Whole Genome Sequencing in Psychiatric Disorders: the WGSPD Consortium

Supplemental Data

List of Figures

List of Tables

Contents

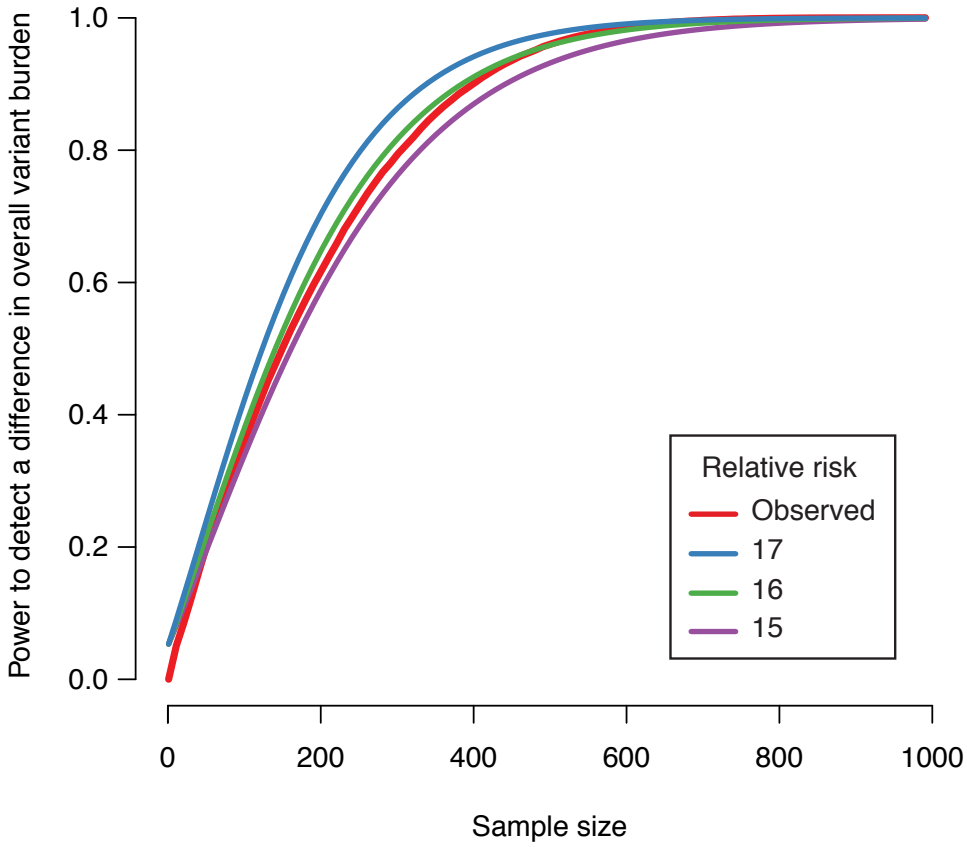

Figure 1. Estimate of statistical power from resampling de novo PTV mutations in ASD exome data.
De novo PTV mutations from exome data of 1,881 quartet families from the Simons Simplex Collection[1] was resampled to estimate the power to detect a difference in burden (alpha = 0.05) between cases and sibling controls (red line). This estimate was compared to predicted power curves based on a relative risk of 15 (purple line), 16 (green line) and 17 (blue line) as described in section 1.5. The closest match was with a relative risk of 16, suggesting that the mean relative risk of PTV mutations in ASD genes was 16, assuming 5% of genes contribute to ASD risk[2,3]. PTV: Protein Truncating Variant.

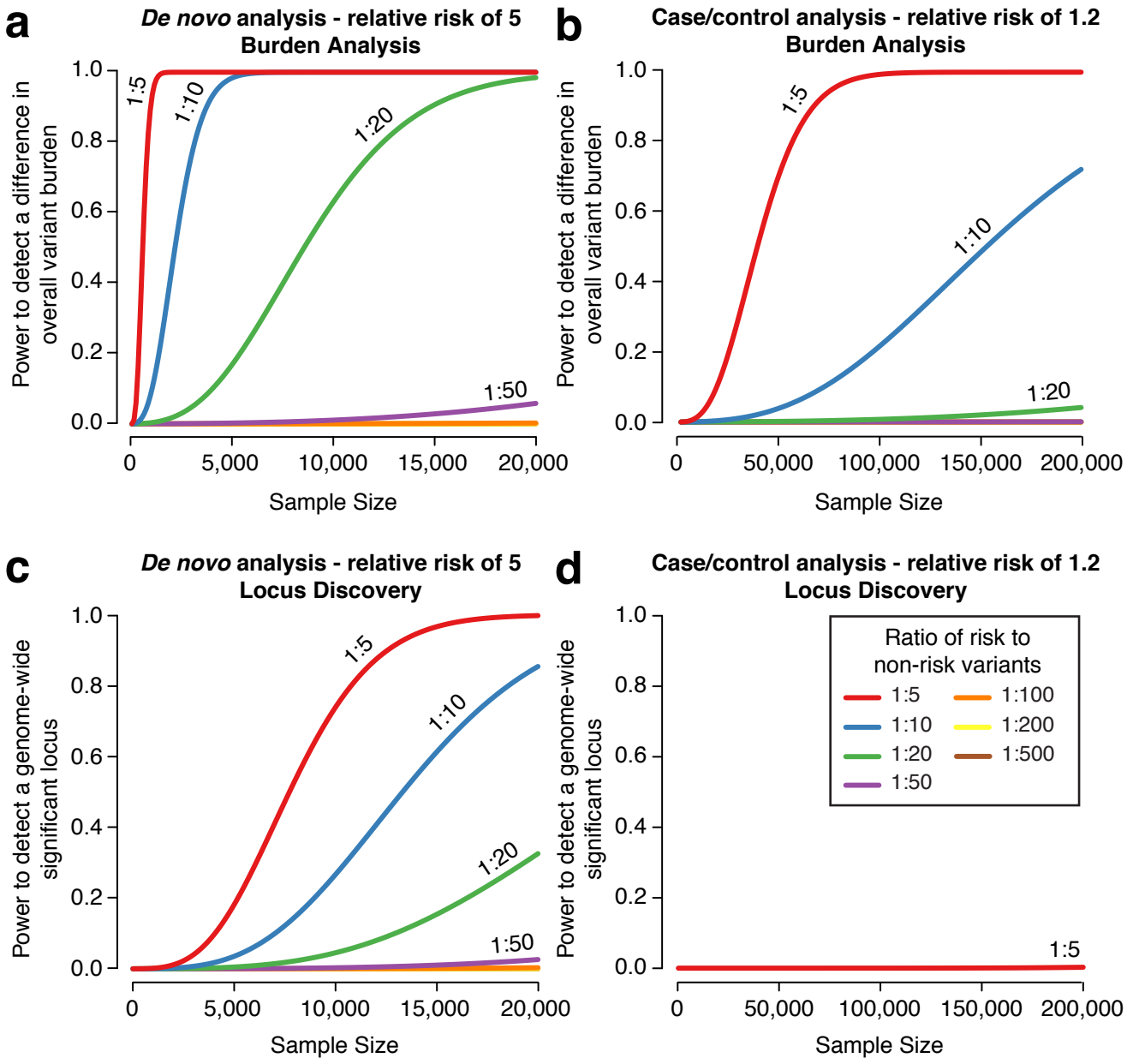

Figure 2. **Statistical power in the non-coding genome by cohort size, Related to Figure 2.**

Figure 2. Statistical power in the non-coding genome by cohort size, Related to Figure 2. We estimated the power at a significance threshold (alpha) of 5 x 10^{-5}, accounting for 1,000 categories of noncoding variants, to detect an excess of noncoding variants at 122,500 risk loci in cases vs. controls as we varied the sample size and risk:non-risk ratio, which represents annotation quality (Table 2). In **a)** we assessed the power for detecting an excess of *de novo* mutations at a relative risk of 5 as sample size increases. With a risk:non-risk ratio of 1:20, approximately equivalent to assessing protein truncating variants in the coding genome, we achieve >80% power with a sample size of 5,000. In **b)** the power to detect an excess burden of rare variants (allele frequency ≤0.1%) is assessed at a relative risk of 1.2. In **c)** we assessed the power to identify an excess of *de novo* mutations at a specific genomic locus, e.g. the noncoding region regulating a single gene. Consequently, we set the significance threshold (alpha) at $2.5x10^{-6}$. In **d)** we assessed the power to identify an excess of rare variants (allele frequency ≤0.1%) at a specific nucleotide (alpha = $1.7x10^{-11}$), since this yielded better power than testing for burden at a locus (alpha = $2.5x10^{-6}$).

Risk:Non-risk variants	Number of risk variants per control	Number of non-risk variants per control	Number of variants per control	Proportion of risk variants
5:1	0.005	0.01992	0.0249	0.2
10:1	0.005	0.04482	0.0498	0.1
20:1	0.005	0.09462	0.0996	0.05
50:1	0.005	0.24402	0.249	0.02
100:1	0.005	0.49302	0.498	0.01
200:1	0.005	0.99102	0.996	0.005
500:1	0.005	2.48502	2.49	0.002

Table 1. Number of variants and proportion of risk variants for *de novo* analysis.

Risk:Non-risk variants	Number of variants (kbp)	Percent of genome	Approximate equivalent for detecting coding variants
1:0	123	0.004%	PTV in disorder associated genes
1:1	245	0.008%	
1:2	368	0.012%	PTV in loss-of-function intolerant (pLI) genes
1:5	735	0.02%	
1:10	1,348	0.04%	All coding variants in disorder associated genes
1:20	2,573	0.08%	PTV in all genes
1:50	6,248	0.2%	All coding variants in loss-of-function intolerant (pLI) genes
1:100	12,373	0.4%	
1:200	24,623	0.8%	All coding variants in all genes
1:500	61,373	2%	
1:1000	122,623	4%	All variants in constrained regions (GERP)
1:2000	245,123	8%	
1:5000	612,623	20%	

Table 2. Ratio of risk to non-risk variants and number of variants considered.

Reference Cohort Size	Minimum heterozygote allele frequency	Minimum homozygote allele frequency	Example
2,000	2.5×10^{-4}	1.6×10^{-8}	1000 Genomes
5,000	1.0×10^{-4}	2.5×10^{-9}	
20,000	2.5×10^{-5}	1.6×10^{-10}	WGSPD
50,000	1.0×10^{-5}	2.5×10^{-11}	ExAC
200,000	2.5×10^{-6}	1.6×10^{-12}	
500,000	1.0×10^{-6}	2.5×10^{-13}	
2,000,000	2.5×10^{-7}	1.6×10^{-14}	
5,000,000	1.0×10^{-7}	2.5×10^{-15}	
20,000,000	2.5×10^{-8}	1.6×10^{-16}	
50,000,000	1.0×10^{-8}	2.5×10^{-17}	

Table 3. Lowest allele frequency distinguishable from a population cohort.

Supplemental Experimental Procedures

1 Estimating statistical power in non-coding regions

To assess the statistical power in the non-coding genome, four variables need to be estimated:

1. Number of cases and controls
2. Frequency of variants per person
3. Proportion of variants mediating risk (risk variants)
4. Mean relative risk mediated by risk variants

The methods for estimating these variables are described separately for family-based *de novo* analysis and case/control rare variant analysis.

1.1 Number of cases and controls - *de novo*

Based on Table 2 in the main text, we anticipate at least 5,000 ASD families with affected cases and matched unaffected siblings being available for *de novo* analysis.

1.2 Frequency of variants per person - *de novo*

De novo mutations occur at a frequency of 1.15×10^{-8} mutations per nucleotide per generation[4] resulting in about 74 *de novo* mutations per diploid genome per individual[5].

1.3 Proportion of variants mediating risk - *de novo*

In neuropsychiatric disorders the protein-truncating variants (PTV, also called loss of function [LoF] or likely gene disrupting [LGD] in the literature) are the best-characterized, therefore we shall base our estimate of the proportion of variants on the PTV distribution. Based on exome analysis in unaffected siblings, 0.0996 *de novo* PTV mutations are observed per individual[1] and 5% of these would be expected to fall in genes associated with ASD risk[2,3], leading to an estimate of 0.005 (0.0966 * 5%) risk mediating PTVs per unaffected individual. This estimate was used to calculate the number of variants and proportion of risk variants (Table 1). A variety of ratios between risk and non-risk variants were considered, representing the ability to distinguish risk mediating variants from non-risk mediating variants (Table 2).

1.4 Mean relative risk mediated by risk variants - *de novo* and rare variants

Since we do not have clear examples of rare non-coding variants that mediate neuropsychiatric risk, it is hard to know what relative risk to estimate. There are examples of non-coding variants mediating Mendelian levels of risk (e.g. SNPs in an enhancer of the gene Oculocutaneous Albinism Type 2 [OCA2] lead to blue eye color; mutations in the zone of polarizing activity regulatory sequence [ZRS] enhancer of Sonic Hedgehog [SHH] lead to polydactyly[6]), however, as with coding variants, these are likely to represent exceptions in neuropsychiatric phenotypes. We estimate that the mean relative risk for *de novo* PTV observed in ASD is 16 (Figure 1). The relative risk at individual loci is likely to follow an exponential distribution[7,8], with the ASD genes identified to date representing the loci with the highest relative risks, probably in excess of 20[1,9].

In contrast the relative risk of common variants associated with schizophrenia is ≤ 1.27[10]. Rare variants are likely to lie between these two extremes, however at an allele frequency $\leq 0.1\%$ the relative risk is probably closer to those observed with common variants, due to natural selection[11–13]. Given this uncertainty, we estimated power across a range of estimates of relative risk (Figure 2, main manuscript). For analyses where the relative risk needs to be fixed, we will use 5 for *de novo* mutations and 1.2 for rare variants.

1.5 Calculating power for overall burden of variants - *de novo*

To estimate the statistical power to detect an overall excess of *de novo* variants in cases, the non-risk variants (Table 1) were distributed equally between cases and controls, while the risk variants (Table 1) were distributed based on the relative risk. Power was estimated based on a Poisson distribution with alpha = 0.05. The standard deviation was estimated based on the distribution of *de novo* PTV mutations in unaffected siblings[1]. The results are shown in the main manuscript (Figure 2) and Figure 2.

1.6 Calculating power for locus detection - *de novo*

De novo mutations are sufficiently rare that we would not expect to observe two at the same nucleotide, even in a cohort of several thousand individuals. To detect an excess at a specific locus we therefore need to group the mutations into functional blocks, for example those that regulate a specific coding gene. The number of comparisons is therefore about 20,000, leading to an alpha of 2.5×10^{-6} (0.05 / 20,000). The number of risk and non risk mutations is shown in Table 1 and a t test was used to estimate power. The t test was shown to be an accurate approximation to permutation testing with a poisson model, see https://github.com/hailianghuang/BurdenPower. The results are shown in the main manuscript (Figure 2) and Figure 2.

1.7 Number of cases and controls - rare variants

Based on Table 1 in the main text, we anticipate at least 20,000 cases and 20,000 matched controls being available for rare variant case/control analysis.

1.8 Frequency of variants per person - rare variants

As with *de novo* mutations, we do not know the number of risk mediating variants in the non-coding genome so we will use PTVs as a proxy. Of the 35 million nucleotides in the Gencode protein coding exome[14], 7% are PTV (35,000,000 * 7% = 2,450,000 nucleotides). Since we estimate that PTVs in 5% of protein coding genes contribute to ASD risk[2,3] this leads to an estimate of 122,500 non-coding nucleotides capable of mediating neuropsychiatric risk (2,450,000 * 5% = 122,500 nucleotides) and 123 risk mediating non-coding nucleotides for each of the 1,000 risk genes.

1.9 Calculating power for overall burden of variants - rare variants

To estimate the statistical power to detect an overall excess of rare variants in cases, the non-risk variants were distributed equally between cases and controls, while the risk variants were distributed based on the relative risk. Power was estimated based on a Poisson distribution with alpha = 0.05. The standard deviation was estimated based on the distribution of rare PTV mutations in unaffected siblings[1]. The results are shown in the main manuscript (Figure 2) and Figure 2.

1.10 Calculating power for locus detection - rare variants

For rare variants we could aim to identify an excess of variants in cases within a functional block (e.g. regulators of a coding gene), leading to about 20,000 comparisons (alpha = 2.5×10^{-6}) or at a specific nucleotide with 3 billion comparisons (alpha = 1.7×10^{-11}). The later approach was consistently better powered. The rare variants were modeled with a exponential distribution of allele frequencies and a mean allele frequency of 0.1%. The risk variants (see section 1.8) were selected from the lower end of this distribution. Power was estimated through 1,000 permutations with a case:control ratio of 1, and population prevalence of 1%. For more details on the methods, and access to the code used to perform these power calculations, see https://github.com/hailianghuang/BurdenPower.

2 Homozygous variant frequencies

The ability to distinguish heterozygous variants at very low population frequencies is directly related to the size of the reference cohort, therefore 50 million individuals are required to identify variants at a similar frequency to a *de novo* mutation (1.2×10^{-8})[4]. In contrast, the frequency of homozygous variants is a function of the size of the reference cohort squared and a comparable population frequency can be distinguished with 2,000 individuals (Table 3). In practice, a slightly higher cohort size is probably needed to reliably make this distinction, since higher frequency variants may be missed by chance and are, by definition, more common.

References

1. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
2. Sanders, S. J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237–41 (2012).
3. He, X. *et al.* Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. *Am J Hum Genet* **92**, 667–680 (2013).
4. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci U S A* **107**, 961–968 (2010).
5. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).
6. Lettice, L. A. *et al.* A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* **12**, 1725–1735 (2003).
7. El-Fishawy, P. & State, M. W. The genetics of autism: key issues, recent findings, and clinical implications. *Psychiatr Clin North Am* **33**, 83–105 (2010).
8. Owen, M. J., Sawa, A. & Mortensen, P. B. Schizophrenia. *The Lancet* 86–97 (2016).
9. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209–215 (2014).
10. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
11. Sanders, S. J. First glimpses of the neurobiology of autism spectrum disorder. *Current Opinion in Genetics & Development* (2015).
12. Krumm, N. *et al.* Excess of rare, inherited truncating mutations in autism. *Nature genetics* **47**, 582–588 (2015).
13. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016). arXiv:030338.
14. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760–1774 (2012).