Phenome-Wide and Genome-Wide Analyses in CLSA Biobank and Interactive Web-Based Platform for the Results Sharing

Mehrdad Kazemi
Department of Human Genetics, Faculty of Medicine and Health Sciences,
McGill University, Montreal, Quebec, Canada

AUGUST 2024

A thesis submitted to the McGill University in partial fulfillment of the requirements of the degree of Master of Science.

To my lovely family and mentors.

Abstract

The advent of genome-wide association studies (GWASs) has significantly advanced our understanding of complex traits—traits influenced by multiple genes and environmental factors—by analyzing the frequency of genetic variants across the genome in thousands of individuals with different health statuses. To date, GWASs have identified over 434,000 significant genetic associations for more than 3,300 complex traits such as obesity, autoimmune diseases, and cancer, influencing a broad range of fields from understanding disease mechanisms to guiding drug development and risk assessment.

Initially, GWASs focused on individual diseases or traits, but it soon became evident that many genetic variants and genes are associated with multiple traits, a phenomenon known as pleiotropy. The emergence of global population-based biobanks like UK Biobank and FinnGen, which compile extensive genetic and phenotypic data on hundreds of thousands of individuals and across hundreds of phenotypic measures, has enabled researchers to examine the impact of a single genetic variant across a plethora of diseases and traits through phenome-wide association studies (PheWAS). This approach not only broadens our understanding of disease mechanisms by revealing shared genetic influences but also aids in identifying targets for new treatments and drug repurposing. The advantages of GWAS and PheWAS highlight the importance of widely sharing summary statistics of genetic variant-trait associations for downstream analyses and applications. This has led to the development of tools like PheWeb, which is an open-source web-based tool used for sharing GWAS and PheWAS summary statistics, enabling researchers worldwide to easily access, visualize, explore, and build upon this wealth of information.

This project aims to perform systematic GWAS and PheWAS scans using all genetic data and binary phenotypes available in the Canadian Longitudinal Study on Aging (CLSA), one of Canada's largest genetic studies, and share results with the broad scientific community using the PheWeb platform. The CLSA PheWeb is built upon 350 binary phenotypes, with over 308 million genetic variants present in approximately 25,000 individuals of European genetic ancestry, providing an invaluable resource for researchers. CLSA PheWeb may foster interdisciplinary research and global

collaboration, including facilitating meta-analysis studies, by providing an accessible platform for the scientific community to explore, validate and compare genetic associations across various datasets. This collaborative potential may accelerate the discovery of novel genetic insights and strengthen the validation of existing associations, enriching our understanding of human genetics.

Résumé

L'avènement des études d'association pangénomique (GWAS) a considérablement avancé notre compréhension des traits complexes—traits influencés par plusieurs gènes et facteurs environnementaux—en analysant la fréquence des variants à travers le génome chez des milliers d'individus avec différents statuts de santé. À ce jour, les GWAS ont identifié plus de 434,000 associations génétiques significatives pour plus de 3,300 traits complexes. Parmi ces traits, on compte l'obésité, les maladies auto-immunes et le cancer. De plus, ces observations ont influencé un large éventail de domaines tels que le mécanisme des maladies, le développement de médicaments et de la médecine personnalisée.

Initialement, les GWAS se concentraient sur des maladies ou des traits individuels, mais il est rapidement devenu évident que de nombreux variants et gènes sont associés à plusieurs traits. Ce phénomène est connu sous le nom de pléiotropie. Des biobanques, comme UK Biobank et FinnGen, compilent les données génétiques et les mesures phénotypiques provenant de centaines de milliers d'individus à l'échelle mondiale. L'émergence de ces biobanques a permis aux chercheurs d'examiner l'impact de chacun des variants à travers une multitude de maladies et de traits grâce aux études d'association phénomiques (PheWAS). Non seulement cette approche élargit notre compréhension des mécanismes de diverses maladies en révélant des influences génétiques partagées, elle aide également à identifier des cibles pour de nouveaux traitements et la reposition de médicaments. Les avantages des GWAS et des PheWAS soulignent l'importance de partager mondialement les statistiques sommaires des associations trait-variant à des fins d'analyses et d'applications dérivées C'est dans cette optique que l'outil de visualisation de données PheWeb a été développé. PheWeb est un logiciel open-source, en ligne, utilisé pour partager des résultats selon les approches GWAS et PheWAS. Il permet ainsi aux chercheurs du monde entier de facilement accéder, visualiser et explorer les données de diverses sources d'informations reconnues.

Ce projet vise à effectuer des analyses systématiques de GWAS et PheWAS en utilisant toutes les données génétiques et les phénotypes binaires disponibles dans l'étude longitudinale canadienne sur le vieillissement (CLSA ÉLCV), l'une des plus grandes études génétiques du Canada, et à partager les résultats avec la communauté scientifique en utilisant la plateforme PheWeb. Le CLSA PheWeb est construit à partir de 348 phénotypes binaires, avec plus de 308 millions de variants présents chez environ 25,000 individus d'ascendance génétique européenne, fournissant une ressource inestimable pour les chercheurs. Le CLSA PheWeb peut favoriser la recherche interdisciplinaire et la collaboration mondiale, incluant les études de méta-analyse, en fournissant une plateforme accessible pour la communauté scientifique afin d'explorer, valider et comparer les associations génétiques à travers divers ensembles de données. Ce potentiel collaboratif peut accélérer la découverte de nouvelles perspectives génétiques et renforcer la validation des associations existantes, enrichissant notre compréhension de la génétique humaine.

List of Figures

List of Tables

Acknowledgements

First and foremost, I extend my deepest gratitude to my supervisor, Dr. Daniel Taliun, for his invaluable guidance, patience, and expert advice throughout this research journey. His mentorship was pivotal in shaping both the direction and success of my work.

I am grateful for the collaboration and support from our esteemed colleagues at the Université de Montréal, Dr. Sarah Gagliano Taliun and Justin Bellavance, whose insights and expertise significantly enhanced our research.

I am also thankful to my MSc Advisory Committee members, Dr. Claude Bhérer and Dr. Sirui Zhou, for their insightful feedback and continuous encouragement throughout my research.

Special thanks to Vincent Chapdelaine, the author of the Regenie_nextflow pipeline, which was instrumental in running our GWAS. His readiness to assist with technical questions profoundly impacted my work. I wish to thank Peyton McClelland, the author of the HGDP_1KG_ancestry_inference pipeline, which was crucial for conducting our ancestry inference. I am also thankful to Justin Bellavance for his assistance in setting up the PheWeb.

This research could not have been conducted without the computational resources provided by the Digital Research Alliance of Canada and all participants of the CLSA who volunteered their time and information for research.

A heartfelt thank you to all the McGill Canada Excellence Research Chair (CERC) in Genomic Medicine team members, whose collaborative environment and shared knowledge greatly enriched my experience.

Lastly, I acknowledge the financial support from the CIHR CGS M program, which funded this research and facilitated my academic and professional growth.

Contribution of Authors

This thesis was written entirely by myself, Mehrdad Kazemi. All research, data analysis, interpretation of results, and manuscript preparation were conducted independently, under the supervision of Dr. Daniel Taliun, who provided guidance throughout the entire project.

Task (section)	**Contributor**
Phenotype and genotype data collection (2.1.1 and 2.1.2)	CLSA team
Ancestry Inference (2.2.1)	Mike Kazemi
Phenotype data analysis (2.3)	Mike Kazemi
LiftOver (2.4)	Mike Kazemi
Imputation for Chr 1-22 (2.1.2)	CLSA team
Imputation for ChrX (2.5.1.3.1)	Mike Kazemi
GWAS (2.5 and 3.1)	Mike Kazemi
Setting up PheWeb (3.2)	Mike Kazemi

Chapter 1: Introduction

The advent of genome-wide association studies (GWASs) marked a paradigm shift in our approach to understanding the genetic underpinnings of complex traits, which are traits influenced by multiple genes and their interactions with the environment. GWAS operates on a hypothesis-free basis and systematically examines the frequency of genetic variants across the entire genome in thousands of individuals to identify those that are associated with a trait. This process results in the identification of statistically significant associations when a genetic variant occurs more frequently in individuals with a particular trait than in those without, indicating a potential genetic influence on the trait. More than 6,000 GWASs have now been conducted for more than 3,300 complex traits resulting in more than 434,000 statistically significant associations across a variety of diseases and disease-related[1] traits such as obesity, autoimmune diseases, Type 2 Diabetes, Schizophrenia, Osteoporosis, Breast Cancer, Asthma and Alzheimer's Disease[1–3]. GWAS findings are instrumental across various domains, including unravelling the biological basis of a trait, advancing clinical risk assessments through polygenic risk scores (PRS), guiding drug development efforts, and exploring potential causal links between risk factors and health outcomes through Mendelian randomization (MR)[1,2].

Typically, early GWASs investigated only one or very few diseases or traits at a time, addressing a specific research question[4]. By 2011, it became clear that some of the genetic variants and genes reported by hundreds of these independent disease or trait-specific GWASs overlap, i.e. are associated with multiple traits, also referred to as pleiotropy[5]. However, the at-scale identification of all pleiotropic effects requires access to thousands of traits, also known as phenotypes, measured in thousands of individuals with genetic data available, which became possible with the development of population-based biobanks worldwide. The UK Biobank[6] (500,000 participants, ~800,000 variants, >4,200 phenotypes), All of Us Research Program[7] (USA) (>1 million participants), China Kadoorie Biobank[8,9] (512,000 participants, 700,701 variants, hundreds of phenotypes), KoGES[10] (72,298 participants, 8,056,211 variants, 136 phenotypes), FinnGen[11] (500,000 Finnish participants, 16,962,023 SNPs and INDELS,

1,932 phenotypes), and BioBank Japan[12] (200,000 participants, up to 964,193 variants[13], 280 phenotypes) illustrate the global scale and diversity of biobank efforts to compile extensive genetic and phenotypic data for research.

The wealth of data provided by these biobanks has enabled researchers to analyze the impact of a genetic variant across a wide array of diseases and traits, using Phenome-Wide Association Studies (PheWAS)[14–16]. PheWAS examines the impact of a known genetic variant across numerous traits, representing a complementary approach to GWAS. This methodology allows for the identification of the pleiotropic effects of a variant, offering insights into shared pathophysiological pathways. This broad approach not only enhances our understanding of disease mechanisms but also identifies potential targets for therapeutic intervention and drug repurposing by shedding light on possible adverse effects[14,15,17].

The numerous benefits of GWAS and PheWAS mentioned earlier, underscore the necessity of sharing summary statistics of genetic variant-trait associations widely within the scientific community. This need is met by PheWeb[18], an open-source web-based tool used for sharing GWAS and PheWAS summary statistics, enabling researchers worldwide to easily access, visualize, explore, and build upon this wealth of genetic information. By providing a user-friendly web-based platform for these findings, PheWeb ensures that the groundbreaking insights from GWAS and PheWAS are fully leveraged, fostering further discoveries and innovations in genomic research. Examples of PheWeb platforms tailored for biobanks around the world include UK Biobank[6] PheWeb, TCGA[19] PheWeb, FinnGen[11] PheWeb, BioBank Japan[12] PheWeb, CARTaGENE[20] PheWeb, COLCORONA[21] PheWeb, COLCOT[22] PheWeb, CHARM[23] PheWeb, SardiNIA[24] PheWeb, KoGES[10] PheWeb and The Qatar Genome Program (QGP)[25] PheWeb. While PheWeb is renowned for its comprehensive approach to displaying PheWAS and GWAS results, other tools such as Genebass[26], GWAS Catalog[3], PhenoScanner[27] and AstraZeneca PheWAS Portal[28], offer similar features that support genetic research.

This project aims to perform systematic GWAS and PheWAS scans using all genotype data and binary phenotypes available in the Canadian Longitudinal Study on Aging (CLSA)[29], one of Canada's largest genetic studies, and share results with the broad scientific community using the PheWeb platform. CLSA represents a significant longitudinal effort, gathering data from over 50,000 individuals aged between 45 and 85 at recruitment[29]. The CLSA's extensive collection of health-related measurements offers a unique lens through which the interplay of genetic and environmental factors on human health can be studied.

The CLSA PheWeb is built upon 350 binary phenotypes, with over 308 million variants and approximately 25,000 individuals of European like genetic ancestry providing an invaluable resource for researchers. CLSA PheWeb fosters interdisciplinary research and global collaboration, including facilitating meta-analysis studies, by providing an accessible platform for the scientific community to explore, validate and compare genetic associations across various datasets. This collaborative potential accelerates the discovery of novel genetic insights and strengthens the validation of existing associations, enriching our understanding of human genetics.

1.1 Genome-wide association study (GWAS)

Genome-wide association study (GWAS) is a hypothesis-free approach for identifying associations between genetic variants and both diseases and traits related to health. For both binary and continuous traits, this approach involves collecting data from thousands of individuals, categorized into cases and controls for binary traits (those affected by the disease versus those unaffected), or measured across a spectrum for continuous traits. Each genetic variant is systematically evaluated for associations, generating millions of association statistics across different phenotypic expressions. Since the first successful GWAS in 2002[30], the number of studies employing this method has grown rapidly. At the time of writing, the GWAS Catalog contains more than 6,000 publications and 415,784 statistically significant gene-disease associations[31]. The

success of GWAS is attributable to several key developments in the field of genomics: the compilation of comprehensive catalogs of human genetic variation, the advent of cost-effective genotyping methods, such as SNP arrays, the availability of large sample sizes that enhance the power of these studies, and the implementation of sophisticated statistical methodologies for data analysis[1,32]. A conventional GWAS workflow includes several critical steps[1]:

- Data Collection: Gathering phenotypic and genetic data from a broad cohort of individuals.
- Genotyping: Determining the genetic variants present in the collected samples using high-throughput sequencing or array-based technologies.
- Quality Control: Filtering the data to remove poor quality or unreliable genetic markers and samples.
- Imputation: Estimating unobserved genotypes to enhance the density of genetic data and facilitate comparisons across studies.
- Statistical Tests: Employing statistical models to identify associations between genetic variants and the traits or diseases of interest.
- Meta-analysis: Combining data from multiple GWAS to increase statistical power and validate findings.
- Replication: Validating significant associations in independent cohorts to confirm findings.
- Post-GWAS Analyses: Conducting further analyses to explore the biological implications of identified associations, including functional studies, gene-environment interactions, and pathway analyses.

The subsequent sections will delve into each of these steps, providing a comprehensive overview of the GWAS process and its implications for understanding the genetic basis of complex diseases and traits.

1.1.1 Data collection

To find replicable genome-wide significant associations, GWAS frequently need very high sample sizes. The needed sample size can be calculated using power estimates in

software tools like CaTS[33] or GPC[34]. When the characteristic of interest is dichotomous, research designs can include cases and controls; otherwise, when the trait is quantitative, the entire study sample can be subjected to quantitative measures. Additionally, there are population-based and family-based design options available. The desired sample size, the experimental question, the availability of pre-existing data, or the ease with which new data may be acquired all influence the choice of data resource and study design for a GWAS[1]. Since Individual cohorts with detailed clinical measures may not be able to meet the necessary sample size, in some cases, "proxy" phenotypes that are simpler to measure and for which there is more data can be used. For instance, educational attainment can be used as a proxy for intelligence or depressive symptoms as a proxy for a clinical diagnosis of depression[1].

Direct-to-consumer studies or data from resources like biobanks or cohorts with disease- or population-based enrollment can be used to conduct GWAS. For a complex trait, a well-powered GWAS needs significant time and financial commitments that are beyond the capabilities of the majority of individual laboratories. As a result, the majority of GWAS are carried out using a number of great public resources that already exist and offer access to large cohorts with both genotypic and phenotypic data. Even when new data are gathered internally, they are frequently co-analyzed with data from pre-existing sources; additional data collection is typically necessary when more accurate phenotyping is wanted[1].

Recruitment tactics must be carefully evaluated for all study designs since they can introduce bias in the collected data. GWAS commonly use genetic and phenotypic data from cohorts based on population surveys, where participants are believed to be randomly selected from the population. As long as the population substructure is taken into account to prevent producing false positive results, several ethnic groups might be included in the same study. Associations between genetic variants, whether genotyped or imputed, can be analyzed for traits that are either continuous or binary. In a typical case-control GWAS design, participants are classified as cases or controls based on whether they exhibit a specific trait. Active recruitment of cases and controls is typically favoured when there are limited financial resources and a need to increase statistical power[1]. A greater effort must be made during quality control and subsequent analysis to

eliminate artifacts if cases and controls are not genotyped together on the same chip. This could mean including the genotyping batches as a covariate in the analyses. It should be highlighted that although samples are thought to be drawn at random from the population, participation bias and asymmetrical socio-demographic features make this assumption untenable[1].

Conducting GWAS in communities that have limited gene exchange with neighbouring populations because of a founder event, such as geographical or cultural barriers, has considerable benefits. One key benefit is that isolated populations may have functional variants that are normally rare in other populations. Therefore, studying such isolated populations can boost the power of association studies for those variants. If even a relatively small number of individuals from the isolated population are included in the reference panel, the long-range linkage disequilibrium expected for isolated populations enhances imputation accuracy and power over similarly sized non-isolated cohorts. Because isolated populations have high levels of relatedness, GWAS frequently adopt a linear mixed model-based approach. Due to the extinction of alleles caused by genetic bottlenecks, isolated populations have a tendency to have high genetic homogeneity. This can boost the power of burden tests by lowering the number of neutral variants. If a variant is too rare, it may be challenging to replicate the discovery in other populations, but other variants implicating the same gene may provide additional evidence.

As GWAS calls for large-scale genotypic and phenotypic data, many national population-based biobanks have been developed worldwide. Researchers have access to various sizable, publicly accessible population biobanks. Data from thousands of genotyped individuals who have undergone extensive phenotyping—either by questionnaires, laboratory tests or linking to electronic health records—and who were not chosen for specific disease traits—can be found in biobanks. A prominent example is the UK Biobank[6], which contains information on roughly 500,000 people and has increased sample sizes for GWAS of common diseases while also enabling well-powered GWAS of hundreds of quantitative traits, including anthropometric traits, blood cell traits, metabolites, cognitive traits, brain imaging traits, and depressive symptoms. Large biobanks of data from people with non-European ancestries are being developed, and many new studies are based on ethnically diverse communities,

motivated by the fact that biobanks have been typically focused on populations with European genetic ancestry. The majority of biobanks have employed imputed genotype data for common variants. However, as the cost of whole-exome sequencing (WES) and whole-genome sequencing (WGS) continues to drop, the field is quickly moving towards WES- and WGS-based GWAS[1].

The table below outlines key details of major biobanks utilized for GWAS, including ancestries, sample sizes, and URLs.

Data set	Ancestry	Sample size	URL
UK Biobank[6]	Predominantly white British	~500,000	https://www.ukbiobank.ac.uk/
BioBank Japan[41]	Japanese	~200,000	https://biobankjp.org/en/
China Kadoorie Biobank[8]	Chinese	~512,000	https://www.ckbiobank.org/
H3Africa[42]	Various African ancestries	~118,000	https://h3africa.org/
TOPMed[43]	Multiple ancestries (USA)	~180,000	https://topmed.nhlbi.nih.gov/
FinnGen[11]	Finnish	260,405	https://www.finngen.fi/en
CARTaGENE[20]	European ancestry and French Canadian heritage	~43,000	https://cartagene.qc.ca/en/index.html

Table 1.
Major Biobanks around the world

1.1.2 Genotyping

Genotyping, the process of determining the genetic variants an individual carries, is a foundational step in GWAS. Genetic variants, also known as alleles, are differences in the DNA sequence among individuals. A "common variant" is identified as a genetic

variation that occurs frequently within a population, with a minor allele frequency often exceeding 5%. This frequency threshold may be adjusted to as low as 1% in studies with larger populations, with the standard being the presence of the minor allele in at least 100 carriers within the study group. These common variants are often implicated in various traits and conditions, making their identification crucial for GWAS[32].

For genotyping in GWAS, two main techniques are predominantly used: microarray-based genotyping for common variants and next-generation sequencing (NGS) methods, such as WES or WGS, for a comprehensive assessment that includes both common and rare variants[1]. Microarray-based genotyping is a cost-effective approach that focuses on known variants across the genome, making it the preferred method for many studies due to its efficiency and lower cost compared to NGS. This method involves the use of chips pre-designed to detect specific variants and is particularly useful for examining genetic variations associated with common diseases or traits[1].

Next-generation sequencing, encompassing WES and WGS, offers a more detailed view by sequencing either the coding regions of the genome (WES) or the entire genome (WGS). While WES focuses on the approximately 1% of the genome that codes for proteins, WGS provides a comprehensive overview, capturing nearly every genomic variation, including rare variants which might have significant effects on phenotype but are missed by microarray-based methods. Although NGS offers deeper insights into the genetic blueprint, its higher costs have traditionally limited its use primarily to studies where the detailed genetic information it provides can justify the expense[1].

The choice of genotyping platform in GWAS often hinges on the study's specific goals, the types of variants of interest, and budgetary considerations. In consortium-led GWAS, for instance, uniformity in genotyping platforms across individual cohorts is usually advised to ensure data comparability and integrity. With the anticipated decrease in the cost of WGS technologies and the increasing recognition of the value of rare variant information, it is expected that WGS will become the predominant choice for genotyping in the coming years, offering unprecedented detail and accuracy in genetic studies[1,44,45].

1.1.3 Quality Control

Anonymized individual ID numbers, coded familial relationships between individuals, sex, phenotype information, covariates, genotype calls for all called variants, and information on the genotyping batch are all included in the input files for a GWAS. Following the data entry, extensive quality control is necessary to produce reliable results from GWAS. The elimination of variants that are not in Hardy-Weinberg equilibrium, the filtering of SNPs that are missing from a portion of the cohort, the identification and removal of genotyping errors, and the assurance that phenotypes are properly matched with genetic data—often by comparing self-reported sex to sex based on the X and Y chromosomes—are some examples of steps[1].

Many of these quality control procedures can be carried out using software tools like PLINK[38], which was created particularly to analyze genetic data. Once GWAS array data has undergone sample and variant quality control, variants are typically phased and imputed. GWAS consortia regularly adhere to pipelines for carrying out quality control stages and imputation, utilizing software like RICOPILI[46] or a similar programme, or they submit their data to imputation servers, where these standardized procedures have been set in place. The utilization of computer clusters or cloud settings that can spread tasks to numerous machines is commonplace since analysis processes can be executed in parallel for genetic data sets which are frequently large[1]. The aforementioned stages are typically carried out independently for numerous different cohorts with varied sample sizes in order to obtain the huge sample sizes characteristic in genetic studies in a logistically practical manner while adhering to data privacy laws[1].

1.1.4 Genotype Imputation

The method of predicting or imputing genotypes that are not directly assayed is known as genotype imputation. This is done by leveraging information from reference panels of individuals with densely genotyped data. Imputation can increase the power and resolution of GWAS by allowing for the testing of a larger set of genetic variants. It also enables meta-analysis of GWAS results from different studies that use different genotyping platforms. Imputed genotypes are typically used to conduct association tests with phenotypes of interest. The accuracy of imputation depends on several factors,

including the density and quality of the genotyping data, the size and representativeness of the reference panel, and the degree of linkage disequilibrium (LD) across the genomic regions of interest. High-quality genotyping results, a large reference panel with diverse ancestries that captures a broad spectrum of genetic variation, and strong LD between known and unknown variants can significantly enhance the precision of imputed genotypes[50,51].

The enhancement of genetic imputation methods is a pivotal aspect of contemporary GWAS, as they facilitate the prediction of ungenotyped variants in a study sample based on a set of reference haplotypes. The most prominent reference panels in the field are the 1000 Genomes Project[52], the Haplotype Reference Consortium (HRC)[53], and the Trans-Omics for Precision Medicine (TOPMed) program[43]. The 1000 Genomes Project provides a comprehensive resource on human genetic variation, which includes data from diverse populations, aiding in the broad representation of global genetic diversity. The HRC compiles high-quality, whole-genome sequencing data, which enhances the imputation of European ancestries. TOPMed, on the other hand, contributes extensive whole-genome sequencing data that supports the precise imputation of rare variants across diverse populations.

For the practical application of imputation, various software tools are utilized, with IMPUTE v2[54], Minimac4[55], and Beagle 5.4[56] being among the most utilized due to their efficiency and accuracy in handling large-scale genomic data. IMPUTE v2 is well-regarded for its robust performance with large reference panels, while Minimac4 offers speedy imputation with minimal computational resources. Beagle 5.4 is commended for its accuracy and speed, as well as its ability to handle both small and large datasets effectively[50,55,57].

Each software has distinct computational requisites and algorithms, which may affect the choice of tool based on the specific needs of the study. For instance, researchers may select a program that is best aligned with the structure of their data or the specific variants of interest. However, the end goal remains the same: to maximize the informativeness of the genetic data, thereby enabling more comprehensive association tests for complex traits. It's important to note that the choice of reference panel and imputation server may depend on the specific population under study, as the accuracy

of imputation can be influenced by the ancestry composition and sample size of the reference panel[50].

Genotype imputation can be done in a more focused region in the context of a fine-mapping study or across the entire genome as part of GWAS. The number of SNPs that can be tested for association can then be increased using these "in silico" genotypes. As a result, the study's power is increased, the causal variant can be settled or fine-mapped, and meta-analysis is made easier[50].

After the completion of genotype imputation, evaluating the quality of the imputed genotypes is crucial, particularly in scenarios where a true genotype dataset for comparison is unavailable. To address this, researchers have developed various post-imputation information measures to gauge the reliability of imputed SNPs, aiming to eliminate low-quality SNPs prior to association testing. These measures, designed to range between 0 and 1, help quantify the certainty of imputed genotypes: a measure of 1 indicates absolute confidence in the imputed genotypes, while a measure of 0 signifies complete uncertainty. The interpretation of these metrics suggests that an information measure value of α across a sample of N individuals approximates the value of having a perfectly observed genotype dataset in a sample size of αN[50].

Among the measures introduced, the MACH $\widehat{r^2}$ metric evaluates the imputed genotype quality by comparing the observed variance of allele dosage against the expected variance under the Hardy-Weinberg equilibrium. BEAGLE recommends utilizing the R^2 between the best-guess genotype and the allele dosage as a proxy for the R^2 between the best-guess genotype and the actual genotype. IMPUTE calculates a measure reflecting the relative statistical information about the SNP allele frequency derived from the imputed data[50].

Marchini and Howie (2010)[50] performed comparative analyses of these metrics, applied to a simulated dataset across a 7 Mb interval on chromosome 22, revealing a high correlation among the MACH, BEAGLE, and IMPUTE measures. However, it's noted that the MACH measure occasionally exceeds 1, and the BEAGLE measure is undefined for nearly 3% of SNPs. This underscores the necessity of choosing appropriate post-imputation quality metrics tailored to the specifics of each study, to

ensure that subsequent analyses are based on reliable and accurate imputed genotype data[50].

1.1.5 Statistical approaches for genetic association testing

The biometrical model, which quantifies the contributions of genetic and environmental factors to phenotypic traits, serves as the foundation for the theory of genetic association[1]. Depending on whether the phenotype is continuous (like height, blood pressure, or body mass index) or binary (like the presence or absence of disease), often in GWAS, linear or logistic regression models are utilized to investigate associations. To account for stratification and prevent confounding effects from demographic characteristics, covariates such as age, sex, and ancestry are added; however, this may reduce the statistical power for binary traits in ascertained samples. It is possible to strengthen control for stratification and enhance statistical power for genome discovery by using an additional random effect term, which is individual-specific in linear or logistic mixed models and accounts for genetic relatedness among individuals. The genotypes of genetic variants that are physically close to one another are not independent since they frequently exhibit linkage disequilibrium; this test dependency should also be taken into account when conducting a GWAS.

The typical linear regression models for GWAS can be written as follows:

$$Y \sim W\alpha + X_s\beta_s + g + e$$

$$g \sim N(0,\ \sigma_A^2\psi)$$

$$e \sim N(0,\ \sigma_e^2 I)$$

Where the phenotype vector is noted by Y for each individual, which is estimated using a set of covariates W with their respective effect sizes α, and the genotype values X_s at a specific SNP s with a fixed effect size β_s. Additionally, the model incorporates a random polygenic effect g, distributed normally with mean 0 and variance $\sigma_A^2\psi$, where the additive genetic variation of the phenotype is measured by σ_A^2 and ψ is the genetic relationship matrix. The model also includes an error term e, representing random

residual errors with a normal distribution with mean 0 and variance $\sigma_e^2 I$, where I is the identity matrix and σ_e^2 measures residual variance[1]. The model outlined above is unsuitable for case-control studies, where the outcome is categorical and not normally distributed. Predicted probabilities in such studies can erroneously fall outside the 0-1 range. Logistic regression rectifies this by employing a logit link function[58].

A strict multiple-testing threshold is needed to examine millions of associations between individual genetic variants and a phenotype without producing any false positives. According to research from the International HapMap Project[59], the average number of independent common genetic variants in the human genome is around 1 million, which results in a Bonferroni testing threshold of $P < 5 \times 10^{-8}$ (indicating a false discovery rate of $0.05/10^6$). In GWAS, the significance threshold to declare genetic associations may need adjustment according to population specifics. For example, populations with larger effective sizes, or studies incorporating rarer alleles due to increased sample sizes, might necessitate more stringent thresholds. This is because rarer variants often do not exhibit linkage disequilibrium with common variants, heightening the challenge of multiple testing. Complex traits such as height, schizophrenia, or blood pressure are typically polygenic, with numerous variants each contributing modestly. Here, the Winner's Curse can be common, leading to inflated effect size estimates for variants at the margin of discovery. This refers to the tendency for the effect sizes of genetic variants to appear larger than they truly are due to the statistical bias toward significant findings[60]. A robust approach to calibrate for false positives and the Winner's Curse involves comparing effect sizes between discovery and independent replication cohorts. While effect sizes cannot be anticipated prior to GWAS, planning for replication at the outset is crucial for sufficient power to address statistical distortions and multiple testing effects. It's essential to compare the effect statistics and their error metrics, like regression coefficients or odds ratios, across cohorts, particularly if different software was utilized for analysis. Moreover, replication cohorts must be strictly independent from the discovery cohort, ensuring no overlap or genetic relationships among participants, to prevent biases in validation efforts.

Manhattan plots and quantile-quantile (QQ) plots are fundamental visualization tools for GWAS result, each serving distinct purposes in the analysis of genetic data. A Manhattan plot displays the -log10 p-values of association tests across the genome, with chromosomes typically delineated by alternating colors, helping researchers quickly identify genomic regions with significant associations.

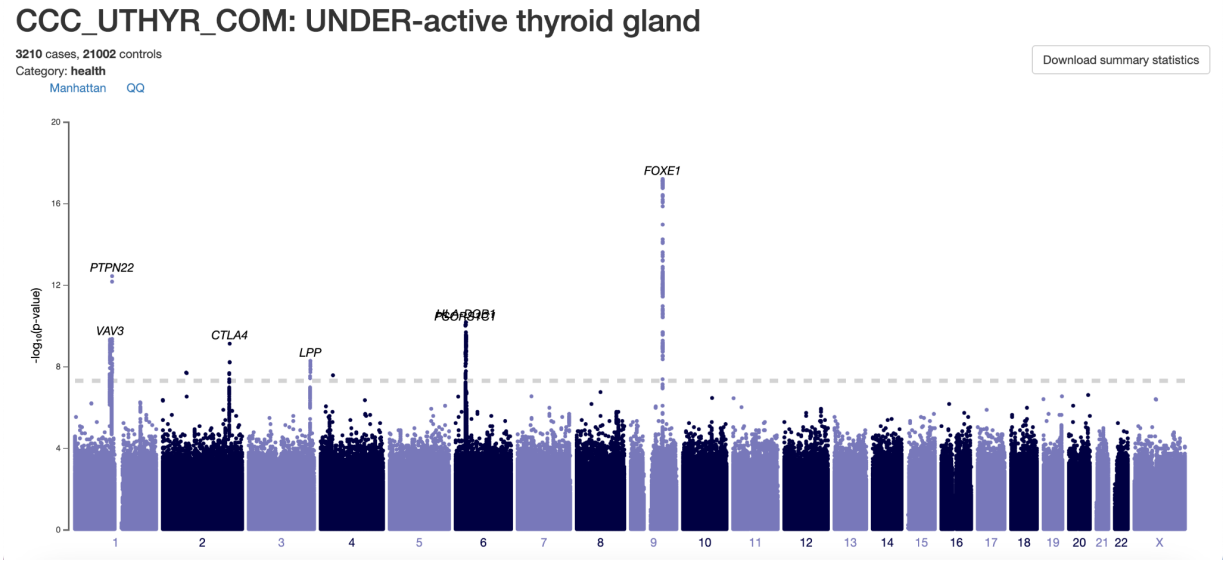


Fig.1

Manhattan plot for GWAS on under-active thyroid gland taken from the CLSA PheWeb

In contrast, QQ plots compare the observed distribution of p-values against the expected distribution under the null hypothesis of no association. Deviations from the expected line in a QQ plot indicate potential issues such as population stratification or inflation of test statistics. Genomic inflation factor lambda (λ), often derived from QQ plots, quantifies the extent of this inflation; a λ close to 1 suggests that the p-values are not excessively inflated, while values significantly greater than 1 may indicate underlying problems in the GWAS analysis or true polygenicity of the trait.

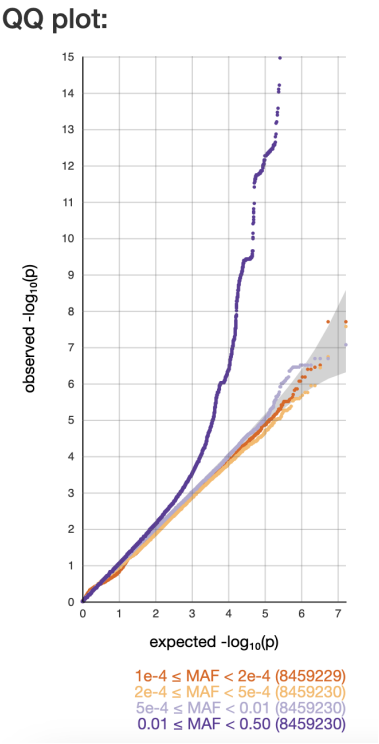


Fig.2

QQ plot for GWAS on under-active thyroid gland taken from the CLSA PheWeb

Together, these plots provide crucial insights into the data quality and the presence of true genetic signals versus artifacts, facilitating more accurate interpretations of GWAS results.

In order to prevent false positives or false negatives and biased test statistics due to population stratification, ancestry and relatedness must be carefully considered and accounted for in GWAS and, indeed, in all genetic studies. This is especially the case in data sets from participants of diverse backgrounds. These signals can result in GWAS with biassed PRSs and inflated SNP-based heritability. The outcomes of Mendelian randomization studies may also be distorted by them.

To prevent confounding, cases and controls should be matched by ancestry. For instance, if cases are defined as "using chopsticks regularly" and controls are defined as "not using chopsticks," cases in a GWAS for chopstick use would likely be drawn more frequently from an East Asian population than controls. In this example, not taking

ancestry into account would simply lead to the identification of associations between chopstick use and certain variants that are more prevalent in East Asian populations than in other populations.

Principal component analysis is typically used in GWAS to examine ancestry; clusters of individuals with related genotypes are generated using data from all individuals' genotypes. This method helps in identifying and adjusting for population stratification, which can confound the association results if not properly accounted for. By projecting the genetic data into a space defined by the principal components, researchers can visualize and correct for the ancestry-related variance in the genetic data. PCA also assists in the identification of outliers who may significantly differ genetically from the rest of the cohort, potentially due to different ancestry backgrounds. Additionally, the principal components can be included as covariates in the association analysis to reduce false-positive findings attributed to population structure differences. The effectiveness of PCA in controlling for population stratification in GWAS has been validated in numerous studies and is considered a standard practice in the field[47–49].

1.1.6 Meta-analysis

Meta-analysis is a robust statistical technique used to integrate results from multiple GWAS datasets, often conducted within the framework of large consortia like the Psychiatric Genomics Consortium (PGC)[61], the Genetic Investigation of Anthropometric Traits (GIANT) consortium[62], or the Global Lipids Genetics Consortium[63]. This approach increases the sample size and enhances the ability to detect genetic associations with traits, thus improving the statistical power of the analyses. Researchers pool data from various studies, applying tools such as METAL[64], N-GWAMA or MA-GWAMA[65], as well as quality control pipelines like those implemented in RICOPILI[46] or EasyQC[66] to ensure consistency in allele frequencies, effect sizes, and study designs.

During meta-analysis, genetic markers are aligned across studies, and models accounting for between-study heterogeneity, such as fixed-effect and random-effects models, are employed. The latter is particularly useful when there is significant variation across studies which might be due to differences in population genetics, phenotype

definitions, or study protocols. A crucial aspect of meta-analysis is the curation and standardization of data, including scaling effect sizes to a typical normal distribution and ensuring that each cohort adheres to a predetermined data analysis plan with standardized phenotypes. One of the major advantages of meta-analysis in GWAS is its capacity to detect genetic variants with small effect sizes that might not be identifiable in individual studies, and to validate previously identified genetic associations with more precise effect size estimates. Overall, meta-analysis serves as a collaborative effort that leverages the collective data of the scientific community, providing insights that would not be achievable through isolated studies alone[67].

1.1.7 Replication

Replication is a cornerstone of GWAS that bolsters the credibility of genetic associations discovered in initial studies[32]. Replication involves conducting an independent study to confirm whether the genetic variants identified as associated with a trait or disease in one population also show similar associations in another, ideally diverse, population[68]. This step is critical for several reasons: it helps to differentiate true genetic associations from those that might have arisen due to chance, population stratification, or technical artifacts[69]. Moreover, replication in diverse populations can provide insights into the generalizability of the findings across different genetic backgrounds and environments, highlighting the robustness and relevance of the genetic markers identified[2]. Successful replication adds a layer of confidence to GWAS findings, paving the way for further biological validation, functional studies, and, ultimately, translational research aimed at developing personalized medicine and interventions[32]. Thus, replication not only serves as a filter for the vast number of potential associations generated in GWAS but also as a fundamental step toward understanding the complex genetics underpinning human traits and diseases[2].

1.1.8 Post-GWAS Analyses

Post-GWAS analysis integrates various computational and experimental methods to clarify the functional effects of genetic findings and their contribution to disease mechanisms. Using in silico approaches, researchers utilize databases and

bioinformatics tools to enhance SNP mapping precision, associate SNPs with specific genes, anticipate gene functions, and investigate involved biological pathways[31]. These efforts often expand to include analyses of genetic correlations, the use of Mendelian randomization to determine causal relationships, and the generation of polygenic risk scores that combine SNP effects to estimate an individual's risk of disease[2]. Subsequent experimental validation steps, such as utilizing CRISPR-Cas9 for genome editing or conducting massively parallel reporter assays, are vital for confirming the biological significance of these findings[70], extending beyond the initial genetic discoveries. Furthermore, correlating GWAS results with data from relevant disease models in humans can shed light on the physiological effects of these genetic variations[71]. Such thorough post-GWAS evaluations are crucial for moving from mere statistical associations to a profound understanding of the genetics behind complex traits and diseases, thereby informing the development of targeted therapies and precision medicine initiatives.

1.2 Phenome-wide association studies (PheWAS)

Pleiotropy is a fundamental concept in genetics and evolutionary biology, describing the phenomenon where a single gene exerts a multifaceted influence on various phenotypic traits. This complex relationship between genes and phenotypes underscores the intricate nature of genetic architecture, with significant implications for understanding disease mechanisms, trait interrelations, and evolutionary processes.

A seminal work by Solovieff et al. (2013)[72] offers a comprehensive examination of pleiotropy in the context of human complex traits and diseases. The authors emphasize the crucial role of pleiotropy in genetic studies, highlighting its potential to reveal the biological pathways that contribute to diverse phenotypes. By considering pleiotropic effects in the design and interpretation of genetic association studies, researchers can gain a deeper understanding of the genetic basis of complex traits and diseases, ultimately informing the development of effective therapeutic strategies[72].

Recent advances in genomic technologies and large-scale GWAS have revealed that pleiotropy is a common feature of the human genome[73]. The study by Watanabe et al. (2019)[73] presents a comprehensive analysis of the genetic architecture of human

complex traits through an extensive compilation of 4,155 GWAS. Focusing on 558 well-powered GWASs, the authors delved into the extent of pleiotropy, which is the influence of single genetic loci, genes, Single Nucleotide Polymorphisms (SNPs), and gene sets on multiple traits. This exploration sheds light on the characteristics of trait-associated variants and the polygenic nature of traits[73].

A striking finding from their analysis is that the total summed length of trait-associated loci for the studied traits encompasses over half of the human genome (60.1%). This revelation underscores the extensive genomic regions implicated in complex trait variation. Even more remarkably, 90% of these loci were found to be associated with multiple traits across different domains, highlighting the pervasive nature of pleiotropy within the genome[73].

Watanabe et al. (2019)[73] identify two distinct scenarios of high locus pleiotropy: one where the same gene within a locus is associated with various traits, and another where different genes or SNPs within the same locus are tied to multiple traits. This differentiation is crucial as it indicates that the same genomic region can influence diverse traits through different genetic mechanisms, either by affecting the same gene in multiple ways or by influencing different genes within a locus. The study found that while locus pleiotropy is widespread (90%), pleiotropy at the gene level (63%) and SNP level (31%) is less common. This suggests that although a gene might be involved in multiple traits, the specific causal SNPs impacting that gene could vary across traits, affecting either its function through coding SNPs or its expression through regulatory SNPs[73].

The concept of pleiotropy holds profound implications for genomic medicine, especially as we advance into the realms of personalized medicine and genome editing. The pervasive nature of pleiotropy underscores the complexity of genetic contributions to diseases and traits, challenging simplistic models of gene-disease associations. This complexity is particularly evident when considering the effects of mutations or genetic polymorphisms, which may exhibit associations with multiple traits in varying directions. Such findings highlight the necessity of a holistic view of genetic variants, considering their multifaceted roles across different physiological contexts[74].

The implications of pleiotropy extend to drug development and the emerging field of genome editing[74]. Identifying molecular targets for therapeutic intervention requires an

understanding of the broader genetic landscape, acknowledging that targeting a specific gene may have unintended consequences due to its pleiotropic effects – adverse effects. This consideration is crucial in the era of CRISPR-Cas systems and other genome editing technologies, where altering a gene to mitigate one condition could inadvertently impact other traits or diseases associated with that gene[74].

The example of diacylglycerol acyltransferase 1 (*DGAT1*) inhibition highlights the critical role of understanding pleiotropy in drug development, particularly in the context of pharmacogenomics. *DGAT1*, targeted as a potential treatment for type 2 diabetes mellitus and obesity, came under scrutiny during a phase I trial[75]. The trial revealed that AZD7687, a reversible and selective *DGAT1* inhibitor, caused severe diarrhoea in more than half of the participants, necessitating drug discontinuation and casting doubt on the drug's viability for further development. This adverse effect aligns with subsequent genetic findings where *DGAT1* mutations were identified as a cause of severe diarrhoea in a family of Ashkenazi Jewish descent[76].

This scenario underscores the importance of genetic insights in predicting drug toxicity and efficacy. Had the pleiotropic effects of *DGAT1* been known earlier, the development strategy for *DGAT1* inhibitors might have been significantly altered, potentially saving considerable time and resources[75].

Beyond its relevance to treatment and intervention strategies, pleiotropy offers valuable insights into the molecular functions of genes and the causal relationships between traits. The case of cystic fibrosis illustrates how a gene known primarily for its role in lung disease also influences reproductive organ development, revealing the *CFTR* protein's shared role in both functions[77]. Similarly, the association between congenital hypercholesterolemia and increased heart disease risk exemplifies how pleiotropy can illuminate causal pathways, in this instance suggesting that lipid levels directly influence heart disease risk[77].

These examples highlight the critical role of pleiotropy in genomic medicine by deepening our understanding of gene functions and disease mechanisms, influencing therapeutic development, and evaluating the extensive effects of genome editing. As genomic medicine progresses, acknowledging and incorporating the pleiotropic characteristics of genes is vital for fully leveraging genetic research to enhance health

outcomes. Utilizing tools such as phenome-wide association studies (PheWAS) is essential for realizing the full potential of pleiotropy, as these studies provide comprehensive insights into the wide-ranging impacts of genetic variants across the phenome, shedding light on the complex roles genes play in health and disease.

PheWAS represents a crucial development in the field of genomic medicine, designed to systematically investigate the association between genetic variants and a wide array of phenotypes. This approach essentially reverses the direction of inquiry typical of GWAS, which traditionally focuses on identifying genetic variants associated with a single disease or trait. PheWAS, in contrast, starts with a known genetic variant and explores its effects across multiple traits or diseases within the phenome—the complete set of phenotypes expressed by an organism. The critical role of PheWAS in genomic medicine cannot be overstated. By elucidating the pleiotropic effects of genetic variants, PheWAS enables a comprehensive understanding of how a single gene can influence multiple biological pathways and clinical outcomes. This holistic view is instrumental in unraveling the complex genetic architecture underlying multifaceted diseases and traits, providing insights that are pivotal for the development of targeted therapeutic interventions and personalized medicine strategies[14,78].

PheWAS also plays a vital role in identifying potential adverse effects of drug targets early in the drug development process. By uncovering the full spectrum of phenotypic expressions associated with genetic variants, PheWAS can help predict the likelihood of off-target effects, thereby informing safer and more effective therapeutic approaches. For instance, a study utilizing the FinnGen biobank identified a missense variant in the *TM6SF2* gene (p.Leu156Pro, rs187429064) that is inversely associated with statin prescription rates, suggesting a protective effect against high cholesterol or cardiovascular diseases. Conversely, the same variant showed a positive association with insulin medication for diabetes and Type 2 Diabetes (T2D) diagnosis, highlighting a complex risk profile that may influence therapeutic decisions and risk assessments in the clinical setting[11]. This exemplifies how PheWAS can delineate the multifaceted influence of genetic variants, potentially preempting the potential adverse effects of drug targets. Moreover, PheWAS contributes to the refinement of disease classifications and

diagnoses by revealing genetic links between seemingly unrelated conditions, fostering a more integrated understanding of human health and disease[15,16].

The advent of large-scale biobanks has enabled PheWAS to be conducted on an unprecedented scale[79]. Biobanks such as the UK Biobank[6] and FinnGen[80] have amassed comprehensive genetic and health-related datasets from vast numbers of individuals, thereby providing the extensive data required for such wide-ranging analyses. These resources have made it possible to explore the effects of genetic variants across numerous phenotypes simultaneously, bringing to light intricate genetic networks and their influence on health and disease. As genomic medicine continues to evolve, the integration of PheWAS into research and clinical practice promises to significantly enhance our ability to interpret the genetic determinants of health and disease comprehensively. In doing so, PheWAS stands as a testament to the importance of recognizing and accounting for the pleiotropic nature of genes, harnessing the full potential of genetic research to improve human health outcomes. Through the systematic exploration of gene-trait associations across the phenome, PheWAS paves the way for novel discoveries and innovations in genomic medicine, emphasizing the critical need to consider the multifaceted roles of genetic variants in shaping human biology[15,16].

1.3 PheWeb

The extensive benefits of GWAS and PheWAS, including the facilitation of meta-analyses, replication studies, Mendelian Randomization (MR), Polygenic Risk Scores (PRS), and the identification of pleiotropic effects, hinge on the efficient sharing and analysis of vast arrays of summary statistics. Manual examination of thousands of studies is impractical; therefore, there is a pressing need for a platform like PheWeb.

PheWeb[18] is an easy-to-use web-based platform specifically designed to visualize, navigate, and share the results of PheWAS alongside GWAS. By aggregating and organizing vast amounts of genetic and phenotypic data, PheWeb enables researchers to systematically explore the relationships between genetic variants and a wide range of diseases and traits. Through an automated data processing pipeline, PheWeb harmonizes association summary statistics, establishes trait relationships using genetic

correlations, and annotates variants. Its interactive web interface offers detailed visualizations, including genome-wide trait summaries, localized regional insights (LocusZoom[81]), and comprehensive phenome-wide variant summaries. PheWeb seamlessly connects with The NHGRI-EBI GWAS Catalog[82], enhancing its informational scope. By enabling URL-based sharing and potential collaborative annotation, PheWeb facilitates accessibility, knowledge dissemination, and collaborative research in the realm of genetic association studies. It emerges as a valuable tool for unravelling the intricacies of human genetics, traits, and biology.

For biobanks, the absence of a tool like PheWeb represents a significant missed opportunity. Without such a platform, the rich datasets housed within biobanks may remain underutilized, as the complexity of navigating and analyzing this information can be prohibitive for many researchers. This limitation can slow the pace of discovery and hinder the translation of genetic research into clinical and therapeutic advancements. Furthermore, the lack of an accessible, user-friendly platform for sharing genetic findings may impede collaboration and the replication of results, critical components of scientific progress. As a result, many biobanks around the world have developed a PheWeb instance. Examples of PheWeb platforms tailored for biobanks around the world include UK Biobank PheWeb[83–85], TCGA PheWeb[86], FinnGen PheWeb[87], BioBank Japan PheWeb[88], CARTaGENE PheWeb[89], COLCORONA PheWeb[90], COLCOT PheWeb[91], CHARM PheWeb[92], SardiNIA PheWeb[93], KoGES PheWeb[94] and The Qatar Genome Program (QGP) PheWeb[95].

PheWeb stands out for its unique ability to display both GWAS and PheWAS results within a single, cohesive interface, making it distinct from other platforms like Genebass[96], GWAS Catalog[82], PhenoScanner[27], and AstraZeneca PheWAS Portal[97]. This integration allows researchers to seamlessly explore and compare comprehensive genetic analyses, providing a more holistic view of the data and enhancing the utility of genomic research. Its user-friendly design ensures that even those without specialized technical knowledge can easily navigate and interpret complex datasets, highlighting its role as a superior tool in the genomic research community. Additionally, PheWeb's open-source nature encourages ongoing development and customization, adapting to the evolving needs of the genomic research community. Below is a walk-through

demonstration of a PheWeb instance (UKBiobank PheWeb based on the Neale lab's GWAS[85]):

To initiate a search on the PheWeb homepage, utilize the search function to explore specific genes (such as *APOB*, *FTO*, *TCF7L2*), variants (via rsID or chromosome:position based on the genome build), or traits/phenotypes. You can access a comprehensive list of traits through the Phenotypes page.
From any section of the site, you can easily return to the homepage by clicking the PheWeb icon located in the upper left corner. For a more dynamic exploration, select the Random icon in the top menu to view a randomly chosen page, or choose Top Hits to display a table of the most significant findings hosted on the PheWeb instance. The About section provides detailed insights into the dataset and how the data was prepared and analyzed. PheWeb presents data through three primary visual formats: Manhattan plots, quantile-quantile (QQ) plots, and LocusZoom plots, alongside PheWAS plots. Below, I illustrate the process by searching for the gene *TCF7L2*:

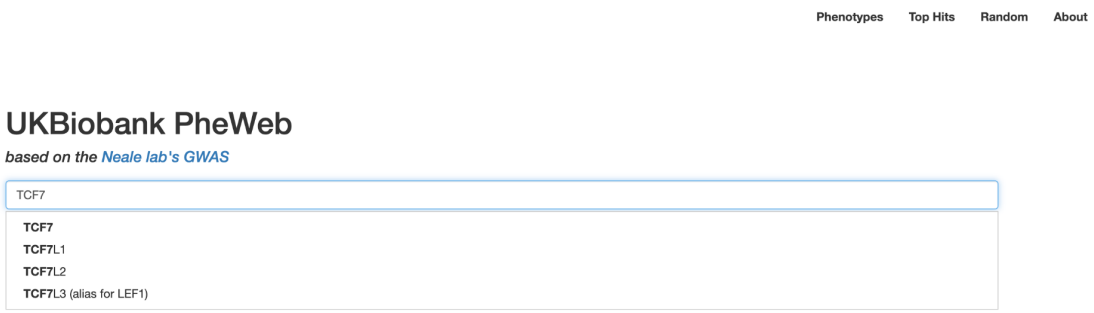


Fig.3
Homepage of PheWeb displaying the search function. The interface shows an example search for the *TCF7L2* gene,

Searching by gene highlights key associations related to the gene in tabular form, accompanied by a LocusZoom plot that visualizes linkage disequilibrium across the surrounding variant region. Selecting different table entries will adjust the LocusZoom plot to reflect the chosen data.

In my exploration of *TCF7L2*, the page displayed the selected table row, "Diabetes diagnosed by doctor" with the corresponding LocusZoom plot depicted below:

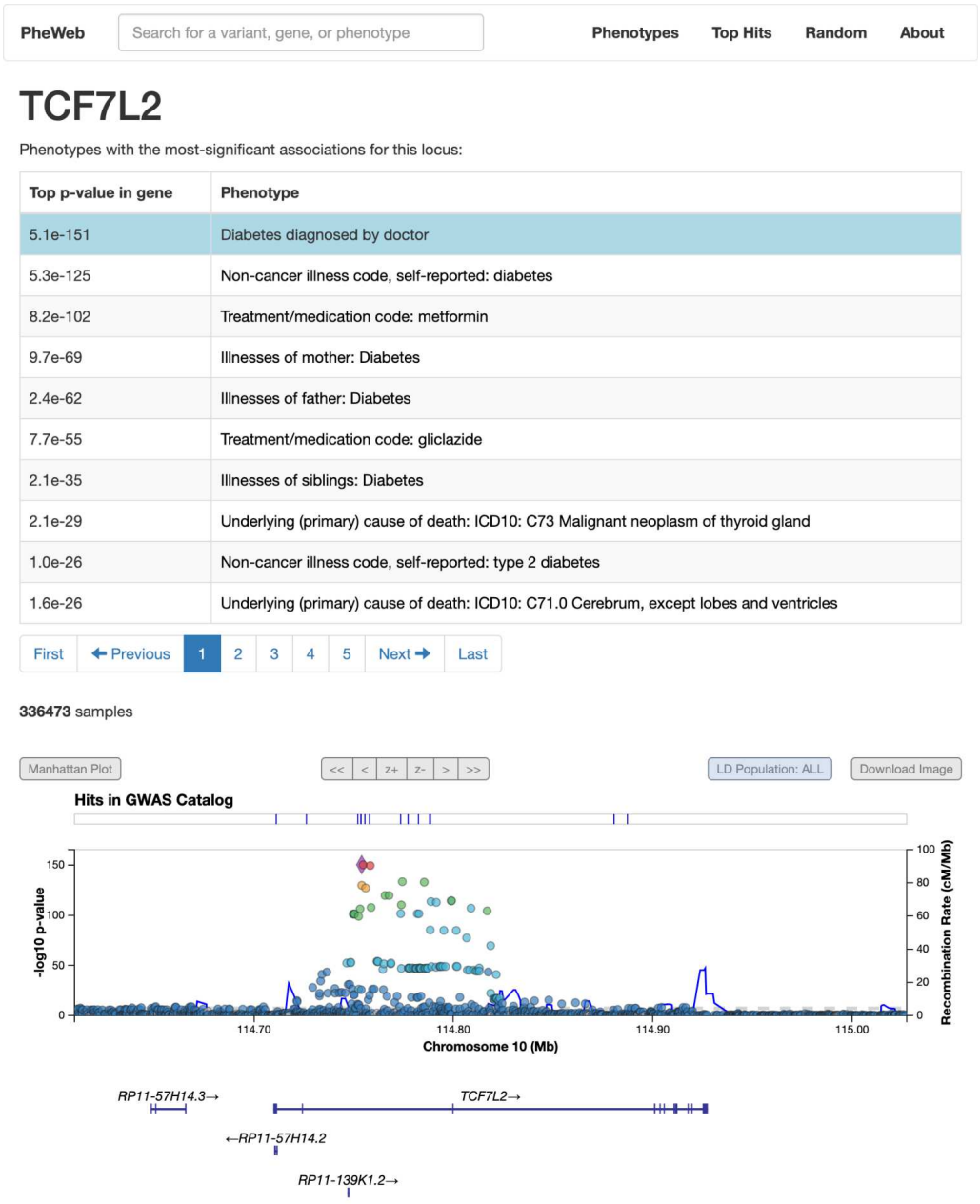
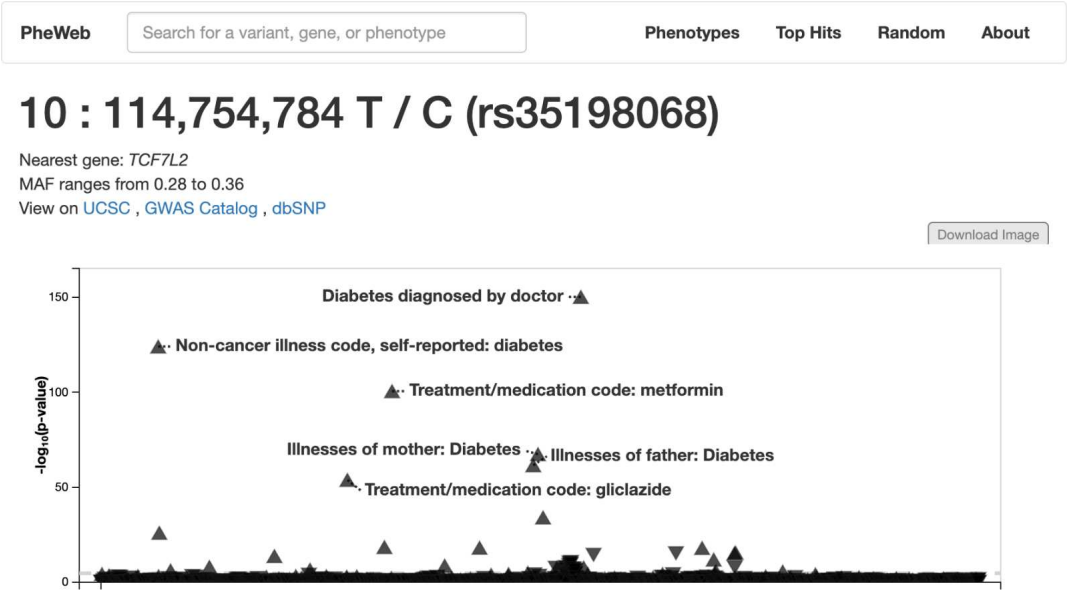


Fig.4

Detailed view of the PheWeb platform showcasing the search results for *TCF7L2*. This page lists the phenotypes with the most significant associations for this locus, with “Diabetes diagnosed by doctor” showing the top p-value.

Interactive elements are integrated into all plots. Hovering over variants within the LocusZoom plot provides additional details about them. Clicking on a variant in the LocusZoom plot will trigger a PheWAS view, demonstrating the variant's association p-value across different phenotypes. In the PheWAS plot, upward-facing triangles indicate a positive variant effect, downward-facing triangles a negative effect, and circles denote variants with imprecise beta estimates (e.g., standard error includes zero). Variants are colored based on biological grouping specified by the host.

After selecting a *TCF7L2* variant (rs35198068) from the previously mentioned screenshot, the following PheWAS view and a summary table were generated:



Search... "427.21", "Diabetes", etc.

2418 total codes

Category	Phenotype	P-value	Number of samples
	Diabetes diagnosed by doctor	7.2e-151	336473
	Non-cancer illness code, self-reported: diabetes	9.3e-125	337159
	Treatment/medication code: metformin	2.9e-101	337159
	Illnesses of mother: Diabetes	4.3e-68	308780
	Illnesses of father: Diabetes	3.0e-62	292053
	Treatment/medication code: gliclazide	1.5e-54	337159
	Illnesses of siblings: Diabetes	9.0e-35	259921
	Non-cancer illness code, self-reported: type 2 diabetes	1.0e-26	337159
	Treatment/medication code: insulin product	3.2e-19	337159
	Treatment/medication code: pioglitazone	7.3e-19	337159

First ← Previous 1 2 3 4 5 Next → Last

Fig.5

PheWAS plot for variant rs35198068 shows its association with several phenotypes. Upward triangles indicate a positive association between the variant and the phenotype, while downward triangles denote a negative association. The table below summarizes the statistical outcomes for each phenotype, including p-values.

In the PheWAS plot, selecting a specific trait takes you to a Manhattan plot of the selected trait. Below the plot, a table lists the most significant associations, and a QQ plot is stratified by minor allele frequency bins and genomic control lambda, calculated from various percentiles of the variants. From the PheWAS view, selecting "Treatment/medication code: gliclazide" directs us to the trait's Manhattan plot where hovering over it provided more details. Selecting this variant from the Manhattan plot will direct me to its regional LocusZoom plot. Scrolling further reveals the QQ plot positioned beneath the table of prominent associations.

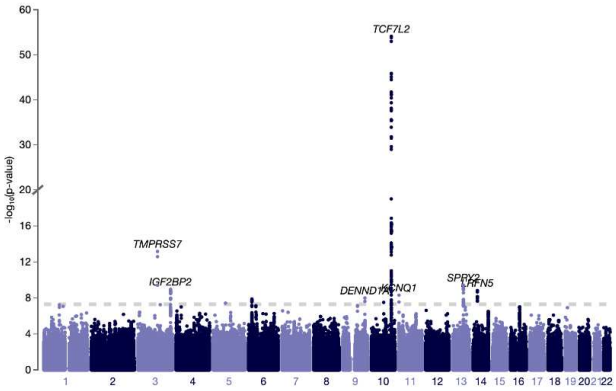
PheWeb | Search for a variant, gene, or phenotype | Phenotypes | Top Hits | Random | About

20003_1140874744: Treatment/medication code: gliclazide

337159 samples

Download summary statistics

Manhattan QQ



Top Loci:

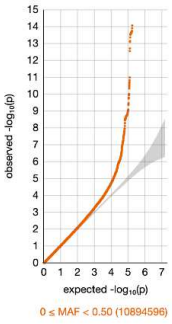
Search... "TCF7L2", "rs1861867", etc.

475 total variants

Variant	Nearest Gene(s)	MAF	P-value
10:114,754,071 T / C (rs34872471)	*TCF7L2*	196485	7.7e-55
3:111,745,882 C / T (rs748863163)	*TMPRSS7*	889.4630000000001	6.8e-14
13:80,760,789 A / G (rs9574587)	*SPRY2*	290645	3.8e-10
3:185,503,456 T / A (rs6780171)	*IGF2BP2*	210491	1.1e-9
14:43,613,474 A / G (rs369369629)	*LRFN5*	1267.33	1.5e-9
11:2,729,947 G / T (rs7925578)	*KCNQ1*	229264	4.9e-9
9:126,591,050 C / T (rs540193236)	*DENND1A*	820.4	1.0e-8
6:20,703,952 A / G (rs6931514)	*CDKAL1*	177508	1.3e-8
10:71,332,301 C / T (rs41277236)	*NEUROG3*	28726.3	3.1e-8
5:71,981,680 T / C (rs193017802)	*TNPO1*	1648.92	3.7e-8

First | ← Previous | 1 | 2 | 3 | 4 | 5 | Next → | Last

QQ plot:



0 ≤ MAF < 0.50 (10894596)

GC lambda 0.5: 1.048
GC lambda 0.1: 1.049
GC lambda 0.01: 1.060
GC lambda 0.001: 1.103
(Genomic Control lambda calculated based on the 50th percentile (median), 10th percentile, 1st percentile, and 1/10th of a percentile)

Fig.6

Manhattan and QQ plots for “Treatment/medication code: gliclazide”. The top loci table lists significant associations and their statistics.

Chapter 2: Materials and Methods

2.1 CLSA

For the purpose of this project, we will use the dataset provided by the Canadian Longitudinal Study on Aging (CLSA)[29]. The Comprehensive Cohort of CLSA was created to offer an opportunity to study the impact of genetic and environmental factors on human disease as well as the aging process. Genome-wide genotyping was performed on a total of 26,622 individuals from the CLSA Comprehensive cohort, which included males and females aged 45 to 85 who were recruited between 2010 and 2015. Genomic information from the CLSA Comprehensive cohort includes whole-genome genotyping data for 794,409 markers and whole-genome imputed data for almost 308 million genetic variants. The TOPMed reference panel was utilized for genotype imputation. Both genetic markers and samples were subjected to extensive quality control metrics. Also, Copy number profiling was used to find sex chromosomal abnormalities. 24,655 (92.6%) of the 26,622 genotyped participants were found to be of European ancestry based on the analysis done by the CLSA team. These genomic data are connected to the CLSA's longitudinally collected physical, lifestyle, medical, economic, environmental, and psychosocial aspects. Potential drawbacks could be the relatively poor genotyping coverage in individuals with non-European ancestry, which can be significantly improved by employing an imputation reference panel with high diversity. The CLSA is still ongoing overall. Participants in the present genomic subcohort will continue to provide follow-up information, such as DNA methylation and metabolomic data. Through the CLSA data access application process, this genetic data resource is accessible upon request.

2.1.1 Phenotype data

CLSA represents a comprehensive initiative aimed at understanding the complex factors underpinning aging through the collection of data from over 51,000 participants. This longitudinal study has a dual approach to data collection, encompassing both the Tracking and Comprehensive assessments, which are instrumental in capturing a wide array of data points from a diverse participant pool at baseline. Our project is specifically focused on analyzing the baseline data, which serves as the foundation for our GWAS within the CLSA biobank PheWeb platform. Below is an overview of the data collection process implemented by the CLSA team.

Tracking Assessment: This component of the CLSA involves data collection from more than 21,000 participants via telephone interviews. These interviews are designed to gather a broad spectrum of information, offering insights into various aspects of health, lifestyle, and aging. The Tracking assessment facilitates the inclusion of participants who may not be able to partake in more extensive in-person evaluations, thereby ensuring a wider representation of the Canadian aging population.

Comprehensive Assessment: The remaining 30,000+ participants contribute data through a more in-depth process that includes in-home interviews and visits to data collection sites. This approach allows for the gathering of detailed information through a combination of interviews, physical assessments, and neuropsychological tests. The Comprehensive assessment is instrumental in collecting rich, multifaceted data that spans biological, medical, psychological, social, lifestyle, and economic domains.

The baseline data collection for the CLSA, conducted between 2011 and 2015, employed a variety of tools to ensure a comprehensive capture of participant information:

Telephone Interviews: Conducted in two sessions, the initial 60-minute interview covered a wide range of topics and was carried out from September 2011 to May 2014.

A subsequent 30-minute follow-up interview was conducted from September 2013 to December 2015 to update and expand on the initial data set.

In-Home, Face-to-Face Interviews: These 90-minute interviews were conducted from May 2012 to May 2015, offering a personal approach to data collection. This method enabled the collection of detailed information directly from participants in the comfort of their homes.

Data Collection Site Visit Interviews: Participants also attended site visits that lasted approximately 2.5 hours, conducted from May 2012 to May 2015. These visits included comprehensive evaluations such as Contraindications, Neuropsychological Battery, and Disease Symptoms assessments, providing a depth of data that complements the information gathered through interviews.

In addition to these primary data collection methods, all participants were contacted by telephone 18 months after their initial interview to complete the Maintaining Contact Questionnaire (MCQ). This questionnaire included additional data collection to maintain engagement with participants and update key information between the major assessment waves.

2.1.2 Genotype data

The genotype data for CLSA participants was carefully collected and processed by the CLSA team to ensure high-quality genetic information for subsequent analyses. The DNA extraction and genotyping procedures were centralized at the McGill and Genome Quebec Innovation Centre, located in Montreal, Canada. Participants' genotypes were determined using the Affymetrix UK Biobank Axiom array, a high-density genotyping platform.

The choice of the Affymetrix UK Biobank Axiom array for genotyping is noteworthy, as this array was specifically crafted for the genotyping of approximately 450,000 individuals within the UK Biobank cohort. The array is well-regarded for its targeted coverage of disease-associated variants, coding variants, and a robust selection of

single nucleotide polymorphisms (SNPs) for imputation, particularly in populations of European descent. The latter is especially relevant given that over 90% of the genotyped participants in the CLSA are of European ancestry. This targeted design enables effective downstream imputation and increases the likelihood of discovering meaningful genetic associations within this population.

The current data release from the CLSA encompasses genotype information for 26,622 participants, featuring 794,409 genetic markers directly genotyped from the array. In addition to these genotyped markers, the dataset was enriched with approximately 308 million genetic variants through imputation using the TOPMed reference panel, further enhancing the breadth of genetic data available for analysis. The TOPMed program, known for its large and diverse reference panels, provides a valuable resource for imputation, potentially improving the fine-mapping of genetic associations and identification of causal variants.

The quality control (QC) measures applied to the genotype data were largely reflective of the rigorous protocols established by the UK Biobank, ensuring consistency and reliability in the dataset. These QC procedures included checks for marker and sample quality, adherence to Hardy-Weinberg equilibrium, call rates, and other standard metrics vital for the integrity of genetic studies. It is important to note that all genomic positions reported in the data align with the GRCh37/hg19 human genome build. Furthermore, the genotype dataset includes data for control samples that were used during the array genotyping process. The inclusion of these controls is a standard practice that aids in the calibration of the genotyping process and serves as an internal check to validate the genotyping results. In summary, the genotype data for CLSA participants is carefully curated and quality-controlled to serve as a solid foundation for uncovering genetic underpinnings of health-related traits.

2.2 Ancestry Inference

2.2.1 PCA analysis

Principal Component Analysis (PCA) is a prevalent statistical technique in GWAS, employed to manage the challenges posed by population stratification, which can

substantially skew study outcomes. Population stratification involves variations in allele frequencies across subpopulations within a broader group, typically arising from diverse genetic backgrounds. Without adequate adjustment for these differences, erroneous correlations may arise between genetic markers and the traits or diseases under investigation, falsely suggesting genetic influences when they may merely reflect underlying population differences[47].

The process of ancestry inference within our GWAS for the CLSA biobank PheWeb platform was executed through a series of computational steps, utilizing the in-house HGDP_1KG Ancestry Inference pipeline[98]. This pipeline leverages the Human Genome Diversity Project (HGDP)[99] along with the 1000 Genomes project[52] to infer the ancestry of individuals. This methodological framework was designed to intersect variant call format (VCF) files of the study data with the reference data, convert genotype data, perform PCA, and ultimately, project study samples onto reference principal components (PCs) before applying a Random Forest model for ancestry classification.

Initially, the pipeline intersects reference and study VCF files for each chromosome using bcftools[100] isec, ensuring that only common variants between the datasets were retained for analysis. The bcftools[100] concat command merged VCFs across all chromosomes, creating comprehensive reference and study VCF files. These files were then converted into genotype and site files using a custom vcf2geno[101] tool, which prepared the data for PCA.

PCA was conducted on the reference dataset using the LASER[102,103] tool, with parameters set to derive 20 principal components, capturing the major axes of genetic variation across the human genome. This step is pivotal for ancestry inference, as it establishes a multidimensional space in which the genetic diversity of the study samples can be compared to known reference populations.

Study samples were then projected onto the reference principal component space. The use of principal components to represent genetic variation allows for a nuanced analysis that accounts for the complex structure of human genetic diversity.

The final step involved the application of a Random Forest model, trained on the reference data, to predict the genetic ancestry of our study samples. We used the train_test_split function from the sklearn library which allows for a random yet

reproducible division of the reference data into test data and training data. We achieved the test data prediction accuracy of 0.997 using 25% of our reference data as test data. By setting the number of principal components to use, a minimum probability threshold for assigning population labels, and a seed for reproducibility, the model classified each study sample into a genetic ancestry group based on its genetic makeup. The Random Forest approach was selected for its efficacy in handling high-dimensional data and its ability to provide an estimate of classification uncertainty, which is crucial for accurate ancestry inference.

The process of determining the optimal minimum probability threshold for assigning population labels was methodically approached by testing various thresholds and assessing the outliers compared to the reference clusters on the PC plots as well as the concordance between our predicted European (EUR) ancestry assignments and those classified as European (EUR) ancestry according to analysis done by the CLSA team. We initiated a series of tests where different threshold values were incrementally tested, ranging from 0.5 to 0.9 in steps of 0.1. For each threshold value tested, the overlap was quantified by calculating the proportion of samples we classified as EUR that were also classified as EUR by the CLSA team. This comparison provided a direct measure of agreement between the two classification approaches, allowing us to assess the sensitivity and specificity of our ancestry predictions in relation to an external standard. The optimal balance between sensitivity and specificity was achieved at a minimum probability threshold of 0.78 which happens to be the same threshold used by the Genome Aggregation Database (gnomAD)[104]. This threshold demonstrated the highest overlap with the CLSA team's EUR predictions. This threshold was therefore selected as the standard for assigning population labels in our ancestry inference analysis, ensuring that our EUR ancestry classifications are both accurate and consistent with established classifications.

Lastly, to ensure the quality and accuracy of our results from the pipeline, we employed a quality control approach focusing on three key measures: the count of nonmissing loci utilized in the analysis, the Procrustes similarity score reflecting the accuracy of the placement of each study individual into the reference ancestry space, and the Z score for each individual. The Z score is particularly crucial as it indicates whether an

individual's ancestry is adequately represented in the reference panel, ensuring the reliability of our ancestry inference[102]. Z score could be especially helpful when a European reference panel is wrongly used for samples of non-European descendant[102]. The Procrustes similarity scores showed an exceptionally high level of agreement between individual PCA maps and the reference panel, with a minimum score of 0.999933. For the nonmissing loci used in the analysis, the count was substantial across all individuals. The minimum number of loci used was 131,639, the maximum was 138,994. The distribution of Z scores, which measure how well each individual's ancestry is represented in the reference panel, indicated some issues. While the mean Z score was 0.1101, the range from -17.7813 to 23.9493 was concerning, as it suggests some individuals' ancestries may not be well-represented by the reference panel used. We then proceeded to exclude samples with an absolute Z score greater than 5. This filtering step led to the removal of 186 samples. With the minimum probability threshold of 0.78, we have:

Ancestry Label	**Count**
EUR	24,505
AFR	146
AMR	98
CSA	267
EAS	296
MID	65
Undefined	1,235
Total	26,622

Table 2.
CLSA inferred genetic ancestry count

The pie chart below depicts that the majority of our dataset is labeled as EUR.

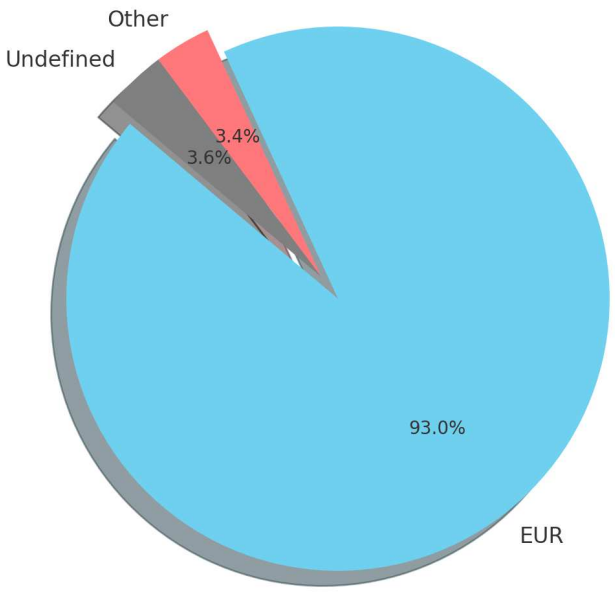


Fig.7

Pie chart depicting the ancestry distribution within the CLSA dataset post-application of the ancestry inference pipeline. The majority, 93%, is classified as having European genetic ancestry (EUR).

Detailed PC plots and analyses for the output of the ancestry inference pipeline can be found in the Supplementary Figures document. Below, we show the tri-panel scatter plots representing principal component analysis (PCA) of individuals from the CLSA labeled as having European genetic ancestry (EUR) with a minimum probability of 0.78, overlaid on a reference panel consisting of the Human Genome Diversity Project (HGDP) and 1000 Genomes (1KG).

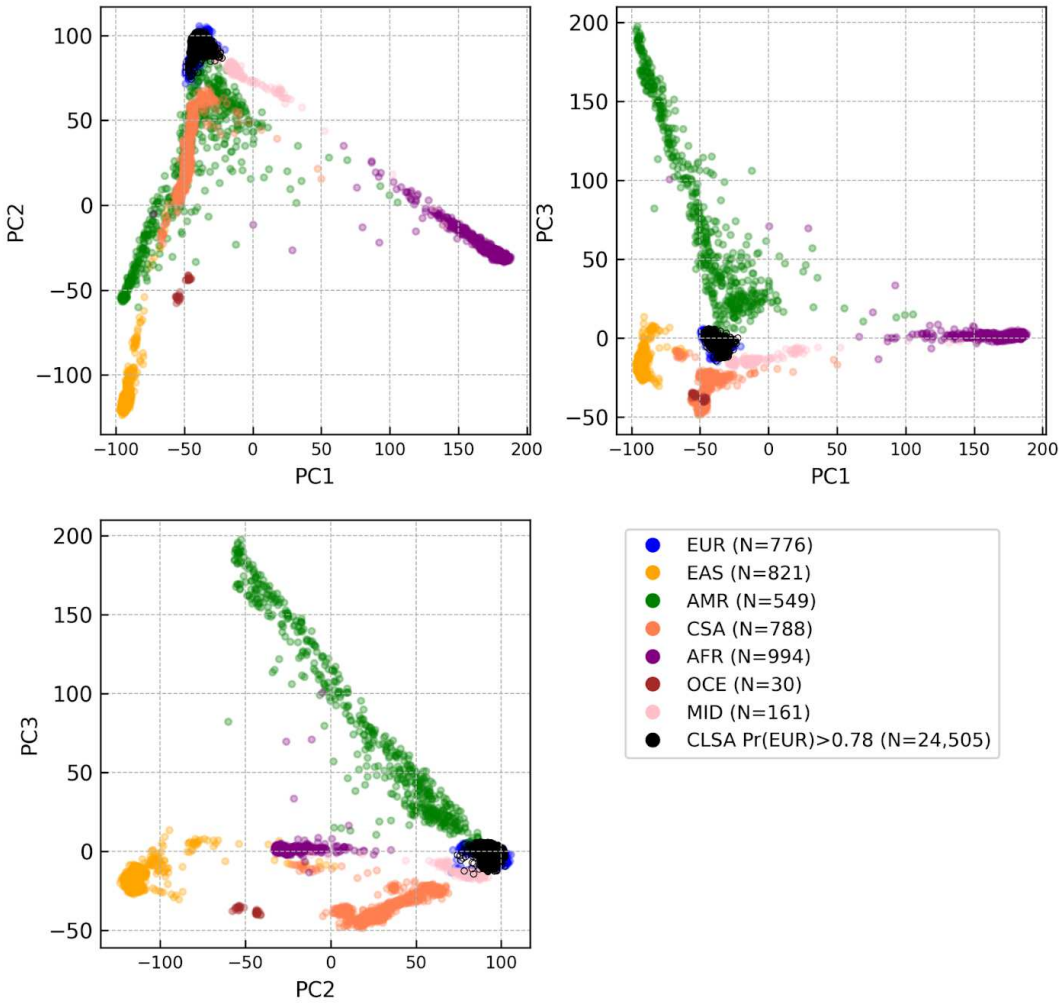


Fig.8

Tri-panel scatter plots representing principal component analysis (PCA) of individuals from the CLSA labeled as having European genetic ancestry (EUR) with a minimum probability of 0.78. These CLSA samples are overlaid on a reference panel consisting of Human Genome Diversity Project (HGDP) and 1000 Genomes (1KG). Each colored cluster represents a distinct ancestry group from the reference panel, while the black cluster highlights the CLSA samples labeled as European, providing insight

into their genetic positioning relative to global genetic diversity.

2.2.2 How CLSA handled population structure

In refining our approach to ancestry inference within the GWAS framework for the CLSA biobank PheWeb platform, we aimed to improve the differentiation of subpopulations by utilizing larger and more diverse reference panels. The CLSA team's methodology, as detailed in their Genome-wide Genetic Data Release (version 3), primarily relied on PCA analysis and self-reported ancestry using the 1000 Genomes Project data. This section will compare our approach with the CLSA's strategy to highlight the differences.

The CLSA team started their population structure analysis by extracting Affymetrix UK Biobank Axiom array markers present within the 1000 Genomes dataset, adhering to stringent criteria including a minor allele frequency (MAF) greater than 0.05, Hardy-Weinberg equilibrium (HWE) p-value exceeding 10^{-6}, and the exclusion of ambiguous strand markers (such as A/T or C/G), among others. Following these criteria, they performed linkage disequilibrium pruning to retain markers that ensured genetic independence and clarity. This process culminated in a set of 87,848 markers, which were then used to compute principal component loadings on the 2504 individuals from the 1000 Genomes Phase 3, subsequently projecting the CLSA cohort onto these principal components.

The CLSA team then utilized PCA analysis to cluster the top four principal components into six distinct clusters. They identified a predominant cluster, referred to as "cluster 4", which closely aligned with European ancestry populations from the 1000 Genomes project and contained the majority of individuals who self-reported European ancestry.

While this methodological approach provided a baseline for understanding population structure within the CLSA cohort, it presented limitations by relying on a single reference dataset and quasi-subjective cluster-based analysis for ancestry inference.

Our methodology addresses these limitations by incorporating both the 1000 Genomes Project and the Human Genome Diversity Project (HGDP) data for reference, expanding the genetic diversity scope accessible for comparison. By using two reference panels, we achieved better differentiation of subpopulations. Furthermore,

instead of quasi-subjective clustering, we employed a Random Forest model for ancestry classification, leveraging the principal component coordinates of study samples alongside pre-defined models of reference population ancestries. This model was trained and validated using the combined reference datasets.

We used the CLSA ancestry labels as a sanity check for our Random Forest approach, ensuring reliability. By comparing our EUR predictions with CLSA classifications and testing various thresholds, we found that a 0.78 threshold provided the best accuracy. Additionally, CLSA labels were not used to create the plot; we only used reference panel labels. Our analysis did not incorporate the CLSA ancestry analysis directly. Our method offered some advantages in certain aspects, particularly in differentiating subpopulations due to the larger and more diverse reference panels.

2.3 Phenotype data analysis

In this project, we focused solely on binary phenotypes to ensure the feasibility of completion within the MSc timeframe, as analyzing continuous traits would have required significantly more time. Our collaborators at Université de Montréal (UdeM) in the Gagliano Taliun Lab handled the analysis of continuous traits, ensuring that the final version of PheWeb contains both binary and continuous traits.

We identified 350 binary phenotypes for our study, with the primary aim of selecting binary phenotypes that not only possess intrinsic relevance for GWAS but also exhibit a potential for yielding statistically powerful results. Given the inherently subjective nature of determining "GWAS relevance" this criterion was deferred to the concluding stage of our analysis, thereby allowing for a more systematic and inclusive preliminary screening of potential binary phenotypes.

We initially categorized the variables into distinct domains: Identity, Socioeconomic, Behavioral, Health, Measurements, Medications, and Diet. These labels are used in the PheWAS plots available in the first version of the CLSA PheWeb.

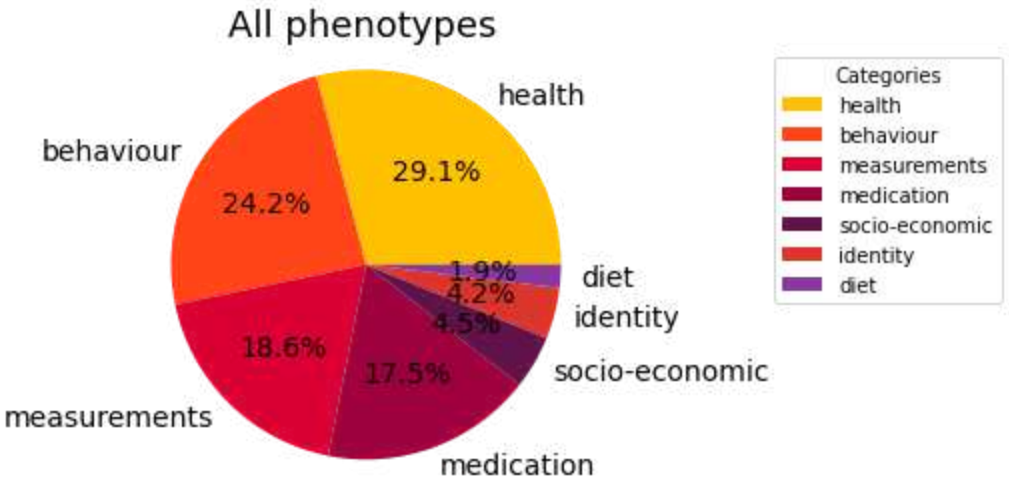


Fig.9

Pie chart showing the distribution of all variable categories in the dataset, with the largest segment representing health-related variables, followed by behavioral factors.

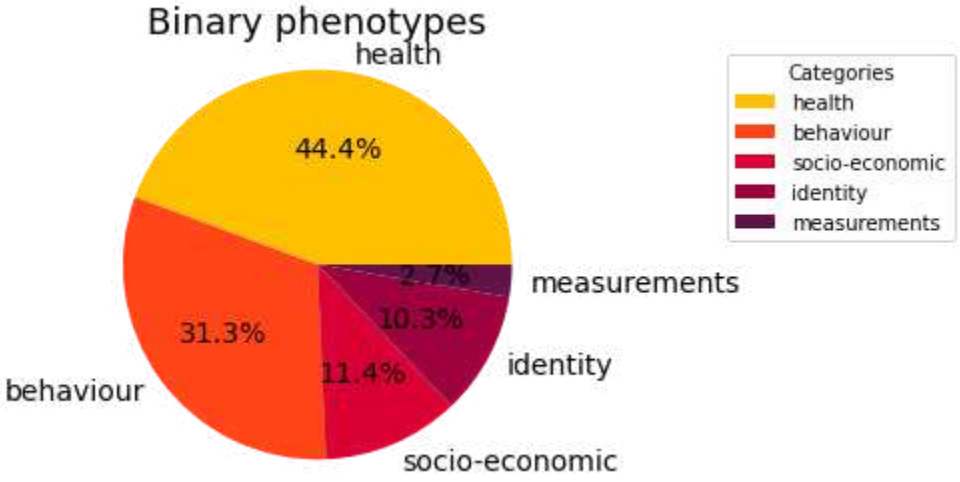


Fig.10

Pie chart showing the distribution of variable categories within binary phenotypes, with the largest segment representing health-related variables, followed by behavioral factors.

We followed the steps below to systematically select our binary phenotypes for running GWAS.

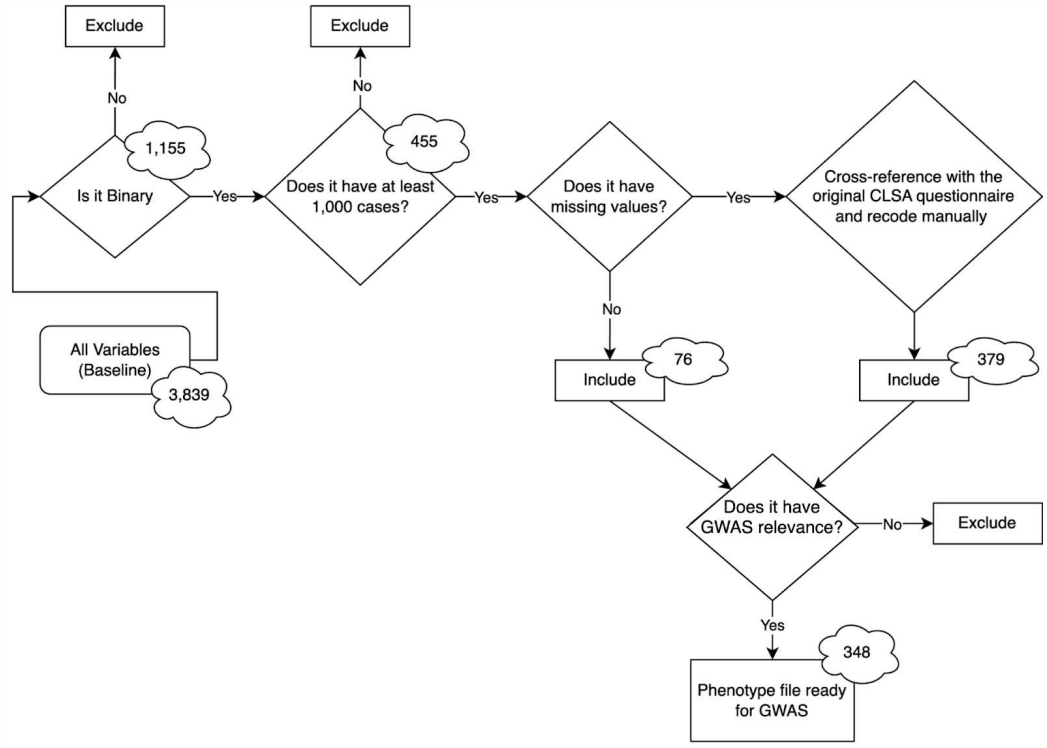


Fig.11

Flowchart depicting the steps implemented to select phenotypes for conducting GWAS. numbers in the clouds represent the number of phenotypes at that stage.

From 3,839 baseline variables, we isolated 1,155 binary variables which includes variables that solely contained values of 0 and 1 (along with "Don't Know" (DK) and "Refused" (RF) responses or missing values) and variables exclusively comprising values of 1 and 2 (along with "Don't Know" (DK) and "Refused" (RF) responses or missing values).

Building upon the initial step of identifying binary variables within the baseline data of CLSA, the second step in our phenotype data analysis focused on selecting those binary variables that also met a minimum case count of 1,000. This threshold was established based on power calculations derived from the University of Michigan Genetic Association Study (GAS) Power Calculator[105], ensuring that our selected

phenotypes would yield studies with sufficient statistical power to detect true genetic associations.

In the context of GWAS, statistical power refers to the likelihood of correctly identifying a significant association between a genetic variant and a phenotypic trait. This concept is crucial for planning and interpreting GWAS because it influences the certainty of genetic association discoveries. Factors affecting statistical power include the size of the sample, the impact magnitude of the genetic variant, its allele frequency, and the chosen significance level for testing hypotheses. The significance level, usually set at 0.05 or lower, manages the Type I error risk— the chance of incorrectly reporting a positive association—considering the extensive number of tests performed in GWAS[2,106].

Increasing the sample size enhances the ability to detect smaller genetic effects, and greater allele frequencies improve the likelihood of identifying genetic influences due to greater variation within the population[107]. Additionally, the effect size, indicating the extent a variant impacts a trait, is positively correlated with statistical power; thus, larger effect sizes allow for the achievement of equivalent power with fewer subjects[108,109].

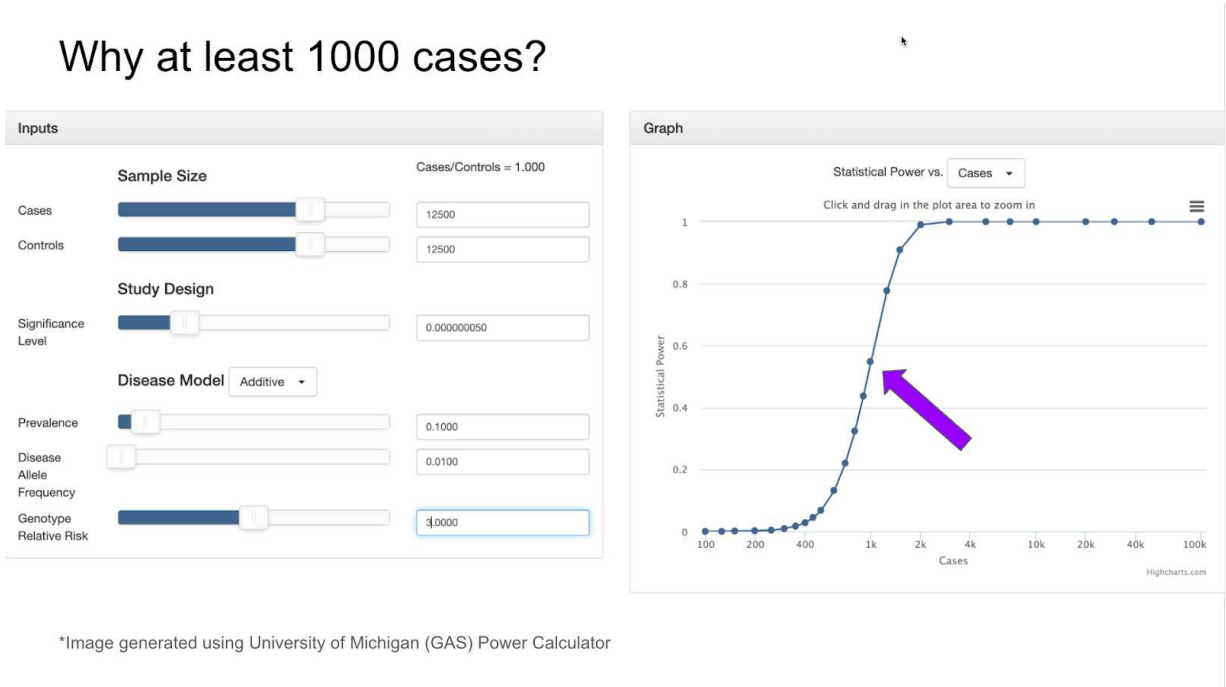


*Image generated using University of Michigan (GAS) Power Calculator

Fig.12

This figure illustrates the relationship between the number of cases and statistical power in genome-wide association studies, highlighting the significant increase in power with a minimum of 1,000 cases. The inputs on the left detail the parameters used in the University of Michigan's Genetic Association Study (GAS) Power Calculator, while the graph on the right shows the power as case numbers rise, with the arrow indicating the power of 0.56 with 1000 cases.

Considering these principles, our selection of binary variables with at least 1,000 cases is designed to optimize the statistical power within the practical constraints of the CLSA dataset. The post-GWAS quality control processes are similarly informed by these concepts, ensuring that our findings are reliable. To inform our decision for the case count threshold, we considered two scenarios using the GAS Power Calculator, varying parameters such as genotype relative risk and disease allele frequency.

From the screenshots provided, which illustrate the inputs and results from the GAS Power Calculator, we can observe that with a sample size of 12,500 cases (the total sample size is about 25,000 people) and an equal number of controls, under the conditions specified, the statistical power of our GWAS ranged from approximately 0.56 to 0.96. This wide range of power underscores the importance of selecting phenotypes with a sufficient number of cases, as achieving high statistical power is essential for the reliability of GWAS findings. By setting the minimum case count at 1,000, we aimed to strike a balance between inclusivity of various phenotypes and the assurance of robust statistical power.

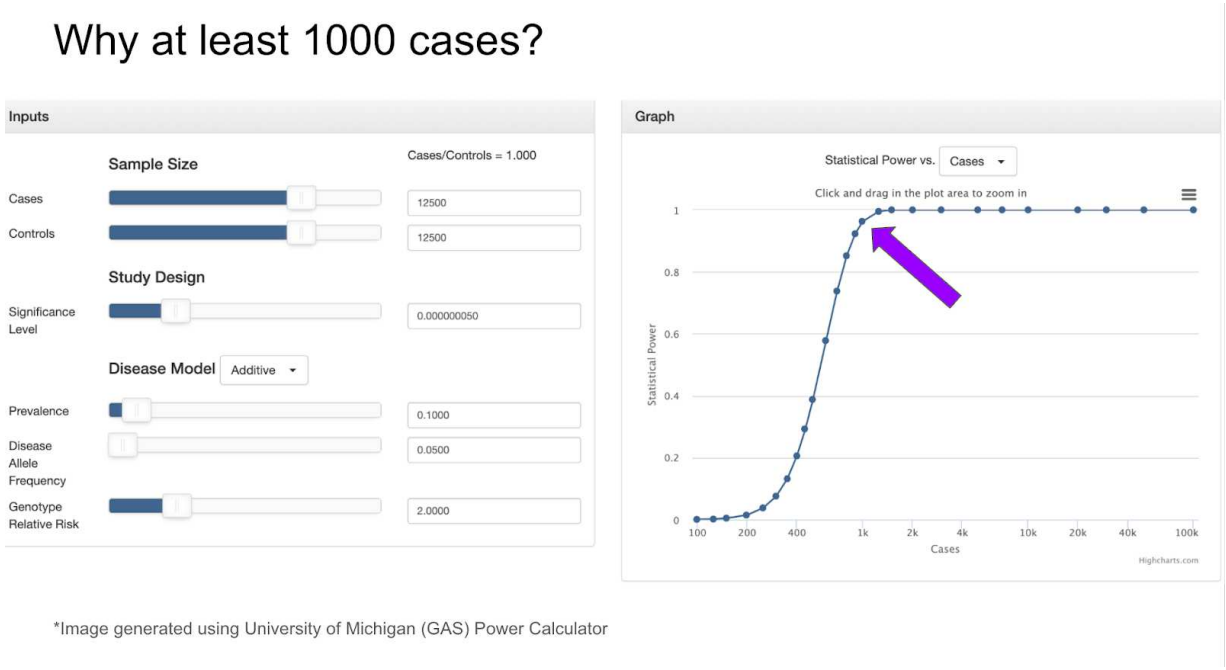


Fig.13

This figure illustrates the relationship between the number of cases and statistical power in genome-wide association studies, highlighting the significant increase in power with a minimum of 1,000 cases. The inputs on the left detail the parameters used in the University of Michigan's Genetic Association Study (GAS) Power Calculator, while the graph on the right shows the power as case numbers rise, with the arrow indicating the power of 0.96 with 1000 cases.

Following the identification of binary variables with at least 1,000 cases, we encountered a significant challenge: a substantial number of these variables were marred by a high incidence of missing data, denoted as 'NA' (not available). This observation prompted a thorough investigation into the structure of the questionnaire from which the data originated, revealing that many of the variables with excessive missing data were, in fact, nested questions contingent on the response to a preceding parent question.

Nested questions are follow-up items presented to participants only if they provide a specific response to an earlier, related question. In the context of the CLSA questionnaire, for example, all participants were required to answer general questions

such as their sex and age. However, more specific queries such as detailed questions about cancer diagnoses, were conditioned upon a participant affirmatively acknowledging a history of cancer. Those who reported not having cancer would not see or respond to the subsequent cancer-specific questions, thereby resulting in 'NA' entries in the dataset for these individuals.

Recognizing the need for accuracy and completeness in our GWAS dataset, we embarked on a labor-intensive task: manually reviewing each variable to discern its relationship with its parent question. By examining the CLSA questionnaire, we determined the context in which 'NA' responses were actually indicative of a participant being a control for the nested question. This meticulous process allowed us to recode 'NA' values to control status where appropriate, thereby significantly reducing the number of missing values and improving the dataset's integrity.

An illustrative example of this process can be seen in the variable "SMK_TYPEOT_PI_COM", which represents whether an individual has ever used tobacco pipes. Originally, this variable had 22,966 'NA' responses. However, upon manual recoding—whereby we identified individuals who should be classified as controls based on their answers to the parent smoking question—the number of missing values was reduced to just 3. The enormity of this undertaking cannot be overstated. Each variable required individual attention, as the conditions underpinning the parent-nested question relationship varied widely, precluding the possibility of a one-size-fits-all automation solution. The commitment to this process was significant, demanding considerable time to ensure that each variable was accurately recoded.

In the final step, we excluded variables irrelevant to genetic association studies. GWAS relevance, though subjective, required each phenotype to have a plausible link to genetic variation. We removed non-informative variables, such as procedural consents and administrative data unrelated to genetics. Additionally, we excluded indecisive or refusing response options like 'Don't know' or 'Refused to answer' because they do not contribute to meaningful genetic data. After this final review, we distilled the list to 350 binary phenotypes as we added two phenotypes for type 1 and type 2 diabetes based on responses to the diabetes question.

2.4 Liftover

The CLSA genotype array data was originally generated in GRCh37. Although the CLSA team used TOPMed for imputation, which automatically performs liftover to GRCh38, we conducted the liftover ourselves to ensure compatibility between the steps of the Regenie algorithm. Specifically, the first step of Regenie relies on common independent variants selected from the genotyping array data. Since all the imputed variants used in the second step of the algorithm are in GRCh38, it was necessary to have our genotyping array data in GRCh38 as well. Therefore, we used the UCSC Genome Browser's LiftOver tool to convert the genomic coordinates of the CLSA genotype array data from GRCh37/hg19 to GRCh38/hg38. It is important to note that the final dataset used for the PheWeb platform is based on the TOPMed imputation and is in the GRCh38 reference genome build.

Liftover is a computational process used to translate genomic coordinates from one human reference genome build to another. Human genome builds are essentially different versions of the human genome sequence. Genome builds differ due to advancements in sequencing technologies, deeper insights into genomic variations, and more precise mapping of genetic loci, leading to updates in nucleotide numbering, new sequence inclusions, and gene placements. Liftover procedures ensure continuity and comparability across different builds, crucial as newer assemblies become standard.

Before initiating the liftover process, we performed stringent QC checks at both the marker and sample levels. These steps are crucial to ensure the integrity of the genetic data before and after the liftover, as the coordinate transformation could potentially exacerbate any pre-existing errors or quality issues. Our marker-based QC involved the exclusion of genetic variants that exhibited discrepancies or failed quality tests in at least one batch of data. This included markers that showed frequency discordance across five genotyping batches, those that failed Hardy-Weinberg equilibrium tests, those with control genotype discordance, and markers that failed sex genotype frequency discordance tests. In addition, we filtered out insertion or deletion polymorphisms (indel), as these variants are not supported by liftover tools. These filters resulted in the removal of 131,892 variants, substantially reducing the potential for data inconsistencies post-liftover. For sample-based QC, we identified and excluded samples

with discrepancies between self-reported sex and chromosomal sex, as well as samples with missing sex data. We also excluded samples flagged for extreme heterozygosity or genotype missingness, which could indicate sample contamination or poor data quality. Following these criteria, we excluded a total of 63 samples.

Subsequently, the filtered dataset was confirmed to contain 667,335 variants and 26,563 individuals, all of which passed our stringent QC filters. This dataset was then poised for the liftover to hg38, which is the latest human reference genome assembly offering the most up-to-date and precise genomic coordinate system. This step is critical, especially considering that the CLSA imputation files were in hg38, and consistency in genomic builds is essential for accurate genomic analyses and subsequent interpretation of GWAS results.

2.5 Testing for associations

Linear mixed models (LMMs) have become a cornerstone in the analysis of complex traits within GWAS, primarily due to their ability to effectively control for population stratification and relatedness among samples[110]. LMMs introduce both fixed and random effects to account for the genetic relationships between individuals, allowing for the accurate identification of genetic variants associated with traits while minimizing false positives.

The application of LMMs in GWAS is particularly advantageous because it enables researchers to correct for the subtle genetic structure within populations that can confound association signals. By incorporating random effects, LMMs can adjust for the kinship matrix, which represents the genetic similarity between pairs of individuals, thus controlling for both known and cryptic relatedness[110]. This aspect is crucial for studies involving samples from diverse backgrounds or family-based cohorts where relatedness can introduce bias.

Moreover, LMMs are utilized in GWAS to enhance the power of detecting true genetic associations under a variety of genetic architectures. They are especially beneficial for traits influenced by many small-effect loci distributed across the genome. The flexibility of LMMs to model multiple layers of random effects makes them suitable for dissecting

the genetic variance attributed to both polygenic effects and specific genetic markers[110,111].

The mathematical framework of LMMs can be described as follows (the model below has the same covariates as the ones that were used in this study):

$$y \sim \beta_0 + \beta_1 x + \beta_2 z_{genotyping\ batch} + \beta_3 z_{sex} + \beta_4 z_{age} + \beta_5 z_{age^2} + \beta_6 z_{PC1} + \ \dots \ + \beta_{25} z_{PC20} + g + \varepsilon \quad (1)$$

The continuous phenotype (y) is modeled as a combination of several components: the fixed effects of genetic variants and covariates, the random polygenic effects from the entire genome, and random non-genetic residual effects. The fixed effects typically include the intercept (β_0), the effect of the genetic variant (β_1) under investigation, and other known covariates ($\beta_2, \beta_3, \dots, \beta_{25}$) such as age, sex, and principal components that account for population structure. The polygenic effect (g) captures the random effect of the entire genome, reflecting the contribution of multiple genetic factors to the phenotype. The residual effect (ε) includes random environmental and other non-genetic factors. Note that, as mentioned earlier in the background section, logistic regression is used for binary phenotypes by employing a logit link function. So the left hand side of equation (1) becomes $logit(y)$ rather than y.

The regression framework employed in GWAS is predicated on modeling the phenotype as a function of individual genotypes across a multitude of genetic markers. In its essence, the linear regression model applied in GWAS can be delineated as follows[112]:

$$y_i = \beta_0 + \sum_{k=1}^{M} \beta_k x_{ik} + \varepsilon_i \quad (2)$$

In this model, which can be expanded to account for additional confounding factors like in equation (1) by replacing β_0 with a matrix that includes columns for each confounding variable, y_i signifies the observed phenotype for the i^{th} individual, β_0 represents the intercept across the population , $\beta_k$embodies the effect size of the k^{th} genetic marker, x_{ik} denotes the minor allele count at the k^{th} locus for individual i and ε_i is the residual error term, which is assumed to follow a normal distribution with a mean of zero and a variance σ_e^2. The goal within this framework is to identify the genetic markers k for which

the effect size β_k significantly deviates from zero, indicating an association with the phenotype in question.

However, the classic linear model proves insufficient for GWAS due to the inherent polygenic nature of complex traits, which entails a multitude of genetic markers exerting small, cumulative effects. Consequently, an error term η_k is incorporated to reflect these polygenic effects:

$$\eta_k = \sum_{s \neq k} \beta_s k_s + \varepsilon \ (3) \qquad \rightarrow \qquad y = \beta_0 + \beta_k x_k + \eta_k \ (4)$$

In this improved model, the error term η_k for each marker k aggregates the effects of all other markers s and the residual error ε, thereby encapsulating the collective influence of the polygenic structure.

In practice, the estimation of η_k assumes a level of independence and identical distribution (i.i.d) that may not hold due to linkage disequilibrium—correlations between genetic markers—and potential stratification within the population. To mitigate this, the kinship matrix K is estimated from genome-wide genotype data, allowing for the use of variance component techniques in LMMs to partition the phenotypic variance into genetic and environmental components effectively.

With dense genome-wide genotype data, the relatedness among individuals can be estimated without detailed genealogical information using K which is the genetic relatedness matrix (GRM). This allows for a variance decomposition approach like:

$$Var(y) = \sigma_g^2 K + \sigma_e^2 I \ (5) \ \rightarrow Var(\eta_k) = \sigma_{g-k}^2 K + \sigma_e^2 I \ (6)$$

Where σ_g^2 is the genetic variance, σ_e^2 is the environmental variance and σ_{g-k}^2 is the genetic variance without SNV k. To use equation (4) to model the effect of individual SNPs, the genetic variance σ_{g-k}^2 and environmental variance σ_e^2 must be estimated for each SNV which can be computationally demanding. However, under the assumption that the contribution of each SNV to the total phenotypic variance is negligible i.e. $\sigma_{g-k}^2 = \sigma_g^2$, we can simplify the model by estimating these variances only once, thereby

computing the GRM K from genetic data and using it to account for the random effects in the mixed model[112].

Efficient Mixed-Model Association eXpedited (EMMAX)[112] streamlines this process by first calculating GRM K from the genotype data, then estimating σ_g^2 and σ_e^2 from equation (5) using Efficient Mixed-Model Association (EMMA)[113]. Lastly, it uses a generalized linear model to test the hypothesis H_0: $\beta_k = 0$ for each SNV k individually.

To avoid proximal contamination, which arises when the variant under test is also factored into the calculation of the polygenic effect i.e. SNV k in $\sigma_{g-k}^2 = \sigma_g^2$, we can carry out the calculations using the leave-one-chromosome-out (LOCO) approach[110,114]. LOCO approach excludes the chromosome containing the test variant during the estimation of the genetic effect to mitigate proximal contamination. Consequently, for each phenotype assessed, 24 separate genetic effect estimations are produced, each corresponding to an analysis where a specific chromosome has been excluded. This ensures that the integrity of the association results is maintained as proximal contamination can lead to false negatives, where true associations are overlooked.

As a way to speed up the process, parallelization is often implemented to increase computational efficiency. However, the initial step of the LMM-based methods, which involves fitting the equation (7) below, presents a computational bottleneck. This step is crucial as it sets the foundation for subsequent analyses and must be executed with the entire dataset to maintain accuracy.

$$y \sim \beta_0 + \beta_2 z_{genotyping\ batch} + \beta_3 z_{sex} + \beta_4 z_{age} + \beta_5 z_{age^2} + \beta_6 z_{PC1} + \ \ldots \ + \beta_{25} z_{PC20} + g + \varepsilon \quad (7)$$

Once equation (7) is fitted, the subsequent step involves testing individual genetic variants across the genome for association with the phenotype. This step can be parallelized effectively across multiple processing units. For instance, the genetic data on a single chromosome can be divided into sections and distributed across different CPUs, enabling simultaneous computation. Each section tests the effect of variants $\varepsilon \sim \beta_1 x$ within its range where ε is the residual error term from equation (7), reducing the overall time required for this phase of GWAS. Despite this, the inherent complexities of step 1 mean that it remains a limiting factor in the overall speed of the LMM-based

GWAS analysis pipeline. Fortunately, a new machine-learning method called Regenie[115] was recently proposed that allows distributed computing in step 1 enabling large scale GWAS and PheWAS across many phenotypes simultaneously.

2.5.1 Regenie

Regenie is a software designed for large-scale GWAS and quantitative trait analysis. It is particularly optimized for the analysis of large datasets, such as those generated by biobank-scale cohorts, which can contain hundreds of thousands of individuals. The core strength of Regenie lies in its ability to efficiently handle such massive datasets while minimizing memory usage and computational time without sacrificing accuracy[115]. Regenie uses a stepwise fitting approach that consists of two main stages. Here's an overview of its underlying algorithm and how it works:

The first step of Regenie's algorithm , as illustrated in figure below, involves partitioning the total number of markers (M) into blocks containing B consecutive markers each. For each block, a ridge regression analysis is conducted using J shrinkage parameters, effectively selecting $J \times (M/B)$ markers across all blocks. Subsequently, these selected markers undergo another round of ridge regression, this time incorporating all $J \times (M/B)$ markers together. To optimize the selection of predictive markers, cross-validation (specifically, 5-fold CV) is employed, determining the optimal subset of markers to be utilized. This step is crucial for enhancing the predictive accuracy of the model. Additionally, the algorithm performs Leave-One-Chromosome-Out (LOCO) predictions for each phenotype across 23 chromosome sets (chromosomes 1-22 and X), excluding one chromosome at a time to account for proximal contamination in the phenotype predictions.

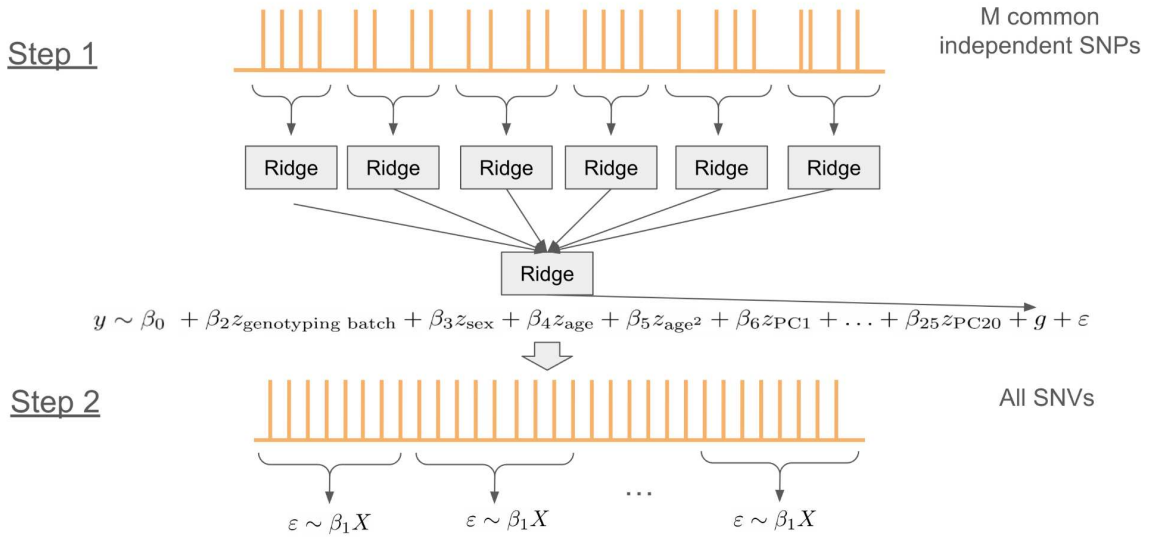


Fig.14

Schematic representation of the two-step Regenie algorithm used in genome-wide association studies.

In the second step, Regenie focuses on the association testing between phenotypes and genetic markers. Initially, covariates are regressed out of both the phenotypes and the genetic markers to control for potential confounding variables. The phenotypes are then adjusted by removing the LOCO predictions derived from the first step, ensuring that the true associations are not overlooked. Linear regression is applied to assess the association between the residualized phenotypes and each genetic marker, allowing for the precise evaluation of a variant's effect on the phenotype.

$$y - LOCO\,prediction = \beta_0 + \beta_2 z_{genotyping\,batch} + \ldots + \beta_5 z_{age^2} + \beta_6 z_{PC1} + \ldots + \beta_{25} z_{PC20} + R \quad (8)$$

$$R = \beta_1 x + \epsilon \quad (9)$$

What sets Regenie apart from other GWAS software tools, such as fastGWA or BOLT-LMM, is its incorporation of ridge regression to estimate the genetic component g. This approach is particularly effective for fitting the null model, which includes the polygenic background effect on a phenotype, thus allowing for a robust control of confounding factors like population stratification and relatedness. Unlike traditional LMM methods, which can become computationally infeasible with large datasets, Regenie is

optimized for efficiency, handling both step 1 and step 2 with scalability in mind. This two-step procedure dramatically reduces computation time without compromising accuracy, especially beneficial when analyzing biobank-scale datasets that contain hundreds of thousands of individuals. Regenie's application of ridge regression also aids in reducing the winner's curse—a common problem in GWAS where effect sizes are often overestimated due to the statistical noise of small sample sizes. As of the writing of this thesis, Regenie has been cited over 500 times and utilized by prominent biobanks, including FinnGen[80], for conducting GWAS.

2.5.1.1 Covariates

To construct the covariates file necessary for our GWAS, we assembled key demographic and technical variables likely to influence the genetic association results[116]. The covariate file included the following elements:

Sex: A binary indicator variable representing the biological sex of each participant, crucial for adjusting the analysis due to potential sex-specific genetic effects.

Age: This continuous variable reflects the age of participants at the time of the genetic data collection. Age is a significant factor in many genetic studies, as the expression of genetic traits can vary with age.

Age Squared: To capture the non-linear effects of age, we included the squared term of the age variable. This allows for the adjustment of the model for the curvature effect that age might have on the phenotype expression.

Genotyping Batch Number: Since the genotyping was conducted in five separate batches, each containing approximately 5,000 samples, a categorical variable indicating the batch number was included to adjust for potential batch effects. Batch effects can arise due to technical variability and can confound the association results if not properly accounted for.

Principal Components (PC) 1-20: The first 20 principal components were included to adjust for population stratification. These principal components are derived from the genetic data and represent major axes of genetic variation across the sampled individuals. Adjusting for these components helps to control for the confounding effects of ancestry, reducing the likelihood of spurious associations due to population structure.

By incorporating these covariates into our analysis, we aim to control for various confounding factors that could otherwise bias our results.

2.5.1.2 Step 1 of Regenie

To prepare our genotype data for the Regenie step 1 analysis, a thorough QC and preparation pipeline was implemented. The initial step involved lifting over our genotype data to the latest human reference genome build (GRCh38/hg38). We then selected individuals of European geneitc ancestry based on our ancestry inference pipeline results to avoid biases due to population stratification for our GWAS.

The QC pipeline was orchestrated to handle VCF files by chromosome, applying rigorous filters. The steps included merging regions of low complexity[117] with those exhibiting high linkage disequilibrium[118], to generate a comprehensive list of variants to be excluded from the analysis. In parallel, the hapmap resource was employed to identify a subset of markers for inclusion, utilizing bedtools to refine our selection.

The key parameters for variant filtering were set as follows: a minor allele frequency (MAF) threshold of 1%, a genotype missingness threshold per variant of 1%, and a strict Hardy-Weinberg equilibrium (HWE) p-value threshold of 10^{-15}. Additionally, an independent set of SNPs was determined using a window size of 1,000 kilobases, a step size of 100, and an r^2 threshold of 0.9 to ensure linkage equilibrium. Upon completing the filtering steps, we merged the chromosome-specific files into a single dataset. This dataset was converted into a PGEN file format, which is optimal for processing with Regenie. This rigorous QC procedure ensured that only high-quality, common, independent SNPs were carried forward into the step 1 of Regenie.

2.5.1.3 Step 2 of Regenie

For the second step of Regenie, our GWAS analysis employed the imputed variant data provided by the CLSA team. This dataset includes approximately 308 million genetic variants that were fed to Regenie without further quality control to facilitate the comprehensive exploration of genetic associations across a wide array of traits.

The imputation process, detailed by the CLSA team in their Genome-wide Genetic Data Release (version 3), began with 26,622 participants who had passed rigorous quality control measures, using 716,347 markers that met stringent criteria, including SNP-wise

missingness below 5% and a minor allele frequency greater than 0.0001. These markers were adjusted to ensure alignment with the human genome GRCh37 reference sequence using the bcftools[100] +fixref plugin, which ultimately pared down the markers to 653,729 for imputation.

The phasing and imputation were performed at the University of Michigan Imputation Service[119] using the TOPMed reference panel version r2, which includes 97,256 reference samples covering 308,107,085 genetic markers. Both autosomal and X chromosome variants were included, with the imputation executed in two separate batches of CLSA samples. The phased and imputed data from these batches were then merged into a unified dataset. In this context, Step 2 of Regenie in our project leveraged this extensively imputed dataset, allowing for a broad and inclusive analysis of potential genetic associations within one of Canada's largest genetic studies. This approach facilitates a deep exploration of the genetic architecture of numerous traits, enhancing our ability to identify genetic associations across a diverse range of phenotypes.

2.5.1.3.1 Chromosome X Imputation

We noticed that the initial imputation done by the CLSA team covered all chromosomes but omitted variants in the pseudoautosomal regions (PAR1 and PAR2) of chromosome X. To address this gap, we proceeded to re-impute chromosome X, using the TOPMed Imputation Server[43,55,120], which is based on the Minimac4 algorithm[121], with the reference panel TOPMed Freeze 2 on the hg38 build. We selected "all" for the population parameter to capture a broad genetic diversity and employed Eagle[122] for phasing, enhancing the accuracy of our imputation in light of chromosome X's unique challenges, including male hemizygosity[50]. This targeted re-imputation effort ensures our dataset now includes comprehensive genetic information for chromosome X, enriching our analysis with the inclusion of the significant PAR1 and PAR2 regions.

Due to the TOPMed Imputation Server's sample size limit of 25,000, we split our CLSA dataset into two batches of 13,313 individuals each for imputation. To consolidate the results, we used an algorithm designed to accurately merge the two output files. This algorithm calculated the combined allele frequency (p) by averaging the haploid dosage scores (HDS) across all samples for each variant. It then recomputed the imputation

quality score (Rsq) using the formula $Rsq = \frac{Var(HDS)}{p(1-p)}$ ensuring a precise measure of imputation accuracy for the merged dataset.

The final imputation output included:

- 375,907 variants in PAR1
- 15,331,044 non-PAR variants
- 39,630 variants in PAR2

Let's use the example below to illustrate how the algorithm works.

Suppose we have two batches of imputed data for a single variant on chromosome X, with the following HDS for a set of individuals with no overlap between the two batches (as was the case with our analysis):

Batch 1 (4 individuals): HDS = [0, 0.2, 0.8, 1]
Batch 2 (4 individuals): HDS = [0.1, 0.9, 1, 0]

First, we calculate the combined allele frequency (p) as the average HDS across both batches:

$$p = \frac{0 + 0.2 + 0.8 + 1 + 0.1 + 0.9 + 1 + 0}{8} = 0.5$$

To recompute Rsq we first need the variance of the HDS values. Given our HDS:
$HDS\ combined = [0, 0.2, 0.8, 1, 0.1, 0.9, 1, 0]$

The variance of these values can be calculated as:

$$Var(HDS) = \sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N}$$

Where x_i are the HDS values, μ is the mean (p=0.5) and N is the number of observations (8). Hence, $Var(HDS) = 0.1875$

Then, recomputing Rsq using the formula:

$$Rsq = \frac{Var(HDS)}{p(1-p)} = \frac{0.1875}{0.5(1-0.5)} = 0.75$$

Chapter 3: Results

3.1 GWAS

Our GWAS examined 350 binary phenotypes, resulting in 7,612 significant associations across 490 genetic loci. The genomic inflation factor (GC lambda) averaged 0.987025, with a range from 0.95 to 1.09, indicating well-calibrated test statistics overall. Among these findings, the top 10 hits span a variety of health-related phenotypes, highlighting key genetic variants linked to specific conditions:

1. Macular degeneration was most strongly associated with a variant at 1:196,697,663 A / C (rs1089033), near the *CFH* gene, showing a P-value of 1.1×10^{-23} and a minor allele frequency (MAF) of 0.38.
2. Type 2 diabetes showed a significant link with 10:113,022,822 G / GCT (rs10659211) near *TCF7L2*, with a P-value of 4.0×10^{-18} and a MAF of 0.30.
3. Condition of borderline diabetes or high blood sugar was associated with 10:112,998,590 C / T (rs7903146) near *TCF7L2*, showing a P-value of 4.5×10^{-18} and a MAF of 0.29.
4. Under-active thyroid gland had a notable variant at 9:97,776,188 A / G (rs7028661) near *FOXE1*, with a P-value of 6.4×10^{-18} and a MAF of 0.34.
5. Type 1 diabetes was linked to 6:32,658,661 A / T (rs9273367) near *HLA-DQB1*, with a P-value of 5.6×10^{-17} and a MAF of 0.28.
6. Another locus for macular degeneration was identified at 10:122,459,759 C / G (rs3793917) near *HTRA1*, with a P-value of 3.6×10^{-14} and a MAF of 0.21.
7. An additional variant for under-active thyroid gland was found at 1:113,834,946 A / G (rs2476601) near *PTPN22*, with a P-value of 3.7×10^{-13} and a MAF of 0.10.
8. High blood pressure or hypertension was associated with 4:80,243,569 C / T (rs1458038) near *FGF5*, with a P-value of 7.1×10^{-13} and a MAF of 0.29.

9. Non-melanoma skin cancer was linked to 6:396,321 C / T (rs12203592) near *IRF4*, with a P-value of 1.6×10^{-12} and a MAF of 0.19.
10. Another significant variant for under-active thyroid gland was at 6:32,669,089 A / T (rs3134996) near *HLA-DQB1*, with a P-value of 6.8×10^{-11} and a MAF of 0.35.

These results highlight the genetic underpinnings of various health conditions, demonstrating the power of GWAS in uncovering associations that could potentially guide future research, diagnosis, and treatment strategies. Below we showcase two examples of our GWAS findings for Type 2 Diabetes (T2D) and Macular Degeneration, which are consistent with well-known genetic associations for these conditions. The findings corroborate well-established associations in these conditions. For T2D, a significant association was detected at locus 10:113,022,822 with a variant in the *TCF7L2* gene, which is known for its substantial impact on insulin secretion and glucose homeostasis.

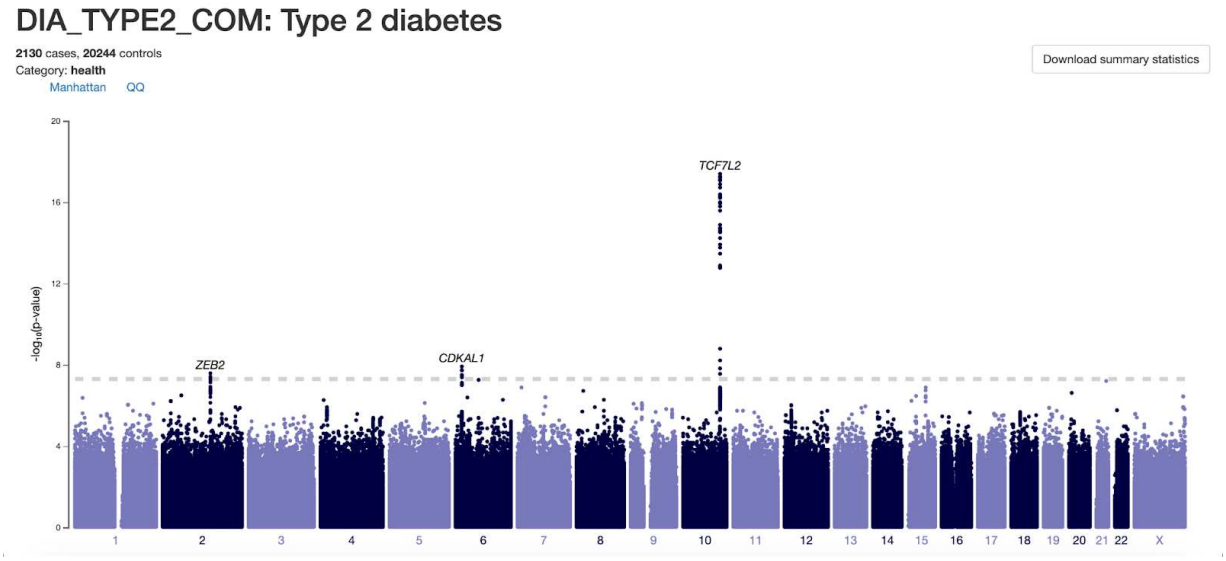


Fig.15

Manhattan plot representing genome-wide association results for Type 2 Diabetes (T2D).

Similarly, Macular Degeneration showed a strong association with the variant at 1:196,697,663 in the *CFH* gene, which plays a pivotal role in the immune response, with dysregulation linked to retinal damage. Additionally, the *HTRA1* gene, implicated in protein degradation, showed significant association at 10:122,459,759, confirming its

involvement in the disease's etiology. These results not only validate the robustness of our GWAS process but also align with well-known genetic influences on these conditions, underscoring the utility of our approach in identifying genetic factors for complex diseases.

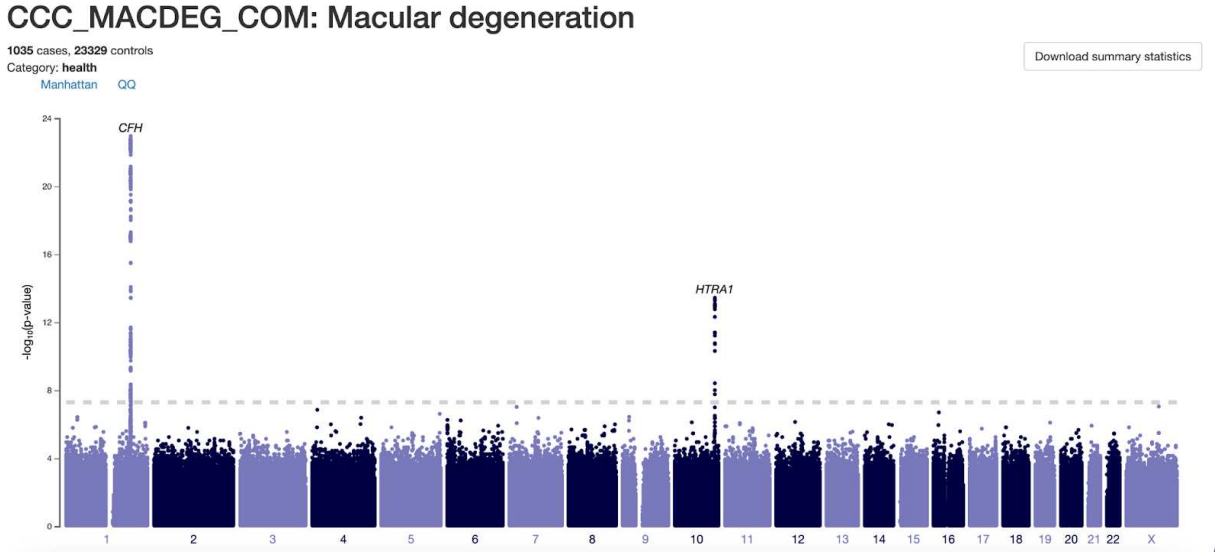


Fig.16

Manhattan plot representing genome-wide association results for Macular Degeneration.

Based on the analysis, the significant findings have been categorized to provide a clearer overview. The breakdown is as follows: medications yielded 4 statistically significant independent (defined as being at least 500 kb apart from each other) hits (p-value < 5×10^{-8}), behavior had 79 hits, health resulted in 163 hits, and socio-economic traits had the highest number with 180 hits. This categorization allows for a more structured interpretation of the data, highlighting areas with the most significant genetic associations. Please note that most of the socio-economic hits tend to be false positives due to the complex nature of socio-economic traits, which are influenced by a myriad of environmental factors and confounding variables, making it challenging to isolate true genetic associations. This comprehensive summary underscores the broad impact of genetic variations across different trait categories and emphasizes the importance of careful interpretation in socio-economic contexts.

To explore the potential of using PheWeb for comparing certain phenotypes across different cohorts, I have conducted a comparative analysis of two phenotypes: macular degeneration and type 2 diabetes. This comparison was performed across the CLSA PheWeb, FinnGen PheWeb, and UK Biobank PheWeb. Due to the differing definitions of phenotypes across studies (e.g., self-reported vs. diagnosed), direct comparisons can be challenging. However, the results of this comparative analysis are presented below, highlighting the Manhattan plots for each phenotype from the three different PheWebs. The Manhattan plots illustrating these findings for each phenotype from the three different PheWebs are shown below. These plots provide a visual representation of the genetic associations and highlight the significant variants identified in each cohort.

Macular Degeneration

The Manhattan plots and top loci for macular degeneration from the FinnGen, UK Biobank (UKBB), and CLSA PheWebs provide insight into the distribution and significance of genetic associations across the genome. Below is a detailed comparative analysis.

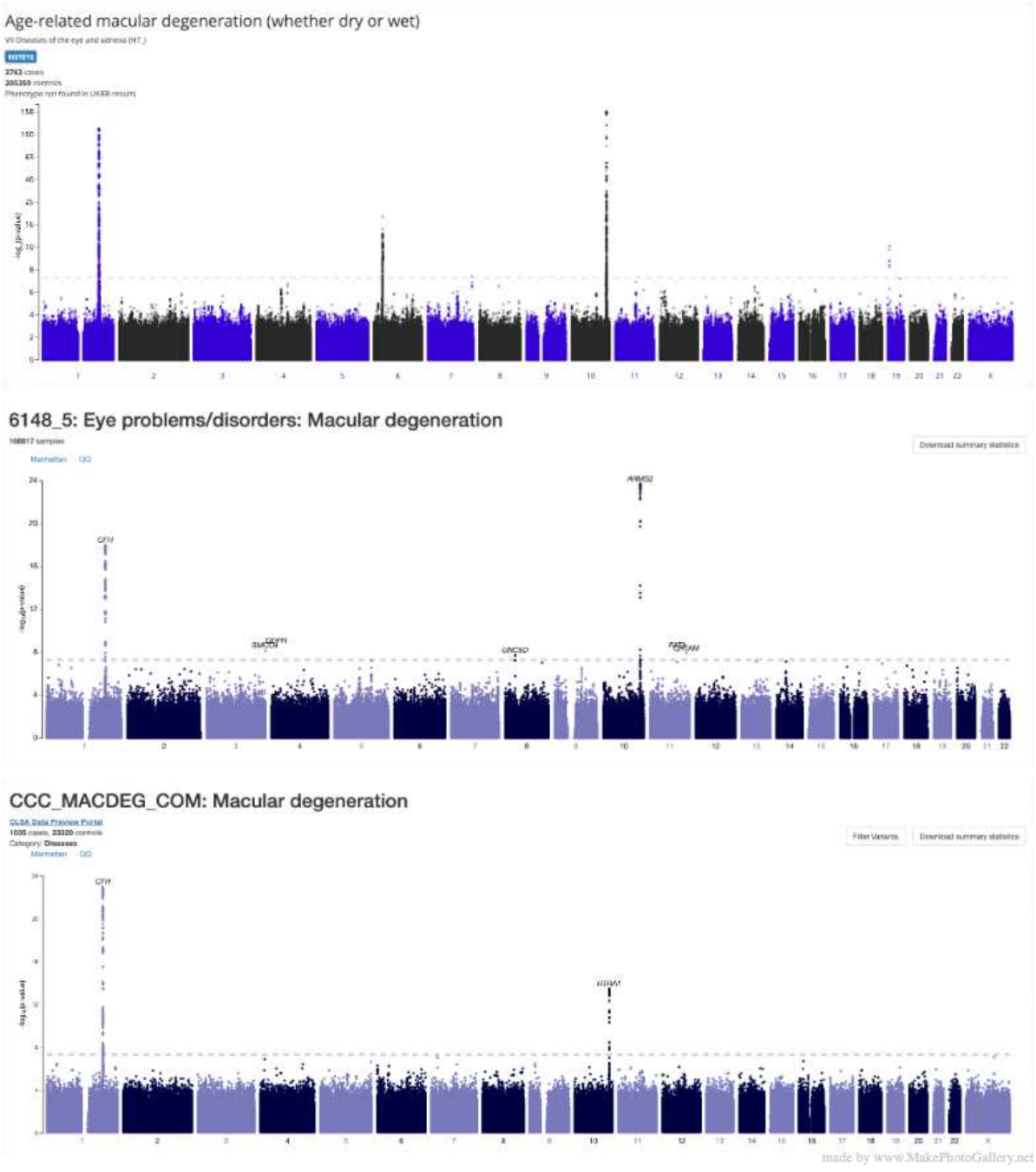


Fig.17
Manhattan plots for macular degeneration across different PheWebs. The plots display the genetic associations for macular degeneration from top to bottom: FinnGen, UK Biobank (UKBB), and CLSA. Each plot highlights the significant variants identified in each cohort, with notable associations near the *ARMS2* gene in FinnGen and UK Biobank, and near the *CFH* gene in CLSA.

PheWeb	Variant	Nearest Gene(s)	MAF	P-value	Effect Size (standard error)
FinnGen	10:122452080:C	*ARMS2*	0.243	3.90×10^{-160}	0.832
	1:196701709:A	*CFH*	0.563	1.80×10^{-113}	-0.577
	1:196662985:T	*CFH*	0.283	1.50×10^{-82}	-0.558
	19:6718376:G	*C3*	0.182	5.50×10^{-11}	0.214
	7:155171928:G	*NA*	0.000463	3.80×10^{-8}	4.04
	19:44908684:T	*APOE*	0.183	6.20×10^{-8}	-0.186
	11:72751271:C	*STARD10*	0.964	1.20×10^{-7}	-0.366
	4:109740713:T	*CFI*	0.0114	1.80×10^{-7}	0.681
UK Biobank	10:124,215,211 T / C (rs36212733)	*ARMS2*	0.470371	1.90×10^{-24}	-
	1:196,660,261 A / G (rs10801555)	*CFH*	0.133361	1.00×10^{-18}	-
	4:17,388,702 C / T (rs111235347)	*QDPR*	0.257596	2.60×10^{-9}	-
	11:91,963,521 G / A (rs191842278)	*FAT3*	0.615494	6.70×10^{-9}	-
	3:196,236,400 C / T (rs199641376)	*SMCO1*	0.294	7.00×10^{-9}	-
	11:122,705,597 C / T (rs559456070)	*CRTAM*	0.331776	1.20×10^{-8}	-
	8:34,400,330 C / T (rs545275191)	*UNC5D*	0.810604	2.00×10^{-8}	-
CLSA	1:196,697,663 A / C (rs1089033)	*CFH*	0.38	1.10×10^{-23}	-0.46 (0.045)
	10:122,459,759 C / G (rs3793917)	*HTRA1*	0.21	3.60×10^{-14}	0.40 (0.051)
	X:99,767,508 T / TA	*PCDH19*	0.00031	8.90×10^{-8}	2.6 (0.50)
	7:23,447,655 AC / A (rs1175748194)	*IGF2BP3*	0.0001	9.40×10^{-8}	5.5 (1.1)
	4:16,435,009 G / A (rs577382363)	*LDB2*	0.0014	1.40×10^{-7}	2.0 (0.33)
	16:17,794,764 G / T (rs1360781733)	*XYLT1*	0.0001	2.00×10^{-7}	5.3 (1.1)
	5:178,440,559 G / A (rs566052659)	*COL23A1*	0.00039	2.40×10^{-7}	3.3 (0.54)

Table 3.

This table summarizes the top genetic loci associated with macular degeneration identified in FinnGen, UK Biobank, and CLSA PheWebs, including the variant, nearest gene(s), minor allele frequency (MAF), p-value, and effect size (if available).

All three PheWebs identify the *CFH* and *ARMS2* loci as significant genetic associations with macular degeneration, demonstrating consistent findings across different populations.

FinnGen: The FinnGen PheWeb shows an extremely strong signal at the *ARMS2* locus with a P-value of 3.9×10^{-160} and several other significant loci, such as *CFH*, *C3*, and *APOE*, reflecting a comprehensive detection of genetic associations.

UK Biobank: The UKBB PheWeb identifies multiple independent loci, including *ARMS2, CFH, QDPR, FAT3, SMCO1, CRTAM,* and *UNC5D*, showing a broad range of significant hits across the genome.

CLSA: The CLSA PheWeb confirms the significance of *CFH* and *ARMS2* and identifies additional loci such as *HTRA1, PCDH19, IGF2BP3, LDB2, XYLT1*, and *COL23A1*, although with fewer significant hits compared to FinnGen and UKBB, likely due to differences in sample size and population structure.

Type 2 Diabetes

The Manhattan plots and top loci for type 2 diabetes from the FinnGen, UK Biobank (UKBB), and CLSA PheWebs reveal the distribution and significance of genetic associations across the genome. Below is a detailed comparative analysis.

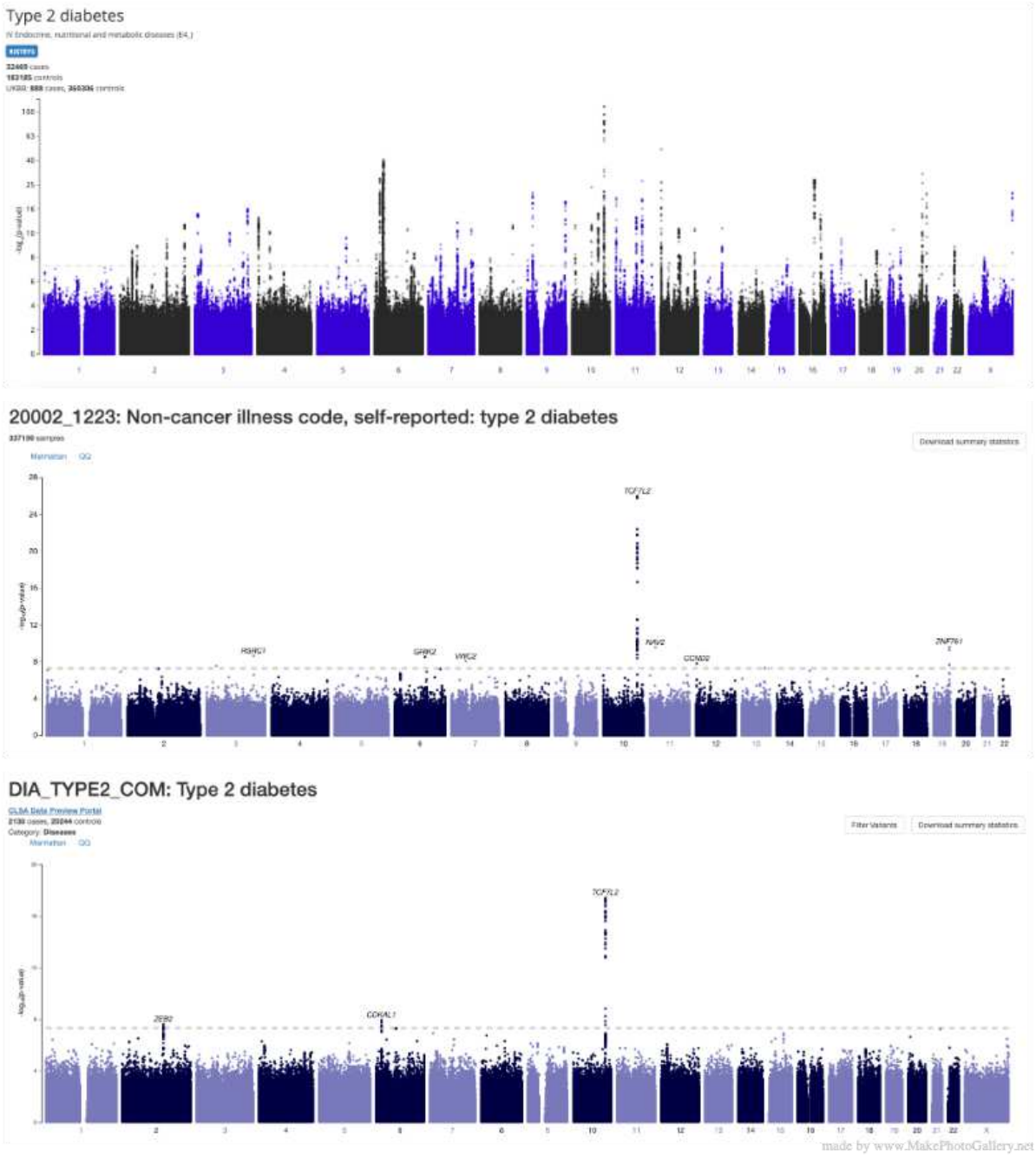


Fig. 18
Manhattan plots for type 2 diabetes across different PheWebs. The plots display the genetic associations for type 2 diabetes from top to bottom: FinnGen, UK Biobank (UKBB), and CLSA. Each plot highlights the significant variants identified in each cohort, with notable associations near the *TCF7L2* gene across all three cohorts.

PheWeb	Variant	Nearest Gene(s)	MAF	P-value	Effect Size (standard error)
UK Biobank	10:114,754,784 T / C (rs35198068)	*TCF7L2*	0.19625	1.00×10^{-26}	-
	19:53,948,824 T / G (rs2617746)	*ZNF761*	3.42×10^{-1}	2.40×10^{-10}	-
	11:19,900,706 A / G (rs182859724)	*NAV2*	2.06×10^{-1}	2.60×10^{-10}	-
	3:157,984,305 A / G (rs577584385)	*RSRC1*	1.07×10^{-1}	2.20×10^{-9}	-
	6:101,736,830 C / A (rs139090836)	*GRIK2*	1.03×10^{-1}	2.90×10^{-9}	-
	7:49,400,754 T / C (rs558253489)	*VWC2*	6.26×10^{-2}	8.50×10^{-9}	-
FinnGen	10:112994312:T	*TCF7L2*	0.202	1.20×10^{-112}	0.296
	12:4275678:T	*CCND2*	0.0312	2.90×10^{-50}	-0.484
	20:44189982:G	*JPH2*	0.012	7.20×10^{-32}	0.566
	6:20680447:T	*CDKAL1*	0.329	4.90×10^{-29}	0.123
	16:53784255:T	*FTO*	0.413	2.60×10^{-28}	0.116
	11:92975544:C	*MTNR1B*	0.357	8.90×10^{-28}	0.118
	10:69554950:G	*NA*	0.0432	1.00×10^{-24}	0.263
	9:22137686:T	*NA*	0.279	2.50×10^{-22}	0.113
	23:153634467:A	*NA*	0.275	2.70×10^{-22}	-0.092
	20:59032308:C	*ATP5E*	0.0499	5.50×10^{-22}	-0.234
CLSA	10:113,022,822 G / GCT (rs10659211)	*TCF7L2*	0.3	4.00×10^{-18}	0.31 (0.035)
	6:20,679,079 T / G (rs1569699)	*CDKAL1*	0.31	1.20×10^{-8}	0.20 (0.035)
	2:145,595,872 C / T (rs10175928)	*ZEB2*	0.4	2.60×10^{-8}	-0.19 (0.033)
	6:71,417,764 C / T (rs142310666)	*OGFRL1*	0.00089	5.50×10^{-8}	2.1 (0.35)
	21:40,108,282 GCTT / G (rs1182705129)	*DSCAM*	0.00027	6.20×10^{-8}	3.4 (0.69)
	15:72,260,074 C / A (rs191031793)	*PARP6*	0.0006	1.30×10^{-7}	2.4 (0.41)
	7:13,263,543 G / A (rs184183327)	*ARL4A*	0.00016	1.30×10^{-7}	4.4 (1.0)

Table 4.

This table summarizes the top genetic loci associated with Type 2 Diabetes identified in FinnGen, UK Biobank, and CLSA PheWebs, including the variant, nearest gene(s), minor allele frequency (MAF), p-value, and effect size (if available).

All three PheWebs identify the *TCF7L2* locus on chromosome 10 as a significant genetic association with type 2 diabetes, demonstrating a consistent finding across different populations.

FinnGen: The FinnGen PheWeb shows a very strong signal at *TCF7L2* with a P-value of 1.2×10^{-112} and several other significant loci, such as *CCND2, JPH2, CDKAL1, FTO,* and *MTNR1B*, reflecting a comprehensive detection of genetic associations.

UK Biobank: The UKBB PheWeb identifies several independent loci, including *TCF7L2, ZNF761, NAV2, RSRC1, GRIK2*, and *VWC2*, showing a broad range of significant hits across the genome.

CLSA: The CLSA PheWeb confirms the significance of *TCF7L2* and identifies additional loci such as *CDKAL1, ZEB2, OGFRL1, DSCAM, PARP6*, and *ARL4A*, although with fewer significant hits compared to FinnGen and UKBB, possibly due to differences in sample size and population structure.

These findings underscore the value of comparing PheWeb results across different cohorts to identify robust genetic associations and to understand the genetic architecture of complex traits. The Manhattan plots illustrating these findings for each phenotype from the three different PheWebs are shown above, providing a visual representation of the genetic associations and highlighting the significant variants identified in each cohort.

3.2 CLSA PheWeb

The CLSA PheWeb is accessible at https://clsa-pheweb.cerc-genomic-medicine.ca/.

Using the CLSA PheWeb, researchers can navigate the rich dataset through several intuitive features. The homepage's search functionality allows for the targeted inquiry of specific genes (e.g., *APOB*, *FTO*, *TCF7L2*), variants (identified by rsID or chromosomal positioning aligned with the designated genome build), or distinct phenotypes and traits. A comprehensive catalog of traits accessible within the PheWeb is detailed on the ‘Phenotypes’ page.

For a broader exploration, one may utilize the 'Random' page located within the top panel to generate a serendipitous selection from the PheWeb database. Conversely, selecting 'Top Hits' reveals a curated table of the most significant genetic associations discovered within the PheWeb framework. To acquire a deeper understanding of the dataset's foundation and how the data used was prepared, the 'About' page offers essential background information.

The PheWeb presents data in three principal visual formats: Manhattan and quantile-quantile (QQ) plots which depict the distribution of p-values across the genome, LocusZoom plots that offer a more granular view of specific genomic regions, and PheWAS plots that illustrate the association of a single genetic variant with a spectrum of traits. These visual tools facilitate a multifaceted analysis of genetic data, enhancing the interpretative process for researchers.

An example of the search utility is demonstrated by entering *FTO* into the search bar, which promptly navigates to the relevant genetic information within the PheWeb.

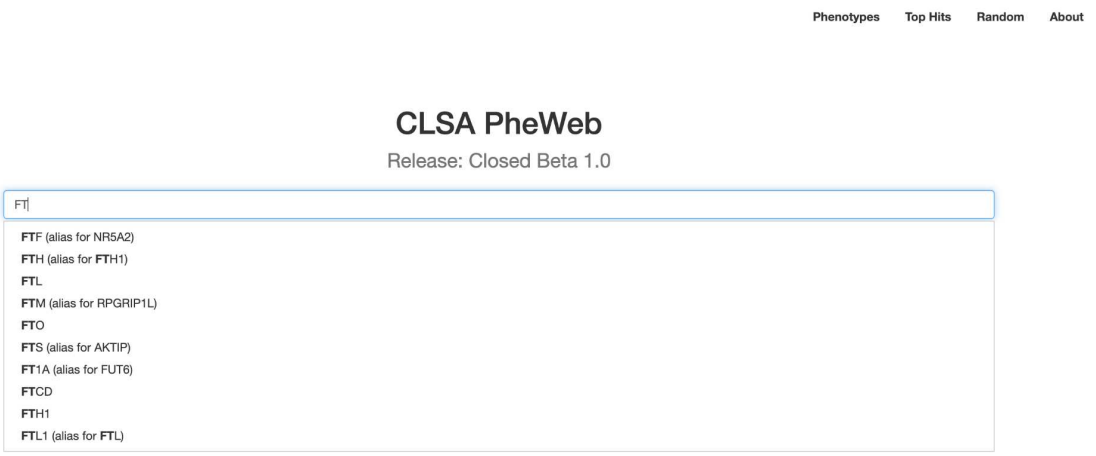


Fig.19

Homepage of the CLSA PheWeb displaying the search function. The interface shows an example search for the *FTO* gene.

In the CLSA PheWeb, initiating a search by gene name yields a table detailing the most significant genetic associations within that gene. Accompanying the table is a

LocusZoom regional plot which illustrates the linkage disequilibrium patterns among variants proximal to the gene of interest. This interactive feature allows users to visualize the genomic context and the extent of correlation between neighboring variants.

For instance, a query for *FTO* leads to a display where one can observe a LocusZoom plot specifically configured to reflect the selected association from the table, in this case, “High blood pressure or hypertension”. The plot dynamically updates to correspond with the association highlighted by the user, providing a detailed visual representation of the genetic landscape surrounding *FTO* and its relationship to the phenotype in question.

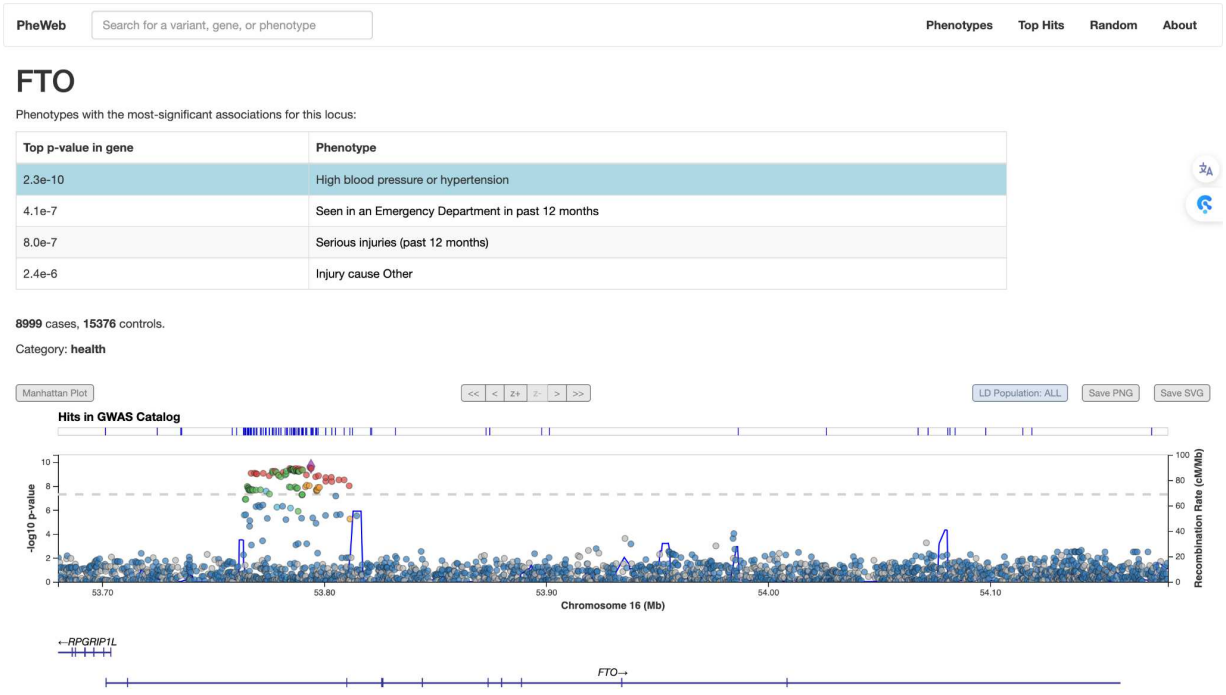


Fig.20

Detailed view of the PheWeb platform showcasing the search results for *FTO*. This page lists the phenotypes with the most significant associations for this locus, with “High blood pressure or hypertension” showing the top p-value.

All visualizations on the CLSA PheWeb are designed to be interactive, enhancing user engagement with the data. By positioning the cursor over variants within plots, such as the LocusZoom plot, users can access detailed information about each variant's properties and context.

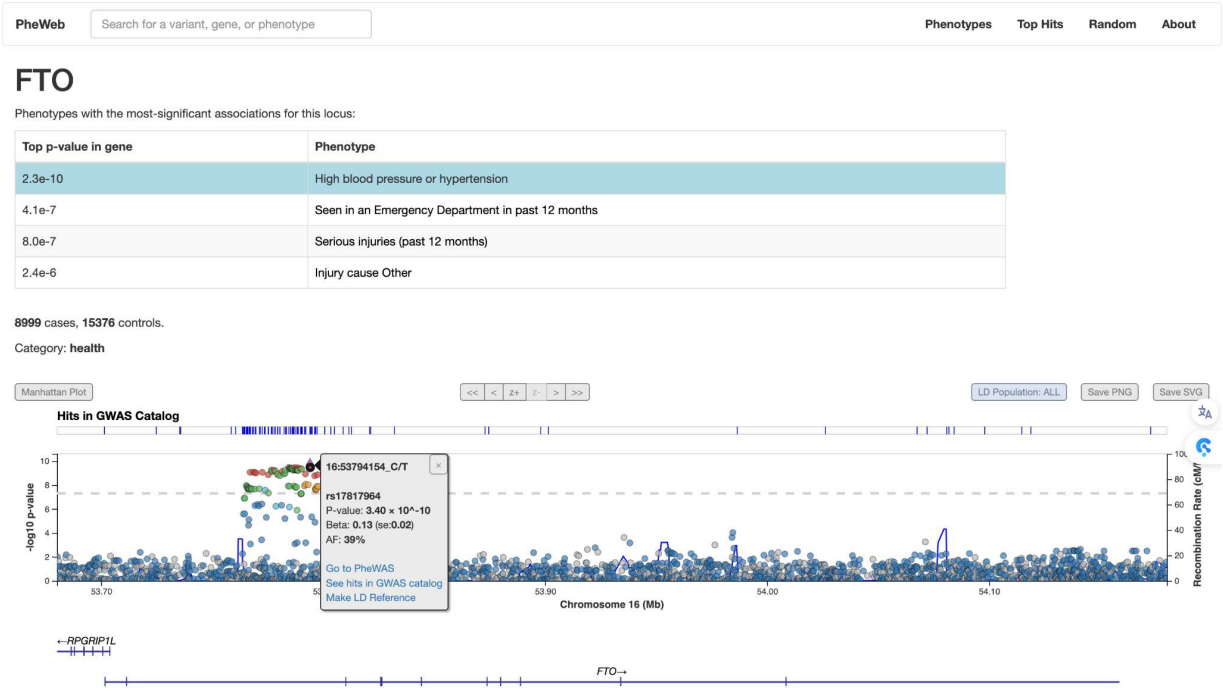
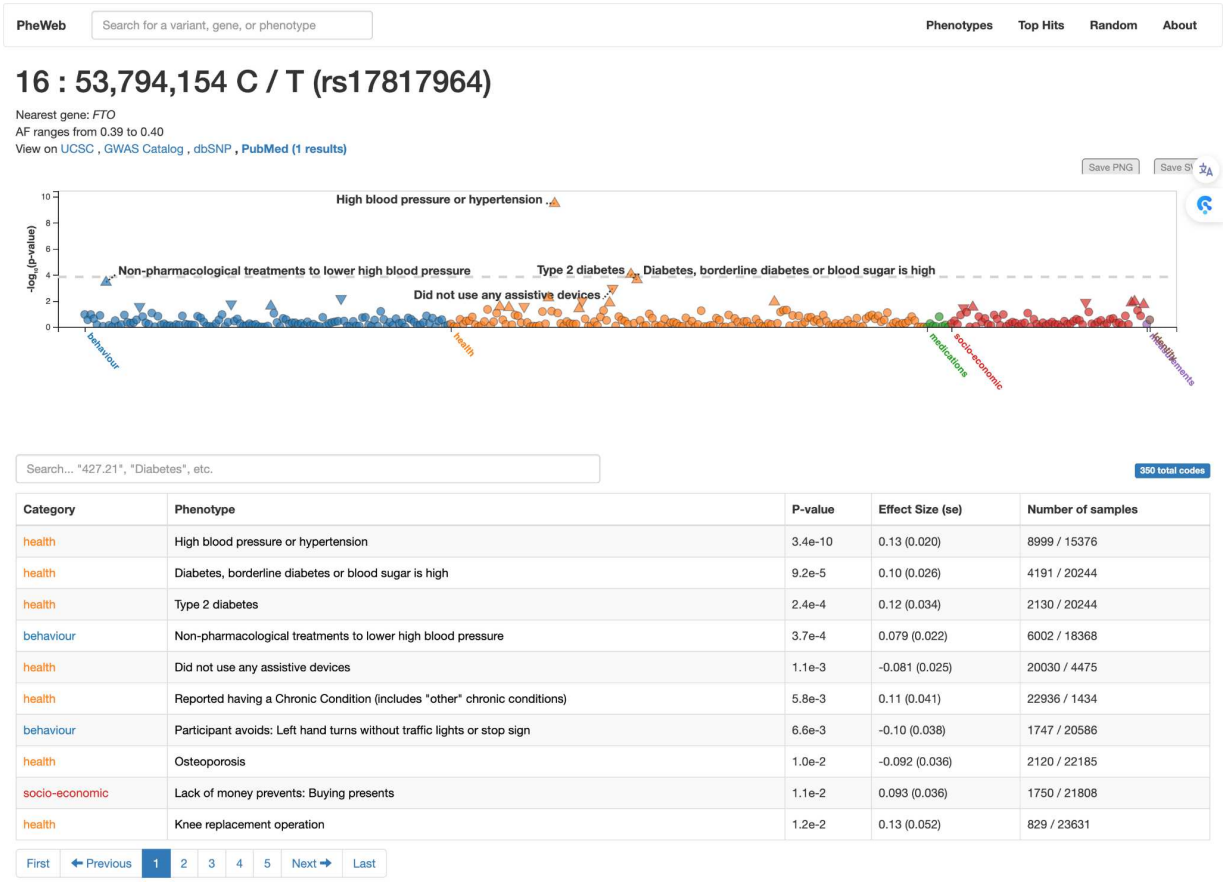


Fig.21

Interactive interface of CLSA PheWeb for the gene *FTO*, presenting a table of phenotypes with significant associations at this locus. Highlighted is “High blood pressure or hypertension” with the most significant p-value. Hovering over a variant in the LocusZoom plot below reveals detailed genetic information, and selecting a variant provides access to its PheWAS view, allowing for a comprehensive analysis of its impact across various phenotypes.

Selecting a variant from the LocusZoom plot in CLSA PheWeb will transition the user to a PheWAS visualization. This displays the variant's association p-values across the

breadth of phenotypes included in the PheWeb. Triangles pointing upwards indicate a variant's positive influence on the phenotype, while those pointing downwards indicate a negative influence. Circles represent variants with less precise beta estimates, such as those with a standard error that includes zero. The coloring of the triangles corresponds to specified categorizations (in the first first version of CLSA PheWeb, we used labels Identity, Socioeconomic, Behavioral, Health, Measurements, Medications, and Diet as outlined in the Phenotype data analysis section). For example, choosing the variant 16:53794154_C/T from the LocusZoom display brings up its PheWAS view, which is accompanied by a summary table detailing its associations.



Category	Phenotype	P-value	Effect Size (se)	Number of samples
health	High blood pressure or hypertension	3.4e-10	0.13 (0.020)	8999 / 15376
health	Diabetes, borderline diabetes or blood sugar is high	9.2e-5	0.10 (0.026)	4191 / 20244
health	Type 2 diabetes	2.4e-4	0.12 (0.034)	2130 / 20244
behaviour	Non-pharmacological treatments to lower high blood pressure	3.7e-4	0.079 (0.022)	6002 / 18368
health	Did not use any assistive devices	1.1e-3	-0.081 (0.025)	20030 / 4475
health	Reported having a Chronic Condition (includes "other" chronic conditions)	5.8e-3	0.11 (0.041)	22936 / 1434
behaviour	Participant avoids: Left hand turns without traffic lights or stop sign	6.6e-3	-0.10 (0.038)	1747 / 20586
health	Osteoporosis	1.0e-2	-0.092 (0.036)	2120 / 22185
socio-economic	Lack of money prevents: Buying presents	1.1e-2	0.093 (0.036)	1750 / 21808
health	Knee replacement operation	1.2e-2	0.13 (0.052)	829 / 23631

Fig.22

PheWAS plot for variant rs17817964 shows its association with several phenotypes. Upward triangles indicate a positive association between the variant and the phenotype, while downward triangles denote a negative association. The table below summarizes the statistical outcomes for each phenotype, including p-values and effect

sizes.

Choosing a trait from the PheWAS plot directs you to its corresponding Manhattan plot, displaying a comprehensive visual of significant genetic associations. Beneath this plot lies a table detailing these key associations, followed by a quantile-quantile (QQ) plot, which is stratified by minor allele frequency bins and features the genomic control lambda derived from a range of variant percentiles. For instance, after selecting the trait "High blood pressure or hypertension" from the PheWAS view above, a user can hover over any variant on the ensuing Manhattan plot to see a detailed LocusZoom regional plot for that variant. Further down the page, below the summary table, lies the QQ plot, providing additional statistical context for the associations displayed above.

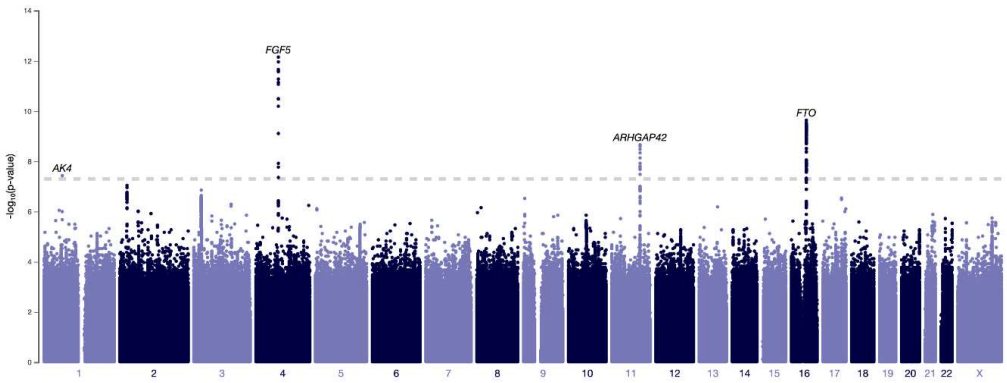
PheWeb | Search for a variant, gene, or phenotype | Phenotypes | Top Hits | Random | About

CCC_HBP_COM: High blood pressure or hypertension

8999 cases, **15376** controls
Category: **health**

Download summary statistics

Manhattan QQ



Top Loci:

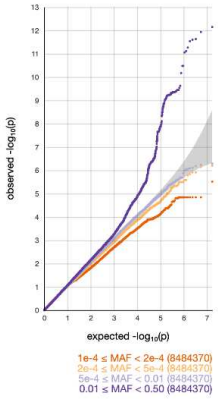
Search... "TCF7L2", "rs1861867", etc.

17 total variants

Variant	Nearest Gene(s)	MAF	P-value	Effect Size (se)
4:80,243,569 C / T (rs1458038)	*FGF5*	0.29	7.1e-13	0.16 (0.022)
16:53,794,050 A / G (rs62033408)	*FTO*	0.39	2.3e-10	0.13 (0.020)
11:100,701,679 A / G (rs670401)	*ARHGAP42*	0.29	2.2e-9	0.13 (0.022)
1:65,187,013 T / C (rs142466008)	*AK4*	0.016	3.7e-8	0.44 (0.079)
2:26,709,928 C / T (rs1275923)	*KCNK3*	0.40	9.1e-8	-0.11 (0.021)
3:27,464,145 G / C (rs11719386)	*SLC4A7*	0.27	1.4e-7	0.12 (0.022)
17:66,781,437 T / C (rs9912671)	*PRKCA*	0.47	2.9e-7	-0.10 (0.020)
9:5,684,411 G / A (rs78470093)	*RIC1*	0.032	3.0e-7	-0.30 (0.058)
3:132,597,795 T / G (rs572631764)	*ACAD11, ACKR4, NPHP3-ACAD11*	0.0011	5.1e-7	1.5 (0.32)
4:186,934,417 G / A (rs533245352)	*FAT1*	0.00088	5.7e-7	1.6 (0.34)

← Previous 1 2 Next →

QQ plot:



GC lambda 0.5: 1.097
GC lambda 0.1: 1.049
GC lambda 0.01: 1.005
GC lambda 0.001: 1.004
(Genomic Control lambda calculated based on the 50th percentile (median), 10th percentile, 1st percentile, and 1/10th of a percentile)

Fig.23

Manhattan and QQ plots for “High blood pressure or hypertension”. The top loci table lists significant associations and their statistics.

The CLSA PheWeb offers a comprehensive platform to explore genetic associations across various phenotypes using PheWAS. To demonstrate the utility and depth of the CLSA PheWeb, we showcase four examples of variants with strong associations across multiple phenotypes. To provide a clear visual representation of the genetic associations, LocusZoom plots have been included for the selected variants. These plots help illustrate the association signals within the genomic context and show the relationships between the primary variant and nearby variants in linkage disequilibrium. By comparing the PheWAS results with the LocusZoom plots, we can visually estimate the relationships between the genetic variants and multiple phenotypes, offering insights into potential pleiotropy and colocalization. These examples highlight the capability of PheWeb to uncover pleiotropic effects and provide insights into the genetic underpinnings of diverse traits:

rs2476601

The variant rs2476601, located in the *PTPN22* gene, shows significant associations with several health-related phenotypes. It is associated with an under-active thyroid gland, with a P-value of 3.7×10^{-13} and an effect size of -0.32, with a standard error of 0.043. This association supports the role of *PTPN22* in autoimmune thyroid disease. The variant is also linked to white blood cell count, with a P-value of 1.7×10^{-5} and an effect size of 0.065, with a standard error of 0.015, suggesting involvement in immune system regulation. Additionally, rs2476601 has a notable association with type 1 diabetes, with a P-value of 4.3×10^{-3} and an effect size of -0.51, with a standard error of 0.17, aligning with its known role in autoimmune disorders. The association with enlargement in the base of the thumbs has a P-value of 5.2×10^{-3} and an effect size of -0.15, with a standard error of 0.054, potentially indicating broader implications for joint health or inflammatory conditions.

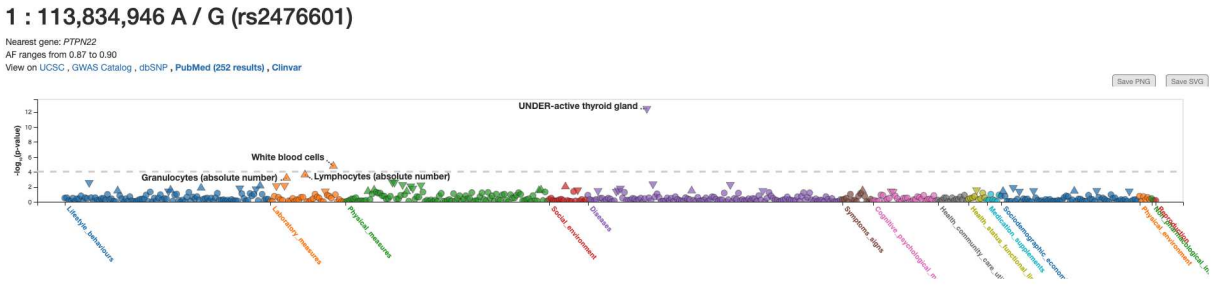


Fig.24

PheWAS plot for the variant rs2476601 taken from the CLSA PheWeb.

These associations highlight the variant's pleiotropic effects, impacting both autoimmune conditions and other health traits, suggesting a broad role for *PTPN22* in immune regulation and inflammatory processes.

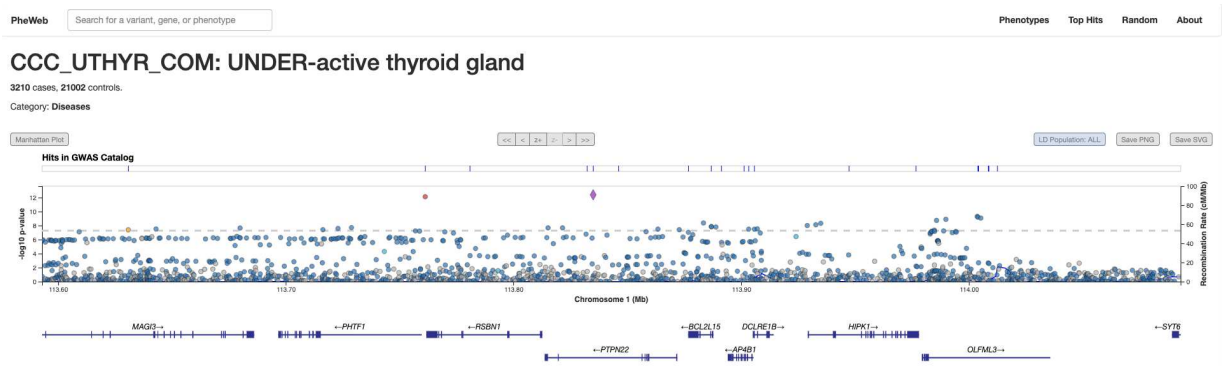


Fig.25

The LocusZoom plot highlights the association signal of rs2476601 with the Under-active thyroid gland, including nearby variants in LD.

rs10774625

The variant rs10774625, located in the *FTO* locus, is associated with several phenotypes beyond its well-known link to obesity. It shows a significant association with platelet count, with a P-value of 5.3×10^{-19} and an effect size of -0.078, with a standard error of 0.0087. This suggests a role in hematological traits. The variant is also linked to TNF-alpha levels, with a P-value of 1.8×10^{-17} and an effect size of -0.12, with a standard error of 0.014, indicating its influence on inflammatory processes. Additionally, rs10774625 is associated with lymphocyte count, with a P-value of 1.1×10^{-14} and an effect size of -0.070, with a standard error of 0.0090, further supporting its role in immune regulation. The association with smoking behavior, specifically having ever smoked 100 cigarettes, has a P-value of 2.2×10^{-6} and an effect size of -0.089, with a standard error of 0.019, pointing to its impact on lifestyle factors.

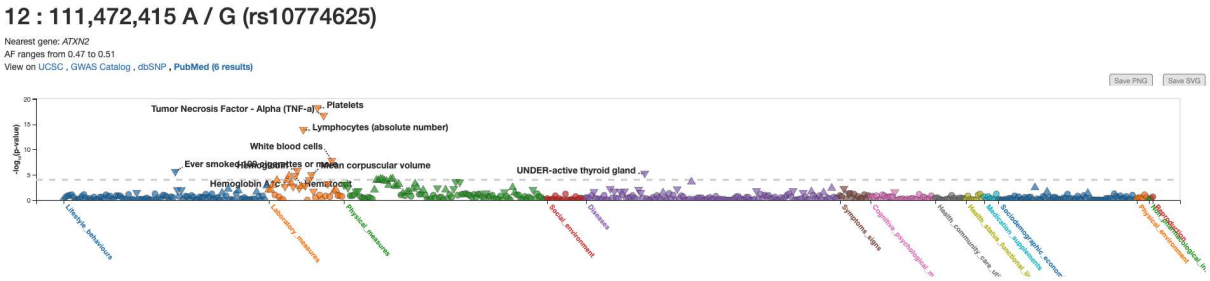


Fig.26

PheWAS plot for the variant rs10774625 taken from the CLSA PheWeb.

These pleiotropic associations highlight the variant's broader influence on immune-related and lifestyle traits, extending its significance beyond obesity to other aspects of health and behavior.

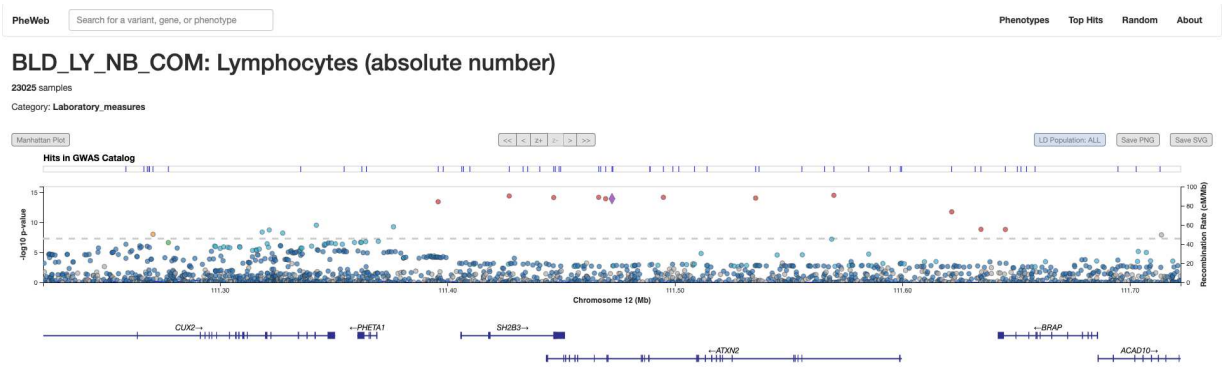


Fig.27

The LocusZoom plot highlights the association signal of rs10774625 with Lymphocytes, including nearby variants in LD.

rs10455872

The variant rs10455872, located in the *LPA* locus, shows significant associations with several phenotypes related to lipid metabolism and cardiovascular health. It is associated with LDL cholesterol, with a P-value of 4.1×10^{-11} and an effect size of 0.11, with a standard error of 0.017. This indicates a notable impact on LDL cholesterol

levels. The variant is also linked to total cholesterol, with a P-value of 1.1×10^{-9} and an effect size of 0.10, with a standard error of 0.016, further supporting its role in lipid metabolism. Additionally, rs10455872 is associated with coronary artery bypass surgery, with a P-value of 2.1×10^{-7} and an effect size of 0.37, with a standard error of 0.069, indicating a significant relationship with severe cardiovascular conditions. The association with heart disease has a P-value of 8.8×10^{-5} and an effect size of 0.22, with a standard error of 0.054, highlighting its broader impact on cardiovascular health.

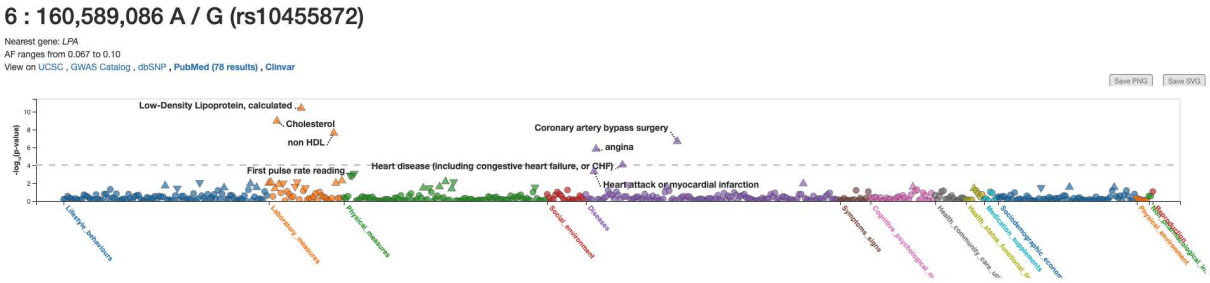


Fig.28

PheWAS plot for the variant rs10455872 taken from the CLSA PheWeb.

These pleiotropic effects underscore the variant's significant role in lipid metabolism and cardiovascular health, contributing to various related disorders.

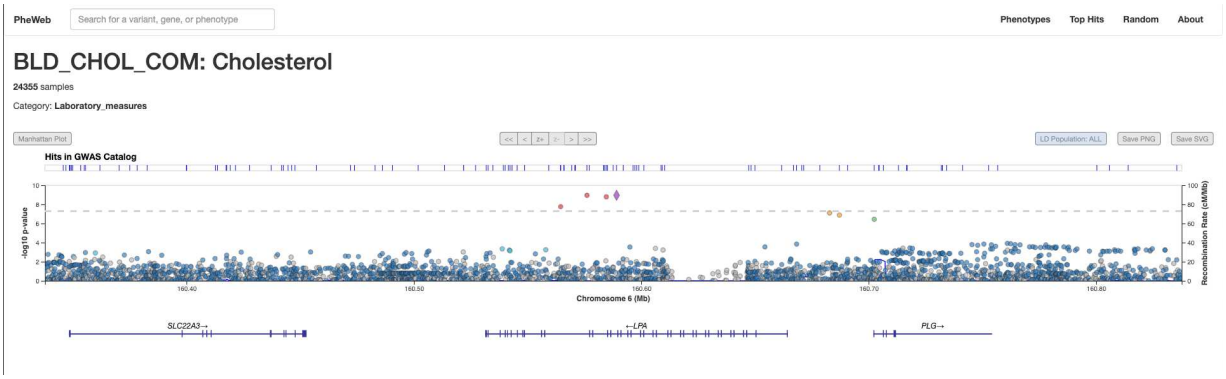


Fig.29

The LocusZoom plot highlights the association signal of rs10455872 with Cholesterol, including nearby variants in LD.

rs1421085

The variant rs1421085, located in the *FTO* locus, shows significant associations with several obesity-related traits and measures of body composition. It is strongly associated with body mass index (BMI), with a P-value of 1.2×10^{-22} and an effect size of 0.088, with a standard error of 0.0090. This variant is also linked to average weight, with a P-value of 9.2×10^{-19} and an effect size of 0.071, with a standard error of 0.0080, further supporting its role in body weight regulation. Additionally, rs1421085 is associated with waist circumference, with a P-value of 1.2×10^{-17} and an effect size of 0.070, with a standard error of 0.0082, highlighting its impact on fat distribution. The variant also shows a significant association with fat tissue in the android region, with a P-value of 1.7×10^{-15} and an effect size of 0.073, with a standard error of 0.0092.

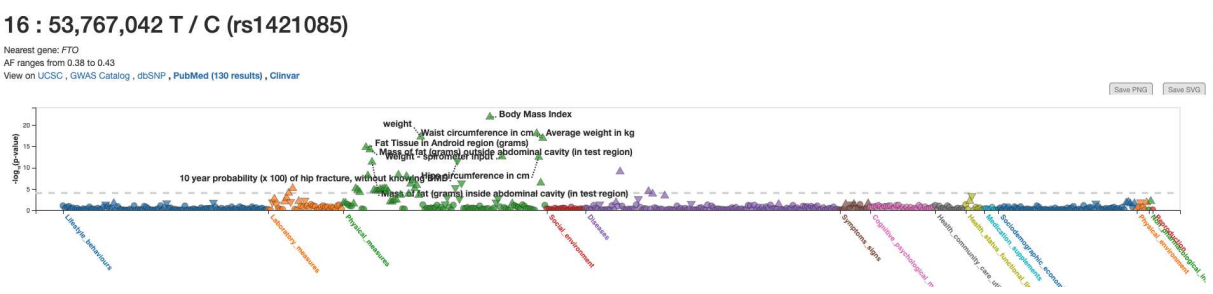


Fig.30

PheWAS plot for the variant rs1421085 taken from the CLSA PheWeb.

These pleiotropic effects demonstrate the variant's strong influence on obesity-related traits and overall body composition.

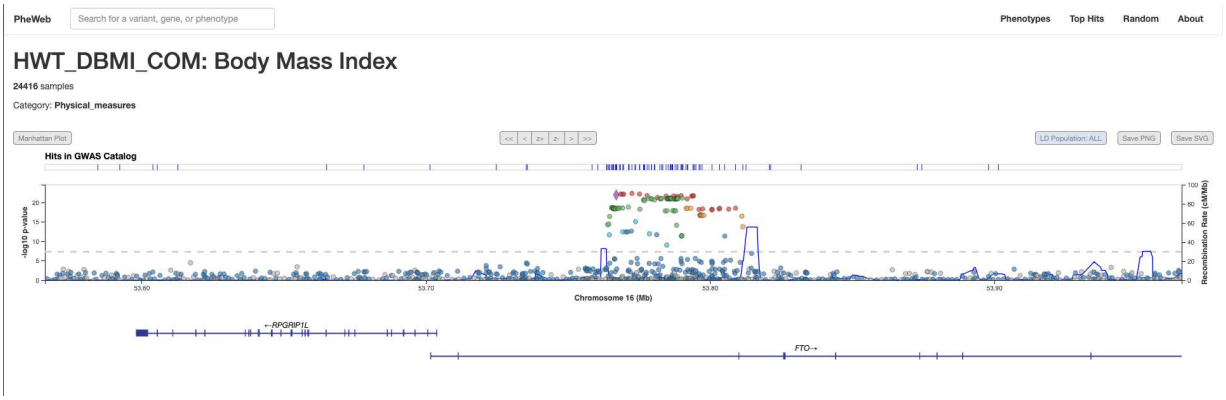


Fig.31

The LocusZoom plot highlights the association signal of rs1421085 with Body mass index , including nearby variants in LD.

These selected variants illustrate the power of the CLSA PheWeb to uncover significant genetic associations across a wide range of phenotypes. The detailed exploration of pleiotropy provided by PheWeb enhances our understanding of the multifaceted roles that specific genetic variants play in health and disease. This comprehensive approach allows researchers to generate new hypotheses and potentially identify novel therapeutic targets, fostering a deeper understanding of the genetic basis of complex traits.

Chapter 4: Discussion

The development and implementation of the CLSA PheWeb platform represented a significant advancement in the sharing and dissemination of GWAS and PheWAS results for the CLSA biobank. Initially, we conducted a thorough data curation process, which included extensive phenotypic and genotypic data from the CLSA biobank. This was followed by quality control procedures to filter out unreliable genetic markers and samples, ensuring the integrity of the dataset. Covariates such as age, age squared, genotyping batch number, sex, and principal components of ancestry were included in the analysis to account for potential confounding factors, and special attention was given to the imputation and analysis of chromosome X. The results from these analyses were then integrated into the PheWeb platform, allowing researchers worldwide to

access, visualize, and explore these findings in depth. This capability not only enhances the reproducibility of genetic research but also facilitates ongoing collaborative efforts, potentially accelerating the discovery of novel genetic insights and their applications in medical research.

The CLSA PheWeb, as a unique platform for the Canadian population, represents a significant contribution to the global landscape of genetic research tools. By offering detailed insights into the genetic determinants of health and disease in a Canadian context, it complements other PheWeb platforms such as those for UK Biobank, TCGA, FinnGen, BioBank Japan, CARTaGENE, COLCORONA, COLCOT, CHARM, SardiNIA, KoGES, and The Qatar Genome Program. While PheWeb is renowned for its comprehensive approach to displaying PheWAS and GWAS results, other tools such as Genebass, GWAS Catalog, PhenoScanner, and AstraZeneca PheWAS Portal, offer similar features that support genetic research. However, the specificity and depth provided by the CLSA PheWeb for the Canadian population underscore its importance and potential impact on both national and international genetic research efforts.

In terms of limitations, by concentrating solely on individuals of European like genetic ancestry, the study might not identify key variants that are particularly significant in other ethnic groups. In the construction of our PheWeb for the CLSA biobank, a pivotal decision was made to subset individuals of European ancestry for downstream analyses, a group constituting more than 90% of our dataset. This decision, while seemingly at odds with the imperative to embrace genetic diversity in genomic studies, was driven by specific analytical necessities and the demographic characteristics of our cohort. It's crucial to underscore that this approach does not undermine the importance of including and analyzing data from minority populations, a concern highlighted by recent discourse in the scientific community advocating for the inclusion of diverse genetic backgrounds to avoid biases and improve the applicability of genomic research across all populations[123].

The rationale behind focusing on the European subset stems from the aim to minimize confounding due to population stratification in GWAS. Population stratification can significantly impact the validity of GWAS findings by introducing spurious associations if not properly controlled. Given that the majority of our dataset comprises individuals of European descent, subsetting this group allows for a more similar genetic background, thereby reducing such confounding and enhancing the reliability of our association analyses. This step is particularly crucial for initial analyses aimed at identifying robust genetic associations that could be obscured by the complexity of analyzing a genetically heterogeneous population.

It's also important to recognize that the demographic makeup of our dataset, with a predominance of individuals of European ancestry, does not fully reflect the genetic diversity of the Canadian population. According to the National Household Survey of 2011, which coincides with the period when the CLSA recruited participants, approximately 20% of Canadians identified as visible minorities meaning persons, other than Aboriginal peoples, who are non-Caucasian in race or non-white in colour[124]. While our cohort's composition aligns to some extent with this demographic distribution, it nevertheless underscores the challenge of achieving representative genetic diversity in biomedical research.

Setting a minimum threshold of 1,000 cases for inclusion in the study helps ensure sufficient statistical power to detect associations. However, this criteria may exclude important insights from less common phenotypes where significant but rare genetic associations might exist, albeit with less detectable power in smaller sample sizes. This approach prioritizes robust statistical outcomes but at the potential cost of broader discovery across rarer conditions. Moreover, the use of solely an additive genetic model may not adequately represent the complex interplay of genetic factors that contribute to many traits and diseases. This simplification might lead to an incomplete understanding of the genetic architecture underlying various phenotypes. Finally, the absence of sex-specific analyses could miss critical differences in genetic associations between

males and females. Such differences could be key to developing more personalized and effective interventions based on sex-specific genetic insights.

In the context of the larger scientific knowledge, our project contributes significantly to the ongoing efforts in genetic epidemiology. The CLSA PheWeb platform not only enhances the transparency and reproducibility of genetic association studies but also provides a valuable resource for meta-analyses and cross-population comparisons. This is particularly relevant in the current era of precision medicine, where understanding the genetic basis of complex traits across diverse populations is paramount. The insights gained from our GWAS and PheWAS can inform public health strategies and healthcare policies, particularly those aimed at aging populations, as the CLSA cohort represents a crucial demographic in understanding age-related diseases and health conditions.

Future directions for this work involve addressing the limitations outlined above. Firstly, we will include individuals from diverse ethnic backgrounds in our analyses to uncover genetic variants that are significant across different populations. This will involve leveraging additional data from biobanks and other genetic studies that focus on underrepresented groups. By doing so, we aim to enhance the applicability and relevance of our findings to a broader range of populations.

Secondly, we plan to lower the threshold for phenotype inclusion to capture associations with less common traits. This will require the development of more sophisticated statistical models and the integration of larger datasets to maintain sufficient power. We will also explore non-additive genetic models to better capture the complexity of genetic influences on traits and diseases. This approach will involve using advanced machine learning techniques and interaction models to identify gene-gene and gene-environment interactions.

Lastly, we will conduct sex-specific analyses to identify genetic associations that differ between males and females. This will help us understand the sex-specific genetic architecture of various traits and diseases, leading to more personalized and effective

healthcare interventions. By integrating these future directions, we aim to provide a more comprehensive understanding of the genetic basis of complex traits and diseases, ultimately contributing to the advancement of precision medicine.

Chapter 5: Conclusions and Future Directions

In conclusion, the project adeptly performed 350 GWASs on binary traits from the CLSA dataset, leading to the inaugural CLSA PheWeb—an accessible platform for disseminating GWAS findings. This resource empowers the broader scientific community, paving the way for advancements in personalized medicine through applications such as Polygenic Risk Scores, Mendelian Randomization, and further replication studies. Looking ahead, recognizing the importance of genetic diversity, future efforts will include a broader range of genetic backgrounds beyond individuals of European descent. This expansion is essential for uncovering significant genetic variants relevant to diverse populations, thereby improving the generalizability and applicability of our research across different ethnic groups. In addition, we aim to broaden our analysis to encompass phenotypes with fewer than 1,000 cases, which were previously excluded due to power constraints. By employing more sensitive statistical methods or innovative data integration techniques, we can explore significant associations in rarer conditions that are less represented in large datasets. This approach will allow for a more comprehensive understanding of genetic influences across a wider array of traits. Further, to capture the complex interplay of genetic factors, future studies will integrate more sophisticated genetic models that consider gene-gene and gene-environment interactions. Such models will better reflect the complexity of genetic architecture, providing deeper insights into how different factors contribute to disease and health outcomes. Lastly, we plan to conduct sex-specific

analyses to uncover potentially crucial differences in genetic associations between males and females. Understanding these differences is vital for developing personalized medical interventions based on genetic insights.

Chapter 6: References

1. Uffelmann, E. *et al.* Genome-wide association studies. *Nat. Rev. Methods Primer* **1**, 59 (2021).
2. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
3. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**, D977–D985 (2023).
4. Denny, J. C., Bastarache, L. & Roden, D. M. Phenome-Wide Association Studies as a Tool to Advance Precision Medicine. *Annu. Rev. Genomics Hum. Genet.* **17**, 353–373 (2016).
5. Sivakumaran, S. *et al.* Abundant pleiotropy in human complex diseases and traits. *Am. J. Hum. Genet.* **89**, 607–618 (2011).
6. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
7. All of Us Research Program Investigators *et al.* The 'All of Us' Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).
8. Chen, Z. *et al.* China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int. J. Epidemiol.* **40**, 1652–1666 (2011).
9. Walters, R. G. *et al.* Genotyping and population characteristics of the China Kadoorie Biobank. *Cell Genomics* **3**, 100361 (2023).
10. Kim, Y. & Han, B.-G. Cohort Profile: The Korean Genome and Epidemiology Study (KoGES) Consortium. *Int. J. Epidemiol.* **46**, e20 (2017).
11. Kurki, M. I. *et al.* FinnGen: Unique genetic insights from combining isolated population and national health register data. 2022.03.03.22271360 Preprint at https://doi.org/10.1101/2022.03.03.22271360 (2022).
12. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *J.*

Epidemiol. **27**, S2–S8 (2017).

13. Akiyama, M. *et al.* Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. *Nat. Genet.* **49**, 1458–1467 (2017).

14. Denny, J. C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinforma. Oxf. Engl.* **26**, 1205–1210 (2010).

15. Pendergrass, S. a. *et al.* The use of phenome-wide association studies (PheWAS) for exploration of novel genotype-phenotype relationships and pleiotropy discovery. *Genet. Epidemiol.* **35**, 410–422 (2011).

16. Pendergrass, S. A. *et al.* Phenome-Wide Association Study (PheWAS) for Detection of Pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. *PLOS Genet.* **9**, e1003087 (2013).

17. Hebbring, S. J. *et al.* A PheWAS approach in studying HLA-DRB1*1501. *Genes Immun.* **14**, 187–191 (2013).

18. Gagliano Taliun, S. A. *et al.* Exploring and visualizing large-scale genetic associations by using PheWeb. *Nat. Genet.* **52**, 550–552 (2020).

19. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).

20. Awadalla, P. *et al.* Cohort profile of the CARTaGENE study: Quebec's population-based biobank for public health and personalized genomics. *Int. J. Epidemiol.* **42**, 1285–1299 (2013).

21. Dubé, M.-P. *et al.* Genetics of symptom remission in outpatients with COVID-19. *Sci. Rep.* **11**, 10847 (2021).

22. Dubé, M.-P. *et al.* Pharmacogenomics of the Efficacy and Safety of Colchicine in COLCOT. *Circ. Genomic Precis. Med.* **14**, e003183 (2021).

23. Dubé, M.-P. *et al.* Pharmacogenomic study of heart failure and candesartan response from the CHARM programme. *ESC Heart Fail.* **9**, 2997–3008 (2022).

24. Pilia, G. *et al.* Heritability of Cardiovascular and Personality Traits in 6,148 Sardinians. *PLOS Genet.* **2**, e132 (2006).

25. Al Thani, A. *et al.* Qatar Biobank Cohort Study: Study Design and First Results. *Am. J. Epidemiol.* **188**, 1420–1433 (2019).

26. Karczewski, K. J. *et al.* Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. *Cell Genomics* **2**, 100168 (2022).

27. Kamat, M. A. *et al.* PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. *Bioinforma. Oxf. Engl.* **35**, 4851–4853 (2019).

28. Wang, Q. *et al.* Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature* **597**, 527–532 (2021).

29. Forgetta, V. *et al.* Cohort profile: genomic data for 26 622 individuals from the Canadian Longitudinal Study on Aging (CLSA). *BMJ Open* **12**, e059021 (2022).

30. Ozaki, K. *et al.* Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nat. Genet.* **32**, 650–654 (2002).

31. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).

32. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9**, 356–369 (2008).

33. Skol, A. D., Scott, L. J., Abecasis, G. R. & Boehnke, M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat. Genet.* **38**, 209–213 (2006).

34. Purcell, S., Cherny, S. S. & Sham, P. C. Genetic Power Calculator: design of linkage andassociation genetic mapping studies of complex traits. *Bioinformatics* **19**, 149–150 (2003).

35. Hernán, M. A., Hernández-Díaz, S. & Robins, J. M. A Structural Approach to Selection

Bias. *Epidemiology* **15**, 615 (2004).

36. Holmes, M. V., Ala-Korpela, M. & Smith, G. D. Mendelian randomization in cardiometabolic disease: challenges in evaluating causality. *Nat. Rev. Cardiol.* **14**, 577–590 (2017).

37. Purcell, S., Sham, P. & Daly, M. J. Parental Phenotypes in Family-Based Association Analysis. *Am. J. Hum. Genet.* **76**, 249–259 (2005).

38. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

39. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).

40. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).

41. Kanai, M. *et al.* Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat. Genet.* **50**, 390–400 (2018).

42. Enabling the genomic revolution in Africa. *Science* **344**, 1346–1348 (2014).

43. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. 563866 Preprint at https://doi.org/10.1101/563866 (2019).

44. Schuster, S. C. Next-generation sequencing transforms today's biology. *Nat. Methods* **5**, 16–18 (2008).

45. van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends Genet. TIG* **30**, 418–426 (2014).

46. Lam, M. *et al.* RICOPILI: Rapid Imputation for COnsortias PIpeLIne. *Bioinformatics* **36**, 930–933 (2020).

47. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).

48. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in

unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).

49. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).

50. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).

51. Naj, A. C. Genotype Imputation in Genome-Wide Association Studies. *Curr. Protoc. Hum. Genet.* **102**, e84 (2019).

52. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

53. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).

54. Howie, B. N., Donnelly, P. & Marchini, J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLOS Genet.* **5**, e1000529 (2009).

55. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).

56. Browning, S. R. & Browning, B. L. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).

57. Pei, Y.-F., Li, J., Zhang, L., Papasian, C. J. & Deng, H.-W. Analyses and Comparison of Accuracy of Different Genotype Imputation Methods. *PLOS ONE* **3**, e3551 (2008).

58. Balding, D. J. A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.* **7**, 781–791 (2006).

59. Altshuler, D., Donnelly, P., & The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).

60. Garner, C. Upward bias in odds ratio estimates from genome-wide association studies.

Genet. Epidemiol. **31**, 288–295 (2007).

61. Sullivan, P. F. The Psychiatric GWAS Consortium: Big Science Comes to Psychiatry. *Neuron* **68**, 182–186 (2010).

62. GIANT consortium - Giant Consortium. https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium.

63. the Global Lipids Genetics Consortium. GLGC. http://lipidgenetics.org/.

64. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).

65. Baselmans, B. M. L. *et al.* Multivariate genome-wide analyses of the well-being spectrum. *Nat. Genet.* **51**, 445–451 (2019).

66. Rangamaran, V. R., Uppili, B., Gopal, D. & Ramalingam, K. EasyQC: Tool with Interactive User Interface for Efficient Next-Generation Sequencing Data Quality Control. *J. Comput. Biol.* **25**, 1301–1311 (2018).

67. Evangelou, E. & Ioannidis, J. P. A. Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet.* **14**, 379–389 (2013).

68. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322**, 881–888 (2008).

69. Ioannidis, J. P. A., Ntzani, E. E., Trikalinos, T. A. & Contopoulos-Ioannidis, D. G. Replication validity of genetic association studies. *Nat. Genet.* **29**, 306–309 (2001).

70. Kleinjan, D. A. & van Heyningen, V. Long-Range Control of Gene Expression: Emerging Mechanisms and Disruption in Disease. *Am. J. Hum. Genet.* **76**, 8–32 (2005).

71. Musunuru, K. *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**, 714–719 (2010).

72. Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.* **14**, 483–495 (2013).

73. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex

traits. *Nat. Genet.* **51**, 1339–1348 (2019).

74. Gratten, J. & Visscher, P. M. Genetic pleiotropy in complex traits and diseases: implications for genomic medicine. *Genome Med.* **8**, 78 (2016).

75. Pirmohamed, M. Pharmacogenomics: current status and future perspectives. *Nat. Rev. Genet.* **24**, 350–362 (2023).

76. Haas, J. T. *et al.* DGAT1 mutation is linked to a congenital diarrheal disorder. *J. Clin. Invest.* **122**, 4680–4684 (2012).

77. Pickrell, J. K. *et al.* Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* **48**, 709–717 (2016).

78. Diogo, D. *et al.* Phenome-wide association studies across large population cohorts support drug target validation. *Nat. Commun.* **9**, 4285 (2018).

79. Huang, J. Y. & Labrecque, J. A. From GWAS to PheWAS: the search for causality in big data. *Lancet Digit. Health* **1**, e101–e103 (2019).

80. Kurki, M. I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).

81. Pruim, R. J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010).

82. GWAS Catalog. https://www.ebi.ac.uk/gwas/.

83. UKBiobank TOPMed-imputed PheWeb. https://pheweb.org/UKB-TOPMed/.

84. UKBiobank ICD PheWeb. https://pheweb.org/UKB-SAIGE/.

85. UKBiobank PheWeb, based on the Neale lab's GWAS. https://pheweb.org/UKB-Neale/.

86. PheWeb of several heritable immune traits in the TCGA data. https://pheweb-tcga.qcri.org/.

87. FinnGen PheWeb. https://r5.finngen.fi/.

88. BioBank Japan PheWeb. https://pheweb.jp/.

89. CARTaGENE PheWeb. https://cerc-genomic-medicine.ca/pheweb/cartagene/.

90. COLCORONA PheWeb. https://pheweb.statgen.org/colcorona.

91. COLCOT PheWeb. https://pheweb.statgen.org/colcot.

92. CHARM PheWeb. https://pheweb.statgen.org/charm.

93. SardiNIA PheWeb. https://sardinia-pheweb.sph.umich.edu/.

94. KoGES PheWeb. https://koges.leelabsg.org/.

95. The Qatar Genome Program (QGP) PheWeb. https://pheweb-qgptraits.qcri.org/.

96. Genebass. *Genebass* https://genebass.org.

97. AstraZeneca PheWAS Portal. https://azphewas.com/.

98. McClelland, P. HGDP_1KG Ancestry Inference pipeline. CERC Genomic Medicine (2024).

99. Cavalli-Sforza, L. *The Human Genome Diversity Project*. https://www.osti.gov/biblio/505327 (1994).

100. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).

101. Eric, F. vcf2geno: Convert from ‘vcf’ to ‘geno’ format in LEA: LEA: an R package for Landscape and Ecological Association Studies.

102. Taliun, D. *et al.* LASER server: ancestry tracing with genotypes or sequence reads. *Bioinformatics* **33**, 2056–2058 (2017).

103. Wang, C., Zhan, X., Liang, L., Abecasis, G. R. & Lin, X. Improved Ancestry Estimation for both Genotyping and Sequencing Data using Projection Procrustes Analysis and Genotype Imputation. *Am. J. Hum. Genet.* **96**, 926–937 (2015).

104. Chen, S. *et al.* A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2024).

105. Johnson, J. L. jenlij/GAS-power-calculator. (2021).

106. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* **9**, e1003348 (2013).

107. Sham, P. C. & Purcell, S. M. Statistical power and significance testing in large-scale genetic studies. *Nat. Rev. Genet.* **15**, 335–346 (2014).

108. Wray, N. R. *et al.* Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14**, 507–515 (2013).

109. Spencer, C. C. A., Su, Z., Donnelly, P. & Marchini, J. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet.* **5**, e1000477 (2009).

110. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **46**, 100–106 (2014).

111. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).

112. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).

113. Kang, H. M. *et al.* Efficient Control of Population Structure in Model Organism Association Mapping. *Genetics* **178**, 1709–1723 (2008).

114. Listgarten, J. *et al.* Improved linear mixed models for genome-wide association studies. *Nat. Methods* **9**, 525–526 (2012).

115. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).

116. Dor, E. *et al.* Selecting Covariates for Genome-Wide Association Studies. 2023.02.07.527425 Preprint at https://doi.org/10.1101/2023.02.07.527425 (2023).

117. Morgulis, A., Gertz, E. M., Schäffer, A. A. & Agarwala, R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* **13**, 1028–1040 (2006).

118. Anderson, C. A. *et al.* Data quality control in genetic case-control association studies.

Nat. Protoc. **5**, 1564–1573 (2010).

119. Michigan Imputation Server. https://imputationserver.sph.umich.edu/index.html#!

120. TOPMed Imputation Server. https://imputation.biodatacatalyst.nhlbi.nih.gov/#!

121. Fuchsberger, C., Abecasis, G. R. & Hinds, D. A. minimac2: faster genotype imputation. *Bioinformatics* **31**, 782–784 (2015).

122. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).

123. Ben-Eghan, C. *et al.* Don't ignore genetic data from minority populations. *Nature* **585**, 184–186 (2020).

124. Government of Canada, S. C. Census Datasets. https://www.recensement2011.gc.ca/datasets/Index-eng.cfm?Temporal=2013 (2015).