



Translating genetic association of lipid levels for biological and clinical application

Bradley Crone¹ · Amelia M. Krause² · Whitney E. Hornsby² · Cristen J. Willer^{1,2,3} · Ida Surakka²

Accepted: 9 February 2021 / Published online: 19 February 2021
© Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract

Purpose of review This review focuses on the foundational evidence from the last two decades of lipid genetics research and describes the current status of data-driven approaches for transethnic GWAS, fine-mapping, transcriptome informed fine-mapping, and disease prediction.

Recent findings Current lipid genetics research aims to understand the association mechanisms and clinical relevance of lipid loci as well as to capture population specific associations found in global ancestries. Recent genome-wide trans-ethnic association meta-analyses have identified 118 novel lipid loci reaching genome-wide significance. Gene-based burden tests of whole exome sequencing data have identified three genes—*PCSK9*, *LDLR*, and *APOB*—with significant rare variant burden associated with familial dyslipidemia. Transcriptome-wide association studies discovered five previously unreported lipid-associated loci. Additionally, the predictive power of genome-wide genetic risk scores amalgamating the polygenic determinants of lipid levels can potentially be used to increase the accuracy of coronary artery disease prediction.

Conclusions Lipids are one of the most successful group of traits in the era of genome-wide genetic discovery for identification of novel loci and plausible drug targets. However, a substantial fraction of lipid trait heritability remains unexplained. Further analysis of diverse ancestries and state of the art methods for association locus refinement could potentially reveal some of this missing heritability and increase the clinical application of the genomic association results.

Keywords Common variants · Exome sequencing · Genome-wide association scan · Rare variants · Drug targets · Prediction

Introduction

Genome-wide association scans (GWAS) and meta-analyses combining information from multiple GWAS datasets have successfully identified common DNA sequence variants (single nucleotide polymorphisms, SNPs) associated with diseases, quantitative traits, and complex phenotypes [1–5]. The number of participants represented in meta-analyses have

increased at an exponential rate since their introduction [6], with recent datasets in atrial fibrillation including >1,000,000 participants [7], although the largest meta-analysis in lipids is among 300,000 multiethnic participants [8] and the largest single cohort study exceeds 390,000 participants [9]. Meta-analyses have expanded our knowledge of specific genes and pathways influencing lipid levels due to the highly polygenic heritability pattern of lipid levels shown by hundreds of associated loci to date [10]. While most of the variation in lipid levels within the general population is due to polygenic variation, single protein-altering variants in known lipid genes can confer extreme lipid levels, generally referred to as dyslipidemias when observed in patients [11].

Technological advances in DNA sequencing have made the interrogation of protein-coding regions of the genome (the exome) more broadly utilized. Exome sequencing has been useful at identifying Mendelian forms of disease [12, 13], although its utility is limited with complex human phenotypes, including lipids, due to the expected low frequency of high impact mutations in the general population and higher costs with concomitant lower sample sizes in sequencing

This article belongs to the Topical Collection: Translating genome medicine to treatments

✉ Ida Surakka
isurakka@umich.edu

¹ Department of Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI, USA

² Department of Internal Medicine, Division of Cardiovascular Medicine, University of Michigan, Michigan Medicine, Ann Arbor, MI, USA

³ Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA

studies relative to other technologies. Several previous reviews of lipid genomics have been published, including in 2015 [14] and 2018 [15] which highlight lipid gene identification using GWAS, exome sequencing approaches, and emphasizing a data-driven approach to therapeutic drug target development. This review focuses on the foundational evidence from the last two decades of lipid genetics, while also illustrating the current status of recent computational approaches for transethnic GWAS, fine-mapping, transcriptome informed fine-mapping, and disease prediction (Fig. 1). Novel genetic insights derived from these methods may provide new plausible candidate genes for drug development, empower disease prediction for earlier identification of high-risk individuals, inform clinical practice for preventative health care, and suggest directions for future research of population level lipid variation in diverse populations.

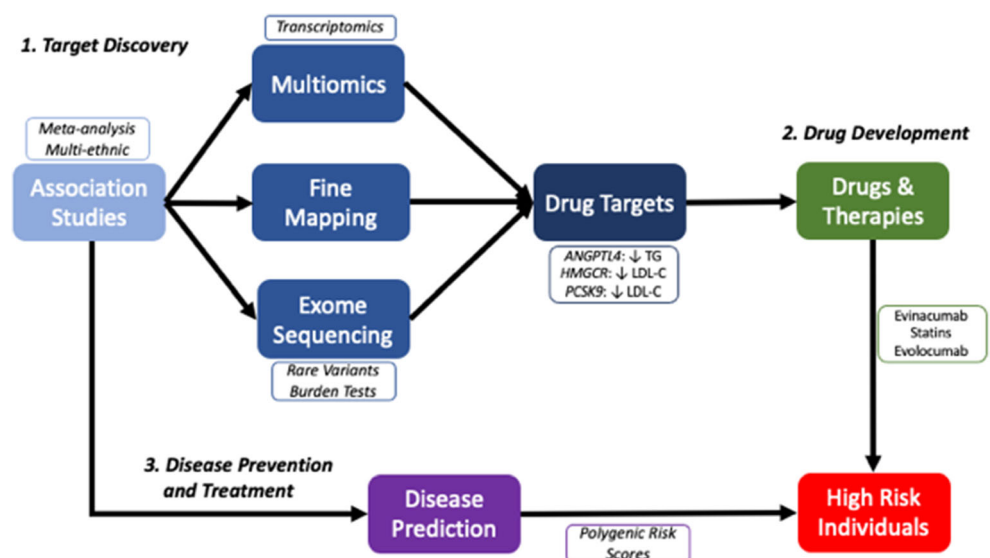
Single variant association discovery

GWAS meta-analyses in large sample sizes capture common variants with small to moderate genetic effects due to enhanced statistical power [16]. In 2010, the Global Lipids Genetics Consortium (GLGC) [2] meta-analyzed 46 lipid GWASs compiling over 100,000 participants of European ancestry. Findings revealed 59 previously unreported genome-wide significant loci across four lipid traits of total cholesterol (TC), low-density lipoprotein cholesterol (LDL-C), high density lipoprotein cholesterol (HDL-C), and triglycerides (TG). A follow-up association test of the meta-analysis lead SNPs in Europeans with coronary artery disease (CAD, $n = 24,607$) and without CAD ($n = 66,197$) identified four loci (*IRS1*, *C6orf106*, *KLF14*, and *NAT2*) significantly associated with CAD and decreased HDL-C and increased TG levels,

indicating potential targets for preventative therapies. Quantitatively combining multiple GWAS cohorts can uncover associated signals not detected by single cohort GWASs, including variants with potential causal effects and candidate loci for disease preventative drugs and therapies.

GWAS meta-analyses within cohorts of the same ancestry have proven to be beneficial in the discovery of common mutations with modest contributions to trait heritability. However, GWAS restricted to single or closely related ancestries only identifies a subset of causative variants. GWAS meta-analyses that target a single ancestry often fail to capture low frequency variants specific to other ancestry groups, which may have significant contributions to complex trait heritability [17], or may point to new therapeutic targets. A recent trans-ethnic analysis of blood lipids and associated traits was conducted on participants across three distinct ancestries: non-Hispanic whites, ($n \sim 216,000$), non-Hispanic blacks ($n \sim 57,000$), and Hispanics ($n \sim 24,000$) from the Million Veterans Project (MVP) [8, 18]. GWAS was conducted within each ancestry cohort and then combined through an inverse variance-weighted fixed effects meta-analysis. The inverse variance-weighted fixed effect approach assumes all studies in the meta-analysis share the same true effect size and minimizes effect size variance by calculating the mean effect size across the GWASs weighted by the inverse variance of each single cohort GWAS. The meta-analysis identified over 46,000 genome-wide significant variants across 188 loci [19]. Subsequent replication in GLGC followed by conditional analysis combining both MVP and GLGC summary statistics for each lipid trait revealed 118 novel loci meeting genome-wide significance. These results demonstrate the benefits of performing a transethnic meta-analysis to isolate trait-specific loci [8].

Fig. 1 A workflow for drug discovery. This diagram demonstrates a general workflow for progressing from variant-trait associations to drugs and therapies. Under (1) target discovery, GWA loci are refined through multiomics approaches, genetic fine-mapping, and WES rare variant analysis. The resulting loci represent potential targets for (2) drug development. Once developed, identifying at-risk individuals for (3) disease prevention and treatment through polygenic risk scores ensures drugs and therapies are administered to the individuals at highest disease risk



Gene-based association discovery

Single cohort and GWAS meta-analysis are successful at identifying common variants with small to moderate effects on disease risk or quantitative traits. However, these approaches fail to capture moderate to large effect rare coding mutations that are integral to explaining the heritability of both Mendelian and complex traits [20]. Whole genome arrays designed to capture common variation are unlikely to include these rare mutations, and imputation from reference population samples rarely has enough occurrences of these mutations for confident imputation of these variants in the GWAS datasets.

Whole exome sequencing is a critical tool for discovering rare trait-associated variants within coding regions of the genome otherwise missed by GWAS arrays and variant imputation. However, traditional single variant tests are generally underpowered to accurately capture rare variant-trait associations. Either large sample sizes or large effect sizes are required for a single variant test to detect rare associations with sufficient power [21]. The former can be difficult to achieve for understudied traits and populations. The latter is typically not observed in polygenic complex traits, where rare variants carry a moderate burden of trait heritability. Unlike familial studies of Mendelian traits where a single mutation is typically implicated in disease inheritance, whole exome sequencing studies of complex traits may find clusters of mutations within a gene that are each associated with a phenotype. To circumvent the issue of insufficient statistical power for rare coding variants, gene-based burden tests are employed to collapse rare variant counts to a single gene across samples. Combining allele counts of low frequency mutations within a gene improves the power of associating trait within a given gene than with single variant testing of rare variants.

Burden tests are particularly powerful for genes with allelic heterogeneity, where a greater proportion of alleles in a gene are causative for the trait or disease [22]. The NHLBI Exome Sequencing Project leveraged whole exome sequencing samples from >2000 participants, including cohorts of extremely high ($n = 307$) and extremely low ($n = 247$) LDL-C levels, to identify rare variants within LDL-C-associated genes [23]. Gene-based burden tests across various allele frequency thresholds—from ultra-rare ($MAF \leq 0.1\%$) to more common ($MAF \leq 5\%$)—and different groupings of mutations based on variant effect classifications including known and predicted missense and loss of function mutations, revealed significant association between three well known lipid genes—*PCSK9*, *LDLR*, and *APOB*—and LDL-C levels. This is in comparison to single-variant tests of common mutations under the same study revealing only significant association with *APOE*. Collapsing rare and potential damaging variants under one gene signal enables discovery of associated genes otherwise missed by traditional single variant tests. The identification of *PCSK9*, *LDLR*, and *APOB* from gene-

based burden tests is validation of dyslipidemia family studies, serving as prime drug targets for lowering LDL-C levels in participants with dyslipidemia traits [24]. *PCSK9* is a target for heterozygous familial hypercholesterolemia treatments [25], whereas *LDLR* and *APOB* are targets for homozygous familial hypercholesterolemia therapies [26–28]. *PCSK9* also serves as a target for lowering atherosclerotic cardiovascular disease risk. This exemplifies how fine-tuned discovery of gene-trait associations can result in actionable drug targets for treatment and prevention (Table 1).

Refining single variant associations

GWAS meta-analyses have identified 167 [33, 34] lipid loci which however only count for approximately 20% [34, 35] of trait variation in the populations studied (less than 50% of the estimated trait heritability). It is hypothesized that the missing heritability may be due to the polygenic heritability pattern of lipid traits, suggesting there are still many genetic loci with small effect sizes to be found by increasing the sample size of the GWAS meta-analysis. Another hypothesis proposes that some of the lipid loci have multiple independent associations in close proximity, which are not considered by the standard approach of defining an associated locus or genetic signal based on physical distance [35]. Finally, some have hypothesized that trait heritability has been overestimated [36].

Fine mapping is one way to find additional associations and to pinpoint the causal genes in the established lipid loci. Fine mapping involves taking the local linkage disequilibrium into account and statistically estimating which variants are the most probable causal variants for the studied trait [37]. In a single cohort GWAS, the secondary signals can also be found using formal conditional analysis where the association test in a locus is adjusted for the lead-SNP association. There have been multiple studies testing different lipid loci showing secondary associations [34, 38, 39] in the coding region of the genome providing a direct link to the biological mechanism of the observed association, and therefore, illuminating potential drug targets.

Identifying relevant genes and pathways associated with noncoding mutations can be achieved with tools such as DEPICT [40] or the Polygenic Priority Score [41]. DEPICT assigns likely causal genes and enriched biological pathways for associated loci and highlights tissues and cell types where causal genes are highly expressed. Polygenic priority score identifies causal genes through integration of GWAS summary statistics with gene expression, biological pathway, and protein-protein interaction prediction data. The latter was successful in prioritizing over 8400 gene-trait associations across 113 complex traits with greater than 75% precision, including correctly identifying a previously discovered association between *SORT1* and LDL-C.

Table 1 Summary of known lipid drug targets validated by rare variant burden tests

Gene	Region	Study	Associated trait	Ancestry	Associated phenotype
PCSK9 [29, 30]	CHR1:55505221-55530525	NHLBI, GLGC	LDL-C, TC	AA, EUR, EAS	Familial hypercholesterolemia
LDLR [23]	CHR19:11200038-11244492	NHLBI, GLGC	LDL-C, TC	AA, EUR, EAS	Familial hypercholesterolemia
APOB [23]	CHR2:21224301-21266945	NHLBI, GLGC	LDL-C, TC	AA, EUR, EAS	Familial hypobetalipoproteinemia
APOC3 [31]	CHR11:116700422-116703788	Exome Sequencing Project	TG	AA, EUR	Familial triglyceridemia
APOA1 [32]	CHR11:116706467-116708666	Dallas Heart Study	HDL	AA, EUR	Familial hypoalphalipoproteinemia

NHLBI National Heart, Lung, and Blood Institute, *GLGC* Global Lipids Genetics Consortium, *HDL* high density lipoprotein cholesterol, *LDL-C* low density lipoprotein cholesterol, *TG* triglycerides, *TC* total cholesterol, *AA* African American ancestry, *EAS* East Asian ancestry, *EUR* European ancestry

It has also been suggested that using samples from different ancestries could identify the true causal variants underlying the association. Trans-ethnic meta-analysis of GWASs accounts for differences in linkage disequilibrium and heterogeneity of allelic effects and frequencies across diverse populations. Assuming a shared causal variant between ancestry groups, the surrounding variants in linkage disequilibrium with the causal variant may differ slightly between ancestries. By taking these slight differences into account in the meta-analysis with proper modeling, rather than excluding the minority ancestries in the analysis to avoid biases, scientists gain statistical power to identify the underlying putative causal variant. Utilization of diverse populations increases fine mapping resolution of the complex trait loci and further isolates the true genetic architecture of the underlying trait [42].

Epigenetic features play an important role in lipid genomics and understanding tissue-specific expression of lipid-associated genes. Recent efforts have been made to elucidate the function of long noncoding RNAs (lncRNAs). lncRNAs are transcribed RNA molecules greater than 200 nucleotides in length that do not encode for the protein. These serve a role in regulating gene transcription and posttranscription modifications and are largely tissue-specific in nature. In the case of lipid metabolism, lncRNA-mediated regulation colocalize primarily within liver and adipose tissues [43]. Previous reviews characterize the role of lncRNAs in cholesterol synthesis and metabolism [44] and diseases associated with lncRNA-mediated cholesterol dysregulation [45], including atherosclerosis, hypoalphalipoproteinemia (low LDL-C), myocardial infarction, and nonalcoholic fatty liver disease [46]. lncRNAs are prime drug and therapy targets because of their role in tissue-specific gene regulation. Other well-studied epigenetic features and their role in lipid-associated gene expression, such as DNA methylation, histone modification, and chromatin accessibility, are highlighted in other published reviews [47–49].

Machine learning and deep learning methods have indeed been implemented in predicting both deleterious coding mutations and prioritizing likely functional non-coding mutations. Predictive models such as CADD [50], PolyPhen [51], and SIFT [52] indicate a given mutation's impact on protein function. These models compile ancestral conservation data, epigenetic information, functional predictions (e.g., amino acid changes), and genetic content to predict the likelihood of deleterious mutations. Other models including RegulomeDB [53] and DeepSEA [54] highlight functional mutations in non-coding regions of the genome. These models compile data from chromatin profiles, transcription factor binding sites, and DNase hypersensitivity sites to predict the likelihood of functionally impactful mutations that affect the expression of target genes.

Refining associations with transcriptomics

Single cohort or GWAS meta-analysis identifies trait-causative loci; however, it is difficult to elucidate biological pathway effects from GWAS associations alone. Transcriptome-wide association studies (TWASs) provide insight into variant effects on gene expression and uncover gene-trait interactions within GWASs [55]. TWASs model the associated effect of variant alleles on nearby gene expression from a reference panel of genotypes and associated expression levels (sourced from public repositories; e.g., the GTEx project) [56]. This model then infers gene expression for participants within the GWAS cohort. From this, we can statistically associate certain expression patterns correlated with the target GWAS trait. The resulting association identifies genes potentially relevant to the trait under investigation (Fig. 2). A common method of modelling TWASs is by summary data-based Mendelian randomization (SMR), which integrates GWAS summary statistics with eQTL data to identify differentially expressed loci associated with complex disease [57]. This circumvents the common issue of unavailable full genotype data for developing well-powered association tests. GWAS2Genes serve as a public database compiling SMR gene-phenotype associations for multiple traits, tissues, and genes. However, this does not prove or disprove causality of a given gene; rather, TWAS methods highlight sets of candidate causal genes that warrant further examination.

TWASs can both confirm and reveal novel gene-trait associations. We revisit the multiethnic MVP cohort blood lipids investigation to demonstrate the utility of TWAS in discovery of novel associations. Four different gene expression reference panels were employed across relevant cell and tissue types: peripheral blood, adipose tissue, liver, and tibial artery tissues. The gene expression profiles derived from these

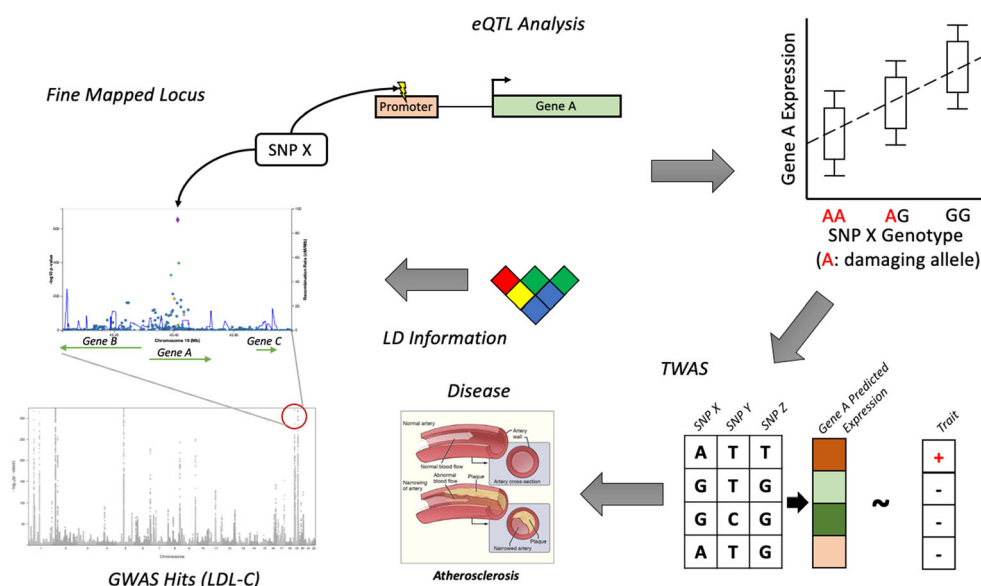
panels were then imputed to predict associated expression within the combined GLGC and MVP GWAS meta-analysis for each lipid trait. 665 gene-lipid associations achieved genome-wide significance within 333 genes. To note, the 333 genes were contained within 122 genomic loci, of which 5 loci were previously unreported. These novel loci represent genomic regions with potential causal impact on lipid traits that were otherwise missed by traditional variant-trait associations [8]. More significant investigation of mutations in non-coding regions of the genome combining different omics data could reveal effects of gene regulation on phenotypic variance where known coding mutations fail to adequately explain the variation between individuals and ancestries.

Translating whole genome information into disease prediction

There are several ways in which lipid genetics could impact clinical practice, but most have not yet been realized. Individuals carrying Mendelian dyslipidemia mutations can be identified based on elevated lipids at a young age, or a family history of premature coronary artery disease, which would allow for earlier and stronger intervention to lower blood lipids. Testing for Mendelian dyslipidemias is not typically used to screen the population. Some challenges with Mendelian testing at scale include the cost of sequencing Mendelian dyslipidemia genes and difficulty in determining between protein-altering variants that cause disease (pathogenic variants) and those that do not (benign).

From a genome-wide perspective, initial discovery efforts were aimed at identifying a large catalogue of lipid genes to enable prioritization of lipid genes for development of new drug therapies. The challenge with this approach is the time

Fig. 2 Applying fine mapping and transcriptomics towards gene prioritization. Significantly associated GWAS loci are identified visually from Manhattan plots. Linkage disequilibrium (LD) information is integrated to identify lead causal SNPs. Measurement of gene expression change (eQTL analysis) for each SNP genotype indicates potential trait causative role of a given allele. Combining multiple candidate SNPs in a transcriptome-wide association study (TWAS) implicates sets of causal SNPs and genes for the target phenotype



and enormous cost required to turn a gene target into a new therapy available to patients. As such, the information produced by genome-wide association studies is not yet applied to clinical practice. This is mainly due to the still ongoing fine mapping efforts (we do not yet know which variants and genes are truly causal), complex biology behind the association (whole gene pathway versus a single gene), and polygenicity (lipids are driven by hundreds or even thousands of genetic loci).

However, an area of intense investigation in the last few years has been on using someone's genetic profile to predict their risk of disease, to again identify individuals who would benefit from lipid-lowering medications or lifestyle modifications. The whole genome information for a susceptibility of a disease/trait level can be summarized using polygenic risk scores (PRS), where the estimated effects for each disease/trait risk allele are summed over the whole genome for each patient carrying those alleles. There are several widely used methods [58–61] to select variants for the score and the different scores created by different research groups worldwide are publicly available (<https://www.pgscatalog.org>).

The predictive power of these scores is promising, especially for CAD [62], as the highest 5% of the CAD PRS appear to have as high a risk of heart disease as those people who carry a Mendelian mutation that causes familial hypercholesterolemia. Moreover, having a high PRS (5% of the population) is more common than carrying a monogenic mutation (less than 1% of the population), which is promising for the advancement of preventive options. Well-developed PRSs for quantitative risk factors (such as LDL-C) could improve the prediction of disease endpoints. For example, prediction of myocardial infarction was increased by combining the disease PRS with the risk factor or biomarker PRSs [63]. Participants in the high-risk group can be identified more accurately with the increased predictive power. It should be noted that CAD is a complex disease with genetic and environmental factors (and possibly interactions between the two) contributing to the overall risk. Hence, thorough evaluation of genetic, environmental and epigenetic factors, together with possible interactions between them, will need to be performed to account for all possible risk factors in the prediction models. This could potentially be accomplished by using machine learning methods that allow for more complex interplay between risk factors [64]. However, to date, machine learning models of this complexity have not yet been applied to CAD risk prediction.

As genetic information is constant throughout lifetime, utilizing genetic information would allow earlier prediction of disease susceptibility, paving the way for prevention rather than treatment after the disease has manifested. In regard to CAD, including the genetic information in the prediction model increases the predictive power to detect early onset cases [65] that would benefit from targeted early adulthood

prevention (Fig. 3). However, this approach is still in early stages, as extensive DNA testing would need to be deployed to identify at-risk patients in clinic for preventive interventions.

There are ongoing efforts to apply genetic risk information to clinical practice. In a Finnish study by Widen et al. [66] CAD risk estimates were returned to study participants, utilizing both traditional risk factors and whole genome genetic risk information, followed by evaluating their lifestyle changes after 6 months. Overall, the results showed positive changes, especially in the high-risk group, suggesting that early prevention with lifestyle changes could be possible with the right tools and easy-to-read risk reports for patients. However, there are well-known challenges in implementing these scores into routine clinical practice, in addition to limitations in patient uptake of recommended behavioral or medication changes.

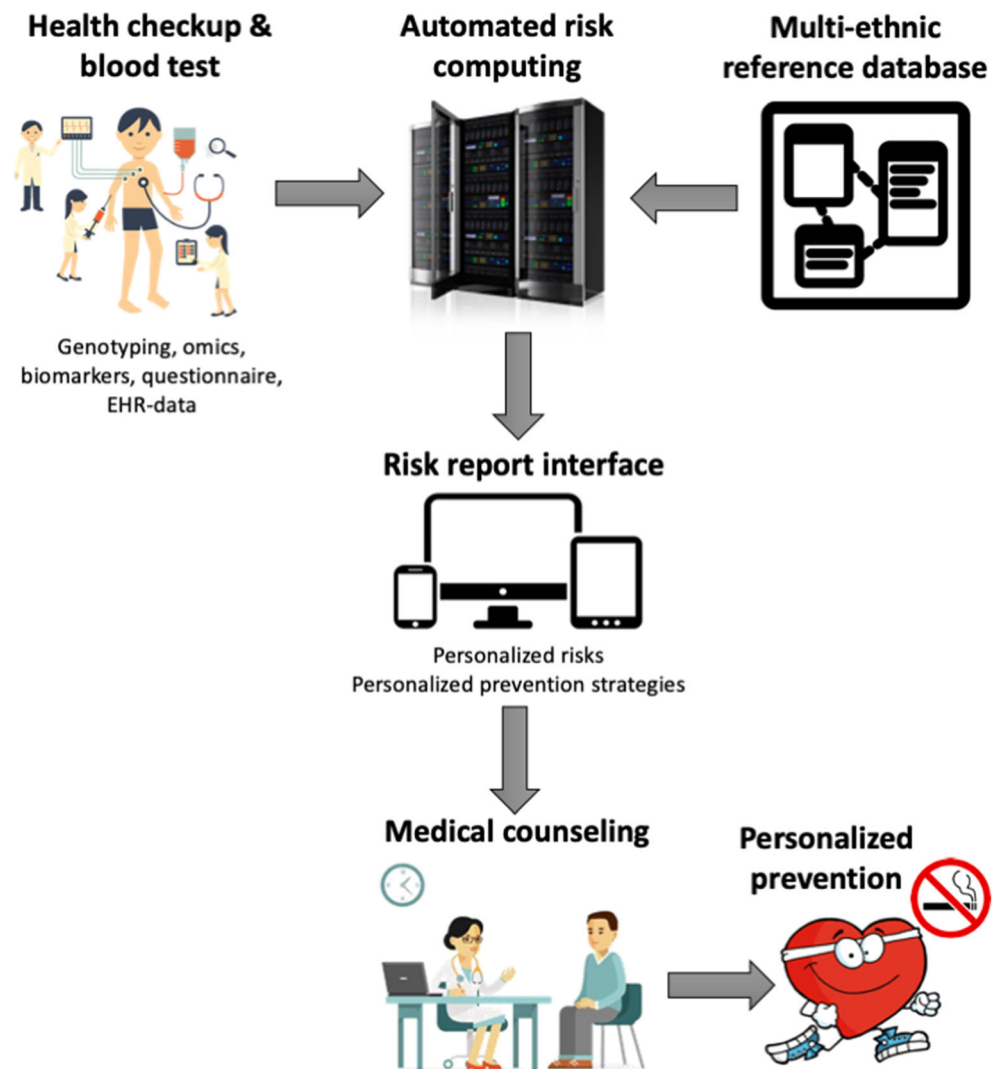
Currently, PRSs are mainly derived from meta-analysis summary statistics that are typically derived from cohorts with a substantial fraction of subjects from European ancestry, which currently limits the utility to predict disease in individuals with other ancestries [67]. There are already multiple haplotype structure and/or genetic variation reference datasets of diverse ancestries available (e.g., HapMap Project [68, 69], 1000 Genomes Project [70], and Haplotype Reference Consortium [71]), but the diversity in datasets with phenotypic data available for association testing or disease prediction remains limited. Additional efforts to build genetic study cohorts with better representation of global ancestries and methodology to better translate results into other ancestry groups may decrease the inequity of disease prediction in the coming years.

Another area of development is to develop best practices and evaluate the overall impact of communicating genetic risks to the patients. Communicating genetic risk requires health care specialists to be able to explain the implications and preventative possibilities to patients in a comprehensible manner. Additionally, the overall predictive power of these scores is still limited. Currently, the biggest challenge in genome analysis and genome-based prediction is the lack of ancestral diversity in the existing study cohort datasets. A proportion of the missing heritability, and therefore lack of predictive power, will most likely be explained by the population specific variants of non-European ancestries currently underrepresented in the GWAS datasets. In addition, we need to be able to create and test disease prediction models for diverse ancestries to equally apply genomic information for participants across the globe.

Future possibilities with large biobanks

There are some suggestive results from phenome wide GWASs to identify drug targets that are less likely to have unanticipated adverse effects by extensively testing

Fig. 3 The aim of complex disease risk prediction. This figure demonstrates how to apply genetic risk to clinical practice. Patients complete a routine DNA test during an annual health exam as well as other laboratory-based tests and basic health questionnaires which are linked to the electronic health record. The risk for multiple complex diseases is calculated and reported through an interface using a secured computing environment. Physicians and/or other health care providers communicate results and recommend tailored preventive actions for the patient



associations between the identified drug target gene or variant across multiple diseases and traits to predict these side-effects that may otherwise be undiscovered until expensive clinical trials are conducted, and human lives may be impacted [72]. In a phenome-wide GWAS, thousands of traits and diseases are tested for association—instead of only one trait of interest—giving clues on all the biological impacts of a single genetic change. These analyses are made possible by the large biobank datasets currently being collected across the world (Biobank Japan, Million Veterans Program, FinnGen, UK Biobank) that combine hospital registries, laboratory measurements, and whole genome information from hundreds of thousands of participants. Integration of transcriptome data and other multiomics approaches can help identify relevant biological pathways and potential targets for new disease therapeutics, as well as avoid creating new medicines that may have unintended negative effects [73]. Large biobank datasets will also allow for testing associations for rarer and population specific genome variations which may quickly highlight

genes as new drug targets. These large datasets with hospital registry data available, some with diverse populations, will also be a powerful tool for creating and testing optimal disease prediction models combining environmental and genetic information.

Discussion

We summarize the currently applied methods used for moving from GWAS summary statistics towards likely drug targets and identification of high-risk individuals for early prevention. While the past two decades have shown an incredible amount of progress, from identifying the first lipid-associated loci using GWAS to uncovering biological mechanisms and drug targets, and more recently translating these discoveries into clinically-meaningful predictions, the gap between observing an association and developing safe drugs for preventing atherosclerotic disease is vast. This gap can be partly narrowed

by using the available methods for fine mapping, aggregating high impact rare variant associations to gene level associations or by combining transcriptomic data on top of the genomic information. As the number of exome sequenced samples is still somewhat limited for burden tests [74], hence having low number of rare variant carriers in the datasets, geneticists are currently capturing genes that have already been found to be plausible candidate genes for drug development in dyslipidemia family-based studies. However, capturing these genes using the current methodology proves that the methods are working, and the lack of novel candidate genes may be due to limited statistical power. In the meantime, prediction may be the key to identifying high risk individuals in the general population, implementing preventive approaches to reduce risk, which will hopefully lead to lower health care costs and, more importantly, reducing the overall number of cardiovascular disease related deaths.

Author contributions B. C.: conception and design; drafting the article and revising for important intellectual content; final approval.

A. M. K.: acquisition of data; drafting the article or revising for important intellectual content; final approval.

W. E. H.: conception and design; drafting the article or revising for important intellectual content; final approval.

C. J. W.: conception and design; drafting the article or revising it critically for important intellectual content; final approval.

I. S.: conception and design; drafting the article or revising it critically for important intellectual content; final approval.

Funding Cristen J. Willer is supported by the National Institutes of Health (R01-HL127564, R35-HL135824, R01-HL142023, and R01-DK075787). Ida Surakka is partially funded by the Michigan Medicine Precision Health Scholarship award.

Declarations

Conflicts of interest Dr. Willer's spouse works for Regeneron.

References

1. Psaty BM, O'Donnell CJ, Gudnason V, et al. Cohorts for heart and aging research in genomic epidemiology (CHARGE) consortium: design of prospective meta-analyses of genome-wide association studies from 5 cohorts. *Circ Cardiovasc Genet*. 2009;2:73–80.
2. Teslovich TM, Musunuru K, Smith AV, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*. 2010;466:707–13.
3. Lango Allen H, Estrada K, Lettre G, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*. 2010;467:832–8.
4. O'Donnell CJ, Kavousi M, Smith AV, et al. Genome-wide association study for coronary artery calcification with follow-up in myocardial infarction. *Circulation*. 2011;124:2855–64.
5. Schunkert H, König IR, Kathiresan S, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat Genet*. 2011;43:333–8.
6. Klein RJ, Zeiss C, Chew EY, et al. Complement factor H polymorphism in age-related macular degeneration. *Science*. 2005;308:385–9.
7. Nielsen JB, Thorolfsdottir RB, Fritsche LG, et al. Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nat Genet*. 2018;50:1234–9.
8. Klarin D, Damrauer SM, Cho K, et al. Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. *Nat Genet*. 2018;50:1514–23.
9. Richardson TG, Sanderson E, Palmer TM, et al. Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: a multivariable Mendelian randomisation analysis. *PLoS Med*. 2020;17:e1003062.
10. Dron JS, Hegele RA. Polygenic influences on dyslipidemias. *Curr Opin Lipidol*. 2018;29:133–43.
11. Hachem SB, Mooradian AD. Familial dyslipidaemias: an overview of genetics, pathophysiology and management. *Drugs*. 2006;66:1949–69.
12. Bamshad MJ, Ng SB, Bigham AW, et al. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet*. 2011;12:745–55.
13. Wolford BN, Hornsby WE, Guo D, et al. Clinical implications of identifying pathogenic variants in individuals with thoracic aortic dissection. *Circ Genom Precis Med*. 2019;12:e002476.
14. Lange LA, Willer CJ, Rich SS. Recent developments in genome and exome-wide analyses of plasma lipids. *Curr Opin Lipidol*. 2015;26:96–102.
15. van der Laan SW, Harshfield EL, Hemerich D, Stacey D, Wood AM, Asselbergs FW. From lipid locus to drug target through human genomics. *Cardiovasc Res*. 2018;114:1258–70.
16. Zeggini E, Ioannidis JP. Meta-analysis in genome-wide association studies. *Pharmacogenomics*. 2009;10:191–201.
17. Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. *Nat Rev Genet*. 2009;10:241–51.
18. Gaziano JM, Concato J, Brophy M, et al. Million Veteran Program: a mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol*. 2016;70:214–23.
19. Lee CH, Cook S, Lee JS, Han B. Comparison of two meta-analysis methods: inverse-variance-weighted average and weighted sum of Z-scores. *Genomics Inform*. 2016;14:173–80.
20. Lacey S, Chung JY, Lin H. A comparison of whole genome sequencing with exome sequencing for family-based association studies. *BMC Proc*. 2014;8:S38.
21. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet*. 2014;95:5–23.
22. Guo MH, Plummer L, Chan YM, Hirschhorn JN, Lippincott MF. Burden testing of rare variants identified through exome sequencing via publicly available control data. *Am J Hum Genet*. 2018;103:522–34.
23. Lange LA, Hu Y, Zhang H, et al. Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. *Am J Hum Genet*. 2014;94:233–45.
24. Hegele RA, Tsimikas S. Lipid-lowering agents. *Circ Res*. 2019;124:386–404.
25. Raal FJ, Kallend D, Ray KK, et al. Inclisiran for the treatment of heterozygous familial hypercholesterolemia. *N Engl J Med*. 2020;382:1520–30.
26. Greig JA, Limberis MP, Bell P, et al. Non-clinical study examining AAV8.TBG.hLDLR vector-associated toxicity in chow-fed wild-Type and LDLR(+/-) Rhesus Macaques. *Hum Gene Ther Clin Dev*. 2017;28:39–50.
27. Greig JA, Limberis MP, Bell P, et al. Nonclinical pharmacology/toxicology study of AAV8.TBG.mLDLR and AAV8.TBG.hLDLR

- in a mouse model of homozygous familial hypercholesterolemia. *Hum Gene Ther Clin Dev*. 2017;28:28–38.
28. Wong E, Goldberg T. Mipomersen (kynamro): a novel antisense oligonucleotide inhibitor for the management of homozygous familial hypercholesterolemia. *P t*. 2014;39:119–22.
 29. Abifadel M, Varret M, Rabès JP, et al. Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. *Nat Genet*. 2003;34:154–6.
 30. Cohen J, Pertsemlidis A, Kotowski IK, Graham R, Garcia CK, Hobbs HH. Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nat Genet*. 2005;37:161–5.
 31. Crosby J, Peloso GM, Auer PL, et al. Loss-of-function mutations in APOC3, triglycerides, and coronary disease. *N Engl J Med*. 2014;371:22–31.
 32. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*. 2004;305:869–72.
 33. Willer CJ, Schmidt EM, Sengupta S, et al. Discovery and refinement of loci associated with lipid levels. *Nat Genet*. 2013;45:1274–83.
 34. Surakka I, Horikoshi M, Mägi R, et al. The impact of low-frequency and rare variants on lipid levels. *Nat Genet*. 2015;47:589–97.
 35. Sanna S, Li B, Mulas A, et al. Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. *PLoS Genet*. 2011;7:e1002198.
 36. Génin E. Missing heritability of complex diseases: case solved? *Hum Genet*. 2020;139:103–13.
 37. Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet*. 2018;19:491–504.
 38. Lu X, Peloso GM, Liu DJ, et al. Exome chip meta-analysis identifies novel loci and East Asian-specific coding variants that contribute to lipid levels and coronary artery disease. *Nat Genet*. 2017;49:1722–30.
 39. Zubair N, Graff M, Luis Ambite J, et al. Fine-mapping of lipid regions in global populations discovers ethnic-specific signals and refines previously identified lipid loci. *Hum Mol Genet*. 2016;25:5500–12.
 40. Pers TH, Karjalainen JM, Chan Y, et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nat Commun*. 2015;6:5890.
 41. Weeks EM, Ulirsch JC, Cheng NY et al. Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. *medRxiv* 2020:2020.09.08.20190561.
 42. Morris AP. Transethnic meta-analysis of genomewide association studies. *Genet Epidemiol*. 2011;35:809–22.
 43. Ruan X, Li P, Chen Y, et al. In vivo functional analysis of non-conserved human lncRNAs associated with cardiometabolic traits. *Nat Commun*. 2020;11:45.
 44. Muret K, Désert C, Lagoutte L, et al. Long noncoding RNAs in lipid metabolism: literature review and conservation analysis across species. *BMC Genomics*. 2019;20:882.
 45. Lee KH, Hwang HJ, Cho JY. Long non-coding RNA associated with cholesterol homeostasis and its involvement in metabolic diseases. *Int J Mol Sci*. 2020;21.
 46. Ruan X, Li P, Ma Y, et al. Identification of human long noncoding RNAs associated with nonalcoholic fatty liver disease and metabolic homeostasis. *J Clin Invest*. 2021;131.
 47. Mittelstraß K, Waldenberger M. DNA methylation in human lipid metabolism and related diseases. *Curr Opin Lipidol*. 2018;29:116–24.
 48. Braun KV, Voortman T, Dhana K, et al. The role of DNA methylation in dyslipidaemia: a systematic review. *Prog Lipid Res*. 2016;64:178–91.
 49. Sayols-Baixeras S, Irvin MR, Arnett DK, Elosua R, Aslibekyan SW. Epigenetics of lipid phenotypes. *Curr Cardiovasc Risk Rep*. 2016;10.
 50. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res*. 2019;47:D886–d894.
 51. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods*. 2010;7:248–9.
 52. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res*. 2003;31:3812–4.
 53. Boyle AP, Hong EL, Hariharan M, et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res*. 2012;22:1790–7.
 54. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods*. 2015;12:931–4.
 55. Wainberg M, Sinnott-Armstrong N, Mancuso N, et al. Opportunities and challenges for transcriptome-wide association studies. *Nat Genet*. 2019;51:592–9.
 56. Carithers LJ, Ardlie K, Barcus M, et al. A novel approach to high-quality postmortem tissue procurement: the GTEx project. *Biopreserv Biobank*. 2015;13:311–9.
 57. Meng XH, Chen XD, Greenbaum J, et al. Integration of summary data from GWAS and eQTL studies identified novel causal BMD genes with functional predictions. *Bone*. 2018;113:41–8.
 58. Vilhjálmsson BJ, Yang J, Finucane HK, et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am J Hum Genet*. 2015;97:576–92.
 59. Ge T, Chen CY, Ni Y, Feng YA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat Commun*. 2019;10:1776.
 60. Mak TSH, Porsch RM, Choi SW, Zhou X, Sham PC. Polygenic scores via penalized regression on summary statistics. *Genet Epidemiol*. 2017;41:469–80.
 61. Inouye M, Abraham G, Nelson CP, et al. Genomic risk prediction of coronary artery disease in 480,000 adults: implications for primary prevention. *J Am Coll Cardiol*. 2018;72:1883–93.
 62. Khera AV, Chaffin M, Zekavat SM, et al. Whole-genome sequencing to characterize monogenic and polygenic contributions in patients hospitalized with early-onset myocardial infarction. *Circulation*. 2019;139:1593–602.
 63. Sinnott-Armstrong N, Tanigawa Y, Amar D et al. Genetics of 38 blood and urine biomarkers in the UK Biobank. *Nat Genet*. 2021;53:185–94.
 64. Dias R, Torkamani A. Artificial intelligence in clinical and genomic diagnostics. *Genome Med*. 2019;11:70.
 65. Mars N, Koskela JT, Ripatti P, et al. Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nat Med*. 2020;26:549–57.
 66. Widen E, Junna N, Ruotsalainen S et al. Communicating polygenic and non-genetic risk for atherosclerotic cardiovascular disease—an observational follow-up study. *medRxiv* 2020:2020.09.18.20197137.
 67. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet*. 2019;51:584–91.
 68. The International HapMap Project. *Nature*. 2003;426:789–96.
 69. A haplotype map of the human genome. *Nature*. 2005;437:1299–320.
 70. Abecasis GR, Auton A, Brooks LD, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491:56–65.
 71. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet*. 2016;48:1279–83.

72. Diogo D, Tian C, Franklin CS, et al. Phenome-wide association studies across large population cohorts support drug target validation. *Nat Commun.* 2018;9:4285.
73. Nielsen JRO, Surakka I, Graham S, Zhou W, Roychowdhury T, Fritsche L, et al. Loss-of-function genetic variants with impact on liver-related blood traits highlight potential therapeutic targets for cardiovascular disease. *Nat Commun.* 2020;11:6417.
74. Van Hout CV, Tachmazidou I, Backman JD, et al. Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature.* 2020;586:749–56.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.