

IMMEDIATE COMMUNICATION

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

V Moskvina^{1,2}, N Craddock¹, P Holmans^{1,2}, I Nikolov^{1,2}, JS Pahwa^{1,2}, E Green¹, Wellcome Trust Case Control Consortium³, MJ Owen¹ and MC O'Donovan¹

¹Department of Psychological Medicine, School of Medicine, Cardiff University, Cardiff, UK and ²Bioinformatics and Statistics Unit, School of Medicine, Cardiff University, Cardiff, UK

Genome-wide association (GWAS) analyses have identified susceptibility loci for many diseases, but most risk for any complex disorder remains unattributed. There is therefore scope for complementary approaches to these data sets. Gene-wide approaches potentially offer additional insights. They might identify association to genes through multiple signals. Also, by providing support for genes rather than single nucleotide polymorphisms (SNPs), they offer an additional opportunity to compare the results across data sets. We have undertaken gene-wide analysis of two GWAS data sets: schizophrenia and bipolar disorder. We performed two forms of analysis, one based on the smallest *P*-value per gene, the other on a truncated product of *P* method. For each data set and at a range of statistical thresholds, we observed significantly more SNPs within genes (P_{\min} for excess < 0.001) showing evidence for association than expected whereas this was not true for extragenic SNPs (P_{\min} for excess > 0.1). At a range of thresholds of significance, we also observed substantially more associated genes than expected (P_{\min} for excess in schizophrenia = 1.8×10^{-8} , in bipolar = 2.4×10^{-6}). Moreover, an excess of genes showed evidence for association across disorders. Among those genes surpassing thresholds highly enriched for true association, we observed evidence for association to genes reported in other GWAS data sets (*CACNA1C*) or to closely related family members of those genes including *CSF2RB*, *CACNA1B* and *DGKI*. Our analyses show that association signals are enriched in and around genes, large numbers of genes contribute to both disorders and gene-wide analyses offer useful complementary approaches to more standard methods.

Molecular Psychiatry (2009) 14, 252–260; doi:10.1038/mp.2008.133; published online 9 December 2008

Keywords: genetics; association; bipolar; schizophrenia; psychosis

Introduction

Until recently, there have been few undisputed genetic associations to non-Mendelian forms of common human diseases, but for many diseases, the advent of genome-wide association (GWAS) technology has recently transformed this position.¹ The attainment of highly significant associations though GWAS reflects in some cases, the availability of large sample sizes,² for others, for example the HLA locus in rheumatoid arthritis and T1D, the existence of at least some common alleles with a greater than average

effect size.³ These conditions may not readily be satisfied for most complex disorders, for example psychotic disorders, where the extremely large sample sizes used for some disorders² are difficult to obtain because diagnosis is labourious and expensive. Moreover, the necessary use of phenotypes entirely defined by symptoms will very likely increase aetiological heterogeneity, and thus the observed correlations between genotypes and phenotypes. Therefore far from having a few common risk genes with higher than expected effect sizes, the observed effect sizes in psychosis might be even smaller than those typical for other complex disorders. Moreover, even for those disorders where successes have been legion, the majority of genetic risk remains unattributed.^{2,4} There is therefore a pressing need for alternative methods for extracting information from GWAS data sets.

GWAS studies to date have focused on single locus tests, which are the simplest to generate and to

Correspondence: Professor MC O'Donovan or V Moskvina, Department of Psychological Medicine, School of Medicine, Cardiff University, Heath Park, Cardiff CF23 6BQ, UK.

E-mails: odonovanmc@cf.ac.uk and moskvinav1@cardiff.ac.uk

³List of members of the Wellcome Trust Case Control Consortium (WTCCC) is given in Supplementary Material online.

Received 23 September 2008; revised 12 November 2008; accepted 12 November 2008; published online 9 December 2008

interpret. There are, however, situations where they might not provide most power. Examples include those where there are multiple common variants at a locus,⁵ and, for replication studies or meta-analysis, where there are differences in the association signals between populations.⁶ Either of these can give rise to complex patterns of association that might not be reflected by association to the same single nucleotide polymorphisms (SNPs) in different samples despite apparently reasonably powered samples.^{7,8} Another consideration is that power to detect association might be enhanced by exploiting information from multiple (quasi) independent signals within genes. Although it is unknown to what extent any of the above concerns occur in practice, arguments such as these have led some to advocate the use of gene-wide statistical approaches.⁵

A final crucial point for GWAS studies is that the likely architecture of genetic risk for the psychoses is a matter of considerable debate. Based on epidemiology, in most cases, risk likely reflects the coaction of several loci but the approximate numbers of loci involved at the individual or the population levels are unknown, as is the spectrum of allele frequencies and effect sizes.^{9,10} At present there is limited direct molecular genetic evidence that favours the existence of common risk alleles. The observations of multiple genome-wide significant or suggestive linkage signals for both disorders that do not readily replicate between studies, but which are not randomly distributed across the genome^{11,12} is compatible with the existence of multiple risk alleles of small-moderate effect. They are not, however, informative with respect to their allele frequencies. Recent papers describing an enrichment of copy number variants (CNV) in schizophrenia^{13–15} and an excess of *de novo* CNV events¹⁶ in that disorder have raised the possibility of a significant contribution from rare events, some of apparently high penetrance. Although it is not yet clear whether the contribution from CNVs is small or substantial, these findings can be interpreted as supporting the hypothesis that common variation may be less important than has generally been assumed.

The current findings from the few published GWAS studies of schizophrenia and bipolar disorder, along with the data from leading candidate genes that predate the era, are supportive of the hypothesis that common variants contribute. However, no locus has yet been reported for schizophrenia that in any single or combined study reaches genome-wide levels of significance.¹⁷ For bipolar disorder, there are three such loci^{18,19} but these have yet to receive support in independent studies. The issue of whether there is wide-scale involvement of common variants is not moot; if the vast majority of genetic risk is conferred almost exclusively by rare alleles, approaches based on indirect genetic association may not be very informative. Given the challenges in obtaining appropriate-sized samples to conduct GWAS studies, it would be preferable to have strong evidence as to

whether doing so is likely to be rewarded with a significant degree of success.

In this study, we have investigated the potential advantages of assessing gene-wide significance in two GWAS data sets, one of schizophrenia²⁰ the other of bipolar disorder.³ Specifically, we aimed to determine if gene-wide analyses resulted in evidence for a marked excess of genes surpassing various thresholds of evidence for association, a finding that would be compatible with a common disease–common variant hypothesis and supportive of more intensive GWAS endeavours. Our secondary aim was to identify at least suggestive evidence for multiple susceptibility genes for each disorder that might support earlier findings, and inform follow-up genetic studies. Also, given the hypothesis of overlap in genetic risk between the two disorders²¹ we wished to determine if there is overlap in the identity of associated genes.

Materials and methods

GWAS data sets

The Bipolar data set was reported by the Wellcome Trust Case Control Consortium (WTCCC)³ and consists of 1868 cases and 2938 controls typed with the GeneChip 500K Mapping Array Set. The UK schizophrenia cases ($n = 479$) were not part of that study but were typed contemporaneously with the WTCCC samples using the same pipeline.²⁰ The full details of the samples and methods for conduct of the GWAS studies are provided in the respective manuscripts. To make our analysis as conservative as possible, we only included autosomal SNPs which passed more stringent quality control criteria than used by the WTCCC, and, unlike that study, additionally corrected all P -values for inflation in the test statistics (see statistical section). Thus we excluded SNPs with Hardy–Weinberg equilibrium $P < 0.001$ in controls or $\hat{P} < 0.00001$ in cases, with minor allele frequencies < 0.01 in each of cases and controls, or with call rates < 0.97 .

SNP assignment

SNPs were assigned to genes if they were located within the genomic sequence corresponding to the start of the first and the end of the last exons of any transcript corresponding to that gene. Functional elements are not restricted to this region but we used this because any other definition is arbitrary. The chromosome and location for all currently known human SNPs and genes and their identifiers was taken from the human genome assembly build 36.2 of the National Center for Biotechnology Information (NCBI) database. All known SNPs and their corresponding chromosomal locations were obtained from the Chromosome Reports data for Taxonomic ID 9606 (that is, humans) available from NCBI's dbSNP. These data were downloaded for chromosomes 1–22 providing information on RefSNPs and chromosome coordinates. The second data source (seq_gene.md) was also downloaded from the NCBI's Genome database giving

information on Gene ID, gene names, and their start and end positions on a chromosome. For the purpose of identification of SNPs in genes we mapped all the SNPs to genes defined by the start and end positions using database techniques. The resulting output file provided information on SNPs for chromosomes 1–22 and the genes in which they are placed. From the chromosome reports data, only reference sequence entries were used. The entries for ‘Celera’ sequence were ignored. In the `seq_gene.md` file also, only reference sequence entries for genes with Taxonomic ID of 9606 were used. The entries for ‘Celera’ sequence and entries of gene types such as ‘PSEUDO’, ‘CDS’, ‘RNA’ and ‘UTR’ were also ignored from this file.

Where a SNP lay within the boundaries of more than one gene, that SNP was arbitrarily assigned to the gene, which included the lowest base number on the chromosome assembly. Arbitrary decisions of this nature and errors in SNP/gene assignment present in the databases have no biological validity, and therefore it is anticipated they will generally generate noise and reduce the power. In total, we retained 145 344 (38.5% of the total) SNPs, which annotated 13 098 unique genes (1–799 SNPs per gene).

Statistics

We used the Armitage trend test (1df) to generate SNP association P -values. In the bipolar and schizophrenia data sets, the association test statistics for the markers we used here are respectively inflated as estimated by the genomic control²² metric λ by 1.11 and 1.08. In neither data set does this appear attributable to population structure because analyses conditional on the principal components derived from multi-dimensional scaling had negligible impact.^{3,20} Nevertheless, to address our aims conservatively, all SNP P -values in the true data sets are corrected for λ . The corrected P -values were then in turn used to generate three types of gene-wide tests. The first was based on the smallest SNP P -value per gene, and the second and third were, respectively, the threshold-truncated²³ product of all SNP P -values within a gene where the truncation thresholds were $P \leq 0.01$ and $P \leq 0.001$. Product analyses were restricted to those genes which had more than 1 SNP.

To calculate empirical gene-wide significance for each gene, we performed 1000 genome-wide permutations for each GWAS data set. For each gene in each permutation we obtained the smallest P and the product of P -values as for the original data set. We then calculated the three empirical P -values for each gene in the observed data by determining the proportion of permuted data sets where the corresponding P -value obtained for each gene was equal to or smaller than was observed in the true data set.

Obvious disadvantages to a gene-wide approach are that we do not know the boundaries of genes, that the presence of linkage disequilibrium (LD) means that association to a physical location may not point to the particular coterminous functional element, and many

important signals will be contained within functional elements that do not correspond to genes. Nevertheless, it seems a reasonable first assumption that SNPs assigned to known functional elements (genes) would have a higher probability of being associated with disease than the remaining SNPs, even though some of the latter will span unknown functional elements. To test this, we counted the total number of SNPs designated by us as within genes surpassing nominal thresholds of $P \leq 0.05$, $P \leq 0.01$ and $P \leq 0.001$ in the observed data set and also the total number of those not designated by us as within genes. We then compared the observed numbers with the null distributions for each as determined from the permutation data sets.

We calculated the significance of the excess number of genes attaining the specified thresholds in two main ways. The first was empirically, the second was based on the assumption that, under the null hypothesis of no association, the number of significant genes in a scan is a normally distributed random variable whose mean and standard deviations can be obtained from the permutations. Given computational restrictions required by genome-wide permutation, we could only perform 1000 permutations and therefore based on the first method, significance can be reported only up to a threshold $P > 10^{-3}$. These were calculated by bootstrapping, as follows. From 1000 permutations, we choose a replicate at random as a ‘true study’. We then calculated the significance for each gene, and therefore the total number of significant genes, by comparing the ‘true study’ with 999 replicates obtained by sampling at random (with replacement) from the remaining 999 permutations. This process was repeated 1000 times. The empirical test of the number of significant genes being higher than expected under the null hypothesis of no association was carried out by comparing the observed number of significant genes to the empirical distribution. The calculated level of significance was based on the distributions of the number of significant genes under the null hypothesis, which formally passed the test for normality (in all instances both the skewness and kurtosis coefficients were between -0.5 and 0.5).

The bipolar and schizophrenia data sets share the same controls, and therefore the test statistics are correlated. This means that we cannot calculate whether genes are associated to both disorders at a rate greater than chance simply based on assumptions of independence, for example the hypergeometric distribution. Instead we used permutations. We pooled together the three groups of individuals (bipolar and schizophrenia cases, and controls) and randomly assigned diagnostic category keeping the numbers of individuals in each group equal to that in the observed data. We assessed genes by comparing one randomly selected permutation (of 1000) with the 999 permutations randomly drawn (with replacement) from the pool of remaining permutations. From this, we generated lists of the top associated genes

that were equal in length to the corresponding list of genes for the observed data sets and then recorded the number of overlapping genes. This process was repeated 1000 times. We then calculated empirical P -values by counting how many times the simulated number of overlapping genes was greater than, or equal to, the observed number of overlapping genes in the true data sets.

Results

In Table 1, we present the numbers of SNPs in the observed data surpassing various nominal thresholds for α within and external to genes, the mean and the standard deviations of the same from the permuted data, and the empirical P -values (two tailed) for the null distribution in the observed data set. For schizophrenia, there is a highly significant excess of SNPs within genes at the two more stringent thresholds whereas this was the case for bipolar disorder only at the intermediate threshold ($P < 0.01$). For neither disorder did we observe a significant excess of associated SNPs located beyond the boundaries of genes. These results suggest that focussing on genes enriches for detectable association signals (in samples of the size we have utilized) providing support for a gene-centric approach.

The results of the gene-wide analyses for schizophrenia and bipolar disorder are shown in Tables 2 and 3. We provide the observed numbers of genes surpassing various thresholds of α in the observed data. The expectations under the Null hypothesis (based on permutation) are described by the mean and the standard deviation. For the product of P , the thresholds at which the data were truncated are given and the resultant product P -value significance threshold in the α column. In the final two columns, we present P -values (empirical and calculated) for the observed data under the one-tailed null hypothesis of no excess of genes surpassing the thresholds. We use one-tailed tests because the

observation of fewer significant genes than chance has no biological meaning. In each disorder, there is a significant, or highly significant, excess in the observed number of genes surpassing virtually all thresholds regardless of the method for calculating gene-wide significance. The data from these tables are depicted graphically in Supplementary Figure S1.

In Table 4, we present the results of our analysis of genes that simultaneously exceed various thresholds of α in the schizophrenia and bipolar data sets. The first two data rows labelled schizophrenia (SZ) and bipolar (BD) give the numbers of genes observed in the respective GWAS data sets at thresholds of α depicted at the top of each row (extracted from the previous two tables). The third row gives the number of genes common to both data sets at these thresholds. The bottom section of the table summarizes the null expectations based on the permutation tests. The distributions of the number of significant genes common to schizophrenia and bipolar disorder in the permuted data are not normal and therefore to characterize the distribution, we present the median, minimum and maximum of the numbers of overlapping genes in the permuted data. We also present empirical P -values for observing the true data set under the null. P -values are two-tailed because observing fewer genes across disorders than expected by chance would have a biological meaning as risk for one disorder could in principle reduce risk for the other.

At the more stringent thresholds ($P < 0.001$), none of the tests were significant although there were a total of three observations in common to both disorders compared with none in the permuted data. At other thresholds, there was a significant excess or a trend to excess of genes in common, with strongest support coming from the smallest P method.

For each of the schizophrenia and bipolar data sets, in Supplementary Tables S1 and S2, we present information about the identity of genes surpassing any of the thresholds corresponding to (best or

Table 1 Comparison of SNPs by genic location in schizophrenia (SZ) and bipolar (BD) datasets

α	Number of SNPs in genes				Number of SNPs outside genes			
	Observed	Permuted		P	Observed	Permuted		P
		Mean	s.d.			Mean	s.d.	
<i>SZ</i>								
0.05	7480	7243.4	180.8	0.098	11534	11594.0	244.4	0.604
0.01	1688	1447.7	75.5	< 0.001	2381	2313.9	101.5	0.256
0.001	222	147.5	22.9	0.002	266	234.7	29.2	0.143
<i>BD</i>								
0.05	7647	7261.18	190.73	0.19	11863	11602.26	246.86	0.135
0.01	1722	1451.03	78.47	< 0.001	2354	2314.51	104.39	0.352
0.001	188	144.01	22.94	0.055	227	231.60	29.19	0.548

Table 2 Overrepresentation of significant genes in schizophrenia

Method of gene assessment	Number of genes surpassing significance level (α)			P-value for observed (one-tailed)		
	α	Permuted		Observed	Empirical	Calculated
		Mean	s.d.			
Smallest P-value per gene	0.05	652.4	31.7	703	0.059	0.056
	0.01	135.1	13.3	175	0.002	0.001
	0.001	20.2	4.8	41	<0.001	1.4×10^{-6}
Product of P truncation ≤ 0.01	0.05	362.4	20.6	425	<0.001	0.0012
	0.01	106.1	11.1	140	0.002	0.0011
	0.001	15.6	4.2	39	<0.001	1.8×10^{-8}
Product of P truncation ≤ 0.001	0.05	77.9	9.2	98	0.023	0.014
	0.01	50.4	7.3	75	0.001	0.0004
	0.001	13.9	4.1	30	0.001	3.9×10^{-5}

Table 3 Overrepresentation of significant genes in bipolar disorder

Method of gene assessment	Number of genes surpassing significance level (α)			P-value for observed (one-tailed)		
	α	Permuted		Observed	Empirical	Calculated
		Mean	s.d.			
Smallest P-value per gene	0.05	660.2	30.8	790	<0.001	1.3×10^{-5}
	0.01	136.8	13.2	197	<0.001	2.4×10^{-6}
	0.001	20.5	5.1	41	<0.001	2.9×10^{-5}
Product of P truncation ≤ 0.01	0.05	365.2	21.2	445	<0.001	8.6×10^{-5}
	0.01	106.8	11.8	159	<0.001	4.5×10^{-6}
	0.001	16.1	4.2	30	0.005	0.0005
Product of P truncation ≤ 0.001	0.05	76.8	9.7	94	0.039	0.038
	0.01	50.0	8.0	63	0.064	0.053
	0.001	14.1	4.3	24	0.018	0.010

Table 4 Number of significant genes observed to overlap in schizophrenia (SZ) and bipolar (BD) datasets

	Smallest P-value per gene			Product of P truncation ≤ 0.01			Product of P truncation ≤ 0.001		
	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
	SZ	703	175	41	425	140	39	98	75
BD	790	197	41	445	159	30	94	63	24
Overlap	85	14	1	37	11	2	5	2	0
Number of overlapping genes (1000 permutations with shared controls)									
Median	60	5	0	33	4	0	2	1	0
Min	27	0	0	12	0	0	0	0	0
Max	93	16	5	60	18	5	8	7	3
Empirical P-value ^a	0.014	0.007	0.402	0.305	0.017	0.09	0.063	0.415	NA

^aComparing overlap on observed and simulated data.

product) $P < 0.001$. For this purpose, we took genes which met any of the thresholds corresponding to $P \leq 0.01$ in Tables 2 and 3 and undertook 100 000 permutations. This is because specific genes assigned $P > 0.01$ should be fairly accurately specified by 1000 permutations whereas those where the gene-wide $P < 0.01$ may not be. We chose to present the identity of genes surpassing the thresholds $P < 0.001$ as at this threshold, the ratio of the number of observed genes to that expected under the null is substantially greater than 1, suggesting that any one gene is more likely to represent a true than a false observation. It is nevertheless important to note that the evidence for any one gene is not compelling, and the specific findings require confirmation in additional samples. Also, the close proximity of a number of the genes suggests that there are instances where more than one gene derives from the same association signal (note, this phenomenon will inflate the estimated numbers of signals in both observed and permuted data sets and therefore does not impact on the main findings). Supplementary Table S3 shows those genes which in each of schizophrenia and bipolar disorder surpassed thresholds of smallest $P < 0.01$ or where for either truncation threshold, the product of P was ≤ 0.01 as these thresholds also correspond to a considerable degree of enrichment in the observed compared with the simulated data.

Discussion

GWAS analyses have identified susceptibility loci for many complex diseases, but the majority of risk for any disorder remains unattributed. Obtaining sufficient samples to extract a high proportion of that component of disease risk attributable to common variants is unlikely to be realized in the near future, and this is particularly true for psychiatric disorders where sample collection is particularly challenging. There is therefore scope for complementary approaches to GWAS data sets.

In this study, we sought to apply one such approach based on gene-wide analysis, which offers a number of possible advantages over single locus tests.⁵ First, if there is more than independent source of an association signal within a gene, for example where there is more than one functional variant, combining these into a single statistic might offer enhanced power over single SNP analysis. This is the main rationale underpinning our use of the truncated product approach.²³ Second, where there are true differences in the associated SNPs between studies, as might occur as a result of allelic heterogeneity, LD heterogeneity, or where this occurs simply as a result of the sometimes unpredictable nature of LD,^{7,8} consistency of associations across studies may be easier to achieve at the gene-wide than at the SNP level. Although such association may not provide a firm basis for implicating specific susceptibility variants, the identification of replicated associations can be of value in identifying candidate genes for further intensive genetic

investigation and also for generating hypotheses concerning aetiology and pathophysiology. Both of the methods we have applied here, best corrected P -value and truncated product, confer this potential advantage. They are also both applicable not just to genes but also to other definitions of functional units, for example groups of genes with related functions.

One caveat is that a gene-wide approach requires at least some of the true association signals to be located within the (arbitrarily) derived boundaries of genes. Our analysis showing an excess in observed association signals within genes but not in sequences beyond gene boundaries supports the view that such a gene-centric enrichment does occur, thereby providing an rationale for targeting the immediate vicinity of genes for analysis. This may be intuitively obvious by analogy with simpler genetic disorders, but with respect to complex diseases, there has been little empirical support for such a strategy. Analogous analyses in multiple phenotypes with genes defined incrementally by adding additional flanking sequence might better define optimal (on average) locations for such endeavours.

Concerning the primary hypothesis, for each disorder, we obtained significant or highly significant evidence for an excess of associated genes at most thresholds. This supports the general validity of the gene-wide approach for detecting true association signals. However, it does not suggest that gene-wide tests make single locus tests redundant; rather that they are useful additional approaches. By way of illustration, we plot the rank order achieved by genes in the schizophrenia data set (the figures for bipolar are similar) based on uncorrected single SNP tests against that achieved with each of the gene-wide tests (Figure 1). As expected, the rank order of genes based on the most significant uncorrected SNP correlates (Spearman $r = 0.8$) with that obtained after correction for the number of independent SNPs in that gene (Figure 1a), but the correlation is much less ($r = 0.37$) between the rankings achieved by the single SNP test and that obtained by the truncated-product approach (Figure 1b). Thus, the approaches are complementary in highlighting the likely involvement of specific genes although the relative merits of each in doing so necessarily awaits the reporting of many more confirmed associations.

That we found significant excess of associated genes at most thresholds provides molecular genetic evidence that has been lacking for the existence of substantial numbers of detectable common alleles of small effect in each of the disorders we tested. Given an assumption of relatively low power to detect genes (which is difficult to estimate given the variation in numbers of SNPs/gene and the number of independent signals/gene, each of different effect size and allele frequency) the excess of associated genes we have observed is likely a small proportion of the total. These data therefore support an important polygenic contribution to liability to both schizophrenia and

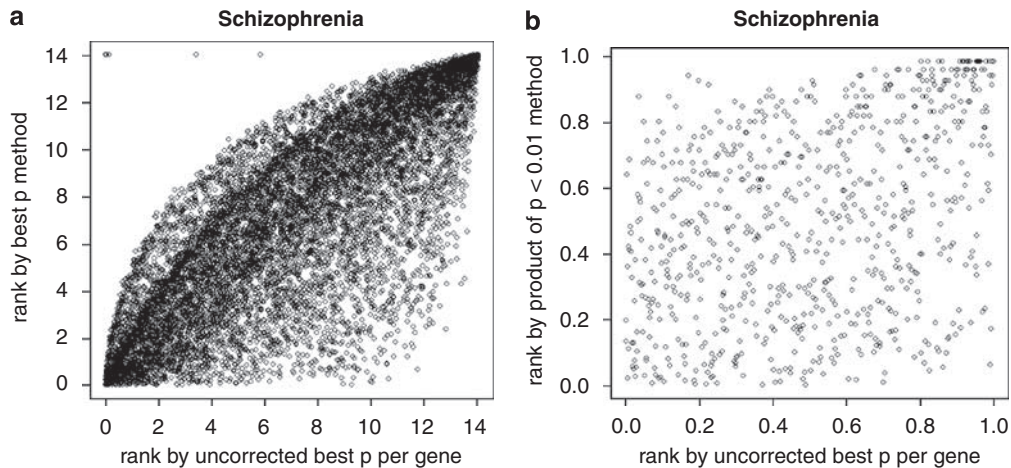


Figure 1 Ranks achieved by genes based on uncorrected single single nucleotide polymorphism (SNP) and gene-wide tests. Ranks are compared based on uncorrected single SNP and gene-wide tests for (a) best P -value per gene and (b) product of P -values truncated at ≤ 0.01 . A higher value corresponds to a more significant rank. Ranks were normalized to lie between 0 and 1. Panel (a) contains all analysed genes. Panel (b) contains only those 733 genes which have more than one SNP and at least one P -value ≤ 0.01 (the threshold inclusion in the analysis).

bipolar disorder. Such a contribution to schizophrenia has previously been suggested by attempts to model the expected prevalence of the disorder among relatives of probands based on various modes of transmission.^{24,25} We also note our data are incompatible with the hypotheses recently advanced elsewhere that there is only a single schizophrenia risk gene or indeed that all risk is epigenetic.²⁶ The existence of multiple genes of small effect suggests that the collection and application of larger samples should deliver many more associations, with the caveat that casting widely for such samples is done in such a way as to avoid the negating effect on power of increasing the (as yet unknown) degree of genetic heterogeneity within diagnostic categories.²⁷

Our secondary objectives were to use the data to identify a number of putative susceptibility genes and also to determine if there is overlap between schizophrenia and bipolar susceptibility. Taking the latter first, although the findings are less robust than the main results, at several thresholds, we observed an excess of associated genes common to schizophrenia and bipolar disorder. These findings support the specific hypothesis that some genes influence risk beyond traditional diagnostic boundaries. That only a small proportion of genes overlapped should not be taken as indicative that this provides an estimate of the extent of shared risk; the small effect sizes inevitably lead to low power for one sample to 'replicate' findings observed in the other. Similar considerations may explain why the excess in overlap was observed at the weaker thresholds of significance; any one gene detected in one study is unlikely to be replicated at similar levels of significance in another.^{2,6} Much larger studies will be required to explore and characterize the extent of the overlap.

With respect to the identity of specific genes, there are important caveats that have to be borne in mind. First the strong statistical evidence in this study is for an excess representation of genes rather than for any individual gene *per se*. Second, association formally implicates regions, not genes, and in some cases, a single 'true' signal may, by LD, result in multiple genes being implicated. However, given the ratio of observed associated genes surpassing the thresholds to those expected, many of those doing so and reported in Supplementary Tables S1 and S2 are likely to represent true associations. A third caveat concerns the joint liability between schizophrenia and bipolar disorder. Our simulation procedures explicitly allow for the use of the same set of controls in the schizophrenia and bipolar GWAS analyses, and therefore our conclusion regarding an excess number of genes showing evidence for association across the two data sets is valid. However, the impact of using shared controls on the association evidence for any specific gene cannot be assessed by those procedures. Additional caution is therefore required with respect to the identities of the specific genes that appear to operate across both disorders (Supplementary Table S3).

It is worth pointing out that the failure of a gene to appear in any of the association lists should also be viewed with caution. As is the case for single SNP analysis, power and gene coverage mean that in a moderate sized GWAS study, however it is analysed, failure to find evidence for association is not strong evidence for absence of involvement of that gene.²⁷ In that context, we simply note that the present analyses provides no support for dysbindin (*DTNBP1*), neuregulin1 (*NRG1*), D-amino-acid oxidase activator (*DAOA*), disrupted in schizophrenia 1 (*DISC1*) or brain-derived neurotrophic factor (*BDNF*), genes which before the advent of GWAS studies were

among the most prominent candidates for either disorder.³

We limit additional comment to genes of particular interest based on existing data. In the gene-wide analysis of schizophrenia, as in the single SNP analysis of the same data set,²⁰ we identified *NOS1*, *RPGRIP1L*, *OPCML*, *TMEM108* and *SIL1* as genes of potential interest. In bipolar disorder, we identified *DPP10*, *RNPEPL1*, *CMTM8*, *DFNB31*, *LAMP3*, *TDRD9*, *PALB2*, *CDC25B* *CAPN6* all of which had at least one polymorphism within the reference sequence that was associated at $P < 10^{-5}$.³ Other than to note that the use of gene-wide tests flagged up many of the putative hits in the original studies, we do not discuss those genes any further because they have been considered in the original manuscripts.

From the perspective of schizophrenia, there are no published genes that meet criteria for genome-wide significant association. Indeed other than the data previously reported from the SNP-based analysis of the present data set,²⁰ there is only one finding from the GWAS literature that meets the criterion for strong evidence for association ($P < 5 \times 10^{-7}$) used by the WTCCC.³ Although that criterion was suggested on the basis of large samples, it is nevertheless interesting that *CSF2RA* encoding colony stimulating factor-2 receptor, α -subunit was reported as a potential susceptibility gene ($P = 3 \times 10^{-7}$) based on the first (small) GWAS study of schizophrenia.²⁸ Among the genes we observed of interest in bipolar disorder was *CSF2RB* (Supplementary Table S2; gene-wide $P_{\min} = 3.5 \times 10^{-4}$). *CSF2RB* is one of only two *CSF2R* genes and encodes the β -subunit with which the other, *CSF2RA*, forms a functional heterodimer. *CSF2RB* itself has been associated with schizophrenia in a case-control and a family-based association sample²⁹ and in addition to its role in forming the CSF2 receptor, it is also a subunit for the interleukin receptors 3 and 5. The present finding therefore clearly warrants attention in additional samples and suggests the possibility that the neuroimmunological hypothesis that has been advanced for schizophrenia might be also relevant to bipolar disorder.³⁰

In addition to the data previously reported from the SNP-based analysis of the present bipolar data set, three genes have been reported that meet criteria for genome-wide significance in bipolar disorder; *ANKK3* (*ankyrin 3*, *node of Ranvier (ankyrin G)*), *CACNA1C* (*calcium channel, voltage-dependent, L type, α -1C subunit*)¹⁹ and *DGKH* (*diacylglycerol kinase, eta*).¹⁸ In this study, we found support for one of these as well as in family members of two of them. *CACNA1C* was not identified as a gene of particular interest when the WTCCC bipolar SNP data were examined³ although those data did subsequently contribute to the meta-analysis.¹⁹ In this study, at the level of the gene, the best corrected SNP P -value in the bipolar analysis was unimpressive ($P = 0.037$) whereas the product method successfully identified this gene as a potential candidate (Supplementary Table S2; gene-wide $P_{\min} = 7 \times 10^{-4}$). Based on a meta-analysis P -value of

7×10^{-8} *CACNA1C*¹⁹ can now be considered very likely to be a true positive. Interestingly, mutations in this gene are already known to cause Timothy syndrome, a disorder whose features include autistic traits (see Ref. 19). Thus, this gene, as well as the other genes whose products encode α 1 subunits (which form the ion pore) of voltage-gated calcium channels, are candidate genes for neuropsychiatric disorders. It is of considerable interest that *CACNA1B*, encoding the subunit typical of the calcium channels which control neurotransmitter release from neurons and one of only 8 *CACNA1* family members represented in our analysis, showed evidence for association (truncated at 0.001, $P = 0.002$) in schizophrenia. This is at a level that just missed our threshold for inclusion in Supplementary Table S1, but is still in the range that is considerably enriched for observed signals compared with the null ($P = 0.0004$). Thus, the calcium channelopathies postulated in bipolar disorder may also operate in schizophrenia.

DGKH was previously implicated in bipolar disorder at genome-wide significance level, $P = 1.5 \times 10^{-8}$.¹⁸ Diacylglycerol kinases are central to an enormous range of signal transduction pathways of potential relevance to neuropsychiatric disorders. Although we did not observe evidence for *DGKH*, *DGKI*, encoding diacylglycerol kinase iota and one of only eight family members represented in this analysis, gave gene-wide evidence for association at a level surpassing the $P = 0.001$ threshold in schizophrenia (Supplementary Table S1; gene-wide $P_{\min} = 6.7 \times 10^{-4}$). Our identification of multiple genes closely related to the handful of genes that have been reported to date by GWAS studies of psychosis points further to the utility of gene-wide analysis of GWAS data sets.

In summary, we present a gene-wide analysis of two GWAS studies, one of schizophrenia and other of bipolar disorder. We show that SNPs within genes are enriched for association signals. We show that the data sets contain substantially more gene-wide signals that surpass nominal significance thresholds than expected by chance, and also but less robustly, that there is an overlap in risk between schizophrenia and bipolar disorder. Genes surpassing thresholds at which there is a considerable enrichment for observed signals are likely to be highly enriched for true associations, and such genes and their family members may form the basis for gene-wide replication studies.

Acknowledgments

This study was supported by the Medical Research Council (UK) and by the Wellcome Trust. This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for that project was provided by the Wellcome Trust under award 076113.

Web Resources

NCBI database NCBI (<http://www.ncbi.nlm.nih.gov/>).
NCBI's dbSNP database (<http://www.ncbi.nlm.nih.gov/SNP/>).
NCBI's Genome database (<http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome>).

References

- 1 Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. *J Clin Invest* 2008; **118**: 1590–1605.
- 2 Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T *et al*. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet* 2008; **40**: 638–645.
- 3 Wellcome Trust Case Control Consortium. Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature* 2007; **447**: 661–678.
- 4 Mathew CG. New links to the pathogenesis of Crohn disease provided by genome-wide association scans. *Nat Genet* 2008; **9**: 10–14.
- 5 Neale BM, Sham PC. The future of association studies: gene-based analysis and replication. *Am J Hum Genet* 2004; **75**: 353–362.
- 6 Ioannidis JP. Non-replication and inconsistency in the genome-wide association setting. *Hum Hered* 2007; **64**: 203–213.
- 7 Terwilliger JD, Hiekkalinna T. An utter refutation of the Fundamental Theorem of the HapMap. *Eur J Hum Genet* 2006; **14**: 426–437.
- 8 Moskvina V, O'Donovan MC. Detailed analysis of the relative power of direct and indirect association studies and the implications for their interpretation. *Hum Hered* 2007; **64**: 63–73.
- 9 Risch N. Linkage strategies for genetically complex traits I: multilocus models. *Am J Hum Genet* 1990; **46**: 222–228.
- 10 Craddock N, Khodel V, Van Eerdewegh P, Reich T. Mathematical limits of multilocus models: the genetic transmission of bipolar disorder. *Am J Hum Genet* 1995; **57**: 690–702.
- 11 Lewis CM, Levinson DF, Wise LH, DeLisi LE, Straub RE, Hovatta I *et al*. Genomescan meta-analysis of schizophrenia and bipolar disorder, part II: schizophrenia. *Am J Hum Genet* 2003; **73**: 34–48.
- 12 Segurado R, Detera-Wadleigh SD, Levinson DF, Lewis CM, Gill M, Nurnberger Jr JI *et al*. Genome scan meta-analysis of schizophrenia and bipolar disorder, part III: bipolar disorder. *Am J Hum Genet* 2003; **73**: 49–62.
- 13 Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM *et al*. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* 2008; **320**: 539–543.
- 14 The International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* 2008; **455**: 237–241.
- 15 Stefansson H, Rujescu D, Cichon S, Pietiläinen OP, Ingason A, Steinberg S *et al*. Large recurrent microdeletions associated with schizophrenia. *Nature* 2008; **455**: 232–236.
- 16 Xu B, Roos JL, Levy S, van Rensburg EJ, Gogos JA, Karayiorgou M. Strong association of *de novo* copy number mutations with sporadic schizophrenia. *Nat Genet* 2008; **40**: 880–885.
- 17 Dudbridge F, Gusnanto A. Estimation of significance thresholds for genomewide association scans. *Genet Epidemiol* 2008; **32**: 227–234.
- 18 Baum AE, Akula N, Cabanero M, Cardona I, Corona W, Klemens B *et al*. A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder. *Mol Psychiatry* 2008; **13**: 197–207.
- 19 Ferreira MAR, O'Donovan MC, Meng YA, Jones IR, Ruderfer DM, Jones L *et al*. Collaborative genome-wide association analysis of 10 596 individuals supports a role for Ankyrin-G (ANKK1) and the alpha-1C subunit of the L-type voltage gated calcium channel (CACNA1C) in bipolar disorder. *Nat Genet* 2008; **40**: 1056–1058.
- 20 O'Donovan MC, Craddock N, Norton N, Williams H, Peirce T, Moskvina V *et al*. Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nat Genet* 2008; **40**: 1053–1055.
- 21 Craddock N, O'Donovan MC, Owen MJ. The genetics of schizophrenia and bipolar disorder: dissecting psychosis. *J Med Genet* 2005; **42**: 193–204.
- 22 Devlin B, Roeder K. Genomic control for association studies. *Biometrics* 1999; **55**: 997–1004.
- 23 Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS. Truncated product method for combining *P*-values. *Genet Epidemiol* 2002; **22**: 170–185.
- 24 Gottesman II, Shields J. A polygenic theory of schizophrenia. *Proc Natl Acad Sci USA* 1967; **58**: 199–205.
- 25 O'Rourke DH, Gottesman II, Suarez BK, Rice J, Reich T. Refutation of the general single-locus model for the etiology of schizophrenia. *Am J Hum Genet* 1982; **34**: 630–649.
- 26 Crow TJ. The emperors of the schizophrenia polygene have no clothes. *Psychol Med* 2008; **21**: 1–5.
- 27 Craddock N, O'Donovan MC, Owen MJ. Genome-wide association studies in psychiatry: lessons from early studies of non-psychiatric and psychiatric phenotypes. *Mol Psychiatry* 2008; **13**: 649–653.
- 28 Lencz T, Morgan TV, Athanasiou M, Dain B, Reed CR, Kane JM *et al*. Converging evidence for a pseudoautosomal cytokine receptor gene locus in schizophrenia. *Mol Psychiatry* 2007; **12**: 572–580.
- 29 Chen Q, Wang X, O'Neill FA, Walsh D, Fanous A, Kendler KS *et al*. Association study of CSF2RB with schizophrenia in Irish family and case - control samples. *Mol Psychiatry* 2008; **13**: 930–938.
- 30 Hanson DR, Gottesman II. Theories of schizophrenia: a genetic-inflammatory-vascular synthesis. *BMC Med Genet* 2005; **6**: 7.

Supplementary Information accompanies the paper on the Molecular Psychiatry website (<http://www.nature.com/mp>)