nature genetics

Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs

S Hong Lee[1,2], Teresa R DeCandia[3,4], Stephan Ripke[5,6], Jian Yang[2,7], The Schizophrenia Psychiatric Genome-Wide Association Study Consortium (PGC-SCZ)[8], The International Schizophrenia Consortium (ISC)[8], The Molecular Genetics of Schizophrenia Collaboration (MGS)[8], Patrick F Sullivan[9], Michael E Goddard[10,11], Matthew C Keller[3,4,12], Peter M Visscher[1,2,7,12] & Naomi R Wray[1,2,12]

Schizophrenia is a complex disorder caused by both genetic and environmental factors. Using 9,087 affected individuals, 12,171 controls and 915,354 imputed SNPs from the Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium (PGC-SCZ), we estimate that 23% (s.e. = 1%) of variation in liability to schizophrenia is captured by SNPs. We show that a substantial proportion of this variation must be the result of common causal variants, that the variance explained by each chromosome is linearly related to its length ($r = 0.89$, $P = 2.6 \times 10^{-8}$), that the genetic basis of schizophrenia is the same in males and females, and that a disproportionate proportion of variation is attributable to a set of 2,725 genes expressed in the central nervous system (CNS; $P = 7.6 \times 10^{-8}$). These results are consistent with a polygenic genetic architecture and imply more individual SNP associations will be detected for this disease as sample size increases.

Schizophrenia is a severe mental disorder with lifetime risk of ~1% and heritability of ~0.7–0.8 (refs. 1–3). Of complex genetic diseases, schizophrenia has perhaps been the subject of the most speculation and debate relating to its genetic architecture[4,5], and the relative importance of common causal variants remains controversial[6,7]. GWAS of schizophrenia have discovered associated variants[8–10] that together explain only a small fraction of heritability[11]. Here, we have applied new methods[12,13] for estimation of the variation explained by genome-wide genotypes to PGC-SCZ data[14]. In these methods, the variance estimate is derived from the average genome-wide similarity between all pairs of individuals determined using all SNPs. Genetic variation is estimated when case-case pairs and control-control pairs are on average more similar across the genome than case-control pairs. We used data only from cases and controls that are 'unrelated' in the classical sense and calculated the variance explained by autosomal SNPs. We partitioned[15] this genomic variation by chromosome, sex, functional annotation and minor allele frequency (MAF).

[1]Queensland Brain Institute, University of Queensland, Brisbane, Queensland, Australia. [2]Queensland Institute of Medical Research, Brisbane, Queensland, Australia. [3]Department of Psychology & Neuroscience, University of Colorado Boulder, Boulder, Colorado, USA. [4]Institute for Behavioral Genetics, University of Colorado Boulder, Boulder, Colorado, USA. [5]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, USA. [6]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard University, Cambridge, Massachusetts, USA. [7]University of Queensland Diamantina Institute, Princess Alexandra Hospital, Brisbane, Queensland, Australia. [8]A full list of members is provided in the **Supplementary Note**. [9]Department of Genetics, University of North Carolina, Chapel Hill, North Carolina, USA. [10]Department of Agriculture and Food Systems, University of Melbourne, Melbourne, Victoria, Australia. [11]Biosciences Research Division, Department of Primary Industries, Melbourne, Victoria, Australia. [12]These authors jointly directed this work. Correspondence should be addressed to N.R.W. (naomi.wray@uq.edu.au).

Received 16 August 2011; accepted 17 January 2012; published online 19 February 2012; doi:10.1038/ng.1108

RESULTS

Genomic variation captured by common SNPs

The PGC-SCZ includes data from the International Schizophrenia Consortium (ISC)[8], the Molecular Genetics of Schizophrenia Collaboration (MGS)[9] and other samples (together referred to as Other) (**Supplementary Table 1**). Using a linear mixed model (see Online Methods), we estimated the proportion of variance in liability to schizophrenia explained by SNPs (h^2) in each of these three independent data subsets (**Table 1**). We use the notation h^2 because the estimates represent a lower bound of narrow-sense heritability that results from the fact that only variation due to association with the SNPs can be estimated. Preliminary analyses were conducted using nonimputed genotypes of the ISC and MGS subsets (**Supplementary Table 2**). The individual estimates of h^2 for the ISC and MGS subsets and for other samples from the PGC-SCZ were each greater than the estimate from the total combined PGC-SCZ sample of h^2 = 23% (s.e. = 1%) (**Table 1**). We investigated this result by conducting bivariate analyses in which we considered cases and controls from one subset to be trait 1 and those from a different subset to be trait 2 (**Table 2**). The two independent subsets were related through the coefficients of genome-wide similarity calculated from SNPs between individuals (Online Methods, Eq. (3)). The estimated correlation coefficients based on SNP genome-wide similarities were <1, consistent with several explanations. Subsets might be more homogeneous—both phenotypically, for example, because of similar and consistent diagnostic criteria, and genetically, because linkage disequilibrium (LD) between causal variants and analyzed SNPs might be higher within than between subsets. Alternatively, subtle artifacts could generate

Table 1 Estimated proportion of variance in liability to schizophrenia captured by SNPs

Data set	Cases	Controls	h^2 (s.e.)
ISC	3,220	3,445	0.27 (0.02)
MGS	2,571	2,419	0.31 (0.03)
Other	3,296	6,307	0.27 (0.02)
ISC and MGS	5,791	5,864	0.25 (0.01)
PGC-SCZ	9,087	12,171	0.23 (0.01)

Estimates are based on 915,354 imputed SNPs. h^2, estimate of proportion of variance in liability to schizophrenia explained by SNPs. The three independent subsets ISC, MGS and Other together comprise the total PGC-SCZ sample.

nonrandom differences in allele frequency between sets of cases and sets of controls from the same study. However, our preliminary analyses using genotyped SNPs for the ISC and MGS subsets and extreme quality control filtering (**Supplementary Table 2**) suggest that this was unlikely to be a major contributor. Furthermore, the correlations between data sets from the bivariate analyses were high (~0.8), indicating that the same genetic signals can explain variance in schizophrenia liability in different case-control samples. As these samples were collected independently with genotyping conducted at different laboratories, it is difficult to envision artifacts that could generate such high correlations. Hence, we conclude that the PGC-SCZ estimate of h^2 represents the lower bound of variance in liability that would be explained by common SNPs in a large phenotypically and genetically homogeneous sample with no genotyping artifacts.

Partitioning of genomic variation by chromosome

Cryptic population stratification has been proposed to be a confounding factor in GWAS[7]. A consequence of population stratification is that segments of ancestry-specific chromosomes segregate together in the population. In this situation, variance attributed to causal variants on one chromosome can be predicted by SNPs from segments derived from the same ancestral population on other chromosomes. To investigate whether population stratification could have contributed to our results (beyond the ancestry principal-component scores included as covariates in the analyses), we performed two kinds of analyses: one in which the similarity matrix for each chromosome was fitted separately (22 analyses estimating one additive genetic variance component per analysis) and a second joint analysis in which the 22 similarity matrices were fitted simultaneously (estimating 22 additive genetic variance components in a single analysis) (Online Methods). Finding higher total variance explained by the 22 individually estimated variances compared to the 22 simultaneously estimated variances would provide evidence of stratification. The total variance explained was determined to be 26% for chromosomes fitted separately compared to a total of 23% when chromosomes were fitted together, thus showing little evidence of population stratification (**Fig. 1a**). The estimates of variance explained by each chromosome were linearly related to the length of the chromosome (correlation $r = 0.89$, $P = 2.6 \times 10^{-8}$), consistent with a highly polygenic model, and the length correlation is very similar to results for human height[12].

Genomic variation by sex

Sex differences have been described for almost all features of schizophrenia (prevalence, incidence, age of onset, clinical presentation, course and response to treatment)[16]. To determine whether the variance in liability captured by SNPs on autosomes differs between the sexes, we undertook a bivariate analysis considering male cases and controls as one trait and female cases and controls as the second trait. The two independent subsets were related through the coefficients of similarity calculated from SNPs (Online Methods, Eq. (3)). The correlation in liabilities captured by SNPs between the sexes was very high (0.89, s.e. = 0.06, not significantly different from 1) (**Table 2**), implying that the majority of additive genetic variance is shared between the sexes. We also investigated variance explained by genotyped SNPs on the X chromosome for the ISC and MGS data sets, and we conclude that the variance explained by the X chromosome is consistent with the expected value given its length (**Supplementary Table 3**).

Partitioning of genomic variation by functional annotation

To assess whether functional annotation of SNPs is associated with the variance they explain, we partitioned the variance explained by SNPs into three components by creating similarity matrices of SNPs in genes expressed in the central nervous system (CNS$^+$), those found in other genes and those not localized to genes (Online Methods). The CNS$^+$ genes included four previously identified subsets[17] comprising genes expressed in the brain (specifically, genes with differential CNS expression) and those with neuronal activity, roles in learning and synapse function. We found that the variance attributable to the CNS$^+$ genes was significantly greater than the proportion of the genome that they represent (31%, s.e. = 2%, versus 20% of the genome represented; $P = 7.6 \times 10^{-8}$) (**Fig. 1b** and **Supplementary Table 4**).

Partitioning of genomic variation by SNP MAF

It has been argued that the low proportion of variance explained by previous GWAS of schizophrenia suggests that common variants are not important to the etiology of the disease[6,7,18]. To evaluate this hypothesis, we undertook an analysis in which we partitioned the variance captured by SNPs into five components defined by MAF (Online Methods). For close relatives (who were excluded from our analyses), estimated similarities based on SNPs with different MAFs would be comparable. However, very distant relatives have inherited chromosome segments from distant common ancestors. If a SNP is more recent than the common ancestor, then the relationship between these individuals would not be reflected by the SNP, and SNPs with low MAF tend to be more recent than SNPs with high MAF. The variance explained by SNPs with MAF of <0.1 was 2% (s.e. = 1%) from a joint analysis of all five MAF bins in the total PGC-SCZ data set (**Fig. 1c** and **Supplementary Table 5**). This low contribution to the total variance explained is likely to partly reflect under-representation of SNPs with low MAF in the analysis (minimum MAF = 0.01) relative to those in the genome. The other four MAF bins each explain approximately equal proportions of the variance (~5%, s.e. = 1%). Analyses of the PGC-SCZ subsets were consistent with these results

Table 2 Bivariate analyses of PGC-SCZ subsets

Subset 1/subset 2	Cases (subset 1/2)	Controls (subset 1/2)	Subset 1 h^2 (s.e.)	Subset 2 h^2 (s.e.)	r (s.e.)
ISC/MGS	3,220/2,571	3,445/2,419	0.26 (0.02)	0.29 (0.03)	0.84 (0.09)
ISC/Other	3,220/3,296	3,445/6,307	0.26 (0.02)	0.27 (0.02)	0.89 (0.07)
MGS/Other	2,571/3,296	2,419/6,307	0.30 (0.03)	0.26 (0.02)	0.79 (0.08)
ISC and MGS/Other	5,791/3,296	5,864/6,307	0.24 (0.01)	0.26 (0.02)	0.87 (0.06)
Male/female	6,031/3,056	5,884/6,287	0.24 (0.01)	0.25 (0.02)	0.89 (0.06)

Estimates are based on 915,354 imputed SNPs. h^2, estimate of proportion of variance in liability to schizophrenia explained by SNPs; r, correlation of liabilities explained by SNPs between subsets 1 and 2. The three independent subsets ISC, MGS and Other together comprise the total PGC-SCZ sample.

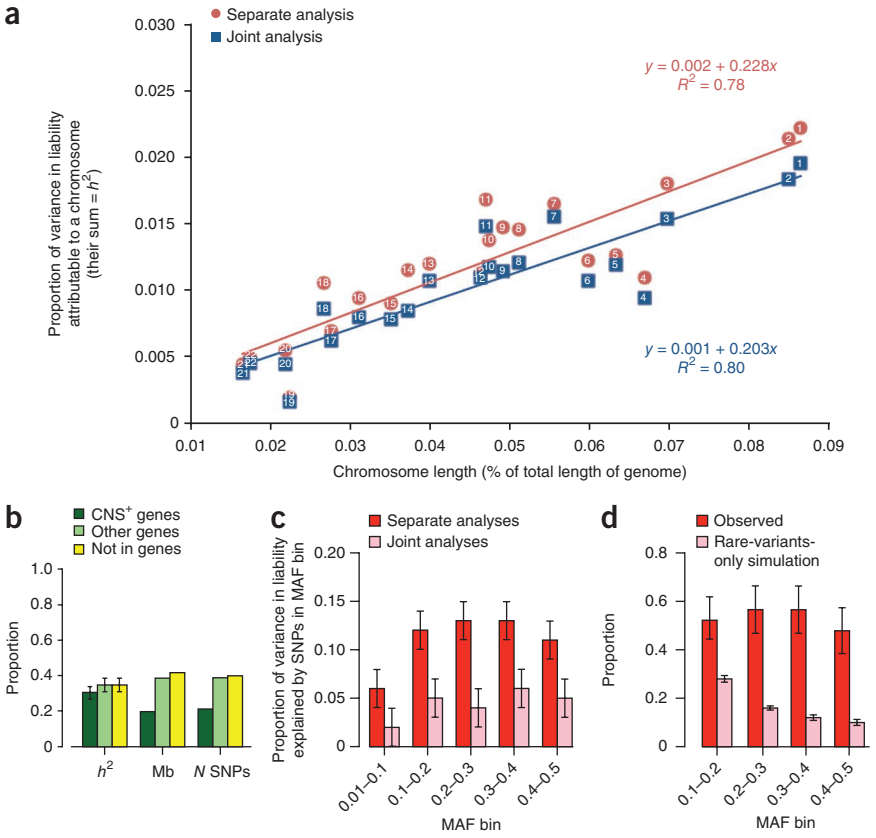

Figure 1 Genomic partitioning of schizophrenia. (**a**) By chromosome. Estimated proportion of the variance in liability to schizophrenia captured by SNPs on individual chromosomes from a joint analysis of all chromosomes simultaneously or separate analyses for each chromosome. The sum of the h^2 value is 0.23 for the joint analysis and 0.26 for the separate analyses. (**b**) By annotation. The total variance explained by SNPs (h^2) found in CNS$^+$ genes and other genes and by those not in genes totals 0.23. Of this, a proportion (0.31) is attributed to SNPs in CNS$^+$ genes, which is greater than expected by chance ($P = 7.6 \times 10^{-8}$), given that the CNS$^+$ genes cover 0.20 of the length of the genome (Mb) and represent 0.21 of the SNP count (*N* SNPs). Error bars, 95% confidence intervals of the estimates. (**c**) By MAF bin from analyses fitting MAF bins jointly or separately. Error bars, 95% confidence intervals. (**d**) By MAF bins compared to simulation under a rare-variants-only model. The variance explained by SNPs in each MAF bin (when MAF bins were fitted in separate analyses) as a proportion of the variance explained by all SNPs. Error bars, 95% confidence intervals. For the simulations (right) calculated using the s.d. across simulation replicates.

(**Supplementary Table 5**). Given the known relationship between allele frequency and LD[19], it is highly unlikely that the estimates of h^2 reported here are explained predominantly by rare causal variants[20]. We performed simulations conditional on PGC-SCZ data and confirmed that a rare-variants-only model could not explain our results. For example, in an analysis of PGC-SCZ data using only SNPs with MAF of >0.4, we found that 11% (s.e. = 1%) of the variance in liability was explained, which is nearly half of the variance explained by all SNPs. However, in simulations that attributed 50% of variation in liability to SNPs with MAF of <0.1, SNPs with MAF of >0.4 explained only 5% (s.e. = 0.3%) of the variance, which is only 10% of the variation captured by all SNPs (**Fig. 1c**,**d** and **Supplementary Tables 5** and **6**). Furthermore, our simulation strategy was a best-case scenario that favored the rare-variants-only model, as our simulation extended the definition of 'rare' variants to those with MAF = 0.1, generating higher LD between the common genotyped SNPs and causal variants than would be expected under a more typical definition of rare variants (MAF < 0.01). Our results are consistent with analyses of the ISC data[8,20]. In the **Supplementary Note**, we compare our methods to the risk-profiling ones used by the ISC and the efficient mixed model association expedited (EMMAX) method[21].

DISCUSSION

We draw four noteworthy conclusions from our results. First, using direct queries of the genome, we quantified the lower limit of the genetic contribution to schizophrenia: approximately one-quarter of the variance in liability is directly explained by common variants represented across the current generation of GWAS arrays[8] (**Table 1**), and this variance is shared between the sexes (**Table 2**). Second, we provide evidence that causal risk variants must include common variants (**Fig. 1d**). Third, we show that the variance explained by chromosomes is linearly related to the length of the chromosome (**Fig. 1b**), consistent with a highly polygenic model (many risk loci). Fourth, we find that the CNS$^+$ gene set explains significantly ($P = 7.6 \times 10^{-8}$) more variation than expected for the proportion of the genome it represents. Taken together, our results provide guidance for the future of genetic studies in schizophrenia. Some have argued[6,7,18] that common variants have only a small role in the etiology of schizophrenia and that the GWAS approach for schizophrenia has been misconceived. Our results refute these claims by showing that at least one-quarter of variation in liability to schizophrenia is explained by SNPs and that common causal variants must be responsible for most of this signal. Therefore, larger sample sizes are likely to achieve the statistical power needed to detect additional effects (in addition to those detected to date) with genome-wide significance. Recently, a GWAS for height[17], considered as a model complex trait, identified 180 robustly associated loci in a total sample size of 180,000 individuals, and the identified variants were concentrated in pathways biologically associated with growth. Samples of ~50,000 schizophrenia cases and 50,000 controls are needed to afford the same power to detect variants that explain an equivalent proportion of phenotypic variance, thereby allowing increased insight into biological pathways as was achieved in the height study[11,12,22]. Our results suggest that GWAS of larger case-control samples will deliver meaningful results for schizophrenia.

In conclusion, we estimate that about one-quarter of variation in liability to schizophrenia, or approximately one-third of genetic variation in liability, is captured by considering all genotyped and imputed SNPs simultaneously. The remaining missing heritability most likely reflects imperfect LD between causal variants and the genotyped and imputed SNPs. The current generation of genotyping chips may explain only ~70% of the total variance attributable to common SNPs (MAF > 0.1) and may explain less of variance attributable to uncommon and rare variants (**Supplementary Fig. 1**). From the analyses we have performed, we cannot estimate a distribution of the allele frequency of causal variants, but the most likely cause of low LD between causal variants and SNPs is that many causal

variants have low MAFs. Nevertheless, from the results presented, we can conclude that common causal variants in LD with genotyped and imputed SNPs must contribute to genetic variation for liability to schizophrenia in the population. Hence, causal risk variants for schizophrenia range across the entire allelic frequency spectrum.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank S.D. Gordon for technical assistance. We acknowledge funding from the Australian National Health and Medical Research Council (389892, 442915, 496688, 613672 and 613601), the Australian Research Council (DP0770096, DP1093502 and FT0991360) and the US National Institute of Mental Health (MH085812). This research utilized the Cluster Computer, which is funded by the Netherlands Scientific Organization (NWO; 480-05-003). Acknowledgments for PGC-SCZ are listed in the **Supplementary Note**.

AUTHOR CONTRIBUTIONS

N.R.W. and P.M.V. devised the study. S.H.L. performed all preliminary analyses on the ISC sample and final analyses on the PGC-SCZ samples. T.R.D. performed preliminary analyses on the MGS sample. M.C.K. directed preliminary analyses on the MGS sample. S.R. undertook the quality control analysis and imputation of the PGC-SCZ samples. M.E.G. and J.Y. advised on analyses and their interpretation. P.F.S. provided interpretation in the context of schizophrenia research. N.R.W., S.H.L. and P.M.V. wrote the first draft of the manuscript. All authors contributed to the final manuscript. The ISC, MGS and PGC-SCZ members collected and genotyped cases and controls.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. 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**, 1187–1192 (2003).
2. Cardno, A.G. & Gottesman, I.I. Twin studies of schizophrenia: from bow-and-arrow concordances to Star Wars Mx and functional genomics. *Am. J. Med. Genet.* **97**, 12–17 (2000).
3. Lichtenstein, P. *et al.* Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet* **373**, 234–239 (2009).
4. McClellan, J.M., Susser, E. & King, M.C. Schizophrenia: a common disease caused by multiple rare alleles. *Br. J. Psychiatry* **190**, 194–199 (2007).
5. Craddock, N., O'Donovan, M.C. & Owen, M.J. Phenotypic and genetic complexity of psychosis. Invited commentary on… Schizophrenia: a common disease caused by multiple rare alleles. *Br. J. Psychiatry* **190**, 200–203 (2007).
6. McClellan, J. & King, M.C. Genetic heterogeneity in human disease. *Cell* **141**, 210–217 (2010).
7. McClellan, J. & King, M.C. Genomic analysis of mental illness: a changing landscape. *JAMA* **303**, 2523–2524 (2010).
8. Purcell, S.M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
9. Shi, J. *et al.* Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature* **460**, 753–757 (2009).
10. Moskvina, V. *et al.* Gene-wide analyses of genome-wide association data sets: evidence for multiple common risk alleles for schizophrenia and bipolar disorder and for overlap in genetic risk. *Mol. Psychiatry* **14**, 252–260 (2009).
11. Visscher, P.M., Goddard, M.E., Derks, E.M. & Wray, N.R. Evidence-based psychiatric genetics, AKA the false dichotomy between common and rare variant hypotheses. *Mol. Psychiatry* published online (14 June 2011), doi:10.1038/mp.2011.65.
12. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
13. Lee, S.H., Wray, N.R., Goddard, M.E. & Visscher, P.M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88**, 294–305 (2011).
14. Ripke, S. *et al.* Genome-wide association study identifies five new schizophrenia loci. *Nat. Genet.* **43**, 969–976 (2011).
15. Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* **43**, 519–525 (2011).
16. Abel, K.M., Drake, R. & Goldstein, J.M. Sex differences in schizophrenia. *Int. Rev. Psychiatry* **22**, 417–428 (2010).
17. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
18. Cirulli, E.T. & Goldstein, D.B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.* **11**, 415–425 (2010).
19. Wray, N.R. Allele frequencies and the r^2 measure of linkage disequilibrium: impact on design and interpretation of association studies. *Twin Res. Hum. Genet.* **8**, 87–94 (2005).
20. Wray, N.R., Purcell, S.M. & Visscher, P.M. Synthetic associations created by rare variants do not explain most GWAS results. *PLoS Biol.* **9**, e1000579 (2011).
21. Kang, H.M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
22. Yang, J., Wray, N.R. & Visscher, P.M. Comparing apples and oranges: equating the power of case-control and quantitative trait association studies. *Genet. Epidemiol.* **34**, 254–257 (2010).

ONLINE METHODS

Data and quality control analysis. The 17 PGC-SCZ case-control sample cohorts (8 ISC, 1 MGS and 8 other subsets) and the common quality control steps applied to genotypes and to individual samples before imputation have previously been described in detail[14] and are briefly summarized in **Supplementary Table 1**. Imputation of autosomal SNPs used the Utah residents of Northern and Western European ancestry (CEU) and Toscani in Italia (TSI) HapMap 3 populations as a reference panel[14]. The total imputed sample included 22,279 individuals. Some individuals were excluded to ensure that all cases and controls were completely unrelated in the classical sense: no genome-wide similarities of >0.05 (equivalent to approximately second-cousin relatedness; see Eq. (3)) were permitted. We calculated the MAF and imputation R^2 (ratio of observed to expected variance) for each SNP in each of the 17 sample cohorts and retained only the SNPs with MAF >0.01 and R^2 >0.6 in all cohorts, a total of 915,354. Preliminary analyses were conducted using only genotyped SNPs from the ISC and MGS subsets (see **Supplementary Table 2**).

Linear mixed model for estimation of variance of case-control status explained by all SNPs. For this model, we used the methods previously presented[13]. Briefly, we estimated the variance in case-control status explained by all SNPs using a linear mixed model

$$\mathbf{y} = X\beta + \mathbf{g} + \mathbf{e} \tag{1}$$

where $\mathbf{y}$ is a vector of schizophrenia case (= 1) or control (= 0) status (the 'observed scale'), β is a vector for fixed effects of the overall mean (intercept), sex, sample cohort and 20 ancestry principal components (see **Supplementary Fig. 2**), $\mathbf{g}$ is a vector of random additive genetic effects based on aggregate SNP information, $\mathbf{e}$ is a vector of random error effects and X is an incidence matrix for the fixed effects that relates these effects to individuals. The variance structure of phenotypic observations was calculated by

$$V = A\sigma_g^2 + I\sigma_e^2 \tag{2}$$

where σ_g^2 is additive genetic variance captured by the SNPs, σ_e^2 is error variance, A is the realized relationship matrix estimated from SNP data and I is an identity matrix. The realized relationship for each pair of individuals was calculated as the sum of the products of SNP coefficients between two individuals scaled by SNP heterozygosity[12] with

$$\hat{A}_{ij} = \frac{1}{L}\sum_{l=1}^{L}\frac{(x_{il} - 2p_l)\cdot(x_{jl} - 2p_l)}{2p_lq_l} \quad (i \neq j) \tag{3}$$

$$\hat{A}_{ii} = 1 + \frac{1}{L}\sum_{l=1}^{L}\frac{x_{il}^2 - (1 + 2p_l)x_{ij} + 2p_l^2}{2p_lq_l}$$

where x_{il} = 0, 1 or 2 according to whether individual i has genotype bb, Bb or BB at locus l (alleles are arbitrarily called b or B), p (q) is allele frequency of B (b) and $2p$ is the mean of x_l. We used imputation best-guess genotypes. Adaptation of Eq. (3) for use with dosage scores produced similar results but was computationally slower. These realized relationships are scaled to be both positive and negative; therefore, for clarity, we have used the term 'similarity' rather than 'relationship'[23]. All variances were on the observed scale and were estimated using restricted maximum likelihood (REML)[24–26]. They were transformed to the liability scale assuming a disease prevalence of 1% (ref. 13), thus generating estimates for h^2.

Similarly, the bivariate analyses implemented a bivariate extension of Eq. (1)

$$\begin{aligned}\mathbf{y}_1 &= X_1\beta_1 + Z_1\mathbf{g}_1 + \mathbf{e}_1 \\ \mathbf{y}_2 &= X_2\beta_2 + Z_2\mathbf{g}_2 + \mathbf{e}_2\end{aligned} \tag{4}$$

where the vectors and matrices follow the definitions from Eq. (1) for each of the two subsets denoted by the subscripts. The Z incidence matrices relate observations with the vectors of random additive genetic effects. If n_1 and n_2 are the total number of cases and controls in subsets 1 and 2, respectively, then $\mathbf{y}_1$ and $\mathbf{e}_1$ have length n_1, $\mathbf{y}_2$ and $\mathbf{e}_2$ have length n_2, and $\mathbf{g}_1$ and $\mathbf{g}_2$ have length $n_1 + n_2$. The variance-covariance matrix of phenotypic observations across the two traits is

$$V = \begin{bmatrix} Z_1AZ_1'\sigma_{g_1}^2 + I\sigma_{e_1}^2 & Z_2AZ_1'\sigma_{g_{12}} \\ Z_1AZ_2'\sigma_{g_{12}} & Z_2AZ_2'\sigma_{g_2}^2 + I\sigma_{e_2}^2 \end{bmatrix} \tag{5}$$

where $\sigma_{g_{12}}$ is the additive genetic covariance captured by SNPs between the two traits. Individuals contributing to the two traits were unrelated, such that the covariance between environmental effects was assumed to be zero. The genetic correlation coefficient r was calculated with

$$r = \sigma_{g_{12}} / (\sigma_{g_1}\sigma_{g_2}) \tag{6}$$

where terms are defined as above.

Genome partitioning linear mixed model. We partitioned the variance explained by the SNPs in several ways using the linear model

$$\begin{aligned}\mathbf{y} &= X\beta + \sum_{t=1}^{n}\mathbf{g}_t + \mathbf{e} \\ V &= \sum_{t=1}^{n}A_t\sigma_{g_t}^2 + I\sigma_e^2\end{aligned} \tag{7}$$

where n is the number of subsets from any nonoverlapping partitioning of SNPs, with n = 22 for the joint analysis by chromosome, n = 5 for the analysis by MAF bin and n = 3 for the analysis of SNPs by gene annotation in which SNPs were classed as being present in CNS$^+$ genes (2,725 genes representing 547 Mb) or other genes (14,804 genes representing 1,069 Mb) or as not being localized to genes. Gene boundaries were set 50 kb up- and downstream from the 5′ and 3′ UTRs of each gene, respectively, and the CNS$^+$ genes included the four previously identified subsets[17] (one comprised genes expressed preferentially in the brain compared to other tissues, and the other three comprised genes annotated to be involved in neuronal activity, learning and synapses). We included these analyses because they showed how the variance explained by SNPs can be partitioned, but they are limited by the current state of the functional annotation of genes.

Under genomic partitioning, each pair of individuals is expected to have different estimates of similarity for each SNP set. For example, when we partitioned SNPs by MAF, genome-wide similarities between all pairs of individuals were calculated using only SNPs allocated to a given MAF bin.

Model comparisons. We used the likelihood ratio test statistic (LR) to evaluate the improved fit of the model for a given variance component term. For example, LR = –2ln (likelihood of the reduced model/likelihood of the full model), where the reduced model excluded the variance component tested. Each comparison excluded a single variance component such that the LR was distributed as a 50:50 mixture of a χ^2 distribution with 1 degree of freedom and point mass of zero[27]. LRs of 2.7, 5.4, 9.5, 13.8, 18 and 32 equated to P values of 0.05, 0.01, 0.001, 1×10^{-4}, 1×10^{-5} and 1×10^{-6}, respectively. We did not report LR values because in all cases the LR was so high that the estimates of h^2 had small standard error, with all showing a difference from zero that was highly significant.

Simulation. To evaluate the hypothesis that common variants have only a small role in the etiology of schizophrenia, we used a simulation based on the PGC-SCZ imputed genotypes to quantify whether the variance explained when fitting common SNPs simultaneously could be attributed to only rare causal variants. We calculated the realized relationship matrix between individuals on the basis of less common SNPs of MAF of <0.1, and we used this matrix to simulate quantitative genetic values. Genetic values were generated from a multivariate normal distribution by multiplying random normal variables by the Cholesky decomposition of this similarity matrix. Quantitative phenotypes were the genetic values added to random error terms drawn from a normal distribution scale such that genetic values explained either 25%, 50% or 80% of the phenotypic variance. We analyzed the simulated data with three models: (i) five separate analyses, with each analysis fitting a

similarity matrix generated with SNPs from one of the five MAF bins, (ii) a single joint analysis fitting five similarity matrices simultaneously for each of the five MAF bins and (iii) a single joint analysis fitting four similarity matrices separately for the four MAF bins with MAF of >0.1. We repeated the simulations but instead associated all variance in liability to SNPs across the entire MAF spectrum. Results were averaged across ten replicates.

23. Powell, J.E., Visscher, P.M. & Goddard, M.E. Reconciling the analysis of IBD and IBS in complex trait studies. *Nat. Rev. Genet.* **11**, 800–805 (2010).
24. Gilmour, A.R., Gogel, B.J., Cullis, B.R. & Thompson, R. *ASReml User Guide Release 2.0* (VSN International, Hemel Hempstead, UK, 2006).
25. 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).
26. Lee, S.H. & Van der Werf, J.H.J. An efficient variance component approach implementing an average information REML suitable for combined LD and linkage mapping with a general complex pedigree. *Genet. Sel. Evol.* **38**, 25–43 (2006).
27. Self, S.G. & Liang, K.Y. Asymptotic properties of maximum-likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Stat. Assoc.* **82**, 605–610 (1987).