Thai medical population genomics based on Brugada syndrome cohort

Mr. John Mauleekoonphairoj

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in Biomedical Sciences (Interdisciplinary Program)
Inter-Department of Biomedical Sciences
GRADUATE SCHOOL
Chulalongkorn University
Academic Year 2022

การศึกษาลักษณะทางพันธุกรรมของมนุษย์ในประชากรไทยจากโครงการวิจัยบรูกาดา

นายจอน เมาฬีกุลไพโรจน์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรดุษฎีบัณฑิต
สาขาวิชาชีวเวชศาสตร์ (สหสาขาวิชา) สหสาขาวิชาชีวเวชศาสตร์
บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2565

Thesis Title	Thai medical population genomics based on Brugada syndrome cohort
By	Mr. John Mauleekoonphairoj
Field of Study	Biomedical Sciences (Interdisciplinary Program)
Thesis Advisor	Professor YONG POOVORAWAN, M.D.

Accepted by the GRADUATE SCHOOL, Chulalongkorn University in Partial Fulfillment of the Requirement for the Doctor of Philosophy

..... Dean of the GRADUATE SCHOOL
(Associate Professor YOOTTHANA CHUPPUNNARAT, Ph.D.)

DISSERTATION COMMITTEE

..... Chairman
(Sissades Tongsima, Ph.D.)

..... Thesis Advisor
(Professor YONG POOVORAWAN, M.D.)

..... Examiner
(Associate Professor DUANGDAO WICHADAKUL, Ph.D.)

..... Examiner
(Professor SUNCHAI PAYUNGPORN, Ph.D.)

..... Examiner
(Professor Apichai Khongphatthanayothin, M.D.)

จอน เมาหีกุลไพโรจน์ : การศึกษาลักษณะทางพันธุกรรมของมนุษย์ในประชากรไทยจากโครงการวิจัยบรูกาดา. (Thai medical population genomics based on Brugada syndrome cohort) อ.ที่ปรึกษาหลัก : ศ. นพ.ยง ภู่วรวรรณ

งานวิจัยทางพันธุกรรมของมนุษย์ส่วนใหญ่ศึกษาในประชากรที่มีลักษณะทางเชื้อชาติจากทวีปยุโรป จึงส่งผลให้ข้อมูลทางพันธุกรรมในประชากรอื่นรวมถึงประชากรไทยมีจำนวลจำกัด ส่งผลให้บางครั้งไม่สามารถผลที่ได้จากงานวิจัยทางพันธุศาสตร์ในประชากรยุโรปมาใช้ในประชากรอื่นเนื่องจากความหลากหลายทางพันธุกรรมที่แตกต่างกัน งานวิจัยนี้จึงศึกษาความหลากหลายทางพันธุกรรมที่พบในประชากรไทยโดยใช้ whole genome sequences (ส่วนที่ 1) เริ่มจากความหลากหลายทางพันธุกรรมที่ส่งผลต่อการใช้ยาหรือ pharmacogenomics (ส่วนที่ 2) ความหลากหลายทางพันธุกรรมที่เกี่ยวข้องกันโรค autosomal recessive และ(ส่วนที่ 3) ความหลากหลายทางพันธุกรรมที่มีรายงานว่าเกี่ยวข้องกับความรุนแรงจากติดเชื้อ COVID-19 นอกจากนี้ (ส่วนที่ 4) ยังได้ศึกษาผลกระทบของความหลากหลายทางพันธุกรรมต่อการเลือก reference panel ที่ใช้ในการคาดการณ์ genotype หรือimputation ในส่วนที่ 1 ผลการศึกษาพบว่าในยีน CYP3A5, CYP2C19, CYP2D6, NAT2, SLCO1B1, และ UGT1A1 มี diplotype ทีส่งผลต่อการตอบสนองต่อยาที่ผิดปกติมากกว่า 25% ของประชากรไทย รวมถึงยังพบ variant CYP3A5*3 *(rs776746), CYP2B6**6 (rs2279343), และ NAT2 (rs1041983) มากกว่าในคนไทยเมื่อเทียบกับชาวตะวันออกและประชากรโลกในฐานข้อมูล GnomAD อย่างมีนัยสำคัญ การศึกษายังพบอีกว่ามี 121 variants ที่ยังไม่เคยมีรายงานแต่ผลวิเคราะห์ชี้ว่าน่าจะส่งผลต่อการการทำงานของโปรตีน โดย 60.3% ของ variant ในกลุ่มนี้ไม่มีรายงานในฐานข้อมูลประชากร gnomAD ใน (ส่วนที่ 2) การศึกษาความหลากหลายทางพันธุกรรมที่เกี่ยวข้องกันโรค autosomal recessive พบว่ามี 263 variants ที่เคยรายงานว่าสามารถก่อให้เกิดโรค โดย 6 variant พบว่ามีผู้ที่เป็นพาหะมากถึง 1% ของประชากรไทย การวิเคราะห์การกระจายตัวของ variants กลุ่มนี้ในประชากรไทยโดยการทำ fine-scale genetic structure analysis พบว่ามีความชุกของผู้เป็นพาหะของโรคธาลัสซีเมีย โรคแกลคโทซีเมีย และ โรคหูหนวกในบางกลุ่มของประชากรไทยจากการศึกษา (ส่วนที่ 3) ความหลากหลายทางพันธุกรรมที่มีรายงานว่าเกี่ยวข้องกับความรุนแรงจากติดเชื้อ COVID-19 พบว่า variant ที่ chromosome 3p21.31 ซึ่งมีความสัมพันธ์สูงกับความรุนแรงของโรคและได้รับการรับรองในหลายการศึกษามีความชุกที่แตกต่างกันในแต่ละประเทศในภูมิภาคเอเชียตะวันออกเฉียงใต้ โดยพบในชาวฟิลิปปินส์ที่ความชุก 0.21 แต่พบแค่ 0.06 ในประชากรไทยและแทบไม่พบเลยในประชากรเอเชียตะวันออกเฉียงเหนือ จากศึกษา(ส่วนที่ 4) ผลกระทบของความหลากหลายทางพันธุกรรมในชาวไทยต่อการเลือก reference panel ใน genotype imputation พบว่า reference panel ที่แตกต่างกันสงผลต่อประสิทธิภาพในการคาดการณ์ โดย TOPMed สามารถคาดการณ์ variants ได้มากที่สุด (~271 ล้าน) ในขนาดที่ GenomeAsia 100K มีความแม่นยำในการคาดการณ์ที่สุด(0.97) ถึงแม้ความแม่นยำลดลงถึง 30.3% ในกลุ่ม rare variants แต่ GenomeAsia 100K ยังให้ความแม่นยำที่สูงกว่า reference panel อื่น ผลจากการศึกษาทั้งหมดนี้แสดงถึงความหลากหลายและความแตกต่างทางพันธุกรรมในประชากรไทยเมื่อเปรียบกับประชากรอื่นในฐานข้อมูล โดยข้อมูลที่ได้จากการศึกษานี้สามารถนำไปใช้เป็นแนวทางการออกแบบการตรวจพันธุกรรมและการออกแบบงานวิจัยเชิงพันธุกรรมในประชากรไทย ถึงแม้ขนาดของตัวอย่างที่ใช้ในงานวิจัยนี้จะมีจำนวลจำกัดเมื่อเทียบกับฐานข้อมูลอื่น แต่พบ variant จำนวนมากมีลักษณะเฉพาะในกลุ่มประชากรไทย แสดงให้เห็นถึงความสำคัญของการจัดตั้งฐานข้อมูลทางพันธุกรรม ของประชากรไทย

สาขาวิชา	ชีวเวชศาสตร์ (สหสาขาวิชา)	ลายมือชื่อนิสิต .....
ปีการศึกษา	2565	ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

6381006320 : MAJOR BIOMEDICAL SCIENCES (INTERDISCIPLINARY PROGRAM)
KEYWORD:

John Mauleekoonphairoj : Thai medical population genomics based on Brugada syndrome cohort. Advisor: Prof. YONG POOVORAWAN, M.D.

Human genomic research has been concentrated in populations of European descent resulted in large portion of the global populations, including Thais, underrepresented. The bias in representation limited transferability of genetics findings to understudied populations and exacerbate health disparities. This study aims to examine medically relevant genetic variation in Thai population uses whole genome sequences. The study examined prevalence of pharmacogenomics variants (part I), variant associated with autosomal recessive disorder (part II) and risk alleles recently identified to associate with severe COVID-19 infection symptoms (part III). The study further examined the effect of genetic variation in Thais on reference panel selection for genotype imputation (part IV). In pharmacogenomics, over 25% of Thais carried a high-risk diplotype in CYP3A5, CYP2C19, CYP2D6, NAT2, SLCO1B1, and UGT1A1 genes. Allele frequencies of CYP3A5*3 (rs776746), CYP2B6*6 (rs2279343), and NAT2 (rs1041983) were significantly higher in Thais than East-Asian and global populations. 121 variants, which is unreported, have potential to exert clinical impact, majority were rare and population-specific, with 60.3% of variants absent from gnomAD database. In examining variants associated with autosomal recessive disorder, 263 likely pathogenic/ pathogenic variants were identified with 6 well-established pathogenic variants have carrier rate of higher than 0.01. Analysis of variant distribution based on genetics structure shows significant enrichment of pathogenic variants associated with thalassemia, galactosaemic and deafness in some subpopulation. When examined prevalence of severe COVID-19 risk alleles, the frequency of risk allele at 3p21.31 locus, which was highly correlated with disease severity and replicated in multiple studies, found to differs vastly among Southeast Asians. Allele frequencies ranging from 0.21 in the Filipino population to 0.06 in the Thai population and are extremely rare in Northeast Asians. Lastly, the choice of reference panel showed to strongly affect imputation performance. While imputation using the TOPMed panel yielded the largest number of variants (~271 million), GenomeAsia 100K achieved the best imputation accuracy with a median genotype concordance rate of 0.97. GenomeAsia 100K also offered the best accuracy for rare variants with 30.3% reduction in concordance rates. In conclusion, this study reports genetic variations in Thai that are clinically relevance in different fields of medical science. This study findings provide an essential information that have wide range of application from the design of genetic testing through to conducting genomic research. In addition to the prevalence of multiple variants in Thai found to differ from other global populations, large number of the variants identified are population-specifics. This stresses the importance of constructing Thai genetic database with larger sample size to enable a better understanding of low frequencies and rare variants in the population that often exert higher clinical impact.

Field of Study:	Biomedical Sciences (Interdisciplinary Program)	Student's Signature .....
Academic Year:	2022	Advisor's Signature .....

ACKNOWLEDGEMENTS

I would like to thank Prof. koonlawee Nademanee, Prof. Yong Poovorawan and Prof. Apichai Khongphatthanayothin for the opportunity to conduct this research and their guidance throughout the project. The committee members, Dr. Sissades Tongsima, Assoc. Prof. Duangdao Wichadakul, Prof. Sunchai Payungporn and Prof. Apichai Khongphatthanayothin, for reviewing this dissertation and their valuable recommendations. Collaborators from Amsterdam University Medical Centre, University of Amsterdam for genotype array data and Connie R. Bezzina, Sean J. Jurgens, Dominic S. Zimmerman for reviewing the manuscript and for their comments.

I would like to acknowledge Duangkamon Ittipcharoen, Boosamas Sutjaporn and Pharawee Wandee for their dedication in collecting samples from various regions throughout Thailand. Members of the Brugada consortium and the Thai Red Cross Society for their contribution in samples collection. CMKL University, PMU-C, and Center of Excellence in Medical Genomics, Chulalongkorn University for computational resources. The Centre of Excellence in Clinical Virology, Faculty of Medicine, Chulalongkorn University for technical assistance in laboratory work.

Lastly, I would like to thank the National Research Council of Thailand for the support of Preventing Lai-Tai among Thais: Discovering the Genetic Causes and Treatments of Lai-Tai (Sudden Unexpected Nocturnal Death Syndrome or Brugada Syndrome), where wholes genomes sequences form the project were used in this dissertation and for the support of the Second Century Fund (C2F), Chulalongkorn University.

John Mauleekoonphairoj

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

Introduction

Aim and Rational

Sequencing human genome had advanced our understanding of human genetics and mechanisms of diseases that led to development of novel diagnostic tools and treatments. The technological advancement and cost reduction of next generation sequencing technology exponentially increase the availability of whole genome sequences (WGS). This led to construction of numerous WGS databases that enable the study of human genetic variation. The availability of WGS at a population level later become a valuable resource in medical research with wide range of applications including facilitate the interpretation of variant identified in rare mendelian disease patients, identifying of novel disease-causing variants or genes, use in studying the prevalence of clinically relevant variants and act as reference panel in genotype imputation.

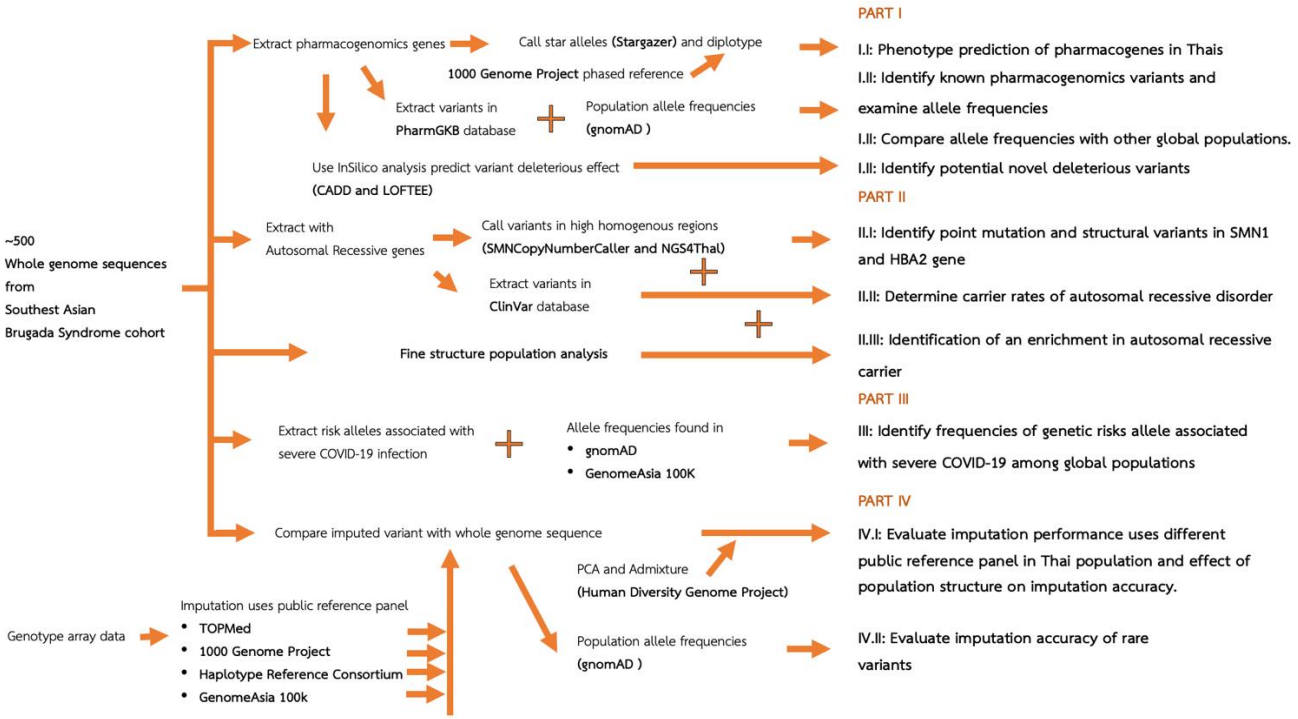
Human genomic research, however, had been previously concentrated in populations of European descent resulted in large portion of the global populations underrepresented. Such bias in representation limited transferability of genetics findings to understudied populations and further exacerbate health disparities. Southeast Asians, including Thai, are often underrepresented in public databases. Despite a global effort to increase representation of diverse population, most East Asian populations currently represented are those of Northeast Asian ancestry. This limited the knowledge on the circulating genetic variation in Thai, including population specific variants.

This study is separated into four parts. The study aims to examine genetic variation in Thai population through WGSs and it effect on different area of medical science. For the first three parts, the study examine genetic variations in Thai population that influences medical diagnostic in different fields including predicting the effect of drug absorption, distribution, metabolism, and excretion based on pharmacogenomics variants (part I), detection of autosomal recessive carrier (part II), and genetic risk

factors associated with severe symptom from COVID-19 infection (part III). The study further evaluate use of currently available reference panel in an imputation, method that are currently widely use in genomic research, of Thai and how genetic variation in Thais effect performance of these panels (part IV).

Different tool and resources were used to comprehensively analyse genetic variation of the Thai population. Information on genotype-phenotype association were extracted from multiple databases for interpretation of genetic variation identified. Different bioinformatic tools were also be used for different purposes from identify multiple variants on the same genomic strain to reanalysis of sequence reads in regions or types of variants that are not accessible using the short read WGS technology traditional pipeline. The study also leverage genetic data from multiple publicly available population databases to represent diverse global populations and to address similarities and differences of variation found in different populations. The study also employ multiple techniques used in examining population structure to study variation within Thai population and the effect it has on disease prevalence and genomic research.

Conceptual framework



Background and Literature Review

Human Genetic Variation in medical diagnostic

Understand a population genetic variation plays an important role in development of diagnostic tools and in human genomic research. Until now over 1.6 million genotype-phenotype submissions have been submitted into ClinVar. Knowing the prevalence of genetic variants within the population allow us to efficiently design diagnostic tools that identify individuals at risk. Different area uses genetic to identify individual at risk include predicting the effect of drug absorption, distribution, metabolism, and excretion based on pharmacogenomics variants and detection of autosomal recessive carrier.

Genome wide association study had been conducted on more than 3000 traits (3). Genetic variation can limit transferability of identified disease risk from one population to another. Moreover, the performance of genotype imputation, which has become a crucial step in conducting Genome wide association study, depends heavily on genetic variation between reference panel and the study population.

Pharmacogenomics

Pharmacogenomics study the effect of genetic variants on drug absorption, distribution, metabolism, and excretion. Studies have shown differences in prevalence of variants found in pharmacogenes between ethnicities and more recently closely related populations. The differences in prevalence of single nucleotide polymorphisms between ethnicities has long been established and commonly used in guidelines. A study examined allele frequencies of single nucleotide polymorphisms use in prediction of drug response and toxicity in 19 global populations found huge variation between populations (Figure 1)(4). Further study show evidence that these variations can be found down to countries when distribution of *CYP2C19* and *CYP2D6* alleles were examined in Europe(Figure 2)(5).

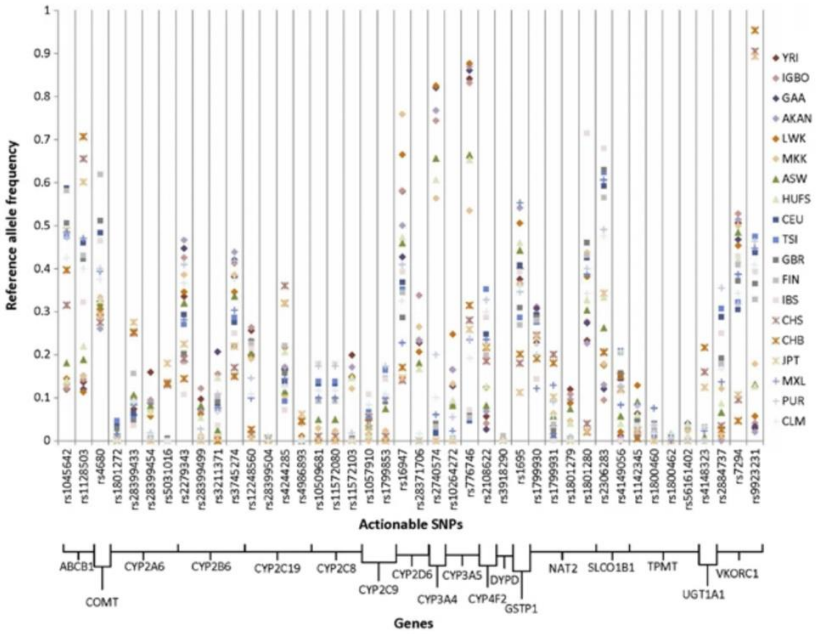


Figure 1 Variation in allele frequencies of actional single nucleotide polymorphisms in 19 global populations.

(4).

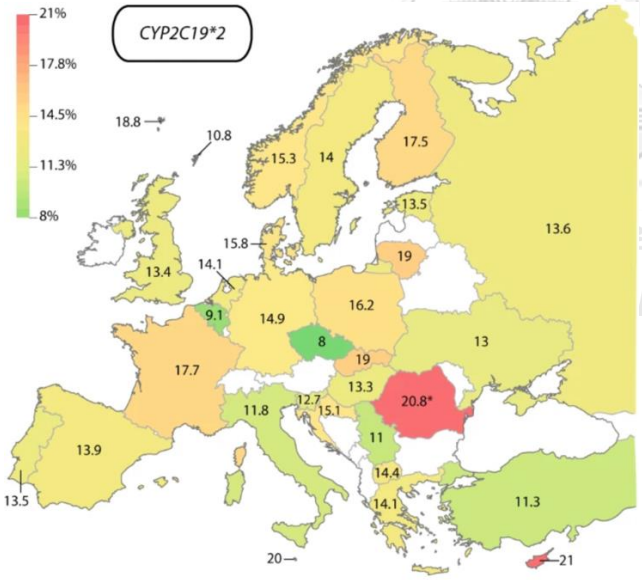


Figure 2**:** Distribution of CYP2C19*2 across different countries in Europe.

(5).

Autosomal recessive disorder

Currently there are over 2,800 know genes in Clinical Genomic Database linked to autosomal recessive disorder with estimate of over 5000 autosomal recessive genes has been proposed(6, 7). Autosomal recessive disorder was estimated to effect 1.4 in 1,000 neonates and could increase to 10-20 per 1000 individuals in geographical region where carrier variant has an evolutionary advantage, such as malaria endemic regions (8). The landscape of AR variants had demonstrated to be highly population specific. A study examined carrier rate in 415 autosomal recessive genes across six major ancestries found a huge variation in gene carrier rate among different ancestries (Figure 3)(1). Within European populations, less than 20% of carrier variants was found to shared between the Dutch and Estonian cohort(9). These findings suggest screening for a carrier using panel that designed for a specific population may better capture autosomal recessive carrier than a universal carrier screening panel. The knowledge of population carrier frequencies of autosomal recessive variants would provide a crucial information in the selection of genes for screening.

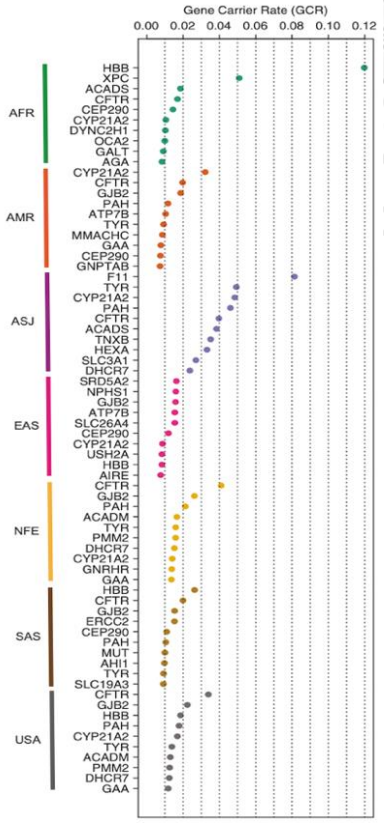


Figure 3 Gene carrier rates of the top ten genes for each ancestry.

AFR African/African American, AMR Hispanic, ASJ Ashkenazi Jewish, EAS East Asian, NFE non-Finnish European, SAS South Asian, USA composite US (1).

Genetic risks and association with severe COVID-19 among global populations

The worldwide pandemic caused by the novel coronavirus infection (COVID-19) has continued unabated as multiple factors have influenced its transmission, morbidity, and mortality. Infected older adults and those with preexisting health conditions are at risk of increased disease severity. Progression to acute respiratory failure accompanies prolonged hospitalization and poor prognosis. Recent genome-wide association studies identified multiple host genetic factors associated with disease susceptibility and severity (10-12).

Chromosomal locus 3p21.31 was highly correlated with disease severity in hospitalized Italian and Spanish COVID-19 patients (rs11385942; 95% confidence interval (CI), p = 1.15x10−10)(figure 4)(10), which was confirmed in the United Kingdom (rs13078854; 95% CI, p = 1.6x10−18)(11) and in a multi-ethnic study (rs73064425; 95% CI, p = 4.77x10−30)(12). This gene-rich locus includes SLC6A20 (encoding sodium-imino acid transporter 1, which interacts with COVID-19 ACE2 receptor) and multiple chemokine receptors (CCR9, CXCR6, CCR1, and CCR2). The frequency of the risk allele at rs657152 located on 9q34.2 (linked to ABO blood group locus) found to be associated with European patients with respiratory failure (rs657152; 95% CI, p = 4.95x10−8)(10). In addition, another study found the same locus to be associated with COVID-19-infected individuals when compared to those uninfected at lower p-value (95% CI, p = 5.3x10−20)(11). Interestingly, three loci (rs11385942, rs74956615 and rs2109069) encode inflammatory response genes (CCR2, TYK2, and DPP9) and are hypothesized to influence COVID-19 severity through hyper-inflammatory response and subsequent organ injury (table 1)(11).

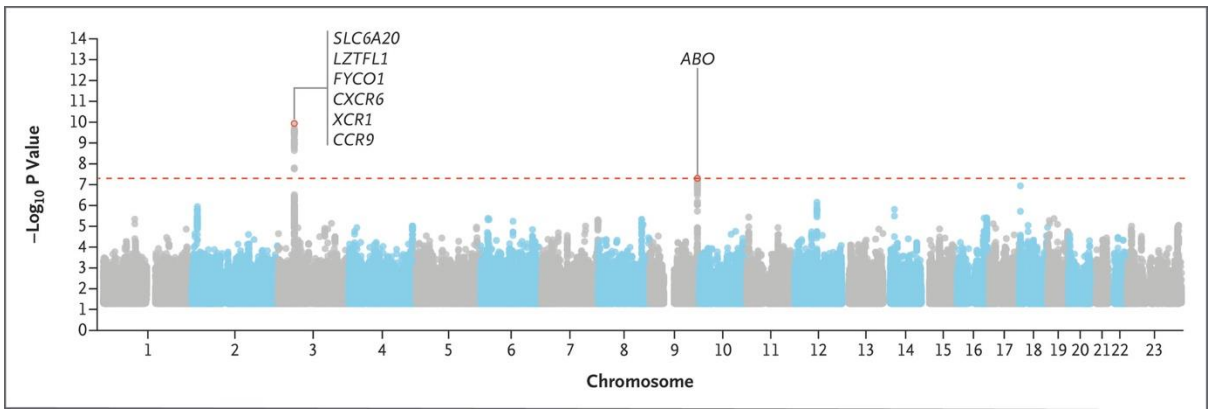


Figure 4: Genome-wide association study of severe Covid-19 with respiratory failure. (Severe Covid et al., 2020)

Table 1: Significant loci from Pairo-Castineira et al., 2021.

SNP	Chr.: pos.	Risk	Alt.	RAF_{gcc}	RAF_{ukb}	OR	CI	$P_{gcc.ukb}$	$P_{gcc.gs}$	$P_{gcc.100k}$	Locus
rs73064425	3: 45,901,089	T	C	0.15	0.07	2.1	1.88–2.45	4.8×10^{-30}	2.9×10^{-27}	3.6×10^{-32}	*LZTFL1*
rs2109069	19: 4,719,443	A	G	0.38	0.32	1.4	1.25–1.48	4×10^{-12}	4.5×10^{-7}	2.4×10^{-8}	*DPP9*
rs74956615	19: 10,427,721	A	T	0.079	0.05	1.6	1.35–1.87	2.3×10^{-8}	2.2×10^{-13}	3.9×10^{-6}	*TYK2*
rs2236757	21: 34,624,917	A	G	0.34	0.28	1.3	1.17–1.41	5×10^{-8}	8.9×10^{-5}	8.3×10^{-7}	*IFNAR2*

Genotype Imputation

Variant imputation has become a mainstay in contemporary genome-wide association studies (GWAS), as the increased exploration and testing of unobserved genotypes improves statistical power(13). Imputation uses haplotype information from a reference panel to infer genetic variation not typed, or typed inaccurately, by genotyping arrays, thereby correcting some genotyping errors and vastly enhancing genome coverage (figure 5). The performance of imputation therefore relies heavily on the specific reference panel used.

Previous studies have demonstrated strong variations in imputation performance when common reference panels were applied to different populations(14, 15). For example, imputation using HRC offered better accuracy among European populations than among the Han-Chinese population(15).

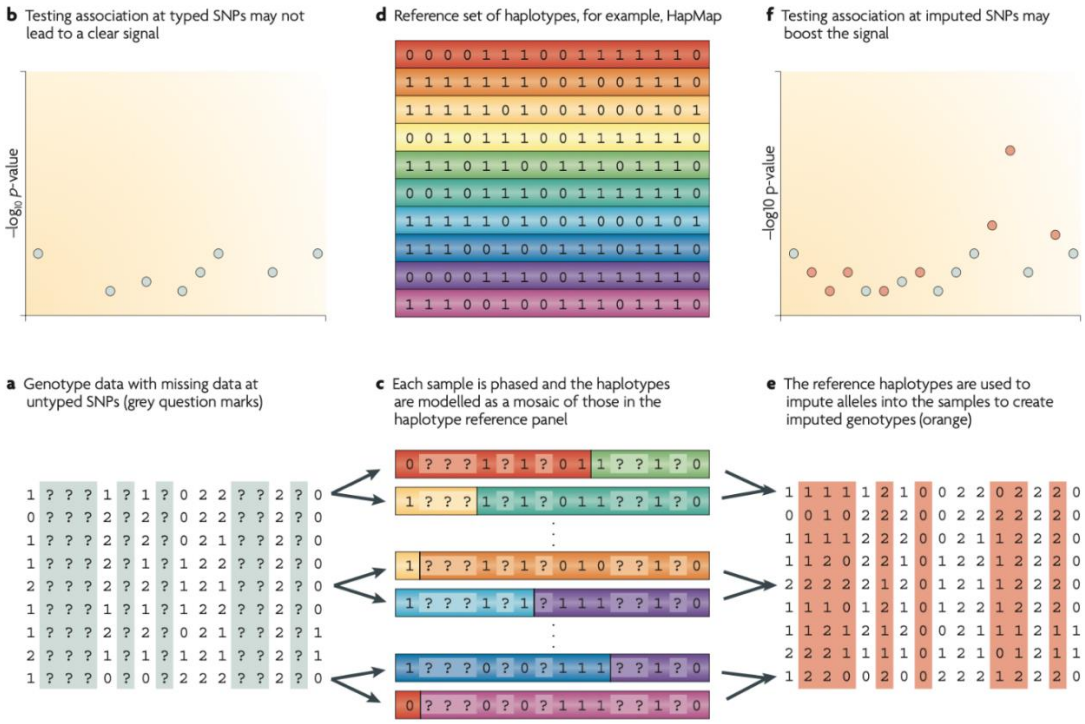


Figure 5 Process of imputation.

Many SNPs are not sequence in genotype array data and were not examined in association study(a and b). In an imputation, array data were phased (c) and match against haplotype reference data(d) to infer untyped genotype (e) allowing association to be tested in untyped genotypes(Das et al., 2018).

Whole Genome Sequencing

The advancement and cost reduction of whole genome sequencing (WGS) technology allow analysis of human genome at population level. The ability to simultaneously capture clinically significant variants in multiple genes gave WGS advantages over other technology. In addition, WGS can identify different types of variants including single point mutation, small insertion or deletion and large structural variations. Furthermore, reanalysis of the WGS data can be conduct when a novel disease associated gene is discovered. These reasons make WGS to be a very desirable method in examining population genetic variations.

Genotype-Phenotype database

Number of databases containing information on association between genotype and phenotype has been developed. These databases provide better understanding of genetic variation observed. Some databases are carefully constructed and reviewed by panel of experts, while some act as a data-sharing platform that is open for submission from wider communities.

In the field of pharmacogenomics, The Clinical Pharmacogenomics Implementation Consortium (CPIC) and the Pharmacogenomics Knowledge Base (PharmGKB) are initiatives which gathered evidence-based, peer-reviewed research and treatment recommendations of pharmacogenes(16, 17). This was done to encourage implementation of pharmacogenomics through efficient extraction and translation of genetic information into clinical action.

In detection of Autosomal recessive carrier, a Return of Results Committee had proposed a recommendation of 728 gene-condition pair for genetic testing of autosomal recessive carriers (Himes et al., 2017). These genes were reviewed by a committee comprise of experts in the field of genetics and public health includes medical geneticist, genetic counselors, molecular laboratory directors and PhD geneticists, a perinatologist, a medical ethicist, and a genetic epidemiologist. Genes were selected based on available evidence including clinical characteristics, associated mortality, and genotype-phenotype correlation(18).

Database such as ClinVar on the other hand is a data sharing database containing information on variant genotype-phenotype association submitted from various medical laboratories. ClinVar database is one of the most widely use database in examining variants' genotype-phenotype association with currently contain over 1.6 million submissions(19). However, high number of variant submissions in ClinVar has conflicting interpretation of variant pathogenicity(20). As interpretating of variant pathogenicity were made from different source and time, conflicting interpretation of

variant pathogenicity often arise due to inconsistency between each laboratory classification system, evidence available at the time of interpretation, and bias toward overestimating variant pathogenicity(21).

Interpreting pathogenic variant

The American College of Medical Genetics and the Association of Molecular Pathology proposed a standard and guideline to standardised classification of variant pathogenicity and encounter the inconsistency of laboratories classification system (22). The standard and guideline involve a scoring system that will categorise variants into 5 categories; pathogenic, likely pathogenic, uncertain significant, likely benign and benign, based on 28 criteria. These criteria are evidence supporting variant pathogenicity including the effect of variant demonstrated in a functional study, segregation analysis, population allele frequency and in silico analysis etc. The guideline is widely adopted among the clinical and molecular laboratories.

As interpreting variant required gathering large amount of data from different sources and the use of various tools, bioinformatic tool, InterVar, had been developed to facilitate variant interpretation (23). The tool involves variant annotation uses annotation tool such as ANNOVAR to classify the variant location and predict the affect variant has on the amino acid sequence, prediction variant deleterious effect uses in silico method that account for evolutionary constrain, position within the protein sequence and changes in biochemical properties and gathered information on previously reported clinical significance and functional study on the variant(24).

Variant misclassification is a known issue in data-sharing databases that could potentially lead reporting of false positive or false negative genetic result. When evaluate frequency of reported variant against expected disease prevalence, it was found that 11.5% of the pathogenic variant examined observed higher frequency when compared to the disease prevalence and up to 92.3% in variant with conflicting interpretation (25). As misclassification often arise from submitters' inconsistent classification system or limited evidence at the time of interpretation, ClinVar attempted to reduce variant misclassification through CLNREVSTAT. CLNREVSTAT is ClinVar's initiative to improve variant interpretation by leveraging information such as reported clinical significance, number of submitters and

evaluating evidence provided by submitters, such as the implementation of the ACMG guideline. By incorporating information on variant submission Shah et al. demonstrated reduction in disease risk inflation which suggest reduction in variant misclassification.

Bioinformatic tools

The "star" nomenclature system

The "star" nomenclature is a system to describe allelic variation and haplotypes of pharmacogenes. It is commonly use in treatment guidelines as it can provide accurate phenotype prediction. Application of "star" nomenclature system involved identification of star alleles, diplotypes and sometime complex structural variation (SV). Assigning star alleles could involve identification of multiple variants on the same haplotype.

Accurately assign alleles has been a challenge as tests designed by different laboratory examined different combination of variants. These differences can result in incorrect allele assignments, hence phenotype prediction (figure 6). For example in Figure 11, if a *CYP2D6* assay were designed to only detect variation at two points, c.2850C>T and c.4180G>C, the assay would not be able to distinguish *2 from *17, *21 or *2XN with duplication. This could later effect the predicted phenotype as *2 and *17 extensive metabolizer, *21 is an intermediate metabolizer and *2XN is an ultrarapid metabolizer. This create disparities in star allele reported for the same sample and further discourage the adaptation of PGx testing (26).

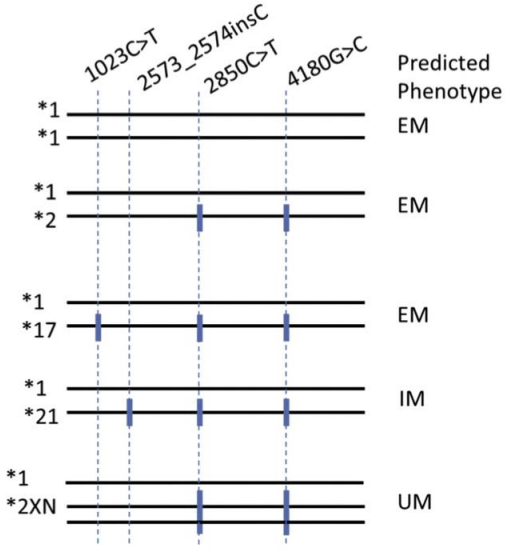


Figure 6 A CYP2D6 assay design. If assay was designed to only detects c.2850C>T and c.4180G>C could miss other star allele with different predicted phenotypes. EM, extensive metabolizer; IM, intermediate metabolizer; UM, ultrarapid metabolizer.

Stargazer

Whole genome sequencing has advantages over other platforms as it identifies all variants required for accurate allele assignment and novel clinically relevant PGx variants, which may account for unexplained differences in drug response. As alleles assignment required identifying large number of variants and detection of SV, bioinformatics tool was developed in facilitate calling of star alleles from next-generation sequencing data(27).

Stargazer perform multiple steps in identification of star alleles and diplotypes (figure 7 left). First, Stargazer identify all variants required for calling of star alleles. Secondly, phasing is performed on genotype data, uses 1000 genome project phased genotype as a reference. The phased genotype data enable identification of variants on the same strain, hence, determine the sample diplotypes. If structural variations are known to effect the phenotype, read depth will be examine in order to call structural variants (figure 7 right).

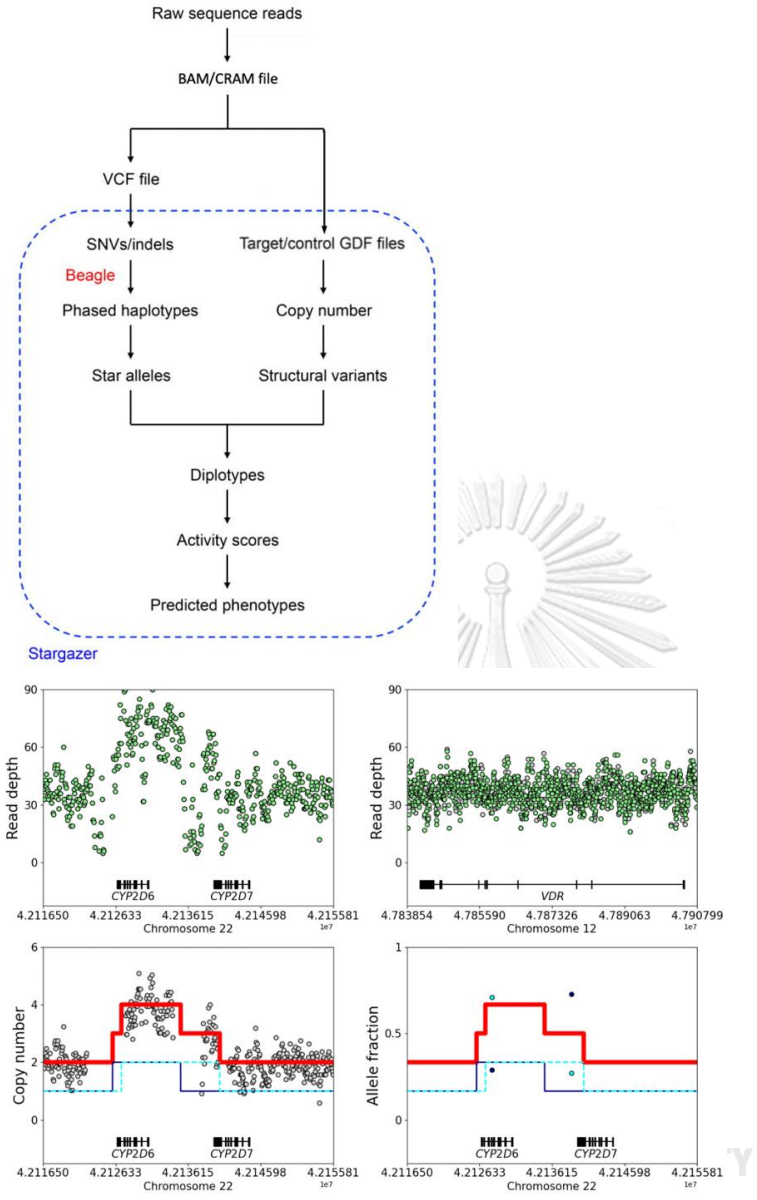


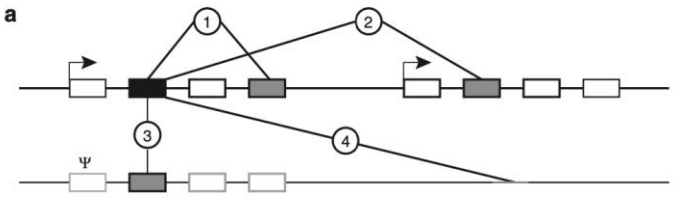
Figure 7 Stargazer workflow

Workflow in calling diplotypes and predict phenotype (left). Illustration of calling complex structural variant using Stargazer (right). Read depth in the gene of interest were examined and standardized uses a control gene.

High sequence homologies regions

Some technical difficulties may arise when using WGS short read technology to identify pathogenic variants located in genomic regions with extensive sequence homologies(28). For example, sequence may inaccurately map to a to the non-functional pseudogene with high sequence homology resulting in reporting of false-positive or false negative result (Figure 8)(29).

Sequences produced by short read WGS are generally 150 bp. In genomic regions with repeated sequences or high sequences homology, short sequences read may have difficulties finding a unique match on the reference genome. This in turn result in variant within extensive sequence homologies regions having lower coverage or mis-mapped sequences. The mis-mapped sequence often led to variant calling that are low in confidence and produced low mapping quality score. These variants with low quality score might be excluded from further analysis during the quality control process. If not carefully assessed, this could lead reporting of false positive or negative results. *SMN1* and *HBA2* genes associated Spinal Muscular Atrophy (SMA) and a-thalassemia, respectively, are two of the commonly screen autosomal recessive disorder genes located in extensive sequence homologies region.



	Homology type	%Match	Query region	Query coordinates	Target region	Target coordinates	psiDR pseudogene
①	Same gene	99.52%	TTN exon 194	chr2:179518479-179518689	TTN exon 176	chr2:179527000-179527210	–
②	Different gene	98.33%	MYH6 exon 33	chr14:23853586-23853991	MYH7 exon 34	chr14:23884199-23884502	–
③	Pseudogene	99.59%	STRC exon 17	chr15:43899996-43900238	–	chr15:43999470-43999712	ENST00000509801.1
④	Non-CDS	99.55%	OTOA exon 21	chr16:21742093-21742316	–	chr16:22558219-22558442	–

Figure 8 Diagram showing sequence read potentially inaccurately map to different part of the genome.
(29)

Spinal Muscular Atrophy (SMA)

Carrier frequency of SMA has been reported to be around 1 in 40 to 80 individuals depending on ancestral group(30, 31). An examination of SMA carrier rate used quantitative PCR-based and MLPA in Thailand reported to be 1.78% when examined(32). In most cases, SMA cause by homozygous deletion of SMN1 gene that lead to loss of alpha motor neurons and result in presentation of muscle atrophy or severe muscle weakness in SMA patients (2, 33, 34).

Identification of SMA carrier include detection of *SMN1* gene copy number. Due to ancestral gene duplication, *SMN1* has a paralogous gene, *SMN2,* that has high sequences similarity and are almost indistinguishable from one another (2, 35). However, one major difference between these two genes is a variant NM_000344.3: c.840C>T found only on *SMN2* gene. The c.840C>T variant disrupt *SMN2* gene splice enhancer and lead to skipping of exon 7. The absence in exon 7 in majority (~90%) of *SMN2* protein causes *SMN2* protein to be unstable and not fully function (Figure 9).

When sequenced with WGS short read technology, the high sequences similarity between 2 genes makes sequences within this region difficult to accurately mapped and make it difficult to detect SMN1 gene copy number. Previous study used short read next-generation sequencing technology as a carrier testing would require additional laboratory work(36). In recent years, supplementary informatic tools targeting the region had demonstrated to improve the identification of sequences and structural variants(37). The SMNCopyNumberCaller target ~30 kb region that cover SMN1 and SMN2 gene. SMNCopyNumberCaller differentiate SMN2 from SMN1 gene by account for 16 bases unique to SMN2, including the variant c.840C>T and its surrounding intronic variants(37).

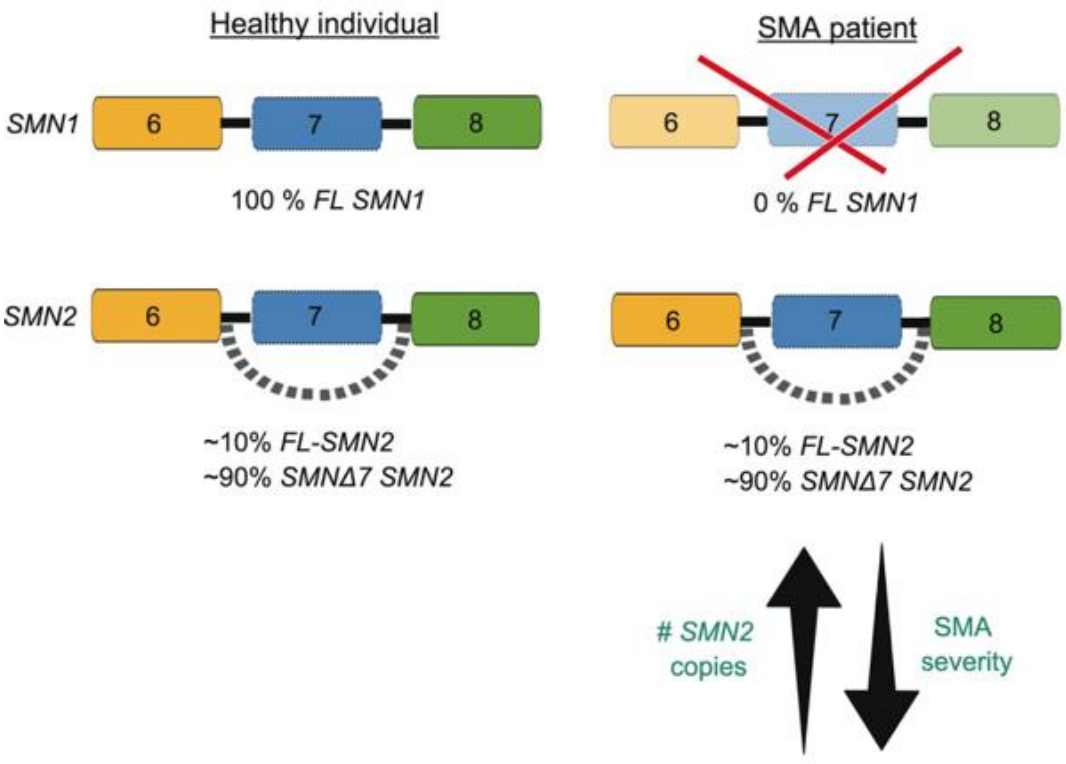


Figure 9 Contribution of SMN1 and SMN2 gene to SMA.
SMA patient loss full length (FL) SMN1 gene while majority (~90%) of SMN2 gene produce a not fully function protein due to the loss of exon 7. SMN2 gene produce some (~10%) functional protein, while not sufficient for survival, it correlates with disease severity (2).

Alpha-Thalassaemia in Thailand

Alpha thalassaemia is a disorder cause by deflect in haemoglobin production due to genetic variation that resulted in an absence or dysfunction in at least one of the four copies of the alpha globin genes(38). Alpha globin genes cluster is located on chromosome 16 (16p13.3). It contains three functional globin genes HBZ, HBA1 and HBA2, the embryonic haemoglobin gene and two foetal/adult haemoglobin gene(38). Over 121 disease causing alpha-globin variants have been identified in HbVar(http://globin.bx.psu.edu). These variants can be separated into three types:

(i) deletions that resulted in the loss of both a-globin genes in cis (a^0-thalassaemia) including –-SEA and -–THAI (Figure 10)

(ii) deletions that resulted in the loss of one of the a-globin gene (a^+-thalassaemia) this include the commonly found 3.7 and 4.2 kb deletion (-$a^{3.7}$ and -$a^{4.2}$) (Figure 10)

(iii) non-deletional, such as point mutations or small insertion/deletion (indels) that interrupt the gene function. For instance, Hb Constant Spring or Hb Pakse that disrupt the stop codon and causes elongation of the alpha globin chain.

(iv)

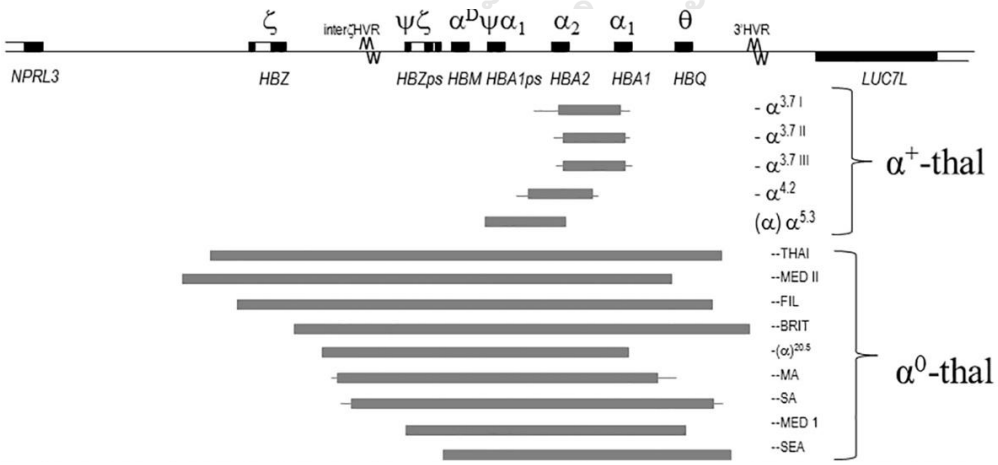


Figure 10 Diagram show common alpha-globin deletion.

Grey bar represent length of the deletion and it relative position on the genome(39).

Currently wide range of techniques are available for haemoglobin variant detection(40). The choice of diagnostic tool required the knowledge of population variant spectrum as variant endemic in each population can be different. Without prior knowledge of population variant frequency, technique uses must be able to detect any point mutation or large deletion in the alpha globin genes. Current gold standard involves the use of sanger sequencing in detecting point mutation and multiples ligation probe for detection of large deletions(40). However, performing both techniques could be labour intensive and require specialized equipment. WGS have benefit over other molecular techniques as it has potential to detect both point mutation and large structural variation simultaneously.

Alpha globin gene clusters is a gene-dense genomic region that is GC-rich and high Alu-repeat. The high homologous sequence within this region causes whole genome short-read sequences to ambiguously mapped to multiple position within the region. This led to the reduction of number of variants confidently called and the ability to detect point mutation. Furthermore, it effects the ability to detect structural variation as this required accurate read depth estimation.

NGS4Thal is a bioinformatics analysis pipeline that designed to detect pathogenic thalassemia variants from next generation sequencing data (Figure 11)(41). By specifically target the alpha-globin cluster and realign poorly mapped sequences, NGS4Thal had demonstrated to improve detection of alpha thalassemia variants. NGS4Thal identify reads with multiple alignment used bwa-based mapping quality score. NGS4Thal kept reads with high mapping quality score, remove read with mapping quality score equal to zero as it likely to map with other position outside of the region and realign read with low mapping score and has less than three base pair mismatches. Using this strategy, NGS4Thal demonstrated to improve the sensitivity of detecting pathogenic variants. The realigned bam files were then use as a template to detect structural variation. Because different structural variation detection tools are specialized at detecting different type of structural variants, NGS4Thal complementarily uses 3 different structural variation callers, including BreakDancer(42), Pindel (43) and CoNIFER (44), to improve detection of diverse type of structural variants.

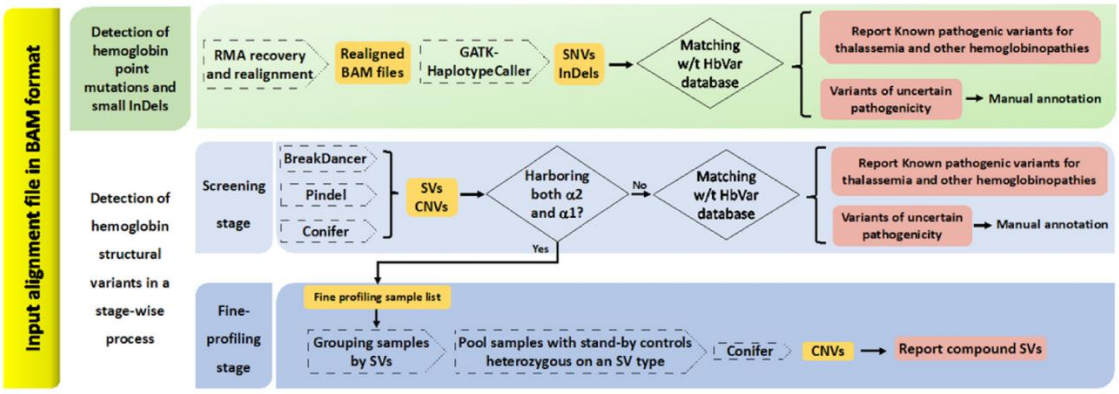


Figure 11 Workflow of NGS4Thal.

NGS4Thal involves realign read with multiple alignment (RMA) and identify single nucleotide variant (SNV) and small insertion and deletion (InDel) under GATK pipeline. NGS4Thal also identify structural variant (SV) and copy number variation (CNV) using multiple SV callers.

Public reference panel

A wide range of public reference panels exists with varying sizes, sequencing coverages, and represented populations(13). These public reference panels include the 1000 Genomes Project phase 3 (1000G), the Haplotype Reference Consortium (HRC), the GenomeAsia 100K project (GenomeAsia), and the Trans-Omics for Precision Medicine (TOPMed) program.

1000G comprises 2,504 ancestrally diverse individuals from 26 global populations (45, 46). HRC covers 32,488 human genomes by combining WGS data from over 20 different studies including 1000G. WGS data from HRC have sequencing coverage of 4x to 8x and are predominantly of European descent (47). GenomeAsia was constructed to address the underrepresentation of Asian populations in the preceding reference panels. GenomeAsia contains WGS data on 1,739 individuals from over 219 populations across Asia, with high depth coverage (~36x)(48). In their most recent release, TOPMed contains WGS of 97,256 individuals publicly available for imputation. TOPMed's WGS data are high-depth coverage (~38x) including individuals from diverse ancestral backgrounds (49).

Population Structure

Studying population carrier frequencies based on self-reported population labels or ethnicity had demonstrated to be unreliable (50, 51). For example, when compared self-reported ancestry written in the requisition form with self-reported ancestry during consultation, the study found that there are inconsistencies between the two sources (50). These inconsistences depend on ancestral group with only 30.3% of individual who self-identified as having Mediterranean ancestry show concordance result between the two sources. Moreover, inconsistency was also found between self-reported ethnicity and genetic ancestry examined used genotype data. Up to 27.5% who of study population self-reported as Southeast Asian has genetic ancestry that are closer to South Asian ancestry rather than East Asian as expected (50). These discordances could arise from multiple reasons such as uncertainty in family origin or self-identification with a particular group due to personal or cultural reason. Furthermore, another study had shown that when examined genetic population structure used PCA method, the first 10 principal component shows number of clusters did not overlap with the reference panel (Figure 12) (51). This suggest that the population structure within a population can be complex and currently available population labels may not provide full description of all subpopulations.

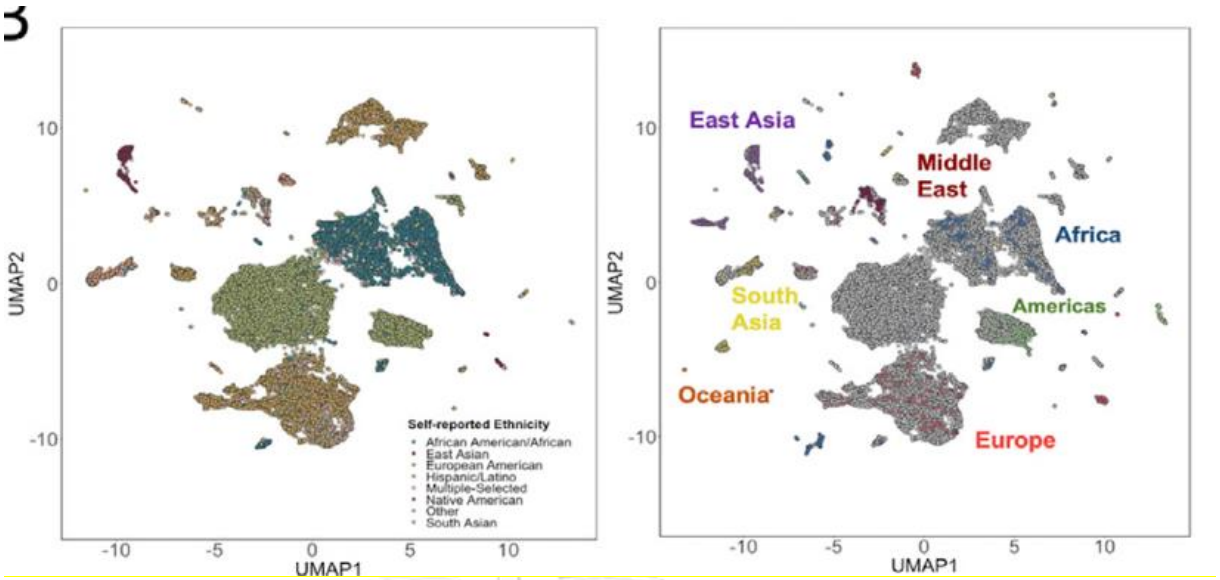


Figure 12 UMAP projection of the first 10 principal components form BioMe participants.

(left) samples were coloured according to self-reported ethnicity. (right) samples form BioMe participants were coloured in gray and reference samples from 87 global populations coloured by their continental region of origin(51).

Human genetic variation at a population level can provide insight into human evolutionary, migration and historical events. One of the widely use method in uncovering population structure is Principal Component Analysis, which uses dimensionality reduction method(52). PCA create a matrix quantifying genetic similarity between each pair of individuals within the cohort and observe grouping of individuals that are genetically close with each other through clustering form after visualisation of principle component. Because PCA projection identifies directions of maximal variance in the data and ignores variation in other directions, finer-scale patterns within population were often obscure and the subtle genomic structure were missed.

Lawson et al. proposed fineSTRUCTURE, a method which took advantages of variants relative position within the genome instead of analysing each variant individually (53). Through haplotype phasing authors were able to exploit linkage disequilibrium pattern. These linkage disequilibrium patterns were then use in an identification of shared haplotype or genomic segments that reflect individual identical descent. Multiple studies demonstrated that when using fineSTRUCTURE to examine genetic population structure, shared haplotype method was able to reveal structure at a much finer resolution when compared to a single-marker PCA method (54, 55) . Shared haplotype method was able to uncover subpopulations that sometime can be differentiate down to provinces (54). The identification of haplotype shared between individual captured shared identity that reflect a much more recent past when compare SNP sharing and enable identification of structure that are more recent and subtle. When applied to 2039 samples from the People of the British Isles collection, fineSTRUCTURE was able to differentiate population up to 53 clusters that correspond with the country geography (55). Clusters identified by fineSTRUCTURE was indistinguishable when uses PCA or admixture (Figure13).

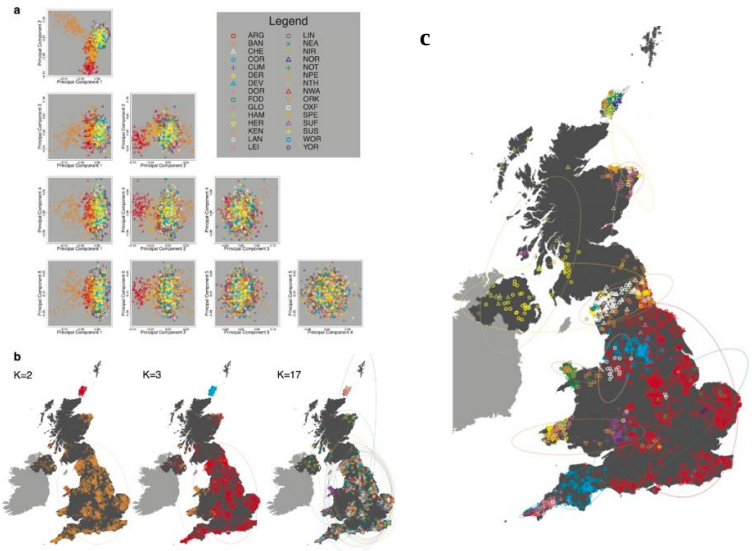


Figure 13 Population structure analysis of UK samples.

a) UK population structure was examined used Principle component analysis. The plot shows a pair of first 5 principle component. b) UK population structure was examined used program ADMIXTURE. The map shows when value of K in ADMIXTURE is set at 2, 3, and 17. Each dot represent individual within the cohort. Dot was plotted according to their grandparent's birthplace and were coloured according to cluster assigned by ADMIXTURE. C) UK population structure was examined used program fineSTRUCTURE(55).

Haplotype Sharing use Whole genome sequences

The haplotype sharing method (ChromoPainter/fineSTRUCTURE) had illustrated to identified fine-scale genetic substructure from genome-wide single nucleotide polymorphism array data in multiple studies (54, 56-58). Lawson et al. showed that performance of ChromoPainter/fineSTRUCTURE improved when applied to genotype data with a more densely packed markers as these markers provided LD pattern at a higher resolution.

WGS deliver a more complete picture of the genomic sequence when compared to genotype array. The more complete genomic sequences could have potential to identify a more accurate size of shared haplotype or identify variants that are private to that population subgroup, which could be missed when used a pre-designed genotype array. This can produce a more accurate clustering and improve resolution of the population from countries to regions within countries.

The high-density WGSs however are exceptionally large. Computational cost of running ChromoPainter depends on the number of individuals within the cohort and the number of SNPs. As ChromoPainter were designed based on genotype array data, running on the high density WGS can be very computational extensive.

Positional Burrows-Wheeler transform (PBWT) is a data compression algorithm that were designed to store haplotypes data (59). PBWT is an extension of the Burrows-Wheeler Transform (BWT), the widely use algorithm for matching read and sequence assembly. PWBT compress haplotypes data and allow efficient search and matching of haplotypes. The efficiency of PBWT reduces processing time and enable work on a much larger data set. A recent study demonstrated that using PBWT-paint, a scalable haplotype sharing algorithm based on the positional Burrows-Wheeler transform, was able to capture genetic structure similar to ChromoPainter (Figure 14)(54). PBWT-paint would allow detection of shared haplotypes in high-density WGS data.

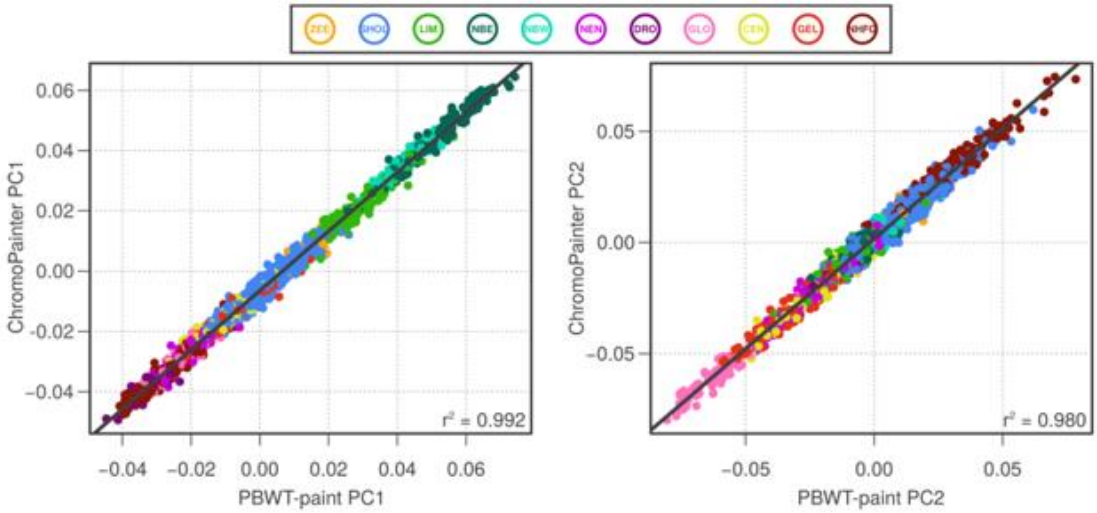


Figure 14 Evaluating ChromoPainter against PBWT-paint.

Principal components (PC) obtained from using ChromoPainter were evaluated against PC obtained from using PBWT-paint.

Part I: Genetic variation in pharmacogenomics

Part I.I: Phenotype prediction of pharmacogenes in Thais from whole genome sequencing

The “star” nomenclature system, commonly use in treatment guidelines, involved identification of alleles, diplotypes and complex structural variation (SV) for accurate phenotype prediction. Whole genome sequencing (WGS) has advantages over other platforms as it identifies all variants required for accurate allele assignment and novel clinically relevant PGx variants, which may account for unexplained differences in drug response. As alleles assignment required identifying large number of variants and detection of SV, bioinformatics tool was developed in facilitate calling of star alleles from next-generation sequencing data4.

Research Questions:

What is the prevalence of star alleles, diplotypes and predicted phenotype of high evidence pharmacogenes in Thai population?

Research Objectives:

To use Stargazer assign star allele and diplotype, which involve identifying multiple variants on the same haplotypes and calling complex structural variation, of 25 high evidence pharmacogenes for accurate phenotype prediction,

To determine prevalence of star alleles in Thai population and predict phenotype of these pharmacogenes.

Expected benefits and application:

The study will demonstrate the utilization of WGS in Pharmacogenomics testing, including accurate phenotype prediction using the “star” nomenclature system. Variations of pharmacogenes in Thai population will facilitate Pharmacogenomics-guided clinical decision making in Thailand for further application of precision public health including dosing guidelines, drug development, clinical trials, and development of population-specific screening.

Methods

All variants within the region specified in Stargazer (version 1.0.8) for 25 pharmacogenes, including CACNA1S, CFTR, CYP2B6, CYP2C8, CYP2C9, CYP2C19, CYP2D6, CYP3A4, CYP3A5, CYP4F2, DPYD, G6PD, GSTM1, GSTP1, IFNL3, NAT1, NAT2, NUDT15, RYR1, SLCO1B1, TPMT, UGT1A1, UGT1A4, UGT2B15, and VKORC1, will be extracted from genome Variant Call Format (gVCF) files using BCFtools (version 1.10.2). Variants will be excluded if they were with locus GQX < 30, with site genotype conflicted with proximal indel call, with locus in the region with conflicting indel calls, and with unbalanced phasing pattern. VCF files of all samples will be merged to generate a single VCF file. Non-variants will be excluded from the final VCF file.

Multidimensional scaling analysis was performed on single nucleotide polymorphism, excluding indels, within 25 pharmacogenes using Plink (version 1.9). Multidimensional scaling plot will be examine if there are separation between cases and control.

Star allele analysis

Stargazer require a VCF file on genome coordinate GRCh37 and a gdf file for SV detection. Genome coordinates, reference, and alternative allele on the VCF file will be converted from GRCh38 to GRCh37 using LiftoverVariants tools available in GATK package (version 4.1.6.0) and VCF file will be use as an input for Stargazer.

To generate the gdf file for SV detection of CYP2D6, first, Bazam (version 1.0.1) will be used to extract CYP2D6-CYP2D7 region from BAM file and realigned on GRCh37 coordinates. Samtools (version 1.9) will be used to extract read depth. Sdf2gdf script, available on Stargazer, was used to generate the gdf files. The haplotype, activity score, diplotype, and predicted phenotype called by Stargazer with VDR as a control gene. Results were combined and visualized using R program (version 3.6.3, dplyr and ggplot2 package).

Result

Star allele analysis

Over 25% of Thais carried high-risk diplotypes in 5 pharmacogenes including CYP3A5, CYP2C19, NAT2, SLCO1B1, and UGT1A1 (Figure 15). CYP3A5*3, loss-of-function allele, was found in heterozygous intermediate metabolizing (IM) diplotype, CYP3A5*1/*3 (48.5%), and homozygous poor metabolizing (PM) diplotypes, CYP3A5*3/*3 (35.1%). CYP2C19 loss-of-function *2 and *3 alleles contributed to the prevalence of IM diplotype, CYP2C19*1/*2 (36%) and *1/*3 (3%), and PM diplotype, CYP2C19*2/*2 (10%) and *3/*3 (1%). CYP2C19 gain-of-function *17 allele was found in rapid metabolizing diplotype, CYP2C19*1/*17 (2.41%). NAT2 slow acetylator *5, *6, and *7 alleles were found in IM diplotypes, NAT2*6/*7 (5.5%), *6/*6 (5.2%), and *5/*6 (3.1%). SLCO1B1*1B/*17, *1B/*15, and *1/*17, which are the most prevalent diplotypes that carried decreased function *5, *15, and *17 alleles, were observed at 3.95%, 3.26%, and 1.72%, respectively. UGT1A1*60/*60, *6/*60, and *28/*60 were among the most prevalent diplotypes at 10.3%, 6.29%, and 5.14%, respectively.

On the other hand, high-risk diplotype frequencies were < 3% in 10 pharmacogenes, which were DYPD, CYP2C8, CACNA1S, RYR1, CFTR, NUDT15, CYP2C9, GTSM1, G6DP, and TPMT (Figure 15). Additionally, the functional effect of over 25% of detected alleles in GSTP1, NAT1, UGT2B15, and VKORC1 was currently unknown (Figure 15).

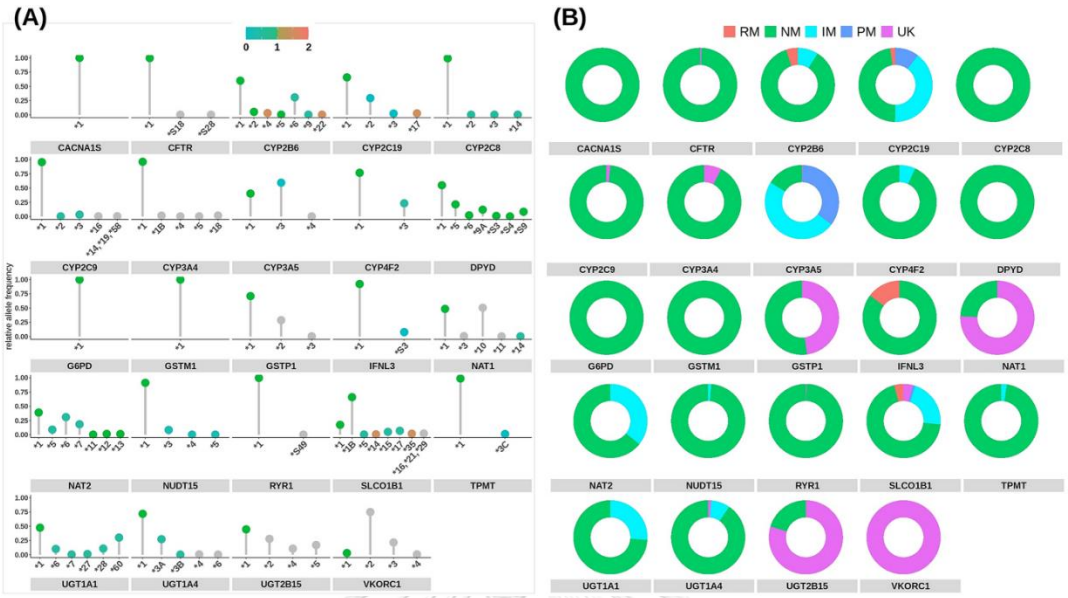


Figure 15 Allele frequencies of star alleles relative to alleles found within this study cohort and predicted phenotypes of 24 CPIC evidence level A pharmacogenes called using Stargazer (version 1.0.8).

(A) Colors on dots represent activity score ranging from blue (no function) to green (normal function) to red (increased function) and grey (unknown function). (B) Predicted phenotypes are presented as rapid metabolizer (RM) or unfavorable response for IFNL3, normal metabolizer (NM) or favorable response for IFNL3, intermediate metabolizer (IM), poor metabolizer (PM), and unknown function (UK).

Twenty different star alleles of CYP2D6 were observed. Among these, 5 were duplication (CYP2D6*1 × 2, CYP2D6*2 × 2, CYP2D6*10 × 2, CYP2D6*34 × 2, CYP2D6*71 × 2), 1 was deletion (CYP2D6*5), and 6 were rearrangement (CYP2D6*S1 + *1, *4N + *4, *36 + *10, *36 × 3 + *10, *68 + *4, *83 + *2), which accounted for 1.9%, 4.5%, and 34.7% of star alleles found, respectively. CYP2D6*36 + *10 and *10 alleles were the most prevalent of CYP2D6 decreased function alleles in this cohort. CYP2D6*1/*36 + *10, *36 + *10/*36 + *10, *10/*36 + *10, and *1/*10 were among the highest diplotypes found at 14.5%, 12.1%, 11.4%, and 9.31%, respectively (Figure 16).

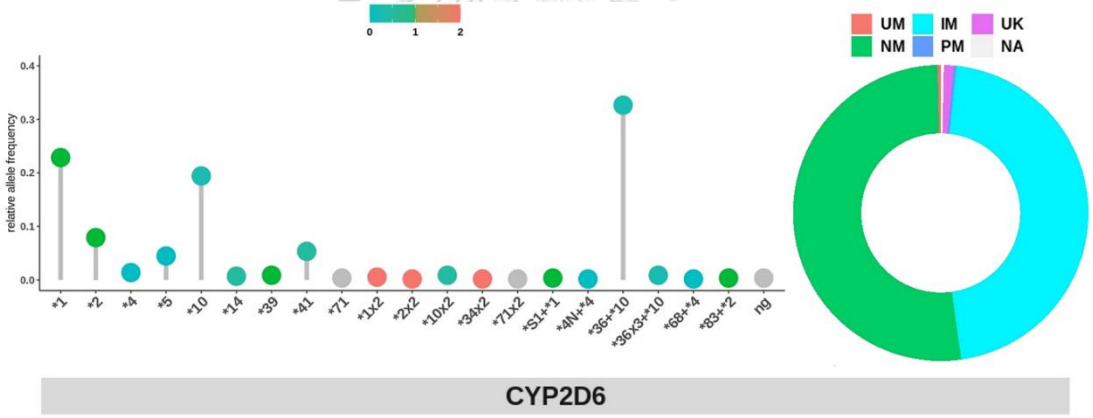


Figure 16 Allele frequencies of star alleles with structural variation relative to CYP2D6 alleles found within this study cohort and predicted phenotypes called using Stargazer (version 1.0.8).

(A) Colors on dots in star alleles plot represent activity score range from blue (no function) to green (normal function) to red (increased function) and grey (unknown function). Uncalled sample is denoted as ng. (B) Predicted phenotypes are presented as ultra-rapid metabolizer (UM), normal metabolizer (NM), intermediate metabolizer (IM), poor metabolizer (PM), unknown function (UK), and not applicable (NA).

Star alleles of frequencies of CYP2D6 called through Stargazer algorithm, were within the range of the previously published East-Asian allele frequencies (Table 2)(60).

Table 2 Distribution of CYP2D6 star alleles in Thais and East-Asian population.

Alleles	**This study**	**Suwannasri *et al.*, 2011[32] (n = 288)**	**Chamnanphon *et al.*, 2013[33] (n = 57)**	**Gaedigk *et al.*, 2017[21] East-Asian (n = 14,816)**	
				Average	Range
***1**	22.93	22.91	35	35.24	17.5-93.79
***2**	7.93	9.7	9.6	13.11	7.65-42.71
***4**	1.38	0.7	0.9	0.59	0-4.35
***5**	4.48	4.3	4.4	5.17	0-9.6
***10**	19.48	44.6	45.6	42.58	8.6-64.1
***14**	0.69	1.04	0.9	0.77	0-3
***36**	-	16.4	0.9	1.52	0-16.4
***39**	0.86	-	-	0.24	0-1.18
***41**	5.34	-	1.8	2.18	0-6.54
***71**	0.34	-	-	0.52	0-1.5
***1x2**	0.52	-	0	0.27	0-0.51
***2x2**	0.17	-	0	0.38	0-0.99
***10x2**	0.86	-	0	0.4	0-1
***71x2**	0.17	-	0	0.03	0-0.2
Other duplication	0.17	0.35	-	1.39	0-6
***36+*10**	32.76	-	-	26.41	22.45-32.65
***36x3+*10**	0.86	-	-	1.02	1.02-1.02
Other rearrangements	1.03	-	-	5.51	5.51-5.51

Discussion

This study report the prevalence of star alleles, diplotypes, and phenotype predictions of 25 clinically relevant pharmacogenes, including CYP2D6 SV, from WGS in the Thai population. The “star” nomenclature system used in this study is a powerful tool for predicting activity or function of enzymes, transporters, or drug targets, as it accounts for a combination effect of multiple variants within an allele(61). We found high clinical relevance cytochrome P450 genes (CYP3A5, CYP2C19, and CYP2D6) exhibiting high variation in predicted phenotype. This could reflect the low evolutionary constraint within these enzymes, as they lack essential endogenous function(62). SV, between CYP2D6 and its pseudogenes (CYP2D7, CYP2D8) established to alter enzymatic activity, found to exerted high importance in the Thai population17. It accounted for 60% of CYP2D6 star alleles detected and 83.8% of all high-risk diplotypes in this study. Interestingly, prevalence of CYP2D6 SV was also previously reported to be highest among Asians when compared to African Americans, Caucasians, and Hispanics17. Our finding emphasizes the importance of detecting CYP2D6 SV for accurate phenotype prediction especially in Thai population.

Part I.II: Phenotype prediction and characterization of pharmacogenes in Thais from whole genome sequencing

Current PGx resources and recommendations are based largely on a population of European descent. Studies have shown differences in pharmacogenes between ethnicities or even in closely related populations (4, 5, 63, 64). As race or ethnicity are often use in guideline for genetic screening recommendation (65). The genetic differences between Thai and other East Asian population in many of pharmacogenes remain uncertain. Furthermore, as Thai population are often underrepresented in genomic studies, there could be pharmacogenetic variants that are population specific to Thai (66).

Research Questions:

What is the allele frequency of well-studied PGx variants in Thai population and are prevalence of these variant different from East Asians and other population?

Are there potential novel deleterious pharmacogenomic variants in the Thai population.

Research Objectives:

To identify known pharmacogenomics variants and examine allele frequencies found in Thai population.

To compare allele frequencies found in Thais with other global population.

To identify potential novel deleterious variants in the Thai population

Expected benefits and application:

The study will determine similarities or differences in allele frequencies of pharmacogenomics variants between Thai and East Asian. This knowledge will help determine if following guideline recommended for East Asian would be suitable.

Methods

Analysis of variants within pharmacogenes

The VCF file will be annotated with gnomAD allele frequencies of the global population using Ensemble Variant Effect Predictor (version 98.3). Annotated variants will be classified into common (MAF ≥ 0.05), low frequency (0.05 > MAF ≥ 0.01), rare (MAF < 0.01), or absent (MAF = 0). Variants within each group will be counted using VCFtools (version 0.1.15).
Number of missense variants per coding sequence was calculated by:

Number of missense variants/Ensembl transcript length

, where Ensembl transcript length will be obtained from BioMart database (https://www.Ensembl.org/biomart/martview/) and for transcript with APPRIS annotation value as “primary assembly”.

Analysis of known pharmacogenomic variants

PGx variants will be retrieved from PharmGKB database (https://www.pharmgkb.org, accessed on 06/06/2020). Variants with evidence level 1A, 1B, and 2A will be identified as known PGx variants.

Allele frequencies of these variant will be compared with those of the population in gnomAD using Chi-square test or Fisher’s exact test. A p-value of < 0.001 was used as a significant level after Bonferroni correction.

Identification of potential novel deleterious variants

Variants will be extracted to examine novel potentially deleterious variants. Variants reported in PharmGKB database or used in star allele analysis will be excluded.

CADD PHRED-normalized scores will be downloaded online. CADD PHRED-normalized scores ≥ 20, or 1% most deleterious single nucleotide variants within the reference genome, will be considered potentially deleterious variants.

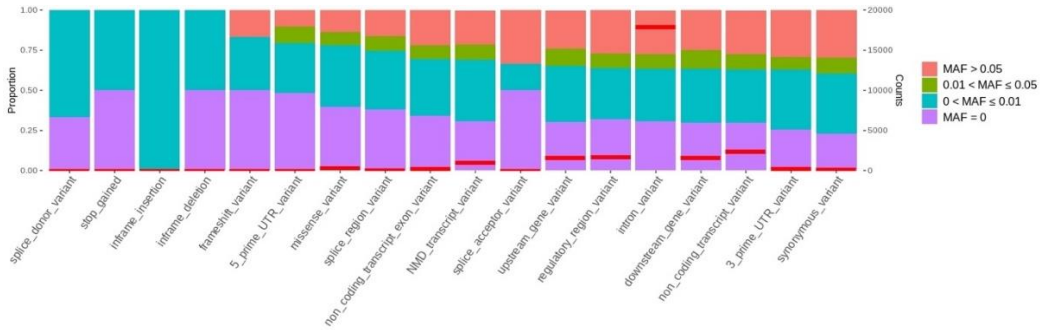
Loss-of-function variants include stop-gained, splice-site disrupting, frameshift insertion, and frameshift deletion variants. LOFTEE algorithm available in VEP-plugin will be used to determine loss-of-function variants, and variants annotated as “high confidence” were considered loss-of-function in this study.

Result and discussion

Variant analysis of 25 pharmacogenes

A total of 18,825 variants were detected within 25 pharmacogenes of 291 individuals. Of 18,825 variants, 12,026 (63.8%) were rare, while 5766 (30.6%) variants were absent from the gnomAD database. An enrichment of rare variants was found within variants that impact protein function. For example, all of in-frame insertion, deletion and stop gained variants found were rare on gnomAD database in compare to 60.5% of synonymous variants and 63.5% of intron variants found were rare (figure 17). IFNL3, UGT1A4, and CYP2D6 reported the highest number of missense variants per coding sequence, while GSTM1, where null genotype link to development of cancers4, was the most conserved (figure 17 B).

(A)



(B)

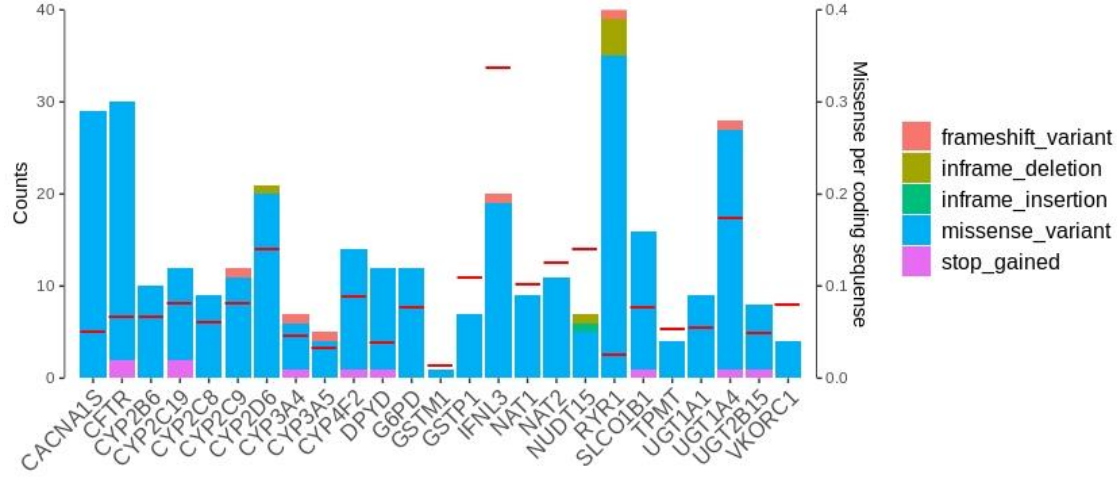


Figure 17. Distribution of variants found within 25 pharmacogenes.

(A) Proportion of variants grouped by allele frequency relative to gnomAD database and number of variants found within each type of variant show in red dash (-). (B) Counts of variant that impact protein function within each gene and missense variant per coding sequence per gene show in red dash (-).

Known PGx variants

The prevalence of 39 high-evidence PGx variants found in Thais compared to East-Asian and global population in gnomAD database were shown in Figure 18. Of these, 19 high-evidence PGx variants were commonly found in Thais, with allele frequency of over 0.1. Fifteen variants were associated with increased risk of toxicity or adverse drug reactions are underlined in Fig. 18 (67). Six variants were associated with increased risk of toxicity were commonly found in Thais, including rs1041983 (NAT2), rs1799930 (NAT2), rs4244285 (CYP2C19), rs1695 (GSTP1), rs4149056 (SLCO1B1), and rs11045879 (SLCO1B1). Fifty-one percent of Thais were carriers of T allele in rs1041983 (NAT2 c.282C>T), which is associated with increased risk of liver toxicity upon treatment of anti-tuberculosis drugs (68, 69). Among the highest evidence level variants (1A), 49% of Thais carried A allele in rs4244285 (CYP2C19c.681G>A), which is associated with an increased risk for secondary cardiovascular events upon clopidogrel usage, and 24% of Thais carried C allele in rs4149056 (SLCO1B1 c.521T>C), which is associated with an increased risk of simvastatin-induced myopathy (70, 71). In comparison to other populations, 26 and 10 variants were significantly different from the global and East-Asian population, respectively (Figure 18). The rs776746 (CYP3A5), rs1041983 (NAT2), and rs2279343 (CYP2B6) were more frequent in Thais than both populations. Multiple variants within VKORC1 in this cohort exhibited a significant degree of deviation from both populations.

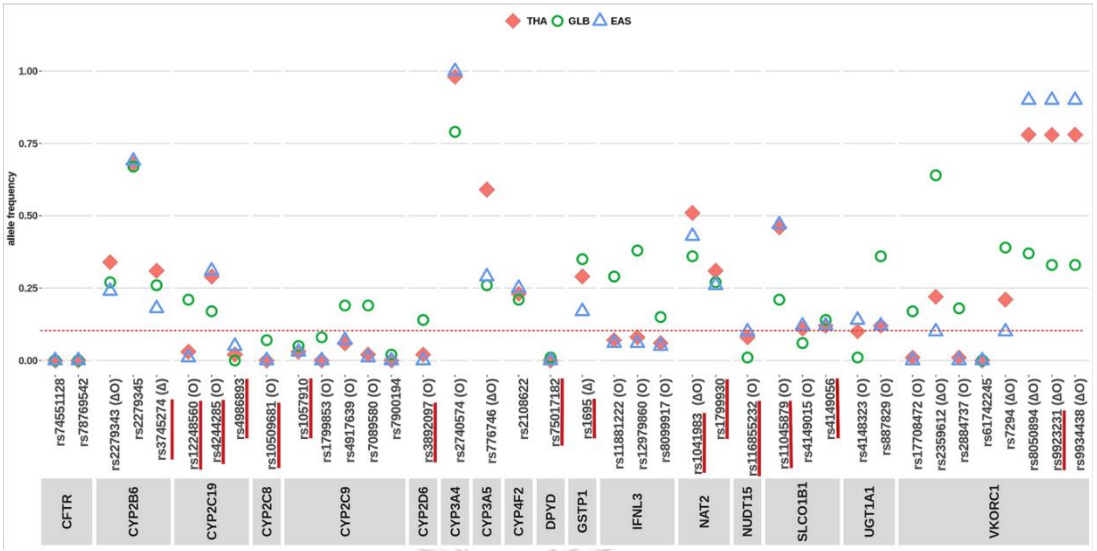


Figure 18: Allele frequencies of 39 high-evidence PGx variants in Thai (THA) compared to East-Asian (EAS) and global population (GLB) in gnomAD database. Variants associated with toxicity are underlined in red. Variants with significant p-value ($p < 0.001$) when comparing Thai allele frequency with gnomAD database, are denoted as (O), and when comparing Thai allele frequency with East-Asian population in gnomAD database, are denoted as (Δ).

Variability in drug response among ethnicities had long been observed, but a recent increase in the number of populations studied unveiled another layer of genetic variability within the sub-population, such as distribution gradient of CYP2C19*17 found from Western to Eastern Europe(72). In Thais, CYP3A5*3 (rs776746), CYP2B6*6 (rs2279343), and NAT2 (rs1041983) were significantly higher compared with East-Asian and global populations. Varying allele frequency of multiple VKORC1 variants to different populations found in this study supported the previously reported variation of rs9923231 among the East-Asian population where allele frequency of 0.96, 0.94, and 0.90 was observed in North-East Asians (Japanese, South Koreans, and Chinese) in comparison to 0.62, 0.69, 0.75 observed in South-East Asians (Filipinos, Malaysians, and Indonesians) (73).

Potentially deleterious PGx variants

Of 305 missense variants found in this cohort, 41 variants were previously reported to associate with drug response in PharmGKB database. Novel potentially deleterious missense variants found in Thais were reported in the Table 3.

One hundred and ten missense variants were considered novel, potentially deleterious, while 5 variants obtained Combined Annotation Dependent Depletion (CADD) PHRED-normalized scores of > 30 (Table 3). Seventy-eight percent (n = 86) of novel potentially deleterious missense variants were only found once in this cohort. Ninety-four percent (n = 103) were rare in gnomAD database, while 61% (n = 67) were absent. Thirty percent (n = 33) had not been reported in dbSNP 150 database. Sixty-two percent of Thais carry up to 4 novel potentially deleterious missense variants.

Table 3 Novel potentially deleterious pharmacogenomics variants.

Ref/Alt: Reference/Alternative nucleotide; **Ref_AA:** Reference Amino Acid; **Alt_AA:** Alternative Amino Acid; **CADD score:** Combined Annotation Dependent Depletion PHRED-normalized scores.

Position (GRCh38)	dbSNP150	Ref/Alt	Allele Count	Gene	Ref_AA	Alt_AA	CADD score
chr1:201062468	.	T/C	1	CACNA1S	E	G	33
chr7:117614633	rs1005269197	G/A	1	CFTR	G	S	32
chr1:201078026	rs780841536	T/C	1	CACNA1S	Y	C	31
chr19:38452983	.	T/G	1	RYR1	L	R	31
chr1:201048591	rs745558537	T/G	1	CACNA1S	K	Q	30
chr19:38516212	rs878984852	G/A	1	RYR1	R	Q	29.8
chr22:42127527	rs769351604	G/A	1	CYP2D6	R	C	29.2
chr7:117592595	rs377447726	A/G	1	CFTR	R	G	28.7
chr1:201083251	.	T/G	2	CACNA1S	Y	S	28.5
chr1:201078005	rs150590855	C/A	1	CACNA1S	R	L	28.5
chr1:201052633	rs555016254	C/T	1	CACNA1S	A	T	28
chr1:201061410	.	G/A	1	CACNA1S	R	C	27.4
chr1:97305348	rs570122671	G/A	1	DPYD	T	I	27.3
chr6:18133821	rs777803269	T/G	1	TPMT	D	A	27.2
chr7:117504290	rs1800073	C/T	1	CFTR	R	C	27.1
chr7:117504306	.	A/C	2	CFTR	D	A	26.9
chr1:201047168	rs3850625	G/A	19	CACNA1S	R	C	26.8
chr1:201060666	rs145039828	C/T	1	CACNA1S	G	S	26.6
chr19:38536011	.	A/G	1	RYR1	N	S	26.6
chr7:117587821	.	T/C	1	CFTR	I	T	26.6
chr1:201065924	rs571902899	C/T	1	CACNA1S	V	M	26.5
chr1:201077922	rs557195329	C/T	1	CACNA1S	V	M	26.4
chr16:31094573	rs781304132	G/T	1	VKORC1	R	S	26.2
chr11:67584499	rs755557033	C/G	1	GSTP1	Q	E	26
chr19:15879844	rs372871763	C/T	1	CYP4F2	R	Q	26
chr19:38444648	.	T/C	1	RYR1	M	T	26
chr19:38512443	.	C/G	1	RYR1	F	L	25.9
chr2:233772309	rs114982090	C/T	5	UGT1A8	P	L	25.9
chr1:201043401	.	A/G	1	CACNA1S	F	S	25.7
chr1:201083173	rs143202536	G/T	1	CACNA1S	T	N	25.7
chr12:21224811	rs377350683	T/C	1	SLCO1B1	C	R	25.7
chr19:38502914	.	C/G	1	RYR1	R	G	25.7
chr19:39243685	rs77379751	G/A	31	IFNL3	R	C	25.6
chr19:38519384	rs201339536	G/A	2	RYR1	E	K	25.6
chr1:201089374	rs186538122	G/A	1	CACNA1S	R	W	25.5
chr19:38519282	rs775895899	G/A	1	RYR1	G	R	25.5
chr19:38499811	rs575780192	C/T	1	RYR1	R	W	25.4
chr19:15892373	rs754089074	G/A	2	CYP4F2	A	V	25.3
chr1:201070353	.	G/A	1	CACNA1S	P	L	25
chr1:201040054	rs12139527	A/G	68	CACNA1S	L	S	24.9
chr13:48041009	rs773719265	C/A	1	NUDT15	S	Y	24.9
chrX:154535348	rs886044847	A/G	1	G6PD	F	S	24.8
chr6:18147901	rs752440908	T/C	1	TPMT	H	R	24.6
chr7:117540282	rs1800086	C/G	1	CFTR	T	S	24.6
chr12:21178618	.	T/A	1	SLCO1B1	F	Y	24.5
chr19:38565511	.	G/A	2	RYR1	G	S	24.5
chr1:201047143	.	C/T	1	CACNA1S	R	Q	24.4
chr1:201110216	rs12406479	G/C	1	CACNA1S	A	G	24.4

chr19:15886018	rs145174239	G/C	1	CYP4F2	L	V	24.3
chr1:201076930	rs142356235	C/T	1	CACNA1S	S	N	24.2
chr19:38502628	rs754579512	T/G	1	RYR1	V	G	24.2
chr19:38448375	rs368711923	G/A	1	RYR1	R	H	24.1
chr8:18222050	.	G/A	1	NAT1	M	I	24.1
chr19:38505076	rs566495420	G/A	3	RYR1	D	N	24
chr7:117592588	rs1800103	A/G	1	CFTR	I	M	24
chr8:18222649	rs768813958	A/T	3	NAT1	D	V	24
chr11:67584472	rs774305853	G/A	1	GSTP1	A	T	23.8
chr7:117535318	rs121909046	A/G	2	CFTR	E	G	23.8
chr8:18400392	.	A/C	1	NAT2	Q	P	23.8
chr19:41006980	.	G/T	1	CYP2B6	R	L	23.7
chr19:41012471	rs201500445	T/C	3	CYP2B6	Y	H	23.7
chr19:38565443	.	G/A	1	RYR1	G	D	23.7
chr4:68663024	.	G/T	1	UGT2B15	A	D	23.7
chr7:117530977	.	T/C	1	CFTR	S	P	23.7
chr19:39243850	rs139076671	G/A	1	IFNL3	H	Y	23.6
chr8:18222271	.	T/C	1	NAT1	L	P	23.6
chr7:117531043	rs145900055	C/T	1	CFTR	P	S	23.5
chr1:201089385	rs35534614	C/T	1	CACNA1S	G	D	23.4
chr19:15878779	rs3093200	G/T	3	CYP4F2	L	M	23.4
chr19:38469044	rs780626994	C/T	1	RYR1	L	F	23.4
chr4:68668066	rs192628779	A/G	5	UGT2B15	C	R	23.4
chr1:201051079	.	G/A	1	CACNA1S	P	S	23.3
chr19:39244019	rs149832972	G/A	1	IFNL3	L	F	23.3
chr19:38504293	.	C/T	1	RYR1	T	I	23.3
chr4:68654253	rs187815441	T/C	1	UGT2B15	H	R	23.3
chr7:117592287	.	C/G	1	CFTR	S	C	23.3
chr7:117627561	.	C/T	2	CFTR	P	S	23.3
chr10:94781959	rs764137538	C/T	1	CYP2C19	R	W	23.2
chr7:117559577	.	T/G	1	CFTR	I	M	23.2
chr19:39244114	rs145428712	G/A	1	IFNL3	T	M	23.1
chr19:38570667	.	A/G	1	RYR1	I	V	23.1
chr19:38485972	rs192863857	C/T	4	RYR1	P	S	23.1
chr10:94775447	rs150152656	C/T	1	CYP2C19	T	M	22.9
chr2:233772416	rs371183955	C/T	4	UGT1A9	H	Y	22.9
chr10:94947843	.	T/G	1	CYP2C9	I	M	22.8
chr2:233718944	rs553189135	C/A	1	UGT1A4	L	I	22.8
chr4:68670516	rs529876617	G/T	1	UGT2B15	H	N	22.8
chr8:18400653	rs568110818	T/A	1	NAT2	F	Y	22.8
chr1:201083231	rs572977674	C/T	1	CACNA1S	V	I	22.7
chr11:67586206	rs4986949	G/T	3	GSTP1	D	Y	22.6
chr1:97193101	rs766833304	G/C	1	DPYD	P	A	22.3
chr19:38492540	rs35364374	G/T	10	RYR1	G	C	22.3
chr19:41004380	rs535039125	C/T	1	CYP2B6	R	W	22.2
chr19:38485976	rs199837883	C/T	2	RYR1	P	L	22.2
chr1:201110258	rs549107212	G/A	1	CACNA1S	T	M	22
chr12:21200625	rs752196141	T/C	1	SLCO1B1	V	A	22
chr13:48041096	.	T/C	1	NUDT15	V	A	22
chr19:38578027	rs373919284	C/T	1	RYR1	P	L	22
chr19:38527689	rs538497899	C/T	3	RYR1	R	W	22
chr1:201089392	rs190152688	T/C	2	CACNA1S	I	V	21.8
chr19:15892398	rs556151888	G/A	1	CYP4F2	R	C	21.8
chr8:18222637	rs1044890902	G/A	1	NAT1	R	Q	21.8
chr19:38527707	rs55876273	G/C	3	RYR1	E	Q	21.5
chr10:95064936	rs750028311	A/G	1	CYP2C8	I	T	21.4
chrX:154532206	.	A/G	1	G6PD	I	T	21.1
chr11:67584478	rs12796085	C/G	1	GSTP1	L	V	21
chr19:38565544	.	G/C	1	RYR1	D	H	20.8

chr7:117594979	rs562851847	A/G	1	CFTR	N	S	20.5
chr7:99660591	.	T/C	1	CYP3A5	S	G	20.5
chr8:18400082	rs765487420	A/C	1	NAT2	I	L	20.4

Eleven high-confidence loss-of-function variants were found in 9 pharmacogenes (Table 4), 8 variants were only found once in this cohort, 2 variants were rare (minor allele frequency [MAF] < 0.01), and 1 variant was found at low frequency (0.01 < MAF < 0.05). According to the gnomAD database, all loss-of-function variants were rare and 6 variants were absent. An enrichment of splice acceptor variant rs373134805 (CYP3A5) was found within the South East-Asian population in GenomeAsia 100 k database(73).

Table 4 Loss of function pharmacogenomics variants.

Ref/Alt: Reference/Alternative nucleotide;**MAF:** Minor Allele Frequency.

Position (GRCh38)	**Ref/Alt**	**dbSNP150**	**GENE**	**Annotation**	**MAF in Thai**	**MAF in gnomAD**		**MAF in 100k GenomeAsia**	
						Global	EAS	NEA	SEA
chr7:99666690	C/G	rs373134805	CYP3A5	splice_acceptor_variant	0.017	3.18E-05	0	0	0.022
chr10:94842889	C/A	rs370320936	CYP2C19	stop_gained	5.15E-03	0	0	0	1.45E-03
chr12:21224840	G/A	rs200994482	SLCO1B1	splice_donor_variant	3.45E-03	1.60E-04	3.22E-03	0	1.45E-03
chr7:117611708	G/A		CFTR	stop_gained	1.75E-03	0	0	0	0
chr7:99666950	A/G	rs55965422	CYP3A5	splice_donor_variant	1.72E-03	4.46E-04	8.99E-03	5.70E-03	1.45E-03
chr10:94941978	AG/A		CYP2C9	frameshift_variant	1.72E-03	0	0	0	0
chr1:97828127	G/A	rs189768576	DPYD	stop_gained	1.72E-03	3.19E-05	6.41E-04	1.42E-03	0
chr7:117559463	G/A	rs397508200	CFTR	splice_acceptor_variant	1.72E-03	0	0	0	0
chr7:117592292	C/T	rs121908760	CFTR	stop_gained	1.72E-03	0	0	0	0
chr19:15897501	C/T	rs752022409	CYP4F2	stop_gained	1.72E-03	3.19E-05	6.42E-04	0	0
chr19:39243908	C/T	rs546666114	IFNL3	splice_acceptor_variant	1.72E-03	0	0	0	0

We identified 110 novel potentially deleterious missense variants and 11 high-confidence loss-of-function variants circulating within the population. Novel potentially deleterious variants were population specific with 94.2% identified were rare in gnomAD database, and 60.3% were absent. This reflect previous finding that high impact variants are often rare and geographically localized as a result of purifying selection (74). For example, potentially deleterious splice acceptor variant c.433-1G>C in CYP3A5 found at a low frequency in Thai (0.017) is population-specific South East-Asian populations including Vietnamese (0.018), Malaysian (0.039), and Indonesian (0.015), while extremely rare in the gnomAD database (73). These variations within subpopulations of East-Asians demonstrate the benefit of PGx testing and highlight the precaution that must be taken when associating PGx with ethnicity labels.

A focus on rare variants in explaining inter-individual variation in drug response is likely to increase as the cost of sequencing is reduced, making WGS more readily available. An important challenge remains in interpreting these rare variants of unknown significant. Repository SPHINX (Sequence, Phenotype, and pHarmacogenomics INtegration eXchange https://emergesphinx.org), that link PGx variants of unknown significance with patients clinical phenotypes would facilitate researchers on studying these variants of unknown significant for future PGx discovery (75).

Limitations

We acknowledge several limitations and ways the study could be improved. A portion of enrolled participants was Brugada syndrome patients. Although none of the genes associated with Brugada syndrome were examined, results could be enhanced with unknown genetic factors influencing the disease. Previous study reported an inconsistent in star alleles calling in samples with complex SV when three bioinformatics tools were compared, this suggest that further confirmation, such as using high-resolution long-read sequencing that allows accurate variant calling and phasing, might be required in some samples with CYP2D6 complex SV (76) . Computational prediction tools like Loss-of-Function Transcript Effect Estimator (LOFTEE) and Combinded Annotation Dependent Depletion (CADD) used in this study and other studies are useful in prioritizing deleterious effects in variants of unknown significance; however, variants must be reported with caution and validated through a functional study before implementation in clinical settings (77, 78).

Summary

In conclusion, we reported a comprehensive overview of the PGx spectrum in a Thai population and its differences with East-Asian populations. We demonstrated the utilization of WGS in PGx testing, including accurate phenotype prediction using the "star" nomenclature system, SV detection, and identification of known and unknown potentially deleterious PGx variants.

The WGS ability to access PGx variants and SV in a single methodology reduced time and labor involved. This study demonstrates WGS to be a highly efficient platform in research and PGx testing. The current high cost and bioinfomatics required to process and translate large data could limit WGS application as a PGx testing platform in routine clinical setting. Development of bioinformatics tools use in translating genotype data are moving toward a more automated manner, such as under developing PharmCAT (79). This would make interpreting WGS data more user-friendly and accessible to wider healthcare provider in the near future. In the meantime, an alternative more cost effective platform such as genotyping arrays could currently be a more applicable (80).

The reported findings and variations within pharmacogenes of the Thai population facilitate PGx-guided clinical decision making in Thailand and contribute to the database of the understudied South-East Asian population for further application of precision public health including dosing guidelines, drug development, clinical trials, and development of population-specific screening.

PART II: Genetic variation in autosomal recessive variants

PART II.I: Identification of point mutation and structural variants in *SMN1* and *HBA2* gene located in high sequence homogenous region.

While whole genome sequencing (WGS) technology can simultaneously capture wide range of clinically significant AR variants, difficulties arise when WGS short read technology were used to identify pathogenic variants located in genomic regions with extensive sequence homologies(28). The extensive sequence homologies cause short read sequences within this region to ambiguously mappings. The poor mapping of sequence resulted in variant calling with low confidence.

SMN1 and *HBA2* are two of Autosomal Recessive genes commonly screen in genetic testing for carrier of Spinal Muscular Atrophy and a-thalassemia, respectively, due to it high incidence rate and disease severity. *SMN1* and *HBA2* are both located in genomic regions with extensive sequence homologies. *SMN1* gene has high sequences similarity to *SMN2* gene making the two genes indistinguishable. *HBA2* gene is located in Alpha globin gene clusters (16p13.3) with high homologous sequences and interspersed repeats. For these reasons detecting carrier uses WGS were not possible for *SMN1* and *HBA2*. However, in recent year reanalysis of WGS data used targeted informatics tool had demonstrated to improve mapping quality and increase variants detection within these regions (37, 41).

Research Questions:

What is the prevalence of *SMN1* and *HBA2* of Spinal Muscular Atrophy and a-thalassemia carrier in Thai population?

Research Objectives:

To use bioinformatic tool in identifying point mutation and structural variants from WGS data in gene associated with Spinal muscular atrophy (*SMN1* gene*)* and alpha thalassemia *(HBA2* gene*)*.

Expected benefits and application:

This study will demonstrate the use multiple bioinformatic tools in facilitating calling variants from WGS data that are unable to confidently call using the standard calling pipeline. This study will further demonstrate the benefit in using WGS in examining population carrier frequency and as a diagnostic tool for carrier testing in the future.

Methods

Study population

WGSs from the Brugada cohort (Clinical Trial Registration Number NCT04232787) will be use in this study. The cohort consist of two groups, patients diagnosed with Brugada syndrome and controls. Controls are volunteers from blood donors at multiple sites of the National Blood Centre, Thai Red Cross Society or visitors for health check-ups and workers at King Chulalongkorn Memorial Hospital. Individual within the control group had no type I Brugada pattern or family history of sudden cardiac arrest. All subjects were of Thai ethnic origin by self-report from 5 major geographical regions: north, northeast, central, east, and south.

Sample size estimation and power of detection

The sample size was estimated used the following equation:

$$n = \frac{Z\alpha^2 P (1-P)}{e^2}$$

n = required a sample size

$Z\alpha$ = standard Z value (e.g. 1.96 for confidence level at 95%, two-tail)

P = Incidence proportion

e = acceptable margin of error at 5% (standard value of 0.05) or confident interval

According to sample size calculation, a sample size of 497 individual would achieve the statistically result in detecting variant at prevalence 3% within the population with marginal error does not exceed than 1.5% with 95% confidence level.

Ethical considerations

Informed consent was obtained from all participants. All methods were performed in accordance with relevant guidelines/regulations. The study was approved by the

Institutional Review Board of the Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand (IRB No. 431/58). Informed consent was obtained from all participants. All methods were performed in accordance with relevant guidelines/regulations.

SMN1 Structural variants analysis

SMN structural variants were analysed using SMNCopyNumberCaller(37). BAM files were provided as input. The detection of full-length SMN1 copy number, full-length SMN2 copy number, deletion of SMN2 Exon7-8 and single nucleotide variant NM_000344.3: c.*3+80T>G were done following the SMNCopyNumberCaller's manual instructions.

HBA2 variants analysis

NGS4THAL pipeline was used to detect pathogenic point mutation, small insertion/deletion, and structural thalassemia variants. As databases in the NGS4THAL pipeline were constructed on genome coordinate GRCh37, Bazam (version 1.0.1) was used to extract haemoglobin regions from BAM files and realigned on GRCh37 coordinates(81). Bam files on GRCh37 genome coordinates were used as inputs into the NGS4THAL pipeline following the manual instructions. NGS4THAL realigned ambiguously mapped NGS sequences, and variant callings were under the GATK framework GATK-HaplotypeCaller version 3.8 to detect pathogenic point mutation and small insertion/deletion. used Complementary structural variant caller BreakDancer version 1.4.5, Pindel version 0.2.5 and CoNIFER version 0.2.2 were used to detect structural thalassemia variants.

Result

Copy number variation of SMA

SMA carriers were identified by calling copy numbers of the SMN1 gene using targeted informatic tools, SMNCopyNumberCaller. 10 (VCR=0.017) individuals carrying one copy of the SMN1 gene were identified as SMA carriers. The copy number of SMN2 that contribute to stable FL-SMN protein and modulate disease severity were then analyzed (33, 82). SMA carriers show variation in SMN2 gene copy numbers from 3(n=2), 2(n=4) to 1(n=3) copy number. One SMA carrier does not carry the SMN2 gene.

Throughout the cohort, 2:2 was found to be the most common SMN1 to SMN2 copy number ratio (50.7%) follow by 2:1 (35.5%) (figure 19). 1 person has partial exon 7 and 8 deletions at the SMN2 gene. The silent carrier was not detected within the cohort, while c.*3+80 T>G that collates with two copies of SMN1 on the same haplotype was detected in 4 samples.

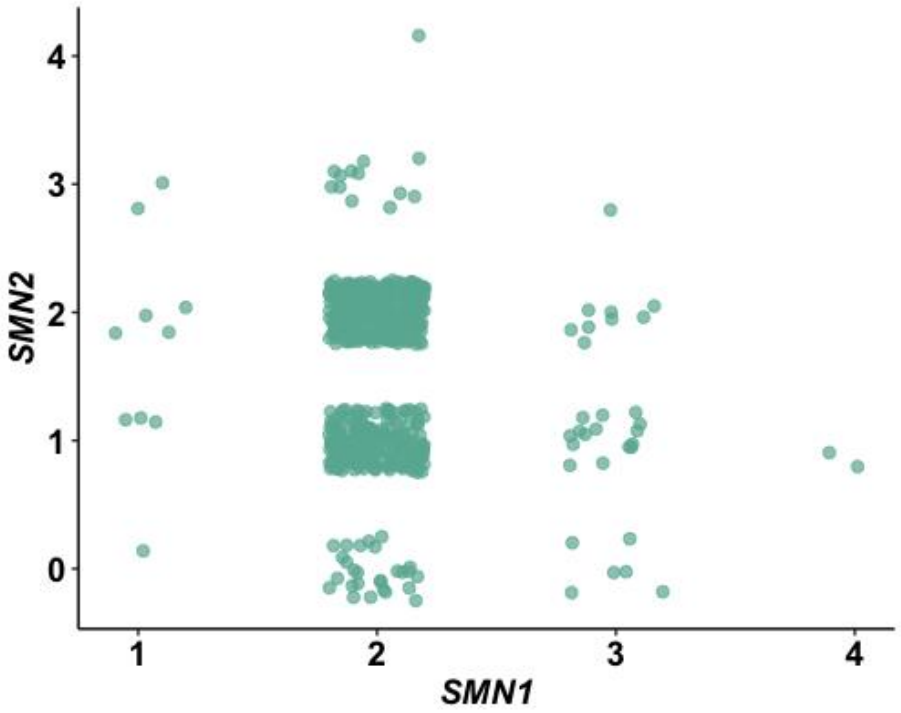


Figure 19 Samples SMN1 gene copy number against SMN2 gene copy number.

Alpha-thalassemia

All three forms of alpha-thalassemia variants were identified from WGS data: deletion in both copies of a-globin (a^0), deletion in one a-globin copy (a^+) and a non-deletional a-globin variant (a^{ND}). 20 individuals (VCR=0.033) are a carrier of $-^{SEA}$ deletion, a ~20 kb a^0-thalassemia deletion (Table 5). 13 individuals (VCR=0.021) are a carrier of $-a^{3.7}$, a 3.7 kb (type I) a^+- thalassemia deletion. For a^{ND}-Thalassemia, Hb CS and Hb Paksé, are found in 43 (VCR=0.057) and 3 (VCR= 0.005) individuals, respectively.

Table 5 Sequence and structural variants in HBA2 gene detected using informatics tools.

Genes	HbVar_Name	Variants	VCR
HBA2	- -(SEA)		0.033
HBA2	3.7 kb (type I) deletion alpha-2		0.021
HBA2	Hb_Constant_Spring_(Hb_CS)	c.427T>C	0.057
HBA2	Hb_Paks&eacute	c.429A>T	0.005

Discussion

1.67% of the cohort identified as SMA carriers used supplementary informatic tool and all three forms of alpha-thalassemia variants (a0, a+ and aND) were identified used NGS4Thal pipeline. Supplementary informatic tools improve the identification of sequences and structural variants in difficult to reach high homology genomic regions that were previously overlooked or required supplementary laboratory work(36). The SMNCopyNumberCaller account for c.840C>T and surrounding intronic variants that are unique to *SMN2* to differentiate its *SMN1* gene(37). 1.67% of the Thai cohort were identified as SMA carriers by SMNCopyNumberCaller. SMA carrier rate is in concordance with previously reported prevalence in Thailand that used quantitative PCR-based and MLPA(32). NGS4Thal realign poorly mapped sequences to identify pathogenic variants in the *HBA2* gene and uses a combination of SV caller to identify partial or whole gene deletion(41). The NGS4Thal realignment of alpha globin gene cluster enables calling structural variant and improve the poor-quality call at the Hb CS position from 3.14% (n=19) of the cohort to 1.98% (n=12). Incorporating specialized bioinformatics for calling structural would increase the economical mean of adapting WGS technology for carrier genetics testing in the future.

PART II.II: Determine carrier rates of autosomal recessive disorder in Thai population.

Carrier genetic testing aims to detect pathogenic variants with the potential to cause autosomal-recessive (AR) disorders. This allows the identification of individuals at risk of having a child with the tested conditions. The testing enables practitioners to provide genetic counselling on reproductive risks and options that aid couples in their family planning.

The landscape of AR variants can be highly population-specific (1). Within European populations, less than 20% of carrier variants were shared between the Dutch and Estonian cohort (9). The knowledge of population carrier frequencies could improve the choice of screening disorders.

Research Questions:
What is the prevalence of autosomal recessive variants circulating in Thai population?

Research Objectives:
To identify variants associated with autosomal recessive disorder circulating in Thai population.
To determine carrier rates of these autosomal recessive variants and if any variants are found at high prevalent.

Expected benefits and application:
The comprehensive overview of population carrier rates of autosomal recessive gene will be a useful resource for the development of carrier testing recommendations and estimation of disease burden.
Demonstration of using WGS in examining population carrier frequency.

Method

Whole genome sequences

Sequencings of paired-end 150 bp fragment read from polymerase chain reaction (PCR)-free sequencing libraries were performed on the HiSeqX (Illumina Ltd, Cambridge, UK). Sequencing, alignment, and variant calling were performed at Illumina Ltd, Cambridge, UK. Reads were aligned to NCBI GRCh38 human reference genome assembly.

Variants quality controls (QC) were performed as previously described15.

Variants were excluded if they were with locus GQX < 30,

with site genotype conflicted with proximal indel call,

with locus in the region with conflicting indel calls

with an unbalanced phasing pattern.

Only variants with GQ > 20 and DP > 10 were included in the analysis. Variants that did not pass QC were set as missing and variants that exceeded 5% missingness in the cohort were excluded from the analysis.

Variants within 672 genes associated with 728 AR disorders previously curated by the NextGen Return of Results Committee (RORC) were extracted and used in this study(18).

Cases and Control

Differences between case and control within AR genes were investigated. Multidimensional scaling analysis was performed in Plink (version 1.9). 174,887 variants within AR genes with a minor allele frequency of higher than 0.01 were selected. The multidimensional scaling plot was done for the first 4 principle components to illustrate no separation between cases and controls. Case and control were then collectively analysed.

Likely Pathogenic/Pathogenic Variant analysis

Variant annotations, including allele frequencies from population database gnomAD version 3.0, were performed using Annovar(24). Clinically relevant likely pathogenic, pathogenic variants and variants with conflicting interpretations of pathogenicity on ClinVar database version 2021-03-08 were extracted for this study analysis. Further variant interpretations were performed on variants with conflicting interpretations of pathogenicity using InterVar, a bioinformatic tool that automatically classified variants based on ACMG-AMP guideline(23).

Variants were then separated into 4 groups (P1, P2, P3 and CoP) based on CLNREVSTAT annotation in ClinVar. P1 group contains likely pathogenic/pathogenic variants that were either reviewed by an expert panel or have multiple submitters with assertion criteria provided. P2 is a superset of P1 that included likely pathogenic/pathogenic variants with only one submitter with assertion criteria provided. P3 included likely pathogenic/pathogenic variants submitted without assertion criteria. CoP included variants with conflicting interpretations of pathogenicity with likely pathogenic/pathogenic variants submissions. Variants that contain a likely benign/benign submission, have minor allele frequencies of more the 0.03 within the cohort or were interpreted as likely benign/begin in InterVar annotation were excluded to reduce the chances of reporting false-positive results.

Carrier rates

Variant carrier rates (VCR) were calculated for each variant according to a previous study (1):

$$\square\square\square = \frac{\square\square - \square\square\square}{0.5 * \square\square},$$

where AC is the total allele count, Hom is the number of homozygous individuals and AN is the total number of alleles.

The collective VCR were then used to calculate the gene carrier rate (GCR) for each gene where:

$$\square\square\square = 1 - \prod_{\square=1}^{\square}(1 - \square\square\square_{\square}),$$

where VCRi is the carrier rate for variant i and v is the number of variants detected for each gene.

Result

Carrier frequency

In the analysis of 672 genes associated with 731 autosomal recessive disorders, we identified 263 likely pathogenic/pathogenic variants in 605 Thai individuals. 198 (75.3%) variants in this group were found in singleton(n=1), where the variant was only detected once throughout the cohort. 60.4% of variants detected in Thais were absent from the East Asian reference population in the gnomAD database and 23.8% of the variants identified were absent from the gnomnAD database.

Likely pathogenic/pathogenic variants were grouped into P1, P2, P3 and CoP according to their level of evidence for pathogenicity, where P1 has the highest level of evidence. 100 variants were identified as P1 (Supplementary table 1). 58.2% of the cohort are a carrier for at least one P1 variant with up to 4 variants identified per person. When accounting for variants with lower evidence for pathogenicity, the number of variants identified increased to 180 (P2), 208 (P3) and 263 (CoP). The percentage of individuals in the cohort carrying at least one variant increased to 64.5% (P2), 68.0% (P3) and 76.7% (CoP). The maximum number of variants detected per person increased to 5 (P2) and 6 (P3 and CoP).

Variant carrier rates (VCR) were calculated for each variant (Supplementary Table 2). Non-singleton variants with high evidence for pathogenicity (P1) are shown in table 6. Four variants have a VCR of higher than 0.01; p.E27K(Hb E) in the HBB gene associated with Beta thalassemia (VCR = 0.26), p.V37I in the GJB2 gene associated with congenital Deafness (VCR = 0.22), p.X143Q(Hb CS) in HBA2 gene associated with Alpha thalassemia (VCR=0.06) and c.-119_-116delGTCA in GALT gene associated with galactosemia (VCR =0.02). Several individuals were identified as homozygotes carrier for variants with high VCR. 9 individuals (1.5%) carry homozygotes p.E27K, 4 individuals (0.7%) carry homozygotes p.V37I and 3 individuals (0.5%) carry homozygotes p.X143Y.

Table 6 Well-established (P1 group) likely pathogenic/pathogenic carrier variants that were detected more than once in the Thai cohort.

GENE	Variants	VCR	Disorder	Disorder Category
	NM_000350:c.G5881A,p.G1961R	0.007		
	NM_000350:c.C1531T,p.R511C	0.003		
AGXT	NM_000030:c.T2C,p.M1?	0.005	HYPEROXALURIA	Serious
BEST1	NM_001139443:c.C404T,p.A135V	0.005	BESTROPHINOPATHY, RETINITIS PIGMENTOSA	Mild
CFTR	NM_000492:c.1393-1G>A	0.003	CYSTIC FIBROSIS	Serious
CYP21A2	NM_001128590:c.G754T,p.V252L	0.003	CONGENITAL ADRENAL HYPERPLASIA	Serious
DHCR7	NM_001163817:c.G725A,p.R242H	0.003	SMITH-LEMLI-OPITZ SYNDROME	Serious
FANCA	NM_000135:c.709+5G>A	0.003	FANCONI ANEMIA COMPLEMENTATION	Serious
GAA	NM_000152:c.C1935A,p.D645E	0.003	GLYCOGEN STORAGE DISEASE	Serious
GALT	NM_000155:c.-119_-116delGTCA,	0.017	GALACTOSEMIA	Serious
GBA	NM_001171811:c.A419G,p.N140S	0.003	GAUCHER DISEASE	Unpredictable
	NM_004004:c.G109A,p.V37I	0.216		
	NM_004004:c.235delC,p.L79Cfs*3	0.005		
	NM_000517:c.T427C,p.X143Q (Hb_CS)	0.055		
	NM_000517:c.A429T,p.X143Y (Hb_Paks)	0.005		
	NM_000518:c.G79A,p.E27K	0.257		
	NM_000518:c.126_129del,p.F42Lfs*19	0.005		
	NM_000518:c.-78A>G	0.005		
PAH	NM_000277:c.284_286del,p.I95del	0.003	PHENYLKETONURIA	Serious
PKHD1	NM_138694:c.T2507C,p.V836A	0.003	POLYCYSTIC KIDNEY DISEASE	Lifespan Limiting
RPGRIP1L	NM_001330538:c.3198_3199insTC,p.A1067Sfs*34	0.005	MECKEL SYNDROME	Lifespan Limiting
SBDS	NM_016038:c.258+2T>C	0.005	SHWACHMAN-DIAMOND SYNDROME	Serious
SLC22A5	NM_001308122:c.C51G,p.F17L	0.007	CARNITINE DEFICIENCY	Unpredictable
	NM_001160210:c.1663_1664ins GAGATTACAGGTGGCTGCCCGGG,p.A555Gfs*17	0.003		

	NM_001160210:c.C958T,p.R320X	0.003			
	NM_001160210:c.852_855del,p.M285Pfs*2	0.003			
SLC26A4	NM_000441:c.1546dupC,p.S517Ffs*10	0.005	DEAFNESS, PENDRED SYNDROME		Mild
UROS	NM_000375:c.T217C,p.C73R	0.003	PORPHYRIA, ERYTHROPOIETIC	CONGENITAL	Serious
USH2A	NM_206933:c.5572+1G>A	0.003	USHER SYNDROME		Serious

VCRs were used to calculate gene carrier rates (GCR) (Supplementary table 5). Genes with the 25 highest GCR are show in figure 20. For the high evidence variants (P1), genes associated with Beta thalassemia (HBB), Deafness (GJB2) and Alpha thalassemia (HBA2), obtained the highest GCR of 0.26, 0.22 and 0.06, respectively (Figure 20). 3 pathogenic variants were identified in the HBB gene including a non-synonymous (p.E27K), a variant in the promoter region (c.-78A>C) and a frameshift deletion (p.F42Lfs*19). In the GJB2 gene, one missense (p.V37I) and one frameshift deletion (p.L79Cfs*3) were identified. In HBA2 gene, Two stop loss variants (p.X143Q and p.X143Y) were identified.

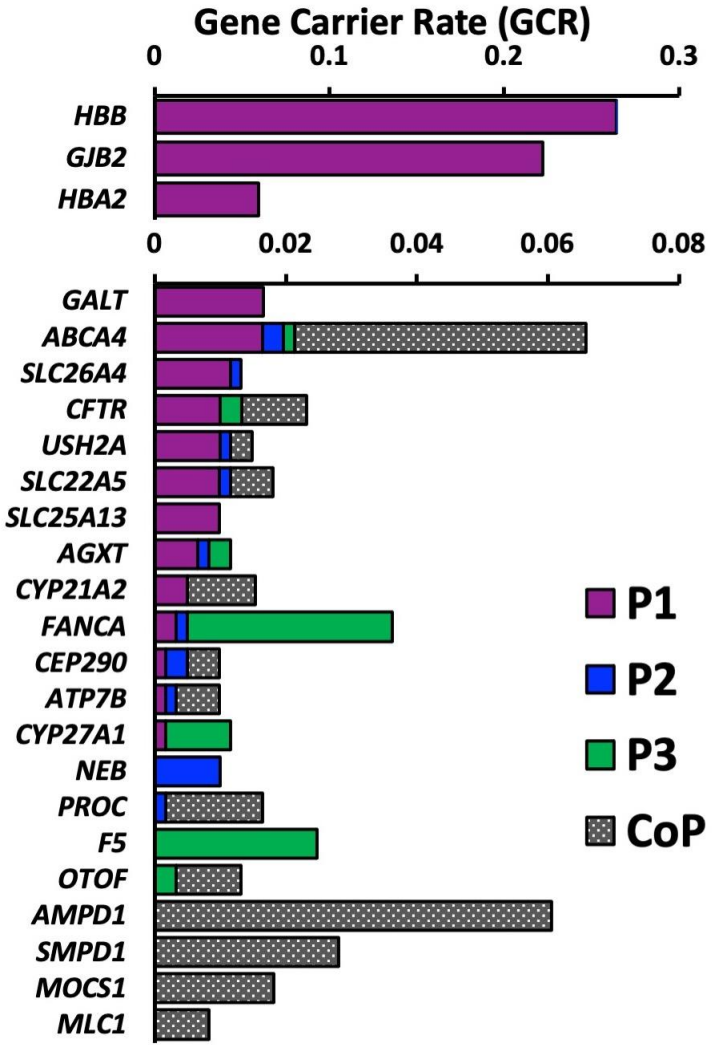


Figure 20: Gene Carrier rate of 25 autosomal recessive genes.

Variants within each gene were classified as P1, P2, P3 and CoP according to their evidence for pathogenicity.

Discussion

Here, we report an estimate of carrier rates in the Thai population for over 672 genes associated with AR disorders and found an enrichment of several AR variants in different subpopulations. Carrier rates for many genes reported in this study are the first to be reported in the Thai population. 263 reported likely pathogenic/pathogenic variants were identified. 62% (n=163) of variants identified showed limited supporting evidence for variant's pathogenicity. 100 AR variants were well-established with 6 variants found prevalent in Thai (VCR > 0.01) and 58.2% of the cohort carry at least one well-established AR variant. 1.67% of the cohort identified as SMA carriers used supplementary informatic tool and all three forms of alpha-thalassemia variants (a0, a+ and aND) were identified used NGS4Thal pipeline. The fine-scale population structure analysis revealed the Thai population complex genetics structure that can be separated into subgroups. Heterogeneity in VCR were observed between subgroups that reflects geographical and ethnic substructure.

p.E27K in HBB (Hb E), p.V37I in GJB2 and p.X143Y in HBA2 (Hb CS) are among the most prevalent AR variants in the Thai cohort with several homozygotes carriers detected. The detected allele frequencies correspond with frequencies reported in the Thai exomes database(83). Frequencies of these variants do not reflect the disease prevalence as these clinically significant variants may not be disease-causing(84-88). Previous studies reported carriers of homozygotes p.V37I to be associated with milder hearing impairment when compared to other pathogenic variants in the GJB2 gene and have a penetrance of only 17%(84, 85). In a longitudinal study, homozygotes p.V37I patients were found to have later age onsets of hearing impairment that progressively deteriorate(88). The variations in phenotypes of AR variant carriers suggested that interpreting variants, especially in carrier genetic testing, must be done with caution. Furthermore, variants can have different clinical outcomes when found in compound heterozygous with another pathogenic variant(85, 89). A study reported an increase in penetrance in patients with compound heterozygous p.V37I when compared to homozygote carriers26.

Differences in clinical outcomes between homozygotes and compound heterozygous states are not usually stated in the mutation database or are unknown. Because the complex relationship between variants on each allele can link to disease severity, the knowledge of alleles' combinational effect could influence reproductive decisions. Studying carriers' phenotypes, especially for variants prevalent in the population, could provide crucial information in couple counselling.

Variant misclassification is a recognised issue in data-sharing databases, such as ClinVar, and can lead to reporting false positive results(90). This study attempted to avoid reporting false positive result by used CLNREVSTAT, ClinVar's initiative to improve variant interpretation. Misclassification often arises from submitters' inconsistent classification system or limited evidence at the time of interpretation(90). CLNREVSTAT encountered the issues by evaluating evidence provided by submitters, such as the implementation of the ACMG guideline. While we focused on well-established likely pathogenic/pathogenic variants (P1) in this study analysis, over a hundred reported likely pathogenic/pathogenic variants were lacking evidence supporting their pathogenicity (P2 and P3). In addition, several variants with a conflicting interpretation of pathogenicity (CoP) show the potential to be clinically significant. For example, p.Val1106Ile in ATP7B gene that encoded for copper-transporting ATPase. While p.Val1106Ile did not disrupt copper transport function in a yeast functional analyse study, later studies found a 44.55% decrease in copper-ATPase activity in a patient carrying compound heterozygotes p.Val1106Ile and the variant obtained an odd-ratio of 10.5 (95%, CI=1.36-79.9) in another case-control study(91-93). Further study into variant pathogenicity would enable effective implementation of genetic data. The ongoing development of the Thai local genetic database is expected to improve interpretations and classifications of AR variants circulating in the Thai population(94). Reanalysis of these genetic results in the future could potentially increase yields of pathogenic variants(95).

Part II.III: Identification of an enrichment in autosomal recessive carrier in Thai subpopulations

An enrichment of carrier variants had been reported in some population subgroups as a result of past migration events or geographical isolation(96). A previous study compiling Thalassemia genetics surveyed in Thailand showed the distribution of Thalassemia variants to be highly geographically heterogeneous with variation observed in neighbouring provinces (97). The resources on population carrier frequencies at a fine scale could improve the estimation of the disease burden and choice of screening disorders. This will assist in guiding public health decisions in the prevention and management of AR disorders.

Studying population carrier frequencies based on self-reported population labels or ethnicity had demonstrated to be unreliable (50, 51). Assessing carrier rates based on genetic structure could provide an insight that was not available in existing population labels. Studies had illustrated the identification of fine-scale genetic substructure using the haplotype sharing method (ChromoPainter/fineSTRUCTURE) from genome-wide single nucleotide polymorphism array data (54, 56-58). WGS could provide a more detailed structure but running high-density genotype data on ChromoPainter can be computational extensive. A recent study demonstrated that PBWT-paint, a scalable haplotype sharing algorithm based on the positional Burrows-Wheeler transform, was able to capture genetic structure identical to ChromoPainter (54). PBWT-paint would allow detection of shared haplotypes in high-density WGS data.

Research Questions:

Are there an enrichment/s of autosomal recessive variant carrier in Thai subpopulation/s?

Research Objectives:

To uses haplotype sharing method in identifying Thai population genetic structure.

To classified Thai subpopulation based on population genetic structure.

To determine there is an enrichment of autosomal recessive variant carrier in any of the Thai subpopulation.

Expected benefits and application:

The information on enrichment of pathogenic variant in Thai subpopulation can be used to facilitate the development of disease prevention and control programs through precision public health approach by prioritizing economic resources and laboratory facilities that are limited to where disease poses the most burden.

Methods

Quality control of WGS samples for population structure analysis

Further QC will be performed for population structure analysis. PLINK2 were used to exclude: -

one individual from a closely related pair with KING kinship coefficients exceeding 0.125

Multidimensional scaling will then be performed to identify if there are any population outliers. Genotype data will be pruned with parameters --indep-pairwise 50 10 0.2. MDS will be performed using --mds-plot function and visualized using R (version 3.6.3). Through visual examination any outliers will be excluded for further analysis.

SNPs with missingness > 0.05.

Fine-scale population structure analysis

The QCed genotype data will be phased using SHAPEIT v2 following default parameter.

Phased genotype data will be used as an input for PBWT-paint.

The outputted PBWT-paint matrix will be used to calculate PCs using fineSTRUCTURE R tools (http://www.paintmychromosomes.com)

visualised in 2 dimensions using t-distributed stochastic neighbour embedding (t-SNE) implemented in the Rtsne package in R version 3.6.

Clustering based on population structure

Sample clustering will be done using a Gaussian mixture model implemented in the R package mclust. t-SNE dimensions will be used as an input for mclust. For each cluster assigned, samples' demographic data will be examined. Each cluster will be label according to the sample majority of geographical region or ethnic group. Variant carrier rate of prevalent variant within Thai population will be calculated for each cluster.

Statistical analysis

Descriptive statistics analysis will be performed using R version 3.6. The statistical chi-square test will be use in comparison between the cohort variant carrier rate and each cluster variant carrier rate. All informative data will be considered statistically significant at p-value less than 0.05.

Results

Fine-scale population structure analysis

Haplotype profiles of 589 Thais were mapped using PWBT-paint to examine the genetic structure of the Thai population. Separations were observed when the first 4 PWBT-paint PCs were projected in 2 dimensions using t-distributed stochastic neighbour embedding (figure 20a). Based on self-reported geographical regions and ethnicities data, the first-dimension display separation corresponds to the country's geographical north-to-south gradient. The second dimension follows the west to east gradient and separates Thai-Chinese ethnic group from the rest of the cohort. Clusters were assigned based on PWBT-paint matrix, at k=9 we observed clustering that segregate along with Thailand's 4 main geographical regions (central, north, north-east and south) and two major ethnic groups (Thai and Thai-Chinese) (figure 20b). Each cluster contains samples ranging from 50 to 89 individuals. Clusters were labelled based on the majority of samples' place of birth or ethnic group.

In the North-East region, sub-regional separation was observed. The population within this region were separated into four clusters (4-NE, 5-NE, 6-NE and 7-NE-N), where each cluster shows a distinct geographical pattern (figure 21). Population from 4-NE found to be located along the border between Thailand’s central, north-east and north region. Majority of population from 5-NE were found in the lower part of north-east region that share boarder with Cambodia. 6-NE population were found the central of northeast region. Lastly, population in 7-NE-N were found in both north-east and north region along Thailand and Laos border.

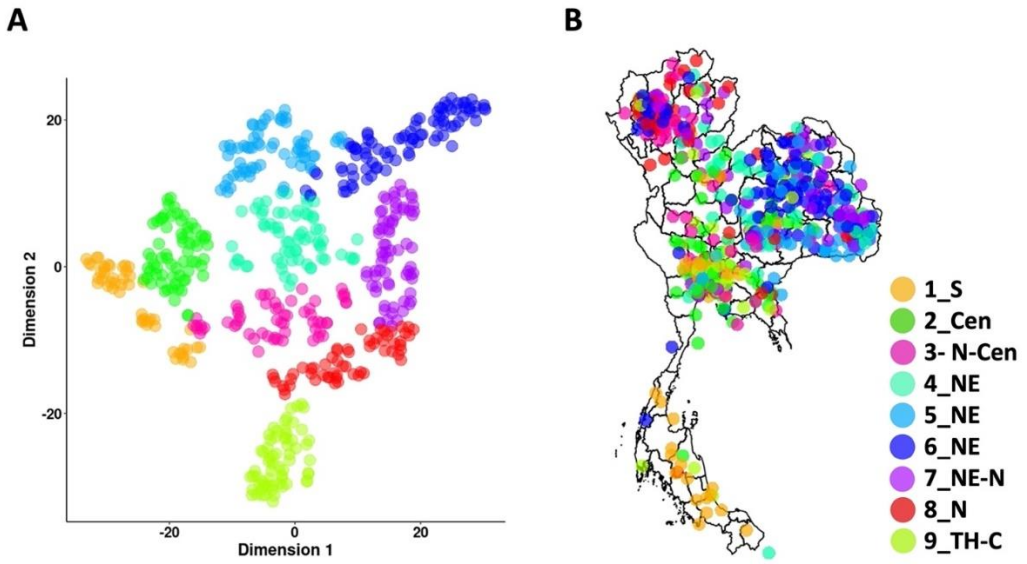


Figure 21: Thai population genetic structure based on PBWT-painting algorithm

(a) t-SNE visualisation of Thai population genetic structure based on PBWT-painting algorithm. Samples were clustered into groups using mclust. (b) Geographical distribution of sample's place of birth. Samples were coloured based on assigned clusters. Source of shapefile: United Nations Office for the Coordination of Humanitarian Affairs https://data.humandata.org/dataset/thailand-administrative-boundaries retrieved on 19 august 2021

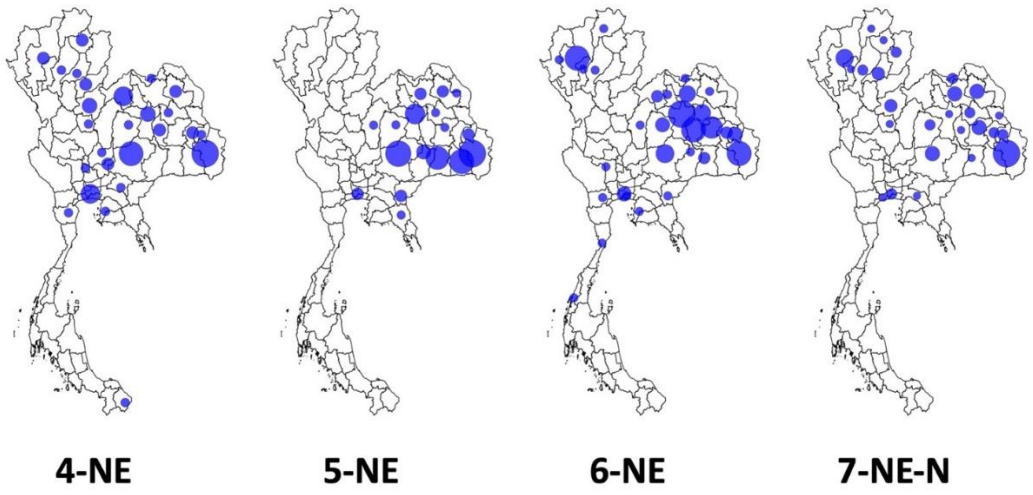


Figure 22 Geographical distribution by provinces of 4 Northeast clusters (4-NE, 5-NE, 6-NE and 7-NE-N) based on sample's place of birth.

The number of samples in each province is represented by the circle diameter.

Enrichment of AR variants in subpopulations

Carrier rates of well-established likely pathogenic/pathogenic AR variants(P1) prevalent in Thai (VCR > 0.01) were examined for each genetic cluster (Table 7). p.V37I(GJB2) VCR vary from 0.070 in the South cluster(1-S) to 0.107-0.138 in North-East clusters(4-NE, 5-NE, 6-NE and 7-NE-N). c.-119_-116delGTCA(GALT) VCR are highest in the North-Central cluster(3-N-Cen) at 0.025 but are absent in the North (8-N) and the Thai-Chinese(9-TH-C) cluster.

Thalassemia variants show the highest enrichment within different clusters. For Hb E, the highest elevation in VCR when compared to the rest of the cohort (VCR = 0.26) was observed in cluster 5-NE in the north-east at 0.49 (OR = 3.6, p < 1.65x10-6) follow by 7-NE-N at 0.34 (OR = 1.8, p < 0.03). Thai-Chinese (9-TH-C) and the north (8-N) cluster show lower carrier rate for Hb E than the rest of the cohort at 0.06 (OR = 0.31, p < 7.06 x10-3) and 0.12 (OR = 0.42, p < 0.03), respectively. For Hb CS, when compared to the rest of the cohort (VCR = 0.06) elevated carrier rates are found in North-East clusters, 5-NE at 0.12 (OR = 3.0, p < 0.01) and 6-NE at 0.11 (OR = 2.7, p < 0.02). Finally, higher VCR for -□3.7 deletion was found in 6-NE at 0.06 (OR = 3.9, p < 0.01) when compared rest of the cohort at 0.02 and higher VCR for Hb Pakse was found in 7-NE-N at 0.03 (OR = 16.3, p < 0.02) when compared rest of the cohort at 0.01.

Table 7 Variant carrier rate of carrier variants separated by population subgroups.

	GJB2	*GALT*	-SEA	-□$^{3.7}$	Hb_CS	Hb_Pakse	Hb_E
1-S	0.070	0.010	0.000	0.000	0.000	0.000	0.260
2-Cen	0.097	0.015	0.045	0.000	0.045	0.000	0.149
3-N-Cen	0.100	0.025	0.033	0.017	0.050	0.000	0.167
4-NE	0.138	0.008	0.031	0.000	0.062	0.015	0.308
5-NE	0.121	0.015	0.030	0.045	0.121	0.000	0.485
6-NE	0.107	0.006	0.034	0.056	0.045	0.000	0.270
7-NE-N	0.086	0.000	0.014	0.029	0.114	0.029	0.343
8-N	0.121	0.000	0.052	0.017	0.052	0.000	0.121
9-TH-C	0.117	0.000	0.063	0.000	0.016	0.000	0.063

Discussion

The fine-scale study of population genetic structure reveals heterogeneity in VCR within the Thai population that reflects geographical and ethnic substructure. Variation in VCR within the region has been previously reported for some AR variants. *Tritipsombut et al.* found Hb E carrier rates to vary from 39.3% to 43.1% in the Northeast when the region was separated based on geographical labels(98). In this study, a more distinct elevation of Hb E was observed in the Northeast (27.0% - 48.5%). Categorised populations based on genetics could reveal a complex substructure that was missed when used self-identified geographical data and enables a better understanding of the disease's burden. Neighbouring countries in close proximity with identified clusters also reported similar elevations. Preah Vihear reported higher Hb E prevalence than other regions of Cambodia(99). Interestingly, Preah Vihear shares border with provinces where 5-NE are located (figure 4). Hb CS that was prevalent in the 5-NE and 7-NE-N clusters also found prevalent in So ethnic group in the south of Laos and the Có-Tu ethnic group in Vietnam(86, 100). High prevalence of Hb E and Hb CS within these regions could be resulted from a founder effect. A study found shared α^0-thalassemia SEA deletion alleles haplotype between the Chinese population and carriers from Thai, Laos, and Cambodian(101). This may

also explain the higher prevalence of SEA deletion within the Thai-Chinese community observed in this study(9-TH-C). The fine-scale population genetic structure analysis identifies population subgroups at risk for carriers of AR variant and provides insight into genetic factors underlying the disease.

Limitations

There are limitations to this study. The carrier rates in this study were calculated from a limited sample size. The reported carrier rate may not capture all rare AR variants circulating within the population and may affect the estimation of VCR in some variants. While the study included samples from multiple regions within Thailand, not all regions were equally represented. Sampled populations from the south only represented 4% of the cohort. An increase in sample size could reveal another layer of genetic structure.

Summary

Population carrier rates are an important resource for the development of carrier testing and estimations of disease burden. Here, we report an estimate of carrier rates in the Thai population for over 672 genes associated with AR disorders and found an enrichment of several AR variants in different subpopulations. Carrier rates for many genes reported in this study are the first to be reported in the Thai population. 263 reported likely pathogenic/pathogenic variants were identified. 62% (n=163) of variants identified showed limited supporting evidence for variant's pathogenicity. 100 AR variants were well-established with 6 variants found prevalent in Thai (VCR > 0.01) and 58.2% of the cohort carry at least one well-established AR variant. 1.67% of the cohort identified as SMA carriers used supplementary informatic tool and all three forms of alpha-thalassemia variants (a0, a+ and aND) were identified used NGS4Thal pipeline. The fine-scale population structure analysis revealed the Thai population complex genetics structure that can be separated into subgroups. Heterogeneity in VCR were observed between subgroups that reflects geographical and ethnic substructure.

Despite the limited sample size, 23.8% of likely pathogenic/pathogenic AR variants reported in this study are absent from the gnomAD population database. Current databases are not extensive with many populations, including Southeast Asians, being underrepresented (102-104). We believe carrier rates reported in this study are an underestimate of the disease-causing variants circulating in the Thai population. Thai population-specific variants may be absent from current mutation databases or are understudied, resulting in the "Variant of Unknown Significance" classification due to limited knowledge on variant pathogenicity.

In conclusion, we demonstrated WGS to be a powerful tool in examining population AR variants. It assists in identifying various types of pathogenic variants from point mutations and small insertion/deletions to large structural variation, which improve the estimation of population carrier rates. The population structure analysis used

WGS identify variant distribution within the population at the finest scale. The comprehensive overview of population carrier rates will be a useful resource for the development of carrier testing recommendations and estimation of disease burden. The information on enrichment of pathogenic variant in Thai subpopulation can be use to facilitate the development of disease prevention and control programs through precision public health approach by prioritizing economic resources and laboratory facilities that are limited to where disease poses the most burden(105, 106).

Part III: Genetic risks and association with severe COVID-19 among global populations

While population demographics and healthcare infrastructure influence mortality, genetic predisposition may also influence clinical severity of COVID-19. Recent genome-wide association studies identified multiple host genetic factors associated with disease susceptibility and severity (10-12). These studies examined mostly European populations, which prompted us to examine these disease-modifying loci in the Asian population.

Research Questions:

What is the allele frequency of severe COVID-19 risk alleles in different global populations?

Research Objectives:

To examine allele frequency of severe COVID-19 risk alleles in different global populations.

Expected benefits and application:

Finding of this study will determine prevalence of COVID-19 risk alleles in different global populations that is essential for studying the effect of COVID-19 risk alleles different global populations.

Methods

The allele frequencies of 5 risk alleles report associated with severe covid-19 (table 2) will be extracted from gnomAD, GenomeAsia 100k, and Brugada syndrome Southeast Asia database.

SNP	Chr.: pos.	Risk	Alt.	Locus
rs73064425	3: 45,901,089	T	C	*LZTFL1*
rs657152	9: 133263862	A	C	*ABO*
rs2109069	19: 4,719,443	A	G	*DPP9*
rs74956615	19: 10,427,721	A	T	*TYK2*
rs2236757	21: 34,624,917	A	G	*IFNAR2*

Different AC and AN will be use to examine allele frequencies of different populations within the database

Form gnomAD database allele frequencies will be calculated for:

East Asia

Africa

Ashkenazi Jewish

European(non-Finnish)

European(Finnish)

Latino

From GenomeAsia 100k database allele frequencies will be calculated for:

Northeast Asia

Southeast Asia

South Asia

Philippines

Indonesia

Malaysia

China

South Korea

Japan

From Brugada syndrome Southeast Asia database allele frequencies will be calculated for:

Control sample of Thai population

Results

Chromosomal locus 3p21.31 was highly correlated with disease severity in hospitalized Italian and Spanish COVID-19 patients (rs11385942; 95% confidence interval (CI), p = 1.15x10−10) (10), which was con- firmed in the United Kingdom (rs13078854; 95% CI, p = 1.6x10−18) (11) and in a multi-ethnic study (rs73064425; 95% CI, p = 4.77x10−30) (12). This gene- rich locus includes SLC6A20 (encoding sodium-imino acid transporter 1, which interacts with COVID-19 ACE2 receptor) and multiple chemokine receptors (CCR9, CXCR6, CCR1, and CCR2). Our analysis found that the frequency of the risk allele rs11385942 at this locus differs vastly among Southeast Asians, ranging from 0.21 in the Filipino population to 0.06 in the Thai population, but it was rare in Northeast Asians.(Figure 22). Surprisingly, frequencies of risk alleles at 19p13.2 (rs74956615) and 19p13.3 (rs2109069) were also low among Northeast Asians relative to other populations. Collectively, these three loci encode inflammatory response genes (CCR2, TYK2, and DPP9) and are hypothesized to influence COVID-19 severity through hyper-inflammatory response and subse- quent organ injury (11).

The frequency of the risk allele at rs657152 located on 9q34.2 (linked to ABO blood group locus) varies from 0.25 in Indonesians to 0.48 in South Koreans. This locus found to be associated with European patients with respiratory failure (rs657152; 95% CI, p = 4.95x10−8) (10). In addition, another study found the same locus to be associated with COVID-19-infected individuals when compare to those uninfected at lower p-value (95% CI, p = 5.3x10−20) (11).. On chromosome 21q22.1 where the interferon receptor gene IFNAR2 is located, the frequen- cies of the risk allele rs2236757 is 0.56 in Southeast and 0.46 in Northeast Asians (higher than 0.29 found in non-Finnish Europeans).

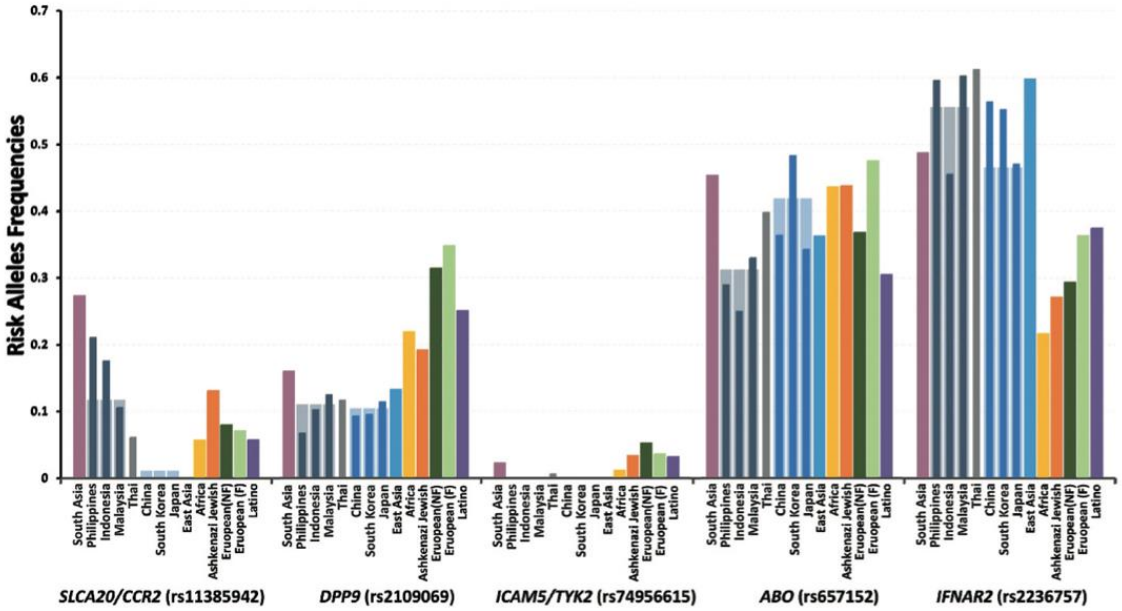


Figure 23 Analysis of the different frequencies of risk alleles known to be associated with the susceptibility and severity of COVID- 19 in different populations.

Allele frequencies available from the gnomAD database, which include East Asia, Africa, Ashkenazi Jewish, European(non-Finnish), European(Finnish) and Latino, GenomeAsia 100k database, which includes South Asia, Philippines, Indonesia, Malaysia, China, South Korea, and Japan, and from a control Thai population (n = 236) were analyzed.

Summary

Along with other factors, lower COVID-19 mortality in East Asian countries may be attributed to lower frequencies of risk alleles. The impact of known risk alleles may not be universal among the different human populations in predicting COVID-19 severity and susceptibility due to differences in the patterns of linkage disequilibrium in some loci. Supplementary studies in Latin America, Africa, and Asia may provide further explanation in the observed unequal disease severity in different populations.

Part IV: The effect of Thai genetic variation on imputation performance

Part IV.I: Evaluate imputation performance.

Previous studies have demonstrated strong variations in imputation performance when common reference panels were applied to different populations(14, 15). For example, imputation using HRC offered better accuracy among European populations than among the Han-Chinese population(15). There are limited data regarding imputation performance when public reference panels are used in populations not widely represented in the reference. In turn, this causes difficulties in the reference selection, in understanding the limitations associated with each reference panel, and created challenges when performing genomic research in populations that are underrepresented. To our knowledge, the Thai population is not represented in any current public reference panel except for GenomeAsia (n=2), and therefore, issues relating to imputation accuracy and panel selection are particularly important to genetic studies in this population.

Research Questions:

What is the genotype yield and accuracy when used 1000G, HRC, GenomeAsia, and TOPMed to impute Thai population?

Does the population structure effect accuracy of imputed variant?

Research Objectives:

To evaluate genotype yields and imputation accuracy when genotyping imputation of Illumina Global Screening Array (GSA) among Thai individuals using four different high-density reference panels (1000G, HRC, GenomeAsia, and TOPMed).

To evaluate the population structure effect on imputation accuracy.

Expected benefits and application:

Finding from this study will facilitate selection of reference panel when imputation is performed in Thai population allow researcher to understand the limitation of each reference panel used.

Methods

Samples enrolled in this study will be selected based on availability of both Genome-wide genotyping and WGS data. Genome-wide genotyping was done using the GSA platform, as previously described24. WGS from South-east Asian Brugada Syndrome cohort will be use as an imputation validating genotype.

Genotype Imputation

Pre-imputation quality controls (QCs) will be performed on genotyping array data following Scelsi et al., 2018 recommendations. PLINK (version 1.9) will be used to exclude samples:-

with discordance between genetically inferred and self-reported sex,

with genotype missingness >0.05, and

with duplicates or first-degree relatives by using the --rel-cutoff command in PLINK (removing one member of each pair of samples with genomic relatedness >0.5)26.

Compatibility at variant level between geno-typing array data and each of the reference panels will be examined using the checking tools by W. Rayner (http://www.well.ox.ac.uk/~wrayner/tools/), to correct consistency of strand, alleles, positions, Ref/Alt assignments, and minor allele frequency differences.

Imputation will be performed on the Michigan Imputation Server (https://imputationserver.sph.umich.edu) using Eagle2 phasing and Minimac imputation. Based on the reference panels, 1000G, HRC, GenomeAsia, and TOPMed, four im-puted genotype datasets will be generated.

Evaluation of genotype yield

Genotypes will be extracted and counted using BCFtools (ver-sion 1.10.2). Minimac-R2 values, ranging from 0 (lowest confidence) to 1 (highest confidence), were used to reflect the imputation confidence for each imputed variant. Imputed variants were clustered according to five Minimac-R2 ranges: [0,0.2), [0.2,0.4), [0.4,0.6), [0.6,0.8), and [0.8,1].

Evaluation of imputation accuracy

Imputation accuracy of the four imputed datasets that used the 1000G, HRC, GenomeAsia, and TOPMed reference panel will be examined. Chromosome 1 variants from each of the imputed datasets will be validated against high coverage genotypes called from WGS (among the same samples).

The WGS data underwent QC using Starling's filtering criteria to filter out sites that have genotype conflicts with proximal indel calls, locus quality score <30, locus quality score <14 for heterozygous or homozygous variant, the fraction of basecalls at a site >0.4, locus read evidence displays unbal-anced phasing patterns, calls with a sample depth three times higher than the chromosomal mean, or genotype calls from variant callers not consistent with chromosome ploidy. Variant sites within the cohort with missingness >0.10 or deviation from Hardy-Weinberg equilibrium (P-value <1 x 10-6) will be excluded. Samples with >0.05 genotype missingness will be removed.

QCed WGS variant sites found in all four imputed genotyping datasets will be selected for evaluation of imputation accuracy. Accuracy will be measured in terms of genotype concordance rate (GCR) between the imputed and validating WGS data for each sample. The underlying GCR for each of the four reference panels will be examined and visualized used ggplot2 package in R (version 3.6.3). Evaluation of imputation accuracy will be further performed using chromosome 21 variants as validation.

Population structure and admixture analysis

The Thai cohort population structure will be examined using a multidimensional scaling (MDS) method implemented in PLINK (version 1.9). Genotyping array data will be pruned with parameters --indep-pairwise 50 10 0.2, leaving 135,661 markers. MDS was performed using --mds-plot function and visualized using R (version 3.6.3) to examine the presence of cohort population sub-structure. Chinese genetic admixture in the study cohort will be examined used genotype dataset of 44 North and South Han-Chinese samples acquired from the Human Diversity Genome Project. Genetic admixture will be estimated using ADMIXTURE software version 1.3 under the setting of K=2(107).

Results

Genotype yield and confidence level

Four different public reference panels (1KGP, HRC, GenomeAsia, and TOPMed) were used to impute SNP-array of 415 Thais from the Southeast Asian Brugada Syndrome cohort. The number of genotypes obtained vary when different reference panels were used. The highest genotypes yield of 271 million (M) achieves when used TOPMed panel (Table 8). TOPMed obtains 6x more genotypes than that of 1KGP (43.8 M), 7x more than HRC (39.1 M), and 13x more than GAsP (21.5 M). In terms of insertion/deletion (INDEL), imputation uses TOPMed obtains 20.9 M INDELs and 1KGP obtains 3.23 M. Due to lack of INDEL in HRC and GenomeAsia reference panels, INDELs could not be infer when these two references were used.

When used Minimac-R2 to examine the number of genotypes obtain at different imputation confidence level, TOPMed offers the highest number of high-confidence imputed genotypes (R^2 > 0.8) at 6.99 M (Table 8). Imputation used 1KGP, GenomeAsia, and HRC obtain lower number of high-confidence genotypes (R^2 > 0.8) at 5.28 M, 5.06 M, and 4.89 M, respectively. The number of genotypes reduce substantially when R^2 cut-offs were applied with the largest reduction presented when used TOPMed. Imputation used TOPMed infer high portion of genotypes with low-confidence. We examined the distribution of imputed genotypes over the range of 0.2 to 1.0 R^2 (Figure 23). Imputation used GenomeAsia shows high concentration of genotypes within the very high-confidence range (R^2 of 0.9-1.0). TOPMed show the lowest density of high confidence genotypes.

Table 8 Number of imputed genotypes when varying their confidence Minimac-R2 levels.

	Number of imputed genotypes in millions (M)							
R^2	GAsP		1KGP		TOPMed		HRC	
Cut-off	#SNP	#INDEL	#SNP	#INDEL	#SNP	#INDEL	#SNP	#INDEL
none	21.50M	n/a	43.80M	3.230M	271.00M	20.900M	39.10M	n/a
0.2	9.87M	n/a	13.10M	1.420M	19.50M	1.460M	12.40M	n/a
0.4	8.26M	n/a	10.10M	1.130M	14.70M	1.090M	9.95M	n/a
0.6	6.86M	n/a	7.88M	0.866M	11.20M	0.815M	7.71M	n/a
0.8	5.06M	n/a	5.28M	0.532M	6.99M	0.496M	4.89M	n/a

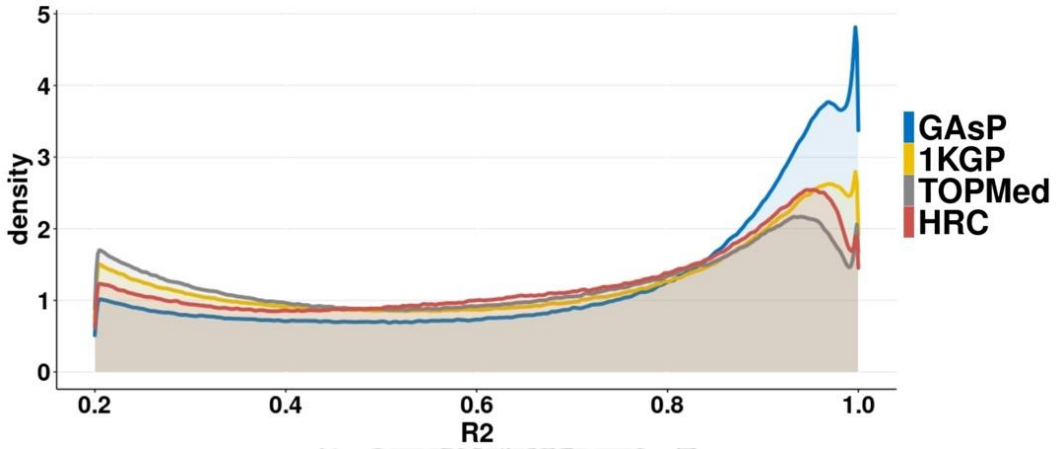


Figure 24 Density plot of genotypes obtaining Minimac R^2 between 0.2 and 1.0 after imputed using GAsP, 1KGP, TOPMed or HRC reference panel.

Imputation accuracy

Imputation accuracies were examined uses genotype concordance rates (GCR). For each sample, GCR was calculated between imputed genotypes and validation genotypes called from WGS. Overall, imputation used GenomeAsia achieves the highest accuracy with cohort median GCR of 0.973 (Figure 24). Median GCRs reduce when used 1000G (0.964), TOPMed (0.945), and HRC (0.931). Imputation accuracies are consistently high for all samples within the cohort when used GenomeAsia (GCRs 0.970–0.978). Higher variation of GCR can be found when used TOPMed (0.935–0.963). When used TOPMed panel, a group of samples achieve high GCR that depart from the cohort mean(outlier). The examination of demographic data shows that high number of individuals within this group self-identified as Thai-Chinese (data not shown).

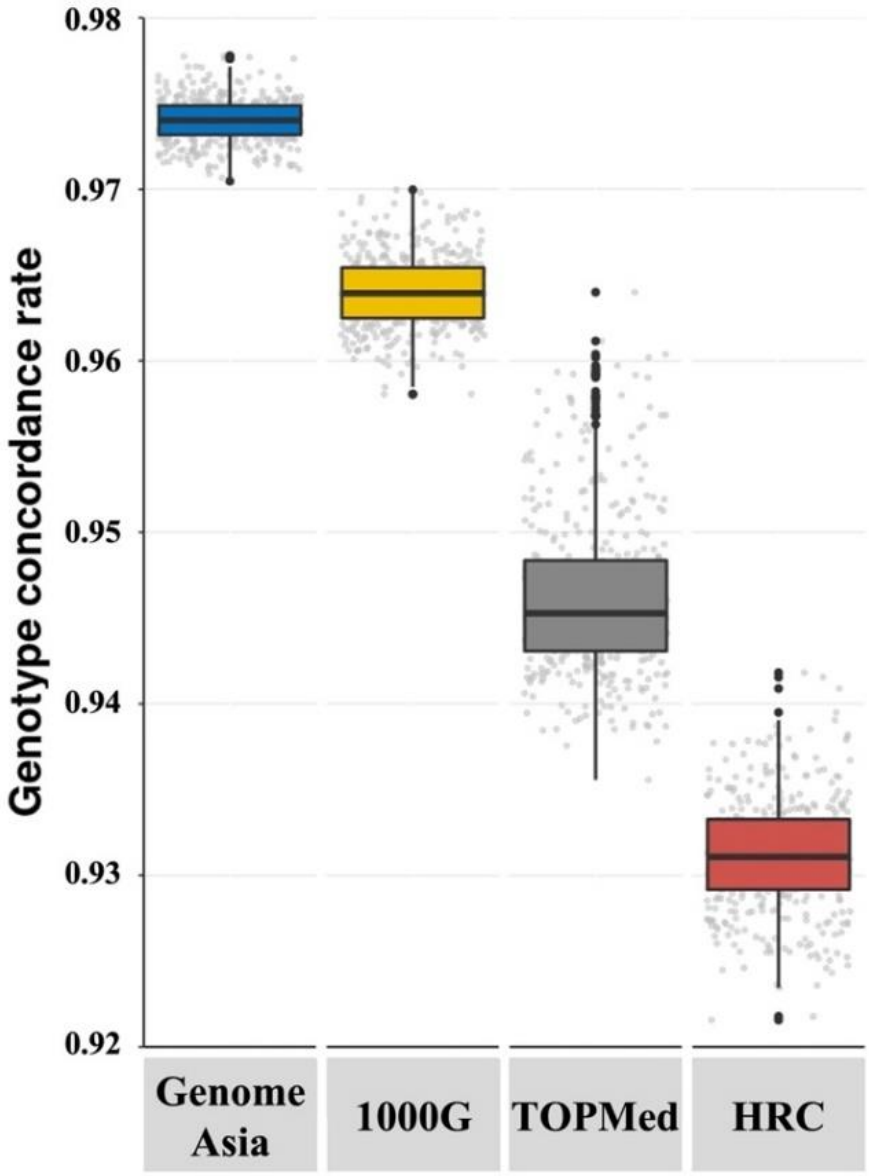


Figure 25 Imputation accuracy measured by genotype concordance rate (GCR) using GenomeAsia (GAsP), 1000 Genomes (1KGP), TOPMed and HRC reference panels.

GCR was evaluated when genotype imputation was done on the known WGS genotypes.

We then investigated the effect of population structure within the Thai cohort on imputation accuracy when used TOPMed panel. From multidimensional scaling analysis, samples outputted GCR correspond with the horizontal axis on the MDS plot (Figure 25a). Individuals obtaining high GCR when used TOPMed clustered together and separated from other samples. Admixture analyses were performed to determine if this cluster are Thai-Chinese as suggested by the demographic data. Using North and South Han-Chinese genotype datasets acquired from the Human Diversity Genome Project, admixture analysis reveal that individuals within the high GCR cluster also have high degree of Han-Chinese admix (Figures 25b and 25c).

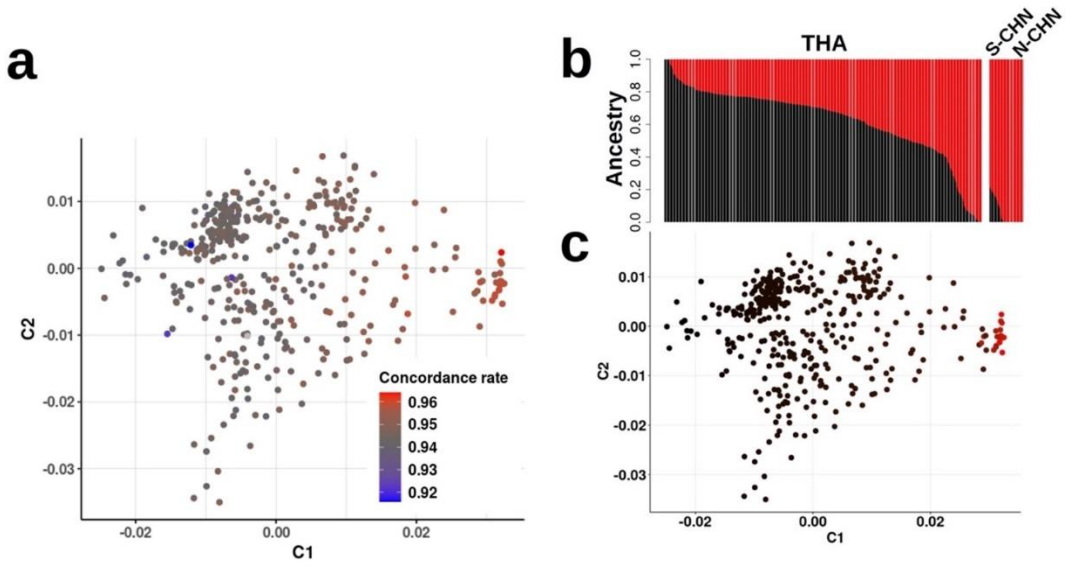


Figure 26 Admixture analysis

(a) Multidimensional scaling plot of 415 individuals coloured with genotype concordance rate obtained when assessed genotypes imputed with TOPMed panel against genotypes from whole genome sequencing. (b) Admixture plot of genome-wide genotype data of Thai and south and north Han-Chinese (S-CHN and N-CHN) acquired from the Human Diversity Genome Project (c) Multidimensional scaling plot of 415 individuals coloured with Q estimate from genome-wide genotype data of Thai and south and north Han-Chinese (S-CHN and N-CHN) from Admixture v. 1.3.

We examined the effect of R2 cut-offs on imputation accuracy. Imputation accuracy increases with more stringent R^2 cut-off (Figure 27). At high-confidence imputed genotypes (R^2 > 0.8), all samples achieve GCR above 0.967 regardless of reference panel used. TOPMed and HRC GCRs significantly improved with the median GCR approaching 0.974 and 0.973, respectively. GenomeAsia achieved the highest median GCR at 0.987.

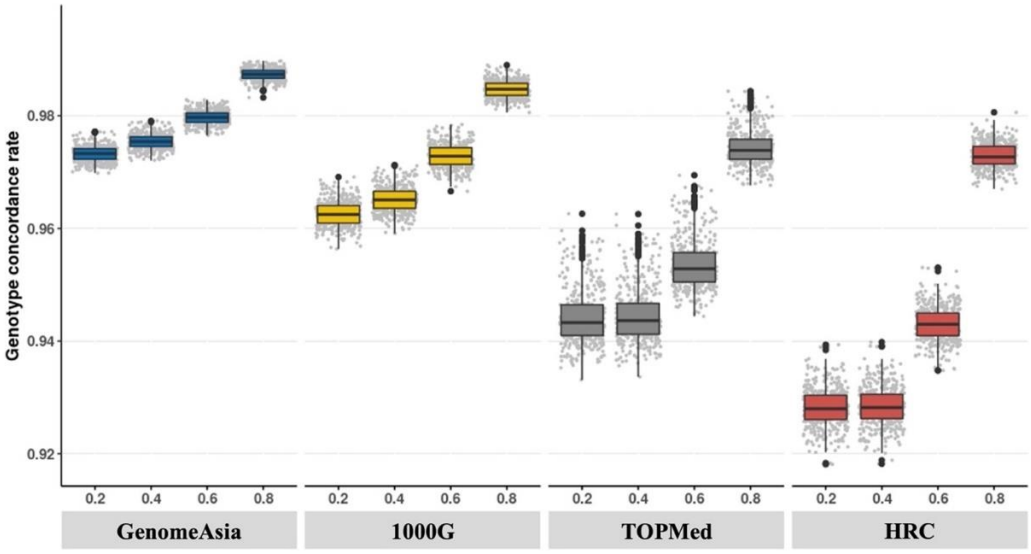


Figure 27 Imputation accuracy of Thai cohort at varying the R^2 cut-offs at 0.2, 0.4, 0.6 or 0.8.

Imputation was performed using the GAsP, 1KGP, TOPMed and HRC reference panels, The imputation accuracies were evaluated using GCR.

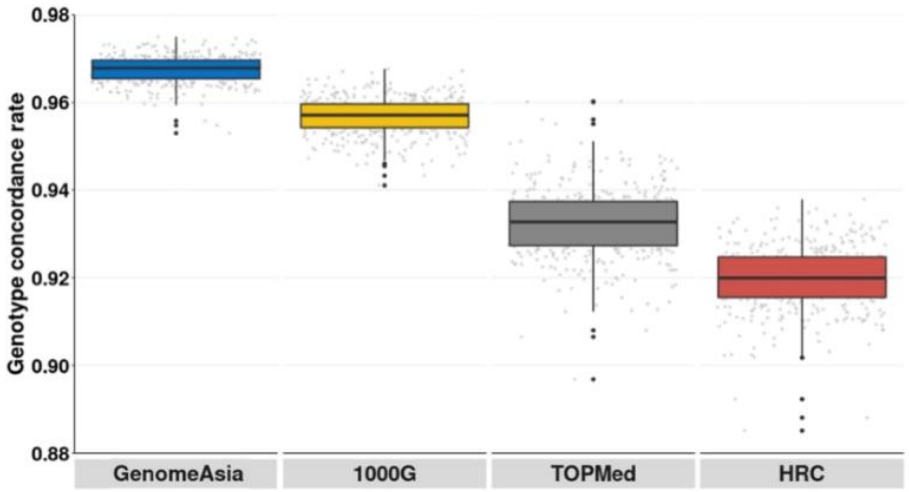


Figure 28 Imputation accuracy of chromosomes 21.

GCR was measured across 412 Thai individuals, using GenomeAsia, 1000G, TOPMed, and HRC reference panels. GCR was computed by comparison of imputed genotypes to validating genotypes from WGS. Data are presented as boxplots with distributions of sample GCR on the y-axis and imputation reference on the x-axis.

Discussion

TOPMed represents an exceptionally large reference sample (N=97,256). In concordance with previous studies, the larger reference size increases variant sites for imputation that can be beneficial in further association analysis(108, 109). Unfortunately, the larger TOPMed and HRC (N=32,488) datasets, when used to impute our Thai cohort, achieved lower imputation accuracy than the smaller 1000G (N=2,504) or GenomeAsia (N=1,739) reference panels. A reduced performance of HRC has previously been described in non-European datasets, including those of Han-Chinese and African ancestry; here, it was suspected that the overrepresentation of European ancestry individuals in the HRC panel may cause bias during phasing and haplotype selection processes(15, 110). While over 1,184 East Asian individuals are represented in TOPMed, it only accounted for 1.22% of the total reference samples. Similar to HRC, the overrepresentation of populations with low genetic similarity to this study cohort in TOPMed may also be responsible for the low accuracy observed.

The high imputation accuracy of GenomeAsia may be attributable to its diverse representation of populations genetically similar to our study cohort. The GenomeAsia reference contains data on >219 Asian populations. Indeed, a previous study demonstrated an improvement in imputation performance when additional populations were added to the reference(111). Thailand is located at the center of mainland Southeast Asia with a high degree of genetic admixture from neighbouring countries through past migrational events(112). While only 2 Thai WGS are represented in GenomeAsia, the diverse representation of genomes from neighboring countries likely provided a useful haplotype reference that benefited different subpopulations within our Thai cohort, leading to a higher accuracy throughout. In contrast, the diversity of Asian populations enrolled in the TOPMed study may not be as extensive with some Thai population subgroups underrepresented, as lower accuracies were observed in some samples within the cohort. The higher accuracy found in Thais with Han-Chinese admixture may reflect the high proportion of Han-Chinese ancestry represented in the East Asian population of the TOPMed database.

Part IV.II: Evaluate imputation accuracy of rare variants.

The advent of next-generation sequencing has led to an increase in whole genome sequencing (WGS) availability, enabling the construction of high-density reference panels. While initially reference panels could accurately infer variants with minor allele frequencies (MAFs) >5%, the increased size and sequencing coverage of recent high-density panels has enabled imputation down to low-frequency, 5% > MAF > 1%, and rare, MAF < 1%, variants(113-115). This has allowed examination of the human genome at a finer resolution, leading to identification of thousands of novel associations in GWAS(116-118).

Research Questions:

Does the accuracy of imputed variant effect by minor allele frequencies?

Research Objectives:

To evaluate the effect of variant minor allele frequencies on imputation accuracy.

Expected benefits and application:

Finding from this study will allow researcher to understand the use and limitation of imputing rare variant in Thai population.

Methods

Imputation accuracy and allele frequencies

Imputation accuracy of variants at different allele frequencies (AFs) were examined used total AF from Genome Aggregation Database (gnomAD) version 2.1.1.

The squared Pearson correlation between imputed and validating WGS variants were used to measure imputation accuracies.

Variants will be classified into AF bins according to gnomAD AFs. Variants were binned at 1, 0.05, and 0.01, to represent common, low-frequency, and rare variants, respectively. Finer examination of rare variants will be performed following AF bins at 0.01, 0.009, 0.008, 0.007, 0.006, 0.005, 0.004, 0.003, 0.002, and 0.001. Square Pearson correlation will be computed for each AF bin used GLIMPSE concordance tools(119).

Result

Imputation accuracy and allele frequency

At different minor allele frequencies (MAFs), imputation used GenomeAsia offers better accuracies than other reference panels (Figure 28a). The common variants (AF ≥ 0.05) and low-frequency variants (0.05 > AF ≥ 0.01) group show similar square Pearson correlation patterns. Accuracy decreases considerably in the rare variants group (AF < 1%) for all four reference-panels. For rare variants, Ge-nomeAsia achieves the highest accuracy with square Pearson correlation of 0.275 follows by 1000G (0.228), TOPMed (0.200) and HRC (0.184). Finer examination of rare variants shows imputation accuracy continue to decrease with AF. GenomeAsia outperforms other reference panels to 0.001 > AF ≥ 0 group (Figure 28b).

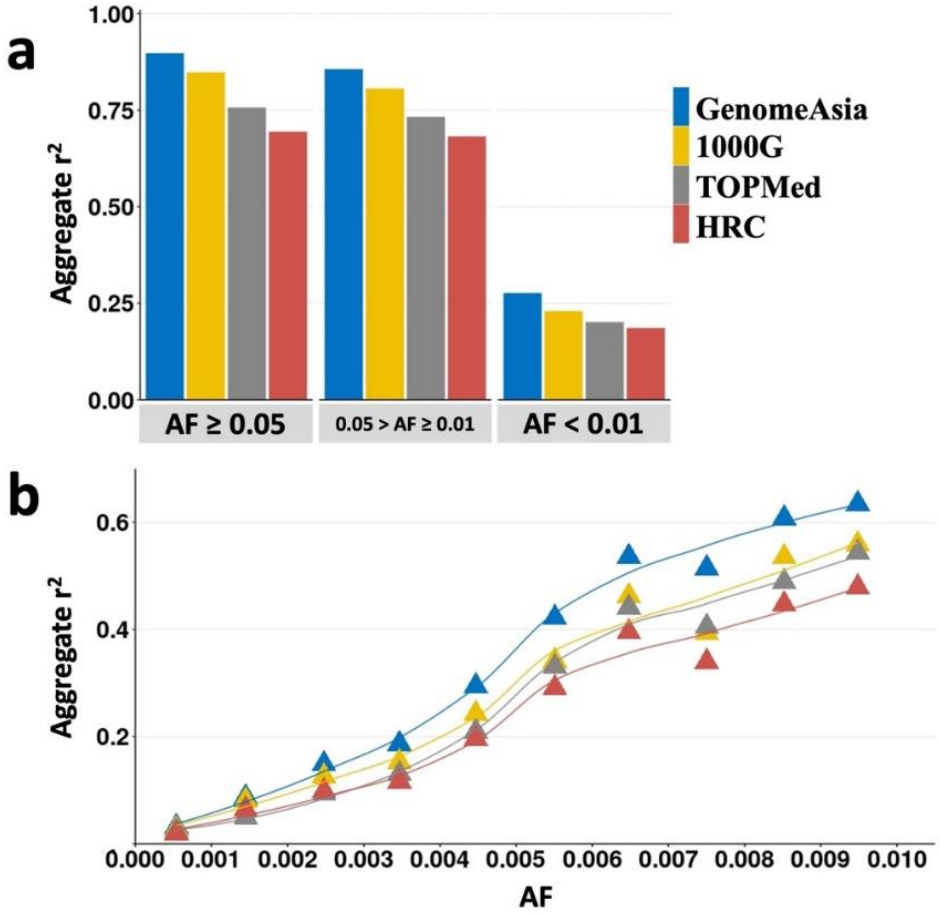


Figure 29 The effect of imputation accuracy based on allele frequencies.

(a) Imputation accuracy of common (AF ≥ 0.05), low-frequency (0.05 > AF ≥ 0.01) and rare (AF < 0.01) variants. (b) Imputation accuracy of rare variants at a finer resolution. Accuracies were measured used the squared Pearson correlation between imputed and validating WGS variants. Variants r2 were aggregated into groups according to AF from gnomAD (version 2.1.1).

Discussion

Although GenomeAsia yielded the best imputation accuracy for all AF bins, imputation accuracy strongly decreased with lower MAF as reported in other populations(120). We found a 30.3% reduction in squared Pearson correlation of rare variants when compared to common variants. Several approaches have been proposed to improve imputation accuracy for rare variants. First and foremost, an increase in reference size strongly benefits rare variant imputation(121, 122). As GenomeAsia currently has the smallest sample size of all four panels studied, an increase in Asian reference samples may vastly improve rare variant imputation accuracy. Secondly, using population-specific reference panels(115, 122, 123). As costs decrease and sequencing becomes more widely accessible, WGS should enable the construction of a Thai population-specific reference panel in the near future.

Limitations

We acknowledge several caveats and limitations of the present study. Imputation accuracy was not examined for all chromosomes, although similar results were obtained for chromosomes 1 and 21 (Figure 27). Evaluation of imputation accuracy was limited to WGS high-coverage regions. The accuracy of INDELs was not evaluated in this study, as this class of variation could only be obtained from imputation using TOPMed and 1000G reference panels.

Summary

This study evaluates the use of four different public reference panels (1000G, GenomeAsia, HRC, and TOPMed) in genotype imputation of Thai SNP-arrays data. The selection of a reference panel affects the number and accuracy of resulting genotypes. Although, TOPMed offers the highest number of genotypes after imputation, imputation used GenomeAsia achieves the best accuracy with low variability within the cohort (GCR from 0.96-0.98). Interestingly, imputation used TOPMed displays slightly higher variation in GCR (0.92-0.96). We demonstrate that the cohort population structure effects imputation accuracies when used TOPMed with Chinese admixed individuals obtain higher accuracy. When considering the accuracy at different MAF groups, imputation used GenomeAsia outperforms other reference panels to the very rare variants ($0.002 > AF \geq 0.001$).

In conclusion, our results demonstrate the benefit of having similar genetic profile between a reference panel and the study cohort in achieving high imputation accuracy. Diverse representation of population in the reference panel facilitates imputation of population not represented in the panel. GenomeAsia harbors a more diverse Asian populations that are genetically similar to the Thai. Hence, GenomeAsia outperformed the other 3 high-density reference panels in terms of imputation accuracy. We speculate that the diverse populations in the GenomeAsia reference panel would result in higher accuracy when using in the imputation of other understudied Asian populations.

Conclusion

The examinations of genome sequences in Thai population illustrated the distinct genetic variation found in Thais. When evaluated against currently available public databases, this study demonstrated that allele frequencies of many variants are unique to Thais. A considerable number of variants found in Thais are population-specifics and are absent form currently available database. A closer examination used fine-scale population structure analysis further revealed the heterogenicity in variant distribution within Thai population itself. Enrichment of several clinically significant variants were found in Thai subpopulation. This demonstrated that assessing prevalence of variants based on super-population label (East-asian) in currently available public database does not provide an accurate overview for many of the variants circulating in Thais.

WGS demonstrated to be a highly efficient platform and a powerful tool in examine population genetic variation. WGS can identify various types of pathogenic variants from point mutations and small insertion/deletions to large structural variation. While some genomic regions or type of variations are not accessible using the standard variant calling pipeline, when use in conjunction with specialised bioinformatic tools it was demonstrated to vastly improved these previously unidentifiable variants. WGS ability to access immense amount of information in a single methodology would reduced time and labor involved.

In summary, this study demonstrated that the knowledge of genetic variations in Thai population would benefit different fields of medical science from the design of genetic testing through to conducting genomic research. Despite a relatively small sample size large number of the variants identified are population-specifics, an increase in sample size would provide a better overview of low frequencies and rare variants within the population that often have clinical significance. This study findings stresses the importance of having Thai population genome database and the sequencing understudied population.

REFERENCES

1. Guo MH, Gregg AR. Estimating yields of prenatal carrier screening and implications for design of expanded carrier screening panels. Genet Med. 2019;21(9):1940-7.

2. Bowerman M, Becker CG, Yanez-Munoz RJ, Ning K, Wood MJA, Gillingwater TH, et al. Therapeutic strategies for spinal muscular atrophy: SMN and beyond. Dis Model Mech. 2017;10(8):943-54.

3. Watanabe K, Stringer S, Frei O, Umicevic Mirkov M, de Leeuw C, Polderman TJC, et al. A global overview of pleiotropy and genetic architecture in complex traits. Nat Genet. 2019;51(9):1339-48.

4. Ramos E, Doumatey A, Elkahloun AG, Shriner D, Huang H, Chen G, et al. Pharmacogenomics, ancestry and clinical decision making for global populations. Pharmacogenomics J. 2014;14(3):217-22.

5. Petrovic J, Pesic V, Lauschke VM. Frequencies of clinically important CYP2C19 and CYP2D6 alleles are graded across Europe. Eur J Hum Genet. 2020;28(1):88-94.

6. Solomon BD, Nguyen AD, Bear KA, Wolfsberg TG. Clinical genomic database. Proc Natl Acad Sci U S A. 2013;110(24):9851-5.

7. Makrythanasis P, Nelis M, Santoni FA, Guipponi M, Vannier A, Bena F, et al. Diagnostic exome sequencing to elucidate the genetic basis of likely recessive disorders in consanguineous families. Hum Mutat. 2014;35(10):1203-10.

8. Antonarakis SE. Carrier screening for recessive disorders. Nat Rev Genet. 2019;20(9):549-61.

9. Fridman H, Yntema HG, Magi R, Andreson R, Metspalu A, Mezzavila M, et al. The landscape of autosomal-recessive pathogenic variants in European populations reveals phenotype-specific effects. Am J Hum Genet. 2021;108(4):608-19.

10. Severe Covid GG, Ellinghaus D, Degenhardt F, Bujanda L, Buti M, Albillos A, et al. Genomewide Association Study of Severe Covid-19 with Respiratory Failure. N Engl J Med. 2020;383(16):1522-34.

11. Pairo-Castineira E, Clohisey S, Klaric L, Bretherick AD, Rawlik K, Pasko D, et al.

Genetic mechanisms of critical illness in COVID-19. Nature. 2021;591(7848):92-8.

12. Shelton JF, Shastri AJ, Ye C, Weldon CH, Filshtein-Sonmez T, Coker D, et al. Trans-ancestry analysis reveals genetic and nongenetic associations with COVID-19 susceptibility and severity. Nat Genet. 2021;53(6):801-8.

13. Das S, Abecasis GR, Browning BL. Genotype Imputation from Large Reference Panels. Annu Rev Genomics Hum Genet. 2018;19:73-96.

14. Huang L, Li Y, Singleton AB, Hardy JA, Abecasis G, Rosenberg NA, et al. Genotype-imputation accuracy across worldwide human populations. Am J Hum Genet. 2009;84(2):235-50.

15. Lin Y, Liu L, Yang S, Li Y, Lin D, Zhang X, et al. Genotype imputation for Han Chinese population using Haplotype Reference Consortium as reference. Hum Genet. 2018;137(6-7):431-6.

16. Thorn CF, Klein TE, Altman RB. PharmGKB: the pharmacogenetics and pharmacogenomics knowledge base. Methods Mol Biol. 2005;311:179-91.

17. Relling MV, Klein TE. CPIC: Clinical Pharmacogenetics Implementation Consortium of the Pharmacogenomics Research Network. Clin Pharmacol Ther. 2011;89(3):464-7.

18. Himes P, Kauffman TL, Muessig KR, Amendola LM, Berg JS, Dorschner MO, et al. Genome sequencing and carrier testing: decisions on categorization and whether to disclose results of carrier testing. Genet Med. 2017;19(7):803-8.

19. Hunter JE, Irving SA, Biesecker LG, Buchanan A, Jensen B, Lee K, et al. A standardized, evidence-based protocol to assess clinical actionability of genetic disorders associated with genomic variation. Genet Med. 2016;18(12):1258-68.

20. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, et al. ClinGen--the Clinical Genome Resource. N Engl J Med. 2015;372(23):2235-42.

21. Amendola LM, Jarvik GP, Leo MC, McLaughlin HM, Akkari Y, Amaral MD, et al. Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium. Am J Hum Genet. 2016;98(6):1067-76.

22. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus

recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405-24.

23. Li Q, Wang K. InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. Am J Hum Genet. 2017;100(2):267-80.

24. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.

25. Shah N, Hou YC, Yu HC, Sainger R, Caskey CT, Venter JC, et al. Identification of Misclassified ClinVar Variants via Disease Population Prevalence. Am J Hum Genet. 2018;102(4):609-19.

26. Pratt VM, Everts RE, Aggarwal P, Beyer BN, Broeckel U, Epstein-Baak R, et al. Characterization of 137 Genomic DNA Reference Materials for 28 Pharmacogenetic Genes: A GeT-RM Collaborative Project. J Mol Diagn. 2016;18(1):109-23.

27. Lee SB, Wheeler MM, Thummel KE, Nickerson DA. Calling Star Alleles With Stargazer in 28 Pharmacogenes With Whole Genome Sequences. Clin Pharmacol Ther. 2019;106(6):1328-37.

28. Trier C, Fournous G, Strand JM, Stray-Pedersen A, Pettersen RD, Rowe AD. Next-generation sequencing of newborn screening genes: the accuracy of short-read mapping. NPJ Genom Med. 2020;5(1):36.

29. Mandelker D, Schmidt RJ, Ankala A, McDonald Gibson K, Bowser M, Sharma H, et al. Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing. Genet Med. 2016;18(12):1282-9.

30. Hendrickson BC, Donohoe C, Akmaev VR, Sugarman EA, Labrousse P, Boguslavskiy L, et al. Differences in SMN1 allele frequencies among ethnic groups within North America. J Med Genet. 2009;46(9):641-4.

31. Sugarman EA, Nagan N, Zhu H, Akmaev VR, Zhou Z, Rohlfs EM, et al. Pan-ethnic carrier screening and prenatal diagnosis for spinal muscular atrophy: clinical laboratory analysis of >72,400 specimens. Eur J Hum Genet. 2012;20(1):27-32.

32. Dejsuphong D, Taweewongsounton A, Khemthong P, Chitphuk S, Stitchantrakul W, Sritara P, et al. Carrier frequency of spinal muscular atrophy in Thailand. Neurol Sci. 2019;40(8):1729-32.

33. Lefebvre S, Burglen L, Reboullet S, Clermont O, Burlet P, Viollet L, et al.

Identification and characterization of a spinal muscular atrophy-determining gene. Cell. 1995;80(1):155-65.

34. Lunn MR, Wang CH. Spinal muscular atrophy. Lancet. 2008;371(9630):2120-33.

35. Rochette CF, Gilbert N, Simard LR. SMN gene duplication and the emergence of the SMN2 gene occurred in distinct hominids: SMN2 is unique to Homo sapiens. Hum Genet. 2001;108(3):255-66.

36. Westemeyer M, Saucier J, Wallace J, Prins SA, Shetty A, Malhotra 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.

37. Chen X, Sanchis-Juan A, French CE, Connell AJ, Delon I, Kingsbury Z, et al. Spinal muscular atrophy diagnosis and carrier screening from genome sequencing data. Genet Med. 2020;22(5):945-53.

38. Galanello R, Cao A. Gene test review. Alpha-thalassemia. Genet Med. 2011;13(2):83-8.

39. Farashi S, Harteveld CL. Molecular basis of alpha-thalassemia. Blood Cells Mol Dis. 2018;70:43-53.

40. Traeger-Synodinos J, Harteveld CL. Advances in technologies for screening and diagnosis of hemoglobinopathies. Biomark Med. 2014;8(1):119-31.

41. Cao Y, Yin Ha S, So CC, For TM, Sze-Man Tang C, Zhang H, et al. NGS4THAL, a one-stop molecular diagnosis and carrier screening tool for thalassemia and other hemoglobinopathies by next-generation sequencing. J Mol Diagn. 2022.

42. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods. 2009;6(9):677-81.

43. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009;25(21):2865-71.

44. Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, Coe BP, et al. Copy number variation detection and genotyping from exome sequence data. Genome Res. 2012;22(8):1525-32.

45. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM,

et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65.

46. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.

47. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016;48(10):1279-83.

48. GenomeAsia KC. The GenomeAsia 100K Project enables genetic discoveries across Asia. Nature. 2019;576(7785):106-11.

49. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature. 2021;590(7845):290-9.

50. Shraga R, Yarnall S, Elango S, Manoharan A, Rodriguez SA, Bristow SL, et al. Evaluating genetic ancestry and self-reported ethnicity in the context of carrier screening. BMC Genet. 2017;18(1):99.

51. Belbin GM, Cullina S, Wenric S, Soper ER, Glicksberg BS, Torre D, et al. Toward a fine-scale population health monitoring system. Cell. 2021;184(8):2068-83 e11.

52. Menozzi P, Piazza A, Cavalli-Sforza L. Synthetic maps of human gene frequencies in Europeans. Science. 1978;201(4358):786-92.

53. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. PLoS Genet. 2012;8(1):e1002453.

54. Byrne RP, van Rheenen W, Project Min EALSGC, van den Berg LH, Veldink JH, McLaughlin RL. Dutch population structure across space, time and GWAS design. Nat Commun. 2020;11(1):4556.

55. Leslie S, Winney B, Hellenthal G, Davison D, Boumertit A, Day T, et al. The fine-scale genetic structure of the British population. Nature. 2015;519(7543):309-14.

56. Takeuchi F, Katsuya T, Kimura R, Nabika T, Isomura M, Ohkubo T, et al. The fine-scale genetic structure and evolution of the Japanese population. PLoS One. 2017;12(11):e0185487.

57. Kerminen S, Havulinna AS, Hellenthal G, Martin AR, Sarin AP, Perola M, et al. Fine-Scale Genetic Structure in Finland. G3 (Bethesda). 2017;7(10):3459-68.

58. Pankratov V, Montinaro F, Kushniarevich A, Hudjashov G, Jay F, Saag L, et al. Differences in local population history at the finest level: the case of the Estonian population. Eur J Hum Genet. 2020;28(11):1580-91.

59. Durbin R. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). Bioinformatics. 2014;30(9):1266-72.

60. Gaedigk A, Sangkuhl K, Whirl-Carrillo M, Klein T, Leeder JS. Prediction of CYP2D6 phenotype from genotype across world populations. Genet Med. 2017;19(1):69-76.

61. Kalman LV. Pharmacogenetic allele nomenclature: International workgroup recommendations for test result reporting. Clin Pharmacol Ther. 2016;99.

62. Fujikura K, Ingelman-Sundberg M, Lauschke VM. Genetic variation in the human cytochrome P450 supergene family. Pharmacogenet Genomics. 2015;25.

63. Hernandez W, Danahey K, Pei X, Yeo KJ, Leung E, Volchenboum SL, et al. Pharmacogenomic genotypes define genetic ancestry in patients and enable population-specific genomic implementation. Pharmacogenomics J. 2020;20(1):126-35.

64. Zhang H, De T, Zhong Y, Perera MA. The Advantages and Challenges of Diversity in Pharmacogenomics: Can Minority Populations Bring Us Closer to Implementation? Clin Pharmacol Ther. 2019;106(2):338-49.

65. Exner DV, Dries DL, Domanski MJ, Cohn JN. Lesser response to angiotensin-converting-enzyme inhibitor therapy in black as compared with white patients with left ventricular dysfunction. N Engl J Med. 2001;344(18):1351-7.

66. Ahn E, Park T. Analysis of population-specific pharmacogenomic variants using next-generation sequencing data. Sci Rep. 2017;7(1):8416.

67. Thorn CF, Klein TE, Altman RB. PharmGKB: The pharmacogenetics and pharmacogenomics knowledge base. Methods Mol Biol. 2005;311.

68. An HR, Wu XQ, Wang ZY, Zhang JX, Liang Y. NAT2 and CYP2E1 polymorphisms associated with antituberculosis drug-induced hepatotoxicity in Chinese patients. Clin Exp Pharmacol Physiol. 2012;39.

69. Chan SL. Association and clinical utility of NAT2 in the prediction of isoniazid-induced liver injury in Singaporean patients. PLoS ONE. 2017;12.

70. Verschuren JJ. Value of platelet pharmacogenetics in common clinical practice of patients with ST-segment elevation myocardial infarction. Int J Cardiol. 2013;167.

71. SLCO1B1 variants and statin-induced myopathy—A genomewide study. N Engl J Med. 2008;359.

72. Petrovic J, Pesic V, Lauschke VM. Frequencies of clinically important CYP2C19 and CYP2D6 alleles are graded across Europe. Eur J Hum Genet. 2020;28.

73. GenomeAsia KC. The GenomeAsia 100K project enables genetic discoveries across Asia. Nature. 2019;576.

74. Nelson MR. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science. 2012;337.

75. Rasmussen-Torvik LJ. Design and anticipated outcomes of the eMERGE-PGx project: A multicenter pilot for preemptive pharmacogenomics in electronic health record systems. Clin Pharmacol Ther. 2014;96.

76. Caspar SM, Schneider T, Meienberg J, Matyas G. Added value of clinical sequencing: WGS-based profiling of pharmacogenes. Int J Mol Sci. 2020.

77. Bush WS. Genetic variation among 82 pharmacogenes: The PGRNseq data from the eMERGE network. Clin Pharmacol Ther. 2016;100.

78. Ingelman-Sundberg M, Mkrtchian S, Zhou Y, Lauschke VM. Integrating rare genetic variants into pharmacogenetic drug response predictions. Hum Genomics. 2018.

79. Sangkuhl K. Pharmacogenomics clinical annotation tool (PharmCAT). Clin Pharmacol Ther. 2020;107.

80. Reisberg S. Translating genotype data of 44,000 biobank participants into clinical pharmacogenetic recommendations: Challenges and solutions. Genet Med. 2019;21.

81. Sadedin SP, Oshlack A. Bazam: a rapid method for read extraction and realignment of high-throughput sequencing data. Genome Biol. 2019;20(1):78.

82. Gennarelli M, Lucarelli M, Capon F, Pizzuti A, Merlini L, Angelini C, et al. Survival motor neuron gene transcript analysis in muscles from spinal muscular atrophy patients. Biochem Biophys Res Commun. 1995;213(1):342-8.

83. Shotelersuk V, Wichadakul D, Ngamphiw C, Srichomthong C, Phokaew C, Wilantho A, et al. The Thai reference exome (T-REx) variant database. Clin Genet. 2021;100(6):703-12.

84. Chai Y, Chen D, Sun L, Li L, Chen Y, Pang X, et al. The homozygous p.V37I variant of GJB2 is associated with diverse hearing phenotypes. Clin Genet.

2015;87(4):350-5.

85. Shen J, Oza AM, Del Castillo I, Duzkale H, Matsunaga T, Pandya A, et al. Consensus interpretation of the p.Met34Thr and p.Val37Ile variants in GJB2 by the ClinGen Hearing Loss Expert Panel. Genet Med. 2019;21(11):2442-52.

86. Nguyen VH, Sanchaisuriya K, Wongprachum K, Nguyen MD, Phan TT, Vo VT, et al. Hemoglobin Constant Spring is markedly high in women of an ethnic minority group in Vietnam: a community-based survey and hematologic features. Blood Cells Mol Dis. 2014;52(4):161-5.

87. Prajantasen T, Teawtrakul N, Fucharoen G, Fucharoen S. Molecular characterization of a beta-thalassemia intermedia patient presenting inferior vena cava thrombosis: interaction of the beta-globin erythroid Kruppel-like factor binding site mutation with Hb E and alpha(+)-thalassemia. Hemoglobin. 2014;38(6):451-3.

88. Wu CC, Tsai CH, Hung CC, Lin YH, Lin YH, Huang FL, et al. Newborn genetic screening for hearing impairment: a population-based longitudinal study. Genet Med. 2017;19(1):6-12.

89. Winichagoon P, Fucharoen S, Chen P, Wasi P. Genetic factors affecting clinical severity in beta-thalassemia syndromes. J Pediatr Hematol Oncol. 2000;22(6):573-80.

90. Yang S, Lincoln SE, Kobayashi Y, Nykamp K, Nussbaum RL, Topper S. Sources of discordance among germ-line variant classifications in ClinVar. Genet Med. 2017;19(10):1118-26.

91. Park S, Park JY, Kim GH, Choi JH, Kim KM, Kim JB, et al. Identification of novel ATP7B gene mutations and their functional roles in Korean patients with Wilson disease. Hum Mutat. 2007;28(11):1108-13.

92. Liu XQ, Zhang YF, Liu TT, Hsiao KJ, Zhang JM, Gu XF, et al. Correlation of ATP7B genotype with phenotype in Chinese patients with Wilson disease. World J Gastroenterol. 2004;10(4):590-3.

93. Dong Y, Ni W, Chen WJ, Wan B, Zhao GX, Shi ZQ, et al. Spectrum and Classification of ATP7B Variants in a Large Cohort of Chinese Patients with Wilson's Disease Guides Genetic Diagnosis. Theranostics. 2016;6(5):638-49.

94. Shotelersuk V, Tongsima S, Pithukpakorn M, Eu-Ahsunthornwattana J, Mahasirimongkol S. Precision medicine in Thailand. Am J Med Genet C Semin Med

Genet. 2019;181(2):245-53.

95. Fung JLF, Yu MHC, Huang S, Chung CCY, Chan MCY, Pajusalu S, et al. A three-year follow-up study evaluating clinical utility of exome sequencing and diagnostic potential of reanalysis. NPJ Genom Med. 2020;5(1):37.

96. Apidechkul T, Yeemard F, Chomchoei C, Upala P, Tamornpark R. Epidemiology of thalassemia among the hill tribe population in Thailand. PLoS One. 2021;16(2):e0246736.

97. Hockham C, Ekwattanakit S, Bhatt S, Penman BS, Gupta S, Viprakasit V, et al. Estimating the burden of alpha-thalassaemia in Thailand using a comprehensive prevalence database for Southeast Asia. Elife. 2019;8.

98. Tritipsombut J, Sanchaisuriya K, Phollarp P, Bouakhasith D, Sanchaisuriya P, Fucharoen G, et al. Micromapping of thalassemia and hemoglobinopathies in diferent regions of northeast Thailand and Vientiane, Laos People's Democratic Republic. Hemoglobin. 2012;36(1):47-56.

99. Munkongdee T, Tanakulmas J, Butthep P, Winichagoon P, Main B, Yiannakis M, et al. Molecular Epidemiology of Hemoglobinopathies in Cambodia. Hemoglobin. 2016;40(3):163-7.

100. Sengchanh S, Sanguansermsri T, Horst D, Horst J, Flatz G. High frequency of alpha-thalassemia in the So ethnic group of South Laos. Acta Haematol. 2005;114(3):164-6.

101. Jomoui W, Fucharoen G, Sanchaisuriya K, Charoenwijitkul P, Maneesarn J, Xu X, et al. Genetic origin of alpha(0)-thalassemia (SEA deletion) in Southeast Asian populations and application to accurate prenatal diagnosis of Hb Bart's hydrops fetalis syndrome. J Hum Genet. 2017;62(8):747-54.

102. Mauleekoonphairoj J, Chamnanphon M, Khongphatthanayothin A, Sutjaporn B, Wandee P, Poovorawan Y, et al. Phenotype prediction and characterization of 25 pharmacogenes in Thais from whole genome sequencing for clinical implementation. Sci Rep. 2020;10(1):18969.

103. Need AC, Goldstein DB. Next generation disparities in human genomics: concerns and remedies. Trends Genet. 2009;25(11):489-94.

104. Kessler MD, Yerges-Armstrong L, Taub MA, Shetty AC, Maloney K, Jeng LJB, et al.

Challenges and disparities in the application of personalized genomic medicine to populations with African ancestry. Nat Commun. 2016;7:12521.

105. Khoury MJ, Iademarco MF, Riley WT. Precision Public Health for the Era of Precision Medicine. Am J Prev Med. 2016;50(3):398-401.

106. Chaibunruang A, Sornkayasit K, Chewasateanchai M, Sanugul P, Fucharoen G, Fucharoen S. Prevalence of Thalassemia among Newborns: A Re-visited after 20 Years of a Prevention and Control Program in Northeast Thailand. Mediterr J Hematol Infect Dis. 2018;10(1):e2018054.

107. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19(9):1655-64.

108. Flannick J, Korn JM, Fontanillas P, Grant GB, Banks E, Depristo MA, et al. Efficiency and power as a function of sequence coverage, SNP array density, and imputation. PLoS Comput Biol. 2012;8(7):e1002604.

109. Nelson SC, Stilp AM, Papanicolaou GJ, Taylor KD, Rotter JI, Thornton TA, et al. Improved imputation accuracy in Hispanic/Latino populations with larger and more diverse reference panels: applications in the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). Hum Mol Genet. 2016;25(15):3245-54.

110. Vergara C, Parker MM, Franco L, Cho MH, Valencia-Duarte AV, Beaty TH, et al. Genotype imputation performance of three reference panels using African ancestry individuals. Hum Genet. 2018;137(4):281-92.

111. Jostins L, Morley KI, Barrett JC. Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets. Eur J Hum Genet. 2011;19(6):662-6.

112. Wangkumhang P, Shaw PJ, Chaichoompu K, Ngamphiw C, Assawamakin A, Nuinoon M, et al. Insight into the peopling of Mainland Southeast Asia from Thai population genetic structure. PLoS One. 2013;8(11):e79522.

113. Deelen P, Menelaou A, van Leeuwen EM, Kanterakis A, van Dijk F, Medina-Gomez C, et al. Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'. Eur J Hum Genet. 2014;22(11):1321-6.

114. Huang J, Howie B, McCarthy S, Memari Y, Walter K, Min JL, et al. Improved

imputation of low-frequency and rare variants using the UK10K haplotype reference panel. Nat Commun. 2015;6:8111.

115. Mitt M, Kals M, Parn K, Gabriel SB, Lander ES, Palotie A, et al. Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. Eur J Hum Genet. 2017;25(7):869-76.

116. de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. Efficiency and power in genetic association studies. Nat Genet. 2005;37(11):1217-23.

117. Mahajan A, Taliun D, Thurner M, Robertson NR, Torres JM, Rayner NW, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat Genet. 2018;50(11):1505-13.

118. Kowalski MH, Qian H, Hou Z, Rosen JD, Tapia AL, Shan Y, et al. Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. PLoS Genet. 2019;15(12):e1008500.

119. Rubinacci S, Ribeiro DM, Hofmeister RJ, Delaneau O. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. Nat Genet. 2021;53(1):120-6.

120. Van Hout CV, Tachmazidou I, Backman JD, Hoffman JD, Liu D, Pandey AK, et al. Exome sequencing and characterization of 49,960 individuals in the UK Biobank. Nature. 2020;586(7831):749-56.

121. Chou WC, Zheng HF, Cheng CH, Yan H, Wang L, Han F, et al. A combined reference panel from the 1000 Genomes and UK10K projects improved rare variant imputation in European and Chinese samples. Sci Rep. 2016;6:39313.

122. Halldorsson BV, Eggertsson HP, Moore KHS, Hauswedell H, Eiriksson O, Ulfarsson MO, et al. The sequences of 150,119 genomes in the UK Biobank. Nature. 2022;607(7920):732-40.

123. Pistis G, Porcu E, Vrieze SI, Sidore C, Steri M, Danjou F, et al. Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs. Eur J Hum Genet. 2015;23(7):975-83.

Supplementary

Supplementary table 1: Likely pathogenic/pathogenic variants detectedt in autosomal recessive genes categorized according to their level of evidence for pathogenicity.

Variants grouped into P1, P2, P3 and CoP, with P1 having the highest evidence for pathogenicity.

Variant number	GROUP	VARIANTS
1	P1	HBB:NM_000518:exon1:c.G79A:p.E27K
2	P1	GJB2:NM_004004:exon2:c.G109A:p.V37I
3	P1	HBA2:NM_000517:exon3:c.T427C:p.X143Q
4	P1	GALT:NM_000155.4:c.-119_-116delGTCA
5	P1	ABCA4:NM_000350:exon42:c.G5881A:p.G1961R
6	P1	SLC22A5:NM_001308122:exon1:c.C51G:p.F17L,SLC22A5:NM_003060:exon1:c.C51G:p.F17L
7	P1	HBA2:NM_000517:exon3:c.A429T:p.X143Y
8	P1	NM_016038:exon2:c.258+2T>C
9	P1	SLC26A4:NM_000441:exon14:c.1546dupC:p.S517Ffs*10
10	P1	HBB:NM_000518:exon2:c.126_129del:p.F42Lfs*19
11	P1	HBB:NM_000518.5:c.-78A>G
12	P1	BEST1:NM_001139443:exon4:c.C404T:p.A135V,BEST1:NM_001300786:exon4:c.C404T:p.A135V,BEST1:NM_001300787:exon4:c.C404T:p.A135V,BEST1:NM_004183:exon5:c.C584T:p.A195V
13	P1	GJB2:NM_004004:exon2:c.235delC:p.L79Cfs*3
14	P1	RPGRIP1L:NM_001330538:exon22:c.3198_3199insTC:p.A1067Sfs*34,RPGRIP1L:NM_015272:exon23:c.3300_3301insTC:p.A1101Sfs*34
15	P1	AGXT:NM_000030:exon1:c.T2C:p.M1?
16	P1	NM_206933:exon27:c.5572+1G>A
17	P1	PKHD1:NM_138694:exon24:c.T2507C:p.V836A,PKHD1:NM_170724:exon24:c.T2507C:p.V836A
18	P1	CYP21A2:NM_001128590:exon6:c.G754T:p.V252L,CYP21A2:NM_000500:exon7:c.G844T:p.V282L
19	P1	ABCA4:NM_000350:exon11:c.C1531T:p.R511C
20	P1	GBA:NM_001171811:exon5:c.A419G:p.N140S,GBA:NM_001171812:exon5:c.A533G:p.N178S,GBA:NM_000157:exon6:c.A680G:p.N227S,GBA:NM_001005741:exon7:c.A680G:p.N227S,GBA:NM_001005742:exon7:c.A680G:p.N227S
21	P1	UROS:NM_000375:exon4:c.T217C:p.C73R,UROS:NM_001324036:exon4:c.T217C:p.C73R,UROS:NM_001324037:exon4:c.T217C:p.C73R,UROS:NM_001324038:exon4:c.T217C:p.C73R,UROS:NM_001324039:exon4:c.T217C:p.C73R
22	P1	DHCR7:NM_001163817:exon7:c.G725A:p.R242H,DHCR7:NM_001360:exon7:c.G725A:p.R242H
23	P1	PAH:NM_000277:exon3:c.284_286del:p.I95del
24	P1	FANCA:NM_000135.4:c.709+5G>A
25	P1	GAA:NM_000152:exon14:c.C1935A:p.D645E,GAA:NM_001079804:exon14:c.C1935A:p.D645E,GAA:NM_001079803:exon15:c.C1935A:p.D645E
26	P1	SLC25A13:NM_001160210:exon16:c.1663_1664insGAGATTACAGGTGGCTGCCCGGG:p.A555Gfs*17,SLC25A13:NM_014251:exon16:c.1660_1661insGAGATTACAGGTGGCTGCCCGGG:p.A554Gfs*17
27	P1	SLC25A13:NM_001160210:exon10:c.C958T:p.R320X,SLC25A13:NM_014251:exon10:c.C955T:p.R319X
28	P1	SLC25A13:NM_001160210:exon9:c.852_855del:p.M285Pfs*2,SLC25A13:NM_014251:exon9:c.852_855del:p.M285Pfs*2

29	P1	NM_000492:exon11:c.1393-1G>A
30	P1	TYR:NM_000372:exon2:c.G896A:p.R299H
31	P1	CFTR:NM_000492:exon20:c.G3197A:p.R1066H
32	P1	NM_138694:exon42:c.6809-2A>T;NM_170724:exon42:c.6809-2A>T
33	P1	LYST:NM_000081:exon6:c.C3310T:p.R1104X,LYST:NM_001301365:exon6:c.C3310T:p.R1104X
34	P1	USH2A:NM_007123:exon13:c.C2209T:p.R737X,USH2A:NM_206933:exon13:c.C2209T:p.R737X
35	P1	ALMS1:NM_015120:exon16:c.C11413T:p.R3805X
36	P1	NM_000441:exon8:c.919-2A>G
37	P1	SLC26A4:NM_000441:exon10:c.C1229T:p.T410M
38	P1	GBA:NM_001171811:exon4:c.C214T:p.R72W,GBA:NM_001171812:exon4:c.C328T:p.R110W,GBA:NM_000157:exon5:c.C475T:p.R159W,GBA:NM_001005741:exon6:c.C475T:p.R159W,GBA:NM_001005742:exon6:c.C475T:p.R159W
39	P1	CEP290:NM_025114:exon50:c.6869dupA:p.N2290Kfs*6
40	P1	CYP21A2:NM_001128590:exon3:c.T428A:p.I143N,CYP21A2:NM_000500:exon4:c.T518A:p.I173N
41	P1	PAH:NM_000277:exon8:c.G890A:p.R297H
42	P1	AGXT:NM_000030:exon1:c.26dupC:p.K12Qfs*156
43	P1	GUSB:NM_001284290:exon5:c.C631T:p.R211X,GUSB:NM_001293105:exon5:c.C412T:p.R138X,GUSB:NM_001293104:exon6:c.C499T:p.R167X,GUSB:NM_000181:exon7:c.C1069T:p.R357X
44	P1	CFTR:NM_000492:exon13:c.G1753T:p.E585X
45	P1	MPL:NM_005373:exon3:c.235_236del:p.L79Efs*84
46	P1	MUTYH:NM_001350650:exon5:c.G38A:p.W13X,MUTYH:NM_001350651:exon5:c.G38A:p.W13X,MUTYH:NM_001048171:exon6:c.G425A:p.W142X,MUTYH:NM_001048172:exon6:c.G386A:p.W129X,MUTYH:NM_001048173:exon6:c.G383A:p.W128X,MUTYH:NM_001048174:exon6:c.G383A:p.W128X,MUTYH:NM_001128425:exon6:c.G467A:p.W156X,MUTYH:NM_001293190:exon6:c.G428A:p.W143X,MUTYH:NM_001293191:exon6:c.G416A:p.W139X,MUTYH:NM_001293192:exon6:c.G107A:p.W36X,MUTYH:NM_001293196:exon6:c.G107A:p.W36X,MUTYH:NM_012222:exon6:c.G458A:p.W153X,MUTYH:NM_001293195:exon7:c.G383A:p.W128X
47	P1	ABCA4:NM_000350:exon46:c.C6316T:p.R2106C
48	P1	ABCA4:NM_000350:exon29:c.G4328A:p.R1443H
49	P1	ABCA4:NM_000350:exon22:c.C3292T:p.R1098C
50	P1	ABCA4:NM_000350:exon21:c.C3056T:p.T1019M
51	P1	NM_001171812:exon2:c.115+1G>A;NM_000157:exon2:c.115+1G>A;NM_001005742:exon3:c.115+1G>A;NM_001005741:exon3:c.115+1G>A
52	P1	NPHS2:NM_001297575:exon6:c.C667T:p.R223W,NPHS2:NM_014625:exon7:c.C871T:p.R291W
53	P1	LAMB3:NM_001017402:exon15:c.2346delC:p.T783Pfs*48,LAMB3:NM_000228:exon16:c.2346delC:p.T783Pfs*48,LAMB3:NM_001127641:exon16:c.2346delC:p.T783Pfs*48
54	P1	USH2A:NM_206933:exon63:c.C13576T:p.R4526X
55	P1	USH2A:NM_206933:exon63:c.13112_13115del:p.Q4371Rfs*19
56	P1	USH2A:NM_206933:exon63:c.C13010T:p.T4337M
57	P1	PRF1:NM_001083116:exon2:c.C160T:p.R54C,PRF1:NM_005041:exon2:c.C160T:p.R54C
58	P1	CYP17A1:NM_000102:exon8:c.1459_1467del:p.D487_F489del
59	P1	ABCC8:NM_000352:exon23:c.C2797T:p.R933X,ABCC8:NM_001287174:exon23:c.C2800T:p.R934X,ABCC8:NM_001351295:exon23:c.C2863T:p.R955X,ABCC8:NM_001351296:exon23:c.C2797T:p.R933X,ABCC8:NM_001351297:exon23:c.C2794T:p.R932X
60	P1	PYGM:NM_001164716:exon12:c.C1462T:p.R488X,PYGM:NM_005609:exon14:c.C1726T:p.R576X
61	P1	GYS2:NM_021957:exon5:c.C736T:p.R246X
62	P1	GNPTAB:NM_024312:exon19:c.C3565T:p.R1189X
63	P1	GNPTAB:NM_024312:exon13:c.2550_2554del:p.K850Nfs*10
64	P1	MMAB:NM_052845:exon7:c.577_578insTGTGCCGCCGGGCCG:p.A192_E193insVCRRA

65	P1	ACADS:NM_000017:exon9:c.A1031G:p.E344G,ACADS:NM_001302554:exon9:c.A1019G:p.E340G
66	P1	GJB2:NM_004004:exon2:c.299_300del:p.H100Rfs*14
67	P1	GJB2:NM_004004:exon2:c.G71A:p.W24X
68	P1	BRCA2:NM_000059:exon15:c.C7558T:p.R2520X
69	P1	ATP7B:NM_001005918:exon12:c.T2822C:p.I941T,ATP7B:NM_001330579:exon14:c.T3191C:p.I1064T,ATP7B:NM_001330578:exon15:c.T3209C:p.I1070T,ATP7B:NM_000053:exon16:c.T3443C:p.I1148T,ATP7B:NM_001243182:exon17:c.T3110C:p.I1037T
70	P1	RDH12:NM_152443:exon4:c.C164T:p.T55M
71	P1	GALC:NM_000153:exon1:c.G136T:p.D46Y,GALC:NM_001201401:exon1:c.G136T:p.D46Y
72	P1	FAH:NM_000137:exon9:c.C782T:p.P261L
73	P1	NM_144672:exon17:c.1880+1G>A;NM_001161683:exon13:c.1643+1G>A;NM_170664:exon8:c.908+1G>A
74	P1	BBS2:NM_031885:exon17:c.C2107T:p.R703X
75	P1	CNGB1:NM_001286130:exon26:c.2526dupG:p.L843Afs*3,CNGB1:NM_001297:exon26:c.2544dupG:p.L849Afs*3
76	P1	NM_001195798:exon5:c.695-1G>A;NM_000527:exon5:c.695-1G>A;NM_001195803:exon4:c.314-1G>A;NM_001195799:exon4:c.572-1G>A
77	P1	CYP27A1:NM_000784:exon6:c.C1072T:p.Q358X
78	P1	COL4A3:NM_000091:exon21:c.C1216T:p.R406X
79	P1	COL4A3:NM_000091:exon48:c.4344_4350del:p.R1450Vfs*77
80	P1	NM_000383:exon5:c.652+1G>T
81	P1	NM_001848:exon23:c.1575+1G>A
82	P1	PLA2G6:NM_001004426:exon13:c.C1741T:p.R581X,PLA2G6:NM_001199562:exon13:c.C1741T:p.R581X,PLA2G6:NM_001349865:exon13:c.C1741T:p.R581X,PLA2G6:NM_001349866:exon13:c.C1741T:p.R581X,PLA2G6:NM_001349868:exon13:c.C1225T:p.R409X,PLA2G6:NM_001349864:exon14:c.C1903T:p.R635X,PLA2G6:NM_001349869:exon14:c.C1207T:p.R403X,PLA2G6:NM_003560:exon14:c.C1903T:p.R635X,PLA2G6:NM_001349867:exon15:c.C1369T:p.R457X
83	P1	PLA2G6:NM_001004426:exon11:c.G1451A:p.R484H,PLA2G6:NM_001199562:exon11:c.G1451A:p.R484H,PLA2G6:NM_001349865:exon11:c.G1451A:p.R484H,PLA2G6:NM_001349866:exon11:c.G1451A:p.R484H,PLA2G6:NM_001349868:exon11:c.G935A:p.R312H,PLA2G6:NM_001349864:exon12:c.G1613A:p.R538H,PLA2G6:NM_001349869:exon12:c.G917A:p.R306H,PLA2G6:NM_003560:exon12:c.G1613A:p.R538H,PLA2G6:NM_001349867:exon13:c.G1079A:p.R360H
84	P1	ARSA:NM_000487:exon8:c.1344dupC:p.G449Rfs*124,ARSA:NM_001085428:exon8:c.1086dupC:p.G363Rfs*124,ARSA:NM_001085425:exon9:c.1344dupC:p.G449Rfs*124,ARSA:NM_001085426:exon9:c.1344dupC:p.G449Rfs*124,ARSA:NM_001085427:exon9:c.1344dupC:p.G449Rfs*124
85	P1	KLHL40:NM_152393:exon4:c.A1516C:p.T506P
86	P1	ACAD9:NM_014049:exon12:c.G1237A:p.E413K
87	P1	SLC26A1:NM_022042:exon2:c.C554T:p.T185M,SLC26A1:NM_134425:exon2:c.C554T:p.T185M,SLC26A1:NM_213613:exon3:c.C554T:p.T185M
88	P1	SLC22A5:NM_003060:exon8:c.C1400G:p.S467C,SLC22A5:NM_001308122:exon9:c.C1472G:p.S491C
89	P1	SLC22A5:NM_003060:exon8:c.G1412A:p.R471H,SLC22A5:NM_001308122:exon9:c.G1484A:p.R495H
90	P1	PEX7:NM_000288:exon7:c.G649A:p.G217R
91	P1	GUSB:NM_000181:exon3:c.C526T:p.L176F
92	P1	POR:NM_000941:exon12:c.G1370A:p.R457H
93	P1	SLC26A4:NM_000441:exon18:c.C2086T:p.Q696X
94	P1	SLC26A4:NM_000441:exon19:c.A2168G:p.H723R
95	P1	CFTR:NM_000492:exon14:c.G1865A:p.G622D
96	P1	CFTR:NM_000492:exon14:c.C2125T:p.R709X
97	P1	CNGB3:NM_019098:exon16:c.C1810T:p.R604X
98	P1	GNE:NM_001190388:exon3:c.G722A:p.R241Q,GNE:NM_001128227:exon4:c.G830A:p.R277Q,GNE:NM_001190383:exon4:c.G737A:p.R246Q,GNE:NM_005476:exon4:c.G737A:p.R246Q
99	P1	FBP1:NM_000507:exon7:c.960_961insG:p.S321Vfs*13,FBP1:NM_001127628:exon8:c.960_961insG:p.S321Vfs*13

100	P1	ASS1:NM_054012:exon13:c.C1087T:p.R363W,ASS1:NM_000050:exon14:c.C1087T:p.R363W
101	P2	NEB:NM_004543:exon143:c.19106_19127del:p.T6369Rfs*36,NEB:NM_001164507:exon176:c.24710_24731del:p.T8237Rfs*36,NEB:NM_001164508:exon176:c.24710_24731del:p.T8237Rfs*36,NEB:NM_001271208:exon177:c.24815_24836del:p.T8272Rfs*36
102	P2	GJC2:NM_020435:exon2:c.C1199A:p.A400E
103	P2	HPS6:NM_024747:exon1:c.155delT:p.V52Efs*6
104	P2	CHKB:NM_005198:exon5:c.598delC:p.Q200Rfs*11
105	P2	NM_000102:exon1:c.297+2T>C
106	P2	MYO15A:NM_016239:exon2:c.3524dupA:p.S1176Vfs*14
107	P2	RPE65:NM_000329:exon14:c.C1543T:p.R515W
108	P2	ALMS1:NM_015120:exon8:c.G7399T:p.E2467X
109	P2	HBB:NM_000518:exon2:c.126delC:p.F43Lfs*19
110	P2	PAH:NM_000277:exon11:c.C1123G:p.Q375E
111	P2	ALMS1:NM_015120:exon16:c.11113_11131del:p.R3705Lfs*11
112	P2	NM_152388:exon6:c.529+1G>A;NM_001044385:exon6:c.553+1G>A
113	P2	FH:NM_000143:exon5:c.T653C:p.L218P
114	P2	CC2D2A:NM_001080522:exon35:c.C4407G:p.S1469R
115	P2	NM_031475:exon7:c.1464+1G>A
116	P2	NM_000478:exon9:c.997+1G>T;NM_001177520:exon7:c.766+1G>T;NM_001127501:exon8:c.832+1G>T
117	P2	FUCA1:NM_000147:exon2:c.T393A:p.Y131X
118	P2	LDLRAP1:NM_015627:exon1:c.65dupG:p.G25Rfs*9
119	P2	RPE65:NM_000329:exon4:c.G272A:p.R91Q
120	P2	ABCA4:NM_000350:exon40:c.G5646A:p.M1882I
121	P2	NM_000350:exon29:c.4352+1G>A
122	P2	GBA:NM_000157:exon3:c.203dupC:p.T69Dfs*12,GBA:NM_001171812:exon3:c.203dupC:p.T69Dfs*12,GBA:NM_001005741:exon4:c.203dupC:p.T69Dfs*12,GBA:NM_001005742:exon4:c.203dupC:p.T69Dfs*12
123	P2	USH2A:NM_206933:exon22:c.C4732T:p.R1578C
124	P2	MYO3A:NM_017433:exon30:c.3498delT:p.S1167Pfs*26
125	P2	ABCC8:NM_000352:exon33:c.G4051A:p.V1351M,ABCC8:NM_001287174:exon33:c.G4054A:p.V1352M,ABCC8:NM_001351295:exon33:c.G4117A:p.V1373M,ABCC8:NM_001351296:exon33:c.G4051A:p.V1351M,ABCC8:NM_001351297:exon33:c.G4048A:p.V1350M
126	P2	BEST1:NM_001139443:exon3:c.C241A:p.R81S,BEST1:NM_001300786:exon3:c.C241A:p.R81S,BEST1:NM_001300787:exon3:c.C241A:p.R81S,BEST1:NM_004183:exon4:c.C421A:p.R141S
127	P2	NDUFV1:NM_001166102:exon9:c.1175dupG:p.D394Gfs*27,NDUFV1:NM_007103:exon9:c.1202dupG:p.D403Gfs*27
128	P2	NM_025114:exon47:c.6358-1G>A
129	P2	NM_025114:exon5:c.251-2A>G
130	P2	NM_024312.5:c.637-6T>G
131	P2	PAH:NM_000277:exon6:c.G516T:p.Q172H
132	P2	BRCA2:NM_000059:exon11:c.G4531T:p.E1511X
133	P2	SLC25A15:NM_014252:exon4:c.407delC:p.M137Cfs*10
134	P2	ATP7B:NM_001005918:exon15:c.G3339C:p.R1113S,ATP7B:NM_001330579:exon17:c.G3708C:p.R1236S,ATP7B:NM_001330578:exon18:c.G3726C:p.R1242S,ATP7B:NM_000053:exon19:c.G3960C:p.R1320S,ATP7B:NM_001243182:exon20:c.G3627C:p.R1209S
135	P2	TGM1:NM_000359:exon6:c.C943T:p.R315C
136	P2	TGM1:NM_000359:exon3:c.A420G:p.I140M
137	P2	NM_001159508:exon4:c.376-2A>G;NM_002225:exon5:c.466-2A>G

138	P2	NM_001159508:exon5:c.470-1G>A;NM_002225:exon6:c.560-1G>A
139	P2	POLG:NM_001126131:exon21:c.C3412T:p.R1138C,POLG:NM_002693:exon21:c.C3412T:p.R1138C
140	P2	VPS33B:NM_001289148:exon3:c.161delT:p.L54Cfs*33,VPS33B:NM_018668:exon4:c.242delT:p.L81Cfs*33
141	P2	TK2:NM_001172644:exon3:c.C193T:p.R65C,TK2:NM_001172643:exon4:c.C175T:p.R59C,TK2:NM_001271935:exon4:c.C175T:p.R59C,TK2:NM_004614:exon4:c.C268T:p.R90C,TK2:NM_001271934:exon5:c.C121T:p.R41C
142	P2	FANCA:NM_000135:exon32:c.G3188A:p.W1063X,FANCA:NM_001286167:exon32:c.G3188A:p.W1063X
143	P2	ALOX12B:NM_001139:exon9:c.C1156T:p.R386C
144	P2	NM_016239:exon37:c.7396-1G>A
145	P2	GAA:NM_000152:exon7:c.G1129C:p.G377R,GAA:NM_001079804:exon7:c.G1129C:p.G377R,GAA:NM_001079803:exon8:c.G1129C:p.G377R
146	P2	GCDH:NM_000159:exon8:c.T797C:p.M266T,GCDH:NM_013976:exon8:c.T797C:p.M266T
147	P2	SLC7A9:NM_001126335:exon5:c.C511T:p.R171W,SLC7A9:NM_001243036:exon5:c.C511T:p.R171W,SLC7A9:NM_014270:exon5:c.C511T:p.R171W
148	P2	ETFB:NM_001014763:exon2:c.G505A:p.A169T,ETFB:NM_001985:exon3:c.G232A:p.A78T
149	P2	FAM161A:NM_032180:exon4:c.1635delA:p.E546Kfs*4,FAM161A:NM_001201543:exon5:c.1803delA:p.E602Kfs*4
150	P2	CNGA3:NM_001079878:exon7:c.G1723A:p.E575K,CNGA3:NM_001298:exon8:c.G1777A:p.E593K
151	P2	PROC:NM_000312:exon9:c.G1000A:p.G334S
152	P2	NM_001257343:exon7:c.889+1G>A;NM_001257342:exon7:c.889+1G>A;NM_001318836:exon5:c.529+1G>A;NM_004328:exon7:c.889+1G>A;NM_001257344:exon6:c.889+1G>A;NM_001320717:exon7:c.889+1G>A;NM_001079866:exon6:c.889+1G>A
153	P2	WNT10A:NM_025216:exon2:c.G311A:p.R104H
154	P2	NM_025216:exon2:c.376+1G>A
155	P2	COL6A3:NM_057164:exon3:c.C604T:p.R202X,COL6A3:NM_057166:exon3:c.C604T:p.R202X,COL6A3:NM_057165:exon4:c.C1207T:p.R403X,COL6A3:NM_057167:exon4:c.C1207T:p.R403X,COL6A3:NM_004369:exon5:c.C1825T:p.R609X
156	P2	AGXT:NM_000030:exon4:c.G481A:p.G161S
157	P2	MKKS:NM_018848:exon3:c.G862A:p.V288I,MKKS:NM_170784:exon3:c.G862A:p.V288I
158	P2	COL7A1:NM_000094:exon51:c.C4888T:p.R1630X
159	P2	NM_014049:exon15:c.1563+1G>A
160	P2	NM_001184:exon12:c.2533-1G>A
161	P2	HPS3:NM_032383:exon2:c.402delG:p.A135Pfs*10
162	P2	PDE6B:NM_001350155:exon9:c.C523T:p.R175C,PDE6B:NM_001145292:exon11:c.C841T:p.R281C,PDE6B:NM_001350154:exon11:c.C841T:p.R281C,PDE6B:NM_000283:exon13:c.C1678T:p.R560C,PDE6B:NM_001145291:exon13:c.C1678T:p.R560C
163	P2	EVC:NM_001306090:exon13:c.C1864T:p.R622X,EVC:NM_153717:exon13:c.C1864T:p.R622X
164	P2	PROM1:NM_001145851:exon10:c.T1211A:p.V404D,PROM1:NM_001145852:exon10:c.T1211A:p.V404D,PROM1:NM_001145847:exon11:c.T1211A:p.V404D,PROM1:NM_001145848:exon11:c.T1211A:p.V404D,PROM1:NM_001145849:exon11:c.T1238A:p.V413D,PROM1:NM_001145850:exon11:c.T1238A:p.V413D,PROM1:NM_006017:exon11:c.T1238A:p.V413D
165	P2	MTTP:NM_001300785:exon12:c.G1700A:p.R567H,MTTP:NM_000253:exon13:c.G1619A:p.R540H
166	P2	ETFDH:NM_001281738:exon3:c.G341A:p.R114H,ETFDH:NM_001281737:exon4:c.G383A:p.R128H,ETFDH:NM_004453:exon5:c.G524A:p.R175H
167	P2	ETFDH:NM_001281738:exon5:c.A587G:p.Y196C,ETFDH:NM_001281737:exon6:c.A629G:p.Y210C,ETFDH:NM_004453:exon7:c.A770G:p.Y257C
168	P2	SLC22A5:NM_001308122:exon1:c.C283G:p.L95V,SLC22A5:NM_003060:exon1:c.C283G:p.L95V
169	P2	RARS2:NM_001350505:exon1:c.T2G:p.M1?,RARS2:NM_020320:exon1:c.T2G:p.M1?
170	P2	NM_000426:exon1:c.112+2T>C;NM_001079823:exon1:c.112+2T>C
171	P2	GUSB:NM_000181:exon2:c.C328T:p.R110X,GUSB:NM_001284290:exon2:c.C328T:p.R110X
172	P2	SLC26A4:NM_000441:exon4:c.349delC:p.L117Sfs*9
173	P2	NM_153704:exon14:c.1413-2A>G;NM_001142301:exon15:c.1170-2A>G
174	P2	TMEM67:NM_153704:exon16:c.C1645T:p.R549C,TMEM67:NM_001142301:exon17:c.C1402T:p.R468C

175	P2	NM_004260:exon5:c.1131+1G>A
176	P2	RMRP:NR_003051.3:n.41G>A
177	P2	GNE:NM_001190384:exon3:c.C457T:p.R153X,GNE:NM_001190388:exon4:c.C772T:p.R258X,GNE:NM_001128227:exon5:c.C880T:p.R294X,GNE:NM_001190383:exon5:c.C787T:p.R263X,GNE:NM_005476:exon5:c.C787T:p.R263X
178	P2	VPS13A:NM_001018037:exon46:c.C6223T:p.R2075X,VPS13A:NM_001018038:exon47:c.C6340T:p.R2114X,VPS13A:NM_015186:exon47:c.C6340T:p.R2114X,VPS13A:NM_033305:exon47:c.C6340T:p.R2114X
179	P2	INVS:NM_001318382:exon15:c.C1909T:p.Q637X,INVS:NM_014425:exon15:c.C2887T:p.Q963X,INVS:NM_001318381:exon16:c.C2599T:p.Q867X
180	P2	POMT1:NM_001136114:exon13:c.1127dupA:p.Y376*,POMT1:NM_001077366:exon14:c.1316dupA:p.Y439*,POMT1:NM_001077365:exon15:c.1478dupA:p.Y493*,POMT1:NM_001136113:exon15:c.1478dupA:p.Y493*,POMT1:NM_007171:exon15:c.1544dupA:p.Y515*
181	P3	FANCA:NM_000135.4:c.710-142_710-141dup
182	P3	F5:NM_000130:exon7:c.A1000G:p.R334G
183	P3	CYP27A1:NM_000784:exon8:c.G1415C:p.G472A
184	P3	FREM2:NM_207361:exon6:c.G5920A:p.E1974K
185	P3	OTOF:NM_001287489:exon13:c.C1273T:p.R425X,OTOF:NM_194248:exon13:c.C1273T:p.R425X
186	P3	CFTR:NM_000492:exon20:c.G3267A:p.W1089X
187	P3	LDLR:NM_001195800:exon15:c.G2026A:p.G676S,LDLR:NM_001195803:exon15:c.G1996A:p.G666S,LDLR:NM_001195799:exon16:c.G2407A:p.G803S,LDLR:NM_000527:exon17:c.G2530A:p.G844S,LDLR:NM_001195798:exon17:c.G2530A:p.G844S
188	P3	GJB3:NM_001005752:exon2:c.421_423del:p.I141del,GJB3:NM_024009:exon2:c.421_423del:p.I141del
189	P3	ABCA4:NM_000350:exon6:c.C763T:p.R255C
190	P3	SCNN1A:NM_001159576:exon10:c.C1699T:p.R567X,SCNN1A:NM_001038:exon11:c.C1522T:p.R508X,SCNN1A:NM_001159575:exon11:c.C1591T:p.R531X
191	P3	RPGRIP1:NM_020366:exon5:c.C799T:p.R267X
192	P3	ZNF469:NM_001127464:exon1:c.C290T:p.P97L
193	P3	USH1G:NM_001282489:exon2:c.G784A:p.D262N,USH1G:NM_173477:exon2:c.G1093A:p.D365N
194	P3	FAM161A:NM_001201543:exon3:c.A943T:p.K315X,FAM161A:NM_032180:exon3:c.A943T:p.K315X
195	P3	CNGA3:NM_001079878:exon7:c.G1714A:p.E572K,CNGA3:NM_001298:exon8:c.G1768A:p.E590K
196	P3	CHRNG:NM_005199:exon2:c.C136T:p.R46X
197	P3	AGXT:NM_000030:exon1:c.G22C:p.V8L
198	P3	AGXT:NM_000030:exon2:c.G175A:p.E59K
199	P3	BTD:NM_000060:exon4:c.G1369A:p.V457M,BTD:NM_001281723:exon4:c.G1375A:p.V459M,BTD:NM_001281725:exon4:c.G1309A:p.V437M,BTD:NM_001323582:exon5:c.G1309A:p.V437M,BTD:NM_001281724:exon6:c.G1375A:p.V459M
200	P3	CRTAP:NM_006371:exon1:c.G3A:p.M1?
201	P3	KLHL40:NM_152393:exon5:c.G1612C:p.A538P
202	P3	HPS3:NM_032383.5:c.2888-1612G>A
203	P3	PDE6B:NM_000283:exon1:c.G293A:p.R98H,PDE6B:NM_001145291:exon1:c.G293A:p.R98H
204	P3	EVC:NM_001306090:exon12:c.C1668G:p.Y556X,EVC:NM_153717:exon12:c.C1668G:p.Y556X
205	P3	PROM1:NM_001145849:exon1:c.139delC:p.H47Ifs*12,PROM1:NM_001145850:exon1:c.139delC:p.H47Ifs*12,PROM1:NM_001145851:exon1:c.139delC:p.H47Ifs*12,PROM1:NM_001145852:exon1:c.139delC:p.H47Ifs*12,PROM1:NM_006017:exon1:c.139delC:p.H47Ifs*12,PROM1:NM_001145847:exon2:c.139delC:p.H47Ifs*12,PROM1:NM_001145848:exon2:c.139delC:p.H47Ifs*12
206	P3	MOCS2:NM_176806:exon1:c.C16T:p.Q6X
207	P3	MAK:NM_001242957:exon6:c.G497A:p.R166H,MAK:NM_005906:exon6:c.G497A:p.R166H,MAK:NM_001242385:exon7:c.G497A:p.R166H
208	P3	CFTR:NM_000492:exon11:c.C1518G:p.I506M
209	P4	MLC1:NM_015166:exon2:c.G65A:p.R22Q,MLC1:NM_139202:exon2:c.G65A:p.R22Q
210	P4	LDLR:NM_001195800:exon10:c.G1217A:p.R406H,LDLR:NM_001195799:exon11:c.G1598A:p.R533H,LDLR:NM_001195803:exon11:c.G1340A:p.R447H,LDLR:NM_000527:exon12:c.G1721A:p.R574H,LDLR:NM_001195798:exon12:c.G1721A:p.R574H
211	P4	GCDH:NM_000159:exon11:c.G1144A:p.A382T,GCDH:NM_013976:exon11:c.G1144A:p.A382T

212	P4	ACADVL:NM_001033859:exon8:c.761_763del:p.E255del,ACADVL:NM_001270448:exon8:c.599_601del:p.E201del,ACADVL:NM_000018:exon9:c.827_829del:p.E277del,ACADVL:NM_001270447:exon10:c.896_898del:p.E300del
213	P4	ACADVL:NM_001033859:exon11:c.C1160T:p.T387M,ACADVL:NM_001270448:exon11:c.C998T:p.T333M,ACADVL:NM_000018:exon12:c.C1226T:p.T409M,ACADVL:NM_001270447:exon13:c.C1295T:p.T432M
214	P4	GAA:NM_000152:exon6:c.C971T:p.P324L,GAA:NM_001079804:exon6:c.C971T:p.P324L,GAA:NM_001079803:exon7:c.C971T:p.P324L
215	P4	SGSH:NM_000199:exon8:c.G1063A:p.E355K
216	P4	MEFV:NM_000243:exon10:c.G2282A:p.R761H
217	P4	NM_001297.5:c.2893-7G>A
218	P4	POLG:NM_001126131:exon20:c.C3139T:p.R1047W,POLG:NM_002693:exon20:c.C3139T:p.R1047W
219	P4	POLG:NM_001126131:exon10:c.G1790A:p.R597Q,POLG:NM_002693:exon10:c.G1790A:p.R597Q
220	P4	ATP7B:NM_001005918:exon11:c.G2695A:p.V899I,ATP7B:NM_001330579:exon13:c.G3064A:p.V1022I,ATP7B:NM_001330578:exon14:c.G3082A:p.V1028I,ATP7B:NM_000053:exon15:c.G3316A:p.V1106I,ATP7B:NM_001243182:exon16:c.G2983A:p.V995I
221	P4	ATP7B:NM_001330579:exon10:c.C2503T:p.R835W,ATP7B:NM_001330578:exon11:c.C2521T:p.R841W,ATP7B:NM_000053:exon12:c.C2755T:p.R919W,ATP7B:NM_001243182:exon13:c.C2422T:p.R808W
222	P4	ATP7B:NM_001005918:exon8:c.G2119A:p.G707R,ATP7B:NM_001330579:exon9:c.G2353A:p.G785R,ATP7B:NM_001330578:exon10:c.G2371A:p.G791R,ATP7B:NM_000053:exon11:c.G2605A:p.G869R,ATP7B:NM_001243182:exon12:c.G2272A:p.G758R
223	P4	SMPD1:NM_000543:exon2:c.C995G:p.P332R,SMPD1:NM_001007593:exon2:c.C992G:p.P331R,SMPD1:NM_001318087:exon2:c.C995G:p.P332R,SMPD1:NM_001318088:exon2:c.C34G:p.P12A
224	P4	RAPSN:NM_005055:exon2:c.C264A:p.N88K,RAPSN:NM_032645:exon2:c.C264A:p.N88K
225	P4	GLDC:NM_000170:exon25:c.A2938G:p.N980D
226	P4	GNE:NM_001128227:exon1:c.T18A:p.Y6X
227	P4	ASL:NM_001024943:exon6:c.C467T:p.P156L,ASL:NM_001024944:exon6:c.C467T:p.P156L,ASL:NM_001024946:exon6:c.C467T:p.P156L,ASL:NM_000048:exon7:c.C467T:p.P156L
228	P4	PEX1:NM_001282677:exon18:c.T2795C:p.I932T,PEX1:NM_000466:exon19:c.T2966C:p.I989T,PEX1:NM_001282678:exon19:c.T2342C:p.I781T
229	P4	NM_000492:c.-34C>T
230	P4	CFTR:NM_000492:exon10:c.C1364T:p.A455V
231	P4	CFTR:NM_000492:exon20:c.G3205A:p.G1069R
232	P4	FARS2:NM_001318872:exon2:c.C467T:p.T156M,FARS2:NM_006567:exon2:c.C467T:p.T156M
233	P4	HFE:NM_139010:exon2:c.G305A:p.C102Y,HFE:NM_139003:exon3:c.G527A:p.C176Y,HFE:NM_139004:exon3:c.G569A:p.C190Y,HFE:NM_139007:exon3:c.G581A:p.C194Y,HFE:NM_139008:exon3:c.G539A:p.C180Y,HFE:NM_000410:exon4:c.G845A:p.C282Y,HFE:NM_001300749:exon4:c.G845A:p.C282Y,HFE:NM_139006:exon4:c.G803A:p.C268Y,HFE:NM_139009:exon4:c.G776A:p.C259Y
234	P4	CYP21A2:NM_001128590:exon8:c.G1084A:p.A362T,CYP21A2:NM_000500:exon9:c.G1174A:p.A392T
235	P4	MOCS1:NM_005943:exon2:c.C394T:p.R132W,MOCS1:NM_001075098:exon3:c.C394T:p.R132W
236	P4	PKHD1:NM_138694:exon46:c.T7280C:p.I2427T,PKHD1:NM_170724:exon46:c.T7280C:p.I2427T
237	P4	EYS:NM_001142800:exon31:c.G6416A:p.C2139Y,EYS:NM_001292009:exon31:c.G6416A:p.C2139Y
238	P4	SLC22A5:NM_003060:exon3:c.C641T:p.A214V,SLC22A5:NM_001308122:exon4:c.C713T:p.A238V
239	P4	BTD:NM_000060:exon4:c.A968G:p.H323R,BTD:NM_001281723:exon4:c.A974G:p.H325R,BTD:NM_001281725:exon4:c.A908G:p.H303R,BTD:NM_001323582:exon5:c.A908G:p.H303R,BTD:NM_001281724:exon6:c.A974G:p.H325R
240	P4	BTD:NM_000060:exon4:c.G1330C:p.D444H,BTD:NM_001281723:exon4:c.G1336C:p.D446H,BTD:NM_001281725:exon4:c.G1270C:p.D424H,BTD:NM_001323582:exon5:c.G1270C:p.D424H,BTD:NM_001281724:exon6:c.G1336C:p.D446H
241	P4	ILDR1:NM_001199800:exon4:c.C505T:p.Q169X,ILDR1:NM_001199799:exon6:c.C772T:p.Q258X
242	P4	OTOF:NM_194322:exon22:c.G3028C:p.E1010Q,OTOF:NM_004802:exon23:c.G2797C:p.E933Q,OTOF:NM_194323:exon23:c.G2797C:p.E933Q,OTOF:NM_001287489:exon40:c.G5098C:p.E1700Q,OTOF:NM_194248:exon40:c.G5098C:p.E1700Q
243	P4	LRPPRC:NM_133259:exon37:c.4128delT:p.E1377Kfs*10
244	P4	PROC:NM_000312:exon7:c.572_574del:p.K193del
245	P4	TTN:NM_001267550.2:c.55432+5G>C
246	P4	NM_133378:exon11:c.1800+1G>A;NM_001267550:exon11:c.1800+1G>A;NM_001256850:exon11:c.1800+1G>A;NM_133379:exon11:c.1800+1G>A
247	P4	ACADM:NM_001286044:exon4:c.A13G:p.N5D,ACADM:NM_001286042:exon6:c.A472G:p.N158D,ACADM:NM_000016:exon7:c.A580G:p.

		N194D,ACADM:NM_001127328:exon7:c.A592G:p.N198D,ACADM:NM_001286043:exon8:c.A679G:p.N227D
248	P4	ACADM:NM_001286044:exon9:c.T680C:p.I227T,ACADM:NM_001286042:exon11:c.T1139C:p.I380T,ACADM:NM_000016:exon12:c.T1247C:p.I416T,ACADM:NM_001127328:exon12:c.T1259C:p.I420T,ACADM:NM_001286043:exon13:c.T1346C:p.I449T
249	P4	ABCA4:NM_000350:exon44:c.G6119A:p.R2040Q
250	P4	ABCA4:NM_000350:exon42:c.G5882A:p.G1961E
251	P4	ABCA4:NM_000350:exon36:c.G5077A:p.V1693I
252	P4	ABCA4:NM_000350:exon33:c.T4685C:p.I1562T
253	P4	ABCA4:NM_000350:exon31:c.C4610T:p.T1537M
254	P4	ABCA4:NM_000350:exon29:c.G4297A:p.V1433I
255	P4	ABCA4:NM_000350:exon19:c.C2827T:p.R943W
256	P4	ABCA4:NM_000350:exon12:c.G1715A:p.R572Q
257	P4	DPYD:NM_000110:exon3:c.C220T:p.R74X,DPYD:NM_001160301:exon3:c.C220T:p.R74X
258	P4	AGL:NM_000028:exon33:c.C4459T:p.R1487X,AGL:NM_000642:exon33:c.C4459T:p.R1487X,AGL:NM_000643:exon33:c.C4459T:p.R1487X,AGL:NM_000644:exon33:c.C4459T:p.R1487X,AGL:NM_000646:exon33:c.C4411T:p.R1471X
259	P4	AMPD1:NM_001172626:exon9:c.G1361A:p.R454H,AMPD1:NM_000036:exon10:c.G1373A:p.R458H
260	P4	AMPD1:NM_001172626:exon6:c.A947T:p.K316I,AMPD1:NM_000036:exon7:c.A959T:p.K320I
261	P4	NPHS2:NM_014625:exon5:c.G686A:p.R229Q
262	P4	USH2A:NM_206933:exon63:c.A13339G:p.M4447V
263	P4	USH2A:NM_007123:exon13:c.T2802G:p.C934W,USH2A:NM_206933:exon13:c.T2802G:p.C934W

Supplementary table 2 Variant carrier rate (VCR), genome coordinate and consequence of likely pathogenic/pathogenic variants detected in autosomal recessive genes.

Variant #	GROUP	GENE NAME	VCR	Genome coordinate on GRCh38					CONSEQUENCE
				CHROM	POS	RS_ID	REF	ALT	
1	P1	*HBB*	0.256622517	chr11	5226943	rs33950507	C	T	nonsynonymous SNV
2	P1	*GJB2*	0.215588723	chr13	20189473	rs72474224	C	T	nonsynonymous SNV
3	P1	*HBA2*	0.054700855	chr16	173598	rs41464951	T	C	stoploss
4	P1	*GALT*	0.016556291	chr9	34646575	rs111033640	CCAGT	C	upstream
5	P1	*ABCA4*	0.006611057	chr1	94008252	rs142253670	C	T	nonsynonymous SNV
6	P1	*SLC22A5*	0.006611057	chr5	132370023	rs115685200	C	G	nonsynonymous SNV
7	P1	*HBA2*	0.005102041	chr16	173600	rs41412046	A	T	stoploss
8	P1	*SBDS*	0.004975124	chr7	66994210	rs113993993	A	G	splicing
9	P1	*SLC26A4*	0.004975124	chr7	107698042	rs786204450	T	TC	frameshift insertion
10	P1	*HBB*	0.004958678	chr11	5226762	rs80356821	CAAAG	C	frameshift deletion
11	P1	*HBB*	0.004958678	chr11	5227099	rs33931746	T	C	upstream
12	P1	*BEST1*	0.004958678	chr11	61956946	rs200277476	C	T	nonsynonymous SNV

13	P1	*GJB2*	0.004958678	chr13	20189346	rs80338943	AG	A	frameshift deletion
14	P1	*RPGRIP1L*	0.004958678	chr16	53622350	rs797045104	C	CGA	frameshift insertion
15	P1	*AGXT*	0.004958678	chr2	240868867	rs138584408	T	C	startloss
16	P1	*USH2A*	0.003338898	chr1	216078088	rs775293551	C	T	splicing
17	P1	*PKHD1*	0.003338898	chr6	52046089	rs199568593	A	G	nonsynonymous SNV
18	P1	*CYP21A2*	0.003333333	chr6	32040110	rs6471	G	T	nonsynonymous SNV
19	P1	*ABCA4*	0.003305785	chr1	94077713	rs752786160	G	A	nonsynonymous SNV
20	P1	*GBA*	0.003305785	chr1	155238215	rs364897	T	C	nonsynonymous SNV
21	P1	*UROS*	0.003305785	chr10	125815061	rs121908012	A	G	nonsynonymous SNV
22	P1	*DHCR7*	0.003305785	chr11	71438985	rs803388857	C	T	nonsynonymous SNV
23	P1	*PAH*	0.003305785	chr12	102894800	rs625087277	TTGA	T	nonframeshift deletion
24	P1	*FANCA*	0.003305785	chr16	89805275	rs759877008	C	T	intronic
25	P1	*GAA*	0.003305785	chr17	80112922	rs289408688	C	A	nonsynonymous SNV
26	P1	*SLC25A13*	0.003305785	chr7	96121928	rs803387225	G	GCCCGGGCAGCCACCTGTAATCTC	frameshift insertion
27	P1	*SLC25A13*	0.003305785	chr7	96184990	rs763191789	G	A	stopgain
28	P1	*SLC25A13*	0.003305785	chr7	96189371	rs803387220	TCATA	T	frameshift deletion
29	P1	*CFTR*	0.003305785	chr7	117559463	rs397508200	G	A	splicing
30	P1	*TYR*	0.00170068	chr11	89191278	rs617543775	G	A	nonsynonymous SNV
31	P1	*CFTR*	0.001697793	chr7	117611638	rs121909019	G	A	nonsynonymous SNV
32	P1	*PKHD1*	0.001677852	chr6	51904044	rs1582470309	T	A	splicing
33	P1	*LYST*	0.001672241	chr1	23580582	rs803386526	G	A	stopgain
34	P1	*USH2A*	0.001666667	chr1	216247185	rs111033334	G	A	stopgain
35	P1	*ALMS1*	0.001666667	chr2	73573290	rs376091780	C	T	stopgain
36	P1	*SLC26A4*	0.001663894	chr7	107683453	rs111033313	A	G	splicing
37	P1	*SLC26A4*	0.00166113	chr7	107690203	rs111033220	C	T	nonsynonymous SNV
38	P1	*GBA*	0.001658375	chr1	155238630	rs439898	G	A	nonsynonymous SNV
39	P1	*CEP290*	0.001658375	chr12	88055666	rs587783017	A	AT	frameshift insertion
40	P1	*CYP21A2*	0.001658375	chr6	32039426	rs6475	T	A	nonsynonymous SNV
41	P1	*PAH*	0.001655629	chr12	102851709	rs62642939	C	T	nonsynonymous SNV
42	P1	*AGXT*	0.001655629	chr2	240868890	rs398122322	A	AC	frameshift insertion
43	P1	*GUSB*	0.001655629	chr7	65974701	rs121918185	G	A	stopgain
44	P1	*CFTR*	0.001655	chr7	117590	rs3975082	G	T	stopgain

			629		426	96			
45	P1	*MPL*	0.001652893	chr1	43338563	rs587778514	CCT	C	frameshift deletion
46	P1	*MUTYH*	0.001652893	chr1	45332955	rs762307622	C	T	stopgain
47	P1	*ABCA4*	0.001652893	chr1	94001072	rs61750648	G	A	nonsynonymous SNV
48	P1	*ABCA4*	0.001652893	chr1	94030452	rs61750142	C	T	nonsynonymous SNV
49	P1	*ABCA4*	0.001652893	chr1	94042797	rs756840095	G	A	nonsynonymous SNV
50	P1	*ABCA4*	0.001652893	chr1	94043470	rs201855602	G	A	nonsynonymous SNV
51	P1	*GBA*	0.001652893	chr1	155240629	rs104886460	C	T	splicing
52	P1	*NPHS2*	0.001652893	chr1	179552605	rs743153468	G	A	nonsynonymous SNV
53	P1	*LAMB3*	0.001652893	chr1	209623516	rs1057516486	TG	T	frameshift deletion
54	P1	*USH2A*	0.001652893	chr1	215674335	rs1003869920	G	A	stopgain
55	P1	*USH2A*	0.001652893	chr1	215674795	rs768161313	CATTT	C	frameshift deletion
56	P1	*USH2A*	0.001652893	chr1	215674901	rs527236137	G	A	nonsynonymous SNV
57	P1	*PRF1*	0.001652893	chr10	70600743	rs200430442	G	A	nonsynonymous SNV
58	P1	*CYP17A1*	0.001652893	chr10	102830761	rs756135168	TGAAAGAGTC	T	nonframeshift deletion
59	P1	*ABCC8*	0.001652893	chr11	17408415	rs570388861	G	A	stopgain
60	P1	*PYGM*	0.001652893	chr11	64751966	rs119103255	G	A	stopgain
61	P1	*GYS2*	0.001652893	chr12	21568952	rs121918419	G	A	stopgain
62	P1	*GNPTAB*	0.001652893	chr12	101753409	rs137852897	G	A	stopgain
63	P1	*GNPTAB*	0.001652893	chr12	101764362	rs281864996	ATTTTC	A	frameshift deletion
64	P1	*MMAB*	0.001652893	chr12	109561046	rs747499304	T	TCGGCCCGGCGGCACA	nonframeshift insertion
65	P1	*ACADS*	0.001652893	chr12	120739141	rs387906950	A	G	nonsynonymous SNV
66	P1	*GJB2*	0.001652893	chr13	20189281	rs111033204	CAT	C	frameshift deletion
67	P1	*GJB2*	0.001652893	chr13	20189511	rs104894396	C	T	stopgain
68	P1	*BRCA2*	0.001652893	chr13	32356550	rs80358981	C	T	stopgain
69	P1	*ATP7B*	0.001652893	chr13	51941194	rs60431989	A	G	nonsynonymous SNV
70	P1	*RDH12*	0.001652893	chr14	67724568	rs766631462	C	T	nonsynonymous SNV
71	P1	*GALC*	0.001652893	chr14	87993029	rs751975987	C	A	nonsynonymous SNV
72	P1	*FAH*	0.001652893	chr15	80173089	rs80338898	C	T	nonsynonymous SNV
73	P1	*OTOA*	0.001652893	chr16	21722979	rs148690740	G	A	splicing
74	P1	*BBS2*	0.001652893	chr16	56484820	rs567573386	G	A	stopgain
75	P1	*CNGB1*	0.001652893	chr16	57904823	rs760430056	G	GC	frameshift insertion

76	P1	*LDLR*	0.001652893	chr19	11106564	rs879254652	G	A	splicing
77	P1	*CYP27A1*	0.001652893	chr2	218814075	rs533885672	C	T	stopgain
78	P1	*COL4A3*	0.001652893	chr2	227263845	rs371334239	C	T	stopgain
79	P1	*COL4A3*	0.001652893	chr2	227307800	rs748026887	TCACCCGA	T	frameshift deletion
80	P1	*AIRE*	0.001652893	chr21	44288459	rs199612115	G	T	splicing
81	P1	*COL6A1*	0.001652893	chr21	45998172	rs1002726737	G	A	splicing
82	P1	*PLA2G6*	0.001652893	chr22	38115658	rs587784339	G	A	stopgain
83	P1	*PLA2G6*	0.001652893	chr22	38120888	rs535486098	C	T	nonsynonymous SNV
84	P1	*ARSA*	0.001652893	chr22	50625330	rs761555167	C	CG	frameshift insertion
85	P1	*KLHL40*	0.001652893	chr3	42688963	rs778022582	A	C	nonsynonymous SNV
86	P1	*ACAD9*	0.001652893	chr3	128906208	rs149753643	G	A	nonsynonymous SNV
87	P1	*IDUA*	0.001652893	chr4	991150	rs139024319	G	A	nonsynonymous SNV
88	P1	*SLC22A5*	0.001652893	chr5	132392565	rs603766244	C	G	nonsynonymous SNV
89	P1	*SLC22A5*	0.001652893	chr5	132392577	rs386134223	G	A	nonsynonymous SNV
90	P1	*PEX7*	0.001652893	chr6	136869905	rs121909152	G	A	nonsynonymous SNV
91	P1	*GUSB*	0.001652893	chr7	65979782	rs121918181	G	A	nonsynonymous SNV
92	P1	*POR*	0.001652893	chr7	75985179	rs28931608	G	A	nonsynonymous SNV
93	P1	*SLC26A4*	0.001652893	chr7	107704382	rs752807925	C	T	stopgain
94	P1	*SLC26A4*	0.001652893	chr7	107710132	rs121908362	A	G	nonsynonymous SNV
95	P1	*CFTR*	0.001652893	chr7	117592032	rs121908759	G	A	nonsynonymous SNV
96	P1	*CFTR*	0.001652893	chr7	117592292	rs121908760	C	T	stopgain
97	P1	*CNGB3*	0.001652893	chr8	86579224	rs200805087	G	A	stopgain
98	P1	*GNE*	0.001652893	chr9	36236864	rs121908629	C	T	nonsynonymous SNV
99	P1	*FBP1*	0.001652893	chr9	94603437	rs757653154	A	AC	frameshift insertion
100	P1	*ASS1*	0.001652893	chr9	130494983	rs121908640	C	T	nonsynonymous SNV
101	P2	*NEB*	0.009917355	chr2	151493386	rs761067911	CTCCATCTCTGGAGTAACAGGTG	C	frameshift deletion
102	P2	*GJC2*	0.004958678	chr1	228158957	rs761261049	C	A	nonsynonymous SNV
103	P2	*HPS6*	0.004958678	chr10	102065628	rs1590262450	GT	G	frameshift deletion
104	P2	*CHKB*	0.004958678	chr22	50580643	rs757369551	TG	T	frameshift deletion
105	P2	*CYP17A1*	0.003305785	chr10	102837063	rs764723654	A	G	splicing
106	P2	*MYO15A*	0.003305785	chr17	18122323	rs766187994	C	CA	frameshift insertion
107	P2	*RPE65*	0.001669	chr1	684298	rs1219177	G	A	nonsynonymous

			449		35	45			SNV
108	P2	*ALMS1*	0.001663894	chr2	73453926	rs1198051503	G	T	stopgain
109	P2	*HBB*	0.00166113	chr11	52267605	rs35755331	AG	A	frameshift deletion
110	P2	*PAH*	0.00166113	chr12	10284372	rs184148104	G	C	nonsynonymous SNV
111	P2	*ALMS1*	0.00166113	chr2	73572989	rs398122992	GAGGTCTAATCAAATTAAAA	G	frameshift deletion
112	P2	*TMEM237*	0.001658375	chr2	20163205	rs80034299	C	T	splicing
113	P2	*FH*	0.001655629	chr1	24150868	rs155334145	A	G	nonsynonymous SNV
114	P2	*CC2D2A*	0.001655629	chr4	15596177	rs587779732	C	G	nonsynonymous SNV
115	P2	*ESPN*	0.001652893	chr1	6445936	rs752649606	G	A	splicing
116	P2	*ALPL*	0.001652893	chr1	21573800	rs129241545	G	T	splicing
117	P2	*FUCA1*	0.001652893	chr1	23865622	rs781230182	A	T	stopgain
118	P2	*LDLRAP1*	0.001652893	chr1	25543762	rs1201229554	T	TG	frameshift insertion
119	P2	*RPE65*	0.001652893	chr1	68444857	rs61752873	C	T	nonsynonymous SNV
120	P2	*ABCA4*	0.001652893	chr1	94010868	rs752160946	C	T	nonsynonymous SNV
121	P2	*ABCA4*	0.001652893	chr1	94030427	rs200967229	C	T	splicing
122	P2	*GBA*	0.001652893	chr1	15523989	rs1170895261	C	CG	frameshift insertion
123	P2	*USH2A*	0.001652893	chr1	21609710924	rs201529124	G	A	nonsynonymous SNV
124	P2	*MYO3A*	0.001652893	chr10	26173761	rs752046945	CT	C	frameshift deletion
125	P2	*ABCC8*	0.001652893	chr11	17396984	rs149331388	C	T	nonsynonymous SNV
126	P2	*BEST1*	0.001652893	chr11	61955891	rs281865236	C	A	nonsynonymous SNV
127	P2	*NDUFV1*	0.001652893	chr11	67612158	rs766830864	A	AG	frameshift insertion
128	P2	*CEP290*	0.001652893	chr12	88060995	rs766670248	C	T	splicing
129	P2	*CEP290*	0.001652893	chr12	88139193	rs951979448	T	C	splicing
130	P2	*GNPTAB*	0.001652893	chr12	10178029292	rs750793712	A	C	intronic
131	P2	*PAH*	0.001652893	chr12	10285532611	rs192592111	C	A	nonsynonymous SNV
132	P2	*BRCA2*	0.001652893	chr13	32338886	rs376338226	G	T	stopgain
133	P2	*SLC25A15*	0.001652893	chr13	40805209	rs780201405	AC	A	frameshift deletion
134	P2	*ATP7B*	0.001652893	chr13	51937337	rs778732681	C	G	nonsynonymous SNV
135	P2	*TGM1*	0.001652893	chr14	24259745	rs397514525	G	A	nonsynonymous SNV
136	P2	*TGM1*	0.001652893	chr14	24261783	rs139208806	T	C	nonsynonymous SNV
137	P2	*IVD*	0.001652893	chr15	40411258	rs771914739	A	G	splicing

138	P2	*IVD*	0.001652893	chr15	40411554	rs1057517043	G	A	splicing
139	P2	*POLG*	0.001652893	chr15	89318611	rs767138032	G	A	nonsynonymous SNV
140	P2	*VPS33B*	0.001652893	chr15	91014430	rs1064793614	CA	C	frameshift deletion
141	P2	*TK2*	0.001652893	chr16	66536981	rs281865489	G	A	nonsynonymous SNV
142	P2	*FANCA*	0.001652893	chr16	89749781	rs1166286386	C	T	stopgain
143	P2	*ALOX12B*	0.001652893	chr17	8077109	rs750066836	G	A	nonsynonymous SNV
144	P2	*MYO15A*	0.001652893	chr17	18150835	rs760461823	G	A	splicing
145	P2	*GAA*	0.001652893	chr17	80108542	rs752002666	G	C	nonsynonymous SNV
146	P2	*GCDH*	0.001652893	chr19	12896366	rs771650894	T	C	nonsynonymous SNV
147	P2	*SLC7A9*	0.001652893	chr19	32862554	rs758242098	G	A	nonsynonymous SNV
148	P2	*ETFB*	0.001652893	chr19	51353275	rs548046212	C	T	nonsynonymous SNV
149	P2	*FAM161A*	0.001652893	chr2	61836057	.	CT	C	frameshift deletion
150	P2	*CNGA3*	0.001652893	chr2	98396947	rs774676415	G	A	nonsynonymous SNV
151	P2	*PROC*	0.001652893	chr2	127428560	rs121918150	G	A	nonsynonymous SNV
152	P2	*BCS1L*	0.001652893	chr2	218662680	rs1553597661	G	A	splicing
153	P2	*WNT10A*	0.001652893	chr2	218882358	rs374910216	G	A	nonsynonymous SNV
154	P2	*WNT10A*	0.001652893	chr2	218882424	rs561503117	G	A	splicing
155	P2	*COL6A3*	0.001652893	chr2	237380987	rs755382829	G	A	stopgain
156	P2	*AGXT*	0.001652893	chr2	240871406	rs180177227	G	A	nonsynonymous SNV
157	P2	*MKKS*	0.001652893	chr20	10412653	rs113032343	C	T	nonsynonymous SNV
158	P2	*COL7A1*	0.001652893	chr3	48581271	rs121912847	G	A	stopgain
159	P2	*ACAD9*	0.001652893	chr3	128909422	rs1936041020	G	A	splicing
160	P2	*ATR*	0.001652893	chr3	142553741	rs755272769	C	T	splicing
161	P2	*HPS3*	0.001652893	chr3	149140187	rs748883997	AG	A	frameshift deletion
162	P2	*PDE6B*	0.001652893	chr4	66219731	rs201541131	C	T	nonsynonymous SNV
163	P2	*EVC*	0.001652893	chr4	5793695	rs132900699 4	C	T	stopgain
164	P2	*PROM1*	0.001652893	chr4	16009012	rs563415711	A	T	nonsynonymous SNV
165	P2	*MTTP*	0.001652893	chr4	99608827	rs199422220	G	A	nonsynonymous SNV
166	P2	*ETFDH*	0.001652893	chr4	158685137	rs121964955	G	A	nonsynonymous SNV
167	P2	*ETFDH*	0.001652893	chr4	158695582	rs780015493	A	G	nonsynonymous SNV
168	P2	*SLC22A5*	0.001652893	chr5	132370255	rs386134191	C	G	nonsynonymous SNV
169	P2	*RARS2*	0.001652	chr6	875899	rs1998620	A	C	startloss

			893		56	50			
170	P2	*LAMA2*	0.001652893	chr6	128883359	rs1211322465	T	C	splicing
171	P2	*GUSB*	0.001652893	chr7	65980292	rs1053785648	G	A	stopgain
172	P2	*SLC26A4*	0.001652893	chr7	107672181	rs1275009555	TC	T	frameshift deletion
173	P2	*TMEM67*	0.001652893	chr8	93787842	rs786205608	A	G	splicing
174	P2	*TMEM67*	0.001652893	chr8	93793267	rs747025617	C	T	nonsynonymous SNV
175	P2	*RECQL4*	0.001652893	chr8	144515987	rs1050860620	C	T	splicing
176	P2	*RMRP*	0.001652893	chr9	35657978	rs1156413585	C	T	ncRNA_exonic
177	P2	*GNE*	0.001652893	chr9	36234115	rs200643106	G	A	stopgain
178	P2	*VPS13A*	0.001652893	chr9	77337499	rs1417854249	C	T	stopgain
179	P2	*INVS*	0.001652893	chr9	100297017	rs1425211517	C	T	stopgain
180	P2	*POMT1*	0.001652893	chr9	131518948	rs727502854	T	TA	stopgain
181	P3	*FANCA*	0.031456954	chr16	89803481	rs172323404	T	TGA	intronic
182	P3	*F5*	0.024793388	chr1	169555300	rs118203905	T	C	nonsynonymous SNV
183	P3	*CYP27A1*	0.009917355	chr2	218814696	rs200883871	G	C	nonsynonymous SNV
184	P3	*FREM2*	0.003305785	chr13	38784709	rs121434355	G	A	nonsynonymous SNV
185	P3	*OTOF*	0.003305785	chr2	26483581	rs397515582	G	A	stopgain
186	P3	*CFTR*	0.001706485	chr7	117611708	rs150020260	G	A	stopgain
187	P3	*LDLR*	0.001655629	chr19	11129653	rs155580961 4	G	A	nonsynonymous SNV
188	P3	*GJB3*	0.001652893	chr1	34785182	rs770247378	CATT	C	nonframeshift deletion
189	P3	*ABCA4*	0.001652893	chr1	94098799	rs6264595 2	G	A	nonsynonymous SNV
190	P3	*SCNN1A*	0.001652893	chr12	6348981	rs137852634	G	A	stopgain
191	P3	*RPGRIP1*	0.001652893	chr14	21303542	rs554396590	C	T	stopgain
192	P3	*ZNF469*	0.001652893	chr16	88427760	rs273585617	C	T	nonsynonymous SNV
193	P3	*USH1G*	0.001652893	chr17	74919743	rs538983393	C	T	nonsynonymous SNV
194	P3	*FAM161A*	0.001652893	chr2	61840061	rs1572879569	T	A	stopgain
195	P3	*CNGA3*	0.001652893	chr2	98396938	rs763041373	G	A	nonsynonymous SNV
196	P3	*CHRNG*	0.001652893	chr2	232540072	rs121912672	C	T	stopgain
197	P3	*AGXT*	0.001652893	chr2	24086887	rs796052057	G	C	nonsynonymous SNV
198	P3	*AGXT*	0.001652893	chr2	24086917	rs767586362	G	A	nonsynonymous SNV
199	P3	*BTD*	0.001652893	chr3	15645225	rs146600671	G	A	nonsynonymous SNV
200	P3	*CRTAP*	0.001652893	chr3	33114080	rs726593557	G	A	startloss

201	P3	*KLHL40*	0.001652893	chr3	42690863	rs397509421	G	C	nonsynonymous SNV
202	P3	*HPS3*	0.001652893	chr3	149170483	rs281865096	G	A	intronic
203	P3	*PDE6B*	0.001652893	chr4	625919	rs776050413	G	A	nonsynonymous SNV
204	P3	*EVC*	0.001652893	chr4	5783656	rs765269619	C	G	stopgain
205	P3	*PROM1*	0.001652893	chr4	16075767	rs747512450	TG	T	frameshift deletion
206	P3	*MOCS2*	0.001652893	chr5	53109714	rs121908607	G	A	stopgain
207	P3	*MAK*	0.001652893	chr6	10803886	rs387906648	C	T	nonsynonymous SNV
208	P3	*CFTR*	0.001652893	chr7	117559589	rs1800092	C	G	nonsynonymous SNV
209	P4	*MLC1*	0.008264463	chr22	50084838	rs184241759	C	T	nonsynonymous SNV
210	P4	*LDLR*	0.003305785	chr19	11116874	rs777188764	G	A	nonsynonymous SNV
211	P4	*GCDH*	0.001652893	chr19	12897764	rs567564095	G	A	nonsynonymous SNV
212	P4	*ACADVL*	0.001652893	chr17	7222250	rs796051913	AAGG	A	nonframeshift deletion
213	P4	*ACADVL*	0.001652893	chr17	7223687	rs113994169	C	T	nonsynonymous SNV
214	P4	*GAA*	0.001652893	chr17	80108305	rs750030887	C	T	nonsynonymous SNV
215	P4	*SGSH*	0.001652893	chr17	80210898	rs766938111	C	T	nonsynonymous SNV
216	P4	*MEFV*	0.001652893	chr16	3243205	rs104895097	C	T	nonsynonymous SNV
217	P4	*CNGB1*	0.001652893	chr16	57901442	rs749199721	C	T	splicing
218	P4	*POLG*	0.001652893	chr15	89319065	rs181860632	G	A	nonsynonymous SNV
219	P4	*POLG*	0.003305785	chr15	89325609	rs1001570418	C	T	nonsynonymous SNV
220	P4	*ATP7B*	0.001652893	chr13	51942482	rs541208827	C	T	nonsynonymous SNV
221	P4	*ATP7B*	0.003305785	chr13	51949772	rs121907993	G	A	nonsynonymous SNV
222	P4	*ATP7B*	0.001652893	chr13	51950132	rs191312027	C	T	nonsynonymous SNV
223	P4	*SMPD1*	0.028099174	chr11	6392060	rs202081954	C	G	nonsynonymous SNV
224	P4	*RAPSN*	0.001652893	chr11	47448079	rs104894299	G	T	nonsynonymous SNV
225	P4	*GLDC*	0.001652893	chr9	6533142	rs772574530	T	C	nonsynonymous SNV
226	P4	*GNE*	0.004958678	chr9	36276927	rs200763627	A	T	stopgain
227	P4	*ASL*	0.001652893	chr7	66086605	rs769017508	C	T	nonsynonymous SNV
228	P4	*PEX1*	0.001652893	chr7	92494357	rs617504227	A	G	nonsynonymous SNV
229	P4	*CFTR*	0.001652893	chr7	117480061	rs756314710	C	T	UTR5
230	P4	*CFTR*	0.001652893	chr7	117548795	rs745511128	C	T	nonsynonymous SNV
231	P4	*CFTR*	0.006791171	chr7	117611646	rs200321110	G	A	nonsynonymous SNV
232	P4	*FARS2*	0.001655	chr6	536903	rs1469884	C	T	nonsynonymous

			629		7	68			SNV
233	P4	*HFE*	0.001652893	chr6	26092913	rs1800562	G	A	nonsynonymous SNV
234	P4	*CYP21A2*	0.010380623	chr6	32040723	rs202242769	G	A	nonsynonymous SNV
235	P4	*MOCS1*	0.018181818	chr6	39925702	rs377167949	G	A	nonsynonymous SNV
236	P4	*PKHD1*	0.001652893	chr6	51883163	rs398124492	A	G	nonsynonymous SNV
237	P4	*EYS*	0.001652893	chr6	64230600	rs749909863	C	T	nonsynonymous SNV
238	P4	*SLC22A5*	0.00661157	chr5	132384290	rs386134199	C	T	nonsynonymous SNV
239	P4	*BTD*	0.003305785	chr3	15644824	rs397507176	A	G	nonsynonymous SNV
240	P4	*BTD*	0.003305785	chr3	15645186	rs13078881	G	C	nonsynonymous SNV
241	P4	*ILDR1*	0.003305785	chr3	12199418 8	rs142746163	G	A	stopgain
242	P4	*OTOF*	0.009917355	chr2	26463969	rs199766465	C	G	nonsynonymous SNV
243	P4	*LRPPRC*	0.001652893	chr2	43889733	rs759052246	CA	C	frameshift deletion
244	P4	*PROC*	0.014876033	chr2	12742612 0	rs199469469	GAGA	G	nonframeshift deletion
245	P4	*TTN-AS1*	0.001769912	chr2	17860165 3	rs754717390	C	G	splicing
246	P4	*TTN*	0.001655629	chr2	17879070 7	rs397517497	C	T	splicing
247	P4	*ACADM*	0.001652893	chr1	75740091	rs773677327	A	G	nonsynonymous SNV
248	P4	*ACADM*	0.001652893	chr1	75762744	rs760892123	T	C	nonsynonymous SNV
249	P4	*ABCA4*	0.018181818	chr1	94005469	rs148460146	C	T	nonsynonymous SNV
250	P4	*ABCA4*	0.003305785	chr1	94008251	rs1800553	C	T	nonsynonymous SNV
251	P4	*ABCA4*	0.001652893	chr1	94019701	rs6175056 3	C	T	nonsynonymous SNV
252	P4	*ABCA4*	0.001652893	chr1	94021934	rs1762111	A	G	nonsynonymous SNV
253	P4	*ABCA4*	0.004958678	chr1	94024978	rs626425775	G	A	nonsynonymous SNV
254	P4	*ABCA4*	0.008264463	chr1	94030483	rs563570600	C	T	nonsynonymous SNV
255	P4	*ABCA4*	0.00331675	chr1	94047010	rs617494446	G	A	nonsynonymous SNV
256	P4	*ABCA4*	0.004958678	chr1	94063157	rs617485559	C	T	nonsynonymous SNV
257	P4	*DPYD*	0.003311258	chr1	97828127	rs189768576	G	A	stopgain
258	P4	*AGL*	0.001655629	chr1	99916709	rs121180558	C	T	stopgain
259	P4	*AMPD1*	0.049586777	chr1	11467746 5	rs121912682	C	T	nonsynonymous SNV
260	P4	*AMPD1*	0.003305785	chr1	11467961 6	rs345261999	T	A	nonsynonymous SNV
261	P4	*NPHS2*	0.00661157	chr1	17955707 9	rs617477228	C	T	nonsynonymous SNV
262	P4	*USH2A*	0.001652893	chr1	21567457 2	rs139474806	T	C	nonsynonymous SNV
263	P4	*USH2A*	0.001663894	chr1	21624659 2	rs201527662	A	C	nonsynonymous SNV

Supplementary table 3: Allele frequencies of likely pathogenic/pathogenic variants detected in autosomal recessive genes in GenomAD database.

	GnomAD Allele Frequency									
Variant number	AF	AF_afr	AF_ami	AF_amr	AF_asj	AF_eas	AF_fin	AF_nfe	AF_oth	AF_sas
1	0.0003	0.0001	0	0	0	0.0013	0	0	0	0.0111
2	0.0035	0.0008	0	0.0033	0.0075	0.0854	0.0014	0.0015	0.0074	0
3	0.0000282	0.0000255	0	0	0	0.0006	0	0	0	0
4	0.0467	0.0128	0.1626	0.0495	0.0472	0.008	0.0679	0.0642	0.0437	0.0618
5	0.0000698	0.0000952	0	0	0	0.0003	0	0.0000775	0	0
6	0.0000488	0	0	0	0	0.0022	0	0	0	0
7	.	.	.	.	.	.	.	.	.	.
8	0.0034	0.0021	0	0.0015	0.0012	0.0061	0.0081	0.0036	0.0032	0.0043
9	0.00000697	0	0	0	0	0.0003	0	0	0	0
10	0.0000698	0	0	0	0	0.0022	0	0	0	0.001
11	0.000014	0	0	0	0	0.0003	0	0	0	0
12	0.0002	0	0	0	0	0.0029	0	0.0002	0	0
13	0.0002	0	0	0	0	0.0093	0	0.0000155	0	0
14	0.0000282	0	0	0	0	0.0013	0	0	0	0
15	0.00000697	0	0	0	0	0.0003	0	0	0	0
16	0.00000698	0	0	0	0	0.0003	0	0	0	0
17	0.0000209	0	0	0	0	0.0006	0	0	0.0005	0
18	0.0139	0.0048	0	0.0288	0.0737	0.0032	0.001	0.0103	0.0156	0.001
19	0.0000558	0	0	0	0	0.0022	0	0.0000155	0	0
20	0.0001	0.0002	0	0.0000733	0	0	0.0000956	0.000031	0.0005	0
21	0.0002	0.0002	0	0.0004	0	0	0	0.0002	0.0009	0
22	0.0000349	0	0	0	0	0	0	0.0000774	0	0
23	0.0001	0	0.0144	0.0000732	0	0	0	0.0000774	0	0
24	0.00000698	0	0	0	0	0.0003	0	0	0	0
25	0.0000418	0	0	0	0	0.0019	0	0	0	0
26	0.000014	0	0	0	0	0.0006	0	0	0	0
27	0.00000698	0	0	0	0	0	0	0.0000155	0	0
28	0.0001	0	0	0.0001	0	0.0061	0	0	0	0
29	0.00000698	0	0	0	0	0	0	0	0	0.0003
30	0.0000628	0.0000477	0	0.0002	0	0.0003	0	0.000031	0.0005	0
31	0.000014	0.0000238	0	0	0	0	0	0.0000155	0	0
32	.	.	.	.	.	.	.	.	.	.
33	0.000014	0.0000238	0	0	0	0	0	0.0000155	0	0
34	.	.	.	.	.	.	.	.	.	.
35	0.0000349	0.0000476	0	0	0	0	0	0.0000465	0	0
36	0.0002	0	0	0	0	0.0077	0	0.0000155	0	0
37	0.0000419	0.0000238	0	0	0.0003	0.0006	0.0000955	0.0000155	0	0
38	0.0000349	0.0000239	0	0	0	0.0003	0	0.000031	0	0
39	.	.	.	.	.	.	.	.	.	.

40	0.002	0.001	0	0.0014	0.0006	0.001	0.0029	0.0011	0.0005	0.0007
41	0.0000279	0.0000238	0	0.0000733	0	0	0	0.0000155	0.0005	0
42	0.0001	0.0000963	0	0.0001	0	0.0003	0.0000966	0.0001	0	0.0003
43	0.0000209	0	0	0	0	0	0	0.0000465	0	0
44	0.000014	0	0	0	0	0	0	0.000031	0	0
45	0.0000488	0.0000476	0	0.0000732	0	0.0006	0	0.000031	0	0
46	0.0000209	0	0	0	0	0.001	0	0	0	0
47	0.0003	0.0009	0	0.0001	0	0	0	0.0001	0	0
48	0.0000279	0.0000238	0	0	0	0	0	0.0000465	0	0
49	0.0000628	0	0	0.0006	0	0	0	0.0000155	0	0
50	0.0000209	0	0	0	0	0	0.0002	0.0000155	0	0
51	0.0000837	0	0	0	0.0003	0	0	0.0000929	0.0005	0.0003
52	0.00000697	0	0	0	0	0.0003	0	0	0	0
53	.	.	.	.	.	.	.	.	.	.
54	0.00000698	0	0	0.0000733	0	0	0	0	0	0
55	0.000014	0	0	0	0	0	0	0.000031	0	0
56	.	.	.	.	.	.	.	.	.	.
57	0.0000209	0.0000238	0	0	0	0.0003	0	0.0000155	0	0
58	0.000014	0	0	0	0	0.0006	0	0	0	0
59	0.00000697	0	0	0	0	0	0	0.0000155	0	0
60	0.00000697	0.0000238	0	0	0	0	0	0	0	0
61	0.0001	0.0000714	0	0	0	0	0	0.0001	0	0.002
62	0.000014	0	0	0	0	0.0003	0	0.0000155	0	0
63	0.00000698	0.0000239	0	0	0	0	0	0	0	0
64	0.00000716	0.0000247	0	0	0	0	0	0	0	0
65	0.000014	0	0	0	0	0.0006	0	0	0	0
66	0.0000279	0	0	0	0	0.0013	0	0	0	0
67	0.0001	0	0	0	0	0	0	0.0000465	0	0.0043
68	0.000014	0.0000476	0	0	0	0	0	0	0	0
69	0.0000349	0.0000476	0	0	0	0	0	0.000031	0.0005	0
70	.	.	.	.	.	.	.	.	.	.
71	0.000014	0	0	0	0	0	0	0.000031	0	0
72	0.0000628	0.0000238	0	0	0.0015	0.0003	0	0.0000155	0	0
73	0.000014	0	0	0	0	0	0	0.000031	0	0
74	0.000014	0.0000476	0	0	0	0	0	0	0	0
75	0.0000628	0.0000238	0	0	0	0	0	0.0001	0	0
76	.	.	.	.	.	.	.	.	.	.
77	0.0000209	0	0	0	0	0.001	0	0	0	0
78	0.0000488	0.0001	0	0	0	0	0	0.0000155	0	0
79	.	.	.	.	.	.	.	.	.	.
80	0.0000418	0.0000238	0	0	0	0.0013	0	0	0.0005	0
81	.	.	.	.	.	.	.	.	.	.
82	0.00000698	0	0	0	0	0	0	0.0000155	0	0
83	.	.	.	.	.	.	.	.	.	.
84	0.000014	0	0	0	0	0.0006	0	0	0	0

85	0.0000209	0	0	0	0	0.001	0	0	0	0
86	0.00000697	0	0	0	0	0	0	0	0	0
87	0.0002	0.0000476	0	0	0	0	0	0.0003	0	0
88	0.0000837	0	0	0	0	0.0035	0	0	0.0005	0
89	.	.	.	.	.	.	.	.	.	.
90	0.0000488	0.0000238	0	0	0	0	0	0.0000929	0	0
91	0.0000697	0.0000713	0	0	0	0	0	0.0001	0	0
92	0.000014	0	0	0	0	0.0006	0	0	0	0
93	.	.	.	.	.	.	.	.	.	.
94	0.0000349	0	0	0	0	0.0016	0	0	0	0
95	0.0004	0.0012	0	0.0000733	0	0.001	0	0	0.0009	0
96	0.00000698	0	0	0	0	0	0	0	0.0005	0
97	0.00000698	0.0000238	0	0	0	0	0	0	0	0
98	0.000014	0	0	0	0.0003	0	0	0.0000155	0	0
99	0.0002	0.0003	0	0.0007	0	0	0	0.0001	0.0005	0
100	0.0000419	0.0000238	0	0	0	0	0	0.0000775	0	0
101	.	.	.	.	.	.	.	.	.	.
102	0.0000488	0	0	0	0	0.0022	0	0	0	0
103	.	.	.	.	.	.	.	.	.	.
104	.	.	.	.	.	.	.	.	.	.
105	0.0000209	0	0	0	0	0.0003	0	0.000031	0	0
106	0.0000628	0	0	0	0	0.0029	0	0	0	0
107	0.0000279	0.0000952	0	0	0	0	0	0	0	0
108	.	.	.	.	.	.	.	.	.	.
109	.	.	.	.	.	.	.	.	.	.
110	.	.	.	.	.	.	.	.	.	.
111	0.00000698	0	0	0	0	0.0003	0	0	0	0
112	0.0000279	0.0000238	0	0	0	0	0	0.000031	0	0.0003
113	.	.	.	.	.	.	.	.	.	.
114	0.000014	0	0	0.0001	0	0	0	0	0	0
115	.	.	.	.	.	.	.	.	.	.
116	0.00000697	0	0	0	0	0.0003	0	0	0	0
117	.	.	.	.	.	.	.	.	.	.
118	0.00000699	0	0	0	0	0	0	0.0000156	0	0
119	0.0001	0.0003	0	0.0000733	0	0	0	0	0	0
120	.	.	.	.	.	.	.	.	.	.
121	0.000014	0.0000238	0	0	0	0	0	0.0000155	0	0
122	.	.	.	.	.	.	.	.	.	.
123	0.0000209	0.0000714	0	0	0	0	0	0	0	0
124	0.000014	0	0	0	0	0.0006	0	0	0	0
125	0.0000279	0.0000238	0.0022	0	0	0	0	0.0000155	0	0
126	0.00000697	0	0	0	0	0	0	0.0000155	0	0
127	.	.	.	.	.	.	.	.	.	.
128	0.000014	0	0	0	0	0.0006	0	0	0	0
129	.	.	.	.	.	.	.	.	.	.

130	0.00000697	0	0	0	0	0.0003	0	0	0	0
131	0.0000349	0	0	0	0	0.0013	0	0	0	0
132	.	.	.	.	.	.	.	.	.	.
133	0.00000698	0	0	0	0	0.0003	0	0	0	0
134	.	.	.	.	.	.	.	.	.	.
135	.	.	.	.	.	.	.	.	.	.
136	0.0000767	0.0002	0	0	0	0.001	0	0	0	0
137	.	.	.	.	.	.	.	.	.	.
138	.	.	.	.	.	.	.	.	.	.
139	.	.	.	.	.	.	.	.	.	.
140	.	.	.	.	.	.	.	.	.	.
141	0.00000698	0	0	0	0	0	0.0000958	0	0	0
142	.	.	.	.	.	.	.	.	.	.
143	0.0000349	0.0000238	0	0	0	0	0	0.000062	0	0
144	0.000014	0	0	0	0	0.0006	0	0	0	0
145	0.000014	0	0	0	0	0	0	0.000031	0	0
146	.	.	.	.	.	.	.	.	.	.
147	0.0000419	0.0000714	0	0	0	0	0	0.0000465	0	0
148	0.0000279	0.0000714	0	0	0	0	0	0.0000155	0	0
149	.	.	.	.	.	.	.	.	.	.
150	0.000014	0	0	0	0	0	0	0.000031	0	0
151	0.00000697	0	0	0.0000732	0	0	0	0	0	0
152	.	.	.	.	.	.	.	.	.	.
153	.	.	.	.	.	.	.	.	.	.
154	.	.	.	.	.	.	.	.	.	.
155	0.000014	0.0000238	0	0.0000732	0	0	0	0	0	0
156	0.000014	0	0	0	0	0.0003	0.0000954	0	0	0
157	0.0000419	0.0001	0	0	0	0	0	0.0000155	0	0
158	0.00000698	0	0	0	0	0	0	0.0000155	0	0
159	.	.	.	.	.	.	.	.	.	.
160	0.000014	0.0000476	0	0	0	0	0	0	0	0
161	.	.	.	.	.	.	.	.	.	.
162	0.00000698	0.0000238	0	0	0	0	0	0	0	0
163	0.00000698	0	0	0	0	0.0003	0	0	0	0
164	.	.	.	.	.	.	.	.	.	.
165	0.0000349	0.0000238	0	0	0	0	0	0.000062	0	0
166	.	.	.	.	.	.	.	.	.	.
167	0.0000349	0.0000238	0	0	0	0.0013	0	0	0	0
168	.	.	.	.	.	.	.	.	.	.
169	.	.	.	.	.	.	.	.	.	.
170	0.00000698	0.0000238	0	0	0	0	0	0	0	0
171	.	.	.	.	.	.	.	.	.	.
172	0.00000698	0	0	0	0	0.0003	0	0	0	0
173	.	.	.	.	.	.	.	.	.	.
174	0.0000279	0	0	0.0001	0	0	0	0.000031	0	0

175	.	.	.	.	.	.	.	.	.	.
176	.	.	.	.	.	.	.	.	.	.
177	.	.	.	.	.	.	.	.	.	.
178	0.0000209	0.0000476	0	0	0	0	0	0.0000155	0	0
179	.	.	.	.	.	.	.	.	.	.
180	.	.	.	.	.	.	.	.	.	.
181	0.0719	0.2076	0	0.0233	0.022	0.0064	0.011	0.0097	0.0515	0.0968
182	0.0002	0	0	0.0004	0	0.0064	0	0.0000155	0.0014	0
183	0.00000697	0	0	0	0	0.0003	0	0	0	0
184	0.00000697	0	0	0	0	0.0003	0	0	0	0
185	0.00000697	0	0	0	0	0.0003	0	0	0	0
186	.	.	.	.	.	.	.	.	.	.
187	0.00000698	0.0000238	0	0	0	0	0	0	0	0
188	.	.	.	.	.	.	.	.	.	.
189	0.0000209	0	0	0.0000733	0	0.0006	0	0	0	0
190	0.000014	0.0000238	0	0	0	0	0	0.0000155	0	0
191	.	.	.	.	.	.	.	.	.	.
192	0.0000419	0.0000476	0	0	0	0	0	0.000062	0	0
193	0.000014	0	0	0	0	0.0003	0	0	0	0.0003
194	.	.	.	.	.	.	.	.	.	.
195	0.00000698	0	0	0	0	0.0003	0	0	0	0
196	0.0000279	0	0	0.0000732	0	0	0	0	0	0.001
197	.	.	.	.	.	.	.	.	.	.
198	0.00000697	0	0	0	0	0	0	0.0000155	0	0
199	0.00000698	0	0	0	0	0	0	0.0000155	0	0
200	.	.	.	.	.	.	.	.	.	.
201	.	.	.	.	.	.	.	.	.	.
202	0.0000698	0.0000715	0	0	0	0	0	0.0001	0	0
203	0.0000209	0	0	0	0	0.0003	0	0.0000155	0	0.0003
204	.	.	.	.	.	.	.	.	.	.
205	.	.	.	.	.	.	.	.	.	.
206	0.00000698	0	0	0	0	0.0003	0	0	0	0
207	0.0000209	0.0000238	0	0	0	0	0	0.000031	0	0
208	0.00000698	0	0	0	0	0	0	0	0	0.0003
209	0.0001	0.00002378	0	0	0	0.0032	0	0.00003097	0.0009	0.002
210	0.00003491	0.00009517	0	0	0	0	0	0.00001549	0	0
211	0.00002793	0.00009527	0	0	0	0	0	0	0	0
212	0.0002	0.00009522	0	0	0	0	0	0.0004	0	0.0003
213	0.00009774	0	0	0.0007	0	0.0003	0	0	0.0019	0
214	0.00002094	0.00004759	0	0	0	0	0	0.00001549	0	0
215	0.00004188	0.00004758	0	0.0001	0	0	0	0.00003097	0	0
216	0.0001	0.0000476	0	0.00007332	0	0.0022	0	0.00006194	0	0.0007
217	0.00002791	0.00004757	0	0	0	0	0	0.00003097	0	0
218	0.00009079	0.00004762	0	0.0004	0	0.0003	0	0.00006197	0	0
219	0.000006979	0	0	0	0	0	0.00009551	0	0	0

220	0.00006279	0	0	0	0	0.0022	0	0	0	0.0007
221	0.00004888	0.0001	0	0	0	0	0	0.00001549	0	0
222	0.001	0.0004	0	0.0006	0	0.0003	0.0000956	0.0018	0.0019	0
223	0.0001	0.00007139	0	0	0	0.0038	0	0.00001548	0	0
224	0.0015	0.0003	0	0.0023	0.0003	0	0.0006	0.0024	0.0028	0.0013
225	0.00001396	0	0	0	0	0.0003	0	0.00001549	0	0
226	0.00004188	0	0	0.00007332	0	0.0016	0	0	0	0
227	0.000006978	0.00002378	0	0	0	0	0	0	0	0
228	0.0000279	0	0	0	0	0.0013	0	0	0	0
229	.	.	.	.	.	.	.	.	.	.
230	0.00007067	0.00002412	0	0.0002	0	0	0	0.00007803	0	0.0003
231	0.0002	0.00009532	0	0.00007352	0	0.0016	0.0006	0.0003	0	0.0003
232	0.00005584	0.0001	0	0	0	0.0003	0	0.00003097	0	0
233	0.038	0.0113	0.0456	0.0189	0.0135	0.0006	0.0352	0.0649	0.0279	0.0026
234	0.0053	0.0019	0	0.0057	0.0066	0.0068	0.0049	0.0068	0.0042	0.0221
235	0.0004	0.001	0	0.00007321	0	0.0038	0.00009546	0.00001549	0	0.002
236	0.00001395	0	0	0	0	0	0	0.00003098	0	0
237	0.00006979	0	0	0.00007331	0	0.0022	0	0.00003098	0	0
238	0.0002	0	0	0	0	0	0	0	0	0.0085
239	0.0003	0.00004757	0	0	0	0	0	0	0.0009	0.0128
240	0.0292	0.0083	0	0.0297	0.028	0	0.0548	0.0402	0.0279	0.0339
241	0.0002	0.00007136	0	0.0002	0	0.0061	0	0	0.0005	0.0007
242	0.0001	0	0	0	0	0.0061	0	0	0.0005	0
243	0.0001	0.00004758	0	0.00007323	0	0.0029	0	0.00001549	0.0005	0.0007
244	0.0002	0	0	0	0	0.0099	0	0	0	0
245	0.00002096	0	0	0	0	0.0007	0	0	0	0.0003
246	0.00005583	0	0	0	0	0.0019	0	0.00001549	0	0.0003
247	.	.	.	.	.	.	.	.	.	.
248	0.00005585	0.00007135	0	0.0000733	0	0.0003	0	0.00003099	0.0005	0
249	0.0003	0.0005	0	0	0	0.0019	0	0.00009291	0	0.0013
250	0.003	0.0006	0	0.0039	0.0232	0	0.001	0.0032	0.0028	0.0155
251	0.0009	0.0027	0	0.0002	0	0	0	0.00007743	0.0005	0
252	0.0013	0.0003	0	0.0006	0.0027	0	0.0011	0.0022	0.0005	0
253	0.00005583	0	0	0	0	0.0006	0	0.00009293	0	0
254	0.0018	0.0009	0.0022	0.0022	0	0.0013	0.0006	0.0028	0.0023	0
255	0.00002095	0	0	0	0	0	0	0.00004647	0	0
256	0.00004887	0.00007139	0	0	0	0.0003	0	0.00004646	0	0
257	0.00006288	0.00004765	0	0.0004	0	0.0006	0	0	0	0
258	0.0001	0.0001	0	0.0003	0	0	0	0.00007747	0	0
259	0.0003	0.00002404	0	0.0006	0	0.0099	0	0	0.0014	0.0007
260	0.027	0.0056	0.1244	0.0167	0.0239	0	0.0839	0.0348	0.0192	0.0118
261	0.0281	0.0062	0.0633	0.0193	0.0566	0.0003	0.0663	0.0374	0.0214	0.0325
262	0.00004188	0	0	0	0	0.001	0.00009549	0.00001549	0	0.0003
263	0.00006281	0	0	0	0.0003	0.0026	0	0	0	0

Supplementary table 4 Clinical significance reported in ClinVar and variant interpretation used InterVar for variant with conflicting interpretation of pathogenicity.

Variant number	CLINVAR			INTERVAR
	CLINSIG	CLNREVSTAT	submissions	
1	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
2	Pathogenic	reviewed_by_expert_panel		
3	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
4	Pathogenic/Likely_pathogenic,_other	criteria_provided,_multiple_submitters,_no_conflicts		
5	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
6	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
7	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
8	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
9	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
10	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
11	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
12	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
13	Pathogenic	reviewed_by_expert_panel		
14	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
15	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
16	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
17	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
18	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
19	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_		

		conflicts		
20	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
21	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
22	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
23	Pathogenic	reviewed_by_expert_panel		
24	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
25	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
26	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
27	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
28	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
29	Pathogenic	reviewed_by_expert_panel		
30	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
31	Pathogenic	reviewed_by_expert_panel		
32	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
33	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
34	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
35	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
36	Pathogenic	reviewed_by_expert_panel		
37	Pathogenic	reviewed_by_expert_panel		
38	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
39	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
40	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
41	Pathogenic	reviewed_by_expert_panel		
42	Pathogenic	criteria_provided,_multiple_submitters,_no_		

		conflicts		
43	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
44	Pathogenic	reviewed_by_expert_panel		
45	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
46	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
47	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
48	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
49	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
50	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
51	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
52	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
53	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
54	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
55	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
56	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
57	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
58	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
59	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
60	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
61	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
62	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
63	Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		

64	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
65	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
66	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
67	Pathogenic	reviewed_by_expert_panel		
68	Pathogenic	reviewed_by_expert_panel		
69	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
70	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
71	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
72	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
73	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
74	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
75	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
76	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
77	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
78	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
79	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
80	Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
81	Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
82	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
83	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
84	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
85	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		

86	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
87	Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
88	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
89	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
90	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
91	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
92	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
93	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
94	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
95	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
96	Pathogenic	reviewed_by_expert_panel		
97	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
98	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
99	Pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
100	Pathogenic/Likely_pathogenic	criteria_provided,_multiple_submitters,_no_conflicts		
101	Pathogenic	criteria_provided,_single_submitter		
102	Pathogenic	criteria_provided,_single_submitter		
103	Pathogenic	criteria_provided,_single_submitter		
104	Pathogenic	criteria_provided,_single_submitter		
105	Pathogenic	criteria_provided,_single_submitter		
106	Pathogenic	criteria_provided,_single_submitter		
107	Pathogenic	criteria_provided,_single_submitter		
108	Pathogenic	criteria_provided,_single_submitter		
109	Pathogenic	criteria_provided,_single_submitter		
110	Likely_pathogenic	criteria_provided,_sing		

		le_submitter		
111	Pathogenic	criteria_provided,_single_submitter		
112	Pathogenic	criteria_provided,_single_submitter		
113	Pathogenic	criteria_provided,_single_submitter		
114	Likely_pathogenic	criteria_provided,_single_submitter		
115	Likely_pathogenic	criteria_provided,_single_submitter		
116	Pathogenic	criteria_provided,_single_submitter		
117	Pathogenic	criteria_provided,_single_submitter		
118	Pathogenic	criteria_provided,_single_submitter		
119	Pathogenic	criteria_provided,_single_submitter		
120	Likely_pathogenic	criteria_provided,_single_submitter		
121	Pathogenic	criteria_provided,_single_submitter		
122	Pathogenic	criteria_provided,_single_submitter		
123	Pathogenic	criteria_provided,_single_submitter		
124	Likely_pathogenic	criteria_provided,_single_submitter		
125	Likely_pathogenic	criteria_provided,_single_submitter		
126	Likely_pathogenic	criteria_provided,_single_submitter		
127	Pathogenic	criteria_provided,_single_submitter		
128	Likely_pathogenic	criteria_provided,_single_submitter		
129	Likely_pathogenic	criteria_provided,_single_submitter		
130	Likely_pathogenic	criteria_provided,_single_submitter		
131	Pathogenic	criteria_provided,_single_submitter		
132	Pathogenic	criteria_provided,_single_submitter		
133	Pathogenic	criteria_provided,_single_submitter		
134	Pathogenic	criteria_provided,_single_submitter		
135	Likely_pathogenic	criteria_provided,_single_submitter		
136	Likely_pathogenic	criteria_provided,_single_submitter		
137	Likely_pathogenic	criteria_provided,_single_submitter		
138	Likely_pathogenic	criteria_provided,_single_submitter		
139	Likely_pathogenic	criteria_provided,_single_submitter		
140	Pathogenic	criteria_provided,_single_submitter		
141	Likely_pathogenic	criteria_provided,_single_submitter		

142	Pathogenic	criteria_provided,_single_submitter		
143	Likely_pathogenic	criteria_provided,_single_submitter		
144	Likely_pathogenic	criteria_provided,_single_submitter		
145	Pathogenic	criteria_provided,_single_submitter		
146	Likely_pathogenic	criteria_provided,_single_submitter		
147	Pathogenic	criteria_provided,_single_submitter		
148	Likely_pathogenic	criteria_provided,_single_submitter		
149	Pathogenic	criteria_provided,_single_submitter		
150	Pathogenic	criteria_provided,_single_submitter		
151	Pathogenic	criteria_provided,_single_submitter		
152	Likely_pathogenic	criteria_provided,_single_submitter		
153	Pathogenic	criteria_provided,_single_submitter		
154	Pathogenic	criteria_provided,_single_submitter		
155	Pathogenic	criteria_provided,_single_submitter		
156	Pathogenic	criteria_provided,_single_submitter		
157	Likely_pathogenic	criteria_provided,_single_submitter		
158	Pathogenic	criteria_provided,_single_submitter		
159	Likely_pathogenic	criteria_provided,_single_submitter		
160	Likely_pathogenic	criteria_provided,_single_submitter		
161	Pathogenic	criteria_provided,_single_submitter		
162	Likely_pathogenic	criteria_provided,_single_submitter		
163	Pathogenic	criteria_provided,_single_submitter		
164	Likely_pathogenic	criteria_provided,_single_submitter		
165	Pathogenic	criteria_provided,_single_submitter		
166	Pathogenic	criteria_provided,_single_submitter		
167	Pathogenic	criteria_provided,_single_submitter		
168	Pathogenic	criteria_provided,_single_submitter		
169	Pathogenic	criteria_provided,_single_submitter		
170	Likely_pathogenic	criteria_provided,_single_submitter		
171	Pathogenic	criteria_provided,_single_submitter		
172	Pathogenic	criteria_provided,_single_submitter		
173	Pathogenic	criteria_provided,_sing		

		le_submitter		
174	Pathogenic	criteria_provided,_sing le_submitter		
175	Likely_pathogenic	criteria_provided,_sing le_submitter		
176	Likely_pathogenic	criteria_provided,_sing le_submitter		
177	Pathogenic	criteria_provided,_sing le_submitter		
178	Pathogenic	criteria_provided,_sing le_submitter		
179	Pathogenic	criteria_provided,_sing le_submitter		
180	Pathogenic	criteria_provided,_sing le_submitter		
181	Pathogenic	no_assertion_criteria_ provided		
182	Pathogenic	no_assertion_criteria_ provided		
183	Pathogenic	no_assertion_criteria_ provided		
184	Pathogenic	no_assertion_criteria_ provided		
185	Pathogenic	no_assertion_criteria_ provided		
186	Likely_pathogenic	no_assertion_criteria_ provided		
187	Pathogenic	no_assertion_criteria_ provided		
188	Pathogenic	no_assertion_criteria_ provided		
189	Pathogenic	no_assertion_criteria_ provided		
190	Pathogenic	no_assertion_criteria_ provided		
191	Pathogenic	no_assertion_criteria_ provided		
192	Pathogenic	no_assertion_criteria_ provided		
193	Likely_pathogenic	no_assertion_criteria_ provided		
194	Pathogenic	no_assertion_criteria_ provided		
195	Pathogenic	no_assertion_criteria_ provided		
196	Pathogenic	no_assertion_criteria_ provided		
197	Pathogenic	no_assertion_criteria_ provided		
198	Pathogenic	no_assertion_criteria_ provided		
199	Pathogenic	no_assertion_criteria_ provided		
200	Pathogenic	no_assertion_criteria_ provided		
201	Pathogenic	no_assertion_criteria_ provided		
202	Pathogenic	no_assertion_criteria_ provided		
203	Likely_pathogenic	no_assertion_criteria_ provided		
204	Pathogenic/Likely_ pathogenic	no_assertion_criteria_ provided		

205	Pathogenic	no_assertion_criteria_provided		
206	Pathogenic	no_assertion_criteria_provided		
207	Pathogenic	no_assertion_criteria_provided		
208	Pathogenic	no_assertion_criteria_provided		
209	Conflicting_interpretations_of_pathogenicity		Likely_pathogenic(2),Uncertain_significance(3)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[0, 1, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
210	Conflicting_interpretations_of_pathogenicity		Likely_pathogenic(4),Uncertain_significance(2)	InterVar: Likely pathogenic PVS1=0 PS=[0, 0, 0, 0, 0] PM=[1, 1, 0, 0, 0, 0, 0] PP=[0, 1, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[1, 0, 0, 0, 0, 0, 0, 0]
211	Conflicting_interpretations_of_pathogenicity		Pathogenic(2),Uncertain_significance(2)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[1, 1, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
212	Conflicting_interpretations_of_pathogenicity		Likely_pathogenic(1),Pathogenic(1),Uncertain_significance(1)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[0, 1, 0, 1, 0, 0, 0] PP=[0, 0, 0, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
213	Conflicting_interpretations_of_pathogenicity		Pathogenic(2),Uncertain_significance(2)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[1, 1, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
214	Conflicting_interpretations_of_pathogenicity		Likely_pathogenic(2),Uncertain_significance(1)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[0, 1, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
215	Conflicting_interpretations_of_pathogenicity		Likely_pathogenic(1),Pathogenic(1),Uncertain_significance(1)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[1, 1, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
216	Conflicting_interpretations_of_pathogenicity		Likely_pathogenic(2),Pathogenic(12),Uncertain_significance(1)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[1, 1, 0, 0, 0, 0, 0] PP=[0, 0, 0, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 1, 0, 0, 0, 0]
217	Conflicting_interpretations_of_pathogenicity		Pathogenic(1),Uncertain_significance(1)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[0, 1, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
218	Conflicting_interpretations_of_pathogenicity		Likely_pathogenic(2),Pathogenic(1),Uncertain_significance(3)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[0, 1, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
219	Conflicting_interpretations_of_pathogenicity		Pathogenic(1),Uncertain_significance(1)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[0, 1, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
220	Conflicting_interpretations_of_pathogenicity		Likely_pathogenic(2),Pathogenic(1),Uncertain_significance(1)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[1, 1, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
221	Conflicting_interpretations_of_pathogenicity		Likely_pathogenic(2),Pathogenic(1),Uncertain_significance(1)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[0, 1, 0, 0, 1, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
222	Conflicting_interpretations_of_pathogenicity		Likely_pathogenic(3),Pathogenic(7),Uncertain_significance(1)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[0, 1, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
223	Conflicting_interpretations_of_pathogenicity		Pathogenic(1),Uncertain_significance(5)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[1, 1, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
224	Conflicting_interpretations_of_pathogenicity		Likely_pathogenic(3),Pathogenic(13),Uncertain_significance(1)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[1, 1, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
225	Conflicting_interpretations_of_pathogenicity		Likely_pathogenic(1),Uncertain_significance(1)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[0, 1, 0, 0, 0, 0, 0] PP=[0, 0, 0, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
226	Conflicting_interpretations_of_pathogenicity		Likely_pathogenic(1),Pathogenic(1),Uncertain_significance(1)	InterVar: Likely pathogenic PVS1=1 PS=[0, 0, 0, 0, 0] PM=[0, 1, 0, 0, 0, 0, 0] PP=[0, 0, 0, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
227	Conflicting_interpr		Likely_pathogenic(1),Unce	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0]

	etations_of_patho genicity		rtain_significance(1)	PM=[1, 1, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
228	Conflicting_interpr etations_of_patho genicity		Likely_pathogenic(1),Path ogenic(1),Uncertain_signifi cance(2)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[1, 1, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[1, 0, 0, 0, 0, 0, 0, 0]
229	Conflicting_interpr etations_of_patho genicity		Likely_pathogenic(1),Unce rtain_significance(1)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[0, 1, 0, 0, 0, 0, 0] PP=[0, 0, 0, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
230	Conflicting_interpr etations_of_patho genicity		Likely_pathogenic(1),Unce rtain_significance(4)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[0, 1, 0, 0, 1, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
231	Conflicting_interpr etations_of_patho genicity		Likely_pathogenic(4),Path ogenic(3),Uncertain_signifi cance(4)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[1, 1, 0, 0, 0, 0, 0] PP=[0, 0, 0, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
232	Conflicting_interpr etations_of_patho genicity		Likely_pathogenic(1),Unce rtain_significance(1)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[1, 1, 0, 0, 0, 0, 0] PP=[0, 0, 0, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
233	Conflicting_interpr etations_of_patho genicity		Pathogenic(14),Uncertain_ significance(1)	InterVar: Likely pathogenic PVS1=0 PS=[0, 0, 0, 1, 0] PM=[1, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[1, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
234	Conflicting_interpr etations_of_patho genicity		Likely_pathogenic(1),Path ogenic(1),Uncertain_signifi cance(1)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[0, 0, 0, 0, 0, 0, 0] PP=[0, 0, 0, 0, 0, 0] BA1=0 BS=[1, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
235	Conflicting_interpr etations_of_patho genicity		Likely_pathogenic(1),Unce rtain_significance(1)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[1, 1, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
236	Conflicting_interpr etations_of_patho genicity		Likely_pathogenic(2),Unce rtain_significance(1)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[1, 1, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
237	Conflicting_interpr etations_of_patho genicity		Pathogenic(5),Uncertain_si gnificance(1)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[1, 1, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[1, 0, 0, 0, 0, 0, 0, 0]
238	Conflicting_interpr etations_of_patho genicity		Likely_pathogenic(3),Path ogenic(1),Uncertain_signifi cance(1)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[1, 1, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
239	Conflicting_interpr etations_of_patho genicity		Likely_pathogenic(1),Unce rtain_significance(5)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[1, 1, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
240	Conflicting_interpr etations_of_patho genicity		Likely_pathogenic(1),Path ogenic(17),Uncertain_signi ficance(1)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[0, 0, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[1, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
241	Conflicting_interpr etations_of_patho genicity		Pathogenic(1),Uncertain_si gnificance(1)	InterVar: Uncertain significance PVS1=1 PS=[0, 0, 0, 0, 0] PM=[0, 0, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
242	Conflicting_interpr etations_of_patho genicity		Likely_pathogenic(1),Path ogenic(1),Uncertain_signifi cance(1)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[0, 0, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
243	Conflicting_interpr etations_of_patho genicity		Pathogenic(1),Uncertain_si gnificance(2)	InterVar: Likely pathogenic PVS1=1 PS=[0, 0, 0, 0, 0] PM=[0, 1, 0, 0, 0, 0, 0] PP=[0, 0, 0, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
244	Conflicting_interpr etations_of_patho genicity		Pathogenic(1),Uncertain_si gnificance(2)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[0, 0, 0, 1, 0, 0, 0] PP=[0, 0, 0, 0, 0, 0] BA1=0 BS=[1, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
245	Conflicting_interpr etations_of_patho genicity		Pathogenic(1),Uncertain_si gnificance(7)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[0, 0, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
246	Conflicting_interpr etations_of_patho genicity		Likely_pathogenic(3),Unce rtain_significance(1)	InterVar: Pathogenic PVS1=1 PS=[0, 0, 0, 0, 0] PM=[0, 1, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
247	Conflicting_interpr etations_of_patho genicity		Likely_pathogenic(1),Path ogenic(1),Uncertain_signifi cance(1)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[1, 1, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
248	Conflicting_interpr etations_of_patho		Pathogenic(1),Uncertain_si gnificance(2)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[1, 1, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0,

	genicity			0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
249	Conflicting_interpretations_of_pathogenicity		Likely_pathogenic(2),Pathogenic(1),Uncertain_significance(1)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[1, 1, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
250	Conflicting_interpretations_of_pathogenicity		Likely_pathogenic(6),Pathogenic(13),Uncertain_significance(1)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[0, 0, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
251	Conflicting_interpretations_of_pathogenicity		Likely_pathogenic(2),Pathogenic(1),Uncertain_significance(1)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[0, 1, 0, 0, 0, 0, 0] PP=[0, 0, 0, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
252	Conflicting_interpretations_of_pathogenicity		Pathogenic(3),Uncertain_significance(6)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[0, 1, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
253	Conflicting_interpretations_of_pathogenicity		Pathogenic(1),Uncertain_significance(3)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[0, 1, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
254	Conflicting_interpretations_of_pathogenicity		Likely_pathogenic(1),Uncertain_significance(4)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[0, 1, 0, 0, 0, 0, 0] PP=[0, 0, 0, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
255	Conflicting_interpretations_of_pathogenicity		Pathogenic(1),Uncertain_significance(1)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[0, 1, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
256	Conflicting_interpretations_of_pathogenicity		Likely_pathogenic(1),Uncertain_significance(4)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[0, 1, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
257	Conflicting_interpretations_of_pathogenicity		Likely_pathogenic(1),Pathogenic(1),Uncertain_significance(1)	InterVar: Pathogenic PVS1=1 PS=[0, 0, 0, 0, 0] PM=[0, 1, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
258	Conflicting_interpretations_of_pathogenicity		Likely_pathogenic(2),Pathogenic(1),Uncertain_significance(1)	InterVar: Pathogenic PVS1=1 PS=[0, 0, 0, 0, 0] PM=[0, 1, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
259	Conflicting_interpretations_of_pathogenicity		Likely_pathogenic(1),Uncertain_significance(1)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[1, 0, 0, 0, 0, 0, 0] PP=[0, 1, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
260	Conflicting_interpretations_of_pathogenicity		Likely_pathogenic(2),Uncertain_significance(1)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[1, 0, 0, 0, 0, 0, 0] PP=[0, 1, 1, 0, 0, 0] BA1=0 BS=[1, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
261	Conflicting_interpretations_of_pathogenicity		Likely_pathogenic(1),Pathogenic(3),Uncertain_significance(3)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[1, 0, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[1, 0, 0, 0, 0] BP=[0, 0, 0, 0, 0, 0, 0, 0]
262	Conflicting_interpretations_of_pathogenicity		Likely_pathogenic(1),Uncertain_significance(2)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[0, 1, 0, 0, 0, 0, 0] PP=[0, 0, 0, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[1, 0, 0, 0, 0, 0, 0, 0]
263	Conflicting_interpretations_of_pathogenicity		Likely_pathogenic(1),Pathogenic(6),Uncertain_significance(2)	InterVar: Uncertain significance PVS1=0 PS=[0, 0, 0, 0, 0] PM=[1, 1, 0, 0, 0, 0, 0] PP=[0, 0, 1, 0, 0, 0] BA1=0 BS=[0, 0, 0, 0, 0] BP=[1, 0, 0, 0, 0, 0, 0, 0]

Supplementary table 5 Gene Carrier Rates (GCR) of genes associated with autosomal recessive disorder

Gene	P1	P2	P3	CoP
HBB	0.26397658	0.26519921	0.26519921	0.26519921
GJB2	0.22205647	0.22205647	0.22205647	0.22205647
HBA2	0.05952381	0.05952381	0.05952381	0.05952381
GALT	0.01655629	0.01655629	0.01655629	0.01655629
ABCA4	0.01642543	0.01967423	0.02129461	0.06579242
SLC26A4	0.01155656	0.01319035	0.01319035	0.01319035

CFTR	0.00992642	0.00992642	0.01324966	0.02318801
USH2A	0.00992574	0.01156222	0.01156222	0.01483794
SLC22A5	0.00989279	0.01152933	0.01152933	0.01806467
SLC25A13	0.00988461	0.00988461	0.00988461	0.00988461
AGXT	0.0066061	0.00824807	0.01152388	0.01152388
GBA	0.00660337	0.00824535	0.00824535	0.00824535
PKHD1	0.00501115	0.00501115	0.00501115	0.00665576
CYP21A2	0.00498618	0.00498618	0.00498618	0.01531504
SBDS	0.00497512	0.00497512	0.00497512	0.00497512
BEST1	0.00495868	0.00660337	0.00660337	0.00660337
RPGRIP1L	0.00495868	0.00495868	0.00495868	0.00495868
PAH	0.00495594	0.00825081	0.00825081	0.00825081
GUSB	0.00330579	0.00495321	0.00495321	0.00495321
FANCA	0.00330579	0.00495321	0.03625435	0.03625435
GAA	0.00330579	0.00495321	0.00495321	0.00659792
DHCR7	0.00330579	0.00330579	0.00330579	0.00330579
UROS	0.00330579	0.00330579	0.00330579	0.00330579
GNPTAB	0.00330305	0.00495049	0.00495049	0.00495049
COL4A3	0.00330305	0.00330305	0.00330305	0.00330305
PLA2G6	0.00330305	0.00330305	0.00330305	0.00330305
TYR	0.00170068	0.00170068	0.00170068	0.00170068
LYST	0.00167224	0.00167224	0.00167224	0.00167224
ALMS1	0.00166667	0.00498339	0.00498339	0.00498339
CEP290	0.00165837	0.00495595	0.00495595	0.00825897
CYP17A1	0.00165289	0.00495321	0.00495321	0.00495321
ATP7B	0.00165289	0.00330305	0.00330305	0.00987918
GNE	0.00165289	0.00330305	0.00330305	0.00824535
ABCC8	0.00165289	0.00330305	0.00330305	0.00330305
ACAD9	0.00165289	0.00330305	0.00330305	0.00330305
BRCA2	0.00165289	0.00330305	0.00330305	0.00330305
CYP27A1	0.00165289	0.00165289	0.01155386	0.01155386
LDLR	0.00165289	0.00165289	0.00330579	0.00660064
KLHL40	0.00165289	0.00165289	0.00330305	0.00330305
NPHS2	0.00165289	0.00165289	0.00165289	0.00825353
CNGB1	0.00165289	0.00165289	0.00165289	0.00330305
ACADS	0.00165289	0.00165289	0.00165289	0.00165289
AIRE	0.00165289	0.00165289	0.00165289	0.00165289
ARSA	0.00165289	0.00165289	0.00165289	0.00165289
ASS1	0.00165289	0.00165289	0.00165289	0.00165289
BBS2	0.00165289	0.00165289	0.00165289	0.00165289

CNGB3	0.00165289	0.00165289	0.00165289	0.00165289
COL6A1	0.00165289	0.00165289	0.00165289	0.00165289
FAH	0.00165289	0.00165289	0.00165289	0.00165289
FBP1	0.00165289	0.00165289	0.00165289	0.00165289
GALC	0.00165289	0.00165289	0.00165289	0.00165289
GYS2	0.00165289	0.00165289	0.00165289	0.00165289
IDUA	0.00165289	0.00165289	0.00165289	0.00165289
LAMB3	0.00165289	0.00165289	0.00165289	0.00165289
MMAB	0.00165289	0.00165289	0.00165289	0.00165289
MPL	0.00165289	0.00165289	0.00165289	0.00165289
MUTYH	0.00165289	0.00165289	0.00165289	0.00165289
OTOA	0.00165289	0.00165289	0.00165289	0.00165289
PEX7	0.00165289	0.00165289	0.00165289	0.00165289
POR	0.00165289	0.00165289	0.00165289	0.00165289
PRF1	0.00165289	0.00165289	0.00165289	0.00165289
PYGM	0.00165289	0.00165289	0.00165289	0.00165289
RDH12	0.00165289	0.00165289	0.00165289	0.00165289
NEB	0	0.00991736	0.00991736	0.00991736
CHKB	0	0.00495868	0.00495868	0.00495868
GJC2	0	0.00495868	0.00495868	0.00495868
HPS6	0	0.00495868	0.00495868	0.00495868
MYO15A	0	0.00495321	0.00495321	0.00495321
RPE65	0	0.00331958	0.00331958	0.00331958
ETFDH	0	0.00330305	0.00330305	0.00330305
IVD	0	0.00330305	0.00330305	0.00330305
TGM1	0	0.00330305	0.00330305	0.00330305
TMEM67	0	0.00330305	0.00330305	0.00330305
WNT10A	0	0.00330305	0.00330305	0.00330305
TMEM237	0	0.00165837	0.00165837	0.00165837
CC2D2A	0	0.00165563	0.00165563	0.00165563
FH	0	0.00165563	0.00165563	0.00165563
CNGA3	0	0.00165289	0.00330305	0.00330305
EVC	0	0.00165289	0.00330305	0.00330305
FAM161A	0	0.00165289	0.00330305	0.00330305
PDE6B	0	0.00165289	0.00330305	0.00330305
PROM1	0	0.00165289	0.00330305	0.00330305
PROC	0	0.00165289	0.00165289	0.01650434
POLG	0	0.00165289	0.00165289	0.00659792
GCDH	0	0.00165289	0.00165289	0.00330305
ALOX12B	0	0.00165289	0.00165289	0.00165289

ALPL	0	0.00165289	0.00165289	0.00165289
ATR	0	0.00165289	0.00165289	0.00165289
BCS1L	0	0.00165289	0.00165289	0.00165289
COL6A3	0	0.00165289	0.00165289	0.00165289
COL7A1	0	0.00165289	0.00165289	0.00165289
ESPN	0	0.00165289	0.00165289	0.00165289
ETFB	0	0.00165289	0.00165289	0.00165289
FUCA1	0	0.00165289	0.00165289	0.00165289
HPS3	0	0.00165289	0.00165289	0.00165289
INVS	0	0.00165289	0.00165289	0.00165289
LAMA2	0	0.00165289	0.00165289	0.00165289
LDLRAP1	0	0.00165289	0.00165289	0.00165289
MKKS	0	0.00165289	0.00165289	0.00165289
MTTP	0	0.00165289	0.00165289	0.00165289
MYO3A	0	0.00165289	0.00165289	0.00165289
NDUFV1	0	0.00165289	0.00165289	0.00165289
POMT1	0	0.00165289	0.00165289	0.00165289
RARS2	0	0.00165289	0.00165289	0.00165289
RECQL4	0	0.00165289	0.00165289	0.00165289
RMRP	0	0.00165289	0.00165289	0.00165289
SLC25A15	0	0.00165289	0.00165289	0.00165289
SLC7A9	0	0.00165289	0.00165289	0.00165289
TK2	0	0.00165289	0.00165289	0.00165289
VPS13A	0	0.00165289	0.00165289	0.00165289
VPS33B	0	0.00165289	0.00165289	0.00165289
F5	0	0	0.02479339	0.02479339
OTOF	0	0	0.00330579	0.01319036
FREM2	0	0	0.00330579	0.00330579
BTD	0	0	0.00165289	0.00824262
CHRNG	0	0	0.00165289	0.00165289
CP	0	0	0.00165289	0.00165289
CRTAP	0	0	0.00165289	0.00165289
GJB3	0	0	0.00165289	0.00165289
MAK	0	0	0.00165289	0.00165289
MOCS2	0	0	0.00165289	0.00165289
RPGRIP1	0	0	0.00165289	0.00165289
SCNN1A	0	0	0.00165289	0.00165289
USH1G	0	0	0.00165289	0.00165289
ZNF469	0	0	0.00165289	0.00165289
AMPD1	0	0	0	0.00165289

SMPD1	0	0	0	0.02809917
MOCS1	0	0	0	0.01818182
MLC1	0	0	0	0.00826446
TTN	0	0	0	0.00342261
DPYD	0	0	0	0.00331126
ILDR1	0	0	0	0.00330579
ACADM	0	0	0	0.00330305
ACADVL	0	0	0	0.00330305
AGL	0	0	0	0.00165563
FARS2	0	0	0	0.00165563
ASL	0	0	0	0.00165289
EYS	0	0	0	0.00165289
GLDC	0	0	0	0.00165289
HFE	0	0	0	0.00165289
LRPPRC	0	0	0	0.00165289
MEFV	0	0	0	0.00165289
PEX1	0	0	0	0.00165289
RAPSN	0	0	0	0.00165289
SGSH	0	0	0	0.00165289

VITA

NAME	John Mauleekoonphairoj
DATE OF BIRTH	19 November 1990
PLACE OF BIRTH	Bangkok
INSTITUTIONS ATTENDED	Chulalongkorn University
HOME ADDRESS	104/20 soi Ronnachai 2, Setsiri Rd. Samsen Nai
PUBLICATION	1: Chitcharoen S, Phokaew C, Mauleekoonphairoj J, Khongphatthanayothin A, Sutjaporn B, Wandee P, Poovorawan Y, Nademanee K, Payungporn S. Metagenomic analysis of viral genes integrated in whole genome sequencing data of Thai patients with Brugada syndrome. Genomics Inform. 2022 Dec;20(4):e44. doi: 10.5808/gi.22047. Epub 2022 Dec 30. PMID: 36617651; PMCID: PMC9847385.
	2: Chimparlee N, Prechawat S, Khongphatthanayothin A, Mauleekoonphairoj J, Lekchuensakul S, Wongcharoen W, Makarawate P, Sahasatas D, Krittayaphong R, Amnueypol M, Anannab A, Ngarmukos T, Vardhanabhuti S, Sutjaporn B, Wandee P, Veerakul G, Bezzina CR, Poovorawan Y, Nademanee K. Clinical Characteristics of SCN5A p.R965C Carriers: A Common Founder Variant Predisposing to Brugada Syndrome in Thailand. Circ Genom Precis Med. 2021

Jun;14(3):e003229. doi:
10.1161/CIRCGEN.120.003229. Epub 2021 Jun 7. PMID:
34092119.

3: Mauleekoonphairoj J, Vongpunsawad S,
Khongphatthanayothin A, Nademanee K,
Poovorawan Y. Genetic risks and association with severe
COVID-19 among global
populations. Pathog Glob Health. 2021 Jun;115(4):209-210.
doi:
10.1080/20477724.2021.1881371. Epub 2021 Feb 3. PMID:
33533704; PMCID:
PMC8168748.

4: Pasittungkul S, Lestari FB, Puenpa J, Chuchaona W,
Posuwan N, Chansaenroj J,
Mauleekoonphairoj J, Sudhinaraset N, Wanlapakorn N,
Poovorawan Y. High
prevalence of circulating DS-1-like human rotavirus A and
genotype diversity in
children with acute gastroenteritis in Thailand from 2016
to 2019. PeerJ. 2021
Feb 26;9:e10954. doi: 10.7717/peerj.10954. PMID:
33680579; PMCID: PMC7919534.

5: Makarawate P, Glinge C, Khongphatthanayothin A, Walsh
R, Mauleekoonphairoj J,
Amnueypol M, Prechawat S, Wongcharoen W,
Krittayaphong R, Anannab A, Lichtner P,
Meitinger T, Tjong FVY, Lieve KVV, Amin AS, Sahasatas D,
Ngarmukos T, Wichadakul

D, Payungporn S, Sutjaporn B, Wandee P, Poovorawan Y, Tfelt-Hansen J, Tanck MWT,
Tadros R, Wilde AAM, Bezzina CR, Veerakul G, Nademanee K. Common and rare
susceptibility genetic variants predisposing to Brugada syndrome in Thailand.
Heart Rhythm. 2020 Dec;17(12):2145-2153. doi: 10.1016/j.hrthm.2020.06.027. Epub
2020 Jun 30. PMID: 32619740.

6: Mauleekoonphairoj J, Chamnanphon M, Khongphatthanayothin A, Sutjaporn B,
Wandee P, Poovorawan Y, Nademanee K, Pongpanich M, Chariyavilaskul P. Phenotype
prediction and characterization of 25 pharmacogenes in Thais from whole genome
sequencing for clinical implementation. Sci Rep. 2020 Nov 3;10(1):18969. doi:
10.1038/s41598-020-76085-3. PMID: 33144648; PMCID: PMC7641128.

7: Saprungruang A, Khongphatthanayothin A, Mauleekoonphairoj J, Wandee P,
Kanjanauthai S, Bhuiyan ZA, Wilde AAM, Poovorawan Y. Genotype and clinical
characteristics of congenital long QT syndrome in Thailand. Indian Pacing
Electrophysiol J. 2018 Sep-Oct;18(5):165-171. doi: 10.1016/j.ipej.2018.07.007.
Epub 2018 Jul 20. PMID: 30036649; PMCID: PMC6198685.

8: Thongpan I, Mauleekoonphairoj J, Vichiwattana P, Korkong S, Wasitthankasem R,
Vongpunsawad S, Poovorawan Y. Respiratory syncytial virus genotypes NA1, ON1,
and BA9 are prevalent in Thailand, 2012-2015. PeerJ. 2017 Oct 27;5:e3970. doi:
10.7717/peerj.3970. PMID: 29085762; PMCID: PMC5661434.

9: Chansaenroj J, Auphimai C, Puenpa J, Mauleekoonphairoj J, Wanlapakorn N,
Vuthitanachot V, Vongpunsawad S, Poovorawan Y. High prevalence of coxsackievirus
A2 in children with herpangina in Thailand in 2015. Virusdisease. 2017
Mar;28(1):111-114. doi: 10.1007/s13337-017-0366-8. Epub 2017 Feb 14. PMID:
28466062; PMCID: PMC5377860.

10: Mauleekoonphairoj J, Puenpa J, Korkong S, Vongpunsawad S, Poovorawan Y.
PREVALENCE OF HUMAN ENTEROVIRUS AMONG PATIENTS WITH HAND, FOOT, AND MOUTH
DISEASE AND HERPANGINA IN THAILAND, 2013. Southeast Asian J Trop Med Public
Health. 2015 Nov;46(6):1013-20. PMID: 26867359.

11: Mauleekoonphairoj J, Vongpunsawad S, Puenpa J, Korkong S, Poovorawan Y.
Complete genome sequence analysis of enterovirus 71 isolated from children with
hand, foot, and mouth disease in Thailand, 2012-2014.

Virus Genes. 2015
Oct;51(2):290-3. doi: 10.1007/s11262-015-1239-0. Epub 2015 Aug 25. PMID:
26303899.

12: Puenpa J, Mauleekoonphairoj J, Linsuwanon P, Suwannakarn K, Chieochansin T,
Korkong S, Theamboonlers A, Poovorawan Y. Prevalence and characterization of
enterovirus infections among pediatric patients with hand foot mouth disease,
herpangina and influenza like illness in Thailand, 2012. PLoS One. 2014 Jun
2;9(6):e98888. doi: 10.1371/journal.pone.0098888. PMID: 24887237; PMCID:
PMC4041783.

13: Linsuwanon P, Puenpa J, Huang SW, Wang YF, Mauleekoonphairoj J, Wang JR,
Poovorawan Y. Epidemiology and seroepidemiology of human enterovirus 71 among
Thai populations. J Biomed Sci. 2014 Feb 18;21(1):16. doi: 10.1186/1423-0127-21-16. PMID: 24548776; PMCID:
PMC3937078.