Supplemental Data

AJHG, Volume *85*

Genomic Dissection of Population Substructure of Han Chinese and Its Implication in Association Studies

Shuhua Xu, Xianyong Yin, Shilin Li, Wenfei Jin, Haiyi Lou, Ling Yang, Xiaohong Gong, Hongyan Wang, Yiping Shen, Xuedong Pan, Yungang He, Yajun Yang, Yi Wang, Wenqing Fu, Yu An, Jiucun Wang, Jingze Tan, Ji Qian, Xiaoli Chen, Xin Zhang, Yangfei Sun, Xuejun Zhang, Bailin Wu, and Li Jin

Figure S1 Geographical distribution of Han Chinese sample locations (A), sample size (B), average heterozygosity (C), and average PC1 (D) for twenty-six geographical populations.

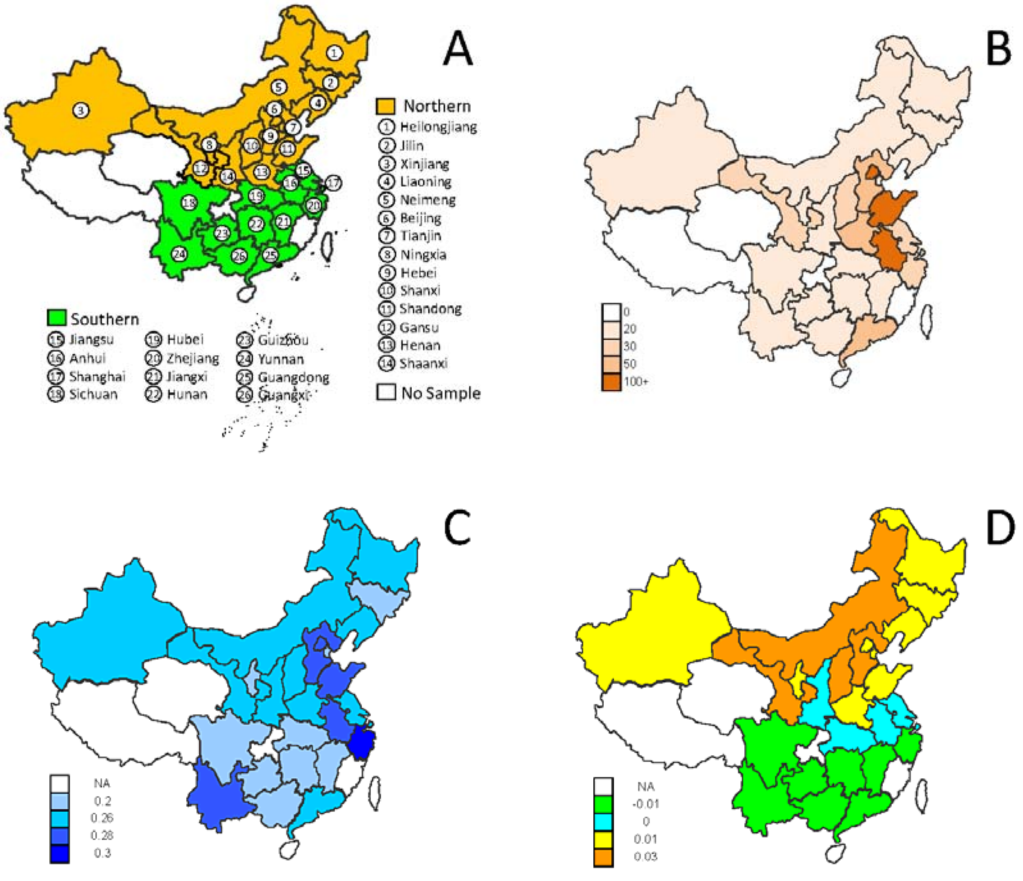

Figure S2 The relationship between genetic distances (F_{ST}) and geographic distances computed using great circle distance. The line in the plot shows the regression line (y = -5E-08x + 0.0015). R^2 for the linear regression of genetic distance on geographic distance is 0.0029 (p=0.23).

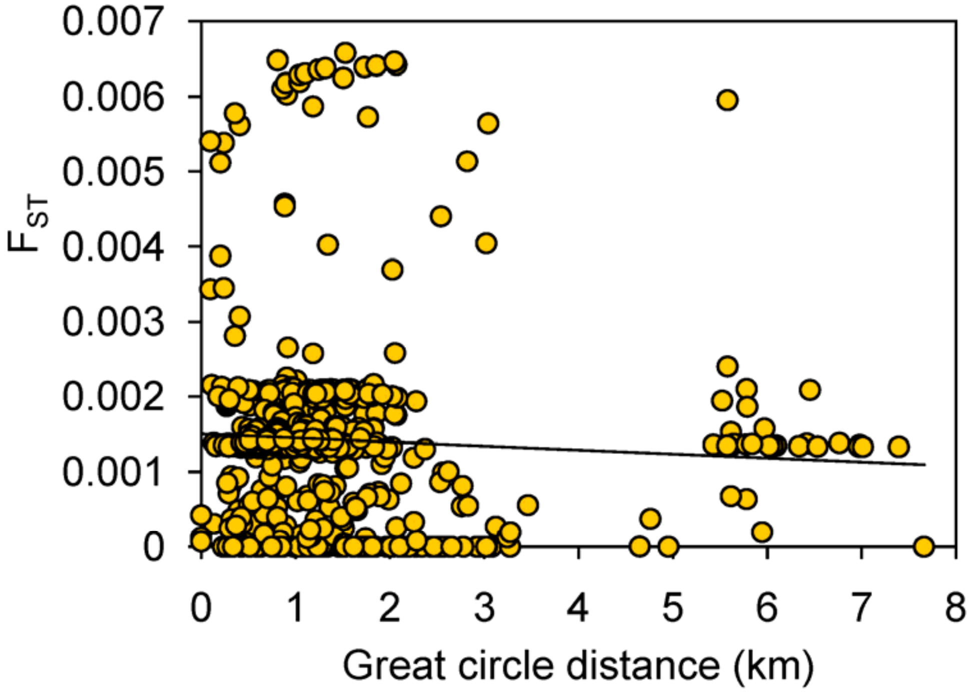

Figure S3 Principal Components Analysis Showing Genetic Affinities of HapMap Han Chinese Samples. **A.** Analysis of the first two Principal Components of CHB and three sub-group of Han Chinese; **B.** Distribution of PC1 for CHB and three sub-group of Han Chinese; **C.** Analysis of the first two Principal Components of CHD and three sub-group of Han Chinese; **D.** Distribution of PC1 for CHD and three sub-group of Han Chinese.

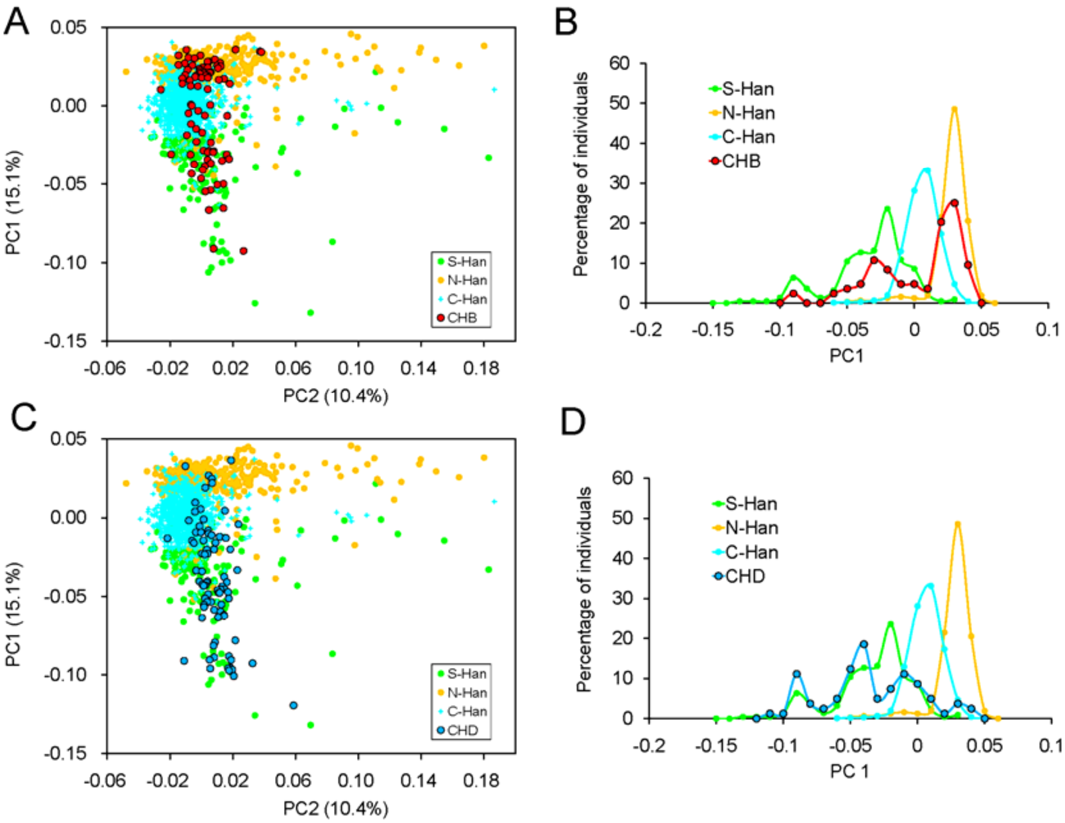

Figure S4 Principal Components Analysis Showing Genetic Affinities of HGDP Han Chinese Samples. A. Analysis of the first two Principal Components of Han-NChina and three sub-group of Han Chinese; **B.** Distribution of PC1 for Han-NChina and three sub-group of Han Chinese; **C.** Analysis of the first two Principal Components of Han and three sub-group of Han Chinese; **D.** Distribution of PC1 for Han and three sub-group of Han Chinese.

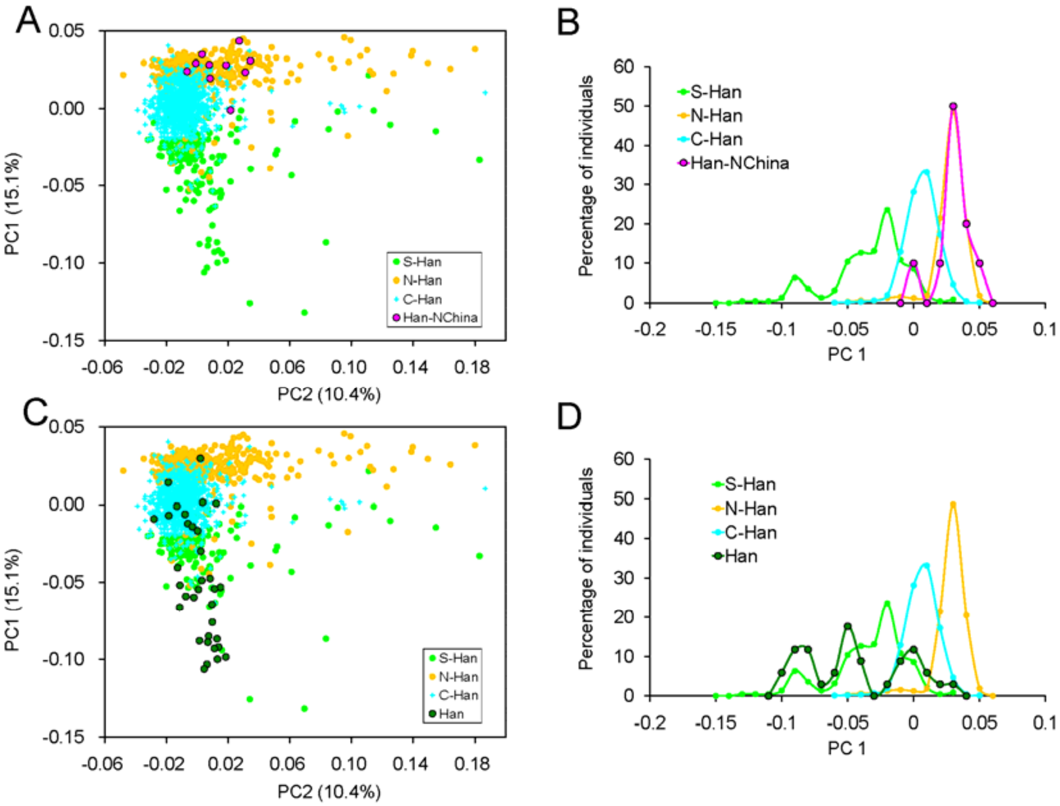

Figure S5 Principal Components Analysis Showing Genetic Affinities of Metropolitan Han Chinese Samples. **A.** Analysis of the first two Principal Components of Shanghai and three sub-group of Han Chinese; **B.** Distribution of PC1 for Shanghai and three sub-group of Han Chinese; **C.** Analysis of the first two Principal Components of Beijing and three sub-group of Han Chinese; **D.** Distribution of PC1 for Beijing and three sub-group of Han Chinese; **E.** Analysis of the first two Principal Components of Guangzhou and three sub-group of Han Chinese; **F.** Distribution of PC1 for Guangzhou and three sub-group of Han Chinese.

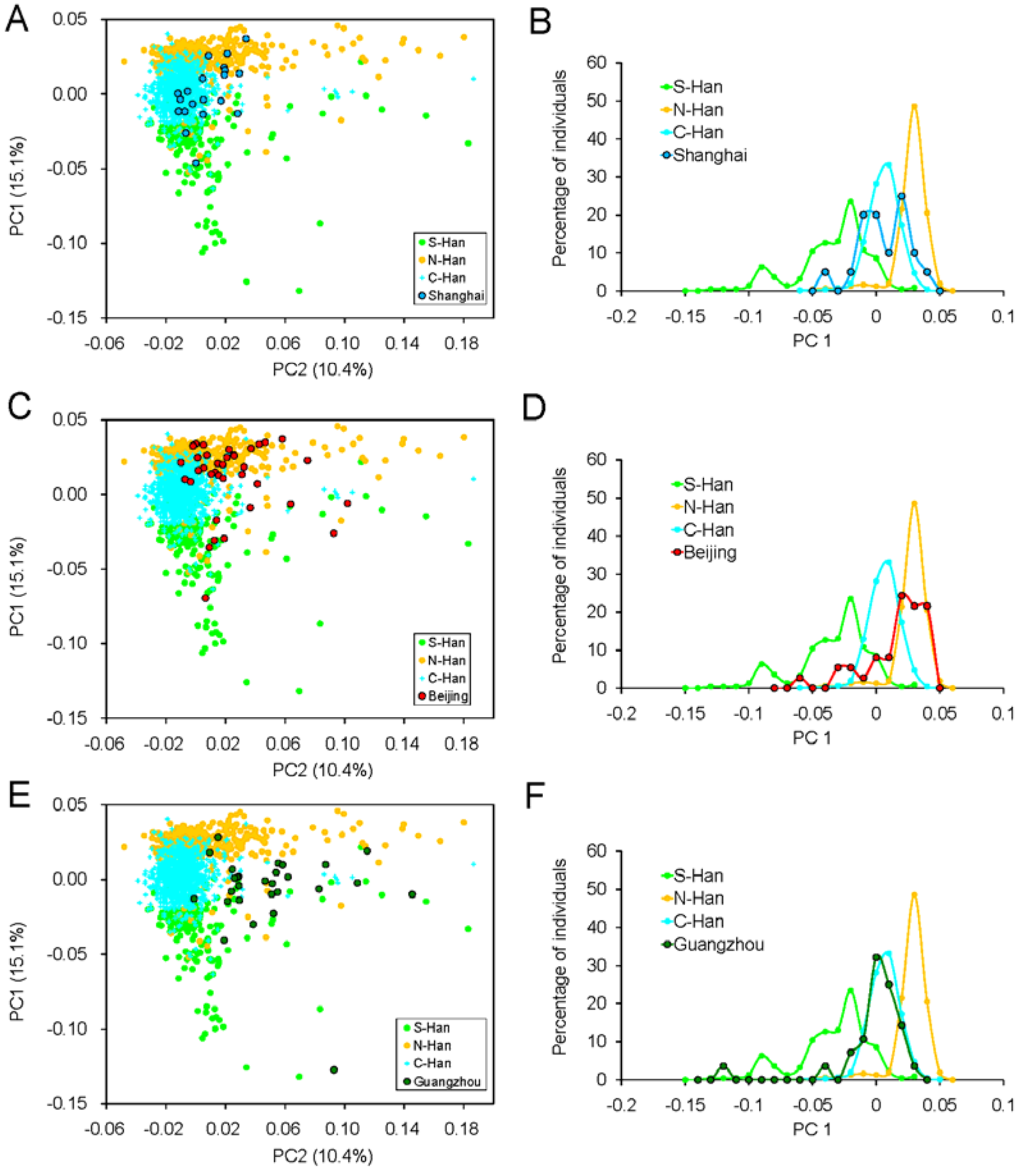

Figure S6 Distribution of F_{ST} between NHC and SHC

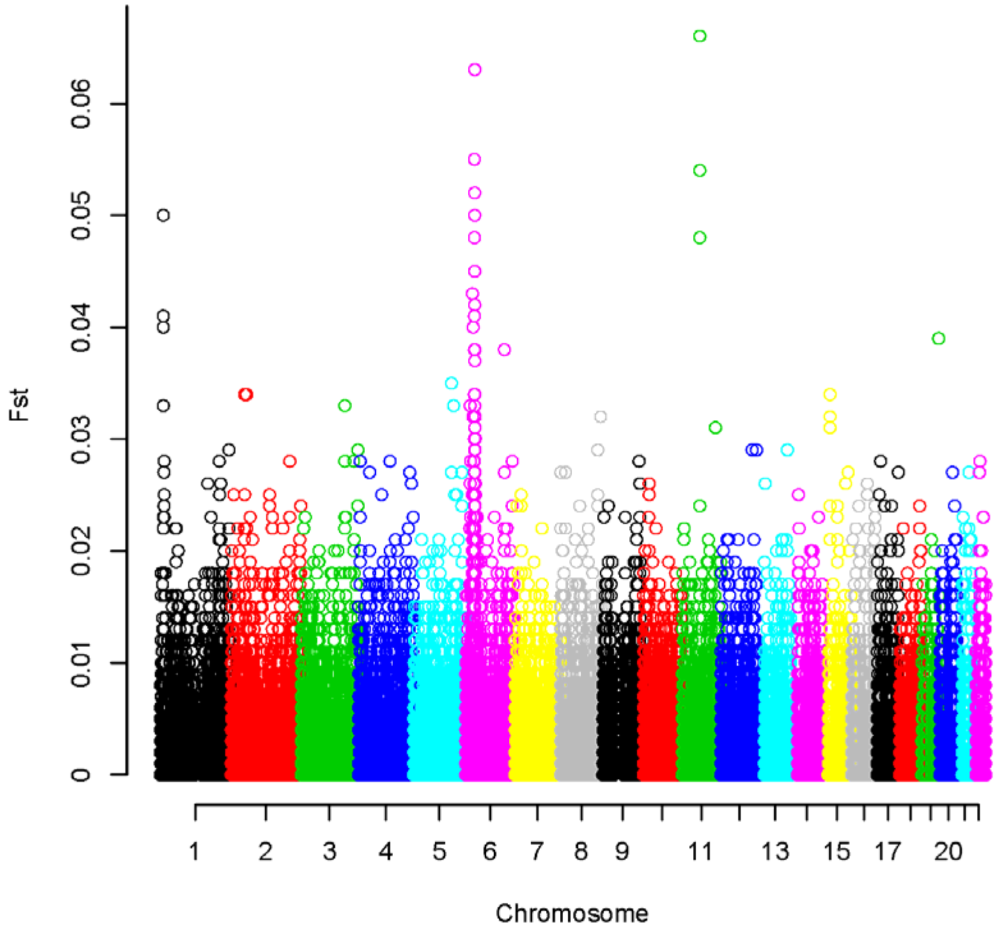

Figure S7 Distribution of F_{ST} between NHC and CHC

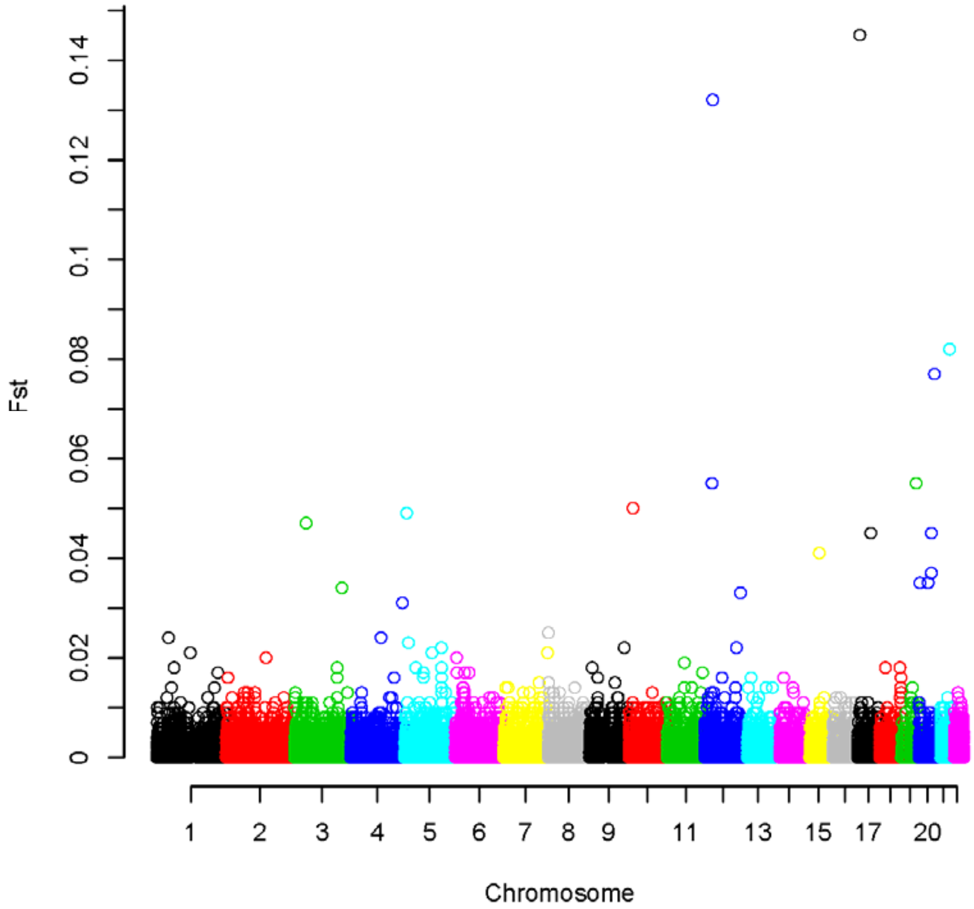

Figure S8 Distribution of F_{ST} between CHC and SHC

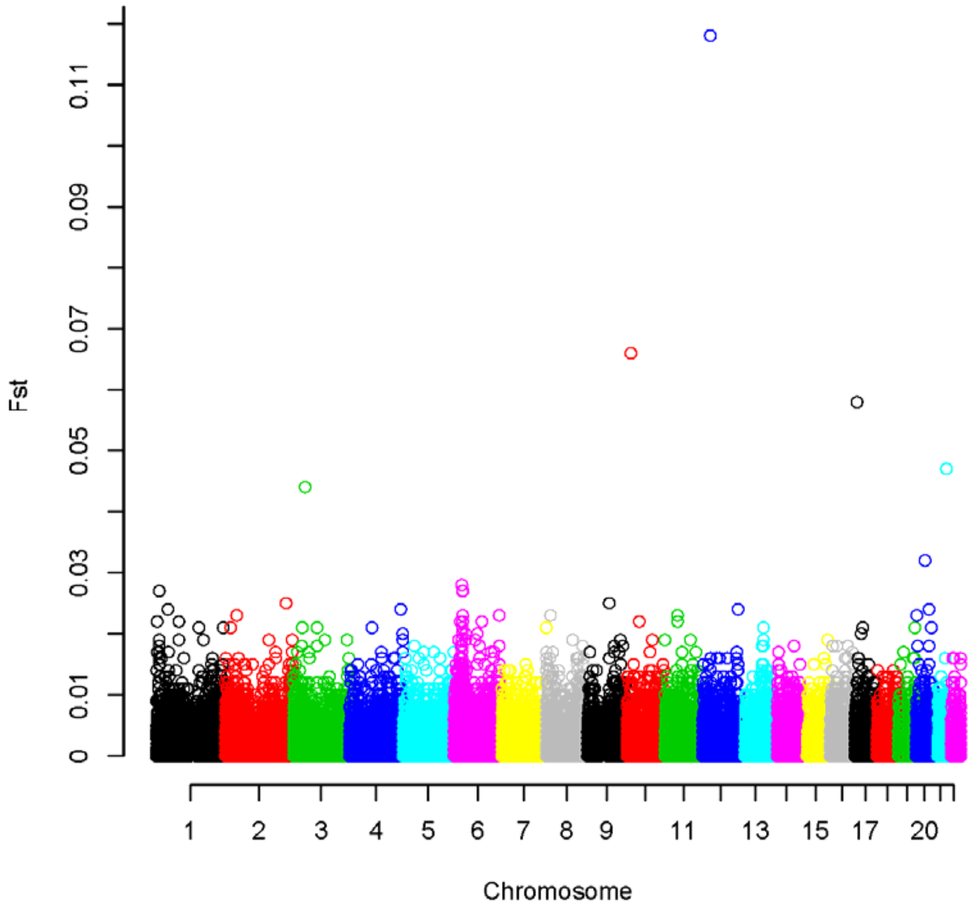

Table S1 Classification of Han Chinese Individuals into the Three Clusters

	NHC	CHC	SHC
N-Han	92.4	4.8	2.8
C-Han	22.5	74.2	3.3
S-Han	1.4	10.9	87.8
Beijing	67.6	18.9	13.5
Shanghai	40.0	50.0	10.0
Guangzhou	17.9	67.9	14.3
CHB	54.8	13.1	32.1
CHD	7.4	24.7	67.9
Han-NChina	90.0	10.0	0.0
Han	5.9	17.6	76.5

The numbers are percentages of individuals belongling to the three clusters as defined in the text.

Table S2 Highly Differentiated SNPs between CHC and SHC

SNP	F_{ST}	*Raw p-value*	*FDR controled p-value*	Chrom	Position	Gene	MIM	Category
rs1975920	0.118	2.79×10^{-24}	5.32×10^{-18}	12	27627253	*PPFIBP1*	603141	intron
rs666101	0.066	4.78×10^{-13}	4.56×10^{-7}	10	14468193	*N/A*	*N/A*	*N/A*
rs3027238	0.058	1.69×10^{-12}	1.08×10^{-6}	17	8075786	*C17orf68*	*N/A*	missense
rs2236479	0.047	2.45×10^{-10}	1.17×10^{-4}	21	45743560	*COL18A1*	120328	intron
rs10510719	0.044	2.64×10^{-7}	0.1007	3	41365143	*ULK4*	*N/A*	intron
rs2378026	0.032	8.59×10^{-7}	0.2731	20	31973312	*N/A*	*N/A*	*N/A*
rs16893673	0.028	1.37×10^{-6}	0.3526	6	28187055	*N/A*	*N/A*	*N/A*
rs11803049	0.027	1.48×10^{-6}	0.3526	1	11851482	*N/A*	*N/A*	*N/A*
rs2517646	0.027	2.69×10^{-6}	0.5699	6	30230554	*TRIM10*	605701	intron
rs2076530	0.027	3.14×10^{-6}	0.5984	6	32471794	*BTNL2*	606000	missense
rs4672775	0.025	3.85×10^{-6}	0.6678	2	216167219	*N/A*	*N/A*	*N/A*
rs1339248	0.025	4.46×10^{-6}	0.7088	9	77815460	*PCSK5*	600488	intron
rs6700677	0.024	6.40×10^{-6}	0.9398	1	42996076	*LEPRE1*	610339	missense
rs10520478	0.024	7.56×10^{-6}	0.9488	4	182532670	*N/A*	*N/A*	*N/A*
rs2662826	0.024	7.72×10^{-6}	0.9488	12	126198334	*N/A*	*N/A*	*N/A*
rs2904065	0.024	8.45×10^{-6}	0.9488	20	46219176	*N/A*	*N/A*	*N/A*
rs11124754	0.023	8.51×10^{-6}	0.9488	2	40729617	*N/A*	*N/A*	*N/A*
rs1265048	0.023	9.90×10^{-6}	0.9488	6	31189388	*PSORS1C1*	*N/A*	nearGene-5
rs9405015	0.023	1.09×10^{-5}	0.9488	6	31263782	*N/A*	*N/A*	*N/A*
rs11964284	0.023	1.09×10^{-5}	0.9488	6	161732616	*PARK2*	602544	intron
rs10503523	0.023	1.13×10^{-5}	0.9488	8	14862012	*SGCZ*	608113	intron
rs751196	0.023	1.19×10^{-5}	0.9488	11	45238288	*SYT13*	607716	intron
rs1810636	0.023	1.21×10^{-5}	0.9488	20	2602925	*N/A*	*N/A*	*N/A*

Table S3 Highly Differentiated SNPs between NHC and CHC

SNP	F_{ST}	*Raw p-value*	*FDR controled p-value*	Chrom	Position	Gene	MIM	Category
rs3027238	0.145	1.43×10^{-40}	2.77×10^{-34}	17	8075786	*C17orf68*	*N/A*	missense
rs1975920	0.132	4.44×10^{-37}	4.29×10^{-31}	12	27627253	*PPFIBP1*	603141	intron
rs2236479	0.082	7.23×10^{-23}	4.66×10^{-17}	21	45743560	*COL18A1*	120328	intron
rs1047972	0.077	8.26×10^{-21}	4.00×10^{-15}	20	54394870	*AURKA*	603072	missense
rs11047889	0.055	9.08×10^{-16}	3.51×10^{-10}	12	25240755	*LYRM5*	*N/A*	intron
rs393412	0.055	6.58×10^{-15}	2.12×10^{-9}	19	53285853	*PLA2G4C*	603602	intron
rs666101	0.05	4.64×10^{-14}	1.28×10^{-8}	10	14468193	*N/A*	*N/A*	*N/A*
rs4702506	0.049	1.55×10^{-13}	3.74×10^{-8}	5	7907289	*N/A*	*N/A*	*N/A*
rs10510719	0.047	2.61×10^{-13}	5.60×10^{-8}	3	41365143	*ULK4*	*N/A*	intron
rs9906298	0.045	2.40×10^{-12}	4.64×10^{-7}	17	47135249	*CA10*	604642	intron
rs2745061	0.045	5.21×10^{-12}	9.15×10^{-7}	20	43505610	*N/A*	*N/A*	*N/A*
rs17238192	0.041	4.41×10^{-11}	7.10×10^{-6}	15	53302238	*RAB27A*	603868	intron
rs2664581	0.037	2.32×10^{-10}	3.45×10^{-5}	20	43237936	*PI3*	182257	missense
rs1810636	0.035	2.66×10^{-10}	3.67×10^{-5}	20	2602925	*N/A*	*N/A*	*N/A*
rs2378026	0.035	5.18×10^{-10}	6.68×10^{-5}	20	31973312	*N/A*	*N/A*	*N/A*
rs10936526	0.034	3.08×10^{-9}	3.72×10^{-4}	3	168470295	*ZBBX*	*N/A*	intron
rs2662826	0.033	5.90×10^{-8}	6.71×10^{-3}	12	126198334	*N/A*	*N/A*	*N/A*

Table S4 Case-Control Simulation Studies

Scenario	Case			Control			λ	False positives (%)				Power (%)			
	NHC	CHC	SHC	NHC	CHC	SHC		Unadjusted	GC	BONF	FDR	Unadjusted	GC	BONF	FDR
II		300 (100%)	0 (0%)		300		1.07	4.95	3.95	0.00	0.00	85.9	84.0	2.6	25.2
XXI		240 (80%)	60 (20%)		300		1.09	5.12	3.96	0.00	0.00	85.9	83.9	2.6	25.2
XXII		210 (70%)	90 (30%)		300		1.13	5.26	3.88	0.00	0.00	85.9	83.5	2.6	25.2
XXIII		180 (60%)	120 (40%)		300		1.14	5.36	3.87	0.00	0.00	85.9	83.2	2.6	25.2
XXIV		150 (50%)	150 (50%)		300		1.26	6.89	3.99	0.00	0.01	85.9	80.2	2.6	26.8
XXV		100 (33%)	200 (67%)		300		1.28	7.25	4.04	0.00	0.01	85.9	80.1	2.6	26.9
VI		0 (0%)	300 (100%)		300		1.30	7.18	3.94	0.00	0.01	85.9	80.0	2.6	26.9
II	0 (0%)	300 (100%)			300		1.07	4.95	3.95	0.00	0.00	85.9	84.0	2.6	25.2
XXVI	60 (20%)	240 (80%)			300		1.09	5.25	4.04	0.00	0.00	85.9	83.9	2.6	25.3
XXVII	90 (30%)	210 (70%)			300		1.11	5.42	4.01	0.00	0.00	85.9	83.5	2.6	25.3
XXVIII	120 (40%)	180 (60%)			300		1.12	5.49	4.06	0.01	0.01	85.9	83.4	2.6	25.4
XXIX	150 (50%)	150 (50%)			300		1.15	5.80	3.95	0.00	0.00	85.9	82.5	2.6	26.0
XXX	200 (67%)	100 (33%)			300		1.18	6.04	3.95	0.01	0.01	85.9	81.5	2.6	26.1
V	300 (100%)	0 (0%)			300		1.20	6.31	3.90	0.02	0.02	85.9	80.9	2.6	26.8

Table S5 Case-Control Simulation Studies

Scenario	Case			Control			λ	False positives (%)				Power (%)			
	NHC	CHC	SHC	NHC	CHC	SHC		Unadjusted	GC	BONF	FDR	Unadjusted	GC	BONF	FDR
III	0 (0%)		300 (100%)			300	1.05	4.81	3.80	0.00	0.00	85.9	84.3	2.6	24.8
XXXI	60 (20%)		240 (80%)			300	1.24	6.53	3.97	0.00	0.01	85.9	80.6	2.6	26.2
XXXII	90 (30%)		210 (70%)			300	1.25	6.68	3.94	0.00	0.01	85.9	80.3	2.6	25.8
XXXIII	120 (40%)		180 (60%)			300	1.27	6.88	3.94	0.00	0.01	85.9	80.1	2.6	26.3
XXIV	150 (50%)		150 (50%)			300	1.28	7.19	3.97	0.00	0.01	85.9	80.1	2.6	26.8
XXXV	200 (67%)		100 (33%)			300	1.35	7.84	3.87	0.00	0.00	85.9	79.7	2.6	27.4
IV	300 (100%)		0 (0%)			300	1.74	11.61	3.69	0.03	0.07	85.9	71.9	2.6	38.4
III		0 (0%)	300 (100%)			300	1.05	4.81	3.80	0.00	0.00	85.9	84.3	2.6	24.8
XXXVI		60 (20%)	240 (80%)			300	1.21	6.45	3.98	0.00	0.00	85.9	80.7	2.6	26.1
XXXVII		90 (30%)	210 (70%)			300	1.23	6.53	3.99	0.00	0.00	85.9	80.7	2.6	26.3
XXXVIII		120 (40%)	180 (60%)			300	1.24	6.67	4.02	0.00	0.00	85.9	80.5	2.6	26.6
XXXIX		150 (50%)	150 (50%)			300	1.24	6.76	3.96	0.00	0.01	85.9	80.3	2.6	26.8
XL		200 (67%)	100 (33%)			300	1.27	6.85	3.85	0.00	0.01	85.9	80.1	2.6	26.8
VI		300 (100%)	0 (0%)			300	1.30	7.18	3.94	0.00	0.01	85.9	80.0	2.6	26.9