Supplementary Information

A Scan-Statistic Based Analysis of Exome Sequencing Data Identifies *FAN1* at 15q13.3 as a Susceptibility Gene for Schizophrenia and Autism

Iuliana Ionita-Laza[1,*], Bin Xu[2], Vlad Makarov[1], Joseph D. Buxbaum[3], J. Louw Roos[4], Joseph A. Gogos[5,6], Maria Karayiorgou[2,*]

[1] Department of Biostatistics, Columbia University, New York, NY USA

[2] Department of Psychiatry, Columbia University, New York, New York, USA

[3] Departments of Psychiatry, Neuroscience and Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, NY USA

[4] Weskoppies Hospital, Pretoria RSA

[5] Department of Neuroscience, Columbia University, New York, New York, USA

[6] Department of Physiology & Cellular Biophysics, Columbia University, New York, New York, USA

[*] Corresponding authors: ii2135@columbia.edu, mk2758@columbia.edu

Scan-statistic to identify clusters of rare disease risk variants with a trio design

Let us assume that N trios with affected offspring have been sequenced in a relatively large region G, such as a CNV or linkage region, or even a large gene. Furthermore let us assume that there are M rare variant positions in the region of interest. We define a rare variant as a variant with MAF in parents less than a fixed threshold (e.g. 0.01). At each position i with $1 \leq i \leq M$ let n_i be the number of minor alleles in heterozygous parents, with y_i of them being transmitted. Then y_i is Binomial(n_i, p_i), where p_i is related to the relative risk at position i (if there is no association between variant i and disease $p_i = 0.5$ according to Mendel's laws; however, the transmission probability in any given region can deviate from 0.5 due, for example, to certain biological mechanisms[1]). In a case-control setting[2], n_i is the number of carriers (both cases and controls) of minor allele at position i, and y_i is the number of cases among them. Then y_i is Binomial(n_i, p_i), where p_i is related to the relative risk at position i; if there is no association between variant i and disease, then p_i is simply the proportion of cases in the dataset (e.g. it is 0.5 if the number of cases is the same as the number of controls). Most of the derivations below for the trio design are similar to the case-control setting described in detail in Ionita-Laza et al.[2].

We are working under the assumption that there is a window W_{dis} such that $p_i = p_{W_{\text{dis}}}$ for $i \in W_{\text{dis}}$ and $p_i = p_0$ for $i \notin W_{\text{dis}}$. W_{dis} is the window that contains a cluster of disease associated variants.

Under the null hypothesis (that disease associated variants do not cluster) $p_{W_{\text{dis}}} = p_0$, while the alternative hypothesis is that $p_{W_{\text{dis}}} > p_0$. Note that the null hypothesis of no clustering does not necessarily imply that there is no association at any of the variants. For a fixed window size we employ a sliding window approach and calculate a likelihood ratio (LR) statistic. To calculate the likelihood ratio statistic we first condition on the window W and calculate for each window W of fixed size w the following LR score:

$$LR_W = \begin{cases} \left(\frac{\widehat{p_W}}{\widehat{r_G}}\right)^{y_W} \left(\frac{1-\widehat{p_W}}{1-\widehat{r_G}}\right)^{n_W - y_W} \left(\frac{\widehat{q_W}}{\widehat{r_G}}\right)^{y_G - y_W} \left(\frac{1-\widehat{q_W}}{1-\widehat{r_G}}\right)^{n_G - n_W - (y_G - y_W)} & \text{if } \widehat{p_W} > \widehat{q_W} \\ 1 & \text{otherwise} \end{cases}$$

where

$$\begin{aligned} y_W = \sum_{i \in W} y_i \quad \text{and} \quad y_G = \sum_{i \in G} y_i \\ n_W = \sum_{i \in W} n_i \quad \text{and} \quad n_G = \sum_{i \in G} n_i, \\ \widehat{p_W} = \frac{y_W}{n_W} \quad \text{and} \quad \widehat{q_W} = \frac{y_G - y_W}{n_G - n_W}, \ \widehat{r_G} = \frac{y_G}{n_G}. \end{aligned}$$

Note that $\widehat{p_W}$ and $\widehat{q_W}$ are the maximum likelihood estimators (MLEs) under the alternative hypothesis, while $\widehat{r_G}$ is the MLE under the null hypothesis. A pseudocount of 1 is added when estimating the proportions p_W, q_W and r_G.

The likelihood ratio statistic for window size w is then computed as: $\Lambda_w = \max_{|W|=w} LR_W$. The window W with the highest value for the LR_W is the most likely region to harbor a cluster of disease risk variants.

Statistical significance

We use permutations to assess the statistical significance of the Λ_w statistic. Since our data consists of trios with affected offspring (i.e. no variation in the offspring phenotype), we can only permute the transmitted and untransmitted haplotypes from the parents. Note that this permutation procedure generates the null hypothesis of no association, i.e. $p_{W_{\text{dis}}} = p_0 = 0.5$, which is stronger than the null hypothesis stated above (namely that $p_{W_{\text{dis}}} = p_0$). However, since we are usually concerned with detection of such clusters in large genomic regions (such as CNVs that are several megabases long) where we expect $p_0 \approx 0.5$, this permutation procedure is approximately correct, as we also show below using simulations.

Simulation Results

Type 1 Error

No Association We first simulate trios assuming there is no association (hence no clustering) at any of the variants, using the simulation software COSI[3]. We simulate $n \in \{100,\ 300,\ 500\}$ trios and sequence data in a $\{50\text{kb},\ 100\text{kb}\}$ genetic region. We use a threshold of 0.01 or 0.05 for the rare variants to be included in the analyses. We use a sliding window size w of $\{10\text{kb},\ 20\text{kb},\ 30\text{kb}\}$. We calculate the p value for the entire region (hence adjusted for the multiple testing in the sliding window procedure), and results are shown in Table S11 for $\alpha = \{0.05,\ 0.01, 0.001\}$. As shown, the empirical Type 1 error is well maintained at the specified significance levels.

Association, but no clustering We also simulate data assuming there is disease association with variants in the region, but that the associated variants are distributed throughout the region (do not cluster in any particular window). This simulation is meant to illustrate that even if there is disease association with some variants in the region, in the absence of clustering, the proposed test does not have good power to detect such a scenario. We show results in Figure S5. As shown, for a large genomic region (e.g. 100 kb or larger), the power is essentially the same as the Type 1 error. This agrees well with the theoretical expectation, namely for a large region we expect p_0 to be ≈ 0.5. However, for smaller regions (e.g. 50 kb or so) there may be increased power to identify "clusters", even if no intentional cluster has been simulated. Even though, the power when there is a true cluster is much higher than when there is no or only weak clustering.

Power and precision to identify clusters in the data

To evaluate the power of the approach we assume the entire region is between 500 kb and 3 Mb. We build our simulations on real exome-sequencing data for 231 SCZ trios. First we randomly select 1000 genetic regions for each length category, $L = \{500\text{ Kb},\ 1.5\text{ Mb},\ 3\text{ Mb}\}$, such that each region contains at least one gene with length between 30 and 60 kb, and such that the gene contains a minimum of 50 variants. From each region we select such a gene as the designated disease gene in the region, and assume it contains all the disease mutations. The number of disease mutations was taken to be $n_{dsv} = 0.1\times$the number of trios (if each disease mutation is unique to a trio, this corresponds to having 10% of trios as being carriers of disease mutation).

In a first disease model, we assume each disease mutation is unique to a trio and has been transmitted from one of the parents to the offspring. We call this model *The Unique Mutation Model*. These mutations are artificially replacing the original sequencing data at n_{dsv} positions in the designated disease gene. This type of disease model resembles what we have observed in the sequencing data for *FAN1*. In a second disease model, we assume a multiple regression model:

$$\text{logit}[P(Y = 1)] = \alpha_0 + \mathbf{X}'\boldsymbol{\beta},$$

where Y is the affection status for the child (we only include children with $Y = 1$), $\mathbf{X}$ is the vector of genotypes at the n_{dsv} positions; and the effect size at each of the n_{dsv} disease variants is determined by its MAF: $\beta = 0.8|\log 10(MAF)|$. We call this model *The Regression Model*. The real sequencing data is replaced by simulated sequencing data for 231 trios (using the software COSI) for these disease variants only. We investigate two cases: (1) strong cluster, where the n_{dsv} disease variants are consecutive positions, hence they cluster closely together, and (2) weaker cluster, where the n_{dsv} disease variants are interspersed among non-disease variants in an alternating fashion.

We estimate power at the $\alpha = 0.05$ level based on the 1000 genetic regions of a given length. For the cluster test we evaluate the power to identify a significant cluster anywhere in the larger genetic region of length L. We also estimate the overlap between the true (simulated) disease cluster and the estimated cluster, using the Jaccard measure of overlap. This measure is between 0 and 1 with 0 meaning no overlap, and 1 meaning complete overlap. We also estimate the power of the conventional sequence-based association tests

for family designs[4], Burden and SKAT, when applied to the designated disease gene and after Bonferroni adjustment for the total number of genes in the region. Hence both the cluster test and the conventional Burden and SKAT tests are region-based tests, and power is directly comparable across these tests.

We show power results in Figure S6 for a strong cluster. The median size of a strong cluster in our simulations is $\sim 13 - 14$ kb, and, as expected, power is highest for a 10 kb window size. In the case of strong clustering of rare disease variants, we also show that the cluster test has substantially more power than conventional gene-based tests, such as Burden and SKAT, since the cluster test makes use of the strong clustering of disease mutations in a small genetic region. In fact, for the model with unique disease mutations, the SKAT test has very little power. However, for the regression model, that includes recurrent disease mutations, SKAT's performance improves and its power is closer to that of the Burden test.

The cluster test starts to lose its advantage as the clustering of disease variants becomes weaker. In Figure S7 we show results for the weaker cluster simulation (median size of the cluster is now $\sim 24 - 26$ kb), and, as can be seen, the power advantage of the cluster test over conventional gene-based tests diminishes. In fact, as the size of the cluster increases, the power of the cluster tests will continue to deteriorate.

We also assess the precision of identifying the true simulated cluster. In Tables S12 and S13 we report the Jacard measure of overlap (size of intersection divided by size of union for two intervals). As expected, the highest overlap is achieved when the scanning window size w is close to the true size of the underlying cluster.

Parent-of-origin effect

We have also performed a parent-of-origin effect analysis. In particular, we restricted analysis to transmissions from heterozygous mothers and from heterozygous fathers, and performed Burden and SKAT tests on these data. Since the variants considered are rare (MAF < 0.05), there is little chance for both parents to be heterozygous in the same family (to be sure, we exclude such trios, if they exist, from the analyses). Therefore, inferring the parent-of-origin for each variant is straightforward.

Mendelian inheritance

To ensure that only high quality variants are included in the analyses, we have removed variants with mendel error rate greater than 0.01. The mendel error rate for a variant is calculated as the total number of mendel errors divided by the number of trios. For all the remaining mendel errors, the genotypes in those implicated trios are set to 0. For the 15q13.3 region, we have looked at the mendel error rates for variants within any of the 7 known genes, plus the 20 kb window within *FAN1* (Figure S4). As shown, only two of 48 variants within *FAN1* have a small, non-zero mendel error.

Network analysis

We have investigated whether other genes that interact with *FAN1*, directly or indirectly, are collectively enriched in rare loss-of-function (LOF) variants in SCZ probands. First, we con-

structed a gene-gene interaction network using the STRING database (http://www.string-db.org). We started with *FAN1* as a seed gene, and then used the STRING database to identify as many genes which are directly or indirectly connected to *FAN1* with a high confidence score (greater than 0.9). This way we have identified a network with 27 genes. Among these genes, there are 10 rare LOF variants being transmitted from parents to SCZ probands (Table S10). The total number of rare LOFs that are transmitted to SCZ probands is 3523. To assess the statistical significance, we have performed a simple re-sampling procedure by randomly sampling 27 genes out of $20,000$ and counting the number of LOFs observed in such a random set. Based on $100,000$ random draws we obtain an empirical p value of 0.037.

References

[1] Zoellner S, Wen X, Hanchard NA, Herbert MA, Ober C, Pritchard JK (2004) Evidence for extensive transmission distortion in the human genome. *Am J Hum Genet* 74: 62–72.

[2] Ionita-Laza I, Makarov V; ARRA Autism Sequencing Consortium, Buxbaum JD (2012b) Scan-statistic approach identifies clusters of rare disease variants in LRP2, a gene linked and associated with autism spectrum disorders, in three datasets. *Am J Hum Genet* 90: 1002–1013.

[3] Schaffner SF et al. (2005) Calibrating a coalescent simulation of human genome sequence variation. *Genome Res* 15: 1576–1583.

[4] Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X (2013) Family-based association tests for sequence data, and comparisons with population-based association tests. *Eur J Hum Genet* in press

[5] Vacic V et al. (2011) Duplications of the neuropeptide receptor gene VIPR2 confer significant risk for schizophrenia. *Nature* 471: 499-503.

[6] Kirov G et al. (2012) De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Mol Psychiatry* 17: 142-153.

[7] Xu B et al. (2012) De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nat Genet* 44: 1365-1369.

[8] Bassett AS et al. (2010) Clinically detectable copy number variations in a Canadian catchment population of schizophrenia. *J Psychiatr Res* 44: 1005-1009.

[9] Kirov G et al. (2008) Comparative genome hybridization suggests a role for NRXN1 and APBA2 in schizophrenia. *Hum Mol Genet* 17: 458-465.

[10] Kirov G et al. (2012) De novo CNV analysis implicates specific abnormalities of post-synaptic signalling complexes in the pathogenesis of schizophrenia. *Mol Psychiatry* 17: 142-153.

[11] Malhotra D et al. (2011) High frequencies of de novo CNVs in bipolar disorder and schizophrenia. *Neuron* 72: 951-963.

[12] Xu B et al. (2008) Strong association of de novo copy number mutations with sporadic schizophrenia. *Nat Genet* 40: 880-885.

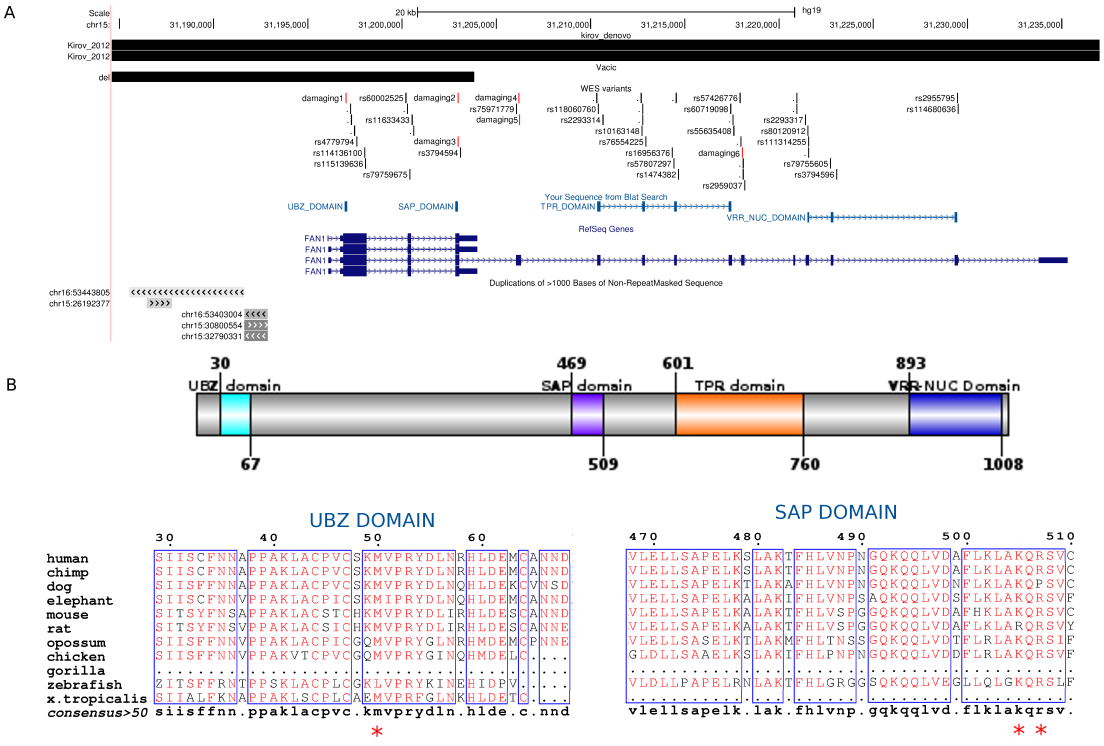

Figure S1: Variants within the *FAN1* region. A. These variants were identified in SCZ datasets. Track1: the de novo CNVs cover the *FAN1* gene in Vacic et al.[5] and Kirov et al.[6]. Track2: Rare variants (MAF$<$ 0.01) identified in our SCZ whole exome sequencing data[7]. Six damaging variants (predicted by the SIFT software) are indicated by red bars. Track3: the genomic coverage of predicted functional domains in the FAN1 protein. Track4: The Refseq gene annotation of all *FAN1* isoforms. Track5: The segmental duplications surrounding the *FAN1* gene. B. The variants that might affect the functional domains of the FAN1 protein. Upper panel: There are four predicted functional domains in the FAN1 protein: UBZ domain, SAP domain, TPR domain and VRR-NUC Domain. Lower panel: The evolutionary conservation of UBZ and SAP domains, which are affected by three damaging variants (red asterisk).

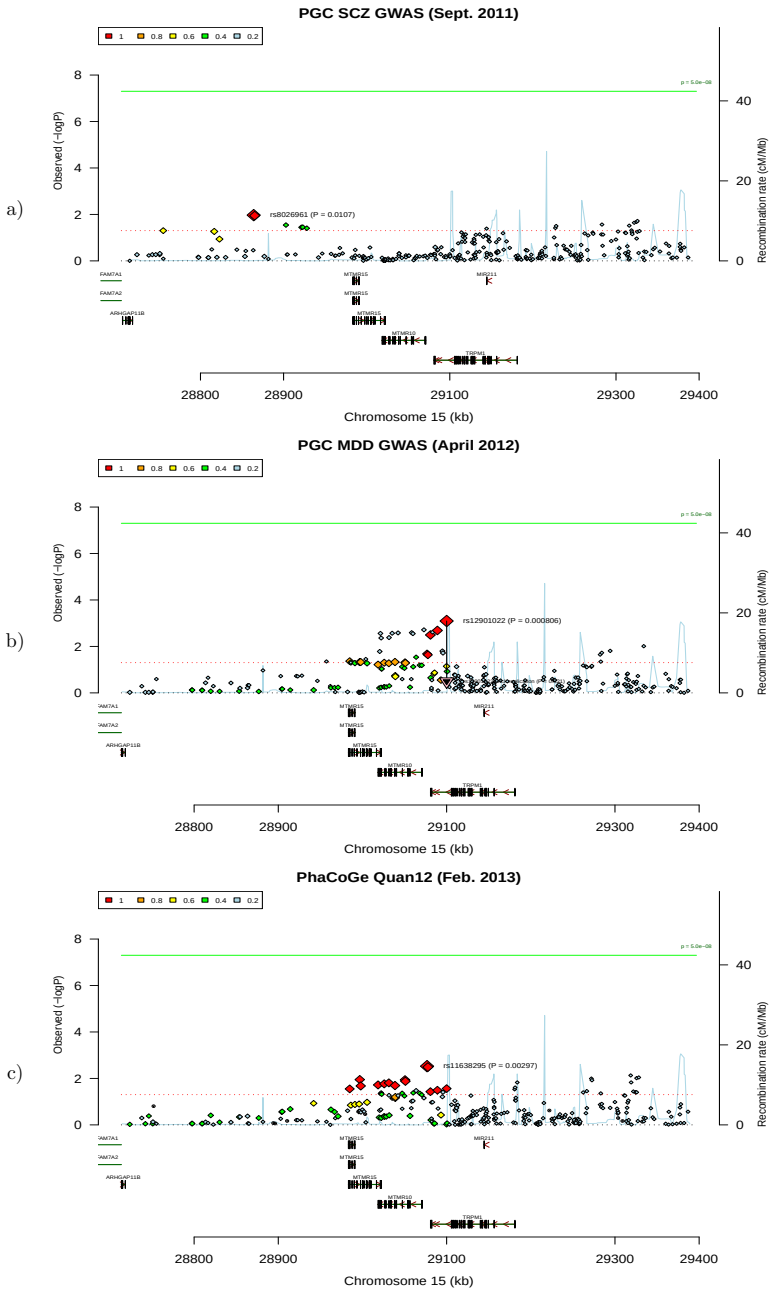

Figure S2: GWAS results in a 400 kb region including *FAN1/MTMR15* for (a) SCZ, (b) Major Depressive Disorder, (c) Antidepressant therapy.

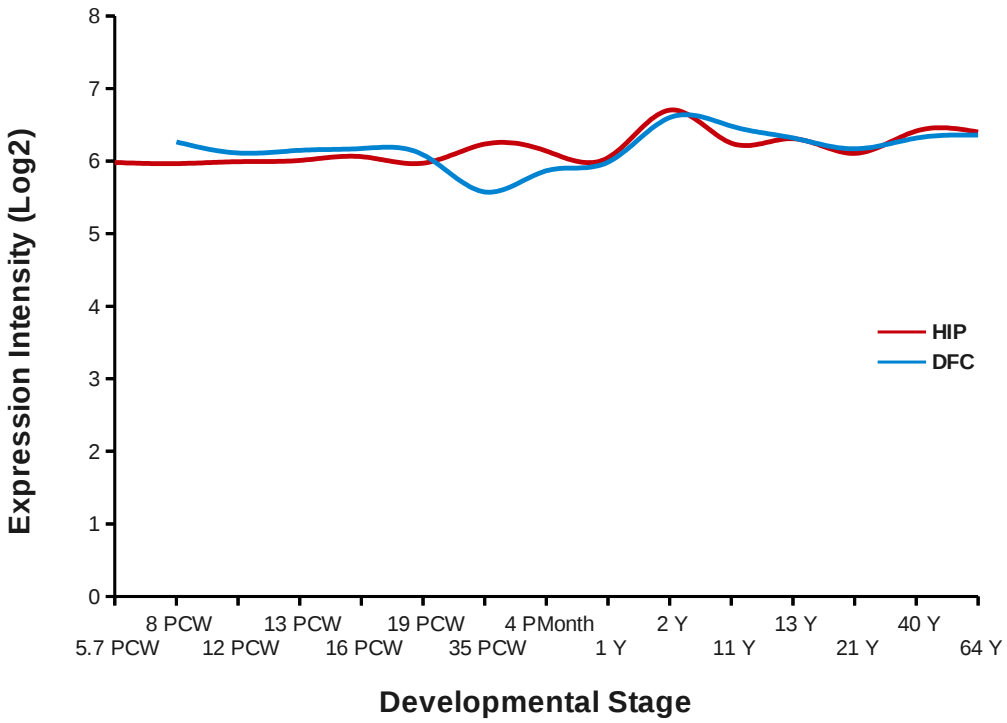

Figure S3: The expression pattern of FAN1 mRNA in Hippocampus (HIP, red line) and dorsal frontal cortex (DFC, blue line) regions along human brain development. PCW = Postconceptual weeks; PMonths = Postnatal months; Y = Year.

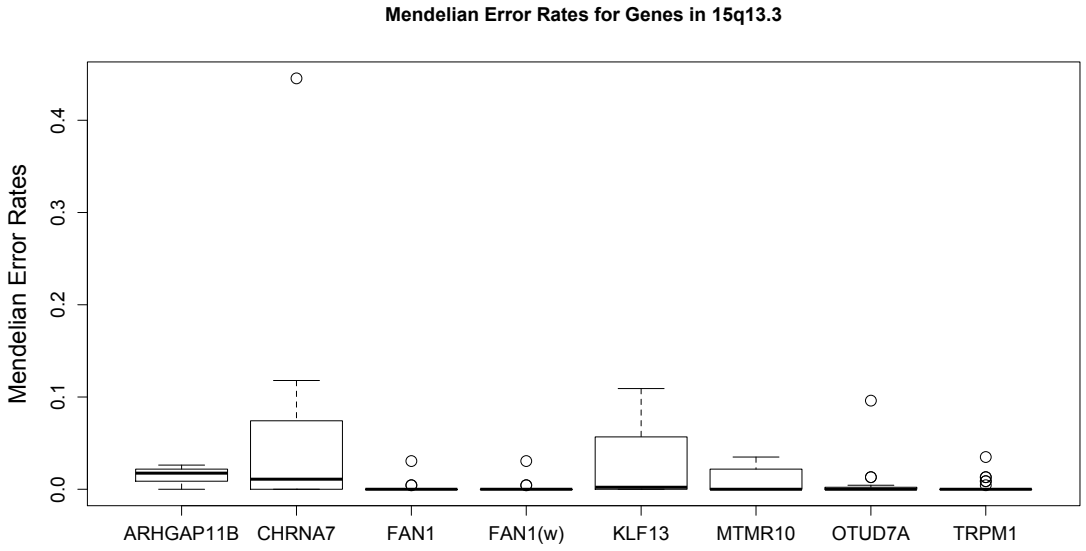

Figure S4: Mendelian error rates for the genes in 15q13.3 CNV. FAN1(w) refers to the 20 kb window with maximum score in the scan statistic procedure. The Mendelian error rate for each variant was calculated as the total number of mendel errors divided by the number of trios.

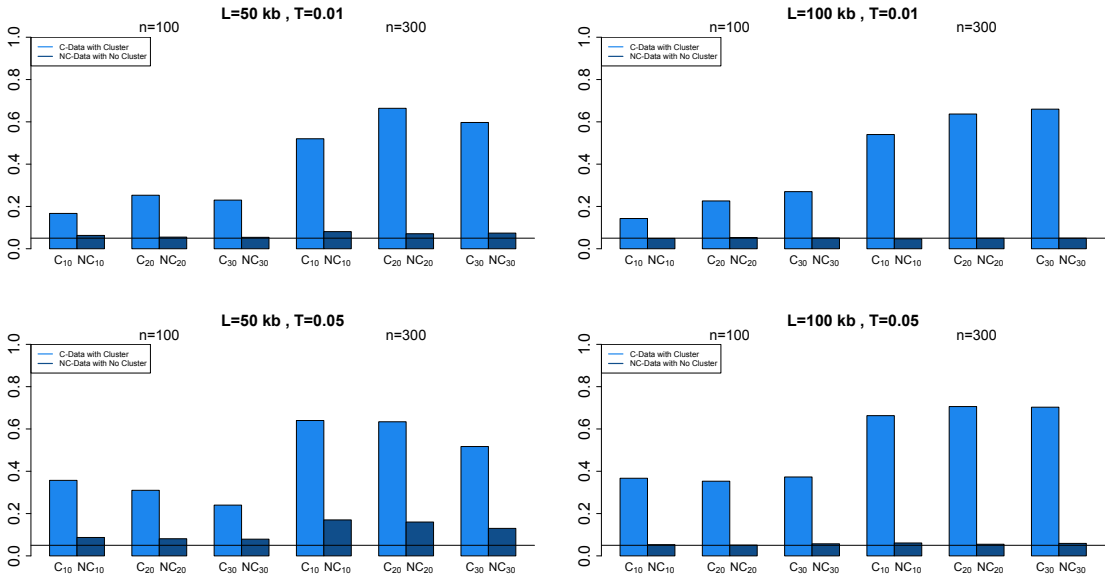

Figure S5: Power of the cluster test on simulated data with clustering (C) and without clustering (NC) of disease risk variants. A scanning window size w of {10kb, 20kb, 30kb} is used. The horizontal line corresponds to the 0.05 level. L is the length of the region, and variants with frequency less than T are included in the study. $n = 100 - 300$ trios are simulated.

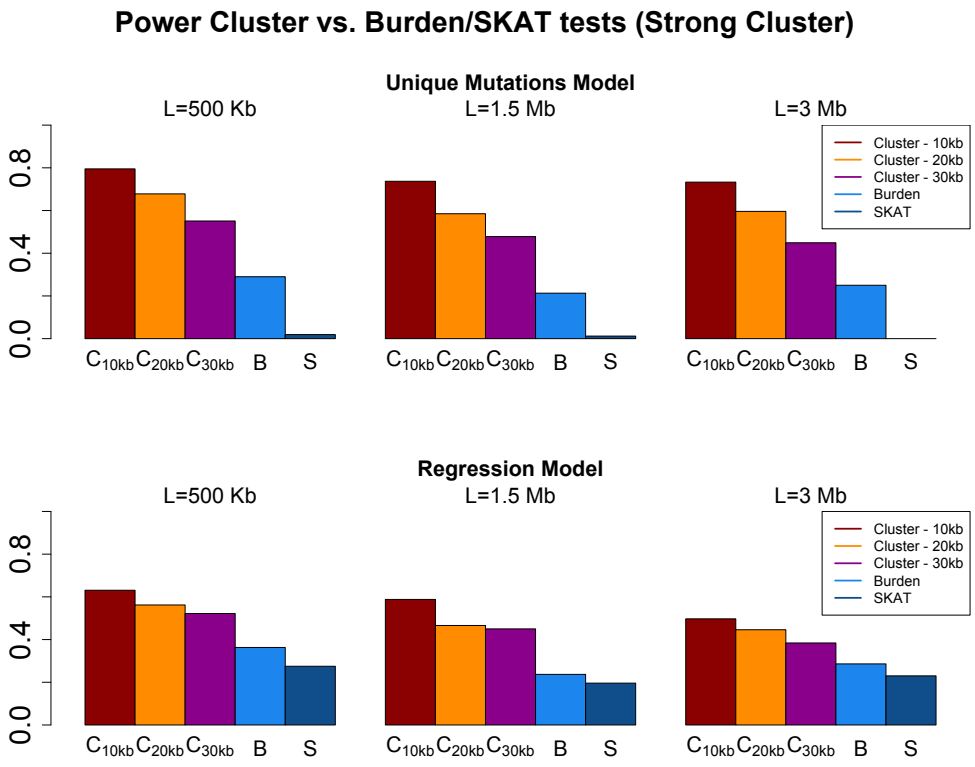

Figure S6: Strong Cluster. Power cluster tests vs. conventional gene-based tests, i.e. Burden and SKAT tests. Two different models have been considered. In the first model, the Unique Mutation Model, 10% of trios ($n = 231$) carry each a unique disease variant that is transmitted from parent to offspring. In the second model, the disease variants are also rare (MAF≤ 0.01), but a multiple regression model is used to generate the phenotype, with the effect size at a disease variant given by $\beta = 0.8|\log 10(MAF)|$. A strong clustering of disease variants is assumed (see text for details).

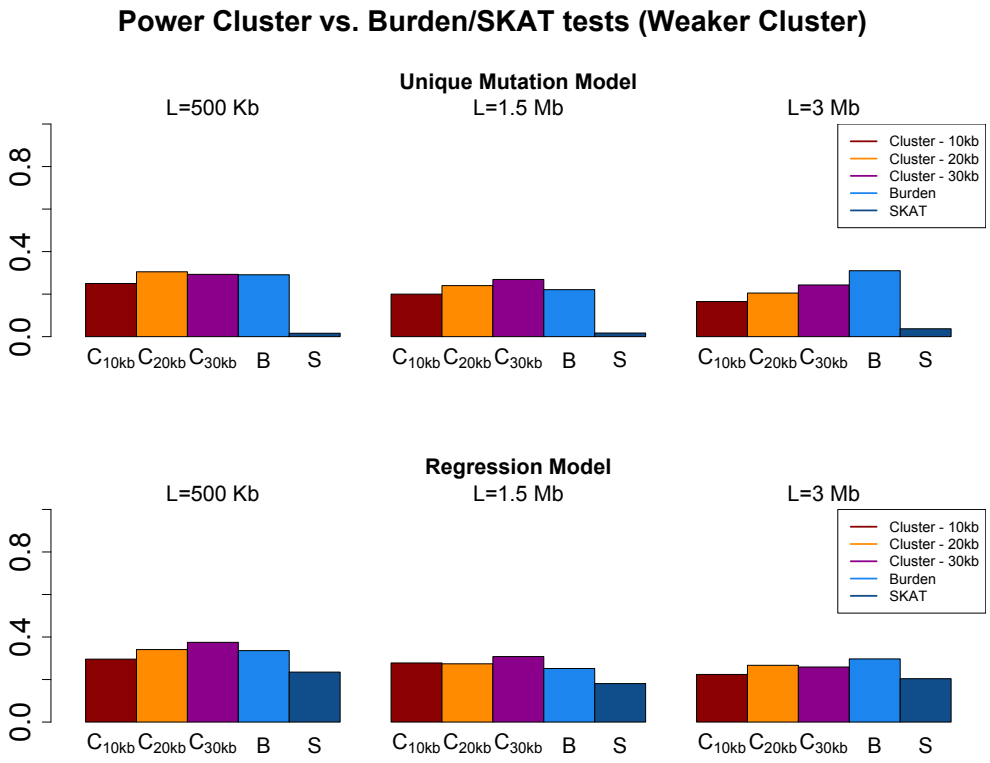

Figure S7: Weaker Cluster. Power cluster tests vs. conventional gene-based tests, i.e. Burden and SKAT tests. Two different models have been considered. In the first model, the Unique Mutation Model, 10% of trios ($n = 231$) carry each a unique disease variant that is transmitted from parent to offspring. In the second model, the disease variants are also rare (MAF≤ 0.01), but a multiple regression model is used to generate the phenotype, with the effect size at a disease variant given by $\beta = 0.8|\log 10(MAF)|$. A weaker clustering of disease variants is assumed (see text for details).

Table S1: De novo CNV regions in SCZ literature (hg19).

Chr	Start	End	Reference
1	765219	1276858	Malhotra et al. Neuron 2011
1	40290696	40904796	Xu et al. Nat Genet 2008
1	145390102	145792052	Kirov et al. Mol Psych 2012
1	237408657	237573021	Kirov et al. Mol Psych 2012
2	133787950	134163308	Kirov et al. Mol Psych 2012
3	28302788	28467088	Xu et al. Nat Genet 2008
3	107847902	108716025	Kirov et al. Mol Psych 2012
3	134282853	136305453	Xu et al. Nat Genet 2008
3	195701151	197340834	Kirov et al. Mol Psych 2012
4	70900915	70934964	Kirov et al. Mol Psych 2012
4	79725588	79862955	Kirov et al. Mol Psych 2012
4	115406126	115518189	Malhotra et al. Neuron 2011
4	187655318	190915739	Malhotra et al. Neuron 2011
5	130835940	130950240	Xu et al. Nat Genet 2008
6	17467091	17831391	Xu et al. Nat Genet 2008
6	17531915	17880119	Malhotra et al. Neuron 2011
6	68619234	68704380	Kirov et al. Mol Psych 2012
7	38294089	38340662	Kirov et al. Mol Psych 2012
7	127488559	127660731	Kirov et al. Mol Psych 2012
7	152011032	159127480	Malhotra et al. Neuron 2011
8	4134560	4312402	Kirov et al. Mol Psych 2012
8	10029452	10118004	Kirov et al. Mol Psych 2012
8	24972985	24990883	Malhotra et al. Neuron 2011
8	113733756	113774509	Malhotra et al. Neuron 2011
9	16320745	16337782	Kirov et al. Mol Psych 2012
9	111819310	112393378	Kirov et al. Mol Psych 2012
9	140642331	140677602	Kirov et al. Mol Psych 2012
9	140649743	140672281	Kirov et al. Mol Psych 2012
11	83795102	84165325	Kirov et al. Mol Psych 2012
11	84328458	84548416	Kirov et al. Mol Psych 2012
12	113239412	113291662	Kirov et al. Mol Psych 2012
12	118581131	118648415	Malhotra et al. Neuron 2011
12	120205095	120229895	Xu et al. Nat Genet 2008
12	131822084	132093577	Kirov et al. Mol Psych 2012
13	41421620	42284276	Kirov et al. Mol Psych 2012
14	35395020	35557969	Kirov et al. Mol Psych 2012
14	94761421	97565021	Xu et al. Nat Genet 2008
15	22673387	23300761	Kirov et al. Mol Psych 2012
15	22751082	23487534	Kirov et al. Mol Psych 2012
15	29212708	30612708	Kirov et al. Hum Mol Genet 2008
15	30920612	32539525	Kirov et al. Mol Psych 2012
16	6948175	6978875	Xu et al. Nat Genet 2008
16	29580611	30191895	Kirov et al. Mol Psych 2012
18	3525935	4342609	Kirov et al. Mol Psych 2012
19	2838838	12423090	Bassett et al. J Psychiatr Res 2010
19	37322979	37501479	Xu et al. Nat Genet 2008
20	14746326	14915051	Kirov et al. Mol Psych 2012
21	23776379	23856373	Kirov et al. Mol Psych 2012
22	20489860	21676782	Malhotra et al. Neuron 2011

Table S2: Top 3 CNV regions significant at the 0.05 level. Only the first CNV, on chromosome 15q13.3, is significant after adjustment for multiple testing. Results are shown for the combined SA and US datasets, with nonsynonymous (NS) variants. 'Window' corresponds to the 20 kb window with the highest score in the scan statistic procedure.

Dataset	n	Variants	CNV	Length	P	Window (hg19)	Gene
SA+US	231	NS	15q13.3	1.6Mb	0.0012	31.197.976-31.217.976	*FAN1*
(SCZ)			22q11.21	2.6Mb	0.051	20.302.276-20.322.276	*DGCR6L*
			3q29	1.6Mb	0.054	196.746.595-196.766.595	*MFI2*

Table S3: Conventional gene-based tests (Burden and SKAT) results for *FAN1* in the SCZ study. All rare variants (All) or only the rare nonsynonymous ones (NS) were analyzed. 'Window' corresponds to the 20 kb window with the highest score in the scan statistic procedure. The p values in the table are gene or window level p values, hence they are unadjusted for multiple genes or windows within a CNV region.

Disease	Dataset	n	Variants	Burden	SKAT
SCZ	SA	146	All	0.21	0.58
			All-window	0.54	0.65
			NS	0.038	0.07
			NS-window	0.064	0.14
SCZ	US	85	All	0.32	0.83
			All-window	0.12	0.53
			NS	0.34	0.47
			NS-window	0.012	0.15
SCZ	SA+US	231	All	0.17	0.68
			All-window	0.06	0.39
			NS	0.025	0.14
			NS-window	0.001	0.022

Table S4: Conventional gene-based tests (Burden and SKAT for family designs) results for genes within 15q13.3, when only rare and nonsynonymous variants are being considered in the SCZ study. *FAN1(w)* refers to the 20 kb window with the highest score in the scan statistic procedure. For three of the genes there were no rare nonsynonymous variants that passed our stringent filtering criteria.

Dataset	n	Variants	Gene	Burden	SKAT
SA+US (SCZ)	231	NS	*FAN1(w)*	0.001	0.022
			FAN1	0.025	0.14
			OTUD7A	0.24	0.08
			TRPM1	0.74	0.040
			MTMR10	0.75	0.69
			ARHGAP11B	NA	NA
			CHRNA7	NA	NA
			KLF13	NA	NA

Table S5: *FAN1* Variants in SCZ study. For each rare variant we report the counts of transmitted (T) and untransmitted (U) minor allele from heterozygous parents to SCZ children. Counts are calculated for mothers (Mo) and fathers (Fa) combined, and separately for mothers and fathers. The functional class for each variant is also reported, with '.' representing non-coding variants. The variants between the double lines reside within the 20 kb window with highest score in the scan statistic procedure.

			Mo+Fa		Mo		Fa	
Chr	Position (hg19)	funcClass	T	U	T	U	T	U
15	31197015	missense	1	1	0	0	1	1
15	31197264	missense	0	1	0	0	0	1
15	31197300	missense	0	1	0	0	0	1
15	31197469	silent	0	2	0	1	0	1
15	31198043	missense	1	0	0	0	1	0
15	31200173	.	1	0	1	0	0	0
15	31200281	.	1	0	1	0	0	0
15	31200396	missense	1	0	1	0	0	0
15	31200602	.	1	2	0	0	1	2
15	31202955	missense	1	0	1	0	0	0
15	31202961	missense	4	0	4	0	0	0
15	31206047	.	2	0	1	0	1	0
15	31206189	missense	1	0	1	0	0	0
15	31206216	missense	1	0	0	0	1	0
15	31210319	.	2	0	1	0	1	0
15	31212667	.	1	0	0	0	1	0
15	31212734	.	1	0	1	0	0	0
15	31214309	.	1	0	1	0	0	0
15	31214385	.	3	1	2	1	1	0
15	31214503	silent	1	0	0	0	1	0
15	31217428	silent	3	1	2	1	1	0
15	31217553	.	2	0	1	0	1	0
15	31217593	.	3	4	2	1	1	3
15	31217935	.	0	4	0	2	0	2
15	31218035	missense	1	1	1	0	0	1
15	31218072	silent	4	3	2	3	2	0
15	31218081	silent	1	1	1	0	0	1
15	31220931	.	0	1	0	1	0	0
15	31221358	.	4	3	3	2	1	1
15	31221528	silent	1	1	0	1	1	0
15	31221552	silent	0	1	0	1	0	0
15	31222702	.	1	1	1	1	0	0
15	31229463	.	1	1	0	0	1	1
Total			45	30	28	15	17	15

Table S6: *FAN1* Variants in SCZ study. For each rare variant we report the counts of transmitted (T) and untransmitted (U) minor allele from heterozygous parents to SCZ children. SIFT score predicts tolerated and deleterious substitutions at each position. $GERP_{RS}$ measures constraint intensity (in terms of rejected substitutions). The functional class for each variant is also reported, with '.' representing non-coding variants. The variants between the double lines reside within the 20 kb window with highest score in the scan statistic procedure.

Chr	Position (hg19)	ID	funcClass	T	U	Prediction	SIFT	$GERP_{RS}$
15	31197015	rs148404807	missense	1	1	damaging	0.01	5.15
15	31197264	.	missense	0	1	.	0.08	-4.32
15	31197300	rs146408181	missense	0	1	.	0.05	-7.76
15	31197469	rs142084532	silent	0	2	.	0.78	1.86
15	31198043	rs115139636	missense	1	0	.	0.54	-3.05
15	31200173	rs60002525	.	1	0	.	.	.
15	31200281	.	.	1	0	.	.	.
15	31200396	rs79759675	missense	1	0	.	0.26	3.36
15	31200602	.	.	1	2	.	.	.
15	31202955	.	missense	1	0	damaging	0.01	-5.43
15	31202961	rs150393409	missense	4	0	damaging	0.04	2.62
15	31206047	rs75971779	.	2	0	.	.	.
15	31206189	.	missense	1	0	damaging	0.01	2.86
15	31206216	.	missense	1	0	damaging	0.02	-0.70
15	31210319	rs189512431	.	2	0	.	.	.
15	31212667	.	.	1	0	.	.	.
15	31212734	.	.	1	0	.	.	.
15	31214309	rs16956376	.	1	0	.	.	.
15	31214385	rs57807297	.	3	1	.	.	.
15	31214503	rs141559766	silent	1	0	.	.	.
15	31217428	rs60719098	silent	3	1	.	.	.
15	31217553	.	.	2	0	.	.	.
15	31217593	rs55635408	.	3	4	.	.	.
15	31217935	.	.	0	4	.	.	.
15	31218035	rs145610507	missense	1	1	damaging	0	5.21
15	31218072	rs143461130	silent	4	3	.	1	0.50
15	31218081	.	silent	1	1	.	.	.
15	31220931	.	.	0	1	.	.	.
15	31221358	rs2293317	.	4	3	.	.	.
15	31221528	rs111314255	silent	1	1	.	.	.
15	31221552	.	silent	0	1	.	.	.
15	31222702	rs79755605	.	1	1	.	.	.
15	31229463	rs114680636	.	1	1	.	.	.

Table S7: *FAN1* rare functional variants in SCZ study, and the study ID for each parent that carries such a variant. Variant annotation has been performed using SnpEff and VariantAnnotator.

Chr	Position (hg19)	SnpEff_Impact	SnpEff_Effect	Parent ID
15	31197015	moderate	NonSyn_Coding	{1011MK0042, 1011MK1032}
15	31197264	moderate	NonSyn_Coding	{1011MK1183}
15	31197300	moderate	NonSyn_Coding	{1011MK0045}
15	31198043	moderate	NonSyn_Coding	{1011MK1195}
15	31200396	moderate	NonSyn_Coding	{1011MK1218}
15	31202955	moderate	NonSyn_Coding	{1011MK0116}
15	31202961	moderate	NonSyn_Coding	{1011MK0719, 1011MK0875, 1011MK0980, 1011MK1227}
15	31206189	moderate	NonSyn_Coding	{1011MK1137}
15	31206216	moderate	NonSyn_Coding	{1011MK1207}
15	31218035	moderate	NonSyn_Coding	{1011MK0003, 1011MK0593}
15	31229463	high	Splice_Site_Donor	{1011MK0294, 1011MK0888}

Table S8: Co-morbid diagnosis (DX) for individuals with nonsynonymous rare variant in the 20 kb window in the SCZ study. 'PofO' refers to the parent-of-origin for the variant (from a heterozygous parent).

Sample ID	Sample	DX	PofO	Co-morbid DX
1011MK0115	SA	SCZ	Mo	DEPR
1011MK0718	SA	SCZ	Mo	DYSTHYMIA
1011MK0874	SA	SCZ	Mo	MDD
1011MK0979	US	SCZAFF, BP type	Mo	
1011MK1136	US	SCZAFF, Depressive type	Mo	Depression-frequent episodes
1011MK1193	US	SCZAFF, BP type	Fa	
1011MK1205	US	SCZ	Fa	Cocaine dep
1011MK1217	US	SCZ	Mo	Depressive episode
1011MK1226	US	SCZAFF, BP type	Mo	Severe depressive episode

Table S9: *FAN1* Variants in ASD study. For each rare variant we report the minor allele frequency in ASD cases (f_A) and controls (f_U). SIFT score predicts tolerated and deleterious substitutions at each position. GERP_{RS} measures constraint intensity (in terms of rejected substitutions). The functional class for each variant is also reported, with '.' representing non-coding variants. The variants between the double lines reside within the 20 kb window with highest score in the scan statistic procedure.

Chr	Position (hg19)	ID	funcClass	f_A	f_U	Prediction	SIFT	GERP_{RS}
15	31196867	rs139312614	missense	0	0.001		0.21	4
15	31196944	.	silent	0.001	0		.	.
15	31197015	rs148404807	missense	0.002	0.003	damaging	0.01	5.15
15	31197040	rs143965941	silent	0.001	0.001		.	.
15	31197300	rs146408181	missense	0.003	0.003		0.31	-7.76
15	31197363	rs146422014	missense	0.001	0.001		0.13	5.22
15	31197469	rs142084532	silent	0	0.001		0.78	1.86
15	31197542	.	missense	0.001	0		0.39	-3.38
15	31197584	rs150748572	missense	0.002	0.004		0.27	4.15
15	31197648	rs137920161	missense	0.001	0		0.5	2.12
15	31197649	rs142437586	silent	0	0.001		.	.
15	31197752	.	missense	0.001	0		0.68	-8.45
15	31197995	rs151322829	missense	0.004	0.004		0.14	-1.21
15	31198043	rs115139636	missense	0	0.001		0.68	-3.05
15	31198044	rs140632948	missense	0	0.001		0.4	2.55
15	31198050	rs201153099	missense	0.003	0	damaging	0.01	5.99
15	31198060	rs141494062	silent	0	0.001		0.25	-8.2
15	31200396	rs79759675	missense	0.001	0.001		0.65	3.36
15	31202787	rs187650218	.	0	0.001		.	.
15	31202794	.	.	0.004	0		.	.
15	31202795	.	.	0.005	0		.	.
15	31202796	.	.	0.001	0.009		.	.
15	31202961	rs150393409	missense	0.018	0.008	damaging	0.04	2.62
15	31206047	rs75971779	.	0	0.003		.	.
15	31206052	.	.	0.001	0		.	.
15	31210319	rs189512431	.	0.002	0.003		.	.
15	31212698	.	.	0.005	0.007		.	.
15	31212734	.	.	0.001	0		.	.
15	31212744	rs200971774	.	0.004	0.007		.	.
15	31212765	rs144081053	missense	0.001	0.001	damaging	0.04	5.7
15	31212777	rs200818425	missense	0.001	0	damaging	0.03	5.89
15	31214425	.	.	0.001	0		.	.
15	31214443	rs144723412	silent	0.001	0.003		.	.
15	31214444	.	missense	0	0.001		0.22	4.78
15	31214488	rs200451712	silent	0	0.001		.	.
15	31214503	rs141559766	silent	0.001	0.001		.	.
15	31217305	.	.	0.001	0		.	.
15	31217339	.	missense	0.001	0	damaging	0.03	5.58
15	31217428	rs60719098	silent	0.001	0.003		.	.
15	31217494	.	silent	0.001	0		0.41	-8.26
15	31217979	.	.	0.001	0.001		.	.
15	31218016	rs144684512	missense	0.001	0		0.22	4.91
15	31218035	rs145610507	missense	0.001	0.001	damaging	0.02	5.21
15	31218036	.	silent	0	0.001		0.06	3.33
15	31218049	.	missense	0.001	0		0.71	-3.61
15	31218057	rs148908472	silent	0.001	0		.	.
15	31218067	rs114441778	missense	0	0.001		0.58	-6.62
15	31218072	rs143461130	silent	0.011	0.003		1	0.505
15	31218180	.	.	0	0.001		.	.
15	31220833	.	missense	0.001	0	damaging	0	3.44
15	31220850	rs200756403	silent	0.002	0		0.05	-7.96
15	31221358	rs2293317	.	0.004	0.004		.	.
15	31221436	.	missense	0	0.001	damaging	0	5.39
15	31221493	rs80120912	missense	0.026	0.009		0.56	-4.95
15	31221498	.	silent	0	0.001		0.59	-10.8
15	31221576	rs142459921	silent	0.001	0		.	.
15	31221582	.	silent	0.001	0		.	.
15	31221647	rs112159557	.	0.001	0.001		.	.
15	31221648	rs191666636	.	0.001	0		.	.
15	31222702	rs79755605	.	0.001	0		.	.
15	31222763	rs200468714	silent	0.001	0		.	.
15	31222916	rs190074079	.	0	0.003		.	.

Table S10: LOF Variants in *FAN1*-based network. Among 27 genes in this network, there are 10 rare LOF variants in SCZ probands from the SA+US dataset. Variant annotation has been performed using SnpEff and VariantAnnotator.

Chr	Position (hg19)	SnpEff_Impact	SnpEff_Effect	Gene
2	58386928	high	Frame_Shift	*FANCL*
3	10128932	high	Frame_Shift	*FANCD2*
7	73649897	high	Stop_Gained	*RFC2*
9	97912338	high	Stop_Gained	*FANCC*
13	32972626	high	Stop_Gained	*BRCA1*
14	45665741	high	Frame_Shift	*FANCM*
14	45667921	high	Stop_Gained	*FANCM*
15	31229463	high	Splice_Site_Donor	*FAN1*
16	89813248	high	Frame_Shift	*FANCA*
17	41276068	high	Frame_Shift	*BRCA1*

Table S11: Type 1 Error Assessment using simulated data. Reported in the table are empirical Type 1 error rates corresponding to α levels of $\{0.05,\ 0.01,\ 0.001\}$. Region length L is $\{50\text{kb},\ 100\text{kb}\}$.

L	n	Threshold	w	α		
(kb)			(kb)	0.05	0.01	0.001
50	100	0.01	10	4.9E-02	9.9E-03	9.7E-04
			20	4.9E-02	1.0E-02	1.1E-03
			30	4.9E-02	1.0E-02	1.0E-03
	100	0.05	10	5.1E-02	1.0E-02	1.0E-03
			20	4.9E-02	9.4E-03	1.0E-03
			30	5.1E-02	1.0E-02	1.4E-03
	300	0.01	10	5.0E-02	1.0E-02	1.3E-03
			20	5.1E-02	9.8E-03	1.1E-03
			30	4.9E-02	1.0E-02	1.1E-03
	300	0.05	10	4.8E-02	9.9E-03	1.4E-03
			20	5.1E-02	1.0E-02	1.0E-03
			30	5.0E-02	9.8E-03	1.0E-03
	500	0.01	10	4.8E-02	9.3E-03	1.1E-03
			20	5.1E-02	1.0E-02	1.2E-03
			30	4.9E-02	1.0E-02	8.1E-04
	500	0.05	10	5.0E-02	1.0E-02	9.5E-04
			20	4.9E-02	9.5E-03	8.7E-04
			30	5.0E-02	1.0E-02	1.1E-03
100	100	0.01	10	4.9E-02	1.0E-02	1.0E-03
			20	4.7E-02	9.5E-03	1.0E-03
			30	4.8E-02	1.0E-02	1.0E-03
	100	0.05	10	4.8E-02	8.1E-03	8.2E-04
			20	4.9E-02	9.8E-03	1.1E-03
			30	4.9E-02	9.3E-03	1.0E-03
	300	0.01	10	4.9E-02	1.1E-02	1.0E-03
			20	4.9E-02	1.0E-02	1.1E-03
			30	4.9E-02	1.0E-02	1.0E-03
	300	0.05	10	5.1E-02	1.0E-02	1.1E-03
			20	5.2E-02	1.1E-02	1.3E-03
			30	5.1E-02	1.0E-02	1.0E-03
	500	0.01	10	4.9E-02	1.0E-02	1.1E-03
			20	4.9E-02	1.0E-02	1.0E-03
			30	4.9E-02	1.0E-02	1.0E-03
	500	0.05	10	5.0E-02	1.0E-02	1.1E-03
			20	5.2E-02	1.0E-02	1.1E-03
			30	4.7E-02	8.0E-03	1.0E-03

Table S12: Jacard measure of overlap for two simulated disease models (strong cluster). 'True Window Size' refers to the median size of the true cluster in the simulations. w is the size of the scanning window. L is the length of the region.

Model	L	True Window Size	w	Overlap
Unique	500 Kb	12.8 Kb	10 Kb	0.55
			20 Kb	0.48
			30 Kb	0.36
	1.5 Mb	13.6 Kb	10 Kb	0.51
			20 Kb	0.47
			30 Kb	0.35
	3 Mb	14.1 Kb	10 Kb	0.48
			20 Kb	0.45
			30 Kb	0.36
Regression	500 Kb	12.6 Kb	10 Kb	0.48
			20 Kb	0.43
			30 Kb	0.33
	1.5 Mb	13.0 Kb	10 Kb	0.45
			20 Kb	0.42
			30 Kb	0.32
	3 Mb	13.7 Kb	10 Kb	0.42
			20 Kb	0.42
			30 Kb	0.32

Table S13: Jacard measure of overlap for two simulated disease models (weaker cluster). 'True Window Size' refers to the median size of the true cluster in the simulations. w is the size of the scanning window. L is the length of the region.

Model	L	True Window Size	w	Overlap
Unique	500 Kb	24.6 Kb	10 Kb	0.30
			20 Kb	0.43
			30 Kb	0.44
	1.5 Mb	24.9 Kb	10 Kb	0.23
			20 Kb	0.35
			30 Kb	0.39
	3 Mb	25.6 Kb	10 Kb	0.21
			20 Kb	0.33
			30 Kb	0.41
Regression	500 Kb	24.5 Kb	10 Kb	0.30
			20 Kb	0.42
			30 Kb	0.44
	1.5 Mb	25.7 Kb	10 Kb	0.26
			20 Kb	0.37
			30 Kb	0.41
	3 Mb	26.0 Kb	10 Kb	0.25
			20 Kb	0.32
			30 Kb	0.37