A set of novel SNP loci for differentiating continental populations and three Chinese populations

Xiao-Ye Jin[1,2,3], Yuan-Yuan Wei[1,2], Qiong Lan[4], Wei Cui[1,2,3], Chong Chen[1,2,3], Yu-Xin Guo[1,2,3], Ya-Ting Fang[4] and Bo-Feng Zhu[1,2,4]

[1] Key laboratory of Shaanxi Province for Craniofacial Precision Medicine Research, College of Stomatology, Xi'an Jiaotong University, Xi'an, China

[2] Clinical Research Center of Shaanxi Province for Dental and Maxillofacial Diseases, College of Stomatology, Xi'an Jiaotong University, Xi'an, China

[3] College of Medicine and Forensics, Xi'an Jiaotong University Health Science Center, Xi'an, China

[4] Department of Forensic Genetics, School of Forensic Medicine, Southern Medical University, Guangzhou, China

Submitted 2 October 2018
Accepted 22 January 2019
Published 29 March 2019

Corresponding author
Bo-Feng Zhu,
zhubofeng@mail.xjtu.edu.cn

Academic editor
Francesc Calafell

Additional Information and Declarations can be found on page 14

DOI 10.7717/peerj.6508

ABSTRACT

In recent years, forensic geneticists have begun to develop some ancestry informative marker (AIM) panels for ancestry analysis of regional populations. In this study, we chose 48 single nucleotide polymorphisms (SNPs) from SPSmart database to infer ancestry origins of continental populations and Chinese subpopulations. Based on the genetic data of four continental populations (African, American, East Asian and European) from the CEPH-HGDP database, the power of these SNPs for differentiating continental populations was assessed. Population genetic structure revealed that distinct ancestry components among these continental populations could be discerned by these SNPs. Another novel population set from 1000 Genomes Phase 3 was treated as testing populations to further validate the efficiency of the selected SNPs. Twenty-two populations from CEPH-HGDP database were classified into three known populations (African, East Asian, and European) based on their biogeographical regions. Principal component analysis and Bayes analysis of testing populations and three known populations indicated these testing populations could be correctly assigned to their corresponding biogeographical origins. For three Chinese populations (Han, Mongolian, and Uygur), multinomial logistic regression analyses indicated that these 48 SNPs could be used to estimate ancestry origins of these populations. Therefore, these SNPs possessed the promising potency in ancestry analysis among continental populations and some Chinese populations, and they could be used in population genetics and forensic research.

Subjects Bioinformatics, Genetics, Data Science, Population Biology

Keywords Ancestry informative markers, Biogeographical origins, SNP, Chinese populations, Continental populations

INTRODUCTION

Ancestry informative markers, which demonstrated distinct allele frequency differences among different populations (*Frudakis et al., 2003*; *Shriver et al., 1997*), could be used to infer biogeographical origins of unknown biological samples and discern population

How to cite this article Jin X-Y, Wei Y-Y, Lan Q, Cui W, Chen C, Guo Y-X, Fang Y-T, Zhu B-F. 2019. A set of novel SNP loci for differentiating continental populations and three Chinese populations. *PeerJ* 7:e6508 http://doi.org/10.7717/peerj.6508

substructure. They are also conductive to forensic investigations by providing investigative clues about the biogeographical origins of the unknown suspect. To date, a great many AIM panels for different purposes have been developed (*Li et al., 2016*; *Phillips, 2015*; *Phillips et al., 2013*). It is noteworthy that AIM panels differentiating different continental populations might not be appropriate to differentiate populations in the same continent, and these panels could even produce prediction errors for ancestry inferences of these populations. The two-tier AIM panels were recommended for this issue: one for estimating ancestry origins of major continental populations and the other for populations in the same continent (*Kidd et al., 2014*). Besides, *Pakstis et al. (2017)* stated that the identification of highly differentiated SNPs in relatively neglected geographical regions or subpopulations from the same continent should be encouraged because these markers could further improve and fine-tune the extant SNP panels.

China, one of the world's earliest civilizations, consists of 56 officially identified ethnic groups. Previous studies found extensive genetic variations among Han populations and minority groups in China (*Wang et al., 2018*; *Zhang et al., 2015*; *Zhao & Lee, 1989*; *Zhao et al., 1991*). Some researchers selected some markers to differentiate Chinese populations. For examples, *Qin et al. (2014)* provided 150 SNPs for ancestry analysis of northern Han and southern Han; *Sun et al. (2016)* presented twelve multi-InDels to differentiate Han and Tibetan populations. Ancestry analysis among more populations in China should be conducted to provide more valuable information for genome-wide association study and forensic investigation. As is known to us, Han people are the largest ethnic group in the world and distribute in many regions. Uygur individuals mainly live in the Xinjiang Uygur Autonomous Region where extensive population movements occurred in history and they possess admixture genetic components of East Asians and Europeans (*Xu et al., 2008*). Mongolian group is one of fifty-six ethnic groups in China. During the period of the Yuan dynasty, the governor of China and his soldiers began their expedition to Europe which might lead to some Mongolian individuals dispersing in different regions (https://en.wikipedia.org/wiki/Mongols). Although genetic differentiations among Han, Uygur and Mongolian populations existed (*Zhang et al., 2015*), no research on ancestry analyses of these populations was conducted.

In this study, based on population data assembled in CEPH-HGDP (*Li et al., 2008*), a great number of SNPs were selected for distinguishing four major regions (Africa, America, Europe and East Asia). Among these SNPs, the SNPs with high differentiations among Han, Uygur and Mongolian populations were further selected to infer ancestry origins of these populations.

MATERIAL AND METHODS

Reference populations

Twenty-five populations within four major geographical regions were chosen as the training set for preliminary assessments of selected SNPs. Six testing populations in three continents included Esan, Yoruba, Finnish, British, Beijing Han and Japanese populations, of which genetic data were downloaded from 1000 Genomes Phase 3 (*Genomes Project*

Table 1 Detailed information of populations used in this study and their corresponding sample sizes.

Datasets	Populations	Abbreviations	Continents	Sources	Sample sizes
Training set	Biaka Pygmy	–	Africa	CEPH-HGDP	22
	Mbuti Pygmy	–	Africa	CEPH-HGDP	13
	Bantu[a]	–	Africa	CEPH-HGDP	19
	Yoruba	–	Africa	CEPH-HGDP	21
	Mandenka	–	Africa	CEPH-HGDP	22
	Brazian[b]	–	America	CEPH-HGDP	22
	Maya	–	America	CEPH-HGDP	21
	Pima	–	America	CEPH-HGDP	14
	Basque	–	Europe	CEPH-HGDP	24
	French	–	Europe	CEPH-HGDP	28
	Italian[c]	–	Europe	CEPH-HGDP	49
	Orcadian	–	Europe	CEPH-HGDP	15
	Adygei	–	Europe	CEPH-HGDP	17
	Russian	–	Europe	CEPH-HGDP	25
	Cambodian	–	East Asia	CEPH-HGDP	10
	Dai	–	East Asia	CEPH-HGDP	10
	Han	–	East Asia	CEPH-HGDP	44
	Miao	–	East Asia	CEPH-HGDP	10
	Mongolian	–	East Asia	CEPH-HGDP	10
	She	–	East Asia	CEPH-HGDP	10
	Tu	–	East Asia	CEPH-HGDP	10
	Tujia	–	East Asia	CEPH-HGDP	10
	Yi	–	East Asia	CEPH-HGDP	10
	Japanese	–	East Asia	CEPH-HGDP	28
	Yakut	–	East Asia	CEPH-HGDP	25
Testing set	Esan in Nigeria	ESN	Africa	1000 Genomes Phase 3	99
	Yoruba in Ibadan, Nigeria	YRI	Africa	1000 Genomes Phase 3	108
	Finnish in Finland	FIN	Europe	1000 Genomes Phase 3	99
	British in England and Scotland	GBR	Europe	1000 Genomes Phase 3	91
	Han Chinese in Bejing, China	CHB	East Asia	1000 Genomes Phase 3	103
	Japanese in Tokyo, Japan	JPT	East Asia	1000 Genomes Phase 3	104
Three subpopulations in China	Uygur	–	Central Asia	CEPH-HGDP	10
	Han	–	East Asia	CEPH-HGDP	44
	Mongolian	–	East Asia	CEPH-HGDP	10

Notes.

[a] Bantu population includes Kenya Bantu and South African Bantu populations.

[b] Brazian population includes Karitiana and Surui populations.

[c] Italian population includes Sardinian, Tuscan and Bergamo populations.

Consortium et al., 2015). Genetic data of three subpopulations in China (Han, Mongolian, and Uygur) was obtained from CEPH-HGDP (*Li et al., 2008*). Detailed descriptions of these populations and their corresponding sample sizes were given in Table 1. Besides, genetic data of 48 SNPs in training, testing and three Chinese populations were presented in Table S1.

Criteria for SNP selection

SPSmart includes the genetic data of 1000 Genomes Phase I, HapMap release #28, Perlegen complete data set and the Stanford University and Michigan University CEPH-HGDP panels, which is developed to help researchers use and combine different datasets and do some statistical analyses of interest (*Amigo et al., 2008*). SNPs were chosen from SPSmart online tool when they met the following criteria: (1) SNP loci should locate in intron regions. (2) SNP loci were bi-allelic genetic markers; (3) SNP loci were located on different chromosomes or at least 10 Mb distances on the same chromosome; (4) Ancestral allele frequency differences between continental populations were at least 0.3; (5) SNP selected must conform to Hardy-Weinberg equilibrium (HWE) in all reference populations. Next, SNPs whose frequency differences among Han, Uygur and Mongolian populations were more than 0.3 were used for further analysis. Besides, we also retained some SNP loci that showed high genetic differentiations among continental populations/three Chinese populations. Finally, forty-eight SNPs were selected to differentiate continental populations and three Chinese populations. General information of 48 SNPs was given in Table 2.

Statistical analysis

HWE tests of SNP loci in 25 training populations were estimated by Genepop software v4.0 (*Rousset, 2008*). Allele frequencies of 48 SNP loci in 25 training populations were calculated by PowerStats software v1.2 (Promega, Madison, WI, USA). The informativeness for assignment (*In*) values of 48 SNP loci in four continental populations (African, American, European, and East Asian) were calculated by Infocalc program v1.1 (*Rosenberg et al., 2003*) based on the genetic data of 25 training populations. Ancestral allele frequency heatmap and the boxplot of *In* values of 48 SNPs were plotted by R software v3.3 (*R Core Team, 2016*). Principal component analysis (PCA) of four continental populations including 25 training populations was conducted by PLINK software v1.9 (http://www.cog-genomics.org/plink/1.9/), and then scatter plot of these populations was plotted by R software v3.3. Population genetic structure of 25 training populations at $K = 2$–5 and cross-validation error of each K value were performed by ADMIXTURE software v1.3 (*Alexander, Novembre & Lange, 2009*). Graphical results of estimated ancestry proportions were conducted with the CLUMPAK online tool (*Kopelman et al., 2015*).

To further evaluate discrimination efficiencies of 48 SNPs for continental populations, ancestry components of six testing populations were estimated by ADMIXTURE software v1.3 and their results were shown in the form of beeswarm by R software v3.3. Next, twenty-two training populations (excluding from American populations) were treated as reference populations and six testing populations were blind samples. PCA and Naïve Bayes analysis of these populations including reference populations and six testing populations were conducted with PLINK software v1.9 and the Snipper App suite v2.5 (http://mathgene.usc.es/snipper), respectively.

Ancestry analyses among Uygur, Han and Mongolian populations were conducted based on the 48 SNPs. First of all, allele frequencies of 48 SNPs in these three populations were calculated by PowerStats software v1.2. *In* values of 48 SNP loci in Uygur, Han and Mongolian populations were also calculated with Infocalc program v1.1. Secondly,

Table 2 General information and ancestral allele frequencies of 48 SNP loci in different continental populations.

Rs numbers	Alleles[a]	Chromosomes[a]	Positions (bp)[a]	African[b]	American[b]	European[b]	East Asian[b]
rs10918196	C/T	1	165478920	0.8351	0.3947	0.6108	0.2429
rs2801178	G/T	1	14881716	0.8608	0.2368	0.7753	0.4774
rs4652825	A/G	1	183824189	0.7320	0.4737	0.8101	0.6215
rs10779958	A/C	2	74528651	0.3763	0.5000	0.8481	0.1921
rs1161474	C/T	2	239209361	0.7320	0.6053	0.3829	0.5424
rs7570426	A/C	2	3325638	0.5567	0.8947	0.9051	0.4915
rs11716005	C/T	3	18394010	0.9742	0.9474	0.8101	0.5621
rs301927	A/G	3	97627774	0.7423	0.8947	0.0918	0.6638
rs4533619	A/G	3	42248320	0.2835	0.3246	0.1456	0.7345
rs4894436	G/T	3	171401588	0.6649	0.0789	0.6551	0.0847
rs6446081	C/T	3	59723035	0.8866	0.8070	0.7532	0.5480
rs12650562	C/T	4	23799564	0.4897	0.7456	0.5032	0.5593
rs3762894	C/T	4	99144933	0.7268	1.0000	0.8291	0.3757
rs2400219	C/T	5	146472334	0.1546	0.2193	0.2057	0.7994
rs277329	A/C	5	54158354	0.8351	0.8158	0.8133	0.3814
rs35414	C/T	5	33969523	1.0000	0.8509	0.3956	0.8644
rs4704322	C/T	5	76526649	0.9639	0.5175	0.7880	0.1299
rs871722	A/G	5	17437385	0.8918	0.2368	0.1361	0.0876
rs1857859	A/G	6	100446711	0.8557	0.4123	0.7025	0.5678
rs4711760	C/T	6	44027931	0.7990	0.7018	0.3291	0.9379
rs947612	A/G	6	73028938	0.8608	0.7018	0.2500	0.7655
rs2373177	C/T	7	147717307	0.6031	0.6053	0.7152	0.7386
rs4646437	G/A	7	99767460	0.8505	0.2456	0.1108	0.1243
rs7795646	A/G	7	120683673	0.5773	0.7632	0.2057	0.4774
rs2595599	A/G	8	92646120	0.4124	0.3947	0.7962	0.2542
rs351554	A/C	8	16149886	0.7680	0.5439	0.1329	0.4972
rs4738110	A/C	8	71193716	0.9588	0.3509	0.7342	0.3136
rs10965206	A/G	9	229826	0.4845	0.5877	0.8766	0.5791
rs4743923	C/T	9	93488779	0.0361	0.3333	0.4019	0.4096
rs16917217	A/G	10	18370276	0.2938	0.2456	0.3513	0.6441
rs4933165	C/T	10	89903145	0.8557	0.5439	0.8956	0.2147
rs11218323	C/T	11	99080380	0.6856	0.8070	0.8734	0.5256
rs1470253	A/G	11	20099911	0.7732	0.2018	0.7437	0.1384
rs1032332	T/C	12	72568351	0.5825	0.2632	0.1361	0.5311
rs17650122	A/G	13	29010756	0.5052	0.5877	0.5696	0.3220
rs3782972	C/T	13	94450792	0.9330	0.2544	0.7468	0.3220
rs7325443	C/T	13	111238209	0.7938	0.3246	0.8070	0.3079
rs10148212	A/C	14	66628895	0.2917	0.5357	0.9209	0.5028
rs10852189	C/T	15	93163645	0.5158	0.2281	0.3228	0.5141
rs9806693	A/G	15	78892045	0.6495	0.5439	0.2437	0.6780
rs170359	A/G	16	57361752	0.6649	0.3333	0.0316	0.1412

(*continued on next page*)

Table 2 (*continued*)

Rs numbers	Alleles[a]	Chromosomes[a]	Positions (bp)[a]	African[b]	American[b]	European[b]	East Asian[b]
rs7219900	C/T	17	15471179	0.7474	0.2105	0.1076	0.3475
rs11152349	A/G	18	62566413	0.7062	0.6316	0.7057	0.3023
rs528438	T/C	18	28093911	0.2165	0.7632	0.3038	0.8079
rs1205357	C/T	20	34289960	0.8711	0.0088	0.1076	0.2147
rs162315	A/G	20	57233824	0.2268	0.6667	0.7025	0.6864
rs2178832	G/A	21	35449600	0.2010	0.4737	0.5443	0.2288
rs361557	A/G	22	18128456	0.9330	0.7456	0.1835	0.5650

Notes.

[a] Information of each SNP locus is shown according to the report of dbSNP build 152.

[b] Ancestral allele frequencies of 48 SNPs in four continental populations are obtained based on the genetic data of 25 training populations in Table 1.

PCA of these populations was also conducted by PLINK software v1.9. Ancestral allele frequency heatmap and the boxplot of *In* values for 48 SNPs, and scatter plot of these three populations were generated by R software v3.3 (*R Core Team, 2016*). Genetic structure analyses of these populations at $K = 3$ were conducted by ADMIXTURE software v1.3 and graphical results were plotted by CLUMPAK online tool. In the end, multinomial logistic regression analyses of these three populations were conducted by Snipper App suite v2.5.

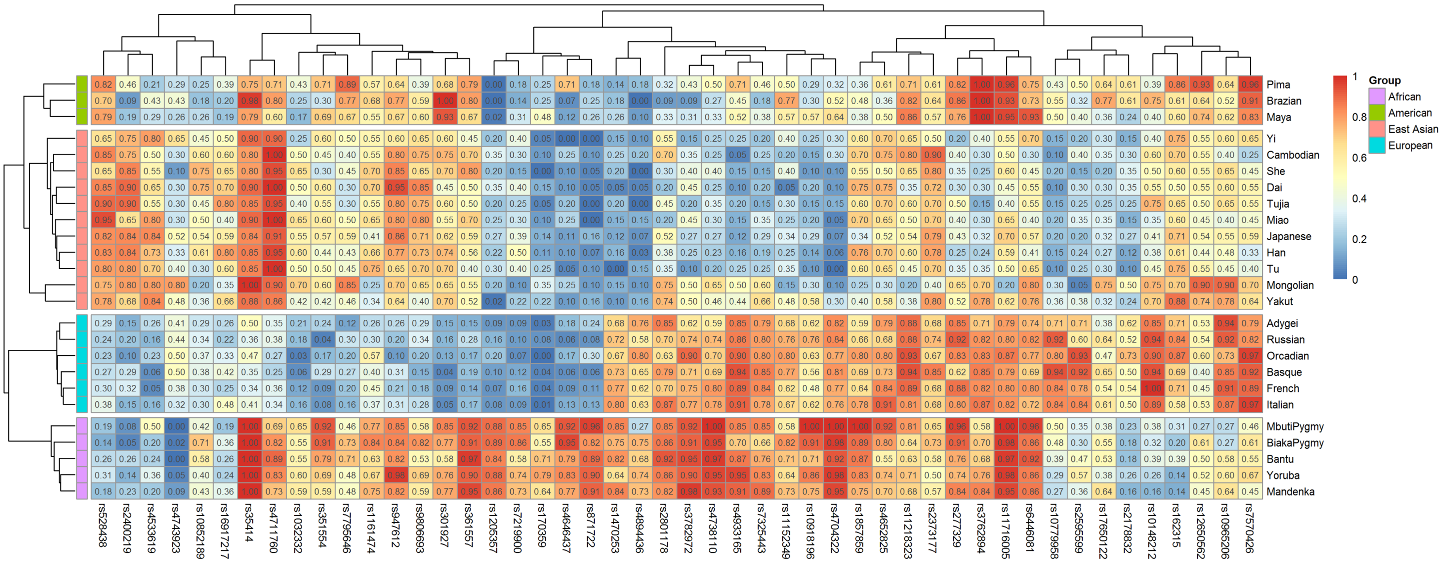
RESULTS

Frequency distributions and population specific *In* values of 48 SNPs

After applying Bonferroni correction (the significant level = 0.05/48 = 0.00104), the selected 48 SNPs all conformed to HWE in 25 training populations (Table S2).

Ancestral allele frequencies of 48 SNP loci in 25 reference populations were shown in Fig. 1. Color contrasts of SNPs in pairwise populations reflected genetic differentiations of pairwise populations: more apparent color contrasts of pairwise populations were, larger genetic differentiations pairwise populations possessed, and vice versa. Besides, the SNPs with distinct color contrasts in pairwise continental populations contributed to differentiating these populations. For example, ancestral allele frequencies of the rs10918196 locus in African, American and East Asian populations were 0.8351, 0.3947 and 0.2429 (Table 2), implying the locus was beneficial to distinguish African populations from the other two continental populations. The phylogenetic tree above the heatmap (Fig. 1) revealed the relationships of SNPs: SNPs showed similar frequency distributions in continental populations tended to locate in the same sub-branches, and vice versa. The phylogenetic tree in the left part of the graph reflected genetic divergences of different continental populations: populations with the same biogeographical origins located in the same sub-branches.

In values of 48 SNPs in African, American, European and East Asian populations were shown in Fig. 2 and Table S3. Similar with allele frequency differences, *In* values were also used to evaluate the degree of population differentiations: genetic markers with high *In* values in the certain population contributed to differentiating the population from the other populations (*Phillips, 2015*). For these 48 SNPs, there were 12, 11, 10, and three SNPs with relatively high *In* values (≥0.1) in African, East Asian, European, and American



	rs528438	rs2400219	rs4533619	rs4743923	rs10852189	rs16917217	rs35414	rs4711760	rs1032332	rs351554	rs7795646	rs1161474	rs947612	rs9806693	rs301927	rs361557
Pima	0.82	0.46	0.21	0.29	0.25	0.39	0.75	0.71	0.43	0.71	0.89	0.57	0.64	0.39	0.68	0.79
Brazian	0.70	0.09	0.43	0.43	0.18	0.20	0.98	0.80	0.25	0.30	0.77	0.68	0.77	0.59	1.00	0.80
Maya	0.79	0.19	0.29	0.26	0.26	0.19	0.79	0.60	0.17	0.69	0.67	0.55	0.67	0.60	0.93	0.67
Yi	0.65	0.75	0.80	0.65	0.45	0.50	0.90	0.90	0.25	0.60	0.50	0.55	0.60	0.65	0.60	0.45
Cambodian	0.85	0.75	0.50	0.30	0.60	0.60	0.80	1.00	0.50	0.45	0.40	0.55	0.80	0.75	0.75	0.70
She	0.65	0.85	0.55	0.10	0.75	0.65	0.80	0.95	0.65	0.30	0.45	0.70	0.85	0.65	0.70	0.80
Dai	0.85	0.90	0.65	0.30	0.75	0.70	0.90	1.00	0.50	0.60	0.30	0.70	0.95	0.85	0.45	0.50
Tujia	0.90	0.90	0.55	0.30	0.45	0.80	0.85	0.95	0.40	0.55	0.30	0.55	0.80	0.75	0.60	0.50
Miao	0.95	0.65	0.80	0.30	0.50	0.40	0.90	1.00	0.65	0.60	0.50	0.65	0.80	0.80	0.55	0.70
Japanese	0.82	0.84	0.84	0.52	0.54	0.59	0.84	0.91	0.55	0.57	0.59	0.41	0.86	0.71	0.62	0.59
Han	0.83	0.84	0.73	0.33	0.61	0.80	0.85	0.95	0.60	0.44	0.43	0.66	0.77	0.73	0.74	0.56
Tu	0.80	0.80	0.70	0.40	0.30	0.60	0.85	1.00	0.50	0.50	0.45	0.75	0.65	0.70	0.70	0.40
Mongolian	0.75	0.80	0.80	0.80	0.20	0.35	1.00	0.90	0.70	0.60	0.85	0.25	0.70	0.65	0.65	0.55
Yakut	0.78	0.68	0.84	0.48	0.36	0.66	0.88	0.86	0.42	0.42	0.46	0.34	0.64	0.40	0.70	0.52
Adygei	0.29	0.15	0.26	0.41	0.29	0.26	0.50	0.35	0.21	0.24	0.12	0.26	0.26	0.29	0.15	0.15
Russian	0.24	0.20	0.16	0.44	0.34	0.22	0.36	0.38	0.18	0.04	0.30	0.30	0.20	0.34	0.16	0.28
Orcadian	0.23	0.10	0.23	0.50	0.37	0.33	0.47	0.27	0.03	0.17	0.20	0.57	0.10	0.20	0.13	0.17
Basque	0.27	0.29	0.06	0.50	0.38	0.42	0.35	0.25	0.06	0.29	0.27	0.40	0.31	0.15	0.04	0.19
French	0.30	0.32	0.05	0.38	0.30	0.25	0.34	0.36	0.12	0.09	0.20	0.45	0.21	0.18	0.09	0.14
Italian	0.38	0.15	0.16	0.32	0.30	0.48	0.41	0.34	0.16	0.08	0.16	0.37	0.31	0.28	0.05	0.17
MbutiPygmy	0.19	0.08	0.50	0.00	0.42	0.19	1.00	0.69	0.65	0.92	0.46	0.77	0.85	0.58	0.85	0.92
BiakaPygmy	0.14	0.05	0.20	0.02	0.71	0.36	1.00	0.82	0.55	0.91	0.73	0.84	0.84	0.82	0.77	0.91
Bantu	0.26	0.26	0.24	0.00	0.58	0.26	1.00	0.89	0.55	0.79	0.71	0.63	0.82	0.53	0.58	0.97
Yoruba	0.31	0.14	0.36	0.05	0.40	0.24	1.00	0.83	0.60	0.69	0.48	0.67	0.98	0.69	0.76	0.90
Mandenka	0.18	0.23	0.20	0.09	0.43	0.36	1.00	0.73	0.59	0.59	0.48	0.75	0.82	0.59	0.77	0.95

	rs1205357	rs7219900	rs170359	rs4646437	rs871722	rs1470253	rs4894436	rs2801178	rs3782972	rs4738110	rs4933165	rs7325443	rs11152349	rs10918196	rs4704322	rs1857859
Pima	0.00	0.18	0.25	0.71	0.18	0.14	0.18	0.32	0.43	0.50	0.71	0.46	0.50	0.29	0.32	0.36
Brazian	0.00	0.14	0.25	0.07	0.25	0.18	0.00	0.09	0.09	0.27	0.45	0.18	0.77	0.30	0.52	0.48
Maya	0.02	0.31	0.48	0.12	0.26	0.26	0.10	0.33	0.31	0.33	0.52	0.38	0.57	0.57	0.64	0.38
Yi	0.20	0.40	0.05	0.00	0.00	0.15	0.10	0.55	0.15	0.25	0.25	0.20	0.40	0.25	0.30	0.60
Cambodian	0.35	0.30	0.10	0.25	0.10	0.25	0.20	0.70	0.35	0.25	0.05	0.25	0.20	0.15	0.30	0.70
She	0.20	0.15	0.00	0.10	0.05	0.20	0.00	0.30	0.40	0.40	0.15	0.15	0.40	0.10	0.10	0.55
Dai	0.35	0.40	0.15	0.10	0.05	0.05	0.05	0.20	0.45	0.25	0.10	0.20	0.05	0.20	0.10	0.75
Tujia	0.20	0.25	0.05	0.20	0.00	0.20	0.00	0.45	0.20	0.35	0.20	0.15	0.30	0.15	0.15	0.60
Miao	0.25	0.30	0.10	0.25	0.00	0.15	0.15	0.20	0.45	0.30	0.15	0.35	0.25	0.20	0.05	0.70
Japanese	0.27	0.39	0.14	0.11	0.16	0.12	0.07	0.52	0.27	0.27	0.12	0.29	0.34	0.29	0.07	0.34
Han	0.22	0.50	0.11	0.10	0.07	0.16	0.03	0.38	0.25	0.23	0.16	0.19	0.25	0.14	0.06	0.76
Tu	0.30	0.40	0.25	0.05	0.10	0.00	0.15	0.35	0.10	0.20	0.25	0.25	0.35	0.10	0.00	0.60
Mongolian	0.20	0.10	0.35	0.25	0.10	0.15	0.10	0.75	0.50	0.65	0.50	0.60	0.15	0.30	0.10	0.25
Yakut	0.02	0.22	0.22	0.10	0.16	0.10	0.16	0.74	0.50	0.46	0.44	0.66	0.48	0.58	0.30	0.40
Adygei	0.09	0.09	0.03	0.18	0.24	0.68	0.76	0.85	0.62	0.59	0.85	0.79	0.68	0.62	0.82	0.59
Russian	0.16	0.10	0.08	0.06	0.08	0.72	0.58	0.70	0.74	0.74	0.86	0.80	0.76	0.76	0.84	0.66
Orcadian	0.20	0.07	0.00	0.17	0.30	0.67	0.80	0.63	0.90	0.70	0.90	0.80	0.80	0.63	0.77	0.80
Basque	0.10	0.12	0.04	0.06	0.06	0.73	0.65	0.79	0.71	0.69	0.94	0.85	0.77	0.56	0.81	0.69
French	0.07	0.16	0.04	0.09	0.11	0.77	0.62	0.70	0.75	0.80	0.89	0.84	0.62	0.48	0.77	0.64
Italian	0.08	0.09	0.01	0.13	0.13	0.80	0.63	0.87	0.77	0.78	0.91	0.78	0.67	0.62	0.76	0.78
MbutiPygmy	0.88	0.85	0.65	0.92	0.96	0.85	0.27	0.85	0.92	1.00	0.85	0.85	0.58	1.00	1.00	1.00
BiakaPygmy	0.89	0.86	0.55	0.95	0.82	0.75	0.75	0.86	0.91	0.95	0.70	0.66	0.82	0.91	0.98	0.89
Bantu	0.84	0.58	0.71	0.79	0.89	0.82	0.68	0.92	0.95	0.97	0.87	0.76	0.71	0.71	0.92	0.87
Yoruba	0.88	0.74	0.79	0.83	0.90	0.64	0.74	0.86	0.90	0.95	0.95	0.83	0.64	0.86	0.98	0.83
Mandenka	0.86	0.73	0.64	0.77	0.91	0.84	0.73	0.82	0.98	0.93	0.91	0.89	0.73	0.75	0.95	0.75

	rs4652825	rs11218323	rs2373177	rs277329	rs3762894	rs11716005	rs6446081	rs10779958	rs2595599	rs17650122	rs2178832	rs10148212	rs162315	rs12650562	rs10965206	rs7570426
Pima	0.61	0.71	0.61	0.82	1.00	0.96	0.75	0.43	0.50	0.64	0.61	0.39	0.86	0.93	0.64	0.96
Brazian	0.36	0.82	0.64	0.86	1.00	0.93	0.73	0.55	0.32	0.77	0.61	0.75	0.61	0.64	0.52	0.91
Maya	0.50	0.86	0.57	0.76	1.00	0.95	0.93	0.50	0.40	0.36	0.24	0.40	0.60	0.74	0.62	0.83
Yi	0.70	0.65	0.50	0.20	0.40	0.55	0.70	0.15	0.20	0.30	0.15	0.40	0.75	0.55	0.60	0.65
Cambodian	0.75	0.80	0.90	0.40	0.30	0.50	0.30	0.10	0.40	0.35	0.30	0.60	0.70	0.55	0.40	0.25
She	0.50	0.65	0.80	0.35	0.25	0.60	0.45	0.20	0.15	0.20	0.35	0.50	0.60	0.55	0.65	0.45
Dai	0.75	0.35	0.72	0.30	0.30	0.45	0.55	0.10	0.30	0.30	0.35	0.55	0.50	0.55	0.60	0.55
Tujia	0.70	0.60	0.70	0.50	0.15	0.40	0.55	0.15	0.25	0.25	0.25	0.75	0.65	0.50	0.65	0.55
Miao	0.65	0.50	0.70	0.40	0.30	0.65	0.40	0.20	0.35	0.35	0.15	0.35	0.60	0.45	0.40	0.45
Japanese	0.52	0.54	0.79	0.43	0.32	0.70	0.57	0.20	0.20	0.32	0.27	0.41	0.71	0.54	0.55	0.59
Han	0.70	0.60	0.78	0.25	0.24	0.59	0.41	0.15	0.24	0.27	0.12	0.38	0.61	0.44	0.48	0.33
Tu	0.65	0.45	0.70	0.35	0.35	0.50	0.65	0.10	0.25	0.30	0.10	0.45	0.75	0.55	0.45	0.40
Mongolian	0.30	0.20	0.40	0.65	0.70	0.20	0.80	0.30	0.05	0.75	0.50	0.75	0.70	0.90	0.90	0.70
Yakut	0.58	0.38	0.80	0.52	0.78	0.62	0.76	0.36	0.38	0.32	0.24	0.70	0.88	0.74	0.78	0.64
Adygei	0.79	0.88	0.68	0.85	0.71	0.79	0.74	0.71	0.71	0.38	0.62	0.85	0.71	0.53	0.94	0.79
Russian	0.68	0.88	0.74	0.92	0.82	0.80	0.82	0.92	0.60	0.64	0.52	0.94	0.84	0.54	0.92	0.82
Orcadian	0.80	0.93	0.67	0.83	0.83	0.87	0.77	0.80	0.93	0.47	0.73	0.90	0.87	0.60	0.73	0.97
Basque	0.73	0.94	0.85	0.62	0.85	0.79	0.69	0.94	0.92	0.65	0.50	0.94	0.69	0.40	0.85	0.92
French	0.84	0.89	0.68	0.88	0.82	0.80	0.80	0.84	0.78	0.54	0.54	1.00	0.71	0.45	0.91	0.89
Italian	0.91	0.81	0.68	0.80	0.87	0.82	0.72	0.84	0.84	0.61	0.50	0.89	0.58	0.53	0.87	0.97
MbutiPygmy	0.92	0.81	0.65	0.96	0.58	1.00	0.96	0.50	0.35	0.38	0.23	0.38	0.31	0.27	0.27	0.46
BiakaPygmy	0.80	0.64	0.73	0.91	0.70	0.98	0.86	0.48	0.30	0.55	0.18	0.32	0.20	0.61	0.27	0.61
Bantu	0.55	0.63	0.58	0.76	0.68	0.97	0.92	0.39	0.47	0.53	0.18	0.39	0.39	0.50	0.58	0.55
Yoruba	0.74	0.71	0.50	0.74	0.76	0.98	0.86	0.29	0.57	0.38	0.26	0.26	0.14	0.52	0.60	0.67
Mandenka	0.70	0.68	0.57	0.84	0.84	0.95	0.86	0.27	0.36	0.64	0.16	0.16	0.14	0.45	0.64	0.45

Figure 1 **Ancestral allele frequency heatmap of 48 SNPs in 25 training populations from different continents.** Different colors represent for different levels of frequency values: blue for low value, red for high value.

Full-size DOI: 10.7717/peerj.6508/fig-1

populations, respectively. Besides, some SNPs which had similar frequency distributions among different continental populations showed low *In* values in these four continental populations (Fig. 2). Besides, we also calculated the cumulative *In* values of 48 SNPs in African, European, East Asian, and American populations, which were 3.4312, 3.0343, 2.7727 and 1.3028, respectively (Table S3). The low cumulative *In* values in American populations indicated these SNPs might possess lower power to differentiate American populations from the other populations.

PCA and ancestral component analysis of 25 training populations based on 48 SNPs

PCA of four continental populations comprising 25 training populations was shown in Fig. 3. Results revealed Africans and Europeans formed two distinct population clusters at quite some distances from the other populations. However, some American individuals clustered closely with East Asian individuals, which might be related to the shared ancestries before the divergences of East Asian populations and American populations (*Li et al., 2008*).

Next, we assessed genetic components of 25 training populations (Fig. 4A). At $K = 2$, the populations in Africa and Europe showed similar genetic components which could be distinguished from the populations in East Asia and America. At $K = 3$, African populations and European populations exhibited their distinct ancestry components, respectively. When K became 4, specific ancestry components in American populations could be observed and all training populations could be classified into four apparent clusters. No further distinctions among these populations could be discerned at $K = 5$. Given these results, apparent distinctions among these continental populations could be achieved by the 48 SNPs. Cross-validation error of each K value was estimated by ADMIXTURE software v1.3

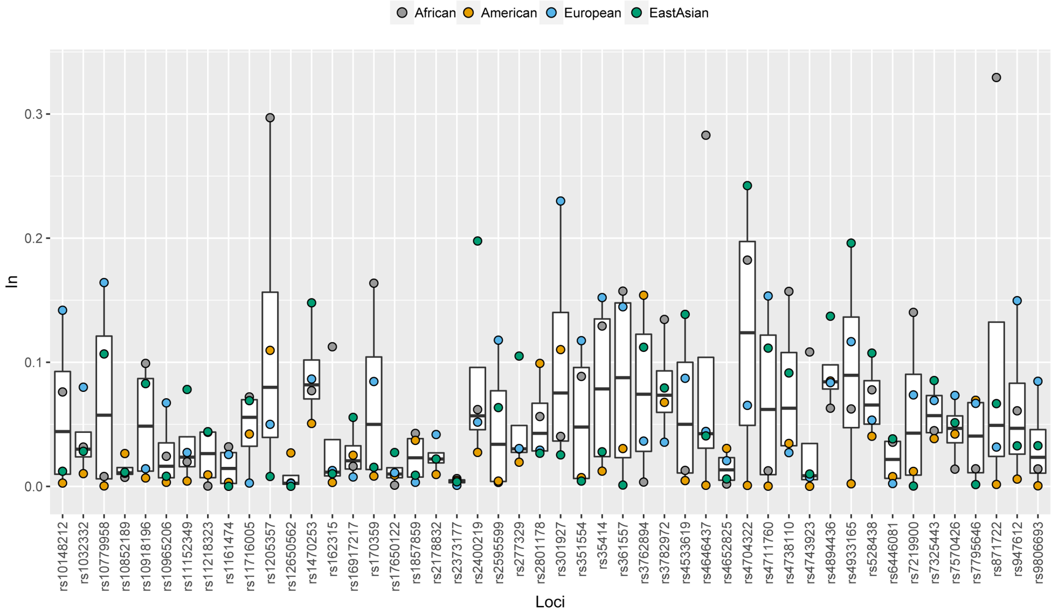


Figure 2 **Population specific *In* values of 48 SNPs in African, American, European and East Asian populations.**

Full-size DOI: 10.7717/peerj.6508/fig-2

to determine the optimum K value, as presented in Fig. 4B. Results revealed that the lowest cross-validation error was observed at $K = 4$, indicating $K = 4$ was the most appropriate for these data.

Ancestry analysis of six testing populations based on 48 SNPs

Since American populations collected in 1000 Genomes Phase 3 showed different degrees of admixture components of European, American and African ancestries (*Genomes Project Consortium et al., 2015*), some populations from East Asia, Europe and Africa were employed to evaluate the efficiency of 48 SNPs for differentiating continental populations. The testing set included two African populations (ESN and YRI), two European populations (GBR and FIN) and two East Asian populations (CHB and JPT). Firstly, we estimated ancestry components of these six populations by ADMIXTURE software v1.3, as shown in Fig S1A. Results indicated that populations within the same continents showed similar genetic component distributions and could be separated from the populations in the other continents (Fig S1A). Furthermore, cross-validation plot revealed $K = 3$ was the best value for these testing populations (Fig S1B). Therefore, estimated genetic components of each individual for six testing populations at $K = 3$ were presented in Fig. 5A. We found FIN and GBR individuals showed high genetic components of European ancestry; CHB and JPT individuals possessed high genetic components of East Asian ancestry; ESN and YRI individuals demonstrated high genetic components of African ancestry. Consequently, these six testing populations could be assigned into their corresponding continental origins by these 48 SNPs.

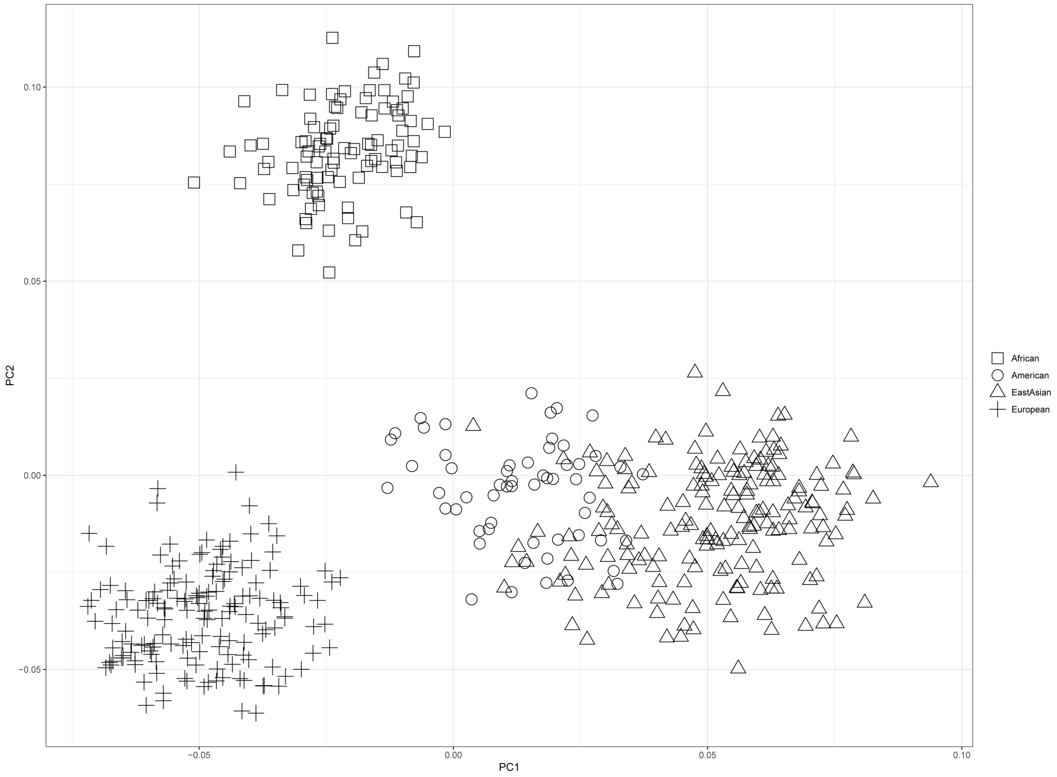


Figure 3 **Principal component analysis of four continental populations comprising 25 training populations.**

Full-size DOI: 10.7717/peerj.6508/fig-3

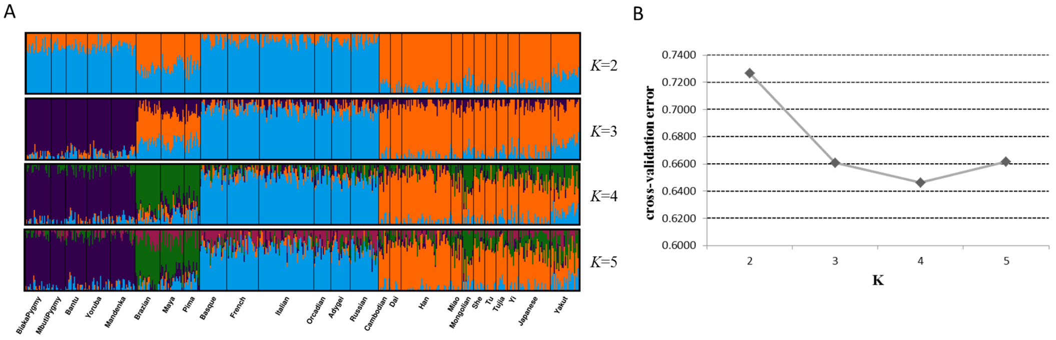


Figure 4 **Genetic structure analyses of 25 training populations at $K = 2$–5 (A) and cross-validation error of each K value (B) based on 48 SNPs.**

Full-size DOI: 10.7717/peerj.6508/fig-4

Next, twenty-two training populations were classified into three known populations (African, European and East Asian) according to their biogeographical origins; six testing populations were treated as unknown individuals. PCA of these populations was conducted, as shown in Fig. 5B. Result demonstrated that African, European and East Asian individuals formed three population clusters. Moreover, we found that these testing individuals were

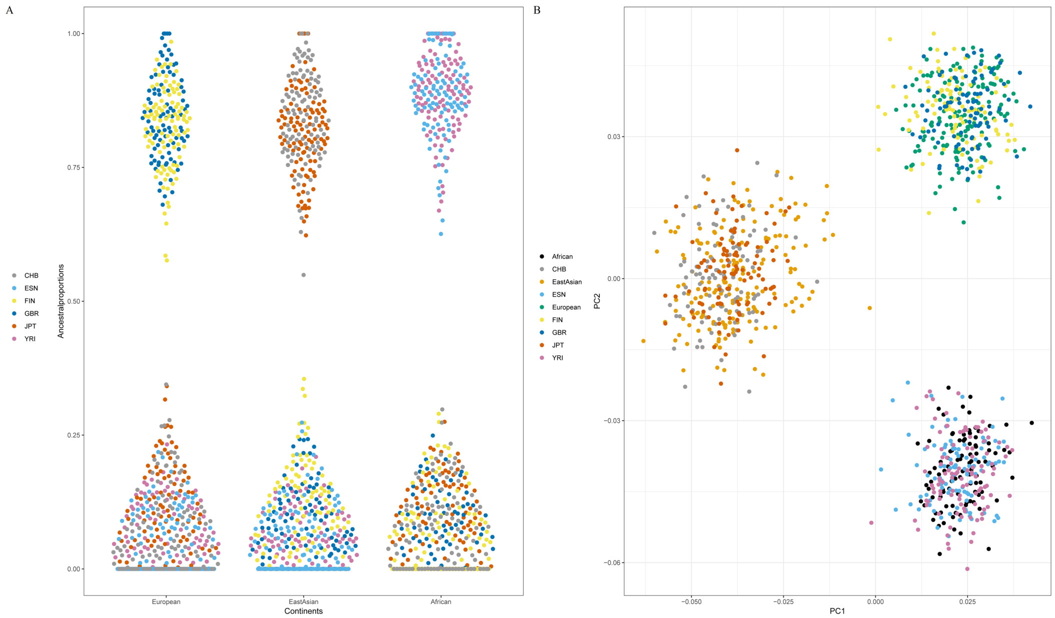


Figure 5 Ancestral origin analyses of six testing populations based on 48 SNPs. (A) genetic components of six testing populations by ADMIXTURE software v1.3. (B) Principal component analysis of six testing populations and three continental populations. Population abbreviations (CHB, ESN, FIN, GBR, JPT and YRI) are explained in Table 1.

Full-size DOI: 10.7717/peerj.6508/fig-5

superimposed onto the correct population clusters in Fig. 5B. Results of Naïve Bayes analysis also revealed all testing samples could be assigned to their corresponding continental regions (Table S4). For example, individual HG02922 was classified into African individuals with more than one billion times than European and East Asian individuals. From the above results, these 48 SNP set performed well for ancestry origin predictions of three continental populations (African, European and East Asian).

Discrimination efficiencies of 48 SNP loci for three Chinese populations

Ancestral allele frequencies of 48 SNP loci in Uygur, Han and Mongolian populations were given in Supplementary Fig. 2. Distinct frequency differences of most SNPs could be observed among pairwise populations. As an example, ancestral frequencies of rs1857859 locus in Uygur, Han and Mongolian populations were 0.65, 0.76 and 0.25, respectively, indicating the locus was good for distinguishing Mongolians from Uygurs and Hans. Population specific *In* values of 48 SNPs in these three Chinese populations were presented in Fig. 6 and Table S3. Results revealed that 10, 7 and 5 SNP loci displayed relatively high *In* values ($\geq$0.1) in Uygur, Han and Mongolian populations, respectively. Additionally, some SNP loci which had low *In* values in four continental populations showed relatively high *In* values in one of the three Chinese populations. For example, rs10852189 locus whose

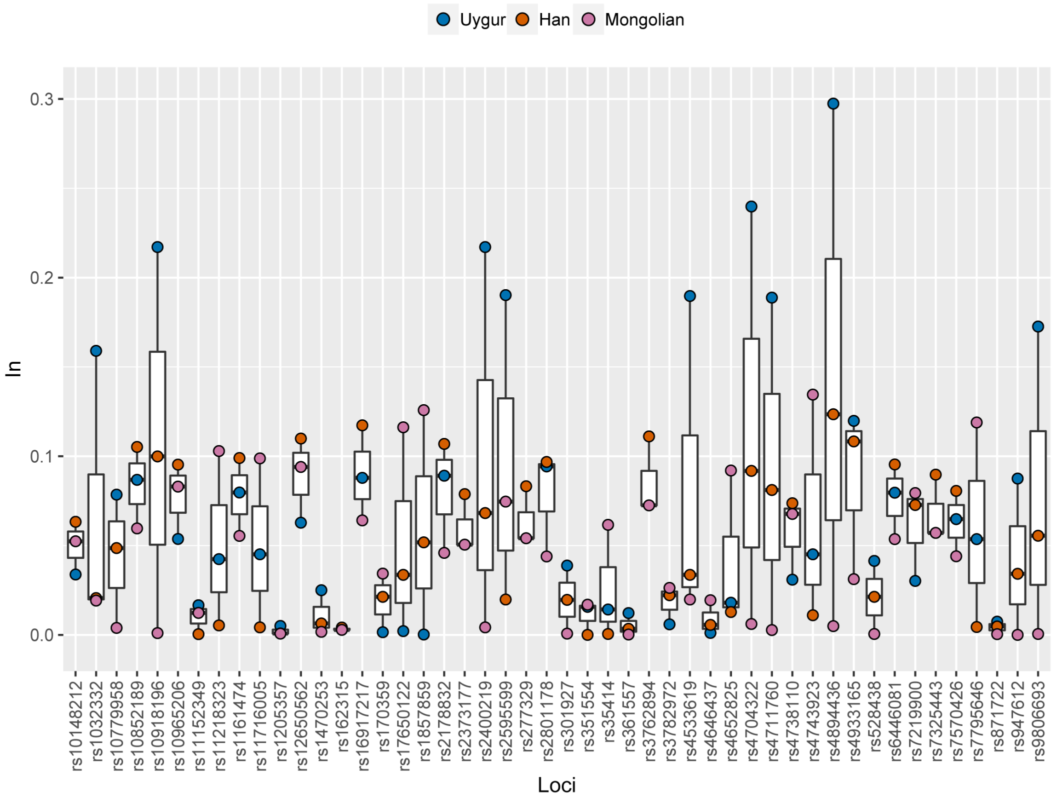


Figure 6 Population specific *In* values of 48 SNPs in Han, Uygur and Mongolian populations.
Full-size DOI: 10.7717/peerj.6508/fig-6

In values in four continental populations were less than 0.05 had *In* value with being more than 0.1 in Han Chinese population.

Next, we assessed the power of 48 SNPs for ancestry analyses of Han, Mongolian and Uygur populations. Firstly, population structure analysis of three Chinese populations at $K = 3$ was given in Fig. 7A. Different ancestral component distributions were seen among these populations: Han population showed high blue proportions; Mongolian group displayed high purple proportions; Uygur group exhibited high orange proportions. We also found some individuals showed admixture ancestry proportions, which might result from the recent admixtures of these populations. Nevertheless, individuals from three Chinese populations formed three distinct clusters in the PCA plot (Fig. 7B). Moreover, multinomial logistic regression analyses of three Chinese populations were conducted to further evaluate the efficiencies of 48 SNP loci (Table S5). Results indicated Uygur, Han and Mongolian individuals were correctly classified into their corresponding populations with the probability values of 1.0000, reflecting these 48 SNP loci could estimate biogeographical origins of these Chinese populations well.

DISCUSSION

Ancestry origin predictions of different continental populations could be achieved by some SNP assays (*Kidd et al., 2011*; *Kidd et al., 2014*; *Phillips et al., 2007*). However,

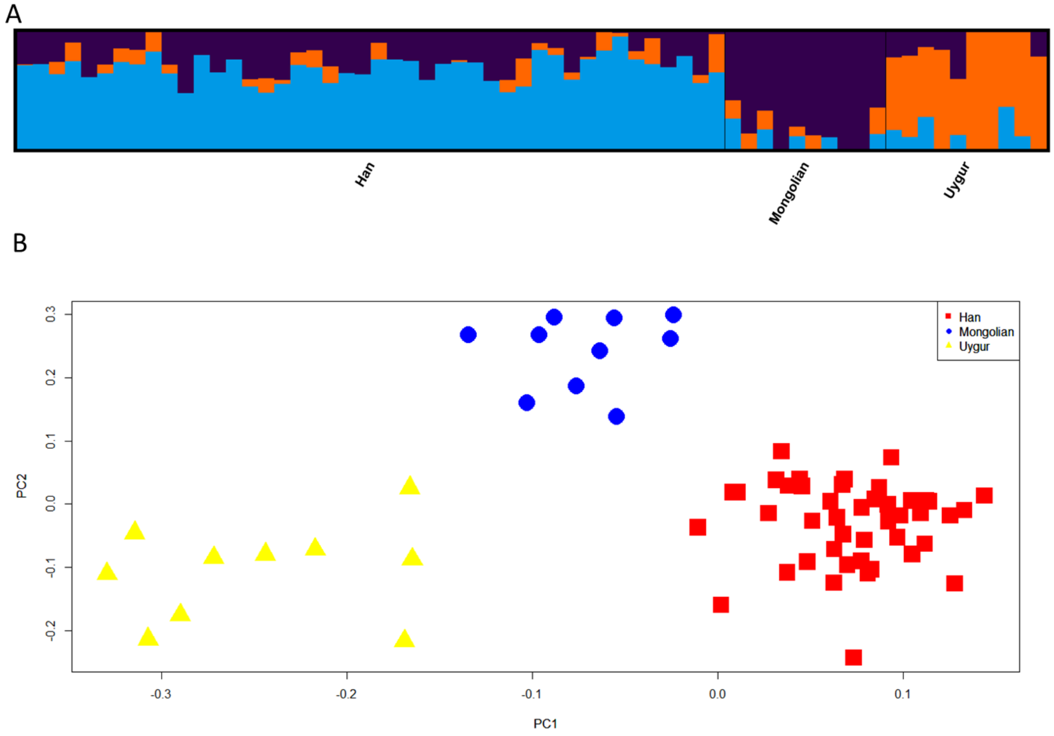


Figure 7 Genetic differentiation analyses among Han, Mongolian and Uygur populations in China. (A) Genetic structure analyses among these populations. (B) Principal component analysis of these populations.

Full-size DOI: 10.7717/peerj.6508/fig-7

Soundararajan et al. (2016) stated that ancestral resolution among continental populations might be insufficient for the forensic application. Therefore, several SNP panels which were used to improve the ancestral resolution of regional populations have been developed by some researchers (*Bulbul et al., 2018*; *Li et al., 2016*; *Phillips et al., 2013*). In the current study, we presented one set of SNPs which were utilized to differentiate different continental populations and three Chinese populations.

For continental populations (African, American, East Asian and European), we found these 48 SNPs could perform well for ancestry analyses of individuals from Africa, East Asia and Europe although some SNPs were not informative in these continental populations. Moreover, some American individuals were observed to be overlapped onto East Asian individual cluster (Fig. 3). Nonetheless, distinct genetic components among these populations were discerned from Fig. 4A. To obtain better ancestral resolutions among Americans and other continental populations, more highly differentiated SNPs in American populations should be selected. What's more, future research should be paid more attention to ancestry inferences of within-continental populations. The 1000 Genomes Project has assessed genetic variations of 2,504 individuals from five continents and found that some variants were unique to some populations (*Genomes Project Consortium et al., 2015*). Therefore, the variations private to one population should be selected to enhance resolution power among within-continental populations.

For Uygur, Han and Mongolian populations, previous research investigated their genetic variations based on different genetic markers (*Mei et al., 2016*; *Tao et al., 2018*). *Tao et al. (2018)* assessed genetic polymorphisms of 12X-STRs and found that the Mongolian group showed indistinguishable genetic component distributions when compared to the components of Han populations in different regions. *Mei et al. (2016)* conducted genetic differentiation analyses between Uygur and other reference populations based on 30 InDels and found that Uygur group was far from Han populations and other Chinese populations in the PCA plot. In this study, we selected 48 SNPs to differentiate Han, Mongolian and Uygur populations. Compared with the study for the identification of Japanese people, *Yuasa et al. (2018)* selected the SNPs with the Japanese-specific alleles to differentiate Japanese individuals from the other East Asian populations. The similar method could be employed to select the SNPs with population-specific alleles so as to obtain better ancestral distinctions among different ethnic groups in China. Moreover, some research concerning on genomic analysis of a great many of Chinese individuals has been reported (*Chiang et al., 2018*; *Liu et al., 2018*). We will make full use of these data to further screen those highly differentiated genetic markers in different ethnic groups in China in the future.

Although these 48 SNP loci perform well for ancestry origin inferences of three continental populations (African, East Asian and European) and three Chinese populations (Han, Mongolian and Uygur), some loci of the 48 SNPs who had lower *In* values were not suggested to perform ancestry analyses of these populations so that we could obtain more accurate results for ancestry origin predictions. For example, SNPs whose *In* values in continental populations were less than 0.1 should not be employed to infer biogeographical origins of continental populations given that high genetic differentiations among different continental populations existed; SNPs whose *In* values in subpopulations from the same regions were less than 0.05 should not be utilized to perform ancestry analysis among these populations. Furthermore, there were some limitations in the current study. On one hand, the validation of 48 SNP loci in new samples was not conducted, especially for Han, Mongolian and Uygur populations. On the other hand, sample sizes of Uygur (10), Han (44) and Mongolian (10) were relatively small. Accordingly, it is necessary for us to further validate the efficiency of these 48 SNP loci for ancestry analysis in larger samples.

CONCLUSIONS

To conclude, forty-eight SNPs were provided to differentiate different continental populations, which could provide valuable clues for forensic investigations. Furthermore, these SNPs could also be utilized to differentiate three populations residing in China, which achieved fine-scale resolutions in regional populations. Further validation for these 48 SNPs should be conducted in the large sample set. What's more, the SNPs with population-specific alleles should be screened to obtain better ancestral resolutions among regional populations.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the Guangdong Province Universities and Colleges Pearl River Scholar Funded Scheme (GDUPS, 2017). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Guangdong Province Universities and Colleges Pearl River Scholar Funded Scheme (GDUPS, 2017).

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Xiao-Ye Jin analyzed the data, approved the final draft.
- Yuan-Yuan Wei and Qiong Lan contributed reagents/materials/analysis tools, approved the final draft.
- Wei Cui and Chong Chen prepared figures and/or tables, approved the final draft.
- Yu-Xin Guo and Ya-Ting Fang authored or reviewed drafts of the paper, approved the final draft.
- Bo-Feng Zhu conceived and designed the experiments, authored or reviewed drafts of the paper, approved the final draft.

Data Availability

The following information was supplied regarding data availability:

Genetic data of 48 SNPs used in this study were given in Table S1.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.6508#supplemental-information.

REFERENCES

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19**:1655–1664 DOI 10.1101/gr.094052.109.

Amigo J, Salas A, Phillips C, Carracedo A. 2008. SPSmart: adapting population based SNP genotype databases for fast and comprehensive web access. *BMC Bioinformatics* **9**:428 DOI 10.1186/1471-2105-9-428.

Bulbul O, Speed WC, Gurkan C, Soundararajan U, Rajeevan H, Pakstis AJ, Kidd KK. 2018. Improving ancestry distinctions among Southwest Asian populations. *Forensic Science International-Genetics* **35**:14–20 DOI 10.1016/j.fsigen.2018.03.010.

Chiang CWK, Mangul S, Robles C, Sankararaman S. 2018. A comprehensive map of genetic variation in the world's largest ethnic Group-Han Chinese. *Molecular Biology and Evolution* **35**:2736–2750 DOI 10.1093/molbev/msy170.

Frudakis T, Venkateswarlu K, Thomas MJ, Gaskin Z, Ginjupalli S, Gunturi S, Ponnuswamy V, Natarajan S, Nachimuthu PK. 2003. A classifier for the SNP-based inference of ancestry. *Journal of Forensic Sciences* **48**:771–782.

Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. 2015. A global reference for human genetic variation. *Nature* **526**:68–74 DOI 10.1038/nature15393.

Kidd JR, Friedlaender FR, Speed WC, Pakstis AJ, De La Vega FM, Kidd KK. 2011. Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples. *Investigative Genetics* **2(1)**:1 DOI 10.1186/2041-2223-2-1.

Kidd KK, Speed WC, Pakstis AJ, Furtado MR, Fang R, Madbouly A, Maiers M, Middha M, Friedlaender FR, Kidd JR. 2014. Progress toward an efficient panel of SNPs for ancestry inference. *Forensic Science International Genetics* **10**:23–32 DOI 10.1016/j.fsigen.2014.01.002.

Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I. 2015. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Molecular Ecology Resources* **15**:1179–1191 DOI 10.1111/1755-0998.12387.

Li CX, Pakstis AJ, Jiang L, Wei YL, Sun QF, Wu H, Bulbul O, Wang P, Kang LL, Kidd JR. 2016. A panel of 74 AISNPs: improved ancestry inference within Eastern Asia. *Forensic Science International Genetics* **23**:101–110 DOI 10.1016/j.fsigen.2016.04.002.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**:1100–1104 DOI 10.1126/science.1153717.

Liu S, Huang S, Chen F, Zhao L, Yuan Y, Francis SS, Fang L, Li Z, Lin L, Liu R, Zhang Y, Xu H, Li S, Zhou Y, Davies RW, Liu Q, Walters RG, Lin K, Ju J, Korneliussen T, Yang MA, Fu Q, Wang J, Zhou L, Krogh A, Zhang H, Wang W, Chen Z, Cai Z, Yin Y, Yang H, Mao M, Shendure J, Wang J, Albrechtsen A, Jin X, Nielsen R, Xu X. 2018. Genomic analyses from non-invasive prenatal testing reveal genetic associations, patterns of viral infections, and chinese population history. *Cell* **175**:347–359 DOI 10.1016/j.cell.2018.08.016.

Mei T, Shen CM, Liu YS, Meng HT, Zhang YD, Guo YX, Dong Q, Wang XX, Yan JW, Zhu BF, Zhang LP. 2016. Population genetic structure analysis and forensic evaluation of Xinjiang Uigur ethnic group on genomic deletion and insertion polymorphisms. *Springerplus* **5(1)**:1087 DOI 10.1186/s40064-016-2730-3.

Pakstis AJ, Kang L, Liu L, Zhang Z, Jin T, Grigorenko EL, Wendt FR, Budowle B, Hadi S, Al Qahtani MS, Morling N, Mogensen HS, Themudo GE, Soundararajan U, Rajeevan H, Kidd JR, Kidd KK. 2017. Increasing the reference populations for the 55 AISNP panel: the need and benefits. *International Journal of Legal Medicine* **131**:913–917 DOI 10.1007/s00414-016-1524-z.

Phillips C. 2015. Forensic genetic analysis of bio-geographical ancestry. *Forensic Science International-Genetics* **18**:49–65 DOI 10.1016/j.fsigen.2015.05.012.

Phillips C, Freire Aradas A, Kriegel AK, Fondevila M, Bulbul O, Santos C, Serrulla Rech F, Perez Carceles MD, Carracedo A, Schneider PM, Lareu MV. 2013. Eurasiaplex: a forensic SNP assay for differentiating European and South Asian ancestries. *Forensic Science International Genetics* **7**:359–366 DOI 10.1016/j.fsigen.2013.02.010.

Phillips C, Salas A, Sánchez JJ, Fondevila M, Gómez-Tato A, Alvarez-Dios J, Calaza M, De Cal MC, Ballard D, Lareu MV. 2007. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Science International Genetics* **1**:273–280 DOI 10.1016/j.fsigen.2007.06.008.

Qin P, Li Z, Jin W, Lu D, Lou H, Shen J, Jin L, Shi Y, Xu S. 2014. A panel of ancestry informative markers to estimate and correct potential effects of population stratification in Han Chinese. *European Journal of Human Genetics* **22**:248–253 DOI 10.1038/ejhg.2013.111.

R Core Team. 2016. R: a language and environment for statistical computing. Version 3.3.0. Vienna: R Foundation for Statistical Computing. *Available at https://www.R-project.org/*.

Rosenberg NA, Li LM, Ward R, Pritchard JK. 2003. Informativeness of genetic markers for inference of ancestry. *American Journal of Human Genetics* **73**:1402–1422 DOI 10.1086/380416.

Rousset F. 2008. genepop’007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources* **8**:103–106 DOI 10.1111/j.1471-8286.2007.01931.x.

Shriver MD, Smith MW, Jin L, Marcini A, Akey JM, Deka R, Ferrell RE. 1997. Ethnic-affiliation estimation by use of population-specific DNA markers. *American Journal of Human Genetics* **60**:957–964.

Soundararajan U, Yun LB, Shi MS, Kidd KK. 2016. Minimal SNP overlap among multiple panels of ancestry informative markers argues for more international collaboration. *Forensic Science International-Genetics* **23**:25–32 DOI 10.1016/j.fsigen.2016.01.013.

Sun K, Ye Y, Luo T, Hou Y. 2016. Multi-InDel analysis for ancestry inference of sub-populations in China. *Scientific Reports* **6**:39797 DOI 10.1038/srep39797.

Tao R, Zhang J, Bian Y, Dong R, Liu X, Jin C, Zhu R, Zhang S, Li C. 2018. Investigation of 12 X-STR loci in Mongolian and Eastern Han populations of China with comparison to other populations. *Scientific Reports* **8(1)**:4287 DOI 10.1038/s41598-018-22665-3.

Wang Z, Lu B, Jin X, Yan J, Meng H, Zhu B. 2018. Genetic and structural characterization of 20 autosomal short tandem repeats in the Chinese Qinghai Han population and its genetic relationships and interpopulation differentiations with other reference populations. *Forensic Sciences Research* **3**:145–152 DOI 10.1080/20961790.2018.1485199.

Xu S, Huang W, Qian J, Jin L. 2008. Analysis of genomic admixture in Uyghur and its implication in mapping strategy. *American Journal of Human Genetics* **82**:883–894 DOI 10.1016/j.ajhg.2008.01.017.

Yuasa I, Akane A, Yamamoto T, Matsusue A, Endoh M, Nakagawa M, Umetsu K, Ishikawa T, Iino M. 2018. Japaneseplex: a forensic SNP assay for identification of Japanese people using Japanese-specific alleles. *Legal Medicine* **33**:17–22 DOI 10.1016/j.legalmed.2018.04.008.

Zhang YD, Tang XL, Meng HT, Wang HD, Jin R, Yang CH, Yan JW, Yang G, Liu WJ, Shen CM, Zhu BF. 2015. Genetic variability and phylogenetic analysis of Han population from Guanzhong region of China based on 21 non-CODIS STR loci. *Scientific Reports* **5**:8872 DOI 10.1038/srep08872.

Zhao T, Lee TD. 1989. Gm-Allotypes and Km-Allotypes in 74 Chinese populations—a hypothesis of the origin of the Chinese nation. *Human Genetics* **83**:101–110 DOI 10.1007/Bf00286699.

Zhao TM, Zhang GL, Zhu YM, Zheng SQ, Gu WJ, Chen Q, Zhang X, Liu DY. 1991. Study on immunoglobulin allotypes in the Chinese: a hypothesis of the origin of the Chinese nation. *Yi Chuan Xue Bao* **18**:97–108.