NyuWa Genome Resource: Deep Whole Genome Sequencing Based Chinese Population Variation Profile and Reference Panel

Running title: NyuWa Chinese Population WGS and Reference Panel

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

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

2 College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China,

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

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

# These authors contributed equally to this work

* Correspondence: Shunmin He heshunmin@ibp.ac.cn or Tao Xu xutao@ibp.ac.cn

§ Full list of participants (collaborators) of the Han100K Initiative can be found online via http://www.pgghan.org/HCGD/about.

Key words: Whole genome sequencing, Chinese population, haplotype reference panel

Abstract

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

Introduction

Comprehensive catalogues of genetic variation are fundamental building blocks in studying population and demographic history, medical genetics and genotype-phenotype association. Since the first assembly of human genome released in 2003 (International Human Genome Sequencing 2004), many large-scale whole genome sequencing (WGS) projects have been launched in Western countries and recently in Asia, and have created large and diverse population genetic variation resources. Constructing haplotype reference panel from these large cohort WGS variation resources is meaningful and cost-effective to facilitate genome-wide association study (GWAS) by imputation of unobserved genotypes into samples that have been assayed using relatively sparse genome-wide microarray chips or low coverage sequencing (Asimit and Zeggini 2012; McCarthy et al. 2016). However, as the largest ethnic group in the world, the Chinese specific reference panel is absent.

A remarkable milestone of population genome project is the 1000 Genomes Project, which released an important resource of ~7.4X WGS data of 2,504 individuals from 26 populations, and constructed a reference panel (1KGP3) of 5,008 haplotypes at 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, thus enabling a deeper understanding of the genetic architecture of complex diseases (Timpson et al. 2018). Nevertheless, rare and low-frequency variants tend to be population- or sample-specific (Auton et al. 2015), and many disease related variants are very rare and population specific (Maher et al. 2012; Saint Pierre and Genin 2014; Bomba et al. 2017). The GWASs missed a proportion of potential trait-associated variants that were poorly imputed with current reference

panels (Asimit and Zeggini 2012; Hoffmann and Witte 2015; Bomba et al. 2017). So, a number of projects have focused on specific populations, attempting to capture the 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 large number of novel rare variants, and constructed a reference panel with 998 haplotypes (Francioli et al. 2014). In addition, based on the GoNL panel, researchers found a rare variant rs77542162 to be associated with blood lipid levels in Dutch population (van Leeuwen et al. 2015). Afterwards there were more such projects including UK10K in United Kingdom population (Walter et al. 2015), SISu in Finnish (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 Japanese (Nagasaki et al. 2015), 219 population groups across Asia by GenomeAsia 100K project (GAsP) (Wall et al. 2019), and three Singapore populations by SG10K project (Wu et al. 2019). Some studies have also focused on Chinese population, but the sample sizes (Lan et al. 2017; Du et al. 2019) or geographical coverage (Lin et al. 2017) were limited, or genotyping methods were mainly low coverage WGS (1.7X or 0.1X) (Liu et al. 2018a; Cai et al. 2020; Gao et al. 2020). In a most recent study, the China Metabolic Analytics Project (ChinaMAP) has presented deep WGS (40.8X) dataset of 10,588 Chinese individuals mainly involved in metabolic disease (Cao et al. 2020). However, the reference panel is not yet constructed in the study. The Han Chinese population comprises about 1.23 billion people, which is the largest ethnic group in East Asia and in the world, accounting for ~20% of the global human population and ~92% of Mainland Chinese (Xu et al. 2009b). Constructing an integrated, large cohort and high quality Han Chinese population

genetic variation database and reference panel is imperative, which will help to understand the population structure, population history, and facilitate genetics studies in the world's largest population.

Here we released the genome resource named NyuWa based on high depth (median 26.2X) WGS of 2,999 Chinese individuals from 23 out 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 71.1M single nucleotide polymorphisms (SNPs) and 8.2M small insertions or deletions (indels), of which 25.0M are novel. More importantly, we constructed the NyuWa reference panel of 5,804 haplotypes and 19.3M variants, which is the first publicly available Chinese population specific reference panel with thousands of samples, and has currently the best performance for imputation of Han Chinese. We also found 1,140 pathogenic variants, 18,711 loss-of-function protein truncating variants (PTVs) and 3,793 long non-coding RNA (lncRNA) splicing variants, of which 11,493 were novel compared with existing genome resources. In a word, 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 26.2X) whole-genome sequences (WGS) of 2,999 different Chinese samples including diabetes and control samples collected from hospitals or 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 (provinces for short, Figure 1A), which can be summarized into several geographical divisions of China (Supplementary information, 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 large incoming population from external provinces and rich sample diversity. The ethnicities for the samples were currently not available. Since national minorities are usually geographically clustered in China and not in our sampling areas, we estimated that the Han Chinese is the overwhelming majority in our samples.

Most of the samples were sequenced more than 30X (median 38.9, Supplementary information, Figure S1A). After genome alignment and removal of duplicates, the median of actual genomic coverage is 26.2X (Figure 1B; Supplementary information, Figure S1B). Samples with contamination levels alpha >= 0.05 were removed (Supplementary information, Figure S1C). Based on the genomic coverage of sex chromosomes, sample sex could be clearly identified except one potential XO type (Figure 1C). The ploidy of chrX for the sample also supported the XO type, which was classified as female. There were in total 1,335 females and 1,664 males in The NyuWa dataset. After identification of close relatives within 3rd degree (Supplementary information, Figure S1D), a maximum of 2,902 independent samples can be obtained in NyuWa dataset.

25.0M Novel Variants were Discovered in NyuWa Resource

SNPs and indels were called and genotyped using GATK cohort pipeline (Ryan Poplin 2017) with

human reference genome version GRCh38/hg38 as reference. After site quality filtering, a total of 76.4M variant sites were identified, including 2.5M multi-allele sites (Supplementary information, Figure S2A). After splitting of multi-allele sites, the final dataset contained 71.1M SNPs and 8.2M indels (Supplementary information, Figure S2B), including 2.5M SNPs and 0.3M indels from sex chromosomes (Supplementary information, Table S2). The transition-to-transversion ration (Ts/Tv) is 2.107 for all bi-allelic 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 to other public variant repositories including ExAC (Lek et al. 2016), gnomAD (v2 & v3) (Lek et al. 2016), 1KGP3 (Auton et al. 2015), ESP (NHLBI GO Exome Sequencing Project), dbSNP (v150) (Sherry et al. 2001), GAsP (Wall et al. 2019), 90 Han (Lan et al. 2017) and TOPMed (Taliun et al. 2019), the NyuWa dataset discovered 25.0M novel variants, including 23.1M SNPs (32.5%) and 1.9M indels (23.3%) (Figure 2A). The ChinaMAP resource (Cao et al. 2020) only provided website for variant search, but did not make 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 randomly selected from NyuWa dataset (10 per chromosome), and there were only 21.3% variants that also exist in ChinaMAP dataset. Another set was novel variants in 906 cancer related genes collected from ClinGen database and literature (Rehm et al. 2015; Huang et al. 2018; Mirabello et al. 2020). There were a total of 959k novel variants in these genes, and only 27.3% of these variants overlapping with ChinaMAP. We estimated that there were about 73% novel variants remain (~18M) compared with ChinaMAP. As expected, most novel variants were extremely rare, with singletons, doubletons and tripletons accounting for 86.8%, 10.1%, 1.9% of novel variants, respectively (Figure 2A). This is not

surprising since rare variants are usually sample- and population-specific (Francioli et al. 2014). The absolute number of novel variants with minor allele frequency (MAF) > 0.1% is still large (77.2k). These variants are frequent enough to be subject to large scale genetic association studies, and may bring new biological discoveries (Piton et al. 2013; Walter et al. 2015). The overall large number of novel variants indicates severe underrepresentation of variants in Chinese population in current genetic studies.

On average, a NyuWa sample carries a median number of 3.51M SNPs and 523k indels in autosomes. These numbers are close to East Asia samples in 1KGP3 (3.55M SNPs, 546k indels) (Auton et al. 2015). The detected SNPs and indels with MAF > 0.1% per sample had slightly positive correlation with genomic coverage (R^2 = 0.075 and 0.11, respectively) (Supplementary information, Figure S2C and S2D), indicating that the WGS quality can still be improved by increasing sequencing depth to higher than 30X, especially for indels. This could be explained by that although there is sufficient coverage for the whole genome, there are still regions lack coverage randomly or are difficult to amplify, which will be improved by increasing the sequencing depth. The median of MAF < 0.1% SNPs and indels in a genome were 26.4k (0.75%) and 2.57k (0.49%), respectively. The very rare SNPs and indels showed no positive correlation with sequencing depth (Supplementary information, Figure S2E and S2F), probably because the number of rare variants in different samples vary more largely (~±10 %) compared with MAF > 0.1% variants (~±1%), and the positive correlation is submerged by the large fluctuation.

To evaluate variant discovery by increasing sample size, we randomly down-sampled NyuWa dataset to different sizes with 100 samples intervals, and estimated the total variants and variant increase at different sample sizes (Supplementary information, Figure S2G-J). We found that the

number of both SNPs and indels continued to increase with the increasing sample size (Supplementary information, Figure S2G and S2H), but the growth rate decreased, from the initial average increase of 39.4k and 5.7k per sample to the final 13.0k and 1.0k for SNPs and indels, respectively (Supplementary information, Figure S2I and S2J).

There were a total of 31.9M variants in protein coding genes, including 857k CDS, 1.10M UTR, 8.60k splicing and 30.0M intron variants (Figure 2B; Supplementary information, Figure S2K and Table S3). For lncRNAs, variants were also annotated with NONCODE v5 (Fang et al. 2017), which has the largest collection of lncRNAs. There were in total 4.78M variants in lncRNA exon regions (Figure 2C; Supplementary information, Table S4). Focusing on variants in protein coding exons, among 501k non-synonymous SNPs, 315k were annotated as deleterious by at least two of ten selected prediction algorithms provided by dbNSFP (Liu et al. 2016) (Figure 2D). The number of novel non-synonymous and deleterious SNPs were 149k and 101k, respectively (Table 1). Other functional protein coding variants included 311k synonymous SNPs, 15.3k frameshift indels, 12.7k non-frameshift indels, 11.9k stop gains and 613 stop losses (Supplementary information, Table S5). Compared to adjacent frameshift indels, there are more in-frame indels in coding region (Supplementary information, Figure S2L), consistent with previous report (Lek et al. 2016).

We have designed a companion database (http://bigdata.ibp.ac.cn/NyuWa_variants/) to archive SNPs and indels in NyuWa resource, and to comprehensively catalogue the variants on allele frequencies in our Chinese dataset and external datasets including 1KGP3 and gnomAD v3. Besides, variant quality metrics, genome region annotations, non-synonymous impact prediction, loss-of-function prediction, clinical annotation and pharmacogenomics annotation are also

collected and presented.

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, which is cost-effective for GWAS with SNP arrays when compared with whole exome sequencing (WES) or WGS. NyuWa haplotype reference panel (http://bigdata.ibp.ac.cn/refpanel/) was constructed using 19.3M SNPs and indels with minor allele count >= 5 (MAC5, approximately MAF > 0.1%) in 2,902 independent samples, including 73.3k non-synonymous and 33.5k deleterious SNPs (Table 1). Compared with 4 other publicly available reference panels including 1KGP3 (Auton et al. 2015), Haplotype Reference Consortium release 1.1 (HRC.r1.1) (McCarthy et al. 2016), GAsP (Wall et al. 2019) and TOPMed r2 (Taliun et al. 2019), NyuWa reference panel had 3.25M specific variants not included in other panels, including 7.05k non-synonymous and 3.32k deleterious SNPs (Table 1). These NyuWa panel specific variants may bring new discoveries in future association studies. To evaluate the imputation performance, array genotyping data for 58 worldwide populations from the Human Genome Diversity Project (HGDP) (Li et al. 2008) were used as a testing dataset. We focused on 16 Chinese populations and 11 other Asian populations in HGDP. NyuWa outperformed 1KGP3, HRC.r1.1 and TOPMed r2 in all Chinese populations except Uygur (Figure 3A; Supplementary information, Figure S3A and S3B). This can be explained by that the Uygur population belongs to the Central Asia and was seldom included in our sampled areas. For Han Chinese, imputation with NyuWa reduced the error rate by 38.1%, 50.8% and 30.4% compared with 1KGP3, HRC.r1.1 and

TOPMed r2, respectively. NyuWa also achieved better performance in most other East Asian and Northeast Asian populations (Figure 3A; Supplementary information, Figure S3A-D). 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 (Qamar et al. 2002; Majumder 2010). Comparing to GAsP, a newly released reference panel for Asian populations, NyuWa also has advantage in Chinese populations including Han, She, Tujia, Miaozu, Yizu, Tu and Naxi (Figure 3B; Supplementary information, Figure S3C). For Han Chinese, imputation with NyuWa reduced the error rate by 33.2% compared with GAsP. Nevertheless, NyuWa performed worse in some of Chinese minorities and the Pakistan Central/South Asian populations, possibly due to the overwhelming proportion of Han population in NyuWa. These results indicated that additional minority samples were needed to improve the imputation performance of certain Chinese minorities. We further compared the number of high-quality imputed variants in total imputed variants among these panels. NyuWa showed the largest number and proportion of high-quality imputed variants (Rsq > 0.8) across all MAF bins in Chinese and Han Chinese population compared with GAsP and HRC.r1.1 (Figure 3C and 3D). Compared to 1KGP3, NyuWa had larger number of high-quality imputed variants in low MAF regions (Rsq > 0.8, MAF < 0.05), and similar number in high MAF regions (MAF > 0.05) (Figure 3C and 3D). While TOPMed r2 had slightly less numbers in high MAF regions and the largest numbers in low MAF regions, the percentage of high-quality imputed variants were very low in all MAF regions, due to the largest number of total variants (an order of magnitude higher than other 4 panels) (Figure 3C and 3D).

To optimize imputation performance, we also combined NyuWa reference panel with 1KGP3

panel using the reciprocal imputation strategy (Huang et al. 2015). The combined panel (NyuWa + 1KGP3) included 5,406 samples and 40.2M variants, which improved imputation in all tested Asian populations (Figure 3A; Supplementary information, Figure S3). The imputation accuracy was obviously improved by about 10% for Chinese minorities of Mongolian, Dai, Daur, Xibo, Tu, Oroqen and Uygur, and outperformed GAsP in more Chinese minority populations (Figure 3B). For Chinese and Chinese Han population, NyuWa+1KGP3 could impute more high-quality variants (Rsq > 0.8) across all MAF bins, with significant increase in low MAF variants (MAF < 0.01) (Figure 3C and 3D). In brief, NyuWa+1KGP3 is an excellent alternative of NyuWa.

Applicability of One Integrated Reference Panel for Both Northern and Southern Chinese

Due to the awareness of north-south genetic differences in Han Chinese people (Xu et al. 2009a; Chiang et al. 2018), we asked if it is adequate to use one integrated reference panel for both north and south Han populations. To answer this question, we analyzed NyuWa dataset from the perspective of population structure and imputation simulation tests.

In order to verify the ethnic authenticity of NyuWa samples, principal component analysis (PCA) of 200 randomly selected NyuWa samples together with 1KGP3 samples showed that NyuWa samples were clustered together with 1KGP3 Han Chinese samples (Supplementary information, Figure S4A and S4B), indicating that NyuWa samples are truly Chinese samples and little batch effect is observed. Y chromosome analysis of male samples in NyuWa population showed that majority (77.5%) of Y-chromosome haplogroups was O group, which is the dominant group in Han Chinese population. The following groups were C (9.0%) and N (7.5%). The Y

haplogroup distribution was consistent with previous analysis of Chinese populations (Yan et al. 2014) (Supplementary information, Figure S5A). The distribution of Y haplogroups in different provinces were shown in Supplementary information, 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 Asia populations (Figure 4A; Supplementary information, Figure S6). Consistent with CHB (Han Chinese in Beijing, China) and CHS (Southern Han Chinese) samples in 1KGP3, the most predominant component in NyuWa samples was the ancestral component 1 (red). In the view of sample origins, a clear difference between people in north and south provinces was that south people have more proportion of ancestral component 3 (blue, Figure 4B), which was also the case between CHB and CHS samples in 1KGP3. The component 3 was also the major ancestral component for Dai (CDX) and Vietnamese (KHV) people (Figure 4A and 4B). The component 2 (green) was the major ancestral component for Japanese (JPT) people, and was minor in Chinese samples (Figure 4A and 4B).

The above ADMIXTURE results indicated that north and south Chinese share two major ancestral components, and are different at the proportions of these components, which is consistent with the historical migration and partial mix within the past two to three millennia (Wen et al. 2004; Chen et al. 2009). Using primary component analysis (PCA), we found that the primary component 1 (PC1) of NyuWa samples represented the trend of north to south differentiation (Figure 4C), which is consistent with previous studies for Han and Chinese minorities (Chiang et al. 2018; Liu et al. 2018a; Cao et al. 2020). Other PCs does not show differentiation between north and south (Supplementary information, Figure S7A). We observed

that variants with high absolute weights in PC1 also showed high AF differences between ancestral components 1 and 3 (Supplementary information, Figure S7B). F_{st}, another analysis for genetic differentiation between north and south NyuWa samples classified with the classic geographical demarcation of Qinling Mountains-Huaihe River, also showed that north-to-south differential sites are also different between ancestral components 1 and 3 (Supplementary information, Figure S7C). These results showed that the genetic differences already existed between the ancestries 1 and 3, which is consistent with the partial mix of ancestry components. Collectively, since north and south Chinese share the same major ancestral components, we reason that one integrated reference panel is applicable for both north and south Han Chinese.

To test the speculation, we divided samples from NyuWa reference panel into different north or south subsets based on sample positions on PC1, which represents differentiation between north and south (Figure 4C). North/south Han Chinese specific panels were then constructed using these sample subsets, and imputation error rates were compared on independent public datasets including north Han Chinese (Han North China in HGDP) and south Han Chinese (CHS, Chinese Han South in 1KGP3). As expected, given the same sample sizes, the north or south matched specific 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, no matter added samples are matched or unmatched (Figure 4D; Supplementary information, Figure S8A). The error rates of the integrated panel were always the lowest. The imputation results for Han Chinese Beijing (CHB) samples in 1KGP3 also showed lower error rates for panels with larger sizes (Supplementary information, Figure S8B), while the differences between north and south panels were not obvious, probably because there are also many south

samples in CHB (Supplementary information, Figure S4B). Another classification way using geographical demarcation of Qinling Mountains-Huaihe River showed similar results (Supplementary information, Figure S8C&D). These results confirmed the applicability of one integrated panel for both northern and southern Chinese.

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

Clinical Annotations for Variants

To demonstrate the value of NyuWa resource in improving human health, we further evaluated the utility of NyuWa in disease genetic studies and medical applications. We annotated all the variants with ClinVar (Landrum et al. 2018), and found 1,140 pathogenic variants (Supplementary information, Figure S10A and S10B). As expected, most of the pathogenic variants were singletons or rare variants in NyuWa and public datasets (Figure 5A). Each sample had a median of 4 homozygous pathogenic variants and 7 heterozygous pathogenic variants (Supplementary information, Figure S10C). We noticed that there were 32 pathogenic variants with allele frequency (AF) > 1% (Figure 5A and Supplementary information, Data S1). Pathogenic variants

are usually rare, especially for rare diseases, and pathogenic variants with high AFs may relate to common diseases, or their pathogenicity are subject to further examinations. We also found some variants annotated as conflicting interpretations of pathogenicity by ClinVar showing specific higher AFs in NyuWa resource (Figure 5B and Supplementary information, Data S1). For example, taking AF 1% as threshold, two variants rs182677317 and rs369849556 were annotated as conflicting for a rare disease ciliary dyskinesia, while the high AFs (> 1%) in NyuWa dataset suggested these variants may not be pathogenic (Figure 5C). These results showed that variant AFs in NyuWa dataset can provide additional reference to assist 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 China provinces and worldwide regions (Supplementary information, 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 World Health Organization (WHO) in the treatment of tuberculosis (TB), is metabolized primarily by the NAT2 (N-Acetyltransferase 2) enzyme. *NAT2*12* refers to rs1208, and the reference allele A dampens the enzyme activity (Vatsis et al. 1991). The homozygous reference genotype will cause drug accumulation and poisoning, while heterozygous and homozygous alternative genotypes have less toxic side effects (Toure et al. 2016). We detected consistently high AFs (near 100%) of *NAT2*12* in different China provinces and East Asians, while relatively lower frequencies in other populations (Figure 5D). This suggested that testing the *NAT2*12* genotype before using isoniazid for Chinese is not as necessary as for other populations. For other examples, the AFs were not close to 0% or 100%, and vary in different China provinces (Figure 5D), hence it is recommended

to take genetic tests before using certain drugs for individualized treatment.

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

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. We first screened high confidence loss-of-function protein-truncating variants (PTVs), especially novel variants. We found 18,711 PTVs in 7,696 genes, in which most PTVs were singletons (Figure 6A and 6B), in line with PTV data from ExAC (67% singletons) (Lek et al. 2016). There were 9,994 novel PTVs found in NyuWa dataset, and 1,381 PTVs can be imputed by NyuWa reference panel (Table 1). The number of homozygous PTVs were 21 (Figure 6B, Supplementary information, Figure S10D). For each sample, there was a median of 24 homozygous PTVs and 58 heterozygous PTVs (Supplementary information, Figure S10E). We detected 1,138 PTVs in 385 of 906 cancer related genes, in which 636 are novel. Focusing on 9 well studied cancer-associated genes (*BRCA1*, *BRCA2*, *TP53*, *MEN1*, *MLH1*, *MSH2*, *MSH6*, *PMS1* and *PMS2A*) (Wall et al. 2019), there were 5 novel PTVs and 48 known ones in *BRCA2*, *BRCA1*, *PMS1*, *TP53* and *MSH6* (Figure 6C). Both *BRCA1* and *BRCA2* are involved in maintenance of genome stability, specifically the homologous recombination pathway for DNA double-strand break repair. Inherited mutations in *BRCA1* and *BRCA2* confer increased lifetime

risk of developing breast or ovarian cancer. There were 10 known PTVs in *BRCA1* and *BRCA2*, in which 9 have been annotated as pathogenic and related to breast-ovarian cancer in ClinVar (Landrum et al. 2018), and 1 has not been collected in dbSNP yet. The uncollected and novel PTVs in *BRCA1* and *BRCA2* may also increase the risk of breast and ovarian cancer.

Since lncRNAs do not contain consensus CDS regions, for possible lncRNA loss-of-function variants, splicing variants become the most important class. Splicing variants may cause intron retention or exon skipping, and greatly change the lncRNA sequence and structure (Ulitsky et al. 2011). 230 lncRNA genes were 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 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 (Supplementary information, Figure S10F and S10G). Each sample had a median of 61 homozygous and 91 heterozygous lncRNA splicing variants (Supplementary information, 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 were relatively smaller in the 20 essential lncRNA genes compared with all lncRNA splicing variants (Figure 6E and 6F), suggesting that splicing variants can really affect functions of these lncRNAs. In general, the loss-of-function variants for both protein-coding and non-coding genes identified in NyuWa dataset may be associated with disease causing or trait tendency, which will provide novel insights into disease and genetic studies.

Discussion

Chinese population accounts for about 20% of the global human population, with 56 ethnic groups and great diversity of disease types. Constructing a comprehensive genome resource platform of Chinese population empowers medical genetics discoveries in the world's largest population, and also contributes to the diversity of worldwide human genetic resource. Here we presented the NyuWa resource of large cohort deep WGS data for Chinese population, and constructed a companion database to comprehensively catalogue the variants. The 25.0M novel variants identified in NyuWa resource will greatly benefit studies of human diseases, especially for Chinese people. Although ChinaMAP has also published a resource of Chinese population, the variant data files were not available for downloading. By comparing of manually selected variants, we estimated ~18M variants would remain novel after subtracting variants in ChinaMAP.

Another important contribution of this work is that the NyuWa resource filled the blank of WGS based haplotype reference panel for Chinese population. Previously, the most commonly used imputation panels were constructed by the 1KGP3 (Auton et al. 2015) and HRC (McCarthy et al. 2016). The recently released TOPMed reference panel included the largest number of haplotypes (Taliun et al. 2019) so far. However, imputation performance of these panels for Chinese and East Asian populations are limited, as East Asian samples are underrepresented. In addition, large number of genome variants are population- or sample-specific, especially for rare variants, imputation of which can be challenging (Carmi et al. 2014). Our NyuWa reference panel contains 19.3M variants (approximately MAF > 0.1%) with 3.25M specific variants not included in other panels, which contains a large proportion of low frequency alleles. The imputation performance of NyuWa outperformed that of 1KGP3, HRC and TOPMed for Chinese population

(Figure 3A; Supplementary information, Figure S3A-B). Furthermore, the combined reference panel of NyuWa and 1KGP3 outperformed 1KGP3, HRC and TOPMed for nearly all the Asians (Figure 3A; Supplementary information, Figure S3). Compared to GAsP, a newly public Asian reference panel (Wall et al. 2019), NyuWa also has advantage in Chinese populations including Han, She, Tujia, Miaozu, Yizu, Tu and Naxi, and possesses larger number of high quality imputed variants (Rsq > 0.8) across all MAF bins. However, due to the lack of samples from certain Chinese minority populations, the performance of NyuWa reference panel can still be improved by including more minority samples in the future.

We also found that the genetic differences between north and south Chinese are consistent with differences between two major ADMIXTURE components, suggesting that the north-south differences mainly result from partial population mix in recent history. In the ADMIXTURE results, the difference was mainly the proportion of north Han like component (ancestry 1, red) and south Dai or Vietnamese like component (ancestry 3, blue) (Figure 4A and 4B). The north samples have very large proportion of component 1 and small component 3, while component 3 reaches to about a half in the south samples. This population structure implies a partial mix of two ancestral components of north and south, which is also consistent with the history of China. The earliest center of Chinese civilization located in the central to north of 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 unitary multiethnic country beginning at Qin and Han Dynasty facilitated the mix of early Chinese population with south ancestral populations. The mix has still not achieved equilibrium up to now. Despite the lack of native place identities for many samples in NyuWa resource, we could still detect a clear difference between north and south

Chinese samples, indicating that the hospitals collecting these samples were good approximations.

An ideal reference panel for a population needs to cover all major ADMIXTURE components in the population. Each major component is required to have sufficient and balanced sample size to cover most haplotypes in the component. As described above, both north and south Han Chinese consist of the same two major components, though the proportions of these components are different. So the same reference panel that cover these major components can be used to impute both north and south Han populations. Imputation tests using north or south subset panels confirmed the speculation. These results are based on the current sample size. In future when sample size is large enough, which panel works better still depends on the specific situation.

The current knowledge and guidelines about medical genomics are mainly from Eurocentric genetic and genomic resources, and may miss information about non-European ancestry. Our study provides a large and high-quality WGS resource for Chinese populations, which will help to examine 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 both protein-coding and lncRNA genes in this study expands the catalogue of loss-of-function variants in nature. When combined with phenotype information, this resource will provide important biological insights into gene functions.

Methods

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 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). Adaptor sequences and low quality bases were removed with Trimmomatic v0.36 (Bolger et al. 2014).

Variant calling

The variant calling pipeline followed GATK (Ryan Poplin 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).

Sample and site filtering

Duplicate sequencing data for same persons were removed. verifyBamID2 (Zhang et al. 2020) version 1.0.6 was used to check the contamination. Samples with contamination level alpha >= 0.05 were removed. 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) 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. 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 excluded close relatives, samples with more relatives were excluded with priority, and a maximum of 2,902 unrelated samples were kept.

Haplotype phasing

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 rate 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.

Reference panel

The 2,902 independent samples were extracted from phased data above. 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 40,196,029 variants in total.

Imputation performance

The chromosome 2 of HGDP data was used to test imputation performance 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 panel 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 rate was computed for each population as the genotype discordance rate of the masked SNPs. In addition, for Chinese and Han Chinese samples in HGDP dataset, we compared the Rsq statistic for total imputed variants in different MAF bins (MAF: <=0.01, 0.01-0.05, 0.05-0.2, and 0.2-0.5).

The imputation error rates of reference panels constructed with sample subsets of 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 (Supplementary information, 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 (Supplementary information, 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.

Introgression of 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/ and http://cdna.eva.mpg.de/neandertal/altai/Denisovan/. 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 colour the tree and label main branches manually.

Protein-truncating variants (PTVs) and lncRNA loss-of-function splicing 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) 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). 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.

Data access

The datasets generated and/or analysed during the current study are available at http://bigdata.ibp.ac.cn/NyuWa/.

Acknowledgments

We thank Weiwei Zhai for thoughtful discussions and valuable comments on the population

structure analysis.

Funding

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences [XDA12030100 and XDB38040300]; the National Key R&D Program of China [2017YFC0907503 and 2016YFC0901002]; National Natural Science Foundation of China [91940306, 31871294, 31701117 and 31970647].

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 analysis. 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 the primary authors reviewed, edited, and approved manuscript.

Disclosure Declaration

The authors declare that they have no competing interests.

References

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

Ardlie KG DeLuca DS Segre AV Sullivan TJ Young TR Gelfand ET Trowbridge CA Maller JB Tukiainen T Lek M et al. 2015. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**: 648-660.

Asimit JL, Zeggini E. 2012. Imputation of Rare Variants in Next-Generation Association Studies. *Hum Hered* **74**: 196-204.

Auton A Abecasis GR Altshuler DM Durbin RM Bentley DR Chakravarti A Clark AG Donnelly P Eichler EE Flicek P et al. 2015. A global reference for human genetic variation. *Nature* **526**: 68-74.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114-2120.

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

Cai N, Bigdeli TB, Kretzschmar WW, Li YH, Liang JQ, Hu JC, Peterson RE, Bacanu S, Webb BT, Riley B et al. 2020. 11,670 whole-genome sequences representative of the Han Chinese population from the CONVERGE project (Retraction of 10.1038/SDATA.2017.11). *Sci Data* **7**.

Cao Y, Li L, Xu M, Feng Z, Sun X, Lu J, Xu Y, Du P, Wang T, Hu R et al. 2020. The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. *Cell research* doi:10.1038/s41422-020-0322-9.

Carmi S, Hui KY, Kochav E, Liu XM, 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. *Nature communications* **5**.

Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**.

Chen JM, Zheng HF, Bei JX, Sun LD, Jia WH, Li T, Zhang FR, Seielstad M, Zeng YX, Zhang XJ et al. 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 CWK, Mangul S, Robles C, 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 CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST et al. 2011. The variant call format and VCFtools. *Bioinformatics* **27**: 2156-2158.

Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M et al. 2016. Next-generation genotype imputation service and methods. *Nature Genetics* **48**: 1284-1287.

Delaneau O, Zagury JF, Robinson MR, Marchini JL, Dermitzakis ET. 2019. Accurate, scalable and integrative haplotype estimation. *Nature Communications* **10**.

Du ZL, Ma L, Qu HZ, Chen W, Zhang B, Lu X, Zhai WB, Sheng X, Sun YQ, Li WJ et al. 2019. Whole Genome Analyses of Chinese Population and De Novo Assembly of A Northern Han Genome. *Genom Proteom Bioinf* **17**: 229-247.

654 Edge P, Bafna V, Bansal V. 2017. HapCUT2: robust and accurate haplotype assembly for diverse
655 sequencing technologies. *Genome Research* **27**: 801-812.
656 Fang S, Zhang L, Guo J, Niu Y, Wu Y, Li H, Zhao L, Li X, Teng X, Sun X et al. 2017. NONCODEV5: a
657 comprehensive annotation database for long non-coding RNAs. *Nucleic acids research*
658 doi:10.1093/nar/gkx1107: gkx1107-gkx1107.
659 Francioli LC, Menelaou A, Pulit SL, Van Dijk F, Palamara PF, Elbers CC, Neerincx PBT, Ye K, Guryev V,
660 Kloosterman WP et al. 2014. Whole-genome sequence variation, population structure and
661 demographic history of the Dutch population. *Nature Genetics* **46**: 818-825.
662 Gao Y, Zhang C, Yuan LY, Ling YC, Wang XJ, Liu C, Pan YW, Zhang XX, Ma XX, Wang YC et al. 2020.
663 PGG.Han: the Han Chinese genome database and analysis platform. *Nucleic Acids Res* **48**:
664 D971-D976.
665 Hoffmann TJ, Witte JS. 2015. Strategies for Imputing and Analyzing Rare Variants in Association
666 Studies. *Trends in Genetics* **31**: 556-563.
667 Huang J, Howie B, McCarthy S, Memari Y, Walter K, Min JL, Danecek P, Malerba G, Trabetti E, Zheng HF
668 et al. 2015. Improved imputation of low-frequency and rare variants using the UK10K
669 haplotype reference panel. *Nature Communications* **6**.
670 Huang KL, Mashl RJ, Wu YG, Ritter DI, Wang JY, Oh C, Paczkowska M, Reynolds S, Wyczalkowski MA,
671 Oak N et al. 2018. Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell* **173**: 355-+.
672 International Human Genome Sequencing C. 2004. Finishing the euchromatic sequence of the human
673 genome. *Nature* **431**: 931-945.
674 Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis
675 across Computing Platforms. *Mol Biol Evol* **35**: 1547-1549.
676 Lan TM, Lin HX, Zhu WJ, Laurent TCAM, Yang MC, Liu X, Wang J, Wang J, Yang HM, Xu X et al. 2017.
677 Deep whole-genome sequencing of 90 Han Chinese genomes. *Gigascience* **6**.
678 Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu BS, Hart J, Hoffman D, Jang W et
679 al. 2018. ClinVar: improving access to variant interpretations and supporting evidence.
680 *Nucleic Acids Research* **46**: D1062-D1067.
681 Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ,
682 Cummings BB et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans.
683 *Nature* **536**: 285-+.
684 Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform.
685 *Bioinformatics* **26**: 589-595.
686 Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M,
687 Cavalli-Sforza LL et al. 2008. Worldwide human relationships inferred from genome-wide
688 patterns of variation. *Science* **319**: 1100-1104.
689 Lin JC, Fan CT, Liao CC, Chen YS. 2017. Taiwan Biobank: making cross-database convergence possible in
690 the Big Data era. *Gigascience* **7**.
691 Liu SY, Huang SJ, Chen F, Zhao LJ, Yuan YY, Francis SS, Fang L, Li ZL, Lin L, Liu R et al. 2018a. Genomic
692 Analyses from Non-invasive Prenatal Testing Reveal Genetic Associations, Patterns of Viral
693 Infections, and Chinese Population History. *Cell* **175**: 347-+.
694 Liu XM, Wu CL, Li C, Boerwinkle E. 2016. dbNSFP v3.0: A One-Stop Database of Functional Predictions
695 and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Hum Mutat* **37**: 235-241.
696 Liu Y, Cao Z, Wang Y, Guo Y, Xu P, Yuan P, Liu Z, He Y, Wei W. 2018b. Genome-wide screening for
697 functional long noncoding RNAs in human cells by Cas9 targeting of splice sites. *Nat*

698 *Biotechnol* doi:10.1038/nbt.4283.
699 Loh PR, Danecek P, Palamara PF, Fuchsberger C, Y AR, H KF, Schoenherr S, Forer L, McCarthy S,
700 Abecasis GR et al. 2016. Reference-based phasing using the Haplotype Reference Consortium
701 panel. *Nat Genet* **48**: 1443-1448.
702 Maher MC, Uricchio LH, Torgerson DG, Hernandez RD. 2012. Population Genetics of Rare Variants and
703 Complex Diseases. *Hum Hered* **74**: 118-128.
704 Majumder PP. 2010. The Human Genetic History of South Asia. *Current Biology* **20**: R184-R187.
705 Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. 2010. Robust relationship inference
706 in genome-wide association studies. *Bioinformatics* **26**: 2867-2873.
707 Maretty L, Jensen JM, Petersen B, Sibbesen JA, Liu S, Villesen P, Skov L, Belling K, Theil Have C,
708 Izarzugaza JMG et al. 2017. Sequencing and de novo assembly of 150 genomes from
709 Denmark as a population reference. *Nature* **548**: 87-91.
710 McCarthy S Das S Kretzschmar W Delaneau O Wood AR Teumer A Kang HM Fuchsberger C Danecek P
711 Sharp K et al. 2016. A reference panel of 64,976 haplotypes for genotype imputation. *Nat*
712 *Genet* **48**: 1279-1283.
713 Mirabello L, Zhu B, Koster R, Karlins E, Dean M, Yeager M, Gianferante M, Spector LG, Morton LM,
714 Karyadi D et al. 2020. Frequency of Pathogenic Germline Variants in Cancer-Susceptibility
715 Genes in Patients With Osteosarcoma. *Jama Oncol* **6**: 724-734.
716 Nagasaki M, Yasuda J, Katsuoka F, Nariai N, Kojima K, Kawai Y, Yamaguchi-Kabata Y, Yokozawa J, Danjoh
717 I, Saito S et al. 2015. Rare variant discovery by deep whole-genome sequencing of 1,070
718 Japanese individuals. *Nat Commun* **6**.
719 Okonechnikov K, Conesa A, Garcia-Alcalde F. 2016. Qualimap 2: advanced multi-sample quality control
720 for high-throughput sequencing data. *Bioinformatics* **32**: 292-294.
721 Ozdemir BC, Dotto GP. 2017. Racial Differences in Cancer Susceptibility and Survival: More Than the
722 Color of the Skin? *Trends in cancer* **3**: 181-197.
723 Piton A, Redin C, Mandel JL. 2013. XLID-Causing Mutations and Associated Genes Challenged in Light
724 of Data From Large-Scale Human Exome Sequencing (vol 93, pg 368, 2013). *Am J Hum Genet*
725 **93**: 406-406.
726 Poznik GD. 2016. Identifying Y-chromosome haplogroups in arbitrarily large samples of sequenced or
727 genotyped men. *bioRxiv* doi:10.1101/088716: 088716.
728 Price AL, Weale ME, Patterson N, Myers SR, Need AC, Shianna KV, Ge D, Rotter JI, Torres E, Taylor KD et
729 al. 2008. Long-range LD can confound genome scans in admixed populations. *Am J Hum*
730 *Genet* **83**: 132-135; author reply 135-139.
731 Qamar R, Ayub Q, Mohyuddin A, Helgason A, Mazhar K, Mansoor A, Zerjal T, Tyler-Smith C, Mehdi SQ.
732 2002. Y-chromosomal DNA variation in Pakistan. *American journal of human genetics* **70**:
733 1107-1124.
734 Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, Ledbetter DH, Maglott DR,
735 Martin CL, Nussbaum RL et al. 2015. ClinGen - The Clinical Genome Resource. *New Engl J*
736 *Med* **372**: 2235-2242.
737 Ryan Poplin VR-R, Mark A. DePristo, Tim J. Fennell, Mauricio O. Carneiro, Geraldine A. Van der Auwera,
738 David E. Kling, Laura D. Gauthier, Ami Levy-Moonshine, David Roazen, Khalid Shakir, Joel
739 Thibault, Sheila Chandran, Chris Whelan, Monkol Lek, Stacey Gabriel, Mark J. Daly, Benjamin
740 Neale, Daniel G. MacArthur, Eric Banks. 2017. Scaling accurate genetic variant discovery to
741 tens of thousands of samples. *bioRxiv*.

742 Saint Pierre A, Genin E. 2014. How important are rare variants in common disease? *Brief Funct*
743 *Genomics* **13**: 353-361.
744 Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI
745 database of genetic variation. *Nucleic Acids Research* **29**: 308-311.
746 Sud A, Kinnersley B, Houlston RS. 2017. Genome-wide association studies of cancer: current insights
747 and future perspectives. *Nat Rev Cancer* **17**: 692-704.
748 Taliun D Harris DN Kessler MD Carlson J Szpiech ZA Torres R Taliun SAG Corvelo A Gogarten SM Kang
749 HM et al. 2019. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program.
750 *bioRxiv* doi:10.1101/563866: 563866.
751 Tang H, Choudhry S, Mei R, Morgan M, Rodriguez-Cintron W, Burchard EG, Risch NJ. 2008. Long-range
752 LD can confound genome scans in admixed populations - Response to Price et al. *Am J Hum*
753 *Genet* **83**: 135-139.
754 Timpson NJ, Greenwood CMT, Soranzo N, Lawson DJ, Richards JB. 2018. Genetic architecture: the
755 shape of the genetic contribution to human traits and disease. *Nat Rev Genet* **19**: 110-124.
756 Toure A, Cabral M, Niang A, Diop C, Garat A, Humbert L, Fall M, Diouf A, Broly F, Lhermitte M et al.
757 2016. Prevention of isoniazid toxicity by NAT2 genotyping in Senegalese tuberculosis patients.
758 *Toxicology reports* **3**: 826-831.
759 Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP. 2011. Conserved Function of lincRNAs in Vertebrate
760 Embryonic Development despite Rapid Sequence Evolution. *Cell* **147**: 1537-1550.
761 van Leeuwen EM, Karssen LC, Deelen J, Isaacs A, Medina-Gomez C, Mbarek H, Kanterakis A, Trompet S,
762 Postmus I, Verweij N et al. 2015. Genome of the Netherlands population-specific imputations
763 identify an ABCA6 variant associated with cholesterol levels. *Nat Commun* **6**.
764 Vatsis KP, Martell KJ, Weber WW. 1991. Diverse point mutations in the human gene for polymorphic
765 N-acetyltransferase. *Proceedings of the National Academy of Sciences of the United States of*
766 *America* **88**: 6333-6337.
767 Wall JD, Stawiski EW, Ratan A, Kim HL, Kim C, Gupta R, Suryamohan K, Gusareva ES, Purbojati RW,
768 Bhangale T et al. 2019. The GenomeAsia 100K Project enables genetic discoveries across Asia.
769 *Nature* **576**: 106-+.
770 Walter K Min JL Huang J Crooks L Memari Y McCarthy S Perry JRB Xu C Futema M Lawson D et al. 2015.
771 The UK10K project identifies rare variants in health and disease. *Nature* **526**: 82-+.
772 Wang K, Li MY, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from
773 high-throughput sequencing data. *Nucleic Acids Research* **38**.
774 Weir BS, Cockerham CC. 1984. Estimating F-Statistics for the Analysis of Population Structure.
775 *Evolution* **38**: 1358-1370.
776 Wen B, Li H, Lu DR, Song XF, Zhang F, He YG, Li F, Gao Y, Mao XY, Zhang L et al. 2004. Genetic evidence
777 supports demic diffusion of Han culture. *Nature* **431**: 302-305.
778 Wu DG, Dou JZ, Chai XR, Bellis C, Wilm A, Shih CC, Soon WWJ, Bertin N, Lin CB, Khor CC et al. 2019.
779 Large-Scale Whole-Genome Sequencing of Three Diverse Asian Populations in Singapore. *Cell*
780 **179**: 736-+.
781 Xu SH, Yin XY, Li SL, Jin WF, Lou HY, Yang L, Gong XH, Wang HY, Shen YP, Pan XD et al. 2009a. Genomic
782 Dissection of Population Substructure of Han Chinese and Its Implication in Association
783 Studies. *Am J Hum Genet* **85**: 762-774.
784 Xu SH, Yin XY, Li SL, Jin WF, Lou HY, Yang L, Gong XH, Wang HY, Shen YP, Pan XD et al. 2009b. Genomic
785 Dissection of Population Substructure of Han Chinese and Its Implication in Association

786 Studies. *American journal of human genetics* **85**: 762-774.
787 Yan S, Wang CC, Zheng HX, Wang W, Qin ZD, Wei LH, Wang Y, Pan XD, Fu WQ, He YG et al. 2014. Y
788 Chromosomes of 40% Chinese Descend from Three Neolithic Super-Grandfathers. *Plos One* **9**.
789 Zhang F, Flickinger M, Taliun SAG, Abecasis GR, Scott LJ, McCaroll SA, Pato CN, Boehnke M, Kang HM,
790 Genetics IP. 2020. Ancestry-agnostic estimation of DNA sample contamination from sequence
791 reads. *Genome Research* **30**: 185-194.
792 Zhao H, Sun ZF, Wang J, Huang HJ, Kocher JP, Wang LG. 2014. CrossMap: a versatile tool for coordinate
793 conversion between genome assemblies. *Bioinformatics* **30**: 1006-1007.

794

Figure Legends

Figure 1. Overview of NyuWa dataset

(A) Distribution of samples in NyuWa resource. Samples were assigned to provinces based on the native places or hospitals where samples were collected. The map was downloaded from the standard map service website (http://bzdt.ch.mnr.gov.cn/).

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

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

Figure 2. Variants statistics in NyuWa resource

(A) Number 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 & v3, 1KGP3, ESP, dbSNP, TOPMed, 90 Han and GAsP. INS: small insertion, DEL: small deletion.

(B) Number (upper) and novel rate (lower) of variants in different RefSeq annotation regions.

(C) Number (upper) and novel rate (lower) of variants in different NONCODE annotation regions.

(D) Number of non-synonymous 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).

Figure 3. Performance of NyuWa haplotype reference panel

(A) Fold change (FC) of imputation error rate in different Asia populations in HGDP dataset between 1KGP3 panel and NyuWa (left) or NyuWa+1KGP3 (right) panel. Lower fold change values represent better performance in NyuWa or NyuWa+1KGP3 panels. EA: East Asian; NEA: North East Asian; CA: Central Asian; CSA: Central South Asian. Significances of error rate differences were calculated by chi-squared test. *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

(B) Fold change of imputation error rate between GAsP panel and NyuWa (left) or NyuWa+1KGP3 (right) panel. Colors representing regions in (A) and (B) are consistent.

(C-D) Number (upper) and proportion (lower) of high-quality (Rsq > 0.8) variants imputed for All Chinese populations (C) and Han Chinese population (D) in HGDP dataset. Variants were grouped in different MAF intervals.

Figure 4. Chinese population structure based on NyuWa dataset

(A) ADMIXTURE analysis of NyuWa samples with East Asia samples in 1KGP3. Number of ancestries K = 3 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) Proportion 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) were also shown.

(C) Top 2 primary components (PC1 & 2) of NyuWa samples. Each point represents a sample. Samples were marked by provinces and areas of China. PC1 represents the difference of north and south Chinese.

(D) Imputation error rates of two test datasets representing north (Han N. China in HGDP, upper) and south (CHS in 1KGP3, lower) Han Chinese. Each point represents a reference panel constructed with a certain sample subset of NyuWa reference panel. The color of red represents North (N) specific panels from samples in the left part of PC1 shown in (C), while blue represent South (S) specific panels in the right part of PC1. The gray triangles represent reference panels with randomly (R) selected samples. The numbers 1k and 1.5k represent the proportions 1/3 and 1/2 of the 2902 total samples in NyuWa panel. Dotted lines represent addition of more samples.

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 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 NyuWa dataset, only provinces with sample sizes >= 20 were shown. 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.

(E) Allele frequencies of known cancer risk loci (row) that vary in different populations or regions (column). For NyuWa dataset, only provinces with sample sizes >= 20 were shown. The AF color bar is consistent with (D).

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

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

(B) Number of protein truncating variants (PTVs) grouped by novel, known, heterozygous and homozygous.

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

(D) Number of lncRNA splicing variants grouped by 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.

Tables

Table 1. Number of total and new variants in NyuWa resource and reference panel.

Type	**All variants [a]**		**Reference panel [b]**	
	Total	**Novel [c]**	**Total**	**Specific [d]**
All	79,226,351	25,014,646	19,256,267	3,246,071
Non-synonymous	500,966	149,343	73,260	7,048
Non-synonymous deleterious	315,016	101,407	33,526	3,323

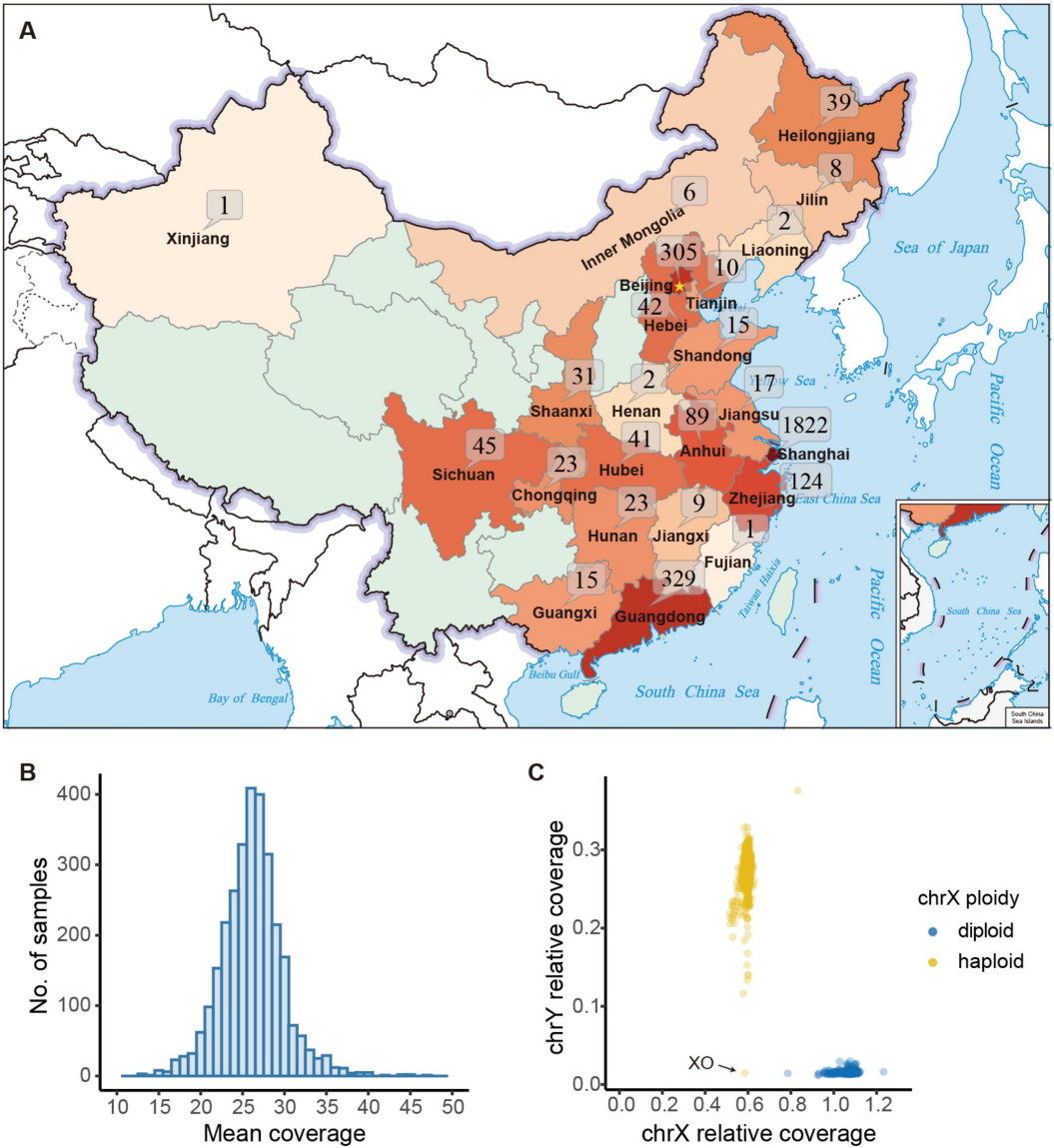
PTV	18,711	9,994	1,381	334
lncRNA splicing	3,793	1,499	743	80

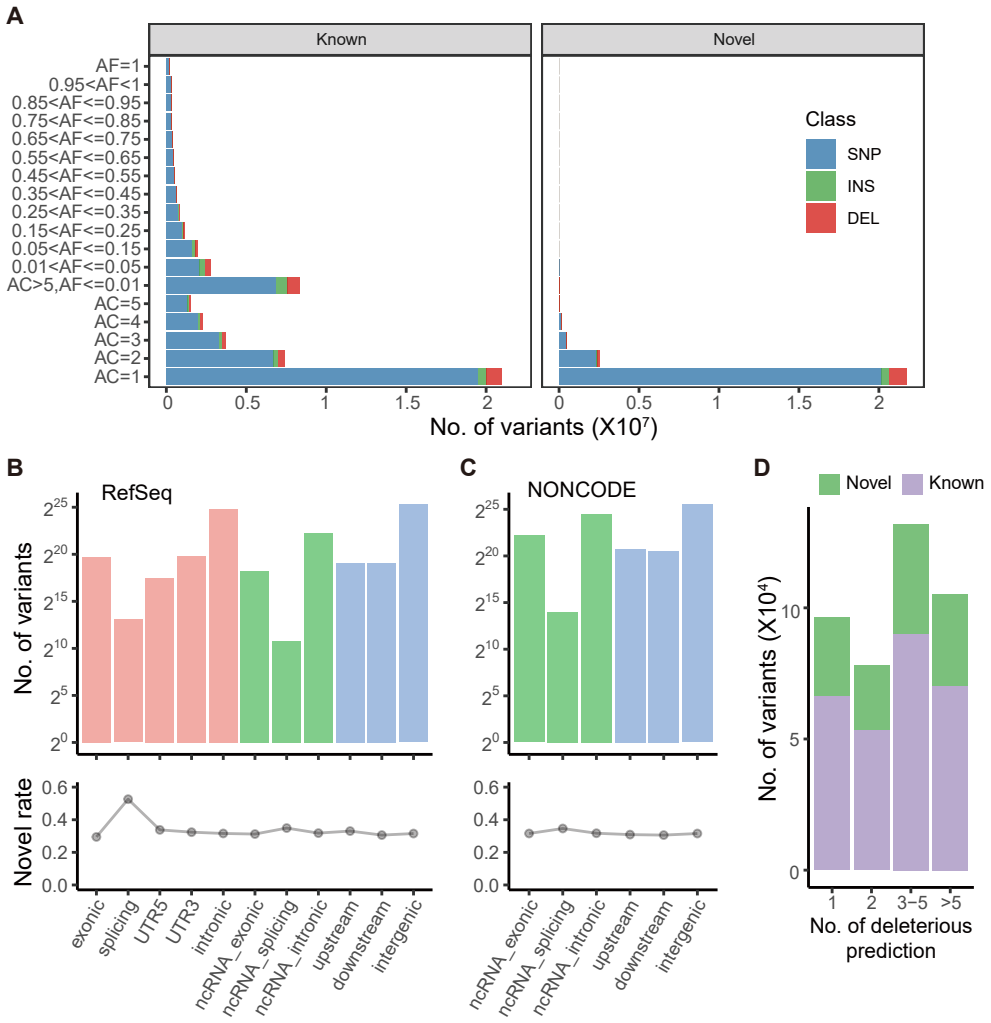
a. All variants refers to variants in NyuWa resource.

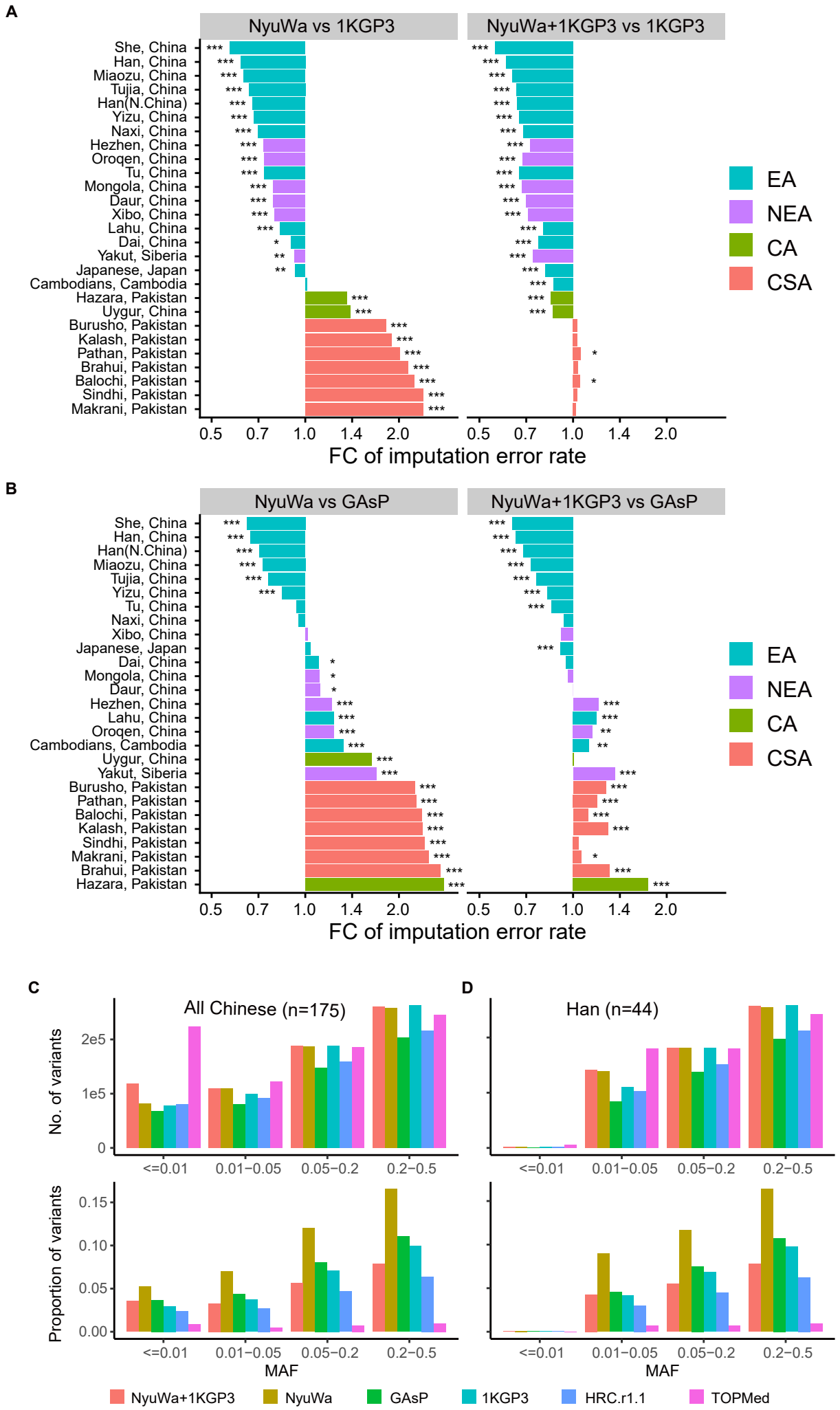
b. Reference panel refers to variants in NyuWa reference panel.

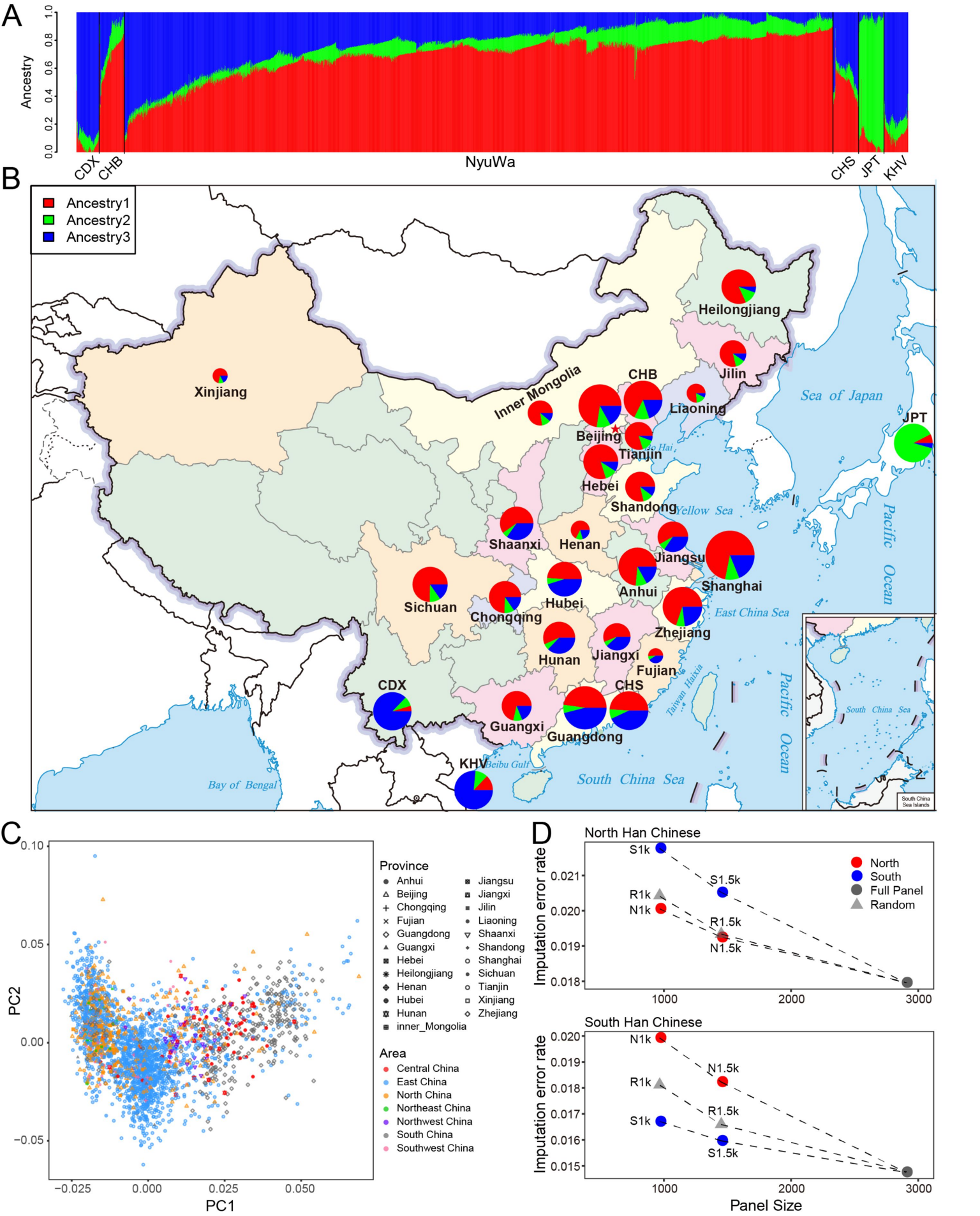
c. Novelty of variants was compared with dbSNP, 1KGP3, gnomADV2.1, EXAC, ESP, gnomADV3, TOPMed, 90 Han and GAsP

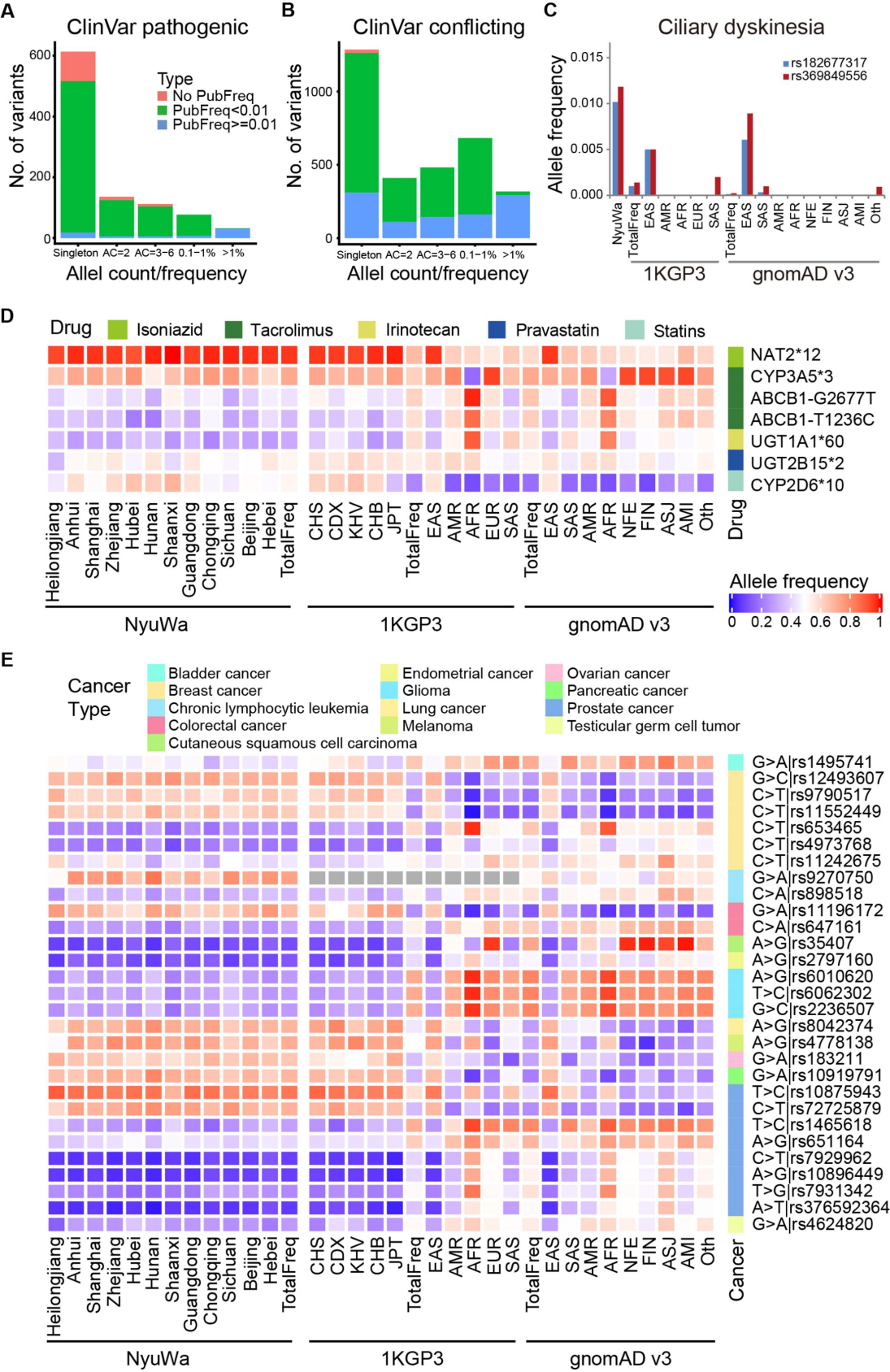
d. Variants in the other 4 public available haplotype reference panels (1KGP3, HRC.r1.1, GAsP and TOPMed) were excluded.

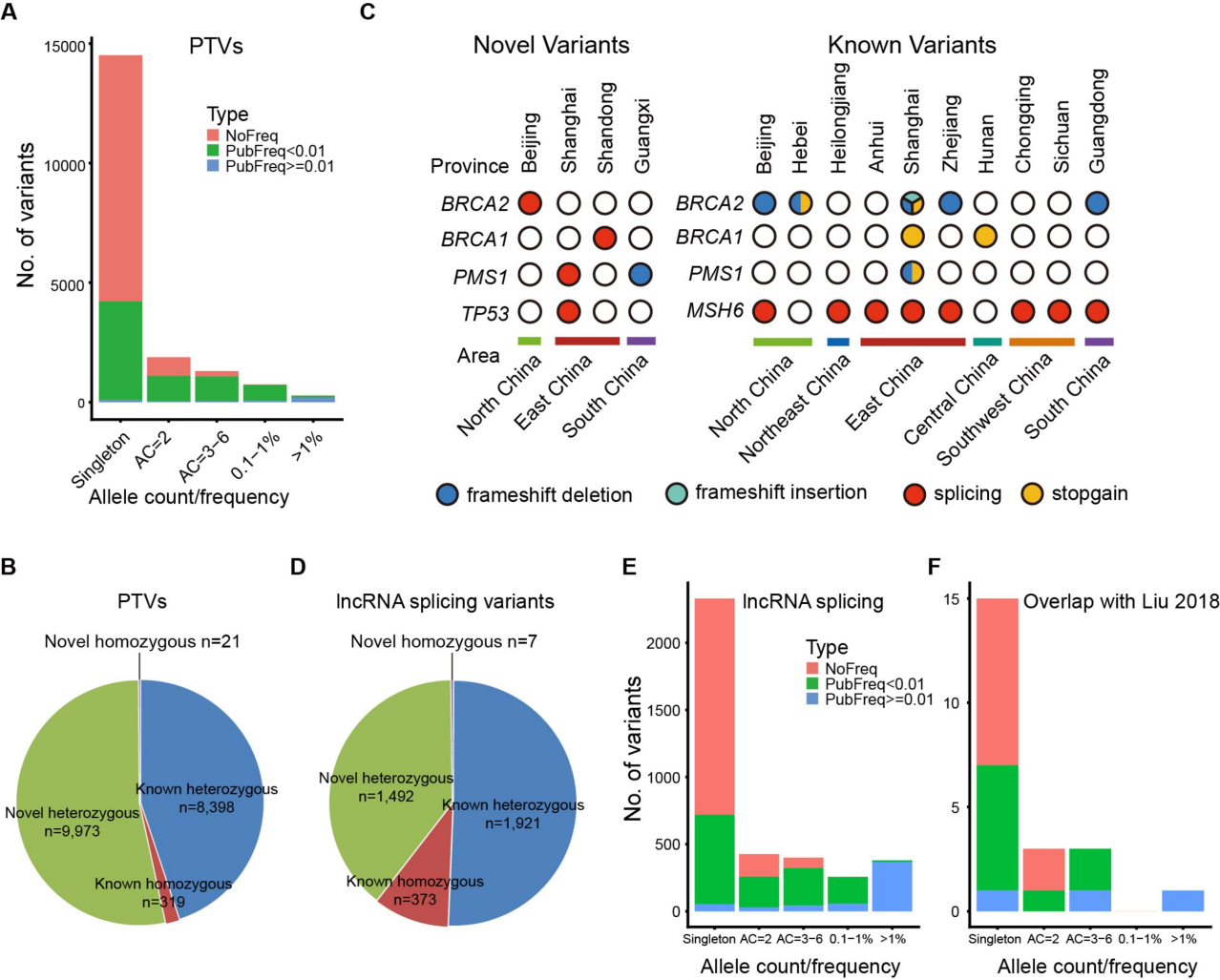

A
Xinjiang
1
Inner Mongolia
6
Heilongjiang
39
Jilin
8
Liaoning
2
Beijing
305
Tianjin
10
Hebei
42
Shandong
15
Shaanxi
31
Henan
2
Jiangsu
17
Anhui
89
Shanghai
1822
Sichuan
45
Hubei
23
Chongqing
41
Zhejiang
124
Hunan
23
Jiangxi
9
Fujian
1
Guangxi
15
Guangdong
329
Sea of Japan
Pacific Ocean
East China Sea
Taiwan Haixia
Beibu Gulf
South China Sea
Bay of Bengal
B
No. of samples
Mean coverage
C
chrY relative coverage
chrX relative coverage
XO
chrX ploidy
diploid
haploid


A
Known
Novel
AF=1
0.95<AF<1
0.85<AF<=0.95
0.75<AF<=0.85
0.65<AF<=0.75
0.55<AF<=0.65
0.45<AF<=0.55
0.35<AF<=0.45
0.25<AF<=0.35
0.15<AF<=0.25
0.05<AF<=0.15
0.01<AF<=0.05
AC>5,AF<=0.01
AC=5
AC=4
AC=3
AC=2
AC=1
Class
SNP
INS
DEL
No. of variants (X10⁷)
B
RefSeq
No. of variants
Novel rate
exonic
splicing
UTR5
UTR3
intronic
ncRNA_exonic
ncRNA_splicing
ncRNA_intronic
upstream
downstream
intergenic
C
NONCODE
D
Novel
Known
No. of variants (X10⁴)
No. of deleterious prediction


A
NyuWa vs 1KGP3
NyuWa+1KGP3 vs 1KGP3
She, China
Han, China
Miaozu, China
Tujia, China
Han(N.China)
Yizu, China
Naxi, China
Hezhen, China
Oroqen, China
Tu, China
Mongola, China
Daur, China
Xibo, China
Lahu, China
Dai, China
Yakut, Siberia
Japanese, Japan
Cambodians, Cambodia
Hazara, Pakistan
Uygur, China
Burusho, Pakistan
Kalash, Pakistan
Pathan, Pakistan
Brahui, Pakistan
Balochi, Pakistan
Sindhi, Pakistan
Makrani, Pakistan
0.5
0.7
1.0
1.4
2.0
FC of imputation error rate
EA
NEA
CA
CSA
B
NyuWa vs GAsP
NyuWa+1KGP3 vs GAsP
She, China
Han, China
Han(N.China)
Miaozu, China
Tujia, China
Yizu, China
Tu, China
Naxi, China
Xibo, China
Japanese, Japan
Dai, China
Mongola, China
Daur, China
Hezhen, China
Lahu, China
Oroqen, China
Cambodians, Cambodia
Uygur, China
Yakut, Siberia
Burusho, Pakistan
Pathan, Pakistan
Balochi, Pakistan
Kalash, Pakistan
Sindhi, Pakistan
Makrani, Pakistan
Brahui, Pakistan
Hazara, Pakistan
0.5
0.7
1.0
1.4
2.0
FC of imputation error rate
EA
NEA
CA
CSA
C
All Chinese (n=175)
No. of variants
2e5
1e5
0
<=0.01
0.01−0.05
0.05−0.2
0.2−0.5
Proportion of variants
0.15
0.10
0.05
0.00
MAF
D
Han (n=44)
NyuWa+1KGP3
NyuWa
GAsP
1KGP3
HRC.r1.1
TOPMed


A
Ancestry
0.0
0.2
0.4
0.6
0.8
1.0
CDX
CHB
NyuWa
CHS
JPT
KHV
B
Ancestry1
Ancestry2
Ancestry3
Heilongjiang
Jilin
Liaoning
Xinjiang
Inner Mongolia
CHB
Beijing
Tianjin
Hebei
Shandong
Shaanxi
Henan
Jiangsu
Shanghai
Anhui
Hubei
Sichuan
Chongqing
Zhejiang
Hunan
Jiangxi
Fujian
CDX
CHS
Guangxi
Guangdong
KHV
JPT
Sea of Japan
Yellow Sea
East China Sea
South China Sea
Pacific Ocean
Bay of Bengal
Taiwan Haixia
Beibu Gulf
C
PC1
PC2
Province
Anhui
Beijing
Chongqing
Fujian
Guangdong
Guangxi
Hebei
Heilongjiang
Henan
Hubei
Hunan
inner_Mongolia
Jiangsu
Jiangxi
Jilin
Liaoning
Shaanxi
Shandong
Shanghai
Sichuan
Tianjin
Xinjiang
Zhejiang
Area
Central China
East China
North China
Northeast China
Northwest China
South China
Southwest China
D
North Han Chinese
South Han Chinese
Imputation error rate
Panel Size
North
South
Full Panel
Random
S1k
S1.5k
R1k
N1k
R1.5k
N1.5k


A
ClinVar pathogenic
No. of variants
Type
No PubFreq
PubFreq<0.01
PubFreq>=0.01
Singleton AC=2 AC=3−6 0.1−1% >1%
Allel count/frequency
B
ClinVar conflicting
No. of variants
Singleton AC=2 AC=3−6 0.1−1% >1%
Allel count/frequency
C
Ciliary dyskinesia
Allele frequency
rs182677317
rs369849556
NyuWa TotalFreq EAS AMR AFR EUR SAS TotalFreq EAS SAS AMR AFR NFE FIN ASJ AMI Oth
1KGP3
gnomAD v3
D
Drug
Isoniazid
Tacrolimus
Irinotecan
Pravastatin
Statins
NAT2*12
CYP3A5*3
ABCB1-G2677T
ABCB1-T1236C
UGT1A1*60
UGT2B15*2
CYP2D6*10
Heilongjiang Anhui Shanghai Zhejiang Hubei Hunan Shaanxi Guangdong Chongqing Sichuan Beijing Hebei TotalFreq
CHS CDX KHV CHB JPT TotalFreq EAS AMR AFR EUR SAS TotalFreq EAS SAS AMR AFR NFE FIN ASJ AMI Oth
NyuWa
1KGP3
gnomAD v3
Allele frequency
0 0.2 0.4 0.6 0.8 1
E
Cancer Type
Bladder cancer
Breast cancer
Chronic lymphocytic leukemia
Colorectal cancer
Cutaneous squamous cell carcinoma
Endometrial cancer
Glioma
Lung cancer
Melanoma
Ovarian cancer
Pancreatic cancer
Prostate cancer
Testicular germ cell tumor
G>A|rs1495741
G>C|rs12493607
C>T|rs9790517
C>T|rs11552449
C>T|rs653465
C>T|rs4973768
C>T|rs11242675
G>A|rs9270750
C>A|rs898518
G>A|rs11196172
C>A|rs647161
A>G|rs35407
A>G|rs2797160
A>G|rs6010620
T>C|rs6062302
G>C|rs2236507
A>G|rs8042374
A>G|rs4778138
G>A|rs183211
G>A|rs10919791
T>C|rs10875943
C>T|rs72725879
T>C|rs1465618
A>G|rs651164
C>T|rs7929962
A>G|rs10896449
T>G|rs7931342
A>T|rs376592364
G>A|rs4624820
Heilongjiang Anhui Shanghai Zhejiang Hubei Hunan Shaanxi Guangdong Chongqing Sichuan Beijing Hebei TotalFreq
CHS CDX KHV CHB JPT TotalFreq EAS AMR AFR EUR SAS TotalFreq EAS SAS AMR AFR NFE FIN ASJ AMI Oth
NyuWa
1KGP3
gnomAD v3
Cancer


A
PTVs
No. of variants
15000
10000
5000
0
Type
NoFreq
PubFreq<0.01
PubFreq>=0.01
Singleton
AC=2
AC=3~6
0.1~1%
>1%
Allele count/frequency
C
Novel Variants
Known Variants
Province
Beijing
Shanghai
Shandong
Guangxi
Hebei
Heilongjiang
Anhui
Zhejiang
Hunan
Chongqing
Sichuan
Guangdong
BRCA2
BRCA1
PMS1
TP53
MSH6
Area
North China
East China
South China
Northeast China
Central China
Southwest China
frameshift deletion
frameshift insertion
splicing
stopgain
B
PTVs
Novel homozygous n=21
Known heterozygous n=8,398
Novel heterozygous n=9,973
Known homozygous n=319
D
lncRNA splicing variants
Novel homozygous n=7
Novel heterozygous n=1,492
Known heterozygous n=1,921
Known homozygous n=373
E
lncRNA splicing
No. of variants
2000
1500
1000
500
0
Type
NoFreq
PubFreq<0.01
PubFreq>=0.01
Singleton
AC=2
AC=3~6
0.1~1%
>1%
Allele count/frequency
F
Overlap with Liu 2018
15
10
5
0
Singleton
AC=2
AC=3~6
0.1~1%
>1%
Allele count/frequency