The American Journal of Human Genetics, Volume *108*

Supplemental information

A unified framework for cross-population trait prediction by leveraging the genetic correlation of polygenic traits

Mingxuan Cai, Jiashun Xiao, Shunkang Zhang, Xiang Wan, Hongyu Zhao, Gang Chen, and Can Yang

1 Supplementary Figures

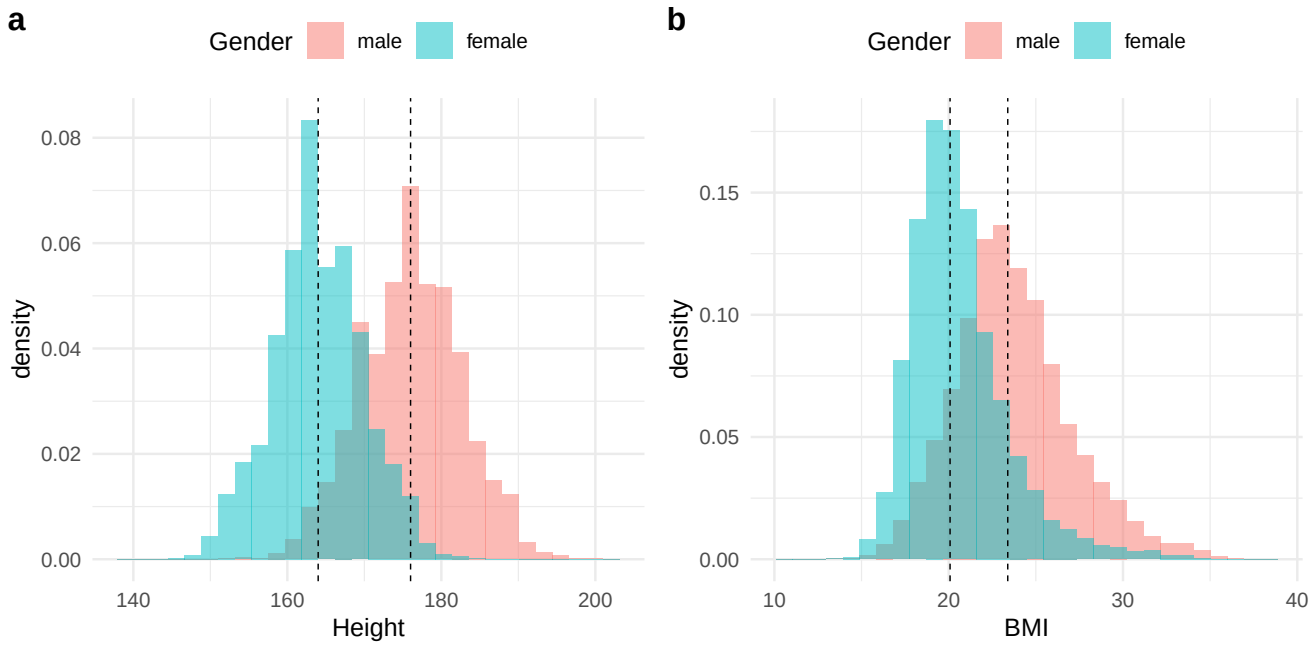

Figure S1: The distribution of height (**a**) and BMI (**b**) in Chinese dataset.

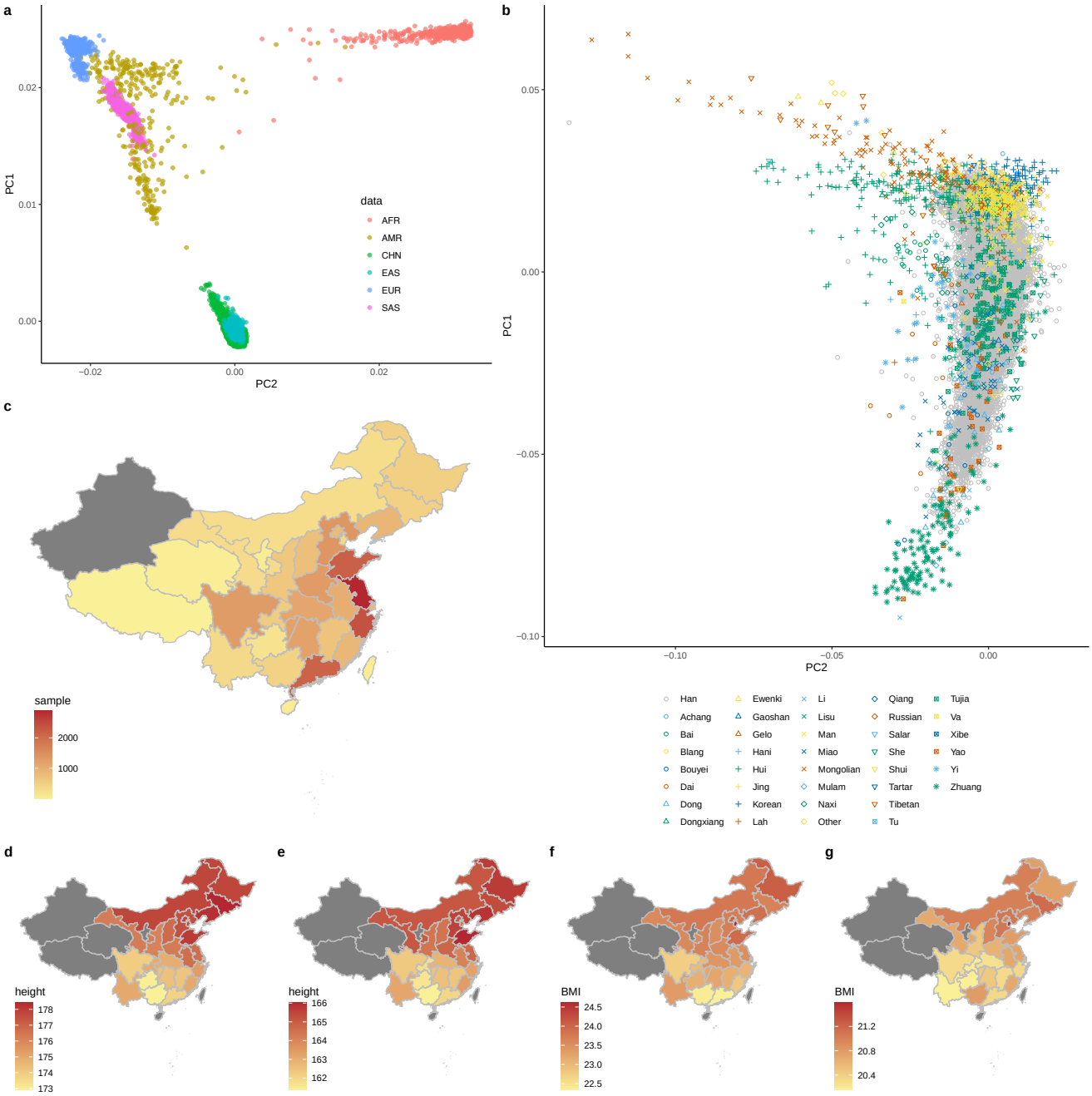

Figure S2: **a**, PCA of the combined samples from Chinese cohort and the 1000 Genomes Project. Chinese are genetically closest to East Asians in the 1000 Genomes project. **b**, PCA of Chinese participants only. The first two principal components reflect the longitudinal and latitudinal differentiation behind Chinese genetic structure. **c**, Distribution of genotyped individuals by province. The majority of genotyped samples are from the southeastern area of China. **d-g**, Average phenotypic values of male height (**d**), female height (**e**), male BMI (**f**) and female BMI (**g**) for provinces with more than 50 samples. Four administrative divisions Xinjiang, Tibet, Qinghai and Ningxia are shown in grey because their sample sizes are less than 50.

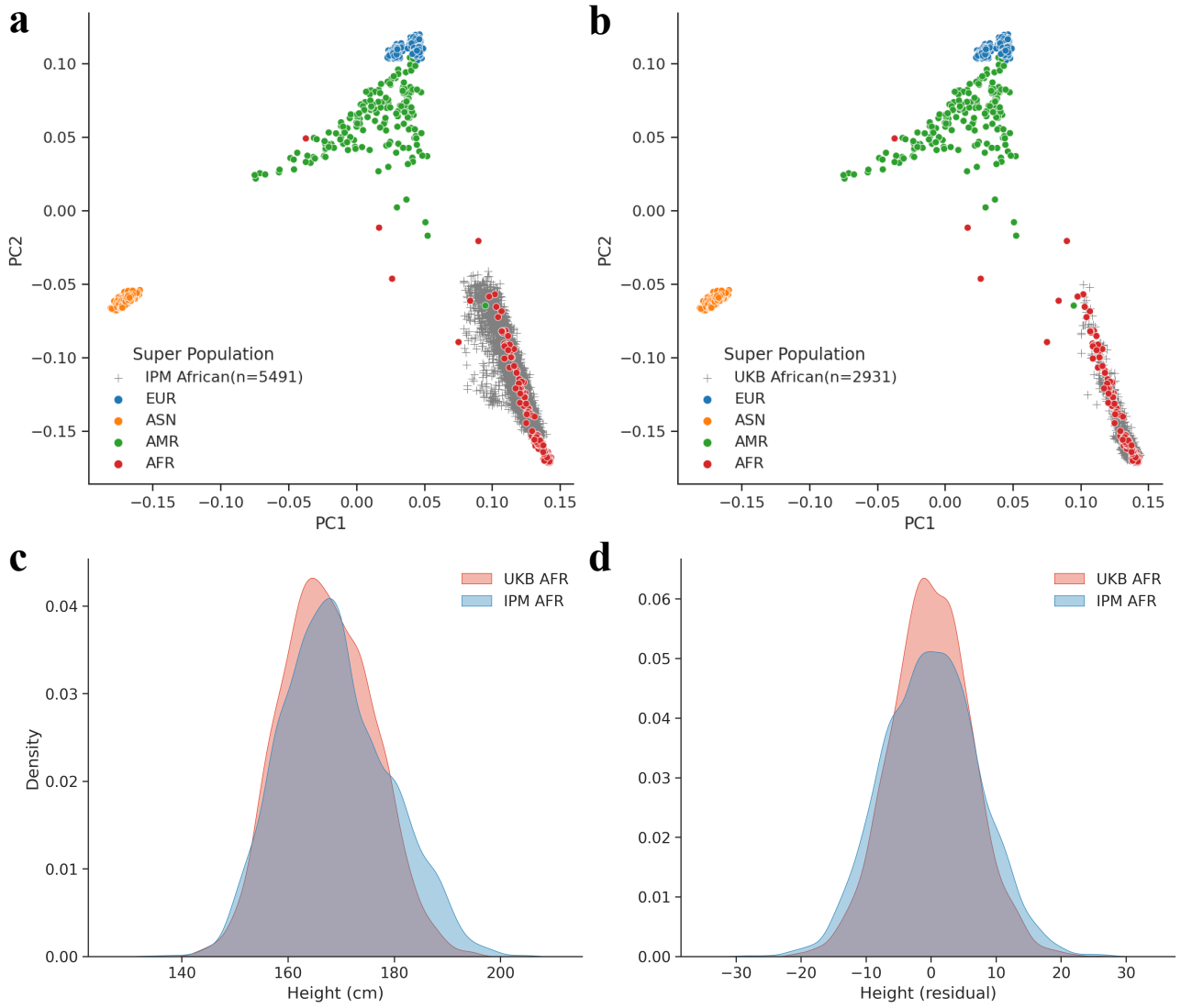

Figure S3: The PCA projection of IPM Afican participants (a) and UKBB African participants (b) to the 1000 Genome Project dataset. Kernel density estimation (KDE) of height (c) and its residual (d).

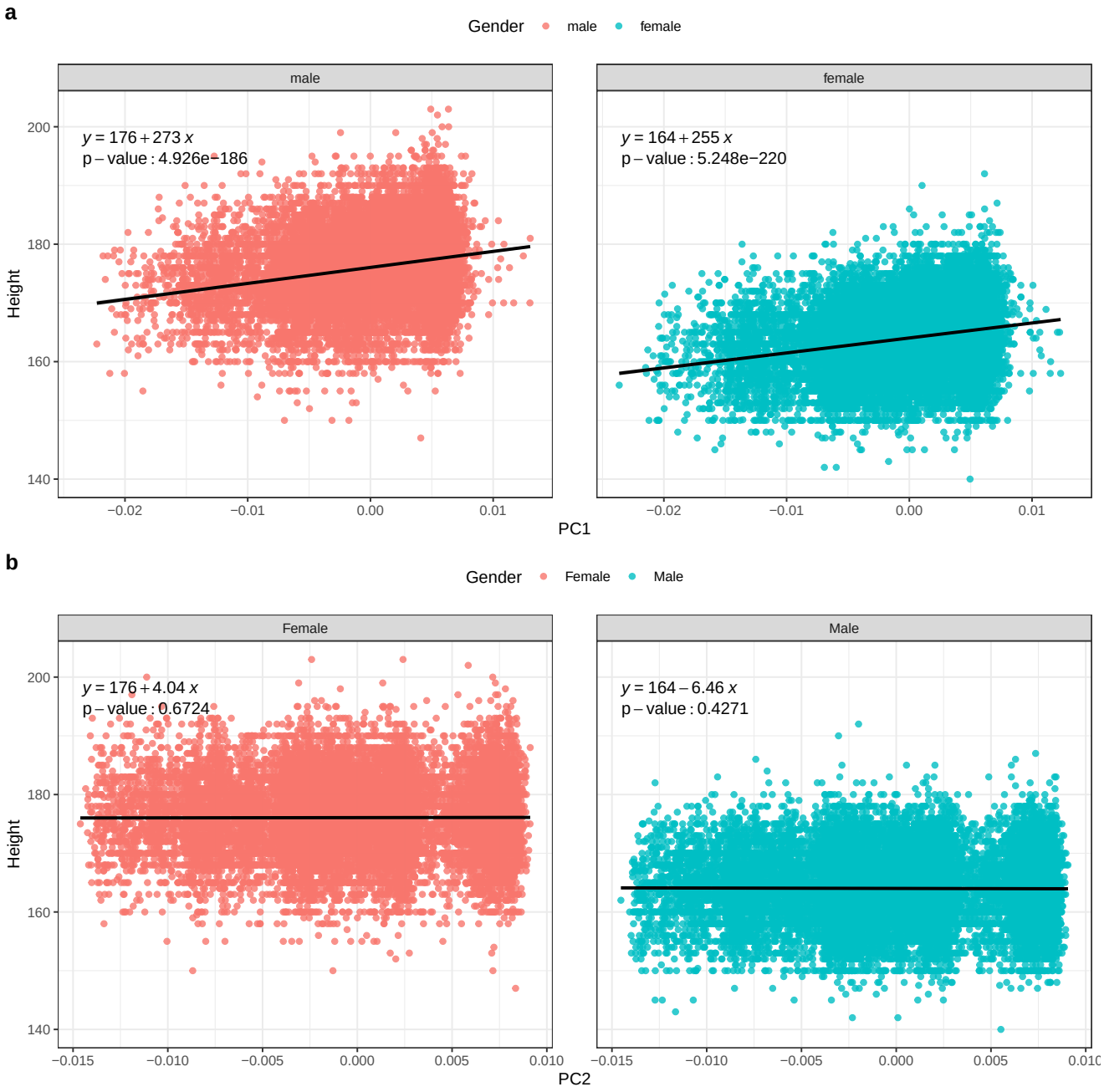

Figure S4: The relationship between height and first two principal components in Chinese dataset. (**a**) Height against the first principal component grouped by sex; (**b**) Height against the second principal component grouped by sex. The black lines represent the fitted regression between height and corresponding PCs. There is an increasing trend of height along the gradient of the first PC.

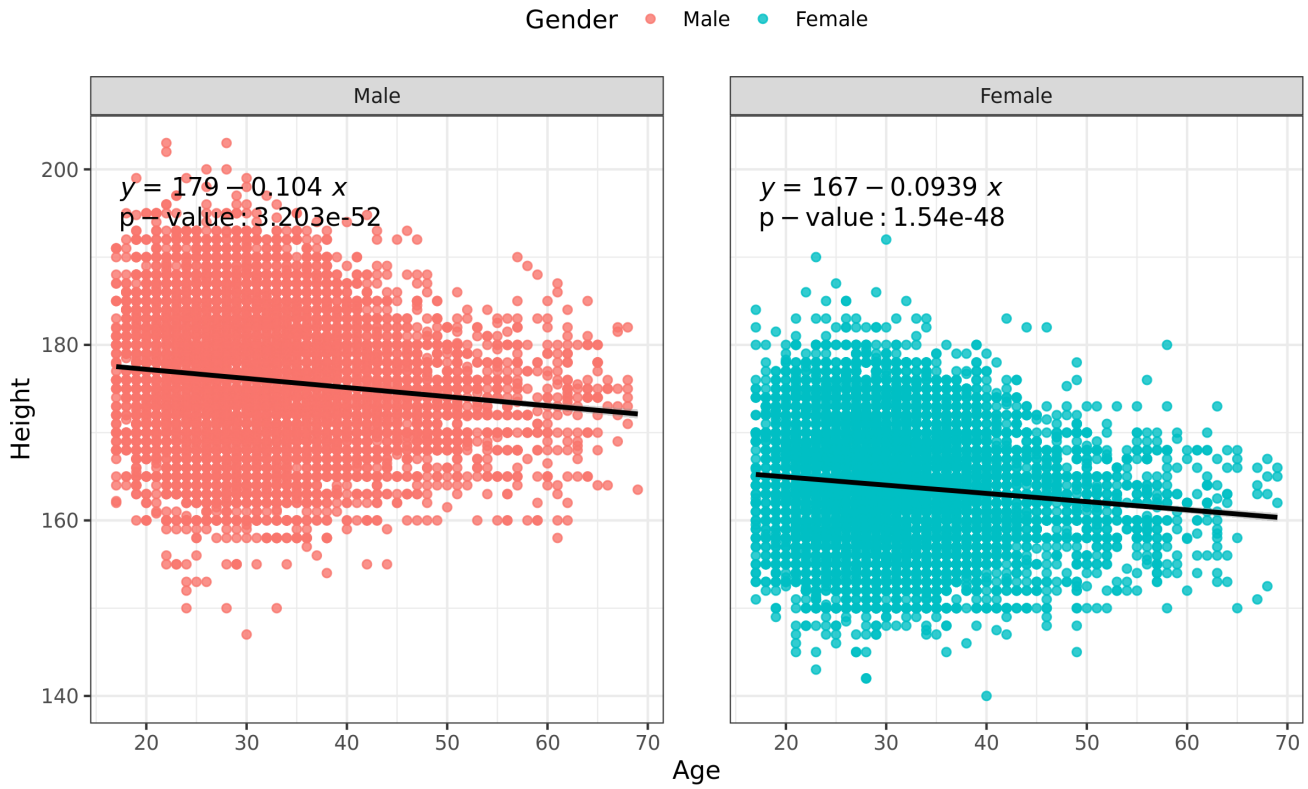

Figure S5: The relationship between height and age in Chinese dataset. (**a**) Height against age for males; (**b**) Height against age for females. The black lines represent the fitted regression between height and age. There is an decreasing trend of height for older people.

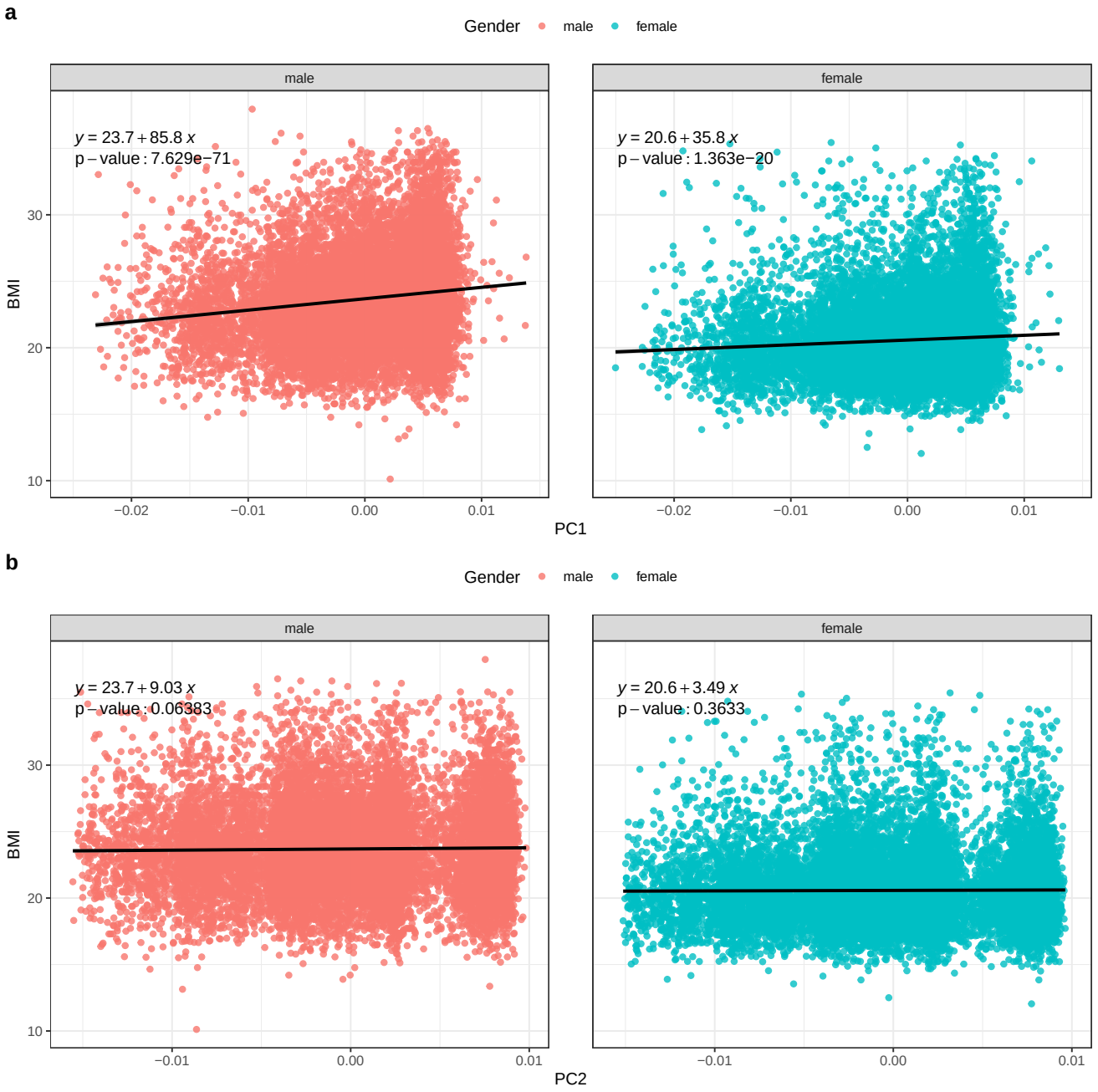

Figure S6: The relationship between BMI and first two principal components in Chinese dataset. (**a**) BMI against the first principal component grouped by sex; (**b**) BMI against the second principal component grouped by sex. The black lines represent the fitted regression between BMI and corresponding PCs. There is an increasing trend of BMI along the gradient of the first PC.

Figure S7: The relationship between BMI and age in Chinese dataset. (**a**) BMI against age for males; (**b**) BMI against age for females. The black lines represent the fitted regression between BMI and age. There is an increasing trend of BMI for older people.

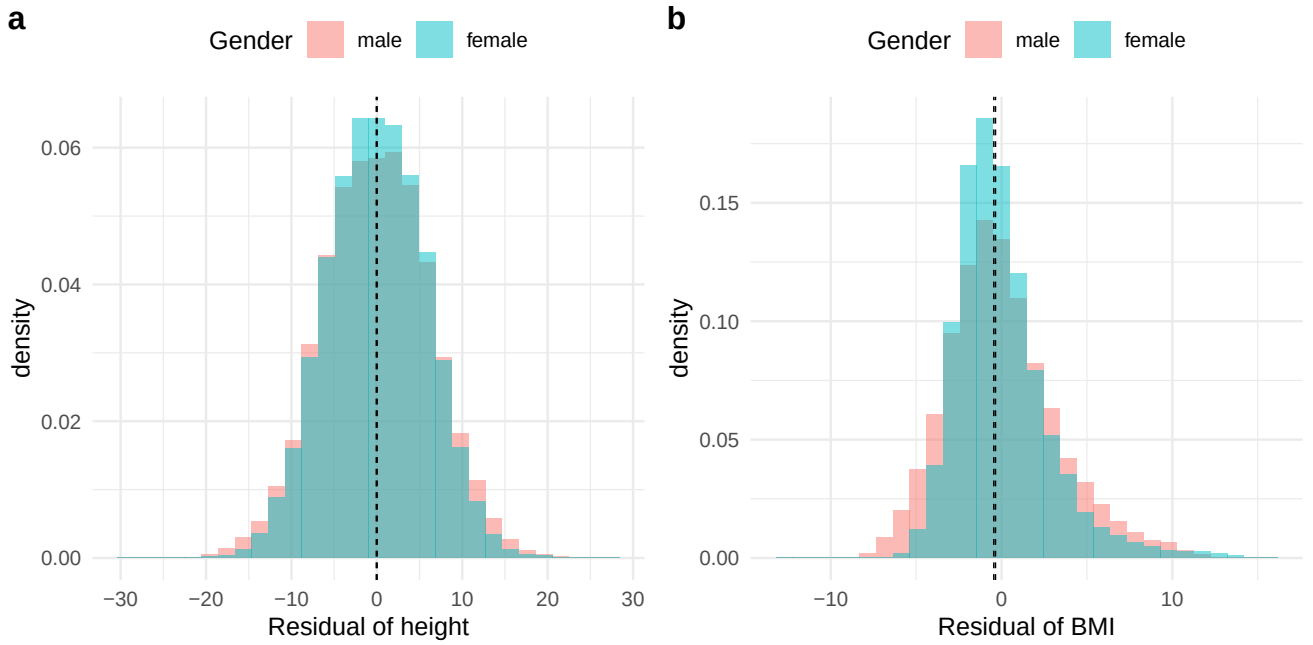

Figure S8: The distribution of residuals of height (**a**) and BMI (**b**) after adjusting for covariates in Chinese dataset.

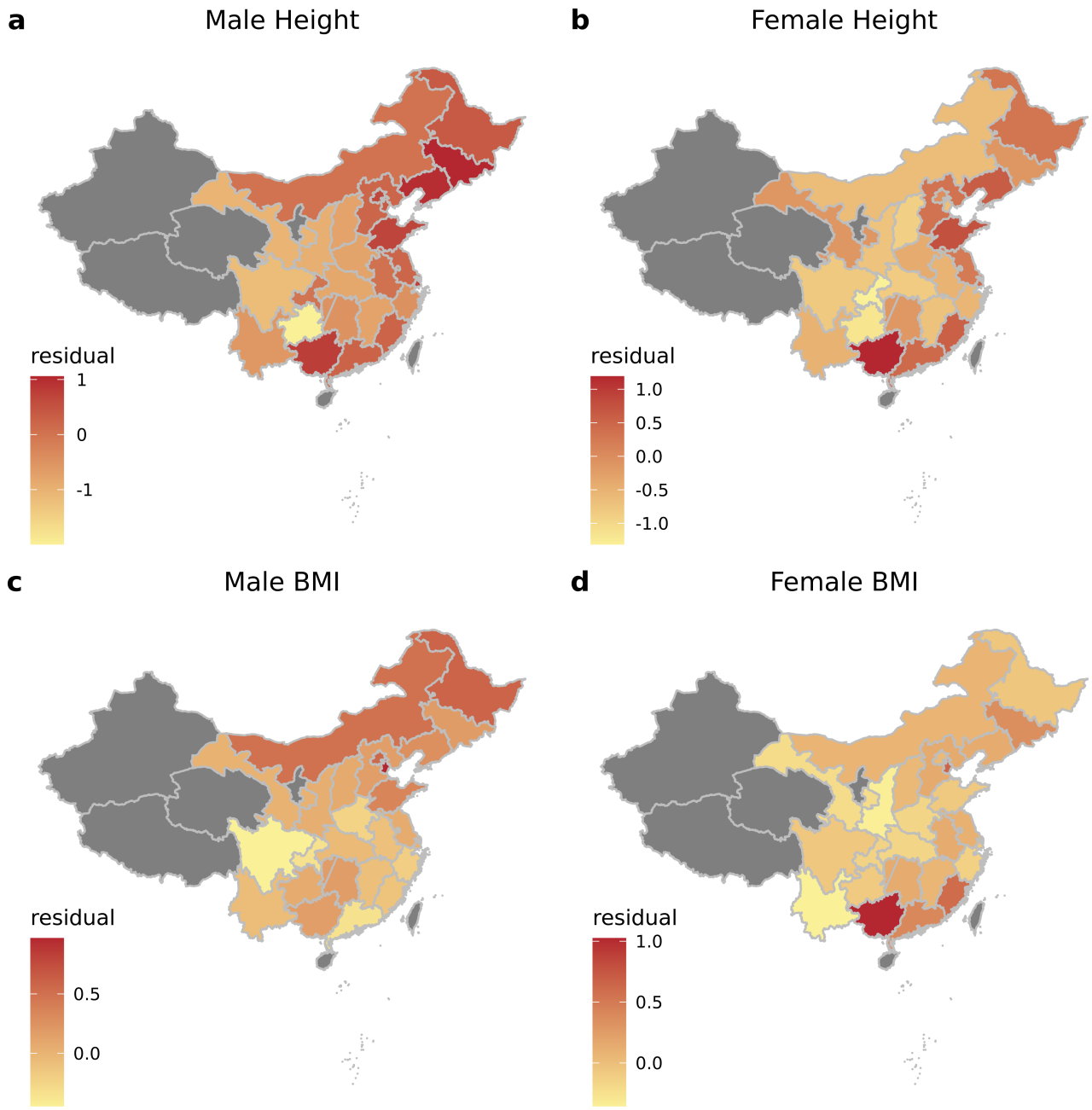

Figure S9: Average residual values after adjusting for the covariates of male height (**a**), female height (**b**), male BMI (**c**) and female BMI (**d**) in the provinces with more than 50 samples.

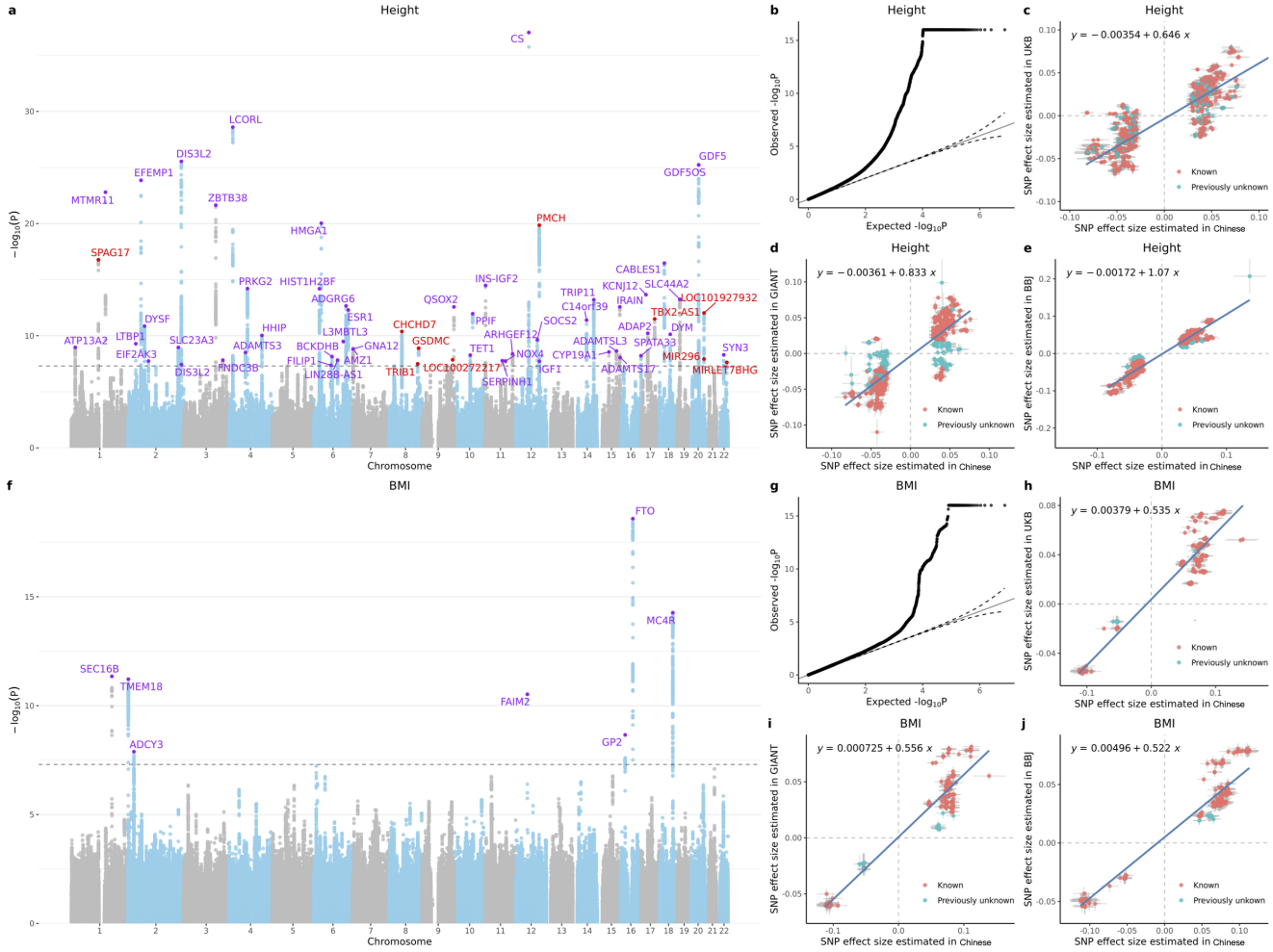

Figure S10: Manhattan plots of Chinese height (**a**) and BMI (**f**). The x axis shows chromosomal position, and the y axis shows significance on the $-\log_{10}$ scale. The dashed line marks the threshold for genome-wide significance (p-value$= 5 \times 10^{-8}$). Previously unknown associations are highlighted with purple dots, with the nearest gene names printed in purple. Known associations are highlighted with red dots, with the nearest gene names in red text. QQ plots of Chinese height (b) and BMI (f). **c-e** Comparison of the effect sizes for the genome-wide significant SNPs identified from the GWAS of Chinese height versus those identified in previous studies. **h-j** Comparison of the effect sizes for the genome-wide significant SNPs identified from the GWAS of Chinese BMI versus those identified in previous studies.

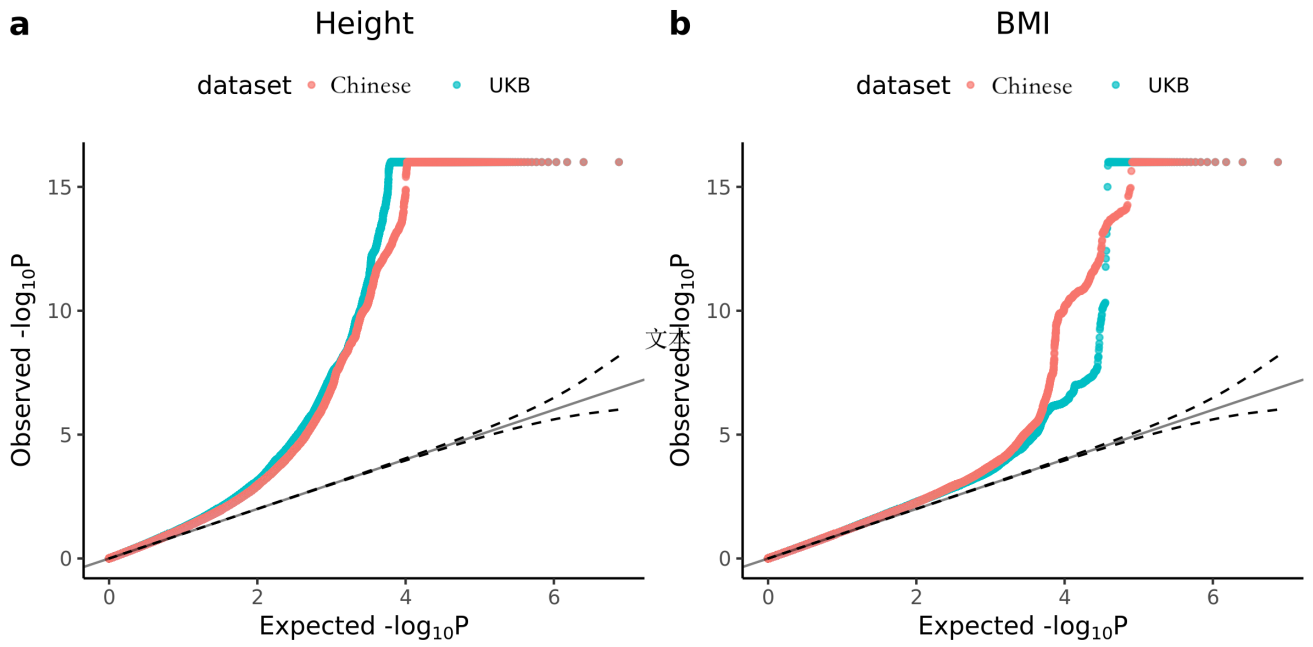

Figure S11: The Q-Q plot of GWAS p-values in height (**a**) and BMI (**b**) derived from Chinese dataset and 33,000 UKBB samples. We used the BOLT-LMM v2.3.2 to test for associations between phenotypes and SNPs. For Chinese population, we included age, sex and first 10 principal components as covariates. For UKBB, we used the top 20 principal components, age, squared age, sex, genotyping arrays and sequencing platforms as covariates.

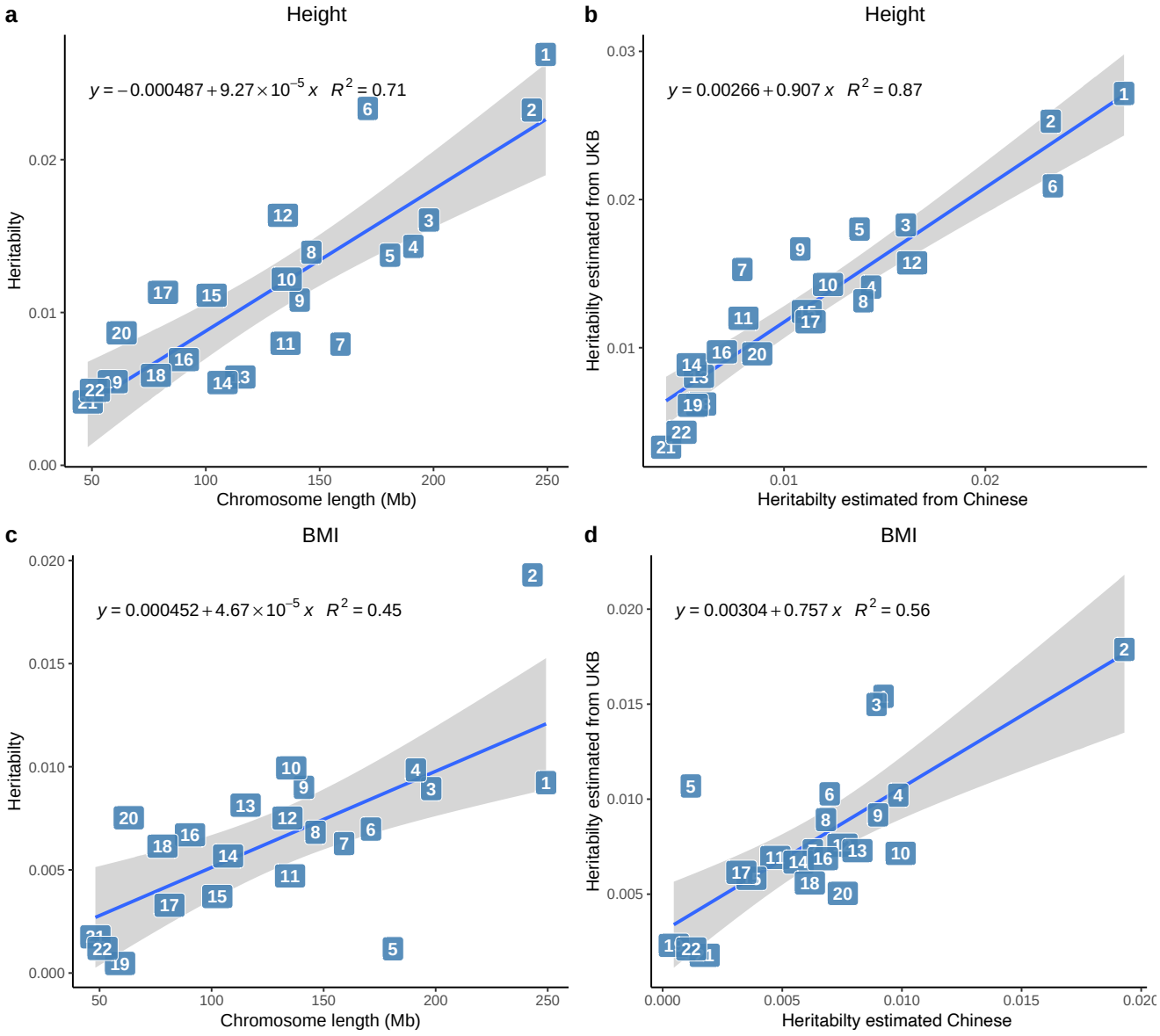

Figure S12: (**a**) The heritability of height against chromosome length in million base pair. (**b**) The chromosome heritabilities of height estimated from Chinese cohort against those estimated from UKBB. (**c**) The heritability of BMI against chromosome length in million base pair. (**d**) The chromosome heritabilities of BMI estimated from Chinese cohort against those estimated from UKBB.

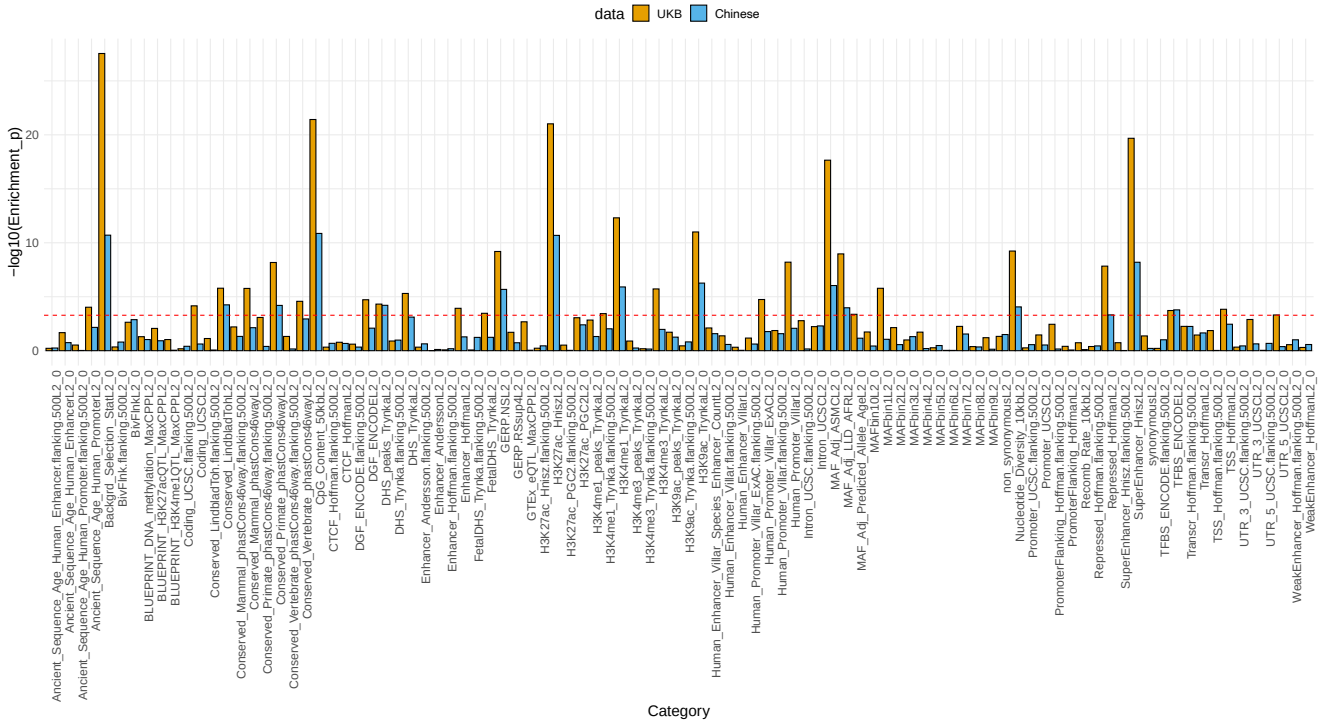

Figure S13: Enrichment of heritability for height in 95 functional annotations. The dashed line represents the significance threshold after Bonferroni correction (0.05/95). The LDSC software v1.0.0 was used to identify the heritability enrichment for the genome partitions in baseline model [1]. We used the LD scores provided by the Alkes Price group (`https://data.broadinstitute.org/alkesgroup/LDSCORE/`) in the analyses of Chinese cohort and UKBB.

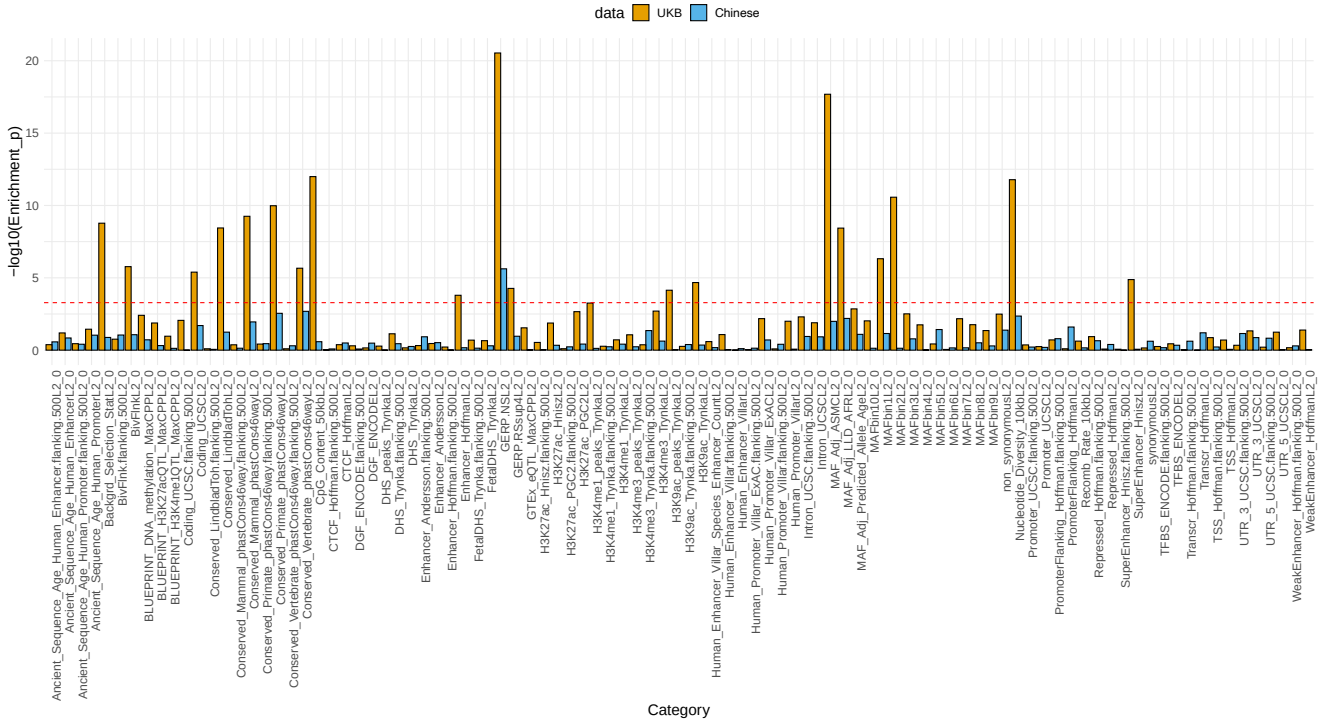

Figure S14: Enrichment of heritability for BMI in 95 functional annotations. The dashed line represents the significance threshold after Bonferroni correction (0.05/95).

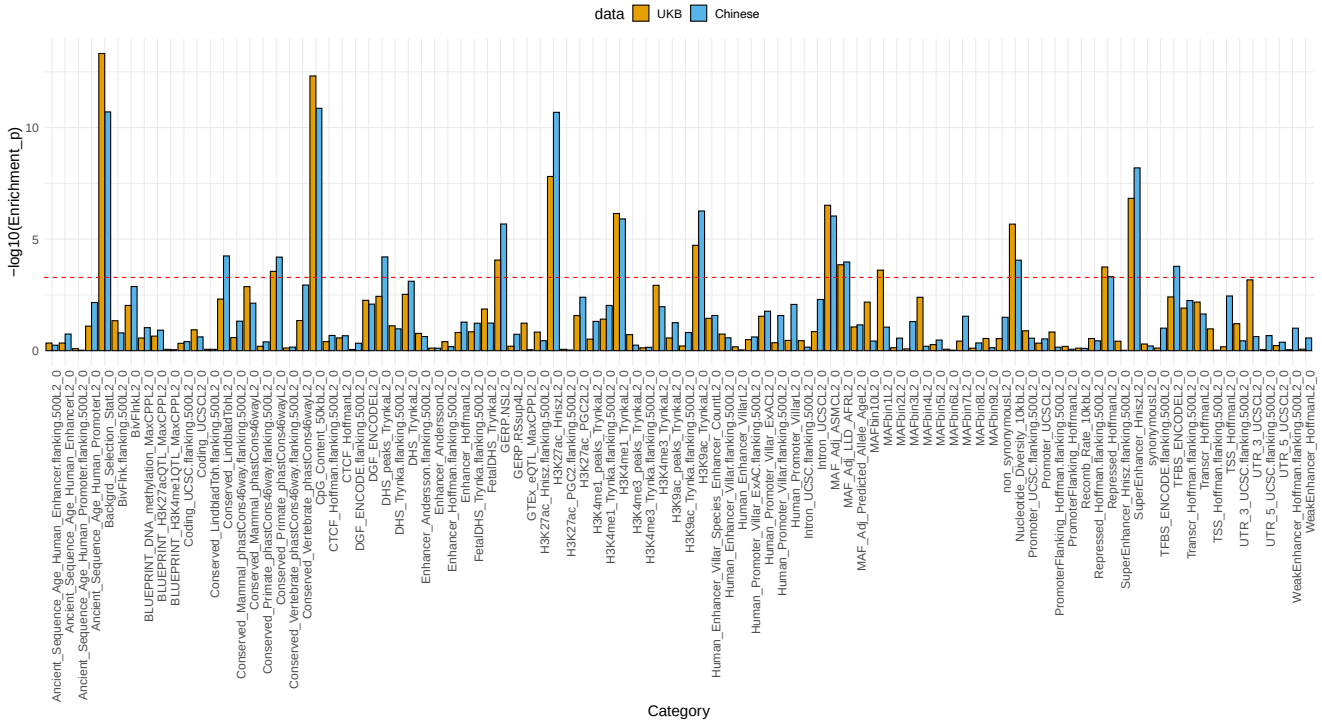

Figure S15: Enrichment of heritability for height in 95 functional annotations. We have randomly subsampled 20,000 individuals from UKBB to make the sample size comparable with Chinese cohort. The dashed line represents the significance threshold after Bonferroni correction (0.05/95).

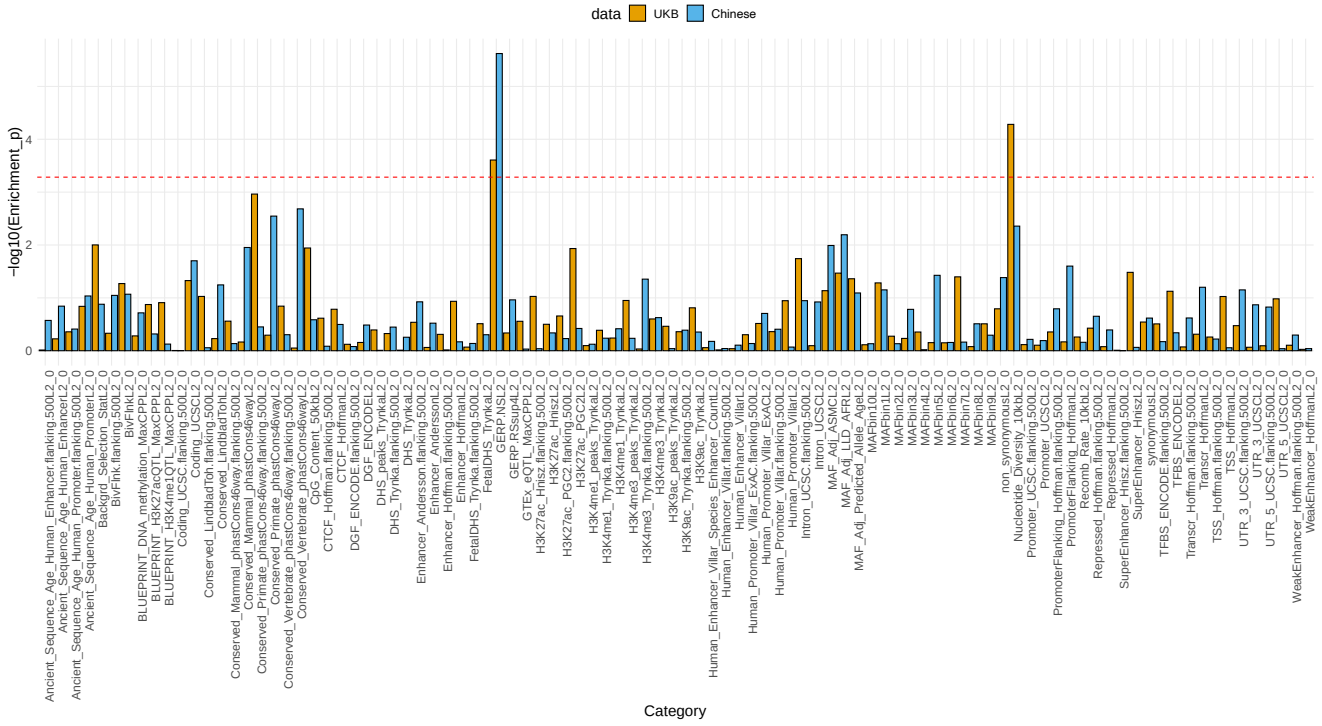

Figure S16: Enrichment of heritability for BMI in 95 functional annotations. We have randomly subsampled 20,000 individuals from UKBB to make the sample size comparable with Chinese cohort. The dashed line represents the significance threshold after Bonferroni correction (0.05/95).

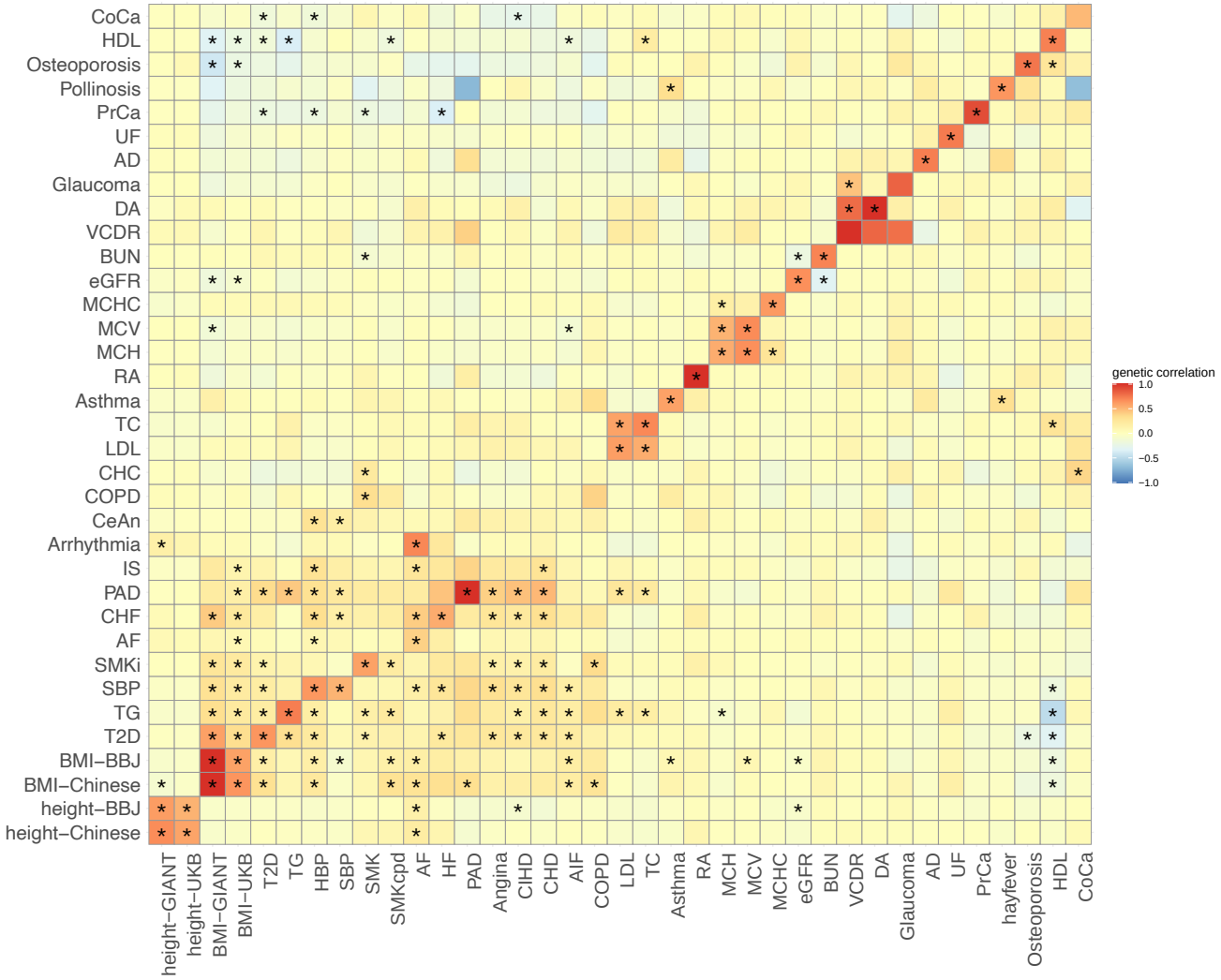

Figure S17: Trans-ancestry genetic correlations estimated by XPA. Positive correlations are colored in red. Negative correlations are colored in blue. Genetic correlations that are significantly different from zero after Bonferroni correction for the 1295 tests (p-value$< 0.05/1295$) are marked with asterisk.

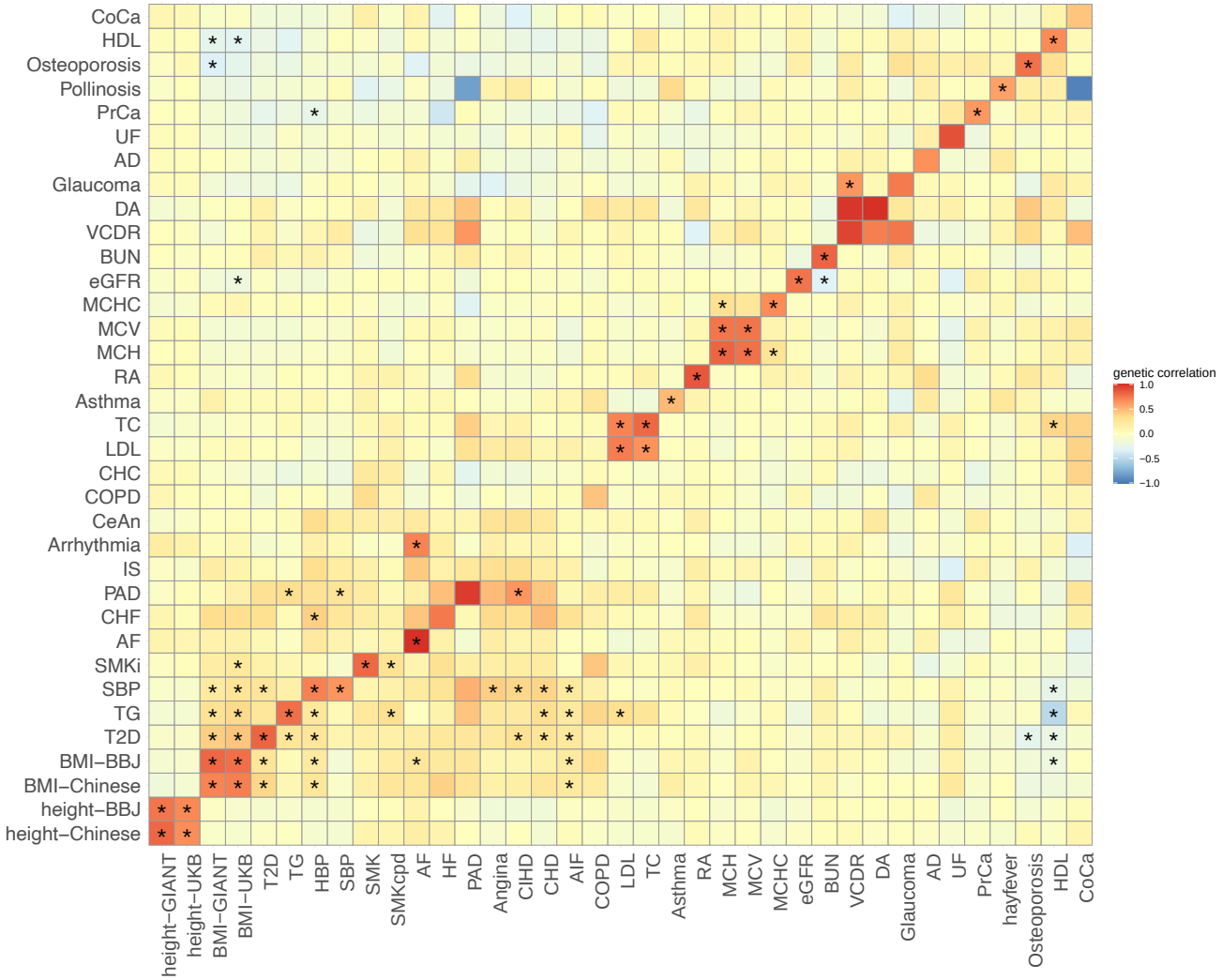

Figure S18: Trans-ancestry genetic correlations estimated by popcorn. Positive correlations are colored in red. Negative correlations are colored in blue. Genetic correlations that are significantly different from zero after Bonferroni correction for the $37 \times 35 = 1295$ tests (p-value$< 0.05/1295$) are marked with asterisk.

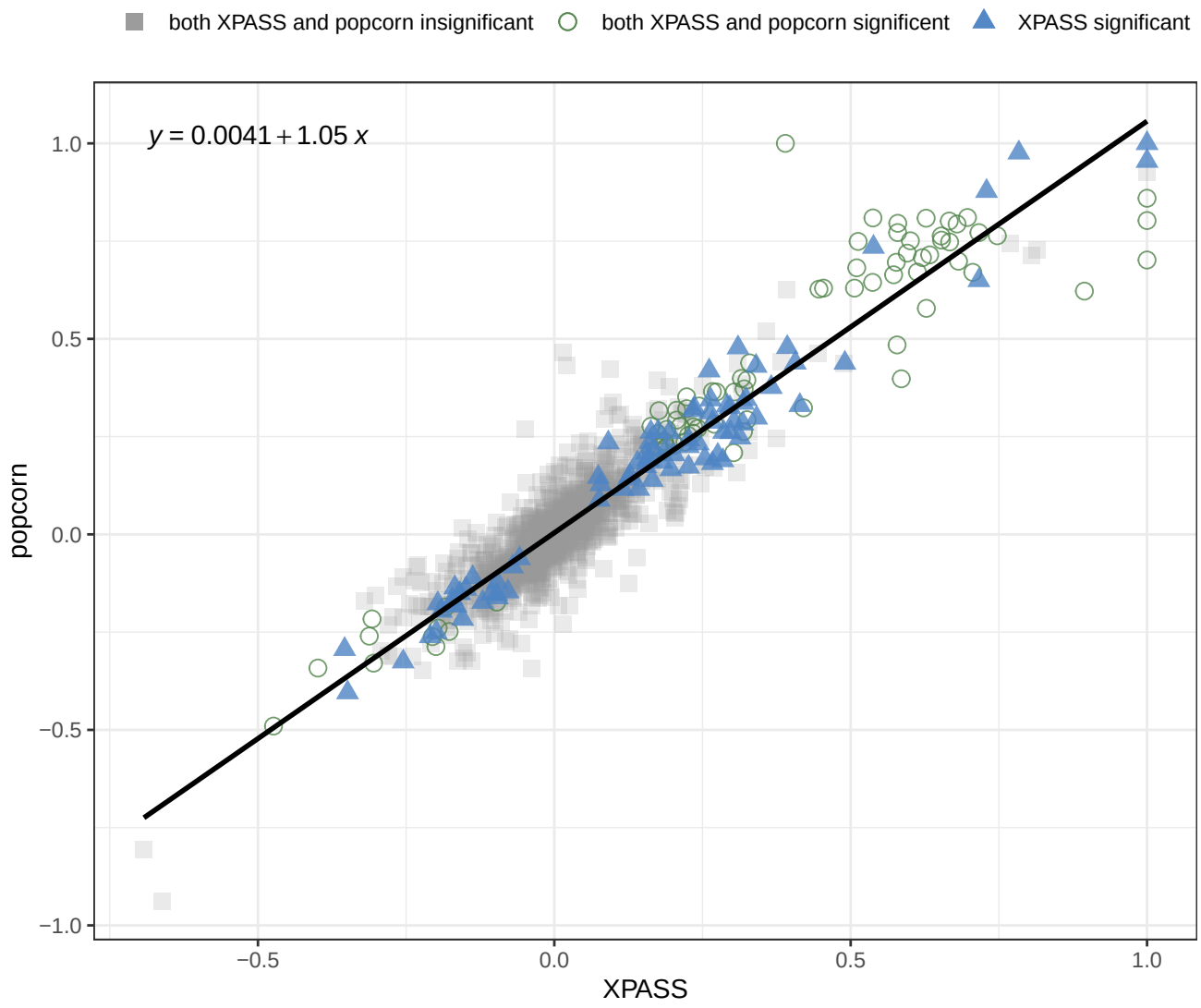

Figure S19: Genetic correlation estimates generated by popcorn versus those generated by XPASS. A regression line between the two sets of estimates is added.

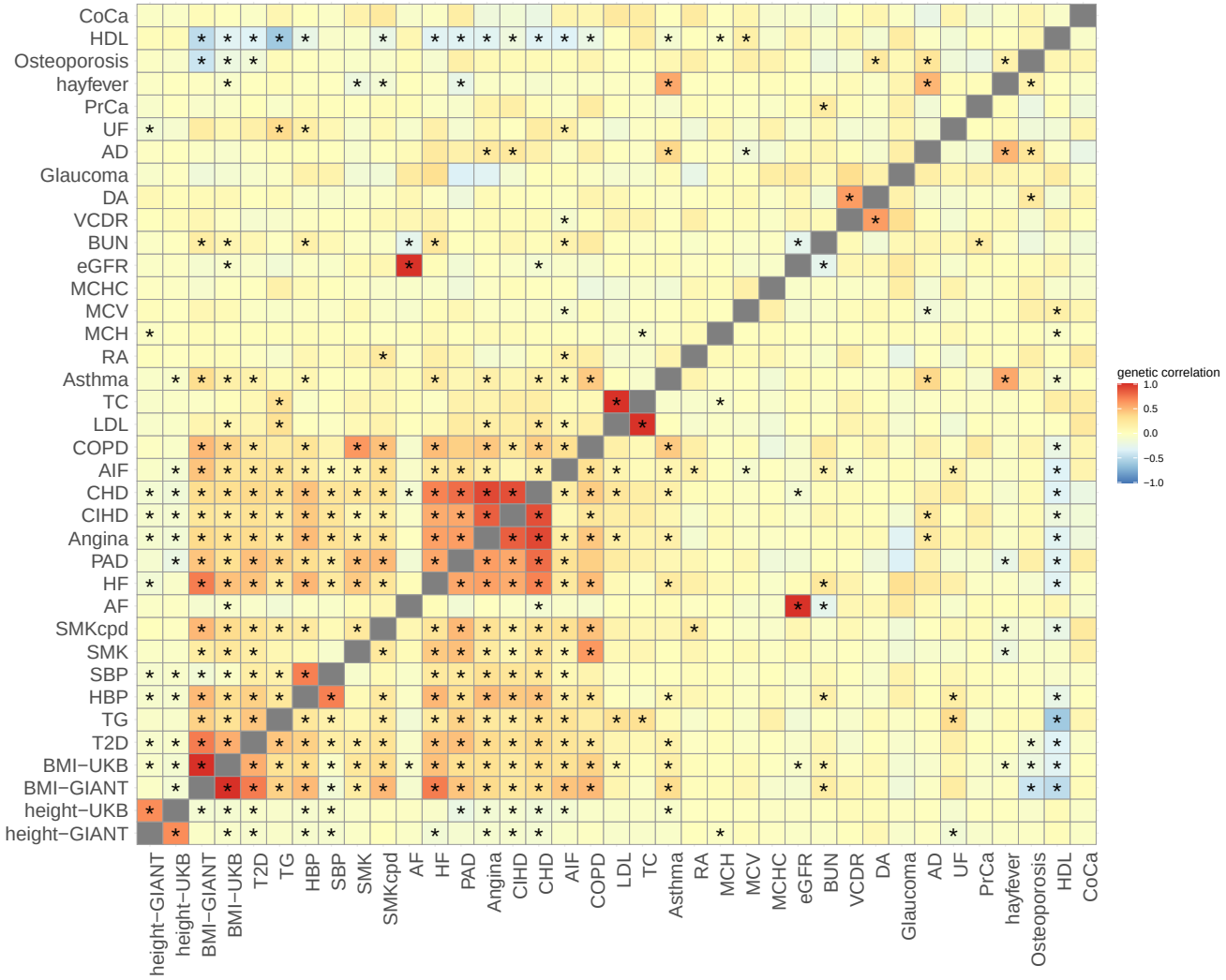

Figure S20: Genetic correlation of 37 traits in Europeans estimated by GNOVA. Positive correlations are colored in red. Negative correlations are colored in blue. Genetic correlations that are significantly different from zero after Bonferroni correction for the $(37 \times 36/2) = 666$ tests (p-value$< 0.05/666$) are marked with asterisk.

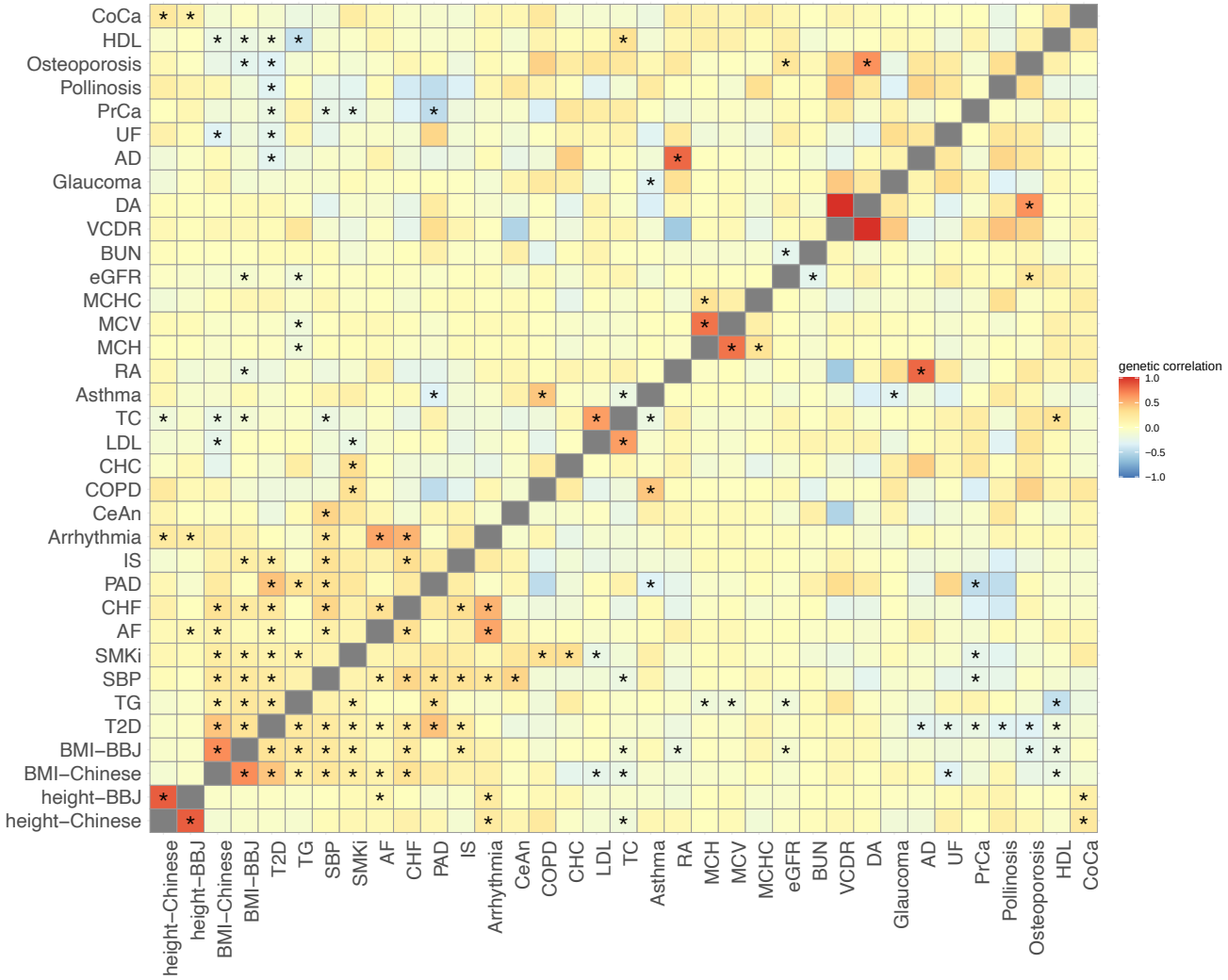

Figure S21: Genetic correlations of 35 traits in East Asians estimated by GNOVA. Positive correlations are colored in red. Negative correlations are colored in blue. Genetic correlations that are significantly different from zero after Bonferroni correction for the $(35 \times 34/2) = 595$ tests (p-value$< 0.05/595$) are marked with asterisk.

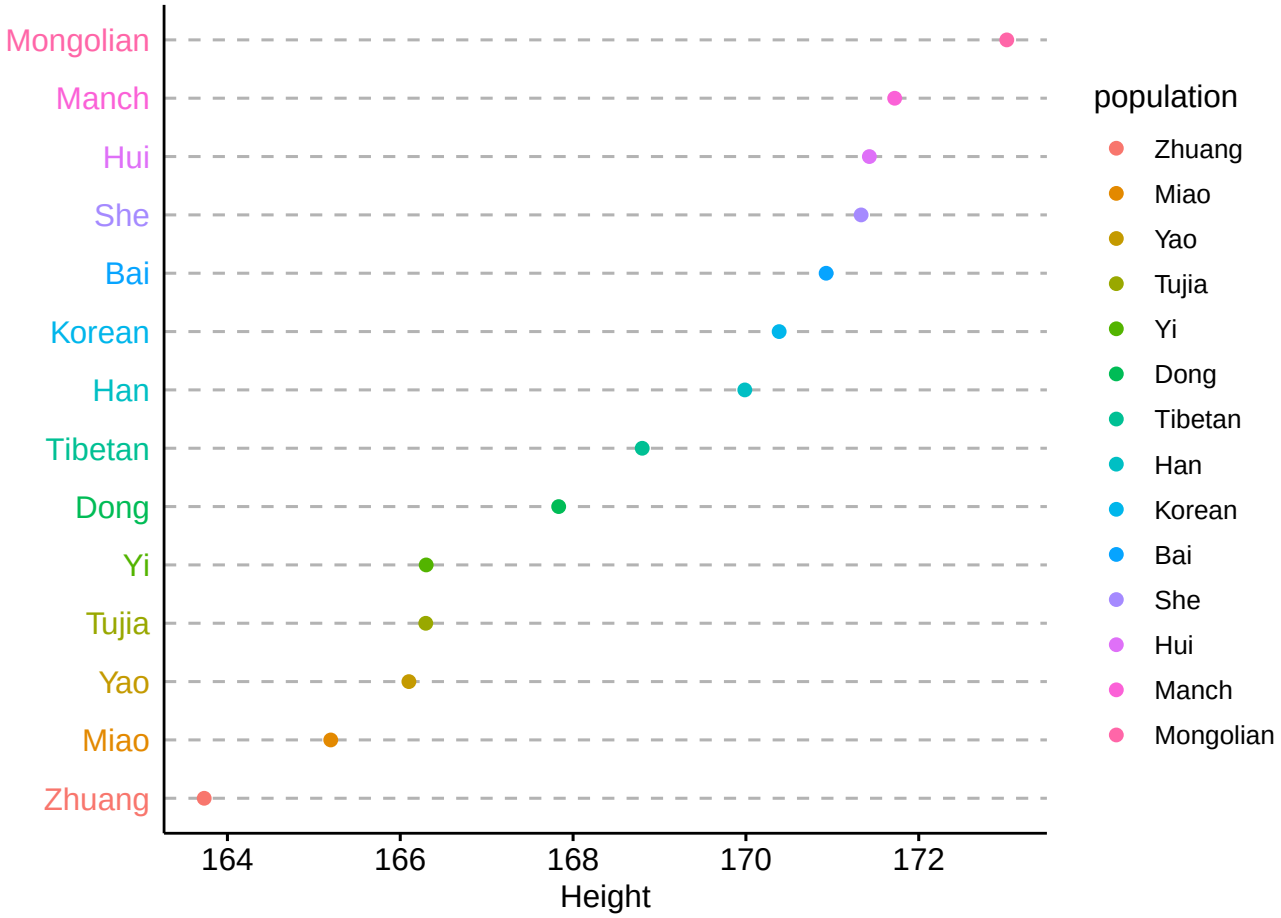

Figure S22: The average height among minority ethnic groups with ≥ 50 samples.

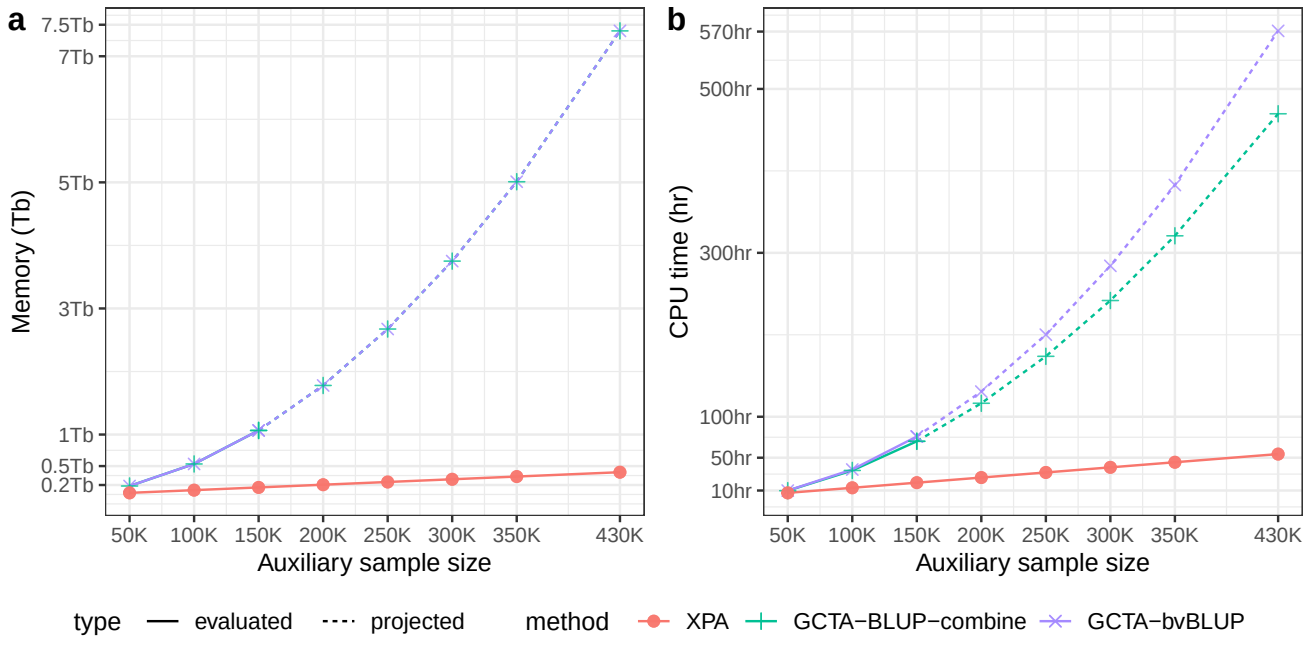

Figure S23: Memory usage (a) and CPU time (b) for XPA, GCTA-BLUP-combine, and GCTA-bvBLUP are shown for increasing auxiliary sample sizes when combining Chinese cohort and UKBB data to construct PRS for height. XPA used only 54.5 hours (including 9 hours for loading data, 3 hours for estimating variance components, and 42.5 hours for computing the posterior means and estimating fixed effects) and 385Gb to analyze all 430K Chinese and UKBB samples. In contrast, GCTA-bvBLUP required 1.07Tb when only 150K UKBB samples were included in the analysis, reaching the memory limit of our server. We note that the memory requirement exceeding this value is also infeasible for most high performance computational platforms. Therefore, we projected its CPU time and memory by fitting a quadratic curve using the recorded values. Our projection suggests that it would cost 570.8 hours and 7.5 Tb memory for GCTA-bvBLUP to integrate all 430K UKBB samples. Note that the memory of a node is often about 512Gb at Yale high-performance computing server, and the maximum memory of a node at the Hong Kong University of Science and Technology is about 1.5 Tb. Given above observations, we believe that XPA has advantage over GCTA-bvBLUP in practice as it can leverage the bio-bank scale dataset from the European population to construct more accurate PRS in minor populations. Both computational time and memory cost of XPA were linear to the auxiliary sample size, which was consistent with our observations in the simulation study. We evaluated all approaches with 32 CPU threads on the platform of Intel Xeon Gold 6152 CPU.

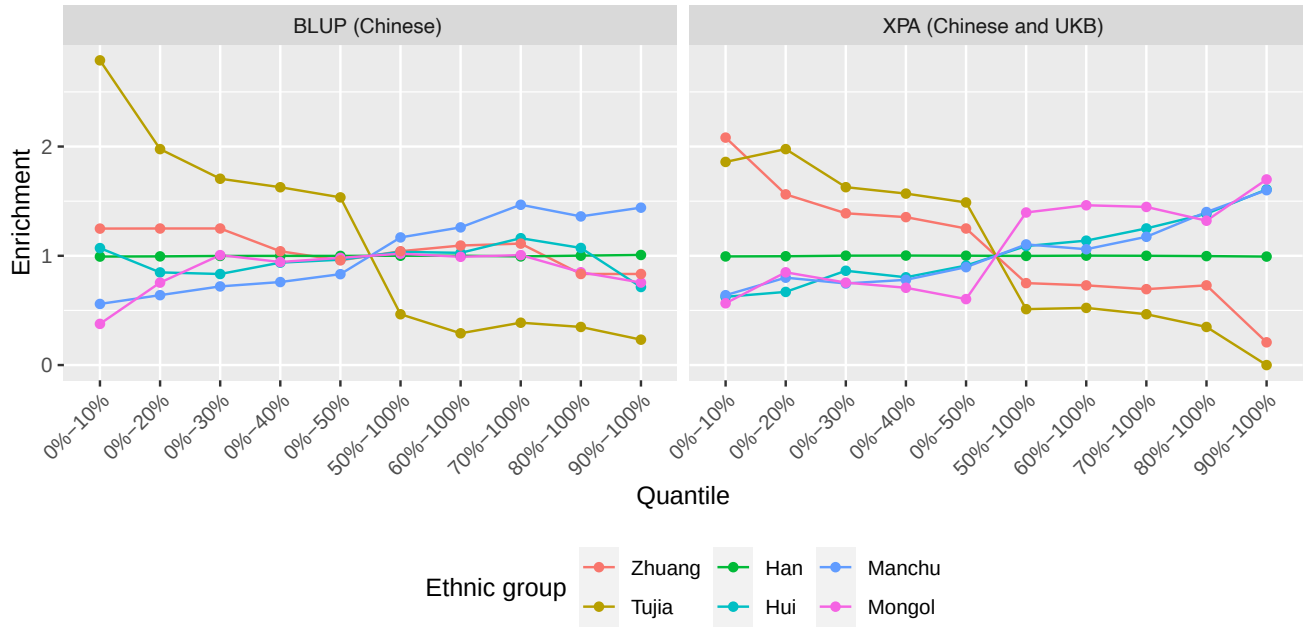

Figure S24: Proportion enrichment of ethnic groups in the top and bottom PRS quantiles. XPA successfully prioritizes the heights of the five minor ethnic groups with more than 50 samples in the test set, whilst BLUP can only predict Tujia and Manchu.

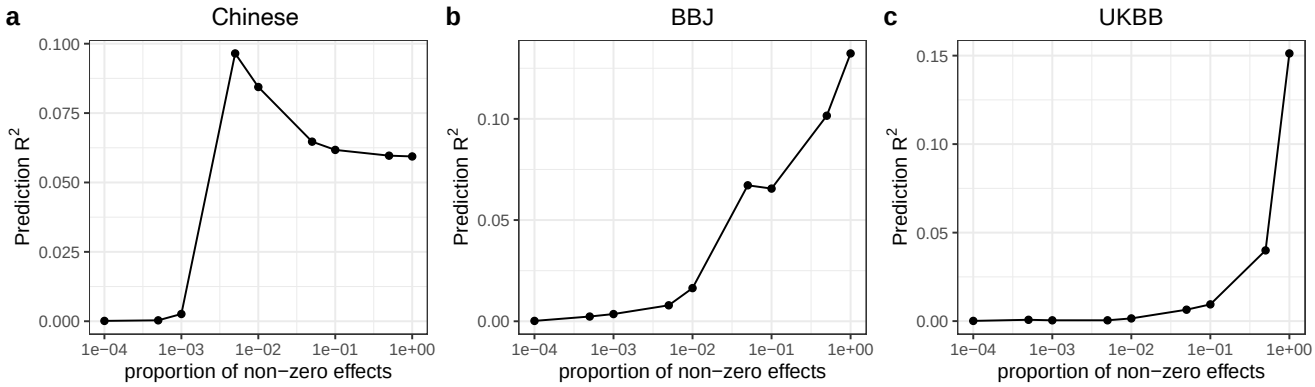

Figure S25: Tuning the proportion of non-zero effects in LDpred for height: (a) Chinese, (b) BBJ and (c) UKBB.

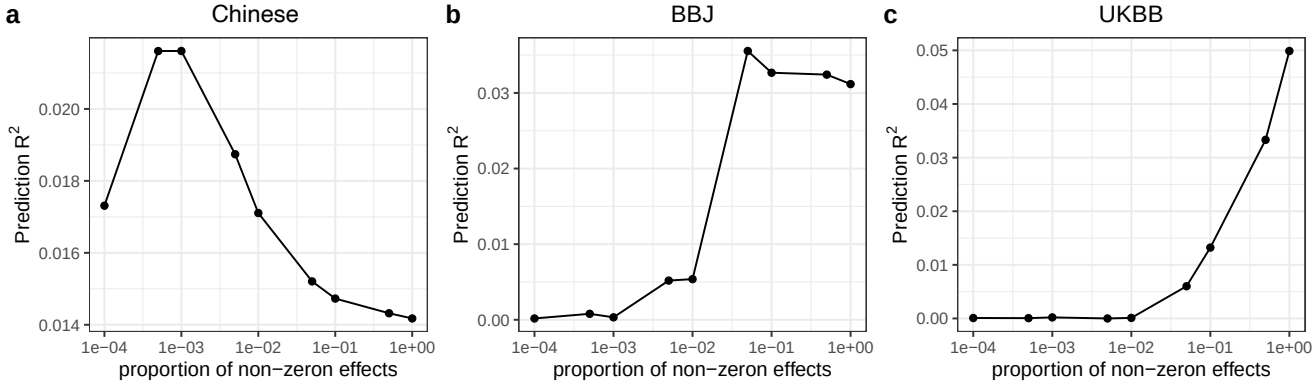

Figure S26: Tuning the proportion of non-zero effects in LDpred for BMI: (a) Chinese, (b) BBJ and (c) UKBB.

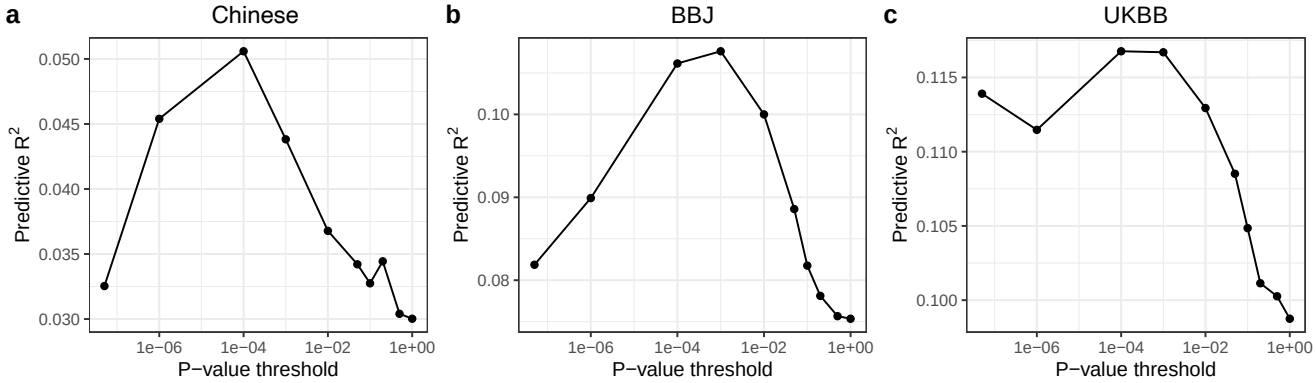

Figure S27: Tuning the p-value threshold for height: (a) Chinese, (b) BBJ and (c) UKBB.

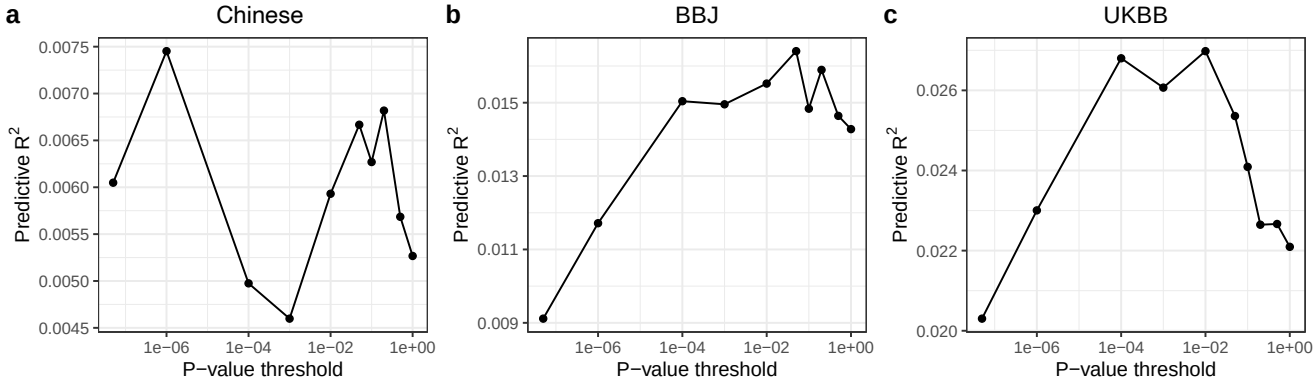

Figure S28: Tuning the p-value threshold for BMI: (a) Chinese, (b) BBJ and (c) UKBB.

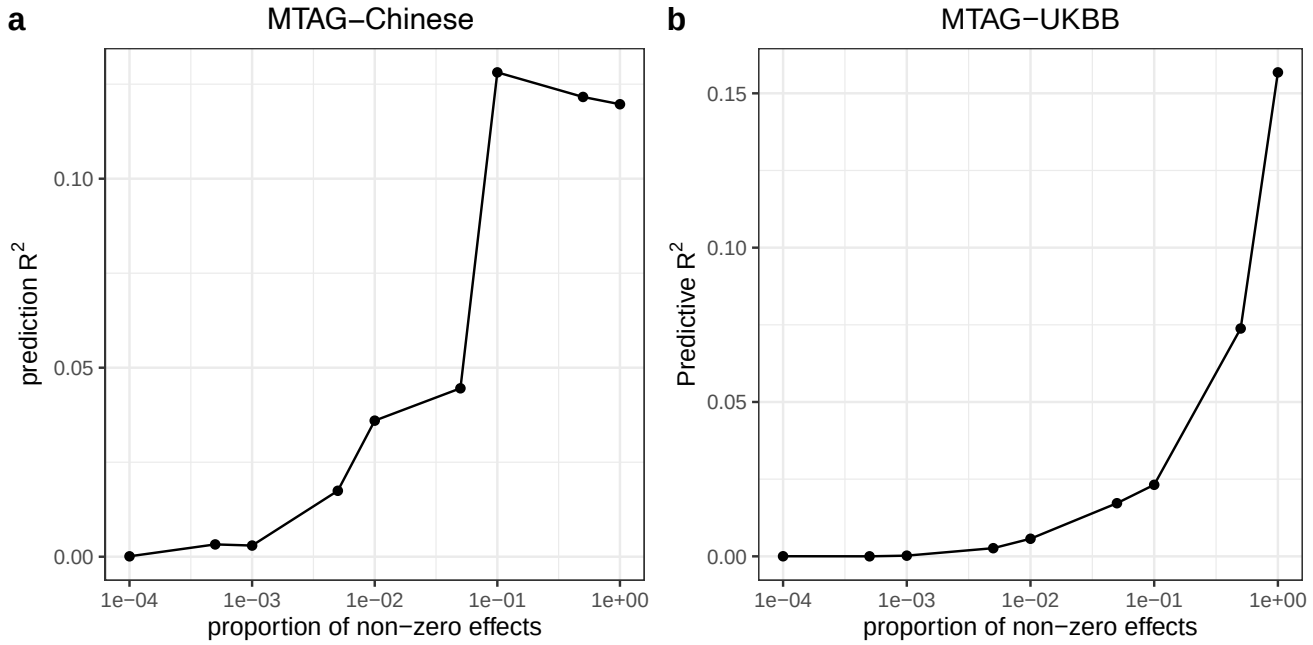

Figure S29: Tuning the proportion of non-zero effects in LDpred for height: (a) MTAG-Chinese, (b) MTAG-UKBB.

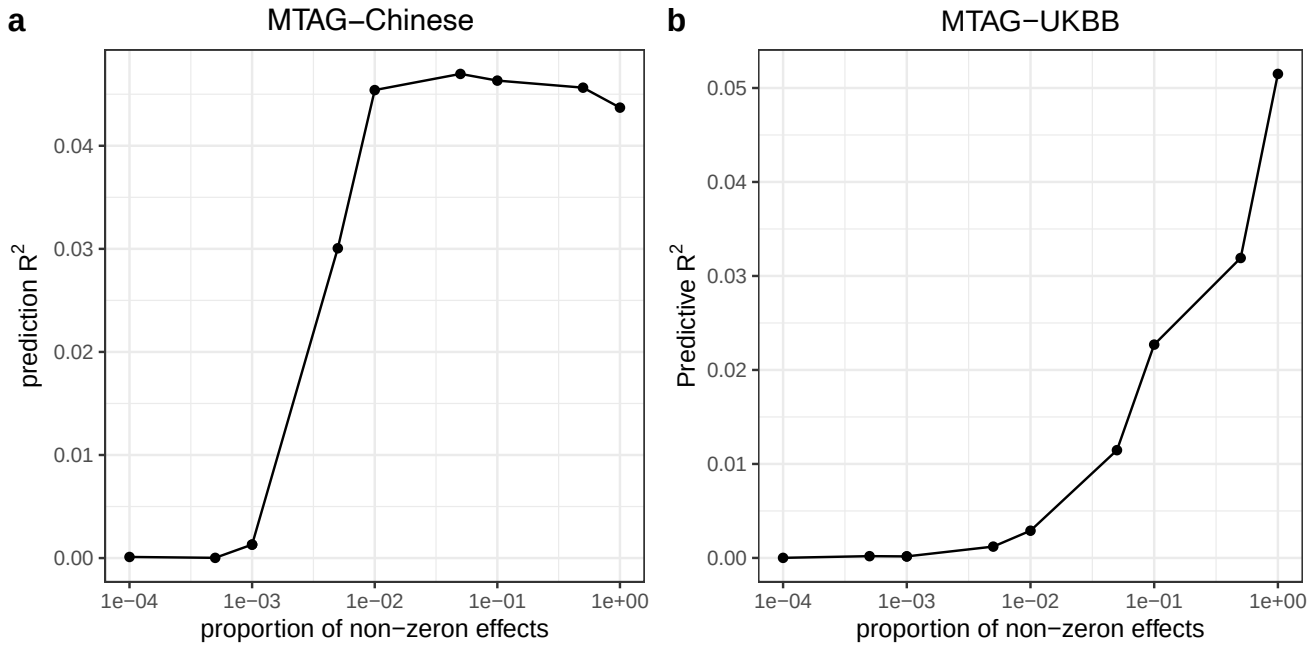

Figure S30: Tuning the proportion of non-zero effects in LDpred for BMI: (a) MTAG-Chinese, (b) MTAG-UKBB.

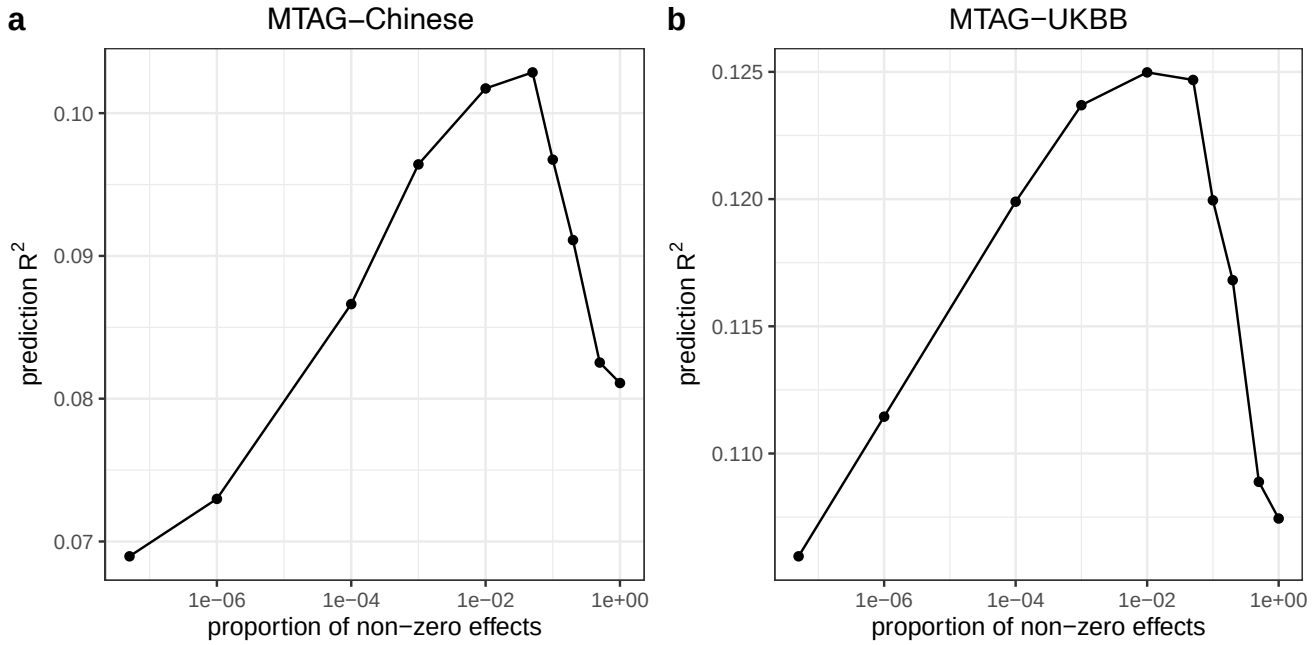

Figure S31: Tuning the p-value threshold for height: (a) MTAG-Chinese, (b) MTAG-UKBB.

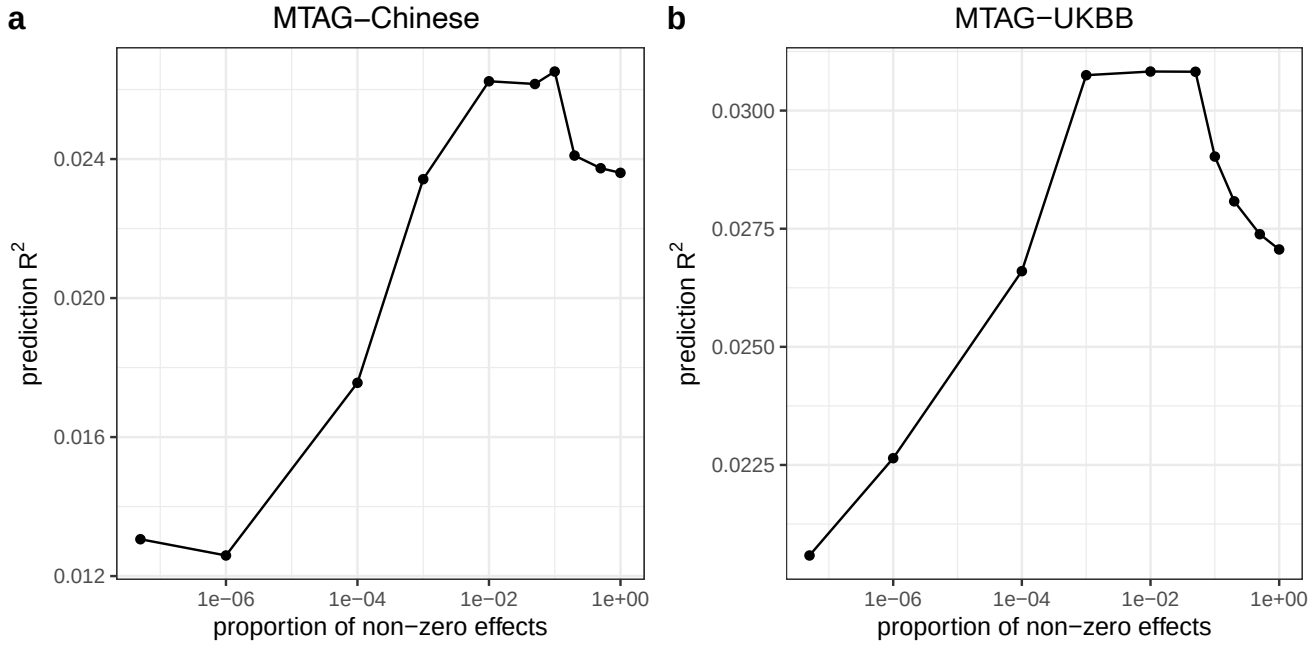

Figure S32: Tuning the p-value threshold for BMI: (a) MTAG-Chinese, (b) MTAG-UKBB.

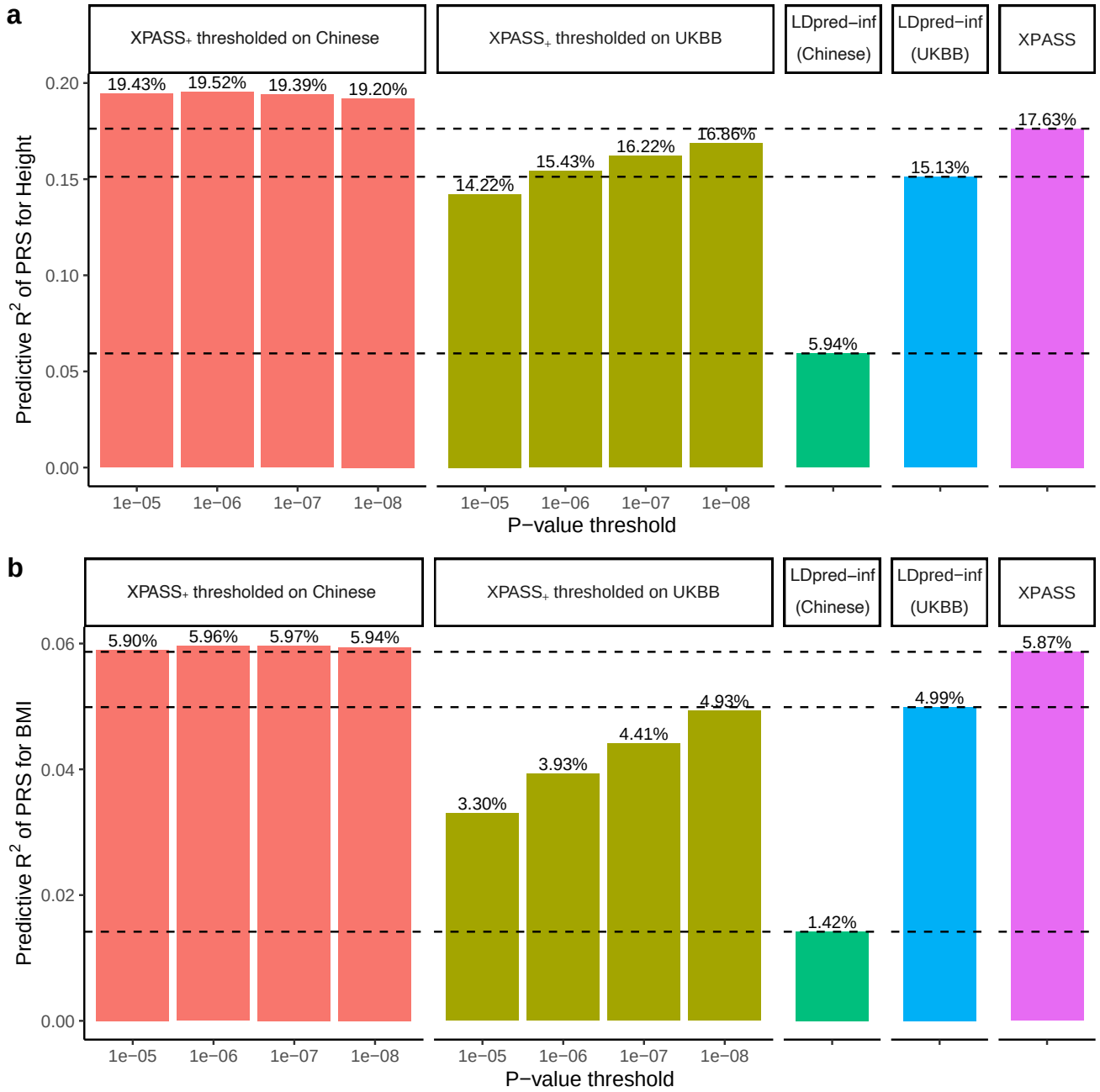

Figure S33: Predictive R^2 of height (a) and BMI (b) when XPASS$_+$ was applied with different p-value thresholds. The SNPs to be included in the covariates were selected by applying the p-value threshold to either the Chinese data or the UKBB data. The LD threshold was set as r2 = 0.1. The predictive R^2 of LDpred-inf (Chinese), LDpred-inf (UKBB) and XPASS were also shown as reference. When the P+T procedure was applied to the target dataset, including the selected SNPs as fixed effects further improved the prediction accuracy. In contrast, when the P+T procedure was applied to the auxiliary dataset, the predictive R^2 decreased as the p-value threshold increases. This observation suggests that when the pre-selected SNPs are specific to the target population, XPASS$_+$ can effectively utilize these signals to improve prediction accuracy.

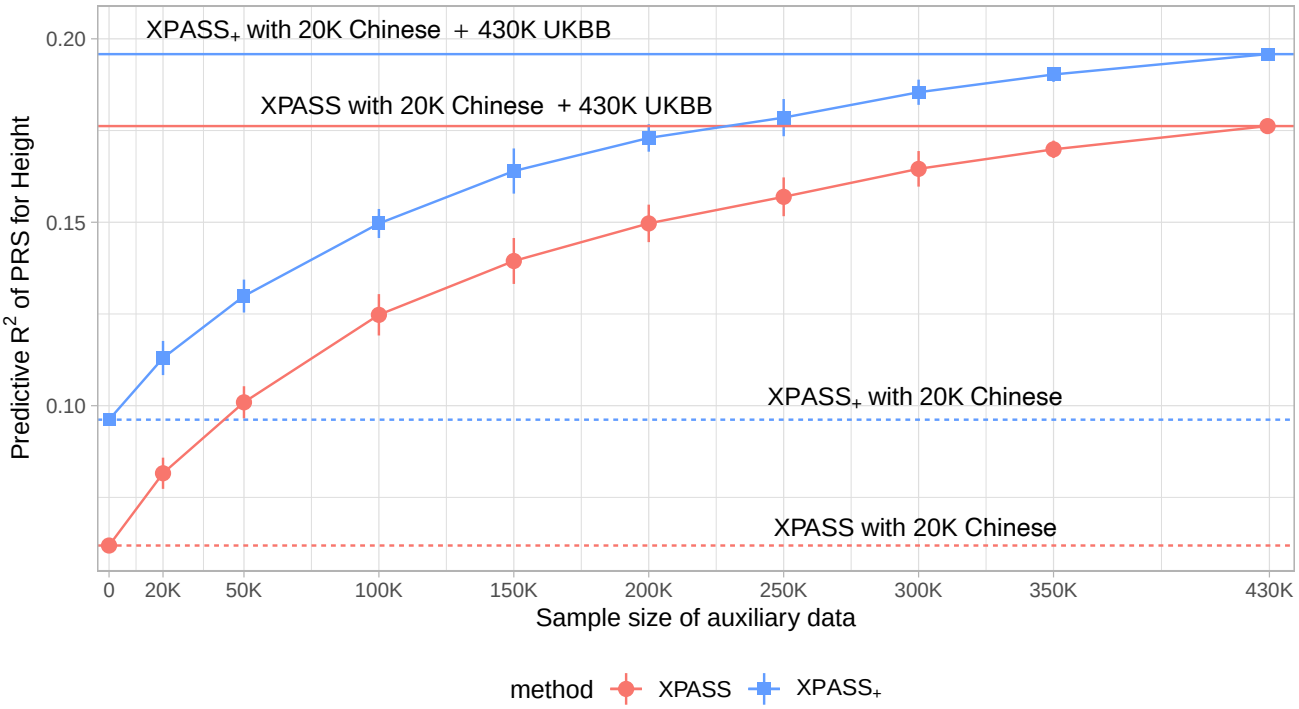

Figure S34: Influence of the auxiliary sample size on the prediction performance of XPASS and XPASS$_+$ for predicting height. We trained XPASS and XPASS$_+$ by integrating $21,069$ Chinese training samples with $20,000 \sim 300,000$ random subsamples drawn from UKBB, where samples from UKBB could be viewed as the auxiliary dataset. The results are summarized from 10 replications. Dashed horizontal lines mark the results obtained by training with only Chinese cohort using XPASS (red) or XPASS$_+$ (blue). Solid horizontal lines in mark the results obtained by combining 20K Chinese samples and all 430K UKBB samples using XPASS (red) or XPASS$_+$ (blue).

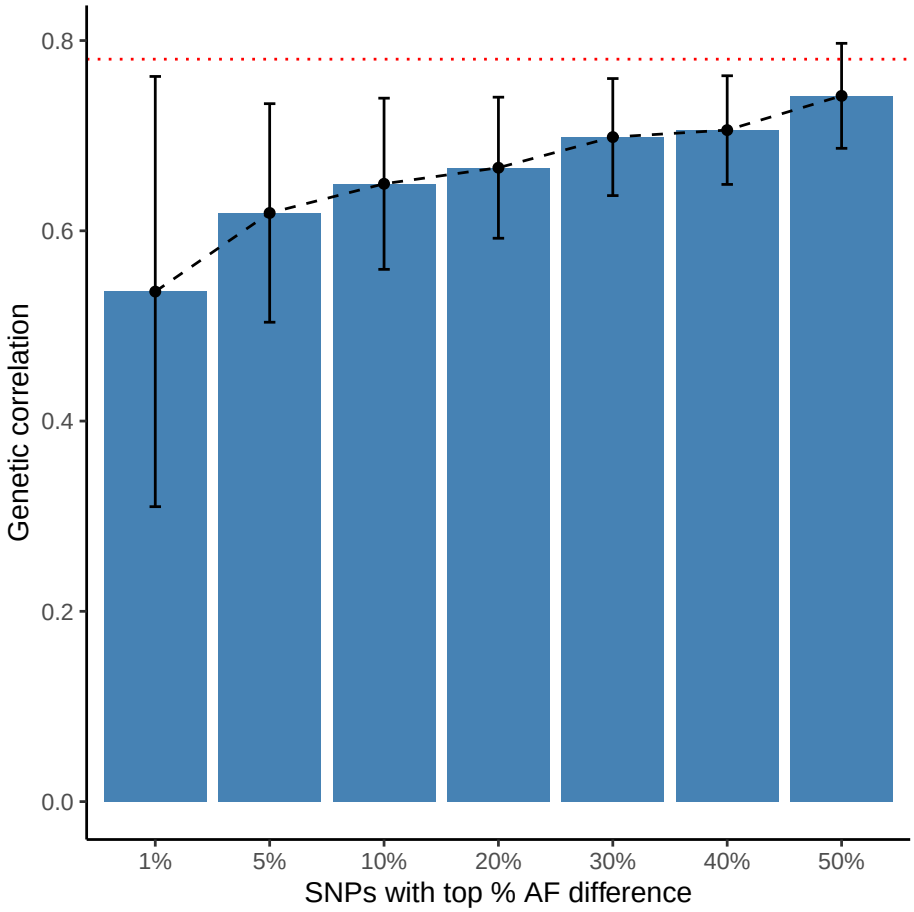

Figure S35: Genetic correlation of height estimated using the top 1% $\sim$ 50% SNPs sorted by the frequency (AF) difference between EAS and EUR. The AF difference is measured by $\frac{|f_1 - f2|}{\sqrt{2(f_1(1-f_1))+2(f_2(1-f_2))}}$. The red dotted line represents the genetic correlation estimated using all SNPs.

2 Supplementary Tables

		1%	5%	10%
Height	h^2_{1A}	0.747% (0.408%)	2.103% (0.721%)	4.244% (1.086%)
	Enrichment	2.135 (1.147)	1.205 (0.405)	1.213 (0.295)
	Predictive R^2	17.18%	16.87%	16.63%
BMI	h^2_{1A}	-0.301%	0.370% (0.466%)	1.395% (0.700%)
	Enrichment	-	0.451 (0.566)	0.845 (0.414)
	Predictive R^2	5.86%	5.68%	5.62%

Table S1: Application of the extended model to the Chinese and UKBB data. Estimated heritability and the enrichment of heritability explained by the 'heterogeneous' SNPs in Chinese samples are summarized in the table, with the corresponding standard errors given in the parentheses. The heritability explained by the 'heterogeneous' SNPs is computed as $h^2_{1A} = \hat{\sigma}^2_{1A}/(\hat{\sigma}^2_{1A} + \hat{\sigma}^2_{1B} + \hat{\sigma}^2_{\epsilon})$, and its enrichment is obtained as $(\hat{\sigma}^2_{1A}/p_A)/((\hat{\sigma}^2_{1A} + \hat{\sigma}^2_{1B})/p)$, where p_A and p are the number of SNPs in $\mathcal{A}$ and the total number of SNPs, respectively. The standard errors are obtained by applying the Jackknife approach with approximately independent LD blocks derived from the EAS population. Top 1%, 5% and 10% SNPs with highest diff_j were considered as 'heterogeneous' SNPs. The predictive R^2 were also computed for corresponding partition strategies.

Trait name	Full name	sample size (case+control)	paper link
RA-EAS	Rheumatoid Arthritis	4,873+17,642	https://www.nature.com/articles/nature12873?message-global=remove
RA-EUR	Rheumatoid Arthritis	14,361+43,923	https://www.nature.com/articles/nature12873?message-global=remove
T2D-EAS	Type 2 Diabetes	36,614+155,150	http://jenger.riken.jp:8080/pheno/Type_2_Diabetes
T2D-EUR	Type 2 Diabetes	459324	https://www.nature.com/articles/s41588-018-0144-6
BMI-BBJ	Body Mass Index	158284	https://www.nature.com/articles/ng.3951
BMI-Chinese	Body Mass Index	29147	http://www.sribd.cn/index.php/cn/%E7%A7%91%E5%AD%A6%E7%A0%94%E7%A9%B6/%E8%BD%AF%E4%BB%B6%E4%B8%8E%E6%95%B0%E6%8D%AE.html?layout=edit&id=322
BMI-UKB	Body Mass Index	457824	https://www.nature.com/articles/s41588-018-0144-6
BMI-GIANT	Body Mass Index	485,648∼795,640	https://academic.oup.com/hmg/article/27/20/3641/5067845
height-BBJ	height	159095	https://www.nature.com/articles/s41467-019-12276-5#Ack1
height-Chinese	height	32921	http://www.sribd.cn/index.php/cn/%E7%A7%91%E5%AD%A6%E7%A0%94%E7%A9%B6/%E8%BD%AF%E4%BB%B6%E4%B8%8E%E6%95%B0%E6%8D%AE.html?layout=edit&id=322
height-UKB	height	458303	https://www.nature.com/articles/s41588-018-0144-6
height-Giant	height	50,003∼253,280	https://www.nature.com/articles/ng.3097
HDL-EAS	High-density-lipoprotein cholesterol	70657	https://www.nature.com/articles/s41588-018-0047-6
HDL-EUR	High-density-lipoprotein cholesterol	188577	https://www.nature.com/articles/ng.2797
LDL-EAS	Low-density-lipoprotein cholesterol	72866	https://www.nature.com/articles/s41588-018-0047-6
LDL-EUR	Low-density-lipoprotein cholesterol	188577	https://www.nature.com/articles/ng.2797
MCH-EAS	Mean corpuscular hemoglobin concentration	108054	https://www.nature.com/articles/s41588-018-0047-6
MCH-UKB1+BIL	Mean corpuscular hemoglobin	132224	https://www.cell.com/cell/references/S0092-8674(16)31463-5
MCHC-EAS	Mean corpuscular hemoglobin concentration	108728	https://www.nature.com/articles/s41588-018-0047-6
MCHC-UKB1+BIL	Mean corpuscular hemoglobin concentration	132586	https://www.cell.com/cell/references/S0092-8674(16)31463-5
MCV-EAS	Mean corpuscular volume	108256	https://www.nature.com/articles/s41588-018-0047-6
MCV-UKB1+BIL	Mean corpuscular volume	132353	https://www.cell.com/cell/references/S0092-8674(16)31463-5
SysBP-EAS	Systolic blood pressure	136597	https://www.nature.com/articles/s41588-018-0047-6
SysBP-EUR	Systolic blood pressure	422771	https://www.nature.com/articles/s41588-018-0144-6
TC -EAS	Total cholesterol	128305	https://www.nature.com/articles/s41588-018-0047-6
TC-EUR	Total cholesterol	188577	https://www.nature.com/articles/ng.2797
TG-EAS	Triglyceride	105597	https://www.nature.com/articles/s41588-018-0047-6
TG-EUR	Triglyceride	188577	https://www.nature.com/articles/ng.2797
eGFR-EAS	Estimated glomerular filtration rate	143658	https://www.nature.com/articles/s41588-018-0047-6
eGFR-EUR	Estimated glomerular filtration rate	480698	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6377354/
BUN-EAS	Blood urea nitrogen	139818	https://www.nature.com/articles/s41588-018-0047-6
BUN-EUR	Blood urea nitrogen	480698	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6377354/
AF-BBJ	Atrial Fibrillation	8,180+28,612	https://www.nature.com/articles/ng.3842
AF-UKB	Atrial Fibrillation	10,986+233,904	https://www.gov.uk/government/publications/atrial-fibrillation-prevalence-estimates-for-local-populations
VCDR-EAS	vertical cup-disc ratio	8373	https://www.nature.com/articles/s41467-017-01913-6
VCDR-EUR	vertical cup-disc ratio	23899	https://www.nature.com/articles/s41467-017-01913-6
DA-EAS	Disc Area	7307	https://www.nature.com/articles/s41467-017-01913-6
DA-EUR	Disc Area	22504	https://www.nature.com/articles/s41467-017-01913-6
SMKi-BBJ	Smoking Initiation	83,810+81,626	https://www.nature.com/articles/s41562-019-0557-y
SMK-UKB	Current tobacco smoking	386150	https://www.nature.com/articles/s41588-019-0481-0
SMKcpd-UKB	Cigarettes per day	90143	https://www.nature.com/articles/s41588-019-0481-0
Angina-UKB	Angina	12114+373585	https://www.nature.com/articles/s41588-019-0481-0
CIHD-UKB	Chronic ischemic heart disease	14456+286335	https://www.nature.com/articles/s41588-019-0481-0
CHD-finngen	Major coronary heart disease event	10157+351037	http://www.nealelab.is/uk-biobank/
AIF-UKB	Alcohol intake frequency	360726	http://www.nealelab.is/uk-biobank/
HBP-UKB	High blood pressure	103381+282318	https://www.nature.com/articles/s41588-019-0481-0
hayfever-UKB	Hayfever, allergic rhinitis or eczema	83407+277120	http://www.nealelab.is/uk-biobank/
Asthma-UKB	Asthma	41633+318894	http://www.nealelab.is/uk-biobank/
glaucoma-UKB	Glaucoma-ICD10(H40)	1715+359479	http://www.nealelab.is/uk-biobank/
HF-UKB	heart failure	6504+387652	https://doi.org/10.1161/CIRCULATIONAHA.118.035774
CoCa-UKB	malignant neoplasm of colon-IDC10(C18)	2226+358968	http://www.nealelab.is/uk-biobank/
COPD-UKB	Other chronic obstructive pulmonary disease-ICD90(J44)	1531+359663	http://www.nealelab.is/uk-biobank/
Osteoporosis-UKB	Osteoporosis	5736+355405	http://www.nealelab.is/uk-biobank/
PAD_UKB	Peripheral artery disease	1230+359964	http://www.nealelab.is/uk-biobank/
PrCa-finngen	Prostate cancer	6321+160699	http://www.nealelab.is/uk-biobank/
UF-UKB	Uterine fibroids	5514+188639	http://www.nealelab.is/uk-biobank/
AD-UKB	Atopic dermatitis	9321+351820	http://www.nealelab.is/uk-biobank/
Arrhythmia-BBJ	Arrhythmia	17861+194592	https://www.nature.com/articles/s41588-020-0640-3
Asthma-BBJ	Asthma	8216+201592	https://www.nature.com/articles/s41588-020-0640-3
Cataract-BBJ	Cataract	24622+187831	https://www.nature.com/articles/s41588-020-0640-3
CHC_BBJ	Chronic hepatitis C	5794+206659	https://www.nature.com/articles/s41588-020-0640-3
CHF_BBJ	Congestive heart failure	9413+203040	https://www.nature.com/articles/s41588-020-0640-3
CoCa_BBJ	Colorectal cancer	7062+195745	https://www.nature.com/articles/s41588-020-0640-3
COPD_BBJ	Chronic obstructive pulmonary disease	3315+201592	https://www.nature.com/articles/s41588-020-0640-3
Glaucoma_BBJ	Glaucoma	5761+206692	https://www.nature.com/articles/s41588-020-0640-3
IS_BBJ	Ischemic stroke	17671+192383	https://www.nature.com/articles/s41588-020-0640-3
Osteoporosis_BBJ	Osteoporosis	7788+204665	https://www.nature.com/articles/s41588-020-0640-3
PAD_BBJ	Peripheral artery disease	3593+208860	https://www.nature.com/articles/s41588-020-0640-3
Pollinosis_BBJ	Pollinosis	5746+206707	https://www.nature.com/articles/s41588-020-0640-3
PrCa_BBJ	Prostate cancer	5408+103939	https://www.nature.com/articles/s41588-020-0640-3
UF_BBJ	Uterine fibroids	5954+95010	https://www.nature.com/articles/s41588-020-0640-3
AD_BBJ	Atopic dermatitis	2385+209651	https://www.nature.com/articles/s41588-020-0640-3

Table S2: Sources of 37 traits from EUR and 35 traits from EAS.

3 Supplementary Note

3.1 Sample quality control of Chinese cohort

We first removed non-Chinese and individuals without height records. We also excluded the related individuals with genetic relatedness exceeding 0.025 to ensure that heritability estimation and PRS construction are not influenced by related individuals. Only individuals with reported age between 16 and 70, and height between 130 cm and 220 cm were retained. Individuals with the genotyping rate less than 5% were also removed. Next, we excluded SNPs with one or more of the following properties: minor allele frequency less than 1%; missing genotypes in more than 5% of the samples; Hardy-Weinberg equilibrium (HWE) p-value below 0.0001. Finally, we took the overlap of SNPs between the Chinese dataset and the UKBB dataset. After these steps, we had $32,921$ individuals with $3,776,575$ SNPs for GWAS and the individual-level PRS analysis. We computed the genetic relatedness matrix (GRM) based on genome-wide genotype data, and then performed a randomized approximation of principal component analysis using plink v2.00 to extract the first 10 principal components for GWAS and cross-population analysis.

For BMI, we further removed individuals with extreme BMI values (larger than 38 or less than 10). This step results in $29,147$ participants with $3,777,871$ SNPs for GWAS and the individual-level PRS analysis. We conducted approximated PCA using plink v2.00 on these genotypes and used the first 10 principal components in data analysis.

3.2 Sample quality control of UKBB data

The full UKBB data were downloaded from `https://www.ukbiobank.ac.uk`. We first extracted the European whites who have reported their height and age. Then the relatives were removed by a genetic relatedness threshold 0.025. Only the individuals with reported height between 130 cm and 220 cm were retained. Individuals with genotyping rate less than 5% were also removed. SNPs were removed if at least one of the following is satisfied: minor allele frequency less than 1%; missing genotypes in more than 5% of the samples; Hardy-Weinberg equilibrium (HWE) p-value below 0.0001. Finally, we took the overlap of SNPs between the Chinese dataset and the UKBB dataset. At the end, we had $429,312$ individuals with $3,776,575$ SNPs for GWAS and the individual-level PRS analysis. Using plink v2.00, the approximate PCA was carried out on these genotypes and the first 20 principal components were included as covariates for data analysis.

For BMI, the same QC steps were applied, resulting in $428,846$ samples with $3,777,871$ SNPs for GWAS and the individual-level PRS analysis. We conducted approximate PCA using plink v2.00 on these genotypes and used the first 20 principal components for data analysis.

3.3 Sample quality control of IPM data

To obtain the ancestries of samples from IPM, we first projected their genotypes to the PC coordinates derived from the 1000 Genomes Project. The samples with the coordinate of the first PC > 0.09 and the coordinate of the second PC < -0.1 were identified as Africans, roughly corresponding to the boundary of African ancestry suggested by the AFR from the 1000 Genomes Project. We applied the same threshold to remove the ancestry outliers in the self-reported Africans

from the UKBB dataset. For both datasets, samples that have phenotypic values more than 4 standard errors away from the mean phenotypic values were identified as outliers and excluded from the analysis.

SNP-level (Rsq score$\geq$0.3) and genotype-per-participant-level (genotype probability$\geq$0.9) filters were used to exclude poorly imputed variants. Genotype QC was performed in PLINK V2.0 after excluding SNPs with a high missing call rate ($\geq$5%), a low minor allele frequency ($\leq$0.01) and deviation from Hardy-Weinberg equilibrium (p-value$\leq 1 \times 10^{-6}$). After phenotype and genotype quality control process (with details given in the Supplementary Note), we first merged two African datasets together, leading to 8,422 confirmed African samples with a total of 2,690,737 overlapping SNPs. Then we randomly selected 1K samples as testing data and used the remaining 7.4K samples as training data.

3.4 Height and BMI associations in Chinese population

To analyze the PRS performance in multi-ancestry datasets, we have collected more than 30k Chinese samples. Here, to study the genetic basis of height and BMI in Chinese population, we conducted GWAS to identify associations from 3.7 million SNPs in the Chinese population. Covariates including age, sex, and first 10 principal components were incorporated in the linear mixed model. Using LD score regression (LDSC) [1], we observed genomic control factor $\lambda_{gc} = 1.20$ and LDSC intercept= 1.026 with standard error (SE=0.014) for height, $\lambda_{gc} = 1.10$ and intercept= 0.998 with (SE=0.012) for BMI, respectively. Considering the polygenicity and the sample size, these statistics suggested no evidence of inflation in our GWAS analysis (Q-Q plots in Figure S10b and g, and Figure S11). After adjusting for the covariates, the residuals of both BMI and height show no correlation with either sexual or geometric factors, suggesting the confounding factors were well-controlled (Figure S8 and S9).

We used the BOLT-LMM v2.3.2 to test for associations between phenotypes and SNPs. We first identified the genome-wide significant SNPs using the p-value threshold 5×10^{-8}. Next, we conducted LD clumping on the significant SNPs using PLINK v2.0 with the LD threshold of 0.1 and clumping radius of one million base pairs. The nearly independent index SNPs were then annotated by the ANNOVAR software [2].

The GWAS identified 58 and 7 genome-wide significant loci (i.e., with leading SNP p-value$<$ 5×10^{-8}) for height and BMI, respectively (Figure S10a and f). Among the 58 height associated loci, 50 loci were previously known, and 36 of them were reported in EAS [3]. The eight novel loci include three intragenic ones (*TBX2-AS1*, *LOC101927932* and *GSDMC*), one located in the exonic area of gene *MIRLET7BHG* and six at the intergenic regions with nearby genes *SPAG17*, *PMCH*, *MIR296*, *TRIB1*, *CHCHD7* and *LOC100272217*. All the seven loci of BMI were previously reported and six of them were found in EAS [3].

To validate the associations identified from the Chinese data, we considered the summary-statistics datasets released from UKBB, the GIANT consortia [4, 5] and BBJ [6, 7] as validation. We compared the effect sizes of the genome-wide significant SNPs in our discovery study with those from the validation studies. For height associated SNPs (Figure S10c-e), all the effects in BBJ were in the same direction with Chinese cohort. In contrast, a number of SNP effects showed opposite directions between EAS and EUR. Besides, the slopes obtained by regressing the effect sizes of the Chinese data on those from the other studies were higher for EAS than for EUR (1.07 for BBJ

compared with 0.65 for UKBB and 0.83 for GIANT), suggesting a more similar genetic architecture within the EAS population and attenuated sharing of genetic basis between EAS and EUR. For BMI (Figure S10h-j), the effect sizes were consistent in directions across all studies, with similar slopes in regression analysis for all non-Chinese populations (0.54 for UKBB, 0.56 for GIANT and 0.52 for BBJ).

By partitioning the genome by chromosomes, we found the heritability of height explained by a chromosome was largely proportional to the chromosome length (Figure S12), consistent with previous studies conducted in EUR [4]. We further conducted a heritability enrichment analysis using the baseline model in the stratified LDSC. We found that all the significantly enriched functional regions in EAS are also enriched in Europeans (Figure S13 and S14). By subsampling the UKBB to the same sample size with Chinese, the enrichment patterns are very similar for the Chinese and UKBB datasets (Figure S15 and S16). The comparative study of GWAS results suggest that the genetic architectures of height and BMI are largely overlapped between EAS and EUR.

3.5 PRS performance in different ethnic groups of the Chinese population

Because the Chinese population is comprised of individuals from various ethnic backgrounds (Supplementary Fig.S1), the PRS performance may also vary across ethnic groups. To study the behavior of PRS in different minority-ethnic groups, we computed the enrichment of each ethnic group in different PRS-defined groups as the ratio between the proportion of an ethnic group in each PRS quantiles to its proportion in the whole test set. The results from six ethnic groups with more than 50 samples in the testing dataset are summarized in Fig.S24. For all the PRS models, there is no enrichment for Han Chinese as the ratio is nearly one across the PRS range. For the PRS derived by BLUP using the Chinese data only, Tujia and Manchu were enriched in the bottom and top quantiles, respectively. This is consistent with the relative height of these two ethnic groups in the population (Supplementary Fig.S22). However, BLUP failed to stratify the other ethnic groups based on the Chinese training data. By incorporating the UKBB dataset in training, XPA not only stratified the Tujia and Manchu people, but also captured the enrichment of Mongols (the highest group) and Hui people (the third highest group) in the top quantile and Zhuang people (the shortest group) in the bottom quantile. These results suggest that the PRS derived by XPA can effectively stratify the subgroups in Chinese population, despite their different ethnic backgrounds.

3.6 Trans-ancestry genetic correlations estimated by XPASS

The success of XPA and XPASS relies on the robust estimate of trans-ancestry genetic correlation. In addition to risk prediction, the trans-ancestry genetic correlation has the value of representing the shared genetic basis between populations.

Here, we applied XPASS to estimate trans-ancestry genetic correlations for a wide spectrum of complex phenotypes, including complex traits/diseases as well as cellular and organismal phenotypes, to provide a global picture of genetic architecture shared between EAS and EUR. Our analysis includes 37 traits from EUR and 35 traits from EAS, where 28 of them are matched pairs (Figure

S17). We also estimated the pair-wise genetic correlations of the phenotypes within each population using GNOVA [8] (Figure S20 and S21). We used the individuals from the 1000 Genomes project as external reference panels. For Europeans, 417 independent samples with $1,313,833$ SNPs were used for constructing the reference panel. For East Asians, 337 independent samples with $1,209,411$ SNPs were used in analysis. Because the sets of variants vary across studies, we only considered the SNPs from the third phase of the International HapMap project phase 3 (HapMap3), resulting in $850,000$ SNPs on average included for estimating the genetic correlation after overlapping procedure. For XPASS, we included the first 5 and 20 principal components as covariates for EAS and EUR reference panels, respectively. The summary statistics of GWAS used in the analysis are summarized in Supplementary Table 1.

Out of the the 28 matched traits, XPASS identified 27 traits that are significantly correlated between the two populations (p-value $< 0.05/28$). Six traits, including type-2 diabetes (T2D), systolic blood pressure (SBP), low-density lipoprotein (LDL), mean corpuscular hemoglobin (MCH), Disc Area (DA) and Glaucoma, were highly correlated between EAS and EUR ($\rho \geq 0.9$). The estimated glomerular filtration rate (eGFR) had the lowest genetic correlation. We estimated the trans-ancestry correlation of height as 0.67 (SE=0.018) and BMI as 0.63 (SE=0.034), consistent with previous findings [9].

Among all 1,295 trans-ancestry pairs of traits, 171 were significantly correlated after Bonferroni correction (p-value $< 0.05/1521$), suggesting pervasive shared genetic basis between the two populations. In particular, multiple pairs of traits strongly correlated within EUR largely remain between EAS and EUR. Examples include positive genetic correlations between triglyceride levels (TG) and T2D, BMI and heart-related diseases, and BMI and smoking behaviors as well as negative genetic correlations between height and chronic ischemic heart disease (CIHD), high-density lipoprotein (HDL) and TG, and eGFR and BMI [10].

We compared the estimates generated by XPASS with those generated by popcorn [11] and summarized the results in Figure S19. We found that the estimated correlations were highly consistent between XPASS and popcorn. Besides, XPASS identified 164 pairs of significantly correlated traits in total, including all 81 significant correlations reported by popcorn.

3.7 Extended Variance Component Model for Accounting for Allele Frequency Difference

To assess the effect of allele frequency difference on the prediction accuracy, we extended the XPASS model to include an additional genetic component that captures the effects of SNPs with large allele frequency differences across populations. From our real data analysis, we did not observe significant enrichment of heritability among these SNPs. As a result, we did not obtain a better PRS by modeling the effect sizes of these SNPs as an additional variance component in the extended model.

We first partitioned the p SNPs into two disjoint sets according to the frequency difference $\text{diff}_j = \frac{|\mathbf{f}_{1,j} - \mathbf{f}_{2,j}|}{\sqrt{2\mathbf{f}_{1,j}(1-\mathbf{f}_{1,j}) + 2\mathbf{f}_{2,j}(1-\mathbf{f}_{2,j})}}$. The set $\mathcal{A}$ included all the 'heterogeneous' SNPs with large allele differences and the set $\mathcal{B}$ contains the remaining SNPs that are not in $\mathcal{A}$. Let $\mathbf{X}_1^A \in \mathbb{R}^{n_1 \times p_A}$ and $\mathbf{X}_2^A \in \mathbb{R}^{n_2 \times p_A}$ denote the standardized genotype matrices of 'heterogeneous' SNPs and $\mathbf{X}_1^B \in \mathbb{R}^{n_1 \times p_B}$ and $\mathbf{X}_2^B \in \mathbb{R}^{n_2 \times p_B}$ denote the standardized genotype matrices of the remaining SNPs for populations one and two, respectively. We related the phenotypes and genotypes using the extended linear

models:

$$\mathbf{y}_1 = \mathbf{Z}_1\boldsymbol{\omega}_1 + \mathbf{X}_1^A\boldsymbol{\beta}_1^A + \mathbf{X}_1^B\boldsymbol{\beta}_1^B + \boldsymbol{\epsilon},$$
$$\mathbf{y}_2 = \mathbf{Z}_2\boldsymbol{\omega}_2 + \mathbf{X}_2^A\boldsymbol{\beta}_2^A + \mathbf{X}_2^B\boldsymbol{\beta}_2^B + \boldsymbol{\xi},$$

where $\boldsymbol{\omega}_1 \in \mathbb{R}^{c_1}$ and $\boldsymbol{\omega}_2 \in \mathbb{R}^{c_2}$ are fixed effects of covariates, $\boldsymbol{\beta}_1^A = [\beta_{1,1}^A, \beta_{1,2}^A, \ldots, \beta_{1,p_A}^A]^T \in \mathbb{R}^{p_A}$ and $\boldsymbol{\beta}_2^A = [\beta_{2,1}^A, \beta_{2,2}^A, \ldots, \beta_{2,p_A}^A]^T \in \mathbb{R}^{p_A}$ are vectors collecting the effect sizes of the 'heterogeneous' SNPs from the two populations, $\boldsymbol{\beta}_1^B = [\beta_{1,1}^B, \beta_{1,2}^B, \ldots, \beta_{1,p_B}^B]^T \in \mathbb{R}^{p_B}$ and $\boldsymbol{\beta}_2^B = [\beta_{2,1}^B, \beta_{2,2}^B, \ldots, \beta_{2,p_B}^B]^T \in \mathbb{R}^{p_B}$ are vectors collecting the effect sizes of the remaining SNPs from the two populations, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_\epsilon^2\mathbf{I}_{n_1})$ and $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma_\xi^2\mathbf{I}_{n_2})$ are independent errors. We considered probabilistic structures to the SNPs in sets $\mathcal{A}$ and $\mathcal{B}$ as

$$\begin{pmatrix}\beta_{1,j}^A \\ \beta_{2,j}^A\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_{1A}^2 & \rho_A\sigma_{1A}\sigma_{2A} \\ \rho_A\sigma_{1A}\sigma_{2A} & \sigma_{2A}^2\end{pmatrix}\right), \quad j = 1, \ldots, p_A,$$
$$\begin{pmatrix}\beta_{1,j}^B \\ \beta_{2,j}^B\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_{1B}^2 & \rho_B\sigma_{1B}\sigma_{2B} \\ \rho_B\sigma_{1B}\sigma_{2B} & \sigma_{2B}^2\end{pmatrix}\right), \quad j = 1, \ldots, p_B,$$

respectively, where σ_{1A}^2 and σ_{2A}^2 are the variance components of the 'heterogeneous' SNP effects in the two populations, respectively, ρ_A is the trans-ancestry genetic correlation of the 'heterogeneous' SNP effects, σ_{1B}^2 and σ_{2B}^2 are the variance components of the remaining SNP effects in the two populations, respectively, and ρ_B is the trans-ancestry genetic correlation of the remaining SNP effects. With this flexible statistical structure of genetic effects, the variance and genetic correlation of 'heterogeneous' SNPs are allowed to be different from the remaining SNPs. We can estimate the parameters and obtain the posterior means of $\boldsymbol{\beta}^A$ and $\boldsymbol{\beta}^B$ using GWAS summary statistics in the similar way as in XPASS. To evaluate the impact of the 'heterogeneous' SNPs on prediction performance, we applied this extended model to the height and BMI datasets. We estimated the heritability explained by the two components, evaluated the enrichment of heritabilities of the 'heterogeneous' component, and constructed PRS from the extended model. Recall that the predictive R^2 of the original XPASS model is 17.63%. As summarized in Table S1, we observed neither significant enrichment of heritability in the 'heterogeneous' SNPs, nor improvement of prediction performance when these SNPs were introduced as an additional component in the model. Our results suggest that modeling the effect sizes of SNPs with large allele frequency difference may not be the key to improve PRS.

References

[1] Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, Benjamin M Neale, Schizophrenia Working Group of the Psychiatric Genomics Consortium, et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*, 47(3):291–295, 2015.

[2] Kai Wang, Mingyao Li, and Hakon Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16):e164–e164, 2010.

[3] Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, 47(D1):D1005–D1012, 2019.

[4] Andrew R Wood, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, Karol Estrada, Zoltán Kutalik, Najaf Amin, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics*, 46(11):1173–1186, 2014.

[5] Adam E Locke, Bratati Kahali, Sonja I Berndt, Anne E Justice, Tune H Pers, Felix R Day, Corey Powell, Sailaja Vedantam, Martin L Buchkovich, Jian Yang, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206, 2015.

[6] Masato Akiyama, Yukinori Okada, Masahiro Kanai, Atsushi Takahashi, Yukihide Momozawa, Masashi Ikeda, Nakao Iwata, Shiro Ikegawa, Makoto Hirata, Koichi Matsuda, et al. Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. *Nature genetics*, 49(10):1458–1467, 2017.

[7] Masato Akiyama, Kazuyoshi Ishigaki, Saori Sakaue, Yukihide Momozawa, Momoko Horikoshi, Makoto Hirata, Koichi Matsuda, Shiro Ikegawa, Atsushi Takahashi, Masahiro Kanai, et al. Characterizing rare and low-frequency height-associated variants in the Japanese population. *Nature communications*, 10(1):1–11, 2019.

[8] Qiongshi Lu, Boyang Li, Derek Ou, Margret Erlendsdottir, Ryan L Powles, Tony Jiang, Yiming Hu, David Chang, Chentian Jin, Wei Dai, et al. A powerful approach to estimating annotation-stratified genetic covariance via GWAS summary statistics. *The American Journal of Human Genetics*, 101(6):939–964, 2017.

[9] Alicia R Martin, Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M Neale, and Mark J Daly. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature genetics*, 51(4):584–591, 2019.

[10] Brendan Bulik-Sullivan, Hilary K Finucane, Verneri Anttila, Alexander Gusev, Felix R Day, Po-Ru Loh, Laramie Duncan, John RB Perry, Nick Patterson, Elise B Robinson, et al. An atlas of genetic correlations across human diseases and traits. *Nature genetics*, 47(11):1236–1241, 2015.

[11] Brielin C Brown, Chun Jimmie Ye, Alkes L Price, Noah Zaitlen, Asian Genetic Epidemiology Network Type 2 Diabetes Consortium, et al. Transethnic genetic-correlation estimates from summary statistics. *The American Journal of Human Genetics*, 99(1):76–88, 2016.