Hypothesis
Based Prioritization Algorithm and Neurodevelopmental Hypothesis

Zhongming Zhao1,2,3, Bradley T. Webb3,4,5, Peilin Jia1, T. Bernard Bigdeli3, Brion S. Maher2, Edwin van den Oord3,4, Sarah E. Bergen6,7, Richard L. Amdur8, Francis A. O’Neill9, Dermot Walsh10, Dawn L. Thiselton3, Xiangning Chen11,12, Carlos N. Pato1,14, The International Schizophrenia Consortium,1, Brien P. Riley3,11,12, Kenneth S. Kendler3,11,12, Ayman H. Fanous3,8,11,13,14

Abstract

Integrating evidence from multiple domains is useful in prioritizing disease candidate genes for subsequent testing. We ranked all known human genes (n = 3819) under linkage peaks in the Irish Study of High-Density Schizophrenia Families using three different evidence domains: 1) a meta-analysis of microarray gene expression results using the Stanley Brain collection, 2) a schizophrenia protein-protein interaction network, and 3) a systematic literature search. Each gene was assigned a domain-specific p-value and ranked after evaluating the evidence within each domain. For comparison to this ranking process, a large-scale candidate gene hypothesis was also tested by including genes with Gene Ontology terms related to neurodevelopment. Subsequently, genotypes of 3725 SNPs in 167 genes from a custom Illumina iSelect array were used to evaluate the top ranked vs. hypothesis selected genes. Seventy-three genes were both highly ranked and involved in neurodevelopment (category 1) while 42 and 52 genes were exclusive to neurodevelopment (category 2) or highly ranked (category 3), respectively. The most significant associations were observed in genes PRKGC1, PRKCE, and CNTN4 but no individual SNPs were significant after correction for multiple testing. Comparison of the approaches showed an excess of significant tests using the hypothesis-driven neurodevelopment category. Random selection of similar sized genes from two independent genome-wide association studies (GWAS) of schizophrenia showed the excess was unlikely by chance. In a further meta-analysis of three GWAS datasets, four candidate SNPs reached nominal significance. Although gene ranking using integrated sources of prior information did not enrich for significant results in the current experiment, gene selection using an a priori hypothesis (neurodevelopment) was superior to random selection. As such, further development of gene ranking strategies using more carefully selected sources of information is warranted.


Editor: Huiping Zhang, Yale University, United States of America

Received: August 28, 2012; Accepted: May 28, 2013; Published: July 29, 2013

This is an open-access article, free of copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the Creative Commons CC0 public domain dedication.

Funding: Two dataset(s) used in the analyses described in this manuscript were obtained from the database of Genotype and Phenotype (dbGaP) found at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number (GAIN: phs000021.v2.p1, nonGAIN: phs00167.v1.p1). For the GAIN dataset, the genotyping of samples was provided through the Genetic Association Information Network (GAIN). The CATIE trial was funded by a grant from the National Institute of Mental Health (N01 MH90001) along with MH074027 (PI PF Sullivan). Genotyping was funded by Eli Lilly and Company. The CATIE dataset was approved to use in this analysis through the authors’ application. This work was supported by Department of Veterans Affairs Merit Review Program (101CX000278, to AHF), 2009 NARSAD Maltz Investigator Award (to ZH), and 2010 NARSAD Young Investigator Award (to PJ). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: Genotyping was funded by Eli Lilly and Company. The CATIE dataset was approved to use in this analysis through the authors’ application. These genotypes are available online and made available to qualified investigators. This does not alter the authors’ adherence to all the PLOS ONE policies on sharing data and materials.

* E-mail: ayman.fanous@va.gov
† These authors contributed equally to this work.
‡ Membership of The International Schizophrenia Consortium is provided in the Acknowledgments.
Introduction

A wealth of information relevant to the genetics of complex disorders is available via a wide variety of platforms such as gene expression, protein-protein interactions (PPIs), biological pathways, and Gene Ontology (GO). It was hoped that the advent of large scale genome-wide association studies (GWAS) would eliminate the need to utilize this data as a means to uncover susceptibility loci. However, psychiatric GWAS have shown that there are likely many loci of small effect and few results are significant after corrections for multiple testing [1,2,3]. Furthermore, the loci that do survive only account for modest proportions of heritability. Therefore, novel methods are still needed to identify additional causative loci. The use of multiple, existing sources of information could increase statistical power to detect susceptibility genes and minimize the risk of pursuing false positives in follow up investigations. However, due to the large amount of information plus heterogeneity among data sources, the task of combining such information in an optimal way is complex and difficult, either intuitively or manually.

Schizophrenia is a disorder that is particularly suitable to this type of approach. While other complex disorders and traits such as type 2 diabetes and height have been gathering a rapidly growing list of replicated and validated susceptibility loci, several features of schizophrenia will arguably make such success less likely. Although its heritability is higher than many complex disorders such as type 2 diabetes, its prevalence is lower. This makes very large studies with tens or hundreds of thousands of participants much more challenging (albeit necessary in order to detect an effect). There is also phenotypic and diagnostic heterogeneity which is arguably less present in other complex disorders and which may reflect genetic heterogeneity as well. Moreover, for schizophrenia, there is increasing evidence suggesting a complex genetic architecture comprising a mixture of rare highly-penetrant mutations such as large deletions in gene NRG1 [4] as well as common single nucleotide polymorphisms (SNPs) [5]. Furthermore, well developed animal models or the availability of patient tissue are very rare. Additional causative loci. The use of multiple, existing sources of information could increase statistical power to detect susceptibility genes and minimize the risk of pursuing false positives in follow up investigations. However, due to the large amount of information plus heterogeneity among data sources, the task of combining such information in an optimal way is complex and difficult, either intuitively or manually.

Materials and Methods

Ethics statement

This research was approved by the Institutional Review Boards of Virginia Commonwealth University School of Medicine and the Washington VA Medical Center. All subjects gave verbal assent to participate in research, as this represented the ethical standard in Ireland at the time these data were collected. This strategy was specifically approved by the Health Research Board, Dublin. Permission was received to use the data in this study, and the data we de-identified prior to analysis.

Subjects and phenotypes

The Irish Study of High Density Schizophrenia Families (ISHDSF) sample consists of 265 high-density schizophrenia families with 1408 individuals available for genotyping [8]. All participating individuals gave appropriate informed consent to the study. The sample was divided into 4 concentric diagnostic categories for analysis purposes, ranging from core schizophrenia (D2, 625 affected individuals), through narrow spectrum (‘intermediate phenotype’ D5, 804 affected individuals), broad (D6, 888 affected individuals) and very broad spectrum disease (D9, 1172 affected individuals). Phenotypic details of these subcategories are given briefly in Thielson et al. [9].

Linkage regions

We first limited the ranking to genes in regions with evidence for linkage in the ISHDSF. These regions were obtained from an autosomal genome-wide scan using over 4000 SNPs as part of the Multicenter Genetic Studies of Schizophrenia (PI, Douglas F. Levinson, MD) [10]. Regions were defined as genomic segments with nonparametric linkage (NPL) maximum score of at least 2.0 and telomeric and centromeric boundaries of NPLs of 1.0. The detailed genomic locations were provided in File S1. A bioinformatics search of these regions yielded 3819 human protein-coding genes.

Prior sources of information

For each of the 3819 genes, we obtained a separate p-value pertaining to each of 3 domains: 1) gene expression, 2) protein-protein interaction (PPI) subnetwork, and 3) high-throughput literature search, as illustrated in Figure 1. First, the p-value for fold-change in each gene’s expression level was obtained from the Stanley Brain Expression Database (http://www.stanleygenomics.org), which contains meta-analysis results using data from 12 different studies and 988 arrays. A False Discovery Rate (FDR) procedure [11] was applied to the uncorrected p-values and used to generate a corrected ranked p-value. Second, assuming disease susceptibility genes may be functionally connected, we identified the genes whose proteins interact closely with proteins encoded by three established schizophrenia susceptibility genes DTNBP1 [12,13], NRG1 [14,15], and AKT1 [9] in the PPI network. A comprehensive human PPI network was generated using human PPI data retrieved from NCBI Entrez Gene (February 2007) which summarizes interactions from multiple sources including HPRD (http://www.hprd.org) [16], BioGrid (http://www.thebiogrid.org) [17,18], and BIND (http://www.bind.ca) [19]. After removing redundant and problematic interactions, 52,288 unique human PPI pairs remained in the network. The program Pajek (http://vlado.fmf.uni-lj.si/pub/networks/pajek/) [20] was used to determine the minimum number of steps between the proteins encoded by DTNBP1, NRG1, and AKT1 and every other human gene in the PPI network. Each of the 3819 genes was assigned a rank based p-value based on the number of steps (lowest to highest). The hypothesis is that a gene closer in the network to a probable susceptibility gene is more likely to harbor susceptibility alleles. Finally, high-throughput literature searching was performed using a Perl script which automatically queried the PubMed database (http://www.ncbi.nlm.nih.gov/pubmed/) for each of the 3819 genes along with 29 schizophrenia-related search terms that we assembled (3819×29=110,751 searches). These search terms were divided into several categories: disease states (e.g., “schizophrenia”, “psychosis”), neurotransmitters (e.g., glutamate, dopamine), neuronal features (e.g., “dendrite”, “axon”), brain development, and brain structures (e.g., “cortex”). Genes were ranked...
For the genomic regions, we attempted to select all haplotype-tagging genic SNPs within each gene using computer program Tagger [23] \((r^2 = 0.8, \text{ minor allele frequency (MAF)} = 0.1)\) and the HapMap data \((\text{phase 2}, \text{ http://hapmap.ncbi.nlm.nih.gov/})\). Genotyping was conducted by Illumina, Inc. using a custom iSelect array, which employs the Infinium assay. In total, genomic DNA for 1128 individuals was submitted for genotyping. Average genotyping completion rate across all SNPs was 99.97%. Of 1128 samples, 21 failed to yield usable genotypes. Genotypes were examined for apparent Mendelian incompatibilities using PEDCHECK v1.1 [24] and removed for entire families where appropriate. After excluding SNPs failing quality control, 3725 SNPs were available for analysis.

Association analyses

Association analysis for categorical diagnoses of schizophrenia was performed using PDTPHASE (UNPHASED v.2.4.04), an implementation of the pedigree disequilibrium test (PDT) with extensions to deal with uncertain haplotypes and missing data [25,26]. The PDT is an extension of the transmission disequilibrium test (TDT) to examine general pedigree structures and is similarly a test of association in the presence of linkage.

GWAS datasets

The International Schizophrenia Consortium (ISC) samples were collected from eight study sites in Europe and the US [1]. The samples were genotyped using Affymetrix Genome-Wide Human SNP 5.0 and 6.0 arrays. This data was initially analyzed by ISC [1] and was used here for evaluation. A total of 3322 patients with schizophrenia, 3587 normal controls of European ancestry, and a total of 739,995 SNPs were included in our analysis. To account for potential population sub-structure associated with collection sites, the Cochran-Mantel-Haenszel test was used for a single marker association test [1].

We used two GWAS datasets from the Molecular Genetics of Schizophrenia (MGS): The Genetic Association Information Network (GAIN) dataset for schizophrenia and nonGAIN. The GAIN dataset was genotyped using Affymetrix Genome-Wide Human SNP 6.0 array. Our access to this dataset was approved by the GAIN Data Access Committee (DAC request #4532-2) through the NCBI dbGaP. For optimal comparison with the Irish samples (ISHDSF) genotyped in this study, we used only the GAIN samples of European ancestry. We performed quality control (QC) as follows. For individuals, those with a high missing genotype rate (>5%), extreme heterozygosity rate (±3 s.d. from the mean value of the distribution), or problematic gender assignment were excluded. PLINK [27] was used to compute the identify-by-state (IBS) matrix to pinpoint duplicate or cryptic relationships between individuals. We retained the sample with the highest call rate for each pair of samples with an identity-by-descent (IBD) being greater than 0.185. Principle component analysis (PCA) was performed using the smartpca program in EIGENSTRAT [28] to detect population structure and to allow removal of outlier individuals. Eight significant PCs with the Tracy–Widom test p-value < 0.05 were used as covariates for logistic regression (additive model). For genotyped SNPs, those with a missing genotype rate >5%, MAF < 0.05, or departing from Hardy-Weinberg equilibrium \((p<1 \times 10^{-6})\) were removed. The final analytic dataset included 1158 schizophrenia cases and 1377 controls and a total of 654,271 SNPs. The genomic inflation factor \((\lambda)\), which was defined as the ratio of the median of the empirically observed distribution of the test statistic to the expected median and an indication of the extent of excess false positive rate [29], was 1.04. This value indicates little (if any) inflation.
The MGS - nonGAIN dataset was genotyped in the same laboratory as the MGS-GAIN, but in different phases. Access to this dataset was approved by dbGaP (DAC request #4533-3). Similar QC and PCA processes were performed as described for GAIN. These processes retained 1086 cases and 1268 controls and 623,059 SNPs for subsequent analyses. Fifteen significant PCs with the Tracy-Widom test p-value<0.05 were used as covariates for logistic regression (additive model) using PLINK. The genomic inflation factor (λ) was 1.04.

CATIE. (Clinical Antipsychotic Trials of Intervention Effectiveness) is a multi-phase randomized controlled trial of antipsychotic medications involving 1460 persons with schizophrenia. CATIE GWAS included 492,900 SNPs genotyped in a total of 738 cases and 733 group-matched controls using the Affymetrix 500K two-chip genotyping platform plus a custom 164K fill-in chip [30]. Access to this dataset was approved by the National Institute of Mental Health (NIMH) Schizophrenia Genetics Initiative.

Imputation and meta-analysis

The three GWAS datasets, ISC, GAIN, and nonGAIN, were genotyped on the same Affymetrix platform. To make the data from these GWAS datasets comparable with our custom-design SNPs, we conducted imputation analysis using the HapMap genotyping data for CEU population (release 24) as reference panel. We predicted the genotyping data for a total of 66 SNPs involved in 22 genes using the tool impute2 [31]. Frequentist association test was then conducted for SNP association using the tool snptest [32] by the option “-frequentist 1”, and a missing data likelihood score test for the imputed genotypes by the option “-method score”.

We conducted meta-analysis of candidate SNPs using the imputed data. We performed inverse-variance weighted meta-analysis based on the fixed-effects model using the tool meta (http://www.stats.ox.ac.uk/~jsliu/meta.html). This method combines study-specific beta values under the fixed-effects model using the inverse of the corresponding standard errors as weights. Between-study heterogeneity was tested based on I² and Q statistics. SNPs having possible evidence of heterogeneity (I²-heterogeneity<0.05) were removed.

Gene set simulations

The three GWAS datasets, ISC, GAIN, and nonGAIN, were genotyped on the same Affymetrix platform. To make the data from these GWAS datasets comparable with our custom-design SNPs, we conducted imputation analysis using the HapMap genotyping data for CEU population (release 24) as reference panel. We predicted the genotyping data for a total of 66 SNPs involved in 22 genes using the tool impute2 [31]. Frequentist association test was then conducted for SNP association using the tool snptest [32] by the option “-frequentist 1”, and a missing data likelihood score test for the imputed genotypes by the option “-method score”.

We conducted meta-analysis of candidate SNPs using the imputed data. We performed inverse-variance weighted meta-analysis based on the fixed-effects model using the tool meta (http://www.stats.ox.ac.uk/~jsliu/meta.html). This method combines study-specific beta values under the fixed-effects model using the inverse of the corresponding standard errors as weights. Between-study heterogeneity was tested based on I² and Q statistics. SNPs having possible evidence of heterogeneity (I²-heterogeneity<0.05) were removed.

Gene set simulations

In order to determine how often the observed enrichment in p-values would occur, 100,000 simulations were performed where the same number of genes was randomly chosen from the CATIE and GAIN GWAS results. Then, p-values less than 0.05 and 0.005 for genotyped SNPs that mapped to the randomly chosen genes were counted. Due to the great variation in gene size, SNP density per gene, and difference in arrays used in each GWAS, the number of SNPs in each iteration of the simulations could vary. Therefore, we examined whether the observed number of SNPs for the real set of ranked genes was similar to randomly selected sets. The empirical significance for SNP count, minimum p-value, and p-values below a given threshold were calculated both using all simulations and restricted to those simulations where the SNP count was not significantly different from observed. The empirical significance was calculated using the number of simulations greater than or equal to the observed plus one divided by the total number of simulations as per North et al. [33].

Results

Figure 1 summarizes the data process, algorithm for gene ranking and selection, custom-based genotyping and association analysis. We analyzed 3725 SNPs covering 167 prioritized genes whose genotypes were examined in 1107 individuals from the 265 high-density schizophrenia families using a custom Illumina iSelect array. This gene list included 115 genes selected by neurodevelopmental hypothesis since it ranked only 954th of the 3819 genes. There was an enrichment of significant p-values in the mental hypothesis and highly rank categories. Gene rank was not correlated with the minimum p-value observed in the tested genes. There was an enrichment of significant p-values in the neurodevelopmental category; however, surprisingly, the ranked genes...
prKCE heterogeneity test. These SNPs are rs2176348 in three schizophrenia GWAS datasets (ISC, GAIN and non-GAIN). We examined them in a meta-analysis using Meta-analysis Comparison with published schizophrenia GWAS results

We compared the current results to two published schizophrenia GWAS datasets, the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) GWAS dataset [30] and the GAIN dataset from Molecular Genetics of Schizophrenia (MGS) [2] (see Materials and Methods). Simulated gene selection was used to determine how often the observed enrichment would occur if the whole genome was assessed. Lists of genes were randomly selected 100,000 times using the same numbers of genes selected in the three selection categories (73, 42, and 52 for categories 1, 2, and 3, respectively). Unfiltered simulation showed that the number of SNPs per gene in the current study was significantly higher than for randomly selected genes from either the CATIE or GAIN study. Therefore, the tests for significant enrichment of p-values below 0.05 and 0.005 are biased in the unfiltered simulations. To reduce the bias, random gene sets were ranked based on the total number of SNPs. Different rank filtering thresholds were tested until there was no significant difference in total number of SNPs between the observed and simulated sets. The rank filter thresholds necessary to achieve non-significance were quite different for the CATIE and GAIN studies with the top 500 and 10,000 being used respectively. The filtered simulations showed the ‘neurodevelopment only’ category to be significantly different (p = 0.012) and marginally (p = 0.058) enriched for p-values less than 0.005 in the CATIE and MGS-GAIN samples, respectively. Further details of the simulation results are in Table 3.

Meta-analysis

In our SNP list, there were 66 SNPs with p<0.01. These SNPs belonged to 22 genes. We examined them in a meta-analysis using three schizophrenia GWAS datasets (ISC, GAIN and non-GAIN). Using the inverse-variance weighted meta-analysis method, we identified 3 SNPs in 3 genes that showed nominal significance (p<0.05) (Table 4). None of them had significant heterogeneity by heterogeneity test. These SNPs are rs2176348 in prKCE (p-value = 0.044), rs352551 in MGMT (p-value = 0.044), and rs2616591 in CNTN4 (p-value = 0.048). Another SNP, rs2043534 in NPAS2, had marginal significance (p-value = 0.062). However, none of these SNPs passed Bonferroni multiple testing correction.

We further examined the association signals of these SNPs using the data from the Psychiatric Genomics Consortium (PGC), the largest and most comprehensive cohort dataset for schizophrenia association studies so far [34]. Among the 66 SNPs, 24 were not available in the public release of the PGC dataset (https://pgc.unc.edu/Sharing.php); thus, they could not be imputed due to the lack of access to PGC’s raw genotyping data. For the 42 SNPs that had p-values in the PGC dataset, we found that three were nominally significant (p<0.05), including one SNP (rs2616591) in Table 4. Of note, for the four SNPs that were significant or marginally significant in the meta-analysis of our SNPs (Table 4), two were available in the PGC dataset including SNP rs2616591 that had a small p value (1.68×10⁻⁷).

Discussion

In this study, we attempted to develop gene ranking strategies based on either evidence from multiple domains (meta-analysis of gene expression, proteins closely interacting with well-studied schizophrenia susceptibility genes, and a systematic literature search) or the neurodevelopmental hypothesis and then applied them to the genes under linkage peaks in the Irish Study of High-Density Schizophrenia Families. For the top ranked genes, we tested their associations with schizophrenia using a custom Illumina iSelect array. The association signals were further evaluated using three GWAS datasets (ISC, GAIN, and non-GAIN). Although none of the SNPs were robustly associated, clusters of significant SNPs were found in several large genes including prKG1, CNTN4, and prKCE. These genes were tested not due to rank but as part of the neurodevelopmental hypothesis. This category showed enrichment for significant association signals, and simulations showed this enrichment is unlikely to be due to chance.

There is additional evidence that makes the top results of interest in addition to the reason they were originally tested. CNTN4 (contactin 4) is a neural cell adhesion molecule whose gene has been reported to be associated with autism and developmental delay in multiple studies [35,36,37]. Interestingly, another member of the contactin family, CNTNAP2, has been found to be associated with both schizophrenia and autism [38]. PRKG1 and PRKCE are known as protein kinase cGMP-dependent, type I and protein kinase C epsilon, respectively. Although they are both protein kinases, they are functionally distinct and activated via different mechanisms. PRKG1 is dependent on cyclic GMP for activation while PRKCE is activated by calcium and the second messenger diacylglycerol. prKG1 has previously shown its association with schizophrenia with the 21st most significant SNP in the CATIE GWAS [30]. PRKG1 also interacts with RGS2 and GABBR1, which have

Table 2. Summary of genes and number of SNPs per test category.

<table>
<thead>
<tr>
<th>Category</th>
<th># genes</th>
<th># SNPs</th>
<th>Highly ranked Neuro-development</th>
<th>Mean p-value</th>
<th>SNPs with p&lt;0.05</th>
<th>SNPs with p&lt;0.005</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>73</td>
<td>1271</td>
<td>Yes</td>
<td>0.521</td>
<td>42</td>
<td>63.6</td>
</tr>
<tr>
<td>2</td>
<td>42</td>
<td>1525</td>
<td>No</td>
<td>0.480</td>
<td>103</td>
<td>76.3</td>
</tr>
<tr>
<td>3</td>
<td>52</td>
<td>929</td>
<td>Yes</td>
<td>0.488</td>
<td>63</td>
<td>46.5</td>
</tr>
<tr>
<td>1+2</td>
<td>115</td>
<td>2796</td>
<td>-</td>
<td>0.486</td>
<td>145</td>
<td>139.8</td>
</tr>
<tr>
<td>1+3</td>
<td>125</td>
<td>2200</td>
<td>Yes</td>
<td>0.507</td>
<td>105</td>
<td>110.0</td>
</tr>
<tr>
<td>All</td>
<td>167</td>
<td>3725</td>
<td></td>
<td>0.507</td>
<td>208</td>
<td>186.3</td>
</tr>
</tbody>
</table>

Category 1: genes are both highly ranked and involved in neurodevelopment. Category 2: genes are exclusive to neurodevelopment. Category 3: genes are exclusively highly ranked (see details in Materials and Methods).

doi:10.1371/journal.pone.0067776.t002
shown modest association with schizophrenia symptoms [39] and schizoaffective disorder [40], respectively. Finally, PRKG1 can attenuate beta-catenin expression [41], which is a known downstream target of antipsychotics [42]. PRKCE interacts with several proteins encoded by genes of potential relevance to psychiatric disorders, including the glutamate decarboxylases (GAD1, GAD2), NMDA receptors (GRIN2D, GRIN1), and a metabotropic glutamate receptor (GRM5). PRKCE is also activated by the stimulation of nicotinic receptors [43]. Although these genes were not highly ranked, prior evidence makes them all plausible candidates for schizophrenia. Therefore, each could be chosen using expanded sources of prior information and a refined ranking procedure.

There are several limitations to the current work that could be potentially improved in future application. First, the primary filtering of genes in the genome was done using linkage results from the ISHDSF. Due to the large number of risk variants in schizophrenia, there are likely to be many true associations outside of these regions. The second limitation is the small number of genes and minimum step approach used for the PPI network sets. We used three well-studied genes (DTNBP1, NRG1, and AKT1) in this work. More informative genes including microRNA genes (e.g., miR-137 [34,44] and TCF4 [3,34]) were recently reported to be associated with schizophrenia and could improve this approach. Larger networks or results from more comprehensive network analyses are probably superior; nevertheless, this work proves the concept using more closely related genes in the PPI network sets.

<table>
<thead>
<tr>
<th>Category</th>
<th>Simulation Method</th>
<th>Observed in ISHDSF</th>
<th>CATIE GWAS 100,000</th>
<th>Top 500</th>
<th>GAIN GWAS 100,000</th>
<th>Top 10,000</th>
<th>Top 500</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>SNP count</td>
<td>3741</td>
<td>0.0097</td>
<td>0.194</td>
<td>0.066</td>
<td>0.659</td>
<td>0.998</td>
</tr>
<tr>
<td></td>
<td>min p-value</td>
<td>0.000582</td>
<td>0.618</td>
<td>0.896</td>
<td>0.787</td>
<td>0.879</td>
<td>0.964</td>
</tr>
<tr>
<td></td>
<td># SNPs with p&lt;0.05</td>
<td>208</td>
<td>0.0019</td>
<td>0.208</td>
<td>0.158</td>
<td>0.767</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td># SNPs with p&lt;0.005</td>
<td>23</td>
<td>0.0258</td>
<td>0.234</td>
<td>0.241</td>
<td>0.614</td>
<td>0.924</td>
</tr>
<tr>
<td>1</td>
<td>SNP count</td>
<td>1271</td>
<td>0.079</td>
<td>0.655</td>
<td>0.389</td>
<td>0.806</td>
<td>0.892</td>
</tr>
<tr>
<td></td>
<td>min p-value</td>
<td>0.001975</td>
<td>0.729</td>
<td>0.872</td>
<td>0.859</td>
<td>0.920</td>
<td>0.956</td>
</tr>
<tr>
<td></td>
<td># SNPs with p&lt;0.05</td>
<td>42</td>
<td>0.586</td>
<td>0.914</td>
<td>0.887</td>
<td>0.976</td>
<td>0.986</td>
</tr>
<tr>
<td></td>
<td># SNPs with p&lt;0.005</td>
<td>4</td>
<td>0.556</td>
<td>0.836</td>
<td>0.780</td>
<td>0.895</td>
<td>0.934</td>
</tr>
<tr>
<td>2</td>
<td>SNP count</td>
<td>1525</td>
<td>0.0222</td>
<td>0.104</td>
<td>0.017</td>
<td>0.121</td>
<td>0.267</td>
</tr>
<tr>
<td></td>
<td>min p-value</td>
<td>0.000582</td>
<td>0.214</td>
<td>0.349</td>
<td>0.319</td>
<td>0.398</td>
<td>0.445</td>
</tr>
<tr>
<td></td>
<td># SNPs with p&lt;0.05</td>
<td>103</td>
<td>0.00042</td>
<td>0.032</td>
<td>0.017</td>
<td>0.102</td>
<td>0.232</td>
</tr>
<tr>
<td></td>
<td># SNPs with p&lt;0.005</td>
<td>17</td>
<td>0.00294</td>
<td>0.012</td>
<td>0.018</td>
<td>0.058</td>
<td>0.142</td>
</tr>
<tr>
<td>3</td>
<td>SNP count</td>
<td>929</td>
<td>0.088</td>
<td>0.591</td>
<td>0.358</td>
<td>0.707</td>
<td>0.822</td>
</tr>
<tr>
<td></td>
<td>min p-value</td>
<td>0.003559</td>
<td>0.797</td>
<td>0.926</td>
<td>0.883</td>
<td>0.932</td>
<td>0.972</td>
</tr>
<tr>
<td></td>
<td># SNPs with p&lt;0.05</td>
<td>63</td>
<td>0.049</td>
<td>0.427</td>
<td>0.264</td>
<td>0.558</td>
<td>0.719</td>
</tr>
<tr>
<td></td>
<td># SNPs with p&lt;0.005</td>
<td>2</td>
<td>0.701</td>
<td>0.896</td>
<td>0.837</td>
<td>0.914</td>
<td>0.954</td>
</tr>
</tbody>
</table>

*aCategory 1: genes are both highly ranked and involved in neurodevelopment. Category 2: genes are exclusive to neurodevelopment. Category 3: genes are exclusively highly ranked (see details in text).
*bWe performed simulations by four methods: 1) based on the count of SNPs, 2) based on the minimum p-value, 3) based on the number of SNPs with p<0.05, and 4) based on the number of SNPs with p<0.005.
*c100,000 simulations (see text).
*dTo reduce bias, simulations were filtered with top 500 or 10,000 SNPs being used (see Materials and Methods).

doi:10.1371/journal.pone.0067776.t003

There are several limitations to the current work that could be potentially improved in future application. First, the primary filtering of genes in the genome was done using linkage results from the ISHDSF. Due to the large number of risk variants in schizophrenia, there are likely to be many true associations outside of these regions. The second limitation is the small number of genes and minimum step approach used for the PPI network sets. We used three well-studied genes (DTNBP1, NRG1, and AKT1) in this work. More informative genes including microRNA genes (e.g., miR-137 [34,44] and TCF4 [3,34]) were recently reported to be associated with schizophrenia and could improve this approach. Larger networks or results from more comprehensive network analyses are probably superior; nevertheless, this work proves the concept using more closely related genes in the PPI network sets.

**Table 4.** Four SNPs from the meta-analysis of 66 SNPs using GAIN, nonGAIN, and ISC GWAS datasets.

<table>
<thead>
<tr>
<th>Gene</th>
<th>SNP ID</th>
<th>Chr</th>
<th>Position (bp)</th>
<th>Allele</th>
<th>Meta-analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p-value</td>
<td>Beta</td>
<td>s.e.</td>
<td>$P_{\text{heterogeneity}}$</td>
<td>$I^2$</td>
</tr>
<tr>
<td>PRKCE</td>
<td>rs2176348</td>
<td>2</td>
<td>45798033</td>
<td>A/G</td>
<td>0.044</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MGMT</td>
<td>rs552551</td>
<td>10</td>
<td>1312771915</td>
<td>C/T</td>
<td>0.044</td>
</tr>
<tr>
<td>CNTN4</td>
<td>rs2616591</td>
<td>3</td>
<td>2614861</td>
<td>C/T</td>
<td>0.048</td>
</tr>
<tr>
<td>NPAS2</td>
<td>rs2043534</td>
<td>2</td>
<td>100847317</td>
<td>C/T</td>
<td>0.062</td>
</tr>
</tbody>
</table>

Chr: chromosome. GAIN, nonGAIN and ISC are three GWAS datasets for meta-analysis. ISHDSF min p was the smallest p-value in the gene from the IHDS dataset (this study). NA: this SNP was not analyzed in ISC due to missing genotyping data in samples.

doi:10.1371/journal.pone.0067776.t004
network. Third, while our keyword-based literature search seemed to be useful, it might include underpowered studies, negative findings, or studies with methodological flaws or reported false positive results. This is a common problem in literature mining, which could be improved by careful manual check or advanced literature mining technologies like natural language processing (NLP). In our study, gene ranking was performed by the combined evidence from three domains (gene expression meta-analysis, PPI subnetwork, and literature mining). This strategy might help reduce the noisy data from literature mining. Finally, besides the well-supported neurodevelopmental hypothesis, we may test other hypotheses or the candidate genes for samples with refined characterization of phenotypic spectrum. For example, Greenwood et al. [45] recently tested a set of schizophrenia candidate genes in schizophrenia-related endophenotypes, suggesting both converging and independent genetic pathways mediating schizophrenia risk and pathogenesis.

There are several ways to improve or expand the gene selection and prioritization approaches. First, we may develop a more comprehensive data integration approach. This includes the integration of data from multiple domains such as gene expression, copy number variation (CNV), methylation, microRNA, association results, etc. This has been demonstrated in our weight matrix approach for evidence scores [46], as well as other approaches like convergent analysis [47,48,49] and microRNA regulatory network analysis [50]. Of note, TCP4 gene, along with three other genes reported in the PG Meta-analysis (C4/CNA1E, CSMD1 and C16orf29), has predicted miR-137 target sites [34]. This makes the microRNA-mediated regulatory analysis promising in schizophrenia. In terms of the algorithm, we may apply Bayesian approach, or a comprehensive network and pathway approach, to those multi-domain datasets, since the underlying biological information and regulation is expected to be much related in a complex disease. For example, we recently demonstrated our network approach in schizophrenia [51]. Such approach can be expanded in future by including transcriptional (transcription factors, methylation) and post-transcriptional (microRNA) regulation. Second, with the rapid advances in high throughput technologies, such as Exome chip or next-generation sequencing, we may prioritize the candidate genes that show association signals detected by both common and rare variants and that are involved in disease-related altered genomic regions such as CNVs or structural variants (SVs). This approach benefits from cross-platform and cross-study validation.

In summary, we did not find compelling association evidence for any individual gene selected either by evidence-based gene-ranking or by the rank based on its relevance to the neurodevelopmental hypothesis. However, the neurodevelopmental set of genes showed enrichment for significant associations when examined en masse. Finally, several tested genes have additional independent evidence not used in the ranking that make them attractive candidates for further investigation.

Supporting Information

File S1 This file includes 26 linkage peaks (genomic regions) with nonparametric linkage (NPL) maximum score being at least 2 and telomeric and centromeric boundaries of NPLs of 1.0. (XLS)

File S2 This file includes details of the 167 genes ranked by three categories and their SNPs with association results in The Irish Study of High Density Schizophrenia Families (ISHDSF) samples. (XLSX)

Acknowledgments

Two dataset(s) used in the analyses described in this manuscript were obtained from the database of Genotype and Phenotype (dbGaP) found at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number [GAIN: phs000021.v2.p1, nonGAIN: phs001676.v1.p1]. For the GAIN dataset, the genotyping of samples was provided through the Genetic Association Information Network (GAIN). The CATIE dataset was approved to use in this analysis through our application. We thank the Schizophrenia Psychiatric Genomics Consortium (PGC) for sharing the meta-analysis results. We thank the families for their participation in this research.

The members of the International schizophrenia consortium:


1Psychiatric and Neurodevelopmental Genetics Unit, 2Center for Human Genetic Research, Massachusetts General Hospital, 15 Cambridge Street, Boston, Massachusetts 02114, USA. 3Stanley Center for Psychiatric Research, The Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. 4The Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA. 5Queensland Institute of Medical Research, 300 Herston Road, Brisbane, Queensland 4006, Australia. 6MRC Centre for Neuropsychiatric Genetics and Genomics, Department of Psychological Medicine, School of Medicine, Cardiff University, Cardiff C14 4XN, UK. 7Departments of Genetics, Psychiatry, and Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. 8Molecular Psychiatry Laboratory, Research Department of Mental Health Sciences, University College London Medical School, Windeyer Institute of Medical Sciences, 46 Cleveland Street, LondonW1FJF, UK. 9NeuropsychiatricGenetics Research Group, Department of Psychiatry and Institute of Molecular Medicine, Trinity College Dublin, Dublin 2, Ireland. 10Division of Psychiatry, University of California San Diego, and Department of Psychiatry, First Psychiatric Clinic, Alexander University Hospital, Sofia 1431, Bulgaria. 11Department of Psychiatry, University of Edinburgh: 12Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, SE-171 77 Stockholm, Sweden. 13Department of Neurosciences, Psychiatry, Ulleraker, Uppsala University, SE-750 17 Uppsala, Sweden. 14Center for Genomic Psychiatry, University of Southern California, Los Angeles, California 90033, USA. 15Institute of Medical Sciences, 16Department of Mental Health, University of Aberdeen, Aberdeen AB25 2ZD, UK. 17Department of Medical Genetics, University Hospital Maria Inmaculada Dom, Sofia 1431, Bulgaria. 18Department of Psychiatry, First Psychiatric Clinic, Alexander University Hospital, Sofia 1431, Bulgaria. 19Department of Psychiatry, University of Oxford, Warneford Hospital, Headington, Oxford OX3 7JX, UK. 20Ravenscourt Hospital, Inverkip Road, Greenock PA16 9HA, UK. 21State University of
New York – Upstate Medical University, Syracuse, New York 13210, USA. 25Washington VA Medical Center, Washington, DC 20422, USA. 26Department of Psychiatry, Georgetown University School of Medicine, Washington DC 20057, USA. 27Department of Psychiatry, Virginia Commonwealth University, Richmond, Virginia 23298, USA. 28Department of Psychiatry, Sao Miguel, 9500-310 Azores, Portugal. 29Department of Psychiatry University of Coimbra, 3004-504 Coimbra, Portugal.

Author Contributions
Conceived and designed the experiments: AHF BTW ZZ BSM RLA. Performed the experiments: AHF KSK DW FAO RLA. Analyzed the data: PJ TBB ISC DLT. Wrote the paper: ZZ BTW AHF PJ.

References