Rare variants discovery by extensive whole-genome sequencing of the Han Chinese population in Taiwan: Applications to cardiovascular medicine

Jyh-Ming Jimmy Juanga, Tzu-Pin Lub, Ming-Wei Suc, Chien-Wei Lin, Jenn-Hwai Yangd, Hou-Wei Chuc, Chien-Hsiun Chen, Yi-Wen Hsiaob, Chien-Yueh Lee, Li-Mei Chiang, Qi-You Yub, Chuhsing Kate Hsiaob, Ching-Yu Julius Chen, Pei-Ei Wud, Chien-Hua Pai, Eric Y. Chuange, Chen-Yang Shen.

Introduction: A population-specific genomic reference is important for research and clinical practice, yet it remains unavailable for Han Chinese (HC) in Taiwan. Objectives: We report the first whole genome sequencing (WGS) database of HC (1000 Taiwanese Genome (1KTW-WGS)) and demonstrate several applications to cardiovascular medicine.
Introduction

The human genome contains an enormous amount of information about human evolution, development, and medicine. To discover the underlying genetic causes of diseases, comprehensive whole-genome sequencing (WGS) is needed to interrogate all types of genetic variation, including single-nucleotide variants (SNVs) and structural and de novo variants. High-throughput genotyping and sequencing have shown how diversity in the sequence of the human genome affects human diversity. A huge number of variants identified by WGS presents new opportunities and challenges for researchers and clinicians.

Several large-scale WGS projects have been completed. Foremost among them are the 1000 Genomes (1000G) Project; the National Heart, Lung and Blood Institute's GO Exome Sequencing Project for European Americans and African Americans; the integrative Japanese Genome Variation Database (iJGVD); and gnomAD. All of them have provided valuable information about human genome diversity across different ethnic populations worldwide. In the past few decades, studies based on WGS and whole-exome sequencing have identified rare variants associated with diseases. Therefore, it is essential to study large samples to build a complete database of genetic variations for each population.

Taiwan is a small but densely populated island in East Asia (~23.5 million inhabitants in 36,543 km², a population larger than that of the Netherlands but with a smaller geographic area). The majority of the population (>95%) in Taiwan is of Han Chinese (HC) ancestry [1]. In the Taiwan Biobank (TWB), a total of 20,117 relatively healthy HC individuals were genotyped using a customized Axiom Genome-Wide Association Study (GWAS) array. Although the GWAS chip was specifically designed for Taiwanese people, it only contains around 600,000 SNV loci. This number is dramatically lower than the number of bases in a whole human genome (~24 billion), and most of them are in non-coding regions.

The TWB used a customized Axiom GWAS array that includes the integrative Japanese Genome Variation Database (iJGVD); and gnomAD. All of them have provided valuable information about human genome diversity across different ethnic populations worldwide. In the past few decades, studies based on WGS and whole-exome sequencing have identified rare variants associated with diseases. Therefore, it is essential to study large samples to build a complete database of genetic variations for each population.

Taiwan is a small but densely populated island in East Asia (~23.5 million inhabitants in 36,543 km², a population larger than that of the Netherlands but with a smaller geographic area). The majority of the population (>95%) in Taiwan is of Han Chinese (HC) ancestry [1]. In the Taiwan Biobank (TWB), a total of 20,117 relatively healthy HC individuals were genotyped using the customized Axiom Genome-Wide Association Study (GWAS) array. Although the GWAS chip was specifically designed for Taiwanese people, it only contains around 600,000 SNV loci. This number is dramatically lower than the number of bases in a whole human genome (~24 billion), and most of them are in non-coding regions.

The TWB used a customized Axiom GWAS array that includes the integrative Japanese Genome Variation Database (iJGVD); and gnomAD. All of them have provided valuable information about human genome diversity across different ethnic populations worldwide. In the past few decades, studies based on WGS and whole-exome sequencing have identified rare variants associated with diseases. Therefore, it is essential to study large samples to build a complete database of genetic variations for each population.

Taiwan is a small but densely populated island in East Asia (~23.5 million inhabitants in 36,543 km², a population larger than that of the Netherlands but with a smaller geographic area). The majority of the population (>95%) in Taiwan is of Han Chinese (HC) ancestry [1]. In the Taiwan Biobank (TWB), a total of 20,117 relatively healthy HC individuals were genotyped using the customized Axiom Genome-Wide Association Study (GWAS) array. Although the GWAS chip was specifically designed for Taiwanese people, it only contains around 600,000 SNV loci. This number is dramatically lower than the number of bases in a whole human genome (~24 billion), and most of them are in non-coding regions.

The TWB used a customized Axiom GWAS array that includes the integrative Japanese Genome Variation Database (iJGVD); and gnomAD. All of them have provided valuable information about human genome diversity across different ethnic populations worldwide. In the past few decades, studies based on WGS and whole-exome sequencing have identified rare variants associated with diseases. Therefore, it is essential to study large samples to build a complete database of genetic variations for each population.

Taiwan is a small but densely populated island in East Asia (~23.5 million inhabitants in 36,543 km², a population larger than that of the Netherlands but with a smaller geographic area). The majority of the population (>95%) in Taiwan is of Han Chinese (HC) ancestry [1]. In the Taiwan Biobank (TWB), a total of 20,117 relatively healthy HC individuals were genotyped using the customized Axiom Genome-Wide Association Study (GWAS) array. Although the GWAS chip was specifically designed for Taiwanese people, it only contains around 600,000 SNV loci. This number is dramatically lower than the number of bases in a whole human genome (~24 billion), and most of them are in non-coding regions.

The TWB used a customized Axiom GWAS array that includes the integrative Japanese Genome Variation Database (iJGVD); and gnomAD. All of them have provided valuable information about human genome diversity across different ethnic populations worldwide. In the past few decades, studies based on WGS and whole-exome sequencing have identified rare variants associated with diseases. Therefore, it is essential to study large samples to build a complete database of genetic variations for each population.

Taiwan is a small but densely populated island in East Asia (~23.5 million inhabitants in 36,543 km², a population larger than that of the Netherlands but with a smaller geographic area). The majority of the population (>95%) in Taiwan is of Han Chinese (HC) ancestry [1]. In the Taiwan Biobank (TWB), a total of 20,117 relatively healthy HC individuals were genotyped using the customized Axiom Genome-Wide Association Study (GWAS) array. Although the GWAS chip was specifically designed for Taiwanese people, it only contains around 600,000 SNV loci. This number is dramatically lower than the number of bases in a whole human genome (~24 billion), and most of them are in non-coding regions.

The TWB used a customized Axiom GWAS array that includes the integrative Japanese Genome Variation Database (iJGVD); and gnomAD. All of them have provided valuable information about human genome diversity across different ethnic populations worldwide. In the past few decades, studies based on WGS and whole-exome sequencing have identified rare variants associated with diseases. Therefore, it is essential to study large samples to build a complete database of genetic variations for each population.

Taiwan is a small but densely populated island in East Asia (~23.5 million inhabitants in 36,543 km², a population larger than that of the Netherlands but with a smaller geographic area). The majority of the population (>95%) in Taiwan is of Han Chinese (HC) ancestry [1]. In the Taiwan Biobank (TWB), a total of 20,117 relatively healthy HC individuals were genotyped using the customized Axiom Genome-Wide Association Study (GWAS) array. Although the GWAS chip was specifically designed for Taiwanese people, it only contains around 600,000 SNV loci. This number is dramatically lower than the number of bases in a whole human genome (~24 billion), and most of them are in non-coding regions.

The TWB used a customized Axiom GWAS array that includes the integrative Japanese Genome Variation Database (iJGVD); and gnomAD. All of them have provided valuable information about human genome diversity across different ethnic populations worldwide. In the past few decades, studies based on WGS and whole-exome sequencing have identified rare variants associated with diseases. Therefore, it is essential to study large samples to build a complete database of genetic variations for each population.

Taiwan is a small but densely populated island in East Asia (~23.5 million inhabitants in 36,543 km², a population larger than that of the Netherlands but with a smaller geographic area). The majority of the population (>95%) in Taiwan is of Han Chinese (HC) ancestry [1]. In the Taiwan Biobank (TWB), a total of 20,117 relatively healthy HC individuals were genotyped using the customized Axiom Genome-Wide Association Study (GWAS) array. Although the GWAS chip was specifically designed for Taiwanese people, it only contains around 600,000 SNV loci. This number is dramatically lower than the number of bases in a whole human genome (~24 billion), and most of them are in non-coding regions.

The TWB used a customized Axiom GWAS array that includes the integrative Japanese Genome Variation Database (iJGVD); and gnomAD. All of them have provided valuable information about human genome diversity across different ethnic populations worldwide. In the past few decades, studies based on WGS and whole-exome sequencing have identified rare variants associated with diseases. Therefore, it is essential to study large samples to build a complete database of genetic variations for each population.
identified with a $P$ value of $5.72 \times 10^{-8}$ in our previous GWAS, which was conducted to identify important and predictive SNVs related to hyperlipidemia. However, novel variants associated with hyperlipidemia were not detectable due to the limitations of using an SNP microarray. Therefore, to demonstrate the application of the 1KTW-WGS database, we explored novel variants associated with rs7115242. Initially, a linkage disequilibrium (LD) analysis was performed for rs7115242 using the single nucleotide polymorphism annotation and proxy search (SNAP) web site [6]. To focus on the East Asian population, we selected the Chinese Han Beijing (CHB) and Japanese in Tokyo (JPT) populations as reference groups. The genotyping data for the reference panel were obtained from the 1000G Pilot 1 project [7], and confidence interval testing (used D prime as the parameter) in Haploview 4.2 software was used to partition the chromosome into blocks. The physical coordinates for the two SNPs closest to the boundary of the LD block were determined according to the $R^2$ value ($R^2 \geq 0.8$) [8–10]. Subsequently, the chromosomal locations for the LD block were utilized to select nonsynonymous mutations identified from the next-generation sequencing (NGS) data in the TWB. Three bioinformatics algorithms—Sorting Tolerant From Intolerant (SIFT) [11], Protein Variation Effect Analyzer (PROVEAN) [12], and Polymorphism Phenotyping v2 (PolyPhen-2) [13]—were used to assess the functional significance of the nonsynonymous mutations.

Development of a prediction model for hypertension. Hypertension was defined as (i) systolic blood pressure $\geq 140$ mmHg, or (ii) diastolic blood pressure $\geq 90$ mmHg, or (iii) self-reported hypertension patients on medications; all other subjects were classified as non-hypertensive. We genotyped 18,450 samples with the TWB genotyping array. We excluded samples with (i) sample call rate $< 95\%$, (ii) heterozygosity rate $> 5$ standard deviations from the population mean, (iii) closely related individuals (identity by descent rating $> 0.1875$), and (iv) non–East Asian outliers identified with PLINK software [4] for variants, excluding those with (i) SNP call rate $> 95\%$, (ii) minor allele frequency $< 5\%$, and (iii) Hardy–Weinberg equilibrium $P < 1 \times 10^{-4}$, and (iv) different missing rate between cases and controls ($P < 0.00001$). Subsequently, for each of the remaining SNVs, a logistic regression model was used to evaluate the association of the SNP with hypertension after adjustment for age, gender, and body mass index (BMI). PCA was used for the assessment of population stratification. The first 8 principal components were used to adjust the GWAS analysis. The genome-wide $P$ value significance threshold ($P < 10^{-8}$) was defined to select SNVs significantly associated with hypertension. Lastly, a prediction model for hypertension was developed based on the significant SNVs using a logistic regression model. We selected candidate SNVs to construct a polygenic risk score after LD clumping of the variants with $P < 5 \times 10^{-5}$. The score was created using the equation $\sum_{i=1}^{\text{SNVs}} k_i \beta_i$, where $k_i$ is the number of selected SNVs and $\beta_i$ is the regression coefficient for each SNV, that was derived from logistic regression.

Development of an online database for identifying SCN5A variants in sudden arrhythmia death syndrome. Because sudden arrhythmia death syndrome is inheritable and is one of the primary causes of sudden death in children and young adults, databases of disease-relevant mutations (e.g., Human Gene Mutation Database (HGMD), ClinVar) are often used to identify potentially pathogenic variants for clinical genetic testing or basic research. We annotated variants in the HC WGS data that were listed as disease-causing in the HGMD and ClinVar [14,15], and established an online gene database to easily explore racial differences and the potential functional impact of identified variants. The design of the database is summarized in Supplementary Fig. 4. Briefly, we collected all nucleotide variants of SCN5A from the TWB [16] and 3 online databases, including the 1000G database (phase 3) [7], the iJGVD [17], and gnomAD [18]. The detailed characteristics of the four datasets and procedures are summarized in Supplementary Table 1 and the Supplementary Note, and each population was classified into Asian or non-Asian groups according to their ethnic background. The proportion test embedded in the R language was performed to assess whether DNA variants in SCN5A show different allele frequencies between the two groups (proportion test: $P < 0.001$). The online database for SCN5A variants was developed using two programming languages, Python Flask and MySQL.

Statistical analysis

For the comparison of minor allele frequency (MAF) across different groups, a proportion test was conducted, and $P$ values less than 0.05 were reported as significant. For SNVs analyzed for hypertension and hyperlipidemia, a logistic regression model was utilized, and confounding factors including age, gender, BMI, and principal components were adjusted. The LD associations were evaluated to identify tag SNVs within a LD block.

Data availability

The URL for the online database for SCN5A variants is http://140.112.136.14:8000. The 1KTW-WGS database is open to the public (the Taiwan Biobank website is https://taiwanview.twbiobank.org.tw/browse38).

Results

We randomly selected 997 unrelated HC individuals from the TWB to construct 1KTW-WGS, identified novel variants in their genomes, and applied the WGS data to clinical practice and patient care in the cardiovascular field.

Sequencing and variant discovery

After considering the quality and abundance of DNA samples, uncertain kinship, and ethnicity outliers (Supplementary Table 2), we sequenced the whole genomes of 997 HC and identified a total of 26,051,907 SNVs and 3,592,314 indels with at least 30X coverage. Of the SNVs, 13,624,601 (32.7%) were novel compared to gnomAD, while 2,266,663 of the indels (43.7%) were novel. Because HC in both Taiwan and Japan are East Asian, we compared our WGS data with that of 1,070 Japanese individuals (1KJP) from iJGVD [17]. Numbers of SNVs, novel SNVs, indels, and novel indels were similar to those in 1KJP (Supplementary Table 3). At first, two different cutoffs (20X or 30X) of the read depth were utilized to identify DNA variants in this study (Supplementary Table 3). Since the average depth of the NGS data in this study was around 30X and the number of identified variants was comparable to that from the 1KJP Biobank, we decided to use 30X as the cutoff in this study. Per individual, the mean total number of SNVs with at least 30X coverage was 1,894,528.7 (range 585,773–2,740,034), and the mean number of novel SNVs was 52,394.1 (16,037–91,629), whereas the mean total number of indels was 133,869 (31,182–306,432) and the mean number of novel indels was 8,746.7 (3,403–51233) (Fig. 1). The two platforms detected similar numbers of SNVs relative to dbSNP build 152 and the combined datasets of dbSNP and 1000G, whereas more novel SNVs were detected by the Ion Torrent-Proton platform than the Illumina platform (Supplementary Table 4).
Comparing the number of indels between protein-coding regions and non-coding regions using the dbSNP 152 and 1000G phase 3 public databases as references, we found that there were fewer indels in protein-coding regions than in non-coding regions, but the percentage of novel indels was higher in protein-coding regions than in non-coding regions (Fig. 2).

We inspected the length and number of indels by genomic location. The overall shape of the distribution for size frequency showed that larger indels were less frequent than smaller indels (Fig. 2), irrespective of their location in a coding or non-coding region. Comparison of the length distribution of indels between protein-coding regions and non-coding regions showed that indel lengths were similar, from ~30 to 30 base pairs. However, the number of large indels (>20 bp) was higher in non-coding regions than in protein-coding regions (Supplementary Fig. 5). Compared with the distribution of indel lengths inside and outside protein-coding regions of the Icelandic population (the only population that provided this data) [18], our data also showed more deletions were called than insertions. However, Gudbjartsson et al. observed a deficit of deletions and insertions that were not multiples of three, which was different from our finding.

In protein-coding regions, we found that the percentage of indels that were multiples of three was higher compared to that in non-coding regions, regardless of the MAF (Supplementary Fig. 6). The MAF of most large indels was relatively low (<0.5%) compared to the MAF of small indels in coding regions, but this difference in MAF distribution between large indels and small indels was not found in non-coding regions (Supplementary Fig. 6).

Comparison of 1KTW-WGS with existing databases to reveal novel variants

As expected, the majority of SNVs (65.8%) in the 1KTW-WGS database were present in the Asian subset of gnomAD-genome (EAS) or gnomAD-genome-all, which implied that common alleles segregated across Asian populations. Of these variants, 16.1% of SNVs in the 1KTW-WGS database were novel relative to dbSNP.
build 152, 53.4% were novel relative to the combined dataset of dbSNP and 1000G phase 3, and 34.2% were novel relative to gnomAD-genome-all (Fig. 3, Supplementary Table 5). This demonstrates the value of performing WGS in individual populations in greater depth. On the other hand, we also compared the results generated by two different platforms (Illumina vs. Ion Torrent-Proton), which showed similar results.

**Variants categorized by the burden of loss of function**

In a clinical setting, predicting the biological consequences of variants is an important application of sequencing data. To particularly evaluate and compare the burden of loss-of-function (LOF) variants, we classified these variants into four categories as described in previous studies: (1) LOF, including stop-gain or -loss variants, frameshift indels, splice donor or acceptor variants, and initiator codon variants; (2) moderate impact, including missense variants, in-frame indels and splice-region variants; (3) low impact, including synonymous variants and 3′-and 5′-UTR variants; and (4) others, including deep intronic and intergenic variants. Table 1 shows the frequency distribution of variants based on functional annotation. We found that the percentage of variants with a MAF below 0.5% was 76.3%, 75.5%, 62.3% and 63.2% in the categories for the LOF, moderate impact, low impact, and other categories, respectively. In each category, the proportion of novel variants with an MAF below 0.5% was similar to the proportion of total variants with an MAF below 0.5%. We also found that the proportion of total variants with a MAF below 0.5% was similar in the categories of LOF, moderate impact, low impact, and other, between the two NGS platforms. The exception was indels, which were higher in the Ion Torrent-Proton platform than the Illumina platform (Supplementary Table 6). Notably, in Supplementary Table 6, we only included indels with MAF<0.5% to do the comparisons, i.e., rare indels. We note that the indels identified in the Ion Torrent-Proton system had higher MAFs and thus under such criteria (MAF < 0.5%), the number of novel indels from the Ion Torrent-Proton system was less than that from the Illumina system (888,033 versus 1045,081). However, if we calculate the proportion of novel indels versus total indels, the Ion Torrent-Proton system still has a slightly higher number (57.57% versus 56.63%). This might be explained by the presence of false positive indels, as reported by previous studies [19,20].

Identifying a causal variant in patients with an inherited Mendelian disease is crucial in clinical genetic testing. Online Mendelian Inheritance in Man (OMIM) is a public and popular database of reported genes and mutations of rare diseases and mutations with high penetrance [21]. For Mendelian diseases, the causal variants in the OMIM database are most often LOF or moderate-impact mutations [21]. Reliable estimates of genotype frequency in a population play an extremely important role when filtering candidate variants in a patient with an inherited Mendelian disease. The 1KTW-WGS database provides the distribution of mean genotype counts per individual by frequency and variant impact (Table 2, Supplementary Fig. 7). Our sequenced individuals carried on average 149 LOF variants, of which 1.4 were only seen in 1 or 2 of the 997 sequenced individuals (MAF < 0.04%) and are thus likely candidates for dominant determinants of rare traits. In the context of rare recessive traits, we found that 3.3 individuals were homozygous for a LOF variant with MAF < 3.2%, a threshold that would correspond to 1 in 997 individuals being homozygous under Hardy-Weinberg equilibrium (Supplementary Table 7).

**Functional variants**

We found that the overall patterns and per-individual distributions of LOF SNVs (variants introducing premature stop codons or variants interrupting splice sites) and missense variants in 1KTW-WGS were consistent with those found in the 1000G Project. Regarding the individual variant load of coding mutations, each individual carried an average of 218.07 LOF SNVs, 99.49 LOF indels, and 0.67 LOF large deletions. Interestingly, 78.4% of the LOF SNVs for each individual are common variants. In contrast, considering rare LOF SNVs (MAF < 0.5%) alone, each individual carried an average of 5.67 nonsense variants, 10.88 variants interrupting a splice site, 11.34 frameshift indels, and 0.03 larger deletions (Table 2). Because databases of disease-relevant mutations are often employed to identify potential variants of interest, we annotated variants in 1KTW-WGS that were listed as disease-causing mutations (DMs) in the HGMDB [14]. We identified 1700 variants with at least 30X coverage as DMs, and each individual carried an average of 16.42 (SD = 3.72) DM variants whose GERP score was greater than 2.0 (Table 2, Supplementary Fig. 8). ClinVar is another well-known and popular disease-relevant database (https://www.ncbi.nlm.nih.gov/clinvar/). In addition to annotating disease-causing variants in 1KTW-WGS based on the data in the HGMD database, we listed some variants in 1KTW-WGS that were classified as pathogenic in ClinVar (Table 3).

As the strength of negative selection increases, we expect a greater fraction of very rare variants (FVRVs), defined the same as in previous work [17], which are not usually included on GWAS chips. In FVRV analyses, we only considered positions with a coverage of at least 30X. A detailed breakdown of FVRVs by variant annotation category is shown in Fig. 4. More than 40% of the total identified missense variants were FVRVs, which was the highest among all mutation types and genomic regions (Fig. 4a). In the Japanese population [17], the FVRV was highest in intergenic regions. This finding suggested that although Japanese and HC in Taiwan are both East Asian, the effect of negative selection is different in the two ethnic populations. In addition, we were surprised to find that variants with a greater impact on gene-encoded proteins tended to have a higher FVRV. Approximately 40–50% of the total identified variants predicted to be probably/possibly damaging by PolyPhen-2 or LOF variants by the Variant Effect Predictor were very rare variants (Fig. 4b).

**Genotyping consistency between 1KTW-WGS and a customized GWAS array and imputation performance**

Since we used two platforms of NGS, we tested the genotyping consistency of each platform separately. Four hundred ninety-nine individuals underwent both WGS by Illumina and the customized GWAS array, whereas 498 individuals underwent both WGS by Ion Torrent-Proton and the customized GWAS array. There were 610,225 and 585,535 overlapping SNVs, and the consistency rates were 98.7% and 95.6%, respectively. We also compared 1KTW-WGS and the samples in 1000G (phase 3: EAS) for genotyping imputation. The highest R² value (the measure of imputation accu-
racy) was achieved with 1KTW-WGS plus 1000G (phase 3: EAS), irrespective of a high or low MAF; the mean R^2 values were 0.7 for rare SNVs (MAF 0.01–1%), 0.7–0.85 for low-frequency SNVs (MAF 1–5%), and 0.88–0.94 for common SNVs (MAF > 5%) (Supplementary Fig. 9).

In addition, we compared the number of imputed SNVs using 1KTW-WGS and 1000G (phase 3: EAS) separately versus 1KTW-WGS plus 1000G (phase 3: EAS). In general, the combination of 1KTW-WGS plus 1000G (phase 3: EAS) could impute more variants, especially rare variants, than the use of 1KTW-WGS or 1000G (phase 3: EAS) alone (Supplementary Fig. 10). The significant improvement in imputation accuracy and the number of imputed SNVs using the 1KTW-WGS data suggest the importance of constructing and examining a population-specific reference panel.

Applying 1KTW-WGS sequencing data to the discovery of novel genetic variants associated with hyperlipidemia

The results of a LD analysis using the SNAP web site indicated that the nucleotide variants located within chr11:116264865–116432950 (GRCh 37/hg19) had R^2 values higher than 0.5. Based on their chromosomal coordinates, 43 non-synonymous variants were reported from the WGS data in the TWB (Supplementary Table 8).

### Table 1
The frequency distribution of variants based on functional annotations*

| Type                        | SNV  | All  |
|-----------------------------|------|------|
|                            | MAF  | > 5% | 0.5–5% | < 0.5% | All  |
|                            | Total| Mean | Mean   | Mean   | Mean |
| 3’ or 5’ UTR, upstream, downstream, synonymous | 217,916 | 143,562 | 618,151 | 979,629 | 33,692 | 30,607 | 86,387 | 150,686 |
| Novel^*                     | 2674 | 17,717 | 245.81 | 54.11 | 171.05 | 2032.78 | 1452.5 |
| % of total variants         | 0.005 | 0.015 | 0.029 | 0.028 | 0.009 | 0.017 | 0.029 | 0.031 |
| % of novel variants         | 0.005 | 0.015 | 0.029 | 0.028 | 0.009 | 0.017 | 0.029 | 0.031 |
| Stop gain or loss, frameshift variant, splicing variant | 674 | 605 | 4943 | 6195 | 937 | 1163 | 5463 | 7067 |
| Novel^*                     | 4 | 16 | 1171 | 1191 | 0.097 | 0.201 | 0.312 | 0.241 |
| % of total variants         | 0.005 | 0.015 | 0.029 | 0.028 | 0.009 | 0.017 | 0.029 | 0.031 |
| % of novel variants         | 0.005 | 0.015 | 0.029 | 0.028 | 0.009 | 0.017 | 0.029 | 0.031 |
| Missense, in-frame variant  | 18,893 | 18,097 | 17,752 | 18,114 | 64 | 134 | 1668 | 1866 |
| Novel^*                     | 32 | 330 | 17,532 | 18,114 | 0.049 | 0.055 | 0.116 | 0.089 |
| % of total variants         | 0.003 | 0.03 | 0.041 | 0.042 | 0.026 | 0.03 | 0.11 | 0.09 |
| % of novel variants         | 0.003 | 0.03 | 0.041 | 0.042 | 0.026 | 0.03 | 0.11 | 0.09 |
| Intronic, intergenic        | 5,647,133 | 3,494,935 | 16,385,854 | 25,527,922 | 929,818 | 875,413 | 619,043 | 1,469,126 |
| Novel^*                     | 78,362 | 102,894 | 3,935,234 | 4,116,490 | 96,363 | 96,121 | 96,115 |
| % of total variants         | 0.003 | 0.047 | 0.043 | 0.042 | 0.026 | 0.03 | 0.11 | 0.09 |
| % of novel variants         | 0.003 | 0.047 | 0.043 | 0.042 | 0.026 | 0.03 | 0.11 | 0.09 |
| Overall                     | 5,884,589 | 3,657,199 | 17,123,130 | 26,664,938 | 964,917 | 844,626 | 4,247,891 | 4,237,434 |
| Novel^*                     | 81,072 | 107,285 | 4,091,937 | 4,280,294 | 242,525 | 376,733 | 1,440,680 | 2,437,434 |

*Not observed in dbSNP build 152 and 1000 Genomes Project Phase 3; MAF: minor allele frequency; SNV: single nucleotide variant; Indel: insertion or deletion.

### Table 2
Variant load of coding mutations and disease-associated variations per individual in 1KTW-WGS.

| Type                              | Mean (s.d.) | Mean (s.d.) | Mean (s.d.) |
|-----------------------------------|-------------|-------------|-------------|
| Rare variant (<0.5%)              | 19.76 (13.79) | 7.27 (5.52) | 171.05 (26.31) |
| Low frequency variant (0.5–5%)    | 135.67 (27.57) | 245.81 (54.11) | 2032.78 (415.2) |
| Common variant (>5%)              | 0.29 (0.55) | 0.36 (0.62) | 1.47 (0.97) |
| Stops loss of function             | 10.88 (3.71) | 3.5 (1.32) | 10.39 (3.51) |
| Non-synonymous                     | 5.67 (2.34) | 6.55 (2.57) | 70.42 (18.71) |
| Probably damaging                  | 51.37 (9.79) | 114.13 (25.97) | 1123.95 (242.7) |
| Splicing variant                   | 0.03 (0.17) | 0.99 (0.74) | 11.25 (3.62) |
| HGMD (only disease-causing mutations) | 0.23 (0.47) | 0.25 (0.5) | 1.25 (0.88) |
| Loss of function >20BP deletion  | 0.03 (0.07) | 0.11 (0.32) | 0.53 (0.75) |
| Indel frameshift >20 BP           | 11.34 (13.05) | 15.44 (5.82) | 72.04 (15.32) |
| Indel non-frameshift >20 BP       | 3.47 (16.87) | 4.91 (3.15) | 68.2 (13.48) |
| Synonymous                        | 599.8 (126.1) | 1008.38 (239.15) | 9982.2 (2230) |
| All SNVs                          | 26.42 (12.4) | 5.58 (3.98) | 3.98 (5.97) |
| Novel^*                           | 211.49 (36.61) | 365.23 (79.53) | 3221.63 (671.2) |

a Not observed in dbSNP build 152 and 1000 Genomes Project Phase 3; bOMIM, Online Mendelian Inheritance in Man; HGMD, The Human Genetic Mutation Database professional version.

Only SNV sites at which ancestral state can be assigned with high confidence and that are highly conserved (GERP > 2.0) are reported. OMIM, Online Mendelian Inheritance in Man; HGMD, The Human Genetic Mutation Database professional version.

The results of a LD analysis using the SNAP web site indicated that the nucleotide variants located within chr11:116264865–116432950 (GRCh 37/hg19) had R^2 values higher than 0.5. Based on their chromosomal coordinates, 43 non-synonymous variants were reported from the WGS data in the TWB (Supplementary Table 8).
HGMD or ClinVar reported disease-causing or pathogenic mutations in sudden arrhythmic death syndrome or cardiomyopathy in 1KTW-WGS individuals.

| Phenotypic Inheritance | Database | 1000G | gnomAD | 2KJPN | ChinaMap | Mutation | Reference/ prevalence | manifestation | pattern | Mutational allele | AF | EA | SA | AFR | EA | SA | AFR | EA | SA | AFR | EA | SA | AFR | EA | SA | AFR | EA | SA | AFR | EA | SA | AFR | EA | SA |
|------------------------|----------|-------|--------|--------|---------|----------|----------------------|--------------|---------|------------------|----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|----|-----|----|---
VKORC1 gene (rs9923231) have been reported to be associated with the dose requirement for warfarin [24–26]. Table 4 shows that the MAFs of these three SNVs in 1KTW-WGS were significantly higher or lower than those in Caucasians and African Americans. This genetic difference may be one of the reasons explaining the different dosage requirements for warfarin across populations (Supplementary Fig. 13).

Clopidogrel is the standard of care for patients receiving coronary stenting [27–29]. It becomes an active metabolite after processing by cytochrome P-450 (CYP) enzymes to achieve its antiplatelet effect. LOF alleles (*2 or *3) of the CYP2C19 gene have been reported to be associated with clopidogrel response and cardiac ischemic events in patients with coronary stenting or acute coronary syndrome [30,31]. It is reported that the prevalence of the CYP2C19 LOF variants is 35% to 45% among blacks, and 25% to 35% in whites, whereas it is 55% to 70% among Asians. The prevalence of CYP2C19 poor metabolizers, which means individuals carrying 2 LOF alleles, is 5% among blacks and whites, whereas it is higher among Asians (10–20%). In addition to the CYP2C19*2 allele, 10% to 20% of Asians also carry another defective allele, CYP2C19*3 [32–34]. The differences in the MAFs of SNVs in CYP2C19*2 and CYP2C19*3 across different populations are shown in Table 4, showing substantial ethnic differences in the distribution and type of CYP2C19 LOF alleles.

A very recent project entitled ‘ChinaMap’ reported the genomes of 10,588 individuals (http://www.mbiobank.com/) [35]. We compared ChinaMap with the TWB, and found that in general, the results between them are very similar, and only a few loci showed minor differences. For example, rs199473009, associated with long QT syndrome, was not reported in ChinaMap but showed a MAF of 0.002 in the TWB (Table 3). Another SNV, rs4244285 in Table 4, had 35% frequency in the TWB but around 31% in ChinaMap. Thus, minor differences do exist between these two datasets.

Applying 1KTW-WGS to the investigation of the impact of ethnic genetic differences on the SCN5A gene and development of an online database

It is well known that SCN5A (cardiac voltage-gated sodium channel α-subunit) is an important cardiac gene that can cause several inheritable life-threatening arrhythmic diseases, including long QT syndrome, Brugada syndrome, and dilated cardiomyopathy [36–39]. Mutations in SCN5A can occur in sudden cardiac death, especially in children or young adults. For physicians and patients in clinical practice, genetic testing for patients with inheritable arrhythmic diseases and family members is strongly recommended to obtain an early diagnosis, a treatment plan, and risk stratification [40]. The SCN5A gene is an example of the usefulness of 1KTW-WGS in clinical genetic testing. When a SCN5A variant is identified in a patient, a physician could use our online database to easily explore the potential functional impact of the identified SCN5A variant in a race-specific way. An online tool was developed to demonstrate this application of the 1KTW-WGS database.

A total of 1877 DNA variants in SCN5A were identified in 1KTW-WGS. A proportional test revealed that 349 variants showed significant differences in allele frequencies between the Asian and non-Asian groups (proportion test: P < 0.001), and among them, 191 variants had higher allele frequencies in the Asian group (Supplementary Table 12). Intriguingly, none of the 349 significant variants were reported in the HGMD database, suggesting that these significant SCN5A variants identified in healthy volunteers are not causal mutations in inheritable arrhythmic diseases or cardiomyopathy.

As shown in Supplementary Fig. 14, we developed an online database for SCN5A variants that accesses five datasets derived from different populations (Supplementary Table 1). Briefly, users can input the chromosomal coordinates of interest to query SCN5A variants (Supplementary Fig. 14a), and the database is able to output a spreadsheet including the genetic structure, allele frequency in different populations, and the pathogenic level in previous studies, if available (Supplementary Fig. 14b). A tutorial and an example of the use of the database are described in the Supplementary Note.

With advancements in high-throughput technology, such as microarrays and NGS, more SCN5A variants in Brugada syndrome are expected to be identified, and thus this SCN5A database can serve as an important reference system. Users can easily explore racial differences and the potential functional impact of identified SCN5A variants. Importantly, the results indicate that the development of a reference database in healthy general populations can facilitate the screening of important and pathogenic DNA variants associated with inheritable arrhythmic diseases or cardiomyopathy, since these variants seldom overlap between the 1KTW-WGS and HGMD databases.

Discussion

We have deeply sequenced the whole genomes of 997 HC individuals in Taiwan and constructed 1KTW-WGS, the first large reference database for the HC population. The results presented here reflect the wealth of knowledge that can be gleaned from WGS data, providing a comprehensive understanding of the genetic structure of the HC population and a basis to uncover associations between DNA variants and phenotypes and develop clinical applications. The observed proportion of novel variants that were not identified in the 1000G and gnomAD databases demonstrated the value of in-depth population-specific WGS studies in HC.

Large-scale investigation of human genetic variations suggested that recently increased human population growth has caused an excess of rare genetic variants, most of which possibly arose in these rare variants are population-specific. As a result, it would be difficult to impute these variants from a reference database generated from diverse genetic backgrounds. Here, we showed that the number of imputed variants increased with the combined use of both a population-specific reference (1KTW-WGS data) and a non-specific reference (1000G phase 3). These
### Table 4
Examples of pharmacogenetics in clinical scenarios and comparisons of minor allele frequency across different populations.

| Drug name | Gene | Genetic variants | Minor Allele Frequency | Clinical application or impact |
|-----------|------|-----------------|------------------------|-------------------------------|
| Warfarin  | VKOR1 CYP2C9 | rs9923231 C>T | 0.865 0.956* | Weekly warfarin dosage |
|           |      | rs1057910 A>C | 0.034 0.038 | Affects the metabolism of clopidogrel and the risk of coronary stenting thrombosis or cardiovascular events |
|           |      | rs1799853 C>T | 0.001 0.004 | |
| Clopidogrel | CYP2C19*2 | rs4244285 G>A | 0.348 0.335 | |
|           |      | rs3758580 C>T | 0.305 0.335 | |
|           |      | rs181297724 G>C | 0.0048 <0.000 | |
|           |      | rs1778459 G>C | <0.000 <0.000 | |
|           |      | rs778258371 G>A | <0.000 <0.000 | |
|           |      | rs144036596 G>A | <0.000 <0.000 | |
|           |      | rs550517959 A>T | <0.000 <0.000 | |
| CYP2C19*3 |      | rs4986893 G>A | 0.055 0.044 | |
|           |      | rs763625282 T>A | <0.000 <0.000 | |
|           |      | rs144036596 G>A | <0.000 <0.000 | |
|           |      | rs144036596 G>T | <0.000 <0.000 | |

* Indicates P value < 0.01 compared with 1000-WGS; AFR: African/African American; CEU: European; CHB: Chinese-Beijing; CHS: Chinese-South; JPT: Japanese; EAS: East Asian; NA: not available; SAS: South Asian; Ref: reference; Alt: alternate; CHC: chronic hepatitis C; <0.000 means the actual number is not provided in the databases, but it should be less than < 0.000001.
results demonstrated that combining sequencing datasets within and across populations can maximize sensitivity and resolution for the discovery of DNA variants.

The density and frequency of variants in genes and genomic regions afford important information about the strength of natural selection acting on them [42,43]. Because deleterious mutations generally disappear from populations faster than neutral mutations via natural selection, SNVs observed at lower frequencies in a population are indicative of natural selection. Furthermore, the strength of natural selection (purifying selection) differs among various functional genomic categories. We evaluated the relative influence of natural selection on SNVs of each functional category in the same way as previous large sequencing projects [41,44,45]. In 1KTW-WGS, we found that the non-synonymous FVRVs was higher in all categories than in the 1000G dataset [17]. Furthermore, the non-synonymous FVRVs in both 1KTW-WGS and 1000G was higher than the FVRVs of intergenic regions in both datasets, implying this region is under weak natural selection. However, the FVRVs in 3’-UTRs and introns in 1KTW-WGS was significantly higher than the FVRVs in intergenic regions, whereas the FVRVs in 3’-UTRs and introns in the 1000G dataset was lower than the FVRVs in intergenic regions [17]. This might be due to low coverage in these regions in the 1000G dataset rather than a signature of weak purifying selection in UTRs and introns. With regard to structural variants, the overall shape of the distribution for size-frequency showed that larger structural events are less frequent than smaller ones, presumably reflecting the relatively deleterious nature of larger structural changes. The difference between the length distributions of coding and noncoding indels is possibly the result of negative selection against frameshift indels [46,47]. Overall, our results are consistent with previous reports [42] showing that variants with a greater impact on gene products tended to have a higher FVRVs, illustrating that these variants can be associated with diseases and could be useful in capturing causal variants in future studies.

In the 1KTW-WGS database, we observed that a relatively healthy individual carries an average of 16.42 disease-causing variants reported in HGMD. This raises the question of how any person remains disease-free. One possible explanation is that the existence of modifier alleles induces incomplete penetrance or variable expression of DMs, which is dependent on the carrier’s genetic background [48]. An alternative explanation is that HGMD contains a large number of false-positive DMs [49]. Of the 1139 DMs identified in 1KTW-WGS, 30% had a MAF greater than 1%, which is higher than the prevalence of many of the diseases described in HGMD. The majority of these mutations were common in 1KTW-WGS, suggesting that these variants are not subject to strong selective pressure and are likely phenotypically benign. In other words, given the inheritance patterns of the diseases conferred by these variants, many individuals in 1KTW-WGS should be reliably compared each other. Second, although 1KTW-WGS was established, population-wide sequencing with a larger sample size (e.g., > 100 K) is warranted to discover the full spectrum of SNVs and structural variants. Third, all participants who were enrolled in TWB were volunteers. They were not randomly sampled from the HC population. Lastly, in our study, Torrent Variant Caller (TVC) was utilized for Ion Torrent-Proton data, whereas ISAAC Variant Caller was used for Illumina data. No study has directly compared the callers of the two types of the data, but a previous study [56] suggested that the consistency of different callers used with Illumina is up to 91.7%, whereas the variations of the different callers used with Ion Torrent-Proton data ranged from 1.3% to 34.6%. Furthermore, some previous reports indicated that the false discovery rate of SNVs and indels was relatively higher in the Ion Torrent-Proton system than in the Illumina system [57,58]. In addition, several studies suggested that higher false positive rates were observed when using the Ion Torrent-Proton system versus the Illumina system [59–61]. One plausible explanation for this is that the data from the Illumina system have better sequencing quality and stable read length. These advantages give the Illumina system better performance in identification of
Conclusions

In this study, we used the 1KTW-GWS reference database to discover novel HC population-specific variants, identified novel hyperlipidemia-related variants, developed a hypertension prediction model, and established an online SCN5A tool for genetic testing. Taken together, the results demonstrate the necessity of constructing a population-specific genomic reference, which can pave the way for precision medicine and population health initiatives.

The cost of WGS is gradually decreasing. Thus, we expect that a more comprehensive interrogation of genetic variants will advance basic research and support development of diagnostic tools, healthcare customization, and preventive therapeutics for human diseases.

Competing financial interests

The authors declare that they have no competing interests.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jare.2020.12.003.

References

[1] Executive Yuan rTROC. The Republic of China Yearbook; 2016.
[2] Executive Yuan rTROC. The Republic of China Yearbook; 2013.
[3] Chen CH, Yang JH, Chiang CCK, Hsiung CN, Wu PE, Chang LC, et al. Population structure of Han Chinese in the modern Taiwanese population based on 10,000 participants in the Taiwan Biobank project. Hum Mol Genet 2016;25(24):5321–31. doi: https://doi.org/10.1093/hmg/ddw346. PubMed PMID: 27798100.
[4] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75. Epub 2007/08/19. doi: 10.1086/519795. PubMed PMID: 17701901; PubMed Central PMCID: PMC1950838.
[5] Browning BL, Browning SR. Genotype imputation with millions of reference samples. Nat Genet. 2009;41(6):366–71. doi: https://doi.org/10.1038/ng.346. PubMed PMID: 19760539; PubMed Central PMCID: PMC2720775.
[6] Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Haywood LB, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74. doi: 10.1038/nature15393. PubMed PMID: 26432245; PubMed Central PMCID: PMC4750478.
[7] Adzhubei I, Schmidt S, Peshkin L, Ramensky V, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2013;10(2):929–31. doi: https://doi.org/10.1038/nmeth.726. PubMed PMID: 23883674; PubMed Central PMCID: PMC3786708.
[8] Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: a database of humans germline genetic variation. Nucleic Acids Res. 2016;44(D1):D1062-8. Epub 2015/11/20. doi: 10.1093/nar/gkv1078. PubMed PMID: 26582928; PubMed Central PMCID: PMC4702885.
[9] McKusick VA. Mendelian Inheritance in Man and its online version. OMIM. Am J Human Genet 2007;80(4):588–604. doi: https://doi.org/10.1086/514346. PubMed PMID: 17523400; PubMed Central PMCID: PMC1852721.
[10] Niki M, Kishino H, Hasegawa M. Trends in the theory of nucleotide substitution. J Theor Biol 2005;230(2):224-31. doi: 10.1016/j.jtbi.2004.02.003. PubMed PMID: 15749944; PubMed Central PMCID: PMC1030463.
[11] Yuferov VV, Trifonov EN. Estimation of physical and chemical properties of amino acids: a database approach. J Mol Model. 2008;14(6):2393–9. doi: https://doi.org/10.1007/s00894-007-0106-1. PubMed PMID: 18074271; PubMed Central PMCID: PMC2720775.
