Received: 10 May 2021 | Revised: 31 August 2021 | Accepted: 2 September 2021

DOI: 10.1111/cge.14060

ORIGINAL ARTICLE

CLINICAL GENETICS WILEY

The Thai reference exome (T-REx) variant database

Vorasuk Shotelersuk[1,2] | Duangdao Wichadakul[3] | Chumpol Ngamphiw[4] | Chalurmpon Srichomthong[1,2] | Chureerat Phokaew[1,2] | Alisa Wilantho[4] | Sujiraporn Pakchuen[4] | Vorthunju Nakhonsri[4,5] | Philip James Shaw[6] | Rujipat Wasitthankasem[4] | Jittima Piriyapongsa[4] | Pongsakorn Wangkumhang[4] | Adjima Assawapitaksakul[1,2] | Wanna Chetruengchai[1,2] | Keswadee Lapphra[7] | Athiphat Khuninthong[4] | Pattarapong Makarawate[8] | Kanya Suphapeetiporn[1,2] | Surakameth Mahasirimongkol[9] | Nusara Satproedprai[9] | Thantrira Porntaveetus[10] | Prapaporn Pisitkun[11] | Verayuth Praphanphoj[12] | Piranit Kantaputra[13] | Wichittra Tassaneeyakul[14] | Sissades Tongsima[4]

[1]Center of Excellence for Medical Genomics, Medical Genomics Cluster, Department of Pediatrics, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand

[2]Excellence Center for Genomics and Precision Medicine, King Chulalongkorn Memorial Hospital, the Thai Red Cross Society, Bangkok, Thailand

[3]Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand

[4]National Biobank of Thailand, National Science and Technology Development Agency, Pathum Thani, Thailand

[5]Department of Biochemistry, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand

[6]National Center for Genetic Engineering and Biotechnology, National Science and Technology Development Agency, Pathum Thani, Thailand

[7]Department of Pediatrics, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand

[8]Department of Medicine, Faculty of Medicine, Khon Kaen University, Khon Kaen, Thailand

[9]Genomic Medicine Center, Division of Genomic Medicine and Innovation Support, Department of Medical Sciences, Ministry of Public Health, Nonthaburi, Thailand

[10]Genomics and Precision Dentistry Research Unit, Department of Physiology, Faculty of Dentistry, Chulalongkorn University, Bangkok, Thailand

[11]Division of Allergy, Immunology, and Rheumatology, Department of Medicine, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand

[12]Center for Medical Genetics Research, Rajanukul Institute, Department of Mental Health, Ministry of Public Health Bangkok, Bangkok, Thailand

[13]Division of Pediatric Dentistry, Department of Orthodontics and Pediatric Dentistry, Faculty of Dentistry, Chiang Mai University, Chiang Mai, Thailand

[14]Department of Pharmacology, Faculty of Medicine, Khon Kaen University, Khon Kaen, Thailand

Correspondence

Vorasuk Shotelersuk, Center of Excellence for Medical Genomics, Department of Pediatrics, Sor Kor Building 11th Floor, Faculty of Medicine, Chulalongkorn University, Bangkok 10330, Thailand.
Email: vorasuk.s@chula.ac.th

Sissades Tongsima, National Biobank of Thailand, National Science and Technology Development Agency, Thailand Science Park, Phahonyothin Road, Khlong Nueng, Khlong Luang, Pathum Thani 12120 Thailand.
Email: sissades.ton@nstda.or.th

Funding information

Chulalongkorn University, Grant/Award Number: 764002-HE01; Health Systems Research Institute, Grant/Award Number: 63-113; National Research Council of

Abstract

To maximize the potential of genomics in medicine, it is essential to establish databases of genomic variants for ethno-geographic groups that can be used for filtering and prioritizing candidate pathogenic variants. Populations with non-European ancestry are poorly represented among current genomic variant databases. Here, we report the first high-density survey of genomic variants for the Thai population, the Thai Reference Exome (T-REx) variant database. T-REx comprises exome sequencing data of 1092 unrelated Thai individuals. The targeted exome regions common among four capture platforms cover 30.04 Mbp on autosomes and chromosome X. 345 681 short variants (18.27% of which are novel) and 34 907 copy number variations were found. Principal component analysis on 38 469 single nucleotide variants present worldwide showed that the Thai population is most genetically similar to East and Southeast Asian populations. Moreover, unsupervised clustering revealed six Thai

© 2021 John Wiley & Sons A/S. Published by John Wiley & Sons Ltd

Thailand; National Science and Technology Development Agency, Grant/Award Number: P-19-51269; Thailand Research Fund, Grant/Award Numbers: DPG6180001, RSA6280001; TSRI Fund, Grant/Award Numbers: CU_FRB640001_01_30_10, CU_FRB640001_01_32_3, CU_FRB640001_01_32_4

subpopulations consistent with the evidence of gene flow from neighboring populations. The prevalence of common pathogenic variants in T-REx was investigated in detail, which revealed subpopulation-specific patterns, in particular variants associated with erythrocyte disorders such as the HbE variant in *HBB* and the Viangchan variant in *G6PD*. T-REx serves as a pivotal addition to the current databases for genomic medicine.

KEYWORDS

database, genomic medicine, population genetics, precision medicine, rare disease, Thai

1 | INTRODUCTION

Genomic medicine is an emerging area in which diagnosis and treatment can be tailored to individuals by using their genomic information. With the advancement of high throughput next generation sequencing (NGS), genome sequencing could soon become a common tool for public health services. Exome sequencing (ES) concentrates on sequencing human protein-coding regions to identify disease-causing variants. To exclude common variants not associated with disease, large-scale aggregated variant databases and functional prediction software tools are used for variant prioritization. Generally, variants predicted as deleterious are good disease candidates if they are very rare in a reference variant database. The gnomAD[1] reference variant database harbors a dense collection of variants collected from ES and genome sequencing. However, accurate estimation of allele frequencies (AF) for Asian peoples is limited by inadequate sampling in gnomAD. To address this discrepancy, ethnic-specific variant databases have been constructed for Asian populations, including the very recently described pilot phase of the GenomeAsia 100 K project,[2] Korean,[3,4] and Singapore[5] databases. Notably, the GenomeAsia 100 K project, which included 1739 individuals of 219 population groups and 64 countries, did not include Thai individuals. Here, we report protein-coding variants analyzed from ES of 1092 Thai individuals and compiled them as the Thai Reference Exome (T-REx) variant database.

2 | METHODS

2.1 | Cohort and sample preparation

A cohort of 1119 unrelated Thais was recruited from six different hospitals for ES (Supplementary Table S1). The cohort comprises pediatric patients with various rare diseases or unaffected parents. The majority of patients attended the Genetics Clinic of the King Chulalongkorn Memorial Hospital (KCMH), tertiary referral center-accepting patients from all over the country. For trios, we sampled the unaffected parents. We also obtained a minority of patient samples from non-trios. This project was approved by the Institutional Review Board of each group with the written informed consent provided by the participants. De-identification of all participants is ensured, conforming to the relevant guidelines and regulations. The data were processed using quality controls as follows: ≥95% of genotyping quality, removal of first-degree relatives (IBD PIHAT >0.5), and removal of de-identified non-Thai outliers based on principal component analysis (PCA). Data of 1092 unrelated individuals passed quality controls (Supplementary Figure S1).

2.2 | Variant calling and filtering

We employed GATK best practices (version 3.7[6]) to analyze the genotypic data. Reads were aligned to the reference sequence (hg19) using BWA (version 0.7.15) and duplicate reads were removed with Picard (version 2.9.0). HaplotypeCaller was used to identify individual variants in GVCF format. Genotyping was done using the GenotypeGVCFs tool on the combined GVCF file. Chromosome X variants from 557 males were identified by setting ploidy equal to one in HaplotypeCaller. The utility "bedtools" was used to identify regions common to the four capture kits employed with different exome coverages, that is, SureSelect V5 (36.79 Mbp), Nextera exome v.1.2 (45.39 Mbp), SureSelect V4 (51.32 Mbp), and SureSelect V6 + UTR (90.78 Mbp). The common regions span 30 040 841 bp from 248 528 exons (60 250 exons are not included). We excluded the mitochondrion from analysis as the exome capture libraries used in our experiments do not uniformly contain probes for this genomic region.

2.3 | Population structure analysis

A total of 211 378 autosomal biallelic single nucleotide variant (SNVs) were used in population structure analysis. We combined genotypic data from 2504 individuals in 1000G with our dataset of 1092 individuals in population analyses. We removed SNVs in linkage disequilibrium (r^2 > 0.2), with minor allele frequency <1%, missing genotype >2% or not in Hardy-Weinburg equilibrium (p-value < 0.001) as shown in Supplementary Figure S1. 38 469 SNVs were retained after filtering. 3596 individuals from the combined datasets were clustered using Iterative Pruning to CAPture fine-scale Structure (IPCAPS[7]) with a stopping threshold of 0.1. ADMIXTURE[8] was used to profile ancestry mixture ratios.

2.4 | Variant effect prediction

Ensembl Variant Effect Predictor[9] (VEP version 95) and dbNSFP[10] (version 3.5) were used to annotate effects of variants based on the hg19 reference sequence.

2.5 | Copy number variation analysis

We grouped the ES data into four subsets according to the capture platform library, each of which was copy number variation (CNV)-called using CODEX2 (https://github.com/yuchaojiang/CODEX2) with default settings and later combined the CNV results. Known CNVs were obtained from Database of Genomic Variants (DGV)[11] based on the hg19 reference sequence. The overlaps between CNV segments called from the data and DGV variants were identified using an in-house script (https://github.com/cucpbioinfo/T-REx-CNV-scripts). If a called CNV segment overlapped ≥50% with a DGV variant, it was considered “known” and “novel” otherwise. We extracted the frequencies of amplifications and deletions within gene regions annotated by GENCODE[12] (release 30).

3 | RESULTS AND DISCUSSION

3.1 | Data quality control

We collected ES data of 1119 unrelated Thai individuals recruited from six centers with majority located in the central region (Supplementary Table S1). In Thailand, we have limited number of clinical geneticists and majority of them are working in Bangkok.[13,14] Thus, most rare-disease patients would attend hospitals in Bangkok either by walk-in or via a referral system from their local healthcare providers. The cohort is therefore representative of individuals residing in different parts of the country.

Four different exome capture libraries were employed which showed negligible batch effect (Supplementary Figure S2). The average sequencing depths per capture library were 56x for Nextera exome v1.2, 61x for SureSelect V4, 63x for SureSelect V5 and 45x for SureSelect V6 + UTR. From reads aligned to the common regions of the capture libraries, 8 661 168 variants on autosomes and chromosome X were initially identified, and after variant quality adjustment in GATK (VQSR), the number of variants passing filter was 7 762 541.

We performed population quality control (Pop-QC) on 1119 sampled individuals, in which no first-degree relative was detected. Three individuals were excluded owing to poor genotyping quality. Additionally, IPCAPS[7] identified 24 outlier individuals whose projections on the first two principal components (PC1 and PC2) show clear separation from the majority of the Thai individuals, that is, suspected non-Thai or admixed genetic background. These 24 non-Thai outliers were also removed from further analyses. Therefore, 1092 individuals (557 males and 535 females) were qualified for further analyses. From these individuals, we identified 338 544 autosomal variants (average GATK's Genotype Quality [GQ] of 91.35) and 7137 variants on chromosome X (average GQ of 95.69). The summarized workflow of Pop-QC is shown in Supplementary Figure S1.

Since we included participants who were either patients with suspected genetic diseases or their parents, the frequency of pathogenic variants in T-REx could be inflated over what is present among the general population. We identify pathogenic/likely pathogenic variants present among known rare disease genes related to the indication for testing in 108 patients out of 1092 individuals. However all of these variants were private except for variants among six genes (*ELANE*, *SCN1A*, *TGFBI*, *COLQ*, *ITGA2B*, and *PIGA*) (Supplementary Table 2). Moreover, among these genes, only the c.393 + 1G > A variant in *COLQ* has an allele count greater than two with the population frequency of 0.003. Therefore, we believe the T-REx database is valid for filtering in genomic medicine if alleles with frequencies of >0.01 are considered as common variants not causal of rare diseases.

3.2 | Population genetic analyses

To determine the genetic relationship of the Thai population with respect to others, we employed several population genetic analyses. Using a set of 38 469 markers found throughout world populations, PCA[15] showed that 1092 Thai individuals occupied positions in the PCA space distinct from other populations in a continuous gradient or a cline, suggesting a non-homogenous, sub-structured population (Figure 1A).

Using all available markers qualified for population analysis (minor allele frequency greater than 1%), we employed unsupervised clustering software, IPCAPS, to assign individuals collated from multiple datasets into homogeneous clusters. Multiple runs of IPCAPS revealed a consensus topology of 52 clusters (Supplementary Table 3). Thai individuals were assigned to 25 clusters of which nine were major (≥20 individuals). Clusters with fewer than 20 individuals could represent subpopulations with insufficient sampling and were excluded from further analyses. The finding of subpopulation structure among Thais is consistent with a previous study using SNP array data;[16] however, the exome data with more informative markers reveals finer differences in population structure. We performed ADMIXTURE[8] analysis (varying from K = 1–35) to generate Q-matrices representing each individual's admixture ratio. K = 10 gave the minimum cross-validation error (Supplementary Figure S3), suggesting that 10 admixture components best describe the admixture patterns in the data. We plotted the Q-matrices from K = 10 analysis with individuals grouped according to the clusters assigned by IPCAPS (Figure 1B). The groups assigned by IPCAPS show distinct admixture patterns that generally follow ethno-geographic labels of assigned individuals.

Next, we examined the Thai population substructure in more detail. The admixture patterns among six Thai subpopulations (A, B, C, D, E, and X) inferred from IPCAPS cluster assignments are shown in Figure 2. Subpopulations C and X are composed solely of Thai individuals and show unique admixture patterns with no conspicuous similarity to individuals from any other Asian population. However, subpopulations A, B, D, and E include Thais and other Asian

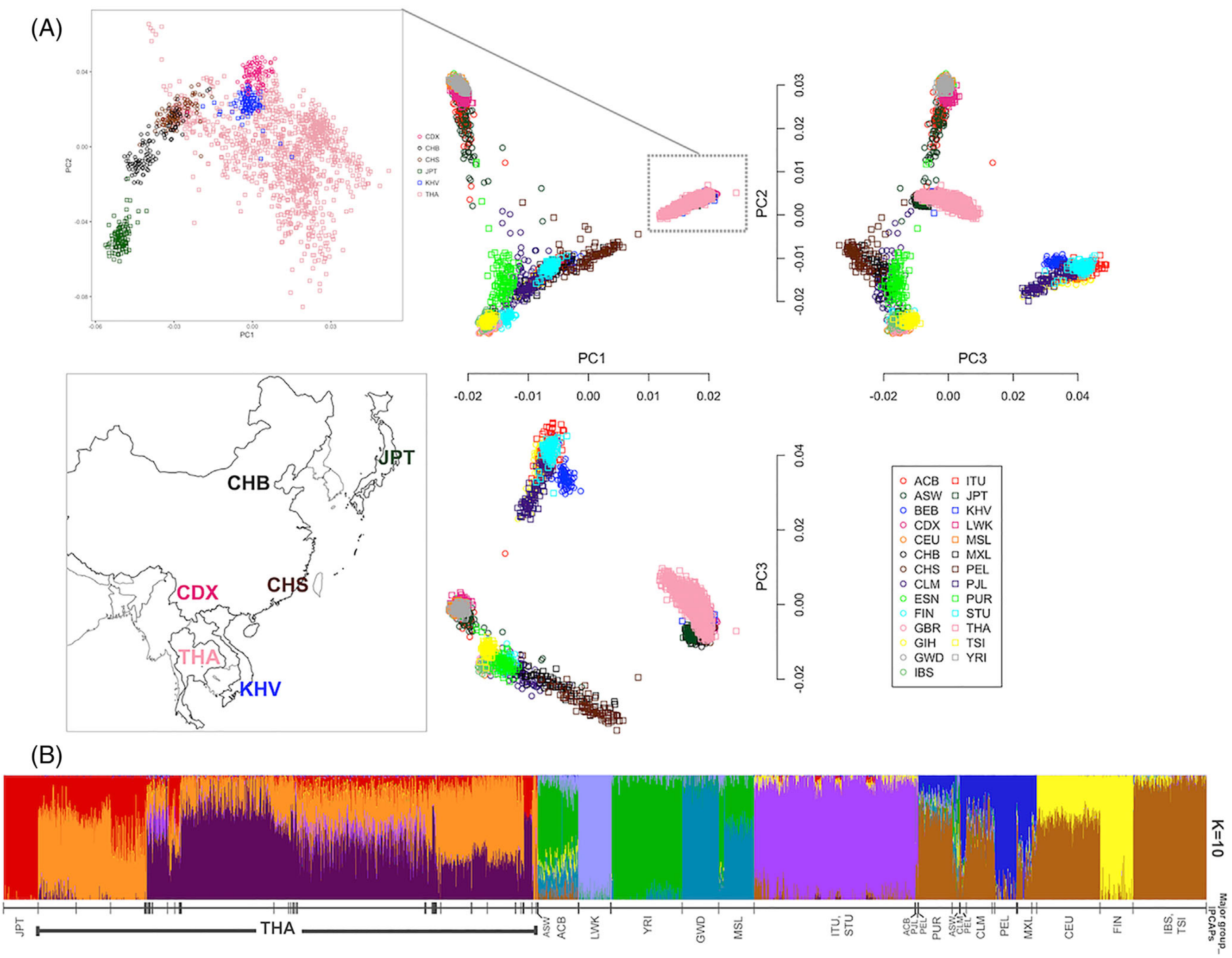

FIGURE 1 Population analysis of 1092 T-Rex individuals compared with 2504 individuals sampled from 26 populations in 1000G. (A) Principal component analysis (PCA) with separate plots shown for PC1 versus PC2, PC2 versus PC3, and PC1 versus PC3. Individuals are color-coded according to their ethno-geographic labels as indicated in the key on the bottom right. An expanded view of the PCA space occupied by East and Southeast Asian populations (boxed region in the PC1 vs. PC2 plot) is shown on the top left, and the geographical locations of these populations are shown on the bottom left. The IPCAPS tool was used for unsupervised cluster assignment of individuals. (B) Graphical representation of ancestry modeled by ADMIXTURE. The results are shown for the minimum cross-validation error reported by ADMIXTURE ($K = 10$), in which the proportion of each inferred ancestral component is indicated by different colors for each individual. Individuals were grouped according to clusters assigned by IPCAPS (Supplementary Table 2). Clusters with assigned Thai individuals are indicated by the brace (THA) below the plot. Pophelper version 2.2.6 implemented in R was used for visualization of admixture patterns. ACB, African Caribbean in Barbados; ASW, African Ancestry in Southwest US; BEB, Bengali in Bangladesh; CDX, Chinese Dai in Xishuangbanna, China; CEU, Utah residents with Northern and Western European ancestry; CHB, Han Chinese in Beijing, China; CHS, Southern Han Chinese, China; CLM, Colombian in Medellin, Colombia; ESN, Esan in Nigeria; FIN, Finnish in Finland; GBR, British in England and Scotland; GIH, Gujarati Indian in Houston, TX; GWD, Gambian in Western Division, The Gambia; IBS, Iberian Population in Spain; IPCAPS, Iterative Pruning to CAPture fine-scale Structure; ITU, Indian Telugu in the UK; JPT, Japanese in Tokyo, Japan; KHV, Kinh in Ho Chi Minh City, Vietnam; LWK, Luhya in Webuye, Kenya; MSL, Mende in Sierra Leone; MXL, Mexican Ancestry in Los Angeles, California; PEL, Peruvian in Lima, Peru; PJL, Punjabi in Lahore, Pakistan; PUR, Puerto Rican in Puerto Rico; STU, Sri Lankan Tamil in the UK; THA, Thai in Thailand; T-REx, Thai reference exome; TSI, Toscani in Italy; YRI, Yoruba in Ibadan, Nigeria [Colour figure can be viewed at wileyonlinelibrary.com]

individuals. In addition to 114 Thais, subpopulation A includes the majority of Han Chinese (CHS) and Han Chinese from Beijing (CHB) individuals. Thais assigned to subpopulation A are thus likely to be Chinese descendants, who constitute approximately 14% of the Thai population.[17] The Chinese descent can be attributed to large-scale migration from southern Chinese ports to Bangkok, in which over 1 million immigrants arrived from 1882 to 1910.[18,19] Subpopulations B, D, and E contain Xishuangbanna (CDX) and Vietnamese Kinh (KHV) individuals in addition to Thais, with the majority of CDX and KHV assigned to subpopulation E. The distinctive admixture patterns of subpopulations B, D, and E support the cluster assignments by IPCAPS. The presence of genetically distinct subpopulations comprising Thais and individuals of other Asian ethno-geographic origins suggests migration is a major factor for population stratification. Population expansion across mainland

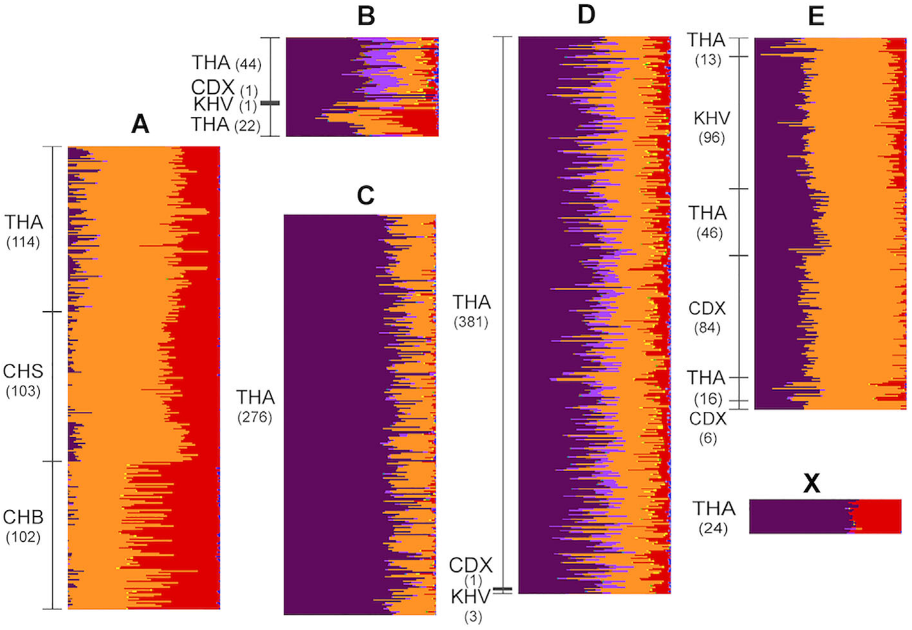

FIGURE 2 Thai subpopulation structure patterns. Six Thai subpopulations (A, B, C, D, E, and X) were inferred from unsupervised clustering analysis using IPCAPS (Supplementary Table 2). Ancestry components modeled by ADMIXTURE (K = 10) as shown in Figure 1 are shown here in an expanded view to highlight differences in admixture patterns and ethno-geographic origins of assigned individuals among the subpopulations. The numbers of assigned individuals with the same ethno-geographic origin are shown, including Thai (THA), Southern Han Chinese (CHS), Han Chinese in Beijing (CHB), Chinese Dai in Xishuangbanna (CDX), and Kinh in Ho Chi Minh City, Vietnam (KHV) [Colour figure can be viewed at wileyonlinelibrary.com]

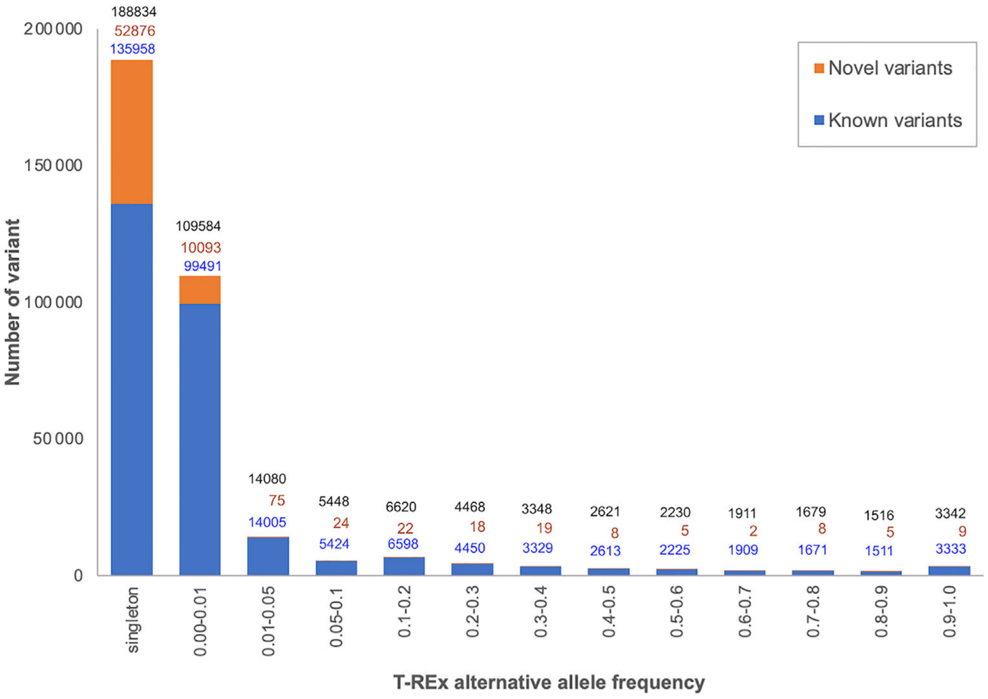

FIGURE 3 Distribution of variants according to alternative allele frequencies. Variants are binned according to alternative allele frequencies. The number of known, novel and total variants are indicated above each bin [Colour figure can be viewed at wileyonlinelibrary.com]

Southeast Asia is historically attributed to the migration of Dai (Tai) people. Dai-speaking peoples originated in the sixth century BCE south of the Yangzi River and migrated into northern Thailand settling in mountain river basins.[19]

3.3 | Variant analysis

Next, we surveyed AF of all 345 681 qualified variants on autosomal and X chromosomes. The variants comprise 337 200 SNVs, 2661 insertions, 4929 deletions, and 891 sequence alterations that were classified according to Sequence Ontology.[20] The number of variants and their alternative allele frequency distributions called from GATK HaplotypeCaller[6] are shown in Figure 3. Comparison of the 345 681 variants with dbSNP database build 151 (http://www.ncbi.nlm.nih.gov/SNP/), NHLBI Exome Sequencing Project (ESP) release 2014110 (http://evs.gs.washington.edu/EVS/), COSMIC v86,[21] and gnomAD release 2.1[1] revealed that 63 164 (18.27%) are novel. 52 876 (83.71%) of the novel variants are found in only one individual and 10 093 (15.98%) are rare with alternative allele frequency ≤ 0.01.

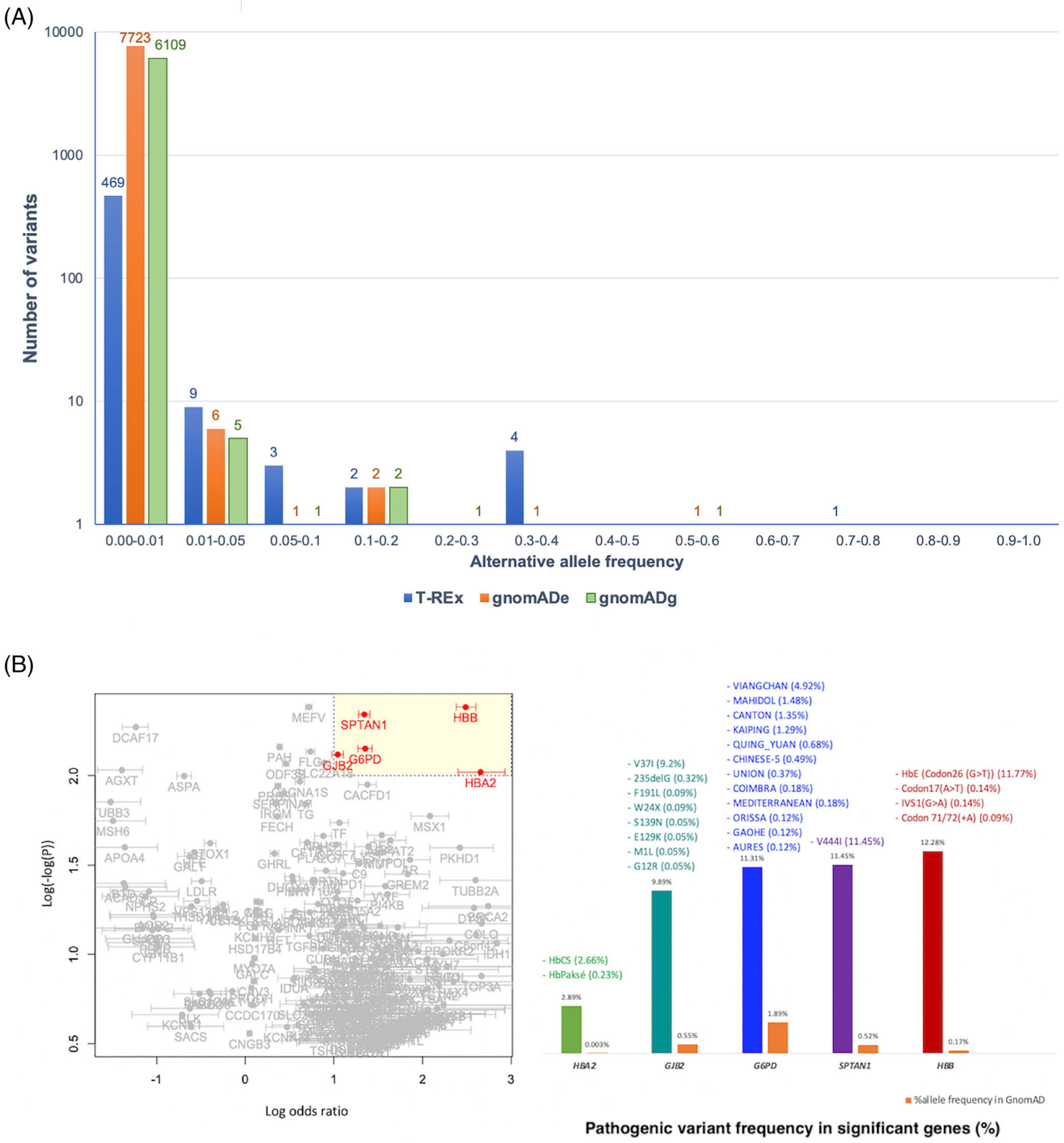

FIGURE 4 High-frequency pathogenic variants in the Thai population. (A) Distribution of variants in T-REx classified as pathogenic in ClinVar, compared with those in gnomAD. (B, left) Comparison of predicted pathogenic variant allele frequencies among Thai and gnomAD populations. The combined log odds ratio of variants for each gene comparing prevalence among the sampled Thai population (N = 1092) with aggregated gnomAD populations is shown on the X-axis. The Y-axis represents the *p*-value (log(−log[*p*]) from Fisher's exact test. Genes highlighted in red harbor pathogenic variants that are significantly more prevalent in the Thai population compared with gnomAD and with log odds ratio greater than 1. (B, right) Allele frequencies of pathogenic variants in genes with significantly higher prevalence in the Thai population. The combined frequency of all pathogenic variants is shown by barplots for each gene and the frequencies of known pathogenic variants in the Thai population are indicated above. T-REx, Thai reference exome [Colour figure can be viewed at wileyonlinelibrary.com]

3.4 | Functional annotation of coding variants

We investigated the phenotypic impact of the 345 681 variants, of which 341 831 are biallelic (Supplementary Table 4 and Supplementary Figure S4). We categorized them according to putative impact on protein function and functional consequence.[9,22] 8368 variants (of which 2815 [33.64%] are novel) are annotated as splice acceptor, splice donor, stop gained, frameshift, stop lost, or start lost. These high-impact variants could cause protein truncation, loss of protein function, or trigger nonsense-mediated RNA decay. 186 128 variants were predicted as moderate impact including in-frame insertion, in-frame deletion, missense (non-synonymous), and protein-altering variants of which 36 332 (19.52%) are novel. This variant annotation catalog includes not only descriptions of known variants that can be found elsewhere, but also the frequencies that are important for genomic medicine of the Thai population. The rare missense variants show more damaging effects on protein function (reduced SIFT score and increased PolyPhen score)[10] (Supplementary Figure S4b). The majority of novel missense variants are also rare (Supplementary Figure S4c).

3.5 | Ethno-specific pathogenic variants

The analysis to delineate and compare the distribution of variants classified as pathogenic in ClinVar in T-REx, gnomAD exome (gnomADe) and gnomAD genome (gnomADg) databases was performed. To be comparable, only the common regions of the four exome capture libraries used in T-REx were investigated. Variations defined as pathogenic were those with more than 50% reported as pathogenic or likely pathogenic by the ClinVar submitters. Overall, there are 488 alleles (sum of all variant AF: 3.04), 7734 alleles (sum of all variant AF: 1.69), and 6119 alleles (sum of all variant AF: 1.67) in T-REx, gnomADe and gnomADg, respectively, as shown in Figure 4A. The T-REx database is more enriched for pathogenic variants than gnomAD. This is probably due to its recruitment criteria, which enrolled parents of patients with rare diseases. Therefore, ones must be cautious when using T-REx for variant-interpretation in the context of Mendelian diseases.

Notably, 469 out of 488 (96%) pathogenic alleles in T-REx have AF < 0.01. Therefore, if AF < 0.01 is used as a variant filtering criterion when looking for an underlying variant for a rare Mendelian disorder, these 469 alleles would not be filtered out. The remaining 19 pathogenic variants in T-REx with AF > 0.01 are shown in Supplementary Table 5.

In addition, to explore possible genetic disorders with high-local prevalence, all known pathogenic variants from ClinVar 2019[23] were investigated in terms of their prevalence among the Thai population compared with others from gnomAD.[1] Fisher's exact test was used to identify variants with statistically higher prevalence (odds ratio) than expected (Figure 4B). From the test, we grouped these variants by genes and there are five genes (*HBA2*, *HBB*, *G6PD*, *SPTAN1*, and *GJB2*) with pathogenic variant frequencies higher than other populations reported in gnomAD (cut-offs: log odds ratio > 1 and *p*-values

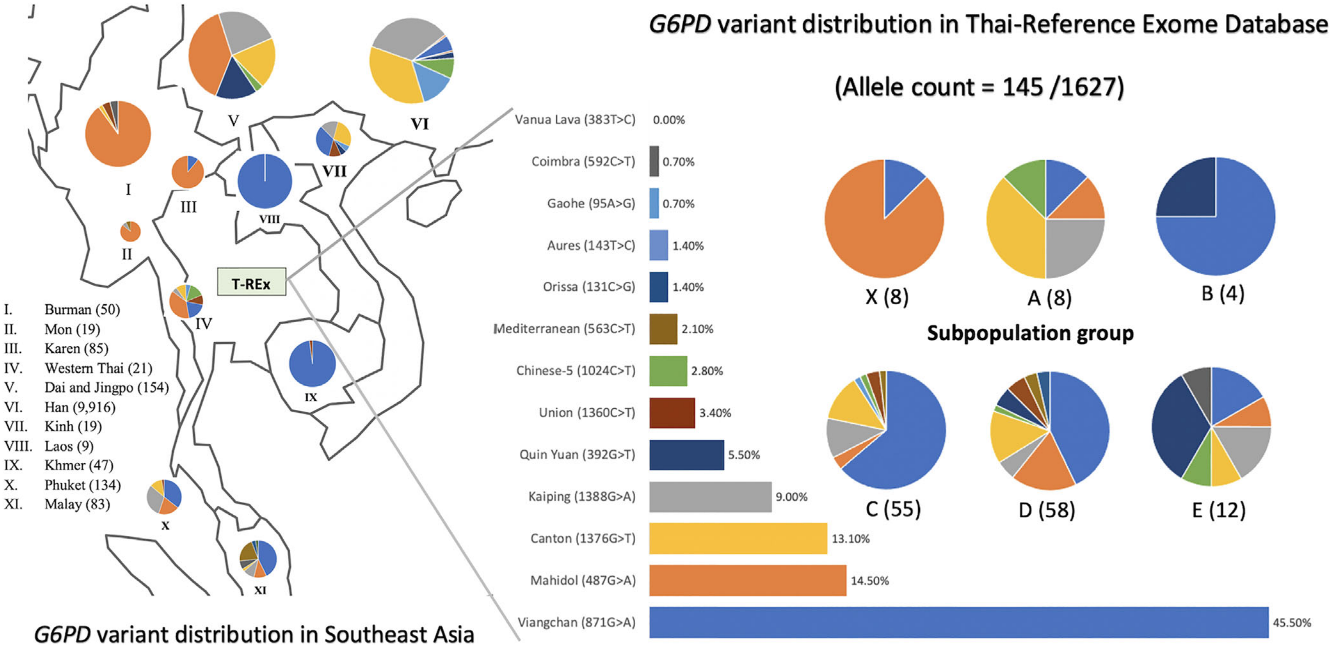

FIGURE 5 G6PD-variant distribution in the Thai population compared with neighboring populations. Frequencies of major *G6PD* variants among 1092 T-REx individuals are plotted in the main bar chart. To the right of the bar chart, variant frequencies are shown among A, B, C, D, and X subpopulations as pie charts. Pie-chart patterns of G6PD variants from the neighboring countries (data obtained from previous reports) are shown on the left with color matching of the variants. The map was created using Thematic Maps version 2.3–1 implemented in R. T-REx, Thai reference exome [Colour figure can be viewed at wileyonlinelibrary.com]

$<1 \times 10^{-100}$). Pathogenic variants in *SPTAN1* causes intellectual disability and the only variant of *SPTAN1* identified in T-REx was rs77358650 (p.V444I) with the allele frequency of 11.5%. Given its high frequency, it is unlikely that p.V444I is pathogenic as previously suggested.[24] The pathogenicity of the p.V37I variant in *GJB2* with similarly high frequency (9.2%) is equivocal.[25] If this allele is excluded, the frequency of pathogenic variants in *GJB2* is in to that of gnomAD. Interestingly, several high-prevalence pathogenic variants include those in genes with well-known variants associated with malaria-related hematological disorders in this region, namely glucose-6-phosphate dehydrogenase (*G6PD*) deficiency and thalassemia associated with *HBA2* and *HBB* gene variants (Figure 4B). The details of known G6PD-deficiency and thalassemia variant prevalence in T-REx are shown in Supplementary Table 6.

We explored the prevalence of *G6PD* in more detail among southeast Asain populations (Figure 5), owing to the great diversity in pathogenic variants known for this gene.[26] G6PD Viangchan is the most common pathogenic variant in the Thai population, in which subpopulation C habors the greatest frequency. This variant is commonly found among non-Chinese populations in Cambodia and Laos. Subpopulation X (N = 24) comprises individuals with the highest prevalence of the G6PD Mahidol variant (7 out of 24 individuals) among all subpopulations with assigned Thai individuals. The G6PD Mahidol variant is predominant among individuals from Myanmar and the western part of Thailand.

Subpopulation E harbors rare G6PD deficiency variants, including Mahidol, Canton, Kaiping, and Quing Yan. The Quing-Yan G6PD variant is highly prevalent among the Dai people (CDX) in Yunnan province,[27] further supporting the population genetic analyses (Figure 2) that subpopulation E originated from Dai migrants to northern Thailand.

3.6 | CNV in T-REx

Among the 1092 Thai individuals, we found a total of 34 907 putative CNV segments. The average number of CNVs per individual is 15.3 for amplification and 16.7 for deletion (Supplementary Table 7). Major CNVs were found only in one individual (Figure 6A). Most of the remaining CNVs were found in 2–5 individuals. CNV length varies

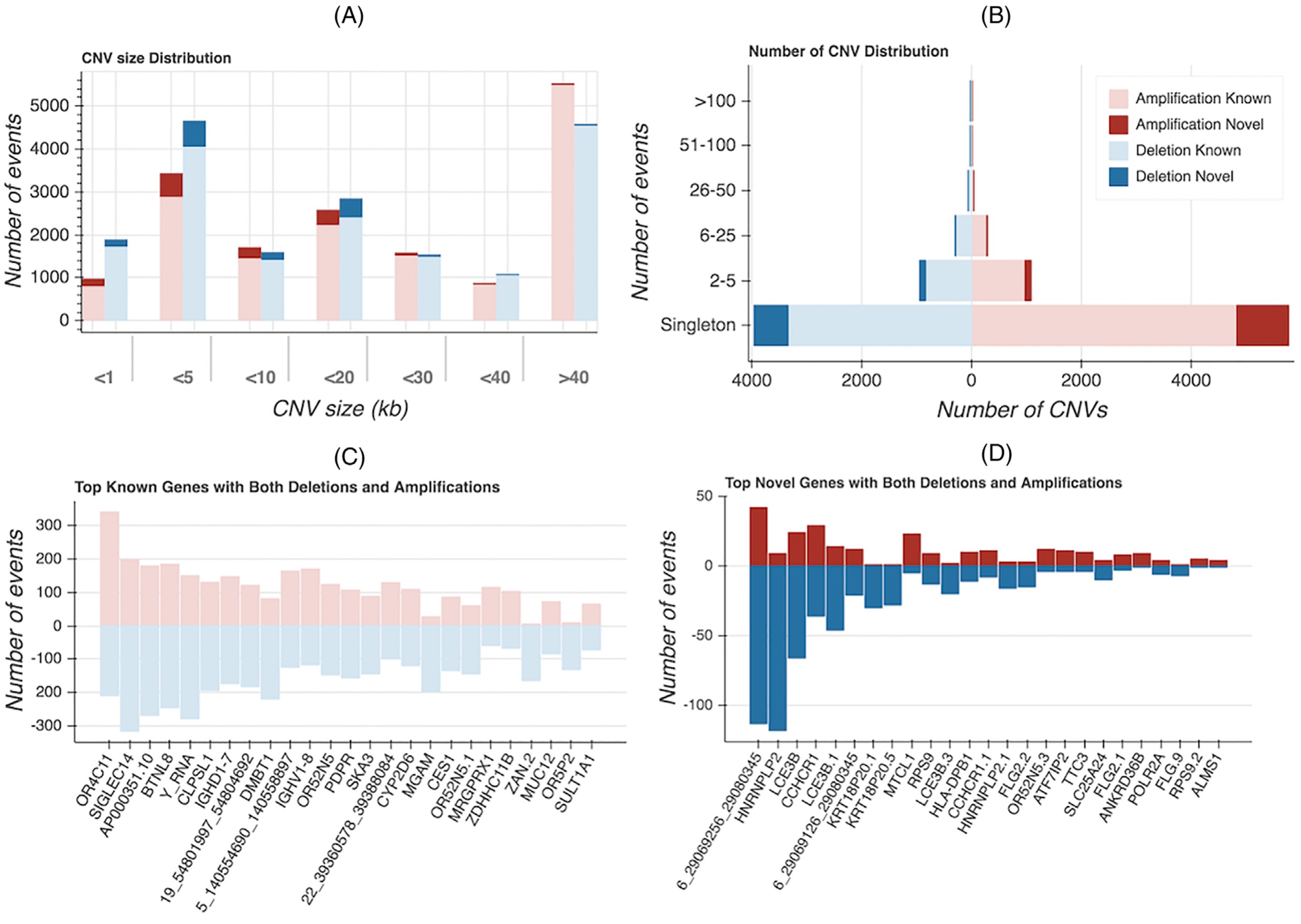

FIGURE 6 Copy number variants (CNVs) in the T-REx database. (A) Distribution of CNV sizes. (B) Frequency of CNVs by number of events. (C) Highly polymorphic CNV genes overlapping with those reported in DGV. (D) Highly polymorphic CNV genes that do not overlap with DGV variants. DGV, Database of Genomic Variants; T-REx, Thai reference exome [Colour figure can be viewed at wileyonlinelibrary.com]

(Figure 6B), and 8.5% of CNVs in T-REx are novel compared with DGV[11] (Supplementary Table 8). Novel CNVs are mostly small (less than 5 kb).

OR4C11, *SIGLEC14*, *AP000351.10*, *BTNL8*, *Y_RNA* are the most highly polymorphic genes in T-REx (Figure 6C). *OR4C11* is located in a cluster of three OR genes (*OR4C11*, *OR4P4*, and *OR4S2*) and two pseudogenes (*OR4V1P* and *OR4P1P*) previously reported to encompass a large common biallelic deletion on chromosome 11.[28] The specific pattern of deletions, however, varies and can distinguish African from non-African populations.[28] *SIGLEC14* was reported as the most highly polymorphic CNV in the Korean population,[3] and its allelic frequency is more common among Asian populations compared with African and European (Supplementary Figure S5).[29] *AP000351.10* is a pseudogene. *BTNL8* was reported as a molecule involved with stimulating the primary immune response.[30] A previous study reported that the deletion allele (*BTNL8_BTNL3-del* allele) is common among Asian, American, and European populations, but is infrequent among African and Oceanic populations.[31] CNVs overlapping the *Y_RNA* gene are common in T-REx, although the phenotypic consequences of these variants are unknown. The immunoglobulin heavy-chain locus variable (IGHV) and density (IGHD) genes were also among the most highly copy-number polymorphic loci. Several of them were shown to be highly polymorphic between African and Asian/European populations.[32] Novel genes with high copy-number polymorphism are shown in Figure 6D. The deletions of *LCE3B* and *LCE3C* were reported as a susceptibility factor for psoriasis[33] and rheumatoid arthritis.[34] *CCHCR1* was also reported to be associated with Psoriasis susceptibility.[35] Its function, however, is still unknown.

In conclusion, the T-REx database represents the first large-scale survey of exome variants for the Thai population, including frequencies of variants of functional relevance. The data support previous Thai population genetic study, in which the two largest subpopulations in this study (C and D) are distinct from other Asian populations in terms of admixture pattern. However, subpopulations of Chinese and Dai descendants were identified, suggesting migration is an important factor for substructure. The prevalence of pathogenic variants differs markedly among subpopulations, in which known variants associated with G6PD deficiency are highly prevalent in some subpopulations. Novel variants were also identified, including CNVs in highly polymorphic genes. The catalog of variants in the T-REx database is a valuable resource for population genetic and genomic medicine, especially for rare and undiagnosed disease variant prioritization, for not only Thais but also for other Southeast Asian populations.

ACKNOWLEDGMENTS

This study was partially funded by Health Systems Research Institute (HSRI) (63-113) under the Genomics Thailand project, TSRI Fund (CU_FRB640001_01_30_10, CU_FRB640001_01_32_3, CU_FRB640001_01_32_4), the Ratchadapisek Sompoch Endowment Fund (2021) of Chulalongkorn University (764002-HE01), National Research Council of Thailand, Thailand Research Fund (DPG6180001, RSA6280001) and the NSTDA grant number P-19-51269. We thank Mr. Krittin Phornsiricharoenphant for setting up computing environment for deploying the T-REx database.

CONFLICT OF INTEREST

The authors declare no competing interests.

AUTHOR CONTRIBUTIONS

Vorasuk Shotelersuk and Sissades Tongsima initiated the project and conceived the idea in which pooled genetic variations from Thai population could impact Thailand's genomic medicine. Duangdao Wichadakul, Chumpol Ngamphiw, Chureerat Phokaew, Sujiraporn Pakchuen, Wanna Chetruengchai, and Athiphat Khuninthong implemented bioinformatic pipelines to extract genetic variations. Alisa Wilantho, Pongsakorn Wangkumhang, Philip James Shaw, and Sissades Tongsima conducted population genomic analyses to resolve population structure. Rujipat Wasitthankasem and Jittima Piriyapongsa constructed the concept in which evolutionary forces shape rare diseases specific to certain subpopulations in Thailand. Vorthunju Nakhonsri, Philip James Shaw, and Sissades Tongsima performed statistical analyses to identify common genetic disorders in the region. Sissades Tongsima, Philip James Shaw, and Vorasuk Shotelersuk organized and took the lead in drafting the manuscript. Duangdao Wichadakul, Sujiraporn Pakchuen, Alisa Wilantho, Pongsakorn Wangkumhang, Chumpol Ngamphiw, Vorthunju Nakhonsri, and Thantrira Porntaveetus drafted methods and reported the results. Chalurmpon Srichomthong, Adjima Assawapitaksakul, Piranit Kantaputra, Keswadee Lapphra, Verayuth Praphanphoj, Prapaporn Pisitkun, Nusara Satproedprai, Wichittra Tassaneeyakul, Pattarapong Makarawate, Surakameth Mahasirimongkol, and Kanya Suphapeetiporn provided exome sequencing data and related information that were used in this study. All authors critically read and commented on the manuscript.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/cge.14060.

DATA AVAILABILITY STATEMENT

All aggregated exomic variant information to support the findings of this study and the supplementary information are freely accessible from https://trex.nbt.or.th. Further data are available from the corresponding authors upon reasonable request.

ORCID

Vorasuk Shotelersuk https://orcid.org/0000-0002-1856-0589
Kanya Suphapeetiporn https://orcid.org/0000-0001-5679-7547
Nusara Satproedprai https://orcid.org/0000-0001-8754-231X
Piranit Kantaputra https://orcid.org/0000-0001-9841-0881

REFERENCES

1. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020; 581(7809):434.

2. Wall JD, Stawiski EW, Ratan A, et al. The GenomeAsia 100K project enables genetic discoveries across Asia. *Nature*. 2019;576(7785):106-111.
3. Lee S, Seo J, Park J, et al. Korean variant archive (KOVA): a reference database of genetic variations in the Korean population. *Sci Rep*. 2017;7(1):4287.
4. Kim J, Weber JA, Jho S, et al. KoVariome: Korean National Standard Reference Variome database of whole genomes with comprehensive SNV, indel, CNV, and SV analyses. *Sci Rep*. 2018;8(1):5677.
5. Wu D, Dou J, Chai X, et al. Large-scale whole-genome sequencing of three diverse Asian populations in Singapore. *Cell*. 2019;179(3):736-749.e715.
6. Poplin R, Ruano-Rubio V, DePristo MA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*. 2018; 201178. https://doi.org/10.1101/201178
7. Chaichoompu K, Abegaz F, Tongsima S, et al. IPCAPS: an R package for iterative pruning to capture population structure. *Source Code Biol Med*. 2019;14:2.
8. Alexander DH, Lange K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinform*. 2011;12:246.
9. McLaren W, Gil L, Hunt SE, et al. The ensembl variant effect predictor. *Genome Biol*. 2016;17(1):122.
10. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Hum Mutat*. 2016;37(3):235-241.
11. MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res*. 2014;42(Database issue:D986-D992.
12. Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res*. 2012;22(9):1760-1774.
13. Shotelersuk V, Limwongse C, Mahasirimongkol S. Genetics and genomics in Thailand: challenges and opportunities. *Mol Genet Genomic Med*. 2014;2(3):210-216.
14. Shotelersuk V, Tongsima S, Pithukpakorn M, Eu-Ahsunthornwattana J, Mahasirimongkol S. Precision medicine in Thailand. *Am J Med Genet C Semin Med Genet*. 2019;181(2):245-253.
15. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006;38(8): 904-909.
16. Wangkumhang P, Shaw PJ, Chaichoompu K, et al. Insight into the peopling of mainland Southeast Asia from Thai population genetic structure. *PLoS One*. 2013;8(11):e79522.
17. West BA. *Encyclopedia of the Peoples of Asia and Oceania*. Facts on File; 2009.
18. Thavan IPU, Anucha U, Aphipol L. Migration route and livelihood of Chinese-Thai in Yao wa rat and Pak nam pho: Department of Geography, Faculty of Social Science, Kasetsart University; 2019.
19. Phongpaichit CBaP. *A History of Thailand*. 3rd ed. Cambridge University Press; 2014.
20. Eilbeck K, Lewis SE, Mungall CJ, et al. The sequence ontology: a tool for the unification of genome annotations. *Genome Biol*. 2005; 6(5):R44.
21. Tate JG, Bamford S, Jubb HC, et al. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res*. 2019;47(D1):D941-D947.
22. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res*. 2003;31(13):3812-3814.
23. Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*. 2018;46(D1):D1062-D1067.
24. Karaca E, Harel T, Pehlivan D, et al. Genes that affect brain structure and function identified by rare variant analyses of Mendelian neurologic disease. *Neuron*. 2015;88(3):499-513.
25. Wattanasirichaigoon D, Limwongse C, Jariengprasert C, et al. High prevalence of V37I genetic variant in the connexin-26 (GJB2) gene among non-syndromic hearing-impaired and control Thai individuals. *Clin Genet*. 2004;66(5):452-460.
26. Louicharoen C, Patin E, Paul R, et al. Positively selected G6PD-Mahidol mutation reduces plasmodium vivax density in southeast Asians. *Science*. 2009;326(5959):1546-1549.
27. He M, Lin K, Huang Y, et al. Prevalence and molecular study of G6PD deficiency in the Dai and Jingpo ethnic groups in the Dehong prefecture of the Yunnan Province. *Hum Hered*. 2018;83(2):55-64.
28. Waszak SM, Hasin Y, Zichner T, et al. Systematic inference of copy-number genotypes from personal genome sequencing data reveals extensive olfactory receptor gene content diversity. *PLoS Comput Biol*. 2010;6(11):e1000988.
29. Yamanaka M, Kato Y, Angata T, Narimatsu H. Deletion polymorphism of SIGLEC14 and its functional implications. *Glycobiology*. 2009;19(8): 841-846.
30. Chapoval AI, Smithson G, Brunick L, et al. BTNL8, a butyrophilin-like molecule that costimulates the primary immune response. *Mol Immunol*. 2013;56(4):819-828.
31. Aigner J, Villatoro S, Rabionet R, et al. A common 56-kilobase deletion in a primate-specific segmental duplication creates a novel butyrophilin-like protein. *BMC Genet*. 2013;14:61.
32. Watson CT, Steinberg KM, Huddleston J, et al. Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *Am J Hum Genet*. 2013;92(4):530-546.
33. de Cid R, Riveira-Munoz E, Zeeuwen PL, et al. Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. *Nat Genet*. 2009;41(2):211-215.
34. Docampo E, Rabionet R, Riveira-Munoz E, et al. Deletion of the late cornified envelope genes, LCE3C and LCE3B, is associated with rheumatoid arthritis. *Arthritis Rheum*. 2010;62(5):1246-1251.
35. Capon F, Munro M, Barker J, Trembath R. Searching for the major histocompatibility complex psoriasis susceptibility gene. *J Invest Dermatol*. 2002;118(5):745-751.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Shotelersuk V, Wichadakul D, Ngamphiw C, et al. The Thai reference exome (T-REx) variant database. *Clinical Genetics*. 2021;100(6):703-712. doi:10.1111/cge.14060