Psychological Medicine (2012), **42**, 607–616. © Cambridge University Press 2011
doi:10.1017/S0033291711001607

ORIGINAL ARTICLE

Hypothesis-driven candidate genes for schizophrenia compared to genome-wide association results

A. L. Collins[1], Y. Kim[1], P. Sklar[2]; International Schizophrenia Consortium, M. C. O'Donovan[3] and P. F. Sullivan[1]*

[1] *Department of Genetics, University of North Carolina at Chapel Hill, NC, USA*
[2] *Department of Psychiatry, Mt Sinai School of Medicine, New York, NY, USA*
[3] *MRC Centre for Neuropsychiatric Genetics and Genomics, School of Medicine, Cardiff University, UK*

Background. Candidate gene studies have been a key approach to the genetics of schizophrenia (SCZ). However, the results of these studies are confusing and no genes have been unequivocally implicated. The hypothesis-driven candidate gene literature can be appraised by comparison with the results of genome-wide association studies (GWAS).

Method. We describe the characteristics of hypothesis-driven candidate gene studies from the SZGene database, and use pathway analysis to compare hypothesis-driven candidate genes with GWAS results from the International Schizophrenia Consortium (ISC).

Results. SZGene contained 732 autosomal genes evaluated in 1374 studies. These genes had poor statistical power to detect genetic effects typical for human diseases, assessed only 3.7% of genes in the genome, and had low marker densities per gene. Most genes were assessed once or twice (76.9%), providing minimal ability to evaluate consensus across studies. The ISC studies had 89% power to detect a genetic effect typical for common human diseases and assessed 79% of known autosomal common genetic variation. Pathway analyses did not reveal enrichment of smaller ISC *p* values in hypothesis-driven candidate genes, nor did a comprehensive evaluation of meta-hypotheses driving candidate gene selection (SCZ as a disease of the synapse or neurodevelopment). The most studied hypothesis-driven candidate genes (*COMT*, *DRD3*, *DRD2*, *HTR2A*, *NRG1*, *BDNF*, *DTNBP1* and *SLC6A4*) had no notable ISC results.

Conclusions. We did not find support for the idea that the hypothesis-driven candidate genes studied in the literature are enriched for the common genetic variation involved in the etiology of SCZ. Larger samples are required to evaluate this conclusion definitively.

Received 24 March 2011; Revised 18 July 2011; Accepted 25 July 2011; First published online 19 August 2011

Key words: Candidate gene association, comparative study, genetics, genome-wide association, neurodevelopmental, schizophrenia, synapse.

Introduction

Candidate gene studies have been a major focus in schizophrenia (SCZ) research with the SZGene database listing more than 1400 studies since 1965 (Allen *et al*. 2008). By contrast, there are around 2200 PubMed citations for 'schizophrenia randomized controlled trials'. Until 5 years ago, genetic studies could investigate only an extremely small proportion of the genome because of genotyping and cost limitations. Investigators thus had to focus on a limited number of genetic markers, genes and samples. In hypothesis-driven candidate gene studies, targets were selected based on ideas about pathophysiology or gene location under a linkage peak (Cichon *et al*. 2009). For most biomedical disorders (including SCZ), the results of these studies were inconsistent or confusing (Ioannidis *et al*. 2001). It is unclear whether this reflects poor choices of candidate genes or inadequate assessment (i.e. small samples or incomplete coverage of common genetic variation).

Genotyping and cost improvements now permit routine assessment of a million or more genetic variants distributed across the genome (Beaudet & Belmont, 2008). Genome-wide association studies (GWAS) can extract information from most common genetic variants in the genome either through direct assessment of single nucleotide polymorphisms (SNPs) or indirectly through linkage disequilibrium (LD) between genotyped SNPs and unmeasured but correlated genetic variants. Despite the advantages of

* Address for correspondence: P. F. Sullivan, M.D., FRANZCP, Department of Genetics, CB#7264, 5097 Genomic Medicine, University of North Carolina, Chapel Hill, NC 27599-7264, USA.
(Email: pfsulliv@med.unc.edu) [P.F.S.]
(Email: collin@med.unc.edu) [A.L.C.]

genome-wide genotyping (Hindorff *et al*. 2009), stringent adjustment for multiple comparisons is required, which necessitates the use of large sample collections.

Ten GWAS for SCZ have been published (Lencz *et al*. 2007; O'Donovan *et al*. 2008; Shifman *et al*. 2008; Sullivan *et al*. 2008; Kirov *et al*. 2009; Need *et al*. 2009; Purcell *et al*. 2009; Shi *et al*. 2009; Stefansson *et al*. 2009; Athanasiu *et al*. 2010). Given that some GWAS had larger samples and more comprehensive genotyping than typical for the candidate gene literature, GWAS may be better placed to capture true associations than earlier studies. Indeed, GWAS have yielded highly significant and replicated associations for SCZ, including genetic variation in the major histocompatibility complex (MHC) region and within *TCF4* and *ZNF804A* (O'Donovan *et al*. 2008; Purcell *et al*. 2009; Shi *et al*. 2009; Stefansson *et al*. 2009). A lack of congruity has been noted between the hypothesis-driven candidate genes for SCZ and the best findings from GWAS. This may be typical for biomedical diseases, where results from large GWAS infrequently correspond to *a priori* candidate genes (www.genome.gov/gwastudies; Hindorff *et al*. 2009).

For SCZ, there are multiple reasons for the lack of overlap between GWAS and candidate gene studies. A key possibility is that prior hypotheses about the genetics of SCZ are incorrect. However, alternative explanations require exploration before accepting such an important conclusion. First, current GWAS chips provide coverage for most but not all of the genome (International HapMap Consortium, 2005), so particular regions and non-SNP genetic variation may be covered poorly. Second, power may be insufficient. Although GWAS tend to have large sample sizes by historical standards, the necessity to adjust for around a million statistical tests could result in low power. If that is the explanation, support for the hypotheses underpinning previous candidate genes might be obtained by a more systematic analysis of the GWAS data for evidence for over-representation of smaller p values than expected by chance (Holmans *et al*. 2009). Third, individually rare genetic variants of strong effect might also be missed by GWAS (although these would also go undetected by most prior candidate gene studies).

The main purpose of this study was to compare hypothesis-driven candidate genes for SCZ from SZGene (Allen *et al*. 2008) with the largest SCZ GWAS published to date (International Schizophrenia Consortium, 2008; Purcell *et al*. 2009). First, we characterized the hypothesis-driven candidate gene studies. Second, we conducted quantitative comparisons to determine whether the set of hypothesis-driven candidate genes were either enriched for lower p values in the International Schizophrenia Consortium (ISC) results or contained markers with predictive power for SCZ. Finally, we performed a qualitative comparison of the most studied hypothesis-driven candidate genes with the ISC results.

Method

Hypothesis-driven candidate genes

Candidate genes for SCZ were drawn from SZGene (Allen *et al*. 2008) (courtesy of Dr L. Bertram, 11 September 2009). SZGene included studies evaluating associations between a genetic variant and SCZ and published in a peer-reviewed, English-language journal (Allen *et al*. 2008). Studies were identified through PubMed, bibliographies, and tables of contents. To ensure that the list was not 'contaminated' by the results of GWAS, SZGene entries from GWAS were removed, as were genes studied only subsequent to identification in GWAS. As only ISC autosomal results were available, 15 chromosome X genes were dropped. The list of autosomal hypothesis-driven candidate genes for SCZ contained 732 genes from 1374 studies (see Supplementary Table S1, available online).

The purpose in creating this list was to enumerate genes thought to be potentially etiological for SCZ by researchers in the field. The quality of the individual studies was variable. However, our interest was not in the study results *per se* (which can be strongly impacted by quality) but rather in the choice of a gene.

Samples

The ISC study is described elsewhere (International Schizophrenia Consortium, 2008; Purcell *et al*. 2009). In brief, we studied 3322 European cases with DSM-IV or ICD-10 SCZ and 3587 controls from seven sites. All work was approved by institutional review boards. After complete description of the study to the subjects, written informed consent was obtained. Genotyping was performed on DNA extracted from blood using Affymetrix 5.0 or 6.0 arrays. Genotypes were called using Birdsuite (Korn *et al*. 2008) and imputation conducted using Beagle (Browning & Browning, 2007, 2009) against the HapMap3 CEU data resulting in 1 948 385 autosomal SNPs with direct or imputed genotypes. The primary analysis was logistic regression of disease state on the imputed allele dosages with inclusion of covariates to adjust for population stratification effects. The Genetic Association Information Network (GAIN) study genotyped 1230 SCZ cases and 1136 controls of European ancestry on the Affymetrix 6.0 array (Shi *et al*. 2009).

Statistical analysis

We first explored the set of hypothesis-driven candidate genes using a variety of descriptive approaches

https://doi.org/10.1017/S0033291711001607 Published online by Cambridge University Press

(SAS Institute Inc., 2004, 2005). Quanto was used for power calculations (Gauderman, 2002; Gauderman & Morrison, 2006). We used the Database for Annotation, Visualization, and Integrated Discovery (DAVID; Dennis *et al*. 2003; Sherman *et al*. 2007; Huang *et al*. 2009) to characterize hypotheses about the pathophysiology of SCZ reflected in the hypothesis-driven candidate gene list. DAVID identifies Gene Ontology (GO) biological pathways (Harris *et al*. 2004) with chance-corrected over-representation of a given gene list. To account for overlap between pathways, we used the annotation cluster feature in DAVID to focus on higher-level clusters of similar biological processes.

We then compared hypothesis-driven candidate genes for SCZ with ISC GWAS results to assess whether the hypothesis-driven candidate gene list had over-representation of smaller ISC *p* values than expected by chance. These analyses were conducted using ALIGATOR (Holmans *et al*. 2009) and InRich (Lee *et al*. 2011). These programs use different algorithms to assess whether GWAS findings are over-represented for small *p* values with reference to a predefined set of genes (i.e. a pathway). ALIGATOR uses permutation to account for variable numbers of SNPs per gene, different patterns of LD between SNPs (within the same gene), and varying gene sizes. We considered SZGene hypothesis-driven candidate genes as a 'pathway' and used ALIGATOR to estimate the probability that this list contained an over-representation of smaller ISC GWAS *p* values. The ISC GWAS results were input to ALIGATOR, which assigned these SNPs to UCSC hg18 RefSeq genes (Pruitt *et al*. 2005). We determined the significance threshold (generally 0.002–0.004) that designated the top 5% of all genes as 'significant' (Holmans *et al*. 2009). The key statistical comparison is akin to a 2×2 table of whether a gene is in the top 5% by whether a gene is a member of a pathway. Assessing significance is complex because of violation of independence assumptions. ALIGATOR uses an SNP-based permutation algorithm to create a reference distribution for a pathway. InRich controls LD between genes by comparing a gene set of interest to LD-independent regions. Using the same significance thresholds as in ALIGATOR, we identified LD-independent significant regions from the ISC dataset using the clump function within PLINK ($r^2 = 0.5$ over 1 Mb). We then used InRich to determine whether the candidate gene set showed enrichment for these regions.

Polygenic score analysis was conducted as described in the ISC GWAS paper by Purcell *et al*. (2009). We used 14 811 SNPs that were genotyped in both the ISC and SCZ GAIN samples and that mapped to candidate genes. We made a polygenic profile based on the risk alleles within these SNPs in the ISC data, and used this profile to create a polygenic score for each individual within the SCZ GAIN sample (an independent target sample). We used logistic regression between the score of each individual in SCZ GAIN and their case/control status.

Results

Characteristics of hypothesis-driven candidate gene studies of SCZ

Table 1 describes hypothesis-driven candidate genes from SZGene (Allen *et al*. 2008). There were 732 autosomal genes from 1374 hypothesis-driven candidate gene studies (Supplementary Table S1). These genes were studied from one to 81 times. Most genes (563; 76.9%) were investigated in one (60.9%) or two studies (16.0%). Although replication is important in human genetics, there is little capacity to evaluate both false-positive and false-negative findings. Two-thirds of genes were assessed by $\leqslant 3$ markers and had a median SNP density of 15.4 kb/SNP. The median numbers of cases (191) and controls (214) were modest.

We used pathway analysis to characterize the hypotheses that guided candidate gene selection. The 732 hypothesis-driven candidate genes were entered into DAVID to assess the GO biological processes to which these genes belonged. The top four annotation clusters consist of biological processes involved in synaptic transmission, neuronal development and morphogenesis, regulation of synaptic transmission, and response to chemical stimuli (Table 1 and Supplementary Table S2). The full list reflects a diversity of ideas about SCZ etiology; as expected, cluster enrichment scores were particularly strong for candidate genes selected under the hypotheses that SCZ is a disease of the synapse and/or a neurodevelopmental process.

We next evaluated completeness and coverage for the hypothesis-driven candidate genes. First, we estimated statistical power to detect association. Even for relatively large studies (i.e. samples sizes at the 90th percentiles of $n_{\text{case}} = 537$ and $n_{\text{control}} = 628$), and a liberal correction for multiple comparisons ($\alpha = 0.005$, 10 markers), the power was 48% to detect genetic effects typical for GWAS of human diseases (median genotypic relative risk of 1.28 and median minor allele frequency of 0.29 for disease associations with $p < 5 \times 10^{-8}$) (Hindorff *et al*. 2009). Second, we assessed genomic coverage. The 732 hypothesis-driven candidate genes represent 3.7% of RefSeq autosomal genes (Pruitt *et al*. 2005). Marker coverage can be assessed only generally, but included only small proportions of common genetic variation. Finally, of all genes

https://doi.org/10.1017/S0033291711001607 Published online by Cambridge University Press

Table 1. *Characteristics of hypothesis-driven candidate gene studies and the ISC GWAS*

Characteristic	Candidate gene studies	ISC
Studies	1375	–
Genes (autosomal)	732	–
Markers (autosomal)		
Total	6934	1 948 385
Markers per gene, median (IQR)	2 (1–5)	9 (1–34)
Genes with 1, 2 or 3 markers (%)	65.9	N.A.
Marker density per gene as kb/SNP, median (IQR)	15.4 (4.69–46.2)	1.38 (0.69–2.47)
Sample size[a]		
Total subjects	412 (27–5623)	6909
Number of cases	191 (5–2434)	3322
Number of controls	214 (12–4899)	3587
Major annotation clusters from pathway analysis (enrichment score)		
Synaptic transmission	52.2	N.A.
Neuronal development and morphogenesis	32.4	N.A.
Regulation of synaptic transmission	22.6	N.A.
Response to chemical stimuli	22.2	N.A.
Statistical power[b]	0.48	0.89
Proportion of autosomal RefSeq genes studied[c]	0.037	0.902
Proportion of genes in top four DAVID annotation clusters studied	0.067	N.A.

ISC, International Schizophrenia Consortium; GWAS, genome-wide association studies; SNP, single nucleotide polymorphism; IQR, interquartile range; N.A., not applicable; DAVID, Database for Annotation, Visualization, and Integrated Discovery.

[a] Values given as median (range) for candidate gene studies (biased due to subject overlap across publications) or actual number for ISC.

[b] See text for assumptions.

[c] For ISC, assuming gene boundaries expanded by ± 10 kb and SNP density <20 kb/SNP.

comprising pathways in the top four DAVID annotation clusters, only 6.7% had ever been studied. Although these pathways may be overinclusive, the main hypotheses guiding selection of hypothesis-driven candidate genes were evaluated incompletely.

Hypothesis-driven candidate gene studies and the ISC GWAS

The ISC GWAS had 3322 cases, 3587 controls and 1 948 385 genotyped and imputed autosomal SNPs. The sample size was 1.36 times greater than the largest hypothesis-driven candidate gene study. Statistical power was 89% to detect a genetic effect corresponding to that typical for SNPs implicated in human disease GWAS (Hindorff *et al*. 2009) including adjustment for multiple comparisons ($\alpha = 5 \times 10^{-8}$). The ISC reported 1 948 385 associations, which exceeds the total associations in the SZGene database by over 180-fold. The mean marker density over the genome was 1 SNP/1.6 kb. In comparison to HapMap (r27, phases I+II+III), ISC markers assessed 79.0% of known common variants present in individuals of European ancestry either directly or indirectly with $r^2 \geqslant 0.8$.

We next investigated coverage and gene size (using strict gene boundaries, ± 0 kb). The 732 hypothesis-driven candidate genes were markedly larger than the remaining 18 891 autosomal RefSeq genes (median 33.5 *v.* 20.5 kb, Wilcoxon $p = 4 \times 10^{-20}$). Importantly, ISC SNP densities were similar in hypothesis-driven candidate genes in comparison to all other autosomal genes (median 1360 *v.* 1379 bases/SNP, Wilcoxon

https://doi.org/10.1017/S0033291711001607 Published online by Cambridge University Press

Table 2. *Testing for over-representation of smaller ISC GWAS p values in hypothesis-driven candidate genes for SCZ and genes corresponding to major hypotheses*

Gene list	Boundary ±0 kb	Boundary ±10 kb
SZGene hypothesis-driven candidate genes		
Full set of genes	0.05	0.02
Full set, excluding MHC region	0.48	0.49
Full set, LD correction (InRich)	0.63	0.19
Genes studied ⩾3 times	0.19	0.45
Genes studied ⩾3 times, excluding MHC region	0.65	0.52
Genes in DAVID cluster 1 (synaptic transmission)		
Full set of 4808 genes	1	1
Subset of 222 hypothesis-driven candidate genes	0.99	1
Genes in DAVID cluster 2 (neuronal development and morphogenesis)		
Full set of 4834 genes	1	1
Subset of 401 hypothesis-driven candidate genes	0.73	0.95

ISC, International Schizophrenia Consortium; GWAS, genome-wide association studies; SCZ, schizophrenia; MHC, major histocompatibility complex; LD, linkage disequilibrium; DAVID, Database for Annotation, Visualization, and Integrated Discovery.

Empirical *p* values (from ALIGATOR unless otherwise noted) testing for over-representation of smaller ISC GWAS *p* values in a given gene list in comparison to that expected by chance (10 000 permutations). Single nucleotide polymorphisms (SNPs) were mapped to strict (±0 kb) or expanded (±10 kb) gene boundaries. SNP thresholds to select the top 5% of genes varied from 0.002 to 0.004.

$p = 0.25$). A sizable fraction of autosomal RefSeq genes had no ISC SNPs within their boundaries (18.7%). As expected, genes with no SNPs were markedly smaller (median 4.1 *v.* 28.2 kb, Wilcoxon $p \approx 0$). As the 732 SCZ candidate genes were larger, they were significantly less likely to have no coverage than the remaining RefSeq genes (10.0% *v.* 19.0%, $p = 5 \times 10^{-11}$). Although this generation of GWAS chips provides partial genomic coverage of common variation, hypothesis-driven candidate genes for SCZ had better coverage than other RefSeq genes.

Do the ISC GWAS data support hypothesis-driven candidate genes as significant contributors to SCZ risk?

There were no SNPs within the gene boundaries (±0 kb) for 73 hypothesis-driven candidate genes and no SNPs within expanded gene boundaries (±10 kb) for 27 genes (Supplementary Table S3). Hypothesis-driven candidate genes with no ISC coverage were excluded from enrichment analyses. Of the ~1.9 million ISC SNPs, 56 981 mapped within hypothesis-driven candidate genes (±0 kb) and 65 803 mapped within expanded gene boundaries (±10 kb).

We assessed whether hypothesis-driven candidate genes were over-represented for smaller ISC GWAS *p* values using ALIGATOR. The central comparison was whether there was an over-representation of the top 5% of significant genes in the hypothesis-driven candidate gene list in comparison to the remaining annotated genes. In Table 2, there was a nominally significant over-representation of smaller *p* values in the ISC data for the full set of hypothesis-driven candidate genes but these values did not survive multiple testing correction. In addition, there was no evidence for over-representation of small ISC *p* values in the most studied hypothesis-driven candidate genes (⩾3 times, 23.1% of the total).

Pathway analysis can be complex in regions such as the MHC that have extensive disequilibrium between genes (Stenzel *et al*. 2004). When we repeated the ALIGATOR analysis after excluding genes in the MHC region, there was no evidence for over-representation of smaller *p* values ($p \approx 0.48$), indicating that the marginal enrichment was due to bias. We repeated the pathway analysis using InRich, which may be more robust than ALIGATOR in regions of high LD. InRich evaluates regions defined by LD. We tested the full candidate gene dataset, using the same significance thresholds as in ALIGATOR, and found no

https://doi.org/10.1017/S0033291711001607 Published online by Cambridge University Press

evidence for enrichment of significant findings in the candidate genes (Table 2). Therefore, the pathway analyses are consistent with the null hypothesis because all p values were non-significant or marginal and would not survive correction for multiple testing, and because removal of the MHC region and analysis with a program that corrects for LD indicates the results are a false positive resulting from extensive LD in the MHC region.

In addition, we evaluated whether the list of historical candidate genes, as a whole, were making a significant contribution to risk of SCZ by evaluating the polygenic score profile. This approach has provided support for an important polygenic contribution to SCZ (Purcell *et al*. 2009). We created a polygenic score profile for SNPs that mapped to historical candidate genes using the ISC data. We then applied this score to an independent dataset (SCZ GAIN, $n=1230$ cases and 1136 controls). Independent SCZ cases did not have greater risk scores than controls based on these historical candidate genes ($p=0.92$).

Many hypothesis-driven candidate genes were selected from two overarching hypotheses: SCZ as a synaptic or a neurodevelopmental disorder (Table 1). These hypotheses have been incompletely assessed. We used pathway analysis of the ISC data to assess these hypotheses in far more detail than has been possible previously. The set of 4808 genes that comprise the synaptic cluster 1 from DAVID did not show over-representation of lower ISC p values ($p\approx1$). This list may be overinclusive and the candidate genes studied might be more refined and appropriate to SCZ; however, the subset of 222 cluster 1 genes investigated in a hypothesis-driven candidate gene study did not have over-representation of smaller ISC p values ($p\approx1$). Similarly, genes in DAVID cluster 2 (neurodevelopment) did not show enrichment for lower ISC p values for the full set (4834 genes) or the subset evaluated in a candidate gene study (401 genes, all p values non-significant).

Qualitative comparisons

Pathway analysis considers sets of genes in aggregate. Negative aggregate results could miss true over-representation of small p values in a small number of hypothesis-driven candidate genes. Table 3*a* depicts the ISC findings for the 10 most-studied hypothesis-driven candidate genes. There was inadequate coverage for two small genes (*DRD4* and *APOE*), and the remainder had good SNP densities but weak ISC results, with none surviving a liberal gene-wise Bonferroni correction. Supplementary Fig. S2 depicts these genes and highlights regions of conspicuous attention in the literature (*COMT*/val58met, *DRD3*/ser9gly, *DRD2*/Taq1A, *HTR2A*/T102C, *NRG1*/Hap_{ICE}, *BDNF*/val66met, *DTNBP1*, and *SLC6A4*/HTTLPR). The ISC results do not implicate common genetic variation in these genes. Although the region containing *SLC6A4* shows no signal, the widely studied promoter polymorphism HTTLPR was not directly genotyped and neighboring SNPs are in low LD (Konneker *et al*. 2010).

We also investigated the 35 ISC SNPs with associations $<5\times10^{-8}$ and all were located at chr6:31.58–32.77 mb in the MHC region (Purcell *et al*. 2009). These SNPs map to 10 genes (Table 3*b*), five of which had not previously been the subject of a candidate gene study. Four genes had been studied 1–5 times and one extensively (*NOTCH4*). The strong caveat for Table 3*b* is the extensive LD in the MHC region (Supplementary Fig. S1); these genome-wide significant SNPs could reflect one or a few risk variants, which may or may not be a candidate gene.

Discussion

The main aim of this study was to evaluate the hypothesis-driven candidate gene literature for SCZ with respect to a large GWAS dataset. Hypothesis-driven candidate gene studies have been a major approach to the molecular etiology of SCZ. However, we now can perform analyses of several orders of magnitude more detailed than were possible even 5 years ago. We wanted to determine whether the systematic investigations now allowed by GWAS supported this body of work in aggregate and the degree to which GWAS empirical results support the overarching concepts that influenced candidate gene selection. We emphasize that we did not conduct meta-analyses of the findings of hypothesis-driven candidate gene studies as this has been done elsewhere (Allen *et al*. 2008).

We acknowledge the advantages of hindsight. The hypothesis-driven candidate gene literature, a body of work to which the present authors have contributed, contains numerous studies that were state of the art when they were performed and represent considerable effort by teams of investigators. GWAS will undoubtedly be subject to similar scrutiny that performed here for candidate gene studies. Although the ISC study is the largest and most comprehensive of the SCZ GWAS published to date, it was not ideal. Statistical power was high by historical standards, but we now know that greater power is desirable to detect the small genotypic relative risks characteristic of SCZ. In addition, coverage was not necessarily 'genome-wide', as some important regions had inadequate genotyping and rare genetic variation was poorly assessed. With these caveats in mind, several

Table 3*a*. *ISC results for the 10 most studied genes from SZGene*

SZGene studies	Gene	Product	Gene location (hg18)	ISC SNPs (±10 kb)	Density (kb/SNP)	Minimum *p* value	Bonferroni correction
81	*COMT*	Catechol-*O*-methyltransferase	chr22:18309262–18337496	43	1.1	0.042	1
71	*DRD3*	Dopamine receptor D_3	chr3:115330246–115380589	54	1.3	0.082	1
67	*DRD2*	Dopamine receptor D_2	chr11:112785526–112851211	90	1	0.14	1
57	*HTR2A*	Serotonin receptor 2A	chr13:46305513–46368176	118	0.7	0.017	1
45	*DRD4*	Dopamine receptor D_4	chr11:627304–630703	1	N.A.	N.A.	N.A.
41	*NRG1*	Neuregulin 1	chr8:31616809–32742100	1211	0.9	0.019	1
40	*BDNF*	Brain-derived neurotrophic factor	chr11:27633017–27700181	39	2.2	0.0039	0.15
32	*APOE*	Apolipoprotein E	chr19:50100878–50104490	1	N.A.	N.A.	N.A.
32	*DTNBP1*	Dystrobrevin-binding protein 1	chr6:15631017–15771250	140	1.1	0.026	1
32	*SLC6A4*	Serotonin transporter	chr17:25547505–25587080	27	2.2	0.097	1

ISC, International Schizophrenia Consortium; SNP, single nucleotide polymorphism; N.A., not applicable.

Table 3*b*. *Genes with genome-wide significant ISC results and studies from SZGene*

SNP	ISC *p* value	SNP location (hg18)	Gene	Product	SZGene studies
rs1150752	3.3×10^{-9}	chr6:32172704	*TNXB*	Tenascin XB	4
rs3132956	3.7×10^{-9}	chr6:32287416	*NOTCH4*	Notch homolog 4 (Drosophila)	24
rs389884	9.9×10^{-9}	chr6:32048876	*STK19*	Serine/threonine kinase 19	0
rs9268208	1.2×10^{-8}	chr6:32388569	*C6orf10*	Chromosome 6 open reading frame 10	0
rs1270942	1.2×10^{-8}	chr6:32026839	*CFB*	Complement factor B	5
rs3130614	2.5×10^{-8}	chr6:31584437	*MICB*	MHC class I polypeptide-related sequence B	2
rs2734583	2.9×10^{-8}	chr6:31613459	*BAT1*	HLA-B associated transcript 1	1
rs3117577	3.0×10^{-8}	chr6:31835453	*MSH5*	MutS homolog 5	0
rs2187668	3.5×10^{-8}	chr6:32713862	*HLA-DQA1*	MHC, class II, DQ alpha 1	0
rs3135394	4.1×10^{-8}	chr6:32516475	*HLA-DRA*	MHC, class II, DR alpha	0

ISC, International Schizophrenia Consortium; SNP, single nucleotide polymorphism; MHC, major histocompatibility complex.

observations of the historical candidate gene literature emerged from our analyses.

First, hypothesis-driven candidate gene studies had poor statistical power by contemporary standards. Even for a relatively large candidate gene study with power-favorable multiple comparison correction, the power would have been poor to detect the genetic effects typical for GWAS of human diseases. As the genetic effects for SCZ may be smaller than for other human diseases (Purcell *et al*. 2009; Shi *et al*. 2009; Stefansson *et al*. 2009), nearly all hypothesis-driven candidate gene studies were underpowered. Given what we now know about the genetic architecture of SCZ, a typical candidate gene study requires sample sizes of around 11 000 cases plus controls for a single marker, 17 500 subjects for 10 markers, and 24 000 subjects for 100 markers (see Supplemental Methods online). Future association studies of SCZ should use similarly realistic power calculations.

Moreover, we demonstrated that hypothesis-driven candidate gene studies generally had poor coverage of common genetic variation. With the resources provided by the HapMap and 1000 Genomes projects, coupled with decreases in genotyping costs, researchers can ensure that future genotyping covers the majority of common and rare variation present in their samples.

We were surprised to realize that most genes in the hypothesis-driven candidate gene literature for SCZ had been assessed only once or twice (76.9%). Given the importance of replication in genetic studies of

https://doi.org/10.1017/S0033291711001607 Published online by Cambridge University Press

complex traits (Chanock *et al*. 2007), evaluation of false positives and false negatives is not possible. Positive findings from one or two studies cannot be viewed as secure (particularly given the distorting potential of publication bias) and, conversely, negative findings may not provide confident exclusion.

If power and coverage are low, we can anticipate false negatives (i.e. true susceptibility loci with non-significant candidate gene findings). For example, GWAS and replication efforts support *TCF4* as a risk locus for SCZ (Stefansson *et al*. 2009). However, a *TCF4* CAG repeat was studied for association with SCZ in three studies (Vincent *et al*. 1999; Bowen *et al*. 2000; McInnis *et al*. 2000). All reported negative results, which may have led to the inappropriate exclusion of *TCF4* from consideration.

Furthermore, it is possible that the major hypotheses that drove the selection of many candidate genes are incorrect. SZGene candidate genes were selected for many different reasons and some resulted from genome-wide linkage screens (most notably *NRG1* and *DTNBP1*) (Stefansson *et al*. 2002; Straub *et al*. 2002). However, the ISC GWAS results did not lend support for common variation contributing to SCZ, either for candidate genes from the literature as a whole or for the specific pathways from which candidate genes were frequently selected. For the full set of hypothesis-driven candidate genes, there was nominally significant support for an over-representation of small ISC *p* values. However, the effect was marginal, and the results were not significant when corrected for potential bias caused by LD between genes. We found no support for an aggregate effect of hypothesis-driven candidate genes contributing to SCZ risk using a risk profile generated from the SNPs within these genes. This pattern of results is not consistent with robust or notable collective contribution of common variation within the hypothesis-driven candidate genes to SCZ based on the ISC data. However, it is possible that subsets of the heterogeneous list of historical candidate genes are enriched for smaller ISC *p* values. We thus tested the two overarching 'meta-hypotheses' that have been highly influential: notions of SCZ as a disease of the synapse and as a neurodevelopmental disease. To our knowledge, these two larger-scale ideas have not been tested for empirical support in aggregate. We found no evidence to support a genetic basis for these two hypotheses in perhaps the most comprehensive analysis yet conducted. In addition, we specifically evaluated eight of the 10 most studied historical candidate genes and the ISC GWAS results provide no support for common genetic variation associated with SCZ. We note that the strongest ISC GWAS findings were in the MHC region. Genes from the expanded MHC region do appear in the hypothesis-driven candidate gene literature. Most notably, *NOTCH4* had genome-wide significant SNPs in the ISC data and was highly studied (24 times; both positive and negative studies) in the candidate gene literature. However, given the high LD in the region (Supplementary Table S1), we cannot localize the MHC signal more specifically. We cannot therefore either directly validate or exclude *NOTCH4* as being involved in SCZ susceptibility.

Finally, more generally, there are now numerous guidelines for candidate gene studies (Chanock *et al*. 2007; Pearson & Manolio, 2008). If these guidelines are followed at all stages of the scientific process (from study design through review), the published literature will better reflect the genetic architecture of SCZ.

No single study can disprove a meta-hypothesis in psychiatry, and our conclusions are bounded by the statistical power of the ISC sample. However, it is notable that the ISC GWAS results do not support enrichment of SCZ susceptibility loci within the candidate genes. These results suggest, but do not prove, that many traditional ideas about the genetic basis of SCZ may be incorrect. Indeed, the singular advantage of genomic surveys is that they are unbiased by prior knowledge and can yield novel and unexpected findings. Given current knowledge of the genetic architecture of SCZ and the capacity to assess common and rare variations across the genome, it is possible that the next few years will lead to marked changes in major hypotheses about the genetic basis of SCZ.

Note

Supplementary material accompanies this paper on the Journal's website (http://journals.cambridge.org/psm).

Acknowledgments

Funding for this project was from the US National Institutes of Health, who had no role in the design, execution, analysis, and manuscript preparation. This project was funded by R01 MH077139 to Dr P. F. Sullivan.

Declaration of Interest

None.

References

Allen NC, Bagade S, McQueen MB, Ioannidis JP, Kavvoura FK, Khoury MJ, Tanzi RE, Bertram L (2008). Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database. *Nature Genetics* **40**, 827–834.

Athanasiu L, Mattingsdal M, Kahler AK, Brown A, Gustafsson O, Agartz I, Giegling I, Muglia P, Cichon S, Rietschel M, Pietilainen OP, Peltonen L, Bramon E, Collier D, Clair DS, Sigurdsson E, Petursson H, Rujescu D, Melle I, Steen VM, Djurovic S, Andreassen OA (2010). Gene variants associated with schizophrenia in a Norwegian genome-wide study are replicated in a large European cohort. *Journal of Psychiatric Research* **44**, 748–753.

Beaudet AL, Belmont JW (2008). Array-based DNA diagnostics: let the revolution begin. *Annual Review of Medicine* **59**, 113–129.

Bowen T, Guy CA, Cardno AG, Vincent JB, Kennedy JL, Jones LA, Gray M, Sanders RD, McCarthy G, Murphy KC, Owen MJ, O'Donovan MC (2000). Repeat sizes at CAG/CTG loci CTG18.1, ERDA1 and TGC13-7a in schizophrenia. *Psychiatric Genetics* **10**, 33–37.

Browning BL, Browning SR (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American Journal of Human Genetics* **84**, 210–223.

Browning SR, Browning BL (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics* **81**, 1084–1097.

Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, Hirschhorn JN, Abecasis G, Altshuler D, Bailey-Wilson JE, Brooks LD, Cardon LR, Daly M, Donnelly P, Fraumeni Jr. JF, Freimer NB, Gerhard DS, Gunter C, Guttmacher AE, Guyer MS, Harris EL, Hoh J, Hoover R, Kong CA, Merikangas KR, Morton CC, Palmer LJ, Phimister EG, Rice JP, Roberts J, Rotimi C, Tucker MA, Vogan KJ, Wacholder S, Wijsman EM, Winn DM, Collins FS (2007). Replicating genotype-phenotype associations. *Nature* **447**, 655–660.

Cichon S, Craddock N, Daly M, Faraone SV, Gejman PV, Kelsoe J, Lehner T, Levinson DF, Moran A, Sklar P, Sullivan PF (2009). Genomewide association studies: history, rationale, and prospects for psychiatric disorders. *American Journal of Psychiatry* **166**, 540–556.

Dennis Jr. G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology* **4**, P3.

Gauderman W, Morrison J (2006). QUANTO 1.1: a computer program for power and sample size calculations for genetic-epidemiology studies (http://hydra.usc.edu/gxe).

Gauderman WJ (2002). Sample size requirements for matched case-control studies of gene-environment interaction. *Statistics in Medicine* **21**, 35–50.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R; Gene Ontology Consortium (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research* **32**, D258–D261.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences USA* **106**, 9362–9367.

Holmans P, Green EK, Pahwa JS, Ferreira MA, Purcell SM, Sklar P, Owen MJ, O'Donovan MC, Craddock N (2009). Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *American Journal of Human Genetics* **85**, 13–24.

Huang DW, Sherman BT, Lempicki RA (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4**, 44–57.

International HapMap Consortium (2005). A haplotype map of the human genome. *Nature* **437**, 1299–1320.

International Schizophrenia Consortium (2008). Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455**, 237–241.

Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG (2001). Replication validity of genetic association studies. *Nature Genetics* **29**, 306–309.

Kirov G, Zaharieva I, Georgieva L, Moskvina V, Nikolov I, Cichon S, Hillmer A, Toncheva D, Owen MJ, O'Donovan MC (2009). A genome-wide association study in 574 schizophrenia trios using DNA pooling. *Molecular Psychiatry* **14**, 796–803.

Konneker TI, Crowley JJ, Quackenbush CR, Keefe RS, Perkins DO, Stroup TS, Lieberman JA, van den Oord E, Sullivan PF (2010). No association of the serotonin transporter polymorphisms 5-HTTLPR and rs2553 with schizophrenia or neurocognition. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics* **153B**, 1115–1117.

Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K, Lee C, Nizzari MM, Gabriel SB, Purcell S, Daly MJ, Altshuler D (2008). Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nature Genetics* **40**, 1253–1260.

Lee PH, O'Dushlaine C, Thomas B, Holmans P, Purcell S (2011). InRich: Interval-based enrichment analysis for genome-wide association studies (http://pngu.mgh.harvard.edu/~purcell/inrich/).

Lencz T, Morgan TV, Athanasiou M, Dain B, Reed CR, Kane JM, Kucherlapati R, Malhotra AK (2007). Converging evidence for a pseudoautosomal cytokine receptor gene locus in schizophrenia. *Molecular Psychiatry* **12**, 572–580.

McInnis MG, Swift-Scanlanl T, Mahoney AT, Vincent J, Verheyen G, Lan TH, Oruc L, Riess O, Van Broeckhoven C, Chen H, Kennedy JL, MacKinnon DF, Margolis RL, Simpson SG, McMahon FJ, Gershon E, Nurnberger J, Reich T, DePaulo JR, Ross CA (2000).

Allelic distribution of CTG18.1 in Caucasian populations: association studies in bipolar disorder, schizophrenia, and ataxia. *Molecular Psychiatry* **5**, 439–442.

Need AC, Ge D, Weale ME, Maia J, Feng S, Heinzen EL, Shianna KV, Yoon W, Kasperaviciute D, Gennarelli M, Strittmatter WJ, Bonvicini C, Rossi G, Jayathilake K, Cola PA, McEvoy JP, Keefe RS, Fisher EM, St Jean PL, Giegling I, Hartmann AM, Moller HJ, Ruppert A, Fraser G, Crombie C, Middleton LT, St Clair D, Roses AD, Muglia P, Francks C, Rujescu D, Meltzer HY, Goldstein DB (2009). A genome-wide investigation of SNPs and CNVs in schizophrenia. *PLoS Genetics* **5**, e1000373.

O'Donovan MC, Craddock N, Norton N, Williams H, Peirce T, Moskvina V, Nikolov I, Hamshere M, Carroll L, Georgieva L, Dwyer S, Holmans P, Marchini JL, Spencer CC, Howie B, Leung HT, Hartmann AM, Moller HJ, Morris DW, Shi Y, Feng G, Hoffmann P, Propping P, Vasilescu C, Maier W, Rietschel M, Zammit S, Schumacher J, Quinn EM, Schulze TG, Williams NM, Giegling I, Iwata N, Ikeda M, Darvasi A, Shifman S, He L, Duan J, Sanders AR, Levinson DF, Gejman PV, Cichon S, Nothen MM, Gill M, Corvin A, Rujescu D, Kirov G, Owen MJ, Buccola NG, Mowry BJ, Freedman R, Amin F, Black DW, Silverman JM, Byerley WF, Cloninger CR (2008). Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nature Genetics* **40**, 1053–1055.

Pearson TA, Manolio TA (2008). How to interpret a genome-wide association study. *Journal of the American Medical Association* **299**, 1335–1344.

Pruitt KD, Tatusova T, Maglott DR (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* **33**, D501–D504.

Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752.

SAS Institute Inc (2004). *SAS/STAT® Software: Version 9*. SAS Institute, Inc.: Cary, NC.

SAS Institute Inc (2005). *JMP User's Guide* (*Version 6*). SAS Institute, Inc.: Cary, NC.

Sherman BT, Huang DW, Tan Q, Guo Y, Bour S, Liu D, Stephens R, Baseler MW, Lane HC, Lempicki RA (2007). DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. *BMC Bioinformatics* **8**, 426.

Shi J, Levinson DF, Duan J, Sanders AR, Zheng Y, Pe'er I, Dudbridge F, Holmans PA, Whittemore AS, Mowry BJ, Olincy A, Amin F, Cloninger CR, Silverman JM, Buccola NG, Byerley WF, Black DW, Crowe RR, Oksenberg JR, Mirel DB, Kendler KS, Freedman R, Gejman PV (2009). Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature* **460**, 753–757.

Shifman S, Johannesson M, Bronstein M, Chen SX, Collier DA, Craddock NJ, Kendler KS, Li T, O'Donovan M, O'Neill FA, Owen MJ, Walsh D, Weinberger DR, Sun C, Flint J, Darvasi A (2008). Genome-wide association identifies a common variant in the reelin gene that increases the risk of schizophrenia only in women. *PLoS Genetics* **4**, e28.

Stefansson H, Ophoff RA, Steinberg S, Andreassen OA, Cichon S, Rujescu D, Werge T, Pietilainen OP, Mors O, Mortensen PB, Sigurdsson E, Gustafsson O, Nyegaard M, Tuulio-Henriksson A, Ingason A, Hansen T, Suvisaari J, Lonnqvist J, Paunio T, Borglum AD, Hartmann A, Fink-Jensen A, Nordentoft M, Hougaard D, Norgaard-Pedersen B, Bottcher Y, Olesen J, Breuer R, Moller HJ, Giegling I, Rasmussen HB, Timm S, Mattheisen M, Bitter I, Rethelyi JM, Magnusdottir BB, Sigmundsson T, Olason P, Masson G, Gulcher JR, Haraldsson M, Fossdal R, Thorgeirsson TE, Thorsteinsdottir U, Ruggeri M, Tosato S, Franke B, Strengman E, Kiemeney LA, Melle I, Djurovic S, Abramova L, Kaleda V, Sanjuan J, de Frutos R, Bramon E, Vassos E, Fraser G, Ettinger U, Picchioni M, Walker N, Toulopoulou T, Need AC, Ge D, Yoon JL, Shianna KV, Freimer NB, Cantor RM, Murray R, Kong A, Golimbet V, Carracedo A, Arango C, Costas J, Jonsson EG, Terenius L, Agartz I, Petursson H, Nothen MM, Rietschel M, Matthews PM, Muglia P, Peltonen L, St Clair D, Goldstein DB, Stefansson K, Collier DA (2009). Common variants conferring risk of schizophrenia. *Nature* **460**, 744–747.

Stefansson H, Sigurdsson E, Steinthorsdottir V, Bjornsdottir S, Sigmundsson T, Ghosh S, Brynjolfsson J, Gunnarsdottir S, Ivarsson O, Chou TT, Hjaltason O, Birgisdottir B, Jonsson H, Gudnadottir VG, Gudmundsdottir E, Bjornsson A, Ingvarsson B, Ingason A, Sigfusson S, Hardardottir H, Harvey RP, Lai D, Zhou M, Brunner D, Mutel V, Gonzalo A, Lemke G, Sainz J, Johannesson G, Andresson T, Gudbjartsson D, Manolescu A, Frigge ML, Gurney ME, Kong A, Gulcher JR, Petursson H, Stefansson K (2002). Neuregulin 1 and susceptibility to schizophrenia. *American Journal of Human Genetics* **71**, 877–892.

Stenzel A, Lu T, Koch WA, Hampe J, Guenther SM, De La Vega FM, Krawczak M, Schreiber S (2004). Patterns of linkage disequilibrium in the MHC region on human chromosome 6p. *Human Genetics* **114**, 377–385.

Straub RE, Jiang Y, MacLean CJ, Ma Y, Webb BT, Myakishev MV, Harris-Kerr C, Wormley B, Sadek H, Kadambi B, Cesare AJ, Gibberman A, Wang X, O'Neill FA, Walsh D, Kendler KS (2002). Genetic variation in the 6p22.3 gene DTNBP1, the human ortholog of the mouse dysbindin gene, is associated with schizophrenia. *American Journal of Human Genetics* **71**, 337–348.

Sullivan PF, Lin D, Tzeng JY, van den Oord E, Perkins D, Stroup TS, Wagner M, Lee S, Wright FA, Zou F, Liu W, Downing AM, Lieberman J, Close SL (2008). Genomewide association for schizophrenia in the CATIE study: results of stage 1. *Molecular Psychiatry* **13**, 570–584.

Vincent JB, Petronis A, Strong E, Parikh SV, Meltzer HY, Lieberman J, Kennedy JL (1999). Analysis of genome-wide CAG/CTG repeats, and at SEF2-1B and ERDA1 in schizophrenia and bipolar affective disorder. *Molecular Psychiatry* **4**, 229–234.