Promega

Don’t settle for just a supplier. Find a custom manufacturing partner.

Your specifications. Your format.

Our scientists waiting to help.

Let’s TALK CUSTOM

cGMP Manufacturing

15 ISO CERTIFIED Locations

100+ GLOBAL IVD COMPANIES Supplied

Selecting a supplier for your clinical/diagnostic products can be a challenge—especially a supplier who can adapt to your specific needs. Don’t settle for just a supplier. Instead, partner with Promega and work with a custom manufacturer willing to provide you with the scientific expertise, ongoing technical support and quality standards that support your success.

Learn more with our free webinar:
promega.com/CustomWebinar

Received: 30 January 2019 | Revised: 14 May 2019 | Accepted: 5 June 2019

DOI: 10.1002/humu.23835

DATABASES

Human Mutation HGVS HUMAN GENOME VARIATION SOCIETY WILEY

A Vietnamese human genetic variation database

Vinh S. Le[1,2,3*] | Kien T. Tran[1*] | Hoa T. P. Bui[1,2,4] | Huong T. T. Le[1,2] | Canh D. Nguyen[3] | Duong H. Do[1,2] | Ha T. T. Ly[1,2] | Linh T. D. Pham[1] | Lan T. M. Dao[1] | Liem T. Nguyen[1]

[1]Vinmec Research Institute of Stem Cell and Gene Technology, Hanoi, Vietnam

[2]Department of Gene Technology, Vinmec International Hospital Times City, Hanoi, Vietnam

[3]Faculty of Information Technology, University of Engineering and Technology, Vietnam National University Hanoi, Hanoi, Vietnam

[4]School of Environment and Life Science, University of Salford, Manchester, United Kingdom

Correspondence
Vinh Sy Le, University of Engineering and Technology, Vietnam National University Hanoi, 144 Xuan Thuy, Cau Giay, 10000 Hanoi, Vietnam.
Email: vinhls@vnu.edu.vn

Liem Thanh Nguyen, Vinmec Research Institute of Stem Cell and Gene Technology, 458 Minh Khai, Hai Ba Trung, 10000 Hanoi, Vietnam. Email: v.liemnt@vinmec.com

Funding information
Vinmec Healthcare System, Grant/Award Number: ISC.17.03

Abstract

Large scale human genome projects have created tremendous human genome databases for some well-studied populations. Vietnam has about 95 million people (the 14th largest country by population in the world) of which more than 86% are Kinh people. To date, genetic studies for Vietnamese people mostly rely on genetic information from other populations. Building a Vietnamese human genetic variation database is a must for properly interpreting Vietnamese genetic variants. To this end, we sequenced 105 whole genomes and 200 whole exomes of 305 unrelated Kinh Vietnamese (KHV) people. We also included 101 other previously published KHV genomes to build a Vietnamese human genetic variation database of 406 KHV people. The KHV database contains 24.81 million variants (22.47 million single nucleotide polymorphisms (SNPs) and 2.34 million indels) of which 0.71 million variants are novel. It includes more than 99.3% of variants with a frequency of >1% in the KHV population. Noticeably, the KHV database revealed 107 variants reported in the human genome mutation database as pathological mutations with a frequency above 1% in the KHV population. The KHV database (available at https://genomes.vn) would be beneficial for genetic studies and medical applications not only for the Vietnamese population but also for other closely related populations.

KEYWORDS

Asian human genome database, Vietnamese genetic population structure, Vietnamese human genome database, whole genome sequencing

1 | INTRODUCTION

The sequencing cost of a human genome is now around 1,000 US Dollars leading to an era of human genomics. A number of large scale human genome projects have been conducted to build human genome variation databases at both global and specific population levels. Notably, the 1,000 Human Genomes (1KG) Project sequenced 2,504 healthy people from 26 populations (2015 Genomes Project Consortium et al., 2015). The purpose of this project was to detect most common variants with minor allele frequencies of at least 1%. In Asia, the Singaporean human genome project sequenced 100 Malay people to detect low-frequency and rare variants (Wong et al., 2013); the Korean (KR) Personal Genomes Project sequenced 35 KR genomes to decipher the genetic architecture of the KR population (Zhang et al., 2014). Recently, Lan et al. (2017) sequenced the whole genomes of 90 Han Chinese people to investigate Han Chinese human genomes.

A human genome consists of about 4 million variants in comparison to the human reference genome of which a considerable number of variants are missense substitutions and predicted to be damaging (Xue et al., 2012). Normally, variant frequency information

*Vinh S. Le and Kien T. Tran contributed equally to this work.

from human genome databases is used to filter or prioritize potentially deleterious variants (2015 Genomes Project Consortium et al., 2015; Lek et al., 2016; MacArthur et al., 2014; The 2012 Genomes Project Consortium, 2012).

The 1KG project discovered in a total of 88 million variants containing almost all common variants with a frequency of 10% in all populations under the study (The 2012 Genomes Project Consortium, 2012). However, 17% of low-frequency variants (range from 0.5–5%) were only found in a single population. In addition, a large number of variants common in the global populations were rare in a specific population. This was especially true for South East Asia as Lu & Xu (2013) showed that the 1KG project did not have sufficient coverage of the human genetic diversity in this region.

Exome Aggregation Consortium aggregated a large number of genomes and exomes from a variety of large-scale sequencing projects (Lek et al., 2016). Specifically, they combined 15,496 genomes and 123,136 exomes collecting from various disease-specific and population genetic studies to form the gnomAD database. The database plays as a powerful tool for clinical interpretations of variants. As gnomAD used data from various specific disease projects, it is only relevant for interpretations of severe pediatric diseases.

Vietnam has about 95 million people (the 14th largest country by population in the world) of which more than 86% are Kinh people. We started the first Vietnamese human genome project with a Kinh Vietnamese (KHV) trio and discovered a considerable number of novel variants (Hai et al., 2015). Phase 3 of the 1KG project sequenced 99 unrelated KHV people in Ho Chi Minh City, Vietnam at the low coverage level. In this study, we additionally sequenced genomes and exomes of 305 unrelated KHV people in Hanoi, Vietnam, then combined the new data with previously published KHV genomes to create a Vietnamese human genome database. We also analyzed genetic population structures of KHV and other Asian populations to elucidate their genetic relationships. The results provided useful information to utilize the KHV database for genetic studies of KHV and other closely related populations in Asia.

2 | MATERIAL AND METHODS

2.1 | Material

2.1.1 | Sample collection

For whole genome sequencing (WGS), we recruited 105 unrelated people with self-declaration as healthy and Kinh ethnic for at least three generations at Vinmec International Hospital in Hanoi, Vietnam. Approximately 2 ml of peripheral blood from each individual was collected in an anticoagulation tube containing EDTA and stored at −80°C.

For whole exome sequencing (WES), we obtained peripheral blood of 200 healthy parents whose children participated as cases in our autism spectrum disorder study. These parents were self-reported as KHV people.

2.1.2 | Whole genome and exome sequencing

For WGS, the genomic DNA was physically fragmented to expected size of 350 bps by Covaris ME2 (Covaris). The WGS library was prepared by using a TruSeq DNA PCR-Free Library Preps Kit (Illumina), and its concentrations were quantified by using a KAPA Universal Library Quantification Kit (Kapa Biosystems). For WES, the library was prepared by using a Nextera Rapid Capture Kit (Illumina), and its concentrations were quantified by a Qubit double stranded DNA (dsDNA) Broad Range Assay Kit (Invitrogen). The sizes of WGS and WES libraries were measured by using Lab chip 3K Hisense Kits (Perkin Elmer).

The libraries were loaded on a patterned flow cell and subsequently clustered on a cBot system (Illumina). Paired-end sequencings were conducted on HiSeq 4000 machine (Illumina) with an inserted size of 350 bps. As a result, we obtained paired-end short reads of 150 bps from WGS and 75 bps from WES.

2.2 | Methods

2.2.1 | Variant calling and validation

The pair-ended short reads from our newly sequenced genomes and exomes were cleaned and subsequently mapped to the National Center for Biotechnology Information (NCBI) reference genome build 38 (GRCh38) using Burrows-Wheeler aligner (Li & Durbin, 2009) to create alignments. The quality of short reads was measured using the FastQC program (Andrews, 2010). In this study, we focused on determining single nucleotide polymorphisms (SNPs) and short indels. To do that, we followed the best practice guidelines of GATK and Platypus programs to call variants. Variants with a quality Phred-score less than 30 were filtered out from further analyses. All variants from our newly sequenced genomes and exomes have been deposited into the dbSNP database.

Sanger sequencing was used to validate the results from WGS at selected positions. To this end, we performed Sanger sequencing at 64 variant sites in two genomes. Technically, primers were designed by using Primer3plus software (http://www.bioinformatics.nl). PCR reactions were performed with GoTaq DNA Polymerase (Promega, WI). Sanger sequencing was performed by using BigDye Terminator v3.1 cycle sequencing kit (Thermo Fisher Scientific, MA) on an ABI 3500 Dx Genetic Analyzer (Thermo Fisher Scientific).

2.2.2 | Building KHV database

First, we integrated our 105 newly sequenced KHV genomes with 101 previously published KHV genomes (2015 Genomes Project Consortium et al., 2015; Hai et al., 2015) to create a database of 206 KHV genomes, called KHV-G database. We used our 200 newly sequenced exomes to measure the variant detection power of the KHV-G database at the exome regions. Technically, the power to detect variants with a frequency of x% is estimated by the ratio of p to q where q is the number of variants in the 200 exomes with the frequency of x% and p is the number of these variants that are present in the KHV-G database.

Finally, we combined all KHV genomes and exomes to build a comprehensive Vietnamese genetic variation database, namely the KHV database. We used the SnpEff program to annotate and predict genetic effects of variants on genes and proteins (Cingolani et al., 2012). We developed scripts to analyze variant frequencies in the KHV and other global databases.

2.2.3 | Population analysis

Although KHV is one of the largest ethnic groups in Asia, the genomic relationships between KHV and other Asian populations have not been comprehensively investigated. In this study, we analyzed genomic relationships between the KHV population in Hanoi, Vietnam and nine other Asian populations. We also included African (YRI) and European (CEU) populations as outgroups into the study. Specifically, we created an SNP data set of 719 individuals from 12 populations including 94 KHV samples whose depth coverages were at least 10×. The SNP data of KHV people were extracted from their sequenced genomes. The SNP data of other populations were obtained from the HUGO Pan-Asian SNP data set (Abdulla et al., 2009). The data set contained 46,473 autosomal SNPs that appeared in both KHV genomes and the HUGO Pan-Asian SNP data set. We also created a sub-dataset of 4,253 SNPs with minimum interval distance of 500 kb to avoid a high linkage disequilibrium between SNPs.

We conducted several genomic analyses to study the population structures. First, we used the EIGENSOFT program (Patterson, Price, & Reich, 2006) to perform principal component analysis (PCA) based on the sub-dataset of 4,253 SNPs to examine the distribution of individuals and populations. Second, we used the TreeMix program (Pickrell & Pritchard, 2012) to reconstruct evolutionary relationships among the populations. The TreeMix finds the maximum likelihood tree **T** of all populations based on allele frequencies of all 46,473 autosomal SNPs. The block jackknife procedure was applied to assess the confidence of branches in **T** (Reich, Thangaraj, Patterson, Price, & Singh, 2009). Technically, all SNPs were divided into continuous blocks, each containing 400 SNPs. For each replicate, we deleted one block and searched the maximum likelihood tree for remaining blocks. The trees constructed from replicates were summarized to assign support values for branches in **T**. In addition, we used the neighbor-joining method (Nei, 1987) to reconstruct a distanced-based tree using the F_{ST} distances between populations.

Finally, we evaluated the ancestries of the populations. To do that, we used the Bayesian clustering algorithm fastSTRUCTURE (Raj, Stephens, & Pritchard, 2014) with the sub-dataset of 4,253 SNPs to estimate ancestries of all populations. The ancestries of each individual came from *K* different ancestral populations. We conducted fastSTRUCTURE with different *K* values to select the best *K* value. The ancestries of individuals were summarized to determine the ancestries of each population. We also performed F3 statistic tests (Patterson et al., 2012) to examine gene flows between KHV and other populations using the AdmixTools (Patterson et al., 2012). The *F3* (X; B, C) test detects if there was gene flow from two donate populations B and C to the admixture population X. Technically, a significant negative *F3* (X; B, C) value, that is *z* score <−2.58, indicates significant gene flow from B and C to X.

3 | RESULTS AND DISCUSSION

3.1 | Variant calling and validation

We obtained 105 whole genomes sequenced at an average depth of about 17× (from 8× to 36×) and 200 whole exomes sequenced at an average depth of 82× (from 52× to 146×). Almost all short reads have high quality (see Figure S1), that is more than 96.9% of short reads from genomes and 99.7% of exomes have quality Phred-score of at least 20. The high-quality data are sufficient for variant calling.

The GATK and Platypus programs resulted in millions of variants from 105 whole genomes and 200 whole exomes. The results included some discordance, that is some variants called by one caller but not the other. In this study, we considered a variant as reliable for further analyses if it was called by both GATK and Platypus programs. The filter strategy helped reduce false positive variants in our database.

We obtained 10.06 million reliable variants (9.38 million SNPs and 0.68 million indels) from the new genomes and exomes. Examining the appearance of these variants in the 1KG project, gnomAD, and dbSNP databases revealed that 0.71 million variants were novel biallelic variants (0.66 million SNPs and 0.05 million indels). The variants were classified into seven categories corresponding to seven regions in the genome: coding region (CDS), 5′-untranslated region (UTR), 3′-UTR, intron, upstream, downstream, and intergenic (see Table 1). A majority number (~75%) of variants appeared in the intron and intergenic regions. Notably, there were about 0.28 million (~2.7%) variants occurring in the CDS, 5′-UTR, and 3′-UTR regions of which more than 24 thousands of variants were novel. The considerable number of novel variants, including those in the coding and regulatory regions of genes, confirms the necessity of conducting additional genomic studies for Asian populations.

We validated the results from WGS by Sanger sequencing at 64 variant sites in two genomes, that were VIN343 with low coverage (8×) and VIN057 with medium coverage (15×). All the results obtained from WGS and Sanger sequencing were matched except at

TABLE 1 Gene-based annotations of variants from newly sequenced genomes and exomes

Regions	#Variants (million)	#Novel biallelic variants (million)
CDS	0.139	0.013
5′-UTR	0.028	0.002
3′-UTR	0.114	0.009
Intron	3.789	0.272
Upstream	1.233	0.088
Downstream	0.924	0.064
Intergenic	3.834	0.265

Abbreviation: UTR, untranslated region.

one position in VIN057. The variant at that position in VIN057 was detected by Sanger sequencing, but not called by WGS because of a low coverage at the position (i.e., none of 10 short reads sequenced at the position in VIN057 supported the variant). The overall validation rate of the WGS results by Sanger sequencing was about 99.2%.

3.2 | KHV database

The KHV database was built from 206 genomes and 200 exomes of 406 unrelated KHV people (i.e., including variants in our newly sequenced samples and/or previously published in KHV-related genome studies). It contains 24.81 million variants (22.47 million SNPs and 2.34 million indels) with a wide range of allele frequencies (see Figure S2). Specifically, the KHV database consists of 10.97 million (44%) variants with alternative allele frequency ≤1% and 13.84 million (56%) variants with alternative allele frequency >1%. As variants with frequency >1% are typically considered noncausing disease variants, the variant frequency information could be used to evaluate pathological effects of variants in medical studies.

We compared the allele frequencies in the KHV and global populations. The allele frequencies in the global populations were obtained from the 1KG database. The KHV database contained 0.44 and 1.24 million variants that were rare in the global populations with frequencies ≤ 0.1% and ≤ 0.5%, respectively, but common in the KHV population (frequency >1%). We also discovered 1.5 and 0.06 million variants that were common in the global populations with frequencies >1% and >5% respectively, but were rare in the KHV population (frequency <1%).

The discrepancy in variant frequencies between the KHV and global populations implies that the global databases do not sufficiently cover the human genetic diversity in Vietnam. The results confirm the need for regional efforts to develop more comprehensive human genomic databases for Asian, especially Southeast Asian, populations.

We applied the KHV database to examine the frequency of pathological mutations in the KHV population. Pathological and likely pathological mutations were obtained from the human genome mutation database, called HGMD (Stenson et al., 2014). Most of the pathological and likely pathological mutations are rare in the KHV population. However, there are 107 pathological mutations (see Table 2) and 450 likely pathological mutations with frequency >1% in the KHV population. Noticeably, 87 out of the 107 pathological mutations have a frequency smaller than 1% in the 1KG database.

We examined the clinical significance of the 107 pathological mutations with Clinvar database (Landrum et al., 2018). Clinvar annotates only seven of them as pathogenic/likely pathogenic mutations related to thrombosis, steroid 5-alpha-reductase deficiency, retinitis pigmentosa, haemochromatosis, microcephaly global developmental delay, oligodontia, and glucose-6-phosphate dehydrogenase deficiency. Among the seven mutations, six have a frequency below 1% in the 1KG database, and only one related to the haemochromatosis disorder has a frequency of 7.3% in the 1KG database. The mutation is a nonsynonymous single nucleotide variant (NM_000410.3:c.187C>G) in gene *HFE*. We used SIFT (Ng & Henikoff, 2006) and Polyphen-2 (Adzhubei et al., 2010) programs to predict its effects on protein functions and obtained contradict results (i.e., SIFT predicted it as a damaging mutation, but Polyphen-2 predicted it as a benign mutation).

Clinvar annotates the clinical significance of the remaining mutations as benign/likely benign, conflicting interpretations of pathogenicity, uncertain significance, or not provided. More studies must be performed to evaluate the clinical significance of the 107 pathological mutations for general populations and/or for KHV population in particular. The KHV database will be beneficial for the studies.

Finally, we measured the power to detect variants of KHV-G database containing 206 KHV genomes (see Figure 1). The overall power to detect variants with frequency >1% was 99.3% (99.4% for SNPs and 98% for indels). The detection power increased to 99.9% when detecting variants with frequency >5% (99.9% for SNPs and 99.8% for indels). The variant detection power of the KHV-G database was measured for the exome regions. As variants occur less frequently in the exome regions than other regions of the genome, due to protein functional constraints, the overall variant detection power of the KHV-G database for the whole genome is expected to be higher than that for the exome regions. Note that the KHV database contains the KHV-G database, therefore, it has a greater detection power than the KHV-G database, particularly at the exome regions. The high variant detection power makes the KHV database a powerful tool to assess the functional effects of variants in medical studies.

3.3 | Population analysis

3.3.1 | Population relationship analysis

We performed PCA and phylogenetic tree reconstruction to assess the genomic relationships among populations. The KHV and 11 other populations were classified into four main groups: (a) YRI; (b) CEU; (c) South East Asian (SEA) including Malay Malaysia (MY), Filipino Philippine (PI), Javanese Indonesia (ID-JV), Tai Thailand (TAI), Kinh Vietnamese (KHV); and (d) East Asian (EA) consisting of Southern Han Chinese (CHS), Northern Han Chinese (CHB), Korean (KR), Japanese (JPT), Ryukyuan Japanese (JP-RK). Figure 2 shows the geographical locations of the 12 populations.

The PCA result displays relationships among individuals of KHV and other Asian populations in Figure 3. The plot of the first two principal components shows that individuals from the same population were clustered into one group. The KHV and TAI populations are considerably overlapped and separated from other populations. We also observe an overlap between Southern and CHB populations. The Han Chinese populations play as a bridge between SEA and EA populations. The CHS population is close to the SEA populations while the CHB population is close to the EA populations. The positions of populations along the first principal component are in

TABLE 2 Pathological mutations in the human genome mutation database with frequency >1% in the KHV population

Chrom	Position	HGVS	Gene	Phenotype	KHV (%)	1KG (%)	Clinvar
1	169555300	NM_000130.4:c.1000A>G	*F5*	Thrombosis	2.2	0.2	Pathogenic
2	31529325	NM_000348.3:c.680G>A	*SRD5A2*	Steroid 5-alpha-reductase deficiency	1.4	0.1	Pathogenic/likely pathogenic
3	170483441	NM_020949.2:c.988G>A	*SLC7A14*	Retinitis pigmentosa	1.8	0.3	Pathogenic
6	26090951	NM_000410.3:c.187C>G	*HFE*	Haemochromatosis	3.8	7.3	Pathogenic
16	70664131	NM_138383.2:c.1790C>T	*MTSS1L*	Microcephaly global developmental delay	1.6	0.1	Likely pathogenic
X	70035434	NM_001399.4:c.1001G>A	*EDA*	Oligodontia	4.2	0.5	Pathogenic
X	154533122	NM_001042351.2:c.871G>A	*G6PD*	Glucose-6-phosphate dehydrogenase deficiency	1.3	0.2	Pathogenic
1	94113062	NM_000350.2:c.71G>A	*ABCA4*	Stargardt disease	1.6	0.0	Conflicting interpretations
1	114677465	NM_000036.2:c.1373G>A	*AMPD1*	Adenosine monophosphate deaminase deficiency	2.3	0.2	Conflicting interpretations
1	183563302	NM_000433.3:c.1183C>T	*NCF2*	Chronic granulomatous disease	8.9	1.8	Conflicting interpretations
1	216097080	NC_000001.11:g.216097080T>C	*USH2A*	Usher syndrome 2	1.5	0.3	Conflicting interpretations
2	26463969	NM_194248.2:c.5098G>C	*OTOF*	Auditory neuropathy	1.5	0.2	Conflicting interpretations
2	127426114	NM_000312.3:c.565C>T	*PROC*	Protein C deficiency	2.7	0.4	Conflicting interpretations
2	166305834	NM_002977.3:c.554G>A	*SCN9A*	Small fibre neuropathy	2.3	0.6	Conflicting interpretations
2	218890244	NM_025216.2:c.637G>A	*WNT10A*	Ectodermal dysplasia	1.4	0.3	Conflicting interpretations
2	227056031	NM_000092.4:c.2630G>A	*COL4A4*	Alport syndrome	3.8	0.5	Conflicting interpretations
2	233760973	NM_000463.2:c.686C>A	*UGT1A1*	Gilbert syndrome	1.6	0.3	Conflicting interpretations
5	147828115	NM_003122.4:c.101A>G	*SPINK1*	Pancreatitis chronic	2.0	0.6	Conflicting interpretations
5	177404082	NM_000505.3:c.1027G>C	*F12*	Factor XII deficiency	2.3	0.4	Conflicting interpretations
6	135465910	NM_017651.4:c.653A>G	*AHI1*	Retinal dystrophy	2.0	0.2	Conflicting interpretations
7	107248423	NM_006348.3:c.1919T>C	*COG5*	Congenital disorder of glycosylation	2.1	0.3	Conflicting interpretations
7	117548630	NC_000007.14:g.117548630T>G	*CFTR*	Primary ciliary dyskinesia	2.0	1.0	Conflicting interpretations
7	117587820	NM_000492.3:c.1666A>G	*CFTR*	Chronic pulmonary disease	4.2	1.1	Conflicting interpretations
7	117664780	NM_000492.3:c.4056G>C	*CFTR*	Cystic fibrosis	5.1	0.4	Conflicting interpretations
7	155803420	NM_000193.3:c.869G>A	*SHH*	Holoprosencephaly	2.4	0.6	Conflicting interpretations
8	10622877	NM_178857.5:c.324_325insT	*RP1L1*	Retinitis pigmentosa	2.6	0.0	Conflicting interpretations
9	128580928	NM_001130438.2:c.1330G>A	*SPTAN1*	Intellectual disability microcephaly cerebellar atrophy	8.0	1.8	Conflicting interpretations
10	26193228	NM_017433.4:c.4462A>G	*MYO3A*	Sensorineural hearing loss with good cochlear implantation outcomes	6.2	0.9	Conflicting interpretations
10	43105159	NM_020975.5:c.833C>A	*RET*	Hirschsprung disease	1.4	0.4	Conflicting interpretations

(Continues)

TABLE 2 (Continued)

Chrom	Position	HGVS	Gene	Phenotype	KHV (%)	1KG (%)	Clinvar
10	53822914	NM_033056.3:c.4812G>T	*PCDH15*	Sensorineural hearing loss with poor cochlear implantation outcomes	1.8	0.5	Conflicting interpretations
10	53961877	NM_033056.3:c.2884C>T	*PCDH15*	Deafness	1.4	0.1	Conflicting interpretations
13	20189473	NM_004004.5:c.109G>A	*GJB2*	Deafness autosomal recessive 1	9.6	1.5	Conflicting interpretations
13	51937490	NM_000053.3:c.3889G>A	*ATP7B*	Wilson disease	2.5	0.3	Conflicting interpretations
13	100368504	NM_000282.3:c.1676G>T	*PCCA*	Propionic acidaemia	1.5	0.5	Conflicting interpretations
14	64782348	NM_001355436.1:c.4208G>A	*SPTB*	Spherocytosis	3.6	3.3	Conflicting interpretations
15	89320857	NM_002693.2:c.2890C>T	*POLG*	Lactic acidosis	2.4	0.3	Conflicting interpretations
18	31069031	NM_024422.4:c.2368_2370delGGA	*DSC2*	Arrhythmogenic right ventricular dysplasia/cardiomyopathy	2.1	0.4	Conflicting interpretations
19	8305034	NM_016579.3:c.262_264delGAG	*CD320*	Methylmalonic aciduria	1.5	0.7	Conflicting interpretations
19	35839554	NM_004646.3:c.2869G>C	*NPHS1*	Nephrotic syndrome	1.5	0.3	Conflicting interpretations
19	35848142	NM_004646.3:c.1339G>A	*NPHS1*	Congenital nephrotic syndrome Finnish type	3.4	0.7	Conflicting interpretations
19	35851666	NM_004646.3:c.65C>T	*NPHS1*	Congenital nephrotic syndrome Finnish type	1.5	0.3	Conflicting interpretations
20	63414174	NM_172107.3:c.1545G>C	*KCNQ2*	Epilepsy benign neonatal	1.5	0.6	Conflicting interpretations
X	22033015	NM_000444.5:c.10G>C	*PHEX*	Rickets hypophosphataemic	1.4	0.2	Conflicting interpretations
X	30308988	NM_000475.4:c.376G>A	*NR0B1*	Adrenal hypoplasia	4.4	0.6	Conflicting interpretations
X	67723737	NM_000044.4:c.2659A>G	*AR*	Defective spermatogenesis	1.3	0.0	Conflicting interpretations
X	108622766	NM_000495.4:c.2858G>T	*COL4A5*	Alport syndrome	6.1	0.8	Conflicting interpretations
1	169560616	NM_000130.4:c.524A>G	*F5*	Factor V deficiency	2.7	0.2	Uncertain significance
5	178986717	NM_000843.3:c.1537G>A	*GRM6*	High myopia	1.4	0.2	Uncertain significance
7	96321955	NM_014251.2:c.2T>C	*SLC25A13*	Intrahepatic cholestasis neonatal	3.2	0.5	Uncertain significance
9	2717819	NM_133497.3:c.80G>A	*KCNV2*	Cone dystrophy with supernormal rod ERG	2.2	0.3	Uncertain significance
15	68229580	NM_017882.2:c.5A>G	*CLN6*	Neuronal ceroid lipofuscinosis	1.5	0.3	Uncertain significance
17	8003211	NM_000180.3:c.164C>T	*GUCY2D*	Leber congenital amaurosis	1.5	0.2	Uncertain significance
19	35845496	NM_004646.3:c.1802G>C	*NPHS1*	Nephrotic syndrome steroid resistant	2.7	0.4	Uncertain significance
20	18510909	NM_006363.5:c.74C>A	*SEC. 23B*	Anaemia dyserythropoietic congenital type II	1.5	0.3	Uncertain significance
22	24523679	NM_016327.2:c.977G>A	*UPB1*	Beta-ureidopropionase deficiency	2.1	0.4	Uncertain significance
22	31846904	NM_001242896.1:c.3092C>A	*DEPDC5*	Epileptic spasms late-onset	1.4	0.1	Uncertain significance
X	106035377	NM_000354.5:c.631G>A	*SERPINA7*	Thyroxine-binding globulin deficiency partial	2.9	1.1	Uncertain significance

(Continues)

TABLE 2 (Continued)

Chrom	Position	HGVS	Gene	Phenotype	KHV (%)	1KG (%)	Clinvar
X	154860608	NM_000132.3:c.6724G>A	*F8*	Haemophilia A	1.5	0.1	Uncertain significance
1	183567223	NM_000433.3:c.836C>T	*NCF2*	Chronic granulomatous disease autosomal recessive	1.4	0.5	Likely benign
1	186171376	NM_031935.2:c.15614G>A	*HMCN1*	Splenic epidermoid cyst	1.7	0.5	Likely benign
1	215628906	NM_206933.2:c.15427C>T	*USH2A*	Usher syndrome	1.9	0.3	Benign
2	113062148	NM_012275.2:c.140A>G	*IL36RN*	Palmoplantar pustulosis	6.1	1.3	Benign
2	165912511	NC_000002.12:g.165912511T>C	*TTC21B*	Bardet-Biedl syndrome	1.8	0.5	Benign/likely benign
3	49530842	NM_004393.5:c.331G>A	*DAG1*	HyperCKemia & muscular dystrophy	1.4	0.4	Benign
4	102634957	NM_005908.3:c.2246T>A	*MANBA*	Nystagmus	2.1	0.3	Likely benign
4	154586230	NM_021871.3:c.1199C>T	*FGA*	Dysfibrinogenaemia	3.8	0.4	Likely benign
5	1293748	NM_198253.2:c.1138C>T	*TERT*	Cirrhosis	3.5	0.3	Benign
6	112109466	NM_002290.4:c.5422G>A	*LAMA4*	Cardiomyopathy dilated	6.1	1.5	Benign
7	117627815	NC_000007.14:g.117627815G>A	*CFTR*	Cystic fibrosis	2.0	0.3	Benign
9	95467197	NM_000264.4:c.2479A>G	*PTCH1*	Holoprosencephaly	2.2	0.2	Benign
9	133445796	NM_139025.4:c.2708C>T	*ADAMTS13*	Thrombotic thrombocytopaenic purpura	2.2	0.7	Likely benign
10	70435570	NM_018055.4:c.607G>A	*NODAL*	Situs ambiguus	2.6	0.3	Likely benign
11	61960013	NM_004183.3:c.1070C>T	*BEST1*	Retinitis pigmentosa	1.5	0.3	Likely benign
11	68908232	NM_002180.2:c.344C>T	*IGHMBP2*	Charcot-Marie-Tooth disease type 2	2.0	0.2	Benign/likely benign
12	1856044	NM_172364.4:c.2120G>A	*CACNA2D4*	Retinal dystrophy	1.7	0.7	Likely benign
13	24909892	NM_018451.4:c.763A>G	*CENPJ*	Arthrogryposis	2.5	0.5	Likely benign
15	71811966	NM_014249.3:c.361G>A	*NR2E3*	Enhanced S-cone syndrome	4.1	1.1	Likely benign
16	50711322	NM_022162.2:c.1411C>T	*NOD2*	Blau syndrome	3.0	0.3	Likely benign
16	56983380	NM_000078.2:c.1376A>G	*CETP*	Cholesterol ester transfer protein deficiency	4.2	0.6	Likely benign
16	88818027	NM_000512.4:c.1462G>A	*GALNS*	Mucopolysaccharidosis IVa	9.4	1.3	Benign/likely benign
19	7647478	NM_001272034.1:c.1696A>G	*STXBP2*	Haemophagocytic lymphohistiocytosis	2.2	1.2	Benign
19	18869245	NM_001492.5:c.468_470dupGGC	*GDF1*	Double-outlet right ventricle/Tetralogy of Fallot/Right atrial isomerism	12.0	4.1	Benign
19	38499816	NM_000540.2:c.7209C>T	*RYR1*	Multiminicore disease	20.1	3.9	Benign
20	10672955	NM_000214.2:c.133G>T	*JAG1*	Biliary atresia extrahepatic	1.7	0.2	Benign/likely benign
22	19765921	NM_080647.1:c.928G>A	*TBX1*	DiGeorge syndrome	4.4	0.9	Benign

(Continues)

TABLE 2 (Continued)

Chrom	Position	HGVS	Gene	Phenotype	KHV (%)	1KG (%)	Clinvar
X	32362879	NM_004006.2:c.5234G>A	*DMD*	Muscular dystrophy Duchenne	67.9	46.5	Benign/likely benign
X	109624780	NM_012282.3:c.241T>C	*KCNE5*	Idiopathic ventricular fibrillation	1.3	0.1	Benign
1	161213891	NM_004550.4:c.1324C>T	*NDUFS2*	Mitochondrial leukoencephalopathy	1.5	0.1	Not provided
1	171107694	NM_006894.5:c.341A>G	*FMO3*	Trimethylaminuria	1.4	0.2	Not provided
2	178071834	NM_016953.3:c.604C>T	*PDE11A*	Prostate cancer susceptibility to	3.5	1.4	Not provided
3	37519328	NM_002207.2:c.1210G>A	*ITGA9*	Severe chylothorax poor response to therapy	3.2	0.6	Not provided
5	38919015	NM_003999.2:c.1538G>A	*OSMR*	Amyloidosis primary cutaneous	2.3	0.3	Not provided
5	150648252	NM_000154.1:c.−22T>C	*SYNPO*	Glomerulosclerosis focal and segmental	1.5	0.2	Not provided
6	26093233	NC_000006.12:g.26093233G>A	*HFE*	Haemochromatosis	1.4	0.1	Not provided
6	39925702	NM_005943.5:c.394C>T	*MOCS1*	Molybdenum cofactor deficiency	1.4	0.2	Not provided
7	44147797	NM_000162.4:c.716A>G	*GCK*	Diabetes mellitus	1.5	0.0	Not provided
8	6936727	NC_000008.11:g.6936727C>A	*DEFA4*	IgA nephropathy	3.9	1.5	Not provided
9	39102658	NM_033655.3:c.2594T>C	*CNTNAP3*	Autism spectrum disorder	3.0	2.8	Not provided
10	78054951	NM_001142285.1:c.811A>C	*RPS24*	Diamond-Blackfan anaemia	3.9	0.8	Not provided
12	5994567	NM_000552.4:c.6104G>A	*VWF*	Von Willebrand disease 1	1.6	0.2	Not provided
12	55365407	NM_001005497.1:c.305delT	*OR6C75*	Reduced apolipoprotein AII levels	5.8	1.6	Not provided
17	75765158	NM_000154.1:c.−22T>C	*GALK1*	Increased GALK1 activity	1.7	0.1	Not provided
19	57328138	NM_213598.3:c.676C>G	*ZNF543*	IgA nephropathy	4.2	1.4	Not provided
20	3706494	NM_023068.3:c.262G>T	*SIGLEC1*	SIGLEC1 deficiency	1.6	0.6	Not provided
20	41345874	NM_001301860.1:c.71C>T	*LPIN3*	Rhabdomyolysis	1.5	0.2	Not provided
21	44413981	NM_003307.3:c.3053C>T	*TRPM2*	Amyotrophic lateral sclerosis and Parkinson disease	4.7	0.7	Not provided
X	64193202	NM_152424.3:c.85G>A	*AMER1*	Wilms tumour	4.6	0.7	Not provided

Abbreviation: KHV, Kinh Vietnamese.

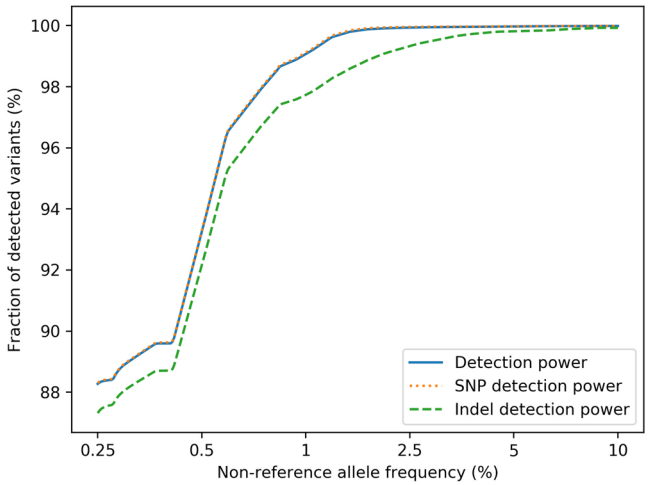

FIGURE 1 The power to detect variants of the KHV-G database. The detection function $f(x)$ represents the fraction of variants detected by the KHV-G database if they have a nonreference allele frequency greater than x% in the KHV population. The KHV-G database has an overall detection power >99.3% for variants with a frequency >1% in the KHV population. Abbreviation: KHV, Kinh Vietnamese

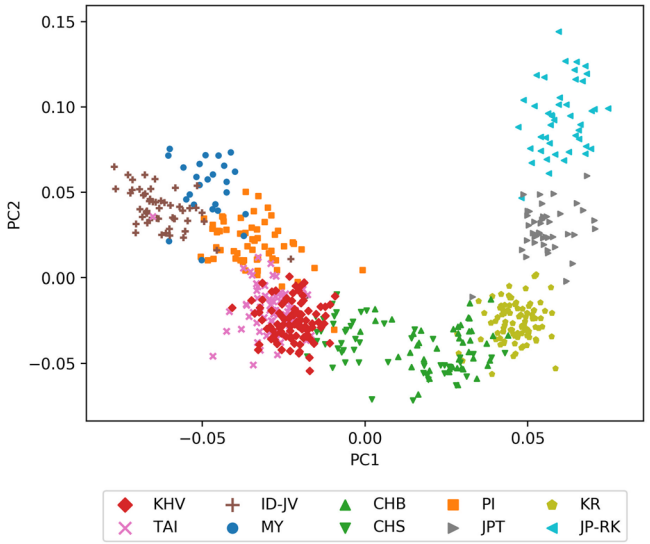

FIGURE 3 Principal component analysis of KHV and other Asian populations. CHB, Northern Han Chinese; CHS, Southern Han Chinese; ID-JV, Javanese Indonesia; JP-RK, Ryukyuan Japanese; JPT, Japanese; KHV, Kinh Vietnamese; KR, Korean; MY, Malay Malaysia; PI, Filipino Philippine; TAI, Tai Thailand

concordance with the South-to-North geographical locations of these populations.

The phylogenetic tree represents the evolutionary relationships among populations. Figure 4 represents the constructed phylogenetic trees of KHV and other populations where YRI is considered as the outgroup. The tree topologies constructed by TreeMix and neighbor joining methods are identical. All branches of the TreeMix tree have bootstrap support values of 100 indicating that the tree structure is highly reliable. The tree structures show that SEA populations are closer to the YRI and CEU than EA populations. The positions of Asian populations in the tree agree with the South-to-North ordering of their geographical locations. The results from both phylogenetic tree reconstruction and PCA support the hypothesis that a population migration from Africa entered Asia along a South-to-North route (Abdulla et al., 2009; Chu et al., 1998).

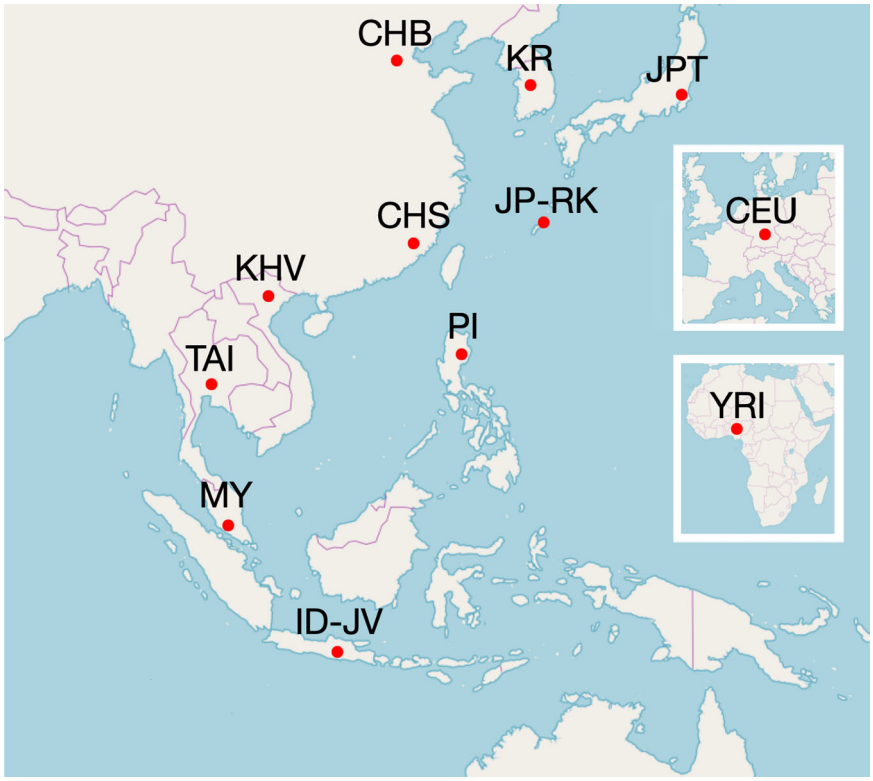

FIGURE 2 The geographical locations of 12 populations under the study: African (YRI), European (CEU), Malay Malaysia (MY), Filipino Philippine (PI), Javanese Indonesia (ID-JV), Tai Thailand (TAI), Kinh Vietnamese (KHV), Southern Han Chinese (CHS), Northern Han Chinese (CHB), Korean (KR), Japanese (JPT), and Ryukyuan Japanese (JP-RK)

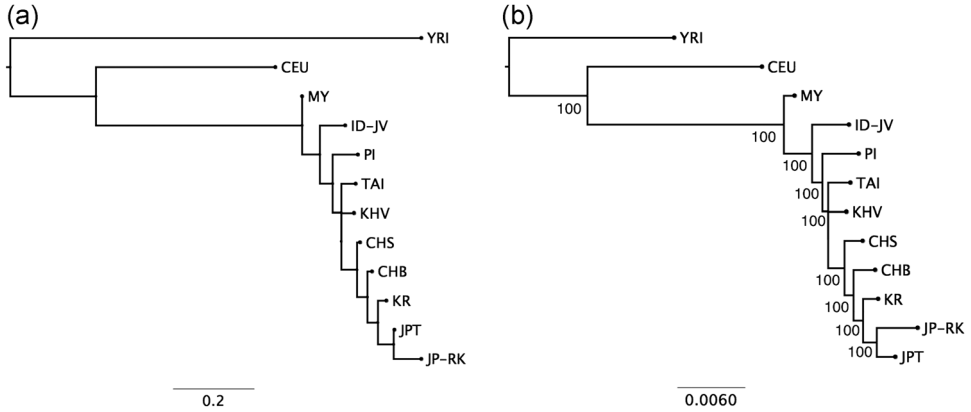

FIGURE 4 The phylogenetic trees of 12 populations where YRI is considered as the outgroup. (a) The neighbor-joining tree based on F_{ST} distances between populations. (b) The TreeMix tree where numbers on branches are bootstrap support values indicating the reliability of clades. The scale bar reflects the amount of genetic drift between populations. CHB, Northern Han Chinese; CHS, Southern Han Chinese; ID-JV, Javanese Indonesia; JP-RK, Ryukyuan Japanese; JPT, Japanese; KHV, Kinh Vietnamese; KR, Korean; MY, Malay Malaysia; PI, Filipino Philippine; TAI, Tai Thailand; YRI, African

We compared phylogenetic trees in the study with two neighbor-joining trees reported by Simons Genome Diversity Project (Mallick et al., 2016). Generally, our trees are concordant with the two neighbor-joining trees. Note that the two neighbor-joining trees are not identical. Particularly, KHV and TAI populations are adjacent in the tree based on F_{ST} distances, but not adjacent in the tree based on the pairwise divergence per nucleotide distances. Our trees support the adjacent of KHV and TAI populations.

3.3.2 | Ancestral population analysis

We analyzed the contribution of ancestral populations to the current populations. To this end, we executed fastSTRUCTURE with different K values from 2 to 10 and determined that $K = 4$ was the best choice to explain the ancestries of all populations under the study. Figure 5 shows the contribution of ancestral populations to the current populations with K values from 2 to 6 (fastSTRUCTURE did not result in any additional meaningful cluster for $K > 6$). For $K = 4$, four ancestral populations are YRI, CEU, SEA, and EA. The KHV and all other SEA populations originated mainly from the SEA ancestry, and partly from the EA and CEU ancestries (the MY population had more CEU ancestral origin than other SEA populations). We found that the KHV and TAI populations had similar ancestral population structures. The KR and JPT populations mainly derived from the EA ancestry while CHB and CHS populations were mixed from both SEA and EA ancestries. For $K = 5$, the JP-RK ancestry was separated from the EA ancestry to form one cluster, and the JP population was mixed from both EA and JP-RK ancestries. The results are generally compatible with that from the 1KG project (2015 Genomes Project Consortium et al., 2015) and the HUGO Pan-Asian SNP Consortium (Abdulla et al., 2009). We realized that the HUGO Pan-Asian SNP Consortium introduced one additional ancestry for ID-JV and MY populations and reported more contribution of the JP-RK ancestry to the KR and CHB populations.

The human population history of SEA has been long debated (Abdulla et al., 2009; Lipson et al., 2018; McColl et al., 2018; and

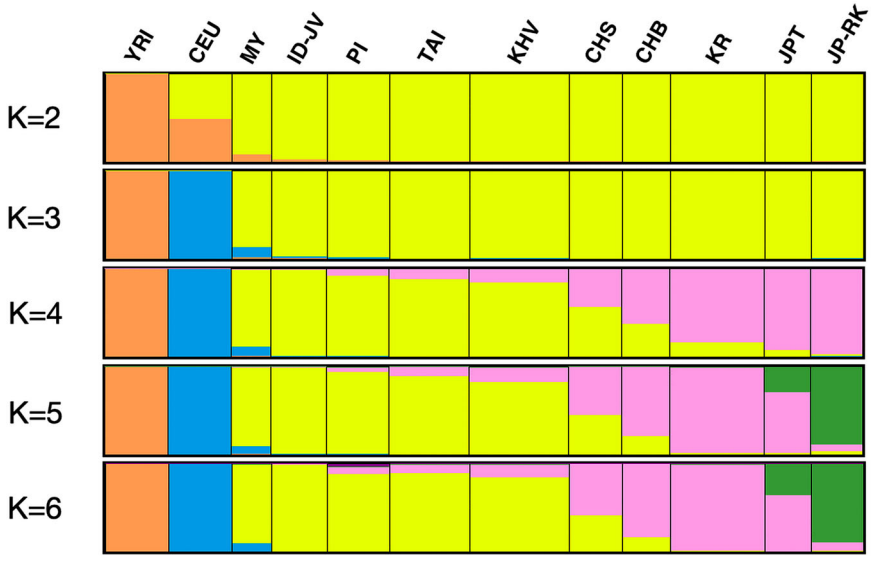

FIGURE 5 The contribution of ancestral populations to the current populations with K values from 2 to 6. The fastSTRUCTURE did not result in any additional meaningful cluster for $K > 6$. CEU, European; CHB, Northern Han Chinese; CHS, Southern Han Chinese; ID-JV, Javanese Indonesia; JP-RK, Ryukyuan Japanese; JPT, Japanese; KHV, Kinh Vietnamese; KR, Korean; MY, Malay Malaysia; PI, Filipino Philippine; TAI, Tai Thailand; YRI, African

references therein). One hypothesis posits a single migration which entered Asia along the southern, coastal route. Another hypothesis suggests additional later southward expansion of EA populations. Current studies of SEA ancient human genomes indicated that both Hòabìnhian hunter-gatherers first recognized about 44 ka years ago and southward expansion of EA farmers about 4,000 years ago influenced the diversity of present-day SEA populations (Lipson et al., 2018; McColl et al., 2018). The findings agree with our ancestral population analyses that the present-day SEA populations were mainly derived from the SEA ancestries and partly from the EA ancestries.

Finally, we examined the gene flows from other Asian populations to the KHV population using the F3 statistic test. A significant negative *F*3 (KHV; B, C) value indicates the existence of significant gene flows from populations B and C to KHV. The F3 statistic tests did not reveal any significant gene flow from Asian populations to the KHV population. The findings explain to some extent the difference between allele frequencies of KHV and other populations.

4 | CONCLUSIONS

The human genetic variation databases are typically used as a reliable tool to examine or prioritize potentially deleterious variants in genetic studies. The global databases such as the 1KG database do not sufficiently cover human genetic diversity in Asia, especially in Southeast Asia. As variant frequencies vary considerably among populations, building population genetic variation databases is needed to precisely evaluate the effects of variants in different populations.

KHV is the main ethnic group in Vietnam and one of the largest ethnic groups in Asia. Our project sequenced whole genomes and exomes of 305 unrelated KHV people and discovered 0.71 million novel variants. We combined the data with other previously published KHV genomes to create the most compressive KHV genetic variation database of 406 unrelated KHV people. The KHV database consists of nearly 25 millions of variants and can detect more than 99% of variants with frequency >1%. Thus, it could be a powerful tool to classify the effects of variants in medical studies.

Our study revealed that a considerable number of variants annotated as pathological mutations in the HGMD database had a frequency above 1% in the KHV population. Most of the mutations have discordant annotations in the Clinvar database, i.e., begin/likely benign, conflicting interpretations of pathogenicity, or uncertain significance. The findings highly suggest that the clinical significance of a variant for a specific population should be comprehensively evaluated based on annotations from different databases, functional predictions from several computational methods, and its frequency in the population under the study. The KHV database will play an important role in clinical studies for both KHV and closely related populations.

In this study, we did not determine structural variants in the KHV population. Although structural variants play an essential role in genomic studies, current computational methods to detect structural variants from genome sequencing data suffer a considerable false positive rate. Thus, called structural variants from genome sequencing data might not be reliable. We are working on other genomic approaches such as microarray-based comparative genomic hybridization to build a structural variant database for KHV and Southeast Asian populations.

Vietnam has a complex history of thousands of years. Our fine-scale genomic analyses of KHV together with other Asian populations elucidated that KHV and other SEA populations mainly derived from the same SEA ancestry. The results from different genomic analyses are generally consistent and support the hypothesis of population migration from Africa to Asia following the South-to-North route. Interestingly, we discovered that KHV and TAI populations had similar genomic structures and close evolutionary relationships. The findings suggest the usefulness of KHV database for Vietnamese as well as other closely related populations in Asia.

ACKNOWLEDGMENTS

This study is financially supported by Vinmec Healthcare System (grant number: ISC.17.03). We would like to thank Lam Nguyen for technical supports.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

ETHICS STATEMENT

The donors gave written consent for the public release of the genomic data for scientific purposes. This study was approved by the Committee on Ethics in Research on Humans of Vinmec International Hospital, Hanoi.

ORCID

Vinh S. Le http://orcid.org/0000-0002-9060-9199

REFERENCES

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Abdulla, M. A., Ahmed, I., Assawamakin, A., Bhak, J., Brahmachari, S. K., Calacal, G. C., & Zilfalil, B. A. (2009). Mapping human genetic diversity in Asia. *Science*, *326*(5959), 1541–1545. https://doi.org/10.1126/science.1177074

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, *7*(4), 248–249. https://doi.org/10.1038/nmeth0410-248

Andrews, S. FastQC: A quality control tool for high throughput sequence data., www.Bioinformatics.Babraham.Ac.Uk/Projects/Fastqc/ § (2010).

Chu, J. Y., Huang, W., Kuang, S. Q., Wang, J. M., Xu, J. J., Chu, Z. T., & Jin, L. (1998). Genetic relationship of populations in China. *Proceedings of the National Academy of Sciences*, *95*(20), 11763–11768. https://doi.org/10.1073/pnas.95.20.11763

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Hai, D. T., Thanh, N. D., Trang, P. T. M., Quang, L. S., Hang, P. T. T., Cuong, D. C., & Vinh, L. S. (2015). Whole genome analysis of a Vietnamese trio. *Journal of Biosciences*, *40*(1), 113–124. https://doi.org/10.1007/s12038-015-9501-0

Lan, T., Lin, H., Zhu, W., Laurent, T. C. A. M., Yang, M., Liu, X., & Guo, X. (2017). Deep whole-genome sequencing of 90 Han Chinese genomes. *GigaScience*, *6*(9), 1–7. https://doi.org/10.1093/gigascience/gix067

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., & Maglott, D. R. (2018). ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, *46*, D1062–D1067. https://doi.org/10.1093/nar/gkx1153

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., & MacArthur, D. G. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291. https://doi.org/10.1038/nature19057

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Lipson, M., Cheronet, O., Mallick, S., Rohland, N., Oxenham, M., Pietrusewsky, M., & Reich, D. (2018). Ancient genomes document multiple waves of migration in Southeast Asian prehistory. *Science (New York, N.Y.)*, *361*(6397), 92–95. https://doi.org/10.1126/science.aat3188

Lu, D., & Xu, S. (2013). Principal component analysis reveals the 1000 Genomes Project does not sufficiently cover the human genetic diversity in. *Asia. Frontiers in Genetics*, *4*, 127. https://doi.org/10.3389/fgene.2013.00127

MacArthur, D. G., Manolio, T. A., Dimmock, D. P., Rehm, H. L., Shendure, J., Abecasis, G. R., & Gunter, C. (2014). Guidelines for investigating causality of sequence variants in human disease. *Nature*, *508*, 469–476. https://doi.org/10.1038/nature13127

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., & Reich, D. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, *538*, 201–206. https://doi.org/10.1038/nature18964

McColl, H., Racimo, F., Vinner, L., Demeter, F., Gakuhari, T., Víctor Moreno-Mayar, J., & Willerslev, E. (2018). The prehistoric peopling of Southeast Asia. *Science*, *361*(6397), 88–92. https://doi.org/10.1126/science.aat3628

Nei, M. (1987). The Neighbor-joining Method: A new method for reconstructing phylogenetic trees. *Molecular Biology Evolution*, *4*(4), 406–425. https://doi.org/citeulike-article-id:93683

Ng, P. C., & Henikoff, S. (2006). Predicting the effects of amino acid substitutions on protein function. *Annual Review of Genomics and Human Genetics*, *7*(61), 80. https://doi.org/10.1146/annurev.genom.7.080505.115630

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., & Reich, D. (2012). Ancient admixture in human history. *Genetics*, *192*(3), 1065–1093. https://doi.org/10.1534/genetics.112.145037

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, *2*(12), e190. https://doi.org/10.1371/journal.pgen.0020190

Pickrell, J. K., & Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics*, *8*(11), e1002967. https://doi.org/10.1371/journal.pgen.1002967

Raj, A., Stephens, M., & Pritchard, J. K. (2014). FastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics*, *197*(2), 573–589. https://doi.org/10.1534/genetics.114.164350

Reich, D., Thangaraj, K., Patterson, N., Price, A. L., & Singh, L. (2009). Reconstructing Indian population history. *Nature*, *461*, 489–494. https://doi.org/10.1038/nature08365

Stenson, P., Mort, M., Ball, E., Shaw, K., Phillips, A., & Cooper, D. (2014). HGMD. *Human Genetics*, *133*(1), 1–9.

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*, 56–65. https://doi.org/10.1038/nature11632

Wong, L.-P., Ong, R. T.-H., Poh, W.-T., Liu, X., Chen, P., Li, R., & Teo, Y.-Y. (2013). Deep whole-genome sequencing of 100 Southeast Asian Malays. *American Journal of Human Genetics*, *92*(1), 52–66. https://doi.org/10.1016/j.ajhg.2012.12.005

Xue, Y., Chen, Y., Ayub, Q., Huang, N., Ball, E. V., Mort, M., & Tyler-Smith, C. (2012). Deleterious—and disease-allele prevalence in healthy individuals: Insights from current predictions, mutation databases, and population-scale resequencing. *American Journal of Human Genetics*, *91*(6), 1022–1032. https://doi.org/10.1016/j.ajhg.2012.10.015

Zhang, W., Meehan, J., Su, Z., Ng, H. W., Shu, M., Luo, H., & Hong, H. (2014). Whole genome sequencing of 35 individuals provides insights into the genetic architecture of Korean population. *BMC Bioinformatics*, *15*(Suppl 11), S6–S6. https://doi.org/10.1186/1471-2105-15-S11-S6

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Le VS, Tran KT, Bui HTP, et al. A Vietnamese human genetic variation database. *Human Mutation*. 2019;40:1664–1675. https://doi.org/10.1002/humu.23835