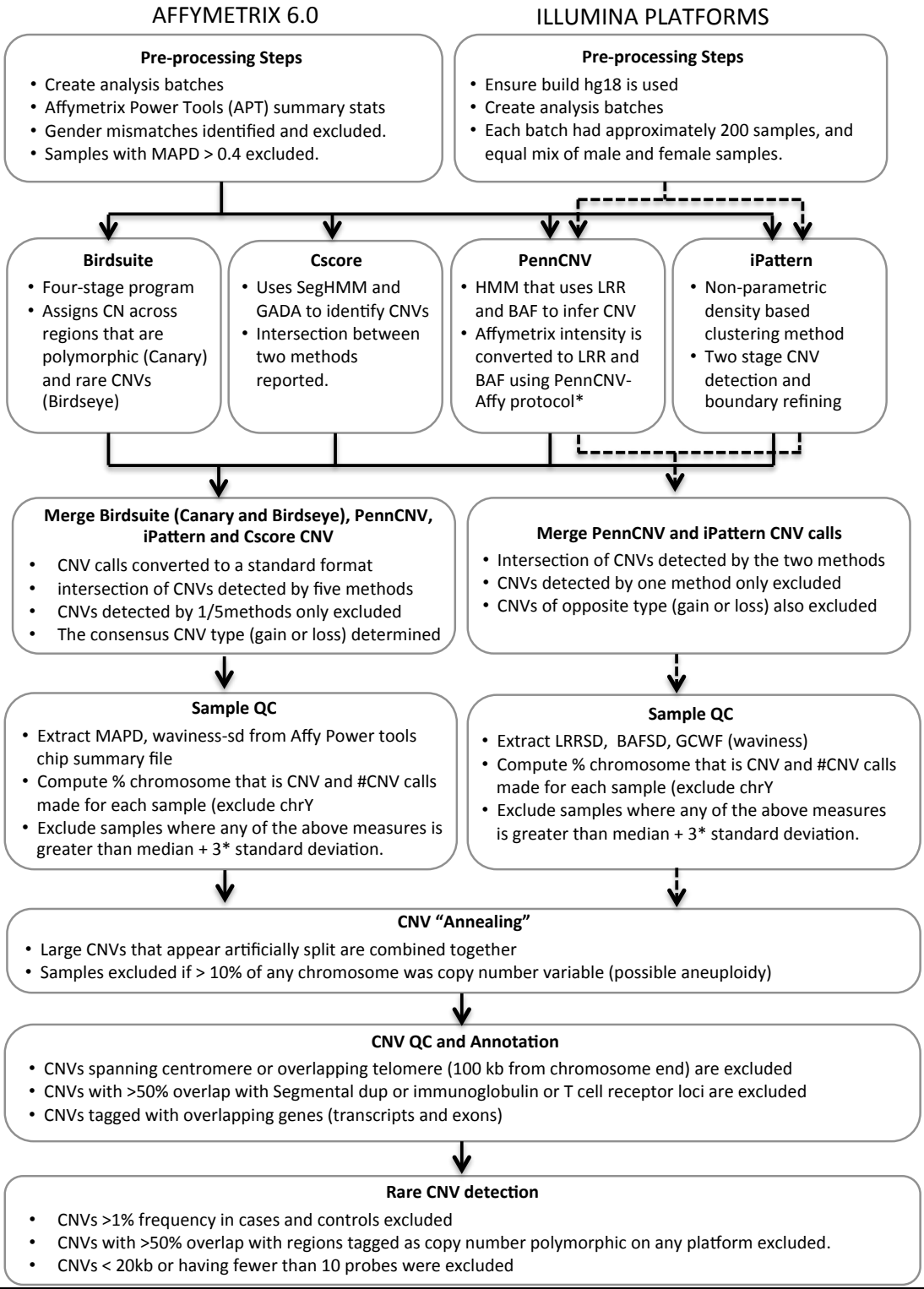
CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects
Authors	Marshall, Christian R.;Howrigan, Daniel P.;Merico, Daniele;Thiruvahindrapuram, Bhooma;Wu, Wenting;Greer, Douglas S.;Antaki, Danny;Shetty, Aniket;Holmans, Peter A.;Pinto, Dalila;Gujral, Madhusudan;Brandler, William M.;Malhotra, Dheeraj;Wang, Zhouzhi;Fajarado, Karin V. Fuentes;Maile, Michelle S.;Ripke, Stephan;Agartz, Ingrid;Albus, Margot;Alexander, Madeline;Amin, Farooq;Atkins, Joshua;Bacanu, Silviu A.;Belliveau, Richard A., Jr.;Bergen, Sarah E.;Ertalan, Marcelo;Bevilacqua, Elizabeth;Bigdeli, Tim B.;Black, Donald W.;Bruggeman, Richard;Buccola, Nancy G.;Buckner, Randy L.;Bulik-Sullivan, Brendan;Byerley, William;Cahn, Wiepke;Cai, Guiqing;Cairns, Murray J.;Campion, Dominique;Cantor, Rita M.;Carr, Vaughan J.;Carrera, Noa;Catts, Stanley V.;Chambert, Kimberley D.;Cheng, Wei;Cloninger, C. Robert;Cohen, David;Cormican, Paul;Craddock, Nick;Crespo-Facorro, Benedicto;Crowley, James J.;Curtis, David;Davidson, Michael;Davis, Kenneth L.;Degenhardt, Franziska;Del Favero, Jurgen;DeLisi, Lynn E.;Dikeos, Dimitris;Dinan, Timothy G.;Djurovic, Srdjan;Donohoe, Gary;Drapeau, Elodie;Duan, Jubao;Dudbridge, Frank;Eichhammer, Peter;Eriksson, Johan;Escott-Price, Valentina;Essioux, Laurent;Fanous, Ayman H.;Farh, Kai-How;Farrell, Martilias S.;Frank, Josef;Franke, Lude;Freedman, Robert;Freimer, Nelson B.;Friedman, Joseph I.;Forstner, Andreas J.;Fromer, Menachem;Genovese, Giulio;Georgieva, Lyudmila;Gershon, Elliot S.;Giegling, Ina;Giusti-Rodriguez, Paola;Godard, Stephanie;Goldstein, Jacqueline I.;Gratten, Jacob;de Haan, Lieuwe;Hamshere, Marian L.;Hansen, Mark;Hansen, Thomas;Haroutunian, Vahram;Hartmann, Annette M.;Henskens, Frans A.;Herms, Stefan;Hirschhorn, Joel N.;Hoffinann, Per;Hofman, Andrea;Huang, Hailiang;Ikeda, Masashi;Joa, Inge;Kahler, Anna K.;Kahn, Rene S.;Kalaydjieva, Luba;Karjalainen, Juha;Kavanagh, David;Keller, Matthew C.;Kelly, Brian J.;Kennedy, James L.;Kim, Yunjung;Knowles, James A.;Konte, Bettina;Laurent, Claudine;Lee, Phil;Lee, S. Hong;Legge, Sophie E.;Lerer, Bernard;Levy, Deborah L.;Liang, Kung-Yee;Lieberman, Jeffrey;Lonnqvist, Jouko;Loughland, Carmel M.;Magnusson, Patrik K. E.;Maher, Brion S.;Maier, Wolfgang;Mallet, Jacques;Mattheisen, Manuel;Mattingsdal, Morten;McCarley, Robert W.;McDonald, Colm;McIntosh, Andrew M.;Meier, Sandra;Meijer, Carin J.;Melle, Ingrid;Mesholam-Gately, Raquelle I.;Metspalu, Andres;Michie, Patricia T.;Milani, Lili;Milanova, Vihra;Mokrab, Younes;Morris, Derek W.;Muller-

	Myhsok, Bertram;Murphy, Kieran C.;Murray, Robin M.;Myin-Germeys, Inez;Nenadic, Igor;Nertney, Deborah A.;Nestadt, Gerald;Nicodemus, Kristin K.;Nisenbaum, Laura;Nordin, Annelie;O'Callaghan, Eadbhard;O'Dushlaine, Colm;Oh, Sang-Yun;Olincy, Ann;Olsen, Line;O'Neill, F. Anthony;Van Os, Jim;Pantelis, Christos;Papadimitriou, George N.;Parkhomenko, Elena;Pato, Michele T.;Paunio, Tiina;Perkins, Diana O.;Pers, Tune H.;Pietilainen, Olli;Pimm, Jonathan;Pocklington, Andrew J.;Powell, John;Price, Alkes;Pulver, Ann E.;Purcell, Shaun M.;Quested, Digby;Rasmussen, Henrik B.;Reichenberg, Abraham;Reimers, Mark A.;Richards, Alexander L.;Roffman, Joshua L.;Roussos, Panos;Ruderfer, Douglas M.;Salomaa, Veikko;Sanders, Alan R.;Savitz, Adam;Schall, Ulrich;Schulze, Thomas G.;Schwab, Sibylle G.;Scolnick, Edward M.;Scott, Rodney J.;Seidman, Larry J.;Shi, Jianxin;Silverman, Jeremy M.;Smoller, Jordan W.;Soderman, Erik;Spencer, Chris C. A.;Stahl, Eli A.;Strengman, Eric;Strohmaier, Jana;Stroup, T. Scott;Suvisaari, Jaana;Svrakic, Dragan M.;Szatkiewicz, Jin P.;Thirumalai, Srinivas;Tooney, Paul A.;Veijola, Juha;Visscher, Peter M.;Waddington, John;Walsh, Dermot;Webb, Bradley T.;Weiser, Mark;Wildenauer, Dieter B.;Williams, Nigel M.;Williams, Stephanie;Witt, Stephanie H.;Wolen, Aaron R.;Wormley, Brandon K.;Wray, Naomi R.;Wu, Jing Qin;Zai, Clement C.;Adolfsson, Rolf;Andreassen, Ole A.;Blackwood, Douglas H. R.;Bramon, Elvira;Buxbaum, Joseph D.;Cichon, Sven;Collier, David A.;Corvin, Aiden;Daly, Mark J.;Darvasi, Ariel;Domenici, Enrico;Esko, Tonu;Gejman, Pablo V.;Gill, Michael;Gurling, Hugh;Hultman, Christina M.;Iwata, Nakao;Jablensky, Assen V.;Jonsson, Erik G.;Kendler, Kenneth S.;Kirov, George;Knight, Jo;Levinson, Douglas F.;Li, Qingqin S.;McCarroll, Steven A.;McQuillin, Andrew;Moran, Jennifer L.;Mowry, Bryan J.;Nothen, Markus M.;Ophoff, Roel A.;Owen, Michael J.;Palotie, Aarno;Pato, Carlos N.;Petryshen, Tracey L.;Posthuma, Danielle;Rietschel, Marcella;Riley, Brien P.;Rujescu, Dan;Sklar, Pamela;St Clair, David;Walters, James T. R.;Werge, Thomas;Siillivan, Patrick F.;O'Donovan, Michael C.;Scherer, Stephen W.;Neale, Benjamin M.;Sebat, Jonathan
Publication date	2017
Original Citation	Marshall, C. R., et al. (2016) 'Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects', Nature Genetics, 49, pp. pages 27–35. doi: 10.1038/ng.3725
Type of publication	Article (peer-reviewed)
Link to publisher's version	https://www.nature.com/articles/ng.3725 - 10.1038/ng.3725

Rights	© 2017, Nature America, Inc., part of Springer Nature. All rights reserved. This is the peer reviewed version of the following article: Marshall, C. R., et al. (2016) 'Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects', Nature Genetics, 49, pp. pages 27–35. doi: 10.1038/ng.3725, which has been published in final form at https://doi.org/10.1038/ng.3725
Download date	2026-11-05 21:44:17
Item downloaded from	https://hdl.handle.net/10468/6517

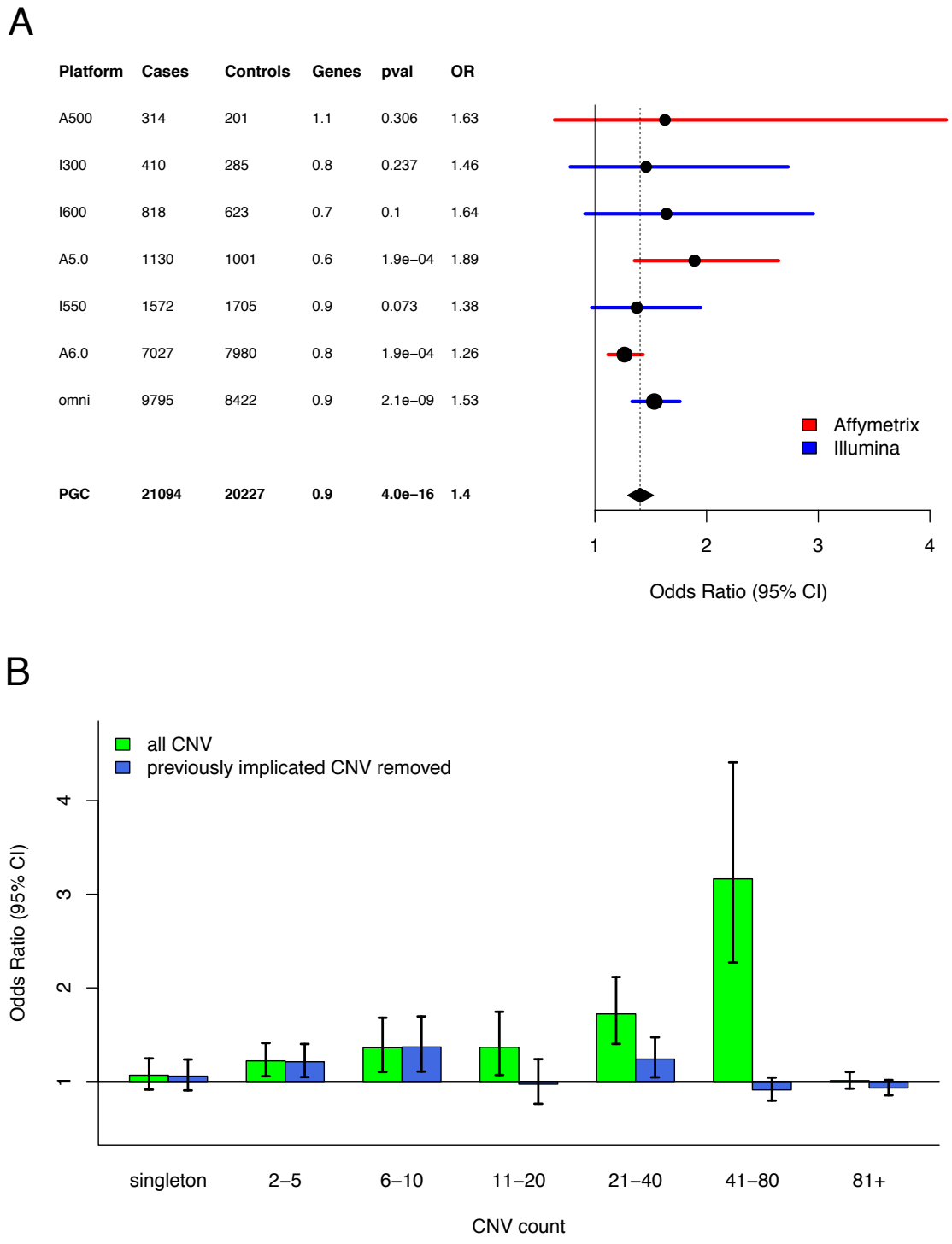




Supplementary Figure 1

Copy number variation (CNV) analysis pipeline workflow

Copy number variation (CNV) analysis pipeline for Affymetrix and Illumina Arrays.



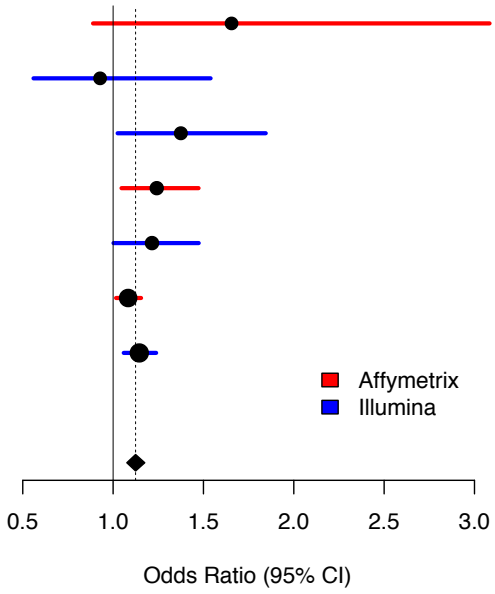
Supplementary Figure 2

CNV burden for losses.

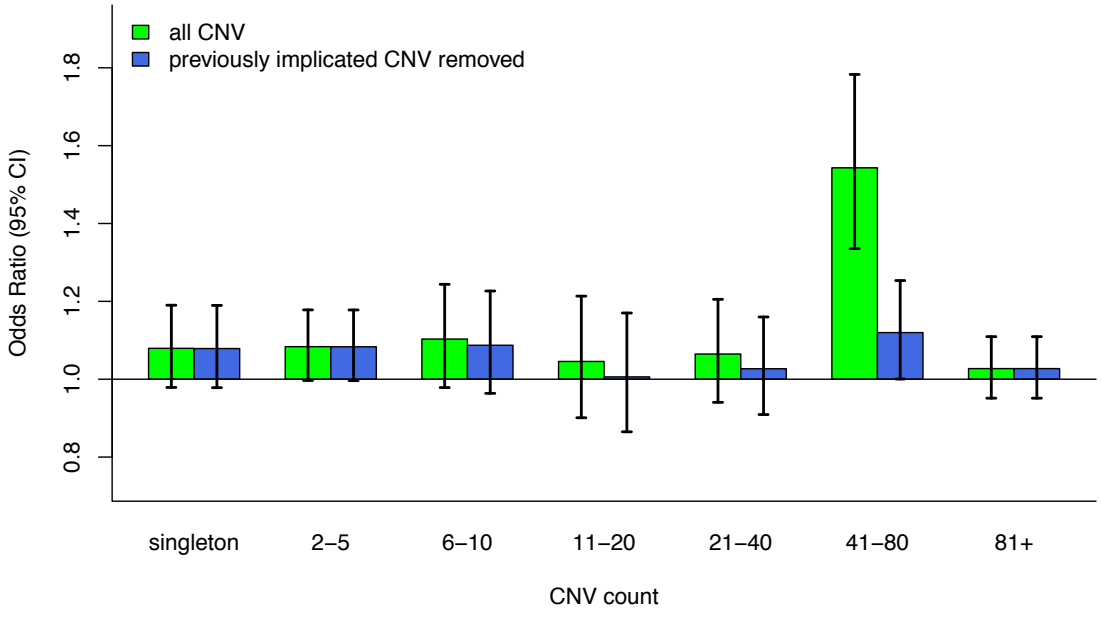
(A) Forest plot of CNV burden (genes affected) partitioned by genotyping platform. (B) CNV burden partitioned by CNV frequency.

A

Platform	Cases	Controls	Genes	pval	OR
A500	314	201	1.6	0.112	1.66
I300	410	285	2.4	0.772	0.93
I600	818	623	1.4	0.034	1.38
A5.0	1130	1001	1.8	0.013	1.24
I550	1572	1705	1.9	0.047	1.22
A6.0	7027	7980	2.2	0.014	1.08
omni	9795	8422	2.1	0.001	1.15
PGC	**21094**	**20227**	**2.1**	**2e−07**	**1.12**



B

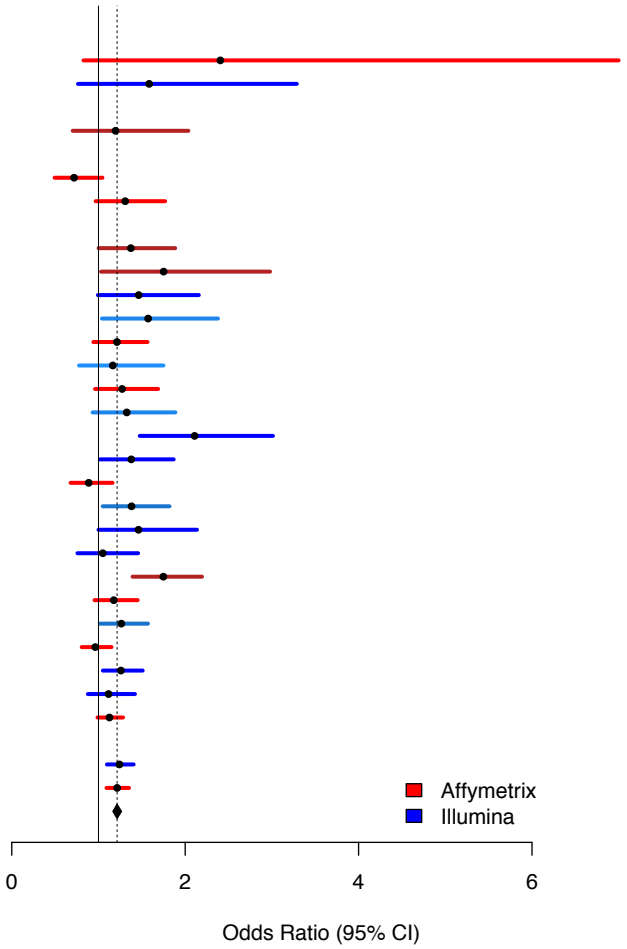


Supplementary Figure 3

CNV burden for gains

(A) Forest plot of CNV burden (genes affected) partitioned by genotyping platform. (B) CNV burden partitioned by CNV frequency.

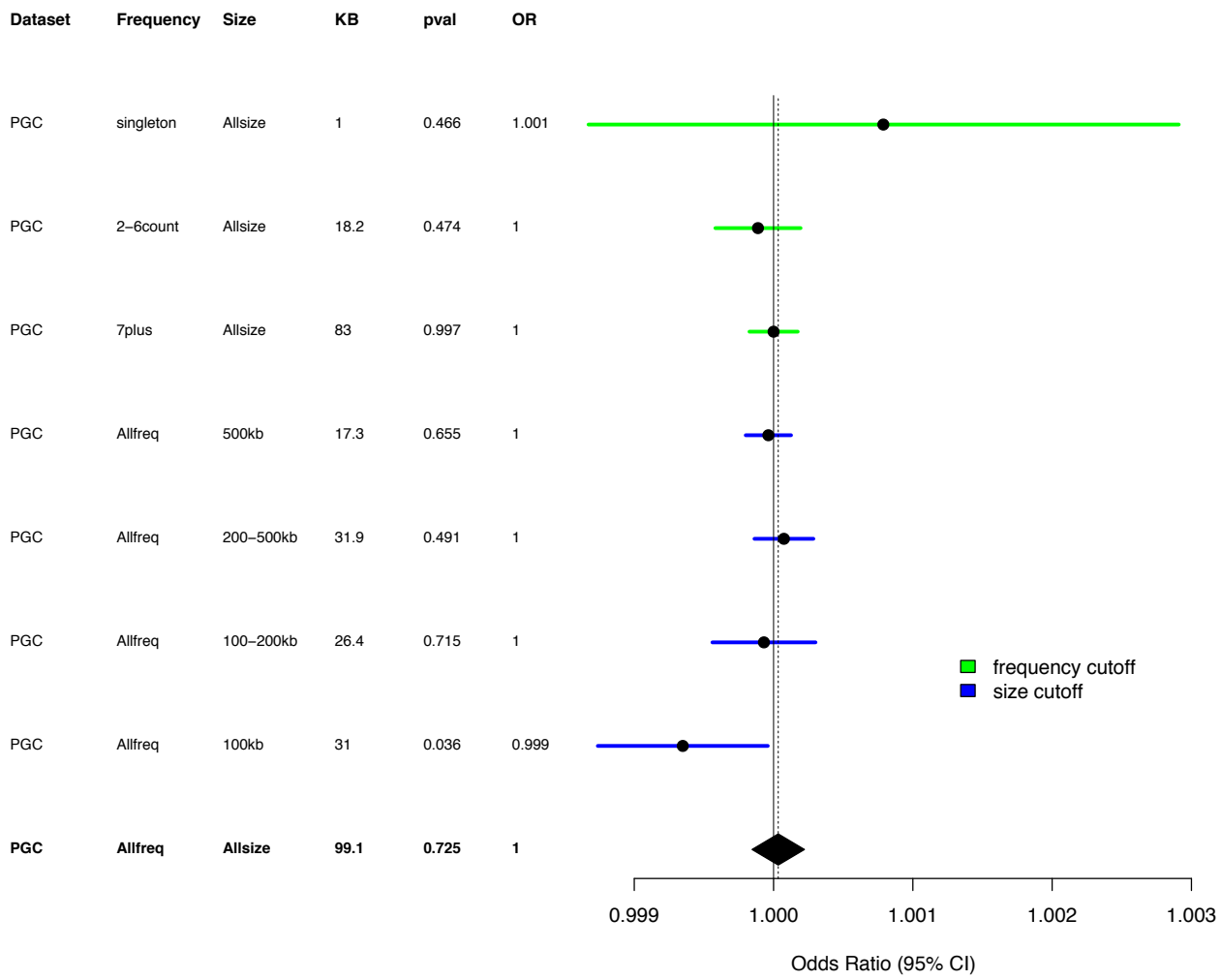
Dataset	Cases	Controls	Genes	pval	OR
uktr	39	0			
pews	42	18	1.5	0.105	2.41
cims	63	64	2.3	0.214	1.58
lktu	172	0			
swe1	140	125	2.1	0.506	1.2
butr	293	0			
msaf	274	125	2.9	0.084	0.72
top8	166	265	1.6	0.081	1.31
uclo	461	0			
port	281	191	2	0.049	1.37
cati	314	201	1.7	0.039	1.75
ersw	260	317	1.4	0.055	1.46
asrb	367	244	1.3	0.032	1.57
edin	341	276	1.9	0.137	1.21
munc	410	285	2	0.461	1.16
buls	182	585	2.4	0.093	1.27
denm	451	379	1.3	0.117	1.33
umes	186	662	1.8	4.5e−05	2.11
umeb	325	525	1.9	0.038	1.38
dubl	252	806	2.6	0.387	0.89
ucla	672	452	1.6	0.022	1.38
cou3	526	601	1.7	0.05	1.46
egcu	229	1118	2	0.769	1.05
aber	690	685	1.6	1.6e−06	1.75
pewb	365	1265	2	0.128	1.18
boco	458	1253	2	0.036	1.26
irwt	1089	797	2.4	0.677	0.96
swe6	952	1125	2.1	0.013	1.26
clo3	2096	1069	2.6	0.371	1.12
s234	1314	1909	2.5	0.074	1.13
clm2	3418	0			
swe5	1729	2509	2.1	0.001	1.24
mgs2	2537	2376	2.6	3.5e−04	1.22
PGC	**21094**	**20227**	**2.2**	**6.6e−21**	**1.21**



Supplementary Figure 4

CNV burden results for individual cohorts.

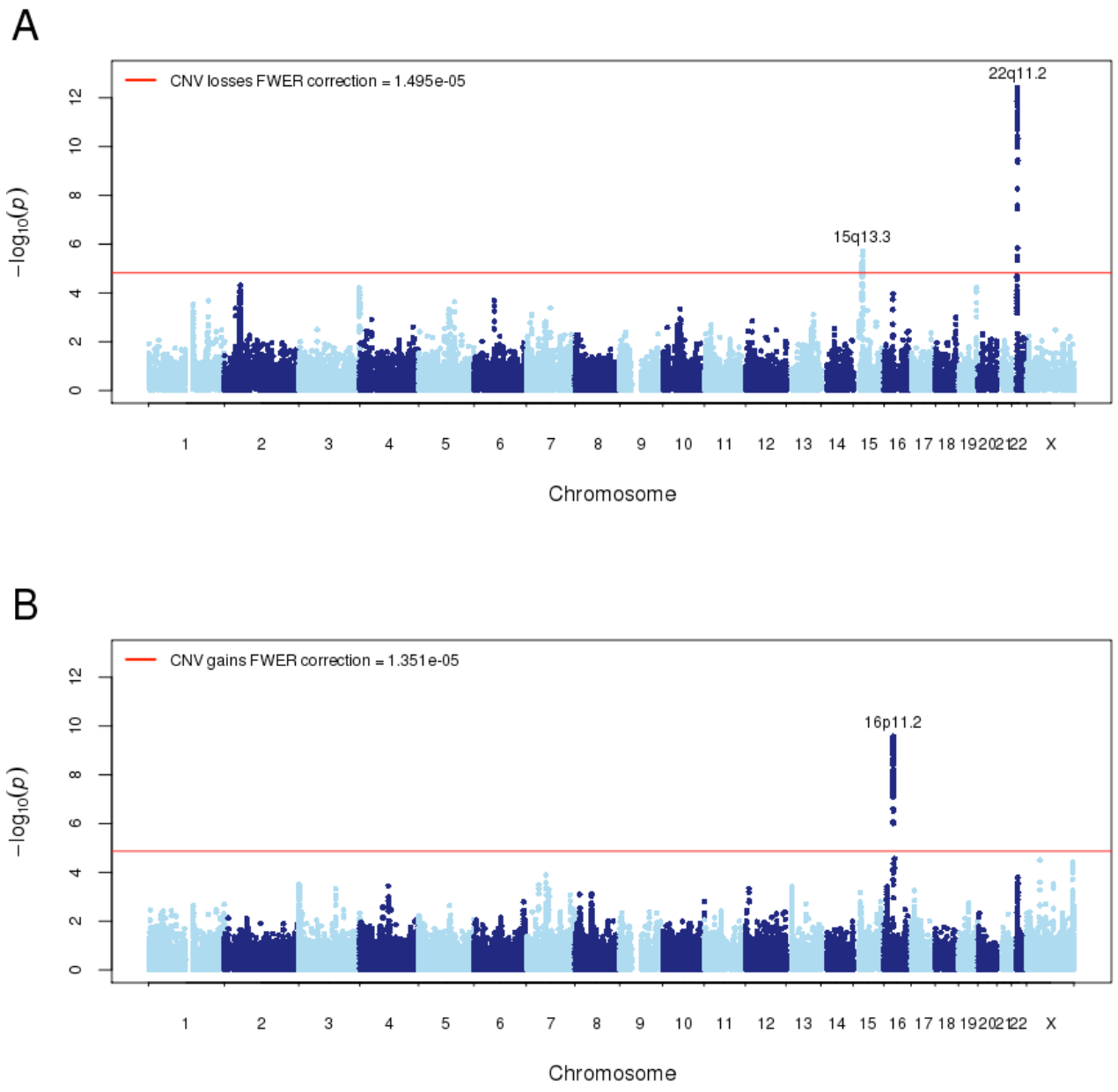
Forest plot of CNV burden (measured here as genes affected by CNV), partitioned by individual cohort, with the full PGC sample at the bottom. CNV burden is calculated by combining CNV gains and losses. Case and control counts are listed, and “genes” is the rate of genes affected by CNV in controls. Three parent-child trio cohorts (uktr, lktu, and butr) and two case-only cohorts (uclo and clm2) did not have matching controls and thus have no burden estimates in the graph, but are included in the full PGC sample. Burden tests use a logistic regression model predicting SCZ case/control status by CNV burden along with covariates (see methods). The odds ratio is the exponential of the logistic regression coefficient, and odds ratios above one predict increased SCZ risk.



Supplementary Figure 5

CNV burden results for non-genic CNV.

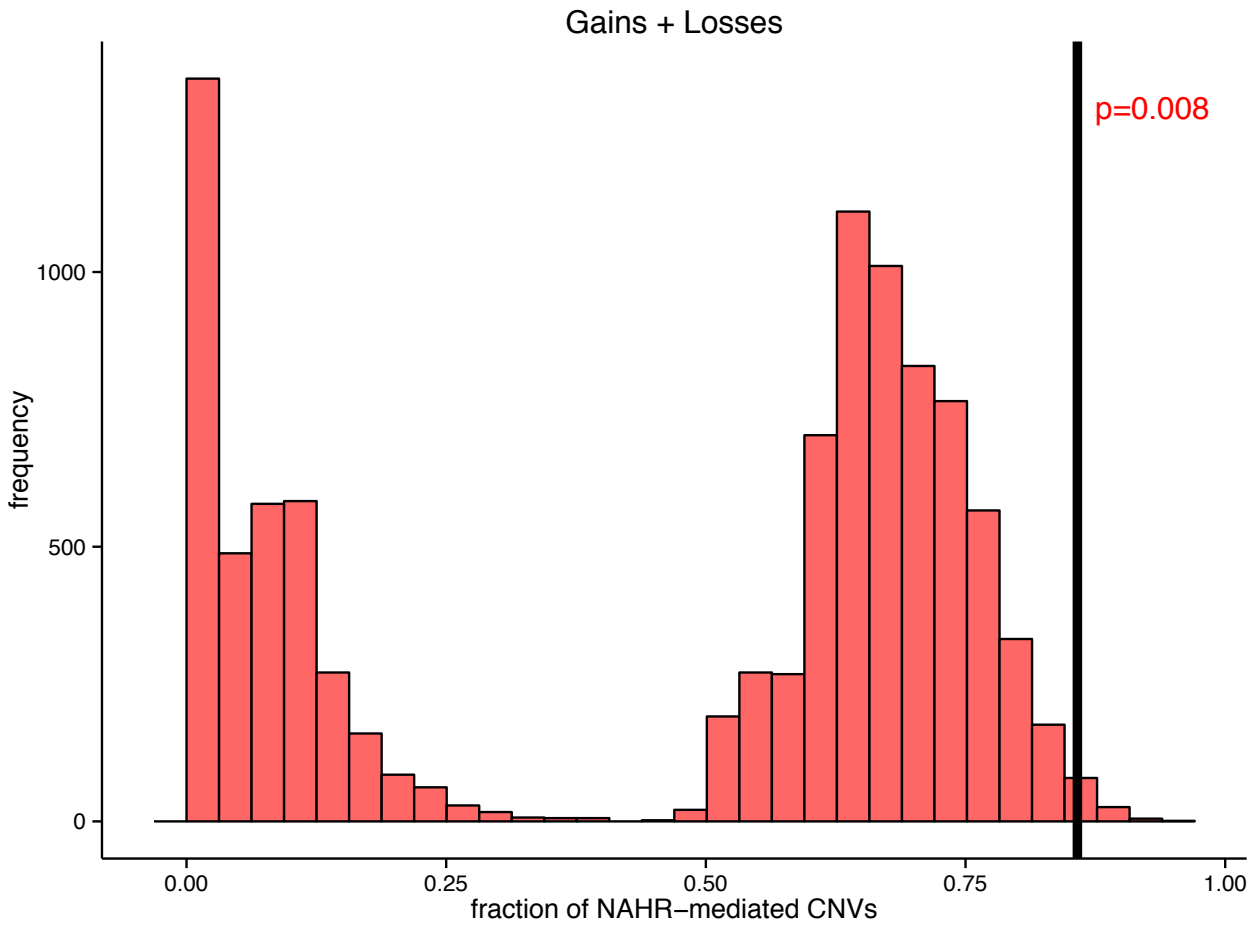
Forest plot of CNV burden from CNV gains and losses occurring outside of any genic region (both intronic and exonic regions). CNV burden is measured by total Kb length, with "KB" being the mean Kb covered by CNV in controls. The PGC cohort is split separately by CNV frequency and size. Burden tests use a logistic regression model predicting SCZ case/control status by CNV burden along with covariates (see methods). The odds ratio is the exponential of the logistic regression coefficient, and odds ratios above one predict increased SCZ risk.



Supplementary Figure 6

CNV probe level Manhattan plot.

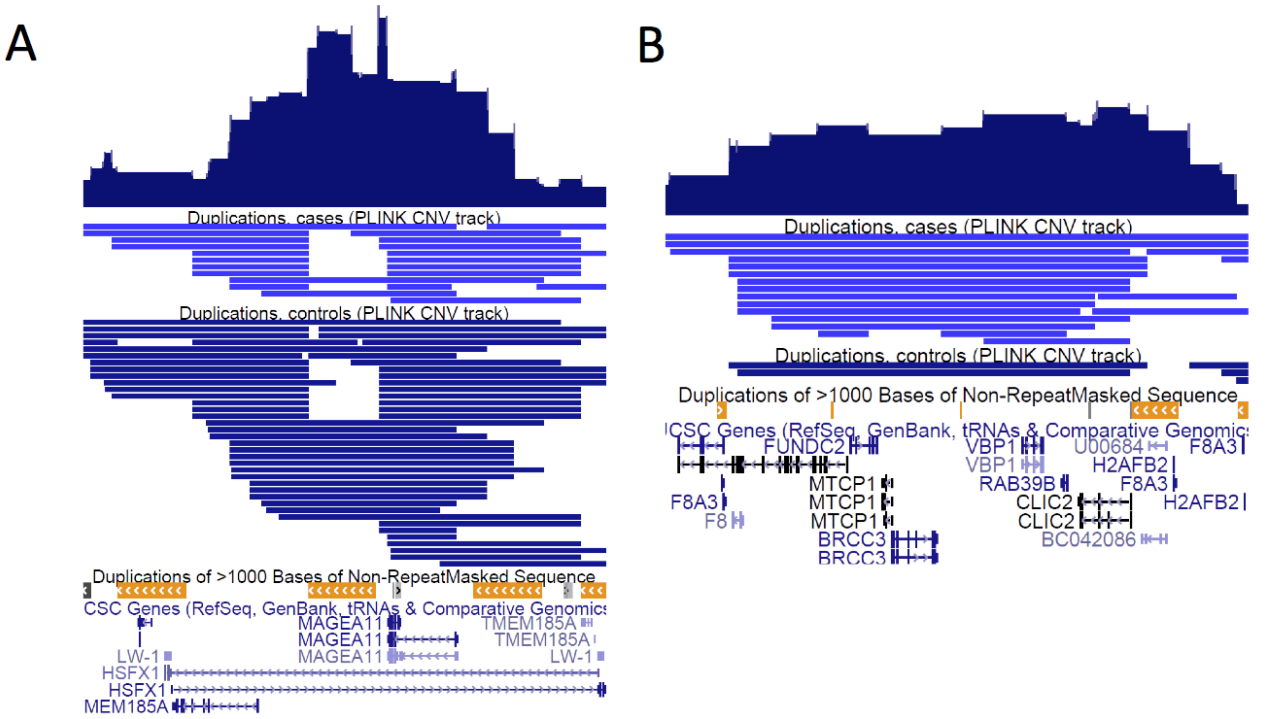
Manhattan plot of probe-level association results from the SCZ residual phenotype for CNV losses (A) and gains (B). Genome-wide correction was determined using the family-wise error rate (FWER) drawn from permutations.



Supplementary Figure 7

Permutation of NAHR-mediated CNVs.

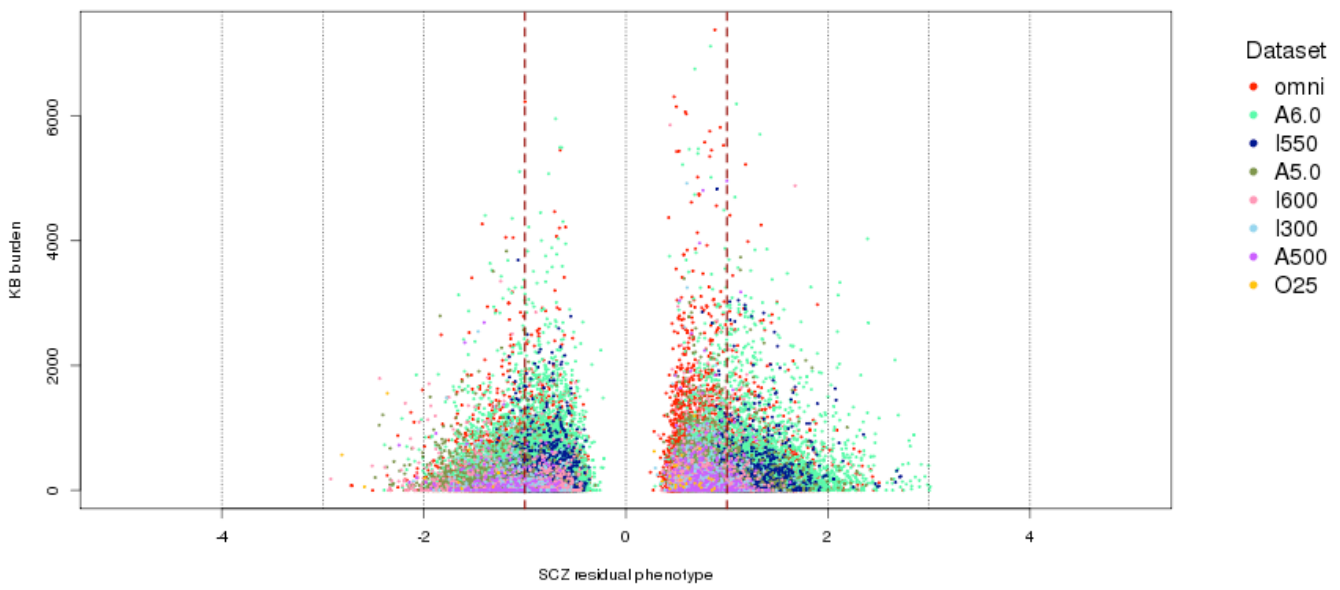
Permutation results from drawing frequency-match CNV loci and testing for fraction of NAHR-mediated CNVs. To test for the enrichment of NAHR-mediated loci in our suggestive results from the gene-based test, each permutation selected an equivalent number of independent CNV loci and tested the faction of NAHR-mediated CNVs.



Supplementary Figure 8

Xq28 CNV hotspot.

(A) Protective CNV gain association peak around the MAGEA11 and TMEM185A gene, both within an intron of the HSFX1 gene. (B) Risk CNV gain association peak at the distal end of Xq28 covering 10 separate gene. Mahanttan plot, CNV and Gene tracks from the UCSC genome browser are shown.



Supplementary Figure 9

SCZ phenotype residual distribution.

X-axis: Distribution of phenotype residual values after regressing case/control status on selected covariates. Plotted against overall CNV burden (Y-axis) to visually inspect if large residuals have an excess of CNV burden, which can lead to higher false positive associations. SCZ cases have positive residual values and controls negative residual values.

PGC Schizophrenia CNV analysis – Supplementary Note

Supplementary Methods and Results

- **CNV post-processing and QC**
- **CNV burden between sexes**
- **Breakpoint level power analysis**
- **Controlling for population stratification**
- **Gene-set burden enrichment analysis: selection of gene-sets and pre-processing of CNV calls**
- **Genome-wide correction for multiple testing**
- **Gene-based network analysis**
- **Follow up of significant CNV loci**
- **Proportion of variance in SCZ explained by top CNV loci**
- **NAHR enrichment in significant novel gene loci**
- **Identifying SCZ risk loci that have been previously implicated in the literature**

Subsets of the PGC schizophrenia dataset that have been published previously

Consortium Membership

Acknowledgements

PGC Data Sharing Policy

Supplementary Tables

Nature Genetics: doi:10.1038/ng.3725

Supplementary Results

CNV post-processing and QC

Merging of CNV data from multiple CNV callers: For each subject, CNV calls were generated using multiple algorithms. After standardization of outputs from each algorithm, CNV calls from each algorithm were merged at the sample level to increase specificity. For CNVs generated from Affymetrix 6.0 array, we took the intersection of the four outputs (Birdsuite, iPattern, C-Score, PennCNV) at the sample level to create a consensus CNV. For the Affymetrix 500, Affymetrix 5.0, and Illumina platforms, CNV merging was performed by taking the intersection of the calls made by the two algorithms (PennCNV and C-Score for Affymetrix 500, Birdsuite and PennCNV for Affymetrix 5.0, and iPattern and PennCNV for Illumina) at the sample level. CNV calls that were made by only one of the algorithm were excluded. Calls discordant for type of CNV (gain or loss) were also excluded.

Quality control filtering: Following merging we applied filtering criteria for removal of arrays with excessive probe variance or GC bias and removal of samples with mismatches in gender or ethnicity or chromosomal aneuploidies. For Affymetrix data, we extracted the MAPD and waviness-sd from the APT summary file. We then retained experiments if each of these measures was within 3 SD of the median. We also calculated the proportion of each chromosome (excluding chrY) tagged as copy number variable and excluded samples where > 10% of the chromosome was copy number variable as possible aneuploidies. We applied the same criteria to the Illumina datasets using similar QC metrics LRRSD, BAFSD, GCWF (waviness) from PennCNV log files. Individual Large CNVs that were fragmented into multiple calls were merged together if one of the methods detected a CNV spanning the gap. Further, we excluded CNVs that: 1) spanned the centromere or overlapped the telomere (100 kb from the ends of the chromosome); 2) had > 50% of its length overlapping a segmental duplication; 3) had >50% overlap with immunoglobulin or T cell receptor. The final filtered CNV dataset was annotated with Refseq genes (transcriptions and exons). After this stage of quality

Nature Genetics: doi:10.1038/ng.3725

control (QC), we had a total of **52,511** individuals, with **27,034** SCZ cases and **25,448** controls.

Filtering for rare CNVs: To make our final dataset of rare CNVs for all subsequent analysis we universally filtered out variants that were present at >= 1% (50% reciprocal overlap) frequency in cases and controls combined. CNVs that overlapped > 50% with regions tagged as copy number polymorphic on any other platform were also excluded. CNVs < 20kb or having fewer than 10 probes were also excluded.

Post CNV calling QC: A number of steps were undertaken after CNV calling and initial filtering QC to minimize the impact of technical artifacts and potential confounds. In summary, we removed individuals not present in the PGC2 GWAS analysis [1], removed datasets with non-matching case or control samples that could not be reconciled using consensus platform probes, and removed any additional outliers with respect to overall CNV burden, CNV calling metrics, or SCZ phenotype residuals. All steps are described in more detail below. Lastly, we retained only individuals that also passed QC filtering from the companion PGC GWAS study in Schizophrenia [1]. This step filtered out samples with low-quality SNP genotyping, related individuals, and repeated samples across cohorts.

Individual dataset removal: Some datasets submitted to the PGC consisted of only case or control samples, affected trios, or recruited external samples as controls. This asymmetry in case-control ascertainment and genotyping can present serious biases for CNV analysis, as the sensitivity to detect CNV will vary considerably across genotyping platforms, as well as within dataset and genotyping batch. Unlike imputation protocols commonly used for SNP genotyping, there is no equivalent process to infer unmeasured probe intensity from nearby markers. We took a number of steps to identify and remove datasets that showed strong signs of case-control ascertainment or genotyping asymmetry:

Nature Genetics: doi:10.1038/ng.3725

1) Identify genotyping platforms where case-control ratio was not between 40-60%
2) Where possible, merge similar genotyping platforms using consensus probes prior to CNV-calling pipeline in order to improve case-control ratio.
3) Examine overall CNV burden and association peaks for spurious results
4) Remove datasets that remain problematic due to unusual CNV burden or multiple spurious CNV associations.

The genotyping platforms identified and processed are listed in **Supplementary Table 2**. We were able to combine the Illumina OmniExpress and Illumina OmniExpress plus Exome Chip platforms with success by removing probe content specific to the Exome chip platform. We removed the *caws* Affymetrix 500 datasets due to a number of strong CNV association peaks not seen in any other dataset. We also remove the *fii6* dataset due to a 2-fold CNV burden in cases relative to controls. In order to improve case-control balance, we had to remove the affected proband trio datasets (*boco*, *lacw*, and *lemu*) in the Illumina 610 platform, and the control-only *uclo* dataset in the Affymetrix 500 platform.

Individual sample removal: We re-analyzed CNV burden estimates in the reduced sample to flag any remaining outliers missed in the initial QC. We identified outliers for CNV count and Kb burden in the autosome (> 30 CNVs or 8 Mb, respectively) and in the X chromosome (> 10 CNVs or 5 Mb, respectively), removing an additional 15 individuals.

CNV burden between sexes

Following recent evidence that ostensibly healthy females carry an increased burden of rare CNVs [2], we examined whether this increased female burden existed in the current PGC dataset. We used a logistic regression model predicting sex using CNV burden and controlling for study covariates, as well as the Wilcoxon rank-sum test comparing male to female CNV burden [2]. Focusing on the significant findings in the previous paper, we examined the burden in autosomal CNV count and genes affected among PGC controls

Nature Genetics: doi:10.1038/ng.3725

(9856 males and 10371 females). We do see an elevated CNV count in control females (1.90 autosomal CNV rate) to males (1.87 autosomal CNV rate), however this difference is not significant in either the regression model (OR = 1.004, p = 0.66) or the Wilcoxon rank-sum test (p = 0.1). We do, however, observe a marginally significant enrichment when focusing on CNV loss count, where control females (0.99 autosomal CNV loss rate) show a higher burden than control males (0.94 autosomal CNV loss rate; logistic regression OR = 1.03, p = 0.05; Wilcoxon rank-sum test $p = 3 \times 10^{-3}$). No single genotyping platform seemed to drive the enrichment in females (data not shown), and we don't observe any difference in CNV count when looking at CNV gains (logistic regression OR = 0.98, p = 0.18; Wilcoxon rank-sum test p = 0.56). Finally, no significant differences between sexes were found using either test when examining the number of genes affected, or when we include SCZ cases and controls (all $p > .05$).

Breakpoint level power analysis

By restricting analysis to rare CNV in the population (MAF < 0.01), many loci do not have enough CNV to surpass genome-wide correction for multiple testing, prompting pathway and gene level analyses to achieve sufficient statistical power. To use a specific example, the 3q29 deletion is fully penetrant in the current sample, with 16 SCZ carriers and 0 controls (MAF = 3.8×10^{-4}) at the peak of association. Assuming no platform bias, this leads to an uncorrected chi-square p-value of 8.9×10^{-5}, and a permuted p-value of 6.2×10^{-5} testing association using SCZ phenotype residuals. Neither p-value, however, surpasses their respective genome-wide significance cutoff for CNV deletion. While permutation methods used to generate genome-wide cutoffs accurately reflect the testing space among observed CNVs (very rare CNVs have little to no contribution to the family-wise error rate), we wanted to estimate the proportion of CNV detectable at the breakpoint level. Under our current analytical design and sample size, we calculated the power to detect associated CNV across various MAFs and effect sizes and determine the proportion of association tests capable of surpassing genome-wide correction.

Nature Genetics: doi:10.1038/ng.3725

We simulated CNVs within our dataset (21,094 cases and 20,227 controls) and regressed them using the same association design with SCZ residual phenotypes. We simulated various effect sizes by randomly sampling cases and controls at different probabilities as CNV carriers, and rounded to the nearest CNV count to reflect the MAF of each CNV in the sample. For each combination of effect size and MAF, we ran 1000 simulations, retrieving the *t*-test *p*-value of CNV carriers from the SCZ residual phenotype. Simulated *p*-values behaved in much the same way as the Z-score correction on permutated *p*-values used in the primary test (data not shown).

We define statistical power as the proportion of simulations surpassing genome-wide significance. For a fully penetrant risk CNV, we require a MAF of ~6 $x10^{-4}$ (or about 25 CNV) to achieve 80% detection power. For CNV with a genotype relative risk (GRR) of 10, we require a MAF of $1e^{-3}$ (or at least 41 CNV) to achieve 80% detection power. Looking across the landscape of CNVs tested, on the whole about 10% of deletion or duplication CNV breakpoints reach a frequency greater than $1e^{-3}$ in the sample. On the other extreme, a CNV with MAF of .005 (or at least 206 CNV) and a GRR of 2 will only be detected 58% of the time.

Controlling for population stratification

We accounted for population stratification across cohorts using principal components based on SNP genotypes derived from the same genotyping arrays. After the post-CNV calling quality control steps described below, we re-calculated principal components using the Eigenstrat software package [3]. Sample information and subsequent CNV and GWAS filtered sample sets are presented in **Supplementary Table 2**. In the process of matching to the GWAS-specific cohort, all individuals of non-European ancestry were removed from analysis (~5.8% of the post-QC sample comprising three separate datasets). We also removed 42 samples that had discordant phenotype designations between the GWAS analysis and CNV genotype submission.

Nature Genetics: doi:10.1038/ng.3725

Gene-set burden enrichment analysis: selection of gene-sets and pre-processing of CNV calls.

Gene-sets with an a priori expectation of association to neuropsychiatric disorders were compiled based on gene annotations (Gene Ontology and curated pathway databases, downloaded June 2013) and published article materials (for details, see **Supplementary Table 3**). Gene-sets based on brain expression were compiled by processing the BrainSpan RNA-seq gene expression data-set (http://www.brainspan.org/static/download.html, downloaded Sept 2012). Four roughly equally sized gene-sets (about 4600 genes each) were derived to represent four expression tiers (very high, medium-to-high, medium-to-low, very low or absent); genes were selected if they passed a fixed expression threshold in at least 5/508 experimental data points (corresponding to different regions of donor brains, different donor ages corresponding to different developmental brain stages, and different donor sexes). Gene-sets based on mouse phenotypes were assembled by downloading MPO (Mammalian Phenotype Ontology) annotations from MGI (www.informatics.jax.org, downloaded August 2013), up-propagating annotations following ontology relations, and mapping to human orthologs using NCBI Homologene (www.ncbi.nlm.nih.gov/homologene); finally, top-level organ systems with fewer genes were aggregated while striving to preserve biological homogeneity, so to have roughly equal-sized sets (2,600-1,300 genes). For all gene-sets, gene identifiers in the primary source were mapped to Entrez-gene identifiers using the R/Bioconductor package *org.Hs.eg.db*.

Subjects were restricted to the ones with at least one rare CNV. For copy number gains and losses, we separately calculated the following subject-level totals: variant number, variant length and number of genes impacted; these covariates are then used to model global burden and correct gene-set burden to ensure it is specific (i.e. not a mere reflection of genome-wide burden with some stochastic deviation due to sampling). The subject-level total number of genes impacted was also calculated for each gene-set,

Nature Genetics: doi:10.1038/ng.3725

again separately for gains and losses. Subjects were flagged if they carried at least one CNV matching a locus previously implicated in schizophrenia (see section "Identifying previously implicated CNV loci in the literature"); this was then used to analyzed gene-set burden for all subjects, or excluding subjects with an already implicated CNV.

Genome-wide correction for multiple testing

Breakpoint association test: Beyond identifying significant CNV at the breakpoint level, we also estimated the genome-wide testing space for rare CNV analysis. With the large PGC cohort being called through a consistent pipeline, we saw an opportunity to characterize the null expectation of segregating and recurrent *de novo* rare CNV in populations of European ancestry. Accepted thresholds for significance among published risk CNV have been limited in scope, as accurate population estimates of rare CNV frequency and distribution across the genome require large representative samples.

Genome-wide significance thresholds were calculated using the 5% family-wise error rate from 5,000 permutations in both the SCZ residual phenotype and CMH test. Specifically, we selected the 95^{th} percentile of the minimum *p*-values obtained across permutations. Below are the genome-wide correction *p*-value thresholds determined in this manner:

SCZ residual phenotype FWER correction:

CNV losses and gains: 6.73×10^{-6}

CNV losses: 1.5×10^{-5}

CNV gains: 1.35×10^{-5}

CMH test FWER correction:

CNV losses and gains: 3.65×10^{-5}

CNV losses: 8.25×10^{-5}

Nature Genetics: doi:10.1038/ng.3725

CNV gains: $7.8x10^{-5}$

This method differs slightly from those used in Levinson et al. [4] to estimate the multiple test correction for rare CNV, however their genome-wide correction of $p = 1x10^{-5}$ corresponds quite closely to the estimates observed using the SCZ residual phenotype. The observed family-wise error correction serves as good approximation of the independent rare CNV signals found among European ancestry populations for array-based CNV detection, but as sample sizes increase, so too will the effective number of tests, necessitating further evaluation of the multiple testing burden. We also applied a Benjamini-Hochberg False Discovery rate (BH-FDR) to test the proportion of previously reported loci reaching suggestive significance. BH-FDR was applied to CNV losses and gains, losses, and gains separately. Drawing form the gene association analysis, we restricted our BH-FDR correction to gains with at least 12 CNV carriers (21% of CNV gain breakpoints), and losses with at least 8 CNV carriers (28% of CNV loss breakpoints).

Gene-based association test: Association was tested for loci rather than for genes, to avoid the strong correlation between test introduced by multi-genic CNVs; for the same reason, it is more useful to count false positives as loci rather than genes. We followed a greedy step-down procedure:

- start from gene with most significant deviance p-value G1, create locus L1
- remove from the gene list all genes that share at least 50% of their carrier subjects with G1, and add them to locus L1
- do the same for the next gene most significant gene in the list (thus creating a new locus L2), and proceed recursively until there is no gene left
- define locus p-value as the smallest deviance p-value of its genes

We computed permutation-based FDR by permuting subjects' condition labels (schizophrenia, control), but not covariates (as those are expected to correlate to CNV distribution), 1,000 times. The FDR was then defined as the ratio between the average number of tests passing a given p-value threshold across the 1,000 permutations and

Nature Genetics: doi:10.1038/ng.3725

the number of tests passing the same p-value threshold for real data. FDRs were also generated counting only the subset of genes with positive and negative regression coefficients (i.e. risk and presumed protective). The p-value threshold for permutation-based FDR calculation was picked by choosing the maximum nominal p-value corresponding to a given BH-FDR threshold (e.g. 5%). BH-FDR is supposed to be slightly inflated because (i) the deviance test p-value is slightly under-conservative in presence of very few gene indicators different than 0, (ii) we use the smallest gene p-value to define the locus p-value.

Gene-based network analysis

To identify a gene network enriched in schizophrenia risk genes, we queried GeneMANIA [5] using the 17 genes with deletion gene-test Benjamini-Hochberg FDR <= 25% and member of the "GO synaptic" or "ARC complex" sets. We thus created a synaptic protein interaction network of 136 genes, with the most densely connected network core corresponding to post-synaptic density organizers (DLGs, DLGAPs, SHANKs) and ionotropic glutamate receptors (GRIAs, GRIDs, GRINs). NRXN1 is connected to the network core via adhesion partners (NLGN1-3) and CASK. We tested this schizophrenia gene network, and found significant enrichment in genes with evidence of de novo coding variants in sequencing studies of schizophrenia trios [6] (for frameshift, stop-gain and splice-site: Fisher's Exact Test p-value 0.0023; missense and amino acid insertion/deletion: Fisher's Exact Test p-value 0.0004); in addition, we found a greater enrichment for this network, compared to the larger set composed of all "GO synaptic" and "ARC complex" genes. No significant enrichment was found for de novo variants identified in controls.

Follow up of significant CNV loci

Both gene and breakpoint level association follow a uniform testing framework across the genome, however risk loci may exhibit a more nuanced CNV architecture across the entirety of the association peak. All associated loci with FDR < .05 in the gene based test

Nature Genetics: doi:10.1038/ng.3725

were followed up for further testing, along with a small number of candidate loci showing suggestive association in the breakpoint-level association. We visually inspected each association peak and determined the bp coordinates that encapsulate the associated region and determine which CNV segment inclusion, be it covering exons or overlapping a minimum percentage of the total region, most appropriately reflect the association signal. To comprehensively examine the robustness and source of association, we also ran additional tests controlling for individual dataset, splitting by sex, and examining a dosage model, whereby copy number is measured with one copy for deletion, two copies for no CNV, and three copies for duplication. We also examined significant CNV loci in an unfiltered CNV call set, using CNVs called prior to the removal of common CNVs (MAF > 1%) and CNV overlapping segmental duplications.

We further evaluated the associated regions by determining the concordance of calls within the call set with those determined by unsupervised clustering. Call set CNVs were defined as CNVs with at least a 50% overlap with regions in **Table 1**. We restricted this analysis to 26,959 samples across six cohorts (14,419 Affymetrix 6.0, 12,540 Illumina platforms; 1.1:1 case:control ratio). Features for clustering included the median logR ratio (mLRR) and the median logR ratio of the chromosome for which a locus resides in, controlling for large chromosomal abnormalities. We implemented Density-Based Spatial Clustering of Applications (DBSCAN) found in the python scikit-learn library (http://scikit-learn.org) because of high sensitivity to detect outliers in clusters. For each novel region and within each cohort, genotypes were assigned to every sample based on the DBSCAN defined cluster. The cluster with the highest number of samples was designated as reference and assumed to have a copy number of two. Other clusters were flagged as gain or loss based on the average regional mLRR and its relation to the reference regional mLRR. We removed clusters with average chromosomal mLRR outside 3 SD from the reference. CNVs were considered concordant if they were flagged non-reference by DBSCAN and present in the 41k call set, matching on CNV type. We applied a locus based call set concordance filter of >=70%; one region, NPY4R, failed to

Nature Genetics: doi:10.1038/ng.3725

meet this requirement with a concordance of 0.1%. In addition, both proximal and distal loci of ZNF600 were removed due to batch effects, which we defined as a significant deviation from a Poisson distribution of call set calls per plate. Regions that passed both concordance and batch effect filters are reported in **Table 1.**

Proportion of variance in SCZ explained by top CNV loci

To measure the proportion of variance explained on the liability scale of SCZ, we estimated the overall heritability of liability (or logRR genetic variance) explained by the eight CNV loci surpassing genome-wide significance. All eight loci were collapsed into a single signal. Two SCZ affected individuals were found to carry two CNVs in these loci, and their contribution was only counted once. In sum, we observed 298 SCZ patients with a CNV in these regions (1.4% of the total SCZ affected sample), and 29 controls (0.1%; CMH stratified OR = 10.1). To estimate the variance in SCZ liability explained by loci surpassing genome-wide correction, we calculated the heritability of liability using the INDI-V online tool (cnsgenomics.com/software) [7] using an overall disease risk of 1% and a sibling recurrence risk of 8.8 [8].

NAHR enrichment in significant novel gene loci

To test if novel significant loci (FDR<0.05; **Table 1**) were enriched for NAHR events, we performed a permutation test (n=10,000) simulating the null distribution of NAHR-mediated CNVs for a set of random loci. Each simulation randomly selected nine loci taken from CNVs overlapping at least 50% to genes in the gene-set burden analysis. These nine random loci were matched according to CNV call frequency to the nine novel significant loci in **Table 1**. We then created windows for each start and end position for every overlapping CNV to a random locus. Start positions were expanded -50kb and +5kb, and end positions were expanded -5kb and +50kb. We flagged CNVs as NAHR-mediated when both start and end expanded windows overlapped to 1kb segmental duplications obtained from the hg18 build of the UCSC table browser (https://genome.ucsc.edu/cgi-bin/hgTables). Every iteration reported the fraction of

Nature Genetics: doi:10.1038/ng.3725

NAHR-mediated CNVs; that is the ratio of CNVs flagged as NAHR to the total number of overlapping CNVs. We found and enrichment of NAHR mediated CNVs in significant novel loci when compared to the null distribution (86% NAHR-mediated, 6 fold enrichment, p=0.008).

Identifying previously implicated CNV loci in the literature

To delineate CNV burden effects coming from CNV loci that have previously been reported as putative SCZ risk factors from CNV in remainder of the genome, we flagged CNV loci with $p < 0.01$ that have either been reviewed [9,10] or otherwise reported as potential SCZ risk factors in the literature. Previously reported loci meeting inclusion are listed in **Supplementary Table 1**. While a number of CNV loci have been reported in multiple studies, we sought the most recent reports that incorporated the largest sample sizes. To identify putatively associated CNV loci with SCZ from the full list, we applied the genome-wide *p*-value cutoff of $8e^{-5}$, derived from the Cochran-Mantel-Haenzel (CMH) test in the current breakpoint-level analysis as the *p*-value cutoff for inclusion as SCZ implicated CNV loci. While the CMH test is not the primary breakpoint-level test in the current PGC analysis, it corresponds more closely to the tests used in published reports. In all, nine independent CNV loci from published reports surpass genome-wide correction. All published CNV loci, even those excluded as an SCZ implicated regions, are examined in the breakpoint-level association analysis.

Subsets of the PGC schizophrenia dataset that have been published previously

The PGC encourages all participants that provide data to publish their independent studies. While this study was underway, CNV studies were published on overlapping sets of the samples including cohorts from Sweden [11] and Germany [12] and Japan [13]. Cases from the CLOZUK study have been used in recent studies investigating the association of candidate pathways [14] and loci [15]. Other studies that have been published independently including Molecular Genetics of Schizophrenia (MGS) [4] and the international Schizophrenia Consortium (ISC) [16].

Nature Genetics: doi:10.1038/ng.3725

Consortium Membership

Psychosis Endophenotypes International Consortium

Maria J Arranz[137,177], Steven Bakker[92], Stephan Bender[178,179], Elvira Bramon[160], David A Collier[120,163] , Benedicto Crespo-Facorro[43,44], Jeremy Hall[116], Conrad Iyegbe[137], Assen V Jablensky[33,93,159,166], René S Kahn[31], Luba Kalaydjieva[92,93], Stephen Lawrie[116], Cathryn M Lewis[137], Kuang Lin[137], Don H Linszen[82], Ignacio Mata[43,44], Andrew M McIntosh[113,114], Robin M Murray[125], Roel A Ophoff[31,36,73], Jim Van Os [126,136], John Powell[137], Dan Rujescu[77,78], Muriel Walshe[137], Matthias Weisbrod[179], Durk Wiersma[25,180]

[177]Fundació de Docència i Recerca Mútua de Terrassa, Universitat de Barcelona, Spain. [178]Child and Adolescent Psychiatry, University of Technology Dresden, Dresden, Germany. [179]Section for Experimental Psychopathology, General Psychiatry, Heidelberg, Germany. [180]Department of Statistics, University of Oxford, Oxford, UK

Acknowledgements

Data Processing and Statistical analyses were carried out on the Genetic Cluster Computer (http://www.geneticcluster.org) hosted by SURFsara and financially supported by the Netherlands Scientific Organization (NWO 480-05-003) along with a supplement from the Dutch Brain Foundation and the VU University Amsterdam. The GRAS data collection was supported by the Max Planck Society, the Max-Planck-Förderstiftung, and the DFG Center for Nanoscale Microscopy & Molecular Physiology of the Brain (CNMPB), Göttingen, Germany. The Boston CIDAR subject and data collection was supported by the National Institute of Mental Health (1P50MH080272, RWM; U01MH081928, LJS; 1R01MH092380, TLP) and the Massachusetts General Hospital Executive Committee on Research (TLP). ISC – Portugal: CNP and MTP are or have been supported by grants from the NIMH (MH085548, MH085542, MH071681, MH061884, MH58693, and MH52618) and the NCRR (RR026075). CNP, MTP, and AHF are or have been supported by grants from the Department of Veterans Affairs Merit Review

Nature Genetics: doi:10.1038/ng.3725

Program. The Danish Aarhus study was supported by grants from The Lundbeck Foundation, The Danish Strategic Research Council, Aarhus University, and The Stanley Research Foundation. Work in Cardiff was supported by MRC Centre (G0800509) and MRC Programme (G0801418) Grants, the European Community's Seventh Framework Programme (HEALTH-F2-2010-241909 (Project EU-GEI)), the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 279227, a fellowship to JW from the MRC/Welsh Assembly Government and the Margaret Temple Award from the British Medical Association. We thank Novartis for their input in obtaining CLOZUK samples, and staff at The Doctor's Laboratory (Lisa Levett/ Andrew Levett) for help with sample acquisition and data linkage and in Cardiff (Kiran Mantripragada/Lucinda Hopkins) for sample management. CLOZUK and some other samples were genotyped at the Broad Institute (which has a separate acknowledgment) or by the WTCCC and WTCCC2 (WT (083948/Z/07/Z). We acknowledge use of the British 1958 Birth Cohort DNA (MRC: G0000934) and the Wellcome Trust (068545/Z/0/ and 076113/C/04/Z), the UK Blood Services Common Controls (UKBS-CC collection), funded by the WT (076113/C/04/Z) and by NIHR programme grant to NHSBT (RP-PG-0310-1002). Virginia Commonwealth University: BPR and KSK thank all the faculty of the Virginia Institute for Psychiatric and Behavioral Genetics for invaluable insights and discussions over many years. BSM, SAB, BTW, BW, KSK and BPR were supported by National Institute of Mental Health grant R01 MH083094 to BPR. Sample collection was supported by previous funding of National Institute of Mental Health grant R01 MH041953 to KSK and BPR. Genotyping was supported by National Institute of Mental Health grant R01 MH083094 to BPR, National Institute of Mental Health grant R01 MH068881 to BPR and Wellcome Trust Case Control Consortium 2 grant. We thank Novartis for their input in obtaining CLOZUK samples, and staff at The Doctor's Laboratory (Lisa Levett/ Andrew Levett) for help with sample acquisition and data linkage and in Cardiff (Kiran Mantripragada/Lucinda Hopkins) for sample management. Our work was supported by: Medical Research Council (MRC) Centre (G0800509; G0801418), the European Community's Seventh Framework Programme (HEALTH-F2-

Nature Genetics: doi:10.1038/ng.3725

2010-241909 (Project EU-GEI)), the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 279227, a fellowship to JW from the MRC/Welsh Assembly Government and the Margaret Temple Award from the British Medical Association. CLOZUK and some other samples were genotyped at the Broad Institute (which has a separate acknowledgment) or by the WTCCC and WTCCC2 (WT (083948/Z/07/Z). We acknowledge use of the British 1958 Birth Cohort DNA (MRC: G0000934) and the Wellcome Trust (068545/Z/0/ and 076113/C/04/Z), the UK Blood Services Common Controls (UKBS-CC collection), funded by the WT (076113/C/04/Z) and by NIHR programme grant to NHSBT (RP-PG-0310-1002). The recruitment of families in Bulgaria was funded by the Janssen Research Foundation, Beerse, Belgium. We are grateful to the study volunteers for participating in the Janssen research studies and to the clinicians and support staff for enabling patient recruitment and blood sample collection. Informed consent was obtained from all participants or their parents or guardians. We thank the staff in the Neuroscience Biomarkers Genomic Lab led by Reyna Favis at Janssen for sample processing and the staff at Illumina for genotyping Janssen DNA samples. We also thank Anthony Santos, Nicole Bottrel, Monique-Andree Franc, William Cafferty of Janssen Research & Development) for operational support. Funding from the Netherlands Organization for Health Research and Development (ZonMw), within the Mental Health program (to GROUP consortium for collecting patients and clinical data). High-Density Genome-Wide Association Study Of Schizophrenia In Large Dutch Sample (R01 MH078075 NIH/National Institute Of Mental Health PI: Roel A. Ophoff). The Danish Council for Strategic Research (Journ.nr. 09-067048); The Danish National Advanced Technology Foundation (Journ.nr. 001-2009-2); The Lundbeck Foundation (Journ.nr. R24-A3243); EU 7th Framework Programme (PsychGene; Grant agreement nr. 218251); EU 7th Framework Programme (PsychDPC; Grant agreement nr. 286213). The Wellcome Trust supported this study as part of the Wellcome Trust Case Control Consortium 2 project. E. Bramon holds a MRC New Investigator Award and a MRC Centenary Award. The TOP Study was supported by the Research Council of Norway (#213837, # 217776, # 223273), South-East Norway Health

Nature Genetics: doi:10.1038/ng.3725

Authority (# 2013-123) and K.G. Jebsen Foundation. This work was supported by the Donald and Barbara Zucker Foundation, the North Shore – Long Island Jewish Health System Foundation, and grants from the Stanley Foundation (AKM), the National Alliance for Research on Schizophrenia and Depression (AKM), and the NIH (MH065580 to TL; MH001760 to AKM). SynSys, EU FP7-242167, Sigrid Juselius Foundation, The Academy of Finland, grant number: 251704, Sohlberg Foundation. The Swedish Research Council [grant numbers 2006-4472, 2009-5269, 2009-3413] and the County Councils of Västerbotten and Norrbotten, Sweden supported the collection of the scz_umeb_eur and scz_umes_eur samples. The Betula Study, from which the Umea controls were recruited, is supported by grants from the Swedish Research Council [grant numbers 345-2003-3883, 315-2004-6977] and the Bank of Sweden Tercentenary Foundation, the Swedish Council for Planning and Coordination of Research, the Swedish Council for Research in the Humanities and Social Sciences and the Swedish Council for Social Research. The GRAS (Göttingen Research Association for Schizophrenia) data collection has been supported by the Max Planck Society, the Max Planck Förderstiftung, and the DFG (CNMPB). We thank all GRAS patients for participating in the study, and all the many colleagues who have contributed over the past 10 years to the GRAS data collection. We acknowledge support from the North Shore – LIJ Health System Foundation and NIH grants RC2 MH089964 and R01 MH084098. We acknowledge support from NIMH K01 MH085812 (PI Keller) and NIMH R01 MH100141 (PI Keller). EGCUT work was supported by the Targeted Financing from the Estonian Ministry of Science and Education [SF0180142s08]; the US National Institute of Health [R01DK075787]; the Development Fund of the University of Tartu (grant SP1GVARENG); the European Regional Development Fund to the Centre of Excellence in Genomics (EXCEGEN; grant 3.2.0304.11-0312); and through FP7 grant 313010. Milan Macek was supported by CZ.2.16/3.1.00/24022OPPK, NT/13770–4and 00064203 FN Motol. For the scz_tcr1_asn dataset funding from the National Medical Research Council (Grant: NMRC/TCR/003/2008) and the Biomedical Research Council, A*STAR is acknowledged. Genotyping of the Swedish Hubin sample was performed by

Nature Genetics: doi:10.1038/ng.3725

the SNP&SEQ Technology Platform in Uppsala, which is supported by Uppsala University, Uppsala University Hospital, Science for Life Laboratory - Uppsala and the Swedish Research Council (Contracts 80576801 and 70374401). The Swedish Hubin sample was supported by Swedish Research Council (IA, EGJ) and the regional agreement on medical training and clinical research between Stockholm County Council and the Karolinska Insititutet (EGJ). B.J.M., V.J.C., R.J.S., S.V.C., F.A.H., A.V.J., C.M.L., P.T.M., C.P., and U.S. were supported by the Australian Schizophrenia Research Bank, which is supported by an Enabling Grant from the National Health and Medical Research Council (Australia) [No. 386500], the Pratt Foundation, Ramsay Health Care, the Viertel Charitable Foundation and the Schizophrenia Research Institute and the NSW Department of Health. C.P. is supported by a Senior Principal Research Fellowship from the National Health and Medical Research Council (Australia). We acknowledge the help of: Johanna Badcock, Linda Bradbury, Jason Bridge, David Chandler, Janell Collins-Langworthy, Trish Collinson, Milan Dragovic, Cheryl Filippich, David Hawkes, Danielle Lowe , Kathryn McCabe, Tamara MacDonald, Barry Maher, Bharti Morar Marc Seal, Heather Smith, Melissa Tooney, Paul Tooney, and Melinda Ziino. The Danish Aarhus study was supported by grants from The Lundbeck Foundation, The Danish Strategic Research Council, Aarhus University, and The Stanley Research Foundation. The Perth sample collection was funded by Australian National Health and Medical Research Council project grants and the Australian Schizophrenia Research Bank. The Bonn/Mannheim sample was genotyped within a study that was supported by the German Federal Ministry of Education and Research (BMBF) through the Integrated Genome Research Network (IG) MooDS (Systematic Investigation of the Molecular Causes of Major Mood Disorders and Schizophrenia; grant 01GS08144 to M.M.N. and S.C., grant 01GS08147 to M.R.), under the auspices of the National Genome Research Network plus (NGFNplus), and through the Integrated Network IntegraMent (Integrated Understanding of Causes and Mechanisms in Mental Disorders), under the auspices of the e:Med Programme.(GSK control sample; Müller-Myhsok). This work has been funded by the Bavarian Ministry of Commerce and by the Federal Ministry of Education

Nature Genetics: doi:10.1038/ng.3725

and Research in the framework of the National Genome Research Network, Förderkennzeichen 01GS0481 and the Bavarian Ministry of Commerce. M.M.N. is a member of the DFG-funded Excellence-Cluster ImmunoSensation. M.M.N. also received support from the Alfried Krupp von Bohlen und Halbach-Stiftung. M.R. was also supported by the 7th Framework Programme of the European Union (ADAMS project, HEALTH-F4-2009-242257; CRESTAR project, HEALTH-2011-1.1-2) grant 279227. Roche: Thanks are expressed to Olivia Spleiss for great support in genetic data generation, Daniel Umbricht and Delphine Lagarde for their valuable support in clinical and genetic data sharing, and Anirvan Ghosh for continuous encouragement. Authors also wish to thank all investigators and patients who participated in the Roche clinical studies. Jo Knight holds the Joanne Murphy Professor in Behavioural Science. We thank Maria Tampakeras for her work on the samples. The Stanley Center for Psychiatric Research at the Broad Institute acknowledges funding from the Stanley Medical Research Institute. Swedish schizophrenia study (PI CMM, PFS, PS, SM): We are deeply grateful for the participation of all subjects contributing to this research and to the collection team that worked to recruit them: E. Flordal-Thelander, A.-B. Holmgren, M. Hallin, M. Lundin, A.-K. Sundberg, C. Pettersson, R. Satgunanthan-Dawoud, S. Hassellund, M. Rådstrom, B. Ohlander, L. Nyrén and I. Kizling. Funding support for the Sweden Schizophrenia Study (PIs Hultman, Sullivan, and Sklar) was provided by the NIMH (R01 MH077139 to P.F.S. and R01 MH095034 to P.S.), the Stanley Center for Psychiatric Research, the Sylvan Herman Foundation, the Friedman Brain Institute at the Mount Sinai School of Medicine, the Karolinska Institutet, Karolinska University Hospital, the Swedish Research Council, the Swedish County Council, the Söderström Königska Foundation. We acknowledge use of DNA from The UK Blood Services collection of Common Controls (UKBS collection), funded by the Wellcome Trust grant 076113/CI04/Z, by the Juvenile Diabetes Research Foundation grant WT0618S8, and by the National Institute of Health Research of England. The collection was established as part of the Wellcome Trust Case-Control Consortium. We thank the study participants, and the research staff at the study sites. This study was supported by NIMH grant R01MH062276 (to DF Levinson, C

Nature Genetics: doi:10.1038/ng.3725

Laurent, M Owen and D Wildenauer), grant R01MH068922 (to PV Gejman), grant R01MH068921 (to AE Pulver) and grant R01MH068881 (to B Riley). The authors are grateful to the many family members who participated in the studies that recruited these samples, to the many clinicians who assisted in their recruitment. In addition to the support acknowledged for the Multicenter Genetics Studies of Schizophrenia and Molecular Genetics of Schizophrenia studies, Dr. DF Levinson received additional support from the Walter E. Nichols, M.D., Professorship in the School of Medicine, the Eleanor Nichols Endowment, the Walter F. & Rachael L. Nichols Endowment and the William and Mary McIvor Endowment, Stanford University. This study was supported by NIH R01 grants (MH67257 to N.G.B., MH59588 to B.J.M., MH59571 to P.V.G., MH59565 to R.F., MH59587 to F.A., MH60870 to W.F.B., MH59566 to D.W.B., MH59586 to J.M.S., MH61675 to D.F.L., MH60879 to C.R.C., and MH81800 to P.V.G.), NIH U01 grants (MH46276 to C.R.C., MH46289 to C. Kaufmann, MH46318 to M.T. Tsuang, MH79469 to P.V.G., and MH79470 to D.F.L.), the Genetic Association Information Network (GAIN), and by The Paul Michael Donovan Charitable Foundation. Genotyping was carried out by the Center for Genotyping and Analysis at the Broad Institute of Harvard and MIT (S. Gabriel and D. B. Mirel), which is supported by grant U54 RR020278 from the National Center for Research Resources. Genotyping of half of the EA sample and almost all the AA sample was carried out with support from GAIN. The GAIN quality control team (G.R. Abecasis and J. Paschall) made important contributions to the project. We thank S. Purcell for assistance with PLINK. We (DRW, RS) thank the staff of the Lieber Institute and the Clinical Brain Disorders Branch of the IRP, NIMH for their assistance in data collection and management. We thank Ningping Feng and Bhaskar Kolachana for Illumina genotyping and for managing DNA stocks. The work was supported by the Lieber Institute and by direct NIMH IRP funding of the Weinberger Lab. Pfizer is very grateful to the study volunteers for participating in our research studies. We thank our numerous clinicians and support staff for enabling patient recruitment, blood sample collection, and biospecimen administration. Informed consent was obtained from all participants, their parents or guardians. Eli Lilly is grateful to the participants of clinical

Nature Genetics: doi:10.1038/ng.3725

trials and research studies who gave consent for participation in this study. We are also grateful to Philip J Ebert and Jeffrey S Arnold for facilitating our participation in this project. We acknowledge the Irish contribution to the International Schizophrenia Consortium (ISC) study, the WTCCC2 SCZ study & WTCCC2 controls from the 1958BC and UKNBS, the Science Foundation Ireland (08/IN.1/B1916). We thank the Toronto Centre for Applied Genomics for technical and computational assistance and funding from the University of Toronto McLaughlin Centre and Genome Canada. S.W.S. holds the GlaxoSmithKline-CIHR Chair in Genome Sciences at the Hospital for Sick Children and University of Toronto. We acknowledge use of the Trinity Biobank sample from the Irish Blood Transfusion Service & the Trinity Centre for High Performance Computing. Funding for this study was provided by the Wellcome Trust Case Control Consortium 2 project (085475/B/08/Z and 085475/Z/08/Z), the Wellcome Trust (072894/Z/03/Z, 090532/Z/09/Z and 075491/Z/04/B), NIMH grants (MH 41953 and MH083094) and British 1958 Birth Cohort DNA collection funded by the Medical Research Council (grant G0000934) and the Wellcome Trust (grant 068545/Z/02) and of the UK National Blood Service controls funded by the Wellcome Trust. We acknowledge Hong Kong Research Grants Council project grants GRF 774707M, 777511M, 776412M and 776513M. D.A. acknowledges support from the NIH Predoctoral training grant T32 GM008666. W.M.B. acknowledges support from the Autism Science Foundation (REG-15-001). J.S. acknowledges support from the NIMH (MH094421, MH076431), Simons Foundation (SFARI 275724) and the Beyster Family Foundation.

PGC Data Sharing Policy

The PGC utilizes data from a large number of different studies, each funded independently of the PGC and each with their own IRB approval and informed consent documentation and data restrictions. Approval has been obtained from all participating PIs to share summary data from genetic studies (gene-based and breakpoint association). These summary data are provided to the community (see Ricopilli). The CNV calls can be visualized on the PGC CNV browser

Nature Genetics: doi:10.1038/ng.3725

(http://pgc.tcag.ca/gb2/gbrowse/pgc_hg18/). In addition, we the CNV can be obtained via European Genome-Phenome Archive (Study accession #EGAS00001001960). In order to comply with privacy-protection regulations of all institutions and countries that contributed to this study, CNV calls are provided without study or individual identifiers. Sample-level genotype data is the property of the individual studies/PIs. Public release of sample-level genotypes is at the discretion of the individual studies/PIs and their funding agencies.

Nature Genetics: doi:10.1038/ng.3725

References

1. Schizophrenia Working Group of the Psychiatric Genomics, C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-7 (2014).
2. Desachy, G. *et al.* Increased female autosomal burden of rare copy number variants in human populations and in autism families. *Mol Psychiatry* **20**, 170-5 (2015).
3. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-909 (2006).
4. Levinson, D.F. *et al.* Copy number variants in schizophrenia: confirmation of five previous findings and new evidence for 3q29 microdeletions and VIPR2 duplications. *Am J Psychiatry* **168**, 302-16 (2011).
5. Zuberi, K. *et al.* GeneMANIA prediction server 2013 update. *Nucleic Acids Res* **41**, W115-22 (2013).
6. Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179-84 (2014).
7. Witte, J.S., Visscher, P.M. & Wray, N.R. The contribution of genetic variants to disease depends on the ruler. *Nat Rev Genet* **15**, 765-76 (2014).
8. Sullivan, P.F., Kendler, K.S. & Neale, M.C. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Arch Gen Psychiatry* **60**, 1187-92 (2003).
9. Malhotra, D. & Sebat, J. CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell* **148**, 1223-41 (2012).
10. Rees, E. *et al.* Analysis of copy number variations at 15 schizophrenia-associated loci. *Br J Psychiatry* **204**, 108-14 (2014).
11. Szatkiewicz, J. *et al.* Copy number variation in schizophrenia in Sweden. *Molecular Psychiatry* **19**, 762-73 (2014).
12. Degenhardt, F. *et al.* Duplications in RB1CC1 are associated with schizophrenia; identification in large European sample sets. *Transl Psychiatry* **3**, e326 (2013).
13. Kushima, I. *et al.* High-resolution copy number variation analysis of schizophrenia in Japan. *Mol Psychiatry* (2016).
14. Pocklington, A.J. *et al.* Novel Findings from CNVs Implicate Inhibitory and Excitatory Signaling Complexes in Schizophrenia. *Neuron* **86**, 1203-14 (2015).
15. Grozeva, D. *et al.* Independent estimation of the frequency of rare CNVs in the UK population confirms their role in schizophrenia. *Schizophr Res* **135**, 1-7 (2012).
16. Stone, J.L. *et al.* Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* (2008).

Nature Genetics: doi:10.1038/ng.3725

Supplementary Table 1: Previously reported CNV association

Locus	CNV type	Gene or region name	Initial SCZ association reference (see legend)	Tested in Rees et al. 2014	Reported *p*-value	SCZ CNV carrier %	Control CNV carrier %	Reported Odds Ratio
22q11.2	deletion	multigenic	1	yes	**4.40E-40**	0.29	0	Inf
16p11.2	duplication	proximal duplication	2	yes	**2.90E-24**	0.35	0.03	11.52
1q21.1	deletion	multigenic	3,4	yes	**4.10E-13**	0.17	0.021	8.35
2p16.3	deletion	NRXN1 exons	5,6	yes	**1.30E-11**	0.18	0.02	9.01
15q11.2	deletion	multigenic	3	yes	**2.50E-10**	0.59	0.28	2.15
3q29	deletion	multigenic	7,11	yes	**1.50E-09**	0.082	0.0014	57.65
15q13.2-13.3	deletion	multigenic	3,4	yes	**5.60E-06**	0.14	0.019	7.52
15q11.2-13.1	duplication	AS/PWS	8	yes	**5.60E-06**	0.083	0.0063	13.2
8q11.23	duplication	RB1CC1	9	no	**1.29E-05**	0.106	0.014	8.58
16p13.11	duplication	multigenic	8	yes	5.70E-05	0.31	0.13	2.3
7q11.23	duplication	Williams-Beuren	10	yes	6.90E-05	0.066	0.0058	11.35
1q21.1	duplication	multigenic	11	yes	9.90E-05	0.13	0.037	3.45
16p13.2	duplication	C16orf72/USP7	11	no	1.00E-04	0.254	0.0197	12.9
1p36.33	duplication	multigenic	12	no	5.00E-04	0.065	0.0075	8.66
22q11.2	duplication	multigenic	13	no	8.60E-04	0.014	0.085	0.17
17p12	deletion	HNPP	14	yes	1.20E-03	0.094	0.026	3.62
9q34.3	duplication	intergenic	15	no	1.40E-03	1.47	0.43	3.38
16p12.1	deletion	multigenic	12	no	1.60E-03	0.15	0.057	2.72
15q21.3	duplication	CGNL1	12	no	1.90E-03	0.32	0.19	1.71
11q25	deletion	GLB1L3/GLB1L2	11	no	3.00E-03	0.38	0.123	3
2q37.3	duplication	AQP12A/KIF1A	12	no	3.00E-03	0.34	0.24	1.43

Nature Genetics: doi:10.1038/ng.3725

17q12	deletion	RCAD	16	yes	0.0072	0.036	0.0054	6.64
9p24.2	deletion	GLIS3	12	no	8.40E-03	0.033	0	Inf
9p24.2	deletion	SLC1A1	12	no	9.80E-03	0.047	0.0075	6.19
16p11.2	deletion	distal deletion	17	yes	0.017	0.063	0.018	3.39
7q36.3	duplication	WDR60/VIPR2	11,18	yes	0.27	0.11	0.069	1.54

Legend: We assembled a list of 26 reported CNV associations to SCZ where an odds ratio and *p*-value were available. At each CNV locus, we list the odds ratios and *p*-values from the largest sample collection available in the literature. Results from all CNV loci meta-analyzed in Rees et al. (2014), when available, were used. Throughout this article we refer to this entire list as "previously reported" loci. Reported *p*-values for nine loci shown in bold surpass the multiple testing threshold drawn from the current dataset using a Cochran-Mantel Haenszel test stratified by genotyping platform. Throughout this article we refer to these nine loci as "previously implicated".

References

1. Shprintzen, R.J., Goldberg, R., Golding-Kushner, K.J. & Marion, R.W. Late-onset psychosis in the velo-cardio-facial syndrome. *Am J Med Genet* **42**, 141-2 (1992).
2. McCarthy, S.E. *et al.* Microduplications of 16p11.2 are associated with schizophrenia. *Nat Genet* **41**, 1223-7 (2009).
3. Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* (2008).
4. The International Schizophrenia, C. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature.* **455**, 237-241 (2008).
5. Kirov, G. *et al.* Comparative genome hybridization suggests a role for NRXN1 and APBA2 in schizophrenia. *Hum Mol Genet* **17**, 458-65 (2008).
6. Rujescu, D. *et al.* Disruption of the neurexin 1 gene is associated with schizophrenia. *Hum Mol Genet* (2008).
7. Mulle, J.G. *et al.* Microdeletions of 3q29 confer high risk for schizophrenia. *Am J Hum Genet* **87**, 229-36 (2010).
8. Ingason, A. *et al.* Maternally derived microduplications at 15q11-q13: implication of imprinted genes in psychotic illness. *Am J Psychiatry* **168**, 408-17 (2011).

Nature Genetics: doi:10.1038/ng.3725

9. Degenhardt, F. *et al.* Duplications in RB1CC1 are associated with schizophrenia; identification in large European sample sets. *Transl Psychiatry* **3**, e326 (2013).
10. Mulle, J.G. *et al.* Reciprocal duplication of the williams-beuren syndrome deletion on chromosome 7q11.23 is associated with schizophrenia. *Biol Psychiatry* **75**, 371-7 (2014).
11. Levinson, D.F. *et al.* Copy number variants in schizophrenia: confirmation of five previous findings and new evidence for 3q29 microdeletions and VIPR2 duplications. *Am J Psychiatry* **168**, 302-16 (2011).
12. Rees, E. *et al.* CNV analysis in a large schizophrenia sample implicates deletions at 16p12.1 and SLC1A1 and duplications at 1p36.33 and CGNL1. *Hum Mol Gen* **23**, 1669-76 (2014).
13. Rees, E. *et al.* Evidence that duplications of 22q11.2 protect against schizophrenia. *Mol Psychiatry* **19**, 37-40 (2014).
14. Kirov, G. *et al.* Support for the involvement of large cnvs in the pathogenesis of schizophrenia. *Hum Mol Genet* (2009).
15. Bergen, S.E. *et al.* Genome-wide association study in a Swedish population yields support for greater CNV and MHC involvement in schizophrenia compared to bipolar disorder. *Molecular Psychiatry* **17**, 880-6 (2012).
16. Moreno-De-Luca, D. *et al.* Deletion 17q12 is a recurrent copy number variant that confers high risk of autism and schizophrenia. *Am J Hum Genet* **87**, 618-30 (2010).
17. Guha, S. *et al.* Implication of a rare deletion at distal 16p11.2 in schizophrenia. *JAMA psychiatry* **70**, 253-60 (2013).
18. Vacic, V. *et al.* Duplications of the neuropeptide receptor gene VIPR2 confer significant risk for schizophrenia. *Nature* **471**, 499-503 (2011).

Nature Genetics: doi:10.1038/ng.3725

Supplementary Table 2: Description of Datasets used (see Excel file)

Supplementary Table 3: Summary of GeneSets used (see Excel file)

Nature Genetics: doi:10.1038/ng.3725

Supplementary Table 4: Gene Set Burden stepdown

	STEPDOWN		INDEPENDENT	
NEUROFUNCTIONPHENO GENE-SETS SORTED BY SIGNIFICANCE	**Coeff_U**	**Pvalue_U_dev**	**Coeff_U**	**Pvalue_U_dev**
Synaptic GO	0.521	2.82E-11	0.521	2.82E-11
Neurof. Str.	0.127	9.76E-02	0.322	1.33E-08
Neuron Proj. GO	-0.142	2.58E-01	0.260	4.37E-05
ARC Kirov	0.269	1.47E-01	0.609	1.81E-04
FMR1 Targ. Darnell	0.121	1.31E-01	0.255	1.88E-04
NEUROFUNCTIONPHENO GENE-SETS SORTED BY EFFECT SIZE				
ARC Kirov	0.609	1.81E-04	0.609	1.81E-04
Synaptic GO	0.479	1.86E-08	0.521	2.82E-11
Neurof. Str.	0.127	9.40E-02	0.322	1.33E-08
Neuron Proj. GO	-0.179	1.61E-01	0.260	4.37E-05
FMR1 Targ. Darnell	0.121	1.31E-01	0.255	1.88E-04
NEUROF., MOUSEPHENOTYPE, BRAINEXPRESSION GENE-SETS SORTED BY SIGNIFICANCE				
Synaptic GO	0.521	2.82E-11	0.521	2.82E-11
MousePheno. Neurol. Behav.	0.109	9.96E-03	0.162	7.35E-05
BrainSpan Brain-expressed equal pre/post-natal	0.066	1.02E-01	0.152	4.89E-05

Legend: Gene-set burden analysis with global burden correction: beta coefficients and p-values for step-down logistic regression analysis (STEPDOWN) as well as original results without step-down correction (INDEPENDENT). For the step-down procedure, gene-sets were sorted by deviance p-value (SIGNIFICANCE) or beta-coefficient (EFFECT SIZE).

Nature Genetics: doi:10.1038/ng.3725

Supplementary Table 5: Digital PCR results (see Excel file)

Nature Genetics: doi:10.1038/ng.3725

Supplementary Table 6: CNV probe-level results – Previously reported CNVs

Locus	CNV type	Gene or region	Reference (see ED table 1)	Z-score *p*-value	Z-score *q*-value	CNV carriers	SCZ Case count	Control count	CMH test Odds Ratio
						N = 41321	N = 21094	N= 20227	
22q11.21	**Deletion**	**Multigenic**	1	**6.2e-13**	**4e-10**	**58**	**58**	**0**	**NA**
16p11.2	**Duplication**	**Proximal duplication**	2	**2.6e-10**	**7e-7**	**67**	**63**	**4**	**13.8**
15q13.2-13.3	**Deletion**	**Multigenic**	3,4	**6.1e-6**	**5.6e-4**	**33**	**30**	**3**	**10.55**
3q29	**Deletion**	**Multigenic**	7	**6.2e-5**	**4.4e-3**	**16**	**16**	**0**	**NA**
2p16.3	**Deletion**	**NRXN1**	5,6	**9.4e-5**	**6.2e-3**	**27**	**23**	**4**	**5.87**
16p11.2	Deletion	Distal deletion	17	1.0e-4	6.7e-3	12	11	1	12.68
22q11.21	Duplication	Multigenic	12	1.6e-4	0.03	27	4	23	0.18
1q21.1	**Deletion**	**Multigenic**	3,4	**2.9e-4**	**0.01**	**39**	**33**	**6**	**5.42**
16p13.2	Duplication	C16orf72/USP7	11	3.8e-4	0.06	27	24	3	9.02
7q11.23	Duplication	Williams-Beuren	10	4.9e-4	0.06	13	13	0	NA
15q11.2-13.1	**Duplication**	**AS/PWS**	8	**6.6e-4**	**0.07**	**15**	**15**	**0**	**NA**
8q11.23	**Duplication**	**FAM150/RB1CC1**	9	**9.2e-4**	**0.09**	**14**	**14**	**0**	**NA**
15q11.2	**Deletion**	**Multigenic**	3	**1.7e-3**	**0.06**	**142**	**95**	**47**	**1.8**
1q21.1	Duplication	Multigenic	11	2.0e-3	0.14	21	19	2	6.28
16q22.1	Duplication	WWP2	19	3.2e-3	NA	5	5	0	NA
7q36.3	Duplication	WDR60/VIPR2	11,18	4.1e-3	0.18	14	13	1	12.12
17q12	Duplication	RCAD duplication	19	0.009	0.24	20	16	4	3.81
9q33.1	Deletion	NA	19	0.02	0.35	11	9	2	4.02
22q11.23	Duplication	Multigenic	19	0.02	0.34	20	15	5	3.28
5q21.2	Deletion	NA	19	0.03	0.39	32	22	10	2.16
8p22	Duplication	SGCZ	19	0.03	NA	5	5	0	NA
9p24.2	Deletion	SLC1A1	12	0.03	0.38	8	8	0	NA
16p12.1	Deletion	Multigenic	12	0.03	0.40	33	26	7	3.22
15q21.3	Duplication	CGNL1	12	0.04	0.42	103	69	34	1.99
17q12	Deletion	RCAD deletion	16	0.04	NA	4	4	0	NA
16p13.11	Del/Dup	Multigenic	8	0.08	0.59	139	84	55	1.49
7q11.21	Duplication	NA	19	0.09	0.58	64	26	38	0.76
12q23.1	Duplication	ANKS1B/UHRF1BP1L	19	0.1	0.62	28	16	12	1.23
1p36.33	Duplication	Multigenic	12	0.11	0.63	15	12	3	3.98
5q33.1	Deletion	NA	12	0.11	0.62	11	9	2	4.19
9q21.33	Duplication	AGTPBP1	11	0.2	0.78	22	15	7	1.94
9q34.3	Duplication	C9orf62	20	0.23	0.81	409	190	219	0.8
6q24.2	Duplication	PHACTR2	12	0.26	NA	10	8	2	4.03

Nature Genetics: doi:10.1038/ng.3725

3q26.1	Deletion	NA	11	0.27	NA	5	4	1	3.5
4q35.2	Deletion	TRIML1/TRIML2	12	0.35	0.82	26	17	9	1.82
18q21.31	Duplication	NEDD4L	11	0.39	NA	2	2	0	NA
11q25	Deletion	GLB1L3/GLB1L2	11	0.42	0.85	58	34	24	1.44
9p24.2	Deletion	GLIS3	12	0.43	0.85	10	5	5	0.99
18q23	Duplication	GALR1	12	0.57	NA	7	4	3	1.22
4q35.2	Duplication	FAM149A/CYP4V2	12	0.69	0.97	12	5	7	0.71
2q37.2	Duplication	AQP12A/KIF1A	12	0.72	0.97	125	72	53	1.34
17p12	Deletion	HNPP	14	0.82	0.96	22	12	10	1.06
4q25	Duplication	ELOVL6	12	0.9	0.998	13	7	6	1
10q11.21	Duplication	Likely common CNV	19	NA	NA	NA	NA	NA	NA

Legend: Probe-level association results for all previously reported CNVs from genome-wide scans of SCZ. We report association results from the SCZ residual phenotype and use Odds Ratios from a CMH test stratified by genotyping platform. CNV loci in bold make up previously implicated loci, in which the most recent published *p*-value surpassed genome-wide correction. We also report the BH-FDR corrected *q*-value as suggestive criteria for significance. Using cutoffs from the gene-based association, BH-FDR was calculated when considering CNV loss breakpoints with at least 8 CNV carriers, and CNV gain breakpoints with at least 12 CNV carriers.

Nature Genetics: doi:10.1038/ng.3725

Supplementary Table 7: Chromosome X regional tests by sex

locus (GENE)	locus	CNV test	Sex	Cases	Controls	Deviance test *p*-value	Odds Ratio
Xp21.2-21.1 (DMD)	chrX:31047266-33139594	loss	male	5	0	0.04	Inf
			female	4	2	0.22	3.0 [0.5-18.5]
			combined	9	2	0.04	4.4 [0.9-22.4]
Xq28 (MAGEA11)	chrX:148575477-148580720	gain	male	4	14	0.001	0.17 [0.05-0.54]
			female	8	22	0.12	0.54 [0.23-1.2]
			combined	12	36	0.001	0.35 [0.18-0.68]
Xq28 (distal)	chrX:153800000-154225000	gain	male	9	0	0.01	Inf
			female	9	2	0.01	6.3 [1.2-30.4]
			combined	18	2	0.0004	8.87 [2.0-40.0]

Nature Genetics: doi:10.1038/ng.3725