nature neuroscience

Genome-scale neurogenetics: methodology and meaning

Steven A McCarroll[1,2], Guoping Feng[1,3,4] & Steven E Hyman[1,5]

Genetic analysis is currently offering glimpses into molecular mechanisms underlying such neuropsychiatric disorders as schizophrenia, bipolar disorder and autism. After years of frustration, success in identifying disease-associated DNA sequence variation has followed from new genomic technologies, new genome data resources, and global collaborations that could achieve the scale necessary to find the genes underlying highly polygenic disorders. Here we describe early results from genome-scale studies of large numbers of subjects and the emerging significance of these results for neurobiology.

Schizophrenia, bipolar disorder and autism are neuropsychiatric syndromes that produce severe symptoms and significant, often lifelong, disability. Unfortunately, knowledge of disease mechanisms has been scant, and existing treatments benefit only a subset of the disease symptoms and even then only incompletely. For example, antipsychotic drugs, all of which act by antagonizing D2 dopamine receptors, are partly effective for alleviating hallucinations, delusions and other psychotic symptoms that occur in schizophrenia, bipolar disorder, severe forms of depression and neurodegenerative disorders. However, there is no effective pharmacologic treatment for the disabling deficit symptoms of schizophrenia (such as blunted affect, decreased motivation and impoverished thought and speech) or for cognitive impairments (such as deficits in attention, working memory, verbal fluency, declarative memory and social cognition). All widely used psychiatric medications are based on prototype drugs that were serendipitously discovered in the period between 1949 (lithium) and 1957 (the antidepressants and benzodiazepines); no significant new class of drugs has been introduced to the clinic since then. Contributing to this lull in discovering new therapeutic agents is the lack of insight into the molecular mechanisms of these disorders. Across all of medicine, molecular insight has proven to be the most important information needed to identify and validate drug targets and to discover biomarkers[1,2]. The advent of genomic technologies of the early twenty-first century has allowed, for the first time, detailed examination of the genetic basis of many non-Mendelian diseases. Here we review advances in the genetics of psychiatric disorders, what neuroscientists can take away from these studies, and how best to apply the resulting information toward investigating the neurobiology of these complex and heterogeneous disorders.

[1]Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. [2]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [3]McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [4]Department of Brain and Cognitive Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [5]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts, USA. Correspondence should be addressed to S.A.M. (mccarroll@genetics.med.harvard.edu) or S.E.H. (stevehy@broadinstitute.org).

Received 27 January; accepted 9 April; published online 27 May 2014; doi:10.1038/nn.3716

Psychiatric genetics before 2009

Unlike the situation for some neurodegenerative disorders such as Alzheimer's disease, brains of patients with neuropsychiatric disorders such as schizophrenia, bipolar disorder and autism have not yielded predominant biochemical abnormalities (such as plaques) that could be exploited scientifically, nor have rare Mendelian forms of these diseases been identified except for rare syndromal forms of autism. Nonetheless, these neuropsychiatric disorders run in families, and twin and adoption studies show them to be highly influenced by genes. On the basis of these observations, attempts to identify disease-associated DNA variation began more than two decades ago. However, the non-Mendelian patterns of segregation in families, and the non-Mendelian ratios seen when comparing concordance for disease between monozygotic twin pairs (who share 100% of their DNA sequences) and dizygotic twin pairs (who share, on average, 50% of their DNA sequences), portended a degree of polygenicity that stymied early attempts to use genetics to discover molecular mechanisms of pathogenesis.

Twin studies indicate that schizophrenia, bipolar disorder, autism and attention deficit hyperactivity disorder (ADHD) are among the most heritable of any common, genetically complex medical disorders. (Heritabilities for these four disorders fall in the range of 0.65–0.80; ref. 3.) In twin studies, the rate of concordance for a phenotype is compared between monozygotic and dizygotic twin pairs, permitting an estimate of the contribution of inherited and non-inherited factors to phenotypic variance. More recently, specific components of heritability[4] have also been estimated on the basis of common DNA sequence variants across the genome shared by unrelated individuals. Significant lower bounds have been set for the heritabilities of schizophrenia, bipolar disorder, autism, ADHD and major depressive disorders[5].

A problem in older genetic studies was the widely held hypothesis that each named psychiatric disorder represented a natural, independent category that might be expected to breed true. In fact schizophrenia and bipolar disorder often co-occur in the same families[6]. On the basis of prevailing models, such families were often considered to be anomalous and were often excluded from genetic studies. It is now apparent, however, that psychiatric disorders, as currently defined, share a high percentage of risk-associated genes with one another[5,7], a connection that at least partly explains the high frequencies of shared symptoms across diagnoses, of patients being diagnosed with multiple disorders, and of patients receiving different diagnoses at different times in their lifespans.

Box 1 Linkage studies

Linkage analysis attempts to identify genomic segments that are shared by individuals in a family who are affected by the same disorder. The underlying hypothesis is that shared DNA segments contain a sequence variant that is highly penetrant in regard to the disease phenotype.

Linkage studies have been most successful for diseases that have a simple, monogenic architecture—i.e., that segregate with a single genetic variant within a family and with variants (which may be different) within the same gene in other families. This allows evidence for a gene to accumulate across families, as was the case for the *CFTR* gene in cystic fibrosis, since a single family is generally too small to make it possible to zero in on one locus.

Linkage analysis has failed when extended to genetically complex disorders (such as many psychiatric disorders) that are influenced by different combinations of variants in the same family or by different constellations of genes in different families. In such cases, linkage studies have failed to produce results that are statistically strong or broadly replicated. Today, we understand that psychiatric disorders are influenced by both rare and common variants in hundreds of genes; that the observation of multiple affected individuals in the same family does not represent the action of a shared, highly penetrant variant; and that such variants, when they do exist, are present in only a small percentage of families.

There is today wide recognition of the scientific limitations of psychiatric disorder definitions within current standard nosologies, the *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition* (DSM-5)[8] and the *International Classification of Diseases, 10th Edition*[9]. These manuals erroneously represent psychiatric disorders as categories qualitatively distinct from normalcy and from one another. Recent data from genetics, cognitive neuroscience, imaging, and population-based epidemiology of symptoms rather than syndromes argue that psychiatric disorders are better understood as heterogeneous spectra that deviate from health quantitatively but not qualitatively, in analogy to hypertension or dyslipidemias[10]. Moreover, lacking objective biomarkers, the DSM system defines disorders in terms of fairly arbitrary constellations of symptoms. As a result many distinctions between disorders are not empirically supported, and unsurprisingly, many symptomatic individuals fall outside specific diagnostic categories. Such limitations in phenotyping would seem inconsistent with the possibility that DSM-defined psychiatric disorders could exhibit high heritabilities or aid in successful gene-discovery efforts. In fact, for schizophrenia, bipolar disorder and autism, the symptoms, typical ages of onset and courses of illness have been well documented apart from DSM diagnoses. Nonetheless, the limitations of the DSM system have likely contributed to the 'noisiness' of patient data sets and thus, along with the complexity of the genetic architecture, have contributed to the need for larger population samples than are required for genetic analysis of many polygenic diseases outside of psychiatry. Ultimately genetics will contribute, along with other measures, to significant revisions in psychiatric diagnosis. However, the pleiotropy of many risk-associated genes suggests that genetics alone will not prove sufficient to make diagnoses for common psychiatric disorders.

With the advent of genetic linkage methods (**Box 1**), many scientists in the 1980s and 1990s took up the challenge of trying to identify genomic segments that co-segregated with neuropsychiatric disorders in families with multiple affected individuals. The hypothesis underlying linkage analyses is that family members who share a disease phenotype do so as a result of one or a few shared, highly penetrant risk alleles that co-segregate with the phenotype.

Why did linkage analysis fail for psychiatric disorders when it was successful for so many diseases with simple Mendelian patterns of inheritance, including familial Alzheimer's disease? The genetic architecture of psychiatric disorders may be more complex, such that even when a family carries a risk variant of large effect, it tends to be far less penetrant than loci that 'cause' familial Alzheimer's disease or cystic fibrosis. Moreover, even in the situation in which a family carries a rare variant of large effect, other superficially similar families are likely to carry different rare variants. Thus linkage evidence will not accumulate for any specific locus. Second, individuals affected by psychiatric disorders may carry *many* alleles that influence their risk. Such risk variants may aggregate in individuals and families by chance such that even in the same family affected individuals may have both shared and unshared risk alleles. Evidence from the study of common variants (**Fig. 1**) supports this model. Genetic risk arises from many different genes, both within individuals and across families, and from rare and common variants at the same time. When disorders arise in different families from the effects of alleles in many different genes—the definition of a polygenic trait—the assumptions underlying linkage analysis are violated.

Another unsuccessful effort to identify disease-associated genes in psychiatric disorders involved biological candidate-gene studies (**Box 2**). These studies, which gained much attention during the 1990s and into the 2000s, attempted to use knowledge of biology to nominate candidate genetic variants for association studies. Candidate-gene studies ask whether a particular allele of a nominated gene is statistically associated with a disease (or other phenotype of interest) by comparing the genotypes of affected subjects to those of healthy control subjects. The underlying assumption was that the molecular biology of mental disorders (or of interindividual differences in cognitive capacities or emotion processing) was well enough understood to make it possible to nominate candidate genes with substantial prior probabilities of association. In schizophrenia, more than 1,400 candidate-gene studies were published, of which half focused on the same 18 genes (http://www.szgene.org/). Owing perhaps to an inadequate understanding of multiple hypothesis testing, candidate-gene studies were often reported as having positive results whenever any tested association reached a nominal significance level (such as $P = 0.05$). In a field in which thousands of such hypotheses are tested each year, it

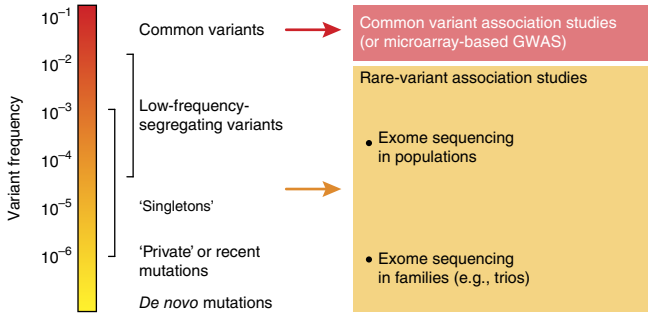

Figure 1 Allele frequency represents a continuum in human populations, with cutoffs between 'common' and 'rare' being somewhat arbitrary. On the left we display the frequency of common variants (common polymorphisms) and the large range of frequencies of rare variants. On the right we display the methodologies used to detect different types of variation. For rare variants, design is based not only on allele frequency but also on whether the variant is transmitted or *de novo*. Whole-genome sequencing is not depicted because it has not yet been broadly incorporated into published studies in psychiatric genetics. Common-variant association studies (CVAS) can detect both coding and noncoding variation (as will whole-genome sequencing). By definition, exome sequencing is focused on protein-coding regions of the genome.

Box 2 Candidate-gene association studies

Candidate-gene association studies are hypothesis-driven evaluations of specific variants in a particular gene, most often chosen on the basis of a biological hypothesis. This approach contrasts with unbiased genome-scale approaches represented by common- or rare-variant association studies. Candidate-gene association studies have been most successful for diseases in which the relevance of the selected gene to the biology of the disorder was already well established. For example, candidate-gene studies have identified both common and rare alleles in genes already known to influence serum lipid phenotypes such as levels of low-density lipoprotein and triglycerides[40]. For psychiatric disorders in which the underlying biology is not yet understood, candidate-gene studies involve more speculation about the relevance of the selected gene and thus have a smaller prior probability of detecting a true association. Across a field in which different laboratories are collectively testing hundreds of candidate hypotheses, *P* values of 0.01 or 0.001 will very frequently be reached by chance, even if the alleles in question bear no true underlying relationship to the phenotype.

Moreover, publication bias has almost certainly favored the preparation, submission and publication of positive studies over negative studies. Thus, even the presence of multiple positive studies in the literature does not necessarily imply a true genetic effect on phenotype. In fact, many biological candidate genes that have been reported to be associated with psychiatric disorders show no evidence of association in today's well-powered, unbiased genome-wide searches.

is expected that—even in the absence of any true underlying genetic relationships—hundreds of such analyses will arrive by chance at *P* values that reach this modest threshold. For example, while a PubMed search can find large literatures on *DISC1*, *NRG1* and other candidate genes in schizophrenia, the actual results appear random and show little or no consensus on the genetic variants reported as positive or on the directionality of their putative effects. Moreover, today's genome-wide studies, having examined these same alleles in large, well-powered cohorts (see below), do not find even modest evidence for association at most of these candidate loci.

Neuropsychiatric genetics since 2009: accelerating progress

Unbiased analyses of the entire genome, pursued in large cohorts of affected and unaffected individuals, have recently demonstrated associations of many specific genomic loci with schizophrenia, bipolar disorder and autism. Many results have already been replicated in other patient populations and are statistically conclusive. Several factors have made this success possible.

The first is a new understanding of human genome sequence, structure and variation. Over the decade since the Human Genome Project was completed, extensive study has revealed how the sequence and structure of genomes vary from person to person[11–14]. Although it may be unsurprising that healthy human beings are not each an isogenic 'wild-type' organism in which functional mutations clearly stand out, the degree of variation among individuals is remarkable. A key insight is that functional variation abounds in every human genome: each human genome contains thousands of alleles that alter the protein sequence or expression of genes. This large pool of functional variation across human populations provides the grist for polygenic inheritance. On a practical level, databases of genome sequences and sequence variants support the design of technologies for genome-wide SNP genotyping (**Box 3**), analysis of copy number variations (CNVs, **Box 4**) and exome sequencing (**Box 5**).

A second important factor has been *technological innovation* in the molecular tools used to analyze genome sequence variation. A microarray-based analysis can type the common single-nucleotide polymorphisms (SNPs; **Box 3**) and identify the large CNVs (**Box 4**) that are present in an individual's genome for about $100 today. The design of these microarrays is informed by data from the 1000 Genomes Project, making results extensible to untyped SNPs through the use of well-accepted statistical techniques that exploit the tendency of genomically nearby SNP alleles to segregate together as a haplotype[15]. The cost of sequencing has fallen dramatically, by 5–7 orders of magnitude since the Human Genome Project. Today the protein-coding parts of the genome (the 'exome', **Box 5**) are routinely sequenced for well under $1,000. With recently announced innovations in sequencing technology, whole genomes will soon be sequenced for about $1,000. Such innovations make it both technically possible and affordable to conduct studies of rare sequence variation in thousands of individuals[16]. Although genome-sequencing platforms are costly to set up and maintain, there is enough capacity in many research communities for individual investigators to gain access at low cost.

Box 3 Common-variant association studies

Genome-wide common-variant association studies (CVAS) involve the analysis of millions of common sequence polymorphisms (single-nucleotide polymorphisms, SNPs) in a genome-wide search for alleles that are more common in affected than in unaffected individuals. CVAS have more often been referred to as genome-wide association studies (GWAS), though this term can be confusing, as rare-variant studies by sequencing (**Box 5**) also represent genome-wide searches for association[41]. (Since the use of the term 'CVAS' is new[41], we use both acronyms here.)

Human populations expanded from a few thousand to 7 billion individuals over fewer than 100,000 years (5,000 generations). The sequence variation that was present in those small, ancestral populations 100,000 years ago is today common and found throughout the world; this allows such polymorphisms to be evaluated, on thousands of genetic backgrounds and in diverse environmental contexts, for relationship to disease. Human populations contain about 10 million common sequence polymorphisms, almost all of which trace back to these ancestral populations. Whole-genome sequencing has allowed these polymorphisms to be systematically catalogued, as by the 1000 Genomes Project[13]. Moreover, polymorphisms near one another on a chromosome often carry strong statistical relationships to one another—a phenomenon called linkage disequilibrium (LD)—such that, by genotyping only a few hundred thousand well-selected polymorphisms across the genome, one can statistically infer ('impute') the state of most of the others. Thus, in a CVAS (GWAS) large numbers of samples can be genotyped by low-cost SNP arrays (~$100/sample) designed to detect several hundred thousand common variants with other variants statistically imputed[15]. (The allelic states of common variants can also be ascertained with sequencing; common-variant studies may eventually use sequencing when its cost begins to approach that of array-based analyses, but for now the use of arrays supports the analysis of larger cohorts.)

Because a genome-wide search comes with a high burden of multiple hypothesis testing, CVAS (GWAS) demand a high level of statistical evidence for relationship to disease: a typical threshold is 5×10^{-8}. This level is hard to reach when a variant is just one of many common risk factors, and thus the primary determinant of the success of CVAS (GWAS) has been sample size. CVAS (GWAS) for psychiatric disorders initially appeared unsuccessful, though with expanded sample sizes they are now implicating large numbers of specific loci[25].

A SNP that is reported in a CVAS as associated with disease is almost always a marker for a set of sequence polymorphisms that are nearby in the genome and that segregate together as a haplotype. Haplotypes are sets of sequence alleles that travel together in human populations on a short chromosomal segment that are seldom disrupted by recombination. It is really this haplotype 'tagged' by the SNP that is associated with a disease; the haplotype is likely to contain one or more functional variants that influence a biological process relevant to the disease.

Box 4 Studies of copy number variation (CNV)

The early array-based tools used to genotype common single-nucleotide polymorphisms (SNPs) across the genome in CVAS (GWAS) were redesigned for the simultaneous analysis of copy number variation (CNV)—deletions and duplications of genomic segments[29]. Such efforts yielded surprisingly quick results; CNVs turned out to be principal early discoveries of genome-wide association studies (GWAS) for autism and schizophrenia in early studies with sample size in the low thousands[42,43]. As such studies have expanded, more than a dozen loci have been found at which large (>200 kb), recurring deletions or duplications are present in 0.1–1% of affected individuals but are 3–30 times rarer in the general population[44]. Most such results are by now well replicated. The challenges in using CNVs for neurobiology are twofold: their intrinsic complexity and the frequency with which they contribute to diverse disease phenotypes. Because many CNVs encompass large numbers of genes, it has proven difficult to determine whether their damaging effects are mediated by a single gene or by interactions among genes within the CNV. The most penetrant central nervous system phenotype across recurring CNVs is intellectual disability, with a fraction of those affected also having attention deficit hyperactivity disorder, autism or schizophrenia. Indeed, the relationship of a CNV to phenotype in any individual is very likely dependent on genetic background as well as nongenetic factors.

Analysis of CNVs, despite their high penetrance, supports polygenic models of autism and schizophrenia. Large (500 kb), *de novo*, genic CNVs found at *any* genomic locus are observed several times more frequently in patients with autism than in controls[45]. However, these CNVs are distributed across a large number of genomic sites, giving rise to an estimate that at least 130–234 regions of the human genome can mutate to substantially increase autism risk[45]. Of these disease-associated CNVs, only a small minority recur at specific sites within the genome.

Another important lesson from CNVs relates to partial penetrance. Large deletions and duplications (of 4–50 genes) are strong genetic perturbations, yet the large, recurring CNVs implicated in schizophrenia carry odds ratios of 3–30—i.e., they may increase risk from a background rate of, say, 1% to 3%–30%. The great majority of carriers therefore do not have the phenotype in question, though a recent study shows that these CNVs often have subclinical effects on cognition and IQ.

The third key factor has been a scientific understanding of the *scale* required to be successful. Scale is important for two reasons. First, consistent with the complexity of the molecular, cellular and circuit-based substrates of all cognitive processes and behavior, the heritability of psychiatric disorders appears to be distributed across many hundreds of loci, with each individual locus explaining only a small fraction of the overall heritability of the phenotype in populations. Second, an unbiased genome-wide search comes with a tremendous burden of multiple hypothesis testing: an association is meaningful only if it is very unlikely to have arisen by chance. High levels of evidence are therefore necessary to definitively implicate any individual genetic locus. (GWAS routinely use a significance threshold of 5×10^{-8}.) In common-variant association studies (**Box 3**), clear drivers of recent success are meta-analyses by large consortia sharing data from across the field. The current analysis of schizophrenia risk by the Psychiatric Genomics Consortium encompasses data from 52 study cohorts and from 82,000 individual human genomes. The size of such studies allows them to implicate individual loci with genome-wide significance and insulates them from many potential artifacts involving genotyping error and population substructure. Rare-variant association studies based on whole-exome or whole-genome sequencing are also beginning to reach substantial scale. In schizophrenia, the largest such studies to date included a case-control cohort of 5,000 individuals and a separate study of 600 father-mother-proband trios[16,17], and studies of autism trios by exome sequencing have covered about 1,000 trios collectively[18–21]. Increased scale will almost certainly be enabled by recent and future innovations in sequencing technology.

Interpreting emerging genetic results: common variants

Genetic discoveries with strong statistical support are emerging abundantly today from common-variant association studies (CVAS), which are most often called genome-wide association studies (GWAS). As described in **Box 3**, the convincing levels of evidence emerging from these studies reflect the ability to evaluate common alleles in thousands of subjects with diverse genetic backgrounds. Large-scale meta-analyses by international consortia have confidently identified 20 loci carrying risk of late-onset Alzheimer's disease (in a study of 72,000 genomes)[22], 10 or more loci affecting bipolar disorder (15,000 genomes) and 100 or more loci affecting schizophrenia (82,000 genomes). These results reflect high levels of statistical significance; many also are supported by replication analyses in multiple cohorts[23,24]. Moreover, the strong aggregation of these results in specific molecular complexes and biological pathways has the ring of scientific truth. In schizophrenia, for example, genes encoding the postsynaptic components of excitatory synapses and the subunits of the L-type calcium channel are represented disproportionately among the emerging genetic findings.

Biological interpretation

For biologists, the growing torrent of results from CVAS (GWAS) creates an enormous set of challenges. These results have strong levels of statistical evidence, but have qualities that have made them challenging to exploit by traditional methods.

Modest effects. The common haplotypes implicated in common-variant association studies (**Box 3**) are genetic nudges rather than genetic shoves. For example, CVAS (GWAS) for schizophrenia and bipolar disorder both find associations to the same set of common SNPs in the *CACNA1C* gene, encoding a subunit of the L-type calcium channel—a molecular complex that is also implicated in psychiatric disease phenotypes by common polymorphisms in other subunits such as *CACNB2* (refs. 5,23,24). The risk haplotype of *CACNA1C*—a haplotype is a set of alleles on a small chromosomal segment that segregates as a unit in populations—is present at an allele frequency of 35% in populations of European ancestry. Each inherited copy increases risk by about 15%, thus increasing a carrier's risk from the approximately 1.00% average population prevalence to 1.15%. Each such risk allele explains only a small fraction of variation in risk in the population. Moreover, in each person carrying this genetic risk factor, it acts within a genetic background that contains many other genetic influences (some increasing risk, some protective) of a similar magnitude.

Alleles of modest effect are often found in genes of large effect. For example, *CACNA1C* has also been found to harbor rare variants that profoundly disrupt organ development (including Mendelian causes of Timothy syndrome[25]). These large-effect variants so severely disturb the development and function of the brain and other organ systems that they may obscure a psychiatric phenotype. The effects ascertained in CVAS (GWAS) may represent milder and tissue-specific perturbations in the expression of the associated genes, rather than knockouts or, as in the case of Timothy syndrome, a toxic gain of function. CVAS (GWAS) have allowed such partial perturbations to be related to natural variation in phenotypes. Still, the partial and tissue-specific nature of these genetic perturbations makes them challenging to study by traditional biological methods.

Box 5 Rare-variant association studies

Rare-variant association studies (RVAS, often called 'sequencing studies' or 'exome sequencing studies') involve a search for rare sequence variants, including new ('*de novo*') mutations, via sequencing. Most rare variants occur too infrequently to allow association testing of individual variants; analyses of rare variants thus involve aggregating such variants into gene- or pathway-based sets and evaluating their aggregate frequency as a group.

Most rare-variant studies to date have focused on the protein-coding parts of the human genome (the exome)—both because potentially functional variants can be distinguished from neutral or synonymous variants and because the exome (which comprises only about 1.5% of the human genome) requires less sequencing than the whole genome, allowing larger numbers of patients to be sequenced in a study. As sequencing costs fall, exome-sequencing studies may give way to larger numbers of whole-genome sequencing studies.

Rare-variant studies involve one of three principal designs: case-control, trio and other family designs.

Case-control. Case-control studies are the mainstay of disease association in CVAS. For rare variants they involve sequencing, by the same methods at the same time, a large number of affected and unaffected individuals drawn from the same population. Data from the control individuals is used to calibrate how surprising the observation of any set of variants is—for example, whether a given gene (or a functionally related set of genes) harbors significantly more loss-of-function variants in affected than in unaffected individuals. Careful statistical calibration is important, as genomes are replete with rare sequence variation; on average each individual possesses thousands of protein-altering sequence variants, of which about 100 disrupt the function of the encoded protein[32,46].

The design of case-control RVAS involves important methodological issues around choosing samples. First, it is important that cases and controls be well matched—that they be drawn from the same population and analyzed by the same methods at the same time—as ancestry can profoundly shape the number of variants in a genome, and laboratory and computational methods influence the extent to which those variants are ascertained. This holds true for CVAS as well. Second, as studies grow beyond initial explorations in individual populations, an increasingly important decision will involve what populations to sequence in future studies. For example, rare protective alleles of *APP* and *PCSK9* were discovered as a result of their relative enrichment in the study populations[45,47]. Populations with unique histories such as bottlenecks may be particularly useful for uncovering effects of recessive variants.

Trios. Trio-based studies focus on new mutations as a way to filter out the thousands of protein-altering variants that are inherited by each individual. In these studies, DNA is collected from affected individuals and their unaffected parents. Sequencing father, mother and proband allows new mutations to be distinguished from the far larger number of inherited variants. Most genomes contain about 60 new mutations not present in either parent's genome; usually about 0–2 of these new mutations will be present in the exome, and about half of those will affect the sequence of an encoded protein.

The analytical challenge in trio-based mutation studies is that protein-altering mutations arise in a substantial fraction of all individuals (regardless of phenotype); identifying the subset that contributes to disorders in affected individuals is therefore challenging. Trio-based studies thus seek patterns among the mutations (in affected individuals) that appear nonrandom—such as a tendency of such mutations to cluster in specific genes or in constellations of genes with related functions or expression patterns[17–21]. For simple, monogenic diseases, this has often led to quick identification of a culpable gene. For genetically complex psychiatric disorders, the analytical challenge has been much more difficult and involves the gradual accretion of statistical evidence in individual genes or in large constellations of functionally related genes that are defined, for example, by localization or similar functional annotation of the encoded proteins. Generally only a minority of ascertained mutations contribute to such patterns, so statistical power has been key for making such inferences. For example, *CHD8* was implicated as an autism gene through trio-based studies, though *CHD8* mutations are present in less than 1% of cases in these studies; sequencing of hundreds of trios was necessary before this pattern became clear[18,20,21]. In designing trio-based studies for other complex diseases, one can try to increase the odds of finding impactful mutations by selecting sporadic cases (affected individuals without affected relatives). New mutations may also be enriched among patients who have syndromes involving many or more severe phenotypes (such as autism, epilepsy and low IQ) rather than more common presentations of the disorder.

Other family-based designs. Other study designs utilize cohorts that were initially collected for linkage analysis (i.e., families with multiple affected individuals) and attempt to winnow large sequencing data sets by making the kinds of assumptions that linkage analysis makes—for example, that affected individuals drawn from the same family must share a causal variant in common with one another. The challenge in such analyses is that even randomly selected individuals in any family will share substantial fractions of their genomes in common, and this shared part of the genome will include both rare and common variants affecting the sequences of large numbers of genes. The prioritization of individual variants therefore often involves evaluation based on biological candidacy or independent forms of genetic evidence from other studies. An increasingly common design is to see the family analysis as a hypothesis-generating process, after which nominated genes are sequenced and analyzed in larger, well-powered case-control cohorts to test for formal disease association.

Different combinations in different genomes. Common variants appear in different combinations in every human genome. Part of the reason for the modest explanatory power of individual common variants is likely that, in diverse human populations, each is only one of many genetic and environmental influences on the phenotype. As functional alleles begin to be identified and studied experimentally on isogenic backgrounds in controlled environments, their relationship to specific molecular and cellular phenotypes may be clearer than it is today. An important technology for analyzing such variants may involve techniques for precisely editing genomes[26,27] whether in cell lines or in animal models (further described below).

Functional alleles not yet known. A given SNP that is identified in common-variant association studies is generally a molecular proxy (or 'tag') for a set of 5–50 common polymorphisms that co-segregate as a haplotype. The specific allele(s) on that haplotype that influence the phenotype are not initially known—and have been identified in only a few cases to date[28–30]. Moreover, the genetic architecture of the functional alleles may not be identical to that of the associated SNPs: an association to a haplotype could also arise from lower-frequency alleles that segregate on that haplotype, or from combinations of multiple genetic effects at the same locus. Our still limited knowledge of the underlying functional allele(s) makes it challenging to model the genetic perturbation in cell lines or animals.

Noncoding variants and regulatory DNA. Most of the haplotypes implicated in CVAS (GWAS) do not contain protein-altering variants but rather reside within large introns or in sequences upstream of nearby genes, locations that often contain tissue-specific enhancer elements. Thus, while they can often be mapped to recognizable genes, the functional effects of risk-associated sequence variation on those genes may be difficult to establish. One might conclude that their effects arise not from 'broken' proteins but from quantitative variation in expression levels or alterations in cell type–specific expression patterns. Cases in which the functional alleles have been identified support the hypothesis of quantitative, tissue-specific effects on gene expression[28–30]. Noncoding variants in human genes are challenging to model in other organisms, as regulatory sequences appear to evolve on faster time scales and show less conservation than protein-coding sequences do. Though identifying functional alleles within noncoding sequence is a challenging problem, progress will likely be accelerated by data resources that annotate chromatin states in diverse cell types[31]. Overall, noncoding variation may be a way in which natural polymorphism found in populations differs from the severe

mutations that scientists often create in model organisms, which produce phenotypes far outside of the organisms' normal range.

Interpreting emerging genetic results: rare variants

Although less is known today about the contribution of rare variants to polygenic phenotypes, there is a strong theoretical case that some of them are quite significant: alleles that strongly predispose to disease are likely to be kept at low frequencies by purifying selection. Such rare variants must be ascertained in each genome by sequencing. Sequencing DNA from large numbers of affected individuals makes it possible to begin to explore the relationship of such variants to psychiatric and other phenotypes (**Box 5**). Innovations in genome-sequencing technology are starting to allow such studies to be pursued on a large scale.

The basic challenge in any rare-variant association study (RVAS) is that human genomes contain thousands of variants that alter or terminate protein sequences, complicating the assignment of disease causality to any particular variant[32]. Natural variation in human populations contrasts sharply with the isogenic context in which mutations are studied in most model organisms. In diverse populations, the properties of gene-disabling and gene-altering variants in a particular gene become clear only as such variants are observed in many individuals. For many rare, monogenic disorders with recognizable and unusual defining phenotypes (such as Kabuki syndrome), such results yielded quickly to exome sequencing, as most patients have mutations in the same gene. For polygenic disorders, far larger samples have been required before individual genes even begin to stand out from the background of vicissitudes across the genome. Moreover, even where rare variants of large effect have been ascertained—the clearest examples today are CNVs (**Box 4**)—they show only partial penetrance for adult-onset psychiatric phenotypes such as schizophrenia; most carriers have no psychiatric diagnosis, though subclinical effects on IQ and cognition may be more common (**Box 4**).

The strongest influences of rare variants, including *de novo* mutations, have been documented in congenital and early-childhood-onset disorders such as intellectual disability (ID) and autism. In these disorders, affected individuals appear to have increased likelihood of a protein-disrupting mutation, and such mutations have been observed recurrently in specific genes (such as *CHD8*, *SCN2A*, *GRIN2B* and *DYRK1A*) at statistically elevated rates, often with broad phenotypic impacts including seizures and severe ID[18–21]. In contrast, in highly heritable later-onset disorders, such as schizophrenia, affected individuals carry rare mutations in exons at a rate that only modestly exceeds that of the general population and that is substantially lower than in individuals with autism or ID[17]. This subtler enrichment suggests that such mutations play a smaller role in schizophrenia than in very early-onset neuropsychiatric disorders. Indeed these mutations have not yet shown evidence of accumulation in individual genes, although as a group they exhibit patterns of statistical enrichment among sets of functionally related brain-expressed genes[33] such as those that encode postsynaptic components of excitatory synapses[16] and those that encode RNA targets of the fragile X mental retardation protein (FMRP)[17]. As studies based on sequencing are expanded to more patients and controls, a growing set of genes is likely to be implicated by allelic series of rare variants. Such findings would present real opportunities for biological studies by permitting measurement of functional effects across allelic series, exploration of their consequences for cellular phenotypes, and attempts to relate them to clinically significant phenotypes in patients.

Specific protein-altering variants of large effect generally provide an easier starting point for biological studies than the regulatory variants implicated in CVAS (GWAS). However, RVAS are still well short of the sample sizes needed to implicate large numbers of specific genes in a majority of psychiatric disorders. A critical challenge for biologists in interpreting RVAS results, and in deciding when functional studies are merited, will be avoiding premature hypotheses born of biological plausibility and 'Just So' stories. Demonstration that similar concentrations of rare variants (in pathways and specific genes) are not present in cohorts of unaffected individuals will strengthen confidence in such results. The expansion of exome and whole-genome sequencing studies to much larger cohorts of affected and unaffected individuals will be key to arriving at genetic findings that are neurobiologically actionable with reasonable levels of confidence.

The challenge of polygenicity for neurobiology

As described, both common and rare alleles emerging from the analysis of psychiatric disorders have limited penetrance, act in concert with many other alleles (genetic backgrounds) and contribute to multiple phenotypes (are pleiotropic). However, for practical reasons, the predominant modern experimental approach to the functional analysis of disease-associated genes is based on variants of high penetrance[34] introduced into mice. Based on the emerging polygenic architecture of psychiatric disease, we would suggest that current approaches increasingly must be complemented by new approaches that permit far higher throughput, that are more inductive in nature, and that permit analysis of biological systems beyond a single hypothesis about a putatively causal allele. We must go beyond narrow pathophysiological hypotheses derived from small, disconnected parts of the research literature to systematic ways of gathering information and deriving insight from neurobiological systems.

Just as genome-scale experimental tools and data resources and unbiased (hypothesis-free) approaches to data collection and analysis have given new life to genetics, so they may offer new opportunities to molecular and cellular neurobiology. To understand how genetic perturbations affect specific cell populations, biologists will need a better census of the cell populations present in each brain region, along with experimental tools to systematically identify perturbations of their cellular states. This will require development not only of new experimental tools, such as scalable approaches for single-cell expression profiling and cellular models based on stem cell technologies, but also of a new generation of fundamental data resources, such as reference transcriptomes for each of the major cell populations in the brain. Single-cell proteomics is not yet possible, but it will be critical to develop methods that can determine which proteins in a putative pathway or network actually coexist in the same cell at the same time. Ultimately, validation will also require data from postmortem human brains. We do not argue that neurobiologists should await such data before exploring the function of disease-associated alleles; rather we urge that such molecular resources be developed in parallel in order to diminish speculative risks associated with therapeutic target identification and the process of target validation.

Systems neurobiologists will also not lack for challenges. After all, the symptoms and impairments of psychiatric disorders result from abnormal functioning of neural circuits. Thus successful investigation of disease mechanisms is likely to require scalable assays for neuronal structure, function and connectivity applied to primary or reprogrammed neurons *in vitro* (human, rodent, or other species) and to multiple animal models including *Drosophila*, zebrafish, rodents and nonhuman primates.

Studies of genetically engineered mice, currently the most widely used model, for the purposes of translational neurobiology will remain

important when investigating neural circuits that are structurally as well as functionally conserved in evolution, as appears to be the case for basal ganglia circuit dysfunction in repetitive behavior[35]. However, evolutionary conservation must apply not only to the circuits under study but also to the molecular networks by which the engineered gene contributes to circuit function. For many of the disease risk alleles emerging from neuropsychiatric genetics, however, there will be significant limitations to the utility of genetically engineered mouse models. These include the inefficiency of transgenic mice as a system to investigate alleles of limited penetrance against many genetic backgrounds, the poor evolutionary conservation of noncoding DNA sequences being identified by CVAS (GWAS), and significant differences in brain structure (cell types, circuits and regional architecture) between rodents and humans especially in evolutionarily more recent brain structures, such as prefrontal cortex. Given, however, that structural and functional defects in the dorsolateral prefrontal cortex are central aspects of schizophrenia[36], animal models closer to humans in evolution may be needed; specifically, dorsolateral prefrontal cortex is unique to primates[37]. It now appears possible to turn to nonhuman primate models, such as the common marmoset, a small New World monkey with a fast reproduction cycle, to study the neural circuits and biomarkers of psychiatric disorders[38].

The model selected should be based not only on the feasibility of experimentation with current methods but also, more importantly, on the questions being asked. For research that attempts to translate basic biological findings into understanding of pathogenesis or discovery of therapeutics, it is particularly important to attend to evolutionary conservation with humans. What has been called 'face validity'—the similarity of a physiologic finding or a behavior within an animal model to the human disease—is a rather treacherous guide, even if the behavior under study is quite similar from animal model to human[39]. The same structures or behavior can arise by convergent evolution from very different underlying mechanisms. Thus, for translational neurobiology it matters little whether a cognitive or behavioral phenotype appears very similar or identical in an animal model and a healthy or ill human unless the underlying mechanisms are conserved. A drug acts on targets within molecular networks, thus influencing synapses and circuits and only as a result affecting cognition, emotion and behavior.

Convergence of neurobiology with human genetics

Important future progress in psychiatric research may arise at the interface of neurobiology and human genetics. These two fields have enormous potential synergies; yet they have realized few of these synergies to date, and their separate success has involved revealing very different kinds of relationships in biological systems. Biological inquiry is based upon establishing cause-and-effect relationships in well-controlled experiments. Such relationships have often been clearest when they involved experiments with highly penetrant genetic defects (such as homozygous knockouts) and other dramatic experimental perturbations. This work has established powerful models about how the components of neurons and circuits work—though with little understanding of how natural variation in these components could underlie common psychiatric phenotypes. Human genetic studies of psychiatric disorders were organized initially around an unsuccessful quest for similarly simple, general, high-penetrance relationships, and then around an unsuccessful assumption that the relevant candidate genes could be predicted from biological understanding. Most confirmed genetic findings today have come from starting over with unbiased genome-wide searches—and embracing the situation that exists when hundreds of modest, common and often one or a few stronger-impact (but only partially penetrant), rare genetic perturbations inhabit every genome. This is creating a wealth of genetic discoveries for which no facile biological playbook exists.

Such a playbook will need to be written in the coming years. The synthesis of these ways of understanding brain function and dysfunction will offer fundamental intellectual challenges—and, we hope, important scientific insights and medically transformative ideas.

ACKNOWLEDGMENTS

The authors are supported by the Stanley Medical Research Institute.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Hyman, S.E. Revolution stalled. *Sci. Transl. Med.* **4**, 155cm11 (2012).
2. Jack, C.R. Jr. & Holtzman, D.M. Biomarker modeling of Alzheimer's disease. *Neuron* **80**, 1347–1358 (2013).
3. Sullivan, P.F., Daly, M.J. & O'Donovan, M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nat. Rev. Genet.* **13**, 537–551 (2012).
4. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
5. Cross-Disorder Group of the Psychiatric Genomics Consortium and Genetic Risk Outcome of Psychosis (GROUP) Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* **381**, 1371–1379 (2013).
6. Berrettini, W.H. Are schizophrenic and bipolar disorders related? A review of family and molecular studies. *Biol. Psychiatry* **48**, 531–538 (2000).
7. Lee, S.H. *et al.* Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat. Genet.* **45**, 984–994 (2013).
8. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders* 5th edn. (APA, Washington, DC, USA, 2013).
9. World Health Organization. *The ICD-10 Classification of Mental and Behavioural Disorders* (WHO, Geneva, 1992).
10. Hyman, S.E. Can neuroscience be integrated into the DSM-V? *Nat. Rev. Neurosci.* **8**, 725–732 (2007).
11. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
12. Altshuler, D.M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
13. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
14. Handsaker, R.E., Korn, J.M., Nemesh, J. & McCarroll, S.A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.* **43**, 269–276 (2011).
15. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* **10**, 387–406 (2009).
16. Purcell, S.M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185–190 (2014).
17. Fromer, M. *et al. De novo* mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179–184 (2014).
18. Sanders, S.J. *et al. De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237–241 (2012).
19. Iossifov, I. *et al. De novo* gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285–299 (2012).
20. Neale, B.M. *et al.* Patterns and rates of exonic *de novo* mutations in autism spectrum disorders. *Nature* **485**, 242–245 (2012).
21. O'Roak, B.J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* **485**, 246–250 (2012).
22. Lambert, J.C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452–1458 (2013).
23. Psychiatric G.C.B.D.W.G. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat. Genet.* **43**, 977–983 (2011).
24. Ripke, S. *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat. Genet.* **45**, 1150–1159 (2013).
25. Splawski, I. *et al.* Ca(V)1.2 calcium channel dysfunction causes a multisystem disorder including arrhythmia and autism. *Cell* **119**, 19–31 (2004).
26. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
27. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
28. Musunuru, K. *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**, 714–719 (2010).

29. McCarroll, S.A. *et al.* Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat. Genet.* **40**, 1107–1112 (2008).
30. Bauer, D.E. *et al.* An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. *Science* **342**, 253–257 (2013).
31. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
32. MacArthur, D.G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).
33. Gulsuner, S. *et al.* Spatial and temporal mapping of *de novo* mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* **154**, 518–529 (2013).
34. Karayiorgou, M., Flint, J., Gogos, J.A. & Malenka, R.C. The best of times, the worst of times for psychiatric disease. *Nat. Neurosci.* **15**, 811–812 (2012).
35. Welch, J.M. *et al.* Cortico-striatal synaptic defects and OCD-like behaviours in Sapap3-mutant mice. *Nature* **448**, 894–900 (2007).
36. Volk, D.W. & Lewis, D.A. Prefrontal cortical circuits in schizophrenia. *Curr. Top. Behav. Neurosci.* **4**, 485–508 (2010).
37. Preuss, T.M. Do rats have prefrontal cortex? The rose-woolsey-akert program reconsidered. *J. Cogn. Neurosci.* **7**, 1–24 (1995).
38. Okano, H., Hikishima, K., Iriki, A. & Sasaki, E. The common marmoset as a novel animal model system for biomedical and neuroscience research applications. *Semin. Fetal Neonatal Med.* **17**, 336–340 (2012).
39. Nestler, E.J. & Hyman, S.E. Animal models of neuropsychiatric disorders. *Nat. Neurosci.* **13**, 1161–1169 (2010).
40. Cohen, J. *et al.* Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nat. Genet.* **37**, 161–165 (2005).
41. Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proc. Natl. Acad. Sci. USA* **111**, E455–E464 (2014).
42. Weiss, L.A. *et al.* A genome-wide linkage and association scan reveals novel loci for autism. *Nature* **461**, 802–808 (2009).
43. The International Schizophrenia Consortium *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
44. Rees, E. *et al.* Analysis of copy number variations at 15 schizophrenia-associated loci. *Br. J. Psychiatry* **204**, 108–114 (2014).
45. Sanders, S.J. *et al.* Multiple recurrent *de novo* CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70**, 863–885 (2011).
46. Abecasis, G.R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
47. Jonsson, T. *et al.* A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. *Nature* **488**, 96–99 (2012).