ONLINE METHODS

Table of Contents

Overview of Samples

Details of each of the samples (including sample size, ancestry, and whether included in the previous publication by the PGC) are given in **Supplementary Cohort Descriptions**. The core PGC dataset included 90 cohorts for which we had individual level genotype data fully processed under a uniform pipeline. This core dataset contains genotypes on 161,405 unrelated subjects; 67,390 schizophrenia/schizoaffective disorder cases and 94,015 controls, equivalent in power to 73,189 of each. A parent-proband trio is considered to comprise one case and one control. Approximately half (31,914 cases and 47,176 controls) of the samples were not included in the previous GWAS of the PGC[1]. Around 80% of the probands (53,386 cases and 77,258 controls) were of European Ancestry, and the remainder (14,004 cases and 16757 controls) were of East Asian ancestry[2]. We additionally included in the Primary GWAS summary statistics from 9 cohorts comprising African-American (AA; 6152 cases 3918 controls) and Latino (1234 cases, 3090 controls) participants; the combined sample is equivalent in power to 6,551 each of cases and controls. 1249 LD – independent ($r^2 > 0.1$) Variants showing evidence for association ($P< 1x10^{-5}$) were further meta-analysed with an additional dataset of 1,979 cases and 142,626 controls of European ancestry obtained from deCODE genetics, thus the final analysis represents 320,404 diploid genomes.

Association Analysis

Technical Quality Control of the 90 cohorts comprising the primary PGC sample.

Technical Quality control was performed on the core PGC cohorts separately according to standards developed by the PGC[3] including SNP missingness < 0.05 (before sample removal); subject missingness < 0.02; autosomal heterozygosity deviation ($| F_{het} | < 0.2$); SNP missingness < 0.02 (after sample removal); difference in SNP missingness between cases and controls < 0.02; and SNP Hardy-Weinberg equilibrium (HWE: $P > 10^{-6}$ in controls or $P > 10^{-10}$ in cases). For family-based cohorts we excluded individuals with more than 10,000 Mendelian errors and SNPs with more than 4 Mendelian errors. For X-Chromosomal genotypes we applied an additional round of the above QC to the male and female subgroups separately.

Genomic Quality Control: Principal Component Analysis (PCA) and Relatedness Checking in the core PGC dataset

We performed PCA for all 90 cohorts separately using SNPs with high imputation quality (INFO >0.8), low missingness (<1%), MAF>0.05 and in relative linkage equilibrium (LD) after 2 iterations of LD pruning (r2 < 0.2, 200 SNP windows). We removed well known long-range-LD areas (MHC and chr8 inversion). Thus, we retained between 57K and 95K autosomal SNPs in each cohort. SNPs present in all 90 cohorts (N=7,561) were used for robust relatedness testing using PLINK v1.9[4]; pairs of subjects with PIHAT > 0.2 were identified and one member of each pair removed at random, preferentially retaining cases and trio members over case-control members.

To control for false positive associations due to inflated test statistics we evaluated the effectiveness of the primary technical and genomic quality control parameters on the genome-wide inflation of test statistics using the lambda GC (median)[5] and as necessary made the QC parameters more stringent until this value was between 1.0 and 1.4 (before inclusion of principal components as covariates) and/or between 1.0 and 1.15 after inclusion of PCA covariates. Additionally, we applied loose PCA filters for strongly stratified datasets even if we did not observe strong inflation of test statistics in order to retrieve reliable test statistics (see **Supplementary Figure 4**). Since the core PGC cohorts came from many distinct centres, countries, and continents, various measures (e.g., tightening of the technical QC parameters and/or genomic quality control) had to be taken in an iterative process to achieve this goal.

Supplementary Table 22 lists detailed per cohort exclusion numbers for individuals in the non-Asian samples. The Asian cohorts were sufficiently homogeneous as they did not show marked population structure in principal component analyses. The exclusion numbers for individuals during **technical QC** are in most cohorts low. For six cohorts (marked in yellow in **Supplementary Table 22**) it was necessary to exclude more than 100 cases during **genomic QC** so that Lambda GC fell within the window mentioned above. **Supplementary Figure 4** gives details about this process and explains why the excluded cases could not be used with the presently available control cohorts for this manuscript.

Imputation of the core PGC dataset

Genotype imputation of case-control cohorts was performed using the pre-phasing/imputation stepwise approach implemented in EAGLE 2[6] / MINIMAC3[7] (with 132 genomic windows of variable size and default parameters). The imputation reference consisted of 54,330 phased haplotypes with 36,678,882 variants from the publicly available HRC reference, release 1.1[8] Chromosome X imputation was conducted using individuals passing quality control for the autosomal analysis. ChrX imputation and association analysis was performed separately for males and females. For trio-based cohorts, families with multiple (N) affected offspring were split into N parent-offspring trios, duplicating the parental genotype information. Trios were phased with SHAPEIT 3[9]. We created pseudo-controls based on the non-transmitted alleles from the parents. Phased case-pseudo-control genotypes were then taken forward to the IMPUTE4 algorithm[10] into the above HRC reference panel.

Association / Meta-analysis

In each individual cohort, association testing was based on an additive logistic regression model using PLINK[11]. As covariates we used a subset of the first 20 principal components (PCA), derived within each cohort. By default, we included the first 4 PCAs and thereafter every PCA that was nominally significantly associated ($p<0.05$) to case-control status. PCAs in trios were only used to remove extreme ancestry outliers. We conducted a meta-analysis of the results (including the 9 cohorts comprising African-American and Latino participants) using a standard error inverse-weighted fixed effects model. For chrX, gene dosages in males were scored 0 or 2, in females, 0/1/2. We summarised the associations as number of independently associated index SNPs. Index SNPs were LD independent and had $r2 < 0.1$ within 3 Mb windows. We recorded the left and rightmost variant with $r2<0.1$ to an index SNP to define an associated clump. To define loci, we added a 50kb window on each side of the LD clump and combined overlapping LD-clumps into a single locus.

Due to the strong signal and high linkage disequilibrium in the MHC, only one SNP was kept from the extended MHC region (chr6:25-35Mb).

We additionally examined the X chromosome for evidence of heterogeneity between the sexes and X chromosome dosage compensation using the methods described by Lee and colleagues[12,13] (**Supplementary Note**). To minimise possible confounding effects of ancestry on effect sizes by sex, we restricted this analysis to those of European ancestry.

We obtained summary association results from deCODE genetics for 1,228 index SNPs ($P < 1x10^{-5}$) based on 1,979 cases and 142,626 controls of European ancestry. Genotyping was carried out at deCODE Genetics. We used this sample to establish that SNP associations from the primary GWAS replicated *en masse* in an independent sample (see **Supplementary Note**) by showing the directions of effect of index SNPs differed from the null hypothesis of randomly oriented effects and also comparing the expected number of same direction effects with those if all associations were true, taking into account the discovery magnitude of effect, and the replication effect-estimate precision (**Supplementary Note**).

The summary statistics from deCODE were combined with those from our primary GWAS dataset using an inverse variance-weighted fixed effects model. Similarly to the discovery meta-analysis (see above) we merged overlapping LD-clumps to a total of 287 distinct genomic regions (5 on the X-chromosome) with at least one genome-wide significant signal.

Polygenic Prediction

We estimated the cumulative contribution of SNPs to polygenic risk of schizophrenia using a series of leave-one-out polygenic prediction analyses based on LD-clumping and P-value thresholding (P+T)[14] (also known as C+T) using PLINK[11]. For calculating polygenic scores, we included the most significant SNP for any pair of SNPs within <500kb and with LD R^2 >0.1. We included only those with minor allele frequency >1%. We considered a range of P-value thresholds; 5×10^{-8}, 1×10^{-6}, 1×10^{-4}, 1×10^{-3}, 1×10^{-2}, 5×10^{-2}, 1×10^{-1}, 2×10^{-1}, 5×10^{-1} and 1.0. We performed logistic regression analysis within each case-control sample, to assess the relationship between case status and PRS (P+T) quantiles. The same principal components used for each GWAS were used as covariates for this analysis. Whenever the number of controls at a quantile was fewer than 5 times the number of covariates[15], or if the higher bound for the PRS Odds Ratio

(OR) became infinity, Firth's penalised likelihood method was used to compute regression statistics, as implemented in the R package "logistf"[16]. ORs from these calculations were then meta-analysed using a fixed-effects model in the R package "metafor"[17]. To ensure stability of the estimates, meta-analysis was conservatively restricted to case-control samples which contained more than 10 individuals in the top 1% PRS, with at least one of them being a control. Analogous analyses were conducted to assess the ORs between individuals at the top and bottom quantiles. To assess the performance of PRS as a predictor of schizophrenia case status, we calculated liability R^2, Nagelkerke's R^2 following Lee et. al. 2012[18] and a combined area under the receiver operating characteristic curve (AUROC). Both liability R^2 and Nagelkerke's R^2 included any principal components marginally associated with the outcome within each cohort, in the baseline model. AUROC was estimated using the non-parametric meta-analysis implemented in the R package "nsROC"[19]. Polygenic score analysis of the African-American and Latino cohorts were conducted by the authors of the study reporting those datasets[20].

Secondary analyses in core PGC dataset

Some of the secondary analyses (Gene-set enrichments, conditional SNP association analyses, fine-mapping) necessitate access to individual level data, require identical QC and imputation procedures, and/or an accurate LD reference panel meaning these analyses could only be reliably performed in a subset of the dataset. The following analyses focussed on the core PGC dataset for which these conditions are met.

Gene Set Enrichments

Tissue and cell types

We collected bulk RNA-seq data across 53 human tissues (GTEx v8, median across samples)[21]; from a study of 19,550 nuclei from frozen adult human post-mortem hippocampus and prefrontal cortex representing 16 different cell types[22]; from a study of ~10,000 single cells from 5 mouse brain regions (cortex, hippocampus, hypothalamus, midbrain and striatum, in addition to specific enrichments for oligodendrocytes, dopaminergic neurons, serotonergic neurons and cortical parvalbuminergic interneurons) that identified 24 cell types[23]; from a study of~500,000 single cells from the mouse nervous system (19 regions) that identified 265 cell types[24].

Datasets were processed uniformly[25]. First, we calculated the mean expression for each gene for each type of data if these statistics were not provided by the authors. We used the pre-computed median expression (transcript per million (TPM)) across individuals for the GTEx tissues (v8). For the GTEx dataset, we excluded tissues with less than 100 samples, merged tissues by organ (with the exception of brain tissues), excluded non-natural tissues (e.g. EBV-transformed lymphocytes) and testis (outlier in hierarchical clustering), resulting in 37 tissues. Genes without unique names and genes not expressed in any cell types were excluded. We scaled the expression data to 1M Unique Molecular Identifiers (UMIs) or TPM for each cell type/tissue. After scaling, we excluded non-protein coding genes, and, for mouse datasets, genes that had no expert curated 1:1 orthologs between mouse and human (Mouse Genome Informatics, The Jackson laboratory, version 11/22/2016). We then calculated a metric of gene expression specificity by dividing the expression of each gene in each cell type/tissue by the total expression of that gene in all cell types/tissue, leading to values ranging from 0 to 1 for each gene (0: meaning that the gene is not expressed in that cell type/tissue, 1 that 100% of the expression of that gene is performed in that cell type/tissue). We selected the 10% most specific genes per cell type (or tissue) with an expression level of at least 1TPM, or 1 UMI per million, for downstream analyses and used MAGMA v1.08[26] to test whether they were enriched for genetic associations. We performed a one-sided test as we were only interested in enrichments for genetic associations (in contrast with depletions). We also applied partitioned LD score regression (LDSC) as described[27] to the top 10% genes for each cell type for heritability enrichment. We selected the one-sided coefficient z-score p-value as a measure of the association of the cell type/tissue with schizophrenia.

Ontology Gene sets

Gene set analyses were performed using MAGMA v1.08[26]. Gene boundaries were retrieved from Ensembl release 92 (GRCh37) using the "biomaRt" R package[28] and expanded by 35 kb upstream and 10 kb downstream to include likely regulatory regions[29]. Gene-wide p-values were calculated from European and Asian summary statistics separately using the SNP-wise "mean" Imhof method, and meta-analysed within the software. LD reference data files were from the European and East Asian populations of the Haplotype Reference Consortium[30]. Within each gene set analysis, p-values were corrected for multiple testing using the Bonferroni procedure. Specifically, we tested the following gene sets:

(i) Gene ontology: 7,315 sets extracted from the GO database (http://geneontology.org/, accession date: 09/11/2020) curated to include only annotations with experimental or phylogenetic supporting evidence.

(ii) SynGO ontology: Described elsewhere[31], this collection was analysed as two subsets; “biological process” (135 gene sets) and “cellular component” (60 gene sets). We controlled for a set of 10,360 genes with detectable expression in brain tissue measured as Fragments Per Kilobase of transcript per Million mapped reads (FPKM)[32] to detect synaptic signals above signals simply reflecting the property of brain expression. Exploiting the hierarchical structure of SynGO, gene sets were reconstructed using a “roll-up” method, in which parent categories contained all genes annotated to child categories. For stepwise conditional testing[33], we prioritised the most specific child annotations[34] (i.e. the lowest possible level) as regression covariates.

Conditional SNP Association Analyses

We performed stepwise conditional analyses of 248 loci that were genome wide significant in the core PGC dataset looking for independent associations. We performed association testing and meta-analysis across each locus, adding the allele dosages of the index SNP as a covariate. Where a second SNP had a conditional p-value of less than $1x10^{-6}$, we considered this as evidence for a second signal and repeated the process adding this as an additional covariate. We repeated this until no additional SNPs in the region achieved $p<1x10^{-6}$. We also searched for long range dependencies. Here we tested the all pairs of independent signals for conditional independence (**Supplementary Note).**

Fine-mapping

We used FINEMAP[35] to fine-map regions defined by LD clumps ($r^2>0.1$), excluding the MHC locus due to its complex LD structure. Clumps which overlapped (without adding the additional

50kb used to define physically distinct loci) were combined. As fine-mapping requires data from all markers in the region[36] we only performed fine-mapping on regions that attained genome-wide significance (GWS) in the core PGC GWAS. In total, we attempted to fine-map 255 non-overlapping regions (**Supplementary Table 11e**). Further details about the fine-mapping process are given in the **Supplementary Note.**

Summary-data-based Mendelian Randomization (SMR) analysis, FUSION and EpiXcan

We used SMR[37] as our primary method to identify SNPs which might mediate association with schizophrenia through effects on gene expression. The significance for SMR is set at the Bonferroni corrected threshold of 0.05/M where M is the number of genes with significant eQTLs tested for a given tissue. Significant SMR associations imply colocalization of the schizophrenia associations with eQTL. We applied the HEIDI test[37] to filter out SMR associations ($P_{\mathrm{HEIDI}} < 0.01$) due to linkage disequilibrium between SCZ-associated variants and eQTLs. *cis*-eQTL summary data were from three studies: fetal brain (N=120)[38], adult brain (n = ~1,500)[39] and blood (n = ~32,000)[40]. Linkage disequilibrium (LD) data required for the HEIDI test[37] were estimated from the Health and Retirement Study (HRS)[41] (n = 8,557). We included only genes with at least one *cis*-eQTL at $P_{\mathrm{eQTL}} < 5\times10^{-8}$, excluding those in MHC regions due to the complexity of this region. For blood, we included only genes with eQTLs in brain. This left 7,803 genes in blood, 10,890 genes in prefrontal cortex and 754 genes in fetal brain for analysis (see **Supplementary Note** for further details). SMR was performed using data from the primary GWAS. The results were then filtered to exclude significant SMR implicated genes where the eQTLs did not map within our definition of an associated locus in the Extended GWAS meta-analysis of our primary GWAS dataset and the dataset provided by deCODE genetics.

For genomic regions where there were multiple genes showing significant SMR associations, we attempted to resolve these with conditional analysis using GCTA-COJO[42,43]. We selected the top-associated *cis*-eQTL for one gene (or a set of genes sharing the same *cis*-eQTL) ran a COJO analysis in the schizophrenia GWAS data and the eQTL data for each of the other genes conditioning on the selected top *cis*-eQTL. We then re-ran the SMR and HEIDI analyses using these conditional GWAS and eQTL results.

We used FUSION[44] and EpiXcan[45] as tests of robustness of the SMR results. Details are supplied in the **Supplementary Note** as are our approaches to prioritising SMR associated genes.

DATA AVAILABILITY

Summary statistics for the “Extended”, “Core”, ancestry specific and sex-stratified analyses is available at “https://www.med.unc.edu/pgc/download-results/scz/”. Genotype data are available for a subset of cohorts, including dbGAP accession numbers and/or restrictions, as described in the Supplementary Information section “Cohort Descriptions”.

CODE AVAILABILITY

Core analysis code for RICOPILI can be found at “https://sites.google.com/a/broadinstitute.org/ricopili/”. This wraps PLINK (“https://www.cog-genomics.org/plink2/”), EIGENSOFT (“https://www.hsph.harvard.edu/alkes-price/software/”), EAGLE2 (“https://alkesgroup.broadinstitute.org/Eagle/”), MINIMAC3 (“https://genome.sph.umich.edu/wiki/Minimac3”), SHAPEIT3 (“https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html”), METAL (“https://genome.sph.umich.edu/wiki/METAL_Documentation”), LDSR (“https://github.com/bulik/ldsc”). For downstream analyses, FINEMAP can be found at “http://christianbenner.com/”, and our utility for meta-analysing cohort-specific LD matrices can be found at https://github.com/Pintaius/LDmergeFM. MAGMA can be found at ”https://ctg.cncr.nl/software/magma” and the GO gene sets and automated curation pipeline are provided in https://github.com/janetcharwood/pgc3-scz_wg-genesets. SMR is available at “https://cnsgenomics.com/software/smr/” and SbayesS at “https://cnsgenomics.com/software/gctb/”.

METHODS REFERENCES

1. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).

2. Lam, M. *et al.* Comparative genetic architectures of schizophrenia in East Asian and European populations. *Nat. Genet.* (2019) doi:10.1038/s41588-019-0512-x.

3. Lam, M. *et al.* RICOPILI: Rapid Imputation for COnsortias PIpeLIne. *Bioinformatics* (2019) doi:10.1093/bioinformatics/btz633.

4. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L. & Ferreira, M. A. PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet* **81**, (2007).

5. Devlin, B. & Roeder, K. Genomic Control for Association Studies. *Biometrics* **55**, 997–1004 (1999).

6. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).

7. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).

8. Haplotype Reference Consortium r1.1. https://ega-archive.org/datasets/EGAD00001002729.

9. O'Connell, J. *et al.* Haplotype estimation for biobank-scale data sets. *Nat. Genet.* **48**, 817–820 (2016).

10. Bycroft, C. *et al.* Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv* 166298 (2017) doi:10.1101/166298.

11. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

12. Lee, J. J. *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).

13. Tukiainen, T. *et al.* Landscape of X chromosome inactivation across human tissues. *Nature* **550**, 244–248 (2017).

14. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).

15. Vittinghoff, E. & McCulloch, C. E. Relaxing the rule of ten events per variable in logistic and cox regression. *Am. J. Epidemiol.* **165**, 710–718 (2007).

16. Heinze, G. & Ploner, M. *A SAS macro, S-PLUS library and R package to perform logistic regression without convergence problems*. https://cemsiis.meduniwien.ac.at/fileadmin/user_upload/_imported/fileadmin/msi_akim/CeMSIIS/KB/programme/tr2_2004.pdf (2004).

17. Viechtbauer, W. Conducting Meta-Analyses in R with the metafor Package. *J. Stat. Softw.* **36**, 1–48 (2010).

18. Lee, S. H., Goddard, M. E., Wray, N. R. & Visscher, P. M. A Better Coefficient of Determination for Genetic Profile Analysis. *Genet. Epidemiol.* **36**, 214–224 (2012).

19. Martínez-Camblor, P. Fully non-parametric receiver operating characteristic curve estimation for random-effects meta-analysis. *Stat. Methods Med. Res.* **26**, 5–20 (2017).

20. Bigdeli, T. B. *et al.* Contributions of common genetic variants to risk of schizophrenia among individuals of African and Latino ancestry. *Mol. Psychiatry* **25**, 2455–2467 (2020).

21. Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

22. Habib, N. *et al.* Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat. Methods* **14**, 955–958 (2017).

23. Skene, N. G. *et al.* Genetic identification of brain cell types underlying schizophrenia. *Nat. Genet.* **50**, 825–833 (2018).

24. Zeisel, A. *et al.* Molecular Architecture of the Mouse Nervous System. *Cell* **174**, 999-1014.e22 (2018).

25. Bryois, J. *et al.* Genetic identification of cell types underlying brain complex traits yields insights into the etiology of Parkinson’s disease. *Nat. Genet.* 1–12 (2020) doi:10.1038/s41588-020-0610-9.

26. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Comput. Biol.* **11**, e1004219 (2015).

27. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).

28. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/ Bioconductor package biomaRt. *Nat. Protoc.*

4, 1184–1191 (2009).

29. Maston, G. A., Evans, S. K. & Green, M. R. Transcriptional Regulatory Elements in the Human Genome. *Annu. Rev. Genomics Hum. Genet.* **7**, 29–59 (2006).

30. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).

31. Koopmans, F. *et al.* SynGO: An Evidence-Based, Expert-Curated Knowledge Base for the Synapse. *Neuron* **103**, 217-234.e4 (2019).

32. Genovese, G. *et al.* Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nat. Neurosci.* **19**, 1433–1441 (2016).

33. Pardiñas, A. F. *et al.* Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat. Genet.* **50**, 381–389 (2018).

34. Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation. *PLoS One* **5**, e13984 (2010).

35. Benner, C. *et al.* FINEMAP: Efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).

36. Benner, C. *et al.* Prospects of Fine-Mapping Trait-Associated Genomic Regions by Using Summary Statistics from Genome-wide Association Studies. *Am. J. Hum. Genet.* **101**, 539–551 (2017).

37. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).

38. O'Brien, H. E. *et al.* Expression quantitative trait loci in the developing human brain and their enrichment in neuropsychiatric disorders. *Genome Biol.* **19**, 194 (2018).

39. Gandal, M. J. *et al.* Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science (80-.).* **362**, (2018).

40. Võsa, U. *et al.* Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *bioRxiv* 447367 (2018) doi:10.1101/447367.

41. Sonnega, A. *et al.* Cohort profile: The Health and Retirement Study (HRS). *Int. J. Epidemiol.* **43**, 576–585 (2014).

42. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A Tool for Genome-wide Complex Trait Analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

43. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics

identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375 (2012).

44. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).

45. Zhang, W. *et al.* Integrative transcriptome imputation reveals tissue-specific and shared biological mechanisms mediating susceptibility to complex traits. *Nat. Commun.* **10**, (2019).

EXTENDED DATA LEGENDS

Extended Data Figure 1: Primary GWAS Manhattan plot

The x-axis indicates chromosomal position and the y-axis is the significance of association (−log10(P)). The red line represents genome-wide significance level (5×10−8). SNPs in green are in linkage disequilibrium (LD; R2 >0.1) with index SNPs (diamonds) which represent LD independent genome-wide significant associations.

Extended Data Figure 2: Polygenic risk prediction

A) Distributions of liability scale R^2 across 98 left-out-cohorts for polygenic risk scores built from SNPs with different p-value thresholds. Distributions of liability R^2 (assuming schizophrenia life-time risk of 1%) are shown for each p-value threshold, with point size representing size of the left-out cohort and colour representing ancestry. The median liability R^2 is represented as a horizontal black line. B) Liability R^2 of predicted and observed phenotypes in left-out cohorts using variants with p-value threshold p=0.05, from the fixed effect meta-analysis of variant effects, unadjusted for multiple comparisons. The polygenic risk scores are derived from two separate sets of leave-one-out GWAS meta-analyses: y-axis R^2 based on the results of primary GWAS including all ancestries; x axis R^2 based on cohorts of the same ancestry as the test samples. Circles denote core PGC samples. Triangles denote African American and Latino samples processed external to PGC by the providing author.

Extended Data Figure 3: Association between 37 human tissues and schizophrenia.

The mean of the evidence (-$\log_{10}$P) obtained from two methods (MAGMA, LDSC) for testing GWAS data for enrichment of association in genes with high expression in each tissue as determined from bulk RNA-seq[20]. The bar colour indicates whether gene expression in the tissue is significantly associated with both methods, one method or none. The black vertical line represents the significance threshold corrected for the total number of tissues tested in this experiment. We also analysed previous waves of PGC schizophrenia GWAS[11,21] for comparison.

Extended Data Figure 4 Legend: Associations between schizophrenia and cell types from multiple brain regions in human and mouse

The mean of the evidence (-$\log_{10}$P) obtained from two methods (MAGMA, LDSC) for testing GWAS data for enrichment of associations in genes with high expression in cell types. The 15 human cell types (derived from single nuclei) from the cortex and hippocampus.

Extended Data Table1: List of prioritized genes

List of genes meeting prioritisation criteria summarised in Figure 1. Index SNP: index associated SNP for the locus from the GWAS. Ensembl ID: Ensembl gene identifier. Symbol ID: HGNC gene symbol. Gene Biotype: as classified by Ensembl. FINEMAP and SMR priority genes: genes meeting the prioritisation criteria described in the text. Rare priority genes: genes implicated by rare coding variants in schizophrenia, autism spectrum disorders or developmental disorder. Full details regarding the prioritisation criteria for each gene are given in Supplementary Tables 11-18.