Technical University of Munich
TUM School of Medicine and Health

Understanding multimorbidity with big data in genomics

Ana Luiza de Santana Villasboas Arruda

Complete reprint of the dissertation approved by the TUM School of Medicine and Health of the Technical University of Munich for the award of the

Doktorin der Naturwissenschaften (Dr. rer. nat.)

Chair: Prof. Dr. Radu Roland Rad

Examiners:

1. Prof. Dr. Eleftheria Zeggini
2. Prof. Dr. Julien Gagneur

The dissertation was submitted to the Technical University of Munich on 15 March 2024 and accepted by the TUM School of Medicine and Health on 3 July 2024.

Table of Contents

Abstract

This doctoral thesis explores the underlying genetics of multimorbidity using large genomic datasets. Multimorbidity is defined as the co-occurrence of multiple chronic health conditions in one individual and represents a rising public health challenge. Global trends such as increasing average body mass index (BMI) and life expectancy contribute to the escalating prevalence of multimorbidity cases. Beyond the personal burden, including polypharmacy and adverse side effects, multimorbidity poses a challenge to society as a whole by increasing treatment demands that subsequently inflate healthcare expenses. Yet, most health-related and drug development research is focused on treating and/or preventing individual diseases. Consequently, healthcare services are not optimally designed to assist patients suffering from two or more health conditions.

A very prevalent multimorbidity pattern among women and men is a combination of cardiometabolic and osteoarticular diseases. In older adults, a common example is the type 2 diabetes-osteoarthritis comorbidity. Type 2 diabetes affects more than 536 million people worldwide and is characterized by elevated blood glucose levels and insulin resistance. Osteoarthritis is the most common whole-joint chronic disorder, affecting over 520 million people worldwide. A further very common pair of diseases that coexist in adults is a combination of metabolic and severe mental diseases, exemplified here by type 2 diabetes and schizophrenia. Schizophrenia is a major psychiatric disorder with a global prevalence of 1% and typically characterized by problems with perception, cognitive function, and behavior.

Using summary statistics of large-scale genome-wide association studies (GWAS), the genetic intersection between both pairs of comorbidities was investigated separately. Firstly, genome-wide analyses were performed to unveil the genetic correlation and causality via Mendelian randomization between each pair of conditions. Secondly, regional genetic colocalization analysis was conducted to find shared risk signals. By scoring all genes in the vicinity of the identified shared association signals, a list of putative effector genes simultaneously influencing both diseases were derived. Finally, a deeper dive into the genetic insights revealed by the top-scoring putative effector genes was conducted, including pathway analyses and exploration of druggability.

For type 2 diabetes and osteoarthritis, the well-established positive association was corroborated, which was stronger for knee osteoarthritis than hip osteoarthritis. Subsequently, 19 high-confidence effector genes were identified. In addition to the well-established involvement of obesity in this comorbidity, other possible involved biological mechanisms were highlighted including the Wnt/β-catenin signaling pathway and imbalances in bone marrow cell differentiation. Despite the positive epidemiological correlation between type 2 diabetes and schizophrenia, evidence for a negative genetic correlation was found, in addition to no evidence of a causal relationship.

Intriguingly, a protective effect of schizophrenia liability on increased BMI was found, challenging conventional understanding of adiposity-related mechanisms underlying each condition. Highlighted biological pathways possibly underlying this comorbidity pair include cholesterol trafficking and adipogenesis.

The research presented in this doctoral thesis holds promise for advancing personalized medicine through the identification of novel therapeutic targets and opportunities for drug repurposing, contingent upon rigorous validation. The generalizability of findings is partially hampered due to disparities in data availability, particularly from diverse populations.

Zusammenfassung

In dieser Doktorarbeit werden die genetischen Grundlagen der Multimorbidität anhand großer genomischer Datensätze untersucht. Multimorbidität ist definiert als das gleichzeitige Auftreten mehrerer chronischer Gesundheitszustände bei einer Person und stellt eine zunehmende Herausforderung für die öffentliche Gesundheit dar. Globale Trends wie der steigende durchschnittliche (BMI) und die höhere Lebenserwartung tragen zur zunehmenden Prävalenz von Multimorbidität bei. Neben der persönlichen Belastung durch Polypharmazie und unerwünschte Nebenwirkungen stellt die Multimorbidität eine Herausforderung für die Gesellschaft als Ganzes dar, da sie den Behandlungsbedarf erhöht und damit die Kosten im Gesundheitswesen in die Höhe treibt. Der Großteil der gesundheitsbezogenen Forschung und der Arzneimittelentwicklung ist jedoch auf die Behandlung und/oder Vorbeugung einzelner Krankheiten ausgerichtet.

Ein sehr häufig anzutreffendes Multimorbiditätsmuster ist eine Kombination aus kardiometabolischen und osteoartikulären Erkrankungen. Bei älteren Erwachsenen ist die Komorbidität von Typ-2-Diabetes und Osteoarthritis ein häufiges Beispiel. Typ-2-Diabetes betrifft weltweit mehr als 536 Millionen Menschen und ist durch einen erhöhten Blutzuckerspiegel und Insulinresistenz gekennzeichnet. Osteoarthritis ist die häufigste chronische Erkrankung der gesamten Gelenke, von der weltweit über 520 Millionen Menschen betroffen sind. Ein weiteres sehr häufiges Krankheitspaar, das bei Erwachsenen nebeneinander auftritt, ist die Kombination von Stoffwechsel- und schweren psychischen Erkrankungen, hier am Beispiel von Typ-2-Diabetes und Schizophrenie. Schizophrenie ist eine schwere psychiatrische Störung mit einer weltweiten Prävalenz von 1%, die typischerweise durch Probleme mit der Wahrnehmung, der kognitiven Funktion und dem Verhalten gekennzeichnet ist.

Anhand von zusammenfassenden Statistiken groß angelegter genomweiter Assoziationsstudien (GWAS) wurde die genetische Schnittmenge zwischen beiden Komorbiditätspaaren getrennt untersucht. Erstens wurden genomweite Analysen durchgeführt, um die genetische Korrelation und Kausalität zwischen jedem Paar von Erkrankungen aufzudecken. Zweitens wurde eine regionale

genetische Kolokalisationsanalyse durchgeführt, um gemeinsame Risikosignale zu finden. Durch die Auswertung aller Gene in der Nähe der identifizierten gemeinsamen Assoziationssignale wurde eine Liste von mutmaßlichen Effektorgenen erstellt, die beide Krankheiten gleichzeitig beeinflussen. Schließlich wurden die genetischen Erkenntnisse, die sich aus den am besten bewerteten mutmaßlichen Effektorgenen ergaben, vertieft, u. a. durch eine Analyse der Signalwege und die Erforschung der Medikamentenverfügbarkeit.

Für Typ-2-Diabetes und Osteoarthritis bestätigte sich der bekannte positive Zusammenhang, der bei Kniearthrose stärker ausgeprägt war als bei Hüftarthrose. In der Folge wurden 19 Effektorgene mit hoher Wahrscheinlichkeit identifiziert. Neben der bekannten Beteiligung von Fettleibigkeit an dieser Komorbidität wurden weitere mögliche biologische Mechanismen hervorgehoben, darunter der Wnt/β-Catenin-Signalweg und Ungleichgewichte bei der Differenzierung von Knochenmarkzellen. Trotz der positiven epidemiologischen Korrelation zwischen Typ-2-Diabetes und Schizophrenie wurden Beweise für eine negative genetische Korrelation gefunden, aber auch keine Hinweise auf einen kausalen Zusammenhang. Interessanterweise wurde eine schützende Wirkung der Schizophrenie-Haftung auf einen erhöhten BMI festgestellt, was das herkömmliche Verständnis der adipositasbezogenen Mechanismen, die beiden Erkrankungen zugrunde liegen, in Frage stellt. Zu den hervorgehobenen biologischen Wegen, die möglicherweise diesem Komorbiditätspaar zugrunde liegen, gehören der Cholesterinverkehr und die Adipogenese.

Die in dieser Dissertation vorgestellten Forschungsergebnisse versprechen Fortschritte in der personalisierten Medizin durch die Identifizierung neuartiger therapeutischer Ziele und Möglichkeiten für die Umwidmung von Arzneimitteln, vorausgesetzt, die Ergebnisse werden rigoros validiert. Die Verallgemeinerbarkeit der Ergebnisse wird teilweise durch die unterschiedliche Verfügbarkeit von Daten, insbesondere aus verschiedenen Bevölkerungsgruppen, beeinträchtigt.

List of publications

Arruda, A.L., Hartley, A., Katsoula, G. et al. Genetic underpinning of the comorbidity between type 2 diabetes and osteoarthritis. American Journal of Human Genetics (2023). https://doi.org/10.1016/j.ajhg.2023.06.010

Arruda, A.L., Khandaker, G.M., Morris, A.P. et al. Genomic insights into the comorbidity between type 2 diabetes and schizophrenia. Schizophrenia 10, 22 (2024). https://doi.org/10.1038/s41537-024-00445-5

Arruda, A.L., Morris, A.P. and Zeggini, E. Advancing equity in human genomics through tissue-specific multi-ancestry molecular data. Cell Genomics (2024). https://doi.org/10.1016/j.xgen.2023.100485

Suzuki, K., Hatzikotoulas, K., Southam, L. et al. Genetic drivers of heterogeneity in type 2 diabetes pathophysiology. Nature (2024). https://doi.org/10.1038/s41586-024-07019-6

Theoretical background

Brief introduction to human statistical genetics

Statistical genomics is an interdisciplinary field that blends statistical techniques with genomic technologies to interpret large-scale datasets. This field is dedicated to analyzing complex genomic data from the entirety of an organism's genetic information, including the study of gene expression, gene regulation, and the interaction between different genetic loci and environmental factors. This data is generated by modern genomic technologies such as high-throughput sequencing, microarrays, and other forms of genomic profiling[1]. The goal of human statistical genomics is to understand the genomic architecture of complex traits, identify genetic determinants of diseases, and uncover the underlying mechanisms of gene expression and regulation[2]. Statistical genomics is the method of choice to pave the way for personalized medicine, bridging the gap between raw genetic data and meaningful biological insights.

The costs of researching and developing new therapeutics, estimated in the billions of dollars, are primarily fueled by a significant number of clinical trial failures[3]. Statistical genomics can accelerate drug target discovery, making it not only faster but also more cost-effective by prioritizing drug candidates in the initial phases of drug discovery. Two-thirds of the new drugs approved by the FDA in 2021 were supported by human genetic evidence, indicating that drug targets with human genetic support are more likely to succeed in clinical trials[4]. Translation discoveries can not only enhance the efficiency of drug discovery but also hold the promise of developing more targeted and effective therapeutic interventions, ultimately shaping the landscape of precision medicine.

The creation and development of the field of statistical genomics were made possible by the synergetic development of two key elements: the exponential growth in computational power and the emergence of large-scale genomic datasets, facilitated by cutting-edge genomic technologies and methodological tools. The field of genomics witnessed a revolution with the advent of high-throughput sequencing and other genomic profiling technologies. These technologies have made it possible to generate large-scale genomic datasets, providing a comprehensive view of the genome at a resolution that was not possible before[5]. The availability and the ever-diminishing costs of generating extensive genomic datasets has been a game-changer for statistical genomics. A significant milestone was the finishing process of the Human Genome Project in 2003, which published a 99% coverage of the euchromatic human genome[6]. The complete coverage of all genomic regions was completed in 2022[7]. To turn the generated vast amount of raw genomic sequences into meaningful biological information, the development of statistical methods for genetic data analysis has been crucial.

In the early days of genetic research, the limited computational capacity posed a significant challenge in handling and analyzing the complex and voluminous data generated by genomic studies. The rapid advancement in computing technology over the past few decades, concurrently to the advance in high-throughput genomic technologies, has dramatically changed this landscape. Today's high-performance computing systems can process and analyze vast amounts of data at speeds previously unimaginable, enabling the development of sophisticated software capable of dissecting genomic patterns accurately[8].

Using statistical genetics to study complex traits

Complex traits are characterized by their non-Mendelian inheritance patterns. These traits do not follow the simple monogenic model underlying Mendelian traits. Instead, they arise from the interplay of multiple genetic and environmental factors[9]. Environmental factors play a crucial role in the expression of complex traits including lifestyle choices, diet, exposure to toxins, and even social and economic factors. These factors can explain why, even among individuals with similar genetic makeup, such as twins, there can be differences in the expression of traits. Complex traits can be binary, such as diseases, or quantitative traits such as height, and BMI.

Throughout genetics research history different models of trait inheritance have been proposed. The infinitesimal model of inheritance, also known as polygenic model, was developed by Ronald Fisher in 1918[10]. It describes quantitative traits as being influenced by a sum of non-genetic and genetic factors contributing to the trait with a small, infinitesimal effect. These small effects cumulate to influence the phenotype. A more recent concept proposes that all genes expressed in relevant cell types impact complex traits. This omnigenic model suggests that a trait can be influenced by genes outside of traditional disease-specific pathways, highlighting the interconnectedness of genetic networks[11].

One statistical tool that has profoundly reshaped our growing understanding of complex traits is Genome-Wide Association Studies (GWAS). GWAS comprehensively examine genetic variations across many genomes to find variants statistically associated with a particular binary or quantitative complex trait providing insights into its genetic basis[12]. The establishment of large biobanks that combine both genotype and phenotype data from hundreds of thousands of individuals from global populations such as the UK Biobank[13], Biobank Japan[14] and FinnGen[15] propelled large GWAS[2]. Today, the GWAS catalog, a curated collection of GWAS, contains data from more than 6,680 publications encompassing more than 67,000 GWAS summary statistics[16]. After identification of risk variants

associated to traits through GWAS, a significant challenge in the study of complex traits lies in linking the prioritized genetic risk variants to their functional outcomes, known as the variant-to-function challenge[17]. This challenge is at the forefront of current statistical human genetic research, as understanding the functional implications of genetic variations is crucial for translating research into medical clinics and achieving the goal of beneficiating the broad population.

As sample sizes in GWAS increased, the presence of genetic correlation between common traits was revealed. Analysis of GWASs from 588 traits showed that 90% of the loci with association signals were shared between multiple traits across multiple domains, highlighting the interconnected nature of genetic information proposed in the omnigenic inheritance model[18]. This underscores the need for a more nuanced understanding of how genes contribute to the overall phenotype of an organism. The phenomenon wherein a single genetic variant or gene influences multiple, seemingly unrelated traits is called pleiotropy[19]. These pleiotropic effects can manifest in different tissues or organs and are complex to disentangle, as the gene product may have multiple functions or be involved in various pathways within the organism.

Pleiotropy can be classified into biological, mediated or spurious[20]. In biological pleiotropy causal variants influencing distinct traits are located within the same gene or regulatory unit. Mediated pleiotropy occurs when a genetic variant directly influences one trait, which subsequently impacts another. In this case, while GWASs may identify an association between the variant and the second trait, this association disappears when adjusting for the first trait. Finally, spurious pleiotropy refers to the apparent association between a genetic variant and multiple traits that is not reflective of a true biological relationship. Instead, this association arises due to methodological or design artifacts during the planning or analysis stages of the study and can lead to inaccurate or misleading results in genetic research.

The thesis will present research on the interrelation between pairs of diseases and will focus on three human health disorders: type 2 diabetes, osteoarthritis, and schizophrenia. These conditions are exemplary of complex traits, each influenced by a unique interplay of genetic and environmental factors. A brief overview of each condition will be introduced next.

Pathogenesis, prevalence and genetics of type 2 diabetes

Type 2 diabetes is a chronic metabolic disorder characterized by elevated blood sugar levels[21]. The hallmark of type 2 diabetes is an impaired insulin sensitivity leading to higher insulin tolerance. Insulin is a hormone produced by the pancreas that regulates blood sugar levels[21]. In response to this

insulin resistance, the pancreas initially produces more of this hormone to maintain normal blood sugar levels by increasing the cell's glucose intake. Over time, however, this compensatory mechanism fails as the pancreas becomes unable to sustain the high levels of insulin production[22]. This leads to a gradual increase in blood glucose levels, resulting in the onset of type 2 diabetes. It is the most common form of diabetes mellitus, affecting over 521 million of people worldwide in 2021[23]. The increasing global prevalence of this condition underscores the importance of public health efforts focusing on prevention, early detection, and effective management strategies.

There is no curative treatment available for type 2 diabetes and medical interventions are directed towards disease management involving monitoring, lifestyle modifications such as diet and exercise and drug therapies to manage blood sugar levels. Type 2 diabetes usually develops in adulthood and is strongly linked to environmental factors including obesity, sedentary lifestyle, and age as well as genetics serving as a prime example of how lifestyle and genetic factors combine to influence health[22]. Having further chronic health conditions is common in type 2 diabetes patients, with varying profiles of multimorbidity[24].

Genetic heritability of type 2 diabetes was estimated at around 50%[25]. To unravel the genetics underlying this disease, large sample sizes are required. Data must be gathered from global populations to address the ancestry-specific genetic nuances of type 2 diabetes. For instance, at the same BMI level or waist-to-hip ratio, East Asians are at higher absolute risk of type 2 diabetes compared to Europeans[26]. Tackling the genetic contribution to type 2 diabetes in both global and specific populations, the largest GWAS for type 2 diabetes to date was recently published by the Type 2 Diabetes Global Genomics Initiative (T2DGGI). This multi-ancestry GWAS included data from 2,535,601 (428,452 cases) individuals of which almost 40% were from non-European ancestries[27].

Pathogenesis, prevalence and genetics of osteoarthritis

Osteoarthritis, a whole-joint degenerative disorder, is the most common form of arthritis affecting almost 600 million people worldwide[28]. It is characterized by irreversible cartilage degradation eventually followed by its complete loss and synovial inflammation[29]. Its most common symptom is pain accompanied by stiffness and decreased joint mobility. Osteoarthritis is one of the leading causes of disability worldwide affecting weight-bearing joints such as the knee, hip, and spine as well as non-weight-bearing joints such as the hand, and finger[30]. Primarily elderly are affected by this disease, but it can occur in younger individuals as well mainly due to traumas. Similarly to type 2 diabetes, there is no curative treatment for this complex disease and management therapies target

the alleviation of pain[28]. In severe cases, surgical interventions like joint replacement may be necessary.

Genetics plays a significant role in the development of osteoarthritis alongside environmental and lifestyle factors such as obesity, female sex, joint injuries and age. Patients suffering from osteoarthritis are also at higher risk of having other chronic conditions compared to individuals without osteoarthritis[31]. The genetic heritability of this chronic disease was estimated between 20% and 60%[32,33]. Advances in genetic research are shedding light on the molecular mechanisms behind osteoarthritis, offering hope for more targeted therapies and prevention strategies. The largest osteoarthritis GWAS to date is the result of efforts from the Genetics of Osteoarthritis consortium[34]. This meta-analysis has identified 100 independent genetic risk loci for 11 osteoarthritis phenotypes including both weight-bearing and non-weight bearing joints and consisted mostly of samples from European ancestry.

Pathogenesis, prevalence and genetics of schizophrenia

Schizophrenia is a chronic and severe mental health disorder that affects a person's ability to think, feel, and behave[35]. Characterized by episodes of psychosis involving delusions, hallucinations, disorganized thinking, and other cognitive impairments, it typically emerges in late adolescence or early adulthood. Symptoms of schizophrenia are typically divided into three categories: positive symptoms, which include hallucinations and delusions, negative symptoms, such as reduced emotional expression, social withdraw and lack of motivation, and cognitive symptoms, which includes impaired attention and memory[35]. Despite the low global prevalence of less than 1%[36], schizophrenia represents one of the 15 leading causes of disability worldwide[37]. Patients suffering from schizophrenia have high rates of physical comorbidities including cardiovascular and metabolic diseases[38,39]. Diagnosis is based primarily on patient history and observed behavior, for instances episodes of substance abuse. The primary treatment for managing schizophrenia symptoms is antipsychotic medication, which should be complemented by psychosocial interventions, including psychotherapy, social skills training, and supported employment.

Environmental factors including exposure to viruses, malnutrition before birth, problems during birth, and psychosocial factors play a substantial role in the development of this condition. Additionally, genetic epidemiological studies have shown that schizophrenia has an estimated heritability of ~80%[35]. The largest schizophrenia GWAS to date consists of data from 76,755 individuals with schizophrenia and 243,649 controls and has identified several genetic variants associated with risk of schizophrenia in 287 distinct genomic loci[40].

Multimorbidity of complex traits: prevalence, patterns, and implications

Having multiple chronic health conditions is denominated multimorbidity. The overall global prevalence of multimorbidity was estimated from 126 studies including data from over 15 million people at 37.2%. The region with highest prevalence of multimorbidity was South America (45.7%) followed by North America (43.1%), Europe (39.2%) and Asia (35%)[41]. In the world, it is estimated that 65% of the population over 65 years and 85% of the population over 85 years suffer from more than one long-term medical condition simultaneously[42]. This pattern demonstrates the well-known association between multimorbidity and age[43]. Additionally, observational studies have shown that overweight is associated with increased risk of multimorbidity and has been linked to heterogeneous comorbidities including digestive, respiratory, neurological, musculoskeletal, infectious, and malignant diseases[44,45]. A study with 150 Brazilians with severe obesity reported that 90.7% of them suffered from two or more conditions and 76.7% suffered from at least three conditions. The prevalence of three or more health conditions was 90% for the individuals between 45-65 years compared to 65.9% for 18-34 years[46].

Due to the increasing tendency of the world's average BMI and life expectancy, the number of people affected by multimorbidity is predicted to substantially increase on a global scale over the next years[47]. Improved diagnostics capabilities and offer also play a role in the rising number of comorbid cases. By diminishing quality of life and increasing healthcare expenses, multimorbidity represents more than an individual burden influencing the health system and society as a whole[48]. However, despite the rapidly increasing number of patients affected by multimorbidity among global population, most health-related research is employed in preventing and treating diseases individually[49]. Consequently, healthcare services are not designed to assist multimorbid cases. This leads to several negative consequences for both patients and the healthcare services. Patient care is suboptimal and even harmful due to inadequate polypharmacy that increases treatment burden leading to increased healthcare expenses[50].

Understanding multimorbidity requires a paradigm shift from viewing it as a random collection of individual conditions to predictable disease clusters, influenced by an interplay between genetic and environmental factors. To systematically identify these multimorbidity clusters, several methods have been applied. A study using data from the UK Biobank[13] systematically created an atlas of 11,285 multimorbid disease pairs among 438 common diseases[51]. Cardiometabolic and mental health conditions are the most consistently identified clusters, though musculoskeletal and allergic condition clusters have also been observed[42].

Multimorbidity patterns differ strongly across life stages, sexes, ancestries[52] and socio-economic groups[53]. Co-occurrence of multiple diseases tends to appear at a younger age in low- and middle-income countries. A multi-ancestry study showed African American ancestry individuals presented the highest number of multimorbidities at an earlier age than patients of other ancestries[54]. This study also showed that the most common diseases with comorbidities transversing global populations are lipidemia, hypertension, and diabetes regardless of age or obesity level. Multimorbidity increased with age in both with and without obesity groups. This earlier onset can often be attributed to a range of complex socio-economic and environmental challenges. including constrained healthcare systems and social support, environmental and socio-economic stressors related to poverty[42].

As multimorbidity patterns, the correlation between different traits can vary by ancestry, gender and social-cultural environment[55]. Genetic correlation between phenotypes can reveal shared etiology and insights into disease mechanisms. Understanding which diseases cluster together, recognizing patterns across the world, and identifying predictors and determinants to prevent the development of multimorbidity are some of the research questions in the field of multimorbidity. Approaches that target these questions can help design tools to assist clinicians in prevention, early intervention, and treatment of co-occurring health disorders. Moreover, researching the genetic processes underlying multimorbidity can lead to the development of novel drugs that maximize the benefits and limit the risks of treatments, thus preventing the risks of polypharmacy and enabling a more personalized treatment.

Definition of scientific problem and research question

My doctoral thesis focuses on unraveling the shared genetic basis that underlies specific pairs of complex combinations of coexisting chronic conditions. The studied complex combinations of diseases concurrently affect two or more different body systems within a single individual. The rationale behind investigating pairs of diseases, as opposed to larger groups, beyond the relative simplicity of the model, is rooted in the fact that targeting a combination of two diseases can impact a broader population compared to focusing on individuals with a very specific combination of three or more diseases. This approach aims at advancing and democratizing the translation of research findings into clinical applications.

The primary aim of this thesis is to elucidate biological mechanisms and prioritize genes involved simultaneously in the two studied diseases. Leveraging large-scale GWAS data, genetic correlation between the studied complex traits can be estimated, shedding light on the presence of an underlying shared genetic etiology. In addition, genetic causal inference analysis can provide insights into biological pathways mutually implicated in both studied conditions. Once the genome-wide correlation and causality are established, subsequent local analyses can be performed to derive a biologically informed list of prioritized effector genes that concurrently influence both health conditions. The prioritized genes may act on both diseases either in the same or in opposite direction of effect. This phenomenon of genetic variants or genes affecting different diseases is known as pleiotropy, as explained above.

Observational studies can offer valuable insights into the epidemiological correlation between two traits, that can exhibit either a positive or negative direction of association[56]. This research capitalizes on such findings to select common yet understudied comorbid complex traits to be the topic of each presented project. The two studied comorbid conditions were mainly chosen as examples for very frequent co-existing classes of complex diseases, namely cardiometabolic and osteoarticular diseases[57] and cardiometabolic and psychiatric diseases[58]. Another factor that influenced this choice was the availability of large and publicly available GWAS data as well as molecular data from disease-relevant tissues.

In the scope of this doctoral thesis, two peer-reviewed publications tackling the shared genetic etiology between co-occurring diseases pairs were produced. The first project studied the comorbidity between type 2 diabetes and osteoarthritis, which share common risk factors such as age and increased BMI, and most studies report a epidemiological positive correlation between them[59]. Together, type 2 diabetes and osteoarthritis affect more than 950 million patients in the world. The shared genetic etiology underlying type 2 diabetes and schizophrenia was the pair of comorbid conditions chosen for the second project. These two conditions are also observed more often together than by chance, positive correlation results persisting even after adjusting for antipsychotic medications, which at times might increase the patient's BMI[60].

We posit that complex health conditions exhibiting epidemiological associations may, to some extent, have shared genetic aetiology. This doctoral thesis aims to leverage statistical genomics data-based approaches, along with large-scale genomics and multi-omics data from diverse human populations to study these conditions. The overarching goal is to unravel the shared etiology of specific comorbid pairs of health conditions, thereby offering insights into disease biology. Additionally, this research

seeks to identify potential targets for novel candidates or repurposing opportunities in therapeutic intervention.

Literature review and state-of-the-art research

Prioritizing putative causal genes

Following the identification of genetic variants associated with a trait by GWAS, the function of these variants remains unknown, challenging the translation of the findings into meaningful biological insights[17]. Various approaches have been developed to address the well-defined variant-to-function challenge, aiming to pinpoint candidate genes responsible for the observed associations. A very robust method consists of generating a pipeline for gene prioritization by combining different lines of biological evidence that support the involvement of a gene.

In 2021, an open resource called Open Targets was developed to advance translation of genomics discoveries[61]. It integrates GWAS data with molecular and functional genomics, along with drug information in a standardized manner. To prioritize therapeutic target genes for drug discovery, Duffy et al. developed a genetic priority score by integrating eight genetic features with drug indications using the Open Targets databse[62]. The score was further extended by the direction of genetic effect and drug mechanisms, resulting in a directional prioritization score.

Disentangling the shared genetic etiology between complex traits

To uncover links between multimorbid conditions using genetic data, several collaborative projects have been assembled. For instance, a collaborative called GEMINI (Genetic Evaluation of Multimorbidity towards INdividualisation of Interventions) (https://sites.exeter.ac.uk/gemini, accessed on the 1st of February 2024) was formed in the UK and the in the US the Multimorbidity Mechanism and Therapeutics Research Collaborative (MMTRC) has published multiple peer-reviewed papers[52].

A straightforward statistical genetics approach to tackle multimorbidity without resorting to individual level data is to thoroughly explore the genetic risk loci identified by GWAS for different health conditions that are shared among them. This was implemented, for instance, in a study focusing on age-related conditions[63]. Based on individual GWAS, the authors found 22 loci shared between the prioritized age-related phenotypes, including *APOE*. Specifically for type 2 diabetes, the *TCF7L2* locus and the *FTO* locus were shared with different cancer types including breast and prostate cancer, which are epidemiological respectively positively and negatively associated with type 2

diabetes. Another study tackling multimorbidity's association with aging prioritized 995 genes related to multiple age-related diseases using colocalization and clustering analyses[64].

A substantial proportion of studies researching the shared genetic etiology between complex traits focuses on psychiatric traits[65-69]. This is partly due to the high correlation between psychiatric phenotypes arising from overlapping phenotypic definition. Different approaches have been applied in this field. For instance, the genetic association between twelve psychiatric disorders were studied by Romero et al., who conducted cross-trait meta-analysis on these traits to identify pleiotropic genetic variants using publicly available summary statistics[65]. Wingo et al. looked at the overlap between psychiatric and neurodegenerative diseases, finding robust evidence of shared genetic etiology and molecular processes between these traits. The authors performed genetic correlation analysis and identified putative pleiotropic or shared causal proteins and transcripts by integrating GWAS results with brain transcriptomes and proteomes[67].

A further group of highly correlated health disorders exhibiting substantial genetic component are autoimmune diseases, which were investigated by several cross-disorder studies[70]. Noteworthy, a significant challenge in understanding autoimmune diseases lies in deciphering gene-environment interactions, which significantly influence their development. One study, examining 21 autoimmune diseases, revealed that 69% of the genetic loci associated with one disease overlapped with risk loci of other autoimmune diseases, indicating shared genetic mechanisms[71]. The relationship between allergies and autoimmune diseases was the focus of multiple genetic studies. One study observed an enrichment of allergy risk loci among loci associated with autoimmune diseases[72]. Another study performed multi-trait GWAS including six autoimmune or allergy-related diseases using data from large-scale biobanks[73]. This study identified four shared loci between autoimmune and allergic diseases.

Several studies have also conducted research on complex combinations of traits that affect more than one tissue or organ in the body. A popular genetic approach to study these comorbid conditions is to perform multi-trait GWAS combining multiple phenotypes, as mentioned above for autoimmune and allergic diseases. For instance, a multivariate GWAS was conducted by a study researching the shared genetic etiology of psycho-cardiometabolic diseases using genomic data on coronary artery disease, type 2 diabetes and depression[74]. The authors identified genetic variants with cross-trait influence by using genomic structural equation modelling using summary statistics from each univariate GWAS. By making use of the latent multimorbidity factor generated in the genetic factor analysis, a new set of

summary statistics was estimated for the common factor, and 389 SNPs were found to be associated with the investigated psycho-cardiometabolic multimorbidity.

Many other studies have researched the association between cardiometabolic and psychological traits. Leveraging polygenic scores and linkage disequilibrium (LD) score regression, a study investigated the differences in the shared genetic etiology between cardiometabolic traits and earlier or later onset major depressive disorder[75]. All cardiometabolic polygenic scores were associated with depression and significant genetic correlations were found between depression, BMI, coronary artery disease, and type 2 diabetes. Another study examined the multimorbidity between cardiometabolic traits and dementia using individual level genetics data from the UK Biobank to create a cardiometabolic multimorbidity index based on a polygenic score for dementia, which was applied on brain images[76].

A further example of studies leveraging multi-trait GWAS to investigate correlated diseases aimed at investigating the genetic basis of endometriosis with multisite chronic pain and migraine[77]. An additional approach employed to dissect the association between endometriosis and its comorbidities was presented in a study that performed causal inference using Mendelian randomization analyses[78]. This study highlights some potential causes and outcomes of endometriosis, such as depression and ovarian cancer, respectively.

The genetic association between obesity and multiple sclerosis was investigated by Zeng et al., who performed genetic correlation, causal inference, and cross-trait GWAS analyses[79]. The study reports a significant genetic correlation and causal effect between both traits as well as a list of 39 shared genetic variants. A further complex combination of diseases is found between depressive disorder and osteoarthritis. Zhang et al. found evidence of a positive genetic correlation and shared genetic etiology[80]. The shared genetic etiologies of many other disease combinations have been studied through different approaches. However, a clear indication of putative drug targets that can be prioritized in the drug development target is lacking in most studies.

Methodology

The aim of this doctoral thesis is to explore the shared genetic underpinnings between pairs of complex traits by harnessing publicly available genomics datasets. Across the peer-reviewed and published projects outlined in this work, a consistent data-driven genomics approach has been adopted to unravel the common genetic factors associated with the studied conditions. Firstly,

genome-wide analyses were conducted to unveil the genetic correlation and causality between each pair of conditions. Secondly, employing genetic colocalization analysis, a comprehensive list of putative effector genes simultaneously influencing both diseases was derived. To delve deeper into the genetic insights, more exhaustive analyses were undertaken for the top-scoring genes, including pathway enrichment analysis and an exploration of druggability.

Our approach identifies genes acting on both health conditions with the same direction of effect and with opposing direction of effect. Considering the implications for translation and precision medicine, both outcomes are of interest. When a gene acts on two conditions with the same direction, it becomes a potential drug target for bilateral treatment. On the other hand, a gene influencing two conditions with divergent effects highlights opposing biological mechanisms underlying both conditions. In cases where a drug targets such a gene, caution should be exercised for comorbid patients, who should avoid undergoing this gene treatment. Meanwhile, all other patients should be monitored for signs of the condition associated with an increased risk linked to this gene.

Software and coding environment

The main programming languages employed in this research were R and bash. R was used to code most scripts, while bash was used for calling some external software or to perform data processing steps in very large data. RStudio was used to edit R scripts. When writing and compiling R code on the servers, an RStudio singularity container was used. MobaXterm professional was used to establish SSH connections from within a Microsoft Windows environment. Microsoft Word was used to write the manuscripts, conference abstracts and this thesis. The reference manager used was EndNote. Evernote was used to document meetings and manage the open tasks in different project. Microsoft PowerPoint and Excel were used to prepare presentations as well as posters and finalize supplementary tables, respectively. To generate figures we used Microsoft PowerPoint, Inkscape and Biorender.

Several measures were considered to adhere to openness and reproducibility in research. Version control was achieved using git and by backing up data with Microsoft OneDrive. All generated code was deployed in form of scripts and made publicly available on GitHub and Zenodo. Both peer-reviewed publications attached to this thesis were published as open science articles free for the public.

Data

Data generation was not part of this thesis. Instead, publicly available summary statistics from genomics data were used. The choice of using summary statistics instead of individual level genomics data is motivated by three main reasons. The first reason is the power gain of combining multiple GWAS from different cohorts in a meta-analysis instead of using a smaller cohort of genotyped individuals. Secondly, data access is managed differently across studies with some of them having very restrictive data sharing measures. Finally, by using publicly available data, the analysis pipeline developed in this thesis can be reproduced and adapted by other researchers, boosting open science and collaborative works.

For the studied diseases as well as for further diseases and qualitative traits related to them, literature research was conducted for each trait to find the most recent GWAS with the largest summary statistics available. GWAS unadjusted for BMI were prioritized to avoid collider bias. Concordance between ancestries of the main diseases and related traits was not considered in this first step since the largest single-ancestry GWAS available for the studied traits were unanimously from European ancestry.

An extensive literature and databases search was performed to acquire molecular quantitative trait locus (QTL) data specific to tissues associated with each primary disease under investigation. For type 2 diabetes, only pancreatic islets were considered in the first project. For the second project, a broader array of type 2 diabetes-relevant tissues was used including liver, subcutaneous ad well as adipose tissue and brain[27]. Osteoarthritis investigations relied on data extracted from the cartilage and synovium of osteoarthritis patients. In the case of schizophrenia, brain data from different adult regions and dorsolateral data from fetal brains were considered. Notably, for all mentioned tissues except for cartilage and synovium, the molecular data either originated from healthy individuals, or the disease status was not taken into account.

Genome-wide analyses

Leveraging both global genetic correlation and causal inference methods, genome-wide insights about the shared genetic etiology of the studied comorbidity can be gained. To assess the genome-wide genetic correlation between conditions, LD score regression was employed, which requires only GWAS summary statistics instead of individual level data[81]. This method is unbiased against sample overlap and, by restraining the computations to well-imputed HapMap3 SNPs, it achieves computational efficiency.

To infer causality between the studied conditions, we applied Mendelian randomization (MR) analysis[82]. The MR method uses genetic variants as instrumental variables, mimicking a randomized control trial in a natural experiment setting. By relying on non-modifiable genetic information, MR can address confounding and reverse causation challenges common in observational studies. Causal inference is performed from an exposure trait to an outcome, a variable representing the result of interest. Both the exposure and the outcome can be a modifiable life factor, a disease or a biomarker. MR assesses the causal impact of the exposure on the outcome by examining the relationship between genetic variants influencing the exposure and the outcome. The genetic instrumental variables must satisfy three assumptions to be considered valid, which are also visualized in Figure 1:

1) **Relevance**: instrumental variables must strongly associate with the exposure.
2) **Independence**: instrumental variables must be independent of potential confounders that might influence the outcome through mechanisms other than the exposure of interest.
3) **Exclusion restriction**: instrumental variables do not affect the outcome other than through the exposure and do not affect any other trait that has a downstream effect on the outcome of interest.

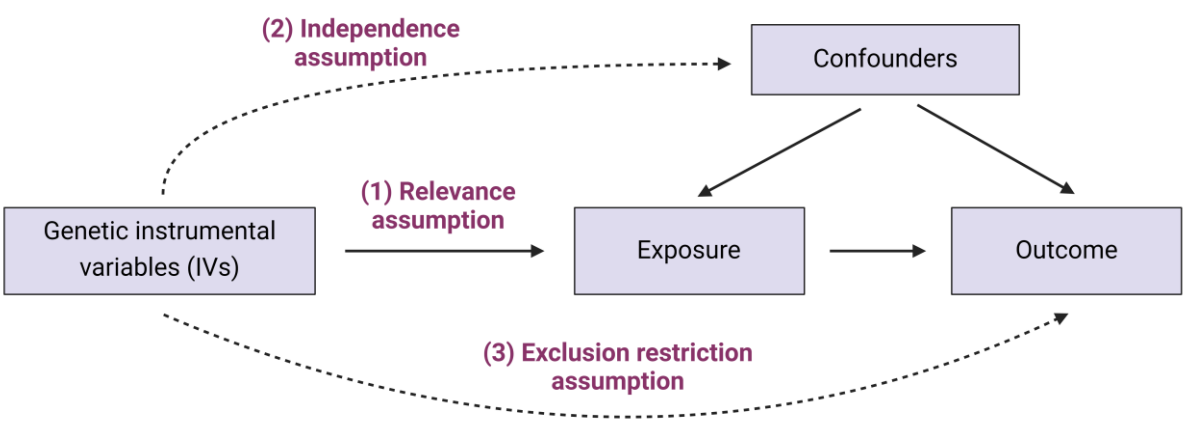


Figure 1: Overview of Mendelian randomization assumptions.

The first condition can be formally tested by calculating the F-statistic of the instrumental variables, which is a measure of the association strength between the genetic variant and the exposure calculated as $beta^2/se^2$, where $beta$ is the effect of the variant on the exposure and se the standard effect of this effect. Weak instrumental variables defined as F-statistic < 10 were removed. The other two conditions can only be assessed through sensitivity analysis aiming at disproving confounding or pleiotropic mechanisms.

As the main analysis, we applied the inverse variance weighted (IVW) method, which performs a random-effects meta-analysis of the Wald ratios for each SNP. As an initial sensitivity analysis, we applied the weighted median and the MR-Egger regression methods to ensure consistency of the

effect size direction. The weighted median method relaxes the relevance assumption by requiring that at least 50% of the variants are valid instruments. MR-Egger combines Wald ratio estimates into a meta-regression using an intercept and a slope parameter to estimate the causal effect adjusted for directional pleiotropy. The intercept of MR-Egger regression is a measure to assess horizontal pleiotropy. Finally, we tested for heterogeneity using the Q-statistic. To account for multiple testing, we corrected the p-values of the IVW results using the Bonferroni method.

Increased BMI plays a significant role in the pathogenesis of type 2 diabetes. Genetic variants associated with increased BMI might exert an effect via genes expressed in brain or adipose tissue[83]. Hence, to determine the tissue-specific role of BMI in the studied conditions, we applied bidirectional MR between BMI and each condition restricted to BMI instruments colocalizing with eQTLs in brain and adipose tissue, respectively.

Regional genetic colocalization analysis

Shifting from a genome-wide perspective to a more local and refined one, we expected to find more specific insights about the biology underlying the studied comorbidities by pinpointing shared signals. To tackle this goal, we conducted pairwise genetic colocalization analyses to find genomic risk loci shared between the studied diseases[84]. This Bayesian method compares signals between two traits in a pre-defined genomic locus and assigns posterior probabilities to five hypotheses regarding the individual trait signals, as depicted in Figure 2.

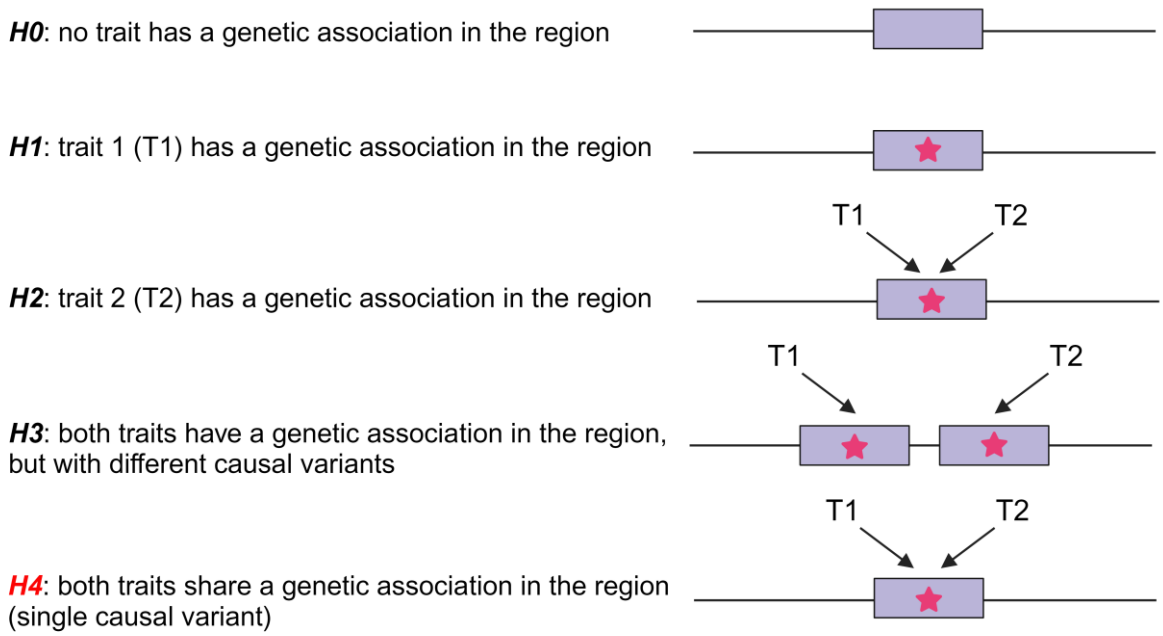


Figure 2: Overview of the five different hypotheses tested in genetic colocalization analysis.

For this work, we are mostly interested in the posterior probability of H4 (PP4). Evidence for colocalization of a shared risk signal was reported at a threshold of PP4>0.8. Since genetic colocalization is a Bayesian method, the concept of type 1 error is not applicable. We kept the default priors suggested by the *coloc* implementation. For each colocalized regions (PP4>0.8) a 95% credible set for the causal variant was computed. The type 1 error refers to the probability of wrongly rejecting the null hypothesis when it is actually true. Some might think that when performing enough regional colocalizations some regions might colocalize (PP4>0.8) by chance. However, this is not the case since test probabilities are assigned to each of the five different hypotheses without prioritizing H4.

Effector gene prioritization

The genetic colocalization analysis outputs regions with putatively shared signals associated with the studied comorbidity along with a set of variants that most likely comprises the true shared causal variant. We have developed a gene prioritization study design to tackle the variant-to-gene challenge and identify putative effector genes that act on both conditions simultaneously. We selected all genes around the colocalized signals and scored them based on orthogonal lines of biological evidence of involvement with each of the studied conditions. It is well-established that regulatory elements might interact with genes further away and enhancer elements can reside in long-range distances to the affected gene[85]. Hence, we have opted to score all genes in a 2Mb window around the lead colocalization variant to not miss out on any long-range interaction.

The practice of scoring genes surrounding risk signals by integrating multiple biological lines of evidence is a widely accepted approach in statistical human genetics, as evidenced by its application in numerous peer-reviewed publications[34,86]. It is essential to note that certain elements of the score may not directly correspond to the genetic factors influencing the human expression of the condition, such as phenotypes observed in knockout mice. However, these components are not evaluated in isolation; instead, they are collectively considered as a composite of small pieces of evidence indicating involvement in the pathogenesis of diseases.

Different tissues are relevant for the studied diseases, these are depicted in Figure 3B. For both studies, we used six biological lines of evidence structured in an orthogonal way to avoid over-representation of any evidence (Figure 3A):

1. Multi-trait genetic colocalization analysis between both studied conditions and molecular quantitative trait loci (QTL) from disease-relevant tissues.
2. Differential gene expression in disease-relevant human tissues.

3. Knock-out mice phenotypes related to each condition.
4. Rare and syndromic human diseases with phenotypes related to each condition.
5. Previously defined high-confidence effector genes in the literature.
6. Existence of missense variants in the set of variants of the colocalizing signals.

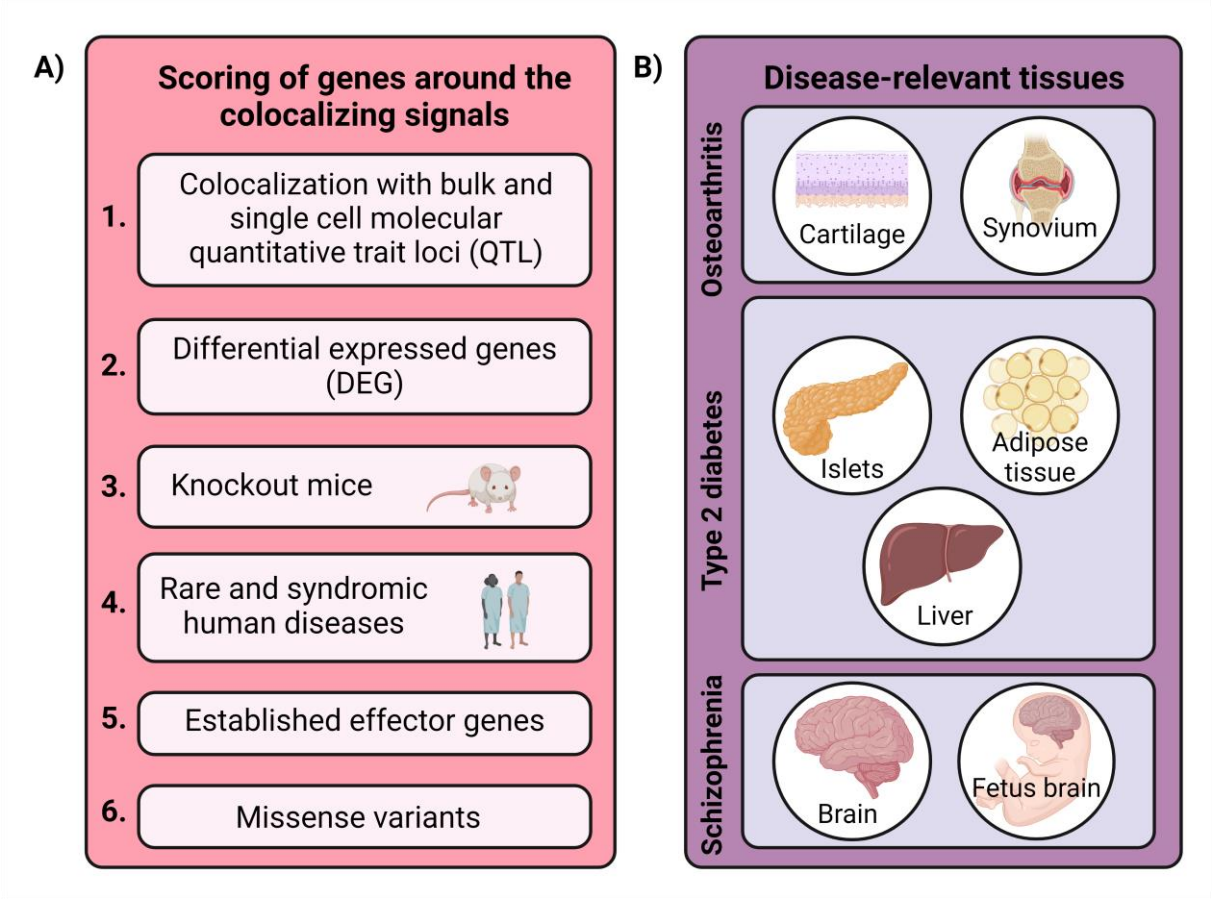


Figure 3: A) Overview of the biological lines of evidence used to score genes around colocalizing signals. B) Overview of the relevant tissues for each studied disease.

Our analysis might overlap with criteria used to define a gene as high confidence for the individual diseases in previous work. Hence, to incorporate the fifth line of evidence orthogonally, we adjusted the gene score. Specifically, if a gene scored zero in our analysis but had been previously designated as high confidence for a particular condition in earlier studies, we updated its score to one.

For each disease, one score was created by summing up the result of each biological line of evidence that indicated involvement in that particular condition. Additionally, we scored the genes in the vicinity of the colocalized signals based on the presence of missense variants associated with the genes within the 95% credible sets of the colocalization. This missense score is disease agnostic. The scores for each condition along with the missense variant score were summed up to generate a final

score for each gene. If a gene only obtained a score in the missense variant category, the total score was maintained at zero since this evidence alone does not indicate involvement in the studied comorbidity.

Analyses on derived list of putatively effector genes

To glean further insight into the biology behind the highest scoring genes, we performed enrichment analysis using curated gene networks in humans and the enrichment software from the ConsensusPathDB (http://cpdb.molgen.mpg.de/)[87]. To identify potential targets for drug repurposing, we assessed the druggability status of these top-scoring genes by consulting the Druggable Genome database[88]. Specifically, if a gene happened to be the target of an investigational or approved drug, we delved into the prescription advice and mechanism of action associated with that drug using the DrugBank database (https://www.drugbank.com).

Additionally, we conducted a deep dive consisting of a more in-depth analysis of the function of the top scoring genes. Firstly, we looked at the association of these genes with endophenotypes of each condition or shared risk factors. Secondly, we looked at the direction of effect of each line of evidence and tried to harmonize across them. Finally, we performed causal inference analysis between the expression of the genes in disease-relevant tissues and the individual conditions using Mendelian randomization.

Discussion

Multimorbidity, characterized by the simultaneous presence of multiple chronic health conditions within an individual, represents a rising public health challenge. Global trends such as increasing average BMI and life expectancy contribute to the escalating prevalence of multimorbidity cases. Beyond the personal burden, including polypharmacy and adverse side effects, multimorbidity poses a challenge to society as a whole by increasing treatment demands that subsequently inflate healthcare expenses. Educating both healthcare practitioners and the general population about multimorbidity requires a shift in research efforts, moving away from looking at individual diseases towards recognizing the interplay among co-existing health conditions.

In this thesis, I present an approach to disentangle the shared genetic aetiology between pairs of co-occurring chronic diseases that affect multiple distinct body parts, with type 2 diabetes-osteoarthritis and type 2 diabetes-schizophrenia comorbidities as examples. We have successfully identified putative genes influencing simultaneously both studied correlated traits and showed interesting

insights into disease biology. For type 2 diabetes and osteoarthritis, our statistical genomics analyses corroborated the well-established epidemiological positive association between traits. We have also shown evidence of a stronger correlation between type 2 diabetes and knee rather than hip osteoarthritis. As expected, this positive association is heavily mediated by obesity, but our results highlight two further potential biological mechanisms underpinning this comorbidity: the Wnt/β-catenin signalling pathway and an imbalance between osteoblasts and adipocytes differentiation in adult bone marrow.

For type 2 diabetes and schizophrenia, our results show evidence of a negative genetic correlation and association, contrasting the observed positive epidemiological correlation. An external factor that introduces further complexity in the relationship between both conditions is the use of antipsychotic medications by schizophrenia patients. Some antipsychotic drugs are known to have an increasing effect on body weight, which in turn is associated with type 2 diabetes. Nevertheless, we showed evidence of a protective effect of schizophrenia liability on increased BMI, replicating previous results[89,90]. This result opposes the well-established causal effect between BMI and type 2 diabetes, pointing to potentially different adiposity-related mechanisms underpinning each condition.

Comparing this work to previous publications, we have focused on using only summary statistics instead of individual level data due to substantial power gain and easier reproducibility of the developed data-based pipeline. Hence, the use of well-powered comorbidity-specific polygenic scores was not possible, which did not pose a major challenge considering that current polygenic scores only partially capture the genetic liability to a disease. Multi-trait GWAS were also not performed in order not to neglect the identification of signals influencing both studied traits in opposing directions. Our approach combined genome-wide and local analyses with a deeper dive into specific top scoring genes identified by our gene prioritization pipeline.

Previous research has shown that multimorbidity tends to manifest a decade earlier in communities with socioeconomic disadvantages, where it is linked to earlier mortality, reduced functional abilities and quality of life, as well as an increased demand for healthcare services[42]. However, the current research in diverse populations faces limitations due to a gap in data availability. There is a notable bias in GWAS towards European ancestry individuals[91]. This bias can be partially attributed to historical disengagement of the scientific community with diverse populations and the consequent mistrust from these communities[92]. The primary focus on Eurocentric populations in GWAS poses a limitation in the extrapolation of findings to a broader demographic. In recent times, there has been a recognition of this bias, leading to the formation of large international collaborative efforts aiming to

broaden the scope of GWAS studies beyond European ancestry populations[93-95]. Notably, the T2DGGI consortium published a multi-ancestry type 2 diabetes GWAS that includes data from 2,535,601 individuals, of which almost 40% belonged to non-European ancestry groups[27]. These collaborative endeavors represent a positive step toward addressing the limitations and promoting inclusivity in genetic research.

Deciphering the function of risk variants identified by GWAS integration with molecular QTL data is a state-of-the-art approach. Whereas efforts have been made to close the ancestry gap of GWAS data, very little has been done to generate molecular data coupled with genetics from diverse population[96]. The gap in data availability from diverse populations is even more striking for molecular data from specific primary tissues relevant to diseases, which are essential for interpreting GWAS signals as different cell types exhibit distinct gene expression profiles and regulatory landscapes[97]. In this thesis, multi-ancestry GWAS were used to study the type 2 diabetes and schizophrenia comorbidity and the lack of molecular QTL data from diverse populations limited the interpretation of the signals shared between both conditions.

GWAS may be susceptible to bias based on how phenotypes are defined and how heterogeneous they are. For instance, overlapping phenotype definitions in psychiatric diseases can lead to misdiagnosis, which will impact the identification of genetic associations. Additionally, GWAS may produce biased results if the control group is not carefully selected or if there are underlying biases in the control population. For instance, in the case of type 2 diabetes, there might be false positives samples in the control groups, who are not yet diagnosed with type 2 diabetes, but will get a formal diagnosis of this disease later in life. Exclusion of these biased controls has the downside of leaning the GWAS results towards individuals with very prevalent symptoms of a disease, the so called “super-cases”. This issue is also known as the liability threshold bias, which can lead to different and even opposing patterns between observational and GWAS-based studies.

Analyzing shared genetic etiology may be biased by the presence of shared risk factors, introducing possible confounding bias. One approach to reduce this bias employed in this thesis was to extend the analyses to endophenotypes or diseases correlated to the studied comorbidities. For instance, both type 2 diabetes and osteoarthritis are heavily mediated by increased BMI. Hence, we expect both diseases to be partially correlated due to this common association. Failure to consider all relevant confounders may lead to misinterpretation of genetic associations and their role in shared etiological pathways. Bias can also arise if there is overlap between samples used in different studies analyzing shared genetic etiology. This overlap may result in inflated estimates of shared genetic

effects, potentially misleading interpretations of the extent of genetic correlation between traits. In this work, the largest GWAS summary statistics were used and these large sample sizes account partially for potential overlapping sample bias.

Considering drug development or repurposing opportunities, the discoveries generated throughout this doctoral thesis should be taken as a filtering step to prioritize interesting putative drug targets. It is crucial to validate the theoretical findings through experimental methods, which ensure the reliability and robustness of the identified associations. To broaden the scope and applicability of the presented research, the investigation should be expanded to include additional pairs of diseases. Examining a diverse range of disease combinations helps in identifying specific common genetic links across various health conditions.

Finally, achieving a more comprehensive understanding of the shared genetic etiology between correlated complex health conditions involves integrating data beyond genetic factors. Incorporating information on medications, imaging results, and environmental influences can provide a holistic view of the complex interactions underlying frequently co-occurring pairs of diseases. This integrated approach can contribute to a more nuanced analysis of the interconnected factors influencing disease associations, ultimately impacting personalized medicine.

Conclusion

In conclusion, this doctoral thesis addressed multimorbidity using large-scale genomic data to conduct an in-depth exploration of the shared genetic etiology between type 2 diabetes and osteoarthritis, and type 2 diabetes and schizophrenia. Valuable insights into disease biology, putative effector genes, shared biological pathways and causal relationships have been uncovered. The findings presented here might advance personalized medicine by paving the way for the identification of novel therapeutic targets and drug repurposing opportunities. Moving forward, it is essential to validate the theoretical findings through experimental methods and to expand the investigation to integrate data beyond genetic factors, including medications and environmental influences.

Post considerations

Amidst the backdrop of the COVID-19 pandemic, my first conference talk in October 2021 at the American Society of Human Genetics conference, took on an unconventional form—it was pre-recorded, and the questions and answers session had no public, just faceless voices. Throughout my first year of doctoral studies, all conferences that I participated in were conducted exclusively in

virtual settings. During this period, remote work became the norm, initially mandatory and later adopted as optional. While this might have led to less interactions within the group at times, it also allowed for high flexibility, which I made use of extensively compared to the first year of my doctoral studies. For the remainder of my studies, traveling to conferences and visiting collaborators was an essential part, which was highly encouraged by my supervisor Eleftheria Zeggini. I started a Ph.D. in data science as a natural progression of my academic trajectory. In the fast-paced and exciting research environment of the Zeggini lab, I have re-encountered a long-lost childhood excitement towards human genetics. I am finishing this chapter in love with the field of statistical genomics and very much looking forward to my next chapters in this research area.

Summary of peer-reviewed publications

Genetic underpinning of the comorbidity between type 2 diabetes and osteoarthritis

A common pattern of multimorbidity is the combination of cardiometabolic and osteoarticular diseases, exemplified by the co-occurrence of type 2 diabetes and osteoarthritis. Osteoarthritis is a whole-joint degenerative disorder that affects over 520 million people worldwide. Type 2 diabetes affects over 430 million people globally and is marked by high blood glucose levels and insulin resistance. Epidemiological studies have shown a positive association between these complex diseases that share risk factors such as increased BMI, which is causally linked to both.

Leveraging large-scale GWAS data, we showed a significant positive genetic correlation between type 2 diabetes and osteoarthritis, which was stronger for the knee compared to the hip. Mendelian randomization analyses showed no evidence of causality between both diseases. Using pairwise Bayesian colocalization analyses, we identified 18 unique genomic loci with a shared signal between type 2 diabetes and osteoarthritis. Integrating multi-omics data and functional information helped pinpoint 72 likely effector genes involved in both diseases, with 19 of these being defined as high-confidence effector genes. To identify potential drug repurposing opportunities, we explored the druggability of these effector genes, finding that 16 out of the 72 genes are part of the druggable genome, including six tier 1 druggable genes already targeted by existing or developing drugs.

We further explored the role of obesity in the studied comorbidity. We found evidence of a causal relationship between adiposity measures and nine high-confidence effector genes. For type 2 diabetes, BMI-associated variants influencing genes expressed in brain tissue showed a stronger impact than those in adipose tissue, a trend also seen in knee osteoarthritis. However, in hip

osteoarthritis, variants associated with gene expression in adipose tissue had a stronger effect, suggesting different underlying biological processes.

Specific high-confidence effector genes were highlighted for their roles in pathways related to type 2 diabetes and osteoarthritis, with some demonstrating opposing causal direction of effect on these diseases. Based on these genes, we highlighted three potential biological mechanisms underpinning the studied comorbidity: obesity, imbalance between osteoblasts and adipocytes differentiation in adult bone marrow and the Wnt/β-catenin signalling pathway.

Together with Eleftheria Zeggini, Ana Arruda has conceptualized the project, including the research questions and analysis steps needed to tackle these. Ana Arruda has performed most analyses and has written the initial draft of the final manuscript, which was circulated to co-authors for comments on the structure, insights, and additional analyses.

Genomic insights into the comorbidity between type 2 diabetes and schizophrenia

Individuals with mental health disorders face an increased risk of multimorbid physical health conditions, impacting life quality and leading to premature death. Here, we have studied the comorbidity between type 2 diabetes and schizophrenia, two co-occurring conditions. Type 2 diabetes, characterized by elevated glucose levels, affects over 536 million people globally, with an estimated heritability of ~50%. Schizophrenia, a major psychiatric disorder, with ~1% global prevalence, has an estimated general heritability of ~80%. Observational studies indicate a positive association between type 2 diabetes and schizophrenia, influenced by sociodemographic factors and antipsychotic medication. Genetic studies suggest a partial shared genetic basis between both conditions.

Using genome-wide data from large-scale GWAS, we investigated the genetic correlation and potential causal relationship between type 2 diabetes and schizophrenia. Despite the positive epidemiological correlation between the two conditions, we find evidence for a negative genetic correlation and no evidence of a causal relationship. Mendelian randomization analyses revealed a protective effect of schizophrenia liability on BMI, contrasting the well-established causal effect of BMI on type 2 diabetes. Further causal inference analyses indicated adulthood but not childhood BMI had a potential protective effect on schizophrenia in both univariate and multivariate analyses.

We identified 11 genomic loci with evidence of shared genetic signals between type 2 diabetes and schizophrenia by performing colocalization analysis. To resolve the identified colocalizing signals, we

incorporated multi-omics and functional biology information and prioritized 15 potential effector genes showing involvement in both conditions. These genes were enriched in pathways associated with lipids and metabolic regulation. Among the 15 identified genes showing evidence of involvement in both diseases, five are part of the druggable genome and four of these are tier 1 druggable targets. The highest scoring genes were *EGR2*, *LAMA4*, and *NUS1*. Our results suggest common genetic pathways underlying both conditions but acting in opposite directions.

The project conceptualization including formulating research questions and delineating the necessary initial analysis steps was conducted by Ana Arruda and Eleftheria Zeggini. Ana Arruda carried out all analyses and wrote the initial draft of the manuscript, which was reviewed by Eleftheria Zeggini. Subsequently, the reviewed draft was shared with co-authors for their input and feedback on the manuscript's organization, insights, and potential supplementary analyses.

List of abbreviations

BMI = body mass index

GWAS = genome-wide association study

T2DGGI = Type 2 Diabetes Global Genomics Initiative

GEMINI = Genetic Evaluation of Multimorbidity towards INdividualisation of Interventions

MMTRC = Multimorbidity Mechanism and Therapeutics Research Collaborative

LD = linkage disequilibrium

QTL = quantitative trait locus

MR = Mendelian randomization

IVW = inverse variance weighted

PP4 = posterior probability of H4

References

1. Montana, G. Statistical methods in genetics. *Briefings in Bioinformatics* **7**, 297-308 (2006). doi: 10.1093/bib/bbl028.
2. Lappalainen, T., Li, Y.I., Ramachandran, S. & Gusev, A. Genetic and molecular architecture of complex traits. *Cell* **187**, 1059-1075 (2024). doi: https://doi.org/10.1016/j.cell.2024.01.023.
3. Ghoussaini, M., Nelson, M.R. & Dunham, I. Future prospects for human genetics and genomics in drug discovery. *Current Opinion in Structural Biology* **80**, 102568 (2023). doi: https://doi.org/10.1016/j.sbi.2023.102568.
4. Rusina, P.V. *et al.* Genetic support for FDA-approved drugs over the past decade. *Nat Rev Drug Discov* **22**, 864 (2023). doi: 10.1038/d41573-023-00158-x.

5. Cirillo, D. & Valencia, A. Big data analytics for personalized medicine. *Current Opinion in Biotechnology* **58**, 161-167 (2019). doi: https://doi.org/10.1016/j.copbio.2019.03.004.
6. International Human Genome Sequencing, C. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945 (2004). doi: 10.1038/nature03001.
7. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44-53 (2022). doi: doi:10.1126/science.abj6987.
8. Qian, D. High performance computing: a brief review and prospects. *National Science Review* **3**, 16-16 (2016). doi: 10.1093/nsr/nww009.
9. Sella, G. & Barton, N.H. Thinking About the Evolution of Complex Traits in the Era of Genome-Wide Association Studies. *Annual Review of Genomics and Human Genetics* **20**, 461-493 (2019). doi: 10.1146/annurev-genom-083115-022316.
10. Fisher, R.A. XV.—The correlation between relatives on the supposition of Mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh* **52**, 399-433 (1919). doi.
11. Boyle, E.A., Li, Y.I. & Pritchard, J.K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177-1186 (2017). doi: 10.1016/j.cell.2017.05.038.
12. Uffelmann, E. *et al.* Genome-wide association studies. *Nature Reviews Methods Primers* **1**, 59 (2021). doi: 10.1038/s43586-021-00056-9.
13. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature 2018 562:7726* **562**, 203-209 (2018). doi: 10.1038/s41586-018-0579-z.
14. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *J Epidemiol* **27**, S2-s8 (2017). doi: 10.1016/j.je.2016.12.005.
15. Kurki, M.I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508-518 (2023). doi: 10.1038/s41586-022-05473-8.
16. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Research* **51**, D977-D985 (2022). doi: 10.1093/nar/gkac1010.
17. Cano-Gamez, E. & Trynka, G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. Vol. 11 424-424 (Frontiers Media S.A., 2020).
18. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nature Genetics* **51**, 1339-1348 (2019). doi: 10.1038/s41588-019-0481-0.
19. Gratten, J. & Visscher, P.M. Genetic pleiotropy in complex traits and diseases: implications for genomic medicine. *Genome Medicine* **8**, 78 (2016). doi: 10.1186/s13073-016-0332-x.
20. Hackinger, S. & Zeggini, E. Statistical methods to detect pleiotropy in human complex traits. Vol. 7 (Royal Society Publishing, 2017).
21. DeFronzo, R.A. *et al.* Type 2 diabetes mellitus. *Nature Reviews Disease Primers* **1**, 15019 (2015). doi: 10.1038/nrdp.2015.19.
22. Galicia-Garcia, U. *et al.* Pathophysiology of Type 2 Diabetes Mellitus. *Int J Mol Sci* **21**(2020). doi: 10.3390/ijms21176275.
23. Sun, H. *et al.* IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Research and Clinical Practice* **183**(2022). doi: 10.1016/j.diabres.2021.109119.
24. Pearson-Stuttard, J. *et al.* Variations in comorbidity burden in people with type 2 diabetes over disease duration: A population-based analysis of real world evidence. *eClinicalMedicine* **52**, 101584-101584 (2022). doi: 10.1016/j.eclinm.2022.101584.
25. Avery, A.R. & Duncan, G.E. Heritability of Type 2 Diabetes in the Washington State Twin Registry. *Twin Res Hum Genet* **22**, 95-98 (2019). doi: 10.1017/thg.2019.11.
26. Huxley, R. *et al.* Ethnic comparisons of the cross-sectional relationships between measures of body size with diabetes and hypertension. *Obes Rev* **9 Suppl 1**, 53-61 (2008). doi: 10.1111/j.1467-789X.2007.00439.x.
27. Suzuki, K. *et al.* Genetic drivers of heterogeneity in type 2 diabetes pathophysiology. *Nature* (2024). doi: 10.1038/s41586-024-07019-6.

28. Steinmetz, J.D. *et al.* Global, regional, and national burden of osteoarthritis, 1990–2020 and projections to 2050: a systematic analysis for the Global Burden of Disease Study 2021. *The Lancet Rheumatology* **5**, e508-e522 (2023). doi: 10.1016/S2665-9913(23)00163-7.
29. Martel-Pelletier, J. *et al.* Osteoarthritis. *Nature Reviews Disease Primers 2016 2:1* **2**, 1-18 (2016). doi: 10.1038/nrdp.2016.72.
30. Hunter, D.J. & Bierma-Zeinstra, S. Osteoarthritis. *The Lancet* **393**, 1745-1759 (2019). doi: 10.1016/S0140-6736(19)30417-9.
31. Swain, S., Sarmanova, A., Coupland, C., Doherty, M. & Zhang, W. Comorbidities in Osteoarthritis: A Systematic Review and Meta-Analysis of Observational Studies. *Arthritis Care Res (Hoboken)* **72**, 991-1000 (2020). doi: 10.1002/acr.24008.
32. Aubourg, G., Rice, S.J., Bruce-Wootton, P. & Loughlin, J. Genetics of osteoarthritis. *Osteoarthritis Cartilage* **30**, 636-649 (2022). doi: 10.1016/j.joca.2021.03.002.
33. Spector, T.D. & MacGregor, A.J. Risk factors for osteoarthritis: genetics. *Osteoarthritis Cartilage* **12 Suppl A**, S39-44 (2004). doi: 10.1016/j.joca.2003.09.005.
34. Boer, C.G. *et al.* Deciphering osteoarthritis genetics across 826,690 individuals from 9 populations. *Cell* (2021). doi: 10.1016/J.CELL.2021.07.038.
35. Owen, M.J., Sawa, A. & Mortensen, P.B. Schizophrenia. *The Lancet* **388**, 86-97 (2016). doi: 10.1016/S0140-6736(15)01121-6.
36. Charlson, F.J. *et al.* Global Epidemiology and Burden of Schizophrenia: Findings From the Global Burden of Disease Study 2016. *Schizophr Bull* **44**, 1195-1203 (2018). doi: 10.1093/schbul/sby058.
37. Vos, T. *et al.* Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet* **390**, 1211-1259 (2017). doi: 10.1016/S0140-6736(17)32154-2.
38. Correll, C.U. *et al.* Prevalence, incidence and mortality from cardiovascular disease in patients with pooled and specific severe mental illness: a large-scale meta-analysis of 3,211,768 patients and 113,383,368 controls. *World Psychiatry* **16**, 163-180 (2017). doi: https://doi.org/10.1002/wps.20420.
39. Vancampfort, D. *et al.* Risk of metabolic syndrome and its components in people with schizophrenia and related psychotic disorders, bipolar disorder and major depressive disorder: a systematic review and meta-analysis. *World Psychiatry* **14**, 339-347 (2015). doi: https://doi.org/10.1002/wps.20252.
40. Trubetskoy, V. *et al.* Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature 2022 604:7906* **604**, 502-508 (2022). doi: 10.1038/s41586-022-04434-5.
41. Chowdhury, S.R., Chandra Das, D., Sunna, T.C., Beyene, J. & Hossain, A. Global and regional prevalence of multimorbidity in the adult population in community settings: a systematic review and meta-analysis. *eClinicalMedicine* **57**(2023). doi: 10.1016/j.eclinm.2023.101860.
42. Skou, S.T. *et al.* Multimorbidity. *Nature Reviews Disease Primers* **8**, 48 (2022). doi: 10.1038/s41572-022-00376-4.
43. Marengoni, A. *et al.* Aging with multimorbidity: A systematic review of the literature. *Ageing Research Reviews* **10**, 430-439 (2011). doi: https://doi.org/10.1016/j.arr.2011.03.003.
44. Kivimäki, M. *et al.* Body-mass index and risk of obesity-related complex multimorbidity: an observational multicohort study. *Lancet Diabetes Endocrinol* **10**, 253-263 (2022). doi: 10.1016/s2213-8587(22)00033-x.
45. Delpino, F.M. *et al.* Overweight, obesity and risk of multimorbidity: A systematic review and meta-analysis of longitudinal studies. *Obes Rev* **24**, e13562 (2023). doi: 10.1111/obr.13562.
46. Rodrigues, A.P.d.S. *et al.* Multimorbidity and complex multimorbidity in Brazilians with severe obesity. *Scientific Reports* **13**, 16629 (2023). doi: 10.1038/s41598-023-43545-5.

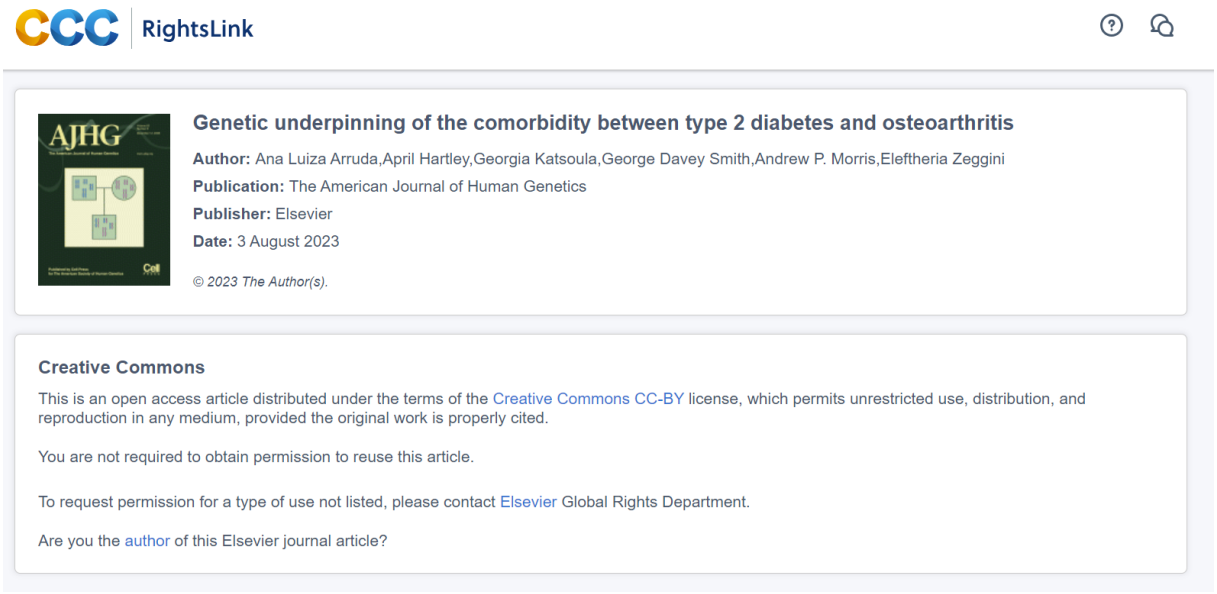
47. Whitty, C.J.M. *et al.* Rising to the challenge of multimorbidity. Vol. 368 (BMJ Publishing Group, 2020).
48. Tran, P.B. *et al.* Costs of multimorbidity: a systematic review and meta-analyses. *BMC Medicine* **20**, 234 (2022). doi: 10.1186/s12916-022-02427-9.
49. Salisbury, C. Multimorbidity: redesigning health care for people who use it. *Lancet* **380**, 7-9 (2012). doi: 10.1016/s0140-6736(12)60482-6.
50. Mair, F.S. & May, C.R. Thinking about the burden of treatment. *Bmj* **349**, g6680 (2014). doi: 10.1136/bmj.g6680.
51. Dong, G., Feng, J., Sun, F., Chen, J. & Zhao, X.-M. A global overview of genetically interpretable multimorbidities among common diseases in the UK Biobank. *Genome Medicine* **13**, 110 (2021). doi: 10.1186/s13073-021-00927-6.
52. Kuan, V. *et al.* Identifying and visualising multimorbidity and comorbidity patterns in patients in the English National Health Service: a population-based study. *The Lancet Digital Health* **5**, e16-e27 (2023). doi: 10.1016/S2589-7500(22)00187-X.
53. Schiøtz, M.L., Stockmarr, A., Høst, D., Glümer, C. & Frølich, A. Social disparities in the prevalence of multimorbidity – A register-based population study. *BMC Public Health* **17**, 422 (2017). doi: 10.1186/s12889-017-4314-8.
54. Alshakhs, M., Jackson, B., Ikponmwosa, D., Reynolds, R. & Madlock-Brown, C. Multimorbidity patterns across race/ethnicity as stratified by age and obesity. *Scientific Reports* **12**, 9716 (2022). doi: 10.1038/s41598-022-13733-w.
55. Elgart, M. *et al.* Correlations between complex human phenotypes vary by genetic background, gender, and environment. *Cell Rep Med* **3**, 100844 (2022). doi: 10.1016/j.xcrm.2022.100844.
56. Peipert, J.F. & Phipps, M.G. Observational Studies. *Clinical Obstetrics and Gynecology* **41**(1998). doi.
57. Bezerra de Souza, D.L. *et al.* Multimorbidity and its associated factors among adults aged 50 and over: A cross-sectional study in 17 European countries. *PLoS One* **16**, e0246623 (2021). doi: 10.1371/journal.pone.0246623.
58. Deste, G. & Lombardi, C.M. Editorial: Cardiometabolic disease and psychiatric disorders. *Frontiers in Psychiatry* **14**(2023). doi: 10.3389/fpsyt.2023.1174055.
59. Williams, M.F., London, D.A., Husni, E.M., Navaneethan, S. & Kashyap, S.R. Type 2 diabetes and osteoarthritis: A systematic review and meta-analysis. Vol. 30 944-950 (Elsevier Inc., 2016).
60. Ward, M. & Druss, B. The epidemiology of diabetes in psychotic disorders. *The Lancet Psychiatry* **2**, 431-451 (2015). doi: 10.1016/S2215-0366(15)00007-3.
61. Mountjoy, E. *et al.* An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nature Genetics* **53**, 1527-1533 (2021). doi: 10.1038/s41588-021-00945-5.
62. Duffy, Á. *et al.* Development of a human genetics-guided priority score for 19,365 genes and 399 drug indications. *Nature Genetics* **56**, 51-59 (2024). doi: 10.1038/s41588-023-01609-2.
63. Melzer, D., Pilling, L.C. & Ferrucci, L. The genetics of human ageing. *Nature Reviews Genetics* **21**, 88-101 (2020). doi: 10.1038/s41576-019-0183-6.
64. West, C.E. *et al.* Integrative GWAS and co-localisation analysis suggests novel genes associated with age-related multimorbidity. *Scientific Data* **10**, 655 (2023). doi: 10.1038/s41597-023-02513-4.
65. Romero, C. *et al.* Exploring the genetic overlap between twelve psychiatric disorders. *Nature Genetics* **54**, 1795-1802 (2022). doi: 10.1038/s41588-022-01245-2.
66. Grotzinger, A.D. Shared genetic architecture across psychiatric disorders. *Psychological Medicine* **51**, 2210-2216 (2021). doi: 10.1017/S0033291721000829.
67. Wingo, T.S. *et al.* Shared mechanisms across the major psychiatric and neurodegenerative diseases. *Nature Communications* **13**, 4314 (2022). doi: 10.1038/s41467-022-31873-5.

68. Lee, P.H. *et al.* Genomic Relationships, Novel Loci, and Pleiotropic Mechanisms across Eight Psychiatric Disorders. *Cell* **179**, 1469-1482.e11 (2019). doi: 10.1016/j.cell.2019.11.020.
69. Solberg, B.S. *et al.* Patterns of Psychiatric Comorbidity and Genetic Correlations Provide New Insights Into Differences Between Attention-Deficit/Hyperactivity Disorder and Autism Spectrum Disorder. *Biological Psychiatry* **86**, 587-598 (2019). doi: 10.1016/j.biopsych.2019.04.021.
70. Harroud, A. & Hafler, D.A. Common genetic factors among autoimmune diseases. *Science* **380**, 485-490 (2023). doi: doi:10.1126/science.adg2992.
71. Farh, K.K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337-343 (2015). doi: 10.1038/nature13835.
72. Kreiner, E. *et al.* Shared genetic variants suggest common pathways in allergy and autoimmune diseases. *Journal of Allergy and Clinical Immunology* **140**, 771-781 (2017). doi: https://doi.org/10.1016/j.jaci.2016.10.055.
73. Yuya, S. *et al.* Multi-trait and cross-population genome-wide association studies across autoimmune and allergic diseases identify shared and distinct genetic component. *Annals of the Rheumatic Diseases* **81**, 1301 (2022). doi: 10.1136/annrheumdis-2022-222460.
74. Baltramonaityte, V. *et al.* A multivariate genome-wide association study of psycho-cardiometabolic multimorbidity. *PLOS Genetics* **19**, e1010508 (2023). doi: 10.1371/journal.pgen.1010508.
75. Hagenaars, S.P. *et al.* Genetic comorbidity between major depression and cardio-metabolic traits, stratified by age at onset of major depression. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **183**, 309-330 (2020). doi: https://doi.org/10.1002/ajmg.b.32807.
76. Tai, X.Y. *et al.* Cardiometabolic multimorbidity, genetic risk, and dementia: a prospective cohort study. *The Lancet Healthy Longevity* **3**, e428-e436 (2022). doi: 10.1016/S2666-7568(22)00117-9.
77. Rahmioglu, N. *et al.* The genetic basis of endometriosis and comorbidity with other pain and inflammatory conditions. *Nature Genetics* **55**, 423-436 (2023). doi: 10.1038/s41588-023-01323-z.
78. McGrath, I.M., Montgomery, G.W. & Mortlock, S. Insights from Mendelian randomization and genetic correlation analyses into the relationship between endometriosis and its comorbidities. *Human Reproduction Update* **29**, 655-674 (2023). doi: 10.1093/humupd/dmad009.
79. Zeng, R. *et al.* Dissecting shared genetic architecture between obesity and multiple sclerosis. *eBioMedicine* **93**(2023). doi: 10.1016/j.ebiom.2023.104647.
80. Zhang, F., Rao, S. & Baranova, A. Shared genetic liability between major depressive disorder and osteoarthritis. *Bone & Joint Research* **11**, 12-22 (2022). doi: 10.1302/2046-3758.111.Bjr-2021-0277.R1.
81. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nature Genetics 2015 47:11* **47**, 1236-1241 (2015). doi: 10.1038/ng.3406.
82. Sanderson, E. *et al.* Mendelian randomization. *Nature Reviews Methods Primers 2022 2:1* **2**, 1-21 (2022). doi: 10.1038/s43586-021-00092-5.
83. Leyden, G.M. *et al.* Harnessing tissue-specific genetic variation to dissect putative causal pathways between body mass index and cardiometabolic phenotypes. *The American Journal of Human Genetics* **109**, 240-252 (2022). doi: 10.1016/j.ajhg.2021.12.013.
84. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genetics* **10**, e1004383-e1004383 (2014). doi: 10.1371/journal.pgen.1004383.
85. Noonan, J.P. & McCallion, A.S. Genomics of Long-Range Regulatory Elements. *Annual Review of Genomics and Human Genetics* **11**, 1-23 (2010). doi: 10.1146/annurev-genom-082509-141651.

86. Stanzick, K.J. *et al.* Discovery and prioritization of variants and genes for kidney function in >1.2 million individuals. *Nature Communications* **12**, 4350 (2021). doi: 10.1038/s41467-021-24491-0.
87. Kamburov, A. *et al.* ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Research* **39**, D712-D717 (2011). doi: 10.1093/NAR/GKQ1156.
88. Finan, C. *et al.* The druggable genome and support for target identification and validation in drug development. *Science Translational Medicine* **9**(2017). doi: 10.1126/SCITRANSLMED.AAG1166/SUPPL_FILE/AAG1166_TABLE_S1.ZIP.
89. Saadullah Khani, N. *et al.* Schizophrenia and cardiometabolic abnormalities: A Mendelian randomization study. *Front Genet* **14**, 1150458 (2023). doi: 10.3389/fgene.2023.1150458.
90. Chen, W. *et al.* Mendelian randomization analyses identify bidirectional causal relationships of obesity with psychiatric disorders. *Journal of Affective Disorders* **339**, 807-814 (2023). doi: https://doi.org/10.1016/j.jad.2023.07.044.
91. Gurdasani, D., Barroso, I., Zeggini, E. & Sandhu, M.S. Genomics of disease risk in globally diverse populations. *Nat Rev Genet* **20**, 520-535 (2019). doi: 10.1038/s41576-019-0144-0.
92. Angelo, F., Veenstra, D., Knerr, S. & Devine, B. Prevalence and prediction of medical distrust in a diverse medical genomic research sample. *Genetics in Medicine* **24**, 1459-1467 (2022). doi: https://doi.org/10.1016/j.gim.2022.03.007.
93. Lyles, C.R., Lunn, M.R., Obedin-Maliver, J. & Bibbins-Domingo, K. The new era of precision population health: insights for the All of Us Research Program and beyond. *Journal of Translational Medicine* **16**, 211 (2018). doi: 10.1186/s12967-018-1585-5.
94. Mulder, N. *et al.* H3Africa: current perspectives. *Pharmgenomics Pers Med* **11**, 59-66 (2018). doi: 10.2147/pgpm.S141546.
95. Zhou, W. *et al.* Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. *Cell Genomics* **2**, 100192 (2022). doi: https://doi.org/10.1016/j.xgen.2022.100192.
96. Yang, G., Mishra, M. & Perera, M.A. Multi-Omics Studies in Historically Excluded Populations: The Road to Equity. *Clin Pharmacol Ther* **113**, 541-556 (2023). doi: 10.1002/cpt.2818.
97. Arruda, A.L., Morris, A.P. & Zeggini, E. Advancing equity in human genomics through tissue-specific multi-ancestry molecular data. *Cell Genom*, 100485 (2024). doi: 10.1016/j.xgen.2023.100485.

Appendix with peer-reviewed publications

1. Genetic underpinning of the comorbidity between type 2 diabetes and osteoarthritis



RightsLink

Genetic underpinning of the comorbidity between type 2 diabetes and osteoarthritis

Author: Ana Luiza Arruda,April Hartley,Georgia Katsoula,George Davey Smith,Andrew P. Morris,Eleftheria Zeggini
Publication: The American Journal of Human Genetics
Publisher: Elsevier
Date: 3 August 2023

© 2023 The Author(s).

Creative Commons

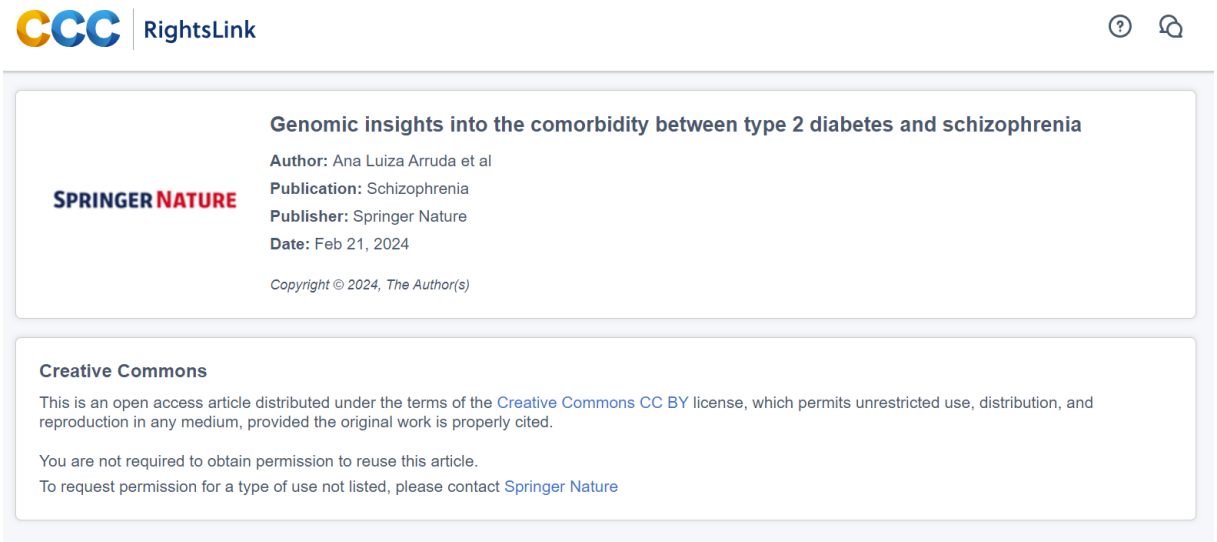
This is an open access article distributed under the terms of the Creative Commons CC-BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.

To request permission for a type of use not listed, please contact Elsevier Global Rights Department.

Are you the author of this Elsevier journal article?

2. Genomic insights into the comorbidity between type 2 diabetes and schizophrenia



RightsLink

SPRINGER NATURE

Genomic insights into the comorbidity between type 2 diabetes and schizophrenia

Author: Ana Luiza Arruda et al
Publication: Schizophrenia
Publisher: Springer Nature
Date: Feb 21, 2024

Copyright © 2024, The Author(s)

Creative Commons

This is an open access article distributed under the terms of the Creative Commons CC BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.
To request permission for a type of use not listed, please contact Springer Nature