Supplemental Data

AJHG*, Volume *85

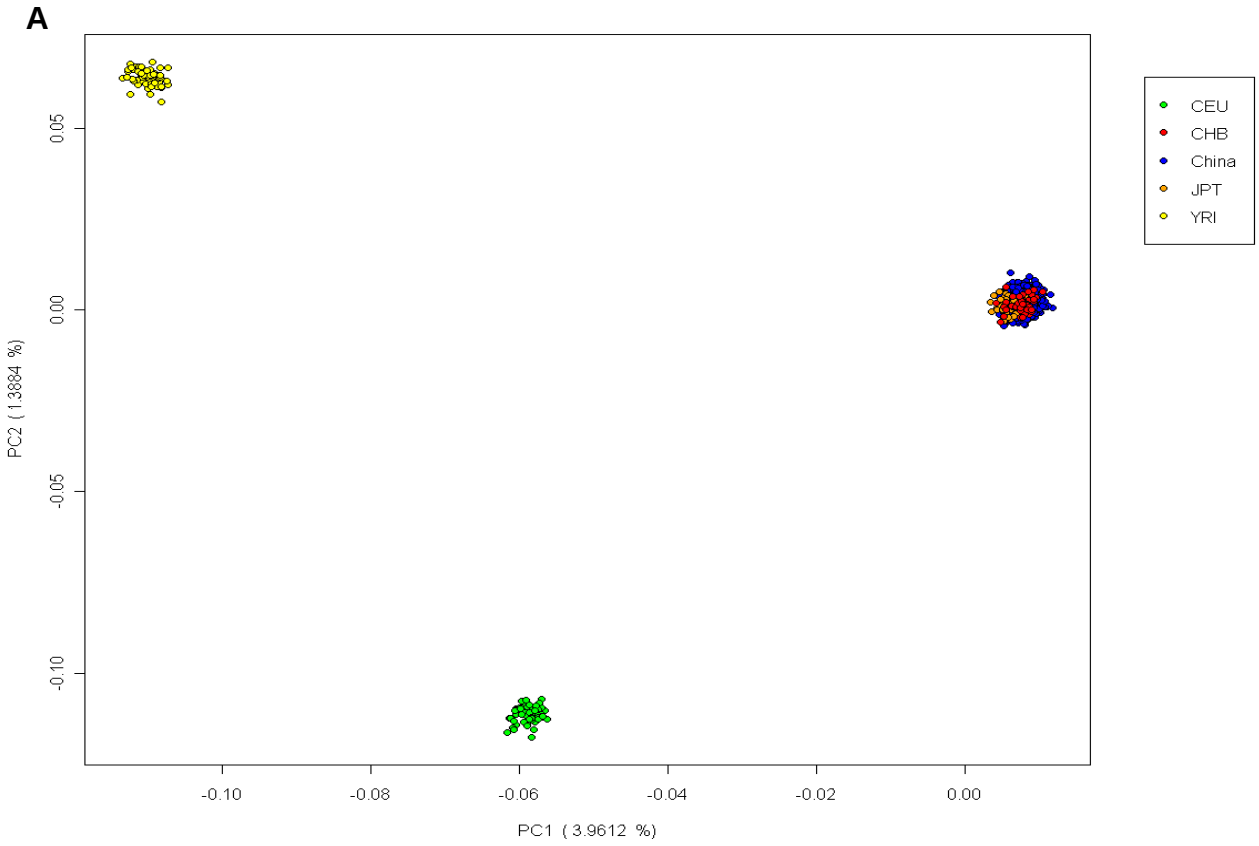
Genetic Structure of the Han Chinese Population

Revealed by Genome-wide SNP Variation

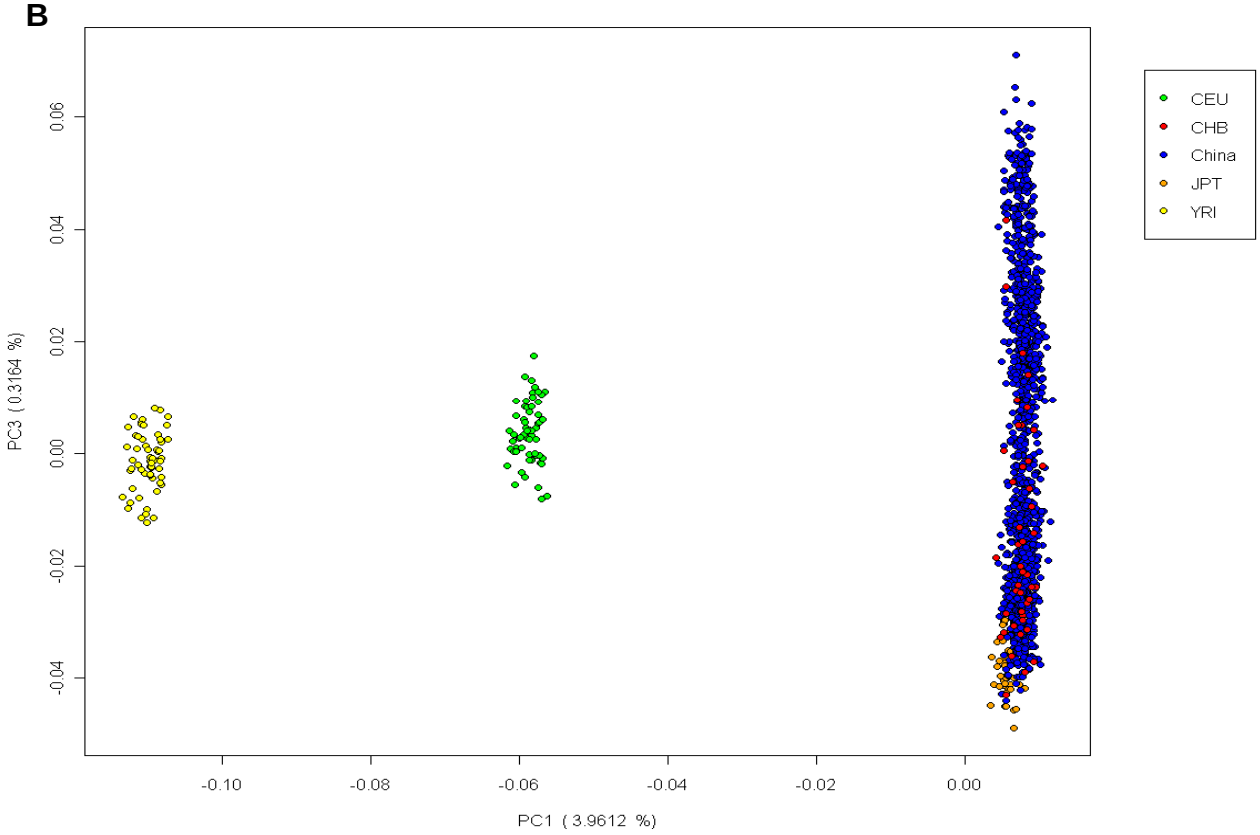
Jieming Chen, Houfeng Zheng, Jin-Xin Bei, Liangdan Sun, Wei-hua Jia, Tao Li, Furen Zhang, Mark Seielstad, Yi-Xin Zeng, Xuejun Zhang, and Jianjun Liu

Figure S1. Relationship between the Chinese and the HAPMAP Populations by PCA

(A) This is a two-dimensional plot based on the principal components (PC) 1 and 2, both of which account for the greatest variations between the populations. The East Asians were clustered together. (B) PC3 accounts for 0.316% of the total variation but failed to separate Japanese from the Chinese. (C) PC4 (0.156%) accounts for the small variation between the Chinese and JPT. All four populations can be seen distinctly here. CHB is situated nicely within the Chinese cluster in each plot below.



B


CEU
CHB
China
JPT
YRI
PC3 (0.3164 %)
0.06
0.04
0.02
0.00
-0.02
-0.04
-0.10
-0.08
-0.06
-0.04
-0.02
0.00
PC1 (3.9612 %)

C

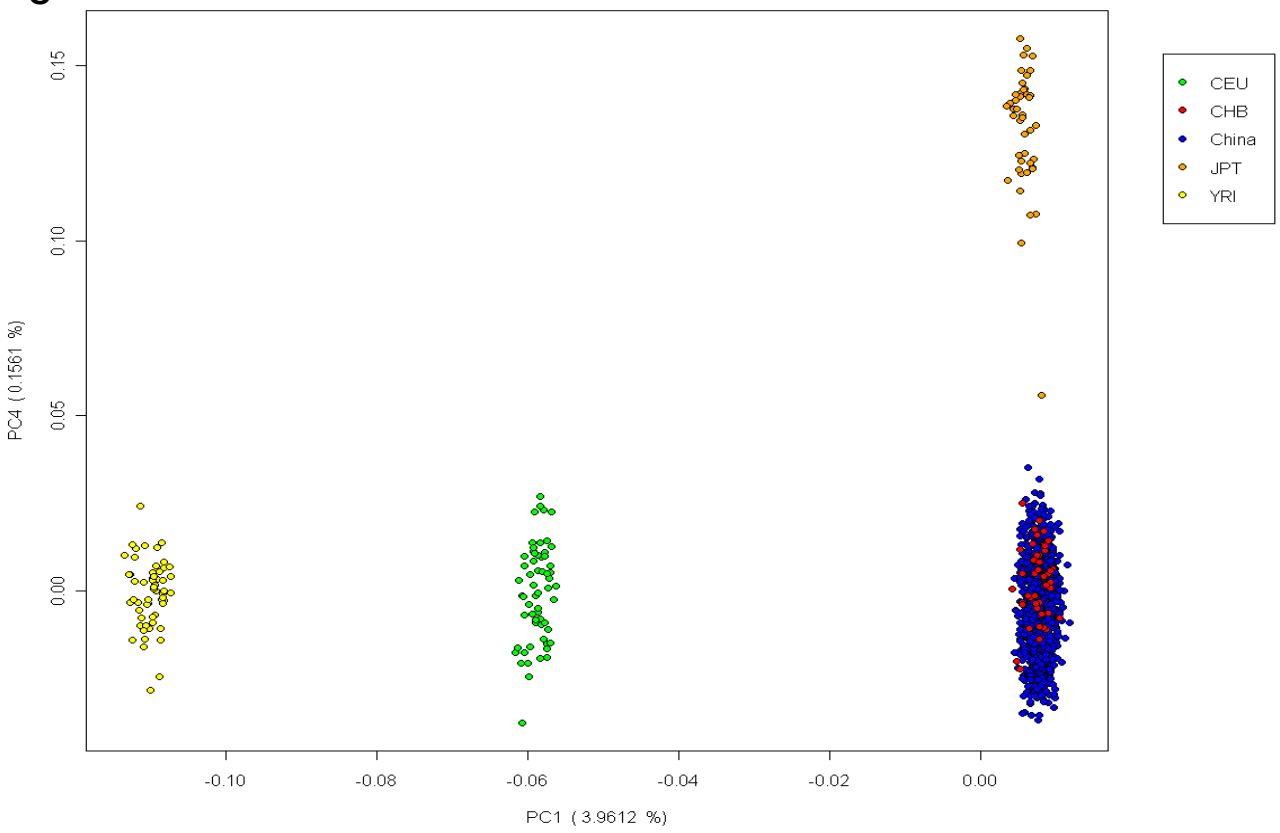

0.15
0.10
0.05
0.00
PC4 (0.1561 %)
-0.10
-0.08
-0.06
-0.04
-0.02
0.00
PC1 (3.9612 %)
CEU
CHB
China
JPT
YRI

Figure S2. Relationship between the East Asians by PCA

PC2 differentiated 44 JPT (orange) from HAPMAP from the Chinese cluster. However, we can see a relationship with the samples from the north of China. 45 CHB (red) from HAPMAP remain in the Chinese cohort (blue).

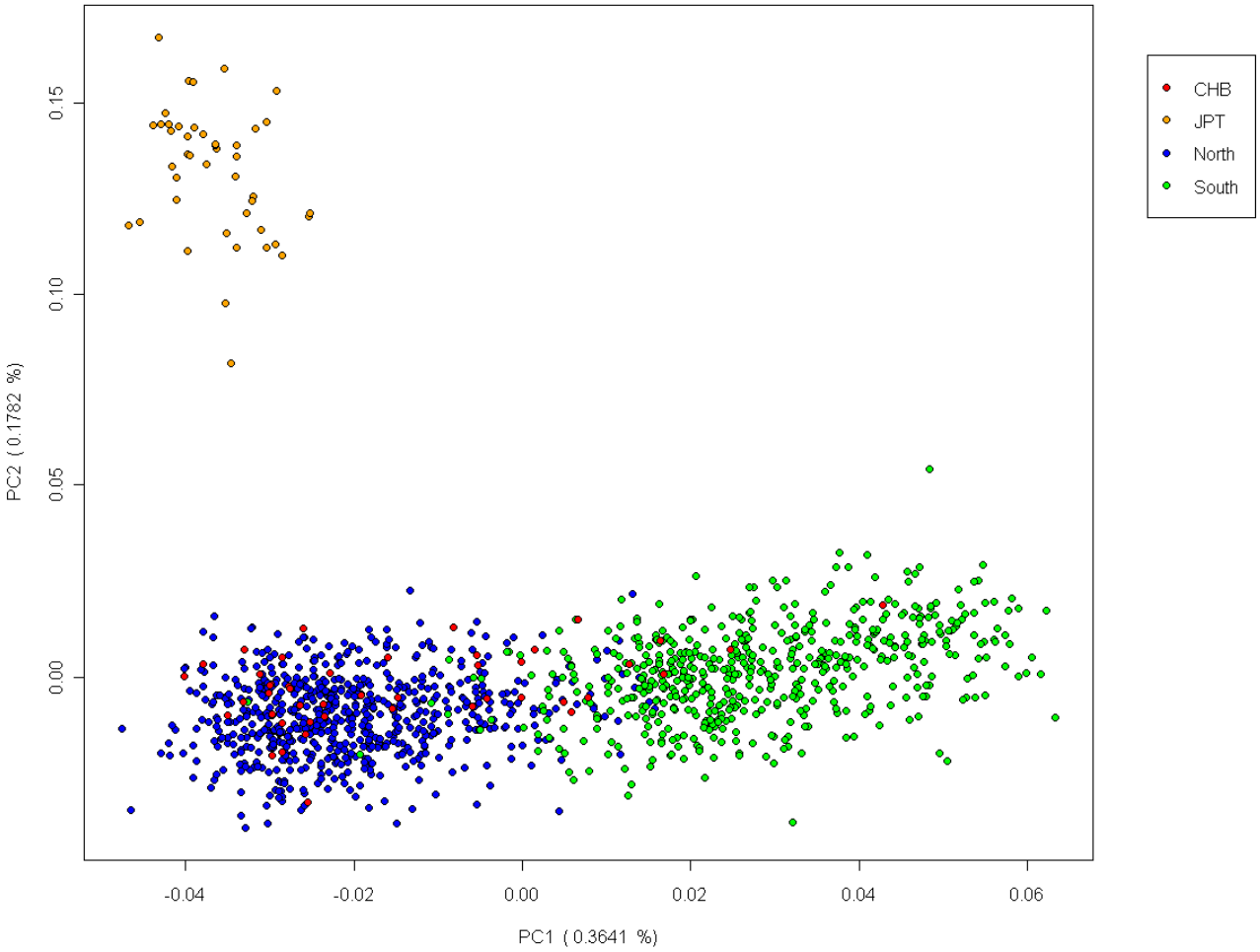


Figure S3. A Pairwise Plot of PC1 to PC10 Derived from 107,565 SNPs in 1109 Samples, Excluding the Three Metropolises

Beyond PC2, the PCs are not informative.

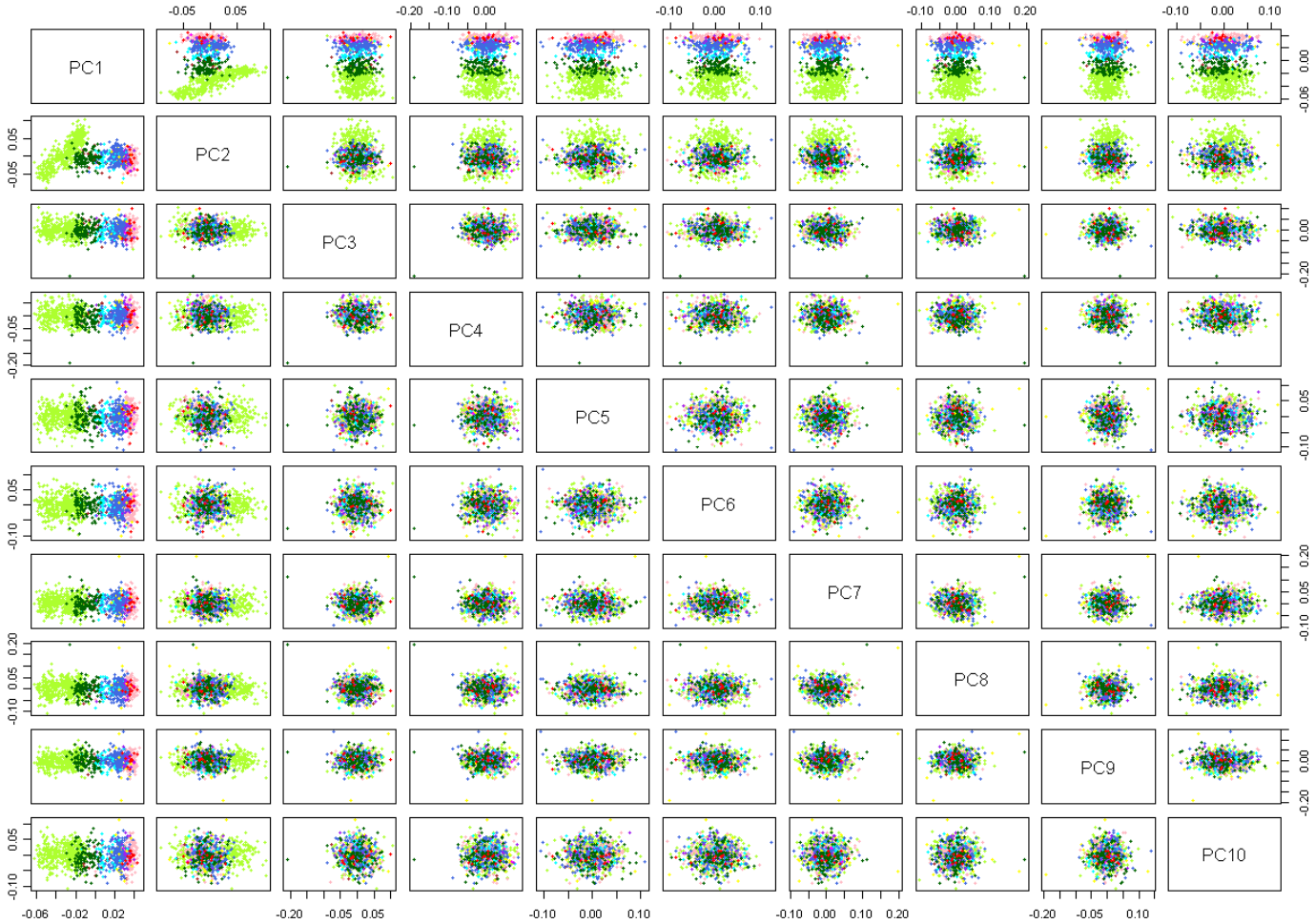


Figure S4. PCA Analysis of the Samples from the Central Provinces

No discernible ‘east-west’ pattern was observed when PC1 is plotted against PC2.

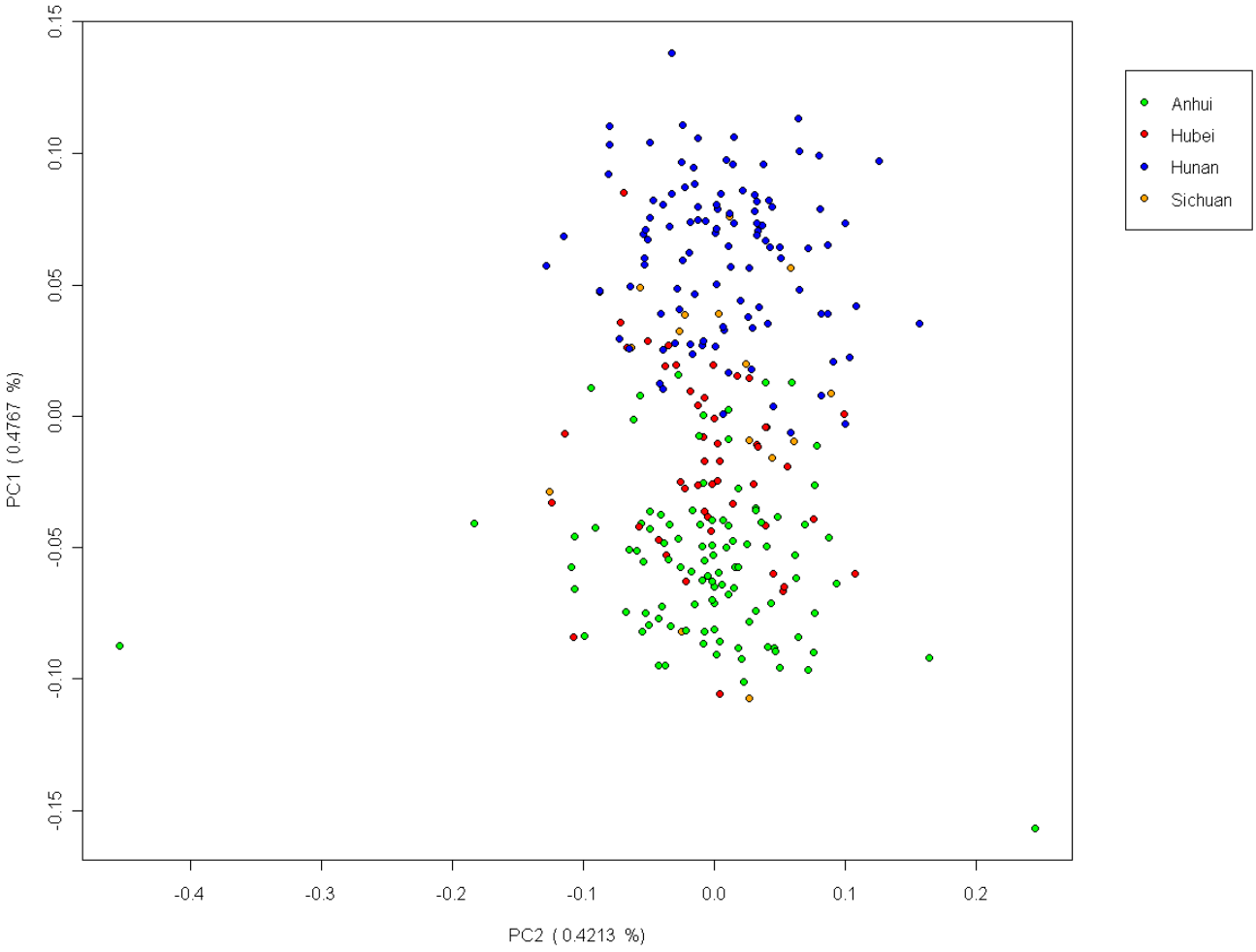


Figure S5. Estimated Population Structure by FRAPPE for K=2 and K=3

We can see that the population structures presented by both STRUCTURE and FRAPPE algorithms produce similar results. Similarly, each individual is represented by a thin vertical line and each province is demarcated by a thick vertical black line. The provinces are arranged from geographical north to south, with JPT on the extreme left, representing the northernmost locality, to Liaoning, the northernmost province, CHB, to Hubei (more central), then to Guangdong, the southernmost province, of which the individuals were rearranged based on its three main dialect groups of Teochew, Hakka and Cantonese. These were then followed by the individuals from the two metropolitan cities of Beijing (CHB) and Shanghai and the Singapore Chinese, of Southeast Asia.

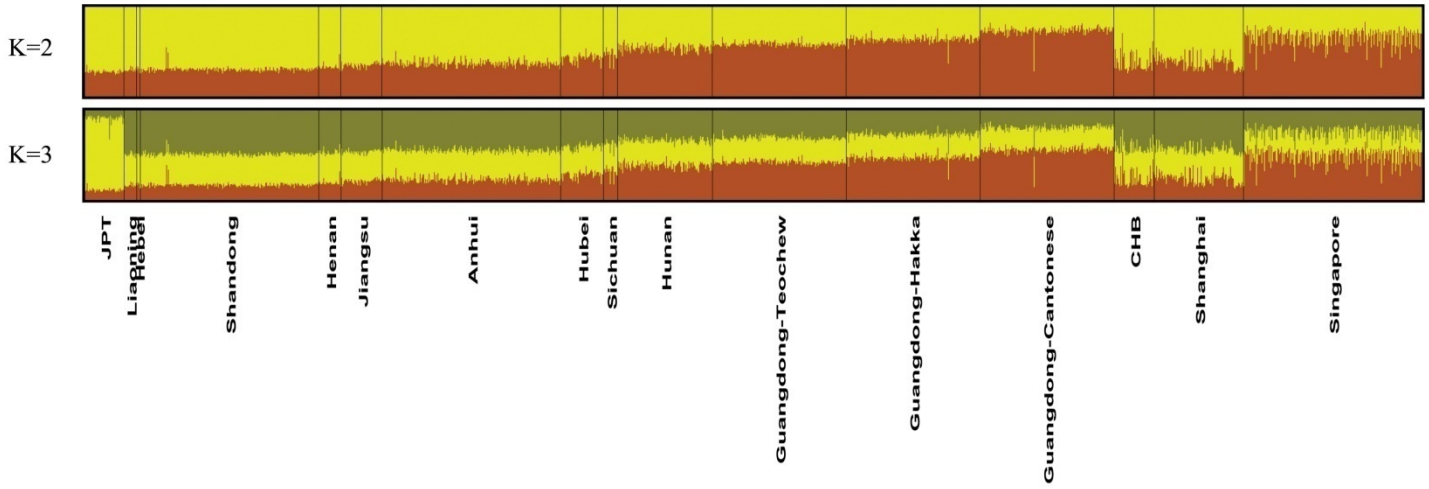


Figure S6. Ordered (A) STRUCTURE and (B) FRAPPE output for K=2 and K=3

All the samples were ordered by province, city or dialect group in terms of their 'brown' membership coefficient, to show the proportion of ancestry of the individuals of the three metropolitan cities.

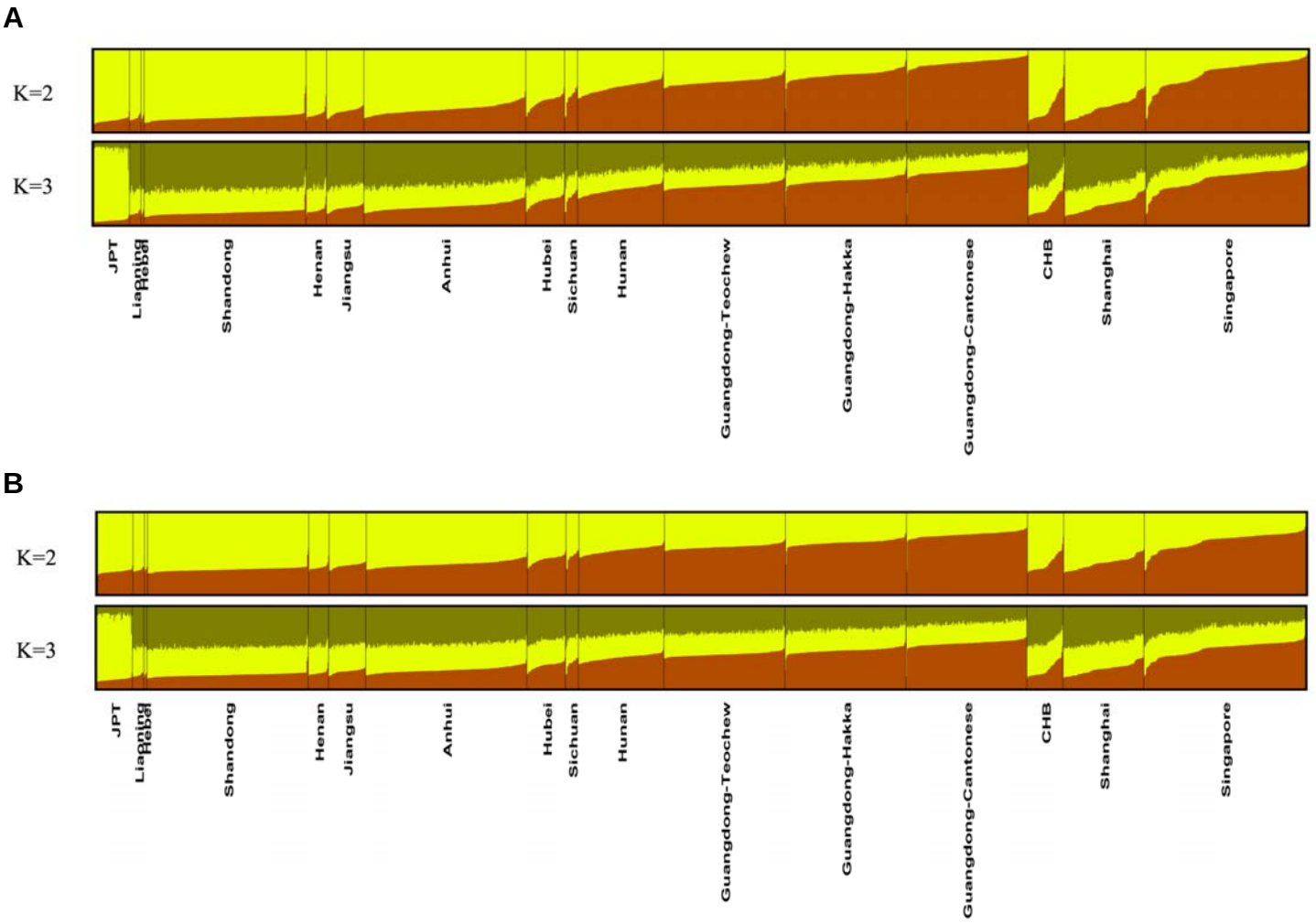


Figure S7. The Graph Shows the Exponential Increase of the λ, with Increasing Proportion of the Chinese from the Souhern Provinces

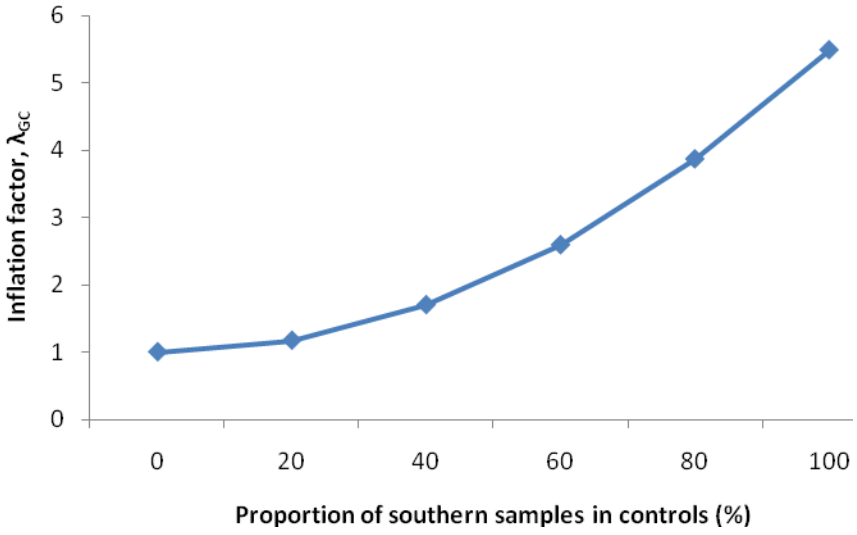


Figure S8. 500N Cases and 500N Controls for the Simulation Study on the Effect of Population Stratification on Case-Control Studies

This serves as a reference, since both sets of individuals are from the same region, so there is supposed to be minimal population stratification. In this case, λ=1.

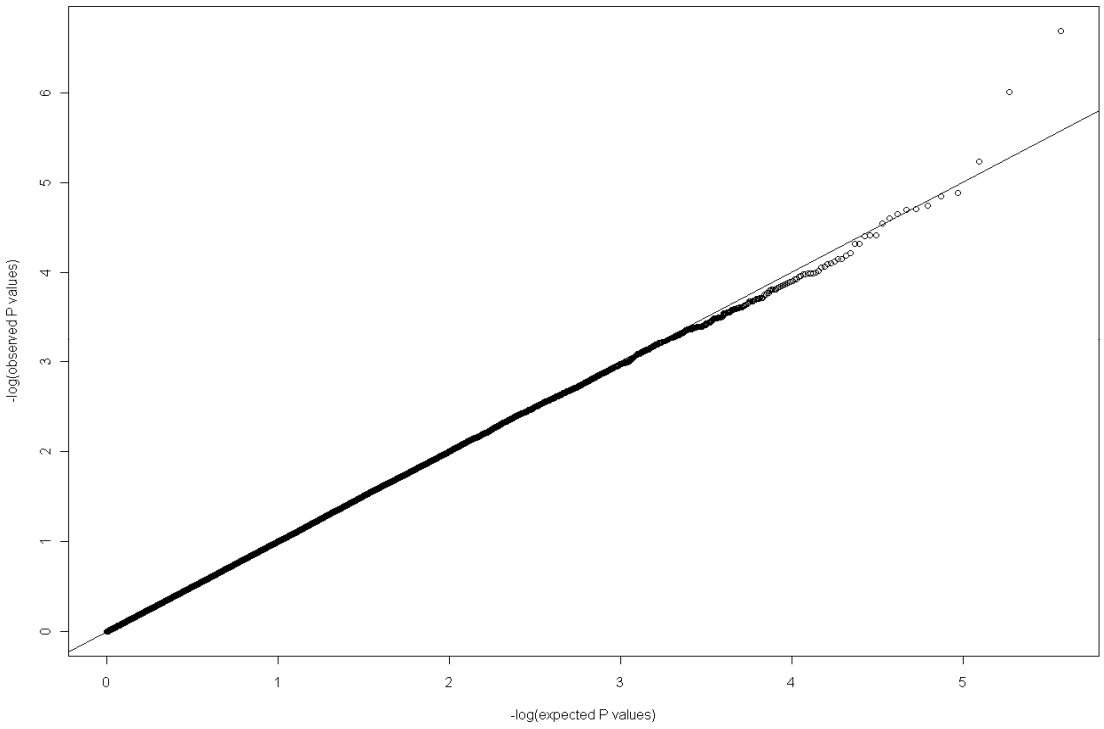


Figure S9. Combination of GC- and PCA-Correction in 100 % Stratification between 500N and 500S

The stratification was rectified only when both GC and PCA corrections were utilized.

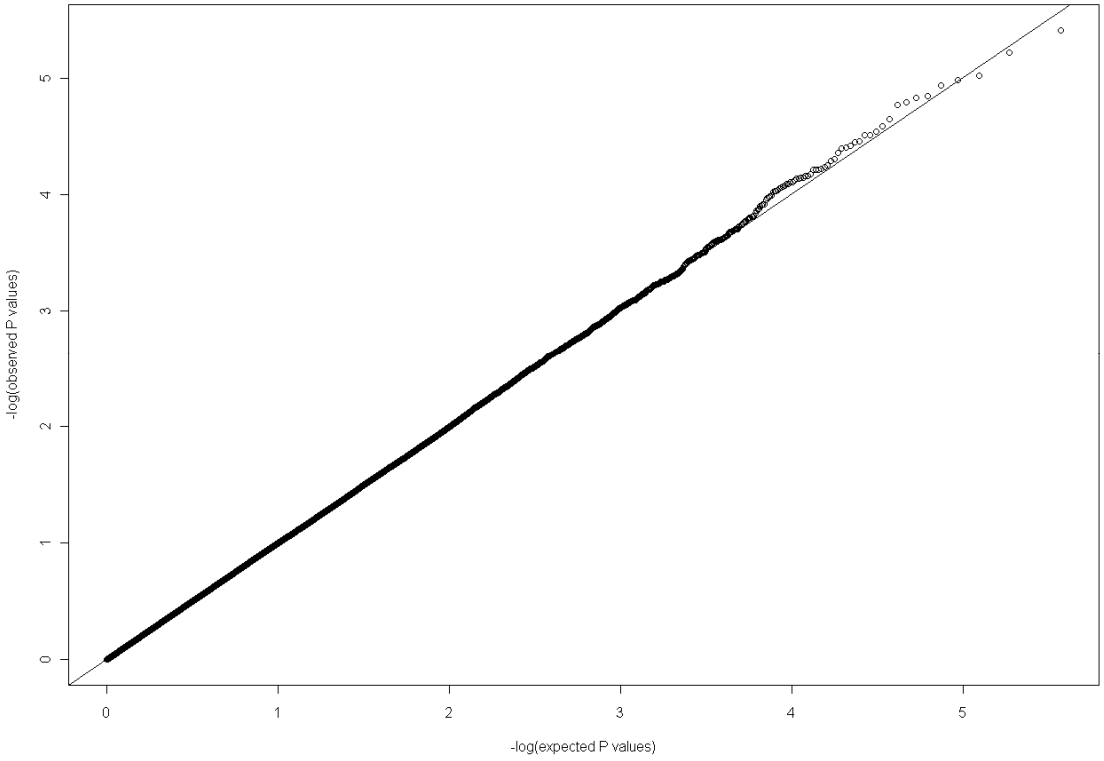


Figure S10. QQ-Plot of the Association Analysis in 1000 Shanghai Samples, Where the Case and Control Samples are Selected based on PC1 Values

The Shanghai samples were first split equally into the 'north' and 'south' clusters based on the PC1 value, and the simulated association analysis was then performed on the 500 case from the 'north' cluster and the 500 control samples from the 'south' cluster of the Shanghai samples.

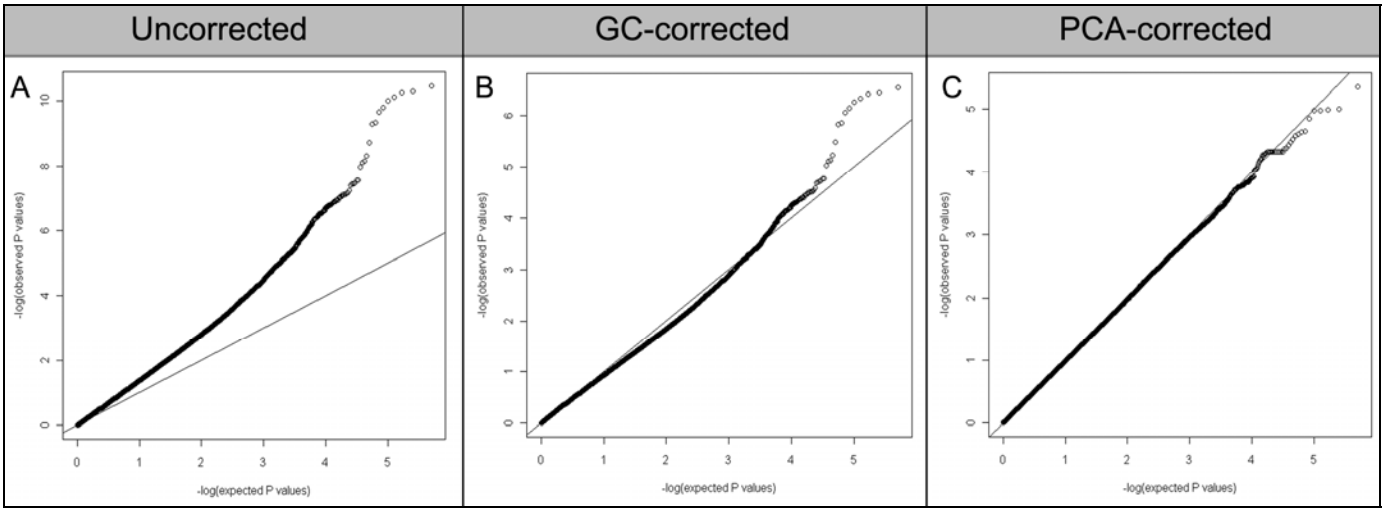


Table S1. Summary of the Genome-wide Datasets Used in this Study

All the samples were combined and the common SNPs from these disease GWAS studies were consolidated into a superset for the purpose of this study. Two datasets were derived from the combined dataset: the PS set for population structure analysis, and the SM set for simulation analysis. The SM set was derived from the GWAS datasets (excluding Psoriasis Singapore) and the Sichuan dataset (marked with *).

	Dataset	**Number of samples**	**Number of SNPs**
GWAS*	Leprosy	706	491,877
GWAS*	Psoriasis	2253	494,901
GWAS*	Nasopharyngeal Carcinoma	1977	469,288
GWAS*	Vitiligo	1041	494,110
GWAS*	Systemic Lupus Erythematosus	587	494,522
GWAS	Psoriasis Singapore	570	482,874
*	Sichuan	16	570,013
	Shanghai	1005	567,933
	CHB (Hapmap)	45	3,731,053
	Set 1 (PS set) – for Population Structure Analysis	**1454**	**420,931**
	Set 2 (SM set) – for Simulation Analysis	**6580 (marked *)**	**400,328**

Table S2. Summary of the Numbers of SNPs Removed during Various Filterings

The initial set of common SNPs pooled from the samples were reduced further after removing SNPs from the regions of long-range LD, significant SNPs from the various GWAS and correlated SNPs. This gives the final number of SNPs used in the respective analyses.

	PS for population structure analysis	**SM for GWAS simulation analysis**
Initial	420,931	400,328
Total number of SNPs significant in various GWAS	Not removed	22,300
Total number of SNPs from regions of long-range LD: Chr5: 44-51.5Mb Chr6: 25-37Mb Chr8: 1-12.7Mb Chr11: 45-57Mb Chr11: 84-86Mb	7,454	4,645
	413,477 (for STRUCTURE/FRAPPE)	**373,383**
Removal of correlated SNPs ($r^2 > 0.2$)	305,912	Not removed
	107,565 (for PCA)	**373,383**

Table S3. Average F_{ST} Values between Three Major Geographic Regions of China and the Three Dialect Groups in Guangdong

	Central	South	Guangdong-Cantonese	Guangdong-Hakka	Guangdong-Teochew
North	0.001016	0.002236	0.003952	0.002911	0.002569
Central		0.001520	0.003044	0.002188	0.001953
South			0.001559	0.001182	0.001381
Guangdong-Cantonese				0.001919	0.002270
Guangdong-Hakka					0.001835

Table S4. The Median Latitude and the Mean PC1 Value of Each Province

The Pearson's correlation between the mean PC1 (of the samples in that province used in population stratification analysis) and the median latitude is 0.93.

	Lowest latitude (°/degree)	Highest latitude (°/degree)	Median latitude (°/degree)	mean PC1
Liaoning	38.72	43.43	41.08	0.03558
Hebei	36.05	42.07	39.06	0.03400
Shandong	34.38	38.40	36.39	0.03544
Henan	31.38	36.40	33.89	0.03067
Jiangsu	30.75	35.33	33.04	0.02691
Anhui	29.68	34.63	32.16	0.02293
Sichuan	26.05	34.32	30.18	0.00513
Hubei	29.42	33.35	31.38	0.01135
Hunan	24.65	30.13	27.39	-0.01036
Guangdong	20.22	25.52	22.87	-0.03075

Table S5. Average Membership Coefficients across the Samples from Each Province, Their Correlation with Latitude and Also the Inter-cluster Correlation between FRAPPE and STRUCTURE Outputs from the Same K Value

(A) Groups (Grp) 1, 2 and 3 were distinguished based on their similarity in membership coefficients. Taking STRUCTURE K=3, it can be observed that group 1 has the largest membership coefficients in the range of 0.50-0.58, group 2 range 0.34-0.45 and group 3 comprising only of Guangdong has 0.2339. Group 2 is very obviously an intermediate region with Sichuan and Hubei more similar to group 1 and Hunan more similar to Guangdong. CHB and Shanghai were excluded in this table because of the highly varied individual membership coefficients. The average membership coefficients of the entire Guangdong province are the averages of the coefficients of the three main dialect groups of Guangdong – Hakka, Cantonese and Teochew. The average absolute Pearson's correlation between each membership coefficient and the median latitude is about 0.94. (B) Table S4B shows the Pearson's correlation between the average membership coefficients of the clusters defined by FRAPPE and STRUCTURE when compared across the same K value. The nomenclature corresponds to the column in Table S4A and is read STRUCTURE (S) and FRAPPE (F), followed by the K value, a dot and then the cluster. So the second cluster of STRUCTURE with K=2, will be denoted "S2.2".

A

Grp	Province	Median Latitude (degree)	STRUCTURE					FRAPPE					Sample size
			K=2		K=3			K=2		K=3			
			S2.1	S2.2	S3.1	S3.2	S3.3	F2.1	F2.2	F3.1	F3.2	F3.3	
1	Anhui	32.16	0.7257	0.2743	0.5084	0.2229	0.2686	0.63	0.37	0.4475	0.2276	0.3249	200
	Jiangsu	33.04	0.7573	0.2427	0.5297	0.1998	0.2705	0.6498	0.3502	0.4621	0.2099	0.328	46
	Henan	33.89	0.7872	0.2128	0.5392	0.1758	0.2851	0.6693	0.3307	0.4688	0.1912	0.34	25
	Shandong	36.39	0.8295	0.1705	0.5679	0.1445	0.2876	0.6972	0.3028	0.4886	0.1668	0.3446	200
	Hebei	39.06	0.8313	0.1687	0.5705	0.1436	0.2859	0.6986	0.3014	0.4903	0.1667	0.3431	4
	Liaoning	41.08	0.8371	0.1629	0.5741	0.1386	0.2873	0.7027	0.2973	0.4922	0.1623	0.3455	14
2	Hubei	31.38	0.6331	0.3669	0.4507	0.2926	0.2567	0.5718	0.4282	0.4083	0.2812	0.3105	48
	Hunan	27.39	0.4528	0.5472	0.3438	0.4293	0.2269	0.4587	0.5413	0.335	0.3826	0.2824	106
	Sichuan	30.18	0.5798	0.4202	0.4154	0.333	0.2516	0.5383	0.4617	0.3833	0.3103	0.3064	16
3	Guangdong	22.87	0.2787	0.7213	0.2339	0.562	0.2042	0.3478	0.6522	0.2574	0.4835	0.2592	450
	-Hakka	-	0.3014	0.6986	0.2481	0.5438	0.2081	0.6365	0.3635	0.2692	0.4692	0.2616	150
	-Cantonese	-	0.1507	0.8493	0.1495	0.6605	0.1899	0.7357	0.2643	0.1945	0.5591	0.2464	150
	-Teochew	-	0.384	0.616	0.3038	0.4814	0.2148	0.5844	0.4157	0.3086	0.4218	0.2695	150
	Correlation with latitude		**0.9353**	**-0.9353**	**0.9348**	**-0.94**	**0.931**	**0.937**	**-0.94**	**0.934**	**0.939**	**-0.94**	

B

Correlation	F2.1	F2.2	F3.1	F3.2	F3.3
S2.1	1	-1			
S2.2		1			
S3.1			1	0.995	-1
S3.2				1	-1
S3.3					1

Table S6. Membership Coefficient of each Individual from the CHB Matched to a Group in Table S4A

There are 28 CHB from group 1, the more northerly individuals, 15 from group 2, the intermediate, and 2 from group 3, the more southerly individuals.

	CHB Sample-id	STRUCTURE					FRAPPE					Grp
		K=2		K=3			K=2		K=3			
1	CHB-NA18524	0.8024	0.1976	0.5311	0.1601	0.3088	0.6792	0.3208	0.4643	0.1861	0.3496	1
2	CHB-NA18526	0.4737	0.5263	0.3475	0.4112	0.2413	0.4711	0.5289	0.3367	0.369	0.2943	2
3	CHB-NA18529	0.8206	0.1794	0.5277	0.1466	0.3258	0.69	0.31	0.4637	0.1716	0.3647	1
4	CHB-NA18532	0.6364	0.3636	0.4249	0.2854	0.2896	0.5743	0.4257	0.3931	0.278	0.3289	2
5	CHB-NA18537	0.7088	0.2912	0.4738	0.2308	0.2954	0.6207	0.3793	0.4232	0.2371	0.3397	1
6	CHB-NA18540	0.8133	0.1867	0.5608	0.1588	0.2804	0.6868	0.3132	0.4861	0.1772	0.3367	1
7	CHB-NA18542	0.8541	0.1459	0.5458	0.1227	0.3315	0.7131	0.2869	0.4743	0.1509	0.3748	1
8	CHB-NA18545	0.4547	0.5453	0.2912	0.4182	0.2906	0.4601	0.5399	0.2909	0.3793	0.3298	2
9	CHB-NA18547	0.8419	0.1581	0.5369	0.1312	0.3319	0.7058	0.2942	0.4693	0.16	0.3707	1
10	CHB-NA18550	0.1877	0.8123	0.1689	0.6275	0.2036	0.2915	0.7085	0.211	0.5346	0.2544	3
11	CHB-NA18552	0.6103	0.3897	0.4302	0.3075	0.2624	0.5579	0.4421	0.393	0.2959	0.3111	2
12	CHB-NA18555	0.5447	0.4553	0.3651	0.3534	0.2814	0.5168	0.4832	0.3496	0.3304	0.3199	2
13	CHB-NA18558	0.8219	0.1781	0.5103	0.1426	0.3471	0.6923	0.3077	0.4477	0.1689	0.3834	1
14	CHB-NA18561	0.835	0.165	0.5879	0.1425	0.2696	0.7006	0.2994	0.5039	0.1622	0.3338	1
15	CHB-NA18562	0.8094	0.1906	0.534	0.1554	0.3106	0.6847	0.3153	0.4613	0.179	0.3597	1
16	CHB-NA18563	0.8895	0.1105	0.5708	0.0939	0.3353	0.7419	0.2581	0.4964	0.1273	0.3764	1
17	CHB-NA18564	0.8582	0.1418	0.552	0.1189	0.3292	0.7148	0.2852	0.4791	0.1498	0.3711	1
18	CHB-NA18566	0.7907	0.2093	0.5434	0.1754	0.2812	0.6714	0.3286	0.4723	0.189	0.3387	1
19	CHB-NA18570	0.5264	0.4736	0.3679	0.374	0.2581	0.5035	0.4965	0.3499	0.3413	0.3088	2
20	CHB-NA18571	0.4752	0.5248	0.3158	0.4071	0.2771	0.4728	0.5272	0.3151	0.3661	0.3188	2
21	CHB-NA18572	0.8487	0.1513	0.561	0.129	0.31	0.7088	0.2912	0.4791	0.1594	0.3615	1
22	CHB-NA18573	0.5168	0.4832	0.3646	0.3773	0.2581	0.4987	0.5013	0.3493	0.3507	0.3001	2
23	CHB-NA18576	0.8187	0.1813	0.539	0.149	0.312	0.6903	0.3097	0.4687	0.1713	0.36	1
24	CHB-NA18577	0.6887	0.3113	0.4612	0.2478	0.291	0.6062	0.3938	0.4204	0.2459	0.3337	2
25	CHB-NA18579	0.8307	0.1693	0.5152	0.1368	0.348	0.6973	0.3027	0.4487	0.1654	0.3859	1
26	CHB-NA18582	0.4821	0.5179	0.3321	0.4045	0.2635	0.4763	0.5237	0.3237	0.3659	0.3104	2

27	CHB-NA18592	0.7856	0.2144	0.53	0.177	0.293	0.6662	0.3338	0.4616	0.1919	0.3465	1
28	CHB-NA18593	0.7527	0.2473	0.5005	0.1993	0.3003	0.6482	0.3518	0.4393	0.2155	0.3452	1
29	CHB-NA18594	0.6333	0.3667	0.4429	0.2906	0.2665	0.572	0.428	0.4012	0.2808	0.318	2
30	CHB-NA18603	0.6218	0.3782	0.439	0.2998	0.2612	0.5648	0.4352	0.401	0.2917	0.3073	2
31	CHB-NA18605	0.8145	0.1855	0.5332	0.1501	0.3167	0.6879	0.3121	0.4614	0.1738	0.3649	1
32	CHB-NA18608	0.8207	0.1793	0.5344	0.1485	0.3171	0.6921	0.3079	0.4662	0.1682	0.3656	1
33	CHB-NA18609	0.8317	0.1683	0.5524	0.1418	0.3058	0.6988	0.3012	0.4786	0.1638	0.3576	1
34	CHB-NA18611	0.8287	0.1713	0.5165	0.1372	0.3463	0.6966	0.3034	0.451	0.1636	0.3854	1
35	CHB-NA18612	0.826	0.174	0.5019	0.1412	0.3569	0.6955	0.3045	0.4456	0.1675	0.3868	1
36	CHB-NA18620	0.8234	0.1766	0.5483	0.1506	0.3011	0.6922	0.3078	0.4758	0.1718	0.3524	1
37	CHB-NA18621	0.8586	0.1414	0.5515	0.1152	0.3333	0.719	0.281	0.4748	0.145	0.3802	1
38	CHB-NA18622	0.6592	0.3408	0.4584	0.2709	0.2707	0.5882	0.4118	0.4131	0.27	0.3169	2
39	CHB-NA18623	0.8085	0.1915	0.4935	0.1526	0.3539	0.6836	0.3164	0.4383	0.1771	0.3846	1
40	CHB-NA18624	0.7757	0.2243	0.5434	0.1868	0.2697	0.6606	0.3394	0.4719	0.2028	0.3253	1
41	CHB-NA18632	0.4527	0.5473	0.3111	0.4268	0.2622	0.4582	0.5418	0.309	0.384	0.307	2
42	CHB-NA18633	0.7147	0.2853	0.4713	0.2278	0.3009	0.6224	0.3776	0.4183	0.234	0.3477	1
43	CHB-NA18635	0.8505	0.1495	0.5264	0.1219	0.3517	0.7112	0.2888	0.4604	0.1516	0.388	1
44	CHB-NA18636	0.329	0.671	0.2473	0.5198	0.2329	0.3817	0.6183	0.27	0.4536	0.2765	3
45	CHB-NA18637	0.5753	0.4247	0.4058	0.3354	0.2588	0.535	0.465	0.3803	0.3124	0.3073	2