Purcell et al. A polygenic burden of rare disruptive mutations in schizophrenia.

Online Methods

1. Samples and ascertainment

All sequenced individuals are part of a larger Swedish cohort, consisting of 11,850 subjects (5,351 cases and 6,509 controls). Genome-wide association study (GWAS) single nucleotide polymorphism (SNP) data have been collected on all samples, using either the Affymetrix 5.0, Affymetrix 6.0 or Illumina Omni Express platforms. From this cohort, we selected 5,091 individuals for exome sequencing. Cases were identified via the Hospital Discharge Register (HDR) that captures >99% of all inpatient hospitalizations in Sweden (Kristjansson et al., 1987; Dalman et al., 2002). The register is complete from 1987 and augmented by psychiatric data from 1973-86. It contains dates and ICD discharge diagnoses (World Health Organization, 1992) for each hospitalization, and captures the clinical diagnosis made by the attending physician. Case inclusion criteria: ≥2 hospitalizations with a discharge diagnosis of schizophrenia, both parents born in Scandinavia, and age ≥18 years. Case exclusion criteria: hospital register diagnosis of any medical or psychiatric disorder mitigating a confident diagnosis of schizophrenia as determined by expert review, and included removal of 3.4% of eligible cases due to the primacy of another psychiatric disorder (0.9%) or a general medical condition (0.3%) or uncertainties in the Hospital Discharge Register (e.g., contiguous admissions with brief total duration, 2.2%). The validity of this case definition of schizophrenia is strongly supported as described in Ripke et al. (2013). Controls were selected at random from Swedish population registers. Control inclusion criteria: never hospitalized for schizophrenia or bipolar disorder (given evidence of genetic overlap with schizophrenia), both parents born in Scandinavia, and age ≥18 years.

Cases were ascertained from across Sweden using the HDR from 2005-11. The sampling frame is thus population-based and covers all hospital-treated patients. Ethical committees in Sweden and in the US approved all procedures and all subjects provided written informed consent (or legal guardian consent and subject assent). We also obtained permissions from the area health board to which potential subjects were registered. Potential cases were contacted directly via an introductory letter followed by a telephone call. If they agreed, a research nurse met them at a psychiatric treatment facility or in their home, obtained written informed consent, obtained a blood sample, and conducted a brief interview about other medical conditions in a lifetime. Controls were contacted directly in a similar procedure as the cases, gave

written informed consent, were interviewed about other medical conditions and visited their family doctor or local hospital laboratory for blood donation.

2. Sequencing protocol and variant calling pipeline

The whole exome hybrid selection sequencing process begins with genomic DNA that is positionally locked in a 96-well plate based format, to reduce sample misidentifications and to increase accuracy of the liquid handling instrumentation. The genomic DNA is then sheared to a target of 150bp using ultrasonic wave technology in a Covaris. The sheared product is then transformed into a library using proprietary barcoded adaptors, and then all the barcoded libraries from one plate are pooled for increased laboratory efficiency. Fragments of interest are isolated from the pool of libraries by hybridization with oligos that are complementary to the target regions of the exome. QC is performed on the isolate using qPCR in order to obtain the optimal concentration for cluster generation and sequencing on Illumina's HiSeq2000 instrument. The resulting data is in 75bp paired-end format. Inspections are performed on the sequencing output in order to assure high quality.

The sample was sequenced in seven waves; the first wave (*N* = 132) employed an earlier version of the hybrid-capture procedure (Agilent SureSelect Human All Exon Kit), that targeted ~160,000 regions (~28Mb), whereas the remaining samples targeted ~190,000 intervals (~32Mb target; Agilent SureSelect Human All Exon v.2 Kit). The initial wave was sequenced using an Illumina GAII instrument rather than a HiSeq. Downstream association analyses controlled for wave. Each wave was matched for the number of cases and controls; within most, but not all, waves, samples were randomly and equally assigned to plates with respect to disease status.

We used the Picard/BWA/GATK (http://picard.sourceforge.net; Li & Durbin, 2009; McKenna et al., 2010) pipeline at the Broad Institute to process, align and call variants from the raw read data. Variant calls were made on the entire sample creating a single Variant Call Format (VCF) file (in contrast to splitting the sample into smaller batches that necessitates downstream merging of VCFs). Raw read data were visualized using the Integrative Genome Viewer (IGV, Thorvaldsdóttir et al., 2012). The BAM and VCF files for this study are available in the dbGaP study phs000473.v1 Sweden-Schizophrenia Population-Based Case-Control Exome Sequencing.

3. Quality control

Individual level QC

We calculated a suite of per-individual metrics based on a) all variants, and b) all "on-target" single nucleotide variants (SNVs) that passed the default GATK filters (defined as the targeted coding regions with five bases extending into introns to cover potential splice site mutations). Metrics included the proportion of variants present in dbSNP (build 137), the total number of alternate alleles, mean heterozygosity, mean X chromosome heterozygosity, mean transition/transversion ratio at heterozygote sites. We also considered technical Picard coverage metrics. From available GWAS data, we estimated per-individual heterozygosity (~inbreeding) and pairwise relatedness using PLINK (Purcell et al., 2007). We also calculated the first four multidimensional scaling (MDS) components to represent ancestry.

We used the same approach to estimate identity-by-state (IBS) and identity-by-descent (IBD) for the same samples based on genotypes from the exome sequence data. We extracted ~5000 SNPs selected based on HapMap3 data for being a) polymorphic in all HapMap3 samples, b) present in the targeted coding regions of most exome capture platforms, c) of good genotyping properties (missingness, Hardy-Weinberg equilibrium), d) autosomal, and e) in approximate linkage equilibrium. The first MDS component correlated very strongly (>0.9) when estimated using SNP microarray versus the exome sequence data. (As a small proportion of samples had GWAS data that had failed QC measures, we used the exome-derived IBS estimates in all downstream association analyses, to control for substructure, rather than remove these sequenced samples.)

We calculated the mean dbSNP percentage per sample and removed clear outliers (*N* = 11) based on dbSNP percentage less than 95%. After removing these 11 individuals, we observed a number of apparent outliers, with a greater than average rate of singletons. However, rather than representing technical artifact, these individuals had disproportionately high rates of singletons that were present in dbSNP, suggesting that these are real variants. By considering estimated ancestry (MDS components), these few individuals appeared to have more southern European ancestry (based on comparison with HapMap samples) and thus correspondingly higher rates of non-novel variants that are singletons against the background of the larger Swedish sample. We retained these individuals in analysis, as all association analyses conditioned on ancestry.

The sample also contains individuals of (partially) Finnish as well as Swedish ancestry, as shown in ED Figure 2a. We used the GWAS and exome MDS components to identify a) the Finnish subset of the sample (*N* = 413) and b) an “ultra-homogeneous” subset of Swedish individuals (*N* = 3,226). We repeated key analyses in the latter homogeneous subset, which did not substantively change the picture of results (data not shown).

After removal of individuals with low dbSNP percentages, the other metrics considered displayed broadly even profiles across individuals, or appeared unrelated to the primary downstream metrics such as the number of alternate alleles called per individual, and so were not used as the basis for further individual-level QC.

Detecting and correcting potential cross-sample contamination

Some samples in waves 6 and 7 experienced low levels of cross-sample contamination due to technical issues during sequencing. Using verifyBamID (http://genome.sph.umich.edu/wiki/VerifyBamID), we estimated that 29 samples had above 5% contamination, and 333 samples had above 2%. A greater number of samples (~2000) had a lower estimate around 1%. High levels of contamination can reduce the accuracy of variant calls and lead to false positives calls. This was addressed by an in house *in silico* decontamination procedure implemented in the GATK pipeline. Specifically, at every targeted site in the exome, the Unified Genotyper proceeds normally by dividing the pileup of bases from the NGS reads spanning the position into separate bins by sample. When running in contamination fixing mode it then performs a “biased down-sampling" of the data for each sample given an expected contamination level (which can be a global value applied to all samples and/or each sample can have its own specified level). Given a contamination level of *N*% for a particular sample, Unified Genotyper will ignore anywhere from 0% - *N*% of the bases for that sample by hypothesizing the genotype of the contaminating sample and choosing the most likely one (based on the resulting Genotype Quality of the likelihoods). Imagine that the sample has 100 bases at a position, 88 As and 12 Cs, with an expected contamination

fraction of 10%. We will hypothesize that the contaminating sample is a C/C homozygote and will ignore 10 Cs (selected randomly) because that maximizes the genotype quality (GQ) – in other words, it contains a better allele balance. This procedure requires that the original read depth is relatively high, as in the current study, and that the level of contamination is relatively constant through the exome for a given sample.

Prior to decontamination, we had indeed observed some individuals showed excess a) chromosome X heterozygosity if male, b) estimated IBD relatedness between multiplexed pairs of individuals (contamination was within pair) and c) a slightly higher rate of discordant genotype calls with GWAS/microarray data. The decontamination procedure rescued approximately 20 samples with relatively high estimated rates of contamination (~5% of reads) that would have otherwise been flagged and removed during standard QC.

Post decontamination, the above metrics indicated that a very low degree of contamination was still impacting the default calls, however. First, in males we observed a correlation between the BAM-derived contamination estimate and chromosome X heterozygosity (r = 0.38, $P<10^{-16}$), which still held for the subset of individuals with estimated contamination rates below 2% (r = 0.34, $P<10^{-16}$). Second, multiplexed pairs showed greater IBD sharing, when estimated from the exome data, compared to all pairs (P = 5e-4) – in contrast, performing the same analysis on the GWAS calls there was no inflation (P = 0.46). Third, comparing the few individuals with 5% or greater estimated contamination against the rest of the sample, we saw a 3.5-fold greater rate of GWAS/exome discordance ($P = 3\times10^{-6}$, 0.99424 versus 0.99834 concordance). We observed a similar effect for individuals with 2% or greater estimated contamination, but not for the greater number of individuals with only 1% estimated contamination.

However, we also noted that, as expected, false-positive calls (compared to GWAS) were supported by only small number of reads and so typically had low Genotype Quality (GQ) values. We further observed that the abovementioned effects were effectively removed by imposing additional genotype-level filters: specifically, we only called non-reference genotypes at sites where individual read-depth (DP) was greater or equal to 10 and the GQ was greater or equal to 90. Genotype filtering reduces, for example, the correlation with male chromosome X heterozygosity to r = 0.03 (P = 0.36). DP/GQ filters retained 96% of all PASS'ing singletons although a higher proportion of more common alleles were removed. The DP/GQ filtered dataset had significantly better properties in terms of dbSNP percentage (99% versus 98.4%, $P < 10^{-16}$) and per-individual transition/transversion ratio (Ts/Tv = 3.38 versus 3.34, $P < 10^{-16}$). All primary rare variant analyses are based on the filtered dataset unless otherwise noted. The particular DP/GQ filtering procedure employed here was chosen because the major focus of this study is on very rare variation. Different strategies may be required to optimize the callset for more common (e.g. >1% frequency) variants. In any case, we performed key analyses on both DP/GQ filtered and unfiltered callsets but did not observe substantive differences in results (data not shown).

Site QC

We retained for analysis only sites that PASS'ed GATK filters. We additionally removed sites that failed a Hardy-Weinberg Equilibrium test (P < 0.001, Fisher's exact test), removing 69,226 sites (many of these were poorly called indels, that were heterozygous in almost all samples). We only retained for the primary analysis variants that were within 5 bases of a targeted region (that is, coding and splice variants).

Sanger and ExomeChip validation of singleton calls

We performed a validation experiment for over 100 singletons, using Sanger sequencing. We selected 126 singletons and obtained Sanger sequencing data for 116 (8 primer design failures, 2 PCR failures). Of the 116, 74 were selected at random; the remaining were associated gene-disruptive mutations (ARC, calcium channel genes and *KYNU*). Two investigators, blind to the targeted site, manually and independently called variants from the Sanger traces.

For all but 3 singletons we observed clear validation of the non-reference allele, yielding an accuracy (the positive predictive value for an exome call indicating a true variant) of 97.4% for the NGS-called singletons. All of the 10 selected frameshifts were confirmed by Sanger sequencing. Based on this validation exercise, the vast majority of singletons in our data are very likely to represent real variant alleles. We repeated all primary association analyses with the three failed sites removed. One of the failures was an apparent dinucleotide polymorphism in *KYNU*. A second instance in which Sanger sequencing did not support the NGS-based call was for the postsynaptic density gene *CYFIP2*. The main text and Tables present results with these three variants removed from the dataset.

To evaluate sensitivity to detect true positive singleton mutations, we considered the available ExomeChip genotype calls on these samples, and compiled a list of sites where precisely one ExomeChip alternate allele was called as a heterozygote and all other individuals had a homozygote reference call (ignoring sites with low ExomeChip genotyping rates). After removing a handful of individuals who were outliers for a high number of ExomeChip singletons, there were 17,154 singleton sites in 4,293 individuals on whom both exome sequence and ExomeChip data were available. Conservatively assuming the chip calls to be the true positive gold standard, we looked at the corresponding calls in the sequence data to assess the sensitivity of our exome sequencing to detect true, singleton variants.

Of all chip singletons, 92.3% also had a heterozygote call in the same individual in the sequence data. Per individual, there was a very high correlation between the number of chip and sequence non-reference alleles at these sites (r=0.96). Importantly, using a linear model for the number of undetected alleles per individual, we found that sensitivity was unrelated to case/control status (P=0.48). As expected, it was however related to coverage: sensitivity was lower in the first experimental wave, which had used an earlier version of the exome target (~160,000 instead of ~190,000 targets). The proportion of targeted bases covered at 10X also independently predicted the number of “missed” alleles. Beyond average coverage, other investigators have also noted that the sensitivity of exome sequencing to detect true variants can be lower than expected from binomial sampling, likely representing “difficult” regions in the exome in which it is harder to call variants (e.g. Meynert et al., 2013).

We also used the comparison with the ExomeChip calls to examine the impact of the stringent genotype-level filtering that we employed on top of the standard GATK filters. (As noted above, this addressed residual low-level contamination between some samples.) We refer to these as the filtered and standard callsets. For the standard callset, we do see a higher sensitivity (94.1% versus 92.3%) suggesting, not unexpectedly, that filtering lost some proportion of true singletons. However, we also used the ExomeChip comparison to assess the joint accuracy considering both specificity and sensitivity: namely, how often an ExomeChip singleton is detected as a singleton in the sequencing data. That is, asking not just whether the expected individual carries the alternate allele, but also that nobody else does. Here we see better performance in the filtered compared to the standard callset, of 91.4% versus 85.5%. In other words, based on filtered

calls, in only ~1% of true singleton sites do we unexpectedly observe a call in one or more of the 4,292 other individuals. The standard callset had a much higher rate of 8.7% however. The discordances were distributed across all individuals, typically reflecting a doubleton call instead of a singleton (i.e. the expected individual and one other person). The 1% rate in the filtered callset is consistent with a very low false positive error rate also (i.e. only 177 out of the ~73 million ((4293-1)*17K) true reference genotypes are called as non-reference genotypes, yielding a specificity of 0.999998).

4. Annotation.

We used PLINK/Seq (http://atgu.mgh.harvard.edu/plinkseq/) to annotate all variants according to RefSeq gene transcripts (accessed from the UCSC Genome Browser, http://genome.ucsc.edu), after removing a handful of "rogue" transcripts (for example: if CDS length was not modulo 3, or if the reference sequence contained a large number of stop codons). We collapsed annotations for all transcripts of the same gene to give a "worst" annotation per gene. Annotation and downstream analyses were allele/gene specific, meaning that different alternate (non-reference) alleles of the same variant were assigned unique annotations; similarly, in the case of overlapping genes, the same allele could for different genes have different annotations. For example, a mutation might be annotated as nonsense for one gene but silent or intronic for a second, overlapping gene. This procedure avoids it being incorrectly counted as a nonsense mutation for both genes. Primary annotations were ordered as follows: nonsense, essential splice site (the conserved first and last two intronic bases in each intron), frameshift indel, start lost, read through, missense, in-frame/codon indel, splice (within 5 bases of an intron/exon junction), silent, intronic and intergenic.

For the purpose of gene-based analysis, we adopted three primary variant sets: 1) disruptive mutations (nonsense, frameshift and essential splice-site mutations), 2) disruptive mutations and missense variants rated as "damaging" by all of five prediction algorithms employed (PolyPhen2 HumDiv and HumVar, LRT, MutationTaster and SIFT), and 3) disruptive mutations and missense rated as damaging by at least one of the above metrics. We obtained prediction data from the dbNSFP database (Liu et al., 2011).

We calculated the correlation across the five missense metrics (based both on quantitative scores and binary predictions of being likely damaging or not). In general, correlations were modest, in the range of 0.2 – 0.5. We used sample allele frequency as an independent metric to evaluate the measures: more deleterious alleles should be less likely to be common due to negative selection pressure. Overall, 58% of all missenses were singletons in Sweden. Across the five measures, missenses rated as "damaging" displayed significantly lower frequencies, with 61-64% being singleton, ED Table 1c. By fitting a multiple logistic regression model of singleton status as a function of the measures, we noted that all five were significantly and independently correlated with the outcome, suggesting they contain independent information.

In the absence of a clear gold standard, we took two approaches to combine these five metrics. A strict definition of damaging required that all five metrics gave a consensus estimate that the allele is damaging (NS_{strict}); a broader definition required only that one of the five ranked the allele as damaging (NS_{broad}). The NS_{strict} definition yielded a set of SNVs with singleton rates even closer to that observed for nonsense mutations: 66% of NS_{strict} mutations were singletons, which was significantly higher than any of the individual metric rates. The substantially more inclusive NS_{broad} set had a singleton rate of 61%. For the primary burden analyses, we present analyses separately for each of the five algorithms (Figure 1 in main text).

5. Single-site and gene-based association analysis

Single-site association

We used Fisher's exact test as well as a standard chi-square test of independence, comparing allele counts between cases and controls, and using a permutation scheme (described below) to control for potential confounding effects of wave, sex and ancestry. Overall, we did not observe case/control differences in the total number of rare mutations (ED Table 1b). With one exception, single site associations observed at $P < 10^{-6}$ were either for common alleles (primarily tagging chromosome 6 major histocompatibility complex (MHC) locus variants already implicated by GWAS in this sample), or for four low frequency variants that, on inspection, appeared of low quality or exhibiting a plate-specific bias. The one exception is a C/T mutation at chr10:135053460, in the gene *VENTX*. This is a missense mutation rated as damaging by PolyPhen2 and SIFT, that appears in 18 cases and 0 controls, and appears to be of high quality, although not significant in the context of a exome-wide scan. However, this allele does show significant population stratification between the identified Finns and homogeneous Swedes in the current sample (6 of 18 carriers are Finnish, $P = 0.002$, odds ratio = 5.5). Although these data may possibly contain nominally significant single variants, we are not in a position to identify them within an exome-wide context.

The most significant *P*-value for a common GWAS variant was $P = 4.9\times10^{-8}$ for a C/A SNP at chr6:31112737. This variant is in the MHC region and tags a large number of SNPs, reproducing the established MHC schizophrenia association that we expected to observe (given the existing GWAS data). Four regions, including the MHC, had variants with $P < 10^{-6}$. Based on the exome data prior to genotype-level filtering, for low frequency variant (MAF<5%) we observed 5 variants with $P < 10^{-6}$. Three of these represented low quality calls: for example, a heterozygote call based on 9 reference and 1 alternate read) and were filtered out in the final genotype-filtered dataset. For another apparent low frequency association (0/28 cases/controls), on inspection it was clear that all carriers were from the same plate, indicating likely experimental artifact. Generally, none of the few striking single site associations involving rare variants appeared to be free from artifact.

Gene-based tests

We evaluated evidence for association at the genic level using two methods implemented in PLINK/Seq (http://atgu.mgh.harvard.edu/plinkseq/). The basic burden test counts the number of minor alleles per gene per individual, summed for all cases and for all controls; the test statistic is the standardized difference in case versus control mutation rate, evaluated by permutation in a one-sided manner (expecting a greater burden of rare mutations in cases compared to controls). This one-sided burden test is the primary test used for most single-gene and all gene-set analyses, for the reason that power will be lower to detect rare protective alleles for a common disease, particularly for a case/control design in which controls are essentially unselected for any trait other than not having a diagnosis of schizophrenia or bipolar disorder. Furthermore, prior work on the burden of copy-number variants (CNVs) and *de novo* mutation in studies of neuropsychiatric disease all point to a predominantly one-sided model, in which most rare alleles are deleterious and increase risk for disease. Note that we intend the word *burden* here to refer to population level (the total extent of alleles in cases) and do not imply that individual cases will necessarily carry multiple risk alleles.

Second, we applied the sequence kernel association test (*SKAT*; Wu et al., 2011) that allows for risk and protective effects in the same gene. We used asymptotic *P*-values and incorporated ancestry covariates (first four MDS components), based on a 5% allele frequency cut-off and the default frequency weights (1 and 25 as the two beta distribution shape parameters, to up-weight lower frequency mutations).

The permutation procedure swapped phenotype label within groups of individuals matched by identity-by-state (IBS) ancestry information as well as sex and experimental wave. We used complete linkage hierarchical cluster analysis as implemented in PLINK (Purcell et al., 2007) to form matched groups based on IBS distance calculated from common variants extracted from the exome genotype calls. Individuals with a pairwise population concordance (PPC) *P*-value < 0.01 were disqualified from being belonging to the same matched group; we further ensured that every group had at least one case and one control. Overall, we obtained 292 matched groups, with a median size of 15 (mean 17.3) individuals; 10 individuals were not matched and removed from the primary analyses. We alternatively implemented a 1:1 matching scheme (every case matched to the nearest control), which did not substantively change the overall profile of results (data not shown).

We additionally performed a series of checks for the main results (data not shown) including: comparing results generated by datasets without genotype-level filtering; restricting analysis only to SNVs that did not change in MAC between filtered and unfiltered datasets; restricting analyses to the homogeneous subset of the Swedish sample; excluding SNVs that were potential multinucleotide polymorphisms or more likely to be badly called (sites at which a second variant site was called within 2 bases). Other genic tests were also applied to these data also, e.g. the variable threshold (Price et al., 2010) and Madsen-Browning (2009) tests. Although results for individual genes fluctuate (particularly for genes with low overall case/control allele counts), none of these alternate analytic decisions appeared to have a noticeable substantive impact on the overall profile of results, particularly for the composite set results, i.e. for the conclusion of a highly polygenic burden primarily driven by disruptive singleton mutations (data not shown).

Considering the exome-wide distribution of *SKAT* and burden test results, we observed distributions consistent with a global null. For example, considering the 19,209 genes with at least one nonsynonymous mutation, we observed 4.92% of tests with a nominal *P*-value less than 0.05, that is, very close to the null expectation; 0.92% of genes tested had nominal *P*-values less than 0.01. Considering the smaller number of genes with one or more disruptive variants, 1.6% and 0.2% of genes had *P*-values less than 0.05 and 0.01 respectively; the deflation in test statistics represents low power and discreteness in the test statistic distribution for the many genes that contain only a single disruptive mutation, for example. Restricted to genes with 5 or more disruptive mutations, the rates were 4.2% and 1.1%, again close to the null expectations. We observed broadly similar results for the burden tests. These results suggest that the overall properties of study, in terms of case/control matching, are appropriate. We did not observe any excess of very small *P*-values ($<10^{-3}$, $<10^{-4}$) in the gene-based results: no gene was significant after correction for 20,000 independent tests.

Estimation of odds ratios

For genes and genesets with very low case and control counts, the standard estimates of effect size (odds ratio) can be biased. For such sets, we report corrected odds ratios in the main text, Tables and Figures, using penalized maximum likelihood logistic regression model (Firth's method, as implemented in the logistf R package). This approach yields point estimates that are numerically very similar to a standard continuity-corrected estimate (obtained by adding 0.5 to all cells, e.g. (Fleiss, 1981); we also report 95% confidence intervals for key genesets (e.g. ED Table 3) from Firth's method.

Winner's curse and the estimation of odds ratios.

Low power and multiple testing can in theory lead to inflated estimates of effect sizes. Here we give a simple, proof-of-principle simulation to quantify the likelihood and extent of any such inflation. Specifically, we simulated case and control counts for a rare (population minor allele frequency, MAF=0.0001) dominant allele, with genotypic relative risk (GRR) of 10.0, for a disease with prevalence 1%. Such a model is broadly consistent with the ARC disruptive variant result in our study. Considering a sample of 2,500 cases and 2,500 controls, we calculated Fisher's exact test for 100,000 random samples generated under this model. We then considered the distribution of the point estimate of the odds ratios as a function of the statistical significance of the result.

Given the MAF and prevalence, a GRR of 10.0 should correspond to a population odds ratio of ~11.0. Calculating the mean case and control counts across all 100,000 replicates and using these means to estimate the odds ratio, this is indeed what we observe (mean of ~5 case and ~0.45 control counts). In practice, for any one sample we must calculate continuity-corrected odds ratios, because control counts are likely to be 0. The mean continuity-corrected odds ratio across all replicates was 8.2, i.e. slightly conservative as noted above.

Because power is low in this scenario, the mean of the odds ratio in significant datasets is indeed significantly higher than the population value:

True effect: 10-fold increase in risk

Significance threshold	**Power**	**Mean estimated OR**
P < 0.05	28.4%	14.3
P < 0.01	8.9%	18.4
P < 0.001	0.9%	23.9
P < 0.0001	~0.01%	31.9

A second round of simulations with a smaller true effect makes a similar point, that it is in theory possible to observe an odds ratio between 10-20 that represents a true 5-fold effect on risk:

True effect: 5-fold increase in risk

Significance threshold	Power	Mean estimated OR
P < 0.05	2.7%	13.7
P < 0.01	0.3%	17.7
P < 0.001	~0%	24.1
P < 0.0001	0%	n/a

Although inflation can in theory be considerable in these cases (up to 3-fold in the first example), the probability of obtaining such an inflated result markedly drops (i.e. the power of the test for the particular significance threshold). In practice, under these models the extent of multiple testing would have to be very large to make it likely to achieve a highly significant result with such low power. (In our analysis, the actual extent of multiple testing is relatively low, as it implicit in the relatively modest correction factors comparing corrected to original significance values.) Extreme, order-of-magnitude inflation in an estimate is therefore unlikely, although some smaller degree of inflation is possible (as is true for almost all association studies).

KYNU disruptive gene-based association

We performed an additional series of analyses to follow-up the phenotypic and haplotypic properties of KYNU carriers. Considering all other ICD9 phenotypes available on these patients, we did not detect any shared or unusual features that clearly co-aggregated with either the single recurrent nonsense allele, or any carrier. We did confirm that carriers were not unusual with respect to ancestry (based on multidimensional scaling analysis of GWAS data) and that carriers were not related to each other (e.g. at the second or third cousin level). The recurrent nonsense mutation (7 cases) did segregate on a large haplotypic background, with a minimum shared haplotype around ~600kb (although we cannot estimate the size of this very accurately). The same haplotype is present at low frequency in other individuals, but is not associated with disease. These results suggest that, rather than being a recurrently mutating site, this nonsense mutation is a low-frequency variant, possibly private to Sweden, segregating on a specific (more common) haplotypic background.

6. Origins of genesets tested

For the **primary test of polygenic burden**, we selected the following sources, all from the recent schizophrenia literature:

- 611 genes with a nonsynonymous *de novo* mutation observed in Girard et al. (2011), Xu et al. (2012) or the parallel Fromer et al. study
- 345 genes spanned by one the schizophrenia-associated rare CNVs tabulated in Sullivan et al. (2012)
- 234 genes spanned by a *de novo* CNV in Kirov et al. (2012)
- 479 genes spanned by a GWAS linkage disequilibrium interval in the Swedish sample (Ripke et al., 2013), for a GWAS $P<10^{-4}$. Intervals were defined by the region spanning region other SNPs with $P<0.05$ and $R^2 > 0.1$ with the index SNP in that region. Because this definition includes the MHC region and other very large, genic intervals, we further excluded intervals containing 5 or more genes, on the assumption that the prior probability of any one gene being a risk gene decreases as the number of genes in the associated interval increases
- 26 genes belonging to the voltage-gated calcium ion channel gene group (all *CACN** RefSeq genes, as defined by the HUGI Gene Nomenclature Committee (HGNC) *CACN* gene-family)

- 446 genes predicted to be high confidence targets of miR-137 by TargetScan 6.2, using the same criteria as Ripke et al. (2011)
- 685 genes from the postsynaptic density (PSD) "human core" geneset used in Kirov et al. (2012), which includes ARC, NMDAR complex, PSD-95 and mGluR5 subsets

In total, this list comprises 2,546 genes that we used as the basis for the test of polygenic burden in the Swedish exome data.

We compiled a secondary list of genes based on **autism and intellectual disability (ID) genes:**

- 132 genes with nonsynonymous *de novo* mutations in ID (de Ligt et al., 2012; Rauch et al., 2012)
- 743 genes with nonsynonymous *de novo* mutations in autism (Issovifov et al., 2012; Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2012)
- 112 genes listed as autism candidates and 196 intellectual disability candidates in Betancur (2011); we considered both all genes, as well as restricting analyses to those with autosomal dominant inheritance (similar patterns of results obtained for both sets)
- 842 genes listed as targets of FMRP (fragile X mental retardation protein) by Darnell et al. (2011) and 939 FMRP targets from Ascano et al. (2012), because targets of FMRP are strongly enriched for *de novo* mutation in autism (Iossifov et al., 2012).

In total, this gives 2,507 candidates, referred to as the "autism/ID" geneset.

Third, to characterize the nature of the burden (with respect to allele frequency, mutation type, brain expression profile, etc) we create a ***composite set*** based on the analysis of the primary SCZ and secondary autism/ID sets. This comprises 1,796 genes: all SCZ *de novos,* PSD genes, calcium channel genes and Darnell FMRP targets.

(Where necessary we mapped genes to human RefSeq symbols from the appropriate, sometimes non-human, source meaning that small numbers of genes may be lost due to mapping inconsistencies or ambiguities.)

In follow-up analyses we tested the following more general genesets, and subsets of the primary sets, based on the following sources:

Ion channel genes: We selected the 237 ion channel genes as enumerated by Harmer et al. (2009) and used by Klassen et al. (2011) in an exome sequencing study of epilepsy.
microRNA targets: We obtained a list of high-confidence microRNA targets from TargetScan version 6.2, http://www.targetscan.org/, for all microRNAs. We used the same definition of confidence as PGC schizophrenia GWAS (Ripke et al., 2011) – the greater number of miR-137 targets (446 versus 300) is due to the more recent version of TargetScan.
Phenotypic subclassification of PSD genes: The classification of PSD genes in terms of associated mouse and human phenotypes was based on Bayés À et al. (2011). The classification based on subcellular localization was based on Kirov et al. (2012).

Gene Ontology (GO): We obtained (Feb 2013) the GOA-Human association files and Genotype Ontology OBO files from http://cvsweb.geneontology.org/, the GO CVS repository. We mapped all child/parent GO terms based on "is_a" and "part_of" relationships and assigned genes to sets based on all evidence codes.
Gene families: HUGO Gene Nomenclature Committee (HGNC) website: http://www.genenames.org/genefamily.html
OMIM disease genes: Based on Table S1 of Goh et al (2006) PNAS, kindly supplied by Marc Vidal & Michael Cusick.
Developmental brain expression trajectories: We classified genes as either prenatally or postnatally biased based on two datasets. See the subsection below for details.
Autism *de novo*/PPI network genes: Protein-protein interaction networks of autism candidates, 1) *CHD8* network (*CHD8*, *DYRK1A*, *GRIN2B*, *TBR1*, *PTEN* and *TBL1XR1*); 2) a 49-gene and 3) extended 74-gene network (O'Roak et al., 2012).
Müller et al. Cav2 genesets: We compiled from Müller et al. (2010) lists of genes in the following categories: Cav2 channel core (7); GPCRs (13); adaptors (20); cytomatrix (10); cytoskeleton (18); enzymes (9); extracellular matrix (4); ion channels transporters (44); ion channels transporters excluding Cav2 (37); kinases & phosphatases (20; modulators & small GTPases (21); protein trafficking (11); unknown function (30). The reported set in the main text is "ion channels transporters excluding Cav2" (which withstands correction for multiple testing across the above sets).

Summary of results from the expanded geneset enrichment analyses

Overall, the extended enrichment analyses (based on GO terms, gene families, OMIM disease genes, targets of microRNAs other than miR-137, and ion channel genes other than the calcium channel) yielded no statistically significant results after correction for multiple testing (data not shown). We considered only singleton disruptive mutations, as this was the set most clearly enriched in the primary, hypothesis-driven analysis.

Nonetheless, calcium channel genes ($P = 0.002$) ranked highest among all 497 gene-families and 18 ion-channel sets tested. Also, the GO term "calcium ion transmembrane transporter activity" (GO:0015085) ranked in the top five of 17,062 GO sets tested, all $P = 7\times10^{-4}$. Given the prior genetic findings, this is consistent with the involvement of calcium channels. We did not observe any enrichment of rare variants in the genes highlighted by Need et al. (2012) or the candidate genes sequenced by Crowley et al. (2013), the latter including *COMT*, *DISC1* and *DRD2*.

Comparisons of FMRP genesets

As noted, we used two definitions of FMRP targets from the recent literature, denoted here as the Darnell and Ascano sets. As others have noted (http://sfari.org/news-and-opinion/news/2013/hunt-for-targets-of-fragile-x-protein-proves-perplexing), there is surprisingly little overlap between Darnell and Ascano targets. Whereas the Darnell et al. list was generated from mouse brain, the Ascano et al. list was generated from cultured human embryonic kidney cells (HEK293) overexpressing FMRP. As shown in ED Table 4b, the two genesets differ in their extent of overlap with PSD genes also, in that the Darnell list is more strongly enriched for PSD genes, in particular ARC, NMDAR network, mGluR5 and PSD-95. The difference between Ascano and Darnell genesets is consistent with respect to a) PSD gene enrichment (ED Table 4b), b) case/control enrichment (Table 2), and c) schizophrenia GWAS enrichment (below). In all cases, the signal we see for Darnell genes is not present in the Ascano geneset.

We also considered whether the Darnell FMRP signal was statistically independent of the PSD enrichment in our data, and vice versa. Considering singleton disruptive mutations, we observed significant case enrichment for non-PSD FMRP targets (*P*=0.0016 from 240 genes with one more disruptive singleton) and from genes that were both PSD genes and FMRP targets (*P* = 0.03, 68 genes). In contrast, there was no enrichment for non-FMRP PSD genes (*P*=0.32 from 151 genes). However, we still observed enrichment for the ARC genes in cases compared to controls (*P* = 0.0098 from 5 genes), despite the small size of the ARC set, and the strong enrichment of FMRP targets in ARC as detailed in ED Table 4b. A similar pattern was found for MAF<0.1% and NS_{strict} mutations. Overall, we conclude that the ARC and FMRP enrichments reported here are not redundant with each other.

Defining developmental trajectories in human brain

We used two datasets to define whether a gene was "brain-expressed" and, if so, whether it showed biased (higher) expression or not, either prenatally or postnatally. The two datasets (based on partially overlapping samples) used were from the BrainSpan (http://developinghumanbrain.org) and Human Brain Transcriptome (http://hbatlas.org/; HBT; Johnson et al., 2009; Kang et al., 2011) studies.

We applied two approaches to the BrainSpan RNA-sequencing data, one based on cluster analysis and one based on pairwise statistical comparisons; both approaches used the genic RPKM (reads per kilobase per million) values as supplied by the study (Gencode v10 analysis, http://download.alleninstitute.org/brainspan/RNASeq_Gencode_v10/). We obtained from the authors of Xu et al (2012) their list of prenatally and postnatally biased genes, as based on the HBT dataset and described in Xu et al. We thank Bin Xu and Maria Karayiorgou for making their classification available to us.

The primary annotation (ED Table 10a, ED Figure 10b) is based on the 8 categories derived from unbiased clustering of BrainSpan data. We also provide results for an alternative "prenatal vs postnatal" classification of the BrainSpan data and the corresponding "prenatal vs postnatal" classification from Xu et al. (2012) based on HBT data. We used data from the same regions – dorsolateral prefrontal cortex (DLPFC) and hippocampus (HPC) – from both BrainSpan and HBT. Although these classifications are statistically very strongly related, there is a considerable degree of differential classification that likely reflects the different technologies and samples used between BrainSpan and HBT studies, as well as the decisions made in assigning labels to individual genes. Nonetheless, despite these differences, the main substantive conclusion (that is, of a predominantly *postnatal* bias in genes showing increased rates of rare mutations in schizophrenia in our dataset) holds over all three classifications.

Clustering of BrainSpan RNA-seq expression data (primary scheme): We applied cluster analysis to BrainSpan RNA-seq data to group genes based on their broad developmental trajectory in the developing human brain. We removed outlying individual samples with very high variance in RPKM values across genes (2 from DLPFC and 6 from HPC datasets). We Winsorized high RPKM values at 50 and transformed to log2(RPKM+1) scale. For each gene we calculated the median scaled expression values across three groups of samples: first trimester (T1, 0-12 weeks), second trimester (T2, 13-24 weeks) and all postnatal samples (the BrainSpan dataset contains only a single sample in the third trimester for the brain regions considered here). We performed this procedure separately for DLPFC and HPC, such that each gene was represented by a vector of 6 expression values (3 time points by 2 brain regions). The heatmap plot in ED Figure 10b is based on these matrices (rendered separately for DLPFC and HPC). We applied K-means clustering to the

6-dimensional gene expression data, to separate distinct developmental trajectories contrasting genes that show either prenatal or postnatal expression biases, genes that are unbiased (showing similar expression across time points) and genes that do not show any expression in brain. Based on inspection of the ratio of between-cluster to total-sample sum of squares and the number of genes assigned to each cluster, we selected a $K = 8$ class solution (that captures 93.6% of the total variance). Broadly similar values of K (e.g. 6 to 10) gave substantively similar results: genes were separated into at least four distinct clusters corresponding to 1) no, 2) unbiased, 3) prenatally-biased, or 4) postnatally-biased gene expression in postmortem brain. Solutions with $K > 8$ considered did not differentiate between first and second trimester; similarly, clustering based on 4 time points per gene/region (splitting postnatal samples into 0-10 years of age and >10 years of age) did not yield substantively different profiles.

Statistical pairwise determination of prenatal versus postnatal genes (BrainSpan): Using the same BrainSpan data, here we aimed simply to generate lists of genes with significantly different RPKM expression levels between prenatal (defined as from conception to the end of the second trimester) and postnatal samples. Specifically, comparing expression from post-mortem brain samples below versus above 6 months from conception, genes with a significant difference ($P < 0.01$ after a Bonferroni correction for multiple tests) in either brain region were rated as PRE or POST, for prenatally or postnatally biased in expression, depending on the direction of effect. In addition, all genes with RPKM values above 5 for either DLPFC and HPC were designated as highly expressed (HIGH); genes with RPKM values below 1 for both DLPFC and HPC were designated as not brain expressed (LOW). Further, for consistency we ensured that the same gene could not be labeled with more than one of PRE, POST or LOW expression (these cases were dropped). This yields 9378 HIGH genes, 6259 LOW, 8052 PRE and 2432 POST.

Statistical pairwise determination of prenatal versus postnatal genes (Xu et al. / HBT): The Supplementary Information in Xu et al. (2012) details the procedure used to obtain lists of prenatally and postnatally biased genes from the HBT microarray dataset. For comparability with Xu et al., we adopted the same strategy of keeping DLPFC and HPC lists separate. The procedure described above based on BrainSpan data yielded 7,872 prenatally biased genes versus 2,365 postnatally biased genes. In contrast, the Xu et al./HBT classification yielded 6,092 versus 7,647 based on DLPFC samples, and 6,260 versus 7,123 based on HPC. That is, the BrainSpan comparison points to significantly fewer postnatally enriched genes. Below we show the classification of prenatally and postnatally biased genes according to the two schemes. For example, 8,771 genes received *either* a BrainSpan (BS) prenatal or a HBT(PFC) prenatal label; of these, 59% were shared by both, 31% were specific to BS and 10% were specific to HBT:

BrainSpan	HBT (Xu et al.)	N genes	Intersection	BS-specific	HBT-specific
Prenatal	Prenatal (PFC)	8,771	59%	31%	10%
	Prenatal (HPC)	8,806	60%	29%	11%
	Postnatal (PFC)	14,722	5%	48%	47%
	Postnatal (HPC)	14,277	5%	50%	45%
Postnatal	Prenatal (PFC)	8,448	0%	28%	72%
	Prenatal (HPC)	8,592	0%	27%	72%
	Postnatal (PFC)	7,813	28%	2%	70%
	Postnatal (HPC)	7,459	27%	5%	68%

Summary of composite set enrichment analyses stratified by developmental brain expression trajectory.

As described above, we used RNA-sequencing expression data from the human dorsolateral prefrontal cortex and hippocampus (BrainSpan) and clustered genes by developmental trajectory profiles to identify those showing prenatal or postnatal expression biases. We also used an alternative statistical comparison to identify developmentally-biased genes as well as the Xu et al. (2012) classifications (from Human Brain Transcriptome microarray data). The strongest pattern from analyses of the composite set stratified by expression profile was that genes with *higher postnatal expression* showed consistently stronger enrichment in cases: this held across all classifications (ED Table 10a) and for MAF < 0.1% as well as singleton disruptive mutations (not shown).

The evidence for case enrichment in *prenatally biased* genes was at best ambiguous: using the Xu et al. classifications, there was no significant enrichment in prenatally biased genes under any condition (ED Table 10a). By contrast, in autism, disruptive *de novos* were clearly enriched in genes with a high prenatal expression bias (ED Figure 10b, based on an analysis using *DNENRICH* as described in the Fromer et al. study).

In contrast to findings in autism, our data point to stronger enrichment in genes showing higher expression postnatally, although this signal is not borne out in the *de novo* analysis of Fromer et al. (no bias detected) or Xu et al. (prenatal bias reported). It is possible that the developmental trajectory of particular disease genes could be a marker for within-case heterogeneity in schizophrenia, although more complete data on gene expression in human brain across development will be required to fully explore the relation between brain expression and risk. We also note that many postnatally-biased genes are still expressed during early development: as such, these results do not necessarily address the neurodevelopmental hypothesis of schizophrenia directly.

Infinium HumanExome BeadChip ("Exome Chip")

For Figure 1, we obtained (accessed Feb 2013) a site list of "exome array" variants from
ftp://share.sph.umich.edu/exomeChip/ProposedContent/codingContent/
for nonsynonymous, splice and stop mutations and excluded flagged sites as tabulated here:
ftp://share.sph.umich.edu/exomeChip/IlluminaDesigns/cautiousSites/cautiousSite.sorted.sites

Note that all analyses of these sets were based on genotype calls using the exome sequence data at these array sites. Of all nonsynonymous variant sites called in the exome data, ~30% of them are present on the exome array in total, although only ~10% of singleton sites.

7. Gene-set association analysis: rationale and approach

Rationale

In many ways, the starting point for robust discovery in schizophrenia genetics has been to establish statistical evidence for a genome-wide burden of particular class of variant, prior to identifying specific genes or alleles of that class as causal risk factors. While there are notable exceptions to this paradigm (for example, the 22q11.2 deletion), for both GWAS and

CNV studies it was important to first establish the genome-wide degree of polygenic burden and to reject the global null hypothesis that the class of variation in question was entirely unrelated to disease risk, e.g. International Schizophrenia Consortium (2008, 2009), both studies that included a subset of the Swedish sample. As well as simply motivating more specific analyses and further, larger studies, an analysis of polygenic burden can flesh out the genetic architecture and indicate the types of variant or gene more likely to yield to focused mapping efforts. This is similar to the approach adopted by several studies of *de novo* mutation (studies reported in the main text) and also rare recessive mutations (e.g. in autism, Lim et al., 2012), that concentrated first on overall rates of mutations before proceeding to map particular genes. In the current study, we aimed to follow an analogous path, to first robustly test the global null hypothesis that rare coding variation plays no role in the aetiology of schizophrenia: if a polygenic burden of rare coding variants can be established, we would then subsequently characterize its nature with respect to the genes and frequency and type of mutations involved. This can be conceptualized as a top-down strategy that stands in contrast to bottom-up approaches that necessarily start by identifying associated alleles and genes.

For rare or *de novo* CNVs and *de novo* SNVs, an individual's polygenic burden can be straightforwardly approximated, by counting the number of alleles they carry, assuming each is deleterious and increases risk. For a common variant, the direction of effect cannot be predicted and so it is first necessary to use an independent GWAS sample to assign which of the two alleles is (very marginally) more likely to increase risk, before counting the number of putative risk alleles per-individual in a second sample. For rare variants from exome sequencing in cases and controls, neither approach will be satisfactory. Compared to CNVs and *de novo* SNVs, the larger universe of mostly neutral, coding inherited SNVs means that we would not expect cases to carry a detectably higher rate of, for example, nonsense mutations averaged across all genes, and the rarity of individuals SNVs also precludes use of an independent sample to estimate effect sizes or prioritize individual alleles. We are however able to leverage the fact that schizophrenia genetics is no longer a blank slate, so to speak, by testing for a polygenic burden of rare, risk increasing alleles specifically in a large set of genes that have been directly or indirectly implicated by recent, large-scale genome-wide schizophrenia genetic studies. Genes directly implicated include those in associated linkage disequilibrium intervals from GWAS, or under specific CNV such as the 22q11.2 microdeletion. Genes indirectly implicated by previous schizophrenia genetic work include all calcium channel genes, for example, and members of the ARC and NMDAR PSD complexes.

Approach to testing for geneset enrichment

We implemented a geneset enrichment procedure (the SMP utility, "statistic/matrix/permutation", part of the PLINK/Seq package, http://atgu.mgh.harvard.edu/plinkseq/) to calculate enrichment statistics for large sets of genes in order to establish whether case-enrichment of rare variants is preferentially concentrated in a particular set of genes, controlling for any exome-wide/baseline difference in case and control rates. The procedure uses single-gene burden statistics and forms sums of these statistics, summing over all genes in a set; the significance of each sum-statistic is evaluated by permutation. Permutation occurs at the level of single-gene tests: the input to this procedure is a matrix of gene-based statistics that contains both observed and permuted single-gene statistics. In the case of 1-sided burden tests, as employed here, evaluating a sum of single-gene statistics is numerically equivalent to evaluating the burden of all variants in a set of genes, although the former is computationally faster. The same series of phenotype permutations is applied similarly to all genes, thereby preserving the correlational structure of tests if there is residual linkage disequilibrium among rare variants in nearby genes.

For all genes in a set, we sum the single-gene statistics (the standardized difference in case minus control rate of rare mutations). For genes in the test set that physically overlap, we conservatively take only a single gene from the overlapping set (to avoid double-counting the same alleles in overlapping genes), selecting the gene with the maximum test statistic, doing similarly for under each permuted set of single-gene statistics as well as for the original data. As well as the sum-statistic for each set, S_{SET}, we evaluate the exome-wide value (i.e. for the "set" that contains all genes), S_{EXOME}. The results quoted in the main text are for the relative enrichment statistic, S_{SET} / S_{EXOME}, with significance evaluated empirically based on the null distribution of this ratio (10,000 replications were used unless otherwise stated).

Note that the reported effect sizes from the geneset enrichment analysis are estimates of the unconditional odds ratio that, unlike the empirical significance values, do not take exome-wide differences in case/control rates into account (or the stratified permutation scheme used to implement the IBS matching).

Composite set enrichment was not driven by a small number of variants or genes.

Within each set, we additionally calculated two series of rank-based statistics (used in ED Figure 9a). Each gene is ranked based on the observed test statistic, from high to low. For the k^{th} gene in the ranked set, we calculate S_k, the sum of the highest ranked k=1,2,3,… genes; this is similarly performed for each null replicate to generate the empirical null distribution of S_k for each set, from which statistical significance values (described as "*P(best)*"). A second statistic ("*P(rest)*" test) can also be calculated, that relates to the significance of the set with the k top-ranked genes removed, indicating whether there is any signal left in the rest of the set after the top k have been accounted for. Unlike the *best* test, here the same top k genes, based on the observed data only, are removed from each null replicate (rather than using the implicit re-ranking within each replicate).

The *best* and *rest* tests are heuristic indicators intended to give a sense of what is driving a significant geneset association. Even for a moderately large set, a set-level association could in theory represent one or two genes of large effect. In this instance, the top few genes should have significant *best* statistics that decline with increasing k; after the top few genes, the *rest* test would not be expected to be significant, as a few genes have explained the entire signal. In contrast, if a significant set-level association instead represents an accumulation of small effects over many genes, k might have to be quite large before the *best* test achieves significance; similarly, the *rest* test may still be significant at a large k, that is, even after discounting the most associated genes.

The trend depicted in ED Figure 9a is consistent with this second, more highly polygenic scenario. Specifically, the horizontal axis shows all genes in the composite set ranked left-to-right by their degree of case-enrichment; the vertical axis represents empirical significance for enrichment of the best k genes, where k = 1, 2, 3, …. The continued upward trend for the composite set (black line) indicates that genes with counts of 1/0 in cases/controls contributed to the signal, rather than the burden reflecting only a subset of more highly enriched genes. For comparison, we artificially constructed a random geneset of similar size and overall enrichment to the composite set, but seeded with the top 25 genes from the individual disruptive gene-based burden tests. Even though 17 of the 25 had only modest P-values ($0.01 < P < 0.05$), the profile in ED Figure 9a is markedly different (orange line) and the best k genes collectively achieved high significance at relatively low values of k (e.g. 10 – 20 genes in this case). If composite set enrichment had a similar architecture, that

represents a relatively concentrated burden across a small number of genes, we would expect a similar profile, but this is not what we observed (black line).

Approach to multiple testing

Correction for multiple testing was performed empirically. In any given comparison (enumerated below), we jointly corrected for all filters (e.g. singletons versus MAF < 0.1% versus MAF < 0.5% alleles; disruptive versus strict versus broad definitions of damaging mutations) and all sets (e.g. the 12 constituent genesets of the primary candidate geneset). For each null permuted version of the data we calculated the minimum empirical significance that would have been obtained, i.e. as if that permuted dataset were in fact the observed one. We then compared the observed empirical value for a given test against the distribution of minimum empirical *P*-values, across all tests and sets considered. This preserves the correlational structure between tests as well as the variable information content across genes and genesets. As such, this procedure (sometimes called "minP" or "maxT") will preserve family-wise error rates without over-correction.

Beyond the single site and gene-based tests (for which we use standard genome-wide threshold, $P < 5 \times 10^{-8}$, or Bonferroni correct based on the number of genes tested), here we enumerate the total extent of tests performed and the level of correction for the geneset analyses: each of the eight lines below is corrected empirically to maintain the family-wise type I error rate across all the tests indicated for the group:

a) Primary geneset (3 MAF bins × 3 variant classes = 9 tests, Table 1)
b) ➔12 subsets, based on omnibus tests, disruptive variants only (12 × 3) = 36 tests
c) Secondary geneset (3 × 3 = 9 tests, Table 2)
d) ➔12 subsets, all variant classes given no significant omnibus test post correction (12 × 3 × 3) = 108 tests
e) All PSD genesets follow-up analyses (17 × 3 × 3 = 153 tests, ED Table 4a)
f) Schizophrenia *de novo* gene follow-up analysis (6 × 3 × 3 = 54 tests, ED Table 7a)
g) Known CNV loci (4 omnibus tests, ED Table 5 upper panel)
h) ➔ Known CNV loci (11 × 4 = 44 tests, ED Table 5 middle panel)

We also present a number of exploratory/descriptive analyses in which we only report uncorrected *P*-values.

a) Intersection with brain expression profiles (ED Table 10a)
b) Stratification of variants by known/novel status and the "case-unique" tests (ED Table 6b)
c) Breakdown of the composite enrichment by type of mutation (Figure 1), damaging missenses only (ED Table 6a) and best *k* genes (ED Figure 9a)

That is, after correction, the substantive core results presented in the main manuscript have an effective testing burden corresponding to only 8 hypothesis tests (a-h). We typically report both corrected and uncorrected *P*-values in the text/tables, because both are informative.

Alternative filtering strategy: case-unique disruptive mutations/genes.

Many of the main enrichment analyses we employed were restricted to alleles that were singletons in our sample, on the premise that this class would be enriched for ultra-rare alleles, and that the alleles with a large effect are likely to be ultra-rare. Nonetheless, a disruptive singleton burden will count a 1/0 variant but ignore a 0/2 doubleton that may exist in the same gene. To the extent that one assumes complete functional equivalence for all disruptive mutations in the same gene, such an accounting is not optimal. On the other hand, different disruptive mutations may have variable effects, including "partial" loss-of-function mutations that do not impact transcripts relevant to a tissue or disease, or those with position-dependent gain-of-function effects. In the absence of gold-standard transcript- and phenotype-specific annotation to guide analysis, we cannot necessarily differentiate these scenarios. To complement the primary singleton analyses, therefore, we also employed a parallel analysis strategy that avoids filtering individual variants based on sample frequency (and also, therefore, avoids the problem alluded to above). In this "case-unique gene-burden", we conservatively exclude genes that have *any* disruptive variants in controls; the test statistic is then simply the number of case alleles observed in the remaining genes. That is, no explicit frequency filter is applied. We used permutation to derive significance for the case burden empirically (i.e. re-filtering genes in each replicate after label-swapping phenotypes among individuals). As shown in ED Table 6b, the composite test remains significant ($P = 0.0006$), as do all of the previously flagged subsets. (We applied this alternate analysis only as a proof-of-principle, to address the above concern over singleton counting. Naturally, as control samples grow larger, filtering genes based on the presence of *any* disruptive variant would become unduly conservative, screening out the majority of genes.) In prioritizing individual genes from singleton burden results, consideration of non-singleton control (and other case) alleles is certainly important: for example, as listed in ED Table 3, for example, calcium channel genes *CACNA1S* and *CACNA2D4* have singletons only in cases but recurrently observed disruptive alleles in controls, and so, all other things being equal, they may be less attractive candidates compared to other calcium channel genes.

8. Linear models of burden: ancestry, sex effects and joint analysis with GWAS/CNV data

We fit a multiple logistic regression model for the likelihood of having one or more singleton disruptive mutations in the *composite* geneset as a function of case/control status, sex, experimental wave and ancestry (an indicator variable for being of Finnish ancestry, four multidimensional scaling components estimated from common variants in the exome data, and indicator variables for county of residence within Sweden). At the nominal $P < 0.05$ threshold, only a single predictor was significant, namely that being a schizophrenia case significantly increased risk ($P = 0.00012$), consistent with the analyses in the main text. This suggests that these confounders (sex, experimental wave or ancestry) do not drive the primary enrichment observed.

Geneset analysis of ancestry

We further explored the potential role of ancestry differences (and in particular the Finnish versus Swedish group difference) in the context of the geneset enrichment procedure used for the case/control analysis. As noted in ED Table 1a, the sample contains 413 individuals of Finnish ancestry (reflecting migration from Finland to Sweden in the last century). Using the composite geneset, we performed two sets of one-sided analyses (Swedes vs Finns and Finns vs

Swedes) to see whether either group had a statistically greater rate of rare mutations, controlling for the exome-wide average. These analyses controlled for sex and phenotype (case/control status) by permuting ancestry label only within those groups.

We observed higher rates of all rare variants exome-wide, including singletons, in the Swede versus Finn comparison (*P*=0.0001). Given we expect population differences to exist across many sites between these two populations, this largely reflects the fact that we have used sample frequency to select "rare" variants (singletons and MAF<0.1%) and have sampled more deeply in the Swedish rare allele spectrum. But, importantly, for disruptive mutations (where we see the primary schizophrenia association), none of the *P*-values for ancestry differences in the composite set (relative to the exome-wide baseline) are significant. For sets of nonsynonymous mutations, we do observe a greater-than-baseline rate of MAF<0.1% (and to a lesser extent, for singletons) in Swedes compared to Finns, i.e. $P = 0.006$ for more NS_{strict} mutations in the Swedes versus Finns for the *composite* set, relative to the overall, exome-wide differences in rare NS_{strict} mutations.

Taken together, these results suggest that ancestry is unlikely to be a confounder in our primary analyses. Firstly, we explicitly controlled for ancestry by IBS-matching, whereas the results here reflect what *uncontrolled* bias might arise. Secondly, irrespective of the factors that underlie the relative increase of composite set mutations in Swedes, the trend is in the opposite direction to the schizophrenia-association, in two senses: 1) the effect is stronger for the more common mutations in the larger variant groups (i.e. $NS_{broad} > NS_{strict} >$ disruptive) whereas for schizophrenia we observed the opposite pattern of (i.e. disruptive > $NS_{strict} > NS_{broad}$); 2) Swedes actually have a *lower* proportion of cases compared to Finns and so any stratification effect here would be working *against* the schizophrenia association signal. Why Swedes have an enriched relative rate of mutations in composite set genes relative to Finns could represent chance, uncontrolled for technical biases, or other unmodeled population-genetic factors beyond the scope of the current analysis.

Sex differences in the burden of composite set mutations

Overall, although we see a trend towards greater case enrichment in females (data not shown), this is not formally significant; in particular, for the class of singleton disruptive mutations, the class that shows the greatest overall case enrichment, we do not see any evidence of sex differences. We fit a logistic regression model to explore the relationship between the burden of disruptive mutations in the *composite* set, sex and other covariates. Fitting the model:

logit(SCZ) = BURDEN × SEX + MDS1 + MDS2 + MDS3 + MDS4 + FINN + COUNTY + WAVE + e

where **BURDEN** is encoded 0/1 for carrying 1 or more disruptive mutations, the formal test of interaction was not significant at the 0.05 level for MAF < 0.1% disruptive mutations ($P = 0.062$), or for disruptive singletons ($P = 0.5$). Restricting analysis to the autosomes yielded very similar results ($P = 0.08$ and 0.6). There are a smaller number of female cases (N = 1,016 females versus 1,520 males, see ED Table 1a).

Age at earliest admission

As a rough proxy for age-at-onset, we analyzed age at first admission for schizophrenia. This measure is distinct from a true population-based estimate of age-at-onset: the HDR only collected psychiatric admission data from 1973 and so will be left-censored for many patients (the mean year of birth for study cases is 1957). Additionally, this study initially did not recruit individuals younger than 18 years of age, although individuals with two admissions prior to that (i.e. adolescent age-of-onset) could have been included. Nonetheless, females had a later age-at-first-admission in the HDR also (33.5 years of age for females versus 30.1 for males, $P = 2\times10^{-11}$).

We regressed as the dependent variable whether each case carried one or more disruptive mutations in the composite set (considering separately both singletons and MAF < 0.1% alleles):

BURDEN = AGE-ADMISSION + SEX + MDS1 + MDS2 + MDS3 + MDS4 + FINN + COUNTY + WAVE + e

and sex-stratified analyses

BURDEN = AGE-ADMISSION + MDS1 + MDS2 + MDS3 + MDS4 + FINN + COUNTY + WAVE + e

In cases, age-at-first-admission was not related to composite set burden, either in all individuals ($P = 0.42$) or females ($P = 0.93$) or males ($P = 0.30$) separately (results for MAF < 0.1% mutations). We observed similar results for singletons ($P = 0.80$, 0.78 and 0.59 respectively).

Joint analysis with GWAS/CNV data

GWAS: To calculate common polygene risk scores for all Swedish samples in the exome dataset (using the available GWAS data on these samples), we obtained summary statistics from the Psychiatric Genomics Consortium schizophrenia GWAS meta-analysis (Ripke et al, 2011) but with all Swedish samples removed. Based on 1000 Genomes imputed results, we clumped based on linkage disequilibrium to extract sets of quasi-independent SNPs in decreasing order of statistical significance, requiring MAF > 2%, imputation $R^2 > 0.9$, removing SNPs within 500kb and LD $R^2 > 0.25$ with a previously selected SNP. Only a single SNP was retained to represent the MHC association. Based on *P*-value thresholds of 0.01, 0.05, 0.1, 0.2 and 0.3 in the discovery (PGC1) sample, we calculated scores for all Swedish individuals in the current study following previously established procedures (International Schizophrenia Consortium, 2009). All five scores were highly significantly enriched in cases ($P<10^{-16}$) with *Z*-statistics of 11.52, 13.36, 14.05, 14.30 and 14.33. We selected the $P<0.2$ score for all subsequent analyses.

CNV: We observed that in this sample genic deletions were most strongly enriched in cases compared to controls (empirical $P = 0.0004$, calculated using PLINK following approach in International Schizophrenia Consortium (2008). We scored individuals as having one or more genic deletion, genome-wide, to create a binary variable for use in subsequent analyses. The mean (median) number of genic deletions carried was 0.17 (0.0) per individual.

EXOME: We extracted the number of rare disruptive mutations in the composite set for all individuals in the current study. We repeated this for MAF<0.1% as well as singleton mutations. We scored individuals as having one or more mutation in this set to create a binary variable for use in subsequent analyses.

We fit the following logistic regression model:

logit(SCZ) = GWAS + CNV + EXOME + MDS1 + MDS2 + MDS3 + MDS4 + SEX + WAVE + GWASGENO + e

The three measures were independently associated with disease risk (GWAS $P < 10^{-16}$; CNV $P = 0.008$; EXOME $P = 5\times10^{-5}$) although the GWAS predictor accounted for an order-of-magnitude more variation as measured by the difference in Nagelkerke R^2 comparing a full logistic regression model versus a reduced model with that term removed (similar results obtained for MAF<0.1% and singleton mutations). The GWAS predictor explained $R^2 = 0.057$ ($P<10^{-16}$); the CNV predictor explained $R^2 = 0.002$ ($P<0.008$); the composite set exome predictor explained $R^2 = 0.004$ ($P<0.00002$).

Because the absolute CNV measure is skewed, we present results in terms of the binarized versisons. We obtained similar results if we scored the CNV and EXOME predictors as the absolute number of rare alleles carried, rather than an indicator for carrying one or more (P=0.01 for CNV and P=0.0001 for EXOME).

As noted, above, neither the burden of disruptive MAF<0.1% nor singleton mutations was significantly related to the first 4 MDS components, county of residence or Finnish ancestry, or to experimental wave ($P > 0.05$ all contrasts).

There was no significant correlation between the exome burden and GWAS polygene ($r = -0.01$, $P = 0.49$) or CNV burden ($r = 0.02$, $P = 0.17$). Further, CNV and GWAS scores were themselves uncorrelated ($r = 0.12$, $P = 0.18$). The previous statistics are for the entire sample; within cases only, the three measures were also uncorrelated ($P > 0.05$ all comparisons). There was also no evidence for non-additive interaction between the three measures (CNV × EXOME $P = 0.41$; CNV × GWAS $P = 0.53$; EXOME × GWAS $P = 0.44$; CNV × GWAS × EXOME $P = 0.48$).

9. GWAS enrichment analysis

We used INRICH (Lee et al., 2012) to test whether genomic intervals representing nominally associated GWAS loci were enriched for certain classes of gene. Based on 1000 Genomes imputed Swedish data, we defined linkage disequilibrium (LD) intervals around index SNPs with $P < 10^{-5}$ and $P < 10^{-6}$ to include all SNPs with $P < 0.05$ in $R^2 > 0.1$, within 500kb. Conservatively, any interval spanning the MHC region (broadly defined as 25-35Mb, hg19) was removed due to the extensive LD in this region and high gene count. The Swedish GWAS sample is described in Ripke et al. (2013), and comprises 5,001 cases and 6,243 controls.

Considering GWAS intervals with an index SNP of $P<10^{-6}$ (similar results obtained for a $P<10^{-5}$ threshold), we observed significant enrichment of GWAS hits for composite set genes (P=0.0006), Darnell FMRP targets (P=0.0004) and voltage-gated calcium channel genes (P=0.02). We did not observe GWAS enrichment for PSD genes (P=0.32) or schizophrenia *de novo* genes (P=0.23) individually however.

10. Methods References

Betancur C (2011) Etiological heterogeneity in autism spectrum disorders: More than 100 genetic and genomic disorders and still counting. Brain Research, 1380, 42-77.

BrainSpan: Atlas of the Developing Human Brain [Internet]. Funded by ARRA Awards 1RC2MH089921-01, 1RC2MH090047-01, and 1RC2MH089929-01. © 2011. Available from: http://developinghumanbrain.org.

Dalman, C., Broms, J., Cullberg, J. & Allebeck, P. Young cases of schizophrenia identified in a national inpatient register--are the diagnoses valid? Social Psychiatry and Psychiatric Epidemiology 37, 527-31 (2002).

Ekholm, B. et al.. Evaluation of diagnostic procedures in Swedish patients with schizophrenia and related psychoses. Nordic Journal of Psychiatry 59, 457-64 (2005).

Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA (URL: http://evs.gs.washington.edu/EVS/) [accessed 5/2013].

Fleiss, J. L. (1981) Statistical methods for rates and proportions. 2nd ed. (New York: John Wiley).

Harmar, A.J., Hills, R.A., Rosser, E.M., Jones, M., Buneman, O.P., Dunbar, D.R., Greenhill, S.D., Hale, V.A., Sharman, J.L., Bonner, T.I., et al.. (2009). IUPHAR- DB: the IUPHAR database of G protein-coupled receptors and ion channels. Nucleic Acids Res. 37(Database issue), D680–D685.

International Schizophrenia Consortium (2008) Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature. 455(7210):237-41.

International Schizophrenia Consortium (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 460(7256):748-52.

Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M, Sousa AMM, Pletikos M, Meyer KA, Sedmak G, Guennel T, Shin Y, Johnson MB, Krsnik Z, Mayer S, Fertuzinhos S, Umlauf S, Lisgo SN, Vortmeyer A, Weinberger DR, Mane S, Hyde TM, Huttner A, Reimers M, Kleinman JE & Sestan N (2011) Spatio-temporal transcriptome of the human brain. Nature 478, 483-489.

Johnson MB, Imamura Kawasawa Y, Mason CE, Krsnik Z, Coppola G, Bogdanovic D, Geschwind DH, Mane SM, State MW, Sestan N (2009) Functional and evolutionary insights into human brain development through global transcriptome analysis. Neuron 62, 494-509.

Klassen T, Davis C, Goldman A, Burgess D, Chen T, Wheeler D, McPherson J, Bourquin T, Lewis L, Villasana D, Morgan M, Muzny D, Gibbs R, Noebels J. (2011) Exome sequencing of ion channel genes reveals complex profiles confounding personal risk assessment in epilepsy. Cell. 145(7):1036-48.

Kristjansson, E., Allebeck, P. & Wistedt, B. Validity of the diagnosis of schizophrenia in a psychiatric inpatient register. Nordisk Psykiatrik Tidsskrift 41, 229-34 (1987).

Lee PH, O'Dushlaine C, Thomas B, Purcell SM. (2012) INRICH: interval-based enrichment analysis for genome-wide association studies. Bioinformatics. 28(13):1797-9.

Li H, Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 25(14):1754-60.

Lichtenstein P, Björk C, Hultman CM, Scolnick E, Sklar P, Sullivan PF. (2006) Recurrence risks for schizophrenia in a Swedish national cohort. Psychol Med. 36(10):1417-25.

Lim ET, Raychaudhuri S, Sanders SJ, Stevens C, Sabo A, MacArthur DG, Neale BM, Kirby A, Ruderfer DM, Fromer M, Lek M, Liu L, Flannick J, Ripke S, Nagaswamy U, Muzny D, Reid JG, Hawes A, Newsham I, Wu Y, Lewis L, Dinh H, Gross S, Wang LS, Lin CF, Valladares O, Gabriel SB, dePristo M, Altshuler DM, Purcell SM; NHLBI Exome Sequencing Project, State MW, Boerwinkle E, Buxbaum JD, Cook EH, Gibbs RA, Schellenberg GD, Sutcliffe JS, Devlin B, Roeder K, Daly MJ. (2013) Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders. Neuron. 23;77(2):235-42.

Liu X, Jian X & Boerwinkle E (2011) dbNSFP: a lightweight database of human non-synonymous SNPs and their functional predictions. Human Mutation, 32:894-899.

Madsen BE, Browning SR. (2009) A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009 Feb;5(2):e1000384.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20(9):1297-303.

Meynert AM, Bicknell LS, Hurles ME, Jackson AP, Taylor MS. (2013) Quantifying single nucleotide variant detection sensitivity in exome sequencing. BMC Bioinformatics. 14:195.

Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR. (2010) Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet. 86(6):832-8.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics, 81(3):559-75.

Statistics Sweden. Multi-Generation Register 2002: A description of contents and quality. (Statistics Sweden, Örebro, Sweden, 2003).

Thorvaldsdóttir H, Robinson JT, Mesirov JP (2012) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics.

World Health Organization. International Classification of Diseases, (World Health Organization, Geneva, 1992).