Journal of Advanced Research xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Advanced Research

journal homepage: www.elsevier.com/locate/jare

Complete genomic profiles of 1496 Taiwanese reveal curated medical insights

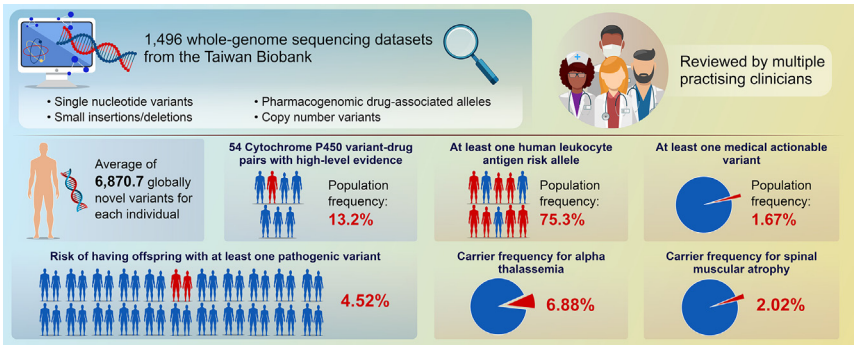
Jacob Shujui Hsu [a,b,1], Dung-Chi Wu [c,1], Shang-Hung Shih [c], Jen-Feng Liu [b], Ya-Chen Tsai [d], Tung-Lin Lee [e], Wei-An Chen [e], Yi-Hsuan Tseng [a], Yi-Chung Lo [f], Hong-Ye Lin [d], Yi-Chieh Chen [a], Jing-Yi Chen [f], Ting-Hsuan Chou [a], Darby Tien-Hao Chang [f,g], Ming Wei Su [h], Wei-Hong Guo [d], Hsin-Hsiang Mao [d], Chien-Yu Chen [c,d,*], Pei-Lung Chen [a,b,c,e,i,*]

[a] *Graduate Institute of Medical Genomics and Proteomics, National Taiwan University College of Medicine, Taipei 100025, Taiwan*
[b] *Institute of Molecular Medicine, National Taiwan University College of Medicine, Taipei 100233, Taiwan*
[c] *Genome and Systems Biology Degree Program, National Taiwan University and Academia Sinica, Taipei 10617, Taiwan*
[d] *Department of Biomechatronics Engineering, National Taiwan University, Taipei 10617, Taiwan*
[e] *Department of Medical Genetics, National Taiwan University Hospital, Taipei 100226, Taiwan*
[f] *Department of Electrical Engineering, National Cheng-Kung University, Tainan 701401, Taiwan*
[g] *Digital Technology Division, SinoPac Holdings, Taiwan*
[h] *Institute of Biomedical Sciences, Academia Sinica, Taipei 115201, Taiwan*
[i] *Graduate Institute of Clinical Medicine, National Taiwan University College of Medicine, Taipei 100233, Taiwan*

HIGHLIGHTS

- WGS from 1496 Taiwanese has been reanalyzed with benchmarked quality scores.
- Taiwanese genomic profiles for SNV, indel, CYP/HLA allele and copy number variant.
- Variant annotation can be accessed at TaiwanGenomes (https://genomes.tw) database.
- 16.5 million variants were found with an average 6,871 novel variants per sample.
- 75.3% and 1.67% had at least one PGx allele and actionable variant, respectively.

GRAPHICAL ABSTRACT



Abbreviations: ACMG, The American College of Medical Genetics and Genomics; ACOG, The American College of Obstetricians and Gynecologists; AN, Allele number; BWA-MEM, Burrows-Wheeler Aligner with Maximal Exact Match algorithm; CNV, Copy number variant; CYP, Cytochrome P450; DP, Depth of coverage; EAS, East Asian; ExAC, The Exome Aggregation Consortium; FDR, False discovery rate; FP, False positive; GIAB, Genome In A Bottle; GVCF, Genomic variant call format; HBA, Hemoglobin subunit alpha; HLA, Human leukocyte antigen; INDEL, small insertion and deletion; JPN, Japanese; LOF, Loss of function; NGS, Next generation sequencing; NSGC, The National Society of Genetic Counselors; OMIM, Online Mendelian Inheritance in Man database; P/LP, Pathogenic/likely pathogenic; PGx, Pharmacogenomic; PKU, Phenylketonuria; PQF, Perinatal Quality Foundation; SFWG, Secondary Findings Maintenance Working Group; SGN, Singaporean; SMA, Spinal muscular atrophy; SMFM, Society for Maternal-Fetal Medicine; SNV, Single nucleotide variant; SV, Structure variant; TN, True Negative; TWB, Taiwan Biobank; VQSR, Variant quality sequence recalibration; WGS, Whole-genome sequencing; gnomAD, The Genome Aggregation Database.

* Corresponding authors at: Graduate Institute of Medical Genomics and Proteomics, College of Medicine, National Taiwan University, No. 2, Xuzhou Rd., Zhongzheng Dist., Taipei 100025, Taiwan (P.-L. Chen). Department of Biomechatronics Engineering, National Taiwan University. No. 1, Sec. 4, Roosevelt Rd., Taipei 106319, Taiwan (C.-Y. Chen).
E-mail addresses: chienyuchen@ntu.edu.tw (C.-Y. Chen), paylong@ntu.edu.tw (P.-L. Chen).
[1] Jacob Shujui Hsu and Dung-Chi Wu have contributed equally to this work.

https://doi.org/10.1016/j.jare.2023.12.018

Please cite this article as: J.S. Hsu, D.-C. Wu, S.-H. Shih et al., Complete genomic profiles of 1496 Taiwanese reveal curated medical insights, Journal of Advanced Research, https://doi.org/10.1016/j.jare.2023.12.018

ARTICLE INFO

Article history:
Received 14 August 2023
Revised 3 December 2023
Accepted 27 December 2023
Available online xxxx

Keywords:
Taiwan Biobank
Whole genome sequence
Population allele frequency
ACMG secondary finding V3 gene list
Carrier rates

ABSTRACT

Introduction: The population of Taiwan has a long history of ethno-cultural evolution. The Taiwanese population was isolated from other large populations such as the European, Han Chinese, and Japanese population. The Taiwan Biobank (TWB) project has built a nationwide database, particularly for personal whole-genome sequence (WGS) to facilitate basic and clinical collaboration nationally and internationally, making it one of the most valuable public datasets of the East Asian population.
Objectives: This study provides comprehensive medical genomic findings from TWB WGS data, for better characterization of disease susceptibility and the choice of ideal treatment regimens in Taiwanese population.
Methods: We reanalyzed 1496 WGS using a PrecisionFDA Truth challenge winner method Sentieon DNAscope. Single nucleotide variants (SNV) and small insertions/deletions (INDEL) were benchmarked. We also analyzed pharmacogenomic (PGx) drug-associated alleles, and copy number variants (CNV). Multiple practicing clinicians reviewed and curated the clinically significant variants. Variant annotations can be browsed at TaiwanGenomes (https://genomes.tw).
Results: We found that each participant had an average of 6,870.7 globally novel variants and 75.3% (831/1103) of the participants harbored at least one PharmGKB-selected high evidence level human leukocyte antigen (HLA) risk allele. 54 PharmGKB-reported high-level instances of evidence of Cytochrome P450 variant-drug pairs, with a population frequency of over 13.2%. We also identified 23 variants in the ACMG secondary finding V3 gene list from 25 participants, suggesting that 1.67% (25/1496) of the population is harboring at least one medical actionable variant. Our carrier status analyses suggest that one in 25 couples (3.94%) would risk having offspring with at least one pathogenic variant, which is in line with rates found in Japan and Singapore. For pathogenic CNV, we detected 6.88% and 2.02% carrier rates for alpha thalassemia and spinal muscular atrophy, respectively.
Conclusion: Our study highlights the overall medical insights of a complete Taiwanese genomic profile.

Introduction

As the field of precision medicine continues to expand, many population-based consortiums have been collecting genomic information for biomedical studies [1]. In Asia, a genetic reference panel based on 3,552 individuals of the Japanese (JPN) population was published in 2019 [2,3]. Another study based on 4,810 Singaporeans (SGN) was published in 2019 by the SG10K project [4]. The results showed that the cohort analysis of a sub-sample was able to capture the population diversity. Even though the ChinaMAP project discloses the genetic profile of 10,588 Chinese [5], Asian genomics data are still relatively underrepresented in the public databases relative to global population percentage. The Taiwan Biobank project (TWB) was established to facilitate biomedical research on the genetic basis of the Taiwanese, a multicultural population. The majority of Taiwanese immigrated from various provinces of China over the past centuries, and there is also a group of Taiwanese aboriginals. As of 30 Sep 2021, the TWB has already recruited 153,543 subjects from the general population. Many types of genomic data are available to users, including single nucleotide variant (SNV) array data from 114,604 subjects, whole-genome sequencing (WGS) data from 2,010 subjects, and human leukocyte antigen (HLA) typing data from 1,102 subjects. Although many genotyping approaches can reveal specific variants that indicate disease susceptibility [6,7], fragmented variant information is not sufficient to fully infer the haplotype of a gene. WGS data have the potential to reconstruct haplotype information at high resolution and cover entire genomic regions, allowing complete determination of disease-causing variants, including SNVs, small insertions/deletions (INDEL), structure variants, and the haplotypes of a gene. For example, a previous study suggested that more than 95 % of hemoglobin subunit alpha (*HBA1/2*) pathogenic or likely pathogenic variants are from –SEA, $-a^{3.7}$, $-a^{4.2}$ in the Chinese population [8]. Most of these are deletions larger than 3 kb in size. The pathogenic variants of fragile X-linked mental retardation gene FMRP translational regulator 1 (*FMR1*) are typically caused by trinucleotide (CGG) repeat expansions, variations of which require specific detection algorithms [9–11]. A medically meaningful complete genomic profile of an individual should also include monogenic disease-causing allele screening [12] and the haplotypes of the human leukocyte antigen (HLA) gene.

Like the HLA gene, the Cytochrome P450 2D6 (*CYP2D6*) gene is associated with the metabolism of many drugs. For functional interpretation, complete haplotype information is needed for both HLA and CPY2D6 genes.

Identifying at-risk individuals with medically actionable information from hereditary diseases may benefit the individual's whole family. Thus, expanded carrier screening for monogenic diseases has become a discipline-wide goal after a joint statement was published by the American College of Medical Genetics and Genomics (ACMG), American College of Obstetricians and Gynecologists (ACOG), National Society of Genetic Counselors (NSGC), Perinatal Quality Foundation (PQF), and Society for Maternal-Fetal Medicine (SMFM) in 2015 [13]. All the disorders noted in abovementioned joint statement involve a cognitive or physical disability, the need for postnatal surgical or medical intervention, or a detrimental effect on the quality of life; most importantly, they are disorders where prenatal intervention could significantly improve perinatal outcomes and delivery management or where prenatal education could meet the unique needs after birth. Each year, the ACMG Secondary Findings Maintenance Working Group (SFWG) evaluates these actionable genes and associated conditions for their actionability, severity, penetrance, and impact or burden of available treatment modalities or screening recommendations. The latest version, ACMG SF v3.0 [14] published in May 2021, contains 73 actionable genes, in contrast to 59 genes in the previous version. Many studies have disclosed the profile of actionable genes in the Asia-Pacific region [15–17], and WGS data allow reanalysis when the gene list has been updated. The advent of WGS helps characterize population genetic structures that can

provide useful information to the public healthcare system, mitigating costs through disease risk prediction, diagnosis, and treatments. In this study, we reanalyzed the WGS data from 1496 participants present in the TWB to provide a medically useful cross-section of the population's complete genomic profile. We also discuss disease carrier status (including all the conditions curated in the ClinVar database), ACMG actionable genes, drug responses from the PharmGKB database, and clinically relevant variants with high minor allele frequency (MAF) in the Taiwanese population. To facilitate the exploration of the variants described here, we designed a web interface for the constructed database TaiwanGenomes (https://genomes.tw) that allows easy browsing of the list of variants with different combinations of filters.

Methods

Ethics statement

All experiments involving human subjects were in accordance with the institutional review board approval from Biomedical Science Research of Academia Sinica, Taiwan (IRB-BM) and the Ethics and Governance Council (EGC) of Taiwan Biobank, Taiwan. All the informed consents were obtained and gave ethical approval for this work (AS-IRB01-18041(N)). All raw sequence data used in this study were generated as part of the Taiwan Biobank project. This study did not reveal any individual participant's information, and none of the results can be used to identify individual participants.

Data source

The Taiwan Biobank (TWB) collected and recruited participants across different cities throughout Taiwan, enrolling participants from the general population. TWB has been enrolling adult volunteers between the ages of 20 and 70 who do not have the diagnosed cancer. For the subset undergoing WGS, participants were selected to match the population distribution as well as to ensure the gender balance. A total of de-identified,496 Taiwanese WGS data sets were sourced from TWB with ethical approval (AS-IRB01-18041(N)). WGS was conducted in three batches (n = 496, 497, and 503) with high-depth mapped reads. The samples were sequenced by next-generation sequencing (NGS) platforms: Illumina HiSeq 2500, 4000, and Novaseq systems. Sequencing of DNA extracted from each blood sample followed by illumina TruSeq DNA PCR-Free HT library preparation kit to generate about 90 GB of data with an average coverage of 30x (https://taiwanview.twbiobank.org.tw/about.php). All pairwise kinships have been checked, and there is no parent-offspring relationship in the cohort.

Genotype calling, validation, and variant annotation

Variant detection and joint genotype calling analyses were conducted based on the Sentieon DNAscope pipeline (Sentieon Inc., version 201808 [18]), a precisionFDA challenge winner implementation of GATK's best practice [19]. The sequence reads of each sample in FASTQ format were aligned against the human reference genome (GRCh37/ucsc.hg19.fasta) using BWA-MEM (Burrows-Wheeler Aligner with Maximal Exact Match algorithm, version 0.7.15-r1140 [20]). The alignment file was sorted using SAMtools [21], and the Sentieon Dedup algorithm was used to mark duplicated reads. The Sentieon Realigner algorithm reinforced local realignment around each indel region, and the Sentieon QualCal algorithm recalibrated base quality scores. SNVs and indels were called in genomic variant call format (GVCF) using Haplotyper. The Sentieon GVCFtyper jointly called 1496 subjects as a cohort, then Sentieon VarCal plus ApplyVarCal, a machine learning-based variant quality sequence recalibration (VQSR) algorithm for variant refinement, were used. The training sets for VQSR were those suggested by GATK's best practice. For SNV, the datasets from 1000G phase1, omni2.5, dbSNP version 138, and HapMap version 3.3 were included as the training sets. For INDEL, the datasets from the 1000G phase1, dbSNP version 138, and Mills-and-1000G gold standard were used as the training sets. VQSR included a list of sequence-level annotations such as QD, MQ, MQRankSum, ReadPosRankSum, and FS. Multi-allelic variants were normalized into multiple bi-allelic variants at the same position through decomposition and left-aligned using bcftools (v1.9) [21,22]. We repeated the same pipeline with an additional seven Genome in A Bottle (GIAB) [23,24] samples (HG001 $\sim$ HG007) and jointly called with 1496 TWB subjects for benchmarking the variant calling pipeline. We stratified the variant call sets according to the VQSR tranches into strata of 100.0, 99.9, 99.8, 99.7, 99.6, 99.5, and 99.0. We further adopted the call rate, allele number (AN), and depth of coverage (DP) criteria for classifying all reference alleles into three quality classifications. Moreover, we randomly validated 149 variants in four samples (NGS2 20150510B, NGS2 20150510C, NGS2 20150510D, NGS2 20150510F) by Sanger sequencing. We were particularly interested in the variants absent, or only present, after joint calling. Due to the limited amount of available DNA, we separated samples into two groups (BC and DF) for cross-validation. We utilized the Sanger sequencing results to calculate the estimated false discovery rate (FDR) for determining the VQSR tranche. In order to provide a clear threshold with higher precision without significantly compromising recall rates, we selected VQSR tranche 99.7 as the PASS criterion based on the comparisons of HG001 (NA12878) true genetic variants (Fig S2) and following Sanger sequencing validation results (Table S1). All identified variants can be accessed from the database **TaiwanGenomes** (https://genomes.tw) together with the corresponded VQSR tranche and annotated information. Only variants with pass quality were considered in the downstream analysis.

Variant annotation was done by ANNOVAR (version 20180416) [25] with updated databases, including RefSeq Gene, UCSC Known Gene, ClinVar (v20210501), avsnp150, NHLBI-ESP 6500 exome, 1000 genome, The Genome Aggregation Database (gnomAD) genome and exome (v2.1.1), The Exome Aggregation Consortium (ExAC) 65,000 exome, Kaviar, cg69, dbnsfp33a, dbscsnv11, gwava, tfbsConsSites, wgRna, and targetScanS. We selected the annotated variants on the autosome and sex chromosome as the final variant call set. We defined nonsynonymous variants as those where a variant's annotation hit exonic or splicing regions. The variations were defined by nonframeshift deletion, frameshift deletion, nonframeshift insertion, frameshift insertion, stopgain, stoploss, or a nonsynonymous SNV with an exonic variant function.

To further elucidate the possible role of structural variant calling tools, we performed a trial using Manta (version 1.6.0) [26] and AnnotSV (version 3.0.4) [27] to identify and annotate the structural variants (SV) in HBA1/2 genes of a subset of the participants (n = 494). Furthermore, the SMN1/SMN2 copy number was analyzed by SMNCopyNumberCaller (version 1.1.1).

Variant filtering and classification

We defined the variants absent in any other public databases (e.g., without allele frequency or effect annotation) as "globally novel variants." Variants that were not found in any previous samples while sequentially examining the 1496 samples we defined as "population novel variants". For functional effect analysis, variants annotated as exonic in refGene were defined as "coding region variants," whereas the variants annotated as nonsense, splicing,

or frameshift in refGene were defined as "loss of function (LOF) variants." We created in-house scripts to conduct the filtering processes. Public users can retrieve the same information by setting corresponding filters in TaiwanGenomes (https://genomes.tw).

Clinical important pharmacogenomic alleles

The HLA gene plays a crucial role in personalized medicine. It is associated with many adverse drug events, including antithyroid drug-induced agranulocytosis [28]. The clinical variant data in the PharmGKB database (clinicalVariants.tsv, version 20210405) detail how genetic variation influences the drug response, including a list of variant-drug pairs with different levels of evidence. TWB has released individual-level HLA alleles by using NXType assay kits (https://taiwanview.twbiobank.org.tw/about.php). To evaluate the allele frequencies of PharmGKB-reported HLA variant-drug pairs in the Taiwanese population, we queried HLA allele typing results from the TWB websites. Here, we annotated the alleles with evidence levels 1A/1B/2A and disclosed their respective frequencies. For CYP genes, we genotyped CYP haplotypes based on 1,017 WGS data sets using ALDY [29] and STARGAZER [30] (a subset of our participants; manuscript in preparation). In calculating CYP variant-drug pairs, we only reported the allele frequency of the variants associated with abnormal drug metabolism. We did not include CYP wild-type allele frequencies.

Carrier status – Cohort MAF and expert reviews

The ClinVar [31] database collects the interpretation of clinically related variants submitted by global researchers. Many known monogenic disease-causing variants are rare and vary across populations [32]. However, only a few records in the database are from the Asian population. To explore medically relevant genetic variants in Taiwanese people, we retained variants annotated as pathogenic, likely pathogenic, or pathogenic/likely pathogenic (P/LP) in ClinVar (v20210501). A high-variant minor-frequency (MAF) allele (>0.5 %) with known pathogenicity may indicate the specificity of the population's genetic architecture. By comparing across the East Asian population (EAS), we singled out pathogenic or likely-pathogenic variants with high frequency to address the specific genetic structure of the Taiwanese.

The variant MAF was further investigated by comparing it with allelic frequencies in the gnomAD (version 2.1.1) and ExAC database. The following filtration criteria were used: Category 1) The minor allele frequency of the variant in TWB1496 was $\geq$0.01, while that in the ExAC database and in the East Asia population of the gnomAD genome and exome database were $\leq$0.01 or absent; Category 2) the allelic frequency was 0.005–0.01 in TWB1496, and $\leq$0.005 or absent in the ExAC database and in the East Asia population of the gnomeAD genome and exome database. The genetic inheritance mode of the variants was manually annotated using the Online Mendelian Inheritance in Man database (OMIM). On the other hand, experts further reviewed the variants within ACMG actionable genes. The concept of medically actionable genes originates from ACMG recommendations on reporting secondary findings from exome or genome sequencing results. To further focus on the East Asian population, we also compared the allele frequency of variants in the carrier sets to two East Asian populations (Singaporean and Japanese). The Singaporean data were acquired from the SG10K project[4], and the Japanese data were obtained from the 3.5KJPN project [33].

To evaluate the clinical impact of the carrier call set, we applied the expanded carrier panel [34,35] to the above data to calculate accumulation frequencies of pathogenic and likely pathogenic variants in monogenic disease-causing genes. We compared these data with the United States cohort [35] (sequencing method) and Taiwan domestic data[6] (array genotyping method).

The overall study design and analysis workflow is shown in Fig S1. In summary, for SNV and INDEL, the analyses were performed for the entire genome of the cohort (N = 1496). For HLA allele, the data was from Taiwan Biobank official release (https://taiwanview.twbiobank.org.tw/hla.php) (N = 1,103), and allele annotation was based on PharmGKB information. For the CYP allele, the analyses were performed on a subset of the WGS cohort (N = 1,017) and only focused on CYP genes. The SV analyses were performed for a subset of the WGS cohort (N = 494), and the analyses were for the entire genome. For SMN CNV calling, the analyses were based on a subset of the WGS cohort (N = 494), and only for SMN genes.

We also used the same carrier call-set data to evaluate other monogenic disease-causing genes as a concept of expanded carrier screening. We compared the P/LP variant carrier frequency with the U.S. cohort and Taiwanese cohort (using TWBv2.0 array genotyping). We also used the Westemeyer et al. (2020) gene list (274 genes) as a virtual panel to estimate how many couples may benefit from the expanded carrier screening in the Taiwanese population.

Web application for TWB1496 cohort

The website for browsing **TaiwanGenomes** is based on the open-source project VASH (https://github.com/mbilab/vash). VASH (a composite word of variant and flash) aims to provide a rapid and fluent user experience while browsing whole genome scale variants. VASH is implemented using Vue.js (https://vuejs.org/) and Django (https://djangoproject.com) as frontend frameworks. MySQL is used as the database engine of Django. All requests by VASH are segmented into small chunks and cached throughout the entire processing pipeline from Vue.js to MySQL. Therefore VASH is capable of processing new chunks while users are viewing previous chunks. VASH enables infinite scrolling, which is more intuitive than pagination browsing and achieves response time in a few seconds regardless of the number of queried variants.

Results

High-quality variant call sets

To evaluate the accuracy of the joint calling, we used the same pipeline to analyze seven samples (HG001 $\sim$ HG007) from GIAB as the internal quality control. We jointly called 1496 TWB subjects with seven reference samples and stratified the joint call set according to the VQSR tranches into strata of 100.0, 99.9, 99.8, 99.7, 99.6, 99.5, and 99.0. We then confirmed both SNV and INDEL variant calling accuracy by comparing HG001 (NA12878) genetic variants as the ground truth to the subset of HG001 results in the joint called cohort for every variant. We defined VQSR tranche 99.7 as the pass criterion to balance the overall sensitivity and specificity Fig S2(a) and S2(b). All variants can be accessed from the TaiwanGenomes database (https://genomes.tw) including their VQSR tranche and annotated information. We only included variants with pass quality in the downstream analyses.

In addition to alternative alleles, we also analyzed reference alleles and used the call rate, allele number (AN), and depth of coverage (DP) filtration for quality classification. We classified all non-alternative positions into A, B, and C categories. If the call rate of a reference allele was above 80 % (AN > 2,400 & DP > 18,000), the reference allele was defined as category A (2,686,586,573 variants). Category B (16,201,674 variants) indicated that the call rate of an allele was below 10 % (AN < 300 & DP < 4,500). Category C

(30,369,949 variants) represents the call rates between categories A and B. Researchers should be careful when interpreting the variants in categories B and C as having a MAF of zero in **TaiwanGenomes**. It actually means missing genotypes or needs more depth of coverage to determine.. The quality information for each genomic locus can be easily browsed at https://genomes.tw/#/supplement.

Variant detection bias or variability in population structure may have driven MAF differences between the two cohorts. We then compared the allele frequencies of 18,624 variants on chromosome 1 in which the variants were identified in both TWB1496 and gnomAD East Asian data. To derive differences in allele frequencies distribution between the two databases, we used the Python package Scipy and Statsmodels.api to fit linear regressions and produce scatter plots. We tested the correlation of MAF between the two databases among different VQSR tranches (Fig S3). Greater VQSR specificity and greater correlation (r-squared) to the current gnomAD East Asian dataset was found.

Variants were rescued by jointly called step and confirmed by Sanger sequencing

The differences between the jointly called VCF file and individual variant calls are worth noting. Theoretically, multi-sample joint calling and followed by VQSR were suggested by GATK best practices to rescue variants with moderate individual-level quality and improve the overall calling accuracy when the sample size increases (https://gatk.broadinstitute.org/hc/en-us/articles/360035890431-The-logic-of-joint-alling-for-germline-short-variants,https://gatk.broadinstitute.org/hc/en-s/articles/360035890411, https://gatk.broadinstitute.org/hc/en-us/articles/360035531612-Variant-Quality-Score-Recalibration-VQSR). Hundreds of variants were checked by Sanger sequencing, but only 109 variants matching the following criteria were included in the accuracy analysis: 1) unambiguous biallelic variants; 2) clear Sanger results; 3) inconsistent results between jointly called and individually called variants. Note that the number of true negative variant calls (TN) was underestimated because the VCF file only included the positive variants. We defined VQSR tranche 99.7 as the pass criterion, which was a greater accuracy than that of the variant calls in all higher VQSR tranches, thereby reducing the false discovery rate (FDR) by 25 % (0.75 to 0.5). Based on the Sanger validation analysis, applying VQSR can remove false positive (FP) variants by 68.6 % (35/51) (Table S1).

Millions of novel variants found in 1496 Taiwanese

Our final variant call set consisted of 59,433,212 variants from autosomes and sex chromosomes (Table 1). Among these variants, there were 49,701,036 variants with MAF < 0.05 and 44,591,573 variants with MAF < 0.01. There were 480,784 variants in coding regions, and 274,265 variants were annotated as nonsynonymous. For the variants satisfying the pass criterion (VQSR 99.7), the number decreased to 51,135,411. Of these, 42,557,774 variants had MAF < 0.05 and 38,599,450 had MAF < 0.01. There were 439,192 variants in coding regions and 250,940 variants annotated as nonsynonymous.

To analyze the unique genetic characteristics of the Taiwanese population, we further filtered out variants that were not present in other databases (see Methods). This resulted in 16,520,159 variants classified as "globally novel." To further investigate the population genetic structure, we sorted the samples by their total variant numbers and calculated the number of unique globally novel variants (Fig. 1). Results from the last 50 samples suggested that a Taiwanese person has an average of 6,870.7 globally novel variants. Overall, there were 16,066,996 variants with MAF < 0.05 and 15,495,346 variants with MAF < 0.01. Most of these novel variants were in the intergenic regions. Of the novel variants, 26,664 were in coding regions, 700 were annotated as missense variants, and 4,159 were annotated as loss-of-function variants, including 151 nonsense variants, 3,174 frameshift variants, and 834 splicing variants.

Loss-of-function and deleterious variants play a crucial role in Mendelian disorders. From the spectrum of variant numbers and the allele frequencies, we observed that there were far fewer loss-of-function variants than others under negative selection, which is in line with our expectations. We also observed that non-frameshift indels occurred at almost the same frequency as loss-of-function variants (Fig. 2).

5.54% of participants had medically actionable variants

We identified 1,136 variants with clinical evidence, which we termed a carrier call-set. Of these, 58 were among the 73 actionable genes recommended by ACMG[13]. These variants were further reviewed by practicing medical geneticists, resulting in 53 secondary findings. We also filtered the carrier call-set by allele frequency and found 78 variants with allele frequency $\geq$ 0.005. For further reduction of ambiguity, we then filtered out variants if the allele frequency was also $\geq$ 0.005 in any East Asian population in public databases, including ExAC EAS, gnomAD genome AF eas, and gnomAD exome AF eas. The final filtered carrier call-set consists of nine variants.

On the TaiwanGenomes website, a user can set the filter combination as "CLNSIG: Pathogenic, Pathogenic/Likely pathogenic, Likely pathogenic," resulting in these 1,136 variants. By adding the filter AF $\geq$ 0.005, 78 variants are retrieved, reduced to 54 by adding the filter AF $\geq$ 0.01.

Most affected samples had only one variant from among the 53 variants considered secondary findings; average occurrence was about 5.54 % (Table 2). However, the high occurrence mainly arises from three disease genes with autosomal recessive inheritance: *MUTYH*, *ATP7B*, and *GAA*. The ACMG guidelines suggest that athogenic or expected pathogenic variants from both alleles should be reported together. None of the samples in the TWB 1496 cohort harbored a second pathogenic variant, indicating an average occurrence of 1.7 %. Notably, one pathogenic variant with high occurrence is *PTEN* (NM 000314.6:c.802-2A > T, rs587782455), a splicing-acceptor variant, which did not pass the VQSR threshold and thus was excluded from further analysis.

75.3% of participants had at least one clinically important pharmacogenomic allele

For PharmGKB-reported clinical variant data with a level of evidence of 1A/1B/2A, there were 13 HLA alleles and 17 clinically significant variant-drug pairs. We evaluated the frequencies of the HLA haplotypes as risk alleles with matched nomenclature fields from 1,103 TWB subjects (Table S2). Most HLA haplotypes in PharmGKB are three or four fields, suggesting that high-resolution genotypes were necessary. Among the 17 variant-drug pairs, we found 16 pairs in the TWB cohort; 13 pairs had haplotype frequencies > 1 %. The most common haplotype was *HLA-C*01:02:01* group alleles (16.05 %), followed by *HLA-A*33:03* (12.93 %). Two alleles were associated with the adverse drug events (ADEs) of taking methazolamide and allopurinol, respectively. However, one known pharmacogenomics allele, *HLA-DRB*08:03* (8 %, 178/2206), was absent from the PharmGKB clinical variant list, suggesting that a systematic review of this resource may be necessary. Among the 1,103 TWB subjects, we found 439 people who carried one risk allele (439/1103 = 39.8 %), 147 people who had at least two, 170 people who had three, 66 people who had four, and nine people who had five risk alleles. This haplotype

Table 1
Variant statistics of 1496 WGS from Taiwan Biobank.

		1496 WGS		
		ALL	Pass	Pass & globally novel
Number of variants after normalization		59,433,212	51,135,411	16,520,159
Minor allele frequency (MAF)	MAF < 0.05	49,701,036	42,557,774	16,066,996
	MAF < 0.01	44,591,573	38,599,450	15,495,346
Coding regions	All	480,784	439,192	26,664
	Nonsynonymous	274,265	250,940	700
Loss of function	Frameshift	10,225	9,528	3,174
	Nonsense	6,402	5,746	151
	Splicing	6,871	5,663	834

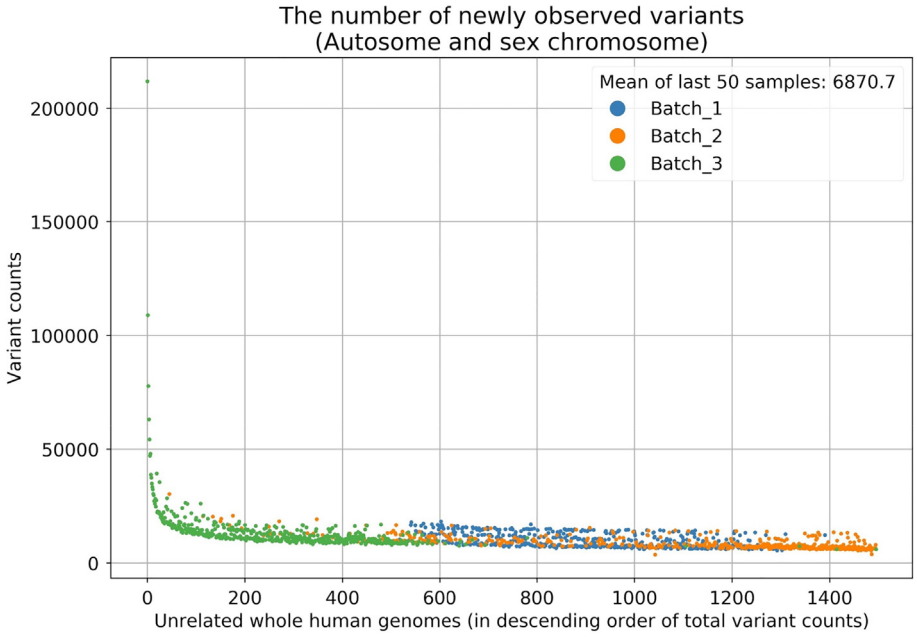


Fig. 1. Novel variants per individual. Novel variants per individual. As more individuals are sequenced, the number of newly observed variants in an individual dramatically decreases. The mean of the globally novel variants among the last 50 individuals was 6,870.7.

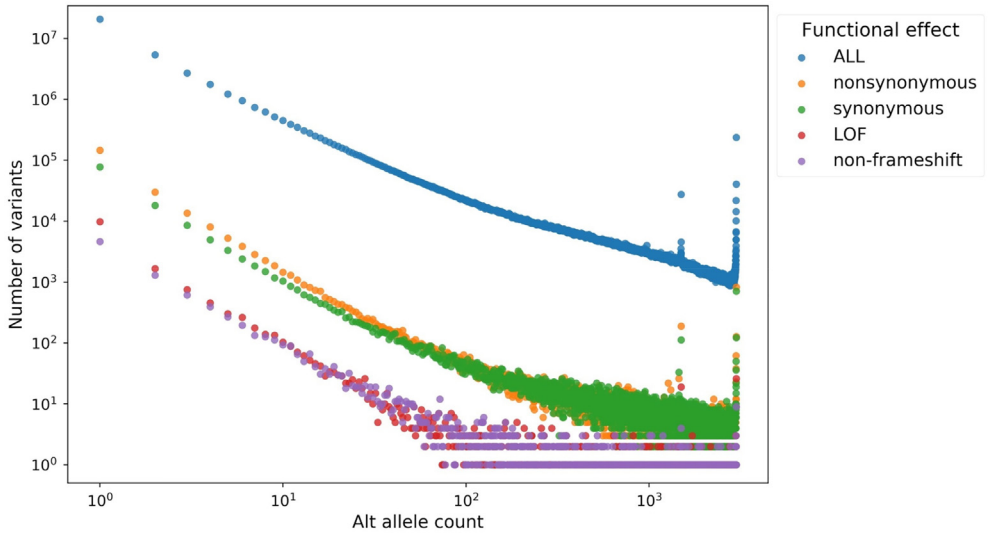


Fig. 2. Annotated functional effect and occurrence of all passed variants. The spectrum between alternative allele counts and number of variants in different categories is shown. Variants were categorized based on the ANNOVAR ExonicFunc.RefGene annotation. LOF: loss of function, including frameshift_deletion, frameshift_insertion, stopgain, and stoploss. non-frameshift: including non-frameshift_deletion and non-frameshift_insertion. nonsynonymous: non-synonymous_SNV. synonymous: synonymous_SNV.

Table 2
Medically actionable (ACMG SF v3.0) variants in the Taiwan Biobank cohort.

Phenotypes	Gene	Inheritance	# of variants	# of samples	Sample percentage
cancer	*PTEN*	AD	1^	16^	0.01070^
	MUTYH	AR	5	7	0.00468
	BRCA2	AD	4	4	0.00267
	MSH6	AD	2	2	0.00134
	PALB2	AD	1	1	0.00067
cardiovascular	*MYBPC3*	AD	3	4	0.00267
	TTN	AD	3	3*	0.00201
	FBN1	AD	2	2	0.00134
	LDLR	AD	2	2	0.00134
	TNNT2	AD	2	2	0.00134
	DSG2	AD	1	1	0.00067
	KCNQ1	AD	1	1	0.00067
	TNNI3	AD	1	1	0.00067
metabolism	*GAA*	AR	8	19	0.01270
	BTD	AR	4	4	0.00267
	GLA	XL	1	2	0.00134
miscellaneous	*ATP7B*	AR	13	28*	0.01872
Overall		AD/XL	23	25*	0.01673
		AR	30	58*	0.03877

* : One sample had two variants (on ATP7B and on TTN); in total there were 82 samples.
^ : The PTEN pathogenic variant NM_000314.6:c.802-2A>T did not pass the VQSR threshold and was not included in the overall sample percentage. Gene symbols that are underlined are genes newly added to the ACMG secondary finding gene list in v3.

frequency analysis suggests that approximately three out of four (831/1103 = 75.3 %) Taiwanese people have at least one risk allele, implying that a considerable proportion of people may benefit from HLA typing whenever a prescription for a corresponding drug is needed.

In addition, we also evaluated the allele frequencies of PharmGKB-reported CYP variant-drug pairs in the sub-sample of 1,017 TWB volunteers (manuscript in preparation). CYP variants analysis revealed 89 pharmacogenetic variant-drug groups with a level of evidence of 1A/1B/2A (Table S3). Of these groups, 54 had allele frequencies (excluding WT) $\geq$ 10 % in the TWB cohort. The most common variant was *CYP3A5*3* (73.2 %), followed by haplotypes of the *CYP2C19* group (*CYP2C19*2*, *CYP2C19*3*, *CYP2C19*9*, *CYP2C19*10*, *CYP2C19*17*, *CYP2C19*24*, and *CYP2C19*26*) (36.2 %). The variants were associated with the abnormal drug metabolism of tacrolimus and omeprazole, implying that CYP typing is helpful for clinical medication safety. Since treatment alterations may occur when a genetic variant alters a treatment's efficacy, dosage, metabolism, or pharmacokinetics or otherwise causes toxicity or an adverse drug reaction (ADR), such information is valuable for both clinicians and patients.

The status of carrier variants in Taiwan, Singapore, and Japan

To evaluate genomic characteristics specific to the Taiwanese, we compared the MAF of 1,136 carrier call-set from our samples with those of two different East Asian populations (Japanese and Singaporean). There were 279 common carrier variants between the Japanese and Taiwanese data sets, and 611 common carrier variants between the Singaporean and Taiwanese data sets (Fig. 3).

Monogenic disease risk in the offspring of 1.2–3.9 % of couples

We compared our carrier data with that of the United States cohort [34,35] and another domestic data set generated using the array genotyping method (TWBv2) [6]. The main results are listed in Table 3. The most striking difference relates to the *GJB2* gene, a well-known causative gene for hearing impairment. The estimated carrier frequency in the Taiwan biobank NGS database was 16.7 %, more than 90 % of which was contributed by the *GJB2* V37I variant (MAF = 8.6 %). *GJB2* P/LP carrier frequency in the US cohort was only 6.25 %, less than half of the Taiwanese frequency. In contrast, the *GJB2* carrier rate was estimated as a much lower 1.59 % with TWBv2 array genotyping.

Another example is the *SLC25A13* gene, related to citrullinemia. P/LP carrier frequency was 2.57 % in our series but only 0.40 % in the U.S. cohort and 1.9 % in the TWBv2 series. Another one is the *PTS* gene, defective variants of which lead to phenylketonuria (PKU). The P/LP carrier frequency of *PTS* was 0.66 % in our cohort and only 0.12 % in the U.S. cohort. The frequency in our samples is compatible with clinical observation of PKU patients in Taiwan, where BH4-deficiency (defective *PTS*) PKU patients account for up to 1/4 of total PKU patients. In contrast, defective *PTS* PKU patients only account for 1–2 % of Caucasian cohorts. No *PTS* variants were shown in the TWBv2 array data.

We applied the method used in Westemeyer et al. [34] to the TWB 1496 NGS cohort data to calculate the combined at-risk couple rate in Taiwan. The employed 270 gene panel (274 genes in Westemeyer et al.) omits 4 genes (*HBA1/2*, *DMD* for Duchenne muscular dystrophy, *SMN1*, and *FMR1*) that would identify the risk for a genetic disorder in the offspring of 1 in 28 couples (3.55 %). If *DUOX2* for hypothyroidism and *G6PD* for glucose-6-phosphate dehydrogenase deficiency were added to the local carrier screening list, the risk ratio would be 1 in 25 couples (3.94 %) (Table 4).

For thalassemia, the SV-related hereditary disease in Taiwan, we used Manta [38] and AnnotSV [27] to identify *HBA1/2* pathogenic variants in a subset of the TWB cohort (N = 494). Alpha thalassemia carrier frequency was 6.88 % (5.06 % ^{0}, 1.81 % $^{+}$) (Table S4). For spinal muscular atrophy (SMA), another high-prevalence hereditary disease, the *SMN1*/*SMN2* copy number was analyzed by SMNCopyNumberCaller [39]. Ten carriers with only one copy of *SMN1* were identified among the 494 samples, equating to an SMA carrier frequency of 2.02 % (Table S5). Validation of NGS-based SMA carrier screening methods in China [39,40] makes a convincing case that these assessments are as accurate as the traditional MLPA method. The comparison of alpha thalassemia carrier and SMA carrier rate in neighboring Asian regions are listed in Table S6 and S7.

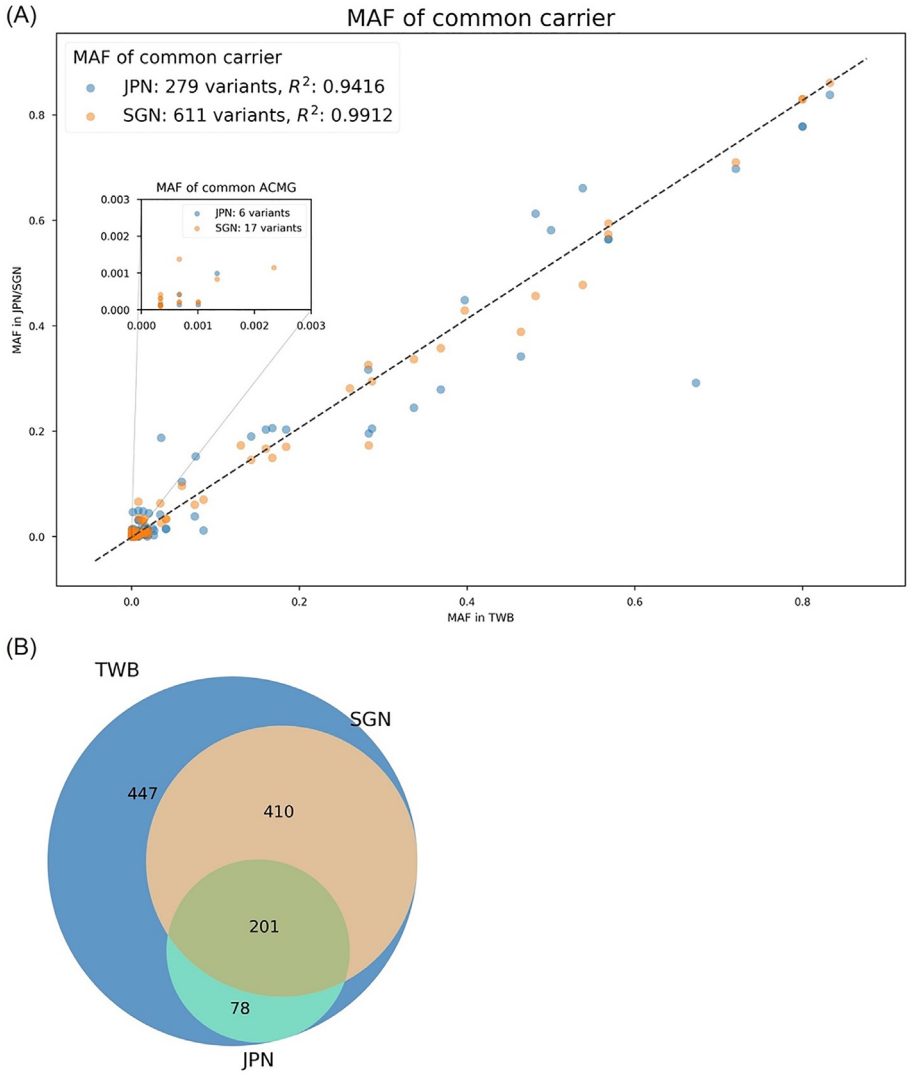


Fig. 3. Comparison of allele frequencies between TWB, JPN, and SGN populations. (A) X-axis: Minor allele frequency (MAF) of ClinVar pathogenic/likely pathogenic (P/LP) variants in the 1496 TWB cohort. Y-axis: MAF in Japan (JPN) and Singapore (SGN) populations. There were 279 carrier variants found in both the TWB and the JPN samples, and 611 carrier variants found in both the TWB and the SGN samples. Six of 54 ACMG secondary finding variants were present in both TWB and JPN samples, and 11 in both TWB and SGN samples. (B) Venn diagram illustrating the distribution of ClinVar P/LP variants among the three populations.

Table 3
Estimated carrier frequency comparison.

Gene	Disease Name	Carrier frequency		
		Our cohort	Westemeyer et al. [35]	C.-Y.Wei et al. [6]
GJB2	Hearing loss (non-syndromic)	16.68 %	6.25 %	1.59 %
VPS13B	Cohen syndrome	2.78 %	0.33 %	
CFTR	Cystic fibrosis*	2.73 %	3.85 %	
GALT	Galactosemia	2.64 %	0.67 %	
SLC25A13	Citrin deficiency	2.57 %	0.40 %	1.94 %
DUOX2	Congenital hypothyroidism	2.34 %		
GALC	Krabbe disease	2.18 %	1.27 %	1.67 %
SLC26A4	Pendred syndrome	2.04 %	1.43 %	1.70 %
ATP7B	Wilson disease	1.98 %	1.52 %	1.77 %
G6PD	G6PD deficiency	1.98 %		2.49 %
PAH	Phenylketonuria	1.32 %	2.50 %	0.48 %
HBB	Beta-Hemoglobinopathies	1.32 %	3.23 %	0.59 %
PTS	PTPS deficiency (PKU)	0.66 %	0.12 %	

* rs551227135 (carrier frequency 1.9 %) is denominanted the *CFTR* IVS8-5T variant in previous literature. It is considered a "mild" *CFTR* mutation. When in trans with a known pathogenic *CFTR* mutation (e.g. ΔF508), the IVS8-5T allele is responsible for a non-classic CF phenotype, such as the bilateral absence of vas deferens, recurrent pancreatitis, and late-onset cystic fibrosis [36]. A domestic study revealed individuals homozygous for the IVS8-5T allele as the sole variation of the whole *CFTR* coding sequence may present as non-classic CF with sinopulmonary disease and male infertility [37].

Table 4
Reproductive risk detection rate with different panels.

Expanded carrier gene panel	Accumulated risk	Couples at risk
270 gene panel	3.55 %	1 in 28
270 gene panel + *DUOX2* + *G6PD*	3.94 %	1 in 25
272 gene panel-adjustment (*CFTR*/*GALT*)*	3.84 %	1 in 26
272 gene panel-adjustment (*CFTR*/*GALT*/*GJB2*)*	1.17 %	1 in 84

* after elimination of *CFTR*(rs551227135)/*GALT*(rs2070074)/*GJB2*(rs72474224) homozygous weak pathogenic variant condition.

Variants with clinical significance and high allele frequency in Taiwan

Initial data filtering with allele frequencies singled out 54 variants with MAF > 0.01 and 24 with MAF 0.005 –0.01 among the carrier call-set from 1496 TWB participants. To address the characteristics of the Taiwanese population, we further filtered these 78 variants by comparing their MAF with those in the gnomAD genome database (version 2.1.1) and the ExAC database, yielding 9 variants with clinical significance and relatively high MAF in Taiwan (Table S8).

High prevalence (MAF ≥ 0.01) clinical significance variants among Taiwan Biobank participants

Splice-donor variant in *CACNA1B*: A splice-donor variant in *CACNA1B* (c.390 + 1 390 + 2insACGACACGGAGC) occurred in the TWB1496 data set with a MAF of 0.019, but was not reported in ExAC and the East Asian population of gnomAD.

Stop-gained variant in *DUOX2*: The stop-gained variant c.1588A > T in *DUOX2* occurred with a MAF of 0.013 among TWB participants, while its MAF in the East Asian population in gnomAD was 0.0071. The protein product of *DUOX2* is an oxidase and part of the peroxide-generating system located at the apical membrane of thyroid follicular cells [41,42]. Prematurely terminated protein products of *DUOX2* of the noted pathogenic variant lead to a lower level of hydrogen peroxide and consequently insufficient thyroid hormone for normal human development. This variant has been identified in patients with transient or permanent hypothyroidism or iodide organification.

Medium prevalence (0.005 ≤ MAF ≤ 0.01) clinically important variants among Taiwan Biobank participants

Missense variant in *OPN1MW*: The missense variant c.989G > A in *OPN1MW* is reported to be causative of deuteranopia. *OPN1MW*, the medium-wave-sensitive opsin-1 gene, is mapped to chromosome Xq28 and encodes the green cone pigment, which is crucial to color vision. In the presence of this variant, the absorbance of the Arg330Gln mutant opsin decreased dramatically compared to normal green opsin, and it has been reported to be causative of deutan color blindness [43]. A precise study on the prevalence of deuteranopia in Taiwan may help explain the relatively high allele frequency of this variant among Taiwan Biobank participants (MAF = 0.008) compared to that in the gnomAD East Asian population. It is also worth to note that less than 50 % of individuals in gnomAD v2.1.1 exomes covered this locus.

Intronic variants in *SLC25A13*: The variant c.615 + 5G > A in *SLC25A13* occurred with a MAF of 0.006 in TWB participants. *SLC25A13* is localized to chromosome 7q21.3 and encodes citrin, which serves as a mitochondrial solute transporter in the urea cycle. The variant is linked to neonatal intrahepatic cholestasis caused by citrin deficiency (NICCD) and adult-onset type II citrullinemia (CTLN2) [44–46].

Frameshift variant in *SERPINB7*: Splice-acceptor variant c.522dup in *SERPINB7* occurred with a MAF of 0.005 among TWB participants, while its allele frequency in the East Asian population in gnomAD2 was 0.0032. This variant is causative of Nagashima-type palmoplantar keratoderma, a skin disorder characterized by hyperkeratosis of the palm and feet of affected individuals [47,48].

Frameshift variant in *TTLL5*: The frameshift variant c.3177 3180del in *TTL5* occurred with a MAF of 0.005 among TWB participants, while its MAF in ExAC and in the East Asian population of gnomAD was 0.0021 and 0.0045, respectively.

TaiwanGenomes web browser

TaiwanGenomes (https://genomes.tw) provides a user-friendly interface to access all variants reported in this study. The'MAIN' tab links to the main table that contains 59,433,212 variants yielded by the joint calling of the 1496 WGS, including 51,135,411 variants passing the VQSR analysis (setting 'VQSR = PASS'). The'SUPPLEMENT' tab links to the supplementary table that provides information on read depth for 2,792,591,408 positions in the human genome. These positions are categorized into four classes. A: Reference allele (MAF = 0); B: Missing (MAF = n.a); C: Uncertain quality (MAF = n.a./0); and the 59,433,212 variants called that have links to the MAIN table. TaiwanGenomes provides users with the flexibility of selecting columns of interest to examine. Among the passed variants, 439,192 variants fall in the coding regions (setting 'FILTER = PASS' and 'fun.refGene = exonic'), and 55,949 have a minor allele frequency ≥ 0.01 (setting 'FILTER = PASS' and 'fun.refGene = exonic' and 'AF ≥ 0.01′). Another example of setting condition combinations of multiple selected columns is examination of all nonsynonymous variants in *BRCA1* by setting 'Gene.refGene = BRCA1′ and 'ExonicFunc.refGene = nonsynonymous SNV', resulting in 40 variants.

Discussion

It is noteworthy that all TWB subjects were 20 year-old adults with no distinct developmental defects or malignant tumors at the time they joined the study. Unlike the genotyping array approach that focuses on detecting the associations between genomic haplotype blocks and phenotypic traits, personal genome sequences directly uncover all hereditary risks. In this study, we reanalyzed 1496 whole-genome sequence data sets from the Taiwan Biobank, which is one of the few existing valuable datasets for the east Asian population. Our reanalysis generated acomprehensive medical genomic profile, which complements previous Taiwan Biobank studies and reinforces the significance of WGS data. We utilized benchmarked data from GIAB and jointly called 1496 samples to determine a cut-off VQSR tranche of 99.7, and provided VQSR quality flags for each locus on the website (**TaiwanGenomes** https://genomes.tw). We further selected hundreds of variants irrespective of VQSR tranche and validated them by the Sanger sequencing. All variants that were removed in the joint calling step had negative Sanger sequencing results, indicating the effectiveness of joint calling to eliminate false positives in individual variant calling. Among variants that were retrieved by joint calling, only some had positive Sanger sequencing results, suggesting that further VQSR filtering criteria may be necessary. For the entile reference genome, we classified all non-alternative allele regions into three categories by overall call rate and depth of coverage. This allowed us to clearly distinguish reference homozygous regions or difficult-to-map regions in the genome reference for the

TWB 1496 cohort. We found that our TWB samples yielded an average of 6,871 globally novel variants for a Taiwanese person.

In this study, we utilized WGS data to estimate carrier risk in Taiwan. For the secondary findings, we checked for presence of the variant in ACMG SF v3.0 gene list. This is the first time ACMG SF v3.0 has been applied to the East Asian population. We found that 5.54 % of the 1496 participants were carriers of ACMG SF v3.0 genes, that 1.67 % carried autosomal dominant and X-linked diseases, and 3.87 % carried recessive diseases. Notably, one pathogenic variant NM000314.6:c.802-2A > T on the *PTEN* gene was caused by a technical error in the TWB 1496 cohort. In our analysis, this variant was located at the end of a polyT region with strand bias evidence, highlighting the necessity of applying VQSR and quality checks in any reanalyses of WGS data. We also analyzed a 270 gene panel to estimate the accumulated monogenic disease risk and found that 3.55 % of offspring would be carriers, implying that approximately 1 in 28 couples might benefit from carrier screening. Two main pathogenic variant contributors were *GJB2*: NP_003995.2:p.Val37Ile (rs72474224, MAF: 8.6 %) and *CFTR*:NM_000492.4:c.1210-11_1210-10insG (rs551227135, MAF: 0.9 %). Both are considered "mild" pathogenic variants. Phenotype penetration of *GJB2*(rs72474224) is highly variable in Taiwan's clinical experience. *CFTR* (rs551227135) is one of the classical *CFTR* IVS8-5 T variants responsible for non-classical cystic fibrosis presentation (CABVD, recurrent pancreatitis, late-onset CF). The regional normalization of variants is complicated and annotations are still controversial. Another caveat to carrier screening is the inability to detect long fragment variants (such as CNV, large deletion, trinucleotide repeats, and gene conversion) from short-reads data. Use of short-read data will often underestimate the carrier frequency of four important genes: *HBA1/2*, *DMD*, *FMR1*, and *SMN1*, of which many known pathogenic variants are relatively large.

Concerning long-fragment genetic variants, we randomly selected 494 samples for deciphering the population structure variation profile on the thalassemia genes *HBA1/2* and the copy number on *SMN1/2* genes. The overall population profile was similar to those found in previous studies using a target panel approach, suggesting that WGS can potentially replace target panels once an analysis pipeline has been built (**Table S4**). Furthermore, our results showed that 75.3 % (831/1103) of the cohort was vulnerable to severe ADRs since they carried at least one risk HLA allele. Although several studies have implied that HLA types can be inferred from SNP genotyping results, full HLA haplotypes from sequence data were found to be unbiased for population-specific rare haplotypes. Similarly, SNVs on CYP genes could be detected by SNP genotyping, but WGS can reveal complete haplotype information, which is more accurate in assessing susceptibility. Our findings reveal that Taiwanese carry a high frequency (MAF > 10 %) of abnormal alleles in more than 60 % of the pharmacogenetic variant-drug groups.

Conclusions

Whole-genome sequencing has become affordable for both research and clinical use. Since a person only needs to be sequenced once in a lifetime, a more comprehensive reanalysis of their genomic profile can be of great clinical value. Our study highlights the potential uses and benefits of a complete genomic profile with medical information for at both the population and individual level.

Funding

This project was supported by the research grants from the Ministry of Science and Technology in Taiwan (MOST 109-2622-B-002-004-CC2, MOST 109-2221-E-002-162-MY3 and MOST 110-2320-B-002-078), National Science and Technology Council in Taiwan (NSTC 111-2320-B-002-091-MY3), and National Taiwan University Hospital (UN109-070). This work was also financially supported by the "Center for Advanced Computing and Imaging in Biomedicine (NTU-112L900701)" from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan.

Compliance with Ethics Requirements

All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1975, as revised in 2008 (5). Informed consent was obtained from all patients for being included in the study.

Ethics approval and consent to participate

The Institutional Review Board Approval from Biomedical Science Research of Academia Sinica, Taiwan (IRB-BM) and the Ethics and Governance Council (EGC) of Taiwan Biobank, Taiwan, were obtained and gave ethical approval for this work (AS-IRB01-18041(N)). All raw sequence data used in this study were generated as part of the Taiwan Biobank project. This study did not reveal any individual participant's information, and none of the results can be used to identify individual participants.

CRediT authorship contribution statement

Jacob Shujui Hsu: Conceptualization, Formal analysis, Funding acquisition, Writing – original draft, Writing – review & editing, Supervision. **Dung-Chi Wu:** Formal analysis, Writing – original draft, Software. **Shang-Hung Shih:** Formal analysis, Writing – original draft. **Jen-Feng Liu:** Formal analysis, Writing – original draft, Data curation. **Ya-Chen Tsai:** Formal analysis, Software. **Tung-Lin Lee:** Formal analysis, Data curation. **Wei-An Chen:** Formal analysis. **Yi- Hsuan Tseng:** Formal analysis. **Yi-Chung Lo:** Software. **Hong-Ye Lin:** Software. **Yi-Chieh Chen:** Formal analysis. **Jing-Yi Chen:** Software. **Ting-Hsuan Chou:** Formal analysis. **Darby Tienhao Chang:** Software, Supervision. **Ming Wei Su:** Data curation. **Wei-Hong Guo:** Formal analysis. **Hsin-Hsiang Mao:** Formal analysis. **Chien-Yu Chen:** Conceptualization, Funding acquisition, Software, Writing – review & editing, Supervision. **Pei-Lung Chen:** Conceptualization, Formal analysis, Data curation, Funding acquisition, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank all the participants, technicians and investigators contributing to this study. We also thank Taiwan Biobank for the supports. We are grateful to National Center for High-performance Computing (NCHC) for providing computational and storage resources. We express our sincere appreciation to Kang-Lin Wong and Hsin-Ta Chan for their invaluable technical support in constructing the database. We thank our funders, including Ministry of Science and Technology in Taiwan, National Science and

Technology Council, National Taiwan University Hospital and National Taiwan University.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jare.2023.12.018.

References

[1] Stark Z et al. Integrating Genomics into Healthcare: A Global Responsibility. Am J Hum Genet 2019;104(1):13–20.
[2] Tadaka S et al. 3.5KJPNv2: an allele frequency panel of 3552 Japanese individuals including the X chromosome. Hum Genome Var 2019;6:28.
[3] Okada Y et al. Deep whole-genome sequencing reveals recent selection signatures linked to evolution and disease risk of Japanese. Nat Commun 2018;9(1):1631.
[4] Wu D et al. Large-Scale Whole-Genome Sequencing of Three Diverse Asian Populations in Singapore. Cell 2019;179(3):736–749 e15.
[5] Cao Y et al. The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. Cell Res 2020;30(9):717–31.
[6] Wei CY et al. Genetic profiles of 103,106 individuals in the Taiwan Biobank provide insights into the health and history of Han Chinese. NPJ Genom Med 2021;6(1):10.
[7] Feng YA et al. Taiwan Biobank: A rich biomedical research database of the Taiwanese population. Cell Genom 2022;2(11):100197.
[8] Yang Z et al. Comparison of gene mutation spectrum of thalassemia in different regions of China and Southeast Asia. Mol Genet Genomic Med 2019;7(6):e680.
[9] Onore ME et al. Linked-Read Whole Genome Sequencing Solves a Double. Genes (Basel) 2021;12(2).
[10] Shang X et al. Rapid Targeted Next-Generation Sequencing Platform for Molecular Screening and Clinical Genotyping in Subjects with Hemoglobinopathies. EBioMedicine 2017;23:150–9.
[11] Chen X et al. Spinal muscular atrophy diagnosis and carrier screening from genome sequencing data. Genet Med 2020;22(5):945–53.
[12] Juang JJ et al. Rare variants discovery by extensive whole-genome sequencing of the Han Chinese population in Taiwan: Applications to cardiovascular medicine. J Adv Res 2021;30:147–58.
[13] Edwards JG et al. Expanded carrier screening in reproductive medicine-points to consider: a joint statement of the American College of Medical Genetics and Genomics, American College of Obstetricians and Gynecologists, National Society of Genetic Counselors, Perinatal Quality Foundation, and Society for Maternal-Fetal Medicine. Obstet Gynecol 2015;125(3):653–62.
[14] Miller DT et al. ACMG SF v3.0 list for reporting of secondary findings in clinical exome and genome sequencing: a policy statement of the American College of Medical Genetics and Genomics (ACMG). Genet Med 2021.
[15] Tang CS et al. Actionable secondary findings from whole-genome sequencing of 954 East Asians. Hum Genet 2018;137(1):31–7.
[16] Kuo CW et al. Frequency and spectrum of actionable pathogenic secondary findings in Taiwanese exomes. Mol Genet Genomic Med 2020;8(10):e1455.
[17] Group, e.C.A.W., Frequency of genomic secondary findings among 21,915 eMERGE network participants. Genet Med, 2020; 22(9): 1470-1477.
[18] Kendig KI et al. Sentieon DNASeq Variant Calling Workflow Demonstrates Strong Computational Performance and Accuracy. Front Genet 2019;10:736.
[19] Van der Auwera GA, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Current protocols in bioinformatics/editoral board, Andreas D. Baxevanis ... [et al.], 2013; **11** (1110): 11.10.1-11.10.33.
[20] Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010;26(5):589–95.
[21] Danecek P et al. Twelve years of SAMtools and BCFtools. GigaScience 2021;10 (2).
[22] Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 2011;27(21):2987–93.
[23] Krusche P et al. Best practices for benchmarking germline small-variant calls in human genomes. Nat Biotechnol 2019;37(5):555–60.
[24] Zook JM et al. An open resource for accurately benchmarking small variant and reference calls. Nat Biotechnol 2019;37(5):561–6.
[25] Dong C et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet 2014;24(8):2125–37.
[26] Fuchsberger C et al. The genetic architecture of type 2 diabetes. Nature 2016;536(7614):41–7.
[27] Geoffroy V et al. AnnotSV: an integrated tool for structural variations annotation. Bioinformatics 2018;34(20):3572–4.
[28] Chen PL et al. Genetic determinants of antithyroid drug-induced agranulocytosis by human leukocyte antigen genotyping and genome-wide association study. Nat Commun 2015;6:7633.
[29] Numanagic I et al. Allelic decomposition and exact genotyping of highly polymorphic and structurally variant genes. Nat Commun 2018;9(1):828.
[30] Lee SB et al. Stargazer: a software tool for calling star alleles from next-generation sequencing data using CYP2D6 as a model. Genet Med 2019;21 (2):361–72.
[31] Landrum MJ et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res 2018;46(D1):D1062–7.
[32] Karczewski KJ et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 2020;581(7809):434–43.
[33] Nagasaki M et al. Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. Nat Commun 2015;6:8018.
[34] Westemeyer M et al. Correction: clinical experience with carrier screening in a general population: support for a comprehensive pan-ethnic approach. Genet Med 2020;22(7):1282.
[35] Westemeyer M et al. Clinical experience with carrier screening in a general population: support for a comprehensive pan-ethnic approach. Genet Med 2020;22(8):1320–8.
[36] Cottin V et al. Late CF caused by homozygous IVS8-5T CFTR polymorphism. Thorax 2005;60(11):974–5.
[37] Wu CC et al. Mutation spectrum of the CFTR gene in Taiwanese patients with congenital bilateral absence of the vas deferens. Hum Reprod 2005;20 (9):2470–5.
[38] Kosugi S et al. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biol 2019;20(1):117.
[39] Feng Y et al. The next generation of population-based spinal muscular atrophy carrier screening: comprehensive pan-ethnic SMN1 copy-number and sequence variant analysis by massively parallel sequencing. Genet Med 2017;19(8):936–44.
[40] Zhao S et al. NGS-based spinal muscular atrophy carrier screening of 10,585 diverse couples in China: a pan-ethnic study. Eur J Hum Genet 2021;29 (1):194–204.
[41] Moreno JC et al. Inactivating mutations in the gene for thyroid oxidase 2 (THOX2) and congenital hypothyroidism. N Engl J Med 2002;347(2):95–102.
[42] Vigone MC et al. Persistent mild hypothyroidism associated with novel sequence variants of the DUOX2 gene in two siblings. Hum Mutat 2005;26 (4):395.
[43] Ueyama H et al. Novel missense mutations in red/green opsin genes in congenital color-vision deficiencies. Biochem Biophys Res Commun 2002;294 (2):205–9.
[44] Kobayashi K et al. Screening of nine SLC25A13 mutations: their frequency in patients with citrin deficiency and high carrier rates in Asian populations. Mol Genet Metab 2003;80(3):356–9.
[45] Lu YB et al. Frequency and distribution in East Asia of 12 mutations identified in the SLC25A13 gene of Japanese patients with citrin deficiency. J Hum Genet 2005;50(7):338–46.
[46] Song YZ et al. Genotypic and phenotypic features of citrin deficiency: five-year experience in a Chinese pediatric center. Int J Mol Med 2011;28(1):33–40.
[47] Kubo A et al. Mutations in SERPINB7, encoding a member of the serine protease inhibitor superfamily, cause Nagashima-type palmoplantar keratosis. Am J Hum Genet 2013;93(5):945–56.
[48] Kabashima K et al. "Nagashima-type" keratosis as a novel entity in the palmoplantar keratoderma category. Arch Dermatol 2008;144(3):375–9.