Title
Whole genome sequencing in psychiatric disorders: the WGSPD consortium.

Journal
Nature neuroscience, 20(12)

ISSN
1097-6256

Authors
Sanders, Stephan J
Neale, Benjamin M
Huang, Hailiang
et al.

Publication Date
2017-12-01

DOI
10.1038/s41593-017-0017-9

Peer reviewed

comment

Corrected: Publisher correction

Whole genome sequencing in psychiatric disorders: the WGSPD consortium

As technology advances, whole genome sequencing (WGS) is likely to supersede other genotyping technologies. The rate of this change depends on its relative cost and utility. Variants identified uniquely through WGS may reveal novel biological pathways underlying complex disorders and provide high-resolution insight into when, where, and in which cell type these pathways are affected. Alternatively, cheaper and less computationally intensive approaches may yield equivalent insights. Understanding the role of rare variants in the noncoding gene-regulating genome through pilot WGS projects will be critical to determining which of these two extremes best represents reality. With large cohorts, well-defined risk loci, and a compelling need to understand the underlying biology, psychiatric disorders have a role to play in this preliminary WGS assessment. The Whole Genome Sequencing for Psychiatric Disorders Consortium will integrate data for 18,000 individuals with psychiatric disorders, beginning with autism spectrum disorder, schizophrenia, bipolar disorder, and major depressive disorder, along with over 150,000 controls.

Stephan J. Sanders, Benjamin M. Neale, Hailiang Huang, Donna M. Werling, Joon-Yong An, Shan Dong, Whole Genome Sequencing for Psychiatric Disorders (WGSPD), Goncalo Abecasis, P. Alexander Arguello, John Blangero, Michael Boehnke, Mark J. Daly, Kevin Eggan, Daniel H. Geschwind, David C. Glahn, David B. Goldstein, Raquel E. Gur, Robert E. Handsaker, Steven A. McCarroll, Roel A. Ophoff, Aarno Palotie, Carlos N. Pato, Chiara Sabatti, Matthew W. State, A. Jeremy Willsey, Steven E. Hyman, Anjene M. Addington, Thomas Lehner and Nelson B. Freimer

Genetic variation is a major contributor to neuropsychiatric disorders. The variants responsible likely include the complete range of sizes, from single nucleotides to large structural variants, and the full spectrum of population frequency, from common variants to rare variants unique to a family or individual. For severe, early-onset neuropsychiatric disorders, such as autism spectrum disorder (ASD) and schizophrenia, natural selection limits the population frequency of variants so that variants with larger effect sizes are extremely rare[1,2]. Over the past decade, genomic technologies have advanced our understanding of neuropsychiatric disorders, yet remaining limitations in technology and cohort sizes have limited progress in identifying inherited rare variants.

Genome-wide association studies (GWAS) using genotyping arrays have detected over 100 regions (loci) at which common genetic variants (population frequency $\geq 2\%$) are associated with a psychiatric diagnosis (Table 1). Individually, these variants exert small effects and thus require very large sample sizes for detection (Table 1). Common risk variants can provide a window into the molecular architecture of these disorders. For example, common variants suggest a previously unrecognized role for the complement cascade in schizophrenia[3].

Exome sequencing, which identifies genetic variants in the ~1% of the genome that encodes proteins, has identified over 50 genes in ASD (Table 1). The majority of this discovery was through de novo protein-truncating variants (PTVs) observed in a patient but not in either unaffected parent. Such mutations are very rare, for example, with population frequency $\leq 0.000002\%$, but they can have large effect sizes, up to a ~50-fold increase in risk. As with common variation, these very rare variants have advanced our understanding of the etiology of these disorders, for example, by implicating chromatin remodeling in ASD[4,5]. Although much remains to be discovered, these results have yielded critical starting points for studies of pathogenesis[6,7], and they indicate the feasibility and importance of discovering sufficient additional variation to fully delineate the key biological pathways underlying these disorders.

Insights from whole genome sequencing

By assaying most of the genome at single-nucleotide resolution, WGS holds the potential to extend rare-variant discovery to the ~99% of the genome that is noncoding (Box 1). While GWAS identifies common noncoding variants, the rare noncoding variants assayed by WGS might have substantially higher effect sizes[1], increasing tractability for biological experimentation. WGS also enables detection of most structural variation, including translocations, inversions, and copy-number variants (CNVs)[8,9]. Furthermore, WGS can improve detection of common variants in existing GWAS by statistically inferring single-nucleotide polymorphisms (SNPs) not directly genotyped (imputation) and by identifying the specific risk variants within a risk region (fine mapping). Similarly, WGS data may allow detection of common structural variants, including CNVs, that can be missed by current SNP-based approaches[10], facilitating common-CNV association studies.

The role of noncoding variation

Noncoding variation influences which exons within a gene are expressed, in which cells, and under what circumstances. There is considerable evidence that noncoding variation influences brain function and neuropsychiatric disorders. Over 90% of disease-associated GWAS loci discovered by assaying common variants map to noncoding regions[11,12]. In humans, at

Table 1 | The largest genomic studies to date in ASD, schizophrenia, bipolar disorder, and major depressive disorder

Study design	Platform	Variant detected	Disorder	Patients	Controls	Genome-wide hits	Reference
Case–control	**Genotyping microarray**	**SNP (GWAS)**	ASD	16,539	157,234	1	Anney et al., 2017[64]
			SCZ	36,989	113,075	108	Ripke et al., 2014[7]
			BPD	11,974	51,792	2	Sklar et al., 2011[39]
			MDD	121,380	338,101	15	Hyde et al., 2016[65]
		CNV	SCZ	21,094	20,227	8	Marshall et al., 2017[66]
			BPD	9,129	81,802	1	Green et al., 2015[67]
			MDD	2,591	8,842	0	Rucker et al., 2015[68]
	Exome sequencing	**Rare PTV**	ASD	5,563	1,881	0	Sanders et al., 2015[6]
			SCZ	2,536	2,543	0	Purcell et al., 2014[69]
		Ultra-rare PTV	SCZ	4,877	6,203	0	Genovese et al., 2016[70]
Family-based	**Genotyping microarray**	**CNV**	ASD	4,687	2,100	8*	Sanders et al., 2015[6]
	Exome sequencing	**De novo PTV**	ASD	5,563	1,881	65*	Sanders et al., 2015[6]
			SCZ	617	731	0	Fromer et al., 2014[71]
Meta-analysis	**Exome sequencing**	**Rare and de novo PTV**	SCZ	7,776	13,028	1	Singh et al., 2016[72]

SCZ, schizophrenia; BPD, bipolar disorder; MDD, major depressive disorder. *False discovery rate (FDR) ≤ 0.1.

least 4% of the noncoding genome has been under strong purifying selection[13]. Additionally, epigenomic studies have identified many functional noncoding elements involved in regulation of gene expression underlying neurogenesis, cell differentiation, and neurodevelopment[14].

Progressing beyond genetics to an understanding of psychiatric neuropathology is likely to require identifying the cell types, developmental periods, and brain regions involved. While such insights can be gained from gene association[15], noncoding variation studies should increase the resolution of such analyses by identifying regulatory regions restricted to fewer spatiotemporal windows or cell types. Given the multiple biological roles (pleiotropy) of genes implicated in psychiatric disorders, such WGS-derived hypotheses may be critical for biological follow-up.

The role of rare noncoding variation

While common noncoding variation clearly plays a role in neuropsychiatric disorders, the role of rare noncoding variation is less clear. A pessimist could note that, in Mendelian disorders, few linkage peaks were resolved to noncoding causal variants and that, for example, systematic deletion of noncoding regions proximate to the *HPRT1* gene (Lesch–Nyhan syndrome) had little impact on protein activity[16]. In contrast, an optimist could argue that fragile X disorder, the first psychiatric linkage peak resolved to a gene, is a triplet repeat expansion in the 5′ untranslated region of the *FMRP* gene and that there are several clear examples of Mendelian traits (e.g., *OCA2* enhancer in eye color) and disorders (e.g., *TBX5* enhancer in congenital heart disease) mediated by noncoding variants[17].

The role and utility of rare variation in the noncoding genome is likely to be a function of the number of noncoding regions that, when altered, disrupt gene expression or function to a high degree. While this can be estimated in model systems, there will be experimental confounds (for example, species, cell type, developmental stage) that limit interpretation. Direct analysis of WGS offers a complementary and irreplaceable approach to identifying and characterizing the role of rare noncoding variants in human disease.

WGS technology is sufficiently novel that we cannot accurately evaluate its potential in neuropsychiatric disorders without generating pilot data in human cohorts. It may implicate novel biological pathways missed by previous genomic efforts and identify disease-associated regulatory elements specific to certain cell types, developmental stages, or brain regions. Alternatively, WGS may prove less efficient than cheaper methods in identifying experimentally actionable disease-associated variation. Optimal allocation of future resources rests on efforts, such as the Whole-Genome Sequencing for Psychiatric Disorders Consortium (WGSPD), that critically evaluate the utility of WGS.

Estimating our ability to find rare noncoding variants

Finding disease-associated loci or variants by WGS will prove more challenging than with GWAS or whole-exome sequencing (WES). With WGS there are two orders of magnitude more sites to consider (~3 billion) compared to potential loci in GWAS (~20 million) or variants in WES (~30 million). Furthermore, we cannot predict functional changes (for example, to transcriptional rate) in the straightforward way we can predict changes to amino acids from coding variation.

To evaluate our power to detect noncoding variants in WGS data, we estimated the power to detect de novo PTVs that contribute to risk in ASD[4,18,19] if they occured in the noncoding genome. Without any additional information to help us distinguish signal from noise, for every one risk-mediating variant in the WGS data there would be about 25,000 non-risk variants (a ratio of 1:25000; Supplementary Table 3). By only considering variants with some evidence of functional effect (for example, conservation) or proximity to a gene with genome-wide significant association to ASD, we would expect to reduce the noise of non-risk variants, making the risk-mediating variant signal easier to detect. We considered a range of annotation scenarios, from an optimistic 1:5 to a pessimistic 1:500 (Supplementary Table 3). Moreover, as we do not know what penetrance to expect for these noncoding variants, we considered a wide range, shown as relative risk. For context, the highest

Box 1 | Types of genetic variation reliably detected by genomic technologies

Karyotype (≤ 1% common; ≤ 1% rare): Chromosomal aneuploidies, massive structural variation (for example, translocations, inversions, CNVs of millions of nucleotides), and some fragile sites with special protocols.

Microarray (~90% common; ~1% rare): Protein coding and noncoding common SNVs and large, rare CNVs (over ~20,000 nucleotides).

Exome sequencing (~1% common; ~1% rare): Protein coding common SNVs and indels, protein coding rare SNVs and indels, and some CNVs.

Low-coverage WGS (~95% common; ~85% rare): Protein coding and noncoding common SNVs, and most protein coding and noncoding rare SNVs.

Deep-coverage WGS (~99% common; ~99% rare): Protein coding and noncoding common SNVs and indels, protein coding and noncoding rare SNVs and indels, rare and common CNVs (over ~1,000 nucleotides), multi-allelic CNVs (for example, over three copies), mobile element insertions, and other structural variation (for example, translocations, inversions).

Long-read (>10,000 bp), deep-coverage WGS (~100% common; ~100% rare): For deep-coverage WGS plus: small CNVs (50–1,000 nucleotides), complex structural variation, variants in repetitive DNA, and direct assessment of phasing (whether two variants are on the same allele).

relative risks for common variants and de novo mutations in psychiatric disorders are about 1.3 and 50, respectively.

We first considered our ability to detect an overall excess of noncoding variants between cases and controls (a burden analysis). Such an analysis could identify a class of variants that mediate risk in psychiatric disorders, for example, promoters in proximity to ASD-associated genes, providing insight into regions of the noncoding genome most likely to yield specific risk variants for neuropsychiatric disorders. Since there is no clear category of noncoding variation equivalent to de novo PTVs, we adjusted for testing 1,000 annotation categories. The results for de novo and case–control analyses are shown in Fig. 1a,b (Supplementary Methods).

We next considered our ability to identify a specific genetic variant, functional element, or group of functional elements (for example, enhancers that regulate one gene) associated with risk that could be assessed in larger patient cohorts. The results for de novo and case–control analyses are shown in Fig. 1c,d (Supplementary Methods). From these analyses, it is clear that we will need (i) large cohorts and (ii) methods to decrease background noise (to obtain a high risk:non-risk ratio), for example, through predicting functional effects or regulation of known risk loci.

Why perform WGS in psychiatric disorders?

Given current uncertainty over the utility of WGS, we could wait until WGS for nonpsychiatric phenotypes provides sufficient insight to enable better power analyses. However, even large case–control cohorts may not be informative of the utility of WGS in ASD, for which de novo mutations have provided a more efficient approach to identifying specific genes and genetic loci[6,20] (Fig. 1). Additionally, there is a pressing need to identify specific cell types, tissues, and developmental stages involved in brain-based disorders, due to the complexity of the nervous system, limited understanding of how molecular changes lead to disorder, and difficulty in interpreting model systems. In short, the potential benefits of WGS in psychiatric disorders may be greater than in other phenotypes, and the availability of family-based cohorts may offer insights otherwise unobtainable.

Implications for neuroscientists

Interpreting the biology downstream of variants identified by existing WES and GWAS analyses remains a challenge; this is especially true in neuroscience due to the inaccessibility and complexity of neural tissue.

The interface of human genetics and neuroscience has typically focused on rare, highly penetrant variants that permit generation of transgenic animals with a robust phenotype[5,21–24]. Neuroscientists now face the challenge of obtaining biological insights through investigation of the multiple weakly penetrant variants, identified through modern genomics, that act through unknown neurological mechanisms in a manner highly dependent on genetic background[25]. Noncoding variants will pose yet harder challenges. Their effect sizes are likely to be small, and the relevant biology is likely to be restricted to specific cell types, developmental stages, or cell states. Analysis of three-dimensional chromatin structure must often be performed to identify the genes that a noncoding variant regulates. Finally, a proportion of noncoding variants may have human-specific functions absent in model organisms. For example, human accelerated regions, which are conserved across multiple species but differ within humans, are enriched for homozygous variants in consanguineous ASD cases[26].

Notwithstanding such challenges, many variants identified by genomic technologies have strong evidence of association with the disorders, creating a foundation for investigating pathogenesis. Furthermore, the presence of numerous variants allows systems analyses that identify biological convergences[5], and generate mechanistic hypotheses.

Strategies for improving locus discovery in WGS

Sample selection. As with other genomic technologies, large sample sizes will be key (Fig. 1 and Supplementary Fig. 1); the simplest way to achieve large cohorts will be through case–control studies (Table 2). However, alternative strategies have been proposed.

Several recent studies have shown an excess of deleterious variants in isolated populations that have expanded rapidly following recent bottlenecks[27–30], including deletions of the *TOP3B* gene, which is associated with schizophrenia and intellectual disability[29], in ~3% of individuals in Northern Finland compared

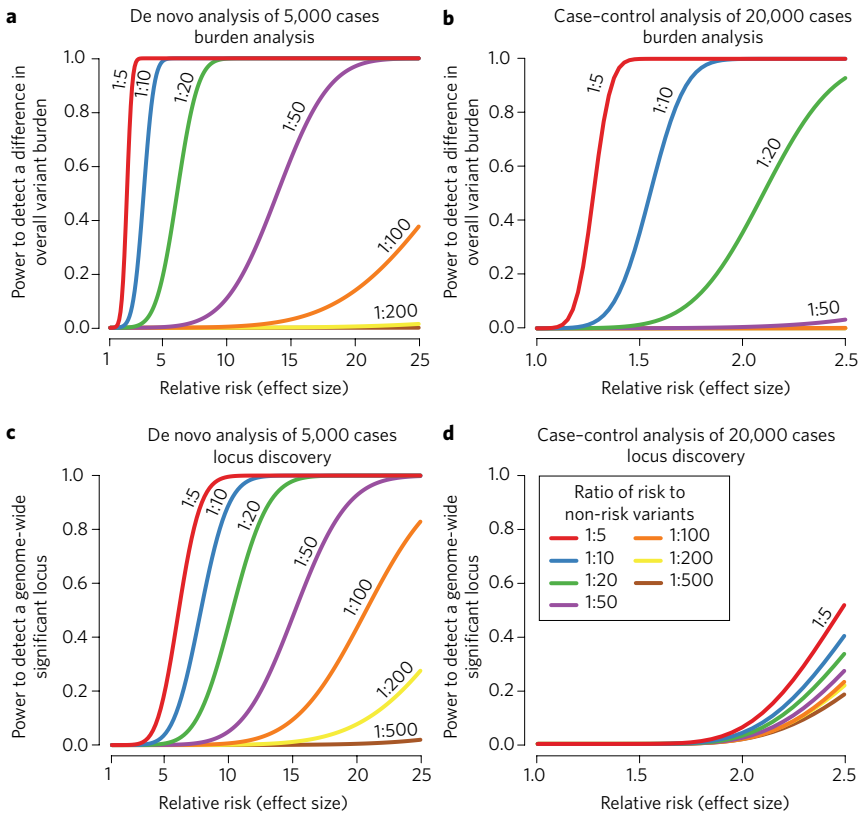

Fig. 1 | Statistical power in the noncoding genome. We estimated the power at a significance threshold (α) of 5×10^{-5}, selected to account for 1,000 categories of noncoding variants, to detect an excess of noncoding variants at 122,500 risk loci in cases vs. controls as we varied the relative risk and risk:non-risk ratio, which represents annotation quality (Supplementary Tables 1–3). **a**, We assessed the power for detecting an excess of de novo mutations in 5,000 cases vs. 5,000 controls as the relative risk increases. With a risk:non-risk ratio of 1:20, approximately equivalent to assessing protein-truncating variants in the coding genome, we achieve >80% power with a relative risk of 5. **b**, Using the same approach, we assessed the power to detect an excess burden of rare variants (allele frequency ≤ 0.1%) in 20,000 cases vs. 20,000 controls. **c**, We assessed the power to identify an excess of de novo mutations at a specific genomic locus, for example, the noncoding region regulating a single gene. Consequently, we set the significance threshold (α) at 2.5×10^{-6}. **d**, We assessed the power to identify an excess of rare variants (allele frequency ≤ 0.1%) at a specific nucleotide ($\alpha = 1.7 \times 10^{-11}$), since this yielded better power than testing for burden at a locus ($\alpha = 2.5 \times 10^{-6}$). The code to perform these analyses is available as an R package: https://github.com/sanderslab/wgsPowerCalc.

to 0.05% in other European populations. Large, multiplex pedigrees with multiple affected individuals may be enriched for rare, inherited variants with high effect sizes[31,32]. Simplex pedigrees, with only one affected individual, are enriched for de novo mutations with very high effect sizes, given the lack of exposure to natural selection. This strategy has succeeded in severe early-onset disorders, including intellectual disability and ASD[4,6,19,33]. Finally, consanguineous pedigrees may be enriched for homozygous variants that, like de novo mutations, are extremely rare with very high relative risks[26,34,35]. Homozygous variants may also play a role in non-consanguineous cases (Supplementary Table 4) and have been found to contribute to risk in some outbred ASD families[36,37]. Determining which of these sample selection strategies will be most successful will require WGS pilot projects under each strategy.

Integrating phenotypic data. Broadly, two contrasting approaches have been employed in integrating phenotypes in genomic studies, both with the aim of improving statistical power: (i) combining clinically or genetically related diagnoses to increase sample size and (ii) subdividing cohorts by shared phenotypes to decrease heterogeneity of the underlying genetics (subtyping). GWAS data demonstrate substantial common-variant sharing across current conventional diagnostic categories, for example, bipolar disorder and schizophrenia[38]. Similarly, genes identified by de novo mutations are frequently shared between ASD, intellectual disability, and developmental delay[4,19]. Thus, combining data from related diagnoses can increase sample size, hastening variant discovery[39].

The alternative approach, dividing by shared phenotypes, was critical for the discovery of Mendelian disorders by linkage methods, in which misclassifying one individual could prevent discovery. However, such an approach is risky for common, non-Mendelian psychiatric disorders given (i) current lack of insight into relevant subtypes and (ii) reduced sample size. A GWAS based on ~2,500 cases in the Simons Simplex Collection ASD cohort showed no improvement in the proportion of genetic heritability explained by the top SNPs, accounting for changes in sample size, for over ten phenotypic characteristics[40]. In contrast, a GWAS of a nonpsychiatric phenotype, bone mineral density, showed benefits of subgrouping, leading to the identification of 16 new loci[41].

Phenotypic subtyping also poses practical challenges. Genetic analysis is comparatively cheap, while deep phenotyping is cumbersome and costly, effectively diminishing sample size. The relative ease of using pre-existing cohorts and registries to inexpensively boost sample size has favored 'phenotype-light' sample collection. This balance could be shifted by the adoption of consistent phenotyping schema[42,43], identification of reliable neuropsychiatric biomarkers, or utilization of electronic medical records. Several large-scale initiatives are already working in this direction, for example deCODE[44], UK Biobank[45], Geisinger[46], and the All of Us Research Program (https://allofus.nih.gov/; formerly the Precision Medicine Initiative).

Identifying functional variants. Our assessment of statistical power (Fig. 1) shows that distinguishing variants that are likely to be functional and risk-mediating (i.e., high risk:non-risk ratio) will maximize discovery of specific noncoding variants. Several strategies might help.

Annotating the noncoding genome. Annotations may predict functional variants, including (i) conservation of DNA sequence across species; (ii) regions of open chromatin, where DNA is exposed, allowing proteins to bind (detected by DNase-seq or ATAC-seq); (iii) regions of active chromatin, where epigenetic marks suggest transcription of a nearby gene (detected by ChIP-seq); (iv) transcription

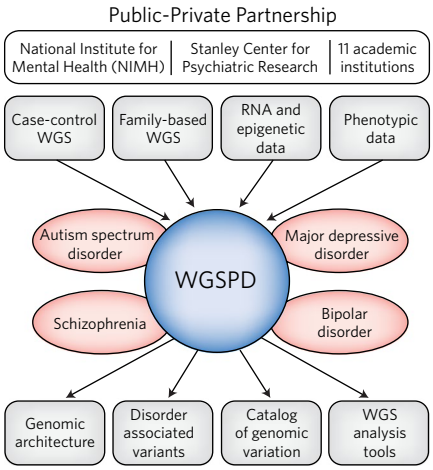

Fig. 2 | Overview of the WGSPD. The WGSPD is a public–private partnership between the NIMH, the Stanley Center for Psychiatric Research, and 11 academic intuitions. The initial focus is on four neuropsychiatric disorders, shown in red, integrating four types of data (grey, top) to achieve four outcomes (grey, bottom).

factor binding sites (detected by ChIP-seq); and (v) predicting the regulatory gene target using proximity to the variant (< 40% accurate[47]) or physical interactions with target loci (for example, ChIA-pet) or genome-wide (for example, Hi-C)[47]. Of note, many of these annotations may be tissue- and developmental-stage-specific[48–51]. For definitions of the genomic techniques mentioned, see Box 2.

Large-scale endeavors, such as ENCODE[52] and the Roadmap Epigenome Consortium (REC)[53], have created a reference for human epigenome annotation. Parallel efforts focused on brain tissue, such as the PsychENCODE Consortium[54], will help extend these resources[55].

Cataloging human variation. Building a database of human variation has proven invaluable in interpreting the coding genome[56], and the Genome Aggregation Database (gnomAD, http://gnomad.broadinstitute.org) extends this approach to WGS. Such data can be used to estimate regions of constraint (with less variation than expected), suggesting functionality[57–59].

Regions associated with psychiatric disorders. GWAS and WES have defined specific regions of the genome that contribute to psychiatric disorders, particularly in ASD[4,6,19] and schizophrenia[7]. It is plausible that noncoding variation in proximity to these regions will be enriched for risk-mediating variants.

Large variants. On average, large variants, especially deletions, have greater potential to mediate risk than small variants[6]. However, while indels (insertions or deletions, i.e., a gain or loss of ≤ 50bp) and CNVs may have a greater impact on noncoding function, there are considerably fewer such variants than SNVs[8]. The utility of this strategy will depend on the balance between these two opposing effects.

Functional validation. Methods have been developed to assess the functional effects of large numbers of potential regulatory regions. These massively parallel reporter assays[60], including self-transcribing active regulatory region sequencing (STARR-seq)[61], assess the function of a regulatory region by its potential to transcribe itself or a specific sequence of DNA (barcode). Of note, this ability to functionally validate noncoding variants en masse is a major benefit over interpreting coding missense variants, for which protein-specific functional assays are usually required.

The Whole Genome Sequencing Consortium for Psychiatric Disorders

The potential for WGS to help understand neuropsychiatric disorders, and the absence of insight into the role of rare noncoding variants, prompted the United States National Institute of Mental Health (NIMH) to fund four pilot projects aimed at generating WGS data in neuropsychiatric disorders to provide a more complete understanding of their genomic architecture.

Big questions in biology are akin to solving problems of similar complexity in other disciplines such as particle physics or astronomy and require a 'Team Science' approach[62]. Recognizing the need for large samples sizes to make progress (Table 1 and Fig. 1), the NIMH, the Stanley Center

Box 2 | Genomic and functional genomic technologies.

Assay for transposase-accessible chromatin (ATAC)-seq: Identifies regions of open (accessible) chromatin by using transposons to insert sequencing adaptors.

Chromatin-interaction analysis by paired-end tag (ChIA-PET)-seq: Identifies physical interactions of DNA bound by a specific protein. ChIP (see below) is used to enrich DNA bound to the protein. DNA ends in proximity are joined, and then the DNA attached to the protein is released and sequenced.

Chromatin immunoprecipitation (ChIP)-seq: Identifies DNA bound by a specific protein, e.g., transcription factors or histone marks. Proteins are crosslinked to DNA, the DNA outside of proteins is digested, and antibodies are used to precipitate the specific protein. The DNA attached to these proteins is released and sequenced.

Hi-C: Identifies physical interactions between DNA loci in the nucleus. Building on the chromosome conformation capture (3C) method, in which DNA ends in proximity are joined and primers specific to sites of interest are used to assess the presence of interactions, the primers are replaced with high-throughput sequencing so that physical interactions between all genomic loci can be detected.

DNase-seq: Identifies regions of open (accessible) chromatin by digesting DNA with the DNase I enzyme and attaching sequencing adaptors to the breakpoints generated.

Whole-exome sequencing (WES): Identifies DNA variants in the exons of protein coding genes.

Whole-genome sequencing (WGS): Identifies DNA variants throughout the entire genome.

Table 2 | Individuals with WGS data generated by, or accessible to, the WGSPD

Data being generated by the WGSPD

Project	Disorder	Cases	Controls	Details
1	Schizophrenia	3,333	1,667	Case–control analysis; African American ancestry
	Bipolar disorder	3,333	1,667	Case–control analysis; African American ancestry
2	ASD	378	1,512	Simplex families with two parents, affected child, unaffected child
	Schizophrenia	281	843	Families with two parents and one or more affected individuals
3	Schizophrenia	1,000	1,400	Case–control analysis of individuals from Finland
	Bipolar disorder	1,000	500	Case–control analysis of individuals from Finland
	Schizophrenia	650	325	Case–control analysis of individuals from Netherlands
	Bipolar disorder	650	325	Case–control analysis of individuals from Netherlands
	Bipolar disorder	62	138	Multiplex families with affected and unaffected individuals from Colombia
	Bipolar disorder	83	170	Multiplex families with affected and unaffected individuals from Costa Rica
4	Schizophrenia	271	280	Multiplex families with affected and unaffected individuals
	Bipolar disorder	299	309	Multiplex families with affected and unaffected individuals
	Major depression	476	492	Multiplex families with affected and unaffected individuals

Data being generated by other funding mechanisms with consistent analysis pipelines

	Disorder	Cases	Controls	Details
	ASD*	5,302	15,856	Families with two parents, affected child, +/− unaffected child
	ASD*	150	150	Multiplex families with affected and unaffected individuals
	Schizophrenia	118	198	Multiplex families with affected and unaffected individuals
	Bipolar disorder	118	198	Multiplex families with affected and unaffected individuals
	Major depression	478	804	Multiplex families with affected and unaffected individuals
	TOPMed†	0	68,950	Heart, lung, blood, and sleep disorders
	CCDG†	0	63,950	Heart, vascular, lung, bowel, neurological, and endocrine disorders
	Totals	**17,957**	**165,834**	

*ASD samples are being generated by several groups: CCDG of the National Human Genome Research Institute (NHGRI), Simons Foundation Autism Research Initiative (SFARI)[73], and Autism Sequencing Consortium (ASC)[74]. †6,100 samples are shared between TOPMed of the National Heart, Lung, and Blood Institute (NHLBI) and CCDG, and therefore the total number of samples was reduced by 3,050 for each cohort. These cohorts are composed of individuals ascertained for nonpsychiatric disorders and individuals whose psychiatric disorder status is generally unknown.

for Psychiatric Research, and researchers at 11 academic institutions across the USA that were funded in the four selected projects have formed a public–private partnership: the WGSPD. This consortium aims to establish a repository of WGS data, processed in a consistent manner, to facilitate large-scale analyses within and across four psychiatric disorders (Fig. 2). This approach can make more efficient use of funding and resources, for example, by using a central data repository, consistent analysis pipelines, and collaborative methods development to help all researchers access and use the data.

The WGSPD will need to expand, both beyond the founding members and beyond these four disorders. Investigators with relevant WGS data will be invited to join the WGSPD and participate in working groups focused on specific disorders or cross-disorder projects. Given the scale of WGS data, the cost of reprocessing the data in a consistent manner and storing the data will be substantial. Establishing a suitable funding strategy for such genomic integration is a key question that needs to be addressed urgently throughout the genomics community. In a first step to improve this, WGS analysis pipelines have been coordinated across several major sequencing centers and consortia (for example, Centers for Common Disease Genomics (CCDG), Trans-Omics for Precision Medicine (TOPMed), and WGSPD) to allow direct comparison of results. To obtain the sample sizes necessary (Fig. 1), a similar consensus will need to be established internationally.

Cloud-based analysis

The sheer scale of WGS datasets necessitates new models for data analysis, since data storage and computation is likely to be beyond the resources of any single institution. Fortunately, the development of cloud-based computing has coincided with the generation of WGS data. Under this model, a single cloud-based data repository can be accessed by teams at each collaborating site, and cloud-based analysis eliminates the need for cumbersome and costly downloads. This approach has the further advantage of facilitating the sharing of preinstalled algorithms and pipelines, encouraging consistent consortium-wide analysis.

The scale of WGS data can make simple analytical tasks overwhelming. Therefore, the WGSPD is committed to developing application program interfaces and software solutions for the wider community to simplify cloud-based data access (for example, Hail[63]). In doing so, computational biologists and analysts can focus on the development and application of methods for analysis, rather than on lower level data management and handling.

The analysis of deidentified genetic data on university-hosted remote servers is common practice, with contributing sites being responsible for securing nongenetic identifying information. So long as cloud environments meet security standards equivalent to those applied to existing remote servers, existing informed consent will cover this use, except in rare instances where the consent specifically excludes this approach. Best practice guidelines for secure

sharing of genomic data have been described by the NIH (https://www.ncbi.nlm.nih.gov/projects/gap/pdf/dbgap_2b_security_procedures.pdf). There is an urgent need for methods that allow such guidelines to be easily adopted and readily vetted across cloud providers and institutions.

The WGSPD projects and data

The four WGSPD projects, developed by independent sets of investigators, encompass the diverse strategies for improving locus discovery and therefore will provide some of the earliest opportunities to assess their relative utility in complex disorders. The four projects are:

1) Case–control analysis of schizophrenia and bipolar disorder in individuals of African American ancestry.
2) Family-based analysis of ASD in families with a single affected child, allowing the detection of de novo mutations.
3) Case–control analysis of schizophrenia or bipolar disorder in isolated populations with recent population bottlenecks.
4) Family-based analysis of schizophrenia, bipolar disorder, or major depressive disorder in families with multiple affected individuals.

Combining these WGS cohorts with consistently processed WGS data from other consortia will yield an initial dataset of 183,000 individuals, including 18,000 cases and 165,000 controls (Table 2). In addition to the genotype data, we are collating phenotype data that are comparable across projects, disorders, and ages to allow in-depth genotype–phenotype analysis.

Conclusion

The noncoding genome remains largely unexplored, and major discoveries undoubtedly await intrepid pioneers. Whole-genome sequencing of neuropsychiatric disorders provides an important avenue in this exploration, potentially offering high-resolution insight into the developmental stages, brain regions, cell types, and biological functions that underlie these disorders. If the cost of sequencing continues to fall, it is inevitable that WGS will ultimately replace both microarray and WES; the key question is, at what price point will this transition offer a good return on investment? Pooling preliminary WGS data between researchers and across disorders offers the most efficient mechanism to make this determination.

The creation of the WGSPD has allowed numerous researchers to pursue diverse scientific approaches on multiple psychiatric disorders, while simultaneously working toward a harmonized dataset for integrated analysis. The pooling of expertise, methods, and data will accelerate progress toward understanding genetic contributions to brain development, function, and pathology and create a resource that will continue to yield scientific and clinical insights for years to come. ❐

Stephan J. Sanders[1], Benjamin M. Neale[2,3], Hailiang Huang[2,3], Donna M. Werling[1], Joon-Yong An[1], Shan Dong[1], Whole Genome Sequencing for Psychiatric Disorders (WGSPD)[4], Goncalo Abecasis[5], P. Alexander Arguello[6], John Blangero[7], Michael Boehnke[5], Mark J. Daly[2,3], Kevin Eggan[3], Daniel H. Geschwind[8,9], David C. Glahn[10], David B. Goldstein[11], Raquel E. Gur[12], Robert E. Handsaker[3], Steven A. McCarroll[3], Roel A. Ophoff[13,14,15], Aarno Palotie[3], Carlos N. Pato[16], Chiara Sabatti[17], Matthew W. State[1], A. Jeremy Willsey[1], Steven E. Hyman[3]*, Anjene M. Addington[5]*, Thomas Lehner[5]* and Nelson B. Freimer[14]*

[1]*Department of Psychiatry, UCSF Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA, USA.* [2]*Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA.* [3]*Stanley Center for Psychiatric Research and Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA.* [4]*A full of list of members appears in the Supplementary Note.* [5]*Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI, USA.* [6]*National Institute of Mental Health, Bethesda, MD, USA.* [7]*South Texas Diabetes and Obesity Institute, University of Texas Rio Grande Valley, School of Medicine, Brownsville, TX, USA.* [8]*Department of Neurology, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA.* [9]*Center for Autism Research and Treatment, Semel Institute, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA.* [10]*Department of Psychiatry, Yale University School of Medicine, New Haven, CT, USA.* [11]*Institute for Genomic Medicine, Columbia University Medical Center, Hammer Health Sciences, New York, NY, USA.* [12]*Department of Psychiatry, Neuropsychiatry Section, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.* [13]*Brain Center Rudolf Magnus, Department of Psychiatry, University Medical Center Utrecht, Utrecht, The Netherlands.* [14]*Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California Los Angeles, Los Angeles, CA, USA.* [15]*Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA.* [16]*Department of Psychiatry, Zilkha Neurogenetic Institute, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA.* [17]*Department of Health Research and Policy, Division of Biostatistics, Stanford University, Stanford, CA, USA. Stephan J. Sanders and Benjamin M. Neale contributed equally to this work.*

**e-mail: seh@harvard.edu; anjene.addington@nih.gov; tlehner@mail.nih.gov; nfreimer@mednet.ucla.edu*

Published online: 28 November 2017
https://doi.org/10.1038/s41593-017-0017-9

References

1. Owen, M. J., Sawa, A. & Mortensen, P. B. *Lancet* **388**, 86–97 (2016).
2. Power, R. A. et al. *JAMA Psychiatry* **70**, 22–30 (2013).
3. Sekar, A. et al. *Nature* **530**, 177–183 (2016).
4. De Rubeis, S. et al. *Nature* **515**, 209–215 (2014).
5. Sanders, S. J. *Curr. Opin. Genet. Dev.* **33**, 80–82 (2015).
6. Sanders, S. J. et al. *Neuron* **87**, 1215–1233 (2015).
7. Schizophrenia Working Group of the Psychiatric Genomics Consortium. *Nature* **511**, 421–427 (2014).
8. Brandler, W. M. et al. *Am. J. Hum. Genet.* **98**, 1–13 (2016).
9. Collins, R. L. et al. *Genome Biol.* **18**, 36 (2017).
10. Chiang, C. et al. *Nat. Genet.* **49**, 692–699 (2017).
11. Hindorff, L. A. et al. *Proc. Natl. Acad. Sci. USA* **106**, 9362–9367 (2009).
12. Maurano, M. T. et al. *Science* **337**, 1190–1195 (2012).
13. Siepel, A. et al. *Genome Res.* **15**, 1034–1050 (2005).
14. Visel, A. et al. *Cell* **152**, 895–908 (2013).
15. Willsey, A. J. et al. *Cell* **155**, 997–1007 (2013).
16. Gasperini, M. et al. *Am. J. Hum. Genet.* **101**, 192–205 (2017).
17. Scacheri, C. A. & Scacheri, P. C. *Curr. Opin. Pediatr.* **27**, 659–664 (2015).
18. Sanders, S. J. et al. *Nature* **485**, 237–241 (2012).
19. Iossifov, I. et al. *Nature* **515**, 216–221 (2014).
20. McRae, J. F. et al. *Nature* **542**, 433–438 (2017).
21. Katz, D. M. et al. *Trends Neurosci.* **39**, 100–113 (2016).
22. Berrios, J. et al. *Nat. Commun.* **7**, 10702 (2016).
23. Erickson, C. A. et al. *J. Autism Dev. Disord.* **44**, 958–964 (2014).
24. de la Torre-Ubieta, L., Won, H., Stein, J. L. & Geschwind, D. H. *Nat. Med.* **22**, 345–361 (2016).
25. Sittig, L. J. et al. *Neuron* **91**, 1253–1259 (2016).
26. Doan, R. N. et al. *Cell* **167**, 341–354.e12 (2016).
27. Lim, E.T. et al. *PLoS Genet.* https://doi.org/10.1371/journal.pgen.1004494 (2014).
28. Service, S.K. et al. *PLoS Genet.* https://doi.org/10.1371/journal.pgen.1004147 (2014).
29. Stoll, G. et al. *Nat. Neurosci.* **16**, 1228–1237 (2013).
30. Gudbjartsson, D. F. et al. *Nat. Genet.* **47**, 435–444 (2015).
31. Cirulli, E. T. & Goldstein, D. B. *Nat. Rev. Genet.* **11**, 415–425 (2010).
32. Leppa, V. M. et al. *Am. J. Hum. Genet.* **99**, 540–554 (2016).
33. Laumonnier, F. et al. *Am. J. Hum. Genet.* **74**, 552–557 (2004).
34. Novarino, G. et al. *Science* **338**, 394–397 (2012).
35. Gamsiz, E. D. et al. *Am. J. Hum. Genet.* **93**, 103–109 (2013).
36. Lim, E. T. T. et al. *Neuron* **77**, 235–242 (2013).
37. Yu, T. W. W. et al. *Neuron* **77**, 259–273 (2013).
38. Lee, S. H. et al. *Nat. Genet.* **45**, 984–994 (2013).
39. Psychiatric GWAS Consortium Bipolar Disorder Working Group. et al. *Nat. Genet.* **43**, 977–983 (2011).
40. Chaste, P. et al. *Biol. Psychiatry* **77**, 775–784 (2015).
41. Saint-Pierre, A. et al. *Eur. J. Hum. Genet.* **19**, 710–716 (2011).
42. Köhler, S. et al. *Nucleic Acids Res* **45**, D865–D876 (2016).
43. Insel, T. et al. *Am. J. Psychiatry* **167**, 748–751 (2010).
44. Stefansson, H. et al. *Nature* **505**, 361–366 (2014).
45. Kendall, K. M. et al. *Biol. Psychiatry* **82**, 103–110 (2016).
46. Dewey, F. E. et al. *Science* **354**, aaf6814 (2016).
47. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. *Nature* **489**, 109–113 (2012).
48. Rao, S. S. P. et al. *Cell* **159**, 1665–1680 (2014).
49. Sahlén, P. et al. *Genome Biol.* **16**, 156 (2015).
50. Schoenfelder, S. et al. *Genome Res.* **25**, 582–597 (2015).
51. Babaei, S. et al. *PLOS Comput. Biol.* **11**, e1004221 (2015).
52. ENCODE Project Consortium. *Science* **306**, 636–640 (2004).
53. Kundaje, A. et al. *Nature* **518**, 317–330 (2015).
54. Akbarian, S. et al. *Nat. Neurosci.* **18**, 1707–1712 (2015).
55. Won, H. et al. *Nature* **538**, 523–527 (2016).
56. Lek, M. et al. *Nature* **536**, 285–291 (2016).
57. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. *PLoS Genet.* **9**, e1003709 (2013).
58. Samocha, K. E. et al. *Nat. Genet.* **46**, 944–950 (2014).

59. Kosmicki, J. A. et al. *Nat. Genet.* **49**, 504–510 (2017).
60. Melnikov, A., Zhang, X., Rogov, P., Wang, L. & Mikkelsen, T. S. *J. Vis. Exp.* **90**, e51719 (2014).
61. Arnold, C. D. et al. *Science* **339**, 1074–1077 (2013).
62. Lehner, T., Senthil, G. & Addington, A. M. *Biol. Psychiatry* **77**, 6–14 (2015).
63. Ganna, A. et al. *Nat. Neurosci.* **19**, 1563–1565 (2016).
64. The Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium. *Mol. Autism* **8**, 21 (2017).
65. Hyde, C. L. et al. *Nat. Genet.* **48**, 1–9 (2016).
66. Marshall, C. R. et al. *Nat. Genet.* **49**, 27–35 (2017).
67. Green, E. K. et al. *Mol. Psychiatry* **21**, 89–93 (2016).
68. Rucker, J. J. H. et al. *Biol. Psychiatry* **79**, 329–336 (2016).
69. Purcell, S. M. et al. *Nature* **506**, 185–190 (2014).
70. Genovese, G. et al. *Nat. Neurosci.* **19**, 1433–1441 (2016).
71. Fromer, M. et al. *Nature* **506**, 179–184 (2014).
72. Singh, T. et al. *Nat. Neurosci.* **19**, 571–577 (2016).
73. Fischbach, G. D. & Lord, C. *Neuron* **68**, 192–195 (2010).
74. Buxbaum, J. D. et al. Autism Sequencing Consortium. *Neuron* **76**, 1052–1056 (2012).

Acknowledgements

The authors acknowledge and thank the study participants and their families. The WGSPD is a public–private partnership between the NIMH, the Stanley Center for Psychiatric Research, and researchers at 11 academic institutions across the USA. This work was supported by grants from the NIMH, specifically U01 MH105653 (M.B.), U01 MH105641 (S.A.M.), U01 MH105573 (C.N.P.), U01 MH105670 (D.B.G.), U01 MH105575 (M.W.S., A.J.W.), U01 MH105669 (M.J.D., K.E.), U01 MH105575 (N.B.F., D.H.G., R.A.O.), U01 MH105666 (A.P.), U01 MH105630 (D.C.G.), U01 MH105632 (J.B.), U01 MH105634 (R.E.G.), U01 MH100239-03S1 (M.W.S., S.J.S., A.J.W.), R01 MH095454 (N.B.F.); by grants from the Simons Foundation (SFARI #385110, M.W.S., S.J.S., A.J.W., D.B.G., SFARI #401457 (D.H.G.)); and by a gift from the Stanley Foundation (S.E.H.).

Author contributions

S.J.S., B.M.N., H.H., and D.M.W. contributed to the power calculation. All authors contributed to the text.

Competing interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41593-017-0017-9.