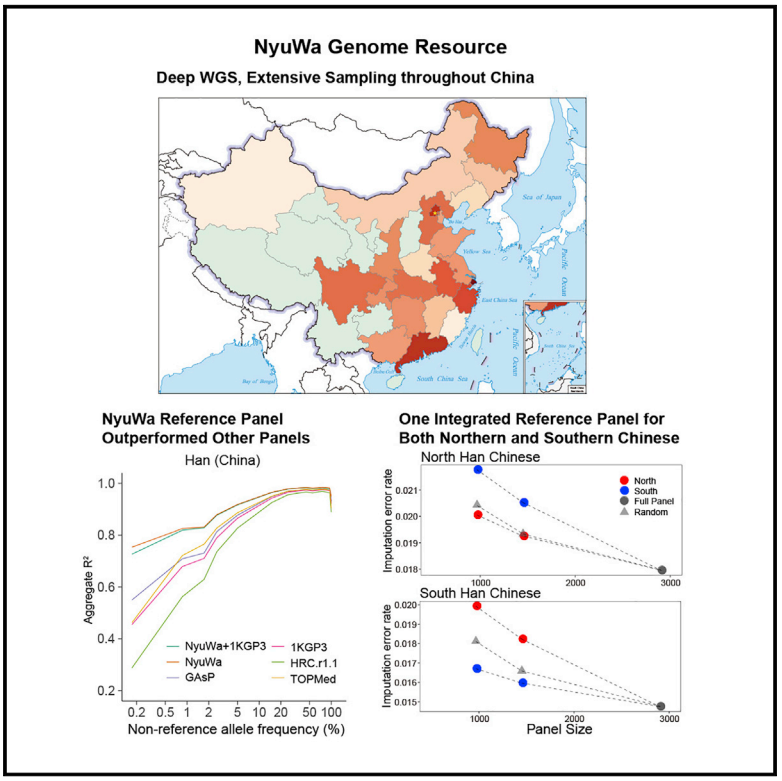
Cell Reports

NyuWa Genome resource: A deep whole-genome sequencing-based variation profile and reference panel for the Chinese population

Graphical abstract



Authors

Peng Zhang, Huaxia Luo, Yanyan Li, ..., The Han100K Initiative, Tao Xu, Shunmin He

Correspondence

xutao@ibp.ac.cn (T.X.), heshunmin@ibp.ac.cn (S.H.)

In brief

Zhang et al. construct the NyuWa genome resource with 79.3 million variants and a reference panel of 5,804 haplotypes based on deep whole-genome sequencing of Chinese individuals, which will help the study of population genetics, medical genetics, and genotype-phenotype association in the world's largest population.

Highlights

- Identification of 25.0 million variants by WGS of 2,999 Chinese individuals
- NyuWa reference panel outperforms public ones for Chinese populations
- A reference panel applicable for northern and southern Chinese populations
- Clinical insights and loss-of-function analysis

Zhang et al., 2021, Cell Reports *37*, 110017
November 16, 2021
https://doi.org/10.1016/j.celrep.2021.110017

Cell Reports

CellPress
OPEN ACCESS

Resource

NyuWa Genome resource: A deep whole-genome sequencing-based variation profile and reference panel for the Chinese population

Peng Zhang,[1,5] Huaxia Luo,[1,5] Yanyan Li,[1,2,5] You Wang,[2,5] Jiajia Wang,[1,3,5] Yu Zheng,[1,3,5] Yiwei Niu,[1,3] Yirong Shi,[1,4] Honghong Zhou,[1] Tingrui Song,[1] Quan Kang,[1] The Han100K Initiative, Tao Xu,[2,*] and Shunmin He[1,3,6,*]

[1]Key Laboratory of RNA Biology, Center for Big Data Research in Health, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China

[2]National Laboratory of Biomacromolecules, CAS Center for Excellence in Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China

[3]College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

[4]University of Chinese Academy of Sciences, Beijing 100049, China

[5]These authors contributed equally

[6]Lead contact

*Correspondence: xutao@ibp.ac.cn (T.X.), heshunmin@ibp.ac.cn (S.H.)

https://doi.org/10.1016/j.celrep.2021.110017

SUMMARY

The lack of haplotype reference panels and whole-genome sequencing resources specific to the Chinese population has greatly hindered genetic studies in the world's largest population. Here, we present the NyuWa genome resource, based on deep (26.2×) sequencing of 2,999 Chinese individuals, and construct a NyuWa reference panel of 5,804 haplotypes and 19.3 million variants, which is a high-quality publicly available Chinese population-specific reference panel with thousands of samples. Compared with other panels, the NyuWa reference panel reduces the Han Chinese imputation error rate by a margin ranging from 30% to 51%. Population structure and imputation simulation tests support the applicability of one integrated reference panel for northern and southern Chinese. In addition, a total of 22,504 loss-of-function variants in coding and noncoding genes are identified, including 11,493 novel variants. These results highlight the value of the NyuWa genome resource in facilitating genetic research in Chinese and Asian populations.

INTRODUCTION

Comprehensive catalogs of genetic variation are fundamental building blocks in research of population and demographic history, medical genetics, and genotype-phenotype associations. Since the first assembly of the human genome was released in 2003 (International Human Genome Sequencing Consortium, 2004), many large-scale whole-genome sequencing (WGS) projects have been launched in Western countries and, more recently, in Asia, creating large and diverse population genetic variation resources. Constructing a haplotype reference panel from large cohort WGS resources is a meaningful and cost-effective way to facilitate genome-wide association studies (GWASs), mainly by imputation of unobserved genotypes into samples that have been assayed using relatively sparse microarrays or low-coverage sequencing (Asimit and Zeggini, 2012; McCarthy et al., 2016). However, there is no specific reference panel for the Chinese population, which is the largest ethnic group in the world.

A remarkable milestone among population genome projects is the 1000 Genomes Project, which released an important resource of ~7.4× WGS data from 2,504 individuals in 26 populations and constructed a reference panel (1KGP3) of 5,008 haplotypes and over 88 million variants (Auton et al., 2015). This resource provides a benchmark for surveys of human genetic variation and has facilitated numerous GWASs through imputation of variants that are not directly genotyped, enabling a deeper understanding of the genetic architecture of complex diseases (Timpson et al., 2018). Nevertheless, rare and low-frequency variants tend to be specific to a population or sample (Auton et al., 2015), and many disease-related variants are very rare and population specific (Bomba et al., 2017; Maher et al., 2012; Saint Pierre and Génin, 2014). The GWASs missed a proportion of potential trait-associated variants that were poorly imputed with current reference panels (Asimit and Zeggini, 2012; Bomba et al., 2017; Hoffmann and Witte, 2015). Therefore, a number of projects have focused on specific populations, attempting to capture population-specific genetic variability and build specific reference panels. For example, the Genome of the Netherlands (GoNL) Project sequenced the whole genomes of 250 Dutch parent-offspring families, found a large number of novel rare variants, and constructed a reference panel with 998 haplotypes (Francioli et al., 2014). Based on the GoNL panel, researchers discovered that a rare variant, rs77542162, was associated with blood lipid levels in the Dutch

population (van Leeuwen et al., 2015). Later, there were other such projects, including UK10K in the United Kingdom population (Walter et al., 2015), SISu in the Finnish population (Chheda et al., 2017), and GenomeDenmark (Maretty et al., 2017). However, these resources are biased toward European populations. Recently, some genomic resources and panels have also been created for Asian populations, including the Japanese population in the work of Nagasaki et al. (2015), 219 population groups across Asia in the GenomeAsia 100K project (GAsP) (Wall et al., 2019), and three Singaporean populations in the SG10K project (Wu et al., 2019). Some studies have also focused on the Chinese population, but these studies had limited sample sizes (Du et al., 2019; Lan et al., 2017) or geographical coverage (Lin et al., 2018) or relied mainly on low-coverage WGS (1.7× or 0.1×) (Gao et al., 2020; Liu et al., 2018a). In a most recent study, the China Metabolic Analytics Project (ChinaMAP) presented a deep WGS (40.8×) dataset of 10,588 Chinese individuals, focusing on metabolic disease (Cao et al., 2020). However, no reference panel has yet been constructed from that study. The Han Chinese population is the largest ethnic group in East Asia and even worldwide, comprising approximately 1.23 billion people. Han Chinese people account for ~20% of the global human population and ~92% of the mainland Chinese population (Xu et al., 2009). Constructing an integrated, large-cohort, high-quality genetic variation database and reference panel for the Han Chinese population is imperative; such a resource would help clarify the population structure and population history and facilitate genetic studies in the world's largest population.

Here we present the genome resource NyuWa, based on deep (median, 26.2×) WGS of 2,999 Chinese individuals from 23 of 34 administrative divisions in China. NyuWa, or Nüwa, is the mother goddess who was the creator of the human population in Chinese mythology. The NyuWa genome resource includes a total of 71.1 million single-nucleotide polymorphisms (SNPs) and 8.2 million small insertions or deletions (indels), of which 25.0 million are novel. More importantly, we constructed the NyuWa reference panel of 5,804 haplotypes and 19.3 million variants; this resource is a high-quality publicly available Chinese population-specific reference panel with thousands of samples and currently has the best performance for imputation in the Han Chinese population. We also found 1,140 pathogenic variants, 18,711 loss-of-function protein-truncating variants (PTVs), and 3,793 long noncoding RNA (lncRNA) splicing variants, of which 11,493 were novel compared with existing genome resources. The NyuWa genome resource can provide useful and reliable support for genetic and disease studies. The NyuWa variant database and imputation server are available at http://bigdata.ibp.ac.cn/NyuWa/.

RESULTS

Large Chinese population cohort of deep WGS data

The NyuWa genome resource included high-coverage (median depth, 26.2×) WGS of 2,999 different Chinese samples, including diabetes and control samples collected from hospitals and physical examination centers. The samples were from 23 administrative divisions in China, including 17 provinces, 2 autonomous regions, and 4 municipalities directly under the central government (termed "provinces" for simplicity; Figure 1A), which can be summarized into several geographical divisions of China (Table S1). The origins of the samples were referenced to the native places or the provinces where samples were collected. The majority of samples were collected from Shanghai, Guangdong, and Beijing (Figure 1A), which all have numerous residents from external provinces. The ethnicities associated with the samples were not available at the time of the study. Because national minorities are usually clustered geographically in China and are not numerous in our sampling areas, we estimated that the Han Chinese ethnicity made up the overwhelming majority of our samples.

Most of the samples were sequenced at a depth of more than 30× (median 38.9; Figure S1A). After genome alignment and removal of duplicates, the median of actual genomic coverage was 26.2× (Figure 1B; Figure S1B). Samples with contamination levels of alpha $\geq$ 0.05 were removed (Figure S1C). Based on the genomic coverage of sex chromosomes, the sex of each subject could be clearly identified except for one potential XO type (Figure 1C). The ploidy of chromosome X (chrX) for the sample also supported the XO type, which was classified as female. In total, there were 1,335 females and 1,664 males. After identification of close relatives within the third degree (Figure S1D), we found that the NyuWa dataset contained a maximum of 2,902 independent samples.

Discovery of 25.0 million novel variants in the NyuWa resource

Variants were called and filtered using NyuWa cohort variant calling pipeline (STAR Methods). SNPs and indels were genotyped jointly using GATK (Poplin et al., 2017) with human reference genome version GRCh38/hg38. After site quality filtering, a total of 76.4 million variant sites were identified, including 2.5 million multiallelic sites (Figure S2A). After splitting of multiallelic sites, the final dataset contained 71.1 million SNPs and 8.2 million indels (Figure S2B), including 2.5 million SNPs and 0.3M indels from sex chromosomes (Table S2). The transition-to-transversion ratio (Ts/Tv) is 2.107 for all biallelic SNPs, which is consistent with previous whole-genome studies such as 1KGP3 (2.09) (Auton et al., 2015) and UK10K (2.15) (Walter et al., 2015).

Compared with other public variant repositories, including ExAC (Lek et al., 2016), gnomAD (v2 and v3) (Lek et al., 2016), 1KGP3, ESP (NHLBI GO Exome Sequencing Project), dbSNP (v150) (Sherry et al., 2001), GAsP, 90 Han (Lan et al., 2017), and TOPMed (Taliun et al., 2019), the NyuWa dataset contained 25.0 million novel variants, including 23.1million SNPs (32.5%) and 1.9 million indels (23.3%) (Figure 2A). The ChinaMAP resource (Cao et al., 2020) merely provided a website for variant search and did not make a full variant list available. To estimate the ratio of novel variants compared with ChinaMAP, we used two variant sets for manual comparison. The first set was 230 novel singletons selected randomly from the NyuWa dataset (10 per chromosome); only 21.3% of variants also existed in the ChinaMAP dataset. Another set consisted of novel variants in 906 cancer-related genes collected from the ClinGen database and literature (Huang et al., 2018; Mirabello et al., 2020; Rehm et al., 2015). There were a total of ~959,000 novel variants in these genes, and only 27.3% of these variants overlapped with

CellPress
OPEN ACCESS

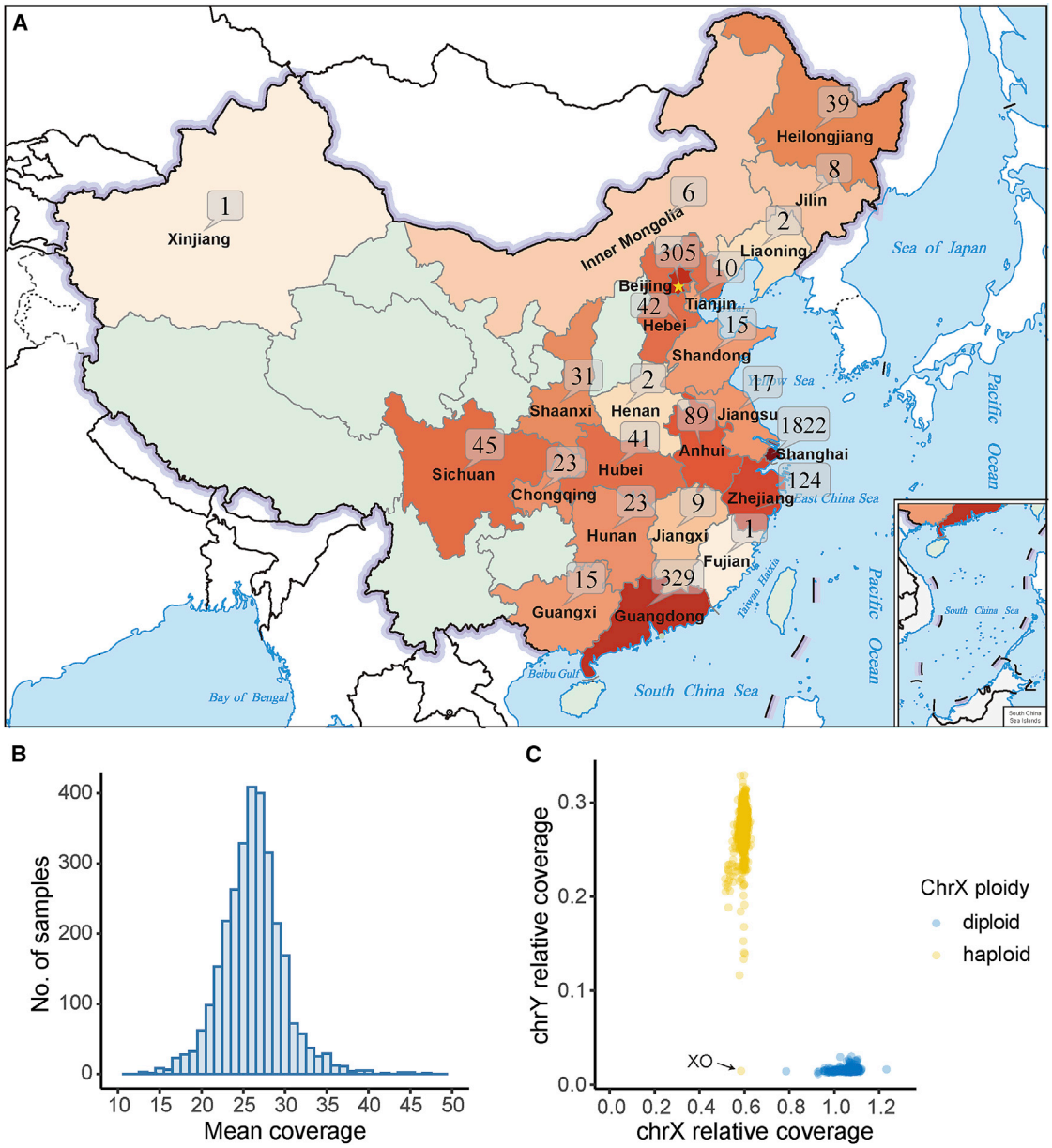


Figure 1. Overview of the NyuWa dataset

(A) Distribution of samples in the NyuWa resource. Samples were assigned to provinces based on the native places or hospitals where the samples were collected.

(B) The distribution of WGS mean genomic coverage after genome alignment and removal of duplicates.

(C) The sex of each sample inferred by sex chromosome coverage and ploidy of the chrX non-pseudo-autosomal region (PAR) estimated by the BCFtools plugin guess-ploidy. The results were consistent for all samples except one with no chrY coverage and a haploid chrX. This special sample was a putative XO type and was classified as female.

See also Figure S1 and Table S1.

ChinaMAP. We estimated that approximately 73% of novel variants would remain (~18 million) after removal of variants in ChinaMAP. As expected, most novel variants were extremely rare, with singletons, doubletons, and tripletons accounting for 86.8%, 10.1%, and 1.9% of novel variants, respectively (Figure 2A). This is not surprising because rare variants are usually specific to a sample or population (Francioli et al., 2014). The absolute number of novel variants with a minor allele frequency (MAF) greater than 0.1% was still large (~77,200). These variants are frequent enough to be subject to large-scale genetic association studies and may lead to new biological discoveries (Piton et al., 2013; Walter et al., 2015). The large overall number of novel variants indicates severe underrepresentation of variants from the Chinese population in recent genetic studies.

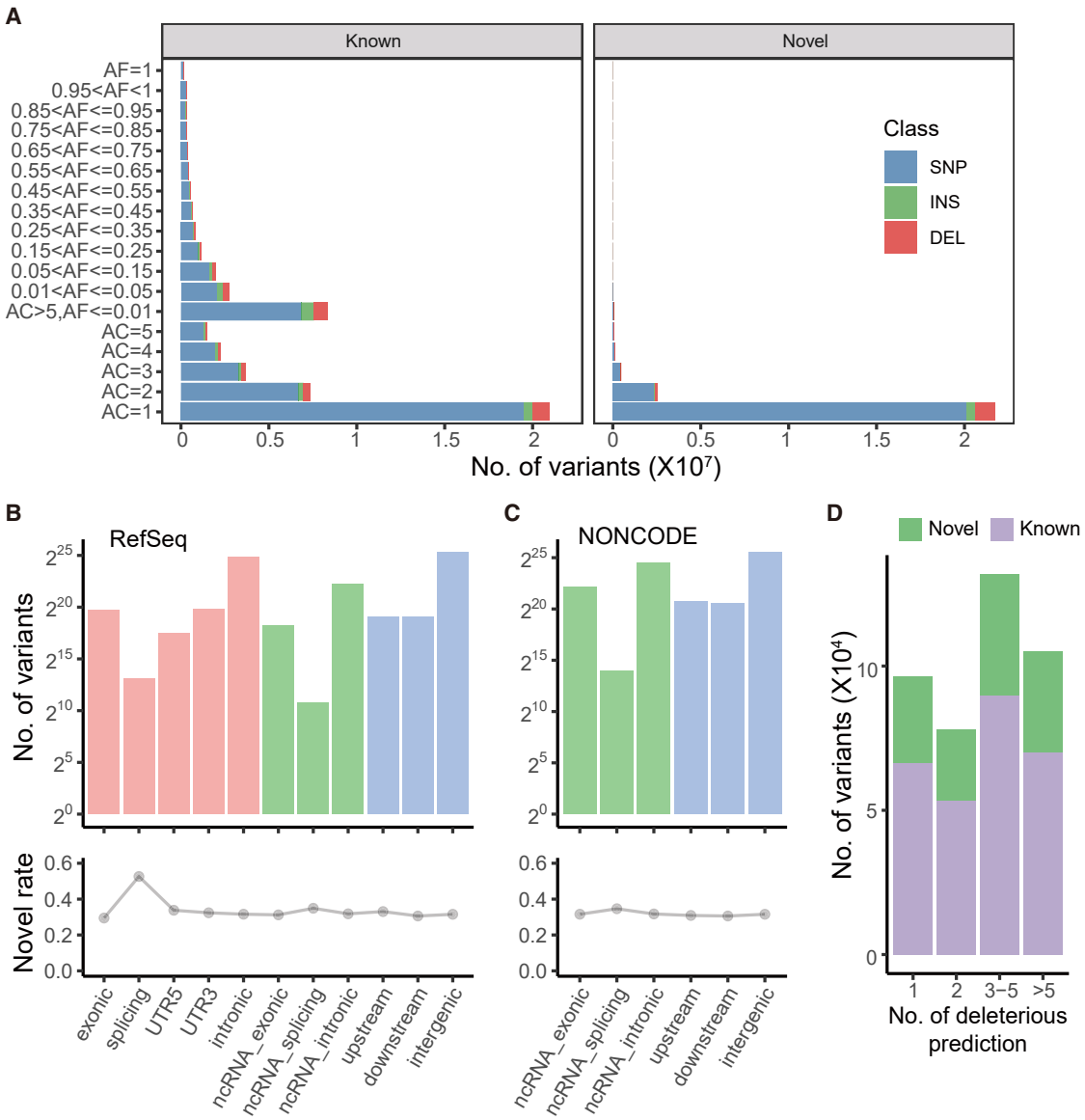


Figure 2. Statistics on variants in the NyuWa resource

(A) Numbers of variants detected in different bins of allele counts or frequencies. Variants were classified as known or novel based on public resources, including ExAC, gnomAD v2 and v3, 1KGP3, ESP, dbSNP, TOPMed, 90 Han, and GAsP. INS, small insertion; DEL, small deletion.

(B) Numbers (top) and novel rates (bottom) of variants in different RefSeq annotation regions.

(C) Numbers (top) and novel rates (bottom) of variants in different NONCODE annotation regions.

(D) Numbers of nonsynonymous SNPs predicted as deleterious by different number of 10 selected prediction algorithms (SIFT, PolyPhen2 HDIV & HVAR, LRT, MutationTaster, FATHMM, PROVEAN, MetaSVM, MetaLR, and M-CAP) provided by dbNSFP. The novel variants are based on results in (A).

See also Figure S2 and Tables S2–S5.

A typical NyuWa sample carries a median number of 3.51 million SNPs and ~523,000 indels in autosomes. These numbers are close to those for East Asia samples in 1KGP3 (3.55M SNPs, ~546,000 indels). The number of detected SNPs and indels with a MAF greater than 0.1% per sample had slightly positive correlations with genomic coverage (R^2 = 0.075 and 0.11, respectively) (Figures S2C and S2D), indicating that WGS quality can still be improved by increasing the sequencing depth beyond 30×, especially for indels. This could be explained by the fact that, although there is sufficient coverage for the whole genome, there are still regions that lack coverage randomly or are difficult to amplify, which will be improved when the sequencing depth increases. The median number of SNPs and indels with a MAF less than 0.1% in a genome were 26,400 (0.75%) and 2,570 (0.49%), respectively. The very rare SNPs and indels showed no positive correlation with sequencing depth (Figures S2E and S2F), probably because the number of rare variants varies more widely (approximately ±10%) in different samples than the number of variants with a

Table 1. Numbers of variants in the NyuWa resource and reference panel

	All variants[a]		Reference panel[b]	
Type	Total	Novel[c]	Total	Specific[d]
All	79,226,351	25,014,646	19,256,267	3,246,071
Nonsynonymous	500,966	149,343	73,260	7,048
Nonsynonymous deleterious	315,016	101,407	33,526	3,323
PTV	18,711	9,994	1,381	334
lncRNA splicing	3,793	1,499	743	80

[a]Variants in the NyuWa resource.
[b]Variants in the NyuWa reference panel.
[c]The novelty of variants was determined by comparison with dbSNP, 1KGP3, gnomAD v2.1, EXAC, ESP, gnomAD v3, TOPMed, 90 Han, and GAsP.
[d]Variants included in the other 4 publicly available haplotype reference panels (1KGP3, HRC.r1.1, GAsP, and TOPMed) were excluded.

MAF greater than 0.1% (approximately ±1%), and the positive correlation is obscured by the large fluctuation.

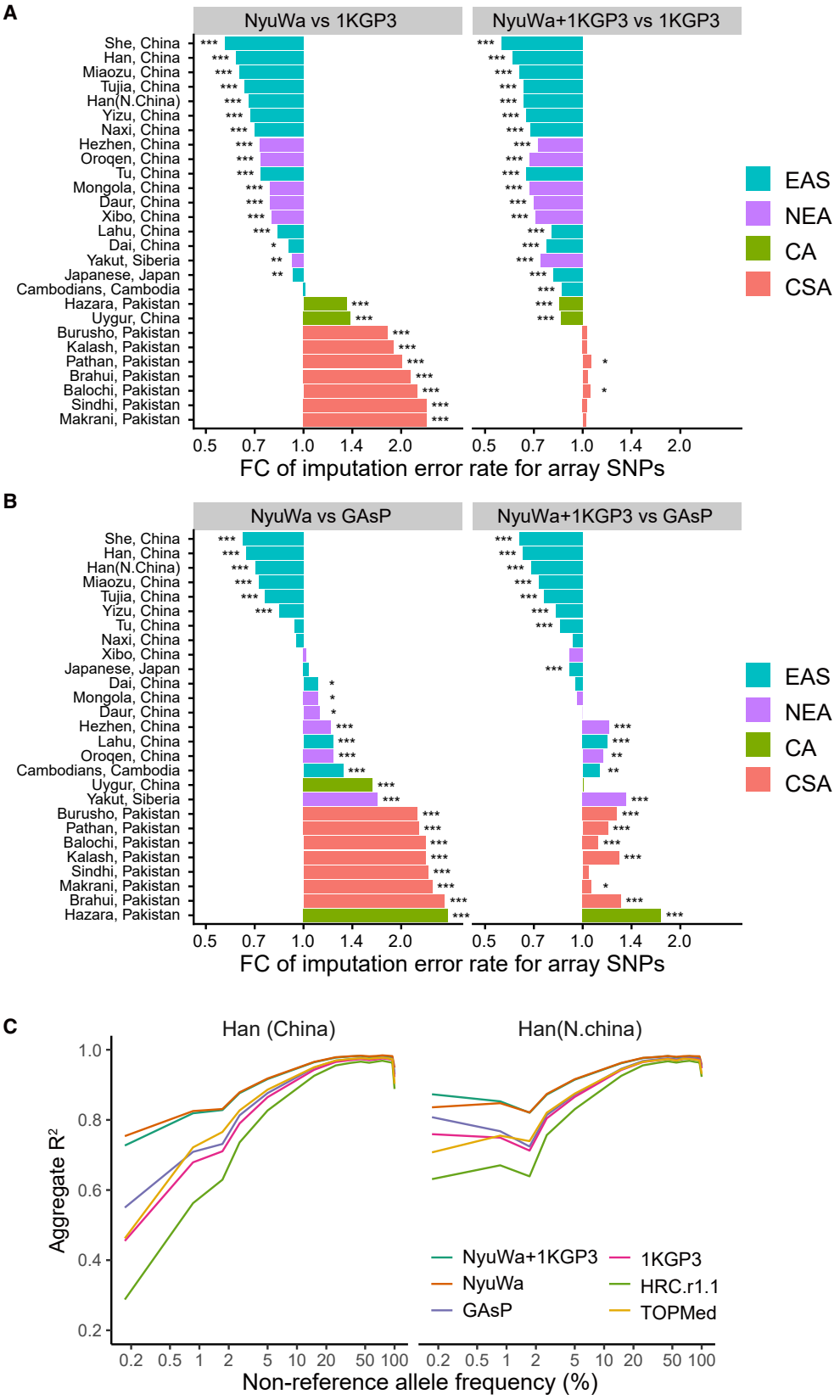
To evaluate the effect of increasing sample size on variant discovery, we randomly downsampled the NyuWa dataset to different sizes and estimated the total number and variant increase at different sample sizes (Figures S2G–S2J). We found that the numbers of SNPs and indels continued to increase with increasing sample size (Figures S2G and S2H), but the growth rate decreased, from an initial average increase of 39,400 and 5,700 per sample to a final average ~13,000 and ~1,000 for SNPs and indels, respectively (Figures S2I and S2J).

There were a total of 31.9 million variants in protein coding genes, including ~857,000 coding sequence (CDS), 1.10 million untranslated region (UTR), ~8,600 splicing, and 30 million intron variants (Figure 2B; Figure S2K; Table S3). For lncRNAs, variants were also annotated with NONCODE v5 (Fang et al., 2018), which has the largest collection of lncRNAs. There were a total of 4.78 million variants in lncRNA exon regions (Figure 2C; Table S4). Focusing on variants in protein-coding exons, ~315,000 of ~501,000 nonsynonymous SNPs were annotated as deleterious by at least two of ten selected prediction algorithms provided by dbNSFP (Liu et al., 2016; Figure 2D). The numbers of novel nonsynonymous and deleterious SNPs were ~149,000 and ~101,000, respectively (Table 1). Other functional protein-coding variants included ~311,000 synonymous SNPs, ~15,300 frameshift indels, ~12,700 non-frameshift indels, ~11,900 stop gains, and 613 stop losses (Table S5). There are more in-frame indels than adjacent frameshift indels in the coding region (Figure S2L), consistent with a previous report (Lek et al., 2016).

We designed a companion database (http://bigdata.ibp.ac.cn/NyuWa_variants/) to archive SNPs and indels in the NyuWa resource and to comprehensively catalog the variants on allele frequencies in our Chinese dataset and external datasets, including 1KGP3 and gnomAD v3. In addition, variant quality metrics, genome region annotations, nonsynonymous variant impact predictions, loss-of-function predictions, clinical annotations, and pharmacogenomic annotations were collected and presented.

The NyuWa reference panel outperformed other publicly available panels for Chinese populations

Genome-wide genotype imputation is a statistical technique to infer missing genotypes from known haplotype information; this technique is more cost-effective for GWAS with SNP arrays than whole-exome sequencing (WES) or WGS. The NyuWa haplotype reference panel (http://bigdata.ibp.ac.cn/refpanel/) was constructed using NyuWa phasing and the reference panel construction pipeline (STAR Methods). The NyuWa panel used 19.3 million SNPs and indels with minor allele count of 5 or greater (MAC5; approximately equivalent to MAF > 0.1%) in 2,902 independent samples, including 73,300 nonsynonymous and 33,500 deleterious SNPs (Table 1). Compared with 4 other publicly available reference panels, including 1KGP3, Haplotype Reference Consortium release 1.1 (HRC.r1.1) (McCarthy et al., 2016), GAsP, and TOPMed r2, the NyuWa reference panel had 3.25 million specific variants not included in other panels, including 7,050 nonsynonymous and 3,320 deleterious SNPs (Table 1). These NyuWa-specific variants may bring new discoveries in future association studies. To evaluate the imputation performance, array genotyping data, and high coverage WGS data for 54 worldwide populations from the Human Genome Diversity Project (HGDP) (Bergström et al., 2020; Li et al., 2008) were used as a test dataset. We focused on 16 Chinese populations and 11 other Asian populations in the HGDP. NyuWa outperformed 1KGP3, HRC.r1.1, and TOPMed r2 in all Chinese populations except the Uygurs (Figure 3A; Figures S3A and S3B). This can be explained by the fact that the Uygurs mainly inhabit Central Asia and were seldom included in our sampled areas. For the Han Chinese population, imputation with NyuWa reduced the error rates by 38.1%, 50.8%, and 30.4% compared with 1KGP3, HRC.r1.1, and TOPMed r2, respectively. NyuWa also achieved superior performance in most other East Asian and Northeast Asian populations (Figure 3A; Figures S3A–S3D). Not surprisingly, NyuWa did not perform as well as 1KGP3 in Central/South Asian populations in HGDP, which are mainly from Pakistan and historically received substantial gene flow from Central Asia and western Eurasia (Majumder, 2010; Qamar et al., 2002). Compared with GAsP, a newly released reference panel for Asian populations, NyuWa also has advantages in several Chinese populations, including the Han, She, Tujia, Miaozu, Yizu, Tu, and Naxi (Figure 3B; Figure S3C). For the Han Chinese population, imputation with NyuWa reduced the error rate by 33.2% compared with GAsP. Nevertheless, NyuWa performed worse in some Chinese minorities and Pakistani Central/South Asian populations, possibly because the Han population makes up an overwhelming majority of subjects in NyuWa. These results indicate that additional minority samples



(legend on next page)

are needed to improve the imputation performance for certain Chinese minorities. Imputation error rates for all other non-Chinese populations are shown in Figures S3D–S3F. We further compared the aggregate R^2 between imputed dosages and true genotypes among panels at different allele frequencies. NyuWa had an absolute advantage over the other panels for the Chinese Han population in all allele frequency bins, with great improvement for low-frequency (allele frequency [AF] < 5%, R^2 > 0.91) and rare (AF < 0.5%; R^2 > 0.81) variants (Figure 3C). NyuWa also achieved the highest aggregate R^2 in some other Chinese populations including She, Miaozu, Tujia, Yizu, and Nanxi (Figure S3G). These results indicated the good overall imputation quality of the NyuWa panel.

To optimize imputation performance, we also combined the NyuWa reference panel with the 1KGP3 panel using the reciprocal imputation strategy (Huang et al., 2015). The combined panel (NyuWa + 1KGP3) included 5,406 samples and 40.2 million variants, which improved imputation in all other tested Asian populations (Figure 3A; Figure S3). The imputation accuracy was improved markedly by approximately 10% for the Mongolian, Dai, Daur, Xibo, Tu, Oroqen, and Uygur and outperformed GAsP in more Chinese minority populations (Figure 3B). For the Han, She, Miaozu, Tujia, Yizu, and Naxi populations, NyuWa combined with 1KGP3 had almost the same aggregate R^2 as NyuWa or a slightly lower R^2 than NyuWa in the rare and low-frequency bins (Figure 3C; Figure S3G). For some other Chinese populations, such as the Mongolian, Dai, Daur, Xibo, Tu, Oroqen, and Uygur, NyuWa+1KGP3 had an improved R^2 compared with NyuWa, which was consistent with the error rate results. In brief, NyuWa+1KGP3 is an excellent alternative to NyuWa.

Applicability of one integrated reference panel for northern and southern Chinese

In light of genetic differences between northern and southern Han Chinese people (Chiang et al., 2018; Xu et al., 2009), we wanted to determine whether it is adequate to use one integrated reference panel for the northern and southern Han populations. To do this, we analyzed the NyuWa dataset from the perspective of population structure and imputation simulation tests.

To verify the ethnic authenticity of NyuWa samples, principal-component analysis (PCA) was performed on 200 randomly selected NyuWa samples together with 1KGP3 samples; the results showed that NyuWa samples were clustered together with 1KGP3 Han Chinese samples (Figures S4A and S4B), indicating that NyuWa samples are truly Han Chinese samples and do not show a large batch effect. Y chromosome analysis of male samples in the NyuWa population showed that the O group, which is the dominant group in the Han Chinese population, accounted for the majority (77.5%) of Y chromosome haplogroups. The next most common groups were C (9.0%) and N (7.5%). The Y haplogroup distribution was consistent with a previous analysis of Chinese populations (Yan et al., 2014; Figure S5A). The distribution of Y haplogroups in different provinces is shown in Figure S5B.

We then analyzed ancestral components of NyuWa samples. Cross-validation of ADMIXTURE analysis for NyuWa with 1KGP3 East Asia samples showed that K = 3 best matched the structure of East Asian populations (Figure 4A; Figure S6). As in CHB (Han Chinese in Beijing, China) and CHS (southern Han Chinese) samples in 1KGP3, the most predominant component in NyuWa samples was ancestral component 1 (red). Regarding the sample origins, a clear difference between people in northern and southern provinces was that southern people had a higher proportion of ancestral component 3 (blue; Figure 4B), which was also the case between CHB and CHS samples in 1KGP3. Component 3 was also the major component for Dai (Chinese Dai in Xishuangbanna, China [CDX]) and Vietnamese (Kinh in Ho Chi Minh City, Vietnam [KHV]) people (Figures 4A and 4B). Component 2 (green) was the major component for Japanese (Japanese in Tokyo, Japan [JPT]) people and was uncommon in Chinese samples (Figures 4A and 4B).

The above ADMIXTURE results indicated that northern and southern Chinese share two major ancestral components and differ in the proportions, which is consistent with the history of migration and partial mixing within the past two to three millennia (Chen et al., 2009; Wen et al., 2004). Using PCA, we found that primary component 1 (PC1) of NyuWa samples represented the trend of north-south differentiation (Figure 4C), which is consistent with previous studies of the Han ethnicity and Chinese minorities (Cao et al., 2020; Chiang et al., 2018; Liu et al., 2018a). Other PCs did not show differentiation between the north and south (Figure S7A). Variants with high absolute weights in PC1 also showed high AF differences between ancestral components 1 and 3 (Figure S7B). F_{st}, another analysis for genetic differentiation between northern and southern NyuWa samples as defined by the classic geographical demarcation of the Qinling Mountains-Huaihe River, also showed that north-south differential variants also differed in ancestral components 1 and 3 (Figure S7C). These results are consistent with the partial mixing of ancestral components. Because northern and southern Chinese people share the same major ancestral components, we reason that one integrated reference panel is applicable to northern and southern Han Chinese.

To test this speculation, we divided samples from the NyuWa reference panel into northern and southern subsets based on sample positions on PC1, which represents differentiation

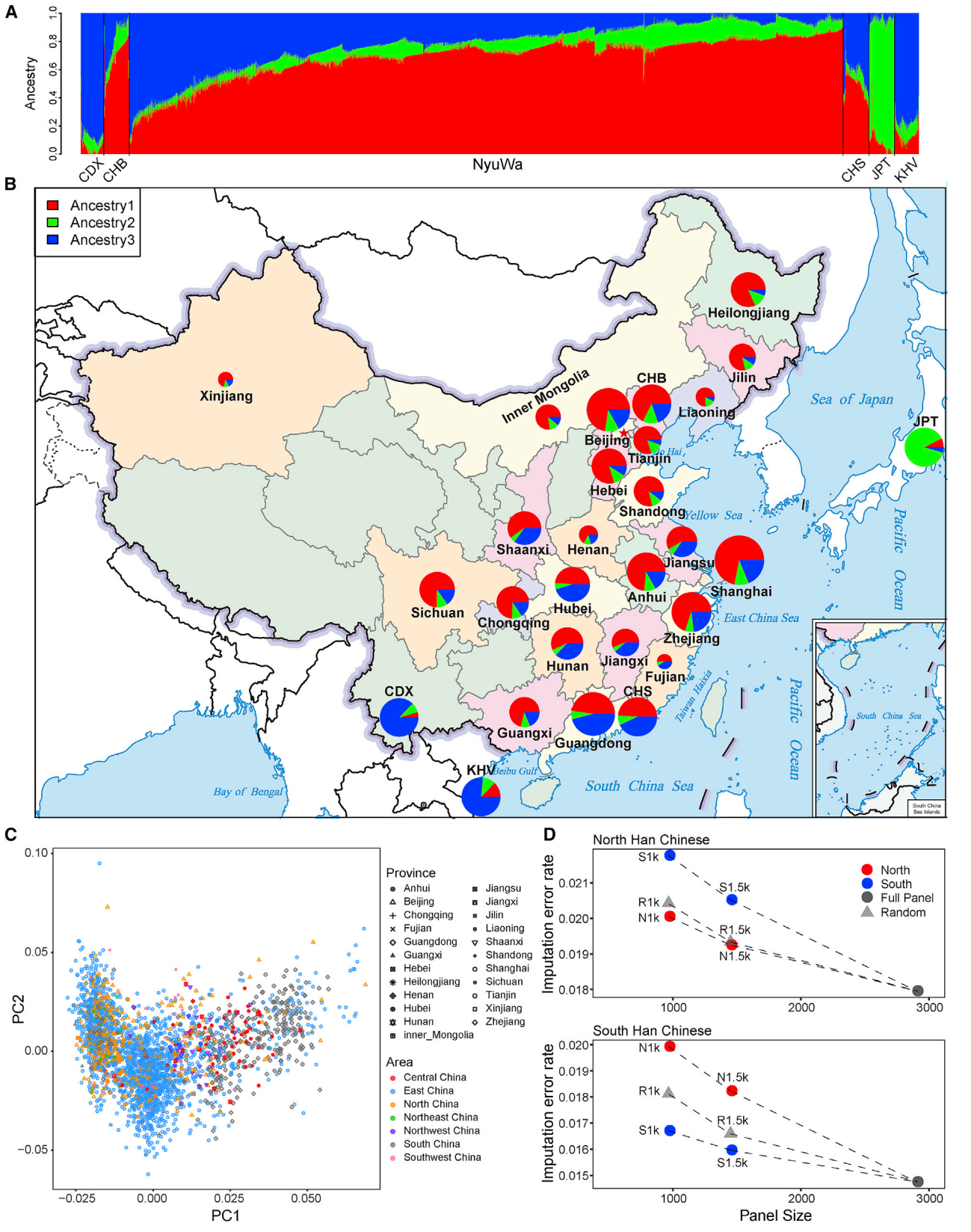
Figure 3. Performance of the NyuWa haplotype reference panel

(A) Fold change (FC) in the imputation error rate in different Asian populations in the HGDP array SNPs between the 1KGP3 panel and the NyuWa (left) or NyuWa+1KGP3 (right) panel. Lower FC values represent better performance with the NyuWa or NyuWa+1KGP3 panel. EAS, East Asian; NEA, Northeast Asian; CA, Central Asian; CSA, Central South Asian. The "Han, China" samples do not include "Han (N. China)" samples in HGDP. The significance of error rate differences was calculated by chi-square test. *p < 0.05, **p < 0.01, ***p < 0.001.

(B) FC in imputation error rate for array SNPs between the GAsP panel and the NyuWa (left) or NyuWa+1KGP3 (right) panel. Colors representing regions in (A) and (B) are consistent.

(C) The aggregate R^2 value between imputed dosage and known genotypes in stratified nonreference AF bins in high-coverage WGS data. Colors represent different reference panels. The "Han, China" and "Han (N. China)" samples are the same as in (A).

See also Figure S3.



(legend on next page)

between the north and the south (Figure 4C). Specific panels for northern and southern Han Chinese were then constructed using these sample subsets, and imputation error rates were compared on independent public datasets, including northern Han Chinese (Han North China in HGDP) and CHS (Chinese Han South in 1KGP3). As expected, given the same sample sizes, the regionally matched panels had lower imputation error rates than unmatched panels (Figure 4D). Panels with randomly selected samples had intermediate error rates. Increasing panel sizes always reduced error rates, regardless of whether the added samples were matched (Figure 4D; Figure S8A). The integrated panel always had the lowest error rates. The imputation results for CHB samples in 1KGP3 also showed lower error rates for panels with larger sizes (Figure S8B), whereas the differences between the northern and southern panels were not obvious, probably because there are also many southern samples in CHB (Figure S4B). Another classification method using the Qinling Mountains-Huaihe River geographical demarcation showed similar results (Figures S8C and S8D). These results confirmed the applicability of one integrated panel for northern and southern Chinese subjects.

We also explored whether there was a difference in the introgression level of Denisovan and Neanderthal ancestries between the northern and southern NyuWa populations (Figure S9). No obvious north-south difference was found, suggesting that the introgression of Denisovan and Neanderthal ancestries occurred before the split of northern and southern ancestral populations, which was far before the current mixing of the population. Additionally, we found no samples with high Denisovan ancestry (>3%) as observed in Melanesians and Aeta (Wall et al., 2019). The top 10 samples with the highest Denisovan ancestry were from Shanghai (5), Beijing (2), Guangdong (1), Shaanxi (1), and Xinjiang (1), with percentages ranging from 0.42%–0.45%.

Clinical annotations for variants

To demonstrate the value of the NyuWa resource in improving human health, we further evaluated the utility of NyuWa in genetic disease studies and medical applications. We annotated all variants with ClinVar (Landrum et al., 2018) and found 1,140 pathogenic variants (Figures S10A and S10B). As expected, most of the pathogenic variants were singletons or rare variants in the NyuWa and public datasets (Figure 5A). Each sample had a median of 4 homozygous pathogenic variants and 7 heterozygous pathogenic variants (Figure S10C). We noticed that there were 32 pathogenic variants with an AF greater than 1% (Figure 5A; Data S1). Pathogenic variants are usually rare, and pathogenic variants with high AFs may relate to common diseases, otherwise, their pathogenicity should be subjected to further examination. We also found some variants annotated with conflicting interpretations of pathogenicity by ClinVar that showed higher AFs specifically in the NyuWa resource (Figure 5B; Data S1). For example, with an AF of 1% as the threshold, two variants, rs182677317 and rs369849556, were annotated as conflicting for a rare disease, ciliary dyskinesia, whereas the high AFs (>1%) in the NyuWa dataset suggested that these variants may not be pathogenic (Figure 5C). These results showed that variant AFs in the NyuWa dataset can provide an additional reference to assist in the study of disease-related variants.

We also assessed the allele frequencies of known pharmacogenomic loci from ADME core genes (http://pharmaadme.org/) that may affect the efficacy and safety of drugs in different Chinese provinces and global regions (Data S2). We found some variants with obvious AF differences in different regions of China as well as in different populations worldwide (Figure 5D). For instance, isoniazid, a drug recommended by the World Health Organization (WHO) for treatment of tuberculosis (TB), is metabolized primarily by the enzyme NAT2 (N-acetyltransferase 2). *NAT2*12* refers to rs1208, and the reference allele (A) dampens enzyme activity (Vatsis et al., 1991). The homozygous reference genotype will cause drug accumulation and toxicity, whereas heterozygous and homozygous alternative genotypes have reduced side effects (Toure et al., 2016). We detected consistently high AFs (near 100%) of *NAT2*12* in different Chinese provinces and East Asians and lower frequencies in other populations (Figure 5D). This suggested that testing the *NAT2*12* genotype before using isoniazid is not as necessary for the Chinese population as for other populations. For other examples, the AFs were not close to 0% or 100% and varied among different Chinese provinces (Figure 5D); hence, genetic tests are recommended before certain drugs are used for individualized treatment.

We also examined cancer risk loci (Sud et al., 2017) in different regions (Data S2). It is generally recognized that there are racial differences in cancer susceptibility and survival, and genetic factors are very important determinants of cancer risk (Özdemir and Dotto, 2017). We also detected obvious AF differences between Chinese and other populations in many cancer susceptibility loci (Figure 5E).

Figure 4. Chinese population structure based on the NyuWa dataset

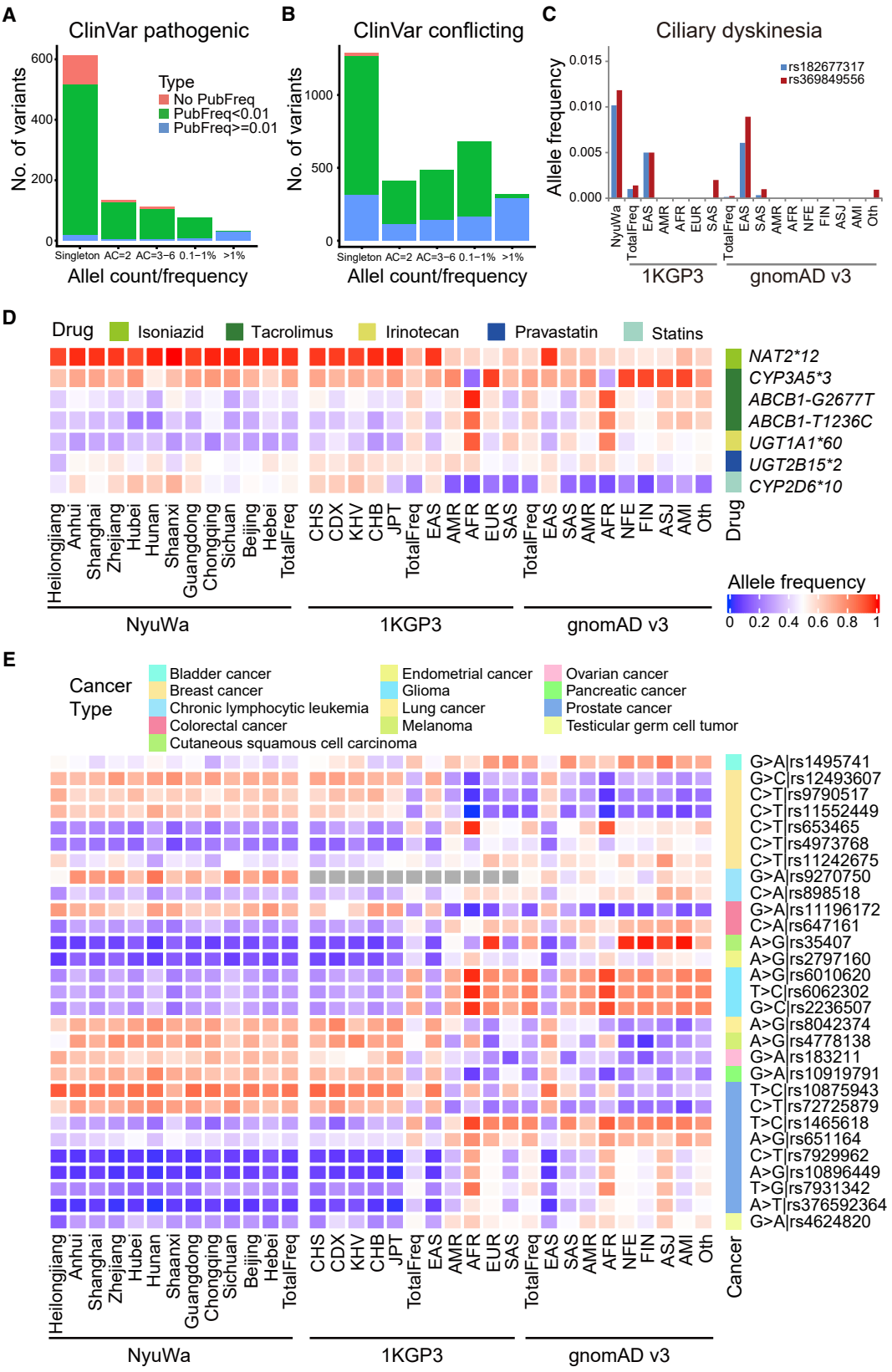
(A) ADMIXTURE analysis of NyuWa samples with East Asia samples in 1KGP3. An assumption of K = 3 ancestries best fits the model. Different colors represent different ancestry components. CHB, Han Chinese in Beijing, China; CHS, Southern Han Chinese; CDX, Chinese Dai in Xishuangbanna, China; JPT, Japanese in Tokyo, Japan; KHV, Kinh in Ho Chi Minh City, Vietnam.

(B) Proportions of ancestry components in different provinces. The ancestry components and colors are consistent with (A). 1KGP3 East Asia populations (CHB, CHS, CDX, JPT, and KHV) are also shown.

(C) Top 2 primary components (PC1 and PC2) of NyuWa samples. Each point represents a sample. Samples are marked with the provinces and areas of China. PC1 represents the difference between northern and southern Chinese.

(D) Imputation error rates of two test datasets representing northern (Han N. China in HGDP, top) and southern (CHS in 1KGP3, bottom) Han Chinese. Each point represents a reference panel constructed with a certain sample subset of the NyuWa reference panel. Red represents north (N)-specific panels from samples in the left part of PC1 shown in (C), and blue represents south (S)-specific panels in the right part of PC1. The gray triangles represent reference panels with randomly (R) selected samples. 1k and 5k, respectively, represent 1/3 and 1/2 of the 2,902 total samples in the NyuWa panel. Dotted lines represent addition of more samples.

See also Figures S4–S9 and Table S1.



(legend on next page)

Loss-of-function variants of protein-coding genes and lncRNA genes

Human loss-of-function variants have profound effects on gene function and are informative for clinical genome interpretation. In this study, we screened high-confidence loss-of-function PTVs, especially novel variants. We found 18,711 PTVs in 7,696 genes, of which most PTVs were singletons (Figures 6A and 6B), in line with PTV data from ExAC (67% singletons) (Lek et al., 2016). There were 9,994 novel PTVs found in the NyuWa dataset, and 1,381 PTVs could be imputed by the NyuWa reference panel (Table 1). The number of homozygous PTVs was 21 (Figure 6B; Figure S10D). There was a median of 24 homozygous PTVs and 58 heterozygous PTVs per sample (Figure S10E). We detected 1,138 PTVs in 385 of 906 cancer-related genes; 636 of these PTVs were novel. Focusing on 9 well-studied cancer-associated genes (*BRCA1*, *BRCA2*, *TP53*, *MEN1*, *MLH1*, *MSH2*, *MSH6*, *PMS1*, and *PMS2A*) (Wall et al., 2019), we identified 5 novel PTVs and 48 known PTVs in *BRCA2*, *BRCA1*, *PMS1*, *TP53*, and *MSH6* (Figure 6C). *BRCA1* and *BRCA2* are involved in maintenance of genome stability. Inherited mutations in *BRCA1* and *BRCA2* confer an increased lifetime risk of developing breast or ovarian cancer. There were 10 known PTVs in *BRCA1* and *BRCA2*, of which 9 have been annotated as pathogenic and related to breast and ovarian cancer in ClinVar (Landrum et al., 2018).

Because lncRNAs do not contain consensus CDS regions, splicing variants become the most important class for possible lncRNA loss-of-function variants. Splicing variants may cause intron retention or exon skipping and greatly change the lncRNA sequence and structure (Ulitsky et al., 2011). A total of 230 lncRNA genes have been reported to affect cell growth after CRISPR editing at lncRNA splicing sites (Liu et al., 2018b), suggesting the importance of lncRNA splicing variants for lncRNA functions. A total of 3,793 splicing variants in 3,544 lncRNA genes were found in the NyuWa dataset (Figure 6D), including 1,454 splicing variants in 1,287 Ensembl lncRNA genes and another 2,339 splicing variants in 2,257 NONCODE lncRNA genes (Figures S10F and S10G). Each sample had a median of 61 homozygous and 91 heterozygous lncRNA splicing variants (Figure S10H). Among 230 lncRNA genes reported to be essential for cell growth (Liu et al., 2018b), we found 22 splicing variants in 20 lncRNA genes. The proportion of AF > 0.1% lncRNA splicing variants was smaller in the 20 essential lncRNA genes than all lncRNA splicing variants (Figures 6E and 6F), suggesting that splicing variants can truly affect the function of these lncRNAs. In general, the loss-of-function variants for protein-coding and noncoding genes identified in the NyuWa dataset may be associated with disease etiology or trait tendency, which will provide novel insights into disease and genetic studies.

DISCUSSION

The Chinese population, which accounts for approximately 20% of the global human population, contains 56 ethnic groups and highly diverse disease types. Constructing a comprehensive genome resource platform of the Chinese population empowers medical genetics discoveries in the world's largest population and contributes to the diversity of worldwide human genetic resources. Here we present the NyuWa resource, consisting of large-cohort deep WGS data for the Chinese population. We also constructed a companion database to comprehensively catalog the variants. The 25 million novel variants identified in the NyuWa resource will greatly benefit studies of human diseases, especially in Chinese people. Although ChinaMAP has also published a resource for the Chinese population, variant data files were not available to download. By comparing manually selected variants, we estimated that ~18 million variants would remain novel after the exclusion of variants in ChinaMAP.

Another important contribution of this work is that the NyuWa resource can fill in the blanks of the WGS-based haplotype reference panel in the Chinese population. Previously, the most commonly used imputation panels were constructed by 1KGP3 and HRC. The recently released TOPMed reference panel included the largest number of haplotypes (Taliun et al., 2019) so far. However, the imputation performance of these panels for Chinese and East Asian populations is limited because East Asian samples are underrepresented. In addition, a large number of genome variants are specific to a population or sample, especially for rare variants, whose imputation can be challenging (Carmi et al., 2014). Our NyuWa reference panel contains 19.3 million variants (approximately MAF > 0.1%) with 3.25 million specific variants not included in other panels, and contains a large proportion of low-frequency alleles. The imputation performance of NyuWa exceeded that of 1KGP3, HRC, and TOPMed for the Chinese population (Figure 3A; Figures S3A and S3B). Furthermore, the combined reference panel of NyuWa and 1KGP3 outperformed 1KGP3, HRC, and TOPMed for nearly all Asians (Figure 3A; Figure S3). Compared with GAsP, a newly public Asian reference panel, NyuWa also has an advantage in Chinese populations, including the Han, She, Tujia, Miaozu, Yizu, Tu, and Naxi, and possesses higher accuracy across all AF bins.

We also found that the genetic differences between northern and southern Chinese are mainly the proportions of two major

Figure 5. Annotation of variants

(A) Allele count and frequency distribution for ClinVar pathogenic variants.

(B) Allele count and frequency distribution for ClinVar variants annotated as conflicting interpretations of pathogenicity.

(C) Allele frequencies of two variants in different repositories. The two variants were annotated by ClinVar as having conflicting interpretations of pathogenicity for ciliary dyskinesia. TotalFreq, the AF of all samples in the corresponding dataset; EAS, East Asian; AMR, American; AFR, African; EUR, European; SAS, South Asian; NFE, non-Finnish European; FIN, Finnish; ASJ, Ashkenazi Jewish; AMI, Amish; Oth, Other.

(D) Allele frequencies of known pharmacogenomic loci (row) that vary in different populations or regions (column). For the NyuWa dataset, only provinces with sample sizes of 20 or greater are shown.

(E) Allele frequencies of known cancer risk loci (rows) that vary in different populations or regions (columns). For the NyuWa dataset, only provinces with sample sizes of 20 or greater are shown. The AF color bar is consistent with (D).

See also Figure S10 and Data S1 and S2.

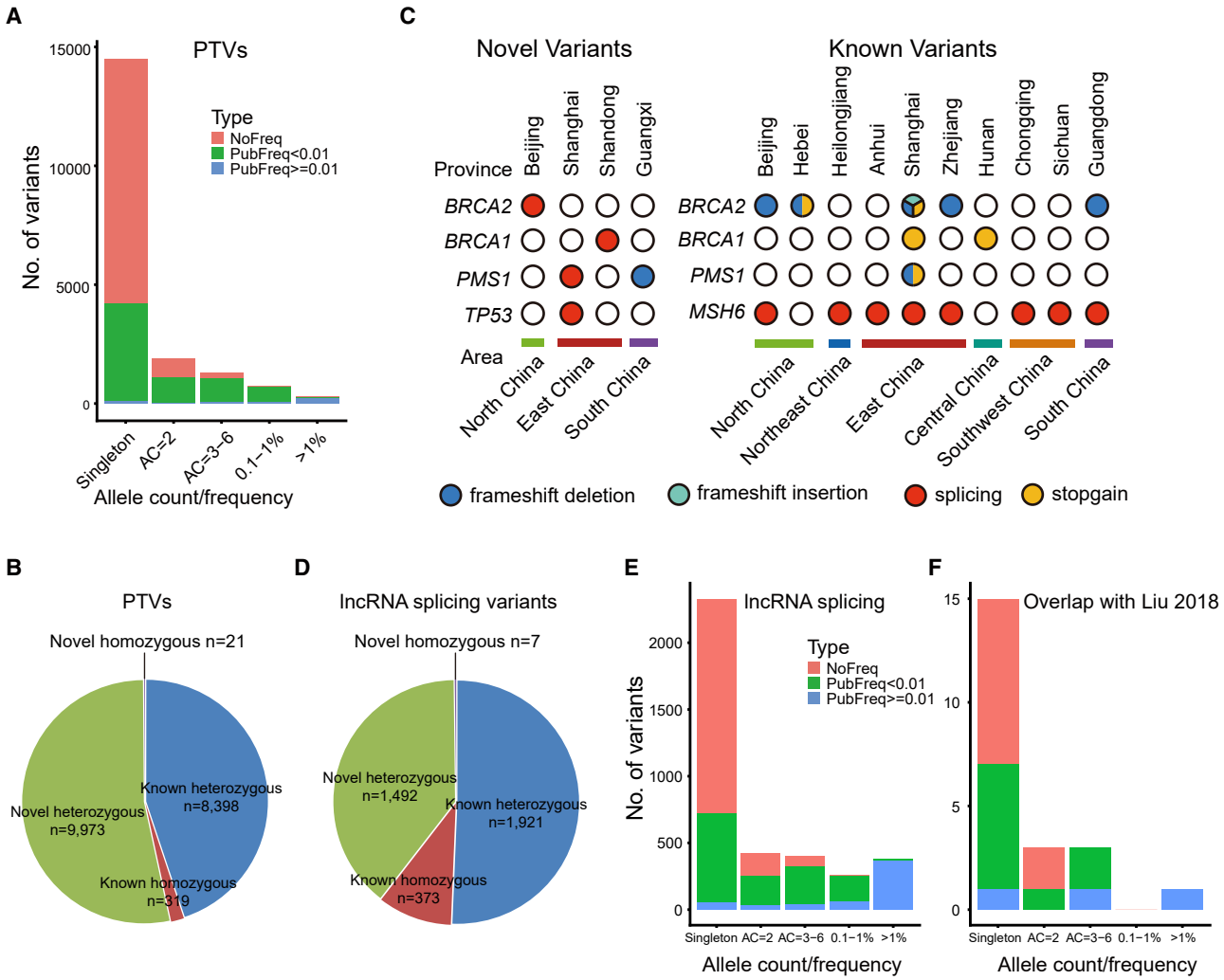


Figure 6. Predicted loss-of-function variants in the NyuWa dataset

(A) Allele count and frequency distribution of protein-truncating variants (PTVs).

(B) Numbers of PTVs classified as novel, known, heterozygous, and homozygous.

(C) Known and novel PTVs identified in selected cancer-associated genes in the NyuWa dataset.

(D) Numbers of lncRNA splicing variants classified as novel, known, heterozygous, and homozygous.

(E) Allele count and frequency distribution for lncRNA splicing variants.

(F) Allele count and frequency distribution for lncRNA splicing variants in 230 lncRNA genes reported to be essential for cell growth.

See also Figure S10.

ADMIXTURE components, suggesting that the north-south differences result mainly from partial population mixing in recent history. In the ADMIXTURE results, the main difference was the proportion of the northern Han-like component (ancestry 1, red) and southern Dai- or Vietnamese-like component (ancestry 3, blue) (Figures 4A and 4B). The northern samples have a very large proportion of component 1 and a small proportion of component 3, whereas component 3 is present in approximately half of the south samples. This population structure implies a partial mix of two ancestral components in the north and south, which is also consistent with the history of China. The earliest center of Chinese civilization was located in central to northern China, ranging from Henan to Shaanxi. Starting from the Eastern Zhou Dynasty, the Chinese territory expanded greatly, especially to the south. Then the foundation of a unitary multiethnic country beginning in the Qin and Han Dynasties facilitated mixing of the early Chinese population with southern ancestral populations. At present, the mix has still not achieved equilibrium.

An ideal reference panel for a population needs to cover all major ADMIXTURE components in the population. Each major component is required to have a sufficient and balanced sample size to cover most haplotypes in the component. As described above, northern and southern Han Chinese have the same two major components, although the proportions of these components are different. Therefore, a single reference panel that covers these major components can be used to impute northern

and southern Han populations. Imputation tests using northern or southern subset panels confirmed this speculation.

The current knowledge and guidelines on medical genomics are mainly from Eurocentric genetic and genomic resources and may be missing information about people of non-European ancestry. Our study provides a large and high-quality WGS resource for Chinese populations, which will be useful in examining the effect of known genetic variants on disease susceptibility and drug responses, and benefit clinical investigations in the future. The identification of loss-of-function variants for protein-coding and lncRNA genes in this study expands the catalog of loss-of-function variants in nature. When combined with phenotype information, this resource will provide important biological insights into gene functions.

Limitations of the study

Because of the lack of samples from certain Chinese minority populations, the performance of the NyuWa reference panel can still be improved by including more minority samples. Currently, ethnic information in the NyuWa resource is not available. Han Chinese are supposed to be the majority in NyuWa samples. The results of better performance using one integrated panel for both northern and southern Chinese are based on the current panel size. When a larger sample size has been accumulated, the specific situation will determine which panel works better.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - DNA extraction and library preparation
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - NyuWa cohort variant calling pipeline
  - Phasing and reference panel construction pipeline
  - Imputation performance
  - Population structure analysis
  - F_{st} between south and north of China
  - Denisovan and Neanderthal ancestry
  - Y chromosome analysis
  - PTVs and lncRNA loss-of-function variants

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2021.110017.

ACKNOWLEDGMENTS

We thank Weiwei Zhai for thoughtful discussions and valuable comments regarding the population structure analysis. This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB38040300 and XDA12030100), the National Natural Science Foundation of China (91940306, 31871294, 31970647, and 81902519, the National Key R&D Program of China (2017YFC0907503, 2016YFC0901002, and 2016YFC0901702), the 13th Five-year Informatization Plan of Chinese Academy of Sciences (XXH13505-05), and the National Genomics Data Center, China.

AUTHOR CONTRIBUTIONS

T.X. and S.H. conceptualized and supervised the project. P.Z., H.L., Y.L., J.W., Y.N., Q.K., Y.S., and H.Z. conducted analyses. Y.W. and T.X. contributed to sample collection and data generation. P.Z., H.L., Y.Z., Q.K., and T.S. made the web server and database. P.Z., H.L., Y.N., and S.H. drafted the manuscript, and all primary authors reviewed, edited, and approved the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 15, 2020
Revised: May 4, 2021
Accepted: October 28, 2021
Published: November 16, 2021

REFERENCES

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

Ardlie, K.G., DeLuca, D.S., Segre, A.V., Sullivan, T.J., Young, T.R., Gelfand, E.T., Trowbridge, C.A., Maller, J.B., Tukiainen, T., Lek, M., et al.; GTEx Consortium (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science *348*, 648–660.

Asimit, J.L., and Zeggini, E. (2012). Imputation of rare variants in next-generation association studies. Hum. Hered. *74*, 196–204.

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.; 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Bergström, A., McCarthy, S.A., Hui, R., Almarri, M.A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., et al. (2020). Insights into human genetic variation and population history from 929 diverse genomes. Science *367*, 1339.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Bomba, L., Walter, K., and Soranzo, N. (2017). The impact of rare and low-frequency genetic variants in common disease. Genome Biol. *18*, 77.

Cao, Y., Li, L., Xu, M., Feng, Z., Sun, X., Lu, J., Xu, Y., Du, P., Wang, T., Hu, R., et al.; ChinaMAP Consortium (2020). The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. Cell Res. *30*, 717–731.

Carmi, S., Hui, K.Y., Kochav, E., Liu, X., Xue, J., Grady, F., Guha, S., Upadhyay, K., Ben-Avraham, D., Mukherjee, S., et al. (2014). Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins. Nat. Commun. *5*, 4835.

Chang, C.C., Chow, C.C., Tellier, L.C.A.M., 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.

Chen, J., Zheng, H., Bei, J.X., Sun, L., Jia, W.H., Li, T., Zhang, F., Seielstad, M., Zeng, Y.X., Zhang, X., and Liu, J. (2009). Genetic structure of the Han Chinese population revealed by genome-wide SNP variation. Am. J. Hum. Genet. *85*, 775–785.

Chheda, H., Palta, P., Pirinen, M., McCarthy, S., Walter, K., Koskinen, S., Salomaa, V., Daly, M., Durbin, R., Palotie, A., et al. (2017). Whole-genome view of

the consequences of a population bottleneck using 2926 genome sequences from Finland and United Kingdom. Eur. J. Hum. Genet. *25*, 477–484.

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.

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.

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. Gigascience *10*, giab008.

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.

Delaneau, O., Zagury, J.F., Robinson, M.R., Marchini, J.L., and Dermitzakis, E.T. (2019). Accurate, scalable and integrative haplotype estimation. Nat. Commun. *10*, 5436.

Du, Z., Ma, L., Qu, H., Chen, W., Zhang, B., Lu, X., Zhai, W., Sheng, X., Sun, Y., Li, W., et al. (2019). Whole Genome Analyses of Chinese Population and De Novo Assembly of A Northern Han Genome. Genomics Proteomics Bioinformatics *17*, 229–247.

Edge, P., Bafna, V., and Bansal, V. (2017). HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. Genome Res. *27*, 801–812.

Fang, S., Zhang, L., Guo, J., Niu, Y., Wu, Y., Li, H., Zhao, L., Li, X., Teng, X., Sun, X., et al. (2018). NONCODEV5: a comprehensive annotation database for long non-coding RNAs. Nucleic Acids Res. *46* (*D1*), D308–D314.

Francioli, L.C., Menelaou, A., Pulit, S.L., Van Dijk, F., Palamara, P.F., Elbers, C.C., Neerincx, P.B.T., Ye, K., Guryev, V., Kloosterman, W.P., et al.; Genome of the Netherlands Consortium (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat. Genet. *46*, 818–825.

Gao, Y., Zhang, C., Yuan, L., Ling, Y., Wang, X., Liu, C., Pan, Y., Zhang, X., Ma, X., Wang, Y., et al.; Han100K Initiative (2020). PGG.Han: the Han Chinese genome database and analysis platform. Nucleic Acids Res. *48* (*D1*), D971–D976.

Hoffmann, T.J., and Witte, J.S. (2015). Strategies for Imputing and Analyzing Rare Variants in Association Studies. Trends Genet. *31*, 556–563.

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.

Huang, K.L., Mashl, R.J., Wu, Y., Ritter, D.I., Wang, J., Oh, C., Paczkowska, M., Reynolds, S., Wyczalkowski, M.A., Oak, N., et al.; Cancer Genome Atlas Research Network (2018). Pathogenic Germline Variants in 10,389 Adult Cancers. Cell *173*, 355–370.e14.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931–945.

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al.; Genome Aggregation Database Consortium (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol. Biol. Evol. *35*, 1547–1549.

Lan, T., Lin, H., Zhu, W., Laurent, T.C.A.M., Yang, M., Liu, X., Wang, J., Wang, J., Yang, H., Xu, X., and Guo, X. (2017). Deep whole-genome sequencing of 90 Han Chinese genomes. Gigascience *6*, 1–7.

Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. *46* (*D1*), D1062–D1067.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al.; Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595.

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.

Lin, J.C., Fan, C.T., Liao, C.C., and Chen, Y.S. (2018). Taiwan Biobank: making cross-database convergence possible in the Big Data era. Gigascience *7*, 1–4.

Liu, X., Wu, C., Li, C., and Boerwinkle, E. (2016). dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Hum. Mutat. *37*, 235–241.

Liu, S., Huang, S., Chen, F., Zhao, L., Yuan, Y., Francis, S.S., Fang, L., Li, Z., Lin, L., Liu, R., et al. (2018a). Genomic Analyses from Non-invasive Prenatal Testing Reveal Genetic Associations, Patterns of Viral Infections, and Chinese Population History. Cell *175*, 347–359.e14.

Liu, Y., Cao, Z., Wang, Y., Guo, Y., Xu, P., Yuan, P., Liu, Z., He, Y., and Wei, W. (2018b). Genome-wide screening for functional long noncoding RNAs in human cells by Cas9 targeting of splice sites. Nat. Biotechnol, 30395134 Published online November 5, 2018. https://doi.org/10.1038/nbt.4283.

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.

Maher, M.C., Uricchio, L.H., Torgerson, D.G., and Hernandez, R.D. (2012). Population genetics of rare variants and complex diseases. Hum. Hered. *74*, 118–128.

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.

Maretty, L., Jensen, J.M., Petersen, B., Sibbesen, J.A., Liu, S., Villesen, P., Skov, L., Belling, K., Theil Have, C., Izarzugaza, J.M.G., et al. (2017). Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. Nature *548*, 87–91.

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.

Meyer, M., Kircher, M., Gansauge, M.T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prüfer, K., de Filippo, C., et al. (2012). A high-coverage genome sequence from an archaic Denisovan individual. Science *338*, 222–226.

Mirabello, L., Zhu, B., Koster, R., Karlins, E., Dean, M., Yeager, M., Gianferante, M., Spector, L.G., Morton, L.M., Karyadi, D., et al. (2020). Frequency of Pathogenic Germline Variants in Cancer-Susceptibility Genes in Patients With Osteosarcoma. JAMA Oncol. *6*, 724–734.

Nagasaki, M., Yasuda, J., Katsuoka, F., Nariai, N., Kojima, K., Kawai, Y., Yamaguchi-Kabata, Y., Yokozawa, J., Danjoh, I., Saito, S., et al.; ToMMo Japanese Reference Panel Project (2015). Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. Nat. Commun. *6*, 8018.

Okonechnikov, K., Conesa, A., and García-Alcalde, F. (2016). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics *32*, 292–294.

Özdemir, B.C., and Dotto, G.P. (2017). Racial Differences in Cancer Susceptibility and Survival: More Than the Color of the Skin? Trends Cancer *3*, 181–197.

Piton, A., Redin, C., and Mandel, J.L. (2013). XLID-Causing Mutations and Associated Genes Challenged in Light of Data From Large-Scale Human Exome Sequencing (vol 93, pg 368, 2013). Am. J. Hum. Genet. *93*, 406, 406.

Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Van der Auwera, G.A., Kling, D.E., Gauthier, L.D., Levy-Moonshine, A., Roazen, D., et al. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. https://doi.org/10.1101/201178.

Poznik, G.D. (2016). Identifying Y-chromosome haplogroups in arbitrarily large samples of sequenced or genotyped men. bioRxiv. https://doi.org/10.1101/088716.

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.

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., de Filippo, C., et al. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. Nature *505*, 43–49.

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.

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.

Saint Pierre, A., and Génin, E. (2014). How important are rare variants in common disease? Brief. Funct. Genomics *13*, 353–361.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311.

Sud, A., Kinnersley, B., and Houlston, R.S. (2017). Genome-wide association studies of cancer: current insights and future perspectives. Nat. Rev. Cancer *17*, 692–704.

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.

Tang, H., Choudhry, S., Mei, R., Morgan, M., Rodriguez-Cintron, W., Burchard, E.G., and Risch, N.J. (2008). Long-range LD can confound genome scans in admixed populations - Response to Price et al. Am. J. Hum. Genet. *83*, 135–139.

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.

Toure, A., Cabral, M., Niang, A., Diop, C., Garat, A., Humbert, L., Fall, M., Diouf, A., Broly, F., Lhermitte, M., and Allorge, D. (2016). Prevention of isoniazid toxicity by NAT2 genotyping in Senegalese tuberculosis patients. Toxicol. Rep. *3*, 826–831.

Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H., and Bartel, D.P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell *147*, 1537–1550.

van Leeuwen, E.M., Karssen, L.C., Deelen, J., Isaacs, A., Medina-Gomez, C., Mbarek, H., Kanterakis, A., Trompet, S., Postmus, I., Verweij, N., et al.; Genome of The Netherlands Consortium (2015). Genome of The Netherlands population-specific imputations identify an ABCA6 variant associated with cholesterol levels. Nat. Commun. *6*, 6065.

Vatsis, K.P., Martell, K.J., and Weber, W.W. (1991). Diverse point mutations in the human gene for polymorphic N-acetyltransferase. Proc. Natl. Acad. Sci. USA *88*, 6333–6337.

Wall, J.D., Stawiski, E.W., Ratan, A., Kim, H.L., Kim, C., Gupta, R., Suryamohan, K., Gusareva, E.S., Purbojati, R.W., Bhangale, T., et al.; GenomeAsia100K Consortium (2019). The GenomeAsia 100K Project enables genetic discoveries across Asia. Nature *576*, 106–111.

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

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.

Weir, B.S., and Cockerham, C.C. (1984). Estimating F-Statistics for the Analysis of Population Structure. Evolution *38*, 1358–1370.

Wen, B., Li, H., Lu, D., Song, X., Zhang, F., He, Y., Li, F., Gao, Y., Mao, X., Zhang, L., et al. (2004). Genetic evidence supports demic diffusion of Han culture. Nature *431*, 302–305.

Wu, D., Dou, J., Chai, X., Bellis, C., Wilm, A., Shih, C.C., Soon, W.W.J., Bertin, N., Lin, C.B., Khor, C.C., et al.; SG10K Consortium (2019). Large-Scale Whole-Genome Sequencing of Three Diverse Asian Populations in Singapore. Cell *179*, 736–749.e15.

Xu, S., Yin, X., Li, S., Jin, W., Lou, H., Yang, L., Gong, X., Wang, H., Shen, Y., Pan, X., et al. (2009). Genomic dissection of population substructure of Han Chinese and its implication in association studies. Am. J. Hum. Genet. *85*, 762–774.

Yan, S., Wang, C.C., Zheng, H.X., Wang, W., Qin, Z.D., Wei, L.H., Wang, Y., Pan, X.D., Fu, W.Q., He, Y.G., et al. (2014). Y chromosomes of 40% Chinese descend from three Neolithic super-grandfathers. PLoS ONE *9*, e105691.

Zhang, F., Flickinger, M., Taliun, S.A.G., Abecasis, G.R., Scott, L.J., McCaroll, S.A., Pato, C.N., Boehnke, M., and Kang, H.M.; InPSYght Psychiatric Genetics Consortium (2020). Ancestry-agnostic estimation of DNA sample contamination from sequence reads. Genome Res. *30*, 185–194.

Zhao, H., Sun, Z., Wang, J., Huang, H., Kocher, J.P., and Wang, L. (2014). CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinformatics *30*, 1006–1007.

Zook, J.M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W., and Salit, M. (2014). Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat. Biotechnol. *32*, 246–251.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Human reference genome (hg38)	GATK resource bundle (Poplin et al., 2017)	https://console.cloud.google.com/storage/browser/genomics-public-data/resources/broad/hg38/v0
1000 Genomes Project Phase 3	Auton et al., 2015	https://www.internationalgenome.org/
HGDP array	Li et al., 2008	http://hagsc.org/hgdp/
HGDP WGS	Bergström et al., 2020	ftp://ngs.sanger.ac.uk/production/hgdp
Neanderthal genome	Prüfer et al., 2014	http://cdna.eva.mpg.de/neandertal/altai/AltaiNeandertal/VCF/
Denisovan genome	Meyer et al., 2012	http://cdna.eva.mpg.de/neandertal/altai/Denisovan/
Human ancestral genome	Ensembl	ftp://ftp.ensembl.org/pub/release-99/fasta/ancestral_alleles/
Genome In A Bottle (GIAB) HG001 v.3.3.2	Zook et al., 2014	ftp://ftp-trace.ncbi.nlm.nih.gov/giab/
dbSNP v150	NCBI	https://ftp.ncbi.nih.gov/snp/
gnomAD v2 & v3	Lek et al., 2016	https://gnomad.broadinstitute.org/
TopMed r2	Taliun et al., 2019	https://imputation.biodatacatalyst.nhlbi.nih.gov/#!pages/home
HRC r1.1	McCarthy et al., 2016	https://imputation.sanger.ac.uk/
GAsP	Wall et al., 2019	https://browser.genomeasia100k.org/
GTEx v8	Ardlie et al., 2015	https://gtexportal.org/home/
NONCODE v5	Fang et al., 2018	http://noncode.org/
NyuWa imputation server	This manuscript	http://bigdata.ibp.ac.cn/refpanel/
NyuWa variant database	This manuscript	http://bigdata.ibp.ac.cn/NyuWa_variants/
NyuWa WGS	This manuscript	NODE: OEP002803
Software and algorithms		
GATK v3.7	(Poplin et al., 2017)	https://gatk.broadinstitute.org/hc
FastQC v0.11.3	Babraham Institute	https://www.bioinformatics.babraham.ac.uk/projects/fastqc
Trimmomatic v0.36	Bolger et al., 2014	http://www.usadellab.org/cms/index.php?page=trimmomatic
BWA-MEM v0.7.15	Li and Durbin, 2010	https://github.com/lh3/bwa
qualimap v2.1.2	Okonechnikov et al., 2016	http://qualimap.conesalab.org/
Picard v 2.9.2	Broad Institute	http://broadinstitute.github.io/picard/
verifyBamID2 v1.0.6	Zhang et al., 2020	https://github.com/Griffan/VerifyBamID
ANNOVAR v2018-04-16	Wang et al., 2010	https://annovar.openbioinformatics.org/en/latest/
LOFTEE v1.0.3	Karczewski et al., 2020	https://github.com/konradjk/loftee
HAPCUT2 v1.0	Edge et al., 2017	https://github.com/vibansal/HapCUT2
SHAPEIT4 v4.1.2	Delaneau et al., 2019	https://odelaneau.github.io/shapeit4/
Minimac3 & 4	Das et al., 2016	https://genome.sph.umich.edu/wiki/Minimac4
Eagle2 v2.4.1	Loh et al., 2016	https://alkesgroup.broadinstitute.org/Eagle/
Plink v2.00	Chang et al., 2015	https://www.cog-genomics.org/plink/2.0/
Bcftools v1.10.2	Danecek et al., 2021	https://samtools.github.io/bcftools/bcftools.html
ADMIXTURE v1.3.0	Alexander et al., 2009	https://dalexander.github.io/admixture/
VCFtools v0.1.15	Danecek et al., 2011	https://vcftools.github.io/
CrossMap v0.5.3	Zhao et al., 2014	http://crossmap.sourceforge.net/
yHaplo v1.0.21	Poznik, 2016	https://github.com/23andMe/yhaplo
FigTree v1.4.4	GitHub	https://github.com/rambaut/figtree

CellPress
OPEN ACCESS

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Shunmin He (heshunmin@ibp.ac.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The raw sequencing data derived from human samples have been deposited at NODE (http://www.biosino.org/node) with accession number: OEP002803. The access and use of the data shall comply with the regulations of the People's Republic of China on the administration of human genetic resources. To request access, contact Shunmin He (heshunmin@ibp.ac.cn). In addition, processed variants derived from these data have been deposited at http://bigdata.ibp.ac.cn/NyuWa_variants/ and are publicly available as of the date of publication.

This paper does not report original code.

Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Whole blood DNA samples of 3064 Chinese samples were collected including diabetes and control samples. This study was approved by the Medical Research Ethics Committee of Institute of Biophysics, Chinese Academy of Sciences. All participants provided written informed consent. The informed consent is used to collect samples for genome studies conducted by Chinese Academy of Sciences. The consent requires participants to be 30-70 years old patients and healthy people with full capacity. Participants voluntarily donate blood samples, provide clinical treatment information and sign informed consent. All their personal information is kept confidential. Participants can choose not to participate in sample donation, or withdraw at any time.

METHOD DETAILS

DNA extraction and library preparation

Genomic DNA was extracted and sequenced by WuXi Apptec Co., Ltd. according to the standard protocols of Illumina on HiSeq X10 platform or NovaSeq 6000. The sequencing reads were paired-end 150 nt and the target depth is 30X. Sequencing quality was checked with FastQC v0.11.3 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc). Adaptor sequences and low quality bases were removed with Trimmomatic v0.36 (Bolger et al., 2014).

QUANTIFICATION AND STATISTICAL ANALYSIS

NyuWa cohort variant calling pipeline

The variant calling followed GATK (Poplin et al., 2017) Best Practices Workflows Germline short variant discovery (SNPs + Indels) joint genotyping cohort mode. In brief, the raw sequencing reads were mapped to human reference genome assembly 38 with BWA-MEM v0.7.15 (Li and Durbin, 2010). Picard (http://broadinstitute.github.io/picard/) was used to sort bam and mark duplicates. Mapping quality was check by qualimap v2.1.2 (Okonechnikov et al., 2016). Indels were realigned and bases were recalibrated with GATK v3.7. Variants were called for each sample using GATK HaplotypeCaller in 'GVCF' mode. GATK GenotypeGVCFs was then used to identify variants for all samples in the cohort. Then GATK VQSR was applied for SNPs and indels with truth sensitivity filter levels 99.7 and 99.0, respectively. Variants were then annotated with annovar v2018-04-16 (Wang et al., 2010).

Duplicate sequencing data for the same persons were removed. verifyBamID2 (Zhang et al., 2020) version 1.0.6 was used to check the contamination. Samples with contamination levels of alpha $\geq$ 0.05 were removed. The sex of each sample was inferred by two ways. Based on whole genome and chromosome coverage results reported by qualimap, the coverage of X and Y chromosomes were divided by the whole genome coverage. The relative coverage of (X, Y) of male is expected to be (0,5, 0.5), and that of female is expected to be (1, 0). The ploidy of non-PAR region of X chromosome were estimated by BCFtools v1.5 (https://samtools.github.io/bcftools/bcftools.html; Danecek et al., 2021) guess-ploidy module. Males are haploid while females are diploid.

To filter low quality sites, variants with VQSR not passed were removed. Additional filters were applied to further exclude low quality variants. Sites with genotype quality (GQ) < 10 in > 50% samples were removed. For Y chromosome, sites were removed if GQ < 10 in > 50% male samples, or GQ > = 10 in > 10% female samples. Sites with no ALT allele in GQ > = 10 samples were also removed. Variants were further filtered with a Hardy-Weinberg Equilibrium (HWE) p value < 10^{-6} in the direction of excessive heterozygosity or ExcessHet > 54.69 in the INFO column calculated by GATK. Multi-allele sites were split using BCFtools norm module.

Some analyses required removal of close relatives. The 3rd degree or closer relationships were identified with the combination of kinship coefficient (Φ) and probability of zero identity-by-descent (IBD) sharing (π_0) (Manichaikul et al., 2010) calculated by plink (Chang et al., 2015). The k-degree relationship was defined as $2^{-k-1.5} < \Phi < 2^{-k-0.5}$. For the 1st degree relationships, parent-offspring was defined as $\pi_0 < 0.1$ and full sibling if $\pi_0 > 0.1$. $\Phi > 2^{-1.5}$ represents monozygotic twin or sample replicates. Relationships more than 3rd degree were treated as unrelated. To determine the list of independent samples, subjects with more relatives were excluded with priority, and a maximum of 2,902 unrelated samples were kept.

Phasing and reference panel construction pipeline

Sequencing reads based haplotype phasing for each sample was carried out with HAPCUT2 (Edge et al., 2017). The local phased sets were then incorporated in population-based phasing of 2,999 samples using SHAPEIT4 (Delaneau et al., 2019) version 4.1.2 with parameter '–use-PS 0.0001'. The information from family trios or duos were converted to phasing scaffold data and used by SHAPEIT4 with '–scaffold' option. Sites with missing call rates greater than 10% were removed. Sites with minor allele count < 2 (MAC2) were also removed. There were no samples with missing call rate greater than 10%. No additional reference panel was used. Only chromosome 1-22 and X were phased, and each chromosome was phased separately. For X chromosome, the pseudo-autosomal regions (PARs) and non-PAR were divided and phased separately. For samples with haploid X chromosome in non-PAR regions (male), the heterozygous genotypes were converted to missing before phasing using SHAPEIT4.

The 2,902 independent samples were extracted from the above phasing results. Sites with minor allele count < 5 (MAC5) in the independent sample set were also removed. The final list included 2,902 samples and 19,256,267 variants. Phased genotypes were then converted to m3vcf. format as imputation reference file using Minimac3 (Das et al., 2016) v2.0.1. The hg38 version of 1KGP3 reference panel was generated similarly with MAC5 sites.

To further improve imputation performance, a combined panel of NyuWa with 1KGP3 panel was generated using the reciprocal imputation strategy (Huang et al., 2015). The missed variants in each panel were imputed with the other with Minimac4 (Das et al., 2016), and the results were combined to form a new panel with all samples and union of variants in NyuWa and 1KGP3 panel. The combined panel had 5,406 samples and 40,196,029 variants in total.

Imputation performance

The chromosome 2 of HGDP genotyping array data was used to test imputation error rates for NyuWa, 1KGP3, GAsP, HRC.r1.1, TOPMed and NyuWa+1KGP3 reference panels. Bi-allele SNPs that exist in all panels were selected. Then every 1 out of 10 of the selected SNPs were masked to evaluate the imputation accuracy. Phasing and imputation of GAsP HRC.r1.1 and TOPMed panels were run on respective web servers. Phasing and imputation of NyuWa, 1KGP3 and NyuWa+1KGP3 panels were run locally with Eagle2 (Loh et al., 2016) and Minimac4, respectively. Imputation error rates were computed for each population as the genotype discordance rates of the masked SNPs.

In addition, for Chinese samples in HGDP dataset, we compared Pearson's R^2 between the genotypes from high coverage WGS of HGDP samples (Bergström et al., 2020) and imputed dosages from the reference panels described above. Sites overlapping with all the compared panels were used, and variants with missing rate > 10% in HGDP WGS were excluded. The imputation accuracy is stratified by the non-reference allele frequencies (AFs) in NyuWa reference panel, and R^2 was calculated for all variants in each bin.

The imputation error rates of reference panels constructed with sample subsets of the NyuWa reference panel were evaluated the same way as NyuWa panel. The 1KGP3 CHS and CHB test samples were already phased, and every 1 out of 10 of the selected SNPs were masked to evaluate the imputation error rates. The samples in the North or South specific panels were divided based on ranks of sample positions on PC1 from PCA or geographical demarcation of Qinling Mountains-Huaihe River (Table S1).

Population structure analysis

NyuWa 2,902 independent samples and 1KGP3 data were merged by extracting overlapped bi-allelic autosomal SNPs. SNPs with missing rate of more than 10% or MAF less than 0.05 were excluded. Linkage equilibrium (LD) was removed by thinning the SNPs to no closer than 2kb using plink. Furthermore, 27 known long-range LD regions were removed according to previous studies (Price et al., 2008; Tang et al., 2008; Wu et al., 2019). The resulted dataset included 901,455 SNPs. The merged data were then used in principal component analysis (PCA) and ADMIXTURE by extracting samples of interest in each analysis. PCA was carried out using plink. ADMIXTURE were carried out using ADMIXTURE Version 1.3.0 (Alexander et al., 2009). For each K, the analysis was repeated 4 to 8 times with different seeds, and the one with the highest value of likelihood was chosen. For ADMIXTURE result display when K > 2, dimensions were reduced to 1-dimension by tSNE and samples were ordered by tSNE values.

F_{st} between south and north of China

SNP-level fixation index (F_{st}) between north and south of China was calculated using the Weir and Cockerham's estimator (Weir and Cockerham, 1984) integrated in VCFtools (Danecek et al., 2011). North and south of China were divided according to the classic demarcation of Qinling Mountains-Huaihe River (Table S1). Henan, Jiangsu, Anhui were excluded because the Huaihe River flows through these provinces. Shanghai was also excluded for the possibility that there may be too many individuals from other provinces.

Denisovan and Neanderthal ancestry

Estimation of Denisovan and Neanderthal ancestry followed methods in GAsP (Wall et al., 2019). In brief, Neanderthal and Denisovan genomes were downloaded from http://cdna.eva.mpg.de/neandertal/altai/AltaiNeandertal/VCF/ (Prüfer et al., 2014) and http://cdna.eva.mpg.de/neandertal/altai/Denisovan/ (Meyer et al., 2012). Human ancestral sequences were downloaded from ftp://ftp.ensembl.org/pub/release-99/fasta/ancestral_alleles/. Potential Neanderthal/Denisovan SNPs were filtered by the following criteria. 1. The REF allele matched the ancestral allele; 2. Neanderthal/Denisovan genotype was homozygous ALT allele; 3. Denisovan/Neanderthal genotype was homozygous REF allele; 4. ALT allele was not found in YRI, GWD, MSL or ESN samples in 1KGP3. Then, for each NyuWa sample, the number of Neanderthal/Denisovan SNP alleles were counted. To correct background, linear models were fit for both Neanderthal and Denisovan SNPs based on allele counts and ancestry percentage in GAsP results. Supposing SNPs called in NyuWa and GAsP were independent for Neanderthal/Denisovan SNPs, allele counts were scaled to make the median of NyuWa samples equal to the average of GAsP HAN samples. The ancestry proportion for each sample was then determined by the linear model using scaled allele count.

Y chromosome analysis

Genotypes of male chrY SNPs in NyuWa dataset were lift over to hg19 using CrossMap (Zhao et al., 2014). Y chromosomal haplogroups were inferred using yHaplo (https://github.com/23andMe/yhaplo; Poznik, 2016). Besides, file of primary tree structure (y.tree.primary.2016.01.04.nwk), file of preferred SNP names (preferred.snpNames.txt) and file of phylogenetically informative SNPs (isogg.2016.01.04.txt) were used.

MEGA X (Kumar et al., 2018) were used to construct a phylogenetic tree based on neighbor joining (NJ) method with 50 bootstrap. FigTree v1.4.4 (https://github.com/rambaut/figtree/releases) was used to color the tree and label main branches manually.

PTVs and lncRNA loss-of-function variants

PTV analysis followed methods in GAsP (Wall et al., 2019). In brief, stop gain, frameshift and splicing sites were selected according to ensGene annotation by annovar (Wang et al., 2010). Splicing variants are variants within 2-bp away from an exon/intron boundary that disrupt the GT-AG boundary pattern. Then multiple filters were applied. Variants out of Genome In A Bottle (GIAB) (Zook et al., 2014) high confidence regions were excluded. Stop gain or frameshift variants in the last exon or the last 50 nt in the second last exon were excluded. Variants in exons with non-classic splice sites were also removed. Splicing variants that locate in introns length < 15 nt or UTRs were excluded. Stop gain and splicing variants with phyloP100way vertebrate rankscore < 0.01 were excluded. Additional filters were applied to filter high quality PTVs. Only variants with GQ > = 20, DP > 7 and ALT DP > DP*0.2 were kept. Only variants affecting transcripts that within top 50% of gene expression in GTEx (Ardlie et al., 2015) were kept. A total of 9,526 PTVs in 4666 genes were obtained.

Loss-of-function variants were also predicted using LOFTEE v 1.0.3 (https://github.com/konradjk/loftee) (Karczewski et al., 2020). A total of 16,910 High confidence loss-of-function variants in canonical transcripts were identified. These variants covered most (7,725) of previously identified PTVs. The results were then combined to get the union set of PTVs.

For lncRNA splicing variants, Ensembl annotation was used first. Splicing variants were filtered similar to PTVs except that the phyloP100way conservation filter was not applied. The remaining splicing variants in NONCODE annotation were also filtered similarly, with GTEx expression replaced with expression data downloaded from NONCODE database.