV-6A/B) from non-integrated circulating HHV-6A/B as noted in previous studies (Moustafa et al., 2017). Human herpesvirus 5 or cytomegalovirus (HHV-5 or CMV) is the fourth most common infection with a prevalence of 0.40% (mean abundance 0.03). This virus is of particular interest in pregnant women as CMV is one of the leading causes of birth defects(Cheeran et al., 2009). Parvovirus B19 has a prevalence of 0.39% (mean abundance 1.68) and is of clinical relevance to pregnant women as active infection can cause fetal anemia and death.

To identify germline polymorphisms associated with viral infections we carried out an association study for each of the major viruses, comparing infected individuals to a control group of 90,531 participants who have no detectable virus in the NIPT sequencing data. We identified an intronic variant, rs73185306, in the *MLC1-MOV10L1*

region that is significantly associated with the presence of high abundance ciHHV-6A/B (N=653) but not low abundance HHV-6A/B (N=1556) (Odds ratio = 3.4, p-value = 7.3e-66) (Figure 5B-C, Figure S6D). rs73185306 is an expression quantitative trait locus (eQTL) for both *MCL1* and *MOV10L1* genes suggesting a functional role (Lonsdale et al., 2013). To ensure that this strong association was not due to alignment error and homology in the *MLC1-MOV10L1* region, we aligned HHV-6A and HHV-6B genomes back to the human genome and found no sequence homology in this genomic region (Table S7). The *MCL1* gene is involved in myeloid cell differentiation and has been shown to be upregulated during herpesvirus infection (CMV, EBV and HHV-8). The *MOV10L1* gene is known to be associated with platelet distribution (Astle et al., 2016) which is also correlated with severity of Hepatitis B infection (Karagoz et al., 2014), though we found no association with circulating HBV DNA. Intriguingly, MOV10L1 is a PIWI interacting RNA helicase that is active during spermatogenesis and functions as a repressor of retrotransposons (Vourekas et al., 2015). We suspect that the PIWI-interacting RNA represses HHV-6A/B integration and polymorphisms in this gene allow for more efficient integration of HHV-6A/B during spermatogenesis. We observed no other significant genome wide SNP associations with other viruses.

Finally, we explored the geographic distribution of detected viruses in the studied population throughout China. We mapped the prevalence of viral sequences among 30 administrative divisions with more than 100 participants. Tibet was excluded due to a small sample size of 13 individuals. We observed different geographic patterns for many viruses (Figure 5D-F, Figure S6E-J). We observed that HBV DNA in serum has a higher

prevalence in southern China compared to central and northern China (Figure 5D), while previous studies have shown higher HBV antibody prevalence in northern China (Yan et al., 2014). We speculate that subgenotype resolution differences in HBV may contribute to circulating DNA levels, a clinically relevant predictor of fetal HBV transmission and tumor progression (Zou et al., 2012). We observed a similar geographic distribution between HBV and HERV-K113 (Figure 5E). HERV-K directly interacts with exogenous viruses such as HIV to reduce the infectivity of the resulting chimeric virions (Zwolińska et al., 2013). HERV-K-HBV co-option may explain the observed geographic co-occurrence. Alternatively, apolio B RNA editing catalytic component (APOBEC) mediated mutation of HERV-K may increase during HBV infection leading to initial alignment errors and subsequent classification by BLAST (Lee et al., 2008; Vartanian et al., 2010).

Discussion

In this study, we develop statistical methods for analysis of NIPT data and illustrate the utility of these data for population genetics, association mapping studies, and studies of the human plasma virome. Despite the low sequencing coverage, we demonstrate that accurate genotype imputation is possible, and we discover novel loci associated with height, BMI and two pregnancy-related traits. The results illustrate the power and feasibility of association mapping using NIPT data.

We also leverage the data for population genetic inferences, and show that the majority Han Chinese have evidence of isolation by distance latitudinally, but not longitudinally, presumably due to recent population movements. In contrast, the genetic

diversity in ethnic minorities roughly mirrors geography. We successfully identify known and novel loci that are under selection based on the small allele frequency differences in the Han population. Finally, we identify circulating DNA of viruses with clinical relevance in pregnancy (HBV, CMV, ParvoB19) and reveal a different viral sequencing distribution spectrum compared to Europeans. We analyze genetic and viromic data together and reveal a highly significant association between suspected ciHHV-6A/B and MOV10L1-MCL1 hinting at a possible germline variant affecting the integration of HHV-6A/B.

Our results illustrate the utility of NIPT data for medical genetic studies, particularly for understanding traits related to fertility and pregnancy. Furthermore, the availability of large samples of shotgun DNA sequencing from blood opens up new avenues for investigating hypotheses regarding interactions between viruses and host DNA genetic variability. As NIPT testing expands to millions of individuals globally, obtaining informed consent for patients and effective digital curation of medical records should be prioritized by the medical community.

Acknowledgement

We are grateful to all the participants and BGI colleagues participating in the project. We thank Miaolan Cen from BGI for organising necessary experimental resources in the project; Dr. Heng Li for useful discussions on development of the BaseVar to call SNP from NIPT sequencing data; Dr. Chuan Li for suggestions on structuring the manuscript. Particularly, we would like to thank the professional technical support service provided by Wei Lin, Ruibo Li, Li Tian, Shaoqing Dai, Qi Zhu and Tiecheng Deng from Alibaba Cloud Corperation Ltd. and the supercomputing capabilities provided by Alibaba Cloud MaxCompute and BatchCompute products that greatly shortened the research process. We thank the Tianhe Supercomputer center for computational support. This project supported by Natural Science Foundation of Guangdong Province, China (No.2017A030306026), Funds for Distinguished Young Scholar of South China University of China (No.2017JQ017), Funds for industrial PhD by Innovation Fund Denmark (4135-00130B).

Author contribution

Conceptualization, X.X., R.N., X.J., A.A., S.L., J.S., M.M., Y.Y., J.W.; Methodology, S.L.,A.A., S.H., R.N., S.F., R.D., T.K., A.K., M.Y., Q.F.; Validation, R.W., K.L., Z. C., F.C., S.L., R.L., S.H.; Investigation, S.L., S.H., F.L.,Y.Z.,H.X., SK.L, Z.C., YY.Y., Q.L.; Formal Analysis, S.L., S.H., Z.L., L.L, R.L, Y.Z., J.J.; Resources, Y.Y., L.Z., H.Z., W.W.; Data Curation, Q.L., YY.Y., H.Z., LJ.Z., J.W.; Writing-Original Draft, S.L.; Writing-Review & Editing, S.L., R.N., A.A., S.F., X.J., J.S., Z.C., M.Y., Q.F.;

Visualization, S.L., S.H., Z.L., L.L, R.L, Y.Z., J.J.; Supervision, X.X., R.N., X.J., A.A., J.W.; Project Administration, X.J., S.H., S.L., F.C.; Funding Acquisition, X.J., S.L.;

Declaration of Interests

The authors declare no competing interests.

Reference

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Astle, W.J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A.L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M.A., et al. (2016). The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell *167*, 1415–1429.e19.

Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Barban, N., Jansen, R., De Vlaming, R., Vaez, A., Mandemakers, J.J., Tropf, F.C., Shen, X., Wilson, J.F., Chasman, D.I., Nolte, I.M., et al. (2016). Genome-wide analysis identifies 12 loci influencing human reproductive behavior. Nat. Genet. *48*, 1–7.

494 Britto, J.M., Lukehurst, S., Weller, R., Fraser, C., Qiu, Y., Hertzog, P., and Busfield, S.J.
495 (2004). Generation and characterization of neuregulin-2-deficient mice. Mol. Cell. Biol.
496 *24*, 8221–8226.

497 Buckley, M.T., Racimo, F., Allentoft, M.E., Jensen, M.K., Jonsson, A., Huang, H.,
498 Hormozdiari, F., Sikora, M., Marnetto, D., Eskin, E., et al. (2017). Selection in Europeans
499 on fatty acid desaturases associated with dietary changes. Mol. Biol. Evol. *34*, 1307–
500 1318.

501 Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly,
502 M.J., Price, A.L., and Neale, B.M. (2015). LD Score regression distinguishes
503 confounding from polygenicity in genome-wide association studies. Nat. Genet.
504 47(3):291-295.

505 Carithers, L.J., and Moore, H.M. (2015). The Genotype-Tissue Expression (GTEx)
506 Project. Biopreserv. Biobank. *13*, 307–308.

507 Cheeran, M.C.J., Lokensgard, J.R., and Schleiss, M.R. (2009). Neuropathogenesis of
508 congenital cytomegalovirus infection: Disease mechanisms and prospects for
509 intervention. Clin. Microbiol. Rev. *22*, 99–126.

510 Chen, J., Zheng, H., Bei, J.X., Sun, L., Jia, W. hua, Li, T., Zhang, F., Seielstad, M., Zeng,
511 Y.X., Zhang, X., et al. (2009). Genetic Structure of the Han Chinese Population Revealed
512 by Genome-wide SNP Variation. Am. J. Hum. Genet. *85*, 775–785.

Chen, Z., Chen, J., Collins, R., Guo, Y., Peto, R., Wu, F., Li, L., Lancaster, G., Yang, X., Williams, A., et al. (2011). China Kadoorie Biobank of 0.5 million people: Survey methods, baseline characteristics and long-term follow-up. Int. J. Epidemiol. *40*, 1652–1666.

DePristo, M. a, Banks, E., Poplin, R., Garimella, K. V, Maguire, J.R., Hartl, C., Philippakis, A. a, del Angel, G., Rivas, M. a, Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

Davies, R.W., Myers, S., Mott, R., and Flint, J. (2016). Rapid genotype imputation from sequence without reference panels. Nat. Genet. *48*, 1–7.

Francioli, L.C., Menelaou, A., Pulit, S.L., Van Dijk, F., Palamara, P.F., Elbers, C.C., Neerincx, P.B.T., Ye, K., Guryev, V., Kloosterman, W.P., et al. (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat. Genet. 46, 818–825

Fumagalli, M. (2013). Assessing the effect of sequencing depth and sample size in population genetics inferences. PLoS One *8*, 14–17.

Fumagalli, M., Moltke, I., Grarup, N., Racimo, F., Bjerregaard, P., Jørgensen, M.E., Korneliussen, T.S., Gerbault, P., Skotte, L., Linneberg, A., et al. (2015). Greenlandic Inuit show genetic signatures of diet and climate adaptation. Science. *349*, 1343–1347.

Galinsky, K.J., Bhatia, G., Loh, P.R., Georgiev, S., Mukherjee, S., Patterson, N.J., and Price, A.L. (2016). Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. Am. J. Hum. Genet. *98*, 456–472.

Gudbjartsson, D.F., Helgason, H., Gudjonsson, S.A., Zink, F., Oddson, A., Gylfason, A., Besenbacher, S., Magnusson, G., Halldorsson, B. V., Hjartarson, E., et al. (2015). Large-scale whole-genome sequencing of the Icelandic population. Nat. Genet. *47*, 435–444.

Huang, R.P., Ozawa, M., Kadomatsu, K., and Muramatsu, T. (1993). Embigin, a member of the immunoglobulin superfamily expressed in embryonic cells, enhances cell-substratum adhesion. Dev. Biol. *155*, 307–314.

Jiang, F., Ren, J., Chen, F., Zhou, Y., Xie, J., Dan, S., Su, Y., Xie, J., Yin, B., Su, W., et al. (2012). Noninvasive Fetal Trisomy (NIFTY) test: an advanced noninvasive prenatal diagnosis methodology for fetal autosomal and sex chromosomal aneuploidies. BMC Med. Genomics *5*, 57.

Karagoz, E., Ulcay, A., Tanoglu, A., Kara, M., Turhan, V., Erdem, H., Oncul, O., and Gorenek, L. (2014). Clinical usefulness of mean platelet volume and red blood cell distribution width to platelet ratio for predicting the severity of hepatic fibrosis in chronic hepatitis B virus patients. Eur J Gastroenterol Hepatol *26*, 1320–1324.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, a. D. (2002). The Human Genome Browser at UCSC. Genome Res. *12*, 996–1006.

552 Korneliussen, T.S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: Analysis of Next
553 Generation Sequencing Data. BMC Bioinformatics *15*, 356.

554 Kothapalli, K.S.D., Ye, K., Gadgil, M.S., Carlson, S.E., O'Brien, K.O., Zhang, J.Y., Park,
555 H.G., Ojukwu, K., Zou, J., Hyon, S.S., et al. (2016). Positive Selection on a Regulatory
556 Insertion-Deletion Polymorphism in FADS2 Influences Apparent Endogenous Synthesis
557 of Arachidonic Acid. Mol. Biol. Evol. *33*, 1726–1739.

558 Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and
559 Maglott, D.R. (2014). ClinVar: Public archive of relationships among sequence variation
560 and human phenotype. Nucleic Acids Res. *42*.

561 Lee, Y.N., Malim, M.H., and Bieniasz, P.D. (2008). Hypermutation of an Ancient Human
562 Retrovirus by APOBEC3G. J. Virol. *82*, 8762–8770.

563 Lek, M., Karczewski, K.J., Minikel, E. V., Samocha, K.E., Banks, E., Fennell, T.,
564 O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of
565 protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

566 Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association
567 mapping and population genetical parameter estimation from sequencing data.
568 Bioinformatics *27*, 2987–2993.

569 Li, H. (2015). FermiKit: Assembly-based variant calling for Illumina resequencing data.
570 Bioinformatics *31*, 3694–3696.

571 Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-
572 Wheeler transform. Bioinformatics *25*, 1754–1760.

573 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G.,
574 Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and
575 SAMtools. Bioinformatics *25*, 2078–2079.

576 Li, Y., Sidore, C., Kang, H.M., and Boehnke, M. (2011). Low-coverage sequencing :
577 Implications for design of complex trait association studies. Genome Res.21(6):940–951.

578 Liang, Z., and Ma, Z. (2004). China's Floating Population: New Evidence from the 2000
579 Census. Popul. Dev. Rev. *30*, 467–488.

580 Liang, Z., and White, M.J. (1996). Internal Migration in China, 1950-1988. Demography
581 *33*, 375.

582 Locke, A., Kahali, B., Berndt, S., Justice, A., and Pers, T. (2015). Genetic studies of body
583 mass index yield new insights for obesity biology. Nature *518*, 197–206.

584 Loh, P.R., Tucker, G., Bulik-Sullivan, B.K., Vilhjálmsson, B.J., Finucane, H.K., Salem,
585 R.M., Chasman, D.I., Ridker, P.M., Neale, B.M., Berger, B., et al. (2015). Efficient
586 Bayesian mixed-model analysis increases association power in large cohorts. Nat. Genet.
587 *47*, 284–290.

588 Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters,
589 G., Garcia, F., Young, N., et al. (2013). The Genotype-Tissue Expression (GTEx) project.
590 Nat. Genet. *45*, 580–585.

591 MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H.,
592 McMahon, A., Milano, A., Morales, J., et al. (2017). The new NHGRI-EBI Catalog of
593 published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. *45*,
594 D896–D901.

595 Marchini, J., and Howie, B. (2010). Genotype imputation for genome-wide association
596 studies. Nat. Rev. Genet. *11*, 499–511.

597 Maretty, L., Jensen, J.M., Petersen, B., Sibbesen, J.A., Liu, S., Villesen, P., Skov, L.,
598 Belling, K., Theil Have, C., Izarzugaza, J.M.G., et al. (2017). Sequencing and de novo
599 assembly of 150 genomes from Denmark as a population reference. Nature. 548 ,1–19.

600 Marouli, E., Graff, M., Medina-Gomez, C., Lo, K.S., Wood, A.R., Kjaer, T.R., Fine,
601 R.S., Lu, Y., Schurmann, C., Highland, H.M., et al. (2017). Rare and low-frequency
602 coding variants alter human adult height. Nature *542*, 186–190.

603 Mathias, R.A., Fu, W., Akey, J.M., Ainsworth, H.C., Torgerson, D.G., Ruczinski, I.,
604 Sergeant, S., Barnes, K.C., and Chilton, F.H. (2012). Adaptive Evolution of the FADS
605 Gene Cluster within Africa. PLoS One *7*, e44926

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*. 122.

Moustafa, A., Xie, C., Kirkness, E., Biggs, W., Wong, E., Turpaz, Y., Bloom, K., Delwart, E., Nelson, K.E., Venter, J.C., et al. (2017). The blood DNA virome in 8,000 humans. PLoS Pathog. *13*. e1006292.

Nkhoma, E.T., Poole, C., Vannappagari, V., Hall, S.A., and Beutler, E. (2009). The global prevalence of glucose-6-phosphate dehydrogenase deficiency: A systematic review and meta-analysis. Blood Cells, Mol. Dis. *42*, 267–278.

Ohashi, J., Naka, I., and Tsuchiya, N. (2011). The impact of natural selection on an ABCC11 SNP determining earwax type. Mol. Biol. Evol. *28*, 849–857.

Pasaniuc, B., Rohland, N., McLaren, P.J., Garimella, K., Zaitlen, N., Li, H., Gupta, N., Neale, B.M., Daly, M.J., Sklar, P., et al. (2012). Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nat. Genet. *44*, 631–635.

Peter, B.M. (2016). Admixture, population structure, and f-statistics. Genetics *202*, 1485–1501.

Pruim, R.J., Welch, R.P., Sanna, S., Teslovich, T.M., Chines, P.S., Gliedt, T.P., Boehnke, M., Abecasis, G.R., Willer, C.J., and Frishman, D. (2011). LocusZoom: Regional visualization of genome-wide association scan results. In Bioinformatics, pp. 2336–2337.

626 Raghavan, M., Skoglund, P., Graf, K.E., Metspalu, M., Albrechtsen, A., Moltke, I.,
627 Rasmussen, S., Stafford, T.W., Orlando, L., Metspalu, E., et al. (2014). Upper
628 palaeolithic Siberian genome reveals dual ancestry of native Americans. Nature *505*, 87–
629 91.

630 Rhoads, E.J.M. (2000). Manchus & Han : ethnic relations and political power in late Qing
631 and early republican China, 1861-1928.

632 Sayers, E.W., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin., V.,
633 Church, D.M., Dicuccio, M., Edgar, R., Federhen, S., et al. (2009). Database resources of
634 the National Center for Biotechnology Information. Nucleic Acids Res. *37*. D5-15.

635 Skotte, L., Korneliussen, T.S., and Albrechtsen, A. (2012). Association Testing for Next-
636 Generation Sequencing Data Using Score Statistics. Genet. Epidemiol. *36*, 430–437.

637 Soldin, O.P. (2006). Thyroid function testing in pregnancy and thyroid disease:
638 Trimester-specific reference intervals. Ther. Drug Monit. *28*, 8–11.

639 Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott,
640 P., Green, J., Landray, M., et al. (2015). UK Biobank: An Open Access Resource for
641 Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age.
642 PLoS Med. *12*. e1001779.

643 Suo, C., Xu, H., Khor, C.C., Ong, R.T.H., Sim, X., Chen, J., Tay, W.T., Sim, K.S., Zeng,
644 Y.X., Zhang, X., et al. (2012). Natural positive selection and north-south genetic diversity
645 in East Asia. Eur. J. Hum. Genet. *20*, 102–110.

Yengo, L., Sidorenko, J., Kemper, K., Z, Z., Wood, Andrew., Weedon, Michael., Frayling, Timothy., Hirschhorn, J., et tal. (2018). Meta-analysis of genome-wide association studies for height and body mass index in ~700, 000 individuals of European ancestry. bioRxiv. https://www.biorxiv.org/content/early/2018/03/02/274654.

Vartanian, J.P., Henry, M., Marchio, A., Suspène, R., Aynaud, M.M., Guétard, D., Cervantes-Gonzalez, M., Battiston, C., Mazzaferro, V., Pineau, P., et al. (2010). Massive APOBEC3 editing of hepatitis B viral DNA in cirrhosis. PLoS Pathog. *6*, 1–9.

Vourekas, A., Zheng, K., Fu, Q., Maragkakis, M., Alexiou, P., Ma, J., Pillai, R.S., Mourelatos, Z., and Wang, P.J. (2015). The RNA helicase MOV10L1 binds piRNA precursors to initiate piRNA processing. Genes Dev. *29*, 617–629.

Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R.B., Xu, C., Futema, M., Lawson, D., et al. (2015). The UK10K project identifies rare variants in health and disease. Nature *526*, 82–90.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. *42*. D1001–D1006.

Wigginton, J.E., Cutler, D.J., and Abecasis, G.R. (2005). A Note on Exact Tests of Hardy-Weinberg Equilibrium. Am. J. Hum. Genet. *76*, 887–893.

664 Wildschutte, J.H., Williams, Z.H., Montesion, M., Subramanian, R.P., Kidd, J.M., and
665 Coffin, J.M. (2016). Discovery of unfixed endogenous retrovirus insertions in diverse
666 human populations. Proc. Natl. Acad. Sci. *113*, E2326–E2334.

667 Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y.,
668 Estrada, K., Luan, J., Kutalik, Z., et al. (2014). Defining the role of common variation in
669 the genomic and biological architecture of adult human height. Nat. Genet. *46*, 1173–
670 1186.

671 Xu, S., Yin, X., Li, S., Jin, W., Lou, H., Yang, L., Gong, X., Wang, H., Shen, Y., Pan, X.,
672 et al. (2009). Genomic Dissection of Population Substructure of Han Chinese and Its
673 Implication in Association Studies. Am. J. Hum. Genet. *85*, 762–774.

674 Yan, Y., Su, H., Ji, Z., Shao, Z., and Pu, Z. (2014). Review Article Epidemiology of
675 Hepatitis B Virus Infection in China : Current Status and Challenges. J. Clin. Transl.
676 Hepatol. *2*, 15–22.

677 Yang, J., Jin, Z.-B., Chen, J., Huang, X.-F., Li, X.-M., Liang, Y.-B., Mao, J.-Y., Chen,
678 X., Zheng, Z., Bakshi, A., et al. (2017). Genetic signatures of high-altitude adaptation in
679 Tibetans. Proc. Natl. Acad. Sci. 114,4189-4194.

680 Yang, L.Q., Tan, S.J., Yu, H.J., Zheng, B.R., Qiao, E.F., Dong, Y.L., Zan, R.G., and
681 Xiao, C.J. (2008). Gene admixture in ethnic populations in upper part of Silk Road
682 revealed by mtDNA polymorphism. Sci. China, Ser. C Life Sci. *51*, 435–444.

683 Zhang, H., Gao, Y., Jiang, F., Fu, M., Yuan, Y., Guo, Y., Zhu, Z., Lin, M., Liu, Q., Tian,
684 Z., et al. (2015a). Non-invasive prenatal testing for trisomies 21, 18 and 13: Clinical
685 experience from 146 958 pregnancies. Ultrasound Obstet. Gynecol. *45*, 530–538.

686 Zhang, H., Gao, Y., Jiang, F., Fu, M., Yuan, Y., Guo, Y., Zhu, Z., Lin, M., Liu, Q., Tian,
687 Z., et al. (2015b). Non-invasive prenatal testing for trisomies 21, 18 and 13: Clinical
688 experience from 146 958 pregnancies. Ultrasound Obstet. Gynecol. *45*, 530–538.

689 Zou, H., Chen, Y., Duan, Z., Zhang, H., and Pan, C. (2012). Virologic factors associated
690 with failure to passive-active immunoprophylaxis in infants born to HBsAg-positive
691 mothers. J. Viral Hepat. *19*. e18-25.

692 Zwolińska, K., Knysz, B., Gasiorowski, J., Pazgan-Simon, M., Gładysz, A., Sobczyński,
693 M., and Piasecki, E. (2013). Frequency of Human Endogenous Retroviral Sequences
694 (HERV) K113 and K115 in the Polish Population, and Their Effect on HIV Infection.
695 PLoS One *8*. e77820.

696

697

Main figure titles and legends

Figure 1. Allele frequency spectrum and imputation accuracy of the 141, 431 humans. (A) Allele frequency spectrum of known and novel variants. (B) Sharing of variants among NIPT calls (CMDB, n=141,431), gnomAD East Asian (gnomAD EAS, n=811) and the Han Chinese population from the 1000 genome project (CHN, n=301). (C) Comparison of frequency of the non-reference allele between the NIPT estimations (CMDB) and Han Chinese estimations in the 1000 genomes project (CHN). (D) Count and imputation accuracy of the known variants as a function of minor allele frequency intervals. The 2.1 million known variants are restricted to those with an imputation info score greater than 0.4 and a test p-value of Hardy Weinberg equilibrium frequencies of greater than 10^{-6}. The error bar in red denotes the 97.5% confidence interval, which is generally very small.

Figure 2. Population structure and distribution of allele sharing with related populations in 1KGP. (A). Geographical distribution of the 36 minorities. Size of the circle reflects the number of minority individuals. (B). Principal component analysis of the 36 minorities. A random selection of an equal number of Han Chinese matching the same city of each minority are included and shown as grey colors. English names of the minorities and the number of randomly selected participants from each ethnic and geographical groups from 96,880 participants after QC on error rate and read length are shown in the legend. (C-F). Private alleles sharings between each administrative division (all ethnic groups or only the Han) and the CEU and ITU reference populations in the 1KGP. Color corresponds to the private allele frequency defined in the main text, i.e. the frequency of sampling an allele from each division that is private to reference CEU or ITU populations.

Figure 3. Genetic adaptation in Han Chinese population

(A). Manhattan plot showing the detected selection signals in Han Chinese population across the first principal component. VEP annotated names of the gene loci under selection are displayed.

(B-G). Derived allele frequency per Chinese administrative division for the lead SNP in loci under selection across latitude. The number and the corresponding color in the legend and map indicate derived allele frequency estimated from NIPT data. *ELK2AP-MIR4507* represents the IGH-gene cluster. (H-O). Allele frequency per administrative division for the ClinVar pathogenic variants with a significant difference of allele frequencies across North, Central and South regions. Number and color in the legend and map represent allele frequency estimated from NIPT data for the risk allele recorded in the ClinVar database.

Figure 4. Genome-wide significant signals for two common quantitative traits and two traits related to reproductive process. Known loci, defined as significant variants with a known association with the investigated trait in the GWAS catalogue (e90_r2017-10-10) within1 Mbp region are marked in black. Novel loci are marked in red. For loci where the lead SNP is located in the intergenic region, the most close gene was plotted. Detailed information about the loci can be found in Table S5-S6.

Figure 5. The viral spectrum in maternal plasma. (A) Prevalence of infection among the investigated population. (B) Distribution of abundance by each virus. Each dot represents the abundance of one individual. (C) Manhattan plot showing results from GWAS of carriers of high abundance ciHHV-6A/B vs. non-carriers. (D) Locus Zoom plot denotes lead snp (rs73185306) and the correlated snps around the region of MCL1 and MOVL1 genes. (E-G) Geographic distribution of prevalence for the three most prevalent virus, i.e. HBV, HERV-K113 and HHV6A/6B.

Main tables and legends

Table 1. Replication status of height- and BMI- associated loci

Infoscore[a]	Number of Variants[b]	Mean R^2[c]	Known Loci[d]	Novel Loci	Known(Replicated |Notreplicated)[e]	Novel(Replicated |Notreplicated)
Height						
0.8	788385	0.95	24	4	24|0	4|0
0.7	1552640	0.92	38	6	38|0	5|1
0.6	1922780	0.9	40	6	40|0	5|1
0.5	2057331	0.89	41	7	40|1	6|1
0.4	2104769	0.89	41	7	40|1	6|1
BMI						
0.8	788385	0.95	6	2	6|0	1|1
0.7	1552640	0.92	8	3	8|0	2|1
0.6	1922780	0.9	10	3	10|0	2|1
0.5	2057331	0.89	10	3	10|0	2|1
0.4	2104769	0.89	10	3	10|0	2|1

[a] Info score is provided by STITCH which measures the ratio of the observed statistical information of the population allele frequency and the complete information (see STAR Methods).

[b] Number of variants refer to number of imputed variants with minor allele frequency greater than 0.01 and p-value of hardy Weinberg equilibrium test greater than 10e-6

[c] Mean R^2 refers to the true imputation accuracy comparing the imputed genotype dosage and the true genotypes of the 40 NIPT samples sequenced to 15x.

[d] Loci is defined as a one megabase window extending 500kbp at both the 5' and 3' ends centering on the snp with smallest p-value in the window. Known refers to the existence of one or more known SNPs in the GWAS catalog found within the one megabase window.

[e] Replicated refer to the number of loci that have p-value less than 0.05 divided by the number of associated loci and same beta direction in any one of the CKB, Giant or UK Biobank test sets. Not replicated denotes the number of loci that are not replicated in all three test sets.

Table 2. Replication statistics of the new loci associated with the height and BMI traits found in NIPT

				NIPT		CKB		Giant		UK Biobank	
Height											
CHR	GENE	SNP[a]	MAF	P	Beta	P	Beta	P	Beta	P	Beta
X	UBQLN2-LINC01420	rs7391861	0.34	1.00E-09	0.04	1.85E-07	0.04	NA	NA	NA	NA
X	MIR325HG-FGF16	rs4892720	0.21	5.67E-13	0.05	4.15E-06	0.04	NA	NA	NA	NA
1	MAST2	rs7520050	0.35	4.18E-09	-0.04	0.01	-0.02	***4.60E-23***	***-0.01***	3.49E-07	-0.01
9	STRBP	rs10818797	0.33	1.41E-09	0.04	0.52	0.01	1.40E-14	0.02	9.08E-06	0.01
11	C11orf24-LRP5	rs450416	0.49	1.04E-08	0.04	0.09	0.01	***1.10E-05***	***-0.01***	1.04E-07	0.01
11	ARHGEF12	rs894839	0.28	3.72E-08	-0.04	6.51E-08	-0.05	***6.70E-04***	***-0.01***	1.57E-03	-0.01
20	LINC00261	rs1203887	0.04	4.07E-08	-0.09	0.02	-0.05	***1.90E-03***	***0.01***	0.90	0.003
BMI											
1	PLD5-LINC01347	rs2780797	0.45	2.57E-08	0.04	0.01	0.02	2.10E-06	0.01	3.69E-03	0.01
11	LOC101054525-LINC01441-	rs12803364	0.26	1.25E-09	-0.04	0.37	0.01	0.66	0.00	0.54	-0.003
20	CBLN4	rs59271815	0.16	3.24E-10	-0.05	2.39E-04	-0.04	***3.60E-08***	***-0.01***	1.67E-05	-0.01

a. proxy SNP is used and marked in bold italic that have similar p-value and effect size in NIPT. NA means no proxy SNP found.

Information about the proxy SNP can be found in Table S5 and Table S6.

Supplemental figure titles and legends

Figure S1. Geographic distribution of 140K participants over 34 administrative divisions in China and sequencing depth, Related to Figure 1, Table S1 and the STAR Methods

(A) Geographic distribution of 140K participants over 34 administrative divisions in China. Color and number in legend indicate the number of participants. Participants from three divisions including Hong Kong, Macau and Taiwan are not involved in the study.
(B) Integrated sequencing depth of all the study participants over accessible and unaccessible regions.
(C) Sequencing depth per individual by chromosomes.
(D) Integrated sequencing depth over the 22 autosomal and the X chromosomes. Sequencing depth for participants with 35bp read length, 49bp read length and a summation of both groups are shown in green, blue and red, respectively.

Figure S2. Quality measurement of variation calling and imputation accuracy, Related to Figure 1, Table S1 and the STAR Methods

(A) Changes in Ti/Tv ratio and the ratio of the remained variants as a function of the increasing filtration threshold on variant recalibration score. In the legend, "positive" and "negative" refers to the status of variants selected as the positive set and negative set in the training process. "All" refers to the status of all variants.
(B-E) Upper bound false positive counts using variants with >=2 alleles in 40WGS validation data. "All", "Known" and "Novel" variants refers to all the variants, variants known in dbsnp147 and variants not known in dbsnp147. False calls are those that are called by BaseVar from the 140K NIPT data and have at least two alternative alleles in the low coverage NIPT sequencing data of 40 validation participants but are not identified as variants in the high coverage sequencing data of those same individuals.
(F) Imputation accuracy for all the variants per individual as a function of fetal fraction.

Figure S3. Principal component analysis and sharing of ancestry with reference populations, Related to Figure 2 and STAR Methods

(A-D) Principal component analysis for the full sample set (N=141, 431). (A) Distribution of eigenvalue for the top eight principal components. (B). PCA colored by read length suggests the first principal component is the batch effect caused by differential mapping of different read length. (C). PCA colored by three latitudinal geographical regions (North, Central and South) suggests the second principal component reflects genetic differentiation across latitude. (D). PCA colored by sequencing error rate suggests the third principal component likely corresponds to the error rate.
(E-F) Principal component analysis of the self-reported Han Chinese (N=43, 387). (E) Geographic locations of the Han. Divisions belong to the "North", "Central" and "South" regions defined by the Chinese government are colored as ""Green", "Red" and "Blue".
(F) Visualization of the top two principal components. Colors correspond to their digital geographical information in Panel A. Solid and dashed ellipses line refers to 95% confidence interval of the PC distributions for individuals from the three geographical regions assuming a multivariate t-distribution or a multivariate normal distribution, respectively.

(G-K) Allele sharing between Han Chinese in each of the 31 administrative divisions with populations in the 1000 Genomes project by the F3 statistic. CHB, CDX, CHS, JPT and KHV refers to Han Chinese in Beijing, Southern Han Chinese, Chinese Dai in Xishuangbanna, Japanese in Tokyo and Kinh in Ho Chi Minh City, respectively.

Figure S4. Quality evaluation in genome-wide association studies, Related to Figure 4, Table 1-2, STAR Methods

(A-B) Distribution of mother's height and body mass index for the 61, 717 participants with height and BMI records.
(C-D) QQ-plot for height and BMI using all 2.1 million variants with Info Score > 0.4, HWE p-value < 10-6 are shown in black. Results excluding all loci known to be associated with the trait is shown in grey line.
(E-K) Correlation of effect size of genome-wide significant variants between discovery set (NIPT) and three test sets (Giant, UK Biobank and CKB) for height(E-(G) and BMI(H-K). Linear regression was performed and the fitting line is shown in red.
(L) Distribution of the age of the mother in the year when taking the test. The four number "13", "28", "35", "48" refer to the minimum, two modes of the binomial distribution and maximum of age.
(M) QQ-plot for maternal age.
(N) QQ-plot for twin pregnancy.

Figure S5. Locus Zoom plot for novel association loci, Related to Figure 4, Table 1-2, STAR Methods

Locus Zoom plots for height (A-F); BMI (G-I); maternal age (J-K); twin pregnancy (L). The rs number and P-value of the most significant primary SNP were labeled on top of the SNP. R2 LD estimated using 1KGP-CHN haplotypes between each SNP and the most significant SNP was color coded.

Figure S6. Extended information for the virome spectrum in plasma, Related to Figure 5, STAR Methods

(A-B) Plasma virome not removing phage and low abundance participants. Shown are the prevalence (A) and the abundance of the viruses (B). Only the top 40 viruses with the highest prevalence are shown.
(C) Medical records for participants with HBV infection (significant hits >=2). Free/NA means no clinical information are available.
(D) QQ-plot for high abundance ciHHV-6A/B phenotype.
(E-J) Geographical distribution of six viruses with sample size greater than 200 in addition to HBV, HERV-K113 and HHV6A/B in Figure 5.

Figure S7. Integrated coverage from 138, 882 participants towards virus genome, Related to Figure 5, STAR Methods

Reads with mapping quality equal to zero were excluded. Only viruses with coverage are shown.

STAR METHODS

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information should be directed to one of the corresponding authors. Summary statistics, including population-level information, such as allele frequencies and genome-wide association summary statistics, are available at website https://db.cngb.org/cmdb/, which can be accessed by qualified researchers by request to bigdata@genomics.cn with an application form from the website. Individual level data including sequencing fastq files, alignment cram files, vcf files that contain individual information are hosted securely in the China National Genebank. Individual genomic data are not made publicly available following Chinese regulations in the Interim Measures for the Administration of Human Genetic Resources after a review by the Human Genetic Resources Administration of China (HGRAC).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

All participants were recruited via the non-invasive fetal trisomy test at BGI between year 2012 and 2013 (Zhang et al., 2015). They underwent pretest counseling and filled in informed written consent before blood sampling. The study was reviewed and approved by the Institutional Review Board of BGI (BGI-IRB17088) in strict compliance with regulations regarding ethical considerations and personal data protection.

The geographic locations of the participants were provided in an anonymous way via the first six digits of the resident identity card number, indicating birth place

information of the participant. Year of birth and the age of the participants were obtained from the written informed consent. The age distribution of 137, 984 participants suggests a bimodal distribution with peak for birth-years of 1977 and 1984 and ages 28 and 35 (Figure S17). The bimodal distribution is likely related to China's first and second child policy and affected by changes to the clinical guidelines for undertaking the NIPT test, where all pregnant women (rather than only high-risk pregnant women) are recommended to take the NIPT test. The height and weight measurements were recorded when the blood sample was taken in the hospital. The BMI is calculated using the standard formula "weight (kg) / height2 (m2)".

Relative fetal fraction can be estimated accurately for a participant if the gender of the child is male based on the proportion of reads that map to Y chromosome relative to the reads that map to the whole genome. Using this proportion, we estimate that the fetal fraction is approximately 3.5% to 30% among all participants, with a median of 8%.

The status of chromosomal aneuploidy was detected using a method we previously developed for screening of fetal chromosomal aneuploidy (Jiang et al., 2012). In addition, both the participant, and if available, the father, reported their karyotype status to the hospital. We removed participants with sequencing error rate greater than 1% (N=3,278) and with potential abnormal chromosomal aneuploidy detected either via the participant's report or detected using the read coverage method (N=502) from further analyses, resulting in 141, 431 participants consisting of 118, 576 participants with 35bp read length and 22, 855 participants with 49bp read length.

METHOD DETAILS

Sequencing and QC

Details of the sequencing protocol were published previously in (Zhang et al., 2015b). In brief, within 8h of blood collection, plasma was extracted from whole blood after two turns of centrifugation. The plasma samples (N=145, 211) were subsequently subjected to library construction, sample quality control and 36-cycle or 50-cycle single-end multiplex sequencing on Illumina Hiseq 2000 platform. The reads were trimmed to 35bp and 49bp before bioinformatic analysis. Filtering of poor quality reads was carried out using SOAPnuke (https://github.com/BGI-flexlab/SOAPnuke). A read was removed if it contained more than 30% low quality bases (Q<=2) or N bases. In general, each participant was whole-genome sequenced to 5-10 million cleaned reads, representing a sequencing depth around 0.06x -0.1x.

Alignment of reads against hg19 and definition of accessible region

For each participant, the cleaned reads were aligned against the hg19 human genome reference with bwa, using the single-end read alignment option(Li and Durbin, 2009). Potential PCR duplicates were removed using samtools rmdup(Li et al., 2009). The indel realignment and base quality recalibration modules in GATK were applied to realign the reads around indel candidate loci and to recalibrate the base quality(DePristo et al., 2011). Finally, the alignment files were stored in the standard CRAM format. The

alignment of all participants was carried out in the Batch Compute system in Aliyun cloud parallelizing 2000 jobs in a batch (~24 hours per job) (https://github.com/aliyun/aliyun-openapi-java-sdk/tree/master/aliyun-java-sdk-batchcompute). Completion of the whole alignment process took around one week.

After alignment, QC statistics were computed using the stats function implemented in samtools which measures sequencing error rate as the proportion of bases that differ from reference base at base positions, i.e. the mismatch rate. The median sequencing error rate was estimated to be 0.3%. Subsequently, we used the samtools depth to estimate the overall coverage of reads with mapping quality >= 30 and bases with base quality >= 20. Reads or bases with quality lower than this threshold were not included in any of the following analysis. We compared the coverage to the mappability uniqueness (wgEncodeDukeMapabilityUniqueness35bp.bedGraph) as well as gene and repeat density information from the UCSC database(Kent et al., 2002) (Figure S4).

Considering potential errors in alignment of short single-end reads, we defined an accessible region for variation calling, population genetics and association mapping analysis. We defined the accessible region as follows: 1) regions that are not in the 35-kmer problematic alignment bed file provided by Heng Li (https://github.com/lh3/sgdp-fermi/releases/download/v1/sgdp-263-hs37d5.tgz). Those regions in the hg19 reference genome were detected by Fermi assembler (Li, 2015) as difficult to map uniquely using a 35bp kmer unit sequence. This filter removed 897,319,085 bp hg19 sequence; 2) regions with a mappability uniqueness score equal to one according to wgEncodeDukeMapabilityUniqueness35bp.bedGraph in UCSC. This filter additionally

removed 6,459,019 bp hg19 sequence; 3) regions where the sequencing depth was between 3000 and 30,000. An additional 4,340,936 bp are excluded in this step. The final length of accessible region in hg19, including the 22 autosomal chromosomes and the X chromosome is 2,128,184,806 bp.

Medium depth sequencing of 40 participants

With informed consent, we sequenced 40 participants out of the total 141, 431 participants using the Hiseq X10 system at a medium depth of 15x. We aligned the reads to the same hg19 human reference genome using bwa-mem(Li and Durbin, 2009) and applied the GATK multi-sample best practice(DePristo et al., 2011) to call and genotype SNPs for the 40 participants. The SNP calls and genotype results were used to benchmark the SNP calling performance using the ultra-low depth sequencing data as well as the genotype imputation accuracy.

Variant discovery and allele frequency estimation

We applied a maximum likelihood approach to identify polymorphic sites and infer allele frequencies (Liu et al., manuscript in preparation). We adopted a single-read sampling strategy, where we sampled only one read for each variant candidate site if there were more than one read. This method ensured that, for each site, all reads derived from different individuals and the allele frequency spectrum is therefore not biased due to existence of fetal DNA in the sample. This maximum likelihood framework is much faster than a representation using diploid genotype likelihoods (DePristo et al., 2011; Li, 2011).

Maximum likelihood estimation of allele frequency

Likelihood Function for a single site

For *N* unrelated individuals with a single read covering the position, the likelihood function for the read data D_i, for a single variant candidate site in individual *i*, of the allele frequency $p = (p_A, p_C, p_G, p_T)$, is defined as:

$$\mathrm{L(p)} = \prod_{i=1}^{N} P(D_i|\, p) = \prod_{i=1}^{N} \sum_{b\in\{A,C,G,T\}} p(b|p)p(D_i|\, b) \quad (1)$$

where $p(b|p) = p_b$ and the genotype likelihood assuming a haploid model is $p(D_i|\, b) = \{1 - \varepsilon_i \, if \;\; D_i = b \; and \;\; \varepsilon_i/3, \, if \; D_i \neq \;\; b$. ε_i corresponds to the GATK corresponds to the GATK-recalibrated error rate converted from the PHRED-scale base quality.

Optimization

We obtain the maximum likelihood estimate $\hat{p} = argmax_p L(p)$ using the EM algorithm with starting value computed by the observed allele frequency:

$$p_b = \frac{\sum D_i = b}{N} \quad (2)$$

In the E step, we compute the posterior probability of allele *b* for individual i at a site *j* as one of the four A/C/G/T bases:

$$P(b|D_i) = \frac{p(b|p)p(D_i|\, b)}{\sum_{b\prime\in\{A,C,G,T\}} p(b\prime|p)p(D_i|\, b\prime))} \quad (3)$$

We compute the updated allele frequency p' in the M step as

$$p'_b = \frac{\sum_{i=1}^{N} P(b|D_i)}{N} \quad (4)$$

When the change in the maximum likelihood is less than 0.001, we terminate the algorithm.

Decision of allelic type and confidence of SNP calling: Likelihood Ratio Test

Formulae (1) – (4) can be used for estimation of allele frequencies of all four nucleotides simultaneously, and may result in tetra-allelic and tri-allelic variant calls. We will use this formulation for SNP calling and for identifying potential tri- and tetra-allelic loci. Denote the likelihood value from the four-allelic model in Equation (1) as f_4. We iteratively set the allele frequency of one of the four nucleotides to zero to obtain models of tri-allelic loci. Let $\hat{f}_3(p_x = 0)$ denote the maximum likelihood value when the frequency of allele *x* is constrained to be zero. We then compute a log likelihood ratio statistic as:

$$LRT_{4vs3} = -2log\left(\frac{\hat{f}_3(p_x = 0)}{\hat{f}_4}\right) \quad (5)$$

The tri-allelic model is nested within the tetra-allelic model and, therefore, the distribution of the LRT_{4vs3} statistic asymptotically follows a chi-square distribution with one degree of freedom, under the assumption of a tri-allelic locus. If the p-values of one of the four LRT_{4vs3} test are significant ($<10^{-6}$), the variant will be classified as a tetra-

allelic loci. If not, we move on to the test a model of a tri-allelic locus versus a bi-allelic locus, where x if the $\widehat{f_3}(p_x = 0)$ is the allele with minimum likelihood (which results in maximum p-value out of LRT_{4vs3}) was set as the alternative-hypothesis and the reduced hypothesis is $\widehat{f_2}(p_x = 0, p_y = 0)$ where $p_y$is the allele frequency for allele y.

$$LRT_{3vs2} = -2log(\frac{\widehat{f_2}(p_x=0,p_y=0)}{\widehat{f_3}(p_x=0)}) \quad (6)$$

Again, the distribution of LRT_{3vs2} asymptotically follows a chi-squared distribution with one degree of freedom under the hypothesis of a bi-allelic locus. If the maximum p-value out of the three LRT_{3vs2} is significant, the variant will be classified as a tri-allelic variant. Otherwise, we continue to test the bi-allelic versus mono-allelic assumption, as defined in the equation below, with *y* being the allele with the highest p-value

$$LRT_{2vs1} = -2log(\frac{\widehat{f_1}(p_x=0,p_y=0,p_z=0)}{\widehat{f_2}(p_x=0,p_y=0)}) \quad (7)$$

Formula (8) is also used to quantify the confidence of the SNP call. We keep variants with p-values less than 10^{-6}.

Note that our method identifies multi-allelic variants. However, since we don't have sufficient validation of the performance for such variants, we focus on reporting results for bi-allelic loci.

Variant quality score recalibration

~32 million raw variants were obtained using a p-value less than 10^{-6} based on the maximum likelihood model in the accessible region (Table S1). However, this set of SNPs may contain false positives due to miscalculated quality scores, alignment errors, or other technical issues such as contamination. We therefore applied a Bayesian Gaussian mixture model, similar to the VQSR model in GATK(DePristo et al., 2011) to assign each variant candidate a Phred-scaled probabilistic score (in short, VQSR score) indicating the probability that the variant is a truly polymorphic variant (Figure S5). The higher the VQSR score, the higher probability that the variant candidate is a true polymorphic variant. The Gaussian mixture model was established by learning technical features of a training set that consists of variant likely to be real. In our case, the training set was defined as a subset of the common known variants (N=50K was randomly chosen). Features include the Fisher exact test statistic for strand bias, sequencing depth, indel density in a 30bp window centered around the variant candidate, and the raw variant quality score using a maximum likelihood model described above. The same likelihood function and expectation and maximization process as that reported in the GATK framework(DePristo et al., 2011) was implemented except that we used prior probability 50% and 50% for all the variants.

The transition versus transversion (Ti/Tv) ratio is high for the raw call set (maximum 3.4 for known variants and 8.9 for novel variants) and decreases as the filtration threshold of VQSR score increases. The final filtration threshold of VQSR score is decided to be 35, which suggests a Ti/Tv ratio of 2.1 for known variants and of 2.4 for

novel variants and a sensitivity of 85% for the common known variants used for training (Figure S6).

Annotation

Annotation of the genes mentioned in the manuscript and the annotation of the existence of the variants in database such as dbSNP, GnomeAD, 1KGP was carried out using Variant Effect Predictor(McLaren et al., 2016).

Imputation

We employed STITCH (version 1.2.7)(Davies et al., 2016) to impute genotype probabilities for all 141,431 individuals in a five megabase window with a 250K buffer assuming 10 ancestral haplotypes. The key parameter *K* (number of ancestral haplotypes) was decided based on tests over a 5Mbp region in chr3 (chr3: 180000000-185000000) via useful discussions with the STITCH author Robert W Davies. Allele frequency information from the Chinese population (CHB+CHS+CDX, N=211) in the 1KG impute2 reference panel was used for the initial values for the EM optimization of the model parameters. 607 jobs were parallelized in the Tianhe 2 supercomputer in Guangzhou city.

The imputed loci are a target of 8.16 million known polymorphic sites in 22 autosomal chromosomes and chrX with a 1KG East Asian allele frequency >=0.01. All the loci recorded in the GWAS catalog are also included for imputation.

For each of the imputed site, there is an IMPUTE2-style info score(Marchini and Howie, 2010) and a P-value for violation of Hardy Weinberg equilibrium (HWE-pvalue

in short)(Wigginton et al., 2005). We used info score greater than 0.4 and HWE-pvalue smaller than 10^{-6} since the remaining variants has greatest power and good replication rate for height and BMI association test.

Principal component analysis

For the PC analysis, we restricted ourselves to known variants with minor allele frequency greater than 0.05 in the data. We continued to use the single read sampling strategy. To compute the covariance between individuals *i* and *j* for *M* loci, we used

$$C_{i,j} = \frac{1}{M}\sum_{m=1}^{M}\frac{(h_m^i - f_m)(h_m^j - f_m)}{f_m(1 - f_m)}$$

where f_m is the minor allele frequency and h_m^i is the haploid genotype coded as either 0 or 1 for the major and minor allele, respectively.

Using the above formulas, we parallelized the distance and the covariance matrix computation in 90 nodes from the Aliyun ODPS system and obtained the full matrix of 141,431 individuals in a few days. We applied the Spectra R package (https://cran.r-project.org/web/packages/RSpectra/index.html) to perform the decomposition of the covariance matrix.

Finally, we visualized the top three principal components and colored the points according to the administrative divisions, the ethnic groups, and also the read length and error rate.

We have carried out several principal component analyses to answer different questions using the workflow described above. First, we carried out a PCA for all 141, 431 participants to identify the main principal components (Figure S18). After noticing that the first principal component reflects read length and the third reflects the estimated sequencing error rate (measured using the stat function in samtools), we only used the 96,880 participants with read lengths of 35bp and sequencing error rate smaller than 0.00325 for further population genetic analysis (Figure 2). In particular, for the PCA in the Han Chinese population (Figure S9), we excluded participants that did not report their ethnicity (although a large majority of them are probably Han). We only used the 45, 387 participants who reported Han ethnicity to understand population structure of the Han.

F3-statistic and private allele frequency analysis

To quantify divergence between populations we use the outgroup F3 statistic, which is a measure of drift time between two populations(Raghavan et al., 2014). The F3 statistic is highly influenced by common alleles, that tend to be older, and we noticed that F3 statistics between CEU and ITU 1KG samples, and samples from each of the Chinese administrative divisions, were highly correlated due to the sharing of ancestry between the CEU and ITU populations after their separation from East Asian populations (Figure S8). We therefore also measured genetic relatedness using a measure based on alleles that are private to either the African (YRI), east Asian (CHB), South Asian (ITU) or Europeans (CEU) 1KG sample. Private alleles are defined as those that were polymorphic in one group and fixed in the other groups. There were in total 3, 485, 371 and 4, 324,376 private alleles in the CEU and ITU samples respectively. We further

applied a filter, using the private alleles that were common in one group with a MAF > 5% and obtained in 66,700 and 45,536 private variants in the CEU and ITU samples, respectively. Those common private variants were used to compute the private allele frequency defined below. We assume that the proportion of allele sharing of these private alleles with any of the administrative division, should be informative regarding genetic exchange at a more recent timescale.

For each administrative division we calculated the fraction of alleles in the NIPT dataset that matched the private allele found in population *K* in the 1KG. We denote the number of sites that contains an allele that is private to population, *K*, as M_k. For each site, *s*, that contains a private allele for population *K*, we count the number of alleles that match the private allele, n_s, and normalize by the total number of alleles N_s. The private allele frequency for population *K* is defined as

$$PAF_k = \frac{\sum_{s=1}^{M_k} n_s}{\sum_{s=1}^{M_k} N_s}$$

Standard errors were estimated using a 5Mb weighted block jackknife where the weights are the number of sites with private alleles within the block.

Detection of selection across PC coordinates

To detect the most extremely differentiated variants, we use a method based on finding deviations from the patterns predicted using the first components of a PCA analysis. We

have adapted the FastPCA statistic(Galinsky et al., 2016) to work on covariance matrix based on PCA (using the NIPT sequencing data). Assuming the eigenvectors obtained by PCA capture the structure in the data in the absence of selection, for each SNP, we use Equation (11) in reference (Galinsky et al., 2016) to calculate a p-value associated with deviations from the genomic pattern. The resulting p-values were visualized using a Manhattan plot.

Detection of clinvar pathogenic variants displaying significant allele frequency differentiation

We investigated allele frequency differences among the three populations defined by the three latitudinally separated geographic divisions in a total of 3,238 bi-allelic potentially pathogenic variants with a clinical significance level of 5 from the total of 246,385 variants in clinvar database(Landrum et al., 2014) (URL: ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/clinvar_20170404.vcf.gz). We note that even for the pathogenic variants with clinical significance level of 5, some of them have high allele frequency greater than 1% in both our data set and the 1KG population. We counted the number of risk and non-risk alleles (B) for each of the North, South and Central populations (Extended Figure 1) using the formula

$$C_b = \sum_{i=1}^{N_b} (1 - 10^{\frac{-q_b}{10}})$$

$$b \in \{A, C, G, T\}$$

where C_b refers to total allele count of b, q_b refers to the base quality of the observed base *b*. For each variant, a two-tailed p-value using Fisher's exact test was calculated for the 2x3 table with data from the 2 alleles and 3 regions.

For each of the pathogenic variants, as well as for all other variants throughout the genome, we estimated allele frequencies using BaseVar. The risk allele frequency of position *j* is defined as R_j. We frequency-matched random SNPs from the whole genome data with SNPs in clinvar by, for SNP *j* in clinvar, randomly selecting 100,000 variants from the non-clinvar variants within a frequency range of $[0.9R_j, 1.1R_j]$. We estimated the rank and percentile of the pathogenic variant comparing to the 100,000 variants. We reported and visualized the top eight loci with a p-value retrieved from the Fisher's exact test was less than 10e-6, and the p-value from the comparison to non-clinvar variants was less than 5e-3.

Identification of genetic variants significantly associated with a trait

We used the score test(Korneliussen et al., 2014) implemented in Angsd(Korneliussen et al., 2014) to detect the association signal between the imputed genotype probabilities and phenotypes, followed by a linear regression for quantitative trait or logistic regression for qualitative trait to compute the effect size of the top SNPs. For height, we applied the top five principal components, the maternal age and the sex of the fetus as covariants. For BMI, we additionally included the gestational age of the fetus as a covariate. We note that the BMI phenotype in the NIPT cohort is not only related to the mothers' non gestational weight but also related to gestational weight gain including the fetus's

weight. We are mostly interested in the genetic effects on the maternal phenotype in this study and therefore, we have used the gestational age and the sex of the fetus as covariates to account for the fetal growth rate and the effect of sex. Even so, there may be some residual variance in the BMI phenotype caused by differences in fetal growth rate. For maternal age, we used the top five principal components and the sex of the fetus as covariants. For the rest of the phenotypes including twins and virus integration and infection, we applied the same covariates as height. Independent loci were defined as significant variants clustered in a 1Mbp window. The lead SNP was defined as the SNP in the 1Mbp window that has most significant, i.e. smallest p-value. The conditional test module in SNPtest(Marchini and Howie, 2010) was used to estimate the number of independent signals for each independent loci. Finally, locuszoom(Pruim et al., 2011) was applied to visualize the loci. The reported loci were determined from the conditional test after the single marker analysis using a significance threshold P value $\leq 5 \times 10^{-8}$.

The genomic inflation factor, GC lambda, attenuation ratio, LD score regression intercept and the SNP heritability were estimated using the LD score regression approach (Bulik-Sullivan et al., 2015).

Replication of significant loci

For replication, we compared all the variants reaching the significance threshold to three independent studies- the China Kadoorie Biobank (CKB cohort)(Chen et al., 2011), the

recent Giant meta-analysis (Yengo et al., 2018, bioRxiv) and the UK Biobank (Ben Neales's website). The CKB cohort has measurements of height and BMI data and has chip-genotyped 32,000 Chinese participants with imputation into the 1000 genomes Phase reference panel. The GC lambdas from the CKB association test using BOLT-LMM(Loh et al., 2015) were 1.10 and 1.17, respectively, for height and BMI. The GIANT and the UK Biobank summary statistics consist of 2.3 million and 10 million SNP markers, respectively. For some associated loci, the lead SNP is not present in the test. For replication purpose, after ascendingly ranking the SNPs by p-value, we chose the first SNP present in the test data as the proxy SNP. In almost all cases the p-values and effect sizes of the lead SNPs are similar to the p-value of the proxy SNPs. When proxy SNP instead of the lead SNP was used for the replication, we marked it cleary in the result.

We defined a locus as replicated if the lead SNP or the proxy SNP 1) has a p-value less than 0.05 divided by the number of loci (for height, N=48; for BMI, N=13) 2) has the same directionality of the effect, in at least one of the CKB, the GIANT or the UK Biobank test set. We note that the genomic inflation factor is high in both the GIANT and the UK Biobank (Supplementary Table4). We also estimated the quantile in Supplementary 5 and Supplementary 6 for reference.

The GWAS catalog database(Welter et al., 2014, e87_r2017-02-20) defines known and novel loci. The b38 coordinates was transferred to the b37 coordinates using

the liftover script from UCSC (http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/liftOver).

Viral sequence analysis

We applied BLAST(Altschul et al., 1990) to align the reads that did not map to the human reference genome (hg19+human decoy sequences) to the NCBI viral reference sequence database(Sayers et al., 2009) (ftp.ncbi.nih.gov/refseq/release/RefSeq-release84/viral/viral.2.1.genomic.fna.gz). For each read, we kept the best alignment with smallest e-value. Only reads with an evalue < 1e-5, identity >=97 and alignment length >=32bp were counted as a hit. After removing alignments to bacteriophages, we found that out of the 138,882 samples analyzed, 48,298 samples (34.8%) have at least one viral hit and 11,351 samples (8.2%) have at least two significant hits mapped to the viral reference database. In Figure 4 Panel A, to reduce false positive, we defined individuals with at least two significant hits aligned to the same virus as carriers of that virus. We then carried out prevalence and abundance analysis of each virus using the top viral hit for each read. For prevalence and abundance analysis, we only used the individuals with at least two hits for a specific virus. Virus abundance was calculated by the following equation adapted from(Moustafa et al., 2017) :

$$Abundance = \frac{2x\frac{numberof\ reads\ mapped\ to\ viral\ genome}{virus\ genome\ size}}{\frac{number\ of\ reads\ mapped\ to\ human\ genome}{human\ genome\ size}}$$

Viruses with multiple strain entries in RefSeq were aggregated for high homology between entries and ease of graphical display. These viruses include: Anellovirus (TTV), HBV, HHV-6A, HHV-6B, HHV-5, Influenza, etc. For virus sequencing coverage analysis, we aggregated the read depth of all the individuals with mapping quality >0 for each virus.

Supplemental table titles and legends

Table S1. Number of variants by chromosomes, Related to Figure 1, STAR Methods

Table S2.Loci under selection across latitude, Related to Figure3, STAR Methods

Table S3. Clinvar pathogenic variants with significantly different allele frequency among Han Chinese population from North, Central and South Regions, Related to Figure 3, STAR Methods

Table S4. Comparison of genomic control lambda and LD score regression statistics between Giant and UK Biobank, Related to Figure 4, STAR Methods

Table S5. Loci associated with height, Related to Figure 4, STAR Methods

Table S6. Loci associated with BMI, Related to Figure 4, STAR Methods

Table S7. Homologous sequence between viral reference genomes and the human reference genomes by bwa alignment, Related to Figure 5, Figure S7 STAR Methods