# 3

# Data Collection

## Next Steps in Psychiatric Genetics

Elise B. Robinson, Heesu Kim, Daniel Weiner,
Alicia Ljungdahl, and Stephan J. Sanders

**Abstract**

Decades of twin and family studies have revealed the high heritability of neuropsychiatric disorders, suggesting an important causal role for genetic factors. Over the last decade, genetic studies have linked hundreds of genes and genetic loci to neuropsychiatric disorders as well as to behavior and cognition more broadly. Most large lists of genetic loci associated with disorders have been generated by consortia. These consortia aggregate data from similarly focused studies from around the globe, with historic emphasis on certain areas. This chapter explores what is necessary to deliver refined genetic insights rigorously and efficiently.

### Identification of Genetic Variants in Neuropsychiatric Disorders

Any two unrelated individuals have about 3 million genetic variants (0.1% of the genome, 1 in a 1,000 bp) that differ between them. The majority of these are common single nucleotide polymorphisms (SNPs); that is, single base pairs of DNA that vary in at least 1% of the population. The remaining variation can be rare (<1% of the population) and/or larger (e.g., insertions, deletions, structural variants). Common variants can be detected using a genotyping array. The role of common variants in a disorder is assessed through genome-wide association studies (GWASs), which identify specific SNPs that are more common than expected in either cases or controls. Rare variants in protein-coding regions can be detected using whole-exome sequencing (WES). Due to their rarity, these variants typically must be grouped together (e.g., within the same gene) to assess whether their frequency differs from expectation between cases and controls. Whole-genome sequencing (WGS) enables the detection of variants across the frequency and size spectra, outperforming

From "Exploring and Exploiting Genetic Risk for Psychiatric Disorders," edited by Joshua A. Gordon and Elisabeth B. Binder. Strüngmann Forum Reports, vol. 31, Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262547383 (paperback) 9780262377423 (pdf)

both genotyping arrays and WES, but it is expensive to conduct. Even with WGS data, moderately rare variants (0.1–2% population frequency) and rare noncoding variants associated with human disorders remain largely undetectable with current cohort sizes and methodologies (Sanders et al. 2017; Singh et al. 2022; Werling et al. 2018). Similarly, only the common variants and rare protein-coding variants with the highest effect sizes for neuropsychiatric disorders have been detected so far.

## Natural Selection and Genomic Architecture

The choice of genetic characterization strategy often reflects the known or hypothesized genomic architecture of a disorder or trait, which is primarily influenced by natural selection. Strong selective pressure, as is present in early-onset neurodevelopmental disorders such as autism spectrum disorder (ASD) or intellectual disability (ID) (Power et al. 2013), imposes a limit on the effect size that a common variant can impart (Devlin et al. 2011), necessitating huge cohorts for GWAS discovery. Accordingly, genetic locus discovery for these disorders has focused on finding rare variants with high effect sizes, e.g., by the analysis of trios composed of both parents and an affected child to identify newly arising *de novo* mutations (Sebat et al. 2007). While each variant is individually rare, many different genes and many different disruptive variants within each gene can lead to symptoms. Combining all of these genes and variants, at least 30%–40% of ID cases and 10% of ASD cases can be explained by a single rare genetic event (Bishop et al. 2021). By comparison, it is rare for a single genetic variant to explain a meaningful fraction of individual risk for late-onset disorders, such as bipolar illness or major depressive disorder (MDD), and locus discovery for this type of disorder has accordingly focused on common variants discoverable through GWAS (Howard et al. 2019; Mullins et al. 2021).

In Table 3.1, we present a set of recent landmark psychiatric genetics studies of developmental (ASD, ID/NDD, and attention deficit hyperactivity disorder), mood (anxiety, MDD, bipolar disorder), and psychotic (schizophrenia) disorders. The studies listed reflect the largest collaborative efforts, in both genotyping and/or sequencing studies, for each outcome over the past five years.

In total, more than three million individuals participated in these studies, and more than 600 genes and genetic loci have been identified. Here, we discuss continuing data collection and association needs in neuropsychiatric statistical genetics, as gaps remain in both publicly available data collections and the published literature. We focus on three themes below: (a) completing discovery, (b) informing biology with genetic associations from across the frequency spectrum, and (c) improving global representation.

From "Exploring and Exploiting Genetic Risk for Psychiatric Disorders," edited by Joshua A. Gordon and Elisabeth B. Binder. Strüngmann Forum Reports, vol. 31, Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262547383 (paperback) 9780262377423 (pdf)

**Table 3.1** Largest and most recent genomic studies across seven major psychiatric disorders: attention deficit hyperactivity disorder (ADHD), autism spectrum disorder (ASD), intellectual disability/neurodevelopmental delay (ID/NDD), anxiety (ANX), major depressive disorder (MDD), bipolar disorder (BD), and schizophrenia (SCZ). Single nucleotide polymorphisms (SNP), de novo (dn), ultrarare (ur), missense (mis), protein truncating variant (PTV), copy number variant (CNV), loss of function (LoF). N includes all samples from discovery and replications, where applicable.

| Disorder | Study Design | Genetic Characterization | Variant | All cases (N) | All controls (N) | Total Subjects (N) | Significant loci/genes | Cohort | Reference |
|---|---|---|---|---|---|---|---|---|---|
| ADHD | Case-control | Genotyping | SNP | 4,208 | 32,222 | 36,430 | 27 | PGC, iPSYCH | Demontis et al. (2019) |
| | | WGS | rare recurrent | 205 | 670 | 875 | 1 | iPSYCH | Satterstrom et al. (2019) |
| ASD | Case-control | Genotyping | SNP | 18,381 | 27,969 | 46,350 | 5 | Danish iPSYCH | Grove et al. (2019) |
| | Parent-child trios, case control | WES | dnPTV, dnCNV, mis | 51,685 | 103,157 | 154,842 | 72 | ASC, SSC, SPARK, iPSYCH | Fu et al. (2022) |
| ID/NDD | Case control, parent-child trios | Genotyping | SNP | 7,715 | 10,726 | 18,441 | 0 | DDD, ADD | Niemi et al. (2018) |
| | Parent-child trios | WES | dn | 31,058 | 62,116 | 93,174 | 285 | GeneDX, DDD, RUMC | Kaplanis et al. (2020) |
| ANX | Case control, continuous trait | Genotyping | SNP, eQTL | 34,189 | 190,141 | 423,941 | 6 | Million Veterans | Levey et al. (2020) |
| | Case control | | SNP | 44,465 | 58,113 | 102,578 | 5 | UK Biobank | Purves et al. (2020) |
| MDD | Case control | Genotyping | SNP | 660,418 | 1,453,489 | 2,113,907 | 87 | UK Biobank, 23andMe, PGC | Howard et al. (2019) |
| | | | SNP | 15,771 | 178,777 | 194,548 | 1 | CONVERGE, TMDD, CKB | Giannakopoulou et al. (2021) |
| | | | CNV | 23,979 | 383,095 | 407,074 | 53 | UK Biobank | Kendall et al. (2019) |
| BD | Case control | Genotyping | SNP | 41,917 | 371,549 | 413,466 | 64 | PGC | Mullins et al. (2021) |
| | | | rare CNV | 6,353 | 8,656 | 15,009 | 0 | ICCBD | Charney et al. (2018) |
| | | WES | urPTV | 38,181 | 111,744 | 149,925 | 1 | BipEx, SCHEMA | Palmer et al. (2022) |
| SCZ | Case control | Genotyping | SNP | 69,369 | 236,642 | 306,011 | 270 | PGC | Trubetskoy et al. (2022) |
| | | | CNV | 21,094 | 20,227 | 41,321 | 8 | PGC | Marshall et al. (2017) |
| | Case control, parent-child trios | WES | dnPTV, mis | 24,248 | 97,322 | 121,570 | 10 | CHEMA, gnomAD | Singh et al. (2022) |



From "Exploring and Exploiting Genetic Risk for Psychiatric Disorders," edited by Joshua A. Gordon and Elisabeth B. Binder. Strüngmann Forum Reports, vol. 31, Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262547383 (paperback) 9780262377423 (pdf)

## Completing Discovery

In our opinion, there is no objective point at which genetic discovery activities for neuropsychiatric disorders will be definitively complete, short of deploying the latest sequencing technology worldwide. Here we present some ideas relevant to the question of whether and how to emphasize future gene discovery.

There are several reasons to believe that identification of risk-conferring variation should continue. First, across the disorders listed in Table 3.1, we continue to identify new loci and do not see evidence of plateauing. Those loci provide new biological inferences and are furthering our understanding of the genetic architecture of neuropsychiatric disease. Schizophrenia provides a useful example, as it is the psychiatric outcome most comprehensively characterized with GWAS (about 70,000 cases have yielded over 250 common variant loci; Trubetskoy et al. 2022) and exome sequencing: about 24,000 cases by SCHEMA[1] have yielded 10 genes from rare variations (Singh et al. 2022). Combining these loci allows the relationship between effect size and population frequency of associated variants to be plotted. In Figure 3.1, we have adapted SCHEMA's figure (Singh et al. 2022) to highlight the search space in which discovery activities remain statistically underpowered. As an example, variants with 0.001 frequency and an odds ratio of 2 cannot be identified with current sample size and will require case samples of approximately 250,000 to be detectable (and at least an equal number of controls). At that same frequency, variants with an OR of 1.5 would be detectable with 900,000 cases.

For schizophrenia genetics, if we achieved a target sample of 250,000 cases, there would be several benefits. First, we would have access to the class of risk variation described above: variants of lower frequency and modest effect size. We do not know the degree to which the biology implicated by that class will overlap with either (a) very rare variants of large effect or (b) common variants of small effect. Second, by increasing schizophrenia case sample size further, we will identify more carriers of specific risk factors, which will permit better powered follow-up studies. For example, 15 SCHEMA individuals (out of 25,000 cases) carried a risk variant in SETD1A, the gene found to have the largest effect size. Expanding this total would enable detailed genotype-phenotype analysis (Sanders et al. 2018), comparison of developmental trajectories (Wickstrom et al. 2021), and the development of valuable cell line resources (Khan et al. 2020). Of course, such analyses and resource development necessitate studies in which the participants are contacted at different stages. Third, as we discuss in detail below, genetic studies of neuropsychiatric disease have, to this point, predominantly included individuals of European ancestry. Additional data collection in schizophrenia is necessary to progress toward equitable polygenic risk score (PRS) performance across ancestral groups, among other imperatives (for further discussion, see Ronald et al., this volume).

[1] The Schizophrenia Exome Sequencing Meta-Analysis

From "Exploring and Exploiting Genetic Risk for Psychiatric Disorders," edited by Joshua A. Gordon and Elisabeth B. Binder. Strüngmann Forum Reports, vol. 31, Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262547383 (paperback) 9780262377423 (pdf)

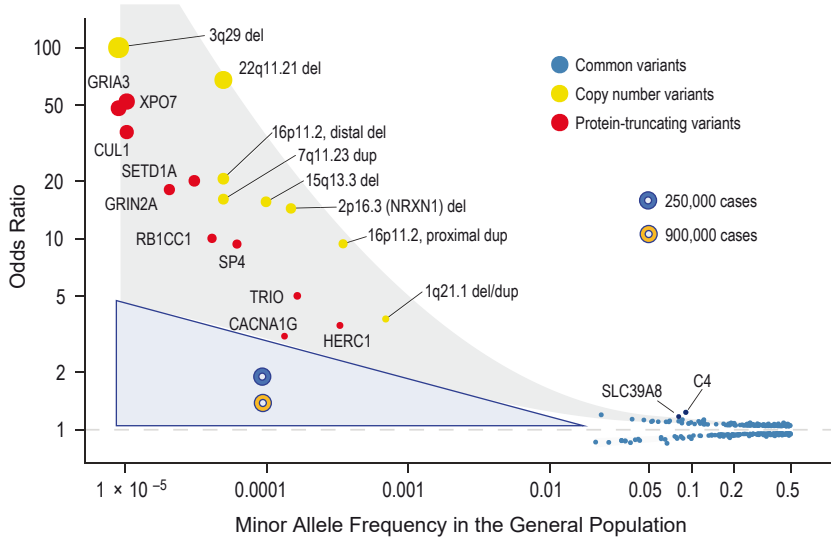


**Figure 3.1** Effect size by allelic frequency for known genetic risk variants in schizophrenia. Updated from Singh et al. (2022).

While schizophrenia provides several useful points for consideration, it cannot be used as a model for all other psychiatric disorders. Schizophrenia has featured a relatively high rate of common variant discovery (as compared to MDD or ASD) and a relatively slow rate of rare variant discovery (as compared to ASD or ID). This pattern may predict average phenotypic differences between individuals with schizophrenia who carry strong-acting risk variants (e.g., in *SETD1A*) versus those who do not, as has been seen for other rare variant-mediated disorders (Ahn et al. 2014; Satterstrom et al. 2020). Such questions, however, can only be answered with larger schizophrenia case sample sizes.

Genetic data highlights other differences between schizophrenia and, for example, ASD and MDD. Schizophrenia has a very high "genetic correlation with itself," which we use to mean the estimated genetic correlation between one schizophrenia case-control data set and another. When one estimates this correlation between various schizophrenia cohorts that have contributed genetic data to the Psychiatric Genomics Consortium, those values almost uniformly exceed 0.9, and are typically close to 1. This can be interpreted to mean that the common variant influences on schizophrenia are extremely similar across different samples, regardless of where or how individuals with the diagnosis are being ascertained.

The same cannot be said for MDD or ASD. The average genetic correlation in MDD with itself is around 0.7. This is similar to the genetic correlation between schizophrenia and bipolar disorder. In other words, there is as much genetic heterogeneity between cohorts of individuals diagnosed with

From "Exploring and Exploiting Genetic Risk for Psychiatric Disorders," edited by Joshua A. Gordon and Elisabeth B. Binder. Strüngmann Forum Reports, vol. 31, Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262547383 (paperback) 9780262377423 (pdf)

MDD as there is between cohorts of individuals diagnosed with schizophrenia and cohorts of individuals diagnosed with bipolar disorder. The picture is similar for ASD. The estimated genetic correlation between ASD in the Danish iPSYCH cohort and ASD in the (predominantly American) Psychiatric Genomics Consortium cohorts is 0.74 (Grove et al. 2019). These low correlations reflect heterogeneity that is likely to slow genetic discovery and increase the frequency of nonreplication between cohorts. Thus, a larger number of cases will be required to reach equivalent points of statistical power. To put it quantitatively, we estimate that *more* than 250,000 individuals will be required to identify a risk variant with OR of 2 and allele frequency of 0.0001 in meta-analyses of MDD and ASD.

## Increasing Genetic Associations across the Frequency Spectrum to Fill in Missing Biology

To understand the biology of a disorder, scientists must mechanistically connect causal factors to phenotypes. With the high heritabilities observed in neuropsychiatric disorders, genetic factors offer a critical causal starting point in this quest for biological understanding. Thus, a key motivation for genetic studies is as an entree into biology. All psychiatric disorders have elements of their genetic architecture that remain uncovered. For many, neither the common nor rare variant influences have been fleshed out. In this section, we link association goals to issues of biological understanding for both rare and common variation.

Rare variation underlies a substantial fraction of ID and ASD, early-onset disorders with high selective pressure. Rare variation is less prominent in late-onset disorders with weaker selective pressures. The majority of variation identified to date is from heterozygous protein-coding variation that disrupts one copy of a specific gene (e.g., *SCN2A*). These rare variants provide tractable targets for experimental model systems. However, the majority of genes associated with neuropsychiatric disorders perform myriad functions across development of the human brain (i.e., they are highly pleiotropic). While identifying a molecular, cellular, or behavioral consequence of disrupting the gene is relatively straightforward, showing that this consequence underlies the human phenotype is not, due to the absence of reliable endpoints in experimental model systems. This translational challenge limits the extent to which analysis of single genes can definitively inform biological understanding of the phenotype.

To overcome this challenge, convergent approaches aim to identify shared features or consequences across multiple genes as an indicator of relevant biology, demonstrating statistical enrichment of a biological measure across genes associated with a disorder compared to a relevant group of comparator genes (e.g., brain-expressed genes without evidence of statistical association).

From "Exploring and Exploiting Genetic Risk for Psychiatric Disorders," edited by Joshua A. Gordon and Elisabeth B. Binder. Strüngmann Forum Reports, vol. 31, Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262547383 (paperback) 9780262377423 (pdf)

This systems biology approach has provided key insights into ASD, including the enrichment of genes with a role in transcriptional regulation and neuronal communication (De Rubeis et al. 2014), prenatal onset (Parikshak et al. 2013; Willsey et al. 2013), and highlights the role of excitatory and inhibitory cortical neurons (Satterstrom et al. 2020). Convergent approaches, however, are very sensitive to the choice of background genes, which, in turn, rely on gene discovery efforts. While current analyses can reliably distinguish genes associated with ASD, they lack the sensitivity to exclude genes that are not associated with ASD. Larger discovery cohorts would improve our ability to distinguish genes associated with ASD versus those which are not, providing deeper insights into the underlying neurobiology.

Meanwhile, common polygenic variation contributes the majority of genetic liability for neuropsychiatric disease and is highly distributed across the genome. For example, more than 71% of 1 Mb blocks of the genome contain at least one schizophrenia-influencing variant (Loh et al. 2015). Significant loci are largely noncoding, which presents a formidable challenge to biological interpretation and identification of causal genes. Here, we consider three strategies for learning risk biology from common polygenic variation, each of which requires further characterization of the common variant signal to be employed effectively.

First, there is the classic approach of (a) fine mapping to identify the causal variant from the GWAS-identified variant and (b) using bioinformatic and experimental approaches to identify the manner in which the causal variant creates risk (e.g., through regulation of a nearby gene). This approach has achieved some notable successes, including identifying expression quantitative trait loci (eQTL), which co-localize with schizophrenia risk loci (Fromer et al. 2016), and the association of specific C4 haplotypes with schizophrenia (Sekar et al. 2016). However, with currently available data, the great majority of neuropsychiatric GWAS loci cannot be resolved to a single variant that alters gene expression or splicing behavior. Efforts to identify brain-related eQTLs and the eGenes that they regulate remain limited by sample size as well as the current state of knowledge of gene expression patterns in specific developmental stages, brain regions, and cell types (Table 3.2). Improved sample size would aid fine mapping of known loci and permit the discovery of new loci as would more specific and complete data regarding gene expression in the brain across the lifespan.

Second, new statistical approaches aim to use neuropsychiatric risk genes uncovered by rare variation to provide an interpretative scaffold for common variation. The degree to which these genes mediate the effects of common variation could be highly informative to questions of rare and common variant convergence. One recently developed approach, the Abstract Mediation Model (Weiner et al. 2022), estimates the fraction of common variant heritability mediated by a set of genes without relying on measured gene expression levels or eQTLs. Most strikingly, more than 90% of heritability for schizophrenia and

From "Exploring and Exploiting Genetic Risk for Psychiatric Disorders," edited by Joshua A. Gordon and Elisabeth B. Binder. Strüngmann Forum Reports, vol. 31, Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262547383 (paperback) 9780262377423 (pdf)

**Table 3.2** Representative quantitative trait locus (QTL) discovery in the human brain. DLPFC: dorsolateral prefrontal cortex.

| QTL Type | Brain Region | Developmental Stage | Total Subjects (N) | eGenes / loci | Cohort | Reference |
|---|---|---|---|---|---|---|
| Expression | DLPFC | Prenatal, postnatal | 176 | 5,728 | BrainVar | Werling et al. (2020) |
| Expression | Cortex | Prenatal | 201 | 6,546 | UCLA | Walker et al. (2019) |
| Expression | Cortex | Adult | 1,433 | 18,433 | AMP-AD (ROSMAP) Mayo TCX | Sieberts et al. (2020) |
| Expression | 10 regions | Adult | 72–103 | 853–3,454 | GTEx | GTEx Consortium (2020) |
| Splicing | 13 regions | Adult | 63–124 | 1,590 | GTEx | Zhang et al. (2020b) |
| Expression | DLPFC | Adult | 1,866 | 32,944 | PsychENCODE, CMC, GTEx | Wang et al. (2018) |
| Chromatin | DLPFC | Adult | 292 | 8,464 | PsychENCODE, CMC, GTEx | Wang et al. (2018) |
| Protein | DLPFC | Adult | 380 | 32 | NA | Yang et al. (2021) |

ASD are mediated by genes other than those implicated in neurodevelopmental disorders through rare coding variation (n = 373 genes; Figure 3.2). In contrast, 21 genes implicated in Mendelian forms of lipid dysregulation are massively enriched for lipid heritability (> 100x), mediating about 25% of overall heritability. These results reflect the differential effect of negative selection on common variation, where negative selection against neuropsychiatric traits limits the allelic effect size of common variants, "flattening" the distribution of heritability across the genome (O'Connor et al. 2019). In contrast, selective pressures act much less strongly on lipid phenotypes, permitting common variants of large effect to co-localize at physiologically important genes.

These observations provide a critical look into the prospects for deriving biological insight from neuropsychiatric polygenicity. First, genome-wide observations such as these do not preclude instances of common and rare overlap; for instance, in schizophrenia, both common and rare associations implicate glutamatergic dysfunction in disease pathogenesis (Singh et al. 2022; Trubetskoy et al. 2022). However, they do suggest that rare variant-implicated genes mediate only a small fraction of polygenic heritability and, consequently, that a very large number of genes mediate heritability for these traits. Sophisticated approaches and biological assays are needed to derive these insights that extend beyond canonical gene set enrichment methods.

A third option is to develop methods to learn risk biology directly from a common polygenic signal. This is a new area of inquiry, becoming more possible through the growth of human or human-derived cell line resources, specifically those paired with genome-wide genetic data. A recent analysis

From "Exploring and Exploiting Genetic Risk for Psychiatric Disorders," edited by Joshua A. Gordon and Elisabeth B. Binder. Strüngmann Forum Reports, vol. 31, Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262547383 (paperback) 9780262377423 (pdf)

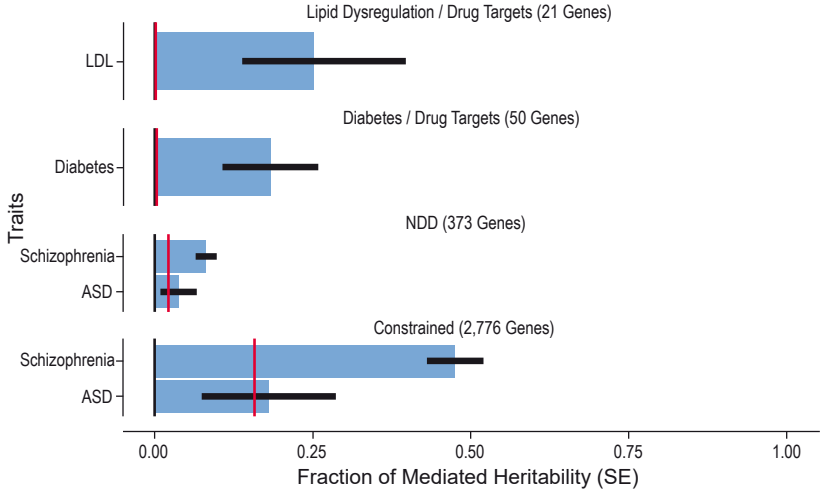


**Figure 3.2** Fraction of common variant heritability explained by specific gene lists (LDL: low-density lipoprotein).

of common variant risk for ASD exemplifies such an approach (Weiner et al. 2022). In their study, Weiner et al. uncovered an unexpected concentration of common variant signal distributed across the 30 Mb p arm of chromosome 16. The region includes one of the best-established neuropsychiatric risk copy number variations (CNVs), deletions, and duplications at 16p11.2. Using human cell line resources, they found that both common polygenic influences on ASD and isogenic deletion of the 16p11.2 CNVs were associated with decreased gene expression across the full 30 Mb p arm. The per gene effects of the deletion and the PRS of ASD were correlated, suggesting convergent functional impact. The region also exhibited an unusually high degree of chromatin contact with itself, potentially explaining the convergent effects of common and rare variation in the region. These findings suggest that biological insight can emerge from highly distributed polygenic liability, though cell line and other omics resources will need to grow substantially to be most successfully used for this purpose.

## Improving Global Representation

Genomic studies have historically relied on data from nations with the resources to fund large-scale research. Reflecting disparities in these resources, Fatumo et al. (2022) estimate that 86% of all genomics studies have been conducted in populations of European descent.

Psychiatric genetics studies are no exception to this trend. Among all studies listed in Table 3.1, only one focuses centrally on a non-European population,

From "Exploring and Exploiting Genetic Risk for Psychiatric Disorders," edited by Joshua A. Gordon and Elisabeth B. Binder. Strüngmann Forum Reports, vol. 31, Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262547383 (paperback) 9780262377423 (pdf)

although other ancestral populations are sometimes represented to some extent. To dissect the ancestral origins of cohorts that have contributed to landmark psychiatric genetics studies, we took a deeper look at the Table 3.1 cohorts. First, we assigned each major cohort in a study to an ancestral group based on the majority of that cohort: if more than half the individuals of the cohort were of European ancestry, we assigned it to European; if the cohort had no majority group, it was assigned to "other"). In some cases (e.g., Autism Sequencing Consortium, SCHEMA, or Psychiatric Genomics Consortium efforts), the goal was to develop a large data set from an aggregate of very small cohorts. Under these scenarios, we counted the full data set as one cohort, particularly because the consortium-led databases continue to be used in later studies as one cohort. Consistent with this, if a study meta-analyzed or replicated across different large cohorts sourced from different databases, each database was deemed a distinct cohort.

In addition to the studies from Table 3.1, we also added non-European GWAS cohorts with more than 50,000 total samples within the past five years, if such studies existed. We found two such papers, both of which studied East Asian cohorts: Lam et al. (2019) conducted a study of schizophrenia and Giannakopoulou et al. (2021) studied MDD. We summarize these results in Figure 3.3.

It will be very difficult to achieve numerically equal ancestral representation in neuropsychiatric genetics collections, particularly in light of the need to drastically increase sample size for many disorders. However, it is possible to achieve parity in the performance of genetic data across populations. Parity has been widely discussed in terms of PRS performance, as PRSs are now being incorporated into clinical care in a number of settings. Should PRS become clinically useful in psychiatry (see Davis et al., this volume), the use of current PRS would exacerbate health disparities (Doan et al. 2019; Martin et al. 2019b). PRS perform best when deployed in the same ancestral populations that provided data for the GWAS activity used to develop the PRS. For example, the schizophrenia PRSs, derived from predominantly European ancestry samples, becomes less associated to schizophrenia liability and less predictive of correlated behavioral outcomes in ancestral groups that are genetically distant from Europeans.

Problems of parity also extend to rare variation. Rates of *de novo* variation do not differ markedly by ancestry, leading to the expectation that similar genes will be associated with neuropsychiatric disorders worldwide through heterozygous rare variation (unless ancestry-specific protective factors exist). Interpreting rare variation for clinical diagnostics, however, relies on using population cohorts, such as gnomAD, to distinguish rare-standing variation from disease-causing new variation. The marked underrepresentation of people of African ancestry extends to gnomAD, in which only 14% of individuals have some African ancestry and these are predominantly from African Americans with a mixture of West African and European ancestry (Gudmundsson et al.

From "Exploring and Exploiting Genetic Risk for Psychiatric Disorders," edited by Joshua A. Gordon and Elisabeth B. Binder. Strüngmann Forum Reports, vol. 31, Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262547383 (paperback) 9780262377423 (pdf)

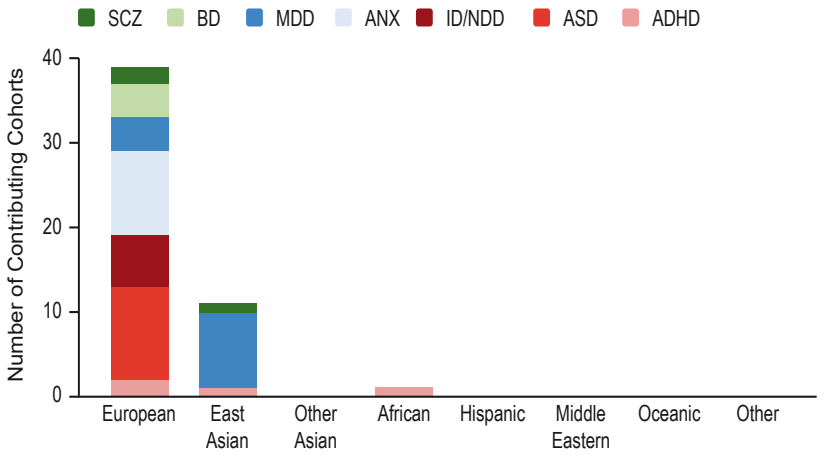


**Figure 3.3** Distribution of majority ancestries of cohorts contributing to major and most recent psychiatric genetic studies.

2022; Karczewski et al. 2020). Measures are underway to improve this, including the NeuroDev Kenya collection, which aims to identify genetic diagnoses in hundreds of East African individuals and families, and improve medical, genetic, and diagnostic pipelines for individuals of all types of African ancestry. Because people of African ancestry carry more genetic variation than people of European or Asian ancestry (groups better represented in genetic studies), they are more likely to carry variants of unknown significance and receive ambiguous results from genetic testing (Popejoy et al. 2018). They are also more likely to receive false positive and false negative genetic diagnoses (Caswell-Jin et al. 2018; Lek et al. 2016; Popejoy et al. 2018). Biallelic rare variation also plays a role in neuropsychiatric disorders (Lim et al. 2013). Since these rely on rare standing variation, we would expect dramatic differences across ancestries, as are seen with disorders such as cystic fibrosis or mucopolysaccharidosis.

## Conclusion

Increasing sample size is critical to further progress in delineating genetic risk for neuropsychiatric disorders; however, advances in assays (e.g., WGS), statistical methodologies, and functional biology will also contribute. Although immense progress has been made, substantial gaps remain, including many psychiatric diagnoses, most major ancestral populations, variants of moderate effect size and/or population frequency, and rare noncoding variation. Further progress will continue to refine our understanding of existing rare and common loci, extend neurobiological insights from these loci, and potentially uncover entirely new biological insights.

From "Exploring and Exploiting Genetic Risk for Psychiatric Disorders," edited by Joshua A. Gordon and Elisabeth B. Binder. Strüngmann Forum Reports, vol. 31, Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262547383 (paperback) 9780262377423 (pdf)

Neuropsychiatric disorder consortia have led the genomics field in terms of building international collaborations to share data across researchers. The Psychiatric Genomics Consortium, Autism Sequencing Consortium, and SCHEMA provide active models. While these groups still generate genomic data, they have increasingly integrated data from other sources, including nonprofit (e.g., Simons Foundation) and for-profit (e.g., 23andMe, GeneDx, Regeneron) organizations. This trend is likely to continue, with consortia increasingly focusing on integrating external data or generating data on ancestrally or phenotypically diverse populations. Providing an environment conducive to these ongoing large-scale international efforts has the potential to deliver refined genetic insights rigorously and efficiently.

From "Exploring and Exploiting Genetic Risk for Psychiatric Disorders," edited by Joshua A. Gordon and Elisabeth B. Binder. Strüngmann Forum Reports, vol. 31, Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262547383 (paperback) 9780262377423 (pdf)