Supplementary information for

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

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

Affiliations:

[1] Division of Birth Cohort Study, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, 510623, China.
[2] School of Public Health (Shenzhen), Sun Yat-sen University, Shenzhen, Guangdong 510006, China.
[3] Provincial Clinical Research Center for Child Health, Guangzhou, 510623, China.
[4] Department of Obstetrics and Gynecology, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, 510623, China
[5] Provincial Key Laboratory of Research in Structure Birth Defect Disease, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, 510623, China.
[6] Department of Pediatric Surgery, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, 510623, China
[7] Department of Women's Health, Provincial Key Clinical Specialty of Woman and Child Health, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, 510623, China

* These authors contributed equally to this work: Shujia Huang, Mingxi Huang, Siyang Liu
# These authors supervised this work: Xiu Qiu and Huimin Xia. Correspondence should be addressed to xiu.qiu@bigcs.org and huimin.xia@bigcs.org

Table of Contents

Supplementary Figures

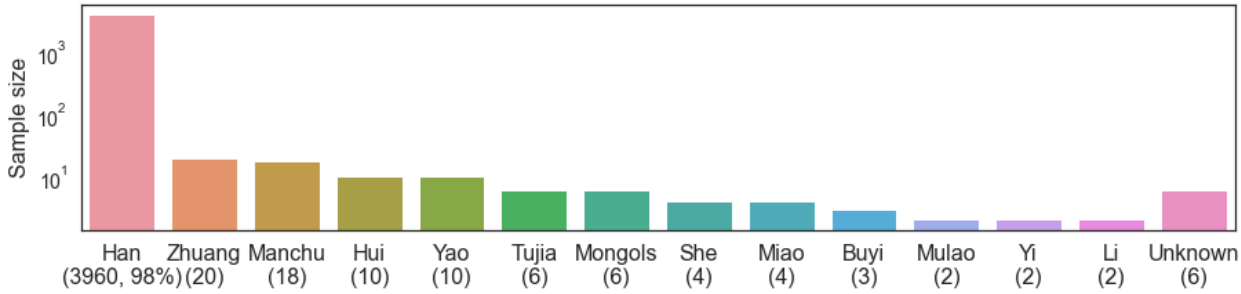


Figure S1. Ethnic distribution of the 4,053 samples of BIGCS.

The y-axis represents the sample size in log-scale and each color bar represents each of the ethnic groups. The number of individuals is indicated in the parenthesis. Related to Figure 1.

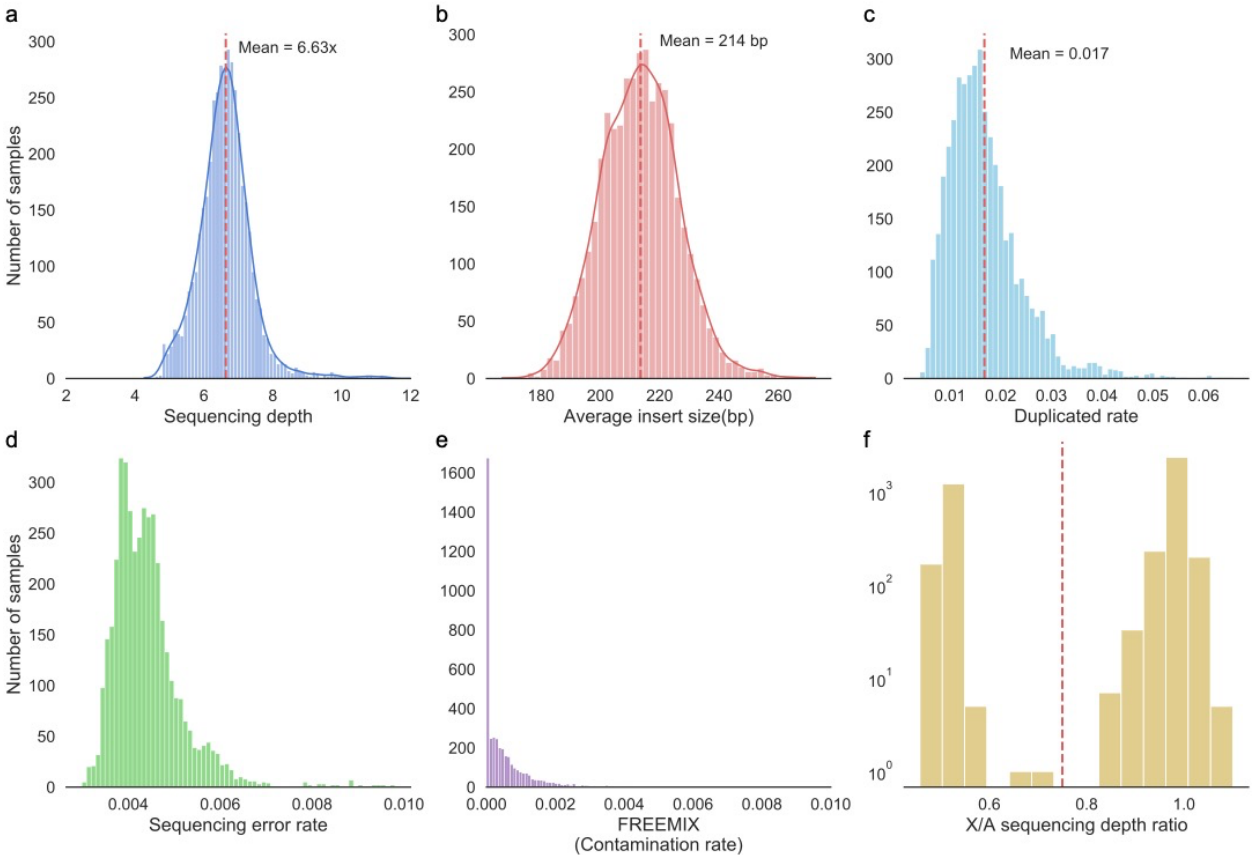


Figure S2. Quality and overview of the sequencing samples.

(a) Distribution of sequencing depth in the WGS data of BIGCS. (b) Distribution of average insert size. (c) Distribution of the duplication rate of the sequencing data. (d) Distribution of sequencing error rate provided by BGI-Shenzhen Co., Ltd. We used 0.01 as the strict cutoff to filter high sequencing error samples. (e) Distribution of contamination rates estimated by verifyBamID2. A strict cutoff of 0.01 was used to filter contaminated samples. (f) Distribution of average sequencing depth between chromosome X and autosomes (X/A ratio). A cutoff of 0.75 was used to infer the gender of the BIGCS samples.

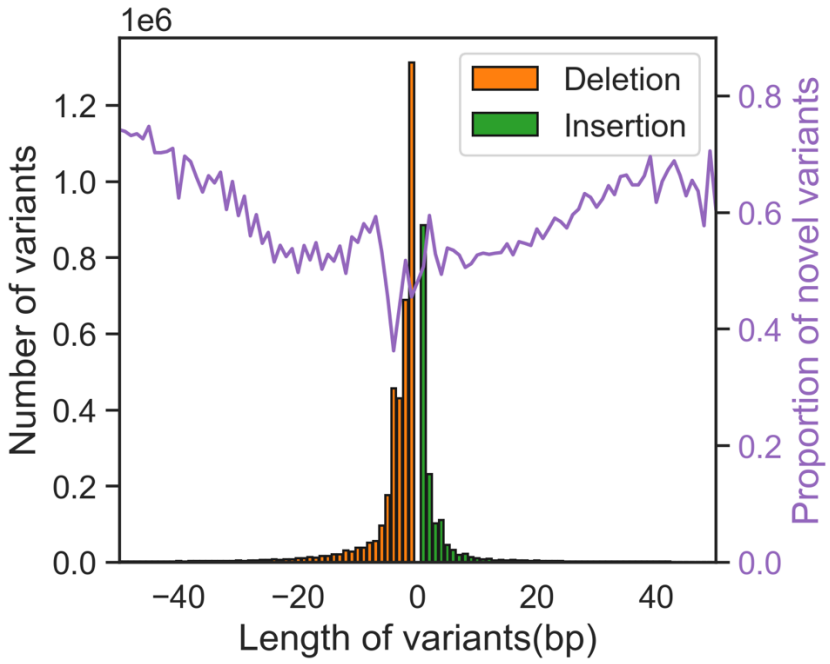


Figure S3. The length distribution of the Indels.

The x-axis represents the length of Indels variants from -50 bp (Deletion) to +50 bp (Insertion). The left y-axis is the count of Indels. The purple line and right y-axis represent the proportion of novel variants**.** Related to Figure 1.

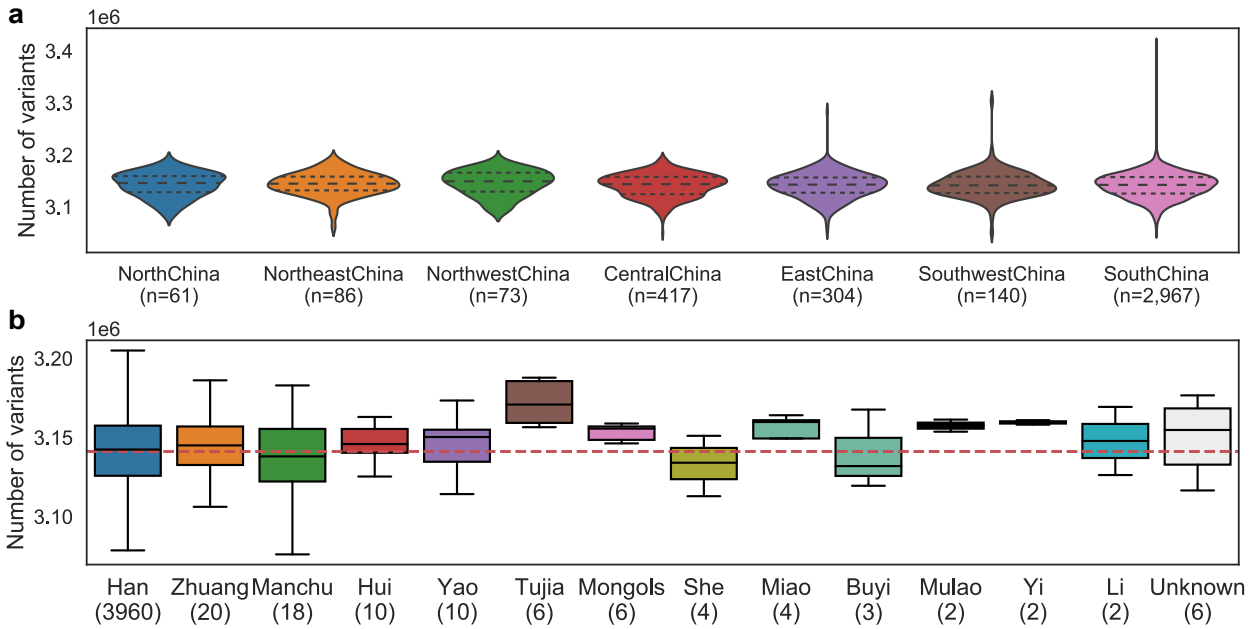


Figure S4. The number of the variants along geographical regions and ethnicities.

The distribution of number of variants (SNPs and Indels) grouped by the seven large geographical regions of China (a) and 13 ethnic groups in BIGCS (b) with sample size indicated in parenthesis. The red dash horizontal line demonstrates the mean of individuals' variants. Related to Figure 1.

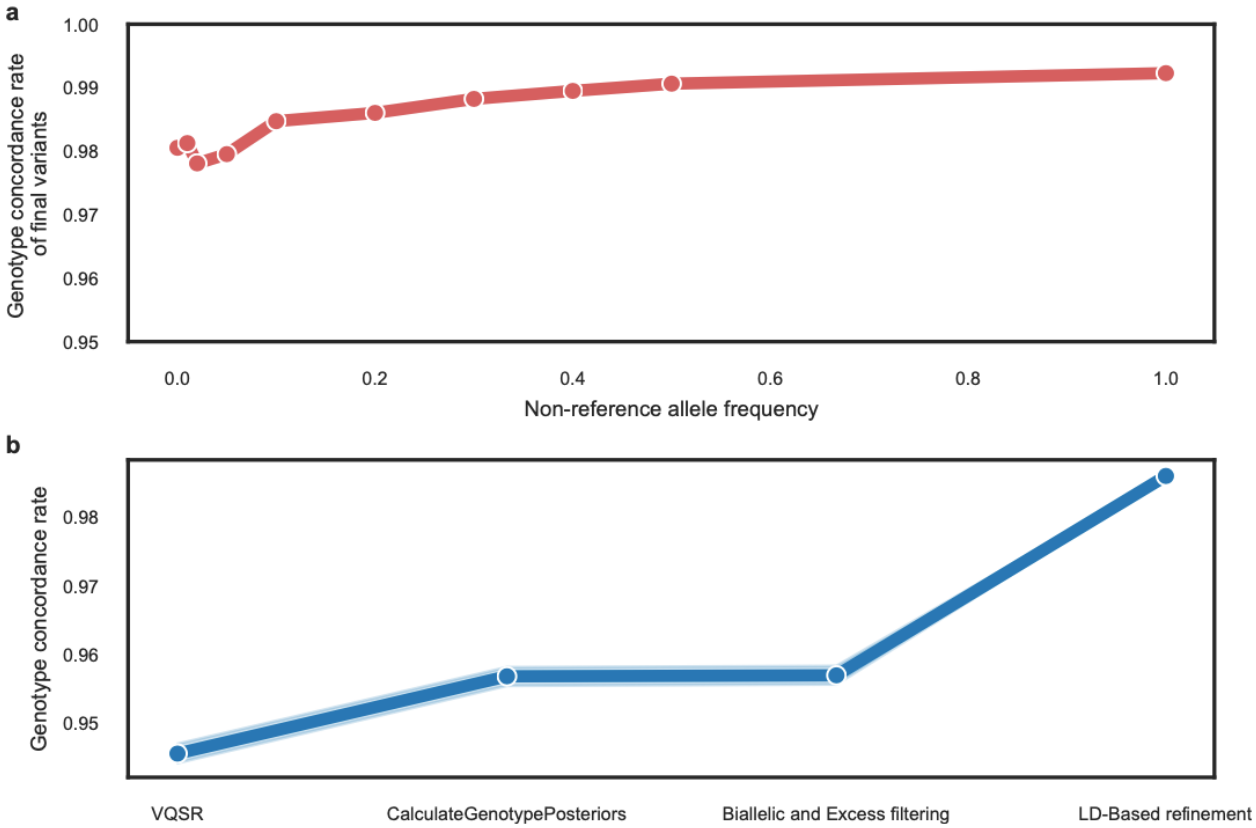


Figure S5. The genotype validation.

Genotype concordance rate of the final variants call set (after LD-based refinement) compare with the genotype data of 240 SNP-array samples in different non-reference allele frequency bins (a) and genotype concordance rate in different variants filtration processes (b). Related to Figure 1.

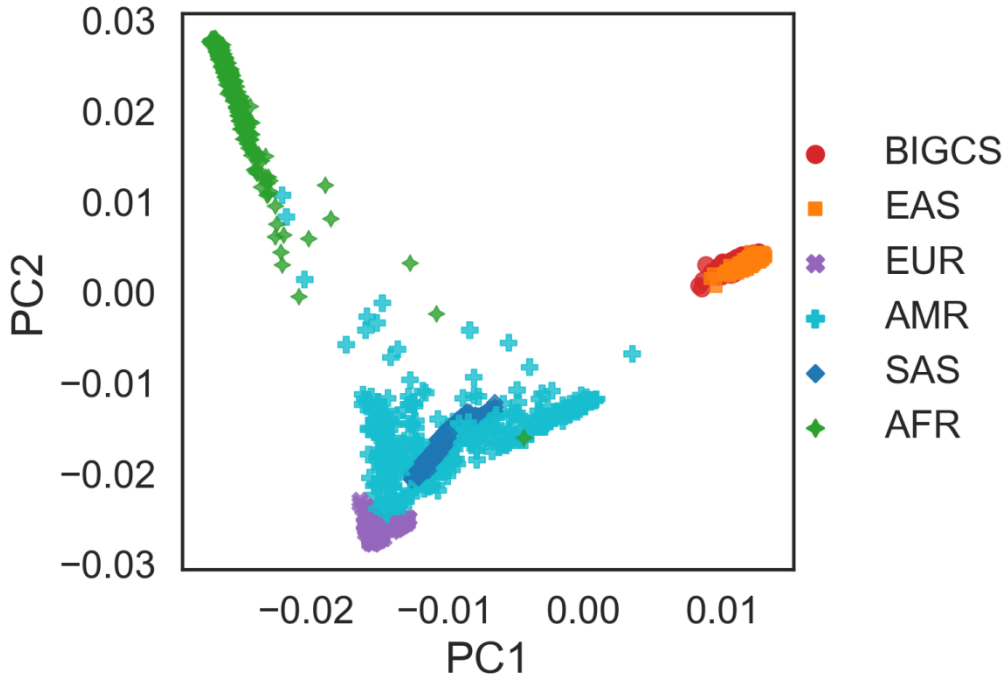


Figure S6. PCA analysis for the BIGCS and the 1KGP3 samples.
Each point represents one participant and is placed according to their PC scores. Related to Figure 2.

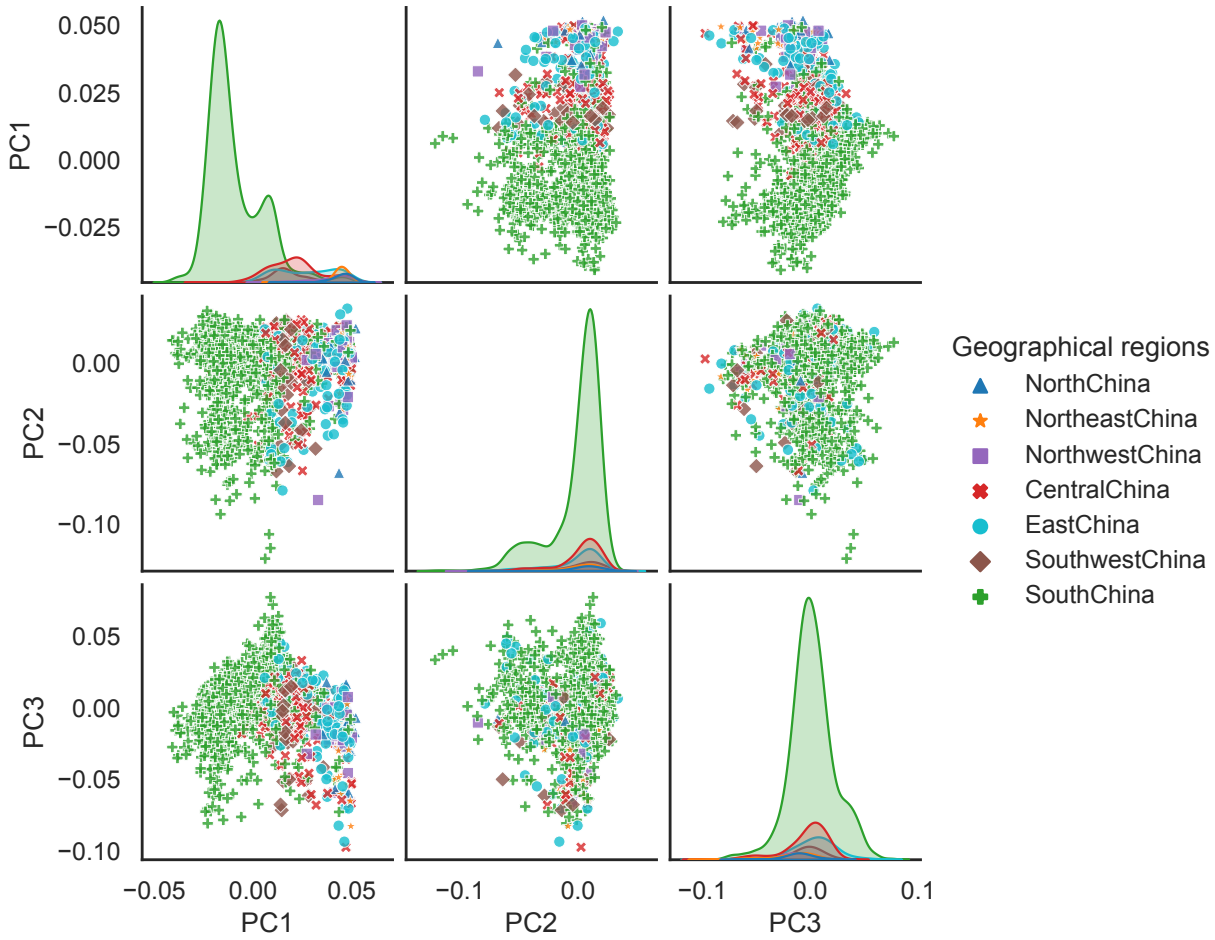


Figure S7. PCA of the BIGCS samples from seven geographical groups of China.
Pair plot for the first three components of PCA analysis. Each point represents one participant and is placed according to their PC scores. Related to Figure 2.

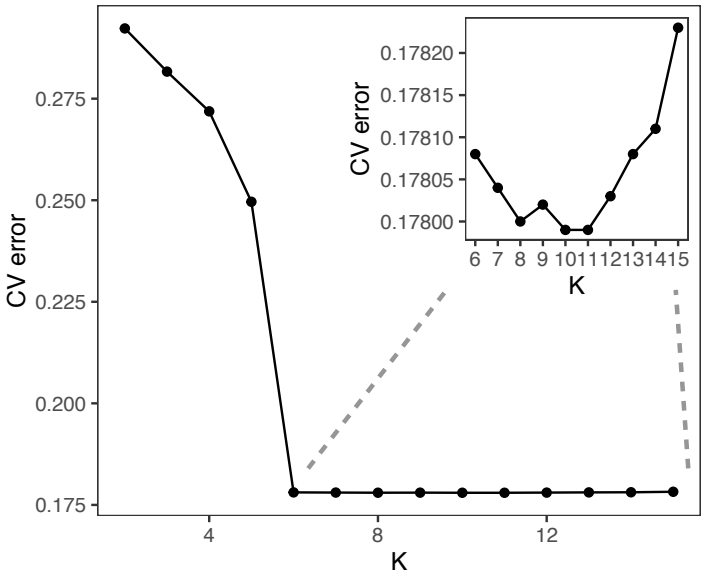
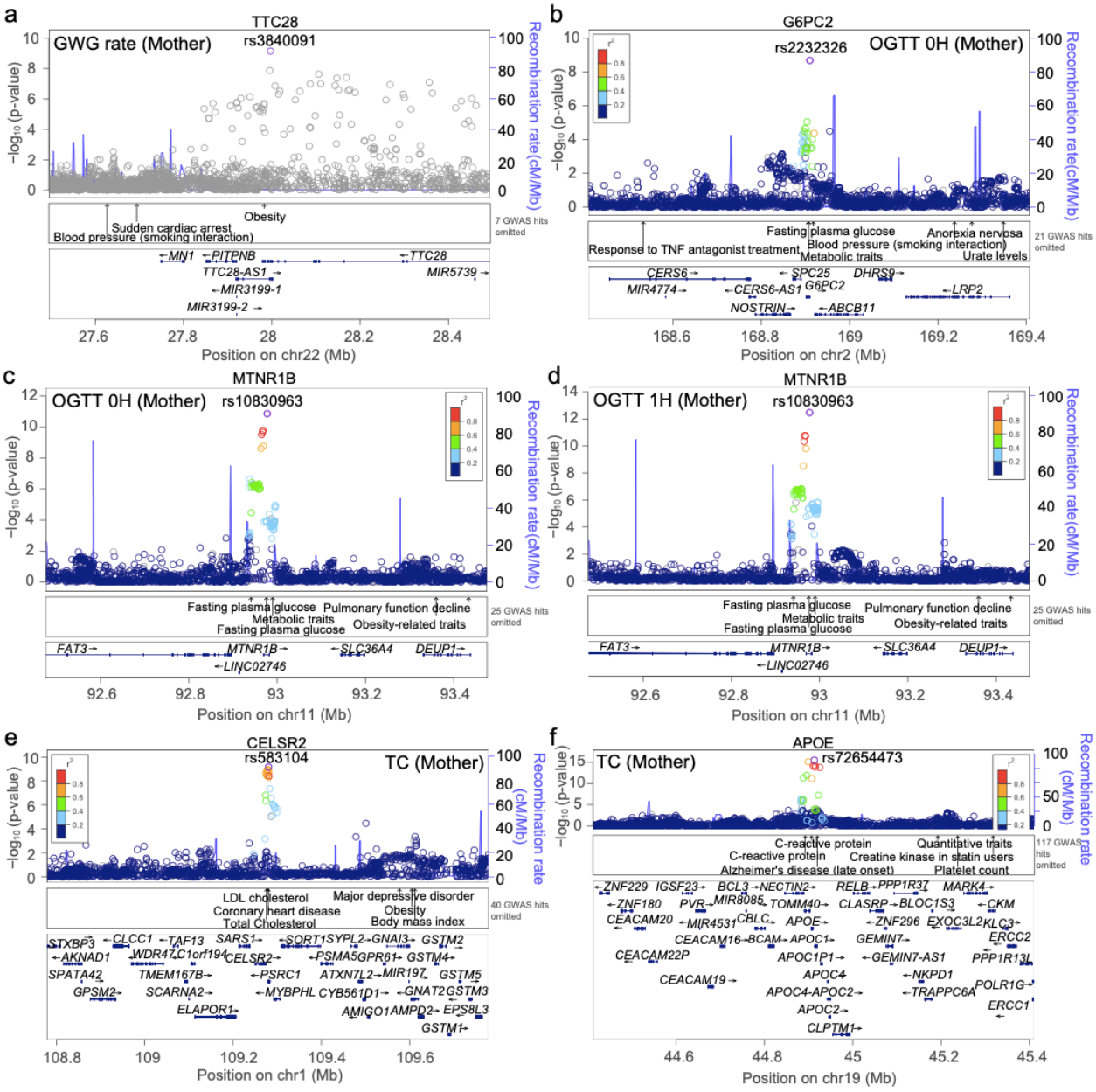
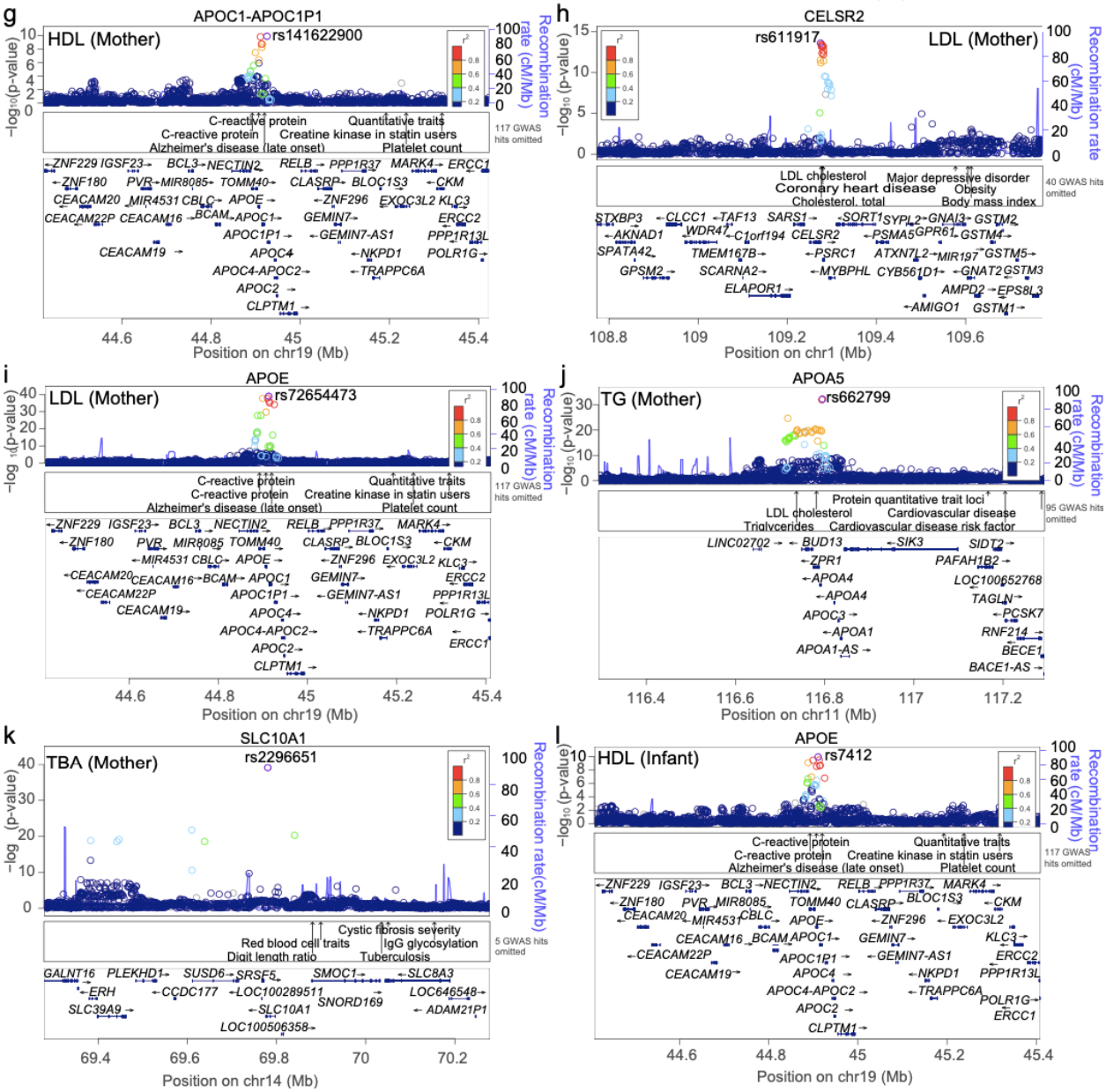


Figure S8. The distribution of cross-validation error of ADMIXTURE analysis. Related to Figure 2.


a
TTC28
GWG rate (Mother)
rs3840091
−log10 (p-value)
Recombination rate (cM/Mb)
Obesity
Sudden cardiac arrest
Blood pressure (smoking interaction)
7 GWAS hits omitted
MN1
PITPNB
TTC28
TTC28-AS1
MIR5739
MIR3199-1
MIR3199-2
Position on chr22 (Mb)
b
G6PC2
rs2232326
OGTT 0H (Mother)
Fasting plasma glucose
Anorexia nervosa
Response to TNF antagonist treatment
Blood pressure (smoking interaction)
Metabolic traits
Urate levels
21 GWAS hits omitted
CERS6
SPC25
DHRS9
MIR4774
CERS6-AS1
G6PC2
LRP2
NOSTRIN
ABCB11
Position on chr2 (Mb)
c
MTNR1B
OGTT 0H (Mother)
rs10830963
Fasting plasma glucose
Pulmonary function decline
Metabolic traits
Fasting plasma glucose
Obesity-related traits
25 GWAS hits omitted
FAT3
MTNR1B
SLC36A4
DEUP1
LINC02746
Position on chr11 (Mb)
d
MTNR1B
OGTT 1H (Mother)
rs10830963
Fasting plasma glucose
Pulmonary function decline
Metabolic traits
Fasting plasma glucose
Obesity-related traits
25 GWAS hits omitted
FAT3
MTNR1B
SLC36A4
DEUP1
LINC02746
Position on chr11 (Mb)
e
CELSR2
rs583104
TC (Mother)
LDL cholesterol
Major depressive disorder
Coronary heart disease
Obesity
Total Cholesterol
Body mass index
40 GWAS hits omitted
STXBP3
CLCC1
TAF13
SARS1
SORT1
SYPL2
GNAI3
GSTM2
AKNAD1
WDR47
C1orf194
CELSR2
PSMA5
GPR61
GSTM4
SPATA42
TMEM167B
PSRC1
ATXN7L2
MIR197
GSTM5
GPSM2
SCARNA2
MYBPHL
CYB561D1
GNAT2
GSTM3
ELAPOR1
AMIGO1
AMPD2
EPS8L3
GSTM1
Position on chr1 (Mb)
f
APOE
rs72654473
TC (Mother)
C-reactive protein
Quantitative traits
C-reactive protein
Creatine kinase in statin users
Alzheimer's disease (late onset)
Platelet count
117 GWAS hits omitted
ZNF229
IGSF23
BCL3
NECTIN2
RELB
PPP1R37
MARK4
ZNF180
PVR
MIR8085
TOMM40
CLASRP
BLOC1S3
CKM
CEACAM20
MIR4531
CBLC
APOE
ZNF296
EXOC3L2
KLC3
CEACAM16
BCAM
APOC1
GEMIN7
ERCC2
CEACAM22P
APOC1P1
GEMIN7-AS1
PPP1R13L
CEACAM19
APOC4
NKPD1
APOC4-APOC2
TRAPPC6A
POLR1G
APOC2
ERCC1
CLPTM1
Position on chr19 (Mb)


g
APOC1-APOC1P1
HDL (Mother)
rs141622900
−log10(p-value)
Recombination rate (cM/Mb)
C-reactive protein
Creatine kinase in statin users
Alzheimer's disease (late onset)
Quantitative traits
Platelet count
117 GWAS hits omitted
ZNF229 IGSF23 BCL3 NECTIN2 RELB PPP1R37 MARK4 ERCC1
ZNF180 PVR MIR8085 TOMM40 CLASRP BLOC1S3 CKM
CEACAM20 MIR4531 CBLC APOE ZNF296 EXOC3L2 KLC3
CEACAM22P CEACAM16 BCAM APOC1 GEMIN7 ERCC2
APOC1P1 GEMIN7-AS1 PPP1R13L
CEACAM19 APOC4 NKPD1 POLR1G
APOC4-APOC2 TRAPPC6A
APOC2
CLPTM1
44.6 44.8 45 45.2 45.4
Position on chr19 (Mb)
h
CELSR2
rs611917
LDL (Mother)
LDL cholesterol
Coronary heart disease
Cholesterol, total
Major depressive disorder
Obesity
Body mass index
40 GWAS hits omitted
STXBP3 CLCC1 TAF13 SARS1 SORT1 SYPL2 GNAI3 GSTM2
AKNAD1 WDR47 C1orf194 CELSR2 PSMA5 GPR61 GSTM4
SPATA42 TMEM167B PSRC1 ATXN7L2 MIR197 GSTM5
GPSM2 SCARNA2 MYBPHL CYB561D1 GNAT2 GSTM3
ELAPOR1 AMPD2 EPS8L3
AMIGO1 GSTM1
108.8 109 109.2 109.4 109.6
Position on chr1 (Mb)
i
APOE
rs72654473
LDL (Mother)
C-reactive protein
Creatine kinase in statin users
Alzheimer's disease (late onset)
Quantitative traits
Platelet count
117 GWAS hits omitted
ZNF229 IGSF23 BCL3 NECTIN2 RELB PPP1R37 MARK4
ZNF180 PVR MIR8085 TOMM40 CLASRP BLOC1S3 CKM
MIR4531 CBLC APOE ZNF296 EXOC3L2 KLC3
CEACAM20 CEACAM16 BCAM APOC1 GEMIN7 ERCC2
CEACAM22P APOC1P1 GEMIN7-AS1 PPP1R13L
CEACAM19 APOC4 NKPD1 POLR1G
APOC4-APOC2 TRAPPC6A ERCC1
APOC2
CLPTM1
44.6 44.8 45 45.2 45.4
Position on chr19 (Mb)
j
APOA5
rs662799
TG (Mother)
Protein quantitative trait loci
LDL cholesterol
Triglycerides
Cardiovascular disease
Cardiovascular disease risk factor
95 GWAS hits omitted
LINC02702 BUD13 SIK3 SIDT2
ZPR1 PAFAH1B2
APOA4 LOC100652768
APOA4 TAGLN
APOC3 PCSK7
APOA1 RNF214
APOA1-AS BECE1
BACE1-AS
116.4 116.6 116.8 117 117.2
Position on chr11 (Mb)
k
SLC10A1
rs2296651
TBA (Mother)
Cystic fibrosis severity
Red blood cell traits
IgG glycosylation
Digit length ratio
Tuberculosis
5 GWAS hits omitted
GALNT16 PLEKHD1 SUSD6 SRSF5 SMOC1 SLC8A3
ERH CCDC177 LOC100289511 SNORD169 LOC646548
SLC39A9 SLC10A1 ADAM21P1
LOC100506358
69.4 69.6 69.8 70 70.2
Position on chr14 (Mb)
l
APOE
rs7412
HDL (Infant)
C-reactive protein
Creatine kinase in statin users
Alzheimer's disease (late onset)
Quantitative traits
Platelet count
117 GWAS hits omitted
ZNF229 IGSF23 BCL3 NECTIN2 RELB PPP1R37 MARK4
ZNF180 PVR MIR8085 TOMM40 CLASRP BLOC1S3 CKM
CEACAM20 MIR4531 CBLC APOE ZNF296 EXOC3L2
CEACAM16 BCAM APOC1 GEMIN7 KLC3
CEACAM22P APOC1P1 GEMIN7-AS1 ERCC2
CEACAM19 APOC4 NKPD1 PPP1R13L
APOC4-APOC2 TRAPPC6A POLR1G
APOC2 ERCC1
CLPTM1
44.6 44.8 45 45.2 45.4
Position on chr19 (Mb)

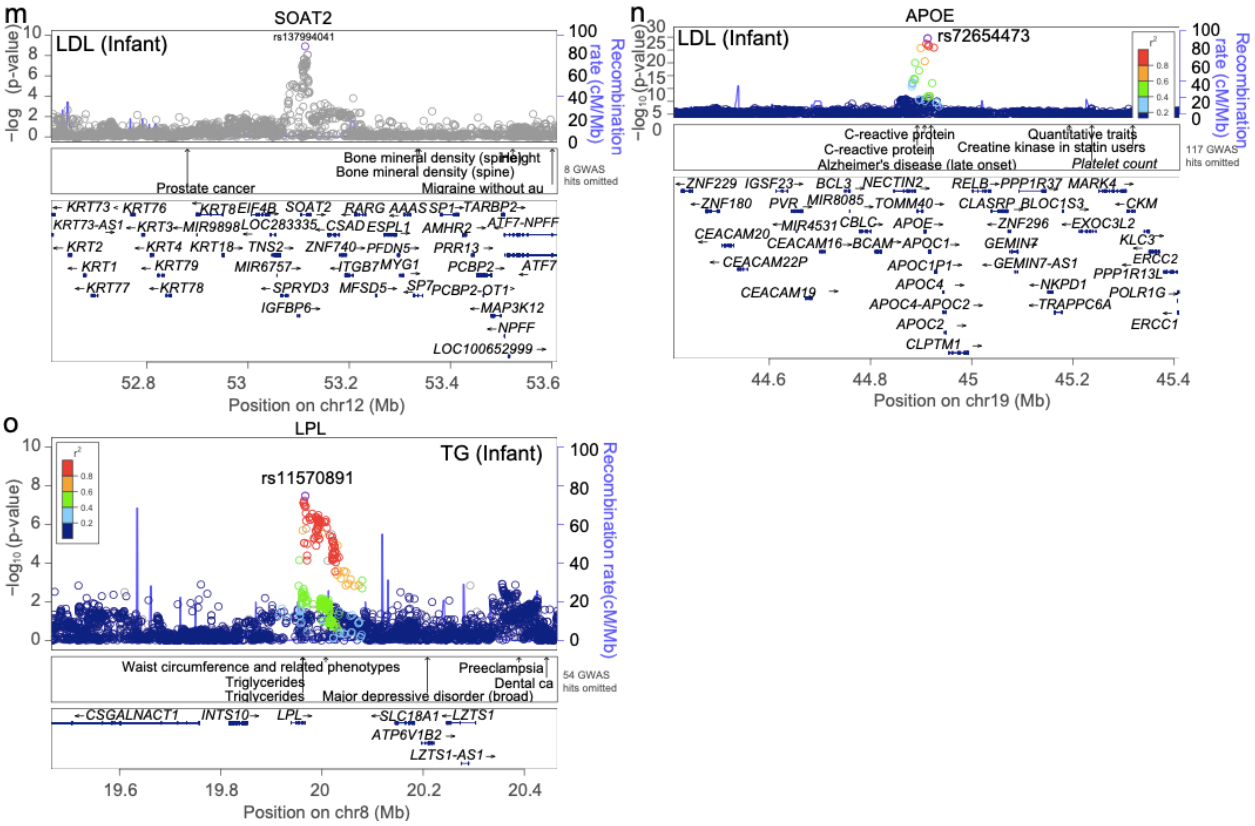


Figure S9. The LocusZoom plots of 15 lead variants for the ten traits with GWAS significant loci revealed in the study.

Related to Figure 3. The detail information of these variants is placed in Table S5.

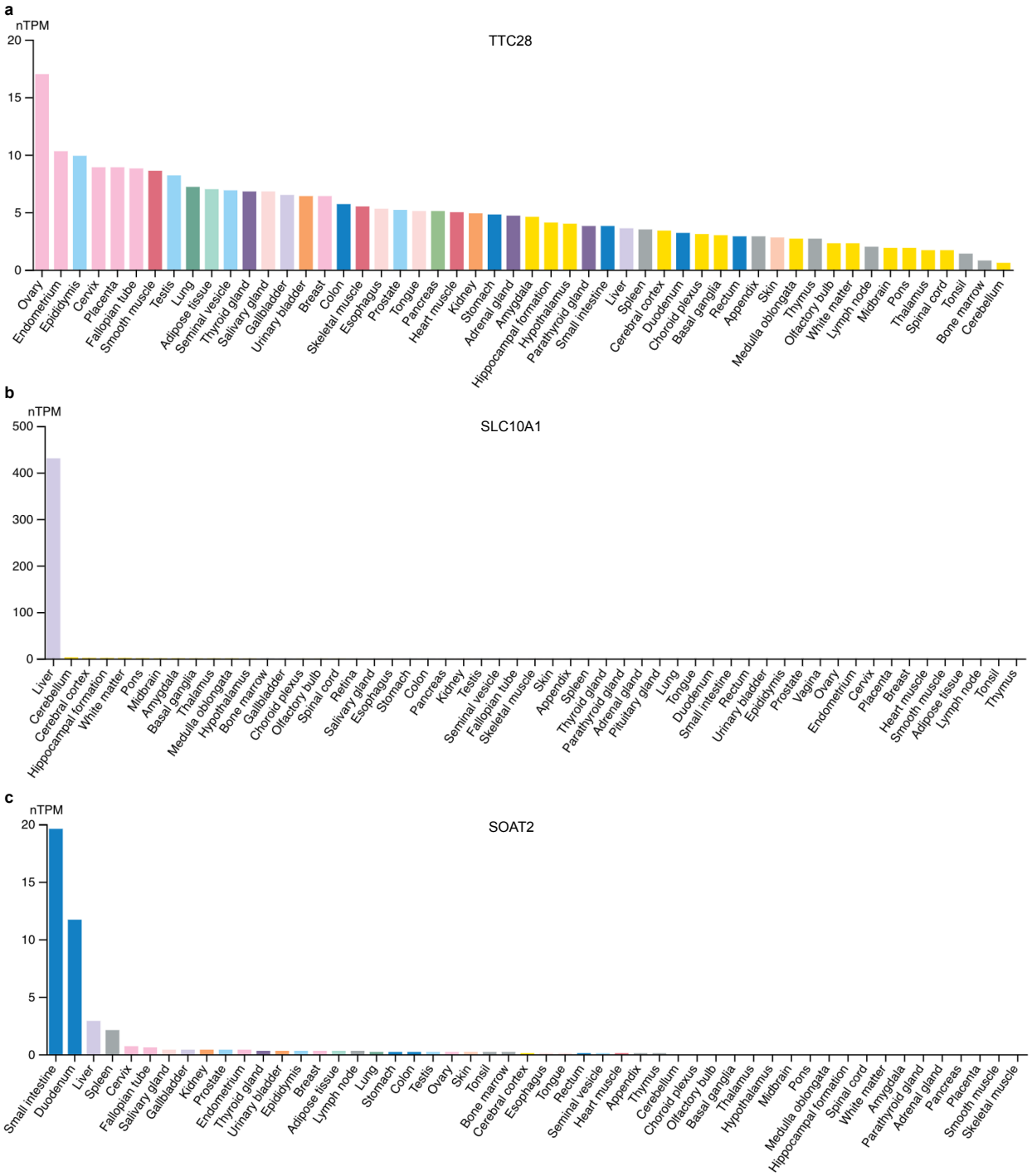


Figure S10. RNA tissue specificity expression of gene *TTC28*, *SLC10A1* and *SOAT2*.
RNA tissue specificity expression of gene *TTC28* (a), *SLC10A1* (b) and *SOAT2* (c) according to the Human Protein Atlas dataset (https://www.proteinatlas.org/). Related to Figure 3.