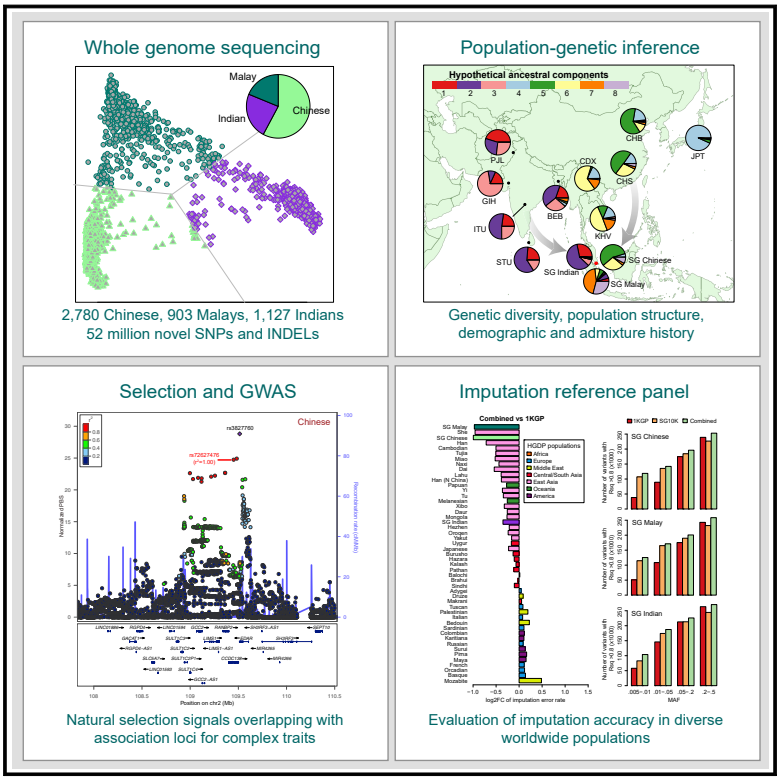
Resource

Cell

Large-Scale Whole-Genome Sequencing of Three Diverse Asian Populations in Singapore

Graphical Abstract



Highlights

- Discovery of 52 million novel variants by 13.7× WGS of 4,810 Singaporeans
- Insights into population structure and evolutionary history of Asians
- Identification of 20 loci under selection that are enriched for GWAS signals
- Substantial improvement of imputation in diverse Asian and Oceanian populations

Authors

Degang Wu, Jinzhuang Dou, Xiaoran Chai, ..., SG10K Consortium, Jianjun Liu, Chaolong Wang

Correspondence

liuj3@gis.a-star.edu.sg (J.L.), chaolong@hust.edu.cn (C.W.)

In Brief

Because of Singapore's unique history of immigration, whole-genome sequence analysis of 4,810 Singaporeans provides a snapshot of the genetic diversity across East, Southeast, and South Asia.

Wu et al., 2019, Cell *179*, 736–749
October 17, 2019
https://doi.org/10.1016/j.cell.2019.09.019

CellPress

Cell

Large-Scale Whole-Genome Sequencing of Three Diverse Asian Populations in Singapore

Degang Wu,[1,2,38] Jinzhuang Dou,[1,3,38] Xiaoran Chai,[1,32,38] Claire Bellis,[4,5] Andreas Wilm,[6] Chih Chuan Shih,[6] Wendy Wei Jia Soon,[7] Nicolas Bertin,[1] Clarabelle Bitong Lin,[4] Chiea Chuen Khor,[4] Michael DeGiorgio,[8] Shanshan Cheng,[2] Li Bao,[2] Neerja Karnani,[9,10] William Ying Khee Hwang,[11,12] Sonia Davila,[13,14] Patrick Tan,[13,15,16,17] Asim Shabbir,[18]

(Author list continued on next page)

[1]Computational and Systems Biology, Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore 138672, Singapore
[2]Department of Epidemiology and Biostatistics, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430030, China
[3]Department of Bioinformatics and Computational Biology, Division of Quantitative Sciences, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[4]Human Genetics, Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore 138672, Singapore
[5]Genomics Research Centre, Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane, QLD 4000, Australia
[6]Bioinformatics Core, Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore 138672, Singapore
[7]Integrated Genome Analysis Platform (iGAP), Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore 138672, Singapore
[8]Department of Computer and Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL 33431, USA
[9]Singapore Institute for Clinical Sciences, Agency for Science, Technology and Research, Singapore 117609, Singapore
[10]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117597, Singapore
[11]Department of Haematology, Singapore General Hospital, Singapore 169608, Singapore
[12]National Cancer Centre Singapore, Singapore 169610, Singapore
[13]SingHealth Duke-NUS Institute of Precision Medicine, Singapore 169856, Singapore
[14]Cardiovascular and Metabolic Disorders Program, Duke-NUS Medical School, Singapore 169857, Singapore
[15]Cancer and Stem Biology Program, Duke-NUS Medical School, Singapore 169857, Singapore
[16]Cancer Science Institute of Singapore, National University of Singapore, Singapore 117599, Singapore
[17]Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore 138672, Singapore
[18]Department of Surgery, National University Health System, Singapore 119228, Singapore
[19]Clinical Research Unit, Khoo Teck Puat Hospital, Singapore 768828, Singapore
[20]Department of Neurology, National Neuroscience Institute, Singapore General Hospital Campus, Singapore 169608, Singapore
[21]Neuroscience and Behavioural Disorders Program, Duke-NUS Medical School, Singapore 169857, Singapore
[22]Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore 308232, Singapore
[23]Personalised Medicine Service, Tan Tock Seng Hospital, Singapore 308433, Singapore
[24]Clinical Research and Innovation Office, Tan Tock Seng Hospital, Singapore 308433, Singapore
[25]Cardiovascular Research Institute, National University Heart Centre, Singapore 119074, Singapore

(Affiliations continued on next page)

SUMMARY

Underrepresentation of Asian genomes has hindered population and medical genetics research on Asians, leading to population disparities in precision medicine. By whole-genome sequencing of 4,810 Singapore Chinese, Malays, and Indians, we found 98.3 million SNPs and small insertions or deletions, over half of which are novel. Population structure analysis demonstrated great representation of Asian genetic diversity by three ethnicities in Singapore and revealed a Malay-related novel ancestry component. Furthermore, demographic inference suggested that Malays split from Chinese ~24,800 years ago and experienced significant admixture with East Asians ~1,700 years ago, coinciding with the Austronesian expansion. Additionally, we identified 20 candidate loci for natural selection, 14 of which harbored robust associations with complex traits and diseases. Finally, we show that our data can substantially improve genotype imputation in diverse Asian and Oceanian populations. These results highlight the value of our data as a resource to empower human genetics discovery across broad geographic regions.

INTRODUCTION

We have gained profound insights into the population history and genetic basis of phenotype diversity and disease etiology by

Angela Moh,[19] Eng-King Tan,[20,21] Jia Nee Foo,[4,22] Liuh Ling Goh,[23] Khai Pang Leong,[24] Roger S.Y. Foo,[4,25,26] Carolyn Su Ping Lam,[27,28,25,29,30] Arthur Mark Richards,[25,26,31] Ching-Yu Cheng,[32,33,34] Tin Aung,[32,33,34] Tien Yin Wong,[32,33,34] Huck Hui Ng,[10,35,36,37] SG10K Consortium, Jianjun Liu,[4,26,*] and Chaolong Wang[2,1,39,*]
[26]Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117597, Singapore
[27]Department of Cardiology, National Heart Centre Singapore, Singapore 169609, Singapore
[28]Cardiovascular Sciences Academic Clinical Program, Duke-NUS Medical School, Singapore 169857, Singapore
[29]Department of Cardiology, University Medical Center Groningen, Groningen 9713GZ, the Netherlands
[30]The George Institute for Global Health, Newtown, NSW 2042, Australia
[31]Christchurch Heart Institute, University of Otago, Christchurch 8011, New Zealand
[32]Singapore Eye Research Institute, Singapore National Eye Centre, Singapore 169856, Singapore
[33]Ophthalmology and Visual Sciences Academic Clinical Program (Eye ACP), Duke-NUS Medical School, Singapore 169857, Singapore
[34]Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117597, Singapore
[35]Stem Cell and Regenerative Biology, Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore 138672
[36]Department of Biological Sciences, National University of Singapore, Singapore 117558, Singapore
[37]School of Biological Sciences, Nanyang Technological University, Singapore 637551, Singapore
[38]These authors contributed equally
[39]Lead Contact
*Correspondence: liuj3@gis.a-star.edu.sg (J.L.), chaolong@hust.edu.cn (C.W.)
https://doi.org/10.1016/j.cell.2019.09.019

mining variation in human genomes (Fan et al., 2016; MacArthur et al., 2017; Nielsen et al., 2017; Timpson et al., 2018). Although earlier efforts were mostly based on array genotyping of common variants, whole-genome sequencing (WGS) has become a powerful approach for understanding human evolutionary and migration history and discovering disease-related genetic variants by surveying the genome in an unbiased and comprehensive fashion. A remarkable milestone is the 1000 Genomes Project (1KGP), which sequenced >2,500 genomes from 26 populations at ~7.4× and cataloged over 88 million variants (1000 Genomes Project Consortium et al., 2015). The resource provided by the 1KGP has empowered numerous genome-wide association studies (GWASs) through imputation, allowing detection of genetic association at low-frequency variants, thus enabling a deeper understanding of the genetic architecture of complex diseases (Das et al., 2016; Timpson et al., 2018). Direct assessment of causal variants enabled by large-scale sequencing analysis has led to the convergence of population and clinical genetics (Ashley, 2016; MacArthur et al., 2014), where population genetic information has become the cornerstone of precision medicine, with applications in the diagnosis of Mendelian diseases (Manrai et al., 2016; Rehm et al., 2015), optimization of medication usage (Relling and Evans, 2015), drug development (Nelson et al., 2015), and disease risk prediction (Chatterjee et al., 2016).

Comparing across populations, most genetic variants are rare and population specific, highlighting the importance of characterizing local population diversity (Hindorff et al., 2018; Manrai et al., 2016). Consequently, many countries have initiated population-based WGS studies (Gudbjartsson et al., 2015; Genome of the Netherlands Consortium, 2014; UK10K Consortium et al., 2015). Nevertheless, the Eurocentric biases in human genetics research have caused intensifying concerns about the exacerbation of health disparities during the implementation of precision medicine (Martin et al., 2019). Although Asia is the largest continent, with ~4.5 billion inhabitants of diverse ethnicities (Abdulla et al., 2009) accounting for 60% of the global population, Asian genomes are underrepresented in public databases. For example, the Haplotype Reference Consortium has brought together 32,488 genomes from 20 studies to construct a large imputation reference panel (McCarthy et al., 2016), which, however, is only suitable for imputing Europeans. Moreover, in the Genome Aggregation Database (gnomAD), only ~5% of its 15,496 genomes are from East Asians and none are from South Asians. Furthermore, in the Trans-Omics for Precision Medicine (TOPMed) Program, only ~10% of its 145,000 samples are Asians (Taliun et al., 2019). In Asia, most published WGS studies have small sample sizes (Bai et al., 2018; Wong et al., 2013, 2014), except for two recent studies of Chinese by 1.7× and 0.1× WGS, respectively (Chiang et al., 2018; Liu et al., 2018). Extremely shallow sequencing, however, has low sensitivity to detect rare variants and, thus, limited power for evolutionary inferences and other population genetics analyses (Li et al., 2011; Han et al., 2014).

Despite its small geographic size, Singapore (SG) has diverse populations because of its immigratory history. Singaporeans are broadly classified into four ethnicities: Chinese, Malays, Indians, and others. 74.3% of Singapore residents are Chinese, descending from several dialect groups from south China with a minority from north China. Malays, representing 13.4% of the population, include descendants of diverse Austronesian-speaking groups in Southeast Asia, primarily from Malaysia and Indonesia. About 9.1% of the population are Indians descending from Indian migrants during the period of British colonization. Most Indians are Telugas and Tamils from southeast India, and a minority are Sikhs and Pathans from north India (Teo et al., 2009). The remaining 3.2% of the population are mainly Europeans and Middle Easterners. The three major ethnicities in Singapore together provide a unique snapshot of the genetic diversity across East Asia, Southeast Asia, and South Asia. Therefore, WGS analysis of Singaporeans has the potential to benefit populations across Asia and the remainder of the globe. Here we describe initial findings from the pilot study of the SG10K project, a national effort aiming to whole-genome-sequence 10,000 Singaporeans.

RESULTS

Data Quality

The SG10K pilot study included 4,810 samples sequenced at a mean depth of 13.7×, including 2,780 Chinese, 903 Malays,

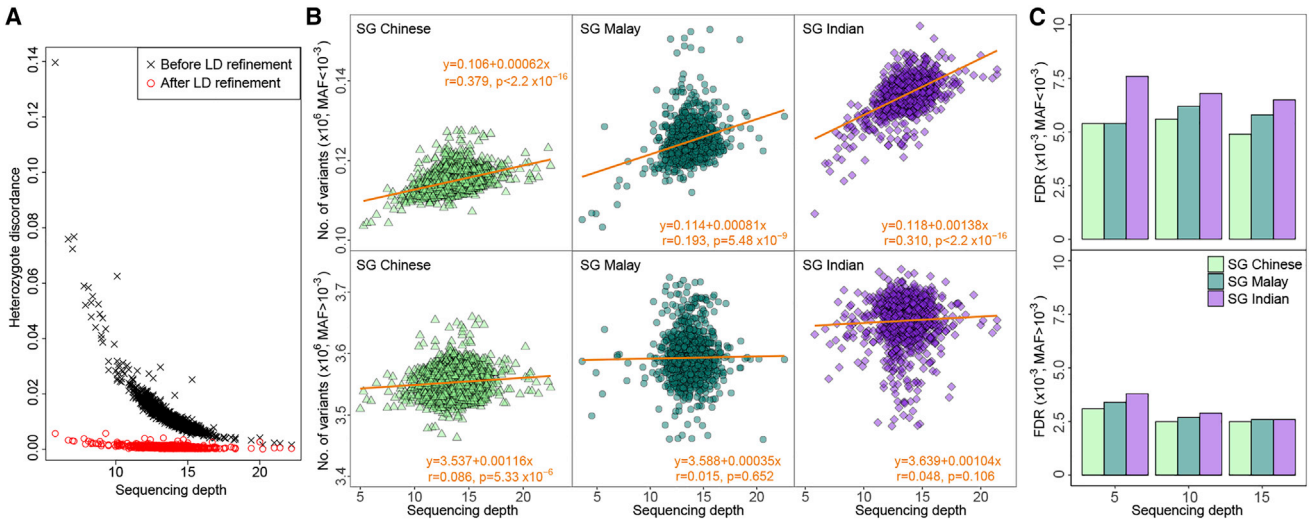


Figure 1. Quality Evaluation of the SG10K Call Set

(A) Heterozygote discordance rate versus sequencing depth for 1,263 array-genotyped samples.

(B) Estimation of non-reference sensitivity by modeling the number of variants detected in each sample as a function of sequencing depth. In each subpanel, outliers more than 5 SD from the mean sequencing depth or the mean number of variants were removed and are not shown.

(C) FDR of variant detection in 9 samples by comparison with deep WGS data.

See also Table S2 and Figures S1 and S2.

and 1,127 Indians (STAR Methods; Table S1; Figures S1 and S2). We performed linkage disequilibrium (LD)-based joint calling, followed by quality controls to remove low-quality variants and by population-based phasing to obtain haplotypes. The final dataset consisted of 89,160,286 SNPs and 9,113,420 insertions or deletions (indels) from 22 autosomes and the X chromosome. By removing relatedness up to the third degree, we obtained a subset of 4,441 unrelated individuals, including 2,535 healthy individuals.

We estimated a transition-to-transversion ratio (Ts/Tv) of 2.07 across 84.7 million bi-allelic autosomal SNPs in our final dataset, consistent with values reported by previous studies (1000 Genomes Project Consortium et al., 2015; UK10K Consortium et al., 2015). We further evaluated the quality of our call set using 1,263 individuals previously genotyped (Cornes et al., 2012). Treating the array data as the gold standard, we achieved a 0.9997 non-reference sensitivity and a 0.9992 heterozygote concordance rate for variants with minor allele frequency (MAF) > 0.01 (Table S2). For low-frequency variants with MAF between 0.01 and 0.05, both the sensitivity and heterozygote concordance rate dropped slightly to 0.9991 and 0.9973, respectively. These statistics indicate the high quality of our call set, mainly because the LD-based refinement step dramatically reduced the discordance rate; e.g., by a 25-fold reduction from ~0.075 to ~0.003 for samples sequenced at ~7× and by 10-fold from ~0.01 to ~0.001 for those sequenced at ~13.7× (Figure 1A).

Because rare variants were poorly covered by GWAS arrays, we designed a novel approach to estimate the sensitivity without external data (STAR Methods; Figures 1B). For each sample, we counted the number of non-reference variants at extremely low frequency (MAF < 0.001). Within each population, we observed that samples sequenced at lower depths carried fewer rare variants (Pearson's $r > 0.2$, $p < 10^{-8}$, t test), indicating lower sensitivity. In contrast, the mutation burden did not have an obvious trend with sequencing depth for variants with MAF > 0.001 ($r < 0.07$). We estimated 0.9329, 0.9225, and 0.8887 non-reference sensitivities at detecting variants with MAF < 0.001 for Chinese, Malays, and Indians, respectively. In contrast, we achieved >0.996 sensitivity to detect variants with MAF > 0.001 in all populations. Furthermore, by direct comparison with deep (>30×) WGS of 9 samples that were originally sequenced at 5×, 10×, and 15×, we estimated the false discovery rates (FDRs) to be ~0.6% and ~0.3% for detecting variants with a MAF of <0.001 and >0.001, respectively (Figure 1C). Compared with Chinese and Malays, the lower sensitivity and higher FDR at detecting rare variants in Indians might be explained by their higher genetic diversity.

Novel Variants and Implications for Genetic Diagnosis

In our call set, 45.6 million SNPs (51%) and 6.3 million indels (70%) were not cataloged in dbSNP (v.150). This higher proportion of novel indels might be partially attributed to multiple representations and ambiguous positions of the same indels in previous studies (Tan et al., 2015). Chromosome X has a higher proportion of novel variants than the autosomes across variant types (Table S3). Unsurprisingly, most novel variants were extremely rare (Figure S3A; 1000 Genomes Project Consortium et al., 2015), with singletons and doubletons accounting for 69.6% and 14.4% of the novel SNPs, respectively, and only 0.5% of the novel variants reaching MAF > 0.01.

Among the common (MAF > 0.01) novel variants, 126 were annotated as "deleterious." These variants were within 113 genes, including 99 harboring at least one pathogenic or likely pathogenic variant according to ClinVar (Figure S3B; Landrum et al., 2014).

Table 1. The Median Number of Autosomal Variants per Genome

Annotation	SG Chinese (n = 1,267, Depth = 13.7×)		SG Malay (n = 454, Depth = 13.8×)		SG Indian (n = 814, Depth = 13.6×)	
	No.	Het/Hom	No.	Het/Hom	No.	Het/Hom
SNP	3,308,882	1.43	3,346,583	1.51	3,406,391	1.73
Insertion	103,206	1.50	106,398	1.61	116,305	1.91
Deletion	158,864	1.85	162,288	1.98	170,818	2.33
SNP not in dbSNP	27,635	12.52	30,155	13.55	30,263	14.02
Insertion not in dbSNP	6,527	23.16	7,089	24.49	8,928	25.53
Deletion not in dbSNP	6,059	25.55	6,478	26.08	8,058	25.79
Synonymous	11,675	1.45	11,844	1.55	12,077	1.78
Missense	11,509	1.51	11,666	1.60	11,911	1.81
Exon	127,268	1.46	128,961	1.56	131,579	1.78
Intron	1,859,670	1.46	1,884,753	1.55	1,925,723	1.77
UTR	53,751	1.46	54,418	1.55	55,564	1.79
TFBS	4,911	1.61	4,995	1.72	5,141	1.99
SIFT: deleterious	1,672	2.81	1,697	3.00	1,715	3.48
PolyPhen: probably damaging	877	3.34	896	3.56	902	4.15
PolyPhen: possibly damaging	1,150	2.92	1,165	3.04	1,182	3.52
ClinVar: pathogenic	30	2.33	31	2.67	33	3.38
ClinVar: association	46	1.80	45	2.14	43	2.28
ClinVar: risk factor	82	1.45	82	1.61	82	1.89

For ClinVar annotations, pathogenic variants are those interpreted for Mendelian disorders. association variants are those identified by GWASs and further interpreted for their clinical significance, and risk factor variants are those interpreted not to cause a disorder but to increase the disease risk. See also Tables S1 and S3 and Figure S3.

However, the high frequencies of these variants in Singapore suggest that they are likely benign or have very low penetrance (Yang et al., 2017). Furthermore, none of the 126 deleterious mutations were specific to one ethnicity, whereas 35 reached MAF > 0.05 in all three ethnicities (Figure S3C).

Variants in a Personal Genome

To compare typical genomes from different ethnicities, we focused on 2,535 healthy individuals. On average, a genome carries autosomal variants at ~3,330,000 SNPs, ~105,000 insertions, and 160,000 deletions (Table 1), including ~11,600 missense variants, ~54,000 variants overlapping with UTRs, ~5,000 variants at transcription factor binding sites (TFBSs), ~2,000 deleterious variants predicted by SIFT (Kumar et al., 2009; Sim et al., 2012) or PolyPhen (Adzhubei et al., 2013), and ~31 pathogenic variants annotated by ClinVar. Moreover, Indians possessed the highest number of variants on average, followed by Malays and Chinese. In addition, the heterozygote to non-reference homozygote ratio (Het/Hom) was also higher in Indians (1.73) than in Chinese (1.43) and Malays (1.51), reflecting a higher level of genetic diversity in Indians. Although Malays have a lower level of genetic diversity than Indians (Teo et al., 2009), sequencing a Malay genome would lead to discovery of ~30.2 thousand novel SNPs, which was close to ~30.3 thousand for an Indian and higher than ~27.6 thousand for a Chinese, reflecting severe underrepresentation of Southeast Asians in current genetic studies. We note that the Het/Hom ratio was very high for novel variants because most of the novel variants were rare and often presented as heterozygotes. Furthermore, except for novel variants, the highest Het/Hom ratio was found among deleterious or pathogenic variants, consistent with negative selection against these variants to reach high frequencies.

The ClinVar pathogenic variants are interpreted for Mendelian disorders and might lead to adverse clinical outcomes when present in the homozygous state, especially for recessive disorders. Even without any known major diseases, an individual carried 3.9 ± 2.0 (mean ± SD), 4.3 ± 2.1, and 4.9 ± 2.2 pathogenic homozygotes in Chinese, Malays, and Indians, respectively. Moreover, individuals with higher inbreeding coefficients tended to have more pathogenic homozygotes (Figure 2), although such a trend might be underestimated because our analysis was restricted to healthy individuals. We also noticed a long tail in the distribution of inbreeding coefficients of SG Indians, consistent with a high level of consanguinity in Dravidian south India and Pakistan (Bittles and Black, 2010) and with an excess of the long runs of homozygosity in South Asians (Auton et al., 2009). Because consanguineous mating can lead to an excess of infant/childhood mortality and extended morbidity (Bittles and Black, 2010), these results have significant implications in public health. We estimated the prevalence of consanguineous mating between second cousins or closer relatives to be 29.1% in SG Indians, followed by 10.8% in SG Malays and 2.6% in SG Chinese.

Population Structure and Genetic Diversity

We analyzed our SG10K data together with the 1KGP populations using principal-components analysis (PCA; Wang et al.,

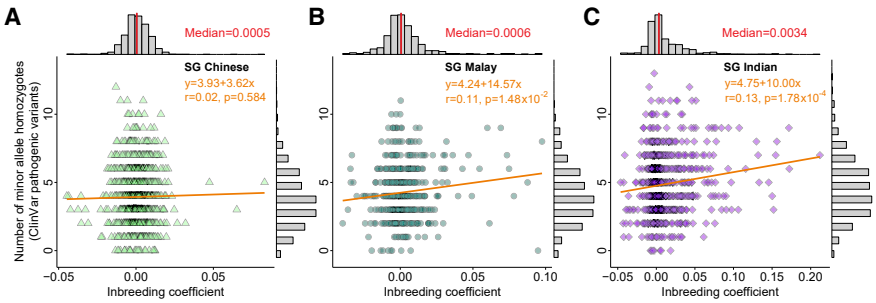


Figure 2. Number of Pathogenic Homozygotes in Each Healthy Individual as a Function of Inbreeding Coefficient for Chinese, Malays, and Indians

(A–C) Number of pathogenic homozygotes in each healthy individual as a function of inbreeding coefficient for Chinese (A), Malays (B), and Indians (C). Distributions of pathogenic homozygotes and inbreeding coefficients are shown on the right side and on top of each panel, respectively. See also Figure S2.

2015). We found that SG Indians and SG Chinese overlapped largely with South and East Asians, respectively, whereas SG Malays formed a distinct cluster (Figure 3A; Figure S4). When restricting PCA to East and Southeast Asians, most SG Chinese overlapped with southern Han Chinese (CHS) from south China, with a minority overlapping with Han Chinese in Beijing (CHB) from north China (Figures S4C and S4D). Similarly, most SG Indians overlapped with Sri Lankan Tamil (STU), Telugu (ITU), and Bengali (BEB) from the south Indian subcontinent, with a small proportion overlapping with Gujarati (GIH) from west India and Punjabi (PJL) from Pakistan (Figures S4E and S4F). Given the clear north-south pattern on both PCA of East Asians and South Asians, we estimated that 96% and 4% of SG Chinese are from south and north China, respectively, and that 81% and 19% of SG Indians are from the south and north Indian subcontinent, respectively (STAR Methods). In addition, neither SG Chinese nor Malays overlapped with Chinese Dai (CDX) and Kinh (KHV) from mainland Southeast Asia. These PCA results were consistent with the genetic distances between populations (Figure S5), in which the SG Malay was relatively distant from the others, with the closest being KHV (F_{ST} = 0.007) and CDX (F_{ST} = 0.009). Finally, we applied PCA on unrelated SG10K samples only and found three distinct clusters consisting of Chinese, Malays, and Indians, with a noticeable number of likely admixed individuals forming clines between clusters (Figure S4G).

We further investigated population structure using ADMIXTURE, which models each genome as a mixture of K hypothetical ancestral components (Alexander et al., 2009). When applying ADMIXTURE to the full 1KGP dataset and 300 SG individuals selected randomly, we inferred the optimal number of ancestral components to be K = 10 (Figure S6A). SG Malays contributed a new ancestral component (Figure 3B) that was also present at moderate levels in KHV and CDX. A noticeable level of ancestral component sharing could be observed between SG Chinese and SG Malays and between SG Indians and SG Malays but not between SG Chinese and SG Indians. It is worth noting that this Malay-related ancestral component appeared at low levels in all Han Chinese but was significantly higher in SG Chinese (0.148 ± 0.006, mean ± SD) than in CHS (0.112 ± 0.005, $p < 10^{-5}$, Wilcoxon rank-sum test) and CHB (0.034 ± 0.005, $p < 10^{-16}$), suggesting potential gene flow between Malays and SG Chinese. When we applied ADMIXTURE to SG10K individuals, K = 3 distinguished the three major ethnicities, with some individuals having apparently higher admixture proportions (Figure 3C). However, we inferred the optimal number of components to be K = 6, indicating fine-scale structure within major ethnic groups (Figure S6B).

We next combined SG10K samples with Asian populations in 1KGP to investigate the geographic distribution of inferred ancestral components. With the optimal K = 8, this analysis yielded results resembling those of K = 6 in the analysis of SG10K alone, with two additional components introduced by GIH and Japanese (JPT) (Figure 3D). Together, there were three South Asian components (1, 2, and 3), three East Asian components (4, 5, and 6), and two Southeast Asian components (7 and 8; Figure 3E). Specifically, SG Indians were predominated by components 1 and 2, which reflect a clear south-north cline among Indian populations (Figure 3E). Similarly, SG Chinese had predominantly component 5, which was prevalent in Han Chinese, and component 6, which was mostly found in CDX from southwest China. Interestingly, both Southeast Asian components (7 and 8) were most prevalent in SG Malays. Although component 7 was also present at moderate levels in mainland Southeast Asians (KHV and CDX), component 8 was largely specific to SG Malays, but with a noticeable presence in SG Chinese that was higher than those in any other East Asian populations. In addition, SG Malays had the highest heterozygosity among East and Southeast Asian populations, whereas SG Indians and SG Chinese had higher heterozygosity than other South and East Asian populations from 1KGP, respectively (Figure 3E; Table S4). These observations together suggest multiple ancestral origins of SG Chinese and Indians and their admixture with indigenous people on the Malay Archipelago after recent settlement from China and India.

Demographic and Admixture History

We inferred demographic histories based on 200 individuals from each ethnicity who were free of recent admixture and inbreeding (STAR Methods). Consistent with previous studies (Terhorst et al., 2017; 1000 Genomes Project Consortium et al., 2015), we observed similar dynamics of the effective population size for three ethnicities before 80 kya, including a steep decline from 200 to 80 kya, revealing their shared demographic history as an ancestral Asian population prior to the split of Indians, Malays, and Chinese (Figure 3F). Chinese and Malays

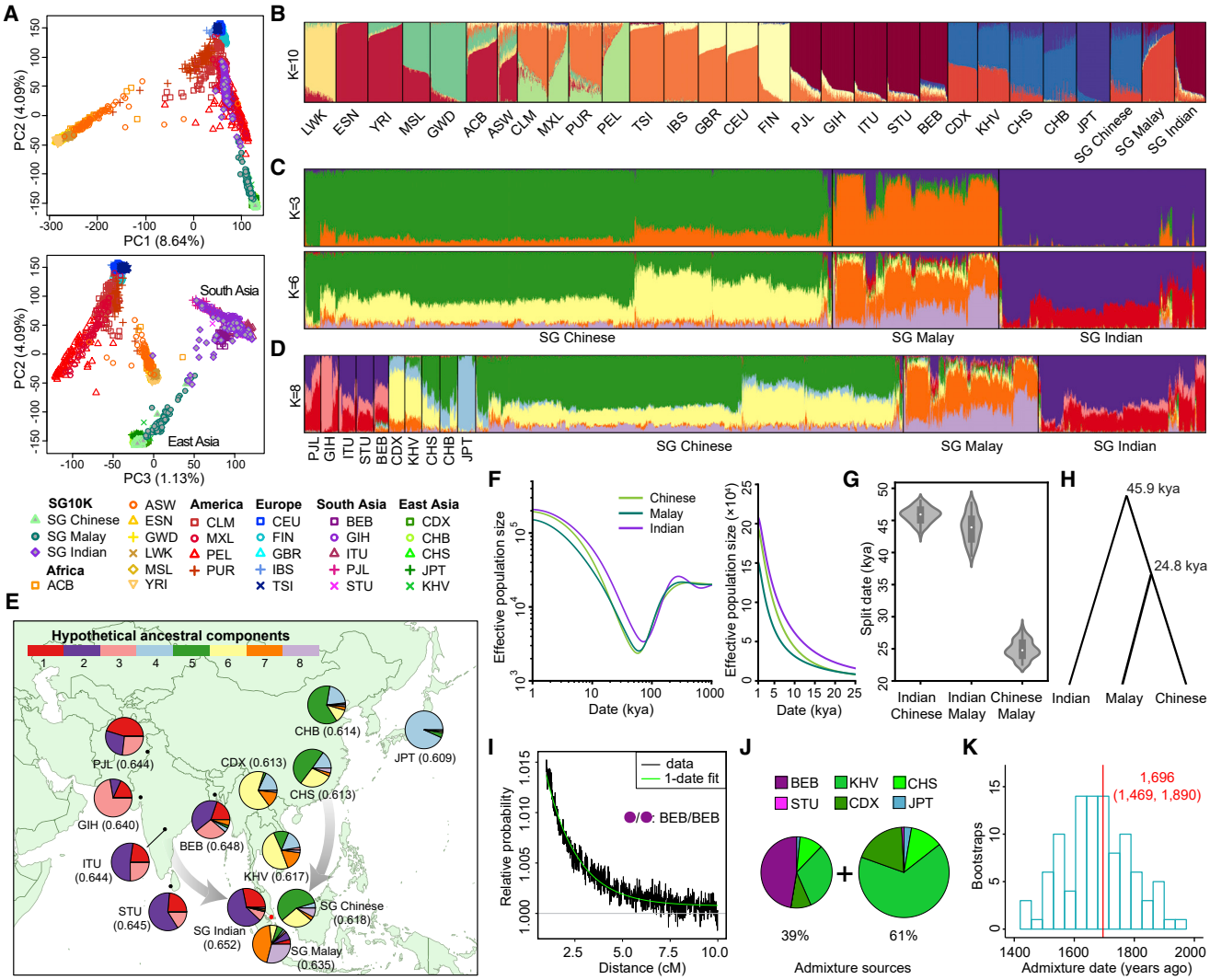


Figure 3. Population Structure and Demographic History of SG Populations

(A) PCA of SG10K and 1KGP samples. The proportion of variance explained by each principal component (PC) is indicated in the axis label.

(B) ADMIXTURE analysis of SG10K and 1KGP samples (K = 10). Each colored bar represents one individual, and colored segments represent proportions of ancestral components.

(C) ADMIXTURE analysis of SG10K individuals (K = 3 and K = 6).

(D) ADMIXTURE analysis of SG10K individuals together with Asians from 1KGP (K = 8).

(E) Geographic distribution of the ancestral components in (D). Each pie chart represents the average ancestral proportions of a population. Haplotype heterozygosity at a 50-kb window size is shown in parentheses.

(F) Dynamics of effective population sizes of SG populations.

(G) Population split dates between SG populations.

(H) Phylogenetic tree based on the estimated population split dates.

(I) Co-ancestry curve for SG Malays, showing the relative probability of finding two segments both copied from BEB at a given genetic distance.

(J) Ancestral composition of the two admixture sources for SG Malays.

(K) Distribution of the estimated admixture dates for SG Malays. The red line indicates the point estimate using the original data.

ACB, African Caribbean; ASW, African American; ESN, Esan; GWD, Gambian; LWK, Luhya; MSL, Mende; YRI, Yoruba; CLM, Colombian; MXL, Mexican; PEL, Peruvian; PUR, Puerto Rican; CEU, northern and western European; FIN, Finnish; GBR, British; IBS, Iberian; TSI, Toscani; BEB, Bengali; GIH, Gujarati; ITU, Telugu; PJL, Punjabi; STU, Sri Lankan Tamil; CDX, Chinese Dai; CHB, Han Chinese in Beijing; CHS, Southern Han Chinese; JPT, Japanese; KHV, Kinh. See also Figures S2, S4, S5, and S6, and Table S4.

experienced a deeper bottleneck that reached its lowest point ~60 kya compared with that of Indians, which reached its lowest point ~80 kya. All three groups experienced rapid population growth after the bottleneck, with Malays having a slower growth rate and, thus, a smaller effective population size in the recent history from ~20 kya to the present. Furthermore, we performed

a joint demographic history analysis to infer the population split date (Terhorst et al., 2017). As shown in Figures 3G and 3H, Indians split from Chinese at ~45.9 kya (95% confidence interval [CI], 43.6–47.9 kya) and from Malays at ~43.9 kya (95% CI, 39.8–47.6 kya), whereas Chinese and Malays split ~24.8 kya (95% CI, 22.6–27.5 kya).

Next we probed the admixture history of Malays, who were severely under-represented in human genetics literature. Although we had selected 200 Malays with the lowest degree of recent admixture, we still detected strong signals of ancient admixture ($p < 0.01$; Hellenthal et al., 2014), as evident by the exponential decay of the co-ancestry curves (Figure 3I). The best-guess admixture model was a one-date-two-party model with 39% contribution from source 1 and 61% contribution from source 2. Source 1 was approximated by a mixture of present-day South Asians (BEB), mainland Southeast Asians (KHV and CDX), and East Asians (CHS), likely representing an ancient Malay population, whereas source 2 was approximated primarily by populations around the South China Sea (KHV, CDX, and CHS; Figure 3J). Both admixture sources were genetically closest to the present-day KHV among all surrogate populations. The admixture date was estimated to be 1.7 kya (95% CI, 1,469–1,890 years ago; Figure 3K). Interestingly, when we randomly selected 100 Malays without excluding recent admixed samples, the best-guess model was reported as multi-date-two-party admixture, indicating a complex admixture history. A two-date-two-party fit suggested an additional admixture event ~233 years ago (95% CI, 150–313), with 11% contribution from a source of both South and East Asians, which was consistent with the recent immigratory history from China and India (Figure S7).

Genomic Regions under Selection

We searched for genomic regions with substantial frequency drift as candidate loci under selection by window-based population branch statistics (PBS) with correction for admixture (STAR Methods; Huerta-Sánchez et al., 2013; Yi et al., 2010). Using a threshold of the top 0.1% maximal PBS values across three populations, we identified 20 independent candidate loci for positive selection, all of which were indexed by SNPs with dramatic allele frequency differences between populations (Table 2; Figure 4A; Data S1). We found seven loci previously hypothesized to be selected in Asians. The top two loci were *EDAR* and *PRSS53*, both related to hair morphology in East Asians (Kamberov et al., 2013; Szpak et al., 2018). Other five well-known loci included *OCA2* for light skin color (Yang et al., 2016), *ALDH2* and *ADH1B* for alcohol metabolism (Li et al., 2007; Oota et al., 2004; Peng et al., 2010), *HYAL2* for cellular response to UV-B irradiation (Ding et al., 2014), and *IL4* for immune response to pathogens (Pillai and Bix, 2011). We also identified 13 loci that have been less studied for selection, among which the strongest signals were *MAGEE2* on chromosome X, *FAM178B* on chromosome 2, and *CENPW* on chromosome 6 (Data S1). Consistent with our findings, two recent studies also suggested selection signals in East Asians at the melanoma-associated gene *MAGEE2* and the gene encoding centromere protein W, *CENPW* (Cheng et al., 2017; Szpak et al., 2018).

Of the 20 loci, 19 were found in either Chinese (10) or Indians (9). There was only one Malay-specific locus (*FAM178B*), whose index SNP (chr2:98058623) was not cataloged in dbSNP (v.150) but reached a high allele frequency in SG Malays (MAF = 0.332 in SG Malays, 0.058 in SG Chinese, and 0.012 in SG Indians). However, because our above PBS analysis was based on an unrooted phylogenetic tree (Figure 4B), if selection occurred in the ancient population prior to the split of Chinese and Malays, then our analysis might incorrectly assign the signal to the outgroup Indian population. We therefore performed additional PBS analyses by introducing northern and western European (CEU), a European population from 1KGP, as an outgroup to help refine the branches under selection (Figure 4C). As expected, the Chinese and Malay loci remained unchanged in the new analyses with CEU. Surprisingly, 8 of the 9 loci initially assigned to the Indian branch were now assigned to the ancestral branch of Chinese and Malays (Figure 4D). These loci include *ALDH2* and *ADH1B*, both of which were well known to be selected for alcohol metabolism in East Asians (Li et al., 2007; Oota et al., 2004; Peng et al., 2010), confirming the validity of our analyses. The only locus with convincing evidence to be selected in Indians is *PRSS53/VKORC1* on chromosome 16. This result, however, was different from a recent study showing that *PRSS53* was selected in East Asians (Szpak et al., 2018).

Last, we performed a systematic analysis to map reported GWAS association to our candidate loci for selection, aiming to understand the direct (target for selection) or indirect (hitchhiking) effect of natural selection on modern human phenotypes. In total, we found that 14 of 20 loci harbored GWAS association signals (Table S5; Data S1), suggesting that selection may have significant direct or indirect contributions to the diversity of many complex traits/diseases across populations. Among 1,260,657 biallelic autosomal SNPs that were at least 2 kb apart from each other, we found 9 of 1,013 GWAS SNPs residing in the selection loci, a chance ~3.5 times of that of non-GWAS SNPs (3,220/1,259,644, $p = 2 \times 10^{-4}$, χ^2 test). Most loci are associated with diverse phenotypes, such as *CENPW* and *ALDH/ATXN2*, where extensive associations with diverse traits and diseases in multiple domains were found. For a few loci, associations were only found with specific domains of traits and diseases, such as *FN1* on chromosome 2, where the associations were mostly cardiovasculature related. Of the 14 loci, the index SNPs for selection and GWAS are in LD ($r^2 > 0.3$) in 7 loci but independent ($r^2 < 0.3$) for the rest. For example, the selected allele rs589278-T (increased frequency in Chinese) within the *CENPW* locus is associated with younger age at menarche; higher risks for androgenetic alopecia, type 1 diabetes, type 2 diabetes, and coronary artery disease; as well as increased body mass index and height. Importantly, we found that the selected allele rs59385041-G (increased frequency in Indians) within the *PRSS53/VKORC1* locus is in strong LD ($r^2 = 0.46$) with the GWAS allele rs749671-G, which has a large effect to lower the drug efficacy of anticoagulants (Parra et al., 2015).

Imputation in Worldwide Populations

We obtained genotyping data for 56 worldwide populations from the Human Genome Diversity Project (HGDP; Li et al., 2008) and the Singapore Genome Diversity Project (SGVP; Teo et al., 2009) and performed imputation using 1KGP and SG10K data as reference panels (STAR Methods). SG10K outperformed 1KGP in all East Asian populations except for Japanese (Figure 5A), which is

Table 2. Candidate Loci for Selection

	Positions	Selection Signal		Index SNP						Candidate Genes	GWAS Traits or Diseases
		PBS_{max}	Branch	ID	Ref/Alt	$AF_{Chinese}$	AF_{Malay}	AF_{Indian}	F_{ST}		
A	chr2:108727043-109625736	22.79	Chinese	rs3827760	A/G	0.925	0.463	0.028	0.748	*EDAR**	morphology
B	chr16:30449906-31155458	16.18	Indian	rs59385041	G/A	0.840	0.640	0.035	0.635	*PRSS53**, *VKORC1**	immunity/ inflammation
C	chrX:73491663-75368879	14.11	Chinese	rs6647668	G/C	0.887	0.565	0.254	0.452	*MAGEE2*	
D	chr2:97477374-98332858	11.40	Malay	chr2:98058623	T/C	0.058	0.332	0.012	0.230	*FAM178B*	psychiatry
E	chr6:126608815-127080700	11.07	Chinese	rs589278	C/T	0.979	0.847	0.595	0.319	*CENPW*	diverse
F	chr2:25945982-26206255	11.03	Chinese	rs78404020	A/G	0.852	0.584	0.120	0.552	*KIF3C*; *ASXL2*	cardiovascular and morphology
G	chr1:172680465-172880359	10.96	Chinese	rs859631	C/T	0.064	0.247	0.567	0.367	*FASLG*	immunity/ inflammation
H	chr12:111994852-113018479	10.66	Indian	rs7962920	A/C	0.872	0.752	0.248	0.477	*ALDH2**, *ATXN2**	diverse
I	chr8:126868840-126961528	10.56	Indian	rs12549315	A/G	0.937	0.816	0.298	0.529	*LINC00861*	
J	chrX:44507058-44562672	10.05	Indian	rs5905847	G/A	0.162	0.294	0.801	0.444		
K	chr2:216258169-216322388	10.04	Chinese	rs1250219	G/C	0.925	0.758	0.210	0.571	*FN1*	cardiovascular and metabolism
L	chrX:133724416-133790275	10.04	Chinese	rs5978021	C/T	0.894	0.707	0.297	0.431	*PLAC1*	
M	chr10:107053687-107568998	9.98	Chinese	rs10786887	G/A	0.953	0.723	0.371	0.441	*LOC101927549*	cancer and morphology
N	chr6:21494076-21632759	9.96	Indian	rs6935474	G/A	0.963	0.840	0.395	0.455	*LINC00581*; *SOX4*	
O	chrX:71841429-72970288	9.73	Chinese	rs201812199	G/A	0.857	0.673	0.194	0.492	*PABPC1L2A/B*	morphology
P	chr17:36192846-36249855	9.72	Indian	rs1963131	A/G	0.161	0.287	0.815	0.485	*YWHAEP7*	
Q	chr15:28179676-28260309	9.39	Chinese	rs4778210	C/T	0.864	0.561	0.130	0.535	*OCA2**	morphology and cancer
R	chr4:99762388-100322106	9.33	Indian	rs1238741	T/C	0.803	0.733	0.122	0.517	*ADH1B**	metabolism and cancer
S	chr3:49523947-50514612	9.16	Indian	rs12494414	C/T	0.663	0.435	0.028	0.435	*HYAL2**, *GNAT1**	diverse
T	chr5:132009154-132128510	9.13	Indian	rs2243289	A/G	0.802	0.653	0.176	0.436	*IL4**	immunity/ inflammation

Candidate genes within each locus are listed with higher priority to genes known for selection (indicated by asterisks), genes reported by the GWAS catalog, and protein-coding genes. See also Table S5 and Data S1.

not surprising given that the Japanese ancestry component was barely found in SG10K samples (Figure 3). Compared with 1KGP, imputation with SG10K reduced the error rates by 50% when imputing SG Malay and SG Chinese and by 10% when imputing SG Indians (Table S6). Beyond Asia, SG10K also improved imputation in Melanesians and Papuans from Oceania, which is likely due to their shared haplotypes with SG Malays. Nevertheless, as expected, SG10K imputation performed worse than 1KGP in other continental groups, especially in African populations (Figure S8A). Surprisingly, SG10K also performed worse in most Central/South Asian populations. We note that Central/South Asian populations in HGDP are mainly from Pakistan, which historically received substantial gene flow from Central Asia and western Eurasia (Majumder, 2010; Qamar et al., 2002). These results reflect the limitation of SG10K in capturing the genetic diversity in Central Asia. For SG populations, we further showed that SG10K substantially improved rare variant imputation, exemplified by a more than 2-fold increase in the number of high-quality imputed rare variants (Rsq > 0.8, MAF < 0.01; Figure 5B). In addition, SG10K imputation had

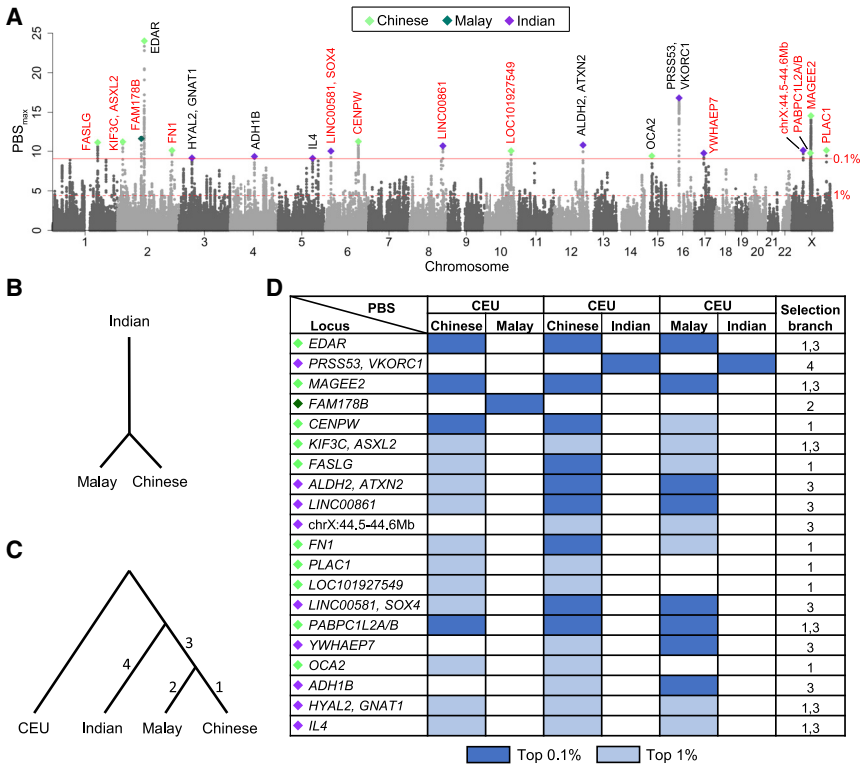


Figure 4. Selection Signals in SG Populations

(A) Manhattan plot of window-based PBS_{max}. Among 20 loci that passed the top 0.1% cutoff, we labeled previously hypothesized candidates in black, whereas the rest are colored in red. The top window of each locus was colored by the population where the maximal PBS was found. The solid and dashed lines indicate top 0.1% and 1% cutoffs, respectively.

(B) Unrooted phylogenetic tree assumed by the PBS analysis in (A).

(C) Rooted phylogenetic tree by introducing CEU as an outgroup.

(D) Examination of the top 20 loci from (A) in three additional PBS analyses, each consisting of CEU and two SG populations. The colored diamond indicates the target population in (A). The heatmap shows whether the top window of a locus was among the top 0.1% or 1% of the normalized PBS values in the corresponding branch. Considering all the PBS results, we identified branches (numbered 1–4) where selection might occur.

See also Table S5 and Data S1.

higher median Rsq than 1KGP imputation across all MAF bins but had fewer high-quality imputed common variants (Rsq > 0.8, MAF > 0.2) because, by sequencing diverse worldwide populations, 1KGP catalogs a larger number of common variants.

To optimize imputation, we sought to create a combined panel using SG10K and 1KGP datasets. Because approximately two-thirds of the variants are unique to either SG10K or 1KGP, a simple intersection would lead to loss of most variants. We instead employed the reciprocal imputation strategy to merge two datasets to their union set of variants (Huang et al., 2015). The combined panel improved imputation over the 1KGP in all Asian and Oceanian populations, except for a minor decrease in Makrani and Balochi, highlighting the potential broad regional effects of SG10K (Figure 5A). The imputation error rate for SG Indians was reduced by 22% compared with use of 1KGP. For SG populations, the combined panel could impute many more high-quality variants (Rsq > 0.8) across all MAF bins than either 1KGP or SG10K alone (Figure 5B) but performed much worse than 1KGP in Africans and Mozabites (Figure S8B) because of the numerous errors introduced by imputing diverse 1KGP populations with the SG10K panel, especially in populations distant from Asians (Figure 5A). Furthermore, we found that, even for sites that were originally genotyped in 1KGP, LD-based imputation with the SG10K panel could mistakenly change the original genotypes of distant samples, such as those from Africa (Figure S8C).

DISCUSSION

We have presented a comprehensive characterization of genetic variation in 4,810 Singaporeans. In contrast to previous WGS studies, which often consist of homogeneous populations, our samples capture the genetic diversity across Asia because of the recent immigratory history of Singapore. The population diversity and large sample size together contributed to the discovery of 52 million novel variants. Although 99.5% of the novel variants are rare (MAF < 0.01), we identified 126 deleterious mutations that are common in Singaporeans but absent from existing public databases. We expect to find many more such variants when we relax the criteria from "absent" to "present at low frequency." Filtering candidate variants by population allele frequencies is an essential step to pinpoint causal mutations in the diagnosis of Mendelian diseases (Whiffin et al., 2017). Our findings reiterate the importance of population-specific reference data for reducing genetic misdiagnoses (Manrai et al., 2016) and support our key objective to set up a reference database for a long-term national precision medicine program in Singapore.

We have gained insights into the population structure and evolutionary history of Asians. We observed a clear north-south clinal pattern of genetic variation in both South Asia and East/Southeast Asia, except for two recent migrant groups in Singapore, reflecting a strong influence of geography on the formation of human population structure (Cavalli-Sforza et al., 1994; Wang et al., 2012). Besides showing that Malays split from Chinese ~24.8 kya and had a slower population growth than Chinese afterward, we inferred a major ancient admixture event in Malays ~1.7 kya, with a source population consisting of present-day KHV, CDX, and CHS. This admixture event is consistent with the Austronesian expansion that originated from Taiwan ~4 kya and reached western island Southeast Asia ~2 kya, as supported by previous studies of genetics, linguistics, and archaeological records (Diamond and Bellwood, 2003; Gray et al., 2009; Lipson et al., 2014; McColl et al.,

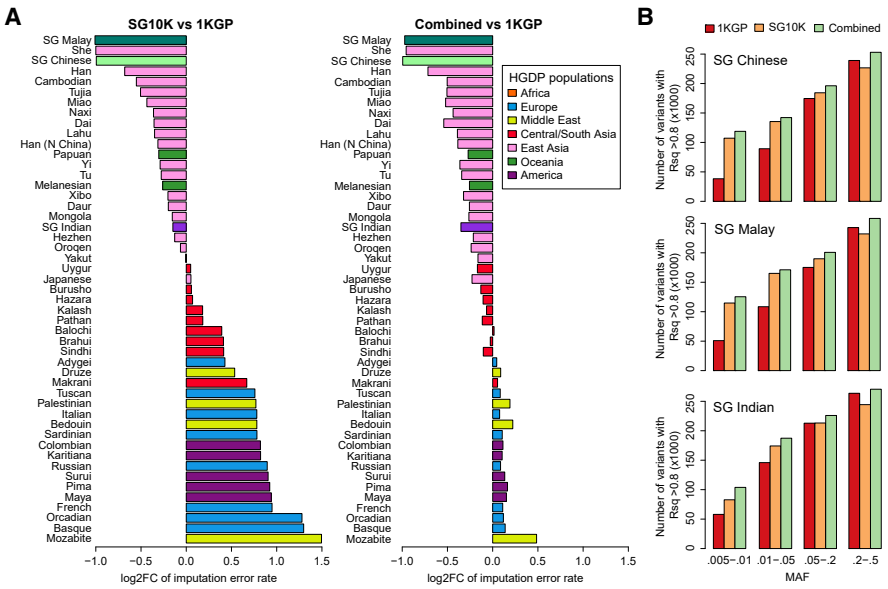


Figure 5. Imputation Accuracy in Worldwide Populations Using Different Reference Panels

(A) Fold change (log_2 scale, log2FC) of imputation error rates for non-African populations using the SG10K and combined panels compared with using the 1KGP panel.

(B) Number of high-quality (Rsq > 0.8) imputed variants in different MAF bins for SG populations.

See also Table S6 and Figure S8.

2018). Considering the admixture in Malays with the Austronesian people, who were more closely related to Han Chinese, the split date of 24.8 kya between Chinese and Malays might represent an underestimate (Terhorst et al., 2017). Although the inferred population composition of ancient Malays is consistent with the hypothesized Austroasiatic farmer expansion from East Asia to mainland Southeast Asia ~4 kya (McColl et al., 2018), we did not detect this admixture event, likely because of the lack of power, because the signal of this earlier admixture might have been masked by the more prominent signal of the subsequent Austronesian admixture event.

In addition, large-scale WGS has enabled us to perform a well-powered genome-wide scan for selection in three ethnicities. We were able to detect many loci reported previously in Asian populations, such as *EDAR* and *OCA2*, even with a stringent criterion. More importantly, we were able to discover additional selection candidates with solid evidence, where all index SNPs showed substantial allele frequency drift. The finding of locus *FAM178B*, which is associated with schizophrenia and bipolar disorder (Amare et al., 2018), is consistent with the hypothesis that schizophrenia might be subjected to positive selection as a maladaptive by-product of adaptation of human cognitive traits (Crespi et al., 2007). Another locus, *CENPW*, is known to play a fundamental role in kinetochore assembly and has been reported by many GWASs, especially for associations with metabolic traits and diseases, suggesting that diet changes such as the domestication of rice in East Asia might be the target for selection. Interestingly, although we had large sample sizes for all three populations, the footprints of selection from Chinese were much more prominent than the ones from Malays and Indians, with 18 of the top 20 loci detected in either Chinese or the ancestral branch of Chinese and Malays but only one locus specific to Malays (*FAM178B*) and another one specific to Indians (*PRSS53/VKORC1*). The large number of loci detected in the ancestral branch of Chinese and Malays might be attributed to the strong gene flow from Austronesian people to Malays, which occurred much more recently than the split of Chinese and Malay ancestors, as suggested by our demographic and admixture inferences. Alternatively, there might be substantial environmental changes after the ancestors of Chinese and Malays split from the ancestors of Indians and migrated to Southeast and East Asia, leading to more opportunities for selection to act.

Of the 20 selection loci, we could find solid association evidence with diverse phenotypes within 14 loci, among which the indexed SNPs for selection and GWAS are in LD for 7 loci. It is worth noting that we only considered LD in SG populations. Because current GWAS findings were largely reported in European populations, and Asian populations are very much underrepresented, future GWASs in Asian populations may discover more genetic variants that are in LD with the selection signals identified in Asians. The significant overlap between the selection and GWAS signals highlights notable direct or indirect effects of natural selection on the diversity of modern human phenotypes. An example of direct influence may be seen at the *FN1* locus, where selection evidence and extensive associations with cardiovascular phenotypes were mapped to the same region around *FN1*. In addition, *FN1* is the only protein-coding gene within the region. These results strongly suggest that *FN1* is likely the direct target for selection. In contrast, many protein-coding genes of functional importance could be found within the locus *PRSS53/VKORC1*, including *PRSS53* and

VKORC1 as well as the tumor suppressor genes *BCL7C* and *PRSS8* (Bao et al., 2019; Kadoch et al., 2013). *VKORC1* is an anticoagulant response gene with a large effect size and extensive geographic differentiation in allele frequencies (Ross et al., 2010; Takeuchi et al., 2009). However, given that humans had only a short period of exposure to anticoagulant medication, *VKORC1* is unlikely to be the direct target for selection. The dramatic frequency difference of *VKORC1*, which has become an important consideration in medical practice, is likely due to a hitchhiking effect of selection within the locus, such as the reported selection on *PRSS53* (Szpak et al., 2018). For many loci, the critical region of selection signal spans over many genes and overlaps with associations of diverse phenotypes, making it challenging to pinpoint specific targets of selection.

Last, because of the genetic diversity of Singapore, our SG10K data have a key advantage over others in improving genotype imputation in Asian and Oceanian populations. Such improvement is evident when using the combined reference panel of our SG10K and the 1KGP data. However, we should be cautious when combining diverse reference panels for better imputation (Huang et al., 2015; McCarthy et al., 2016). In our case, although consisting of worldwide populations from 1KGP, the combined panel was not suitable for imputing samples from Africa, the Middle East, Europe, and America because of errors introduced by the reciprocal imputation approach. Because better imputation can lead to substantial gain in the power of association tests (Huang et al., 2009), it is important to study how to optimally combine published WGS data to generate a mega reference panel that can be applied universally to impute worldwide populations. Coupled with a much better coverage of rare variants and the rapid accumulation of Asian GWAS data, we expect our SG10K data to be a valuable resource to advance genetic studies of heritable traits and complex diseases in Asians and to mitigate the population disparity in current human genetics research.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - DNA extraction and quality check
  - Library preparation and quality check
  - Whole genome sequencing
  - Sequencing quality check
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Quality controls before genotype calling
  - Genotype calling, phasing, and annotation
  - Genetic relatedness, inbreeding coefficient, and consanguinity
  - Evaluation of genotype calling quality
  - PCA, ADMIXTURE, F_{ST}, heterozygosity and nucleotide diversity
  - Demographic inference
  - Scan for selection
  - Imputation experiments
- DATA AND CODE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cell.2019.09.019.

ACKNOWLEDGMENTS

We acknowledge H.M. Kang, S. Das, A. Tan, F. Zhang, J. Terhorst, P.-R. Loh, and G. Hellenthal for helpful discussions and support from all participants and clinical research coordinators of the contributing cohorts and studies: the TTSH Healthy Control Workgroup, the SEED cohort, the Asian Sudden Cardiac Death in Heart Failure Study, the Singapore Heart Failure Outcomes and Phenotypes (SHOP) cohort, the Asian neTwork for Translational Research and Cardiovascular Trials (ATTRaCT), the Parkinson's Disease Study, the Peranakan Genome Study, the Platinum Asian Genomes Project, the Bariatric Surgery Study, the National Heart Centre Singapore Biobank and SingHEART cohorts, and the GUSTO and S-PRESTO study groups. This study was supported by Singapore's A*STAR (core funding and IAF-PP H17/01/a0/007), BMRC (SPF2014/001, SPF2013/002, SPF2014/003, SPF2014/004, and SPF2014/005), NMRC (CIRG/1371/2013, CIRG/1417/2015, CIRG/1488/2018, CSA-SI/0012/2017, CG/017/2013, CG/M006/2017_NHCS, TCR/013-NNI/2014, STaR/0011/2012, STaR2013/001, STaR/014/2013, STaR/0026/2015, TCR/006-NUHS/2013, TCR/012-NUHS/2014, TCR/004-NUS/2008, TCR/012-NUHS/2014, and center grants 2010-13 and 2013-2017), NRF (NRFF2016-03), National University of Singapore, SingHealth and Duke-NUS, and Alexandra Health small innovative grant SIGII/15203 and funding from Huazhong University of Science and Technology, the Tanoto Foundation, the Lee Foundation, the Boston Scientific Investigator Sponsored Research Program and Bayer, the NSF (DEB-1753489), and the Alfred P. Sloan Foundation. The computation was partially performed on resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg).

AUTHOR CONTRIBUTIONS

J.L., C.W., and H.H.N. conceptualized and supervised the project. D.W., J.D., X.C., S.C., L.B., M.D., and C.W. conducted analysis. D.W., J.D., X.C., C.B., A.W., C.C.S., W.W.J.S., N.B., C.C.K., C.W., and J.L. contributed to data generation and curation. N.K., S.D., P.T., A.S., A.M., E.-K.T., J.N.F., L.L.G., K.P.L., R.S.Y.F., C.S.P.L., A.M.R., C.-Y.C., T.A., and T.Y.W. contributed samples. C.W. and J.L. drafted the manuscript, and C.W., J.L., M.D., C.B.L., and all the primary authors reviewed, edited, and approved manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: January 5, 2019
Revised: June 24, 2019
Accepted: September 19, 2019
Published: October 17, 2019

REFERENCES

1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Abdulla, M.A., Ahmed, I., Assawamakin, A., Bhak, J., Brahmachari, S.K., Calacal, G.C., Chaurasia, A., Chen, C.H., Chen, J., Chen, Y.T., et al.; HUGO Pan-Asian SNP Consortium; Indian Genome Variation Consortium (2009). Mapping human genetic diversity in Asia. Science *326*, 1541–1545.

Adzhubei, I., Jordan, D.M., and Sunyaev, S.R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. Curr. Protoc. Hum. Genet. *76*, Chapter 7, Unit 7.20.

Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.

Amare, A.T., Schubert, K.O., Hou, L., Clark, S.R., Papiol, S., Heilbronner, U., Degenhardt, F., Tekola-Ayele, F., Hsu, Y.H., Shekhtman, T., et al.; International Consortium on Lithium Genetics (ConLi+Gen) (2018). Association of polygenic score for schizophrenia and HLA antigen and inflammation genes with response to lithium in bipolar affective disorder: a genome-wide association study. JAMA Psychiatry *75*, 65–74.

Ashley, E.A. (2016). Towards precision medicine. Nat. Rev. Genet. *17*, 507–522.

Auton, A., Bryc, K., Boyko, A.R., Lohmueller, K.E., Novembre, J., Reynolds, A., Indap, A., Wright, M.H., Degenhardt, J.D., Gutenkunst, R.N., et al. (2009). Global distribution of genomic diversity underscores rich complex history of continental human populations. Genome Res. *19*, 795–803.

Bai, H., Guo, X., Narisu, N., Lan, T., Wu, Q., Xing, Y., Zhang, Y., Bond, S.R., Pei, Z., Zhang, Y., et al. (2018). Whole-genome sequencing of 175 Mongolians uncovers population-specific genetic architecture and gene flow throughout North and East Asia. Nat. Genet. *50*, 1696–1704.

Bao, Y., Guo, Y., Yang, Y., Wei, X., Zhang, S., Zhang, Y., Li, K., Yuan, M., Guo, D., Macias, V., et al. (2019). PRSS8 suppresses colorectal carcinogenesis and metastasis. Oncogene *38*, 497–517.

Bhatia, G., Patterson, N., Sankararaman, S., and Price, A.L. (2013). Estimating and interpreting F_{ST}: the impact of rare variants. Genome Res. *23*, 1514–1521.

Bittles, A.H., and Black, M.L. (2010). Evolution in health and medicine Sackler colloquium: Consanguinity, human evolution, and complex diseases. Proc. Natl. Acad. Sci. USA *107* (*Suppl 1*), 1779–1786.

Browning, B.L., and Yu, Z. (2009). Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. Am. J. Hum. Genet. *85*, 847–861.

Cavalli-Sforza, L.L., Menozzi, P., and Piazza, A. (1994). The History and Geography of Human Genes (Princeton, NJ: Princeton University Press).

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7.

Chatterjee, N., Shi, J., and García-Closas, M. (2016). Developing and evaluating polygenic risk prediction models for stratified disease prevention. Nat. Rev. Genet. *17*, 392–406.

Cheng, X., Xu, C., and DeGiorgio, M. (2017). Fast and robust detection of ancestral selective sweeps. Mol. Ecol. *26*, 6871–6891.

Chiang, C.W.K., Mangul, S., Robles, C., and Sankararaman, S. (2018). A comprehensive map of genetic variation in the world's largest ethnic group-Han Chinese. Mol. Biol. Evol. *35*, 2736–2750.

Conomos, M.P., Reiner, A.P., Weir, B.S., and Thornton, T.A. (2016). Model-free estimation of recent genetic relatedness. Am. J. Hum. Genet. *98*, 127–148.

Cornes, B.K., Khor, C.C., Nongpiur, M.E., Xu, L., Tay, W.T., Zheng, Y., Lavanya, R., Li, Y., Wu, R., Sim, X., et al. (2012). Identification of four novel variants that influence central corneal thickness in multi-ethnic Asian populations. Hum. Mol. Genet. *21*, 437–445.

Crespi, B., Summers, K., and Dorus, S. (2007). Adaptive evolution of genes underlying schizophrenia. Proc. Biol. Sci. *274*, 2801–2810.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al.; 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. Nat. Genet. *48*, 1284–1287.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

Diamond, J., and Bellwood, P. (2003). Farmers and their languages: the first expansions. Science *300*, 597–603.

Ding, Q., Hu, Y., Xu, S., Wang, J., and Jin, L. (2014). Neanderthal introgression at chromosome 3p21.31 was under positive natural selection in East Asians. Mol. Biol. Evol. *31*, 683–695.

Dou, J., Sun, B., Sim, X., Hughes, J.D., Reilly, D.F., Tai, E.S., Liu, J., and Wang, C. (2017). Estimation of kinship coefficient in structured and admixed populations using sparse sequencing data. PLoS Genet. *13*, e1007021.

Fan, S., Hansen, M.E., Lo, Y., and Tishkoff, S.A. (2016). Going global by adapting local: A review of recent human adaptation. Science *354*, 54–59.

Faust, G.G., and Hall, I.M. (2014). SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics *30*, 2503–2505.

Genome of the Netherlands Consortium (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat. Genet. *46*, 818–825.

Gray, R.D., Drummond, A.J., and Greenhill, S.J. (2009). Language phylogenies reveal expansion pulses and pauses in Pacific settlement. Science *323*, 479–483.

Gudbjartsson, D.F., Helgason, H., Gudjonsson, S.A., Zink, F., Oddson, A., Gylfason, A., Besenbacher, S., Magnusson, G., Halldorsson, B.V., Hjartarson, E., et al. (2015). Large-scale whole-genome sequencing of the Icelandic population. Nat. Genet. *47*, 435–444.

Han, E., Sinsheimer, J.S., and Novembre, J. (2014). Characterizing bias in population genetic inferences from low-coverage sequencing data. Mol. Biol. Evol. *31*, 723–735.

Hellenthal, G., Busby, G.B.J., Band, G., Wilson, J.F., Capelli, C., Falush, D., and Myers, S. (2014). A genetic atlas of human admixture history. Science *343*, 747–751.

Hindorff, L.A., Bonham, V.L., Brody, L.C., Ginoza, M.E.C., Hutter, C.M., Manolio, T.A., and Green, E.D. (2018). Prioritizing diversity in human genomics research. Nat. Rev. Genet. *19*, 175–185.

Huang, L., Wang, C., and Rosenberg, N.A. (2009). The relationship between imputation error and statistical power in genetic association studies in diverse populations. Am. J. Hum. Genet. *85*, 692–698.

Huang, J., Howie, B., McCarthy, S., Memari, Y., Walter, K., Min, J.L., Danecek, P., Malerba, G., Trabetti, E., Zheng, H.F., et al.; UK10K Consortium (2015). Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. Nat. Commun. *6*, 8111.

Huerta-Sánchez, E., Degiorgio, M., Pagani, L., Tarekegn, A., Ekong, R., Antao, T., Cardona, A., Montgomery, H.E., Cavalleri, G.L., Robbins, P.A., et al. (2013). Genetic signatures reveal high-altitude adaptation in a set of ethiopian populations. Mol. Biol. Evol. *30*, 1877–1888.

Jakobsson, M., Scholz, S.W., Scheet, P., Gibbs, J.R., VanLiere, J.M., Fung, H.C., Szpiech, Z.A., Degnan, J.H., Wang, K., Guerreiro, R., et al. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. Nature *451*, 998–1003.

Jun, G., Flickinger, M., Hetrick, K.N., Romm, J.M., Doheny, K.F., Abecasis, G.R., Boehnke, M., and Kang, H.M. (2012). Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am. J. Hum. Genet. *91*, 839–848.

Jun, G., Wing, M.K., Abecasis, G.R., and Kang, H.M. (2015). An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. Genome Res. *25*, 918–925.

Kadoch, C., Hargreaves, D.C., Hodges, C., Elias, L., Ho, L., Ranish, J., and Crabtree, G.R. (2013). Proteomic and bioinformatic analysis of mammalian SWI/SNF complexes identifies extensive roles in human malignancy. Nat. Genet. *45*, 592–601.

Kamberov, Y.G., Wang, S., Tan, J., Gerbault, P., Wark, A., Tan, L., Yang, Y., Li, S., Tang, K., Chen, H., et al. (2013). Modeling recent human evolution in mice by expression of a selected EDAR variant. Cell *152*, 691–702.

Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. *4*, 1073–1081.

Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. *42*, D980–D985.

Lappalainen, I., Almeida-King, J., Kumanduri, V., Senf, A., Spalding, J.D., Ur-Rehman, S., Saunders, G., Kandasamy, J., Caccamo, M., Leinonen, R., et al. (2015). The European Genome-phenome Archive of human data consented for biomedical research. Nat. Genet. *47*, 692–695.

Lawson, D.J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of population structure using dense haplotype data. PLoS Genet. *8*, e1002453.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv, arXiv:1303.3997. https://arxiv.org/abs/1303.3997.

Li, H., Mukherjee, N., Soundararajan, U., Tarnok, Z., Barta, C., Khaliq, S., Mohyuddin, A., Kajuna, S.L., Mehdi, S.Q., Kidd, J.R., and Kidd, K.K. (2007). Geographically separate increases in the frequency of the derived ADH1B*47His allele in eastern and western Asia. Am. J. Hum. Genet. *81*, 842–846.

Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., and Myers, R.M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. Science *319*, 1100–1104.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, Y., Willer, C.J., Ding, J., Scheet, P., and Abecasis, G.R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol. *34*, 816–834.

Li, Y., Sidore, C., Kang, H.M., Boehnke, M., and Abecasis, G.R. (2011). Low-coverage sequencing: implications for design of complex trait association studies. Genome Res. *21*, 940–951.

Linderman, M.D., Brandt, T., Edelmann, L., Jabado, O., Kasai, Y., Kornreich, R., Mahajan, M., Shah, H., Kasarskis, A., and Schadt, E.E. (2014). Analytical validation of whole exome and whole genome sequencing for clinical applications. BMC Med. Genomics *7*, 20.

Lipson, M., Loh, P.R., Patterson, N., Moorjani, P., Ko, Y.C., Stoneking, M., Berger, B., and Reich, D. (2014). Reconstructing Austronesian population history in Island Southeast Asia. Nat. Commun. *5*, 4689.

Liu, S., Huang, S., Chen, F., Zhao, L., Yuan, Y., Francis, S.S., Fang, L., Li, Z., Lin, L., Liu, R., et al. (2018). Genomic analyses from non-invasive prenatal testing reveal genetic associations, patterns of viral infections, and Chinese population history. Cell *175*, 347–359.e14.

Loh, P.R., Danecek, P., Palamara, P.F., Fuchsberger, C., A Reshef, Y., K Finucane, H., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G.R., et al. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. Nat. Genet. *48*, 1443–1448.

MacArthur, D.G., Manolio, T.A., Dimmock, D.P., Rehm, H.L., Shendure, J., Abecasis, G.R., Adams, D.R., Altman, R.B., Antonarakis, S.E., Ashley, E.A., et al. (2014). Guidelines for investigating causality of sequence variants in human disease. Nature *508*, 469–476.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. *45* (D1), D896–D901.

Majumder, P.P. (2010). The human genetic history of South Asia. Curr. Biol. *20*, R184–R187.

Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873.

Manrai, A.K., Funke, B.H., Rehm, H.L., Olesen, M.S., Maron, B.A., Szolovits, P., Margulies, D.M., Loscalzo, J., and Kohane, I.S. (2016). Genetic misdiagnoses and the potential for health disparities. N. Engl. J. Med. *375*, 655–665.

Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M., and Daly, M.J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. Nat. Genet. *51*, 584–591.

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al.; Haplotype Reference Consortium (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nat. Genet. *48*, 1279–1283.

McColl, H., Racimo, F., Vinner, L., Demeter, F., Gakuhari, T., Moreno-Mayar, J.V., van Driem, G., Gram Wilken, U., Seguin-Orlando, A., de la Fuente Castro, C., et al. (2018). The prehistoric peopling of Southeast Asia. Science *361*, 88–92.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122.

McVean, G. (2009). A genealogical interpretation of principal components analysis. PLoS Genet. *5*, e1000686.

Nei, M., and Li, W.H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA *76*, 5269–5273.

Nelson, M.R., Tipney, H., Painter, J.L., Shen, J., Nicoletti, P., Shen, Y., Floratos, A., Sham, P.C., Li, M.J., Wang, J., et al. (2015). The support of human genetic evidence for approved drug indications. Nat. Genet. *47*, 856–860.

Nielsen, R., Akey, J.M., Jakobsson, M., Pritchard, J.K., Tishkoff, S., and Willerslev, E. (2017). Tracing the peopling of the world through genomics. Nature *541*, 302–310.

Oota, H., Pakstis, A.J., Bonne-Tamir, B., Goldman, D., Grigorenko, E., Kajuna, S.L., Karoma, N.J., Kungulilo, S., Lu, R.B., Odunsi, K., et al. (2004). The evolution and population genetics of the ALDH2 locus: random genetic drift, selection, and low levels of recombination. Ann. Hum. Genet. *68*, 93–109.

Parra, E.J., Botton, M.R., Perini, J.A., Krithika, S., Bourgeois, S., Johnson, T.A., Tsunoda, T., Pirmohamed, M., Wadelius, M., Limdi, N.A., et al. (2015). Genome-wide association study of warfarin maintenance dose in a Brazilian sample. Pharmacogenomics *16*, 1253–1263.

Peng, Y., Shi, H., Qi, X.B., Xiao, C.J., Zhong, H., Ma, R.L., and Su, B. (2010). The ADH1B Arg47His polymorphism in east Asian populations and expansion of rice domestication in history. BMC Evol. Biol. *10*, 15.

Pillai, M.R., and Bix, M. (2011). Evolution of IL4 and pathogen antagonism. Growth Factors *29*, 153–160.

Price, A.L., Weale, M.E., Patterson, N., Myers, S.R., Need, A.C., Shianna, K.V., Ge, D., Rotter, J.I., Torres, E., Taylor, K.D., et al. (2008). Long-range LD can confound genome scans in admixed populations. Am. J. Hum. Genet. *83*, 132–135, author reply 135–139.

Pruim, R.J., Welch, R.P., Sanna, S., Teslovich, T.M., Chines, P.S., Gliedt, T.P., Boehnke, M., Abecasis, G.R., and Willer, C.J. (2010). LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics *26*, 2336–2337.

Qamar, R., Ayub, Q., Mohyuddin, A., Helgason, A., Mazhar, K., Mansoor, A., Zerjal, T., Tyler-Smith, C., and Mehdi, S.Q. (2002). Y-chromosomal DNA variation in Pakistan. Am. J. Hum. Genet. *70*, 1107–1124.

Rashkin, S., Jun, G., Chen, S., and Abecasis, G.R.; Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO) (2017). Optimal sequencing strategies for identifying disease-associated singletons. PLoS Genet. *13*, e1006811.

Rehm, H.L., Berg, J.S., Brooks, L.D., Bustamante, C.D., Evans, J.P., Landrum, M.J., Ledbetter, D.H., Maglott, D.R., Martin, C.L., Nussbaum, R.L., et al.; ClinGen (2015). ClinGen–the clinical genome resource. N. Engl. J. Med. *372*, 2235–2242.

Relling, M.V., and Evans, W.E. (2015). Pharmacogenomics in the clinic. Nature *526*, 343–350.

Ross, K.A., Bigham, A.W., Edwards, M., Gozdzik, A., Suarez-Kurtz, G., and Parra, E.J. (2010). Worldwide allele frequency distribution of four

polymorphisms associated with warfarin dose requirements. J. Hum. Genet. *55*, 582–589.

Sim, N.L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., and Ng, P.C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res. *40*, W452-7.

Staples, J., Qiao, D., Cho, M.H., Silverman, E.K., Nickerson, D.A., and Below, J.E.; University of Washington Center for Mendelian Genomics (2014). PRIMUS: rapid reconstruction of pedigrees from genome-wide estimates of identity by descent. Am. J. Hum. Genet. *95*, 553–564.

Szpak, M., Mezzavilla, M., Ayub, Q., Chen, Y., Xue, Y., and Tyler-Smith, C. (2018). FineMAV: prioritizing candidate genetic variants driving local adaptations in human populations. Genome Biol. *19*, 5.

Takeuchi, F., McGinnis, R., Bourgeois, S., Barnes, C., Eriksson, N., Soranzo, N., Whittaker, P., Ranganath, V., Kumanduri, V., McLaren, W., et al. (2009). A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. PLoS Genet. *5*, e1000433.

Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G., Corvelo, A., Gogarten, S.M., Kang, H.M., et al. (2019). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. bioRxiv. https://doi.org/10.1101/563866.

Tan, A., Abecasis, G.R., and Kang, H.M. (2015). Unified representation of genetic variants. Bioinformatics *31*, 2202–2204.

Teo, Y.Y., Sim, X., Ong, R.T., Tan, A.K., Chen, J., Tantoso, E., Small, K.S., Ku, C.S., Lee, E.J., Seielstad, M., and Chia, K.S. (2009). Singapore Genome Variation Project: a haplotype map of three Southeast Asian populations. Genome Res. *19*, 2154–2162.

Terhorst, J., Kamm, J.A., and Song, Y.S. (2017). Robust and scalable inference of population history from hundreds of unphased whole genomes. Nat. Genet. *49*, 303–309.

UK10K Consortium, Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R., Xu, C., Futema, M., et al. (2015). The UK10K project identifies rare variants in health and disease. Nature *526*, 82–90.

Timpson, N.J., Greenwood, C.M.T., Soranzo, N., Lawson, D.J., and Richards, J.B. (2018). Genetic architecture: the shape of the genetic contribution to human traits and disease. Nat. Rev. Genet. *19*, 110–124.

Wang, C., Zöllner, S., and Rosenberg, N.A. (2012). A quantitative comparison of the similarity between genes and geography in worldwide human populations. PLoS Genet. *8*, e1002886.

Wang, C., Zhan, X., Bragg-Gresham, J., Kang, H.M., Stambolian, D., Chew, E.Y., Branham, K.E., Heckenlively, J., Fulton, R., Wilson, R.K., et al.; FUSION Study (2014). Ancestry estimation and control of population stratification for sequence-based association studies. Nat. Genet. *46*, 409–415.

Wang, C., Zhan, X., Liang, L., Abecasis, G.R., and Lin, X. (2015). Improved ancestry estimation for both genotyping and sequencing data using projection procrustes analysis and genotype imputation. Am. J. Hum. Genet. *96*, 926–937.

Weir, B.S., and Cockerham, C.C. (1984). Estimating F-statistics for the analysis of population structure. Evolution *38*, 1358–1370.

Whiffin, N., Minikel, E., Walsh, R., O'Donnell-Luria, A.H., Karczewski, K., Ing, A.Y., Barton, P.J.R., Funke, B., Cook, S.A., MacArthur, D., and Ware, J.S. (2017). Using high-resolution variant frequencies to empower clinical genome interpretation. Genet. Med. *19*, 1151–1158.

Wong, L.P., Ong, R.T., Poh, W.T., Liu, X., Chen, P., Li, R., Lam, K.K., Pillai, N.E., Sim, K.S., Xu, H., et al. (2013). Deep whole-genome sequencing of 100 southeast Asian Malays. Am. J. Hum. Genet. *92*, 52–66.

Wong, L.P., Lai, J.K., Saw, W.Y., Ong, R.T., Cheng, A.Y., Pillai, N.E., Liu, X., Xu, W., Chen, P., Foo, J.N., et al. (2014). Insights into the genetic structure and diversity of 38 South Asian Indians from deep whole-genome sequencing. PLoS Genet. *10*, e1004377.

Yang, Z., Zhong, H., Chen, J., Zhang, X., Zhang, H., Luo, X., Xu, S., Chen, H., Lu, D., Han, Y., et al. (2016). A genetic mechanism for convergent skin lightening during recent human evolution. Mol. Biol. Evol. *33*, 1177–1187.

Yang, S., Lincoln, S.E., Kobayashi, Y., Nykamp, K., Nussbaum, R.L., and Topper, S. (2017). Sources of discordance among germ-line variant classifications in ClinVar. Genet. Med. *19*, 1118–1126.

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z.X., Pool, J.E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T.S., et al. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. Science *329*, 75–78.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Critical Commercial Assays		
QIAamp DNA Blood Midi Kit (100)	QIAGEN	Cat#51185
abGenix Whole Blood Genomic DNA Extraction Kit (discontinued)	AITbiotech	Cat#800815
Qubit dsDNA HS Assay Kit	Life Technologies	Cat#Q32854
Quantifluor dsDNA System	Promega	Cat#E2670
SeaKem LE Agarose	Lonza Biologics	Cat#50004
GelRed Nucleic Acid Gel Stain 10,000X in water	Biotium	Cat#41003
1kb DNA Ladder	New England Biolabs	Cat#N3232L
Truseq Nano DNA High Throughput Library Prep Kit (96 samples)	Illumina	Cat#20015965
Truseq DNA PCR-Free High Throughput Library Prep Kit (96 samples)	Illumina	Cat#20015963
TruSeq DNA CD Indexes (96 Indexes, 96 Samples)	Illumina	Cat#20015949
NEBNext Ultra II DNA Library Prep Kit for Illumina	New England Biolabs	Cat#E7645L
NEBNext Multiplex Oligos for Illumina (96 Index Primers)	New England Biolabs	Cat#E6609S
Agencourt AMPure XP	Beckman Coulter	Cat#A63882
LabChip GX	Perkin-Elmer	Cat#CLS960013
D1000 ScreenTape	Agilent Technologies	Cat#5067-5582
D1000 Reagents	Agilent Technologies	Cat#5067-5583
Complete Kit (Universal)	Kapa Biosystems	Cat#KK4824
HiSeq 3000/4000 PE Cluster Kit	Illumina	Cat#PE-410-1001
HiSeq 3000/4000 SBS Kit (300 cycles)	Illumina	Cat#FC-410-1003
HiSeq X Ten Reagent Kit v2.5	Illumina	Cat#FC-501-2501
Deposited Data		
Human genome reference (GRCh37)	The 1000 Genomes Project	http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/
1000 Genomes Project (Phase 3, v5a)	The 1000 Genomes Project	https://www.internationalgenome.org/
Human Genome Diversity Project	Li et al., 2008	http://www.hagsc.org/hgdp/
Singapore Genome Variation Project	Teo et al., 2009	https://blog.nus.edu.sg/sshsphphg/singapore-genome-variation/
GWAS Catalog (v1.0.2-associations_e93_r2018-08-15)	MacArthur et al., 2017	https://www.ebi.ac.uk/gwas/
VEP-compiled annotation database (v 91_GRCh37)	McLaren et al., 2016	https://useast.ensembl.org/info/docs/tools/vep/script/vep_cache.html
ClinVar annotation (201706)	Landrum et al., 2014	https://www.ncbi.nlm.nih.gov/clinvar/
SG10K project (Pilot)	This study	EGA: EGAS00001003875
Software and Algorithms		
BWA-MEM (v 0.7.12)	Li, 2013	https://github.com/lh3/bwa
samblaster (v 0.1.22)	Faust and Hall, 2014	https://github.com/GregoryFaust/samblaster
SAMtools (v 1.3)	Li et al., 2009	https://github.com/samtools/samtools
BamUtil (v 1.0.14)	Jun et al., 2015	https://genome.sph.umich.edu/wiki/BamUtil
LASER (v 2.04)	Wang et al., 2015	http://csg.sph.umich.edu/chaolong/LASER/
VerifyBamID (v 1.1.2)	Jun et al., 2012	https://genome.sph.umich.edu/wiki/VerifyBamID

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
GotClound pipeline (TopMed freeze 3a version)	Jun et al., 2015	https://github.com/statgen/topmed_freeze3_calling
BEAGLE (v 4.1)	Browning and Yu, 2009	https://faculty.washington.edu/browning/beagle/beagle.html
PLINK (v 1.9)	Chang et al., 2015	https://www.cog-genomics.org/plink2/
EAGLE (v 2.3.5)	Loh et al., 2016	https://data.broadinstitute.org/alkesgroup/Eagle/
PC-Relate (v 2.2.2)	Conomos et al., 2016	https://www.rdocumentation.org/packages/GENESIS/versions/2.2.2/topics/pcrelate
SEEKIN (v 1.01)	Dou et al., 2017	https://github.com/chaolongwang/SEEKIN
PRIMUS (v 1.90)	Staples et al., 2014	https://primus.gs.washington.edu/primusweb/
ADMIXTURE (v 1.3)	Alexander et al., 2009	http://www.genetics.ucla.edu/software/admixture/
SMC++ (v 1.15.2)	Terhorst et al., 2017	https://github.com/popgenmethods/smcpp
GLOBETROTTER (Nov 8, 2017)	Hellenthal et al., 2014	https://people.maths.bris.ac.uk/~madjl/finestructure/globetrotter.html
ChromoPainter (v 2)	Lawson et al., 2012	https://people.maths.bris.ac.uk/~madjl/finestructure/globetrotter.html
LocusZoom (v 1.4)	Pruim et al., 2010	http://locuszoom.org/
BCFtools (v 1.9)	Li et al., 2009	https://samtools.github.io/bcftools/bcftools.html
Minimac4 (v 1.0.0)	Das et al., 2016	https://github.com/statgen/Minimac4
SIFT web server (v 5.2.2)	Sim et al., 2012	http://sift.bii.a-star.edu.sg/
GATK (v 4.0)	DePristo et al., 2011	https://software.broadinstitute.org/gatk/gatk4

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information may be directed to the Lead Contact, Chaolong Wang (chaolong@hust.edu.cn).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

We collected whole blood DNA samples of 5,424 individuals from nine cohorts in Singapore, including 3303 healthy individuals (no major diseases) contributed by the Singapore Epidemiology of Eye Diseases cohort (SEED, n = 1,536), Tan Tock Seng Hospital (TTSH, n = 971), the Growing Up in Singapore Toward healthy Outcomes study (GUSTO, n = 499), the Singapore's PREconception study of long-Term maternal and child Outcomes (S-PRESTO, n = 72), SingHealth Duke-NUS Institute of Precision Medicine (PRISM, n = 100), the Peranakan Genomes Project (Peranakan, n = 79), and the Platinum Asian Genomes Project (Platinum, n = 46, consisting of 14 trios and 1 quartet). The remaining 2,121 individuals are patients from three disease studies: heart failure (HF, n = 1540), Parkinson's disease (PD, n = 355), and obese patients who had undergone bariatric surgery (Bariatric, n = 226). These studies were approved by the Institutional Review Board of the National University of Singapore (Approvals: N-17-030E and H-17-049), SingHealth Centralized Institutional Review Board (Approvals: 2006/612/A, 2014/160/A, 2018/2570, 2018/2717, 2010/196/C, 2015/2194, 2019/2046, 2009/280/D, 2014/692/D, 2002/008/A, 2002/008 g/A, 2013/605/C, 2018/3081 and 2015/2308), National Health Group Domain Specific Review Board (Approvals: 2007/00167, 2016/00269, 2018/00301, TTSH/2014-00040 and 2009/00021). All participants provided written informed consent. However, 571 samples from the GUSTO/S-PRESTO birth cohort were used only during joint genotype calling. Demographic information of samples from the other eight cohorts was summarized (Table S1).

METHOD DETAILS

DNA extraction and quality check

Genomic DNA was either extracted at the Genome Institute of Singapore using QIAamp DNA Blood Midi Kit (QIAGEN) or abGenix Whole Blood Genomic DNA Extraction Kit (AITbiotech) or extracted by each contributing study prior to delivery to GIS. DNA quantification was performed using Qubit dsDNA HS Assay Kit (Life Technologies) or Quantifluor dsDNA System (Promega). In order to interrogate DNA quality, DNA samples were processed with Qubit dsDNA HS Standard #2 DNA from the Qubit dsDNA HS Assay Kit (Life Technologies) and 1kb DNA ladder (New England Biolabs) on a 1% GelRed (Biotium) stained SeaKem LE (Lonza Biologics) agarose gel. The electrophoresis conditions were at 100-120 V for 60 minutes.

Library preparation and quality check

Library preparation was undertaken as per protocol using either the Illumina TruSeq Nano DNA High Throughput Library Prep Kit, TruSeq DNA PCR-Free High Throughput Library Prep Kit with TruSeq DNA CD Indexes (Illumina) or NEBNext Ultra II DNA Library

Prep Kit for Illumina with NEBNext Multiplex Oligos for Illumina (New England Biolabs). All library clean-up steps were performed using Agencourt AMPure XP (Beckman Coulter). Libraries were quantified and checked using LabChip GX (Perkin-Elmer) or D1000 ScreenTape and Reagents (Agilent Technologies). Quantitative PCR (qPCR) was carried out using the Complete Kit (Kapa Biosystems) before sequencing.

Whole genome sequencing

Paired-end 151bp whole-genome sequencing with an insert size of 350bp was performed on the Illumina HiSeq 4000 platform with HiSeq 3000/4000 PE Cluster Kit and HiSeq 3000/4000 SBS Kit (300 cycles) or HiSeq X platform with HiSeq X Ten Reagent Kit v2.5 (Illumina).

Sequencing quality check

The target depth was 15× for all samples except for 571 samples from the GUSTO/S-PRESTO birth cohort, which were sequenced at 30×. Four samples failed to be sequenced. For the remaining samples, we aligned read pairs to human reference genome GRCh37 using BWA-MEM (v 0.7.12; -M) (Li, 2013). PCR duplicates were removed with samblaster (v 0.1.22) (Faust and Hall, 2014). We sorted and merged aligned read pairs from different sequencing lanes using SAMtools (v 1.3) (Li et al., 2009), followed by base quality recalibration using BamUtil recab (v 1.0.13; --maxBaseQual 40). Finally, BAM files were converted to CRAM files with 8 bins of base quality by BamUtil (v 1.0.14; --binQualS 0:2,3:3,4:4,5:5,6:6,7:10,13:20,23:30). After excluding unmapped reads and PCR duplicates, the mean sequencing depth is 13.7×across 4,849 samples targeted at 15×.

QUANTIFICATION AND STATISTICAL ANALYSIS

Quality controls before genotype calling

We inferred ancestry, sex, and contamination rate for each sample based on sequencing reads. Ancestry estimation was based on the LASER software (Wang et al., 2014, 2015), with reference panels from HGDP (Li et al., 2008) and SGVP (Teo et al., 2009). The HGDP dataset consists of genotypes across 632,958 autosomal SNPs (Illumina 650K array) for 938 individuals from 53 diverse populations worldwide. The SGVP dataset consists of genotypes across 1,285,226 autosomal SNPs (Illumina 1M and Affymetrix 6.0 arrays) for 96 SG Chinese, 89 SG Malays, and 83 SG Indians. Both datasets were filtered by MAF > 0.01.

We first projected our samples into a worldwide ancestry space generated by the top four PCs of the HGDP data (Figure S2A). Three outliers were found: two clustering with Europeans and one close to the Africans. We excluded these outliers and further projected the other samples on a Singapore ancestry map based on the SGVP data (Figures S2B and S2C). Given the missing data and potential errors in self-reported ethnicity, we re-assigned ethnicity to each sample using the ancestry coordinates on the SGVP map. As shown on Figure S2B, the means of PC1 and PC2 of SGVP Chinese, Malays and Indians form a triangle with individuals from each ethnicity group cluster around one vertex. We divided the map into three sectors by medians of the triangle, and assigned ethnicity labels to our study samples based on the sectors they were projected to (Figure S2C). After removing 36 highly contaminated samples, we labeled 2,780 Chinese, 903 Malays, and 1,127 Indians. The overall concordance rate between inferred and self-reported ethnicity is 96.7%. This lower ethnicity concordance rate was mostly due to recent admixture between different ethnic groups.

We inferred sex based on the ratio of sequencing depths between chromosome X and autosomes, denoted as X/A ratio. Males are expected to have X/A ratio equal to 0.5 and females are expected to have X/A ratio equal to 1. For each sample, we computed the sequencing depth per chromosome using SAMtools (Li et al., 2009), discarding reads with mapping quality < 20 or base quality < 20. Samples with a X/A ratio smaller than 0.75 were inferred as males, while the rest were inferred as females (Figure S2D). In total, we inferred 2,462 females (51.2%) and 2,348 males (48.8%). The mismatch rate in comparison to self-reported sex was 0.76%.

We used VerifyBamID (v 1.1.2; --precise --maxDepth 100 --minMapQ 20 --minQ 20 --maxQ 100) to assess the level of DNA contamination for each sample by comparing sequence reads to the allele frequencies of the inferred ethnicity group (Chinese, Malays, or Indians), which were computed using the SGVP data (Jun et al., 2012). 36 samples with an estimated contamination rate > 0.05 were removed (Figure S2E).

Genotype calling, phasing, and annotation

We performed variant detection and joint genotype calling using the GotCloud pipeline (Jun et al., 2015). We used the version that was used to produce the freeze 3 variant call set for the TOPMed Program, adjusting for sample contamination rates. The initial call set consisted of 107,293,300 SNPs and 15,518,308 INDELs on autosomes, and 4,568,579 SNPs and 686,037 INDELs on X chromosome. Multi-allelic variants were coded as multiple bi-allelic variants at the same position. A support vector machine (SVM) classifier was used to filter low-quality variants. We trained the SVM model for autosomes using variants on chromosome 1. Positive labels were defined as known variants genotyped by the 1KGP on Illumina Omni2.5M array and negative labels were defined as variants that failed in > 2 of the six hard filters on variant features (ABE > 0.7, |STZ| > 5, IBC < −0.1, IOR > 2, CYZ < −5, QUAL < 5). Definition of the variant features can be found on the TOPMed pipeline website. For the X chromosome, we excluded the feature of inbreeding coefficient (IBC) from the SVM model and trained the model using variants on the X. The SVM filter removed 5,097,496 SNPs (Ts/Tv = 1.04) and 4,307,458 INDELs on autosomes and 213,308 SNPs (Ts/Tv = 1.00) and 174,871 INDELs on the X chromosome. Following SVM, we further filtered variants with a Hardy-Weinberg Equilibrium (HWE) p value < 10^{-6} in the direction

of excessive heterozygosity (EXHET). The EXHET filter removed 620,625 SNPs (Ts/Tv = 1.03) and 455,285 INDELs on autosomes and 22,973 SNPs (Ts/Tv = 0.86) and 18,802 INDELs on the X chromosome. The low Ts/Tv ratios of the excluded SNPs suggest the effectiveness of both the SVM filter and the EXHET filter.

To improve genotyping accuracy, we used the BEAGLE software (v 4.1), which took genotype likelihoods as inputs, to perform LD-based genotype refinement (Browning and Yu, 2009). We ran BEAGLE in parallel by splitting each chromosome into chunks of 10,000 variants with an overlap of 1,000 variants between neighboring chunks. Splitting and merging were performed using the *splitvcf.jar* and *mergevcf.jar* programs in BEAGLE Utilities. Low quality variants with dosage $R^2 \leq 0.3$ by BEAGLE were filtered, including 7,314,083 SNPs (Ts/Tv = 1.89) and 887,099 INDELs on autosomes and 36,112 SNPs (Ts/Tv = 1.55) and 19,201 INDELs on the X chromosome.

The genotypes from BEAGLE were used to identify cryptic relatedness and duplicated samples (see next section). Based on the inferred trios/duos and the identified MZ/duplicate pairs, we calculated the total Mendelian and duplicate discordances (DISC) for each variant using PLINK (v 1.9; Chang et al., 2015), and filtered variants with DISC > 2, including 582,277 SNPs (Ts/Tv = 0.96) and 509,361 INDELs on autosomes and 30,936 SNPs (Ts/Tv = 1.07) and 37,566 INDELs on the X chromosome.

We performed population-based haplotype phasing of all 5,381 samples using EAGLE (v 2.3.5) with default settings (Loh et al., 2016). For the X chromosome, we first set the heterozygous calls in non-pseudo-autosomal regions (non-PAR) for males to missing and then phased all samples together. After phasing, subsequent analysis was performed on data from eight cohorts (excluding GUSTO/S-PRESTO). We also removed 8,290,266 variants that are monomorphic in the remaining samples, including 1,015,402 variants on X chromosome. Proportionally, many more X chromosome variants were removed due to the extra step to set non-PAR heterozygotes in males to missing before phasing. The final SG10K call set consisted of phased haplotypes from 4,810 WGS samples, covering 98,273,706 SNPs and INDELs from autosomes and the X chromosome. Across autosomes, 84,725,366 bi-allelic SNPs (Ts/Tv = 2.07), 1,112,853 tri-allelic SNPs, 14,669 quad-allelic SNPs, 3,059,717 insertions and 5,708,237 deletions. On the X chromosome, there are 3,275,181 bi-allelic SNPs (Ts/Tv = 1.97), 31,853 tri-allelic SNPs, 364 quad-allelic SNPs, 124,180 insertions and 221,286 deletions.

We annotated variants in our final call set using the Ensembl Variant Effect Predictor (VEP) with the corresponding VEP-compiled annotation database (v 91_GRCh37) (McLaren et al., 2016). Because VEP only annotates PolyPhen (Adzhubei et al., 2013) and SIFT (Kumar et al., 2009) scores for SNPs, we separately annotated INDELs using the SIFT web server (Sim et al., 2012).

Genetic relatedness, inbreeding coefficient, and consanguinity

We used PC-Relate (Conomos et al., 2016) to estimate both kinship coefficient φ and the probability that two individuals share zero identical by descent (IBD, π_0). We focused on the common SNPs (MAF > 0.05) overlapping with the SGVP dataset. We aggressively pruned the SNPs to be at least 100 kb apart from each other, resulting in 25,568 SNPs, in order to accommodate the huge memory demand of PC-Relate. We also applied the SEEKIN software (GT mode with SGVP as the reference) to estimate kinship coefficients without pruning SNPs, and obtained similar results (Dou et al., 2017). We classified pairs as *k*-degree related if $2^{-k-1.5} < \varphi < 2^{-k-0.5}$ (Manichaikul et al., 2010). A zero-degree related pair means monozygotic twins (MZ) or duplicates. First degree related pairs were further split into parent-offspring (PO) if $\pi_0 < 0.1$ and full-sibling (FS) if $\pi_0 > 0.1$. We treated pairs more distant than 3rd degree as unrelated. We used the PRIMUS software to identify duos and trios using the estimated φ and π_0, as well as information on age and sex (Staples et al., 2014). To identify a maximum number of unrelated individuals, we first listed all related pairs, and then removed the individual that appeared most frequently in the list. Once an individual was removed, the corresponding related pairs were also removed from the list. This procedure was repeated until the list was empty. Notably, we inferred 494 pairs of close relatives (3rd degree or closer), including 93 duplicates (or MZ), 17 trios, and 32 PO duos (Figure S2G; Conomos et al., 2016; Staples et al., 2014). The inbreeding coefficient for each sample was estimated as ($2\varphi_{ii}$-1), where φ_{ii} is the self-kinship coefficient for sample i output by PC-Relate. We estimated the prevalence of consanguineous mating between second cousins or closer relatives by the proportion of healthy individuals with an inbreeding coefficient > 0.0156 in each population.

Evaluation of genotype calling quality

1,263 samples from the SEED cohort were previously genotyped by Illumina Quad610 arrays (Cornes et al., 2012). We used the array data across 40,048 SNPs on chromosome 2 to evaluate the quality of our call set. By treating the array data as gold standard, we estimated the sensitivity, non-reference sensitivity, precision, overall genotype concordance, heterozygote concordance, and non-reference concordance of our call set for common variants genotyped on the array (Linderman et al., 2014).

In addition, we designed a novel approach to estimate the sensitivity for variant detection by assuming a linear relationship between the detected number of non-reference variants and the sequencing depth of a sample. Briefly, for all samples from each population, we fitted a linear model $\hat{y}(x) = a + bx$, in which *y* is the detected number of non-reference variants in each sample, and *x* is the sequencing depth of that sample. Because non-reference sensitivity is defined as the proportion of true variants that can be detected in a sample, we further assumed that 99% of the rare variants with MAF < 0.001 and 100% of the common variants with MAF > 0.001 could be detected at 25× WGS (Rashkin et al., 2017). Therefore, the non-reference sensitivity at 13.7× could be estimated as $Se(13.7) = \hat{y}(13.7)/\hat{y}(25) \times Se(25)$, in which $Se(25) = 0.99$ for rare variants with MAF < 0.001 and $Se(25) = 1$ for common variants with MAF > 0.001. We reported the Pearson's correlation *r* between the number of non-reference variants in each sample and the sample sequencing depth, and tested the null hypothesis of $r = 0$ using the Student's t test.

Finally, we assessed the impact of sequencing depth on the FDR of variant detection by comparison to deep WGS. We selected 9 samples from the HF cohort for additional deep WGS at > 30× : one from each ethnicity (Chinese, Malays, and Indians) at each original sequencing depth (5×, 10×, and 15×). The experimental protocol and sequencing data preprocessing were the same as described previously. We used GATK4.0 Haplotypecaller for joint genotype calling of these 9 samples following the GATK best practices (DePristo et al., 2011). We evaluated FDR on sites that satisfied two criteria: 1) the site was polymorphic and bi-allelic in SG10K call set; 2) the read depth of the site in deep WGS must be larger than 10. We defined the number of true positive calls in SG10K (TP) to be the number of sites that were polymorphic in deep WGS call set, and the number of false positive calls (FP) to be the number of sites that were missing or monomorphic. The FDR was computed as FP/(FP+TP).

PCA, ADMIXTURE, F_{ST}, heterozygosity and nucleotide diversity

We merged our SG10K dataset with 1KGP dataset (1000 Genomes Project Consortium et al., 2015) by extracting 26,748,762 biallelic autosomal SNPs called in both datasets, excluding SNPs within five base pairs (bps) of INDELs. We then removed LD by thinning the SNPs to at least 2kb apart using PLINK (Chang et al., 2015), resulting in 1,260,657 SNPs. To avoid potential artifacts, we further excluded SNPs in long-range LD regions, including 24 regions reported by (Price et al., 2008), and three regions identified by our in-house analysis (chr4: 97.9-103.5Mb, chr14: 66.5-67.9Mb, and chr14: 106-106.3Mb). We investigated population structure using PCA (Wang et al., 2015), ADMIXTURE (Alexander et al., 2009), and the F_{ST} and H_e statistics (Weir and Cockerham, 1984).

To avoid the undesirable impacts of oversampling certain populations (McVean, 2009), we randomly selected 100 unrelated individuals from each Singapore population and combined with the 1KGP dataset for PCA. The remaining SG10K samples were projected into the PCA map using LASER (Wang et al., 2014, 2015). Analyses were based on 260,680 SNPs with MAF > 0.05 in the combined dataset. Similarly, in the analyses of East/Southeast Asians (100 SG Chinese, 100 SG Malays, and 1KGP East Asians), South Asians (100 SG Indians and 1KGP South Asians), and Singaporeans (4,441 unrelated SG10K individuals), we filtered SNPs with MAF < 0.05 in the corresponding datasets. Based on the first two PCs of East/Southeast Asian analysis and South Asian analysis, we used SVM classifiers to classify SG Chinese and SG Indians into northern and southern groups. The SVM classifiers were trained using CHB and CHS for classifying northern and southern Chinese, and using PJL and STU for classifying northern and southern Indians.

We performed the unsupervised ADMIXTURE analyses on the same SNPs as PCA, with the number of ancestral components K from 5 to 15 for the dataset combined with 1KGP and from 2 to 10 for SG10K dataset alone. For each K, we repeated the analysis 10 times with different random seeds and picked the one with the highest likelihood to avoid local minimum. We used the five-fold cross-validation approach in ADMIXTURE to select the optimal K.

We calculated genome-wide F_{ST} between pairs of populations using the Weir-Cockerham estimator (Weir and Cockerham, 1984). Hierarchical clustering was applied on the F_{ST} matrix using the complete-linkage method implemented in the *hclust* function in R.

The expected heterozygosity for each population was calculated by $H_e = 1 - 1/L\left(\sum_{l=1}^{L}\sum_{h=1}^{H_l} p_{lh}^2\right)$, where p_{lh} is the population-specific frequency of allele h at locus l, H_l is the total number of different alleles at locus l, and L is the total number of loci. The analyses were based on 237,000 biallelic SNPs with MAF > 0.05 and at least 2kb apart in the combined SG10K and 1KGP Asian dataset. Therefore, for genotype heterozygosity, L = 237,000 and H_l = 2 across all loci. We also computed haplotype heterozygosity using the same formula, in which haplotype loci were defined by various sliding window sizes of 25, 50, 100, and 200 kb with a fixed step size of 10 kb, and H_l is the number of distinct haplotypes at locus l.

We calculated nucleotide diversity π (Nei and Li, 1979) for each population using VCFtools (Danecek et al., 2011). We removed genomic regions that were less accessible to short-read sequencing based on the "strict mask" from 1KGP (1000 Genomes Project Consortium et al., 2015), and the major histocompatibility complex region (chr6:28477797-33448354) which has unusually high genetic diversity. The final analyses were based on 58,328,923 autosomal biallelic SNPs for 1KGP populations and 59,751,679 autosomal biallelic SNPs for SG10K populations. Global average of nucleotide diversity was calculated by summing up the per-site nucleotide diversity and dividing by the total number of bases within regions passed QC (2,061,388,071 bases).

Demographic inference

Because inbreeding might have a strong influence on the haplotype-based demographic inference, we first excluded samples with estimated inbreeding coefficient > 0.0625. In addition, we selected 200 SG Chinese, 200 SG Malays, and 200 Indians with the least inferred admixture proportion in the ADMIXTURE analysis with K = 3 on the SG10K samples (Figure 3C). Therefore, we could exclude the potential influence of the very recent admixture in the past few generations and focused on demographic events before the recent immigratory history. An additional consideration to select 200 individuals per population was to avoid the impractical computational burden of demographic inference based on too many samples.

For each SG population, we used SMC++ (v 1.15.2) to estimate the history of effective population size (Terhorst et al., 2017). We fixed the mutation rate at 1.25×10^{-8} per generation per base pair, and assumed a constant generation time of 29 years. As recommended by the authors of SMC++, we masked low-complexity regions of the genome (downloaded from the 1KGP ftp site) and focused on bi-allelic SNPs without LD pruning or MAF filtering. We repeated SMC++ 10 times for each population by designating 10 samples with the highest sequencing depth as the distinguished lineage, and the results were combined to get the composite likelihood for the final estimates. In addition, we estimated the split dates between pairs of populations by fitting the clean split model

with the marginal SMC++ estimates for each population. Since we had 10 SMC++ repeats for each population, we obtained 100 estimates of split date between each pair of populations. We reported the median value and used the 2.5th and 97.5th percentiles to construct the 95% CI.

We further investigated the admixture history of Malays using the GLOBETROTTER software package (Hellenthal et al., 2014; Lawson et al., 2012). We included all South and East Asian populations from the 1KGP as the surrogate populations for Malays (the target population). The analyses were based on 227,952 biallelic SNPs with MAF > 0.05 and at least 2kb apart in the combined SG Malays and 1KGP Asian dataset. We first used ChromoPainter (v 2), a companion program of GLOBETROTTER, to get the haplotype sample paintings for target individuals and copying vectors for all individuals, which involved two steps. The first step was to infer the switch rate N_e and the global mutation rate θ by running ChromoPainter with 10 expectation-maximization (EM) iterations on chromosomes 2, 10, and 20 for every individual and obtained the final parameter estimates $\widehat{N}_e$ and $\widehat{\theta}$ by the global averaging weighted by the number of SNPs on each chromosome. Then, we ran ChromoPainter again with the estimated parameter values and without EM iterations to paint each haplotype in the target population with haplotypes from the surrogate populations. We sampled 10 paintings per haplotype (i.e., 20 paintings per chromosome) for target individuals and obtained copying vectors for all individuals. Finally, we used the GLOBETROTTER program to estimate admixture dates by fitting the co-ancestry curves with different admixture models. The co-ancestry curves were derived from the sampling paintings and copying vectors from ChromoPainter. We set the grid range for co-ancestry curves to be from 1 cM to 10 cM with bin size equal to 0.01 cM, because 99% of the exponential decay of the co-ancestry curves in our data occurred within 10cM. We applied 100 bootstraps to assess the statistical significance of the admixture event and the 95% CI of the inferred admixture date. We also repeated the GLOBETROTTER analysis on 100 randomly selected unrelated Malay individuals, for which we used the default grid range of 1 cM to 50 cM with bin size equal to 0.1 cM in order to detect both ancient and recent admixture events.

Scan for selection

We computed PBS using 6,822,300 bi-allelic SNPs that are polymorphic in all three populations and have MAF > 0.05 in at least one population. To adjust for the confounding effects of admixture (Huerta-Sánchez et al., 2013), we proposed to compute PBS between the three major ancestral components derived from the ADMIXTURE analysis of 4,441 unrelated SG10K samples at K = 3. The three components represent Chinese, Malays, and Indians respectively. We coded the genotypes as $G_{ij} = 0,\ 1,\ and\ 2$ and the admixture proportions as Q_{ik}, in which $i = 1, 2, \ldots, N$ indexes individuals, $j = 1,\ 2, \ldots,\ M$ indexes SNPs, and $k = 1,\ 2,$ and 3 indexes ancestral components. We estimated SNP-level F_{ST} using the Weir and Cockerham's estimator (Bhatia et al., 2013; Weir and Cockerham, 1984), in which the effective sample size of autosomal haplotypes and the allele frequency of SNP j in ancestral component k were calculated as $n_k = \sum_{i=1}^{N} 2Q_{ik}$ and $f_{jk} = (\sum_{i=1}^{N} Q_{ik}G_{ij})/n_k$, respectively. For non-PAR SNPs on the X chromosome, we used $n_k = \sum_{i=1}^{N} s_i Q_{ik}$, where $s_i = 1$ for male and 2 for female. We then scanned through the genome using a 50 kb sliding window with a step size of 10 kb. The window size was chosen so that LD decays roughly to the background level in East Asians (Jakobsson et al., 2008). We computed window-level F_{ST} from SNP-level statistics using the "ratio of average" approach. Finally, we computed PBS both at the SNP level and the window level from the F_{ST} statistics, for which negative values were set to 0 (Yi et al., 2010). To compare PBS across different populations, we standardized PBS within each ancestral component by subtracting the mean and dividing the standard deviation.

We defined an outlier window if its maximal standardized PBS value across three populations (PBS_{max}) was among the top 0.1% PBS_{max} across all windows. We defined a locus starting from the most outlying window, in which the SNP with the highest PBS was labeled as the index SNP. We then gradually merged adjacent outlier windows into the locus if the windows contain SNPs ranking among the top 0.1% SNP-level PBS and in LD ($r^2 > 0.2$ and < 1 Mb apart) with the index SNP in any of the three populations. The boundaries of a locus were set by the two farthest SNPs in LD with the index SNP. We note that the index SNP does not necessarily have the highest PBS in the locus, because there might be SNPs with higher PBS values in neighboring windows rather than the most outlying window. We visualized each outlier locus using LocusZoom (Pruim et al., 2010), in which LD was defined as the maximal r^2 across three populations.

To gain insights on where selection occurred, we performed three additional PBS scans, in each of which we included CEU as an outgroup, together with two out of the three SG populations. Due to the relatively small sample size of CEU, we relaxed the SNP filtering criteria to biallelic SNPs with MAF > 0.05 in at least one population, resulting in 6,654,606 SNPs for the analysis of (CEU, Chinese, Malays), 6,830,701 for (CEU, Chinese, Indians), and 6,793,994 for (CEU, Malays, Indians). We focused on top windows identified by the PBS analysis of three SG populations, and examined if they were among the top 0.1% or top 1% windows in any branches of the three additional PBS analyses.

We systematically searched literature in PubMed for previously hypothesized selection signals within candidate loci identified by PBS. We used the PubMed search term "(natural selection [All Fields]) OR (positive selection [All Fields]) AND human [All Fields] AND GENE-NAME [All Fields]" for all genes within the candidate loci. In addition, we searched for reported GWAS signals with $p < 5 \times 10^{-8}$ within the PBS loci using the GWAS Catalog (v 1.0.2-associations_e93_r2018-08-15) (MacArthur et al., 2017). We assessed if GWAS hits were more likely to be under selection by a χ^2 test on a 2 × 2 contingency table, in which elements were SNP counts in GWAS

Catalog with $p < 5 \times 10^{-8}$ and/or within the PBS selection loci. To avoid dependency between SNPs due to LD, we restricted this analysis to 1,260,657 biallelic autosomal SNPs, which were shared by 1KGP and SG10K and were thinned to be at least 2kb apart from each other.

Imputation experiments

We evaluated imputation accuracy both in three SG populations from SGVP and in 53 worldwide populations from HGDP. We extracted 46,338 bi-allelic SNPs on chromosome 2, which had consistent alleles in SGVP, HGDP and 1KGP datasets. We then masked genotypes of 4,633 SNPs (1 out of every 10 SNPs sorted by position). The masked genotypes were saved for evaluation of imputation accuracy.

We prepared three imputation reference panels for comparison: 1KGP, SG10K, and a combined panel that merged 1KGP and SG10K datasets. For the 1KGP panel, we used BCFtools to normalize INDELs, split multi-allelic variants into multiple bi-allelic variants, and removed duplicated variants and extremely rare variants (minor allele counts, MAC < 5), resulting in a panel of 2,650,510 variants and 5,008 haplotypes (chromosome 2 only). For the SG10K panel, we took the phased data of 4,441 unrelated samples and removed variants with MAC < 5, leading to a panel of 2,340,867 variants and 8,882 haplotypes. The number of variants in both 1KGP and SG10K panels is 1,233,346, indicating a substantial loss of variants if merging by intersection. Instead, we used the reciprocal imputation approach to merge 1KGP and SG10K to their union set of variants (Huang et al., 2015). Briefly, we used Minimac4 (Das et al., 2016) to impute 1KGP to SG10K and SG10K to 1KGP respectively, and then merged the two imputed datasets to form a combined panel. After removing 10,111 INDELs that have incompatible allele representations in 1KGP and SG10K, the combined panel has 3,747,920 variants and 13,900 haplotypes.

Given a reference panel, we pre-phased each population from HGDP and SGVP using a reference-based phasing algorithm in EAGLE2 (Loh et al., 2016), followed by imputation using Minimac4 (Das et al., 2016). Imputation error rate was computed for each population as the genotype discordance rate of the 4,633 masked SNPs. In addition, for each of the three SGVP populations, we compared the Rsq statistic for imputed variants in different MAF bins (MAF: 0.005-0.01, 0.01-0.05, 0.05-0.2, and 0.2-0.5) (Li et al., 2010). We did not compare the Rsq statistic for HGDP populations because Rsq cannot be estimated accurately when the sample size of the target population is small.

DATA AND CODE AVAILABILITY

The accession number for the sequence data reported in this paper is EGA: EGAS00001003875. Further information about the European Genome phenome Archive (EGA), which is hosted by the EBI and CRG, can be found on https://ega-archive.org (Lappalainen et al., 2015).

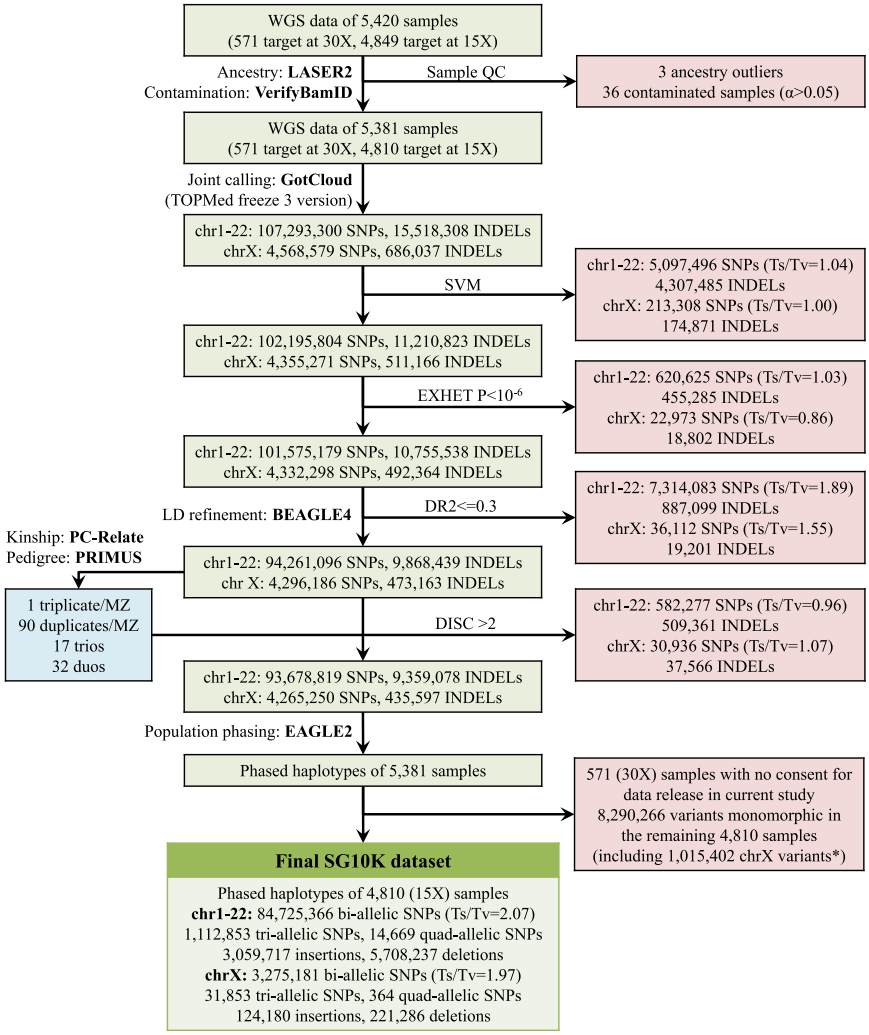


Figure S1. Flowchart of SG10K Data Processing and Quality Controls, Related to Figure 1

*The proportion of chromosome X variants being removed in the last step is substantially higher than that of autosomes due to the extra step of setting heterozygotes in non-PAR regions of chromosome X in males to missing before EAGLE2 phasing.

Abbreviations of variant filters: SVM, support vector machine; EXHET, excessive heterozygosity; DR2, dosage R^2; DISC: sum of Mendelian discordances and duplicate mismatches.

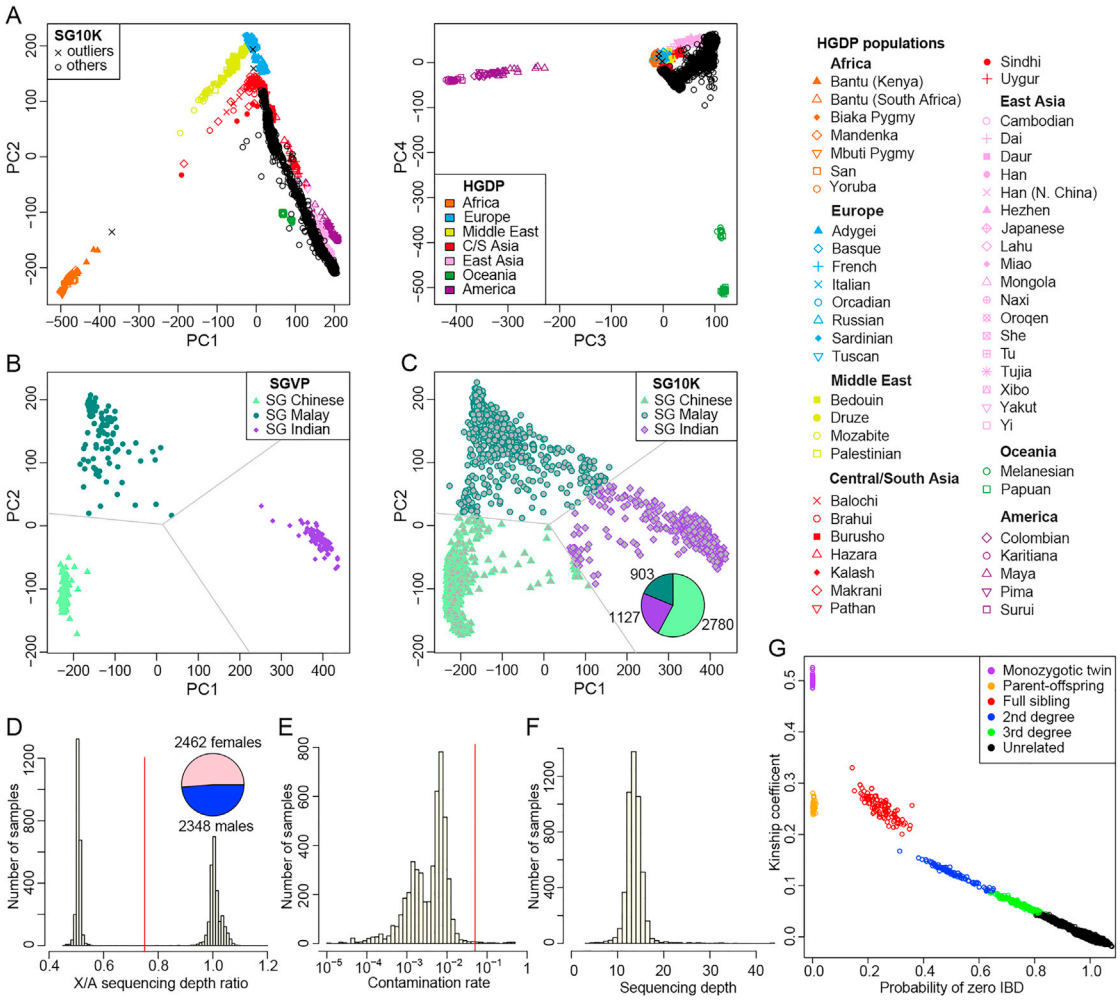


Figure S2. Quality and Sample Overview of SG10K Data, Related to Figures 1, 2, and 3

Results for 571 samples from the GUSTO/S-PRESTO birth cohort were not shown.

(A) Worldwide ancestry estimation of 4,813 sequenced samples (excluding 36 highly contaminated samples) by projecting onto the top 4 PCs of the HGDP panel.

(B) Singapore ancestry space of the top 2 PCs based on the SGVP panel. The space is divided into three ancestry sectors by the median lines (gray lines) of the triangle formed by mean coordinates of SGVP Chinese, Malays, and Indians, respectively.

(C) Inferred ethnicity by projection of 4,810 sequenced samples (further excluding 3 ancestry outliers in A) into the Singapore ancestry space. The pie chart summarizes the sample size of each ethnicity group.

(D) Inferred sex based on the ratio of sequencing depths between chromosome X and autosomes. The X/A ratio is expected to be around 0.5 and 1 for males and females, respectively. The red line indicates a threshold of 0.75, which divided the samples into 2,462 females and 2,348 males as shown in the inserted pie chart.

(E) Distribution of contamination rate. We identified 36 samples with contamination rate $\alpha > 0.05$, indicated by the red line.

(F) Distribution of sequencing depth.

(G) Inferred genetic relatedness based on estimated kinship coefficient ϕ and the probability of zero-IBD-sharing π_0 for pairs of individuals. Relatedness types were determined using the criteria of ϕ and π_0 by Manichaikul et al. (2010).

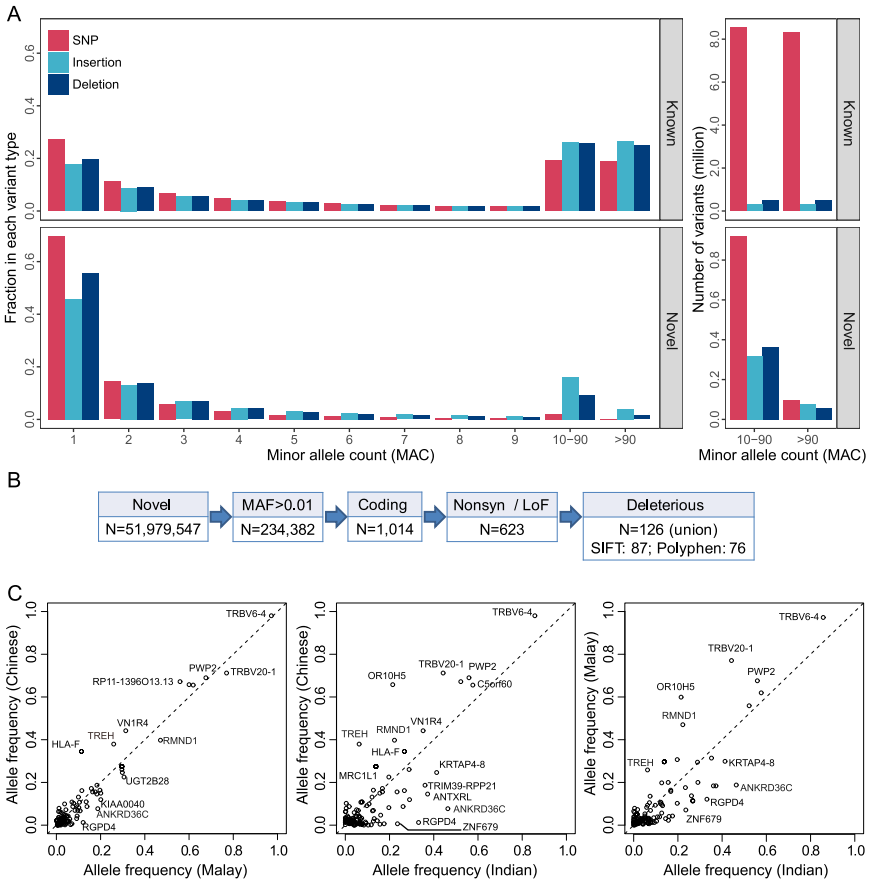


Figure S3. Novel Variants Detected in the SG10K Dataset, Related to Table 1

MAC and MAF were computed based on 4,441 unrelated individuals.

(A) MAC distribution for known and novel variants.

(B) Procedure to identify 126 deleterious novel common variants (MAF > 0.01) in Singaporeans. Deleterious variants are defined as those annotated as "deleterious" by SIFT or "probably/possibly damaging" by PolyPhen.

(C) Allele frequency comparison for 126 high-frequency deleterious novel variants. Corresponding gene names for some variants were labeled.

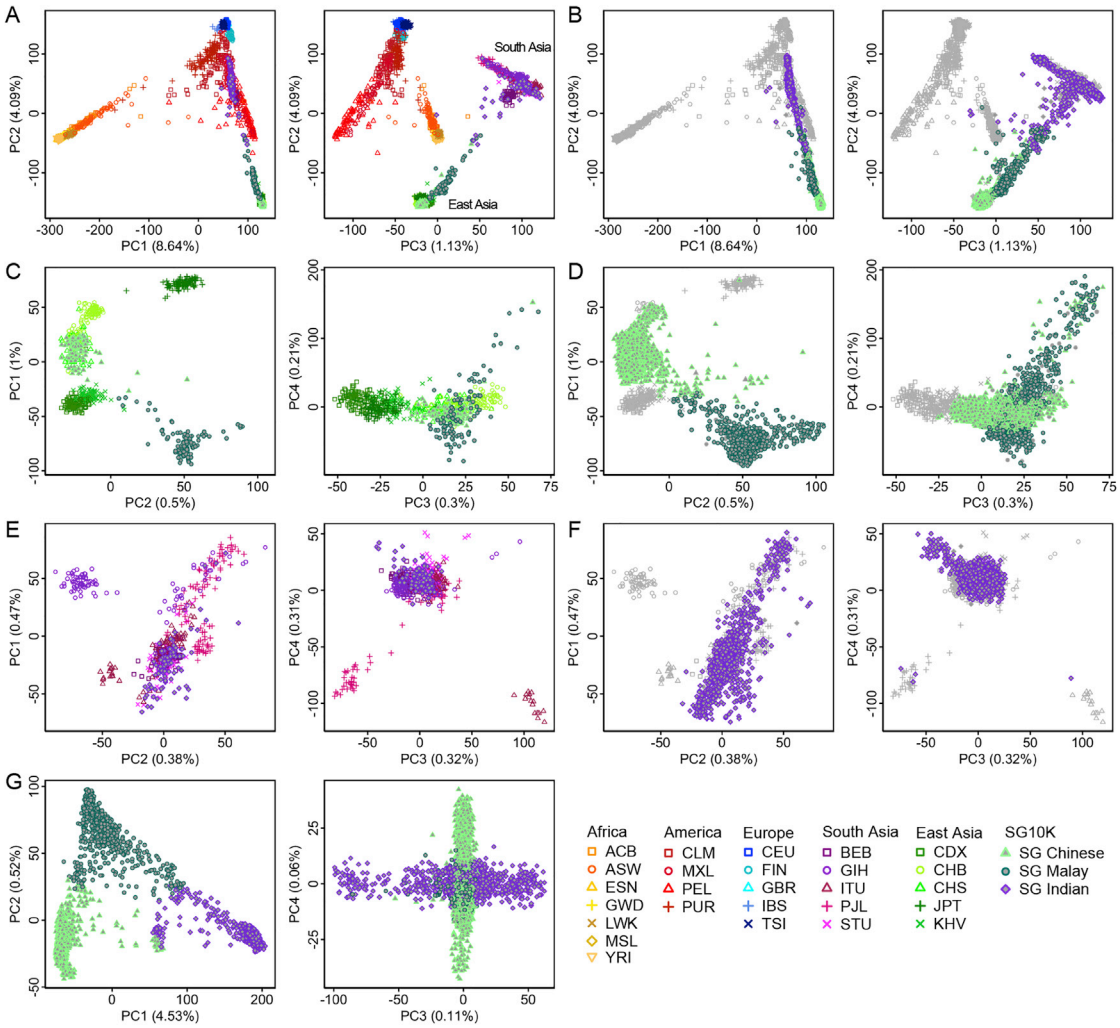


Figure S4. PCA of 1KGP and SG10K Populations, Related to Figure 3

(A) PCA of all 1KGP individuals and 300 SG10K individuals (100 from each population).

(B) Projection of all remaining SG10K individuals into the reference PC space of (A).

(C) PCA of 1KGP East Asians, 100 SG Chinese, and 100 SG Malays.

(D) Projection of all remaining SG Chinese and Malays into the reference PC space of (C).

(E) PCA of 1KGP South Asians and 100 SG Indians.

(F) Projection of all remaining SG Indians into the reference PC space of (E).

(G) PCA of 4,441 SG10K unrelated individuals.

In panels (A-G), proportion of variance explained by each PC is indicated in the axis label. In panels (B), (D) and (F), reference individuals are colored in gray.

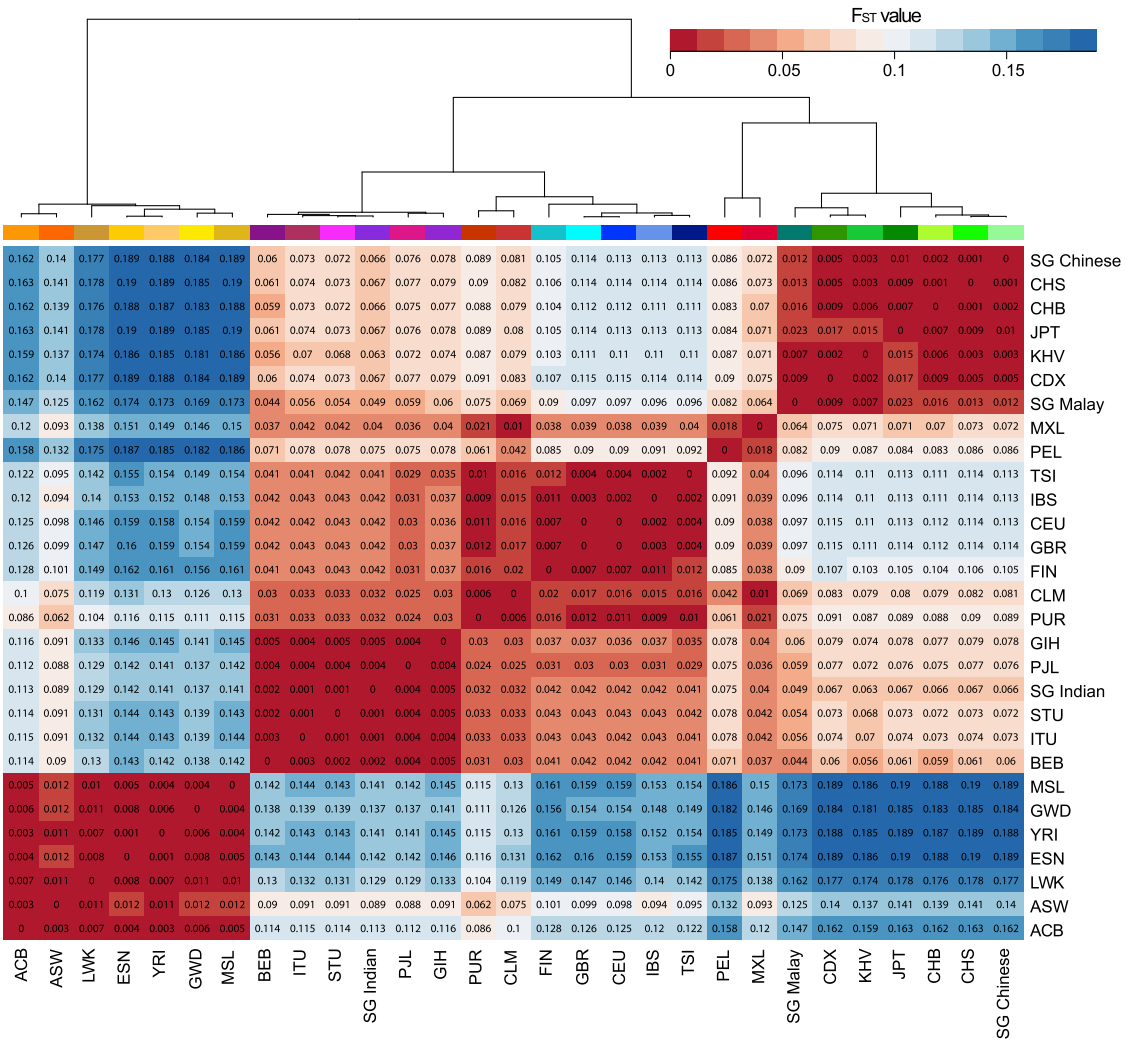


Figure S5. Pairwise F_{ST} between 1KGP and SG10K Populations, Related to Figure 3

Hierarchical clustering was applied on the F_{ST} matrix using the complete-linkage method implemented in the *hclust* function in R.

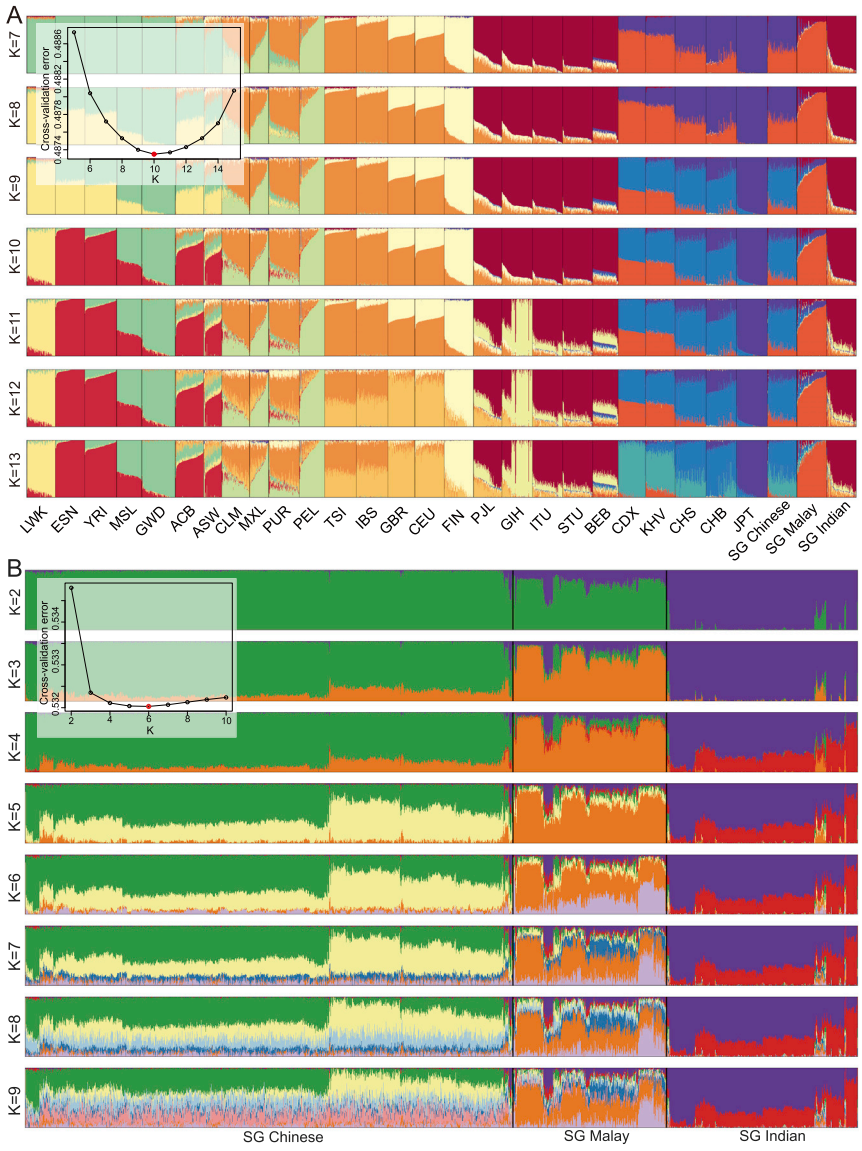


Figure S6. ADMIXTURE Analysis of 1KGP and SG10K Individuals, Related to Figure 3

Unsupervised ADMIXTURE analyses were performed on biallelic autosomal SNPs with MAF > 0.05 and at least 2kb apart from each other. Each colored bar represents one individual and the length of each colored segment represents the proportion of an ancestral component. The inserted panels were five-fold cross-validation error for different *K*, the number of hypothetical ancestral components.

(A) ADMIXTURE analysis of 2,504 1KGP and 300 SG10K individuals.

(B) ADMIXTURE analysis of 4,441 unrelated SG10K individuals.

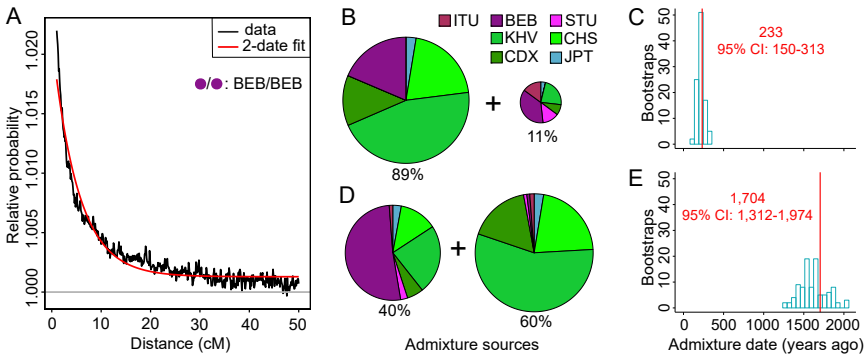


Figure S7. GLOBETROTTER Analysis of 100 Randomly Selected SG Malay Individuals, Related to Figure 3

The analysis was based on biallelic autosomal SNPs with MAF > 0.05 and at least 2kb apart from each other. All East and South Asian populations from 1KGP were included as the surrogate populations. The best guess model was a complex multi-date-two-party admixture model. To gain quantitative insights of the admixture history, we forced GLOBETROTTER to fit a two-date-two-party model.

(A) The co-ancestry curve for SG Malays showing the relative probability of finding two segments both copied from BEB at a given genetic distance.

(B) Ancestral composition of the two sources of the recent admixture event.

(C) Distribution of the estimated dates of the recent admixture event based on 100 bootstraps. The vertical red line indicates the point estimate using the original data.

(D) Ancestral composition of the two sources of the ancient admixture event.

(E) Distribution of the estimated dates of the ancient admixture event based on 100 bootstraps. The vertical red line indicates the point estimate using the original data.

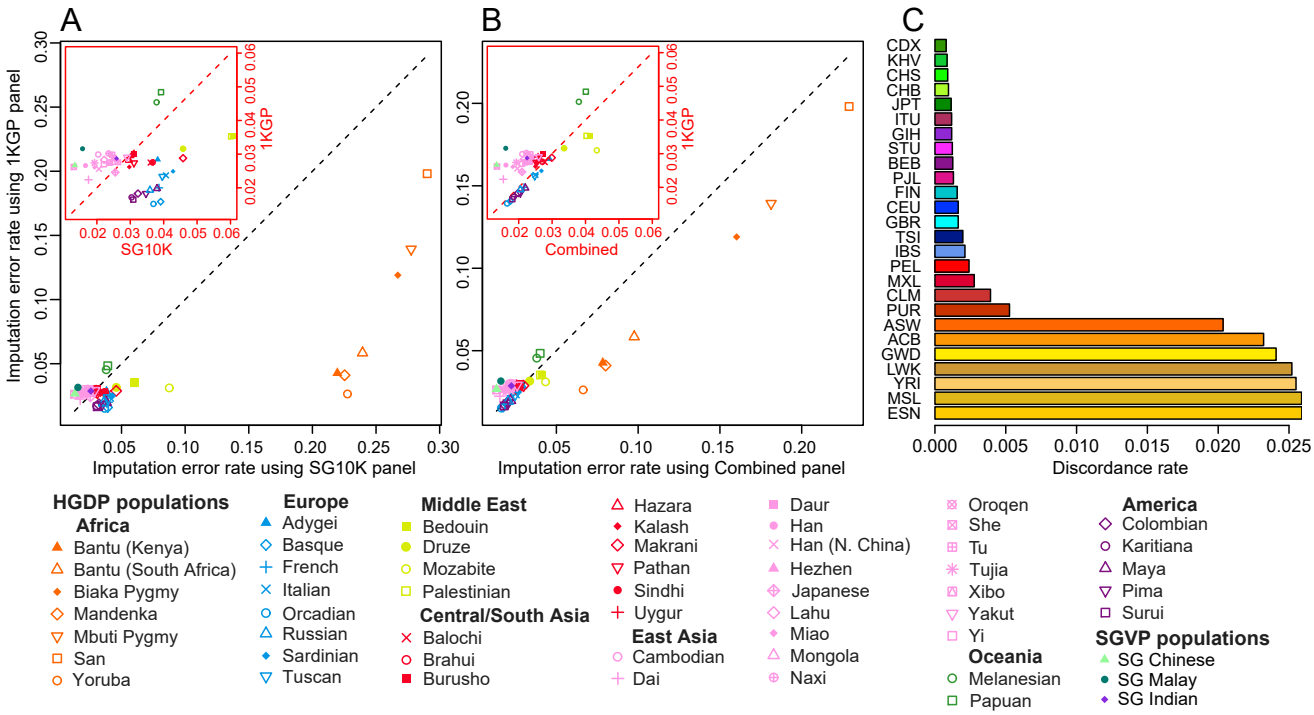


Figure S8. Comparison of Imputation Accuracy in 56 Worldwide Populations Using Different Reference Panels, Related to Figure 5

Three reference panels were compared: 1KGP; SG10K; and the Combined panel by merging 1KGP and SG10K using the reciprocal imputation approach. Imputation error rate was calculated based on 4,633 masked SNPs on chromosome 2.

(A) SG10K versus 1KGP.

(B) Combined panel versus 1KGP. The red inserted boxes are zoom-in plots of the left-bottom corners.

(C) Discordance rate of genotyped SNPs in 1KGP before and after imputation to the SG10K panel. Evaluation was based on 4,633 masked SNPs in the imputation experiments of (A) and (B). Despite that these SNPs were genotyped in 1KGP, imputing 1KGP data to the SG10K panel changed their genotypes, especially for the African populations.