ANNUAL REVIEWS

Annual Review of Genomics and Human Genetics

Using Full Genomic Information to Predict Disease: Breaking Down the Barriers Between Complex and Mendelian Diseases

Daniel M. Jordan and Ron Do

Charles Bronfman Institute for Personalized Medicine and Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; email: ron.do@mssm.edu

Annu. Rev. Genom. Hum. Genet. 2018. 19:289–301

First published as a Review in Advance on April 11, 2018

The *Annual Review of Genomics and Human Genetics* is online at genom.annualreviews.org

https://doi.org/10.1146/annurev-genom-083117-021136

Keywords

polygenic risk score, complex disease, genetic architecture, disease prediction

Abstract

While sequence-based genetic tests have long been available for specific loci, especially for Mendelian disease, the rapidly falling costs of genome-wide genotyping arrays, whole-exome sequencing, and whole-genome sequencing are moving us toward a future where full genomic information might inform the prognosis and treatment of a variety of diseases, including complex disease. Similarly, the availability of large populations with full genomic information has enabled new insights about the etiology and genetic architecture of complex disease. Insights from the latest generation of genomic studies suggest that our categorization of diseases as complex may conceal a wide spectrum of genetic architectures and causal mechanisms that ranges from Mendelian forms of complex disease to complex regulatory structures underlying Mendelian disease. Here, we review these insights, along with advances in the prediction of disease risk and outcomes from full genomic information.

INTRODUCTION

Mendelian disease: a disease caused by a single locus, inherited in a Mendelian pattern (autosomal dominant, autosomal recessive, X-linked recessive, etc.)

Complex disease: a disease caused by multiple heterogeneous risk factors, often including genetic factors, lifestyle, and/or environmental factors as well as unknown confounders and interactions among these

Polygenic risk score (PRS): a quantitative score measuring the combined disease risk contributed by multiple loci genome-wide

Causal risk factor: a risk factor with a definite causal connection to a disease, as opposed to one that is merely correlated with the disease

It has long been one of the core goals of the field of medical genetics to create accurate sequence-based genetic tests that will predict disease risk. This project began even before the complete human genome was sequenced, and the successful mapping of disease genes such as *CFTR* (cystic fibrosis) (53, 56), *HEXA* (Tay–Sachs disease) (3), and *BRCA1* and *BRCA2* (hereditary breast and ovarian cancer) (45, 70) in the late 1980s and early 1990s paved the way for genetic testing and carrier screening for these associated diseases. Since then, the sequencing of the human genome and the rise of whole-exome sequencing studies have accelerated the discovery of disease-causing mutations for rare Mendelian diseases (4). There are now hundreds of known loci that reliably predict risk carrier status for various Mendelian diseases, and many of these tests are routinely performed as part of prenatal, preventative, and diagnostic care (10, 26). In some cases, these tests have had noticeable impacts on disease prevalence and/or outcomes (37, 46).

In parallel with this development of single-locus sequence-based predictors of disease, through the success of genome-wide association studies (GWASs), we have also learned a great deal about the biology of complex diseases that cannot be predicted reliably from a single locus (8, 12, 13, 16, 18, 23, 54, 55, 57–59, 67). These diseases, which include coronary artery disease, type 2 diabetes, irritable bowel disease, schizophrenia, autism, and many others, appear to be caused by contributions from multiple loci all across the genome, which reduces the usefulness of single-locus tests to predict disease risk. In many cases, different causal loci may have different mechanisms of disease causation, which makes predicting disease risk even more difficult. As a result, it has been widely considered infeasible to predict risk for these complex diseases from genetic information at a practical and clinically useful level. However, as the power of GWASs and sequencing studies of these diseases continues to grow, and as clinical genetic testing becomes more and more widely available, the idea of integrating information from multiple loci across the genome to predict complex disease risk becomes more plausible.

Conversely, it has also been thought that using single loci of large effect to predict highly penetrant monogenic disease is a mature field and that adding more genomic information will not help to improve these predictions. In fact, in addition to gaining information about diseases known to be complex, the growing power of GWASs and sequencing studies has shown that many diseases previously thought to be caused by highly penetrant Mendelian alleles exist on a spectrum from Mendelian to complex (14, 21). This includes some hints that diseases previously thought to be Mendelian may be modulated by complex risk factors and controlled by common variation (9). Similarly, some apparently Mendelian diseases, such as familial hypercholesterolemia, contribute to complex networks of risk for diseases thought to be highly polygenic (1, 5, 25, 36, 47, 62).

In this review, we outline the benefits and challenges of using complete genetic information to predict disease risk for both Mendelian and complex diseases. In particular, we discuss the power of polygenic risk scores (PRSs) to predict disease risk, decomposition of disease risk into multiple causal risk factors and/or pathways of disease causation, and the difficulty of accounting for genetic ancestry in these genome-wide genetic testing schemes. Each of these topics is an active subject of current research and has implications in human genetics and genomic medicine. Each also serves to blur the boundaries between Mendelian disease and complex disease and suggests a future where Mendelian disease loci and polygenic estimates of disease risk are interpreted together to generate a prediction of disease risk for each individual patient.

POLYGENIC RISK SCORES

The most common method of integrating information from across the genome into a single estimate of genetic risk is by using a PRS. A PRS is simply a sum of the genetic status at each

associated risk locus weighted by the strength of evidence for the association. Since the loci in question are typically found by GWASs, the weighting of each locus is usually simply the regression coefficient of association for the locus. The PRS acts as a proxy for the total genetic risk or protection from the genome as a whole. This is useful for highly polygenic diseases, in which any individual locus makes a minor contribution to disease risk but the combined genetic contribution across the entire genome may be significant. Such a score was successfully created in 2009 in connection with a GWAS for schizophrenia, where it was shown that a PRS comprising variants across multiple loci could explain a substantial fraction of schizophrenia risk, even if no single variant could on its own (29). Since then, PRSs have been routinely constructed as part of GWAS methodology, and they exist for a wide variety of traits and diseases. They are used to analyze the power of the discovered loci to explain variation and heritability in the trait, analyze the relative contributions of different genes or other loci to disease risk, analyze genetic correlations between traits, and investigate causal relationships between traits using Mendelian randomization, among many other applications (6, 11, 20, 28, 29, 34, 35, 39, 40, 50, 52, 54, 57–59, 63–66, 69).

Genetic architecture: the distribution of genetic risk for a given disease, both among alleles and among the population

In principle, it is also possible to use a PRS to predict disease risk for a single individual; historically, however, they have rarely been used for this purpose. This is because the power of most PRSs to predict disease risk has been very low for most of the history of GWASs. The tone for this field was set by a 2009 study that applied PRSs from current GWASs to predict the case or control status of samples from the Wellcome Trust Case Control Consortium (20). Out of seven diseases investigated, only for type 1 diabetes did the PRS have even modest power to distinguish between cases and controls. In the remaining six diseases investigated—bipolar disorder, coronary artery disease, hypertension, Crohn's disease, rheumatoid arthritis, and type 2 diabetes—the PRS had extremely limited power.

GENETIC ARCHITECTURE: A SPECTRUM FROM MENDELIAN TO COMPLEX

The reason these PRSs are underpowered for these diseases fundamentally has to do with the genetic architecture of the diseases. The genetic architecture of a disease, in this context, has at least two distinct components: (*a*) whether the disease is caused primarily by a few large-effect variants or many small-effect variants and (*b*) whether the disease is caused primarily by a few common variants or many rare variants. A disease that is caused primarily by variants with large effect sizes is easy to predict because each locus carries a great deal of information about disease risk. Meanwhile, a disease that is caused by a combination of variants with small effects may be difficult to predict because many loci must be considered in order to glean any information about disease risk. The population frequency of variants is also vitally important, as it determines the sample size required to detect variants with small effect sizes. Common causal variants may be detectable with sample sizes of 1,000 or less, even if they have weak effects, while causal variants that are private to an individual or a family may be fundamentally impossible to detect with any sample size, especially if their effect sizes are small.

The genetic architecture of a disease therefore determines the number and effect sizes of risk loci found at a given sample size, which in turn determine the power of the PRS constructed from those risk loci. On one extreme are monogenic or Mendelian diseases, where a large amount of genetic risk is contained in one locus, and a simple genotyping or sequencing test for that locus has a great deal of predictive power; on the other extreme are complex and highly polygenic diseases, such as coronary artery disease or type 2 diabetes, where risk is distributed over a large number of loci, each of which individually has a very small effect. For these complex diseases, even polygenic risk scores accounting for thousands of loci have minimal predictive power. In between

are oligogenic diseases like type 1 diabetes, where risk is distributed over a smaller number of loci with intermediate effect sizes.

ESTIMATING AND QUANTIFYING GENETIC ARCHITECTURE

This qualitative description of genetic architecture contains two important quantitative questions: How many loci would be required to have reasonable power for a given disease, and how large a sample size is required to reach it? A 2013 analysis of the predictive power of PRSs expressed this quantitatively, defining predictive power as a function of the number of causal loci and the fraction of disease heritability explained by these loci (17). Using estimates of these parameters, this study concluded that for most complex diseases, GWAS sample sizes would need to reach hundreds of thousands to achieve reasonable power to predict the disease status of a single individual. Furthermore, Agarwala and colleagues from the GoT2D (Genetics of Type 2 Diabetes) consortium (2) conducted a simulation study to estimate the genetic architecture of type 2 diabetes more explicitly. They simulated a variety of different genetic architectures, varying how many variants in the genome could potentially cause type 2 diabetes (which determines the number of causal alleles that will be found) and how strongly natural selection acts against type 2 diabetes (which determines how rare or common the typical causal allele will be). They then compared simulated results to empirical values for quantities like disease prevalence, disease heritability, number of loci discovered by GWASs, and fraction of heritability explained by known loci. Their results generally ruled out a contribution of large-effect common variants but were still consistent with a wide range of genetic architectures; for example, rare variants may still explain as little as 25% or as much as 80% of heritability. This estimate is likely to narrow as sample sizes increase. However, they also concluded that gaining significant power to predict type 2 diabetes would require a GWAS with a sample size in the hundreds of thousands—they estimated a minimum of 250,000, possibly much larger. A subsequent study from the same consortium applied a similar estimation to a larger population of more than 125,000 and was able to narrow the estimate of genetic architecture, ruling out a large contribution of extremely rare variants, but reached the same conclusion that a sample of at least 250,000 will be necessary to reach a comprehensive understanding of the genetics of type 2 diabetes (24).

This kind of direct estimation of the genetic architecture of complex disease is important, since it has a direct impact on the power of GWASs and the possibility of using PRSs to predict disease. Although the study of type 2 diabetes by Agarwala et al. (2) did not produce any concrete recommendations for future studies other than the need to perform larger association studies, such analyses have the potential to inform study design in a significant way, by indicating whether researchers should be targeting rare, large-effect variants; common, small-effect variants; or anything in between.

GENETIC ANCESTRY AND DISEASE PREDICTION: AN UNRESOLVED ISSUE

In addition to genetic architecture, another factor severely limiting the power of PRSs for disease risk prediction is genetic ancestry. The vast majority of published GWASs have been performed on populations of mostly European ancestry, which has led to some concerns about the generalizability of these findings (51). Several examples exist of GWAS loci discovered in European cohorts failing to replicate in trans-ethnic cohorts or of trans-ethnic GWASs explaining less variance than GWASs performed on an ethnically homogeneous population (22, 38, 43, 61). This has several possible explanations. In many cases, it is likely that a variant that is easily detectable in one population

is absent or extremely rare in another. Even if the same causal variant is important in multiple populations, it is also possible that the haplotype structure is different between the populations such that a different set of detectable variants are in linkage disequilibrium with the same causal locus (15). Finally, in some cases, the actual underlying regulatory biology may differ in different populations (41).

A recent study by Martin et al. (42) addressed the question of what effect this may have on prediction of disease risk. The researchers simulated the evolution of the same trait in multiple populations and performed a GWAS in each population to develop a simulated PRS. They demonstrated that predictions were significantly biased when applied to a different population than the population in which the PRS was developed. Even when the causal loci are actually identical across the populations, the contribution of each locus to the PRS will be different in each population, leading to the same bias. The authors demonstrated that this appears to be affecting actual GWASs, as the distribution of PRS values for various traits and diseases differs wildly among populations, far in excess of any difference in actual trait values or disease incidence.

This study highlights the importance of expanding GWASs to non-European populations. This has improved somewhat in recent years, with projects like the PAGE (Population Architecture Using Genomics and Epidemiology) consortium making a concerted effort to generate multiethnic cohorts for research (68), but Europeans still make up a large majority of GWAS populations. This is especially important considering that non-European populations are at significantly higher risk for many of the diseases under discussion, including coronary artery disease, type 2 diabetes, and hypertension. At present, the populations that would most benefit from the use of genome-wide data to predict disease risk would be unable to take full advantage of it, since the loci used to predict risk are discovered using European populations.

STRATIFYING COMPLEX DISEASE RISK

These twin considerations of genetic architecture and genetic ancestry justify the general attitude of the field that using PRSs and GWAS markers to predict disease is difficult. Without sample sizes in the hundreds of thousands or more spanning an extremely heterogeneous group of ancestries, we cannot possibly have enough power to predict disease status reliably using these methods, and these sample sizes have been out of reach until very recently. However, the sample sizes described in these studies are beginning to seem less outlandish. Researchers now have access to resources like the UK Biobank, which includes information on a wide variety of traits for its 500,000 participants, opening the possibility of performing GWASs on any of these traits with sample sizes in the hundreds of thousands (7). Recently released GWASs of coronary artery disease (65) and blood pressure (19) take advantage of this population and others to reach sample sizes of half a million or more. Even without using this population, large meta-analyses of disease genetics now regularly reach sample sizes in the hundreds of thousands. For example, the largest meta-analysis of type 2 diabetes, from the DIAGRAM (Diabetes Genetics Replication and Meta-Analysis) consortium, had a total sample size of 291,748 across all cohorts (60). These sample sizes have not universally reached the level required to give reliable predictive power, and the ancestry issue remains problematic, but they are approaching that level rapidly enough that we must soon begin to reconsider the usefulness of PRSs as predictors of disease risk for an individual patient.

One recent study of more than 55,000 participants by Khera et al. (31) demonstrated that PRSs could be a useful prognostic tool even using current GWAS loci. In this study, the researchers used 50 published associated variants to develop a PRS for coronary artery disease and applied it to well-phenotyped test populations. They then divided the population into quintiles of genetic risk based on this PRS. This transforms genetic risk into a categorical variable, which is commonly done

with risk factors such as diet, obesity, smoking, or exercise. Similar to how one might compute a hazard ratio to quantify how risk of disease is affected by a risk factor like hypertension or obesity, these researchers computed hazard ratios to quantify how risk of disease is affected by genetic risk. Comparing the highest quintile of genetic risk with the lowest, they measured hazard ratios between 1.75 and 1.98, meaning that individuals with high genetic risk are 75–98% more likely to develop coronary artery disease than those without. This was a stronger effect than any of the individual lifestyle factors (diet, obesity, smoking, and exercise) and comparable to an index that combined all four lifestyle factors, which produced hazard ratios between 1.71 and 2.27. Thus, being in the highest quintile of genetic risk has similar predictive power to other factors that are widely used to predict disease risk. Furthermore, the authors concluded that the effect of lifestyle factors was greater in the cohort with higher genetic risk, which raises the possibility of targeting lifestyle interventions based on genetic risk.

Another study from the same group (30) went a step further, directly comparing the predictive power of a PRS with the predictive power of rare monogenic mutations. They found that having a PRS in the top 2.5% of the population for coronary artery disease has equivalent predictive power to being a carrier of a previously published rare monogenic disease mutation, both producing an approximately fourfold increase in risk. They pointed out that this actually makes the PRS substantially more useful for detecting coronary artery disease risk, as the number of individuals in the top 2.5% of PRS is more than six times larger than the number of individuals who carry a rare monogenic mutation.

The approach described in these two studies demonstrates that it is not necessary to have a reliable predictor of disease status in order to gain clinical utility. Instead, using PRSs to stratify individuals into low and high genetic risk may have prognostic utility. Many prognostic tools and risk factors that are of great clinical significance still have only modest predictive power, and PRSs, even at their present power, are comparable to some of them. Furthermore, if we were able to construct a PRS from larger numbers of causal loci, by either relaxing the selection criteria or using a larger population sample, it is plausible that the top quintile of genetic risk might be even more predictive and therefore possibly more useful.

DISEASE SUBTYPES AS A SPECTRUM OF GENETIC ARCHITECTURE

An important additional factor when considering this kind of stratification of disease risk is that, in addition to heterogeneous polygenic risk, many complex diseases have rare subtypes with Mendelian or oligogenic genetic architectures, which are often considered independent from the more common complex form of the disease. A recent review by Flannick et al. (21) enumerated the many monogenic forms of diabetes mellitus and argued that they should be viewed as part of a spectrum ranging from monogenic forms of the disease up to the more common highly polygenic form. This argument is based on the observation that the various forms of the disease appear to act through a shared superset of biological components and risk factors. Thus, rather than each patient being diagnosed with a specific subtype of the disease, we can think of each patient as having a specific disease-causing genotype, which may be monogenic, oligogenic, or polygenic. Each genotype corresponds to a specific profile of disrupted biological function, which in turn leads to different prognoses and different options for treatment. In addition to challenging the categorization of diabetes subtypes specifically, this concept, taken to its logical conclusion, also challenges the categorization of diseases generally into complex and Mendelian, suggesting that many of the diseases we typically think of as complex may have an array of genetic architectures that exist on a spectrum from rare Mendelian forms to common complex forms.

DISEASE SUBTYPES AS A PALETTE OF CAUSAL RISK FACTORS

In addition to highly polygenic genetic architecture, another feature of complex disease is heterogeneity of risk factors and causation. In cases like coronary artery disease, where we do have some understanding of disease biology and some ability to predict disease risk, the specific mechanism that led to the disease can often make a clinically meaningful difference. For example, it may be useful to prescribe cholesterol-lowering drugs such as statins to a patient whose coronary artery disease risk is caused primarily by high low-density lipoprotein (LDL) cholesterol, but the same treatment might be less effective if the primary cause is hypertension or type 2 diabetes.

There is therefore a meaningful difference between a patient with high genetic risk that is caused primarily through a single pathway, such as a patient whose high genetic risk of coronary artery disease comes from a genetic predisposition to high LDL cholesterol, and a patient with high but heterogeneous genetic risk. This principle is outlined in a perspective by Khera & Kathiresan (33), which highlights the clinical difference between patients with monogenic disease, caused through a single biological pathway, and patients with polygenic disease, which is likely to be heterogeneous. Thus, in addition to different subtypes of the diseases existing along a spectrum from complex disease to Mendelian disease, they also exist within a palette of distinct causal pathways and risk factors, each of which corresponds to different prognoses and treatment options.

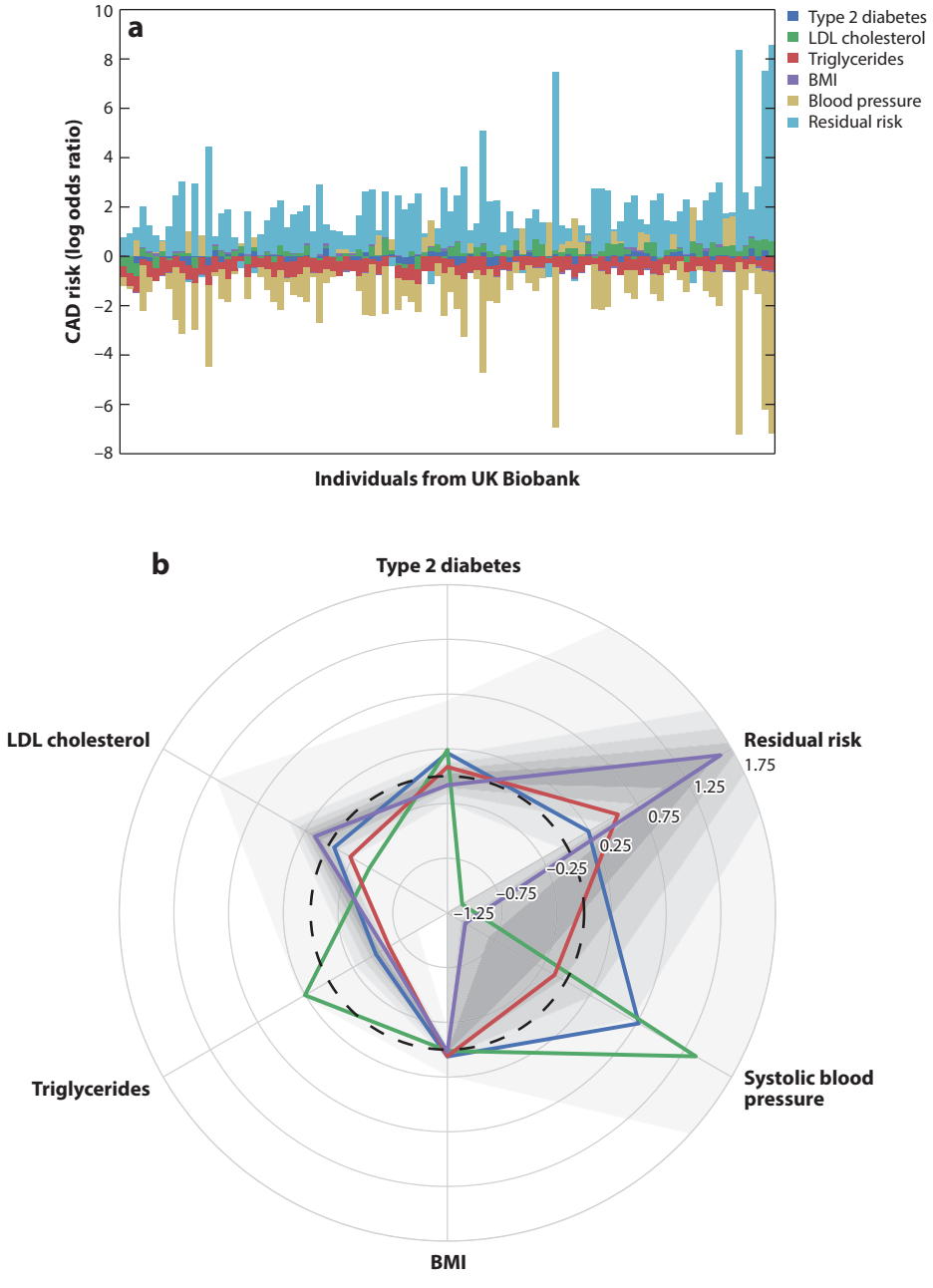
This insight also applies within a highly polygenic disease subtype. A patient with a highly polygenic risk profile for a complex disease has a risk profile composed of contributions from a variety of distinct causal risk factors. These causal risk factors are typically also well-studied traits with their own GWAS and known causal loci, so we can often interrogate the causal palette of a given patient's disease risk directly. In the case of coronary artery disease, meta-analyses with sample sizes in the tens or hundreds of thousands have been performed for each of the five known primary causal risk factors (type 2 diabetes, LDL cholesterol, triglycerides, body mass index, and systolic blood pressure) (32), and each has dozens to hundreds of known risk loci. The known risk loci for coronary artery disease can easily be decomposed into loci that are also known risk loci for these five causal risk factors and loci that cause coronary artery disease idiopathically (**Figure 1**). Once risk loci are decomposed in this way, it is straightforward to identify the palette of risk factors in an individual patient, which may have implications for prognosis and treatment.

IS COMPLEX DISEASE REALLY COMPLEX?

The ideas of decomposing complex disease into a palette of different causal risk factors and decomposing complex disease into a spectrum of genotypes ranging from rare Mendelian disease to common complex disease contribute to a growing sense that supposedly complex diseases may be composed of a variety of distinct disease subtypes, with different genetic architectures, disease etiologies, and clinical features (**Figure 2**). This points to a much less complex conception of complex disease, where patient genotypes can be used to identify which disease subtype a patient has, which can then in turn inform treatment and prognosis.

This can be especially powerful when combined with the approach of stratifying complex disease risk discussed above. The insight that a high PRS confers as much additional risk as we might expect from a single highly penetrant disease allele suggests that there may be a population of currently undiagnosed patients with polygenic risk equivalent to that for Mendelian disease. Particularly if that risk applies to a specific homogeneous disease subtype, this suggests that we should be treating the PRS the way we treat these single highly penetrant disease alleles—for example, by actively screening for patients with high PRSs and targeting clinical interventions toward these patients specifically. Recent studies have begun to examine this approach, particularly

in coronary artery disease and hypercholesterolemia, demonstrating that drugs such as statins may be more effective for individuals with higher genetic risk (44, 48). We anticipate this approach becoming more common for a variety of diseases. Even though genetic risk for these diseases may be complex in terms of genetic architecture, the increasing predictive power of PRSs will allow physicians to treat predictions of genetic risk in a less complex way.



(*Caption appears on following page*)

Figure 1 (*Figure appears on preceding page*)

Decomposition of coronary artery disease (CAD) risk into five causal risk factors for individuals in the UK Biobank. Polygenic risk scores (PRSs) were constructed for CAD, type 2 diabetes, low-density lipoprotein (LDL) cholesterol, triglycerides, body mass index (BMI), and systolic blood pressure (32), each using 65 genome-wide significant CAD variants (49). Causal effects of each risk factor on CAD were estimated using the inverse-variance-weighted regression method of Mendelian randomization (27). The contribution to CAD risk from each risk factor was calculated using the PRS for that risk factor multiplied by the causal effect on CAD; residual risk was calculated by subtracting the five risk factor contributions from the total amount of CAD risk. Note that these values are computed only from CAD risk alleles, and therefore the distributions of individual risk factors in this analysis do not necessarily reflect their true population-level distributions or their overall causal effects on CAD risk. Panel *a* shows the total contributions for 100 randomly selected individuals from the UK Biobank, ordered from smallest total risk (*left*) to largest (*right*). Panel *b* shows four representative individuals in more detail. Gray shaded bands contain the middle 100%, 80%, 60%, 40%, and 20% of individuals. This illustrates that different individuals have not only different levels of risk for CAD but also distinct risk profiles with substantially different contributions from each risk factor. It also illustrates that different risk factors have different levels of variability in contribution. For example, there is wide variation in the contributions of LDL cholesterol and systolic blood pressure but relatively little in the contribution of BMI. This is due to the relative genetic architectures of these traits and their causal effects on CAD. The dashed black circle indicates zero genetic risk, or an odds ratio of 1.0.

IS MENDELIAN DISEASE REALLY MENDELIAN?

In addition to the growing awareness that complex disease may often be composed of a network of less complex subphenotypes, there is a growing understanding that Mendelian disease may

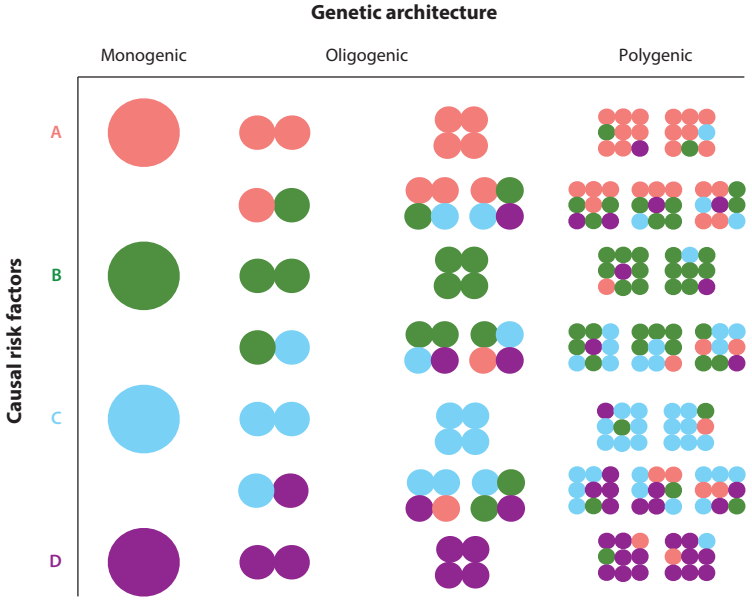


Figure 2

Schematic illustrating two dimensions of disease subtypes: distinct genetic architectures (monogenic versus oligogenic versus polygenic) and distinct causal risk factors. A complex disease may be caused by monogenic contributions acting through any individual risk factor, oligogenic contributions acting through any individual risk factor or subset of risk factors, highly polygenic contributions that predominantly target a particular risk factor or subset of risk factors, or highly polygenic contributions that target a diverse mixture of risk factors. Each individual has a specific genotype that corresponds to a single configuration.

be modulated by modifiers with more complex genetic architectures. One component of this involves the spectrum of disease subtypes mentioned above: Mendelian diseases may be part of a spectrum that includes complex forms of the same disease, or they may interact with similar, complex phenotypes. There are also numerous examples of supposedly Mendelian phenotypes being exacerbated or mitigated by known complex traits. The most intriguing case in current research is that of incompletely penetrant diseases. It has long been known that many well-characterized causal loci for Mendelian disease fail to cause disease in a large number of individuals who carry the disease allele. One possible explanation is that the expression of the disease allele is controlled by a complex network of regulatory variation. One recent study by Castel et al. (9) analyzed expression of the promoters of Mendelian disease genes to demonstrate that regulatory variation has a large effect on the penetrance of a wide range of supposedly Mendelian diseases, supporting this hypothesis. However, this is only a first step toward characterizing these regulatory networks; in most cases, we have very little sense of the regulatory variation that underlies variable penetrance of Mendelian diseases. Deciphering these complex components of Mendelian disease will likely be an exciting area of research in the near future.

CONCLUSION

The idea of using genetic information to predict an individual's disease status has existed for decades, but only recently has it begun to become feasible for complex diseases. As sample sizes grow, we are beginning to understand the genetic architectures of both Mendelian and complex disease and develop tools to apply that knowledge to clinical practice. As we continue to do so, we will likely find that complex disease is not really as intractable as it is often thought to be, and that predictors of genetic risk may translate into clear clinical indications. On the other hand, we may find that many supposedly Mendelian diseases are not so tractable, but are controlled by a currently opaque network of regulators and complex traits. These two insights, combined with the understanding that variation among different genetic ancestries has large effects on both categories of disease, will likely drive the next era of research in disease prediction. In response to these insights, some researchers have begun to suggest a future where patients will be diagnosed and treated based primarily on their genotypes rather than their phenotypic presentations. This genotype-first vision will leverage our understanding of the complexity of genetic risk to identify disease subtypes and quantify genetic risk accurately, resulting in more accurate diagnoses of clinical diseases in patient care and targeted interventions where they will be most useful.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

D.M.J. is supported by grant T32HL007824 from the National Heart, Lung, and Blood Institute of the National Institutes of Health. R.D. is supported by grant R35GM124836 from the National Institute of General Medical Sciences and grant R01HL139865 from the National Heart, Lung, and Blood Institute of the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

LITERATURE CITED

1. Abifadel M, Varret M, Rabès J-P, Allard D, Ouguerram K, et al. 2003. Mutations in *PCSK9* cause autosomal dominant hypercholesterolemia. *Nat. Genet.* 34:154–56
2. **Agarwala V, Flannick J, Sunyaev S, GoT2D Consort., Altshuler D. 2013. Evaluating empirical bounds on complex disease genetic architecture. *Nat. Genet.* 45:1418–27**
3. Arpaia E, Dumbrille-Ross A, Maler T, Neote K, Tropak M, et al. 1988. Identification of an altered splice site in Ashkenazi Tay-Sachs disease. *Nature* 333:85–86
4. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, et al. 2011. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* 12:745–55
5. Berge KE, Tian H, Graf GA, Yu L, Grishin NV, et al. 2000. Accumulation of dietary cholesterol in sitosterolemia caused by mutations in adjacent ABC transporters. *Science* 290:1771–75
6. Brown BC, Asian Genet. Epidemiol. Netw. Type 2 Diabetes, Ye CJ, Price AL, Zaitlen N. 2016. Transethnic genetic correlation estimates from summary statistics. *Am. J. Hum. Genet.* 99:76–88
7. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, et al. 2017. Genome-wide genetic data on ~500,000 UK Biobank participants. bioRxiv 166298. **https://doi.org/10.1101/166298**
8. CARDIoGRAMplusC4D Consort., Deloukas P, Kanoni S, Willenborg C, Farrall M, et al. 2013. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat. Genet.* 45:25–33
9. **Castel SE, Cervera A, Mohammadi P, Aguet F, Reverter F, et al. 2017. Modified penetrance of coding variants by *cis*-regulatory variation shapes human traits. bioRxiv 190397. https://doi.org/10.1101/190397**
10. Ceyhan-Birsoy O, Machini K, Lebo MS, Yu TW, Agrawal PB, et al. 2017. A curated gene list for reporting results of newborn genomic sequencing. *Genet. Med.* 19:809–18
11. Chatterjee N, Shi J, García-Closas M. 2016. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.* 17:392–406
12. Cho YS, Chen C-H, Hu C, Long J, Ong RTH, et al. 2011. Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. *Nat. Genet.* 44:67–72
13. Coron. Artery Dis. (C4D) Genet. Consort. 2011. A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. *Nat. Genet.* 43:339–44
14. Cruchaga C, Haller G, Chakraverty S, Mayo K, Vallania FLM, et al. 2012. Rare variants in *APP*, *PSEN1* and *PSEN2* increase risk for AD in late-onset Alzheimer's disease families. *PLOS ONE* 7:e31039
15. Deo RC, Reich D, Tandon A, Akylbekova E, Patterson N, et al. 2009. Genetic differences between the determinants of lipid profile phenotypes in African and European Americans: the Jackson Heart Study. *PLOS Genet.* 5:e1000342
16. Diabetes Genet. Initiat. Broad Inst. Harv. MIT, Lund Univ., Novartis Inst. BioMed. Res. 2007. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 316:1331–36
17. **Dudbridge F. 2013. Power and predictive accuracy of polygenic risk scores. *PLOS Genet.* 9:e1003348**
18. Erdmann J, Grosshennig A, Braund PS, König IR, Hengstenberg C, et al. 2009. New susceptibility locus for coronary artery disease on chromosome 3q22.3. *Nat. Genet.* 41:280–82
19. Evangelou E, Warren HR, Mosen-Ansorena D, Mifsud B, Pazoki R, et al. 2017. Genetic analysis of over one million people identifies 535 novel loci for blood pressure. bioRxiv 198234. **https://doi.org/10.1101/198234**
20. Evans DM, Visscher PM, Wray NR. 2009. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Hum. Mol. Genet.* 18:3525–31
21. **Flannick J, Johansson S, Njølstad PR. 2016. Common and rare forms of diabetes mellitus: towards a continuum of diabetes subtypes. *Nat. Rev. Endocrinol.* 12:394–406**
22. Franceschini N, Carty C, Bůzková P, Reiner AP, Garrett T, et al. 2011. Association of genetic variants and incident coronary heart disease in multiethnic cohorts: the PAGE study. *Circ. Cardiovasc. Genet.* 4:661–72
23. Franke A, McGovern DPB, Barrett JC, Wang K, Radford-Smith GL, et al. 2010. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* 42:1118–25
24. **Fuchsberger C, Flannick J, Teslovich TM, Mahajan A, Agarwala V, et al. 2016. The genetic architecture of type 2 diabetes. *Nature* 536:41–47**

2. Describes a simulation-based methodology to directly measure the genetic architecture of type 2 diabetes.

9. Detects the presence of widespread regulatory variants modifying the penetrance of Mendelian disease alleles.

17. Analyzes the predictive power of PRSs to predict disease risk.

21. Reviews the concept of disease subtypes existing on a spectrum from Mendelian to complex.

24. Uses the methodology of Agarwala et al. (2) to measure the genetic architecture of type 2 diabetes.

25. Garcia CK. 2001. Autosomal recessive hypercholesterolemia caused by mutations in a putative LDL receptor adaptor protein. *Science* 292:1394–98
26. Himes P, Kauffman TL, Muessig KR, Amendola LM, Berg JS, et al. 2017. Genome sequencing and carrier testing: decisions on categorization and whether to disclose results of carrier testing. *Genet. Med.* 19:803–8
27. Int. Consort. Blood Press. Genome-Wide Assoc. Stud. 2011. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* 478:103–9
28. Int. Mult. Scler. Genet. Consort. (IMSGC). 2010. Evidence for polygenic susceptibility to multiple sclerosis—the shape of things to come. *Am. J. Hum. Genet.* 86:621–25
29. Int. Schizophr. Consort. 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460:748–52

30. Demonstrates that PRS quantiles can have equal predictive power to monogenic variants.

30. Khera AV, Chaffin M, Aragam K, Emdin CA, Klarin D, et al. 2017. Genome-wide polygenic score to identify a monogenic risk-equivalent for coronary disease. bioRxiv 218388. https://doi.org/10.1101/218388

31. Describes a methodology of stratifying patients by PRS quantiles to predict genetic risk.

31. Khera AV, Emdin CA, Drake I, Natarajan P, Bick AG, et al. 2016. Genetic risk, adherence to a healthy lifestyle, and coronary disease. *N. Engl. J. Med.* 375:2349–58
32. Khera AV, Kathiresan S. 2017. Genetics of coronary artery disease: discovery, biology and clinical translation. *Nat. Rev. Genet.* 18:331–44

33. Outlines the importance of identifying complex disease risk as homogeneous or heterogeneous.

33. Khera AV, Kathiresan S. 2017. Is coronary atherosclerosis one disease or many? Setting realistic expectations for precision medicine. *Circulation* 135:1005–7
34. Kooperberg C, LeBlanc M, Obenchain V. 2010. Risk prediction using genome-wide association studies. *Genet. Epidemiol.* 34:643–52
35. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, et al. 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467:832–38
36. Lehrman M, Schneider W, Sudhof T, Brown M, Goldstein J, Russell D. 1985. Mutation in LDL receptor: Alu-Alu recombination deletes exons encoding transmembrane and cytoplasmic domains. *Science* 227:140–46
37. Lerman C, Narod S, Schulman K, Hughes C, Gomez-Caminero A, et al. 1996. BRCA1 testing in families with hereditary breast-ovarian cancer: a prospective study of patient decision making and outcomes. *JAMA* 275:1885–92
38. Li Z, Chen J, Yu H, He L, Xu Y, et al. 2017. Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. *Nat. Genet.* 49:1576–83
39. Machiela MJ, Chen C-Y, Chen C, Chanock SJ, Hunter DJ, Kraft P. 2011. Evaluation of polygenic risk scores for predicting breast and prostate cancer risk. *Genet. Epidemiol.* 35:506–14
40. Márquez-Luna C, Loh P-R, South Asian Type 2 Diabetes (SAT2D) Consort., SIGMA Type 2 Diabetes Consort., Price AL. 2017. Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genet. Epidemiol.* 41:811–23
41. Martin AR, Costa HA, Lappalainen T, Henn BM, Kidd JM, et al. 2014. Transcriptome sequencing from diverse human populations reveals differentiated regulatory architecture. *PLOS Genet.* 10:e1004549

42. Demonstrates the impact of genetic ancestry on the accuracy of current PRS methods.

42. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, et al. 2017. Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.* 100:635–49
43. Medina-Gomez C, Felix JF, Estrada K, Peters MJ, Herrera L, et al. 2015. Challenges in conducting genome-wide association studies in highly admixed multi-ethnic populations: the Generation R Study. *Eur. J. Epidemiol.* 30:317–30
44. Mega JL, Stitziel NO, Smith JG, Chasman DI, Caulfield M, et al. 2015. Genetic risk, coronary heart disease events, and the clinical benefit of statin therapy: an analysis of primary and secondary prevention trials. *Lancet* 385:2264–71
45. Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, et al. 1994. A strong candidate for the breast and ovarian cancer susceptibility gene *BRCA1*. *Science* 266:66–71
46. Mitchell JJ, Capua A, Clow C, Scriver CR. 1996. Twenty-year outcome analysis of genetic screening programs for Tay-Sachs and β-thalassemia disease carriers in high schools. *Am. J. Hum. Genet.* 59:793–98

47. Müller C. 2009. Xanthomata, hypercholesterolemia, angina pectoris. *Acta Med. Scand.* 95:75–84
48. Natarajan P, Young R, Stitziel NO, Padmanabhan S, Baber U, et al. 2017. Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting. *Circulation* 135:2091–2101
49. Nikpay M, Goel A, Won H-H, Hall LM, Willenborg C, et al. 2015. A comprehensive 1,000 Genomes–based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* 47:1121–30
50. Pharoah PDP, Antoniou AC, Easton DF, Ponder BAJ. 2008. Polygenes, risk prediction, and targeted prevention of breast cancer. *N. Engl. J. Med.* 358:2796–803
51. Popejoy AB, Fullerton SM. 2016. Genomics is failing on diversity. *Nature* 538:161–64
52. Rietveld CA, Medland SE, Derringer J, Yang J, Esko T, et al. 2013. GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* 340:1467–71
53. Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, et al. 1989. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* 245:1066–73
54. Ripke S, O'Dushlaine C, Chambert K, Moran JL, Kähler AK, et al. 2013. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat. Genet.* 45:1150–59
55. Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, et al. 2011. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat. Genet.* 43:1066–73
56. Rommens J, Iannuzzi M, Kerem B, Drumm M, Melmer G, et al. 1989. Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science* 245:1059–65
57. Schizophr. Psychiatr. Genome-Wide Assoc. Study (GWAS) Consor. 2011. Genome-wide association study identifies five new schizophrenia loci. *Nat. Genet.* 43:969–76
58. Schizophr. Work. Group Psychiatr. Genom. Consort. 2014. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511:421–27
59. Schunkert H, König IR, Kathiresan S, Reilly MP, Assimes TL, et al. 2011. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat. Genet.* 43:333–38
60. Scott RA, Scott LJ, Mägi R, Marullo L, Gaulton KJ, et al. [Diabetes Genet. Replication Meta-Anal. (DIAGRAM) Consort.]. 2017. An expanded genome-wide association study of type 2 diabetes in Europeans. *Diabetes* 66:2888–902
61. Sim X, Ong RT-H, Suo C, Tay W-T, Liu J, et al. 2011. Transferability of type 2 diabetes implicated loci in multi-ethnic cohorts from Southeast Asia. *PLOS Genet.* 7:e1001363
62. Soria LF, Ludwig EH, Clarke HR, Vega GL, Grundy SM, McCarthy BJ. 1989. Association between a specific apolipoprotein B mutation and familial defective apolipoprotein B-100. *PNAS* 86:587–91
63. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, et al. 2010. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* 42:937–48
64. Stahl EA, Wegmann D, Trynka G, Gutierrez-Achury J, Do R, et al. 2012. Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat. Genet.* 44:483–89
65. van der Harst P, Verweij N. 2018. Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. *Circ. Res.* 122:433–43
66. Wei Z, Wang K, Qu H-Q, Zhang H, Bradfield J, et al. 2009. From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. *PLOS Genet.* 5:e1000678
67. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, et al. 2008. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat. Genet.* 40:161–69
68. Wojcik G, Graff M, Nishimura KK, Tao R, Haessler J, et al. 2017. Genetic diversity turns a new PAGE in our understanding of complex traits. bioRxiv 188094. **https://doi.org/10.1101/188094**
69. Won H-H, Natarajan P, Dobbyn A, Jordan DM, Roussos P, et al. 2015. Disproportionate contributions of select genomic compartments and cell types to genetic risk for coronary artery disease. *PLOS Genet.* 11:e1005622
70. Wooster R, Bignell G, Lancaster J, Swift S, Seal S, et al. 1995. Identification of the breast cancer susceptibility gene *BRCA2*. *Nature* 378:789–92

Annual Review of Genomics and Human Genetics

Volume 19, 2018

Contents

Errata

An online log of corrections to *Annual Review of Genomics and Human Genetics* articles may be found at http://www.annualreviews.org/errata/genom