A framework for the investigation of rare genetic disorders in neuropsychiatry

Stephan J. Sanders1, Mustafa Sahin2, Joseph Hostyk3, Audrey Thurm4, Sebastien Jacquemont5, Paul Avillac6, Elise Douard5, Christa L. Martin7, Meera E. Modi7, Andres Moreno-De-Luca7, Armin Raznahan8, Alan Anticevic9, Ricardo Dolmetsch10, Guoping Feng11,12, Daniel H. Geschwind13, David C. Glahn8, David B. Goldstein3, David H. Ledbetter7, Jennifer G. Mulle14, Sergiu P. Pasca15,16, Rodney Samaco16, Jonathan Sebat17, Anne Pariser18, Thomas Lehner4, Raquel E. Gur19* and Carrie E. Bearden20*

Denovo and inherited rare genetic disorders (RGDs) are a major cause of human morbidity, frequently involving neuropsychiatric symptoms. Recent advances in genomic technologies and data sharing have revolutionized the identification and diagnosis of RGDs, presenting an opportunity to elucidate the mechanisms underlying neuropsychiatric disorders by investigating the pathophysiology of high-penetration genetic risk factors. Here we seek out the best path forward for achieving these goals. We think future research will require consistent approaches across multiple RGDs and developmental stages, involving both the characterization of shared neuropsychiatric dimensions in humans and the identification of neurobiological commonalities in model systems. A coordinated and concerted effort across patients, families, researchers, clinicians and institutions, including rapid and broad sharing of data, is now needed to translate these discoveries into urgently needed therapies.

RGDs are defined as genetic disorders affecting fewer than 200,000 individuals in the United States*, or 1 in 2,000 individuals; however, a few highly penetrant genetic variants that occur more frequently (such as trisomy 21, affecting 1 in 600 individuals*) are often included. RGDs can be caused by the alteration of single nucleotides in one gene, entire chromosomes with hundreds of genes, or anything in between (Fig. 1a). Though they are individually rare, the contribution of RGDs to human morbidity en masse is substantial, with 18% of protein-coding genes currently individually rare, the contribution of RGDs to human morbidity
data sharing has revolutionized the identification and diagnosis of RGDs, presenting an opportunity to elucidate the mechanisms underlying neuropsychiatric disorders by investigating the pathophysiology of high-penetration genetic risk factors. Here we seek out the best path forward for achieving these goals. We think future research will require consistent approaches across multiple RGDs and developmental stages, involving both the characterization of shared neuropsychiatric dimensions in humans and the identification of neurobiological commonalities in model systems. A coordinated and concerted effort across patients, families, researchers, clinicians and institutions, including rapid and broad sharing of data, is now needed to translate these discoveries into urgently needed therapies.

Some RGDs have near-complete penetrance and distinctive features that enable syndromes to be described, affected and unaffected family members to be reliably distinguished, and the underlying genetic locus identified (as for Dravet syndrome, an RGD caused by SCN1A mutations resulting in a severe form of epilepsy6–8). Newer genomic technologies, including chromosomal microarray; whole-exome sequencing (WES) and whole-genome sequencing (WGS) (see Box 1, Glossary), have identified many additional RGDs with less distinctive features, lower penetrance and/or more variable effects on phenotype. Many of these newly described RGDs have milder phenotypic impacts that may be largely brain specific (for example, 15q13.3 deletions increase risk for psychosis 18-fold; Supplementary Table 2), in contrast to those with large effect sizes and multisystem involvement (for example, in cases of trisomy 21, intellectual disability (ID), global developmental delay and short stature are nearly universal†). Genomic technologies have also expanded the known manifestations of previously recognized RGDs (for example, many individuals with SCN2A mutations present with ASD rather than seizures). The full extent to which RGDs contribute to the manifestations of behavioral aberration, evident in several neuropsychiatric disorders,
is still emerging. RGDs appear to contribute significantly to the
genetic architecture of neuropsychiatric disorders characterized
by high heritability, early age of onset, reduced fecundity, impaired
cognition and behavioral deficits\(^{12-14}\). In pediatric patients referred
for developmental delay, ID and/or ASD, the diagnostic yield of
clinical genetic testing (microarray and/or whole-exome sequenc-
ing) is more than 30%\(^{15-19}\). It is important to note, however, that
individuals with ASD with comorbid ID show a significantly higher
rate of rare, de novo damaging variants than those with normal IQ,
whereas there appears to be a greater contribution from common
variants in those with higher IQ\(^{5,20,21}\). Therefore, RGDs contribute
significantly to neurodevelopmental disorders, such as ID, ASD and
epilepsy\(^{22-24}\); contribute modestly to child- or adolescent-onset
neuropsychiatric disorders (attention deficit hyperactivity disorder
(ADHD))\(^{25}\), schizophrenia\(^{26}\) and Tourette’s syndrome\(^{27}\); and con-
tribute less to later-onset neuropsychiatric disorders (bipolar disor-
der and major depressive disorder\(^{8,14,27,28}\)).

Understanding the contribution of RGDs to neuropsychiatric
disorders is critical for patient care and for developing effective
therapeutics. Diagnosing and understanding RGDs enables screen-
ing for specific risks, early detection, informed family planning,
initiation of services and therapies, detailed prognosis and support.
Notably, recent advances in genetically targeted therapies\(^{29,30}\) raise
the possibility of treating the underlying pathology. At the same
time, the importance of this research is not limited to those diag-
nosed with RGDs. These high-penetrance disorders also have the
potential to provide key insights into the underlying biology of idio-
pathic neuropsychiatric disorders.

**Fig. 1 | Overview of rare genetic disorders (RGDs).** a, RGDs may be caused by variants that affect one gene (purple) or many genes (green). Many
aneuploidies and structural variants arise spontaneously at higher rates than single-gene disorders, leading to comparatively high population frequencies
for a given penetrance\(^{12,14}\). b, The CNS is involved in the majority of single-gene RGDs. c, Single-gene RGDs frequently affect multiple neuropsychiatric
domains, as shown by extensive co-occurrence of Human Phenotype Ontology terms (Supplementary Table 1)\(^{125}\). Terms that co-occur in at least 200
RGDs are shown as nodes (colored circles, size determined by the number of RGDs), with edge weight (gray lines) determined by the degree of co-
occurrence of a term between RGDs (203-1,114). Network layout is based on the Compound Spring Embedder algorithm\(^{126}\). OMIM, Online Mendelian
Inheritance in Man; G2P, Gene2Phenotype; HPO, Human Phenotype Ontology; CNS, central nervous system; PNS, peripheral nervous system; UMN, upper
motor neuron. Credit: Debbie Maizels/Springer Nature.
Box 1 | Glossary

**Single-nucleotide variant (SNV)** A genetic variant in which one nucleotide (for example, C) is changed to another (for example, T).

**Single-nucleotide polymorphism (SNP)** A common SNV (for example, found in ≥1% of a population).

**Insertion or deletion (indel)** The gain or loss of 1–50 nucleotides, usually detected by sequencing.

**Missense variant** An SNV in a gene that changes one amino acid critical to protein function or a PTV.

**Loss-of-function (LoF) variant** Genetic variants predicted to seriously disrupt the function of protein-coding genes, for example, a missense variant at an amino acid critical to protein function or a PTV.

**Low probability loss-of-function intolerant (lPI) score** Probability that a given gene is intolerant of PTVs based on lower-than-expected rates of PTVs in the general population, suggestive of selective pressure. Genes with high lPI scores (≥0.9) are extremely LoF intolerant, whereas genes with low lPI scores (≤0.1) may be LoF tolerant or there may be insufficient data to assess their tolerance.

**Copy-number variant (CNV)** The gain (duplication) or loss (deletion) of ≥50 nucleotides (previously ≥1,000 bp), but often thousands to millions of nucleotides. Can be detected by microarray (large CNVs only) or WGS.

**Structural variant (SV)** A large-scale rearrangement of DNA. Can include CNVs, but also inversions, translocations, or more complex rearrangements.

**Germline de novo variant** A new genetic variant observed in a child but not in either parent.

**Somatic de novo variant** A new genetic variant observed in some cells of an individual but not others.

**Chromosomal microarray (CMA)** Technology that enables the number of copies of DNA to be assessed at thousands of locations across the genome, enabling the detection of CNVs. Many SNP genotyping microarrays detect SNPs at these locations too.

**Whole-exome sequencing (WES)** Technology that assesses ~45 million individual nucleotides in regions of DNA that encode proteins (exons), enabling the detection of SNVs and indels.

**Whole-genome sequencing (WGS)** Technology that assesses ~3.2 billion individual nucleotides across the genome, enabling the detection of SNVs, indels, CNVs and other SVs.

**Locus (plural: loci)** A region of DNA ranging from a single nucleotide to an entire chromosome. For RGDs, it often describes a single gene or structural variant.

**Mendelian disorder** A disorder that follows a Mendelian pattern of inheritance. Changes at several loci may elicit the same disorder.

**Complex disorder** A disorder caused by multiple variants with a range of effect sizes at multiple loci, often in combination with environmental factors.

**Neuropsychiatric** Describes a behavioral or emotional disturbance due to an abnormality in the structure or function of the central nervous system.

**Neuropsychiatric domain** A psychological construct relevant to human behavior and mental disorders that can be measured along a continuum in health and disease (e.g., anxiety).

**Autosomes** Chromosomes 1–22 (in humans), in contrast to the sex chromosomes (X, Y) or mitochondria (M).

**Autosomal dominant (AD)** Describes a pattern of Mendelian inheritance in which a single variant on one copy of an autosomal chromosome causes the disorder or trait.

**Autosomal recessive (AR)** Describes a pattern of Mendelian inheritance in which two copies of a variant, one on each copy of an autosomal chromosome, must be present in order to cause the disorder or trait.

**Compound heterozygous** A pattern of Mendelian inheritance in which two different variants at the same locus, one on each copy of a chromosome, cause the disorder or trait.

**X-linked recessive (XLR)** A pattern of Mendelian inheritance in which one variant on chromosome X causes the disorder in males only. It is not uncommon for carrier females to manifest mild or tissue-localized symptoms of disease due to normal or skewed X-inactivation of the functional copy of the gene in individual cells.

---

A US National Institute of Mental Health (NIMH) workshop was held in September 2017 with the goal of identifying strategies for advancing the understanding and management of neuropsychiatric disorders through a focus on RGDs (see Supplementary Table 3 for list of attendees). This Perspective outlines guidelines and recommendations that emerged from this workshop. In particular, three priorities were identified (see Box 2).

**The overlap between rare genetic disorders and psychiatric disorders** Individuals affected by RGDs frequently have multiple neuropsychiatric symptoms, and specific RGDs frequently lead to multiple neuropsychiatric symptoms (grouped into neuropsychiatric domains) in different individuals (Fig. 1c). The co-occurrence of neuropsychiatric phenotypes within single-gene RGDs (Supplementary Table 1) reveals a highly connected network between developmental delay, behavioral abnormalities, seizures and abnormalities of the motor system (Fig. 1c).

Specific RGDs occupy different regions of this co-occurrence network, leading to widespread variation between RGDs in the domains affected (Fig. 2a,b). Notably, some RGDs appear to have strong effects on specific neuropsychiatric domains, as exemplified by the link between the RGD 22q11.2 deletion syndrome and psychosis⁵⁴, and the putative protective effect of the reciprocal 22q11.2 duplication against psychosis⁵⁵ (Fig. 2b, Extended Data Fig. 1, Supplementary Tables 2 and 4). In contrast, the association between the frequency of RGD diagnosis and IQ in ASD suggests that there is a global effect on cognition across many RGDs (Fig. 2a), and this notion is supported by large-scale screens in the general population⁵,⁶,⁵³,⁵⁴,⁵⁶. In addition, recent evidence suggests that common genetic variation may influence the phenotypic expression of RGDs, and thus may contribute to variation in clinical presentation (Fig. 2a)⁵⁵. More data are required to more precisely address the contribution of these RGD-causing mutations to neuropsychiatric traits, ideally from cohorts with consistent genotype and phenotype data ascertained without bias toward selection of individuals with particular characteristics, such as population-based birth cohorts, or with a proper strategy to control for ascertainment bias, such as assessing the same clinically presenting disorder across multiple RGDs and/or using first-degree relatives as controls⁵⁴,⁵⁶ to account for additional genetic and environmental factors (see Fig. 2c,d).

Further confounding our understanding is that the phenotypic severity within each RGD varies widely (Fig. 2a). Presently, our knowledge of individuals with RGDs is biased toward those who are clinically identified and thus likely more severely affected, with milder cases (Fig. 2d) and atypical presentations under-represented⁵⁴,⁵⁶.
At the same time, studies of brain, cognition and behavior in RGDs (i.e., phenotyping studies) often have the opposite bias, assessing only high-functioning individuals. This ascertainment bias complicates genetic counseling, as it is not yet possible to provide an accurate representation of prognosis.

To address Priority 1 (Box 2), we think that proper characterization of RGD effects on brain-related phenotypes requires moving away from categorical diagnoses (such as ASD) to neuropsychiatric domains assessed as quantitative traits (such as social impairment)\textsuperscript{28},\textsuperscript{29}, as advocated for by the NIMH Research Domain Criteria (RDoC) initiative\textsuperscript{30,31},\textsuperscript{56} and assessing traits consistently across carriers who were referred to the clinic (probands) and carriers who were not (family members, individuals from unselected populations) regardless of presenting diagnosis (Fig. 2d)\textsuperscript{36,40,41}. Capturing the full spectrum of phenotypic severity and controlling for ascertainment bias that often leads to inflated estimates of severity (Fig. 2d)\textsuperscript{42} may require conducting population-level analyses of hundreds of thousands of individuals, for example, using birth cohorts or health system registries\textsuperscript{33,34,43}, and applying this experimental design across generations to capture early- and late-onset manifestations\textsuperscript{41,56,41,42}.

Prioritizing RGDs for research

We think that the allocation of resources for research into an RGD should be commensurate with the level of evidence for genetic association to a specific neuropsychiatric domain. In order for resources to be provided for research, genome-wide statistical association should generally be demonstrated for a neuropsychiatric domain\textsuperscript{4,46,47} in a defined cohort\textsuperscript{31,34,36,40,41,56}; however, for very rare neuropsychiatric disorders (such as childhood-onset schizophrenia), a lower level of evidence may be appropriate\textsuperscript{5}. Further, in order to identify supporting mechanistic insights, we assert that it is important that functional data are used to enhance these statistical models (for example, protein-truncating variant (PTV) as opposed to missense mutation data, pLI score; see Box 1, Glossary)\textsuperscript{52}. For genes for which clear functional assays are available, systematic evaluation of observed variants in both cases and controls\textsuperscript{11,53} may also contribute to statistical association. In the absence of clear biomarkers (see below) indicating a role for a genetic variant in the etiology of disease, observing downstream, nonspecific effects for individual variants, such as altered synaptic function or behavioral assays, is not a substitute for identifying a statistical association between the genetic variant and the human condition. For genetic loci containing multiple genes (such as CNVs), association evidence for a locus does not automatically imply that a single gene is driving the effect. For many recurrent CNVs, multiple genes with smaller individual effects appear to contribute to the overall risk (Supplementary Fig. 1)\textsuperscript{55}. The additive contribution of polygenic, common risk should also be considered\textsuperscript{31,34,46,55}.

For the majority of RGDs associated with neuropsychiatric domains, the evidence for statistical association comes from the observation of multiple de novo mutations at the same gene or locus in independent cases\textsuperscript{5,54,23,48–50}. Statistical association is calculated from three metrics: (1) the number of independent cases (affected individuals) with de novo mutations at the gene or locus; (2) the denominator, i.e., the total number of cases assessed to find the mutations; and (3) the mutation rate, i.e. the chance of observing a similar de novo mutation in an unaffected individual. Mutation rate is highly dependent on gene size, with larger genes having higher mutation rates (Fig. 3a)\textsuperscript{4} and therefore requiring more de novo mutations in cases to demonstrate association (Fig. 3b)\textsuperscript{10,66,67}. Defining these three metrics is straightforward in research cohorts\textsuperscript{4,53}; however, when considering multiple clinical reports\textsuperscript{5}, the denominator becomes all neuropsychiatric cases that have undergone genetic testing worldwide. Estimating this worldwide denominator at 100,000 at present, at least 5–10 confirmed and independent

de novo PTV mutations are required to achieve genome-wide association, depending on gene size (Fig. 3b, Extended Data Fig. 2).

Once a genetic locus meets the threshold for genome-wide association, other factors should be considered in selecting which RGDs to study, including the effect size on behavioral domains, level of risk for a categorical disease manifestation, population frequency, tractability for experimentation and therapeutic potential. Given the complexity of neurobiology, there are important roles for both detailed assessment of individual genes and consistent assessments across multiple genes.

Progressing from genotype to mechanisms

The second priority we identified is the design of bottom-up approaches in order to progress from genetic etiology to mechanism. Because RGDs are genetically defined, they provide an unparalleled opportunity to understand the pathophysiology of psychiatric domains and conditions. However, making the link between a genetic variant and the underlying mechanism (Fig. 4a) for a psychiatric domain or disorder presents a significant challenge\textsuperscript{56}. Conceptually, the problem can be divided into two parts: first, to link a genetic variant to a biochemical and/or biological consequence in a specific type of neuron or glial cell, and second, to connect a functional change in a set of cells to a defect in a circuit in a specific anatomic region of the brain at a particular time in development. Although linking these diverse levels seems daunting, the prospects for success are greatly improved by the ongoing development of experimental and conceptual tools, from wet-bench to bioinformatic approaches\textsuperscript{46,59}.

To pinpoint the subcellular location of a specific gene implicated in a disorder, it is possible to use genetic tagging with fluorescent proteins and imaging using ultra-resolution microscopy in human neurons created from stem cells in vitro. The function of the gene can be determined by using CRISPR-Cas9 genome engineering tools to knock out or overexpress the gene and observing the cellular consequences using a combination of high-content microscopy, calcium and voltage imaging, high-throughput electrophysiology and RNA sequencing. Although these tools have traditionally not been amenable to the study of multiple variants, the development of automated microscopes and patch clamps that are used for high-throughput drug screening is rapidly changing this. To establish the time in development and the anatomic location that is affected by a particular genetic variant, a set of gene atlases of the developing human brain have been generated by the Allen Institute and

---

**Box 2 | Identified priorities for advancing neuropsychiatric research through a focus on RGDs**

- **Priority 1:** Develop new strategies to consistently characterize the effects of RGDs on cognitive and behavioral traits that may result in disabilities and clinical referral, across the rapidly expanding spectrum of neuropsychiatric-associated RGDs routinely identified in the clinic.
- **Priority 2:** Prioritize RGDs for exploration of pathophysiology through bottom-up approaches: building experimental models starting from the lower levels of the mechanistic hierarchy, such as gene expression and synaptic function, will help guide investigation and explain phenomena higher up in the hierarchy, such as macroscopic brain and behavioral alterations leading to psychiatric diagnoses.
- **Priority 3:** Integrate the knowledge resulting from Priorities 1 and 2, above, to accelerate the development of new drugs and other therapeutic interventions in those with and without RGDs.
other groups that allow gene expression to be ascribed to specific types of cells in specific regions of the brain at particular times in development (http://portal.brain-map.org). Moreover, recent progress in recapitulating neural development in vitro using pluripotent stem cells offers a powerful platform to manipulate and investigate the roles of specific disease genes in human neurons and glial cells\(^6\). This recent work demonstrates the high reproducibility and validity of these systems for studying many human brain developmental processes.

By looking for cell types that are enriched in the expression of specific genes, it has been possible to determine that the phenotypes associated with particular RGDs are caused by dysfunction in a specific type of cell. For example, phenotypes associated with mutations in the gene \(\text{SCN1A}\) that cause Dravet syndrome are due to defects in fast-spiking interneurons\(^6\), and glutamatergic neurons in the cerebral cortex are implicated in ASD\(^5\). A similar strategy can be used to identify the developmental window during which a particular genetic variant leads to a disorder by narrowing down the expression of the gene to particular periods during development, as is the case for mutations in \(\text{FOXG1}\), which likely lead to neurodevelopmental disorders by affecting the differentiation of cortical glutamatergic neurons during brain development\(^5\). Similar approaches have been applied to structural and functional neuroimaging data, testing multiple brain regions and circuits to identify those that are distinct versus shared across disorders\(^6\), as well as neuroanatomic patterns that predict subsequent disease\(^6\). Advances in data sharing have enabled the analysis of large-scale, multisite cohorts\(^6\). Large-scale harmonized phenotypic data across RGDs could be analyzed in a similar manner and integrated with imaging and genomic data.

### Identifying the circuits

The development of new, effective treatments is hindered by our current very limited understanding of the circuit abnormalities that underlie different neuropsychiatric disorders. RGDs offer a unique roadmap to the identification of these circuits because distinct underlying biochemical defects have now been shown to manifest with common phenotypes. For example, genes with roles in glutamatergic receptors (such as \(\text{GRIN2B}\)), sodium channel activity (such as \(\text{SCN2A}\)), and chromatin remodeling (such as \(\text{CHD2}\)) can all manifest both in the behaviors of ASD and in seizures\(^5,22,62\). In all likelihood, these mechanistically distinct genetic causes of
disorders have common phenotypic manifestations through convergent effects on key circuits. The convergent effects of the mutations responsible for these circuit abnormalities therefore offer a unique set of probes to identify those circuits.

Although in vitro tools are essential for identifying the biochemical and cell biological consequences of genetic variants, identifying the neuronal circuits that lead to changes in behavior requires a more intact system. Well-defined and replicable genetic loci associated with RGDs provide a unique and indispensable opportunity to generate experimental animal models. These experimental models do not replicate the complexity of human neuropsychiatric disorders, but are essential as scientific tools to interrogate the effects of disorder-associated genes and variants on circuit properties in a number of ways. For example, RGDs that rank highest in the prioritization criteria (above) reveal defects associated with circuit-wide integration in RGDs.

Transplantation of human-derived neural cells into animals could generate experimental animal models. These experimental models may not resemble human disorders, the molecular, cellular and circuit phenotypes are more likely to be informative. Cross-species validation of animal models provides evidence for evolutionarily conserved electrophysiological signatures and cellular phenotypes that may enhance confidence in the relevance of the genetic perturbation (Fig. 4b–d)90–92. Regardless of the approach, replicable and rigorously collected data are beneficial, including replication of key findings in a second cohort.

Nonhuman primate (NHP) experimental systems, such as are used in studying depression46 and schizophrenia23,47, offer several advantages, including structural and functional similarity of the prefrontal cortex to humans, presence of similar cortical–subcortical circuits and complex behavior repertoires. These benefits come at the expense of lower throughput and substantially greater costs and ethical concerns. As with other models, validation of resulting findings in humans will be essential.

Alongside animal models, advances in cell reprogramming and differentiation, as well as CRISPR-Cas9 methods, enable the study of human neurons and glial cells in specific RGDs. Although behavioral phenotypes of mouse models may not resemble human disorders, the molecular, cellular and circuit phenotypes are more likely to be informative. Cross-species validation of animal models provides evidence for evolutionarily conserved electrophysiological signatures and cellular phenotypes that may enhance confidence in the relevance of the genetic perturbation (Fig. 4d–g)93–95. Regardless of the approach, replicable and rigorously collected data are beneficial, including replication of key findings in a second cohort.

Fig. 3 | Thresholds for genome-wide significant association. a, Mutation rates vary across genes based primarily on gene size (cDNA) but also on sequence (for example, GC content)294. Genes associated with neurodevelopmental delay, ASD and/or epilepsy to date tend to be large, with higher mutation rates132,52. b, Tens of thousands of individuals with neuropsychiatric disorders have been sequenced to date. To estimate the number of independent de novo protein-truncating variants (PTVs) across multiple case reports that are required for reliable association of a gene with a neuropsychiatric disorder, we used a Poisson distribution. Expected mutation rates in 50,000, 100,000 and 200,000 individuals are estimated from controls across the range of gene sizes (see Supplementary Methods and Extended Data Fig. 2)43,101. ASD, autism spectrum disorder; NDD, neurodevelopmental delay; GC, guanine-cytosine294. Credit: Debbie Maizels/Springer Nature.

Therapeutic development

Aside from some metabolic disorders90, very few RGDs currently have treatments guided by genetic findings. Recent advances in genetic therapy, including the landmark improvement of infant mortality and neurological function in patients with spinal muscular atrophy (SMA) treated with the antisense oligonucleotide nusinersen29 or Zolgensma gene replacement30, offer grounds for cautious optimism. However, potential therapies have yet to modify cognitive or behavioral symptoms in humans in fragile X syndrome91, tuberous sclerosis30 or Rett syndrome101, despite promising results in mice102–104. Gene dosage is important in many such disorders, with too much or too little causing symptoms11,105, presenting challenges in titrating therapy. Consequently, recessive loss-of-function disorders, such as SMA, may be easier to treat than dominant loss-of-function disorders. Delivery of a therapeutic to the brain presents an additional challenge, further complicated if the therapy needs to be targeted to a specific cell type or brain region.

The extent to which neuropsychiatric symptoms are modifiable in humans varies by domain, age and RGD. Although some domains can be modified years after onset (for example, seizures, psychosis, attention), others may require early treatment, as suggested
by the critical period of visual cortex development\textsuperscript{106} or the relationship between early management of phenylalanine hydroxylase deficiency (PKU) and adult IQ\textsuperscript{98}. With this in mind, therapies that have failed in adults and adolescents are being tested in 2–6-year-olds with fragile X syndrome\textsuperscript{107} and infants with tuberous sclerosis\textsuperscript{108,109}. Prenatal screening for many RGDs is feasible with current technology, and there is precedent for neonatal treatment for highly penetrant RGDs\textsuperscript{29}.

Clinical trials require predefined outcome measures. In neuropsychiatric disorders, this usually entails caregiver questionnaires regarding symptom severity and functioning. There is considerable need for more reliable, sensitive and objective measures. Measures of brain function that can be compared to model systems, such as EEG or functional MRI, should be included in trial protocols to help demonstrate target engagement by the therapy and refine dosing for future trials\textsuperscript{94}. Establishing multisite databases of individuals with RGDs, alongside genotype and longitudinal phenotype data, for example, by extracting knowledge from patient’s clinical notes\textsuperscript{110,111}, would reduce the costs of trials and help identify subgroups with the greatest potential to respond. Finally, clinical trials need to be of sufficient size to yield clear positive or null results\textsuperscript{99,112,113}.

**Phenotyping and harmonization**

To date, most phenotypic descriptions have focused on individual RGDs, ranging from case series to quantitative assessments of large patient registries. Such studies provide vital, clinically relevant insights that can have important therapeutic implications\textsuperscript{114,115}. At the same time, the overlap among neuropsychiatric domains, within and between RGDs, provides the opportunity for cross-domain and cross-disorder analyses\textsuperscript{44,116}, which may identify biologically defined subcategories of neuropsychiatric domains\textsuperscript{117}.
Achieving this vision will require a community-wide initiative to collate consistent, harmonized genetic data with quantitative measures of cognition and behavior across multiple RGDs and multiple neuropsychiatric domains. No single set of clinical data elements adequately describes all RGDs. The variables to collect in a natural history study should be broad and based on features of the disorder, including morbidities that are most important to patients, those that are most likely to be life limiting, potential prognostic characteristics and those that may help formulate sensitive clinical endpoints. The resulting data repository should reflect existing National Institutes of Health (NIH) data-sharing policies and standards, and become part of the NIH data sharing ecosystem to secure access for researchers, clinicians and institutions, enabling a wide range of hypothesis-testing with no restrictions on data use and appropriate human subjects and privacy safeguards. This effort will also require standardized phenotypic ontologies, which can be achieved by adapting existing ontologies to ensure that adequate and comprehensive terminology exists for neuropsychiatric symptom manifestations across the lifespan.

Other stakeholder actions

The challenges and opportunities presented above require coordinated efforts among the community of stakeholders to fund, execute, coordinate and communicate research (see Supplementary Table 5). Patient- and family-based organizations fulfill an essential role of providing support and decreasing isolation experienced by patients and caregivers; these organizations also have a key role in forming interactive networks that connect them with researchers and health care systems (for example, to keep families informed about the latest clinical trials and research findings). Although some examples of interactive research and family-based networks for RGDs already exist, they are not all focused on or inclusive of neuropsychiatric disorders, and efforts need to be made to use these networks for such purposes. Such networks will contribute not only to larger sample sizes, which are critical for research progress, but also to organizing healthcare systems to collect consistent phenotypes that include neuropsychiatric measures to enhance data harmonization, as well as implementing standards for genotyping and phenotyping. Academic health systems, in partnership with other health care systems, can play a pivotal role in these networks by expanding the reach of their initiatives and imparting knowledge gained from the research. Some registries and databases will use a federated approach, and some will utilize a centralized repository approach. Both will be useful for specific purposes, with federated approaches allowing for more flexibility in individualization and ‘real-time’ accrual of data, and centralized repositories providing more digested, pre-harmonized data that require less work by the user before analysis, but also restricts usage. Both types of networks will clearly be useful, and will continue to be expanded upon with respect to data types included, with electronic health records (EHRs) quickly becoming an additional data source.

Recently, clinical notes from patient EHRs have been identified as important sources of clinical and demographic information. This has led to the development of a variety of knowledge-extraction tools that transform the raw text content of clinical notes into structured associations between patients and phenotypes defined in terminologies such as the Systematized Nomenclature of Medicine—Clinical Terms and the Human Phenotype Ontology. Knowledge extracted from clinical notes is of increased value to research into RGDs given the scarcity of patient data, necessitating the use of natural language processing engines to extract structured knowledge from clinical notes. In addition, biobanks are increasingly being used to support organized collections of biological specimens, genomic data and associated clinical information—including EHR data—from broadly consented, diverse patient populations contributing to RGD research.

Partnerships among patient and family organizations, academic sites and other health systems will provide an important platform for working with funding agencies to emphasize the need for the implementation of core, cross-diagnostic phenotypic measures and the use of broad neuropsychiatric measures to facilitate data harmonization. This approach will culminate in early collaboration and coordination across groups working in this space, resulting in the design of maximally effective initiatives, data sharing and dissemination of information and knowledge gained. Furthermore, partnerships will create a platform for collaboration with pharmaceutical companies, encouraging alliances in preclinical and translational research toward drug development. However, data harmonization for collaboration necessarily requires data sharing. The community should follow NIH and established standards in the field for phenomic, genomic and biosample sharing. Although this can be done retrospectively, optimal sharing will begin before data collection, and include sharing of variables and even study methodology, to ensure replication and maximal generalizability of findings.

Drug development programs require a strong scientific foundation, including detailed understanding of the natural history of a disorder. Because of the small numbers of patients affected, and because clinical experience is dispersed among clinical referral centers, the natural history of RGDs is rarely well characterized. The US Food and Drug Administration advises sponsors to evaluate the depth and quality of existing natural history knowledge early in drug development. Close collaboration with patient and family groups and all stakeholders is essential to design and conduct impactful natural history studies for future potential drug design and approval.

Several examples (Supplementary Table 5) illustrate the potential contributions of such convergent efforts. The benefits of coordination across stakeholders are immense, and could create resources enabling investigators to merge and integrate available databases, examine existing data across disorders, and define the best data formats for prospective studies that use consistent broad and deep phenotyping to facilitate harmonization efforts. Furthermore, it fosters complementary approaches for building resources and establishing cohorts (for example, mining of electronic medical records, clinical recruitment, patient advocacy) with synergism that propels the field.

Conclusions

We think that RGDs hold important biological clues into the mechanisms underlying complex neuropsychiatric disorders, such as ASD and schizophrenia, that can be leveraged in model experimental systems, with the view to understanding the etiology underlying the full spectrum of these neuropsychiatric conditions. Although the individual features of each RGD are important, it is the commonalities across RGDs that hold the greatest promise to transform our understanding and management of neuropsychiatric disorders. Identifying these commonalities in phenotype and neurobiology will require coordination and sharing of methods, data and resources across individuals and institutions, following the lead of the NIH and genomics community. Although RGDs and the neuropsychiatric domains they impact pose substantial challenges, the opportunity these rare disorders offer for illuminating mechanisms is likely to be transformative for scientists, clinicians and, ultimately, patients.

Received: 7 July 2018; Accepted: 31 July 2019; Published online: 23 September 2019

References

1. US Food and Drug Administration. Orphan Drug Act. (1983).
2. Loane, M. et al. Twenty-year trends in the prevalence of Down syndrome and other trisomies in Europe: impact of maternal age and prenatal screening. *Eur. J. Hum. Genet.* 21, 27–33 (2013).
3. McKusick-Nathans Institute of Genetic Medicine, Online Mendelian Inheritance in Man, OMIM® (Johns Hopkins University, Baltimore, MD, USA) https://omim.org/ (accessed 28 April 2018).

Published online: 23 September 2019
4. McRae, I. F. et al. Prevalence and architecture of de novo mutations in developmental disorders. Nature 542, 433–438 (2017).
5. Sanders, S. J. et al. Insights into autism spectrum disorder genomic architecture and pathology from 71 risk loci. Neuron 87, 1215–1233 (2015).
6. Catterall, W. A., Kalume, F. & Oakley, J. C. NaV1.1 channels and epilepsy. J. Physiol. 588, 1849–1859 (2010).
7. Escayg, A. et al. Mutations of SCN1A, encoding a neuronal sodium channel, in two families with GEFS+4. Nat. Genet. 24, 345–349 (2000).
8. Marshall, C. R. et al. Contribution of copy number variants to schizophrenia from a genome-wide study of 41,331 subjects. Nature Genet. 49, 27–35 (2017).
9. Lukowski, A. F., Miloyevich, H. M. & Eales, L. Cognitive functioning in children with down syndrome: current knowledge and future directions. Adv. Child Dev. Behav. 56, 257–289 (2016).
10. Sanders, S. J. et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature 485, 237–241 (2012).
11. Ben-Shalom, R. et al. Opposing effects on NaV1.2 function underlie differences between SCN2A variants observed in individuals with autism spectrum disorder or infantile seizures. Biol. Psychiatry 82, 224–232 (2017).
12. Pinto, D. et al. Frequency of patients with schizophrenia, autism, bipolar disorder, depression, anorexia nervosa, or substance abuse vs their unaffected siblings. JAMA Psychiatry 70, 22–30 (2013).
13. Stefansson, H. et al. CNVs conferring risk of autism or schizophrenia affect cognition in brains. Nature 505, 361–366 (2014).
14. Sanders, S. J. et al. Whole genome sequencing in psychiatric disorders: the WGSPP consortium. Nat. Neurosci. 20, 1661–1668 (2017).
15. Reuter, M. S. et al. Diagnostic yield and novel candidate genes by exome sequencing in 152 consanguineous families with neurodevelopmental disorders. JAMA Psychiatry 74, 293–299 (2017).
16. Schaefer, G. B. et al. Array comparative genomic hybridization findings in a cohort referred for an autism evaluation. J. Child Neurol. 25, 1498–1503 (2010).
17. Retterer, K. et al. Clinical application of whole-exome sequencing across clinical indications. Genet. Med. 18, 696–704 (2016).
18. Sawyer, S. L. et al. Utility of whole-exome sequencing for those near the road to precision psychiatry. JAMA Psychiatry 73, 1063–1067 (2012).
19. Stessman, H. A., Bernier, R. & Eichler, E. E. A genotype-first approach to studying autism spectrum and related neurodevelopmental disorders. Neuron 73, 1045–1052 (2012).
Acknowledgements

This paper offers a synthesis of the ideas generated at the NIMH-sponsored workshop "Rare Genetic Disease Workshop: Window into Genomic Risk and Resilience of Mental Disorders," held in September 2017, with the goal of discussing research and clinical opportunities presented by recent discoveries of RGDs with high risk for developmental neuropsychiatric disorders. Analyses utilize data generated by the Saguenay Youth Study, OMIM (https://www.omim.org), ExAC (http://exac.broadinstitute.org/) and the DECIPHER Consortium, including the Developmental Disorders Genotype-Phenotype Database (DDG2P, https://decipher.sanger.ac.uk/info/ddg2p). A full list of centers that contributed to the generation of the DECIPHER data is available at http://decipher.sanger.ac.uk and via email from decipher@sanger.ac.uk. We also thank G. Senthil for helpful feedback on the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-019-0581-5.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence should be addressed to R.E.G. or C.E.B.

Peer review information Hannah Stower was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© Springer Nature America, Inc. 2019
**Extended Data Fig. 1 | Impact of RGDs on neuropsychiatric domains.**

**a**, Many RGDs impact cognition, measured by IQ. For CNVs, the decrease in IQ (x axis) can be predicted by considering the pLI score of the genes within the CNV. CNVs that are predicted to markedly reduce IQ are more likely to be de novo (y axis), based on logistic regression (blue line) of 2,743 CNVs detected in patients with neurodevelopmental disorders and the general population (gray distributions at top and bottom). Updated analysis from ref. 54.

**b**, In Fig. 2, we show the odds ratio for ID/NDD, ASD and SCZ across different CNV loci. Here, we show an equivalent plot for single-gene RGDs. Insufficient control data exist to estimate odds ratio, and therefore we show the percentage of cases with ID/NDD, ASD, and IE based on curated publication review applied equally across genes (https://dbd.geisingeradmi.org) with the number of cases are shown in parentheses (see Supplementary Table 2 for numbers). Abbreviations: ID, intellectual disability; NDD, neurodevelopmental delay; ASD, autism spectrum disorder; SCZ, schizophrenia; IE, infantile epilepsy; pLI, probability loss-of-function intolerant.
Extended Data Fig. 2 | Thresholds for genome-wide significant association with de novo PTVs. a. Gene mutability is a function of gene length (cDNA) and sequence context (particularly GC content). b. RGD gene discovery from exome sequencing has been driven by de novo mutations, leading to a bias towards larger genes with higher mutability. c. Thresholds of statistical association (colored lines) are estimated for a given number of de novo PTV mutations (3, 5, 10, and 20) as cohort size (x axis) and gene mutability/size (y axis) varies. P values are estimated based on the rate of de novo PTV mutations in controls and a Poisson distribution (see Methods for details). Abbreviations: pLI, probability of loss-of-function intolerance; ASD, autism spectrum disorder; DDD: Deciphering Developmental Disorders; GC content, guanine-cytosine content.