

Hypotheses of How Common Variants Create Risk for Psychiatric Disorders

Naomi R. Wray

Abstract

Over the last 15 years, genome-wide association studies have demonstrated that psychiatric disorders, like other common diseases, are highly polygenic. The traditional toolbox of approaches used to characterize functional effects of causal genetic variants has been constructed for monogenic disease, where a single variant is associated with a high probability of disease risk at some point in the lifetime. This toolbox has limited utility for studying risk variants of small effect. To develop new experimental paradigms requires a deep understanding of polygenic architectures. First, many risk variants have small effect, which means most people with each variant do not have the disease associated with the risk. Disease is associated with carrying a high burden of risk variants, implying that the function of each risk variant is dependent on its genetic context. Second, each person diagnosed with a common disease is expected to carry a unique, or almost unique, portfolio of risk variants. Yet despite this heterogeneous genetic architecture, diagnostic classes do have some biological validity. Third, as observed for other common diseases, we expect there to be multiple pathways that contribute to increased risk of disease across many cell types and impacting over the lifespan. The key question then is how to penetrate this polygenic complexity.

Introduction

Genome-wide association studies (GWASs) have identified hundreds of loci associated with psychiatric disorders. A key question for the field is how to translate these findings into clinical utility. Traditional laboratory approaches are tailored for studying rare variants of very large effect associated with rare diseases. Even then, when large effect mutations have been identified, the journey from discovery to actionable outcomes for prevention or treatment can be long. For example, causal variants in *HTT*, *SOD1*, *BRCA1*, and

APOE associated with Huntington disease, amyotrophic lateral sclerosis, breast cancer and Alzheimer disease, respectively, were all discovered around two decades ago. While these discoveries have furthered mechanistic understanding of their diseases, translation to new treatments has been limited. New approaches are starting to be developed that aim to penetrate the complexity of how a polygenic architecture creates risk for psychiatric disorders. Here, I explore the fundamental concepts of a polygenic architecture, contrast psychiatric disorders with other polygenic diseases and traits, and consider experimental designs that embrace this polygenicity.

The Different Facets of Polygenicity

The face value description of a polygenic disease, that thousands of DNA variants contribute to risk of disease, seems straightforward. But what does this population description mean for individual people within the population? How do we reconcile the key observations of thousands of risk loci, increased risk in relatives (reflecting the heritability of the disease), and the fact that even common diseases only affect a minority of people? Key to understanding the nature of polygenic disease are some fundamental concepts of quantitative genetics, summarized in this section; for a more in-depth discussion, see Baselmans et al (2021). The term “polygenic” covers many different genetic architectures defined as the number of risk variants, the population frequency of those risk variants, and their effect sizes. How does the polygenicity of psychiatric disorders contrast to that of other diseases and disorders, and can we learn anything from this?

Each Person Has a Unique Portfolio of Risk Variants

A key feature of polygenic disease is that each person with the same disease diagnosis is expected to have a unique portfolio of risk variants, some of which may include variants of relatively large effect. In fact, it is now recognized that even classical monogenic diseases are more accurately classified as polygenic diseases that include a very large effect variant. For example, consider Huntington disease, which is the textbook example of monogenic autosomal dominant disease. All those affected have an expanded copy number of a trinucleotide repeat near the gene *HTT*, with a higher expansion number associated with age of onset. Nonetheless, the Genetic Modifiers of Huntington Disease Consortium has identified 21 independent single nucleotide polymorphisms (SNPs) associated with age of onset from a sample size of 15,000 cases (Lee et al. 2019b), thus illustrating how genetic context of the *HTT* mutation is important. Moreover, genetic architecture signals from these data are consistent with an expectation that more associated SNPs will be identified with larger sample size. In studies of large effect variants associated with common disease (e.g.,

familial hypercholesterolemia variants associated with heart disease, *BRCA1* or *BRCA2* variants associated with breast or ovarian cancer, and colorectal cancer Lynch syndrome), high estimated risk constructed from the polygenic burden is associated with earlier age of onset or greater severity of disease compared to monogenic variant carriers with low estimated polygenic risk (Fahed et al. 2020). Examples more relevant to psychiatry are the copy number variants such as deletions in 1q21.1, 15q13.3, or 16p11.2 seen in the context of epilepsy, autism, mental retardation, or schizophrenia diagnoses. The association with multiple diagnoses likely reflects the genetic background in which the deletion events have occurred. An important consideration for polygenic disease is that all people carry a burden of risk variants for each disease, but only those that carry a high burden of risk variants (and other risk factors) will develop symptoms that lead to disease diagnosis. A polygenic model is consistent with biological robustness since the vast majority of risk variant portfolios carried by individuals does not increase risk of disease.

Additive on Liability Scale but Nonadditive on the Disease Status Scale

Modeling polygenicity suggests that complex biological interactions govern the relationship between polygenic variation and disease. The liability threshold model (LTM) is a working model for common polygenic disease. The model was developed over 70 years ago (Falconer 1965) and is based on the infinitesimal model (all DNA variants contribute a small effect to each trait). The last decades have not provided empirical data to reject the LTM as a useful working model even though estimates from genetic architecture modeling suggest only 1–5% of common SNPs are likely to contribute to each disease (Zeng et al. 2021), but estimates are higher for psychiatric disorders than other diseases discussed below. Indeed, the estimated number of contributing variants is still exceptionally high (tens of thousands) and so the conceptual utility of the model is retained (and in fact the utility of the LTM holds even if there are as few as ten risk variants). The LTM is just one of a suite of models that all imply the same shape of relationship between polygenic burden and risk of disease (Slatkin 2008), but it is usually the model of choice because of its mathematical tractability. The shape of this relationship is very nonlinear, which is the only way to reconcile the two key parameters defining disease: heritability and relatively low lifetime risk of disease. The nonlinear relationship between polygenic burden and disease risk is more nonlinear for diseases that are less common and/or that have heritability. Hence, interaction effects are expected for the biological function of risk variants acting together. However, since all people carry a different portfolio of risk loci, the traditional tools for studying interaction which investigate interactions between only a few loci (usually only two) is unlikely to be useful. The interaction is on a scale of so many variants and many different combinations of variants that disease modeling focuses on additivity to liability disease. Hence a key

concept is that polygenic disease implies additivity of effects on the liability to disease scale and massive nonadditivity on the risk of disease scale. In a recent review (Baselmans et al. 2021), I tried to explain these concepts in depth in the context of psychiatric disorders.

Are Psychiatric Disorders More Polygenic than Other Common Diseases and Disorders?

In the post-GWAS era, many methods have been developed to evaluate genetic architecture from the distribution of SNP effects reported in GWAS summary statistics. For example, for any trait, the SbayesS (summary-based Bayes method estimating the S-parameter) provides estimates of the contribution of common SNP variation to the trait (SNP-based heritability), the proportion of all SNPs that contribute to the trait (polygenicity), and the correlation between minor allele frequency and effect size (“S,” an indicator of selection pressure) (Zeng et al. 2021). These values can be meaningfully compared across traits when the same common SNP set is used. In an SbayesS analysis applied to 18 common diseases (Figure 9.1), GWAS summary statistics for both schizophrenia and bipolar disorder provided strong evidence for negative selection S (~ -0.7). Notably, this was similar to, but not greater than, the estimates for other diseases. It also showed exceptionally high polygenicity (consistent with estimates from other methods, e.g., Ripke et al. 2013). The average polygenicity across the 44 traits studied was $\sim 1\%$, compared to 5% for schizophrenia and 3% of bipolar disorder, which were significantly higher than for other diseases (Zeng et al. 2021).

It could be argued that high levels of polygenicity for disorders of the brain are to be expected; the brain is such an important organ that many backup pathways exist, consistent with biological robustness. However, it is worth considering whether any artifact could contribute to the observation of the exceptionally high levels of polygenicity. Psychiatric disorders are always described as being very heterogeneous. Thus, we must consider the possibility that empirical estimates of polygenicity are a reflection of the same disease labels being attached to biologically distinct disorders. To illustrate this, Wray and Maier (2014) considered the following toy example: Imagine that a disease labeled A with population lifetime risk of $\sim 1\%$ is actually comprised of two biologically distinct diseases, B and C, each with a population lifetime risk of 0.5% that are impossible to differentiate based on clinical presentation. Assume diseases B and C have a heritability of 80% but being biologically distinct have independent risk variants. The heritability of composite disease A would be estimated to be substantial ($\sim 65\%$), using the standard approach of increased risk observed in first degree relatives (the estimate would be lower if increased risks from more distant relatives are used). In a GWAS, however, the composite disease A has much reduced power compared to the biologically distinct diseases of B and C. If the true SNP-based heritabilities of diseases B

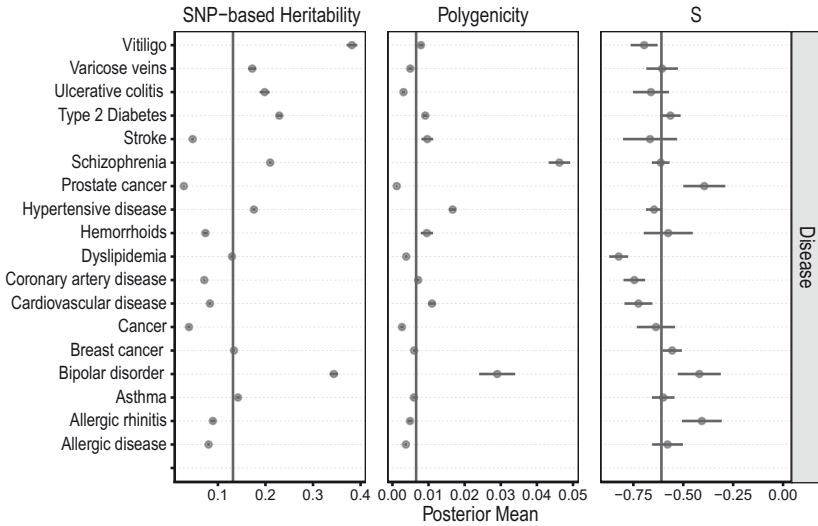


Figure 9.1 Genetic architecture parameters for 18 common diseases, showing the posterior means (dots) and standard errors (horizontal bars) of the parameters for each trait. Data from Zeng et al. (2021; see Figure 2) estimated from GWAS summary statistics using SBayesS. SNP-based heritability is the proportion of variance in liability attributable to common SNPs. The polygenicity parameter is an estimate of the proportion of common SNPs associated with the disease. S is a selection parameter and represents the relationship between allele frequency and effect size (negative values imply selection against the disease).

and C were the same, then the SNP-based heritability of disease A would be estimated as 50% of the value, and it would make sense to infer that estimates of polygenicity would be higher for A than if estimated separately for B and C.

As we consider the genetic architecture of psychiatric disorders, it is worth keeping in mind the potential heterogeneity of biological disease that underpins a diagnosis applied to symptoms. We should not forget to challenge basic assumptions made in our foundational analyses. That said, the toy example described is extreme and, in fact, implies nonadditivity on the liability scale of disease A, which would likely give unusual properties of the joint distribution of GWAS effects which have not been observed. Data sets which include deep phenotyping are needed to provide an evidence base to conclude if the higher polygenicity estimated for psychiatric disorders than for other diseases is inherent or a reflection of biological heterogeneity not reflected in current nosology.

The Relationship between Polygenicity and Disorder

The complexity and polygenicity revealed by genetic studies of psychiatric disorders raise the question whether they are true disorders as opposed to

difficult-to-differentiate collections of multiple conditions, or even behavioral or cognitive phenotypes best described as continua. Considering the structure of the genetic findings pertaining to psychiatric disorders as well as comparing this structure to other disorders suggest that these conditions have the properties of disorders, however heterogeneous.

What Is a Disorder?

The toy example above, where two biologically distinct diseases could not be distinguished in clinical settings, points to an important question: What is a disorder label, and is it meaningful given the extensive discussions of heterogeneity of presentation? In real life, it seems likely that if biological heterogeneity underpins a disease, then this would happen at the level of biologically correlated diseases receiving the same diagnosis rather than biologically independent diseases. In the context of psychiatric disorders, it is notable that estimates of genetic correlations between data sets of the same disorder are consistently higher than estimates of genetic correlations between different psychiatric disorders (Cross-Disorder Group of the Psychiatric Genomics Consortium et al. 2013), implying that the standard nosology does have biological support. Moreover, estimates of genetic correlations between data sets of the same disorder, which are expected to be 1, are notably less than 1 for major depression and attention deficit hyperactivity disorders (Cross-Disorder Group of the Psychiatric Genomics et al. 2013), likely a reflection on the recognized heterogeneity in diagnosis within and between data sets.

Genetic analyses of rare and common variants in developmental delay and autism spectrum disorder clearly demonstrate that genetic architecture and diagnostic labels are not perfectly aligned. For example, girls with a diagnosis of autism are more likely to harbor a large effect copy number variant than boys, despite a higher rate of diagnosis in boys. This is explained by the female protective effect, such that boys who carry the same large effect copy number variant are more likely to be diagnosed with of developmental delay (Robinson et al. 2014). Within those diagnosed with developmental delay, autistic behavior is found to be significantly associated with the polygenic score of autism ($p=2.5 \times 10^{-4}$), and severity of intellectual disability is significantly associated ($p=4.0 \times 10^{-3}$) with the polygene score educational attainment (Niemi et al. 2018).

Learning from Other Common Complex Diseases

Practicing psychiatrists always emphasize the heterogeneity in clinical presentation associated with each diagnostic category and that few individuals fit the classic textbook definitions of disorder diagnoses. Since diagnosis within psychiatry is based on interview criteria rather than any gold standard biological biomarker, there may be a perception that this heterogeneity is not present in other disorders. In fact, heterogeneity in presentation seems to be the norm

in all common diseases. Let us take type II diabetes as an example. Arguably, more is understood about the functional impairment of insulin secretion from the pancreas than is known about functional causes or consequences of psychiatric disorders within the brain. Yet multiple pathways are known to contribute to type II diabetes (Udler et al. 2019), including pathways of insulin secretion, insulin resistance, and dyslipidemia (Figure 9.2). One appealing hypothesis has been to use these pathways to allocate people to diabetes subtypes; however, recent research rejects this as being naïve (McCarthy 2017). To understand a polygenic disorder with many contributing pathways, McCarthy (2017) introduced the concept of the palette model as a visual analogy. An artist’s paint palette comprises a set of primary colors (representing biological mechanisms), and each person has a personal mix of these colors, allowing many combinations to reach the diagnosis (McCarthy 2017). It is unlikely for a person’s color to be dominated by a single color, so by analogy few people can be allocated to specific subtypes defined by these pathways. This conceptualization is consistent with the long-held view of polygenicity and can be equally relevant and helpful when representing psychiatric disorders. Learning from the type II diabetes research community, it would be unwise to invest heavily in research on a pathway-specific model for psychiatric disorders. If pathway-specific considerations are relevant in the identification of personalized treatments, then these

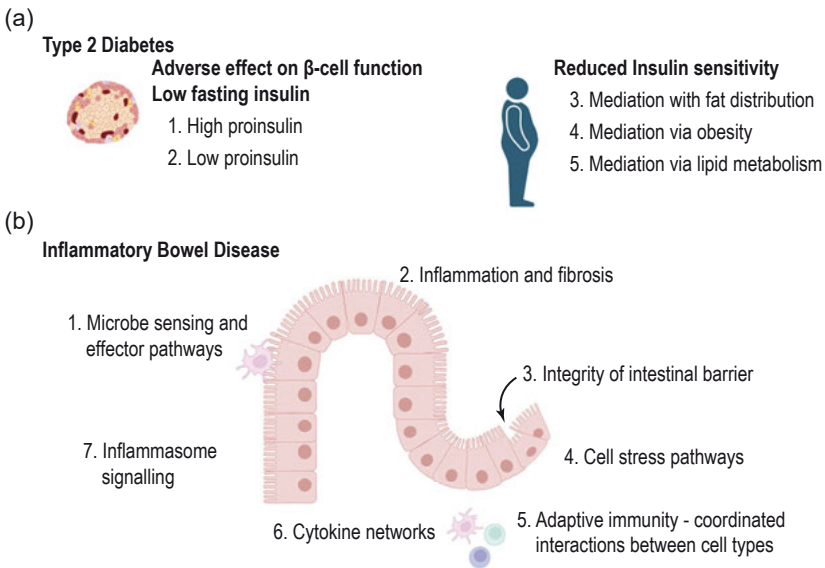


Figure 9.2 Pathway paradigms highlighted by genetics: (a) type II diabetes (Udler et al. 2019) and (b) inflammatory bowel diseases (Graham and Xavier 2020). Created with BioRender.com.

will likely emerge through research that does not specifically set out with this goal in mind.

Another useful benchmark for psychiatric disorders is to consider inflammatory bowel diseases, for which much more is known about the functional mechanisms underlying genetic associations. Graham and Xavier (2020) show seven pathways that contribute to risk of these diseases (summarized in Figure 9.2). We anticipate a similar complexity for psychiatric disorders, implying that the heterogeneous diagnostic construct is likely unavoidable. Research clarifying the contributing mechanistic pathways is emerging and will be important in understanding mechanistic pathways, but it is unlikely that these will contribute to disease subtyping.

Common Diseases Are More Complex than Implied by the Omnigenic Model

For over 100 years the infinitesimal model has been used to conceptualize polygenic traits and diseases, and its utility has not been rejected despite extensive empirical testing. In 2017, the omnigenic model was introduced as a conceptual framework to offer a mechanistic explanation for why so many variants spread across the genome can be responsible for the genetic variation between people observed for complex traits and common disease (Boyle et al. 2017). In brief, the underlying logic is that variants with *cis* effects on genes considered “peripheral” to disease pathways perturb the regulation of a smaller class of core genes via *trans*-regulatory networks. In this way, these “peripheral” genes are important to unravel the biology of a trait only because of their *trans*-regulatory influences on core genes. The model considers the disease-specific core genes as the most important for disease-specific research. A common response to the model was expressed in writing by Nancy Cox (2017):

[M]y first reaction to the swirling discussions of the paper was to wonder what the excitement was all about. How would anyone who understands the nature of polygenic liability and is aware of what we have been learning about the contribution of regulatory variation to common disease heritability think otherwise?

The terms core, key, driver, and peripheral genes were part of the standard vernacular for mechanistic interpretation of genetic studies of polygenic disease prior to the introduction of the omnigenic model. A model that is a conceptual advance over and above previous conceptualizations of polygenicity requires testable hypotheses. Without testable hypotheses the omnigenic model is not distinguishable from conceptual thinking about the genetic basis of complex traits of previous decades.

Sinnott-Armstrong et al. (2021) set out to test one major component of the model, namely the identities and roles of core genes. Recognizing that this cannot be achieved with data from common disease, they took a step back in complexity and looked at molecular traits, where a strong relationship between

trait variation and DNA variation can be expected and for which the omnigenic model is more plausible. They chose three blood traits from the UK Biobank: urate, IGF-1, and testosterone. As has been shown by others for blood-derived measures (e.g., Ruth et al. 2020), many of the lead association signals are very highly significant and interpretable in terms of the known physiology of the traits. Hence, the genes identified can rightly be considered core genes that reflect core pathways. Despite this level of interpretability, these traits are also highly polygenic, with the association signal attributed to core pathways explaining < 10% of the SNP-based heritability. Moreover, Sinnott-Armstrong et al. (2021) conclude that testing the second part of the omnigenic model (that most of the genetic association passes through *trans*-regulatory networks) is intractable given the sample sizes needed, even for these molecular traits. From their work, it seems clear that the omnigenic model is too simplistic to explain common disease, which likely comprises the overlaying of hundreds of these simple molecular traits. There is no need to abandon the existing models of polygenic disease, which better describes the complexity of common disease (Wray et al. 2018b). One reason to speak out against the omnigenic model is that it has been used as a call-to-arms for the identification of core genes for common disease and for reinvigoration of experimental paradigms used for monogenic disease. New experimental designs that embrace the polygenic architecture of disease are needed to further understand common diseases, including psychiatric disorders.

Experimental Designs that Embrace Polygenicity

Drawing on some of the key observations of polygenic disease discussed above, let us consider how these can inform new designs for understanding functional mechanisms of complex diseases:

1. Common diseases are highly epistatic on the scale of disease risk, which means that the function of the variant is dependent on its genetic context. In most genetic backgrounds, the risk variant does not lead to a disease status.
2. Each person diagnosed with a common disease is expected to carry a unique or almost unique portfolio of risk variants. Despite this, individuals share a diagnosis that has some biological validity (i.e., higher genetic correlations between cohorts of the same disorder than between cohorts with different disorders).
3. As observed for other common diseases (e.g., type II diabetes and inflammatory bowel diseases), we expect there to be multiple pathways that contribute to an increased risk of disease across many cell types and impacting over the lifespan.

The question is: How can we penetrate this polygenic complexity?

It is likely that different subsets of risk variants act in different cell types at different time points throughout development and aging. The consequence of any one such subset could be very subtle. For example, some risk variants could set up a vulnerable cellular infrastructure simply through the speed of cell differentiation, the composition of cell types, or cell morphology and function. A vulnerable basic infrastructure could then compound the impact of other genetic risk variants associated with synaptic pruning during adolescence and these effects could be enhanced through environmental stress exposures. While studying one gene at a time (as is traditionally done for monogenic disease) will add value to our knowledge base, I believe that to impact the lives of those affected by psychiatric disorders more quickly (both now and in the future), experimental designs that embrace the polygenicity better must be developed. Below we consider two relevant paradigms.

Extreme of Polygenic Score Design

An experimental design which acknowledges that all people carry risk variants, that those affected carry a higher burden, and that each person carries a unique portfolio requires us to study the cellular/organoid phenotypes of those at the extremes of the polygenic score (PGS) distribution. Based on current GWASs, there is a fortyfold difference in risk between those in the top centile versus bottom centile of the PGS distribution (Trubetskoy et al. 2022). Of course, this approach requires collection of genetic data on very large cohorts of individuals to identify those at the extremes, and the recontacting of participants is likely needed to generate cell lines. The extreme PGS approach has been pioneered in the Brennand lab (Dobrindt et al. 2021) selecting six high and six low PGS lines from existing population-based cell lines (from healthy individuals). Going forward it makes sense to contrast high/low PRS lines using those who do and do not have disease. Given that each individual has a different portfolio of risk variants, cellular phenotypes need to be associated with all (or a high proportion) of those in the group “high PRS and disease” compared to those with low PRS. Induced pluripotent stem cell phenotypes are starting to be associated with clinical phenotype (e.g., circadian rhythm phenotypes in the context of bipolar disorder) (Sanghani et al. 2021).

Integration of GWAS Results with Single-Cell RNA Sequencing

The GWAS era has generated a host of post-GWAS analyses in which many different types of independently collected reference data can be integrated with the GWAS associations linked via SNPs or SNPs annotated to genes (Pasaniuc and Price 2017). At the level of the gene unit, integration allows reference data sets to be generated in animal models. For many other common diseases much is known about the most relevant cell types within specific tissues most affected by the disease (e.g., see Figure 9.2). However, for psychiatric disorders where

the prefrontal cortex is recognized as the brain region of relevance, knowledge of the role of specific cell types is less clear. A number of methods have been proposed that integrate GWAS gene associations with genes enriched with expression in specific cell types (Timshel et al. 2020; Zhu et al. 2020). Disassociation of single cells from human brain tissue is more difficult than for other tissues, so RNA sequencing data sets for brain tend to be from single nuclei if using human brain and only from the whole cell if using mouse brain (which, of course, may not have the complete suite of cell types or cell subtypes present in humans). Application of these methods to diseases where specific, relevant cell types have been identified by histology and other research methods provides verification of these approaches. For example, GWAS for inflammatory bowel disease point to cells in the colon and ileum, whereas for diverticular disease (sac-like protrusions of the colon sigmoid) they point to cells in the colon sigmoid, consistent with known pathology (Wu et al., submitted). When these methods are applied to schizophrenia, pyramidal neurons are identified from both human and mouse tissue, with the evidence strengthening with larger GWASs (Trubetskoy et al. 2022). A limitation of these studies is that they are dependent on the cell types present in the data sets. More reference data sets are needed to build a more complete picture of the cell types in which the genetic risk factors likely operate.

Conclusion

Psychiatric disorders, like other common diseases, have a polygenic genetic architecture. This implies that all individuals likely carry thousands of risk variants for each disease, but those most vulnerable carry a higher burden, each with a unique portfolio. Under this architecture, the consequences of each associated variant or gene seems irrelevant since disease only results when the variants are present in the context of other genetic or nongenetic risk factors. New experimental paradigms are needed to study sets of genetic variants jointly.

Acknowledgment

I acknowledge funding from the Australian National Health & Medical Research Council 1113400, 1173790.

