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 (n)
2. Number of variants per person (v)
3. Proportion of variants mediating risk (f)
4. Mean relative risk mediated by risk variants (RR)

The methods for estimating these variables are described separately for family-based *de novo* analysis and case/control rare variant analysis. The code for performing these analyses is in an R package, available at `https://github.com/sanderslab/wgsPowerCalc` with vignettes that generate the power calculation images shown in this manuscript.

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 Number of variants per person - *de novo*

De novo mutations occur at a frequency of 1.15×10^{-8} mutations per nucleotide per generation[1] resulting in about 74 *de novo* mutations per diploid genome per individual[2].

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 that mediate risk on the PTV distribution. Based on exome analysis in unaffected siblings, 0.0996 *de novo* PTV mutations are observed per individual[3] and 5% of these would be expected to fall in the ~1,000 genes associated with ASD risk[4,5], leading to an estimate of 0.00498 (0.0996 * 5%) risk mediating PTVs per unaffected individual. This estimate was used to calculate the number of variants and proportion of risk variants (Supplementary 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 (Supplementary Table 3).

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. 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 50[3,9].

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 1, 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 (Supplementary Table 1) were distributed equally between cases and controls, while the risk variants (Supplementary Table 1) were distributed based on the relative risk. Power was estimated based on a Poisson distribution with alpha 5×10^{-5} (0.05 / 1,000) to account for the multiple comparisons necessary to identify disease-mediating functional risk in the noncoding genome (plotDnBurdenByRelativeRisk and plotDnBurdenBySampleSize functions in the wgsPowerCalc R package, which call the getDeNovoBurdenPower function). The standard deviation of the distribution of risk variant counts (0.31) was estimated based on the distribution of *de novo* PTV mutations in unaffected siblings[3]. The results are shown in the main manuscript (Figure 1) and Supplementary Figure 1.

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 one 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 (r) and non risk (v) mutations is shown in Supplementary Table 1.

We assume that the expected number of *de novo* mutations per individual genome is Q_0. If we perform the *de novo* analysis on a proportion of the genome (denoted as f, $f \in [0,1]$), the expected number of mutations in this selected region of genome is fQ_0. For cases, assume a proportion (denoted as p, $p \in [0,1]$) of the *de novo* mutations are functional with with relative risk of R. The expected number of *de novo* mutations in cases is $Q_1 = (1-p)fQ_0 + RpfQ_0$ (see proof 1 in section 3).

We simulate the number of *de novo* mutations for controls using a Poisson distribution with the Poisson parameter of fQ_0. We simulate the number of *de novo* mutations for cases using a mixture of two Poisson distributions with Poisson parameters of $(1-p)fQ_0$ and $RpfQ_0$ respectively. We then perform a Poisson regression and get the deviance between the Poisson model (simulated number of *de novo* mutations versus case/control status) and the null model. The deviance is chi-square distributed and will be used to calculate the power (plotDnLocusByRelativeRisk and plotDnLocusBySampleSize functions in the wgsPowerCalc R package, which call the getDeNovoLocusPowerParametric function).

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 ancestry 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[4,5] 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 5×10^{-5} (0.05 / 1,000) to account for the multiple comparisons necessary to identify disease-mediating functional risk in the noncoding genome (plotCcBurdenByRelativeRisk and plotCcBurdenBySampleSize functions in the wgsPowerCalc R package, which call the getCaseConBurdenPower function). The standard deviation of the distribution of rare variant counts was estimated based on the distribution of rare PTV mutations in unaffected siblings[3]. The number of risk (r) and non risk (v) variants is shown in Supplementary Table 2. The results are shown in the main manuscript (Figure 1) and Supplementary Figure 1.

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% .

1.10.1 Power for case/control analysis

Let x_i be the genotype of variant i and y be the phenotype of all individuals. We assume there are s SNPs in a gene and f of them are functional. The allele frequency of the SNPs is sampled from the exponential distribution: $\Pr(AF) = \lambda \exp^{-\lambda AF}$ with $\lambda = 1/\overline{AF}$ (the sampled AF therefore has mean of $\overline{AF}$). The probability for being a risk mediating variant scales with the allele frequency: $\Pr(\text{causal}|AF) = 1/AF$.

1.10.2 Single variant test

The power for a single variant can be calculated using the chi-square distribution with the non-centrality parameter χ^2. This parameter is estimated using the NCPgen function in the wgsPowerCalc R package (basically a 2x2 contingency table test). The power for finding at least one out of the f functional SNPs is $1 - \prod_{i \in f}(1 - \text{power}_{\text{snp}_i})$ (plotCcLocusSingleByRelativeRisk and plotCcLocusSingleBySampleSize functions in the wgsPowerCalc R package, which call the getSingleVarLocusPower function).

1.10.3 Multiple variant test

In the multiple variant test, assuming SNPs are independent, we have $x = \sum_{i \in f} x_i + \sum_{j \in s-f} x_j$. The coefficients are

$$\begin{aligned}
\beta' &= \frac{cov(y,x)}{var(x)} \\
&= \frac{\sum_{i \in f} cov(x_i,y) + \sum_{j \in s-f} cov(x_j,y)}{\sum_{i \in f} var(x_i) + \sum_{j \in s-f} var(x_j)} \\
&= \frac{\sum_{i \in f} cov(x_i,y) + \sum_{j \in s-f} cov(x_j,y)}{\sum_{i \in f} var(x_i) + \sum_{j \in s-f} var(x_j)} \\
&= \frac{\beta \sum_{i \in f} var(x_i) + 0 * \sum_{j \in s-f} var(x_j)}{\sum_{i \in f} var(x_i) + \sum_{j \in s-f} var(x_j)} \\
&= \beta \frac{\sum_{i \in f} var(x_i)}{\sum_{i \in f} var(x_i) + \sum_{j \in s-f} var(x_j)}
\end{aligned}$$

The power for the multiple variant test can be calculated using the chi-square distribution with the non-centrality parameter χ^2 calculated using the new coefficient β' (plotCcLocusMultiByRelativeRisk and plotCcLocusMultiBySampleSize functions in the wgsPowerCalc R package, which call the getMultiVarLocusPower function).

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 ($1.2 \text{x} 10^{-8}$)[1]. 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 (Supplementary Table 4). 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.

3 Proof 1

We prove that under the conditions: a) the population prevalence of the disease is small ($\Pr(case) \ll 1$); and b) the de novo rate is low ($\Pr(A) \equiv \mu \ll 1$),

$$\frac{\Pr(A|case)}{\Pr(A|control)} \simeq R,$$

in which R is the relative risk. Recall that $R \equiv \frac{\Pr(case|A)}{\Pr(case|a)}$

$$\begin{aligned}
\frac{\Pr(A|case)}{\Pr(A|control)} &= \frac{\Pr(case|A)}{\Pr(control|A)} \times \frac{\Pr(control)}{\Pr(case)} \\
&= R \times \frac{\Pr(case|a)}{\Pr(control|A)} \times \frac{\Pr(control)}{\Pr(case)} \\
&= R \times \frac{\Pr(case|a)}{\Pr(case)} \times \frac{\Pr(control)}{\Pr(control|A)}.
\end{aligned} \tag{1}$$

We will then prove $\frac{\Pr(case|a)}{\Pr(case)} \simeq 1$ and $\frac{\Pr(control)}{\Pr(control|A)} \simeq 1$.

$$\begin{aligned}
\frac{\Pr(case|a)}{\Pr(case)} &= \frac{\Pr(case|a)}{\Pr(case|a)\Pr(a) + \Pr(case|A)\Pr(A)} \\
&= (\Pr(a) + \frac{\Pr(case|A)}{\Pr(case|a)}\mu)^{-1} \\
&= (\Pr(a) + R\mu)^{-1} \\
&= (1 - \mu + R\mu)^{-1} \\
&= (1 + (R-1)\mu)^{-1} \\
&\simeq 1 - (R-1)\mu \\
&\simeq 1.
\end{aligned} \tag{2}$$

And

$$\begin{aligned}\frac{\Pr(control)}{\Pr(control|A)} &= \frac{\Pr(control)}{1-\Pr(case|A)} \\ &= \frac{\Pr(control)}{1-R\Pr(case|a)}.\end{aligned} \tag{3}$$

For $\Pr(case|a)$, we have

$$\begin{aligned}\Pr(case|a) &= \frac{\Pr(case)}{\Pr(a)+R\Pr(A)} \\ &= \frac{\Pr(case)}{1+(R-1)\mu} \\ &\simeq K-K(R-1)\mu.\end{aligned} \tag{4}$$

Using equations 3 and 4, we have

$$\begin{aligned}\frac{\Pr(control)}{\Pr(control|A)} &\simeq \frac{\Pr(control)}{1-R(K-K(R-1)\mu)} \\ &\simeq \frac{1-K}{1-RK+R(R-1)K\mu} \\ &\simeq (1-K)(1+RK-R(R-1)K\mu) \\ &\simeq 1+(R-1)K-RK^2-R(R-1)K\mu+R(R-1)K^2\mu \\ &\simeq 1.\end{aligned} \tag{5}$$

Using equations 1, 2 and 5:

$$\frac{\Pr(A|case)}{\Pr(A|control)} \simeq .R \tag{6}$$

References

1. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci U S A* **107**, 961–968 (2010).
2. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).
3. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
4. Sanders, S. J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237–41 (2012).
5. 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).
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).