Interpreting the role of *de novo* protein-coding mutations in neuropsychiatric disease

Jacob Gratten[1,4], Peter M Visscher[1,2], Bryan J Mowry[1,3], Naomi R Wray[1]

[1] The University of Queensland, Queensland Brain Institute, Brisbane, Australia

[2] The University of Queensland Diamantina Institute, Brisbane, Australia

[3] Queensland Centre for Mental Health Research, Wacol, Australia

[4] Corresponding author: j.gratten1@uq.edu.au

Abstract

Pedigree, linkage and association studies are consistent with heritable variation for complex disease, due to the segregation of genetic factors in families and in the population. In contrast, *de novo* mutations make only minor contributions to heritability estimates for complex traits. Nonetheless, some *de novo* variants are known to be important in disease etiology. Identification of risk-conferring *de novo* variants will contribute to discovery of etiologically relevant genes and pathways and may help in genetic counseling. There is presently considerable interest in the role of such mutations in complex neuropsychiatric disease, largely driven by new whole-genome genotyping and sequencing technologies. An important role for large *de novo* copy number variations has been established. Recently, whole-exome sequencing has been used to extend the investigation of *de novo* variation to point mutations in protein-coding regions. Here, we consider several challenges for the interpretation of such mutations in the context of neuropsychiatric disease.

Whole-exome studies of *de novo* mutations in autism, mental retardation and schizophrenia

Nine whole-exome studies have been published to date for neuropsychiatric disorders (see Veltman & Brunner[1] for an overview), all focusing on families with sporadic cases of disease (i.e. those with a negative family history). Autism spectrum disorders (ASD) have received most attention, with five independent studies[2-6], one study has been published for mental retardation[7] and another three for schizophrenia[8-10]. A tenth study reporting whole genome sequencing of both ASD and schizophrenia families was published while this paper was in review[11]. These studies follow others that focused on *de novo* mutations in single or multiple candidate genes (e.g. [12]), which we do not consider here. The largest studies to date, for ASD, have already yielded several candidate genes, highlighting the promise of "*de novo*" experimental designs for gene discovery in neuropsychiatric disease. However, direct comparisons between exome studies are complicated by differences in

experimental design (i.e. the use of matched family controls, population controls or a theoretical control), definition of the targeted exome, sequencing methodology and other features, such as whether insertions and deletions were considered along with point mutations.

The strongest evidence for an involvement of *de novo* mutations exists for ASD, based on the studies by Iossifov *et al.*[2] and Sanders *et al.*[6], which are the largest to date and the only studies to use a comprehensive sample of family controls (in the form of phenotypically discordant siblings). These studies, which along with O'Roak *et al.*[5] sequenced non-overlapping sub-sets of the Simons Simplex Collection (SSC), showed an enrichment of *de novo* gene-disrupting mutations (i.e. nonsense, splice, frameshift) in probands compared to their unaffected siblings[2], particularly for those in brain-expressed genes[6]. O'Roak *et al.*[5] used a similar approach but did not identify an association, perhaps due to the small size of their control sample (N=50). Xu *et al.*'s[9,10] schizophrenia studies are the only others to incorporate controls, in the form of healthy parent-offspring trios. They found no difference in the *de novo* mutation rate in cases versus controls, but there was a significant enrichment of putatively functional mutations in cases. Other studies, in lieu of control samples, have relied on comparisons to the theoretical expectation based on previously reported mutation rates[3,4,7,8], to test for an enrichment of *de novo* mutation in disease. Inferences from these studies, although supportive of a role for *de novo* mutations, are inconsistent, most likely reflecting sampling effects and methodological differences between studies. This highlights the value of matched control samples, and in particular family controls, in studies of *de novo* mutation.

Two studies have demonstrated that protein products of genes harboring highly disruptive *de novo* mutations, together with previously identified ASD genes, form significant (but distinct) protein-protein interaction networks[3,5], providing further evidence for a role of *de novo* mutation in ASD. Iossifov *et al.*[2] also showed that genes harboring disruptive mutations in ASD probands were strongly over-represented in the set of genes known to interact with the fragile X mental retardation gene product FMRP. All studies are consistent with the existence of a large number of genes for these disorders, and for ASD and schizophrenia, there is evidence for a correlation between paternal age and the number of observed *de novo* mutations per child[2,4-6,9,11]. This observation, together with evidence for a paternal bias in the origin of *de novo* mutations[2,3,5,11], is consistent with an accumulation of mutations in the paternal germline[13]. Collectively, these findings represent solid evidence for a contribution from *de novo* protein-coding mutations in the etiology of neuropsychiatric disease. However, studies published to date differ widely regarding the inferred penetrance of identified mutations and the overall contribution of *de novo* variation to disease. The

different conclusions reflect differences of interpretation, although true differences may exist between disorders. Incorrect inference regarding the pathogenicity of *de novo* mutations can have serious consequences, both for genetic counseling[14] and in terms of misplaced investment in candidate genes that fail to replicate, a recurring problem in neuropsychiatric genetics. In this Perspective we focus on several key challenges for the interpretation of *de novo* designs, including the differentiation of risk-conferring mutations (particularly missense mutations) from those that are benign, the estimation of penetrance and the quantification of the relative importance of *de novo* and inherited variation.

Identifying *de novo* protein-coding mutations conferring risk of disease

A major challenge for the interpretation of *de novo* mutations in complex disease is that high levels of functional genetic variation have been shown to exist in the exomes of healthy individuals[15-19]. A typical exome harbors ~100 gene-disrupting variants, including on average 2.8 severe recessive disease alleles in the heterozygous state[20], less deleterious alleles contributing to individual phenotype, and benign variation in redundant genes[21]. In addition, recent studies estimate that each person harbors ~1 *de novo* protein-altering mutation, both in cases and controls[11]. The standard deviation of this number across individuals is also ~1, so that finding two or even three *de novo* mutations in a single genome can occur by chance[6]. Furthermore, if there are up to 1000 genes that can contribute to the risk of, say, ASD[6], then observing a *de novo* mutation in one of those genes by chance is ~5%, assuming a total of 20,000 genes. Consequently, the simple presence of a provisionally functional *de novo* mutation in a proband, even one predicted to have severe effects, is not sufficient evidence that it contributes to risk[6], let alone that it is sufficient to cause disease. Claims to this effect[4,7,8,10] should be viewed with caution.

Multiple observations, either of the same mutation or of independent mutations in the same gene, are required to identify risk alleles or loci. The former is unlikely in exome studies for point mutations in complex disorders, whereas the latter is not unexpected. In theory, per gene counts of functional *de novo* mutations in cases and controls should provide a means of identifying risk genes. In practice, this approach has little statistical power because so many genes contribute to neuropsychiatric disorders that even the largest studies to date are yet to identify a gene harboring more than a few *de novo* mutations (excluding known regions of recurrent deletions/duplications that arise due to non-allelic homologous recombination[22,23]). Two alternative approaches to the interpretation of recurrent *de novo* mutations in single genes have been proposed. O'Roak *et al*.[5] compared the occurrence of multiple *de novo* events in *CHD8*, *GRIN2B*, *LAMC3*, *NTNG1* and

SCN1A to the expected number based on each gene's locus-specific mutation rate. Although all five genes were inferred to have a nominally significant excess of *de novo* events, only the excess in *NTNG1* would survive correction for testing of 14,363 brain-expressed genes. This represents a promising approach to gene discovery, although it is important to be aware that gene-specific mutation rates are estimated with error (for instance, due to uncertainty in the human-chimpanzee divergence time used in the calculation).

Evidence

Sanders *et al*.[6] proposed an alternative approach to the identification of genuine risk mutations, based on the empirical probability, across all brain-expressed genes, of observing multiple independent *de novo* point mutations in the same gene in unrelated individuals. Using simulations that accounted for sample size, gene size and GC content they demonstrated that two or more gene-disrupting mutations were unlikely to be observed by chance in the same brain-expressed gene in unrelated probands in their study, due to the very low rate of *de novo* nonsense and splice site mutations. A single gene (*SCN2A*) satisfied this threshold in their study[6], but five others (*CHD8, DYRK1A*, *GRIN2B*, *KATNAL2*, *POGZ*) have been identified by combining data across the four largest ASD studies[2,3,5,6]. A seventh gene (*CUL3*) has since been identified by the addition of whole genome data from 44 Icelandic ASD trios[11]. These are landmark findings in ASD genetics, not only because GWAS methodology in larger cohorts has not met with comparable success[24] (although we note that ASD GWAS samples sizes have been smaller than for other neuropsychiatric disorders, such as schizophrenia and bipolar disorder, for which GWAS has been successful[25]), but because the identified mutations, with obvious functional effects in single genes, provide valuable insights into disease biology.

An important consideration for future efforts to identify risk genes using this approach is that the probability of a gene harboring multiple gene-disrupting *de novo* mutations by chance increases with sample size. Simulations from Sanders *et al*.[6] suggest that the probability of this occurring exceeds 5% in samples of greater than ~1,400 families. However, as the *de novo* mutation rate is higher in probands than siblings, the false discovery rate (FDR) for the occurrence of two gene-disrupting *de novo* mutations in probands is acceptable for sample sizes in excess of 3000 families[6]. Clearly, the FDR will be lower still for the presence of three or more gene-disrupting mutations in the same brain-expressed gene. It is noteworthy, therefore, that the estimated frequency of such mutations in probands in the ASD candidate genes *CHD8* and *SCN2A* was 0.33% (5 mutations in ~1,500 cases) and 0.2% (3 mutations in ~1,500 cases), respectively[3]. This implies that as sample

sizes grow from hundreds to thousands of families, candidate genes will be identified by the presence of ≥3 *de novo* gene-disrupting mutations in probands.

We note that the simulations of Sanders *et al.*[6], upon which these inferences are based, make a number of assumptions that may not hold for other ASD cohorts, or for other neuropsychiatric disorders. For example, a key simulation parameter estimated from their data is the *de novo* mutation rate in probands. Similar studies of equivalent magnitude in schizophrenia and other disorders will be needed to confirm whether the approach will be more broadly successful. However, on current evidence in ASD, family-based exome sequencing targeting *de novo* gene-disrupting mutations is a promising new paradigm for gene discovery in complex neuropsychiatric disease[6]. The projected yield of ASD risk genes that will be identified in the entire SSC (~2650 ASD families) is large, ranging from ~25-50[6] to >100[2], depending on the total number of ASD risk genes and the penetrance of individual mutations.

The interpretation of *de novo* missense mutations in disease

In contrast to the strong signal from gene-disrupting mutations in ASD, evidence for the involvement of *de novo* missense mutations is inconsistent. Sanders *et al.*[6] reported an enrichment of *de novo* missense mutations (both in isolation and in combination with gene-disrupting mutations) in probands compared to unaffected siblings, particularly in brain-expressed genes, whereas Iossifov *et al.*[2] did not. Sanders *et al.*[6] found no evidence that the severity of missense mutations, as inferred from the degree of evolutionary conservation (e.g. GERP) and other functional prediction methods (e.g. PolyPhen-2), either singly or in combination, was informative with respect to risk. O'Roak *et al.*[5] on the other hand assumed that missense mutations in highly conserved positions were equivalent to gene-disrupting mutations. In the schizophrenia study by Xu *et al.*[10], the strongest evidence for involvement of *de novo* mutations comes from a large excess of missense mutations in cases, a result that would not be expected if this class of variation did not play a role in disease. The role of *de novo* missense mutations in ASD, and more broadly in neuropsychiatric disease, is consequently unclear. Larger studies will be needed to resolve this question because those published to date are small and have limited power to detect (or conversely rule out) realistic contributions from *de novo* missense mutations[2,6]. However, given that a large proportion of *de novo* missense mutations are predicted to be mildly deleterious[26], and that there is evidence for weak purifying selection on coding sequences in human populations[16,18,19], we anticipate that larger family-based exome datasets in neuropsychiatric disorders will converge on a significant role for *de novo* missense mutations.

A number of genes have been found to harbor two *de novo* missense mutations in unrelated ASD probands, but this is not sufficient evidence to implicate them in disease. The probability of observing multiple independent mutations by chance is higher for missense mutations, relative to gene-disrupting mutations, because the mutation rate is ~20-fold higher[6]. As a consequence, the FDR for the presence of multiple missense mutations is strongly dependent on the sample size and underlying genetic model. For studies published to date, involving ~200-300 families, three independent missense mutations are required to implicate a gene in disease, and four are needed as sample size increases, depending on the total number of risk genes. That no gene has been found to harbor this number of *de novo* missense mutations may too be a consequence of inadequate sample size, since the effect size of missense mutations is predicted to be modest. A salient point, given the primacy of the mutation rate in the interpretation of recurrent *de novo* mutations, is that the assumption of a single mutation rate for missense mutations overlooks substantial fine-scale and context-dependent variation in the human mutation rate. The best-known example is the 10 to 20-fold higher rate of C>T and G>A transition mutation at CpG dinucleotides[11], which occurs because cytosines in CpGs are frequently methylated, and methyl-cytosines are prone to undergo spontaneous deanimation to thymine[27]. Other examples of mutation rate variation have also been described[28,29]. Given that this variation is known to exist, more sophisticated analyses that differentiate missense (and for that matter gene-disrupting) mutations into different classes (e.g. mutations in CpG verses non-CpG dinucleotides[11]) and that interpret recurrences in the context of the expected mutation rate for that sub-set of mutations may be insightful. Finally, gene-based testing in case-control exome studies may also be an efficient means of identifying risk genes on the basis of segregating missense variation[30,31]. Such studies benefit from a two-fold lower sequencing cost in comparison to family-based studies focused on *de novo* mutations in probands and unaffected siblings, although preliminary studies indicate that very large sample sizes will be required for success[30].

Estimating the penetrance of *de novo* protein-coding mutations

The distribution of effect sizes for *de novo* protein-coding mutations in neuropsychiatric disorders is yet to be determined. The expectation is that these variants will exhibit moderate to large effect size, since this is the case for identified recurrent copy number variations (CNVs)[32]. Several of the studies mentioned in this article favor models involving highly penetrant mutations. For example Vissers *et al*.[7] proposed that *de novo* point mutations of large effect, in combination with *de novo* CNVs "could explain the majority of all mental retardation cases in the population", and Xu *et al*.[10]

stated that *de novo* mutations, including newly arisen CNVs "account for more than half of the sporadic cases of schizophrenia". In ASD Iossifov *et al.*[2] support a model in which a third or more of all sporadic cases harbor causal *de novo* mutations[33]. These statements imply that single *de novo* mutations may be sufficient to cause disease, consistent with the model of extreme genetic heterogeneity for complex disease advocated by some[34]. While the existence of some fully penetrant mutations is a possibility, particularly for mental retardation (e.g. [35,36]), ASD (e.g. [37]) and other disorders with documented monogenic forms (schizophrenia and bipolar disorder being prominent exceptions[25]), the suggestion that a large proportion of cases for any common neuropsychiatric disorder are due to highly penetrant *de novo* events is not consistent with empirical evidence of the recurrence risk (RR) to relatives[38,39] (a point we discuss in more detail in the next section and Supplementary Note). More specifically, there is currently no way to determine which, if any, *de novo* protein-coding mutations are sufficient to cause disease, because no single event has been observed more than once. This is in contrast to documented recurrent CNVs, many of which have been observed in multiple cases and controls.

The currently available data in ASD are sufficient for effect size estimates of broad classes of mutations (e.g. gene-disrupting), and are consistent with more modest average penetrance. Sanders *et al*.[6] used data from matched family controls to demonstrate that the odds ratio (OR) for gene-disrupting mutations relative to silent mutations in brain-expressed genes in ASD cases compared to unaffected siblings was 5.65 (95% CI 1.44-22.20), similar to that of documented multigenic CNVs[40]. Estimates for *de novo* missense mutations were lower (OR=2.06, 95% CI 1.10-3.85), but nonetheless greater than for identified common variants[24]. These estimates are likely to represent a mix of risk and benign mutations, and thus individual risk mutations may have larger effects[6]. Nonetheless, they imply that the majority of individual *de novo* mutations in ASD are insufficient to cause disease and therefore must combine with other risk factors (including inherited variation and environmental factors) to generate a phenotype. This is a crucial point that in some studies seems under-appreciated[2,7,10,33], with important implications for the clinical application of exome sequencing (e.g. [41]). The field awaits larger exome studies, including studies in multigenerational pedigrees, which should help to refine the distribution of effect sizes for *de novo* protein-coding mutations in ASD and other neuropsychiatric disorders.

Quantifying the overall contribution of *de novo* protein-coding mutations to disease liability

It is now clear that neuropsychiatric disorders are underpinned by a large number of genes[42,43] and that affected individuals harbor multiple risk factors, but the relative importance of *de novo* and

inherited variation is yet to be established. The exome studies that are the focus of this article, although unanimous in establishing a role for *de novo* mutations in neuropsychiatric disorders, differ widely in their conclusions regarding the magnitude of this contribution. As noted above with respect to effect size, three studies in particular conclude that the contribution is very substantial and that *de novo* mutations (including CNVs) may account for a large proportion of all cases of ASD[2], mental retardation[7] and schizophrenia[10]. Other studies have proposed an oligogenic (or "multi-hit") model in ASD[4,5], whereby single *de novo* mutations combine with one or a few deleterious rare inherited variants to confer disease. In contrast, Neale *et al*.[3] and Sanders *et al*.[6] favor a more limited contribution, comprising mutations with effect size similar to those of known CNVs in no more than 10-20% of cases. Clearly, more data will be required to resolve which of these models (if any) is correct, but the suggestion that *de novo* mutations account for a large proportion of the total liability for these disorders is hard to reconcile with heritability estimates that are consistently high (i.e. ~80% for ASD[38] and schizophrenia[44]). A majority of *de novo* mutations are unique to a single individual, and thus will not contribute to the observed recurrence risk (RR) to relatives or to the estimate of heritability. Exceptions include *de novo* mutations originating in the parental germline that are shared by monozygotic (MZ) twins, or by siblings (including dizygotic [DZ] twins) in the case of germline mosaicism. If, as is suspected, mosaicism is relatively infrequent (e.g. Iossifov *et al*.[2] reported the same mutation in ~1 in 50 siblings), then the effect of *de novo* variants on the estimate of heritability will be primarily through their contribution to RR of MZ twin pairs, and only if heritability estimates from pedigree data include MZ twins. Evidence from mutation-accumulation experiments, data on RR to relatives and theory suggest that this contribution is likely to be small, of the order of ~1% (see Supplementary Note and Supplementary Fig. 1). The primary contribution of *de novo* mutations to variance of disease risk in the population will usually be partitioned into the non-heritable component, and because heritability estimates for neuropsychiatric disorders are so high (~80%), there is a ceiling on the total contribution from *de novo* mutations (see Supplementary Note).

Conclusions

Recent family-based exome studies targeting *de novo* mutations in autism[2-6], mental retardation[7] and schizophrenia[8-10] are the likely forerunners of many similar efforts across the breadth of neuropsychiatric disease and complex disease in general. Studies considered in this article support a significant role for *de novo* protein-coding mutations in neuropsychiatric disease, particularly in ASD, which has been the focus of the largest studies to date. The approach established by Sanders *et al*.[6], whereby *de novo* mutations are interpreted in terms of the empirical probability of observing

recurrent mutations in the same gene, has identified a number of promising candidate genes for ASD. These discoveries are particularly important because, in contrast to GWAS and the majority of CNV findings, they involve readily interpretable mutations in single genes, and thus provide immediate biological insight into disease mechanisms. Although we are mindful that equivalent success is yet to be demonstrated for other disorders, and that the method is dependent on assumptions that may not always be satisfied, it is clear that well-powered family-based exome sequencing studies targeting *de novo* mutations represent a promising new paradigm for gene discovery in neuropsychiatric disease. We anticipate that larger studies of this type will make an important contribution to progress in the field.

The emerging empirical evidence on the genetic architecture of neuropsychiatric disease (and that of other complex diseases) from whole-exome sequencing studies (including those reviewed here), CNV analysis and GWAS is one of multiple genetic factors and environmental factors that jointly increase risk of disease. The genetic risk factors include *de novo* CNVs, point mutations and indels, rare inherited mutations and CNVs and both common and rare inherited polymorphisms. This spectrum of variation likely spans both protein-coding and functional non-coding regions. The latter are poorly represented in published exome studies but this is likely to change as the field moves towards whole genome sequencing and findings from the ENCODE project enable detailed annotation of functional non-coding elements[45]. The key challenges for the field are to clarify and quantify the relative importance of the different classes of variation, something that may differ between disorders, and to understand how they act together to cause disease. Very large studies, including family-based exome and/or whole genome sequencing, as well as complementary case-control approaches, will be required to address these questions[31]. Studies that leverage transcriptome-wide expression data to provide insight into the functional consequences of *de novo* mutations, as has been reported for *de novo* CNVs in ASD[46], are also likely to contribute.

In this Perspective article we have addressed a number of challenges for the interpretation of *de novo* mutations in exome sequencing studies of neuropsychiatric disease. A clear message from studies to date is that the number and predicted severity of *de novo* mutations identified in a single individual is insufficient evidence for causality[6]. The burden of proof for causality for a single *de novo* mutation must be set high because incorrect inference can have serious consequences for individuals and their families[14], as well as doing the field a disservice due to misplaced investment in candidate genes that fail to replicate. This burden of proof ('beyond reasonable doubt') is likely to be case specific, and may include knowledge of recurrent gene-disrupting *de novo* mutations in

the same gene in unrelated individuals with the same disorder, together with functional evidence that is consistent with both a direct effect of the mutation[46] and the observed clinical phenotype. The curation of *de novo* protein-coding mutations and their associated phenotypes in databases, as has been instigated for CNVs[14,47], will be essential to identify the genotype-phenotype correlations needed for robust clinical interpretation of *de novo* mutations in exome sequencing studies.[48]

ACKNOWLEDGEMENTS

We acknowledge funding from the Australian Research Council to NRW (FT0991360), the Australian National Health and Medical Research Council to PMV (grants 613672, 613601, 1011506), BJM (grants 631406, 631671), and NRW (grant 613602, 613608), and from Queensland Health to BJM.

AUTHOR CONTRIBUTIONS

All authors conceived and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

REFERENCES

1. Veltman, J.A. & Brunner, H.G. De novo mutations in human genetic disease. *Nat Rev Genet* **13**, 565-75 (2012).
2. Iossifov, I. *et al.* De novo gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285-99 (2012).
3. Neale, B.M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242-5 (2012).
4. O'Roak, B.J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nature Genetics* **43**, 585-589 (2011).
5. O'Roak, B.J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246-50 (2012).
6. Sanders, S.J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-41 (2012).
7. Vissers, L.E.L.M. *et al.* A de novo paradigm for mental retardation. *Nature Genetics* **42**, 1109-1112 (2010).
8. Girard, S.L. *et al.* Increased exonic de novo mutation rate in individuals with schizophrenia. *Nature Genetics* **43**, 860-U65 (2011).
9. Xu, B. *et al.* De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nat Genet* **44**, 1365-9 (2012).
10. Xu, B. *et al.* Exome sequencing supports a de novo mutational paradigm for schizophrenia. *Nature Genetics* **43**, 864-U72 (2011).
11. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471-5 (2012).
12. Awadalla, P. *et al.* Direct measure of the de novo mutation rate in autism and schizophrenia cohorts. *Am J Hum Genet* **87**, 316-324 (2010).
13. Goriely, A. & Wilkie, A.O. Paternal age effect mutations and selfish spermatogonial selection: causes and consequences for human disease. *Am J Hum Genet* **90**, 175-200 (2012).
14. Vermeesch, J.R., Balikova, I., Schrander-Stumpel, C., Fryns, J.P. & Devriendt, K. The causality of de novo copy number variants is overestimated. *Eur J Hum Genet* **19**, 1112-3 (2011).
15. Bamshad, M.J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* **12**, 745-55 (2011).
16. Durbin, R.M. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073 (2010).
17. Li, Y. *et al.* Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat Genet* **42**, 969-72 (2010).
18. Nelson, M.R. *et al.* An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337**, 100-4 (2012).
19. Tennessen, J.A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64-9 (2012).
20. Bell, C.J. *et al.* Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Sci Transl Med* **3**, 65ra4 (2011).
21. MacArthur, D.G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823-8 (2012).
22. Dharmadhikari, A.V. *et al.* Small rare recurrent deletions and reciprocal duplications in 2q21.1, including brain-specific ARHGEF4 and GPR148. *Hum Mol Genet* **21**, 3345-55 (2012).

23. Vacic, V. *et al.* Duplications of the neuropeptide receptor gene VIPR2 confer significant risk for schizophrenia. *Nature* **471**, 499-503 (2011).
24. Wang, K. *et al.* Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature* **459**, 528-533 (2009).
25. Sullivan, P.F., Daly, M.J. & O'Donovan, M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nat Rev Genet* **13**, 537-51 (2012).
26. Kryukov, G.V., Pennacchio, L.A. & Sunyaev, S.R. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am J Hum Genet* **80**, 727-739 (2007).
27. Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. *Nat Rev Genet* **12**, 756-66 (2011).
28. Hodgkinson, A., Ladoukakis, E. & Eyre-Walker, A. Cryptic variation in the human mutation rate. *PLoS Biol* **7**, e1000027 (2009).
29. Green, P., Ewing, B., Miller, W., Thomas, P.J. & Green, E.D. Transcription-associated mutational asymmetry in mammalian evolution. *Nat Genet* **33**, 514-7 (2003).
30. Kiezun, A. *et al.* Exome sequencing and the genetic basis of complex traits. *Nat Genet* **44**, 623-30 (2012).
31. Need, A.C. *et al.* Exome sequencing followed by large-scale genotyping suggests a limited role for moderately rare risk factors of strong effect in schizophrenia. *Am J Hum Genet* **91**, 303-12 (2012).
32. Vassos, E. *et al.* Penetrance for copy number variants associated with schizophrenia. *Hum Mol Genet* **19**, 3477-3481 (2010).
33. Zhao, X. *et al.* A unified genetic theory for sporadic and inherited autism. *Proc Natl Acad Sci U S A* **104**, 12831-6 (2007).
34. McClellan, J.M., Susser, E. & King, M.-C. Schizophrenia: a common disease caused by multiple rare alleles. *Br J Psychiatry* **190**, 194-199 (2007).
35. Krawitz, P.M. *et al.* Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. *Nat Genet* **42**, 827-9 (2010).
36. Najmabadi, H. *et al.* Deep sequencing reveals 50 novel genes for recessive cognitive disorders. *Nature* **478**, 57-63 (2011).
37. Amir, R.E. *et al.* Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat Genet* **23**, 185-8 (1999).
38. Lichtenstein, P., Carlstrom, E., Rastam, M., Gillberg, C. & Anckarsater, H. The genetics of autism spectrum disorders and related neuropsychiatric disorders in childhood. *Am J Psychiatry* **167**, 1357-63 (2010).
39. Lichtenstein, P. *et al.* Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet* **373**, 234-239 (2009).
40. Sanders, S.J. *et al.* Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70**, 863-85 (2011).
41. Need, A.C. *et al.* Clinical application of exome sequencing in undiagnosed genetic conditions. *J Med Genet* **49**, 353-61 (2012).
42. Kim, Y.J., Zerwas, S., Trace, S.E. & Sullivan, P.F. Schizophrenia Genetics: Where Next? *Schizophrenia Bulletin* **37**, 456-463 (2011).
43. State, M.W. & Levitt, P. The conundrums of understanding genetic risks for autism spectrum disorders. *Nat Neurosci* **14**, 1499-506 (2011).
44. Sullivan, P.F., Kendler, K.S. & Neale, M.C. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Arch Gen Psychiatry* **60**, 1187-1192 (2003).

45. Bernstein, B.E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
46. Luo, R. *et al.* Genome-wide transcriptome profiling reveals the functional impact of rare de novo and recurrent CNVs in autism spectrum disorders. *Am J Hum Genet* **91**, 38-55 (2012).
47. Firth, H.V. *et al.* DECIPHER: Database of chromosomal imbalance and phenotype in humans using Ensembl resources. *Am J Hum Genet* **84**, 524-33 (2009).
48. Klassen, T. *et al.* Exome sequencing of ion channel genes reveals complex profiles confounding personal risk assessment in epilepsy. *Cell* **145**, 1036-1048 (2011).

SUPPLEMENTARY NOTE

Quantifying the overall contribution of *de novo* protein-coding mutations to disease liability

The contribution of *de novo* mutations to heritability estimates has received little attention. There are several lines of evidence to suggest that the contribution will be small. Firstly, from mutation accumulation experiments on complex traits across a range of model organisms, the contribution of new mutations to heritability has been estimated to be in the range of 0.001 and 0.01 per generation[1]. Secondly, a multiple variant risk model which is linear and additive on the scale of the logarithm of risk (i.e., multiplicative on the scale of the probability of disease, e.g. [2,3]) predicts that log(RR) is linear in the degree of relationship and that log(RR) for MZ twins is twice that of full siblings (including DZ twins). For schizophrenia, the empirical data largely support these predictions[4,5] (Supplementary Fig. 1), and heritability estimates from relatives other than MZ twins do not differ substantially from those that include MZ twins[6]. For autism, data on recurrence risks in second order (and higher) relatives is lacking, but in the largest available twin studies[7] the log(RR) of MZ twins is slightly less than twice that of DZ twins (e.g. 4.58 and 3.52, respectively[8]). We recognize that these genetic models are necessarily based upon assumptions and other genetic models can also fit the data. In addition, the number of MZ pairs in studies of disease with prevalence of ~1% tends to be low so that the sampling variance of the estimate of RR is high. Thirdly, we have performed theoretical calculations following Kemper *et al.*[9] to assess the likely contribution of *de novo* point mutations (V_m) to the total liability (V_p) under assumptions about the number of contributing genes, the *de novo* mutation rate, the mean number of target sites for mutation per gene and the mean effect size. If we assume that there are 1000 genes each with 500 target sites for major (i.e. gene-disrupting) mutation and a rate for these mutations of ~0.05×10^{-8} per site per generation (following Sanders *et al.*[10]), then $V_m = 2 * 1000 * 500 * 0.05 \times 10^{-8} * a^2 = 5 \times 10^{-4} * a^2$, with a the effect size. If we then assume that *de novo* gene-disrupting mutations have effect size similar to known *de novo* CNVs[10] (e.g. relative risk of ~20), then log(RR) = a ~ 3, and V_m ~0.005. Based on these assumptions, and acknowledging that *de novo* germline mutations are shared by MZ but not DZ twins, the contribution of *de novo* variants with large effects to the estimate of heritability is small but not trivial at 2 * 0.005 = 0.01 (i.e. twice the difference between MZ and DZ similarity). These calculations are for gene-disrupting mutations, and it is important to recognize that other types of *de novo* mutation, including CNVs and missense mutations, will also contribute to V_m. Missense mutations have a higher *de novo* rate (~1×10^{-8}) but the mean effect size is smaller than for gene-disrupting mutations[10] and so their contribution to the heritability is likely to be similar. Precise estimates are problematic because we do not know the total number of target sites for mutation (we have assumed 500,000 in the genome) or the distribution of effect sizes of new mutations. In spite

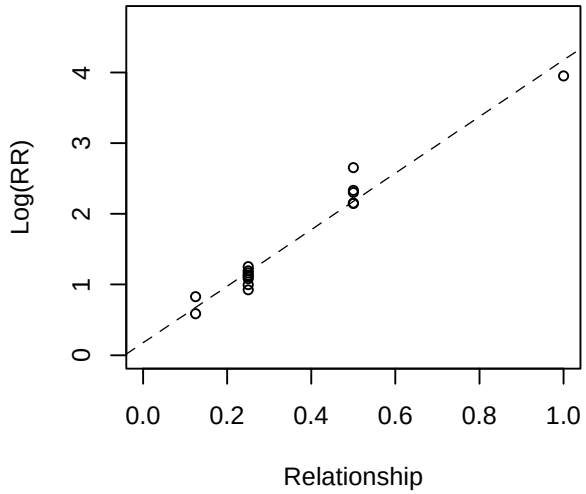
of this, our theoretical estimate closely matches those of Neale *et al.*[11] (i.e. 1.0 - 4.6%). When taken together with the empirical estimates from mutation accumulation experiments and the observations on recurrence risk to relatives, this suggests that the contribution of *de novo* mutations to estimates of heritability is likely to be small.

Quantifying the contribution of *de novo* mutations to variation in liability, as described above, contrasts sharply with measures based on the difference in mean prevalence, which are favored by some recent studies and which imply much larger contributions (e.g. [12-15]). In our opinion, there is confusion in the literature about the definition and interpretation of these measures, which yield statements such as 'the proportion of cases caused by genetic factor X'. For example, in a study of CNVs in children with intellectual disability (ID) and/or developmental delay (DD), Cooper *et al.*[12] state "We estimate that ~14.2% of disease in these children is caused by CNVs >400 Kb". This bold statement is derived from the fact that ~25.7% of 15,767 children with ID and/or DD harbor a CNV of at least this size (and of population frequency <1%), compared to 11.5% of the 8,329 controls. The calculation of causality is then one of P(Exposure|Case) - P(Exposure|Control), 25.7 - 11.5 = 14.2, where 'exposure' here is defined as having a CNV >400Kb that has a frequency <1% in the population. The statement of causality is, in our view, highly misleading because it assumes complete penetrance and ignores other risk variants harbored by the cases that may result from the process of the ascertainment of the case sample. The odds ratio of 2.7 or relative risk of 1.3 is the appropriate way to describe the contribution of CNVs >400kb in their study. These attributable fraction measures imply that the number of cases can be split up into categories that represent mutually exclusive (additive) causal factors. But for genetic factors and complex diseases, such as those discussed in this perspective, mutations may not be fully penetrant and hence on the risk scale the probability of being a case depends on multiple factors that are more likely to act multiplicatively than additively on that scale. Statements of causality are not justified when based solely on burden of a class of variants in cases versus controls.

For the reasons outlined above we interpret the empirical data to favor a model in which *de novo* mutations make an important but minor contribution to the liability for neuropsychiatric disorders, as opposed to accounting for the majority of risk. Our position is similar to that of Neale *et al.*[11], who concluded that the observed data on *de novo* protein-coding mutations was consistent with an overall contribution of <5% of the liability for ASD, and was not consistent with a model comprising a large number of pseudo-Mendelian mutations. It is worth noting that *de novo* mutations of large effect are not required to explain sporadic cases of complex neuropsychiatric

disease. For disorders such as schizophrenia and ASD with high heritability (e.g. ~80%) and low prevalence (e.g. ~1%), a high proportion of sporadic cases are expected under a purely polygenic model involving many contributing loci of individually small effect size[16]. Furthermore, polygenic variation is estimated to account for approximately one third of the genetic risk in genetic studies of schizophrenia in which the proportion of sporadic cases is high[17-19]. The empirical evidence, from multiple sources, therefore implies that *de novo* mutations (where present) generally combine with inherited variants to confer risk, and that the majority of genetic risk is explained by segregating variation. Recent sequencing studies have revealed a tremendous amount of rare functional variation in human populations[20-22], and disease studies suggest that this variation contributes to individual variation[23-27]. Variants that are predicted to be functionally important are overwhelmingly rare[22], and likely comprise many recently arisen (i.e. *de novo*) mutations that have survived selection and are shared by family members. We consider it likely that such variants collectively make a substantial contribution to the heritability, particularly for severe early-onset disorders (e.g. ASD, MR).

A number of ASD studies mentioned in this article, although primarily focused on the role of *de novo* mutations, have tested for a higher burden of rare inherited protein-coding variation in probands compared to controls[10,13,28]. All three studies were negative, as was a smaller case-control study of idiopathic epilepsy[29]. Although these results could be interpreted as evidence that rare inherited variation does not contribute to disease risk, power to detect an excess of functional segregating protein-coding variation in cases is low given the size of the studies. There are three reasons for this: first, only a proportion of all functional variants will be relevant for any particular disorder (e.g. ~5% for ASD assuming 1000 risk genes out of a total of 20,000). Second, these variants will explain only a modest proportion of the overall risk, and third, the relative burden in cases and controls will be "muddied" by the presence of many functional variants in controls that predispose these individuals to any number of complex diseases other than that under investigation. Much larger studies, in ASD and other neuropsychiatric disorders, will be required to fully establish the relative contribution of *de novo* and inherited variation to risk of disease.



Supplementary Figure 1. The logarithm of the recurrence risk (RR) to relatives for schizophrenia is linear with respect to genetic relationship (linear regression, $F_{(1,14)}$ = 347.1, p = 3.12e-11, R^2 = 0.958), as predicted under a multiple variant risk model. Data from Lichtenstein *et al*.[4] and McGue *et al*.[5].

References

1. Lynch, M. The rate of polygenic mutation. *Genet Res* **51**, 137-48 (1988).
2. Risch, N. Linkage strategies for genetically complex traits. I. Multilocus models. *Am J Hum Genet* **46**, 222-8 (1990).
3. Pharoah, P.D. *et al.* Polygenic susceptibility to breast cancer and implications for prevention. *Nat Genet* **31**, 33-6 (2002).
4. Lichtenstein, P. *et al.* Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet* **373**, 234-239 (2009).
5. McGue, M., Gottesman, II & Rao, D.C. Resolving genetic models for the transmission of schizophrenia. *Genet Epidemiol* **2**, 99-110 (1985).
6. McGue, M., Gottesman, II & Rao, D.C. The transmission of schizophrenia under a multifactorial threshold model. *Am J Hum Genet* **35**, 1161-78 (1983).
7. Ronald, A. & Hoekstra, R.A. Autism spectrum disorders and autistic traits: a decade of new twin studies. *Am J Med Genet B Neuropsychiatr Genet* **156B**, 255-74 (2011).
8. Rosenberg, R.E. *et al.* Characteristics and concordance of autism spectrum disorders among 277 twin pairs. *Arch Pediatr Adolesc Med* **163**, 907-14 (2009).
9. Kemper, K.E., Visscher, P.M. & Goddard, M.E. Genetic architecture of body size in mammals. *Genome Biol* **13**, 244 (2012).
10. Sanders, S.J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-41 (2012).
11. Neale, B.M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242-5 (2012).
12. Cooper, G.M. *et al.* A copy number variation morbidity map of developmental delay. *Nat Genet* **43**, 838-46 (2011).
13. Iossifov, I. *et al.* De novo gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285-99 (2012).
14. Vissers, L.E.L.M. *et al.* A de novo paradigm for mental retardation. *Nature Genetics* **42**, 1109-1112 (2010).
15. Xu, B. *et al.* Exome sequencing supports a de novo mutational paradigm for schizophrenia. *Nature Genetics* **43**, 864-U72 (2011).
16. Yang, J., Visscher, P.M. & Wray, N.R. Sporadic cases are the norm for complex disease. *Eur J Hum Genet* **18**, 1039-43 (2009).
17. Lee, S.H. *et al.* Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat Genet* **44**, 247-50 (2012).
18. Purcell, S.M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748-752 (2009).
19. Ripke, S. *et al.* Genome-wide association study identifies five new schizophrenia loci. *Nat Genet* **43**, 969-76 (2011).
20. Durbin, R.M. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073 (2010).
21. Nelson, M.R. *et al.* An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337**, 100-4 (2012).
22. Tennessen, J.A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64-9 (2012).
23. Bonnefond, A. *et al.* Rare MTNR1B variants impairing melatonin receptor 1B function contribute to type 2 diabetes. *Nat Genet* **44**, 297-301 (2012).
24. Raychaudhuri, S. *et al.* A rare penetrant mutation in CFH confers high risk of age-related macular degeneration. *Nat Genet* **43**, 1232-6 (2011).

25. Raychaudhuri, S. *et al.* Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat Genet* **44**, 291-6 (2012).
26. Rivas, M.A. *et al.* Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat Genet* **43**, 1066-73 (2011).
27. Sulem, P. *et al.* Identification of low-frequency variants associated with gout and serum uric acid levels. *Nat Genet* **43**, 1127-30 (2011).
28. O'Roak, B.J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nature Genetics* **43**, 585-589 (2011).
29. Klassen, T. *et al.* Exome sequencing of ion channel genes reveals complex profiles confounding personal risk assessment in epilepsy. *Cell* **145**, 1036-1048 (2011).