Network propagation enhances power of GWAS analysis for PGC schizophrenia cohort

Nadia Arang, Kevin Chau

Genetics underlie susceptibility and causality of human disease across all physiological systems. In particular, the genetic basis of psychiatric disorders has been long established; however, they are among the most debilitating diseases as well as the most challenging to characterize and treat. Schizophrenia is a common and highly heritable psychiatric disorder with a lifetime prevalence rate of approximately 4 cases per 1,000 individuals1. Despite its high heritability, it is also highly complex in its genetic architecture, with over 100 genetic loci associated with disease2.

As genomics technologies have advanced, many have been applied to better understand the genetic basis of psychiatric disorders. In particular, genome wide association studies (GWAS) have been an instrumental tool in assessing genomic variation in large populations of healthy and affected individuals towards the goal of identifying genomic markers for disease3.

Despite the power of GWAS and other related genomic approaches, several factors, including broad diversity in individual genomic variation, have impeded the identification of bona fide genomic makers for schizophrenia3,4. Moreover, despite the fact that genomics technologies have enabled efficient and comprehensive sequencing of patient and control genomes, straightforward sequencing data for identification of drivers of disease has proven challenging to deconvolute.

Recent efforts to better power GWAS analysis have shifted towards the integration of data from biological networks5-7. Indeed, networks are a powerful tool to not only visualize the connectivity within and between cellular processes, they can illustrate and enhance our understanding of key nodes given a particular cellular context. Taking advantage of protein and genetic interaction networks introduce the potential to reorganize large and scattered GWAS datasets to highlight critical nodes and node clusters relevant for driving disease. Towards this end, we have developed a method for network boosting of GWAS and applied our method on data from a previously collected cohort of 36,989 cases and 113,075 controls2. We have developed a composite metric for assessing significance of a given SNP on a gene, factoring in number of SNPs found per gene8. We have further taken advantage of publicly available, brain tissue-specific gene expression networks, in order to highlight the relevance of a particular gene in a network context9. Our method recapitulates previously known genes associated with schizophrenia and implicates novel genes in driving the pathophysiology of schizophrenia.

RESULTS

**Correlation between SNP count and SNP P-value**

SNPs were assigned to genes through a window-based approach, in which SNPs were assigned to genes if they fell within a 10 kilobase (kb) window up- or downstream of the gene’s start and end coordinates, respectively. We observe a Pearson correlation of -0.202 (p = 5.58 × 10-137) between the top SNP p-value and the number of captured SNPs per given gene (Fig. 1). Therefore, genes scores were calculated as the most significant p-value of assigned SNPs normalized by number of captured SNPs. Given that there are approximately 20,000 protein-coding genes under analysis, we derive a gene score threshold of p = 0.05 / 20,000. Initial analysis shows that only 22 genes pass this threshold.

**Personalized PageRank Identifies Novel Schizophrenia-Associated Genes**

Personalized PageRank was used to re-score the initial gene scores. Considering a personalization vector consisting of the initial gene node scores, the algorithm rescores the nodes according to visitation likelihood, as calculated from the input network topology as the probability of visiting each node according to a random walk of the graph either by an edge or a “teleportation” parameter *α*. These new scores are modeled by the following equation.

*A* is a matrix of transition probabilities, *x* is a vector containing the initial distribution of random walkers, and *E* is a personalization vector of initial information. This formulation was performed on the GIANT brain-tissue-specific network utilizing the networkx Python module. Here, we set *α = 0.85*, the default setting in the networkx Python module, and passed the initial gene scores as the personalization vector.

The GIANT brain-tissue-specific genetic network consists of 1,358,435 connections between 14,306 genes9. This version of the genetic network is filtered to only include edges with prior evidence supporting a tissue-specific functional interaction. We chose to use this network given its tissue specificity, given that schizophrenia is a psychiatric disorder.

After transformation, the number of hits increases dramatically to 1,583 genes which is greater than GWAS alone. The network-based rescoring method clearly boosts a majority of the signals (Fig. 2A, B). However, given a lack of perfect representation of all genes studied in the GIANT network, rescoring was not possible for every element, resulting in a noticeable split in the resulting output.

**Validation**

Several analyses validate the robustness of our approach. Using NetWAS, a similar network-guided approach to analyzing GWAS data, and applying the same brain-specific network, yields 84 significant genes9. Compared to our top 84 genes, there is a 38% overlap between these gene lists. Moreover, functional enrichment analysis of our gene list reveals enrichment of genes and gene sets known to play a role in schizophrenia and bipolar disorder as well as general neurodegenerative disease (Fig. 3A, B). We further observe significant KEGG terms including alcoholism, which has known similarities in neuropathology to schizophrenia, and serotonergic signaling, as well as terms related to a heightened immune response which are known to be involved in the pathogenesis of schizophrenia (Fig. 3B)10-13. We also recapitulate known miRNAs with previously described roles in the pathogenesis of schizophrenia, including mir-24-3p (Fig. 3C) 14,15.

Relationships between gender and schizophrenia suggest for a role for gonadal hormones in pathogenesis of disease. It is also known that estrogens participate as neural growth factors in many regions of the central nervous system16. Interestingly, we identify estrogen signaling as a KEGG term and estrogen receptors ESR1 and ESR2 in our dataset, which are a recent preclinical target for the treatment of schizophrenia17(Fig. 3B, D). Enrichment analysis of transcription factors in our dataset reveals a network of transcription factors with known roles in schizophrenia (Fig. 3D). In particular, iPS Transcription Factors, POU5F1 (also known as Oct4) and MYC are known to play a role in synaptic plasticity and also in schizophrenia18,19. Consistent with these findings, we see a significant KEGG term of signaling pathways regulating pluripotency of stem cells (Fig 3B).

Beyond our confirmation of known factors, we observe implication of genes with novel or limited association to schizophrenia. SMARCA4 is a known autism-associated chromatin regulator which is not surprising given that there have been shown to be common genetic mechanisms driving both disorders20{Zhang, 2016 #779}. In the context of schizophrenia, SMARCA2 has been well-characterized; however both genes are involved in synaptic remodeling and activity and are mutually exclusive catalytic subunits in the SWI/SNF complex21. In many cases, we see an expansion of significant genes in a given family. For example, 12 solute carrier (SLC) transport genes are present in the SZDB comprehensive gene list of schizophrenia associated genes; our gene list contains 39 family members22. Similarly there is a significant enrichment of histone cluster genes, expanded, from 2 to 33 genes with varying annotated associations with schizophrenia. Ultimately, our method is a robust and reliable way of boosting significance of genes identified via GWAS, recapitulating genes previously associated with schizophrenia while implicating novel genes.

DISCUSSION

Schizophrenia is a complex neuropsychiatric disorder that has broad physiological impacts from disturbing neural circuitry to impairing synaptic function1. Psychiatric disorders tend to be heritable; however schizophrenia in particular has been established as one of the most highly heritable psychiatric disorders1,23. Given the genetic foundation of schizophrenia, it comes to no surprise that high-throughput, unbiased genetic analyses have been the method of choice for better understanding the complex and genetically heterogeneous architecture of schizophrenia as well as genetic drivers of disease. Numerous studies have used GWAS to identify SNPs that may implicate genes that contribute to disease; however it has been well demonstrated that reproducibility and gene significance between independent GWAS analyses can be poor23. The challenge remains of understanding how hundreds of genes with weak associations interact to cumulatively influence the disease phenotype. The dramatic molecular and physiological neuropathologies driving schizophrenia point at altered cellular networks on several levels and highlight an opportunity to use network biology to better understand the molecular and genetic mechanisms of disease.

Networks of gene or protein-level interactions often illustrate functional associations between groups or pathways; coupled to large GWAS datasets, networks have the potential to reveal underappreciated enrichment of a given pathway or family of genes. This systems approach shifts the focus from pinning individual candidate genes to identifying pathways impacted by the compounding effect of a large set of genes which is critical given the known polygenic nature of schizophrenia. The significance of given genes from GWAS studies can further be used to weigh nodes on a network, and reveal the altered topology defining of the disease state. We applied this rationale in development of our method to better power GWAS analysis. Our method enables the prioritization of genes and therefore sub-networks by normalizing by the number of significant SNPs per gene and further rescoring genes using the Personalized PageRank heat diffusion algorithm.

Several factors can influence the predictions. Here, we have chosen the “teleportation” parameter *α* in the Personalized PageRank based on commonly set values; however this would have a large impact on the number of final significant genes. The choice of network can also certainly impact the final gene list, as differences in base network topology and connectivity will influence network propagation. A systematic assessment of the impact of brain and non-brain-specific networks will be interesting to pursue in the future. Further metrics, such as redefining how SNPs are mapped to genes, and integrating the predicted impact of SNPs on protein function could serve as an additional measure of significance when rescoring genes. As more networks are generated for the study of psychiatric disease, and the number of GWAS studies grow, we will face the challenge of combining numerous datasets, controlling for GWAS on different populations, and data analyzed by different methods or groups. Finding appropriate methods of handling growing data and refining networks will enable further interrogation of genes involved in schizophrenia and other polygenic diseases.

REFERENCES

1 McGrath, J., Saha, S., Chant, D. & Welham, J. Schizophrenia: a concise overview of incidence, prevalence, and mortality. *Epidemiol Rev* **30**, 67-76, doi:10.1093/epirev/mxn001 (2008).

2 Schizophrenia Working Group of the Psychiatric Genomics, C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-427, doi:10.1038/nature13595 (2014).

3 Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* **101**, 5-22, doi:10.1016/j.ajhg.2017.06.005 (2017).

4 Schizophrenia Psychiatric Genome-Wide Association Study, C. Genome-wide association study identifies five new schizophrenia loci. *Nat Genet* **43**, 969-976, doi:10.1038/ng.940 (2011).

5 Chang, S., Fang, K., Zhang, K. & Wang, J. Network-Based Analysis of Schizophrenia Genome-Wide Association Data to Detect the Joint Functional Association Signals. *PloS one* **10**, e0133404, doi:10.1371/journal.pone.0133404 (2015).

6 Leiserson, M. D., Eldridge, J. V., Ramachandran, S. & Raphael, B. J. Network analysis of GWAS data. *Curr Opin Genet Dev* **23**, 602-610, doi:10.1016/j.gde.2013.09.003 (2013).

7 Lee, I., Blom, U. M., Wang, P. I., Shim, J. E. & Marcotte, E. M. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome research* **21**, 1109-1121, doi:10.1101/gr.118992.110 (2011).

8 Lehne, B., Lewis, C. M. & Schlitt, T. From SNPs to genes: disease association at the gene level. *PloS one* **6**, e20133, doi:10.1371/journal.pone.0020133 (2011).

9 Greene, C. S. *et al.* Understanding multicellular function and disease with human tissue-specific networks. *Nat Genet* **47**, 569-576, doi:10.1038/ng.3259 (2015).

10 Wang, Q., Yang, C., Gelernter, J. & Zhao, H. Pervasive pleiotropy between psychiatric disorders and immune disorders revealed by integrative analysis of multiple GWAS. *Hum Genet* **134**, 1195-1209, doi:10.1007/s00439-015-1596-8 (2015).

11 Muller, N., Riedel, M., Gruber, R., Ackenheil, M. & Schwarz, M. J. The immune system and schizophrenia. An integrative view. *Annals of the New York Academy of Sciences* **917**, 456-467 (2000).

12 Batel, P. Addiction and schizophrenia. *Eur Psychiatry* **15**, 115-122 (2000).

13 Selvaraj, S., Arnone, D., Cappai, A. & Howes, O. Alterations in the serotonin system in schizophrenia: a systematic review and meta-analysis of postmortem and molecular imaging studies. *Neurosci Biobehav Rev* **45**, 233-245, doi:10.1016/j.neubiorev.2014.06.005 (2014).

14 Perkins, D. O. *et al.* microRNA expression in the prefrontal cortex of individuals with schizophrenia and schizoaffective disorder. *Genome Biol* **8**, R27, doi:10.1186/gb-2007-8-2-r27 (2007).

15 Beveridge, N. J. & Cairns, M. J. MicroRNA dysregulation in schizophrenia. *Neurobiol Dis* **46**, 263-271, doi:10.1016/j.nbd.2011.12.029 (2012).

16 Gogos, A. *et al.* A Role for Estrogen in Schizophrenia: Clinical and Preclinical Findings. *Int J Endocrinol* **2015**, 615356, doi:10.1155/2015/615356 (2015).

17 Martorell, L. *et al.* Analyses of variants located in estrogen metabolism genes (ESR1, ESR2, COMT and APOE) and schizophrenia. *Schizophr Res* **100**, 308-315, doi:10.1016/j.schres.2007.11.001 (2008).

18 Greenwood, T. A. *et al.* Genetic assessment of additional endophenotypes from the Consortium on the Genetics of Schizophrenia Family Study. *Schizophr Res* **170**, 30-40, doi:10.1016/j.schres.2015.11.008 (2016).

19 Greenwood, T. A. *et al.* Genome-wide linkage analyses of 12 endophenotypes for schizophrenia from the Consortium on the Genetics of Schizophrenia. *Am J Psychiatry* **170**, 521-532, doi:10.1176/appi.ajp.2012.12020186 (2013).

20 Zhang, Z. *et al.* Autism-Associated Chromatin Regulator Brg1/SmarcA4 Is Required for Synapse Development and Myocyte Enhancer Factor 2-Mediated Synapse Remodeling. *Mol Cell Biol* **36**, 70-83, doi:10.1128/MCB.00534-15 (2016).

21 Loe-Mie, Y. *et al.* SMARCA2 and other genome-wide supported schizophrenia-associated genes: regulation by REST/NRSF, network organization and primate-specific evolution. *Hum Mol Genet* **19**, 2841-2857, doi:10.1093/hmg/ddq184 (2010).

22 Wu, Y., Yao, Y. G. & Luo, X. J. SZDB: A Database for Schizophrenia Genetic Research. *Schizophr Bull* **43**, 459-471, doi:10.1093/schbul/sbw102 (2017).

23 Henriksen, M. G., Nordgaard, J. & Jansson, L. B. Genetics of Schizophrenia: Overview of Methods, Findings and Limitations. *Front Hum Neurosci* **11**, 322, doi:10.3389/fnhum.2017.00322 (2017).