Population structure of Han Chinese in the modern Taiwanese population based on 10,000 participants in the Taiwan Biobank project

by Chen et al.

Supplementary Information
(Supplementary tables and figures)

Supplementary Tables

S1 Table. Number of SNPs genotyped in TWB

S2 Table. Breakdown of TWB by genetic clusters and self-reported origins in subgroups under a model of K=4 ancestry origins.

S3 Table The fixation index (Fst)* between subgroups of TWB

S4 Table. Major HLA-A-B haplotypes in an independent sample of Taiwanese (n=442), who have been HLA-genotyped.

S5 Table. Major HLA-A-B-T-allele(rs2233947:T) haplotypes in an independent sample of Taiwanese (n=442).

S6 Table. Power calculations

S1 Table. Number of SNPs genotyped in TWB

Chromosome	Number of SNPs	QC*
1	46936	43949
2	52013	48210
3	46611	43215
4	41551	38154
5	39064	36421
6	46240	42584
7	35555	32969
8	35448	33079
9	30237	28342
10	32777	30792
11	30126	28164
12	31041	28948
13	23435	21696
14	21432	19804
15	21892	20442
16	21161	20011
17	16082	15099
18	20321	18909
19	10153	9426
20	15642	14805
21	9826	9069
22	7316	6960
X+	12721	12721
Y+	1942	1942
Mt+	189	189
Total	649711	605900

*Quality control (QC): a SNP was included if the following three criteria were met: (1) minor allele frequency ≥ 1%, (2) genotyping call rate ≥ 95%, (3) Hardy-Weinberg equilibrium proportion test p-value > 0.0001.

+ All of the data for SNPs on chromosome X and Y and mitochondria (Mt) were included without the quality control stated above.

S2 Table. Breakdown of TWB by genetic clusters and self-reported origins in subgroups under a model of K=4 ancestry origins.

Self-reported origins+	T1	T2	T3	T4	T total	S1	S2	S3	S total	N1	N2	N total	O	Total
N_N	0 (0.0%, 0.0%)	0 (0.0%, 0.0%)	0 (0.0%, 0.0%)	2 (2.7%, 2.0%)	2 (2.7%, 0.1%)	0 (0.0%, 0.0%)	4 (5.4%, 0.5%)	1 (1.4%, 0.0%)	5 (6.8%, 0.1%)	16 (21.6%, 3.6%)	51 (68.9%, 51.5%)	67 (90.5%, 12.3%)	0 (0.0%, 0.0%)	74
S_S	96 (0.9%, 100.0%)	134 (1.3%, 96.4%)	1227 (11.8%, 99.4%)	69 (0.7%, 67.7%)	1526 (14.7%, 97.1%)	125 (1.2%, 99.2%)	5512 (53.2%, 98.6%)	2893 (27.9%, 99.1%)	8530 (82.2%, 98.8%)	257 (2.5%, 57.6%)	11 (0.1%, 11.1%)	268 (2.6%, 49.2%)	47 (0.5%, 94.0%)	10371
Admixture	0 (0.0%, 0.0%)	4 (1.5%, 2.9%)	4 (1.5%, 0.3%)	28 (10.3%, 27.4%)	36 (13.3%, 2.3%)	0 (0.0%, 0.0%)	46 (17.0%, 0.8%)	14 (5.17%, 0.5%)	60 (22.1%, 0.7%)	155 (57.2%, 34.8%)	19 (7.0%, 19.2%)	174 (64.2%, 31.9%)	1 (0.4%, 2.0%)	271
Unknown	0 (0.0%, 0.0%)	1 (1.2%, 0.7%)	3 (3.5%, 0.2%)	3 (3.5%, 2.9%)	7 (8.2%, 0.4%)	1 (1.2%, 0.8%)	27 (31.8%, 0.5%)	12 (14.1%, 0.4%)	40 (47.1%, 0.5%)	18 (21.2%, 4.4%)	18 (21.2%, 18.2%)	36 (42.4%, 6.6%)	2 (2.4%, 4.0%)	85
Total	96	139	1234	102	1571	126	5589	2920	8635	446	99	545	50	10801

* Number of individuals and percentages by row and by column.

+ N_N for subjects with both parents being of northern Han Chinese descent, S_S for subjects with both parents being of southern Han Chinese descent, admixture for subjects with one parent of southern Han Chinese descent and the other being of Northern Han Chinese descent or for subjects with at least one parent being an admixture of southern and northern Han Chinese descent, and unknown for subjects with an unknown origin for at least one parent.

O, Others unable to be defined.

S3 Table The fixation index (F_{st})* between subgroups of TWB

	T1	T2	T3	T4	S1	S2	S3	N1	N2
T1		0.000535	0.001532	0.002661	0.007799	0.006863	0.005101	0.006865	0.007272
T2	0.000667		0.000267	0.001138	0.004481	0.003647	0.002364	0.003815	0.00456
T3	0.001075	0.000207		0.000694	0.002822	0.00192	0.001053	0.002311	0.003334
T4	0.002372	0.001017	0.000686		0.003522	0.002142	0.001335	0.00117	0.001244
S1	0.004733	0.003163	0.002149	0.003062		0.000925	0.001072	0.002372	0.004738
S2	0.003528	0.00264	0.001643	0.002214	0.000888		0.000186	0.000851	0.002645
S3	0.002726	0.001672	0.000832	0.001216	0.001033	0.000244		0.000844	0.002349
N1	0.004047	0.002829	0.00187	0.001027	0.002284	0.000894	0.000833		0.000526
N2	0.005364	0.003755	0.003032	0.001059	0.00453	0.00285	0.002292	0.000515	

- F_{st} was computed based on 591,048 SNPs, using the method introduced in “Estimating F-Statistics for the Analysis of Population Structure” by B. S. Weir and C. Clark Cockerham, Evolution, Vol. 38, No. 6 (Nov., 1984), pp. 1358-1370 , implemented in PLINK 1.9
- Upper triangular matrix is Weighted Fst estimate.
- Lower triangular matrix is Mean Fst estimate.

S4 Table. Major HLA-A-B haplotypes in an independent sample of Taiwanese (n=442), who have been HLA-genotyped.

	HLA	A*33:03-B*58:01		A*11:01-B*40:01		A*02:01-B*40:01		A*02:07-B*46:01		A*11:01-B*13:01		A*11:01-B*15:02		A*11:01-B*27:04		A*24:02-B*40:01	
	Total	N	Freq.[+]	N	Freq.[+]	N	Freq.[+]	N	Freq.[+]	N	Freq.[+]	N	Freq.[+]	N	Freq.[+]	N	Freq.[+]
T1	4	4	1.000	0	0.000	0	0.000	0	0.000	0	0.000	0	0.000	0	0.000	0	0.000
T2	12	8	0.667	0	0.000	1	0.083	0	0.000	0	0.000	0	0.000	0	0.000	0	0.000
T3	90	42	0.467	8	0.089	1	0.011	2	0.022	3	0.033	1	0.011	1	0.011	2	0.022
T4	6	2	0.333	0	0.000	0	0.000	0	0.000	0	0.000	0	0.000	0	0.000	0	0.000
S1	10	0	0.000	1	0.100	0	0.000	3	0.300	0	0.000	0	0.000	1	0.100	0	0.000
S2	406	1	0.002	54	0.133	12	0.030	55	0.136	18	0.044	20	0.049	11	0.027	20	0.049
S3	256	12	0.047	34	0.133	3	0.012	4	0.016	13	0.051	0	0.000	7	0.027	16	0.063
N1	62	0	0.000	3	0.048	1	0.016	6	0.097	0	0.000	3	0.048	0	0.000	2	0.032
N2	26	1	0.038	0	0.000	2	0.077	1	0.038	0	0.000	0	0.000	0	0.000	0	0.000
O[*]	12	3	0.250	1	0.083	0	0.000	1	0.083	0	0.000	0	0.000	1	0.083	0	0.000
T	112	56	0.500	8	0.071	2	0.018	2	0.018	3	0.027	1	0.009	1	0.009	2	0.018
S	672	13	0.019	89	0.132	15	0.022	62	0.092	31	0.046	20	0.030	19	0.028	36	0.054
N	88	1	0.011	3	0.034	3	0.034	7	0.080	0	0.000	3	0.034	0	0.000	2	0.023
Total	844		0.083		0.114		0.023		0.081		0.038		0.027		0.024		0.045

[+]Frequency of specific HLA haplotypes in individual subgroups.

[+]Others, unable to be defined.

S5 Table. Major HLA-A-B-T-allele(rs2233947:T) haplotypes in an independent sample of Taiwanese (n=442).

Haplotype A-B-T	Frequency	StdErr*	LowerCL[+]	UpperCL[+]
11:01-40:01-C	0.11096	0.01057	0.09024	0.13167
33:03-58:01-T	0.08232	0.00925	0.06419	0.10045
02:07-46:01-C	0.08047	0.00915	0.06252	0.09841
24:02-40:01-C	0.04347	0.00686	0.03002	0.05691
11:01-13:01-C	0.03416	0.00611	0.02218	0.04614
11:01-15:02-C	0.02646	0.0054	0.01588	0.03705
02:01-40:01-C	0.02452	0.0052	0.01432	0.03472
11:01-27:04-C	0.02333	0.00508	0.01337	0.03329
11:01-55:02-C	0.01896	0.00459	0.00997	0.02796
11:01-51:01-C	0.01803	0.00448	0.00926	0.02681

*Standard error of estimation

[+]Lower and upper confidence limit of estimation.

S6 Table. Power calculations.

Controls	Cases	Prevalence	Disease allele frequency	Power*	Genotype relative risk
10801	**1000**	**0.01**	**0.05**	0.8181	1.77
			0.1	0.7950	1.54
			0.2	0.8199	1.41
		0.05	**0.05**	0.8146	1.78
			0.1	0.8089	1.55
			0.2	0.8125	1.41
		0.1	**0.05**	0.8057	1.79
			0.1	0.8182	1.56
			0.2	0.8031	1.41
10801	**2000**	**0.01**	**0.05**	0.8185	1.55
			0.1	0.8083	1.39
			0.2	0.8031	1.29
		0.05	**0.05**	0.8024	1.55
			0.1	0.7981	1.39
			0.2	0.7974	1.29
		0.1	**0.05**	0.8096	1.56
			0.1	0.7851	1.39
			0.2	0.7904	1.29
10801	**3000**	**0.01**	**0.05**	0.8169	1.46
			0.1	0.8229	1.33
			0.2	0.8460	1.25
		0.05	**0.05**	0.8034	1.46
			0.1	0.8146	1.33
			0.2	0.8417	1.25
		0.1	**0.05**	0.8192	1.47
			0.1	0.8040	1.33
			0.2	0.8363	1.25

***The calculation was based on CaTS. Skol AD, Scott LJ, Abecasis GR and Boehnke M (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 38:209-13.**

Supplementary Figures

S1 Fig. PCA and admixture analysis of TWB and reference populations from the 1000 Genomes Projects.

S2 Fig. The current residence of the four subgroups (N, S2, T1 and S1).

S3 Fig. Allele frequencies of the major alleles of the top SNPs identified in the N-S GWAS.

S4 Fig. Haplotype similarity in the MHC region of chromosome 6 among different subgroups from the TWB detected by fineSTRUCTURE.

S5 Fig. Proportion of the HLA-A*33:03-B*58:01 haplotype in the subgroups of 10,801 TWB participants.

S6 Fig. Significance levels (P values) of differences in the physical examination indices and biochemical measures between the (S2+S3) and the (T1–T4) subgroups.

S7 Fig. The proportion of HLA-A*33:03-B*58:01 in populations residing in east and southeast regions of Asia.

S8 Fig. PCA plots for TWB and 1000G EAS (East Asian) samples with or without SNPs on chromosome 6.

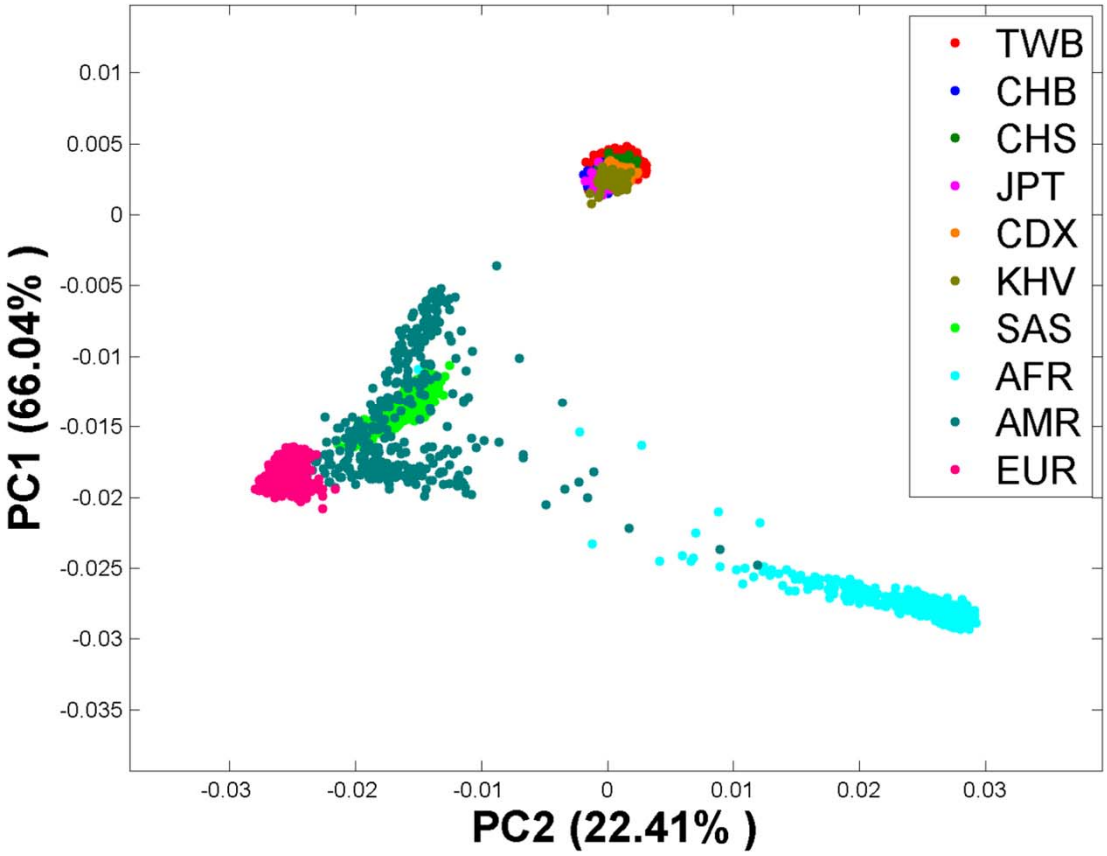
S9 Fig. PCA plots for 1679, 3467 and 7200 subjects from TWB and 1000G East Asian samples.

S10 Fig. Detection of uncertain kinship and ethnicity outliers among the genotyped samples.

S1 Fig. PCA and admixture analysis of TWB and reference populations from the 1000 Genomes Projects.

(A) the scatter plot of the first two PCs showed different continental groups. (B) PCA plot for the TWB (n = 10,801) and 504 Asian samples of 1000G. A north-to-south trend, from the JPT, CHB, CHS, TWB, KHV and CDX samples. (C) the scatter plot of the first two PCs showed that the Min-nan and Hakka groups could not be clearly discriminated.(D) the mean value of PC1 for self-reported origins of the TWB samples, where the the position along PC1 correlates roughly with self-reported N-S cline. (E) the individual ancestry-assigning probabilities under the best model of K=9 ancestry populations. All East Asian groups (TWB, CHB, CHD, JPT, KHV,CDX) shared higher proportions of the ancestry-assigning probabilities of K5 and K9. The abbreviations of individual groups are the same as those in (A).

S1A Fig.



Reference populations in 1000 Genome

AMR (Americas)
CLM Colombian in Medellin, Colombia
MXL Mexican Ancestry in Los Angeles, California
PEL Peruvian in Lima, Peru
PUR Puerto Rican in Puerto Rico

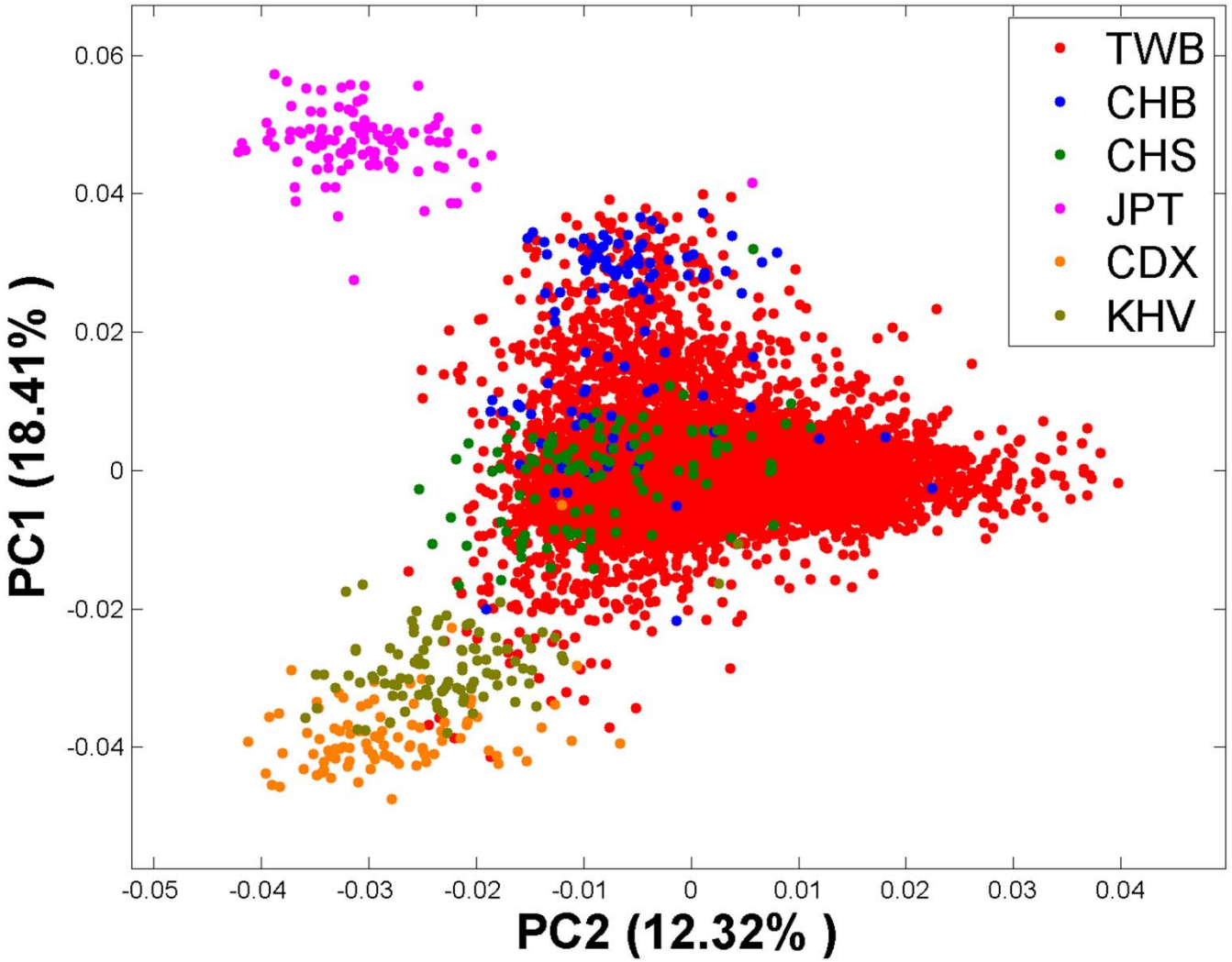
SAS (Southern Asians)
BEB Bengali in Bangladesh
GIH Gujarati Indian in Houston,TX
ITU Indian Telugu in the UK
PJL Punjabi in Lahore,Pakistan
STU Sri Lankan Tamil in the UK

EAS (East Asians)
CDX Chinese Dai in Xishuangbanna, China
CHB Han Chinese in Bejing, China
CHD Chinese in Denver, Colorado
CHS Southern Han Chinese, China
JPT Japanese in Tokyo, Japan
KHV Kinh in Ho Chi Minh City, Vietnam

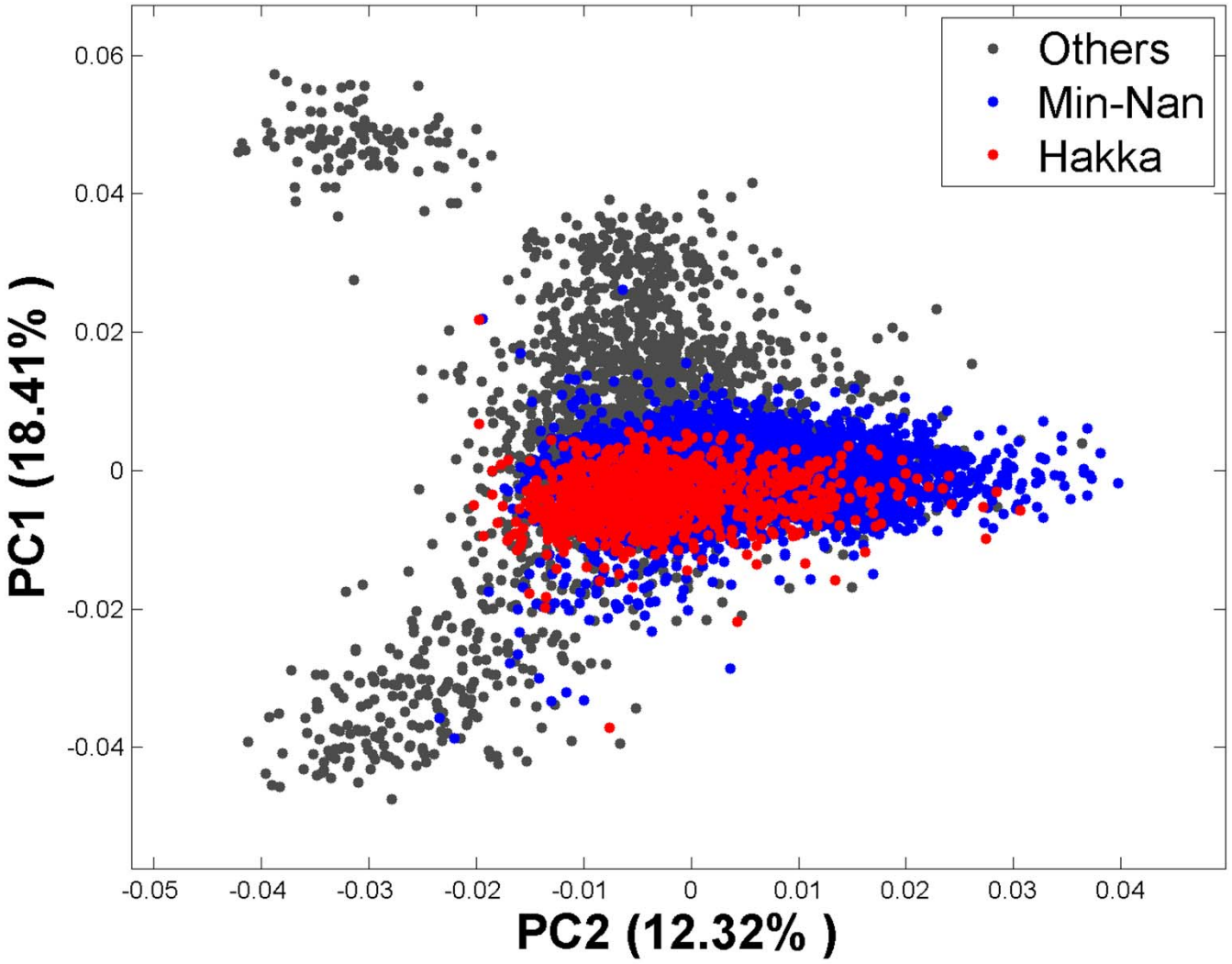
EUR (Europeans)
CEU Utah residents with Northern and Western European ancestry
IBS Iberian populations in Spain
FIN Finnish in Finland
GBR British in England and Scotland
TSI Toscani in Italy

AFR (Africans)
ACB African Caribbean in Barbados
ASW African Ancestry in Southwest US
ESN Esan in Nigeria
GWD Gambian in Western Division, The Gambia
LWK Luhya in Webuye, Kenya
MSL Mende in Sierra Leone
YRI Yoruba in Ibadan, Nigeria

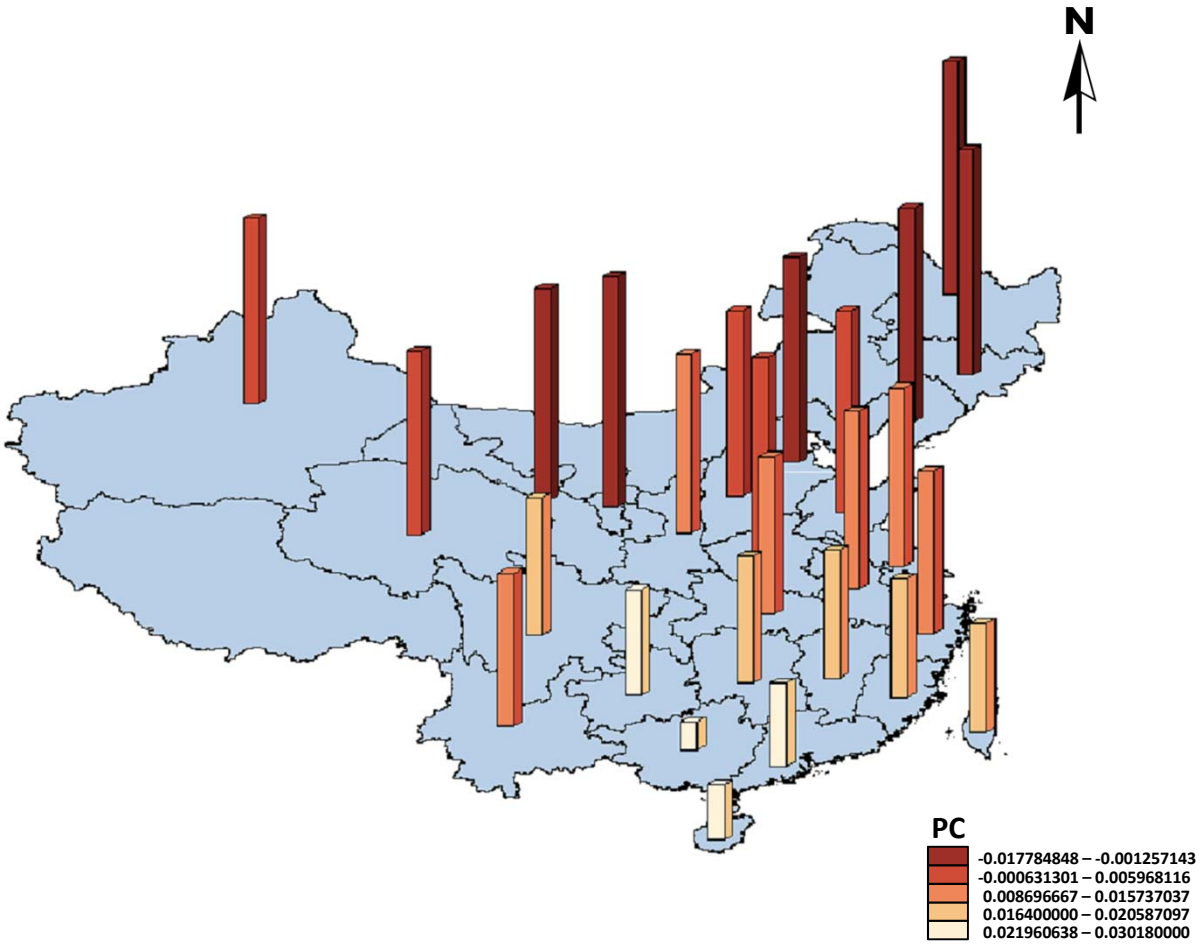
S1B Fig


TWB
CHB
CHS
JPT
CDX
KHV
PC1 (18.41%)
PC2 (12.32%)
0.06
0.04
0.02
0
-0.02
-0.04
-0.05
-0.04
-0.03
-0.02
-0.01
0
0.01
0.02
0.03
0.04

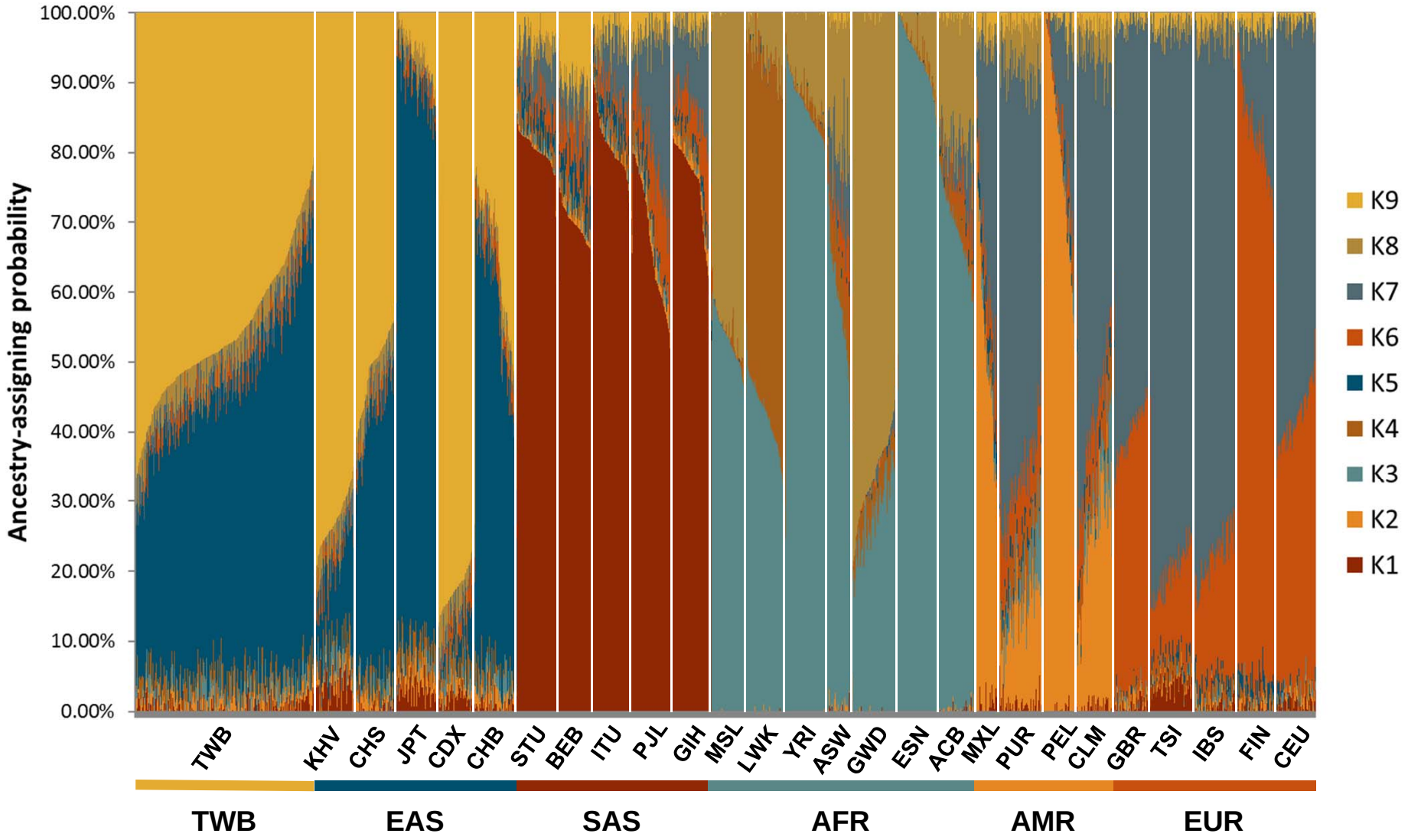
S1C Fig


Others
Min-Nan
Hakka
0.06
0.04
0.02
0
-0.02
-0.04
PC1 (18.41%)
-0.05
-0.04
-0.03
-0.02
-0.01
0
0.01
0.02
0.03
0.04
PC2 (12.32%)

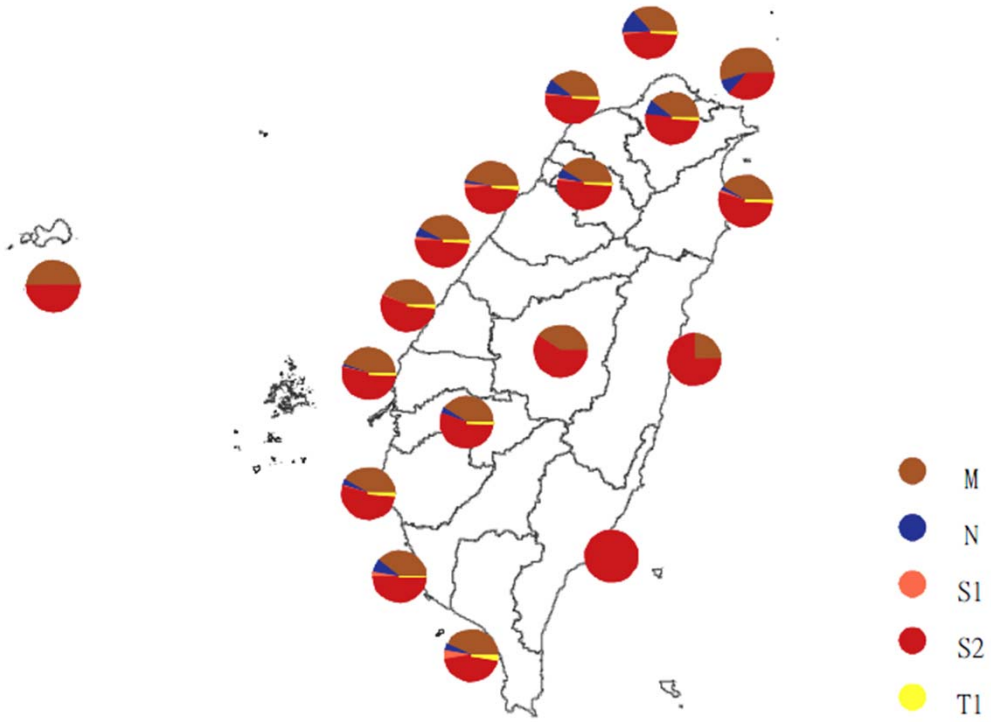
S1D


N
PC
-0.017784848 – -0.001257143
-0.000631301 – 0.005968116
0.008696667 – 0.015737037
0.016400000 – 0.020587097
0.021960638 – 0.030180000

S1E

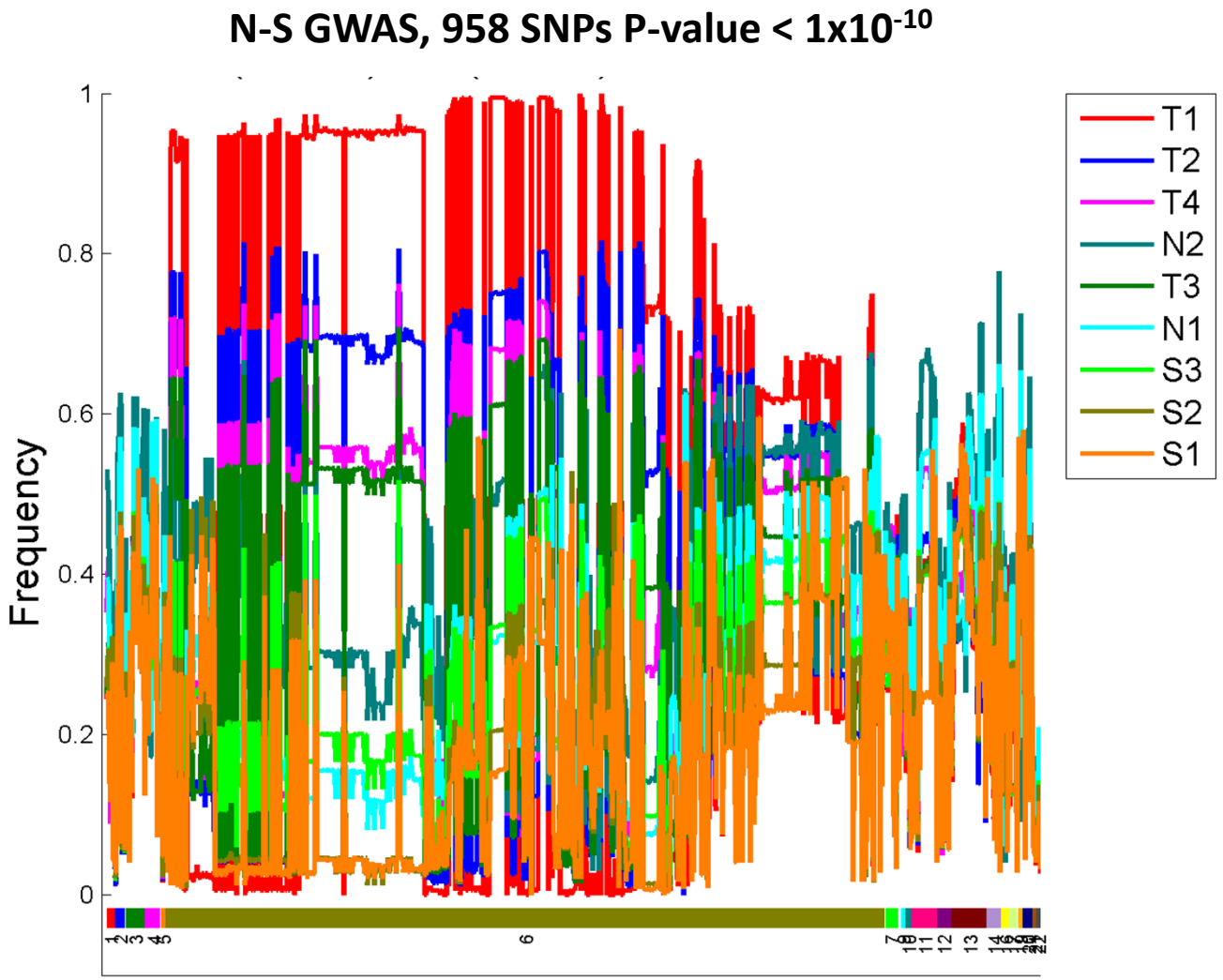

Ancestry-assigning probability
100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
K9
K8
K7
K6
K5
K4
K3
K2
K1
TWB
KHV
CHS
JPT
CDX
CHB
STU
BEB
ITU
PJL
GIH
MSL
LWK
YRI
ASW
GWD
ESN
ACB
MXL
PUR
PEL
CLM
GBR
TSI
IBS
FIN
CEU
TWB
EAS
SAS
AFR
AMR
EUR

S2 Fig. The current residence of the four subgroups (N, S2, T1 and S1)*.

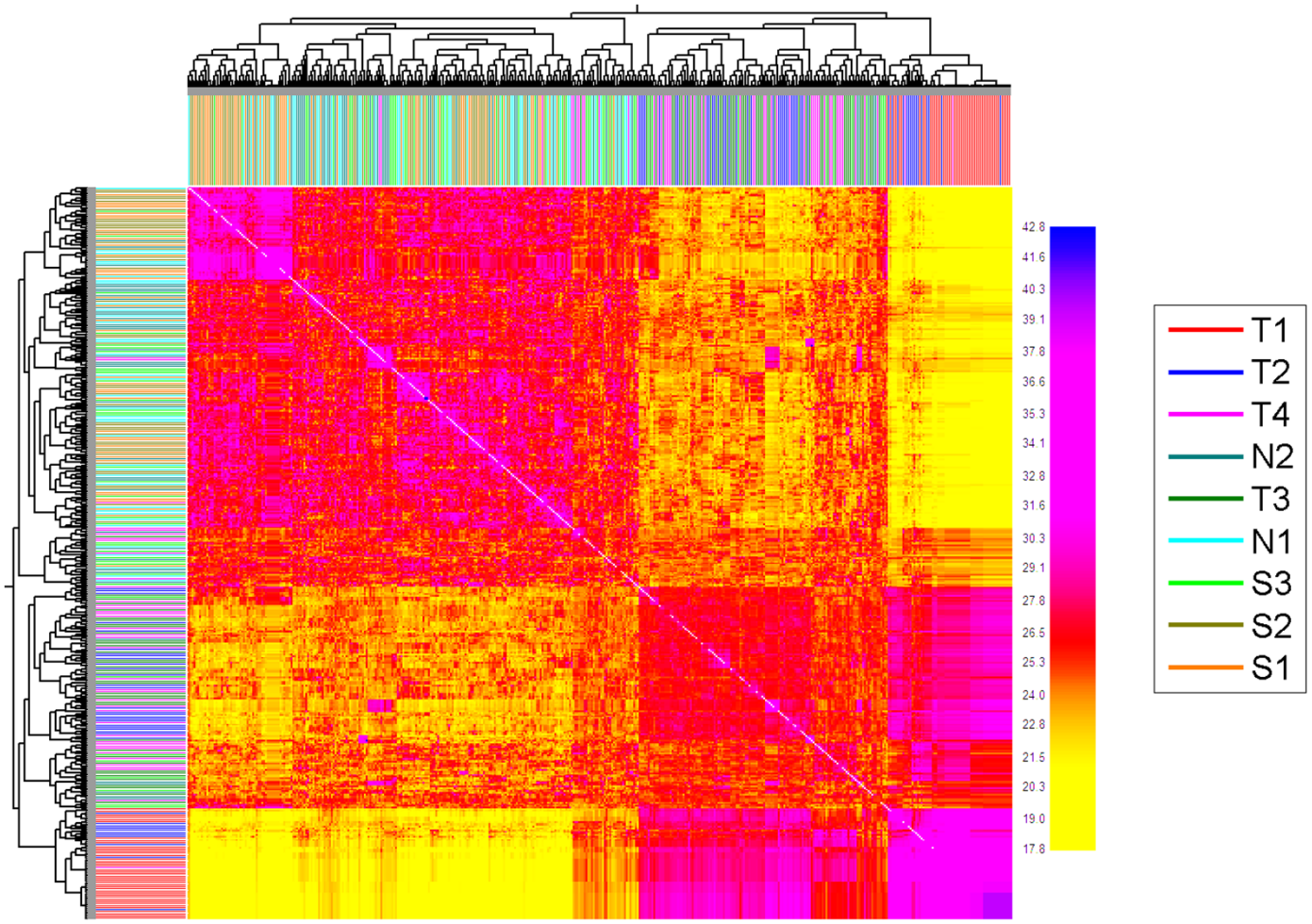


* M represents the mix of the four subgroups.

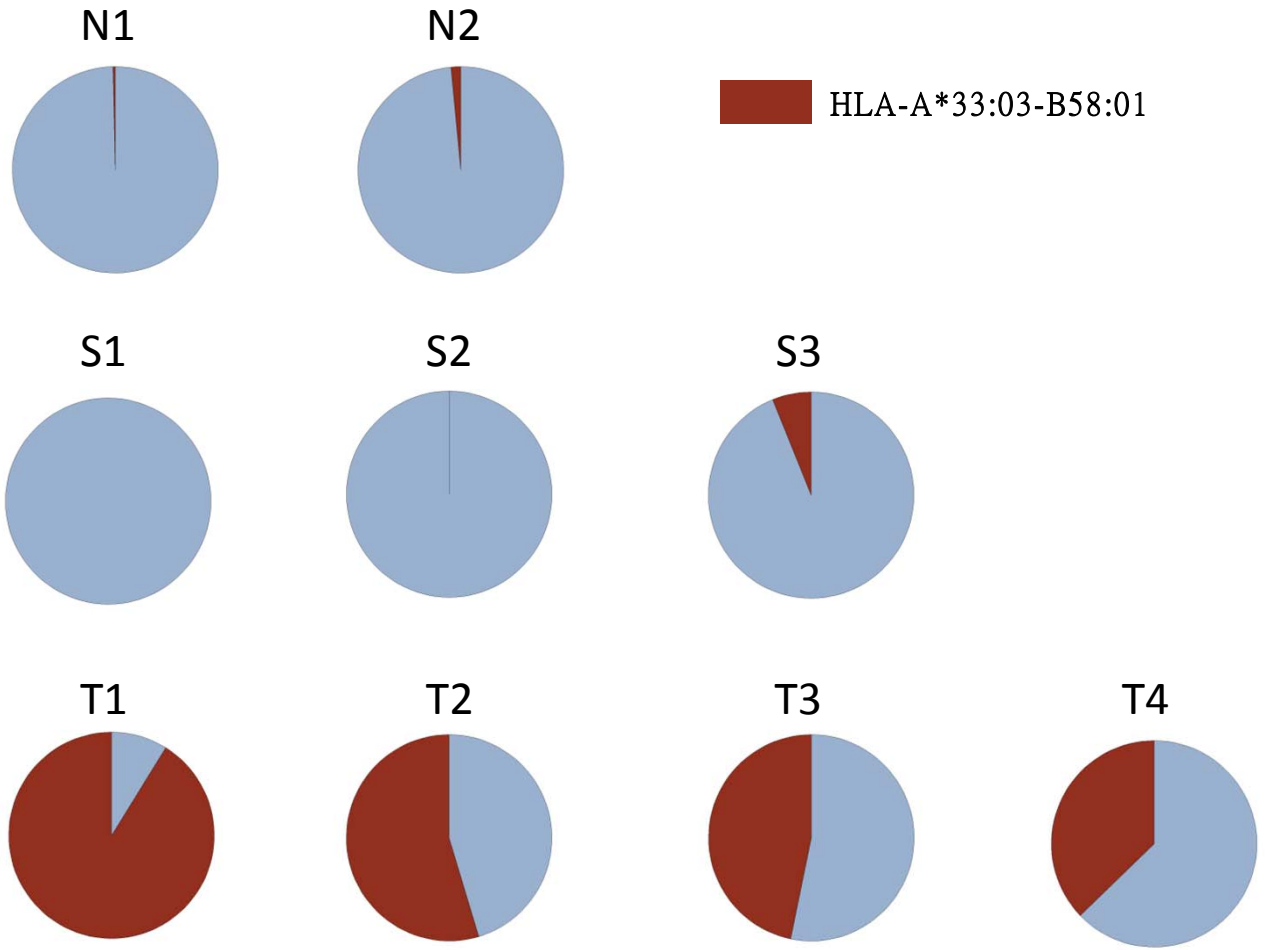
S3 Fig. Allele frequencies of the major alleles of the top SNPs identified in the N-S GWAS.



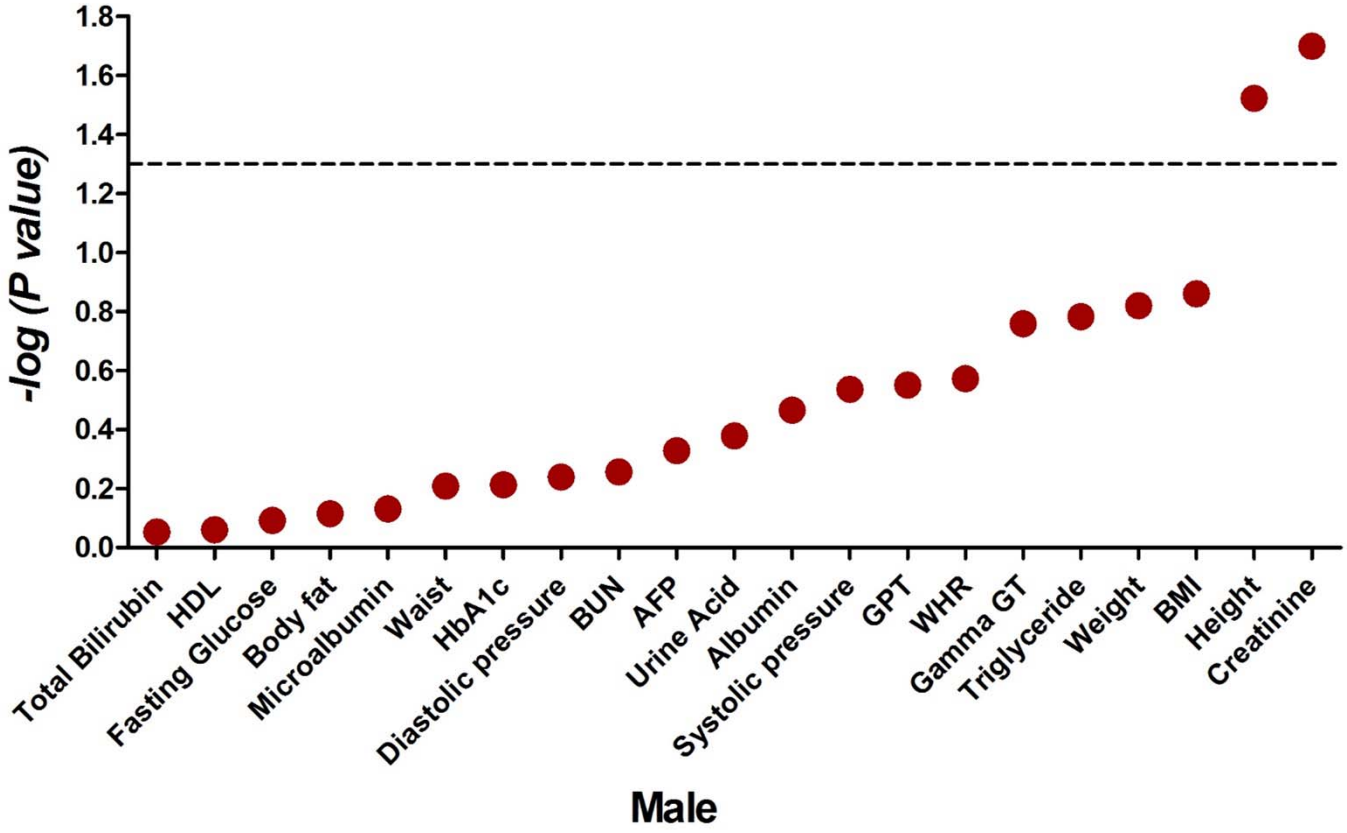
S4 Fig. Haplotype similarity in the MHC region of chromosome 6 among different subgroups from the TWB detected by fineSTRUCTURE. The figure is identical to Fig 4D, but different colors (same as those used in PCA in Fig 1C) were used to represent different subgroups in the S, N and T groups.



S5 Fig. Proportion of the HLA-A*33:03-B*58:01 haplotype in the subgroups of 10,801 TWB participants. Imputation was performed using HIBAG and 36,763 SNPs.

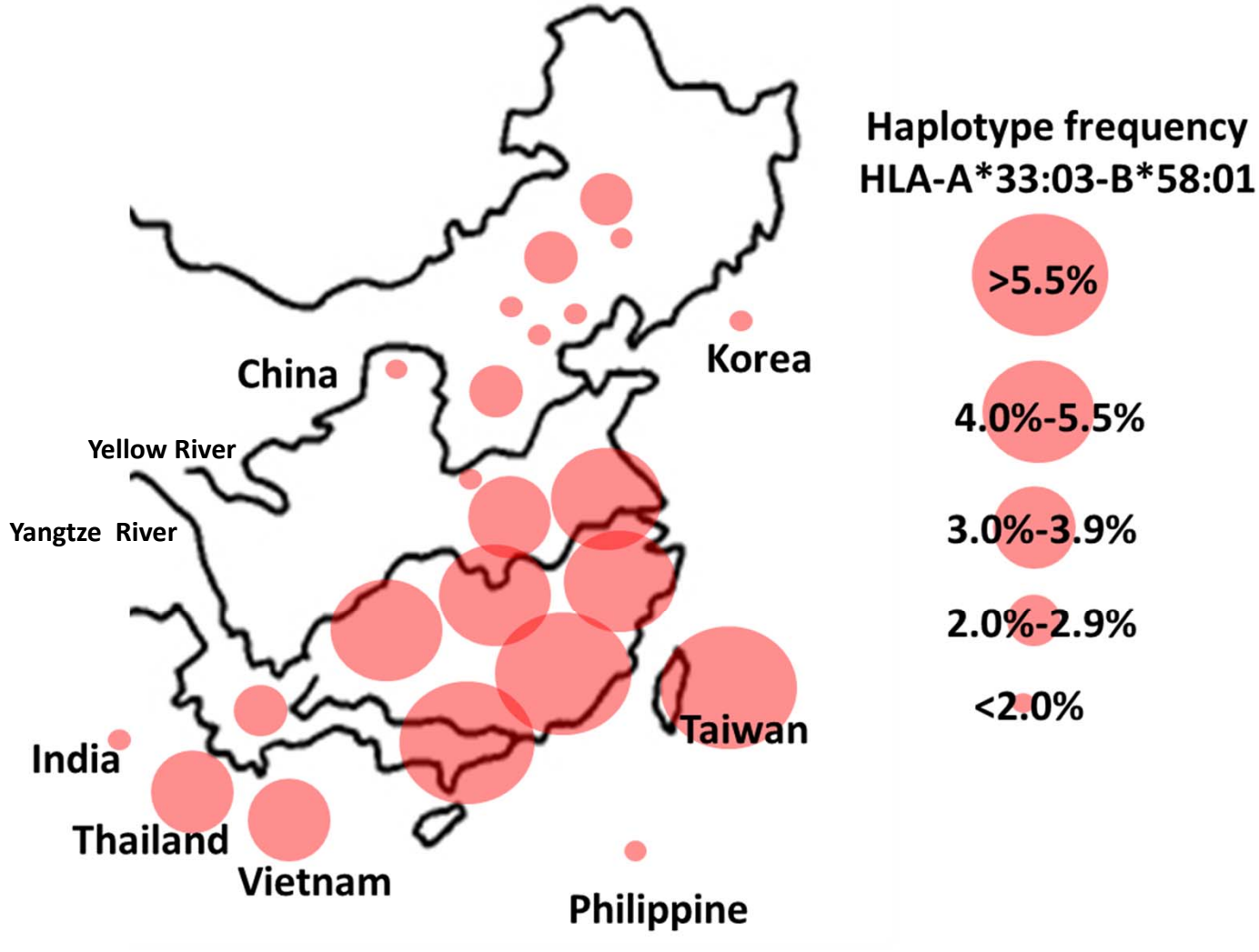


S6 Fig. Significance levels (P values) of differences in the physical examination indices and biochemical measures between the (S2+S3) and the (T1–T4) subgroups.



S7 Fig. The proportion of HLA-A*33:03-B*58:01 in populations residing in east and southeast regions of Asia.

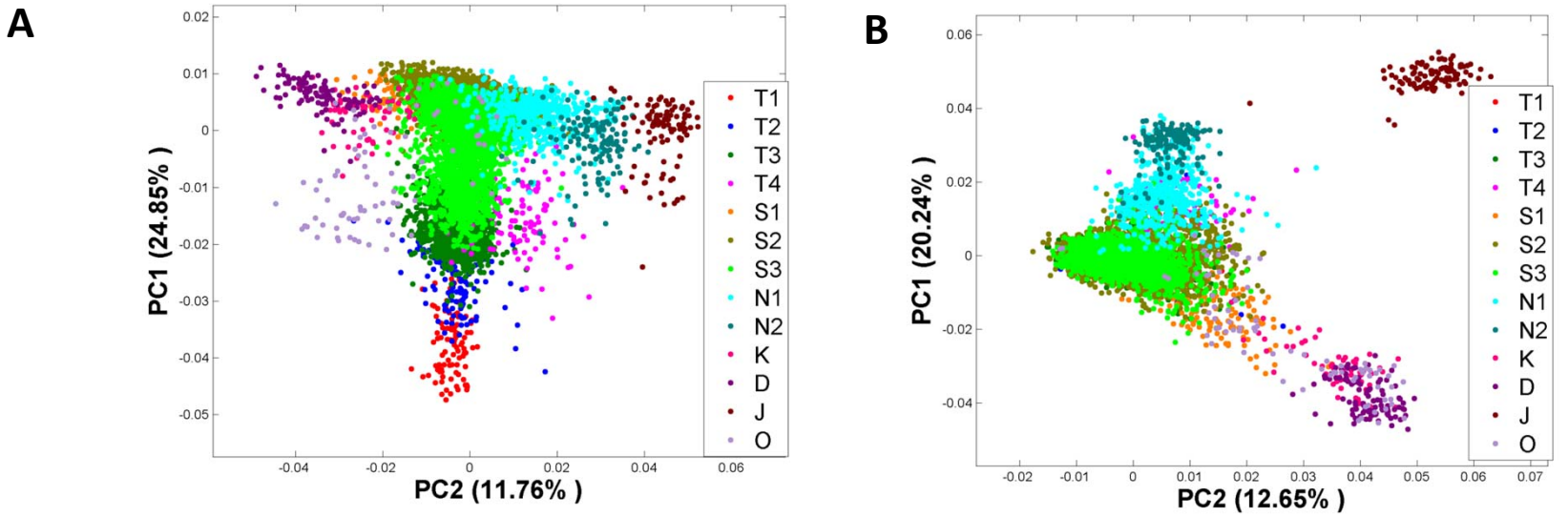
This figure is prepared based on data in http://www.allelefrequencies.net/.



S8 Fig. PCA plots for TWB and 1000G EAS (East Asian) samples with or without SNPs on chromosome 6.

PCA was performed for the TWB samples and for the EAS samples from the 1000G Project. The analysis was based on a set of 172,899 SNPs with MAFs > 5% in both datasets that were selected to be equally spaced across the human genome. The subgroups of TWB were denoted in parentheses whenever applicable; TWB (others) denoted the TWB samples not specified in the plot.

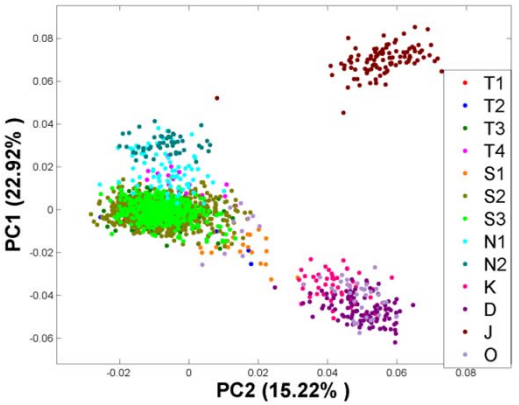
(A) PCA was carried out with the full set of samples and the full set of SNPs. The PCA plot showed a tripartite population structure as found in Fig. 2A, although the relative positions of PC1 and PC2 had been swapped. (B)PCA was carried out with the full set of samples and a subset of SNPs, in which all SNPs on chromosome 6 were excluded. The PCA plot showed that the subgroups distributed along the north-south gradient (PC1) and that the subgroups of S and T clustered together.

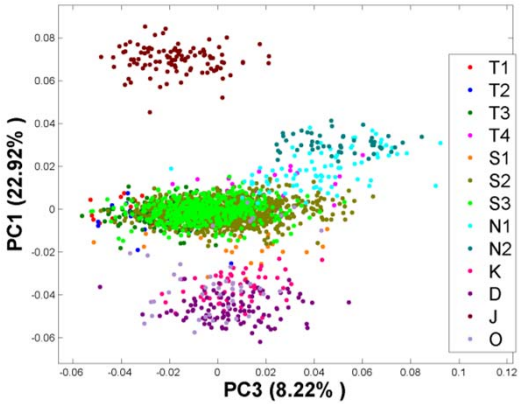


S9 Fig. PCA plots for 1679, 3467 and 7200 subjects from TWB and 1000G East Asian samples.

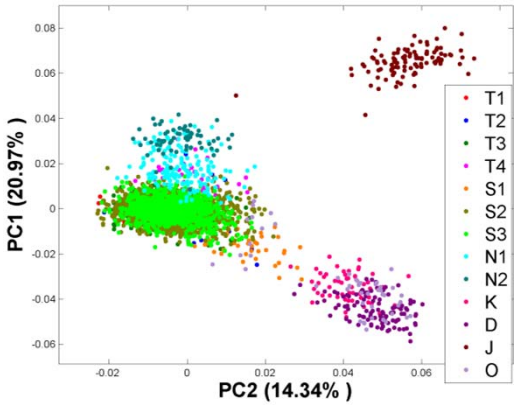
PCA was performed for three subsets of TWB, (A) 1679 subjects (B) 3467 subjects and (C) 7200 subjects, and the 1000G East Asian samples. The analysis was based on a set of 16,824 SNPs with MAFs > 5% in both data sets that were selected to be equally spaced across the human genome. The up-panels are for PC1 and PC2 and the lower panels are for PC1 and PC3. “O “in these figures means others, unable to be defined.

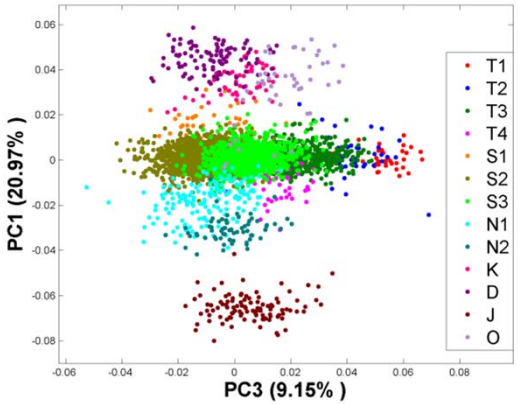
A



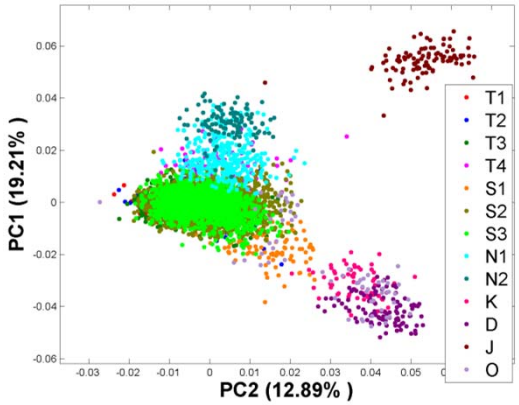


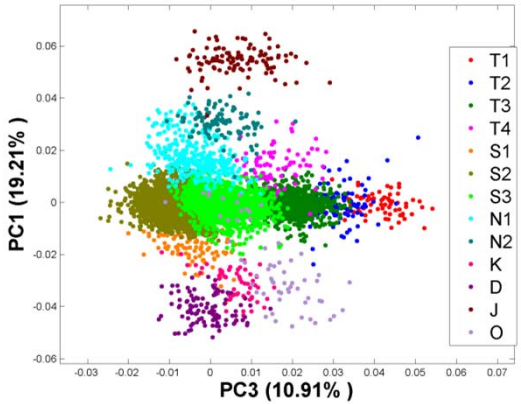
B





C





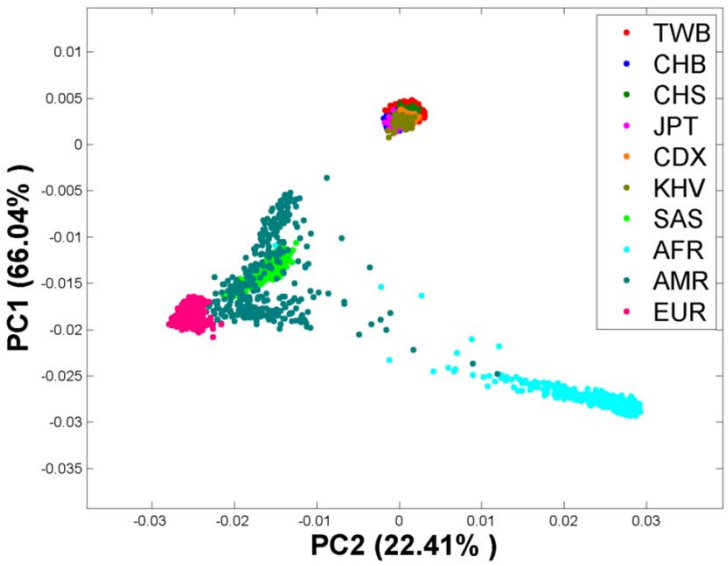
S10 Fig. Detection of uncertain kinship and ethnicity outliers among the genotyped samples.

The detection of uncertain kinship and ethnicity outliers was determined with the following two steps.

A. PCA was performed for TWB samples and 2504 1000G samples. The analysis was based on a set of 11,655 SNPs with minor allele frequencies (MAFs) > 5% in both datasets that were selected to be equally spaced across the human genome. All TWB samples were clustered with the Asian samples of HapMap III. No ethnicity outliers were excluded from further analyses to preserve the maximum information of the genotyped samples.

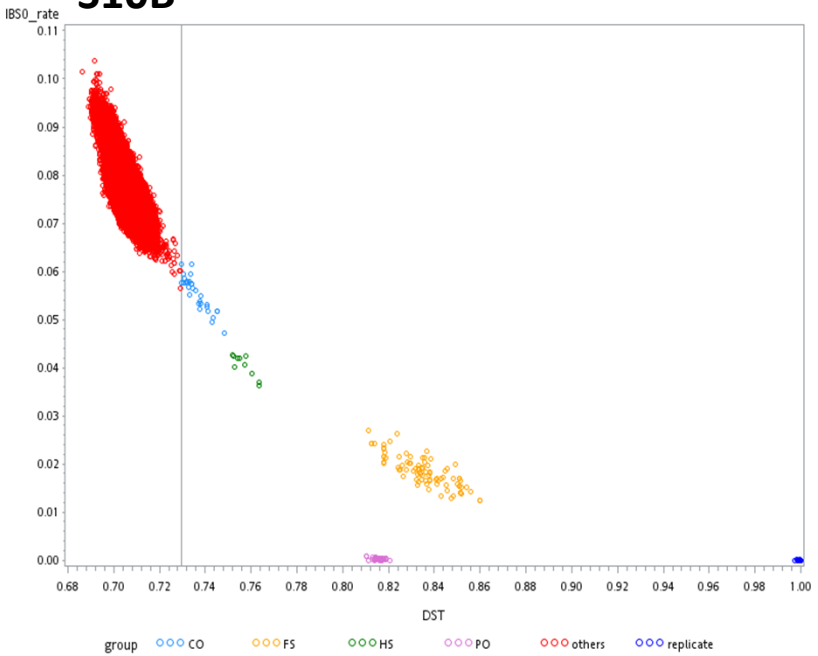
B. IBS plot of the first three batches of TWB samples. Related individuals were clustered based on the pairwise IBS distance between the two individuals of each pair (x axis) and the proportion of SNPs at which the pair shared 0 alleles (IBS0_rate, y axis). The cutoff was determined as the minimum IBS distance score of 100 simulated pairs of cousins based on the allele frequencies of the original TWB data. The samples of the fourth batch were analyzed with the same procedure and the same cutoff. The filtered independent samples from the first three batches and from the fourth batch were pooled together for an additional kinship test.

S10A



TWB	**Taiwan Biobank samples**
CHB	**Han Chinese in Bejing, China**
JPT	**Japanese in Tokyo, Japan**
CHS	**Southern Han Chinese**
CDX	**Chinese Dai in Xishuangbanna, China**
KHV	**Kinh in Ho Chi Minh City, Vietnam**
AFR,	**African**
AMR	**Ad Mixed American**
EAS	**East Asian**
EUR	**European**
SAS	**South Asian**

S10B



Paired kinship groups:

CO: pair of cousins
FS: pair of full siblings
HS: pair of half siblings,
PO: parent-offspring pair
Replicate: DNA samples from the sample subject
Other: pairs not classified