國立臺灣大學電機資訊學院資訊工程學研究所

碩士論文

Department of Computer Science and Information Engineering
College of Electrical Engineering and Computer Science
National Taiwan University

Master's Thesis

透過機器學習結合多重基因風險指數預測心肌病變

Machine Learning Aggregates Polygenic Risk Scores for Cardiomyopathy Disease Prediction

許紅媛

Hung-Yuan Hsu

指導教授：賴飛羆 博士、莊志明 博士

Advisor: Feipei Lai, Ph.D., Jyh-Ming Jimmy Juang, Ph.D.

中華民國 112 年 5 月

May 2023

doi:10.6342/NTU202401923

誌謝

在此向所有支持和鼓勵我的人致以衷心的感謝。過去的兩年中，我在跨領域讀資工所的研究生生涯中經歷了許多挫折，但每一天都充實而有意義。這是因為得到了許多人的支持、鼓勵和指導。 首先，我要感謝我的指導教授 賴飛羆教授。在過去的兩年裡，他不僅在研究方法和態度方面給予我悉心的指導，還關心我的日常生活。他對於為人處世方面的教導對我產生了深遠的影響。

其次，我要感謝我的共同指導教授 莊志明教授。在研究期間他給予我寶貴的醫學和生物基因學方面的專業知識，並提供了寶貴的意見和指正。

此外在口試期間我要感謝口試委員 張哲瑋博士、李妮鍾博士、曾新育主治醫生的細心指正。他們提供了許多意見，使我的論文更加完備。在此向他們致以深深的謝意。

在碩士修業期間，我要感謝學長許靖。他在過去兩年中給予了我許多協助和寶貴建議。同時，也要感謝實驗室的夥伴們，我們相互磨礪，互相關懷。謝謝你們。最後，我要感謝我親愛的家人和秉洋。因為有你們的支持和鼓勵，我能夠毫無後顧之憂地朝著目標前進。你們的存在是我最大的動力。

再次感謝所有支持我的人，你們的幫助和關懷對我來說意義重大。我將永遠感激不盡。

許紅媛 謹誌于

國立臺灣大學資訊工程系

中華民國 112 年 05 月

doi:10.6342/NTU202401923

Acknowledgments

I would like to express my heartfelt gratitude to everyone who has supported and encouraged me. Over the past two years, I have experienced numerous challenges in my graduate career in interdisciplinary computer science. However, each day has been fulfilling and meaningful. This is due to the support, encouragement, and guidance I have received from many individuals.

First and foremost, I would like to thank my advisor, Professor Feipei Lai. In the past two years, he has provided me with meticulous guidance not only in research methods and attitude but also in caring for my daily life. His teachings on personal conduct have had a profound impact on me.

Next, I want to express my gratitude to my co-advisor, Professor Jyh-Ming Jimmy Juang. Throughout the research period, he has imparted valuable knowledge in medical and biological genetics and provided invaluable insights and corrections.

Furthermore, I would like to thank the members of the oral examination committee: Dr. Che Wei Chang, Dr. Ni-Chung Lee, and Dr. Hsin-Yu Tseng. Their meticulous guidance and numerous suggestions have made my thesis more comprehensive. I deeply appreciate their contributions.

During my master's studies, I am grateful to senior student Ching Hsu. He has provided me with much assistance and valuable advice over the past two years. I would also like to express my gratitude to my laboratory colleagues, who have supported and cared for each other. Thank you all.

doi:10.6342/NTU202401923

Lastly, I want to thank my beloved family and BingYang. Your support and encouragement have allowed me to pursue my goals without any worries. Your presence is my greatest motivation.

Once again, I extend my gratitude to everyone who has supported me. Your assistance and care have been of great significance to me. I will forever be grateful.

Hung-Yuan Hsu

Department of Computer Science & Information Engineering

National Taiwan University

doi:10.6342/NTU202401923

中文摘要

本研究旨在運用機器學習技術，結合多基因風險分數，預測心肌病的發生。我們採用臺灣大學醫學院附設醫院（NTUH）和臺灣生物資料庫（TWB）的資料集，首先進行全基因組關聯研究（GWAS），以確定單核苷酸多態性（SNPs）、二元特徵和年齡之間的相關性。隨後在多基因風險分數（PRS）分析中，我們從發現性 GWAS 中獲取具體權重（連續特徵的β值和二元特徵的對數比率）。計算目標樣本中所有個體的 PRS 後，這些分數可以應用於邏輯回歸分析中，預測與感興趣特徵有遺傳重疊的特徵。我們使用先進的機器學習模型和交叉驗證技術，評估 NTUH 和 TWB 數據集中預測心肌病發展的準確性。在評估中，我們考慮了多種心肌病特徵和預測因素，包括 PRS、作為潛在危險因素的臨床參數和 ICD-10 以及 ICD-10-CM。

關鍵字： 心肌病、臺灣生物資料庫、單核苷酸多態性、多基因風險分數、全基因組關聯研究、機器學習

doi:10.6342/NTU202401923

ABSTRACT

The main goal of this study is to utilize machine learning techniques to combine polygenic risk scores and predict the occurrence of cardiomyopathy disease. To achieve this, we employ datasets from the National Taiwan University Hospital and Taiwan Biobank and conduct initial genome-wide association studies to identify correlations between single nucleotide polymorphisms and phenotype [4].

Afterwards, for the analysis of polygenic risk scores, specific weights are derived from discovery genome-wide association studies. These weights are then used to calculate the polygenic risk scores for all individuals in the target sample. These scores can be utilized in a firth regression analysis to predict phenotype that are expected to have genetic overlap with the specific trait of interest, i.e., cardiomyopathy [4].

To evaluate the accuracy of predicting cardiomyopathy development, we use cutting-edge machine learning models and cross-validation techniques on both the National Taiwan University Hospital and Taiwan Biobank datasets. In our evaluation, we take into account various cardiomyopathy features and predictors, including polygenic risk scores, clinical parameters as potential risk factors, as well as ICD-10 and ICD-10-CM codes [9].

Keywords: cardiomyopathy, Taiwan Biobank, single nucleotide polymorphisms, genome-wide association studies, polygenic risk scores, machine learning

doi:10.6342/NTU202401923

CONTENTS

doi:10.6342/NTU202401923

doi:10.6342/NTU202401923

LIST OF FIGURES

doi:10.6342/NTU202401923

LIST OF TABLES

doi:10.6342/NTU202401923

Chapter 1 Introduction

1.1 Background

1.1.1 Cardiomyopathy

Cardiomyopathies (CDM) refer to a group of heart muscle disorders primarily characterized by dysfunction in the electrical or muscular aspects of the heart. Typically, these conditions lead to abnormal structure, function, and stress on the myocardium. Cardiomyopathies, as defined by the American Heart Association (AHA), encompass a wide spectrum of diseases that impact the heart muscle, often characterized by abnormal ventricular enlargement or dilation. These conditions may be localized to the heart or be part of a broader systemic disorder, potentially leading to cardiac-related mortality or gradual heart function decline linked to heart failure [1].

Traditionally, four common of cardiomyopathies have been classified based on their structural and physiological characteristics: hypertrophic cardiomyopathy, dilated cardiomyopathy, restrictive cardiomyopathy, and arrhythmogenic right ventricular cardiomyopathy/dysplasia. These categories include both genetic and non-genetic forms of the disease [1].

Dilated cardiomyopathy (DCM)

DCM is characterized by the enlargement and dilation of at least one ventricle, leading to systolic failure. This condition can arise from various factors, including genetic predisposition, as well as acquired causes like myocardial infarction, certain medications, toxins, inflammatory conditions, chest radiation, valve disorders, and

doi:10.6342/NTU202401923

long-standing severe hypertension. While DCM is typically observed in adults, the age of onset can vary significantly. Its prevalence rate is estimated to be approximately 1 in every 2,700 individuals. DCM exhibits genetic heterogeneity with multiple inheritance patterns. Pathogenic variations in DCM encompass diverse forms, such as missense mutations, nonsense mutations, splicing errors, and minor insertions or deletions [1].

Hypertrophic cardiomyopathy (HCM)

HCM is marked by an augmentation in the quantity of cardiac muscle cells. The root cause of this condition is frequently attributed to mutations in genes responsible for encoding sarcomeric proteins. These mutations give rise to myocyte disarray, which is a prominent characteristic of HCM. This genetic cardiac condition has a relatively high prevalence in the community, estimated to be 1 in every 500 individual Autosomal dominant inheritance is the predominant pattern observed in hypertrophic cardiomyopathy (HCM). Pathogenic variations in HCM can manifest in various forms, including splicing errors, nonsense mutations, missense mutations, and minor insertions or deletions (indels) [1].

Arrhythmogenic right ventricular cardiomyopathy (ARVC)

ARVC is an inheritable form characterized by the presence of fibrosis and fatty infiltration in the myocardium of the right ventricle, along with symptoms of ventricular tachycardia and ventricular fibrillation. The estimated incidence of ARVC in the general population is approximately 1 in every 1,000-1,250 individuals. However, in regions where extensive family screening is conducted, the prevalence appears to be higher [1].

Restrictive cardiomyopathy (RCM)

doi:10.6342/NTU202401923

RCM is a less common type with a significant genetic influence, though genetics can only account for approximately 75% of cases categorized as idiopathic RCM. This condition is distinguished by diastolic dysfunction and impaired ventricular filling, caused by heightened stiffness of the heart muscle, leading to abnormal ventricular relaxation. The prevalence of RCM is low, making up less than 5% of cases in both the United States and Europe [1].

Left ventricular non-compaction cardiomyopathy (LVNC)

LVNC cardiomyopathy is a condition characterized by abnormal development of the lower left chamber of the heart. Instead of having a firm and smooth structure, the left ventricle appears spongy and thickened. This cardiomyopathy is usually present from birth. According to experts, approximately 12 out of every 1 million individuals receive a diagnosis of LVNC cardiomyopathy each year [6].

Fabry disease

In 2008, the Genetic Counseling Center at Taipei Veterans General Hospital made an important finding through their research on newborn screening. They discovered that the occurrence rate of Fabry disease with cardiac involvement in Taiwan is remarkably high, reaching 1 in 1,600 individuals. This particular disease often remains asymptomatic during early stages of life, but many patients start experiencing gradual enlargement of the heart (cardiac hypertrophy) between the ages of 40 and 50. Unfortunately, these patients are frequently misdiagnosed with idiopathic cardiomyopathy, leading to a lack of targeted treatment options. Consequently, some may ultimately require a heart transplant. It is worth noting that highly effective medications for treating Fabry disease already exist. Missing the opportunity for early

doi:10.6342/NTU202401923

intervention and allowing irreversible consequences to unfold would be a source of great regret [5].

Amyloidosis (AL)

Amyloidosis refers to a collection of diseases characterized by the accumulation of protein clusters known as amyloid in various tissues of the body. Gradually, these proteins replace healthy tissue, resulting in the dysfunction and failure of the affected organ. Amyloidosis encompasses multiple forms, each presenting distinct manifestations and underlying causes [8]. In Western countries, the occurrence rate of amyloidosis is 1 case per 100,000 person-years [7].

Cardiac amyloidosis

Cardiac amyloidosis, also known as "stiff heart syndrome," occurs when abnormal amyloid deposits replace the healthy muscle tissue in the heart [8]. Cardiac amyloidosis leads to the development of restrictive cardiomyopathy due to the accumulation of proteins in the myocardium's extracellular space. These proteins have an unstable structure that causes them to aggregate, misfold and form deposits in the form of amyloid fibrils. The estimated incidence of Cardiac amyloidosis in the general population is approximately 3.69 in every 100,000 individuals [3].

1.1.2 Genome-wide Polygenic Risk Scores

The main objective of genome-wide association studies (GWAS) is to establish connections between genetic variations and observable traits. This is achieved by analyzing allele frequency differences among individuals who share a common ancestry but exhibit diverse phenotypes. GWAS involves scanning the entire genome of a large

doi:10.6342/NTU202401923

population to identify genetic variants associated with specific traits or diseases. It encompasses various genetic variations, such as copy-number variants and sequence variations, with single-nucleotide polymorphisms (SNPs) being the most commonly studied markers. The results of GWAS frequently show groups of correlated SNPs that collectively indicate a significant association with the investigated trait. These are commonly known as genomic risk loci [4], [19].

The practical implications of GWAS findings are broad, as they enable the evaluation of an individual's susceptibility to various physical and mental illnesses through genetic profiling. Genome-wide polygenic risk scores (PRS) derived from GWAS data provide insights into complex human phenotypes, including the risk for numerous significant multifactorial diseases. These diseases are often influenced by multiple genetic variants, each contributing a modest effect to the overall risk, similar to the risk prediction methods used for monogenic traits that rely on rare, highly penetrant mutations. Genome-wide polygenic risk scores have been extensively studied in relation to Alzheimer's disease, breast cancer, prostate cancer and coronary artery disease [19].

Subsequently, PRS utilizes the findings from GWAS to evaluate an individual's genetic risk for a specific disease or phenotype. The process of PRS involves combining the effect sizes of the identified or leading variants to calculate a score representing an individual's overall genetic risk. To accomplish this, each variant is assigned a weight based on its effect size, which is determined from separate large-scale GWAS studies. The effect size reflects the strength of the association between a specific variant and the provided phenotypes under investigation. Each individual's personalized PRS is determined by considering the number of risk variants they carry and the respective effect sizes of each variant. The PRS offers a comprehensive assessment of a sample's

doi:10.6342/NTU202401923

cumulative genetic risk by integrating multiple genetic variants, each contributing a small effect.

1.1.3 Objective

Due to the influence of genetic ethnicity and disease types on the results of GWAS, this study focuses on investigating gene variations and risk scores associated with cardiomyopathy in the Taiwanese population. Through GWAS, we will explore the correlation between cardiomyopathy and gene variations. We will utilize a large dataset of genetic information from the Taiwanese population to identify SNPs that may be associated with cardiomyopathy. These SNPs represent specific locations in the genome that can vary between individuals. By analyzing the genetic data of the Taiwanese population, we will determine whether there is a statistical correlation between specific SNPs and cardiomyopathy.

Additionally, we will apply PRS to predict the risk of cardiomyopathy. PRS integrates the impact of multiple SNPs into a unified score, quantifying a sample's genetic risk for cardiomyopathy. By considering the effect size of each SNP and the number of risk variants carried by an individual, we can calculate their polygenic risk score, which predicts their likelihood of developing cardiomyopathy.

The results of this study have significant clinical implications for the prediction and prevention of cardiomyopathy. By combining GWAS and PRS, we can provide personalized assessments of cardiomyopathy risk in the Taiwanese population, offering more accurate predictions and diagnostic insights. Furthermore, these research findings can guide valuable strategies for the prevention and treatment of cardiomyopathy.

doi:10.6342/NTU202401923

Chapter 2 Method

2.1 Dataset

2.1.1 NTUH dataset (case dataset)

The dataset utilized in this study originated from National Taiwan University Hospital (NTUH) and consisted of records from 263 patients. Each patient's diagnosis of cardiomyopathy and determination of the specific type were confirmed by cardiologists from the Department of Cardiology at NTUH. Within the cohort, there were 32 patients identified with dilated cardiomyopathy (DCM), 148 patients diagnosed with hypertrophic cardiomyopathy (HCM), 39 patients diagnosed with arrhythmogenic right ventricular cardiomyopathy (ARVC), 29 patients diagnosed with left ventricular non-compaction (LVNC), 10 patients diagnosed with Fabry disease, and 5 patients diagnosed with amyloidosis (AM). (See Figure 2 and Table 1)

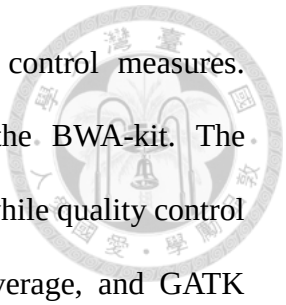
Taiwan Biobank

The Taiwan Biobank provides a comprehensive dataset, comprising demographic and health-related questionnaire, whole-genome sequencing (WGS) data from 1,495 individuals, and genotyping and imputed data from 103,332 individuals typed on the TWBv2 array. These datasets were aligned to GRCh38 mapping by the BWA-kit pipeline and jointly called using GATK. The acquisition of these datasets strictly adhered to ethical guidelines and received approval from the respective ethical committees of Academia Sinica and the Taiwan Biobank [10].

2.1.2 Genotyping of NTUH dataset

The exome sequencing and target sequencing data generated underwent processing

doi:10.6342/NTU202401923



also using GATK pipeline together with supplementary quality control measures. Specifically, the reads were mapped to GRCh38 aligning by the BWA-kit. The recalibration of base quality scores was carried out through GATK, while quality control statistics were generated using FASTQ, Sam Stats, BedTools Coverage, and GATK DepthOfCoverage.

In each sample, variant calling was performed using the GATK HaplotypeCaller, and subsequently joint-call genotyping was conducted across all samples using GATK's GenotypeGVCFs. Variant quality score recalibration (VQSR) was executed, and variants were filtered using a tranche sensitivity threshold of 0.997, along with mapping quality (min-base-quality-score 10) and base quality (minimum-mapping-quality 20) criteria for both SNVs and indels. We carried out further variant annotation using the Ensembl Variant Effect Predictor, ANNOVAR, InterVar, and Nirvana tools [10].

2.2 Imputation

The following process was employed to impute the 231 individuals with whole exome sequencing (WES) in order to expand the set of variants available for Genome-wide Association Studies.

doi:10.6342/NTU202401923

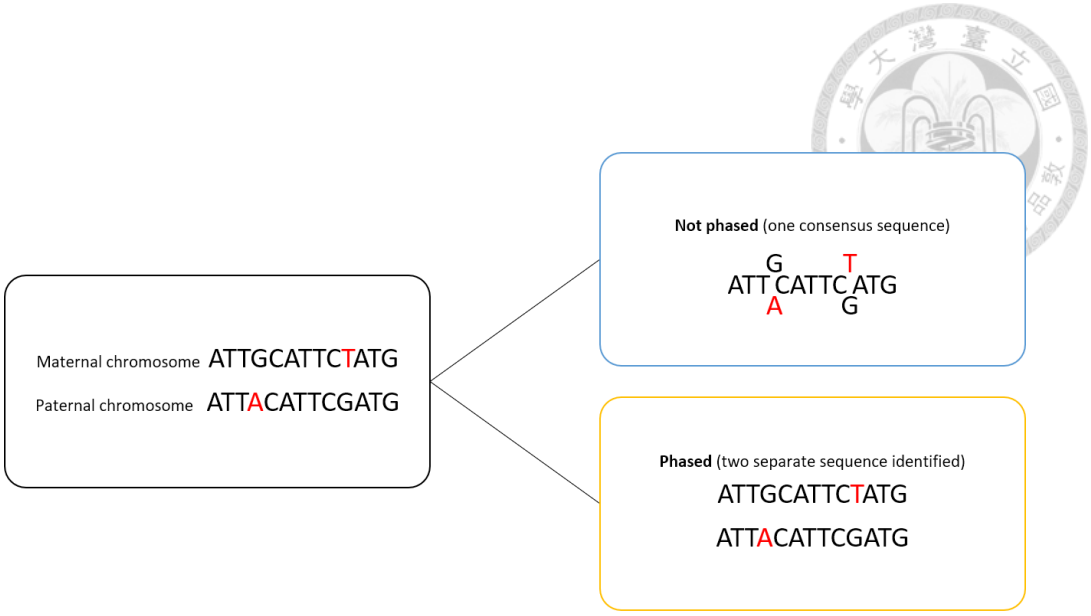


Figure 1 Chromosome sequence with phased or not [20]

Genotype phasing involves accurately assigning alleles in a diploid genome to either the paternal or maternal chromosomes, ensuring that alleles from the same parent are correctly placed on the corresponding chromosome. This process has various applications, including the generation of phased haplotype reference panels for genotype imputation and pre-phasing of study subjects to enhance the accuracy of genotype imputation.

There are three primary methods for phasing, and in this study, linkage disequilibrium (LD) phasing was utilized. LD phasing utilizes a large population of unrelated individuals and leverages the principles of LD and Hidden Markov Models (HMM) to infer haplotypes for each individual in the population. It is the most computationally intensive method among the three.

During meiosis, the cell division process that produces reproductive cells, sister chromatids undergo recombination at a rate of approximately 10^{-8} per generation. In theory, over a long period of time, the two chromosomes originating from a common

doi:10.6342/NTU202401923

ancestor will undergo recombination uniformly. However, due to the relatively small number of generations humans have experienced, chromosome recombination is still limited. Consequently, many adjacent regions in the human genome remain "linked" together and are inherited as blocks, a phenomenon known as linkage disequilibrium. These blocks are referred to as "linkage disequilibrium blocks."

The presence of linkage disequilibrium blocks enables the inference of linkage relationships by constructing hidden Markov models (HMM). However, the accuracy of this method is influenced by the size of the population used. SHAPEIT2 [17], which utilizes HMM, is the chosen tool for pre-phasing in this study.

The primary concept behind genotype imputation is to complete or infer missing genotypes in a dataset using information from one or multiple reference panels, such as UK biobank sequencing data or the Haplotype Reference Consortium (HRC). By leveraging algorithms and utilizing non-missing markers surrounding the missing ones, the imputation process infers the most likely haplotype for the region in question [21].

In this research, we carried out genotype imputation using IMPUTE2 [18]. The reference panel for imputation consisted of whole-genome sequencing (WGS) data from the Taiwan Biobank. This reference panel provided the essential information needed to predict and impute missing genotypes in the target dataset.

2.2.1 Establish a reference haplotype panel exclusively for the Taiwanese population

A dataset of 1,495 individuals with whole genome sequences from the Taiwan Biobank was utilized. From this dataset, a total of 7,801,712 biallelic variant sites with minor allele frequency (MAF) > 1%, and a call rate of > 95% for both variants and individuals were selected for computational phasing. The phasing process was

doi:10.6342/NTU202401923

performed using SHAPEIT2 [17], [10].

2.2.2 Impute the 231 individuals with WES on Taiwanese-specific haplotype panel

To begin the data processing, we performed genotype phasing using SHAPEIT2 [17], followed by imputation for chromosome 1 to 22 using IMPUTE2 [18] based on previous established reference panel. For IMPUTE2, we set the chunk size as 7kb and the buffer size as 250 kb [11]. The allelic dosage was examined, and genotypes were called if the posterior likelihood was greater than 0.9 (using PLINK2.0 --hard-call-threshold [23]). If the likelihood did not satisfy this criterion, the genotype was regarded as missing data. The expected dosages were directly calculated using the posterior genotype likelihoods obtained from IMPUTE2. Variants with an estimated information score of less than 0.3 were excluded from further analysis [10], [12]. Additionally, we filtered out minor allele frequency of SNPs lower than 0.1% and a call rate of SNPs below 95% before proceeding with downstream analysis [10]. (See Table 2, Figure 3, Figure 4 and Figure 5)

2.2.3 Validating imputation process

To assess the accuracy of the imputed genotypes in comparison with whole-genome sequences, a random subset of 231 samples was selected from the TWB WGS cohort and subjected to WES. Following that, the same imputation pipeline utilized previously was applied. Moreover, an independent set of 1,264 samples with phased haplotypes was used as the reference panel, separate from the TWB WGS cohort. Variants with an estimated information score of less than 0.3 were removed from the analysis. The level of agreement between the whole genome sequencing data and the imputed genotypes for every variant served as a validation metric. The validation

doi:10.6342/NTU202401923

process demonstrated a high level of concordance between the imputed genotypes and the sequence data, confirming the accuracy of the imputation pipeline [10]. (See Table 3)

Definition of Datasets

Dataset: (2,090 individuals, 17,944,587 variants; 1,426 males: 664 females; 506 cases: 1,584 controls)

- Cases from NTUH dataset
  - Target sequencing probes:
    - 16 individuals, 7,297 variants
  - WES SURESELECT_V6_COSMIC:
    - 4 individuals, 174,788 variants
  - WES Roche KAPA HyperExome:
    - 12 individuals, 1,133,680 variants
  - Imputed dataset [13]
    - WES SureSelect V6-Post: 231 individuals, 902,617 variants
- Age- and sex-matched Controls: (Exclude individuals who self-reported having cardiomyopathy in the conducted survey.) (See Figure 6 and Figure 7)
  - Selected from TWBv2.0 SNP genotyped_imputed (TWBv2.0_imputed):
    - 1,056 individuals, 16,211,759 variants
  - Selected from TWB Whole genome sequencing (TWB_WGS):
    - 1,052 individuals, 8,447,390 variants

doi:10.6342/NTU202401923

2.3 Genotype Quality Control

In any GWAS aiming to find common variant associations, the incorporation of appropriate quality control (QC) measures is crucial. Thorough QC is essential to ensure the reliability of GWAS results, as raw genotype data may contain inherent imperfections. Various factors can contribute to data errors, including insufficient DNA hybridization to the array, subpar DNA sample quality, potential sample mix-ups or contamination and less effective genotype probes. By implementing robust QC procedures, these issues can be identified and addressed, ensuring the accuracy and validity of the GWAS findings [4]. (See Figure 8 and Figure 9)

The term "missingness of SNPs" (PLINK2.0 [23] --Geno) refers to the proportion of missing values for a specific SNP across all samples, which helps identify low-quality SNPs to be excluded [16]. One of the key characteristics of SNPs is their frequency within a particular population. Different sets of variants may be used for various downstream analyses. For instance, common variants may be used for principal components analysis in order to help correct for population stratification, while rare variants may be used for gene-based tests [16]. SNPs with a lower minor allele frequency are considered rare and are less powerful in detecting associations between SNPs and phenotypes (PLINK2.0 [23] --maf). Additionally, these SNPs are more susceptible to genotyping errors. The choice of the minor allele frequency threshold should depend on the sample size [4]. The Hardy-Weinberg equilibrium is a basic concept that asserts genetic variation in a population remains stable over generations in the absence of disruptive factors [22]. Following the performance of the exact Hardy-Weinberg test (PLINK2.0 [23] --hwe), variants with low p-values often indicate genotyping errors or suggest evolutionary selection for these variants [16].

doi:10.6342/NTU202401923

2.4 Rare-Variant Association Test

Analyzing rare variants presents more significant challenges than studying common variants. Firstly, a considerable sample size is necessary to have a reasonable probability of observing a rare variant. Secondly, the conventional single-variant association analysis lacks the ability to identify associations with uncommon variants due to its limited power [25]. The conventional method for examining the connection between genetic variants and complex phenotypes is the single variant test which was conducted by an additive genetic model. This approach evaluates the association between every variant and a phenotype using logistic regression for binary phenotype. When ample sample sizes are available, standard single variant tests can also detect associations with rare variants [25]. Upon conducting a quality check on our combined sequencing dataset, it is evident that both the allele frequency and allele counts are relatively low. (See Figure 9)

Nevertheless, the power of single-variant tests decreases for rare variants in comparison to common variants that possess the same effect sizes. It is essential to acknowledge that p-value estimates obtained through standard regression methods might lack accuracy when the variant is observed in a limited number of subjects. As an alternative, aggregation tests evaluate the collective impact of multiple genetic variants within a region or Geno, enhancing the statistical power when multiple variants within the group are associated with particular phenotypes. Various approaches have been devised to achieve this goal [25].

Burden tests encompass the combination of data from several genetic variants into a singular genetic score, which is subsequently assessed for its association with a

doi:10.6342/NTU202401923

specific trait. In burden tests, a common approach involves calculating the sum of minor alleles across all variants within a given set [25].

Variance-component tests employ a random effects model to assess the association by analyzing the individual distribution of genetic effects for an assemblage of variants, taking into account their respective weights. In the context of binary traits, estimating p-values based on large sample sizes may lead to inaccurate type I error rates when the number of samples or total minor allele counts is limited. To tackle this concern, SKAT [30] which was optimally combined the burden test and the non-burden sequence kernel association test, which is a variance-component test, adopts a moment-based approach to adjust the null distribution by accurately estimating kurtosis of the test statistic and the variance with small sample size [25].

Variance-component tests demonstrate increased power when the causal variants exhibit diverse directions of association or when a region contains numerous noncausal variants. On the other hand, burden tests exhibit greater power when a region is primarily composed of causal variants showing the same direction of association. Considering the presence of both scenarios, combining burden and variance-component tests is deemed advantageous. One approach to combine the p-values from these two tests is Fisher's method, with permutation being used to assess the significance of the test [25].

The utilization of the dataset involves the integration of data from various individuals originating from different genome probe panels. Additionally, observations within the dataset reveal occurrences of rare allele frequencies and low allele counts. Considering that the patients are affected by cardiomyopathy, which is a rare condition rather than a common disease, the association test provided by PLINK2.0 [23] offers the choice between a logistic regression model and a Firth regression model for binary

doi:10.6342/NTU202401923

phenotypes. The selection of the Firth regression model is particularly appropriate when the minor allele count is less than 400 and a joint analysis is conducted. Notably, the Firth regression model incorporates the Fisher information matrix which addresses the combination of burden and variance-component tests mentioned earlier. Logistic regression, on the other hand, is more prone to generating biased results in separated datasets. On the other hand, Firth regression utilizes a penalized likelihood function to adjust for the first-order asymptotic bias of parameter estimates [25][26].

In the context of the polygenic risk scores model, there are options for handling heterozygous genotypes or SNPs with small effect sizes. Manual adjustments can be applied to the weights assigned to these SNPs. Additionally, researchers have the choice to deviate from the traditional additive genetic model and explore three alternative options: dominant, recessive, and heterozygous models [16]. (See Table 4) These choices have the potential to yield diverse outcomes in the calculation of polygenic scores [27].

For accurate analysis, Prsice-2 [27] offers empirical association P-values that mitigate inflation resulting from overfitting. Empirical P-values are computed using actual observed data, allowing researchers to make informed decisions about accepting or rejecting the null hypothesis [27].

doi:10.6342/NTU202401923

Chapter 3 Result

The study examines two distinct processes. The first process (Figure 10) involves patients diagnosed with cardiomyopathy, comprising samples confirmed by cardiologists from National Taiwan University Hospital. Controls are drawn from the Taiwan Biobank, randomly selected from samples with the exclusion of individuals who responded positively to the "CARDIOMYOPATHY_SELF" questionnaire item. An effort is made to match cases and controls based on age and sex. The combined cases and controls are then partitioned into separate groups for GWAS and PRS analyses.

In the second process (Figure 11), GWAS data are used, with both cases and controls sourced from the Taiwan Biobank. Phenotype classification is carried out based on the "CARDIOMYOPATHY_SELF" questionnaire item. Effect sizes obtained are applied to the polygenic risk model. For the polygenic risk model's target set, cases are selected from samples diagnosed by cardiologists at the National Taiwan University Hospital, while controls are drawn from the Taiwan Biobank (separate dataset from patients used in GWAS), with samples excluded if they responded affirmatively to the "CARDIOMYOPATHY_SELF" questionnaire item, and selected randomly.

Why is the experiment divided into these two parts? The main reason is that we cannot ascertain whether the "CARDIOMYOPATHY_SELF" information from the Taiwan Biobank questionnaire accurately reflects diagnoses made by cardiologists. Therefore, we aim to conduct these two parts of the experiment to make a determination.

The substantial disparity in the GWAS results between these two groups is likely attributed to two main factors. Firstly, in the first process, the genetic makeup of individuals is derived from diverse probe panels, contributing to significant variations in

doi:10.6342/NTU202401923

the results. Secondly, there exists a substantial difference in the overall sample size between the two groups. Sample sizes have a significant impact on GWAS results. (See Figure 12 and Figure 13)

After conducting the genome-wide association test for flowchart-1, we identified 456 leading variants. From the flowchart-1 GWAS results, we observed two common variants, MYH7 and LMNA [1], which are in close proximity to the genes associated with dilated cardiomyopathy (DCM), as well as one gene, TRDN [29], related to hypertrophic cardiomyopathy (HCM) that was reported in HCM GWAS literature. (See Table 5)

Moreover, the flowchart-2 GWAS results revealed two common variants, SGCD associated with DCM, and Plakophilin 2 (PKP2) related to arrhythmogenic right ventricular cardiomyopathy (ARVC) [1], which are nearest to their respective associated genes. Additionally, we observed five genes, namely TTN [29], MAP3K7CL [29], TBX3 [29], SLC35F1 [28], and FNDC3B [28], associated in HCM GWAS literatures.

Regarding the polygenic risk score (PRS) results from flowchart-1, both the validation and test sets exhibited excellent discriminative ability between cases and controls. However, for flowchart-2 PRS results, they did not perform as expected. This could be attributed to significant differences in the sources of GWAS and PRS cases, as well as the usage of distinct probe panels. (See Figure 14)

doi:10.6342/NTU202401923

Chapter 4 Conclusion and Future Work

As quality control was not performed on the sequencing data beforehand, similar to what was mentioned in section 2.3, numerous false-positive SNPs might have been included. Next, we will further utilize machine learning models to analyze which genetic variants are associated with cardiomyopathy from leading variants of GWAS result. Our goal is to determine whether some variants are specific to the Taiwanese population as associated loci. Subsequently, we will incorporate these genetic variants along with clinical measurements obtained from cardiomyopathy patients diagnosed by the cardiology department at the National Taiwan University Hospital. By doing so, we aim to differentiate the severity of the disease among the cases. This approach has the potential to be applied in practical scenarios involving actual cases of cardiomyopathy.

doi:10.6342/NTU202401923

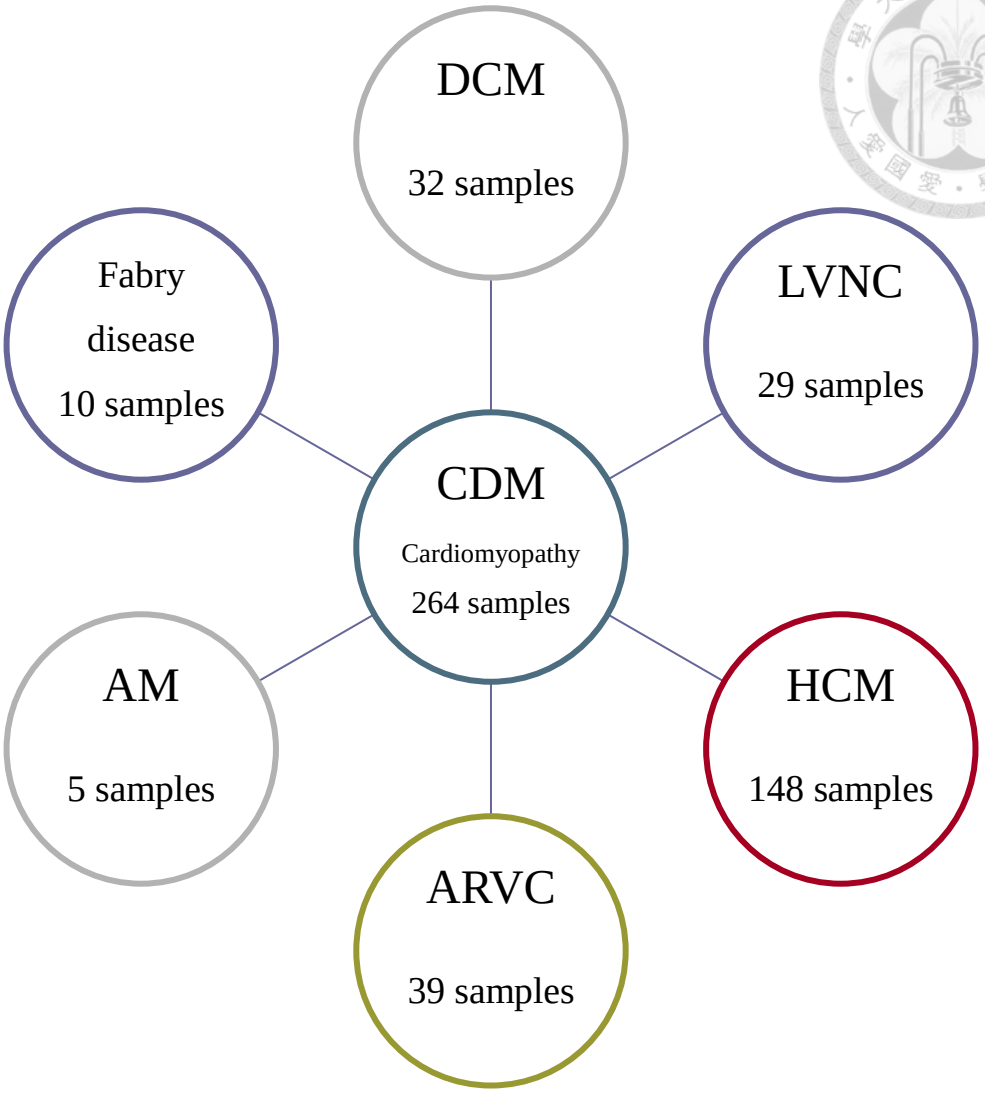


Figure 2 Types of cardiomyopathy in the case dataset

doi:10.6342/NTU202401923

Table 1 NTUH dataset (case dataset) individual information by different probes (WES = Whole-Exome Sequencing, PCV = Target Sequencing, #Avg variants: average number of variants for an individual)

Types	Probes	#Samples	#Males	#Females	#SNPs	# Avg SNPs
WES	SureSelect_V6_Post	231	156	75	25,153,106	**108,888**
	Roche KAPA HyperExome	12	8	3	1,133,680	**94,473**
	SureSelect_V6_ COSMIC	4	3	1	406,462	**101,615**
PCV	PCV_V1	5	4	1	10,227	**2,054**
	PCV_V2	6	4	2	14,832	**2,472**
	PCV_V3	5	3	2	17,060	**3,412**
	Total	**263**	**178**	**84**		

doi:10.6342/NTU202401923

Table 2 Imputation result of SureSelect V6 Post Probes: Number of Variants

WES Type	#Samples	#Variants	#Imputed variants	#Imputed variants (INFO > 0.3)	Amplification Factor
SureSelect V6_Post	231	692,863	7,999,916	6,068,851	8.76

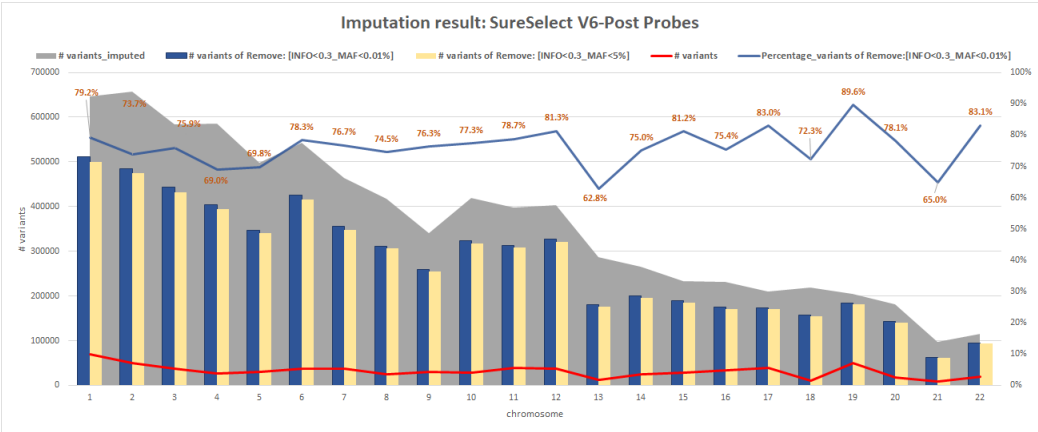


Figure 3 SureSelect V6 Post Probes: number of variants by chromosomes after imputation

doi:10.6342/NTU202401923

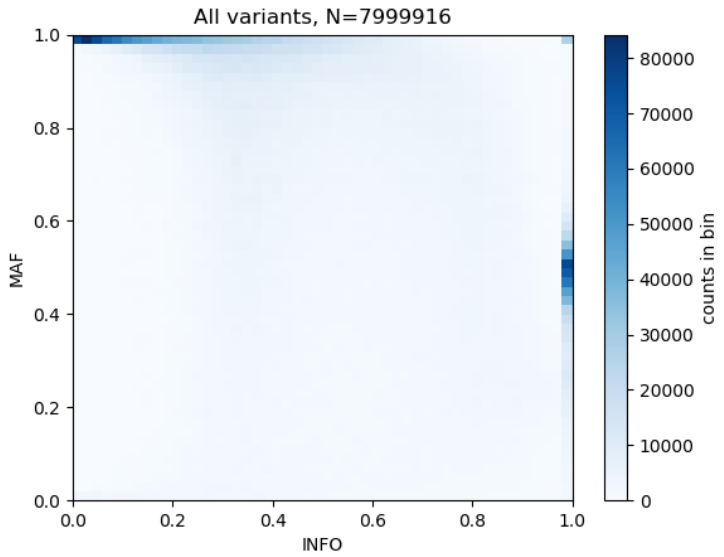


Figure 4 Imputation result of SureSelect V6-Post Probes: Scatterplot of information scores and minor allele frequency at variants in the imputed dataset

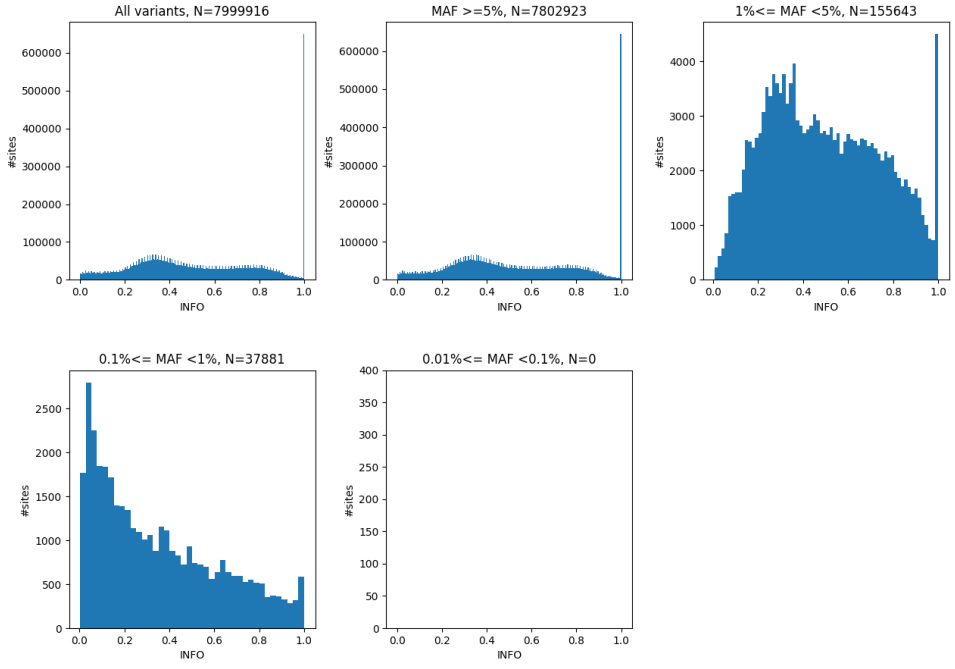


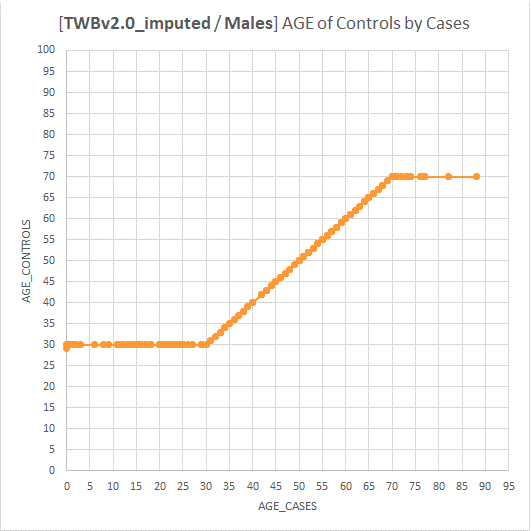
Figure 5 Imputation result of SureSelect V6-Post Probes: Distribution of information scores at variants in the imputed dataset. The x-axis shows the information score on the scale 0 to 1.

doi:10.6342/NTU202401923

Table 3 Comparison the imputed result: A file: raw whole genome sequencing 231 samples and B file: imputed Whole-Exome Sequencing 231 samples, only included sites with INFO score ≥ 0.3. Column 2 means that Compare pairs (REF_A, REF_B) and (ALT_A, ALT_B) on matched sites. Column 3 means that sites may be not imputed or removed by INFO score < 0.3.

CHROM	# different_paires count (only on sites matched)	% sites-skipped-no-match
Chr1	0	2.10%
Chr2	0	2.51%
Chr3	0	1.08%
Chr4	0	0.64%
Chr5	0	3.68%
Chr6	0	1.96%
Chr7	0	4.44%
Chr8	0	3.51%
Chr9	0	5.24%
Chr10	0	0.48%
Chr11	0	1.79%
Chr12	0	0.18%
Chr13	0	2.87%
Chr14	0	2.37%
Chr15	0	6.54%
Chr16	0	10.95%
Chr17	0	10.91%
Chr18	0	5.68%
Chr19	0	4.65%
Chr20	0	2.19%
Chr21	0	12.77%
Chr22	0	6.85%
ChrX	0	9.35%

doi:10.6342/NTU202401923



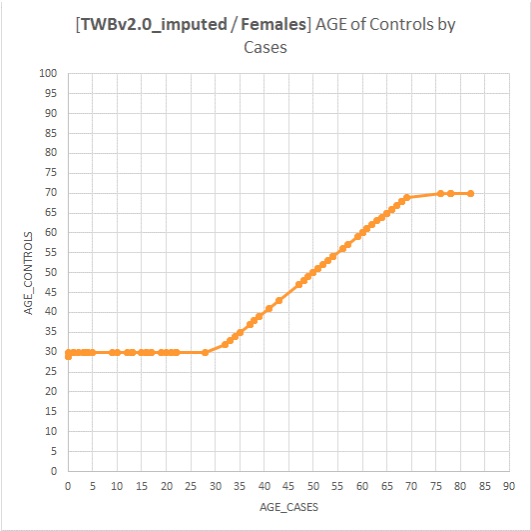
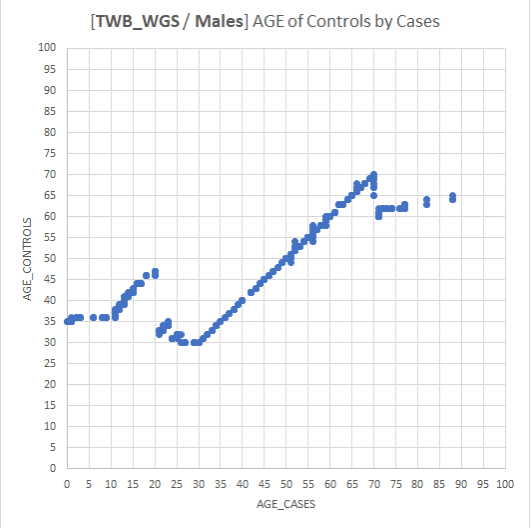


Figure 6 Scatterplot of Ages for Controls and Cases (TWBv2.0_imputed)



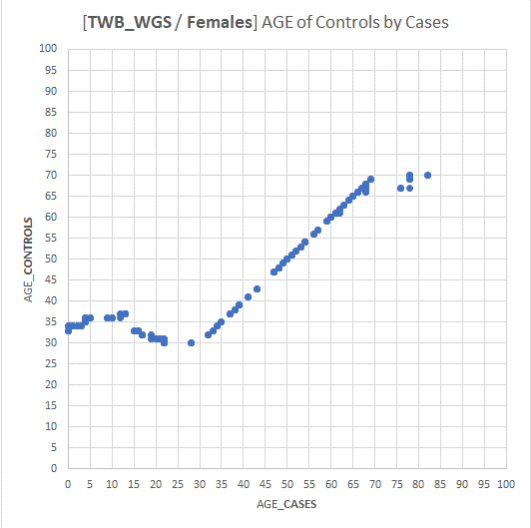


Figure 7 Scatterplot of Ages for Controls and Cases (TWB_WGS)

doi:10.6342/NTU202401923

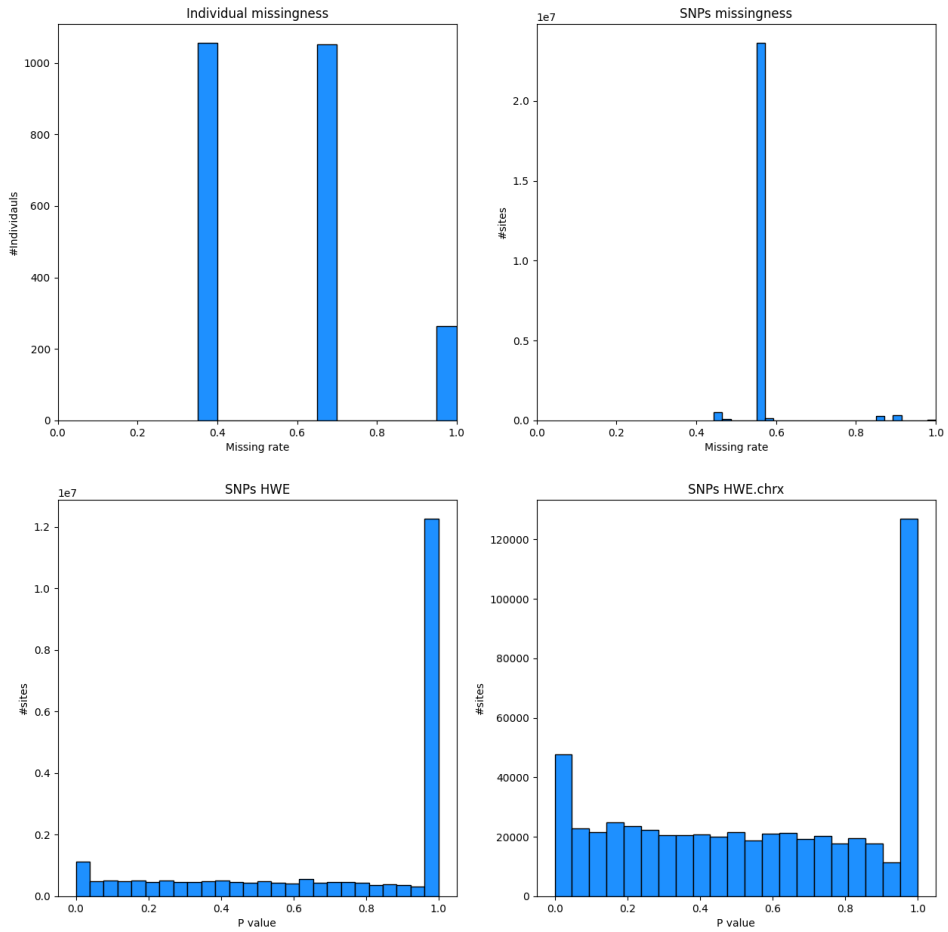


Figure 8 Quality Check of the dataset: [0, 0]: Missingness per individual; [0, 1]: Missingness per SNP; [1, 0]: Hardy-Weinberg equilibrium for autosomal chromosomes; [1, 1]: Hardy-Weinberg equilibrium for chromosome X

doi:10.6342/NTU202401923

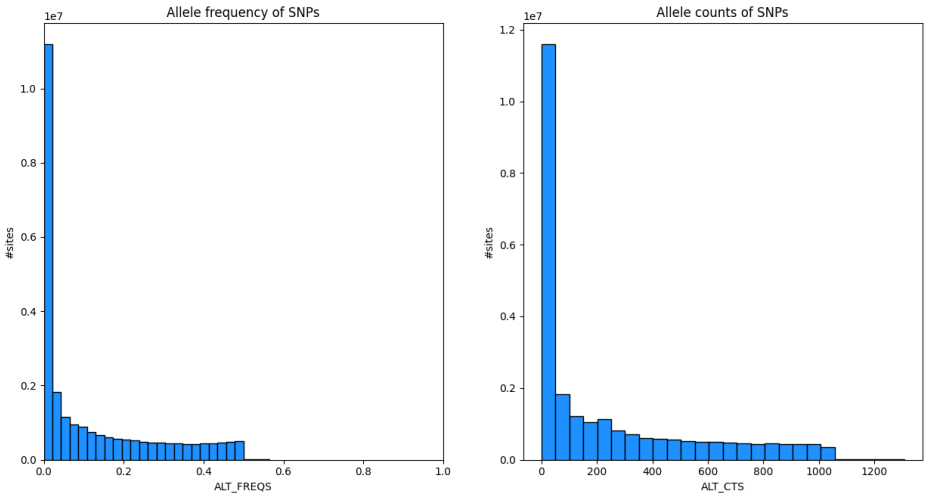


Figure 9 Quality Check of the dataset: [0, 0]: Allele frequency per SNP; [0, 1]: Allele count per SNP

Table 4 Different genetic models: As an illustration, let's consider a biallelic SNP with allele variations, where the reference allele is represented by A, and the alternative allele is denoted as G.

Genetic model	AA	AG	GG
Additive model	0	1	2
Dominant model	0	1	1
Recessive model	0	0	1
Heterozygous model	0	1	0

doi:10.6342/NTU202401923

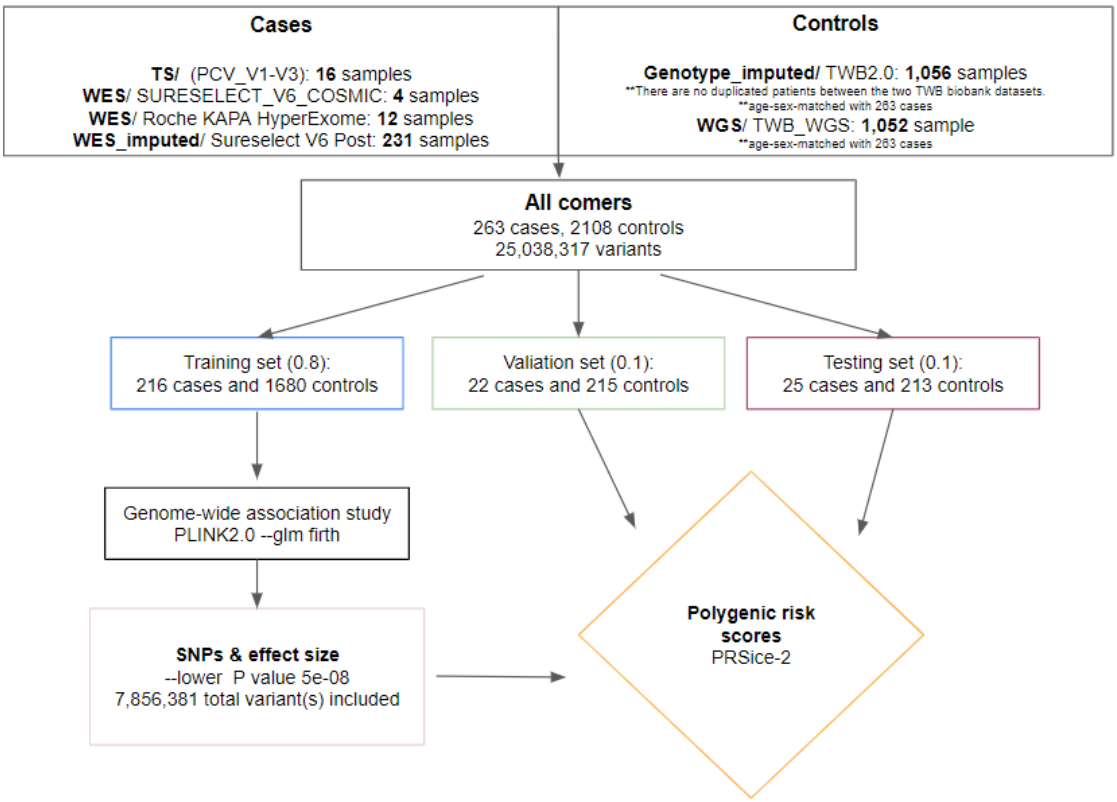


Figure 10 Flowchart-1 (TS: Target sequencing, PLINK2.0 [23], PRSices-2 [24])

doi:10.6342/NTU202401923

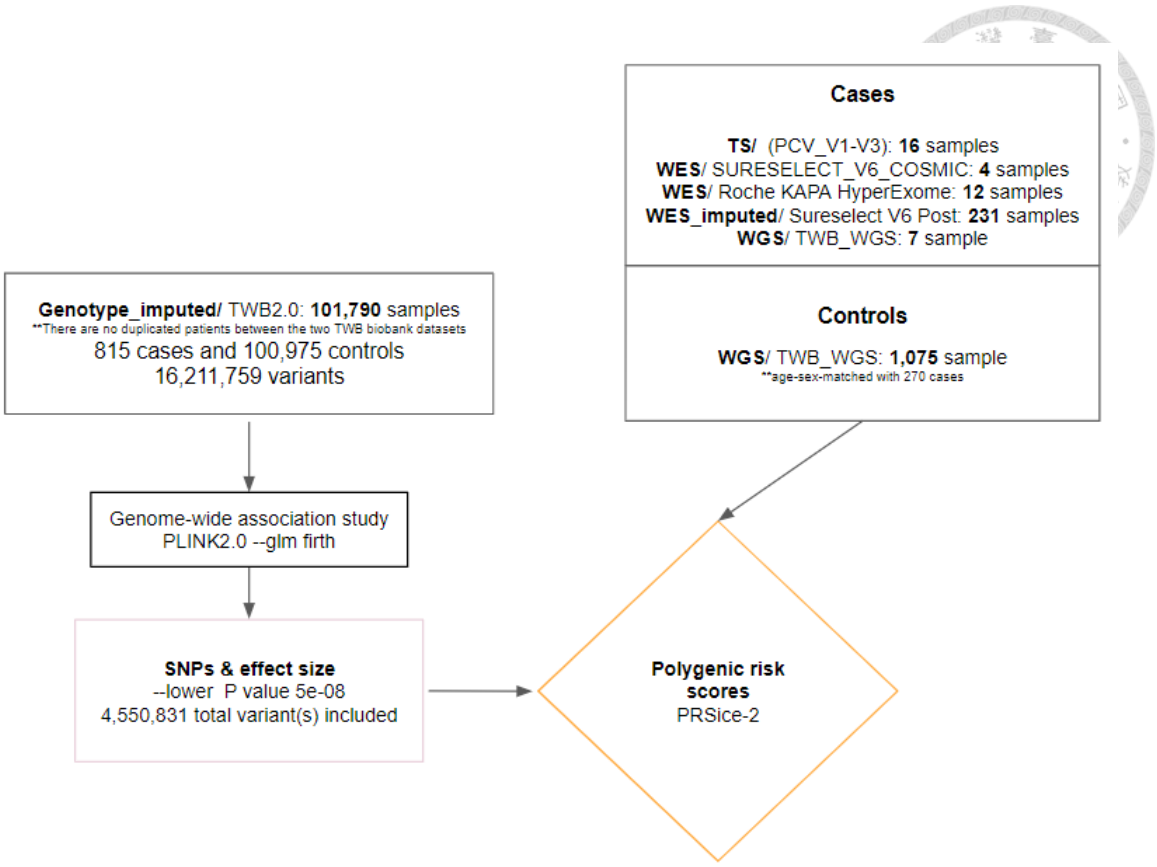


Figure 11 Flowchart-2 (TS: Target sequencing, PLINK2.0 [23], PRSices-2 [24])

doi:10.6342/NTU202401923

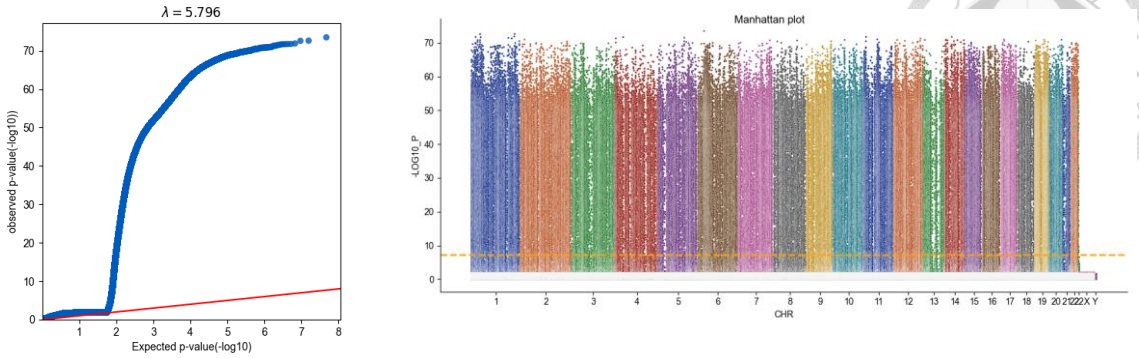


Figure 12 Qqplot and Manhattan plot of flowchat-1

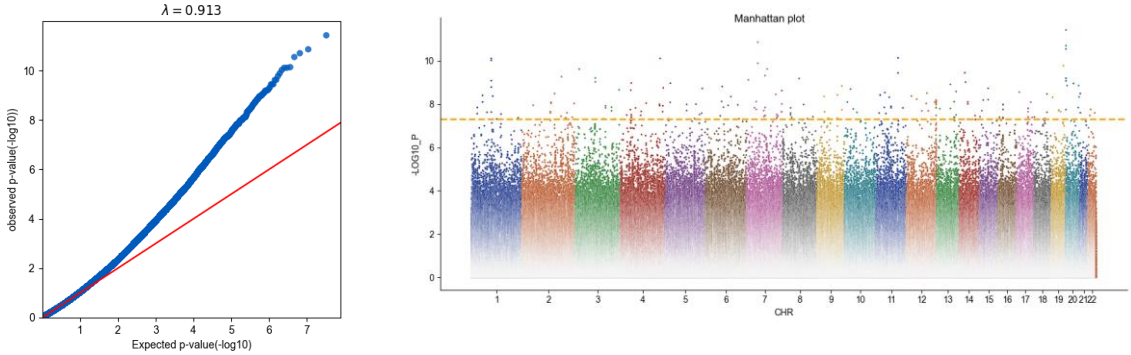


Figure 13 Qqplot and Manhattan plot of flowchat-2

Table 5 Variants associated with cardiomyopathy of flowchart-1: Some of these variants are relevant to the genes mentioned in the research papers (marked *), while the rest share the same nearest gene positions as those found in the flowchart-2 GWAS results.

SNPID	GRCh38	EA	NEA	SE	P	OR	Nearest GENE
rs184473115	1:4769774	G	C	0.267808	2.60E-70	115.285	AJAP1
rs2453058	1:119938086	T	C	0.268766	9.29E-70	115.019	NOTCH2
rs66542898	1:233263172	A	AG	0.318953	1.41E-69	276.894	PCNX2
rs73087440	1:215629059	G	A	0.338717	6.00E-69	381.582	USH2A
rs112542201	1:118086000	C	G	0.262919	3.62E-68	98.2163	SPAG17
rs11264443	1:156134601	T	C	0.318592	1.14E-67	254.077	LMNA*

doi:10.6342/NTU202401923

rs3790387	1:97679300	C	T	0.309029	1.62E-67	213.81	DPYD
rs113173951	1:27550907	A	G	0.313041	2.15E-67	228.089	AHDC1
rs11682884	2:224844919	A	G	0.274944	4.33E-71	134.499	DOCK10
rs191299362	2:235740708	C	A	0.258167	8.37E-70	95.5348	AGAP1
rs11902052	2:140444800	C	G	0.310142	3.09E-69	233.812	LRP1B
rs894197	2:88103041	A	G	0.292127	4.69E-69	169.145	SMYD1
rs61553904	2:164730183	C	T	0.251911	5.87E-69	83.1989	COBLL1
rs149253061	2:158456313	G	A	0.336069	4.47E-68	350.509	CCDC148
rs143460294	2:40175528	A	G	0.263762	5.76E-68	98.9761	SLC8A1
rs62142981	2:50921602	A	G	0.342567	2.42E-67	379.734	NRXN1
rs148827158	2:79026157	G	T	0.259824	8.91E-67	88.7112	REG3G
rs59714365	3:17376363	T	A	0.285886	5.91E-70	156.751	TBC1D5
rs117621227	3:3043335	C	T	0.277787	9.87E-68	125.291	CNTN4
rs62253676	3:24193134	G	A	0.293175	4.55E-67	159.574	THRB
rs67297609	4:30073704	C	T	0.255775	2.17E-72	99.7282	PCDH7
rs67297609	4:30073704	C	T	0.255775	2.17E-72	99.7282	PCDH7
rs147151420	4:20617835	A	T	0.267963	8.48E-67	102.173	SLIT2
rs190229071	5:119240393	A	G	0.278955	3.28E-72	150.374	DMXL1
rs3822376	5:143310270	C	A	0.247503	4.94E-70	79.7186	NR3C1
rs200823856	5:5436345	A	AAAT	0.245507	8.49E-70	76.3792	ICE1
rs180750731	5:65273540	T	A	0.321272	1.14E-69	289.559	ADAMTS6
rs73787785	5:112161631	A	T	0.263585	6.83E-69	101.885	EPB41L4A
rs2973532	5:73923334	C	T	0.308717	1.65E-68	221.414	ARHGEF28
rs142217030	5:9122670	A	G	0.280277	4.49E-68	132.5	SEMA5A
rs325203	5:98860114	G	A	0.315162	5.80E-68	242.301	CHD1
rs370458743	5:90709056	A	ACT	0.293265	7.72E-67	158.395	ADGRV1
rs41271876	6:56627094	T	A	0.252616	1.11E-70	89.1369	DST
rs80348614	6:134115381	G	A	0.282524	3.65E-69	143.47	SGK1

doi:10.6342/NTU202401923

rs577814148	6:123548636	G	A	0.29795	1.88E-67	175.958	TRDN*
rs3734657	6:90516489	A	G	0.245904	2.26E-67	71.1279	MAP3K7
rs142425724	6:121080640	A	G	0.310223	6.14E-67	213.169	TBC1D32
rs569905009	7:7452235	T	G	0.280335	9.69E-71	146.204	COL28A1
rs115049994	7:57450241	A	G	0.289926	2.54E-68	158.263	ZNF716
rs9655774	7:102111780	G	A	0.278765	5.82E-67	123.868	CUX1
rs117600369	8:41661986	T	C	0.26003	1.89E-67	91.1158	ANK1
rs1042701	8:11564536	A	G	0.290317	3.80E-67	152.332	BLK
rs34908836	9:449874	C	G	0.332506	3.50E-70	360.957	DOCK8
rs7863859	9:92188437	G	C	0.297056	9.62E-70	189.423	IARS1
rs74849667	9:20354757	T	C	0.295884	4.91E-68	173.673	MLLT3
rs45519332	10:29491007	A	G	0.255841	4.45E-70	92.5299	SVIL
rs74861203	10:5763180	G	A	0.272329	6.93E-69	118.749	TASOR2
rs74368421	10:20217497	A	G	0.276207	8.03E-69	126.813	PLXDC2
rs77055528	11:62529063	C	T	0.302429	2.57E-70	213.003	AHNAK
rs76879660	12:26065230	A	G	0.249938	1.79E-70	84.4239	RASSF8
rs191400739	12:100622217	G	A	0.255914	1.61E-69	90.9464	GAS2L3
rs192329797	12:113273201	G	C	0.256128	4.52E-69	89.9321	TPCN1
rs3730070	12:48775065	C	G	0.258399	1.96E-68	91.5952	ADCY6
rs3729832	14:23414928	A	T	0.298627	1.38E-71	209.091	MYH7*
rs150410807	14:31157204	C	T	0.288367	1.89E-71	173.16	HECTD1
rs728286	14:90032980	G	C	0.268963	7.09E-71	119.999	TDP1
rs7156821	14:29638980	T	C	0.302775	1.69E-70	215.855	PRKD1
rs139814895	15:89292706	G	A	0.332536	2.75E-68	332.618	FANCI
rs73362147	15:25211955	A	C	0.247625	3.31E-67	72.8879	UBE3A
rs369918248	16:1212372	T	G	0.321907	2.24E-69	289.253	CACNA1H
rs78549091	16:78115232	C	G	0.255178	4.26E-68	85.6047	WWOX
rs1058474	16:1998795	T	C	0.243866	4.35E-68	70.2604	ZNF598

doi:10.6342/NTU202401923

rs2230097	17:42201704	G	A	0.335975	1.92E-68	355.639	STAT5B
rs77805790	17:7356601	T	C	0.267794	1.19E-67	104.998	TMEM95
rs3764494	18:53386145	A	G	0.271774	4.88E-71	126.877	DCC
rs3764494	18:53386145	A	G	0.271774	4.88E-71	126.877	DCC
rs149645969	18:10697698	A	C	0.30181	4.73E-69	200.479	PIEZO2
rs9952711	18:34806376	C	T	0.269865	7.36E-69	113.62	DTNA
rs8091515	18:65859847	T	C	0.240612	6.81E-68	65.9753	CDH7
rs112337232	19:41090177	G	C	0.270272	4.18E-71	123.816	CYP2A13
rs527374014	19:13372038	T	C	0.266251	1.60E-69	109.127	CACNA1A
rs6103631	20:44115519	G	C	0.325135	1.76E-69	307.53	JPH2

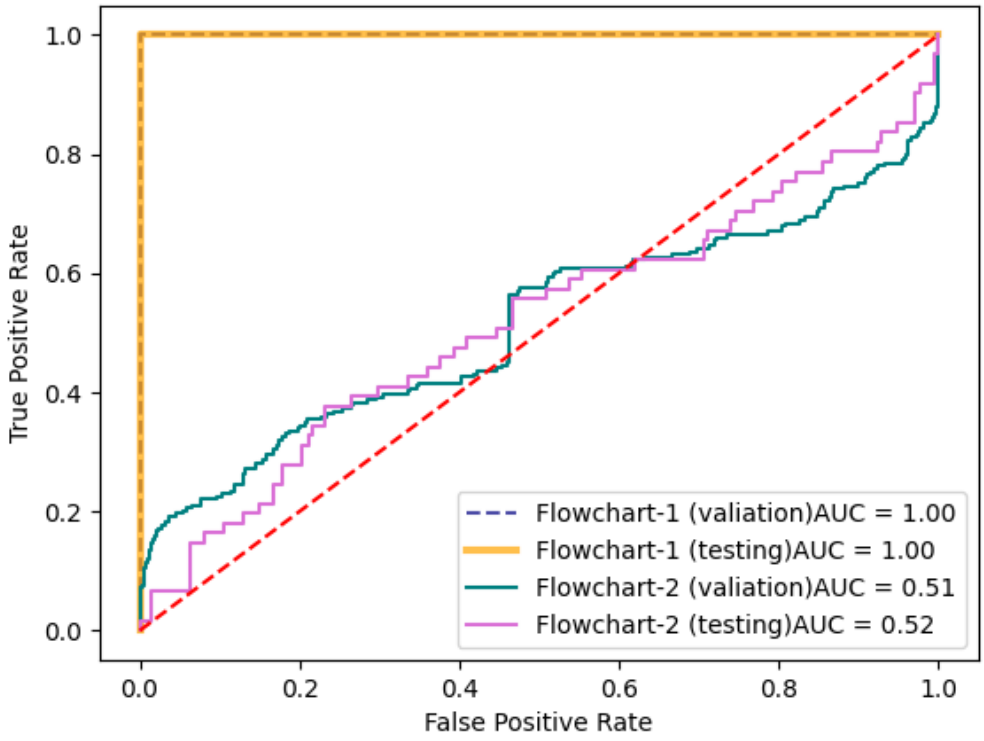


Figure 14 Receiver operator characteristic of the polyenic risk scores result by PRSice-2 [27]

doi:10.6342/NTU202401923

REFERENCE

[1] Kaviarasan, V., Mohammed, V. & Veerabathiran, R. Genetic predisposition study of heart failure and its association with cardiomyopathy. Egypt Heart J 74, 5 (2022). https://doi.org/10.1186/s43044-022-00240-6.

[2] Michelle M. Kittleson, Mathew S. Maurer, Amrut V. Ambardekar, Renee P. Bullock-Palmer, Patricia P. Chang, Howard J. Eisen, Ajith P. Nair, Jose Nativi-Nicolau, Frederick L. Ruberg and on behalf of the American Heart Association Heart Failure and Transplantation Committee of the Council on Clinical Cardiology Originally published 1 June 2020 https://doi.org/10.1161/CIR.0000000000000792 Circulation. 2020;142:e7–e22

[3] Sepehrvand N, Youngson E, Fine N, Venner CP, Paterson I, Bakal J, Westerhout C, Mcalister FA, Kaul P, Ezekowitz JA. The Incidence and Prevalence of Cardiac Amyloidosis in a Large Community-Based Cohort in Alberta, Canada. J Card Fail. 2022 Feb;28(2):237-246. doi: 10.1016/j.cardfail.2021.08.016. Epub 2021 Sep 9. PMID: 34509599.

[4] Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, Derks EM. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. Int J Methods Psychiatr Res. 2018 Jun;27(2):e1608. doi: 10.1002/mpr.1608. Epub 2018 Feb 27. PMID: 29484742; PMCID: PMC6001694.

[5] 肥厚型心肌病變致病原因重大發現－法布瑞氏症，December 05, 2008, https://www1.vghtpe.gov.tw/msg/%E6%B3%95%E5%B8%83%E7%91%9E%E6%B0%8F%E7%97%87971205.pdf

[6] Cleveland Clinic medical professional, Left Ventricular Non-Compaction (LVNC),

doi:10.6342/NTU202401923

June 03, 2022, https://my.clevelandclinic.org/health/diseases/23248-left-ventricular-non-compaction-lvnc

[7] Bustamante JG, Zaidi SRH. Amyloidosis. [Updated 2023 Feb 11]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK470285/

[8] Penn Medicine, Philadelphia, PA, Penn Medicine, Philadelphia, PA, August 05, 2022, https://www.pennmedicine.org/for-patients-and-visitors/patient-information/conditions-treated-a-to-z/cardiac-amyloidosis

[9] Gao, X.R., Chiariglione, M., Qin, K. et al. Explainable machine learning aggregates polygenic risk scores and electronic health records for Alzheimer's disease prediction. Sci Rep 13, 450 (2023). https://doi.org/10.1038/s41598-023-27551-1

[10] Wei, CY., Yang, JH., Yeh, EC. et al. Genetic profiles of 103,106 individuals in the Taiwan Biobank provide insights into the health and history of Han Chinese. npj Genom. Med. 6, 10 (2021). https://doi.org/10.1038/s41525-021-00178-9

[11] Article Source: A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies, Howie BN, Donnelly P, Marchini J (2009) A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLOS Genetics 5(6): e1000529. https://doi.org/10.1371/journal.pgen.1000529

[12] 臺灣人體生物資料庫 TWBv1.0 基因型差補 (genotype imputation) 流程 biobank.org.tw/file_download/實驗資訊/臺灣人體生物資料庫 TWBv1.0 基因型

doi:10.6342/NTU202401923

差補(genotype%20imputation)流程.pdf

[13] König E, Rainer J, Hernandes VV, Paglia G, Del Greco M F, Bottigliengo D, Yin X, Chan LS, Teumer A, Pramstaller PP, Locke AE, Fuchsberger C. Whole Exome Sequencing Enhanced Imputation Identifies 85 Metabolite Associations in the Alpine CHRIS Cohort. Metabolites. 2022 Jun 29;12(7):604. doi: 10.3390/metabo12070604. PMID: 35888728; PMCID: PMC9320943.

[14] Kunkle, B.W., Grenier-Boley, B., Sims, R. et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. Nat Genet 51, 414–430 (2019). https://doi.org/10.1038/s41588-019-0358-2

[15] Ran S, He X, Jiang ZX, Liu Y, Zhang YX, Zhang L, Gu GS, Pei Y, Liu BL, Tian Q, Zhang YH, Wang JY, Deng HW. Whole-exome sequencing and genome-wide association studies identify novel sarcopenia risk genes in Han Chinese. Mol Genet Genomic Med. 2020 Aug;8(8):e1267. doi: 10.1002/mgg3.1267. Epub 2020 Jun 1. PMID: 32478482; PMCID: PMC7434604.

[16] GWAS Tutorial for Beginners designed for the course Fundamental Exercise II provided by The Laboratory of Complex Trait Genomics at the University of Tokyo, https://cloufield.github.io/GWASTutorial/

[17] Delaneau, O., Marchini, J. & The 1000 Genomes Project Consortium. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. Nat Commun 5, 3934 (2014). https://doi.org/10.1038/ncomms4934

[18] Article Source: A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies, Howie BN, Donnelly P, Marchini J (2009) A Flexible and Accurate Genotype Imputation Method for the

doi:10.6342/NTU202401923

Next Generation of Genome-Wide Association Studies. PLOS Genetics 5(6): e1000529. https://doi.org/10.1371/journal.pgen.1000529

[19] Uffelmann, E., Huang, Q.Q., Munung, N.S. et al. Genome-wide association studies. Nat Rev Methods Primers 1, 59 (2021). https://doi.org/10.1038/s43586-021-00056-9

[20] 人类基因组的 Phasing 原理是什么？ - 黄树嘉的文章 - 知乎 https://zhuanlan.zhihu.com/p/36289359

[21] 【文献阅读笔记】(2)：使用 IMPUTES2 和 minimac 软件完成群体特异性的基因型填充 (Imputation) http://t.csdn.cn/DZOko

[22] Scitable by nature EDUCATION - Hardy-Weinberg equilibrium https://www.nature.com/scitable/definition/hardy-weinberg-equilibrium-122/

[23] PLINK2.0 https://www.cog-genomics.org/plink/2.0/

[24] Shing Wan Choi, Paul F O'Reilly, PRSice-2: Polygenic Risk Score software for biobank-scale data, GigaScience, Volume 8, Issue 7, July 2019, giz082, https://doi.org/10.1093/gigascience/giz082

[25] Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet. 2014 Jul 3;95(1):5-23. doi: 10.1016/j.ajhg.2014.06.009. PMID: 24995866; PMCID: PMC4085641.

[26] Ma C, Blackwell T, Boehnke M, Scott LJ; GoT2D investigators. Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. Genet Epidemiol. 2013 Sep;37(6):539-50. doi: 10.1002/gepi.21742. Epub 2013 Jun 20. PMID: 23788246; PMCID: PMC4049324.

[27] Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale

doi:10.6342/NTU202401923

data. Gigascience. 2019 Jul 1;8(7):giz082. doi: 10.1093/gigascience/giz082. PMID: 31307061; PMCID: PMC6629542.

[28] Tadros, R., Francis, C., Xu, X. et al. Shared genetic pathways contribute to risk of hypertrophic and dilated cardiomyopathies with opposite directions of effect. Nat Genet 53, 128–134 (2021). https://doi.org/10.1038/s41588-020-00762-2

[29] Harper, A.R., Goel, A., Grace, C. et al. Common genetic variants and modifiable risk factors underpin hypertrophic cardiomyopathy susceptibility and expressivity. Nat Genet 53, 135–142 (2021). https://doi.org/10.1038/s41588-020-00764-0

[30] Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA; NHLBI GO Exome Sequencing Project—ESP Lung Project Team; Christiani DC, Wurfel MM, Lin X. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet. 2012 Aug 10;91(2):224-37. doi: 10.1016/j.ajhg.2012.06.007. Epub 2012 Aug 2. PMID: 22863193; PMCID: PMC3415556.

doi:10.6342/NTU202401923