HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2012 September 01.

Published in final edited form as:

Nat Genet. ; 44(3): 247–250. doi:10.1038/ng.1108.

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**[1,7], **The Schizophrenia Psychiatric Genome Wide Association Study Consortium (PGC-SCZ)**[11], **The International Schizophrenia Consortium (ISC)**[11], **The Molecular Genetics of Schizophrenia Collaboration (MGS)**[11], **Patrick F Sullivan**[8], **Michael E Goddard**[9,10], **Matthew C Keller**[3,4,12], **Peter M Visscher**[1,2,7,12], and **Naomi R Wray**[1,2,*,12]

[1]Queensland Institute of Medical Research, 300 Herston Road, Brisbane, Australia

[2]Queensland Brain Institute, University of Queensland, Brisbane, Australia

[3]Department of Psychology & Neuroscience, University of Colorado at Boulder, CO, USA

[4]Institute for Behavioral Genetics, University of Colorado at Boulder, CO, USA

[5]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA

[6]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA

[7]University of Queensland Diamantina Institute, Princess Alexandra Hospital, Brisbane, Queensland, Australia

[8]Department of Genetics, University of North Carolina, Chapel Hill, NC, USA

[9]Department of Agriculture and Food Systems, University of Melbourne, Melbourne, Australia

[10]Biosciences Research Division, Department of Primary Industries, Victoria, Australia;

Abstract

Schizophrenia is a complex disorder caused by both genetic and environmental factors. Using 9,087 cases, 12,171 controls and 915,354 imputed SNPs from the Psychiatric GWA Consortium for schizophrenia (PGC-SCZ) we estimate that 23% (s.e. 1%) of variation in liability to

*Corresponding author: Naomi R Wray, Queensland Brain Institute, University of Queensland, St Lucia, QLD, 4072, Australia, naomi.wray@uq,edu,au, Tel: +61 7 3346 6374.

[11]Consortium membership lists are provided in the Supplementary Information.

[12]These authors jointly directed this work

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.DeC performed preliminary analyses on the MGS sample. M.C.K directed preliminary analyses on the MGS sample. S.R. undertook the QC 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 draft of the manuscript.

COMPETING FINANCIAL INTERESTS

None

schizophrenia is captured by SNPs. We show that an important proportion of this variation must be due to common causal variants, that the variance explained by each chromosome is linearly related to its length (r = 0.89, p = 2.6×10^{-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 2725 genes expressed in the central nervous system (CNS) (p = 7.6×10^{-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.

Keywords

heritability; missing heritability; genomic variance; SNPs; GWAS

Schizophrenia is a severe mental disorder with lifetime risk ~1% and heritability of ~0.7 to 0.8[1–3]. Of all complex genetic diseases, the genetic architecture of schizophrenia perhaps has received the most speculation and debate[4,5] and the relative importance of common causal variants remains controversial[6,7]. Genome-wide association (GWA) studies of schizophrenia have discovered associated variants[8–10] that together explain only a small fraction of the heritability[11]. Here, new methods[12,13] for estimation of the variation explained by GWA genotypes are applied to PGC-SCZ data[14]. We use only cases and controls that are 'unrelated' in the classical sense and calculate the variance explained by autosomal SNPs. 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 genome-wide than case-control pairs. We partition[15] this genomic variation by chromosome, by sex, by functional annotation and by minor allele frequency.

RESULTS

Genomic variation captured by common SNPs

The PGC-SCZ includes data from the International Schizophrenia Consortium (ISC)[8], the Molecular Genetics of Schizophrenia (MGS) study[9] and other samples (OTH) (Supplementary Table 1). Using a linear mixed model (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; it is a lower bound because only variation due to association with the SNPs can be estimated. Preliminary analyses were conducted using non-imputed genotypes of the ISC and MGS subsets (Supplementary Table 2). The estimates of h^2 for the PGC-SCZ subsets of ISC, MGS and OTH 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 considering cases and controls from one subset as trait 1 and those from a different subset as trait 2 (Table 2); the two independent subsets are related through the coefficients of genome-wide similarity calculated from SNPs between individuals (Online Methods equation 2). The estimated correlation coefficients based on SNP genome-wide similarities are less than 1, consistent with several explanations. Subsets may be more homogeneous both phenotypically, for

example because of similar and consistent diagnostic criteria, and genetically, because linkage disequilibrium (LD) between causal variants and analysed SNPs may be higher within than between subsets. Alternatively, subtle artefacts could generate non-random differences in allele frequencies between sets of cases and sets of controls from the same study. However, our preliminary analyses using genotyped SNPs for ISC and MGS and extreme QC (Supplementary Table 2) suggest that this is unlikely to be a major contributor. Furthermore, the correlations between data sets from the bivariate analyses are high (~0.8) demonstrating that the same genetic signals can explain variance in schizophrenia liability in different case-control samples collected; given that these samples were collected independently with genotyping conducted at different laboratories, it is difficult to envision artefacts 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 artefacts.

Partitioning of genomic variation by chromosome

Cryptic population stratification has been proposed as a confounding factor in GWA studies[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 contribute to our results (over and above 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 joint analysis which fitted 22 similarity matrices simultaneously (estimating 22 additive genetic variance components in a single analysis) (Online Methods). A 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 26% for chromosomes fitted separately compared to a total of 23% when fitted together, demonstrating little evidence of population stratification (Figure 1a). The estimates of variance explained by each chromosome are linearly related with the length of the chromosome (correlation = 0.89, $p = 2.6 \times 10^{-8}$), consistent with a highly polygenic model and remarkably 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, response to treatment)[16]. To assess if the variance in liability tagged 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 other trait; the two independent subsets are related through the coefficients of similarity calculated from SNPs (Online Methods equation 2). The correlation in liabilities explained 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 concluded that the variance explained by the X chromosome is consistent with expectation given its length (Supplementary Table 3).

Partitioning of genomic variation by functional annotation

To assess if 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 from SNPs in "CNS+" genes, other genes and no genes (Online Methods). The CNS+ genes were the four sets identified by Raychaudhuri *et al.*[28] and comprised the genes in their brain-expressed (specifically, genes with differential CNS expression), neuronal activity, learning and synapse sets. We find that the variance attributable to the CNS+ genes is significantly greater than the proportion of the genome that they represent (31% s.e. 2% vs 20%, $p = 7.6 \times 10^{-8}$) (Figure 1b; Supplementary Table 4).

Partitioning of genomic variation by minor allele frequency of SNPs

It has been argued (e.g.[6,7,18]) that the low proportion of variance explained by previous GWA studies of schizophrenia implies that common variants are unimportant to the etiology of schizophrenia. To evaluate this hypothesis we undertook an analysis partitioning the variance tagged by SNPs into five components defined by minor allele frequency (MAF) (Online Methods). For close relatives (who are excluded from our analyses), estimated similarities based upon SNPs with different MAF will be similar. However, very distant relatives inherit chromosome segments from distant common ancestors. If a SNP is more recent than the common ancestor then the relationship between the individuals will not be reflected by the SNP; low MAF SNPs tend to be younger than high MAF SNPs. The variance explained by SNPs with MAF < 0.1 was 2% (s.e. 1%) from a joint analysis of all five MAF bins in the total PGC-SCZ data (Supplementary Table 5, Figure 1c). 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 explain approximately equal proportions of the variance, ~5% (s.e. 1%) each. Analyses of the PGC-SCZ subsets were consistent with these results (Supplementary Table 5). Based on the known relationship between allele frequencies and LD[19], it is highly unlikely that the estimates of h^2 reported here are caused 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 > 0.4, 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 which attributed 50% of variation in liability to SNPs with MAF < 0.1, SNPs with MAF > 0.4 explained only 5% (s.e. 0.3%) of the variance, which is only 10% of the variation explained by all SNPs (Figure 1c,d; Supplementary Tables 5–6). Furthermore, our simulation strategy is a best case scenario in favour of the rare variants only model since our simulation extends the frequency of "rare" variants to MAF of 0.1 generating higher LD between the common genotyped SNPs and causal variants than would be expected under a more usual MAF definition of "rare". Our results are consistent with analyses of the ISC data[8,20]. In the Supplementary Note we contrast our methods to the risk

profiling methods used by the ISC and the efficient mixed model association expedited (EMMAX) method of Kang *et al*[21].

DISCUSSION

We draw four important conclusions from these results. First, from direct queries of the genome, we quantify the lower limit of the genetic contribution to schizophrenia; approximately one quarter of the variance in liability is directly tagged by common variants represented across the current generation of GWA 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 (Figure 1d). Third, we provide evidence that the variance explained by chromosomes is linearly related to the length of the chromosome (Figure 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 relative to the proportion of the genome it represents. Together our results provide guidance for the future of genetic studies in schizophrenia. Some have argued[6,7,18] that common variants play little role in the etiology of schizophrenia and that the GWA approach for schizophrenia has been misconceived. Our results refute this conjecture that common variants play little role in the etiology of schizophrenia and that the GWA approach for schizophrenia has been misconceived by demonstrating that at least one quarter of variation in liability to schizophrenia is tagged 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 necessary to detect additional effects (over those detected to date) with genome-wide significance. For example, a GWA 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. Sample sizes of ~50,000 schizophrenia cases and 50,000 controls are needed to afford the same power to detect variants that explain the same proportion of phenotypic variance and gain insight into biological pathways achieved in the height study[11,12,22]. Our results imply that the GWA approach applied to larger case-control samples will deliver important 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 tagged when considering all genotyped and imputed SNP 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 explains less of variance attributable to uncommon and rare variants (Supplementary Figure 1). From the analyses we have performed we cannot estimate a frequency 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 MAF. 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 spectrum”.

METHODS

See Online Methods for full details.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We acknowledge funding from the Australian National Health and Medical Research Council (grants 389892, 442915, 496688, 613672 and 613601) the Australian Research Council (grants DP0770096, DP1093502 and FT0991360), the National Institutes of Mental Health (MH085812). This research utilised the Cluster Computer funded by the Netherlands Scientific Organization (NWO 480-05-003). We thank Scott D Gordon for technical assistance.

PGC-SCZ Acknowledgements. We thank the study participants, and the research staff at the many study sites. Over 40 NIH grants (USA), and similar numbers of government grants from other countries, along with substantial private and foundation support enabled this work. We greatly appreciate the sustained efforts of Thomas Lehner (National Institute of Mental Health) on behalf of the Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium (PGC). Detailed acknowledgements, including grant support, are listed in the Supplementary Materials of[14]. Some authors declare competing financial interests: details are available in the Supplementary Materials of[14].

References

1. Sullivan PF, Kendler KS, Neale MC. Schizophrenia as a complex trait -Evidence from a meta-analysis of twin studies. Archives of General Psychiatry. 2003; 60:1187–1192. [PubMed: 14662550]
2. Cardno AG, Gottesman II. Twin studies of schizophrenia: from bow-and-arrow concordances to star wars Mx and functional genomics. Am J Med Genet. 2000; 97:12–17. [PubMed: 10813800]
3. Lichtenstein P, et al. Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. Lancet. 2009; 373:234–239. [PubMed: 19150704]
4. McClellan JM, Susser E, King MC. Schizophrenia: a common disease caused by multiple rare alleles. British Journal of Psychiatry. 2007; 190:194–199. [PubMed: 17329737]
5. Craddock N, O'Donovan MC, Owen MJ. Phenotypic and genetic complexity of psychosis -Invited commentary on … Schizophrenia: a common disease caused by multiple rare alleles. British Journal of Psychiatry. 2007; 190:200–203. [PubMed: 17329738]
6. McClellan J, King MC. Genetic Heterogeneity in Human Disease. Cell. 2010; 141:210–217. [PubMed: 20403315]
7. McClellan J, King MC. Genomic Analysis of Mental Illness A Changing Landscape. Jama-Journal of the American Medical Association. 2010; 303:2523–2524.
8. Purcell SM, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009; 460:748–752. [PubMed: 19571811]
9. Shi JX, et al. Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature. 2009; 460:753–757. [PubMed: 19571809]
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. Molecular Psychiatry. 2009; 14:252–260. [PubMed: 19065143]
11. Visscher PM, Goddard ME, Derks EM, Wray NR. Evidence based psychiatric genetics, AKA the false dichotomy between common and rare variant hypotheses. Molecular Psychiatry. Epub.
12. Yang J, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010; 42:565–569. [PubMed: 20562875]

13. Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. American Journal of Human Genetics. 2011; 88:294–305. [PubMed: 21376301]
14. Ripke S, et al. Genome-wide association study identifies five new schizophrenia loci. Nature Genetics. 2011; 43:969–U77. [PubMed: 21926974]
15. Yang J, et al. Genome partitioning of genetic variation for complex traits using common SNPs. Nature Genetics. 2011; 43:519–U44. [PubMed: 21552263]
16. Abel KM, Drake R, Goldstein JM. Sex differences in schizophrenia. International Review of Psychiatry. 2010; 22:417–428. [PubMed: 21047156]
17. Allen HL, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010; 467:832–838. [PubMed: 20881960]
18. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nature Reviews Genetics. 2010; 11:415–425.
19. Wray NR. Allele frequencies and the r2 measure of linkage disequilibrium: Impact on design and interpretation of association studies. Twin Research and Human Genetics. 2005; 8:87–94. [PubMed: 15901470]
20. Wray NR, Purcell SM, Visscher PM. Synthetic associations created by rare variants do not explain most GWAS results. Plos Biology. 2011; 9:e1000579. [PubMed: 21267061]
21. Kang HM, et al. Variance component model to account for sample structure in genome-wide association studies. Nature Genetics. 2010; 42:348–U110. [PubMed: 20208533]
22. Yang J, Wray NR, Visscher PM. Comparing Apples and Oranges: Equating the Power of Case-Control and Quantitative Trait Association Studies. Genetic Epidemiology. 2010; 34:254–257. [PubMed: 19918758]
23. Powell JE, Visscher PM, Goddard ME. Reconciling the analysis of IBD and IBS in complex trait studies. Nature Reviews Genetics. 2010; 11:800–805.
24. Gilmour, AR.; Gogel, BJ.; Cullis, BR.; Thompson, R. ASReml User Guide Release 2.0. VSN International; Hemel Hempstead, UK: 2006.
25. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for Genome-wide Complex Trait Analysis. American Journal of Human Genetics. 2011; 88:76–82. [PubMed: 21167468]
26. Lee SH, Van der Werf JHJ. 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. 2006; 38:25–43. [PubMed: 16451790]
27. Self SG, Liang KY. Asymptotic properties of maximum-likelihood estimators and likelihood ratio tests under nonstandard conditions. Journal of the American Statistical Association. 1987; 82:605–610.
28. Raychaudhuri S, et al. Accurately Assessing the Risk of Schizophrenia Conferred by Rare Copy-Number Variation Affecting Genes with Brain Function. Plos Genetics. 2010; 6:e1001097. [PubMed: 20838587]

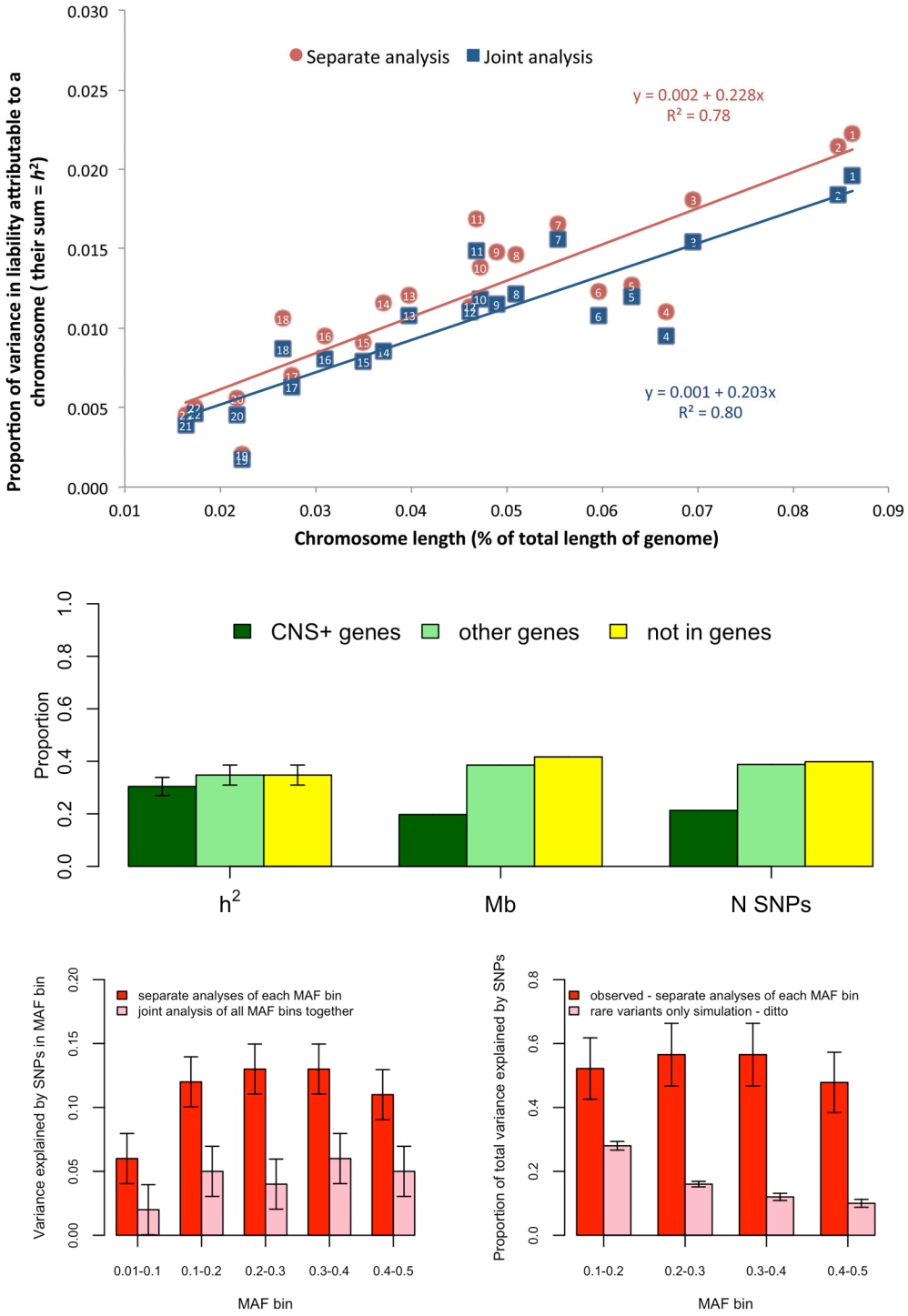

Figure 1.

Genomic partitioning of schizophrenia. **a) By chromosome**: Estimated proportion of the variance in liability to schizophrenia explained bySNPs on individual chromosomes from a joint analysis of all chromosomes simultaneously or separate analyses for each chromosome The sum of the h^2 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) in CNS+ genes, other genes and by

those not in genes totals 0.23. Of this, a proportion 0.31 is attributed to SNPs in brain genes, which is greater than expected by chance ($p = 7.6 \times 10^{-8}$) given that the brain genes cover 0.20 of the length of the genome (Mb) and represent 0.21 of the SNP count (N SNPs). Error bars represent the 95% confidence intervals of the estimates. **c) By MAF bin from analyses fitting MAF bins jointly or each separately d) By MAF bin compared to simulation under a rare variants only model.** The variance explained by SNPs in each MAF bin (when MAF bins are fitted in separate analyses) as a proportion of the variance explained by all SNPs. Error bars represent 95% confidence intervals, for the simulations (right graph) these are calculated using the standard deviation across simulation replicates.

Table 1

Estimated proportion of variance in liability to schizophrenia explained by SNPs (h^2).

Dataset	Cases	Controls	h^2 (SE)
ISC	3220	3445	0.27 (0.02)
MGS	2571	2419	0.31 (0.03)
OTH	3296	6307	0.27 (0.02)
ISC+MGS	5791	5864	0.25 (0.01)
PGC-SCZ	9087	12171	0.23 (0.01)

Estimates based on 915354 imputed SNPs; SE standard error of h^2. PGC-SCZ is comprised of independent subsets ISC, MGS and OTH

Table 2

Bivariate analyses of PGC-SCZ subsets

Subset 1/Subset 2	Cases Subsets 1/2	Controls Subsets 1/2	Subset 1 h^2 (SE)	Subset 2 h^2 (SE)	r (SE)
ISC/MGS	3220/2571	3445/2419	0.26 (0.02)	0.29 (0.03)	0.84 (0.09)
ISC/OTH	3220/3296	3445/6307	0.26 (0.02)	0.27 (0.02)	0.89 (0.07)
MGS/OTH	2571/3296	2419/6307	0.30 (0.03)	0.26 (0.02)	0.79 (0.08)
ISC+MGS/OTH	5791/3296	5864/6307	0.24 (0.01)	0.26 (0.02)	0.87 (0.06)
Male/Female	6031/3056	5884/6287	0.24 (0.01)	0.25 (0.02)	0.89 (0.06)

Estimates based on 915354 imputed SNPs; h^2 estimate of proportion of variance in liability to schizophrenia explained by SNPs; SE standard error of h^2. r correlation of liabilities explained by SNPs between subset 1 and subset 2. PGC-SCZ is comprised of independent subsets ISC, MGS and OTH.