Prioritizing gene-candidates associated with schizophrenia based on a summary score utilizing network properties of GWAS data.

Edahi González-Avalos, Elly Poretsky, Anya Umlauf.

Background / Introduction

Schizophrenia is a disease that affects 24 million people worldwide. It is a disruptive cognitive disorder, leaving many unable to lead a normal life, having friendships, or holding jobs [Bouwmans 2015]. People suffering from the disorder are at higher risk of homelessness, substance abuse, suicide and accidental death [Silverstein 2008, Lee 2015, Hellemose 2018]. Finding exact mechanisms for development of schizophrenia is therefore an important step in treating and preventing this brain disease. While multiple studies implicated that perinatal, developmental, and environmental factors play an important role in disease progression [Van Os 2009], the heritability of the disease is believed to be as high as 81% [Sullivan 2003]. Persons with a family history of schizophrenia are at higher risk of developing it themselves with risks ranging based on the degree of relationship to the affected family member [Piccioni 2007]. Thus, exploring genetic risks of schizophrenia is of at most importance in finding the cause.

The primary difficulty in establishing the genetic pathway to schizophrenia is the fact that the association is complex and non-Mendelian, implying that schizophrenia is a polygenic disorder [Harrison 2015]. Genome-wide association studies (GWAS) have identified up to 108 loci statistically associated with schizophrenia [Ripke 2014], but it is estimated that 8300 single nucleotide polymorphisms (SNPs) may contribute to the disease development [Ripke 2013]. And yet, the implication of specific genes in the disease progression remains uncertain. Some genes, identified in some earlier studies, failed to have significant associations with the schizophrenia in the follow-up studies [Harrison 2015].

One of the difficulties facing researches is the immensity of genome data and consequential necessity to control for false positive identifications of disease-related genes. Multiple approaches have been proposed and used to address these challenges. The common approaches range from simple, such as assuming a direct effect of genes based on their proximity to a significant SNP, to more evolved such as using Hi-C datasets that allow linking physically two stretches of DNA and therefore identifying associations between SNPs and regions leading to function assignment based on the proximity.

Molecular networks provide a thorough process of identifying genetic pathways to a disease, a process that attempts to combine various aspects of gene interactions and to maximize the collective information to improve predictions. And, it has been noted that larger molecular networks tend to perform better [Huang 2018].

In this exercise, we aimed to improve identification of schizophrenia-related genes through enrichment of information available about each gene. Our focus was not on improving the

assignment of SNPs to genes, but on the refinement of the candidate genes through network analysis.

Results

The initial data consisted of association statistics from a schizophrenia GWAS of up to 36,989 cases and 113,075 controls, in which 128 linkage-disequilibrium-independent SNPs were identified as genome-wide significant ($P \leq 5 \times 10^{-8}$) [Ripke 2014]. The summary statistics were available on 1252901 SNPs and 14965 genes. We initially prioritized gene candidates with p-values < 0.0001, which resulted in a list of 196 genes. This was followed with creation of an overall ranking score for the selected genes by summing 5 components. The individual components consisted of ranks based on gene attributes, including (1) gene-wise p-values from the original summary statistics data and (2) the average protein levels in brain sample tissues, as well as ranks based on network attributes, including (3) shortest path between the selected genes, (4) node centrality based on the PPI network, and (5) node centrality based on a tissue co-expression network. The ranks were summed and points were assigned to each in the reverse order so that higher values corresponded to lower (better) ranks. This ranking method is known as Borda count, a consensus-based ranking. We hypothesize that genes with the ranks closer to the top would have a higher overall probability of being related to the schizophrenia (higher Borda score).

Component (1): gene-wise p-values. Naturally, the gene-level p-values for association with schizophrenia are an important part of identifying correct genes. Undoubtedly, some of the significant p-values are likely to be false positive results even after correcting for multiple testing. This can simply be due to the gene's location and proximity to the true positive results or due to other inaccuracies in translating SNP significance to gene significance. Still, the large sample size of the original study, including a sizable case cohort, suggests that the set of significant results will include true positives. The top 100 genes in this significance set included some that are thought to play role in the physiology and development of neurons (CALN1), some that have been linked to neurodegenerative disease as Alzheimer's (APP) or other brain disorders (CNNM2, FEZ1, TCF4), but the role of others in development of schizophrenia and other brain-related abnormalities is less clear, if any. **Component (2): the average protein levels in brain sample tissues.** It seems natural that genes related to the schizophrenia would also be associated with brain-derived proteins, thus using the Human Protein Atlas (HPA) a ranking was devised to give more weight to genes with proteins related to the brain. It happens that the brain has one of the largest number of tissue-specific genes, with 315 of them classified as tissue-enriched genes [Begcevic 2016]. **Component (3): shortest path between the genes.** If a given gene has a great impact over other genes, the former should be relatively close to the latter; furthermore, if a given gene has a huge impact in a series of other genes, therefore the distance to these genes should also be relatively short. We used this to our advantage and calculated the 200x200 (or 2000x2000 in the second submission) shortest distance matrix between any pair of genes in our candidate gene list (either the starting 196 [rounded to 200] genes or the expanded 2000 top genes) and for each gene, the shortest distance (from said

genes to the other genes) was calculated and placed on its corresponding position on the matrix. Afterwards the row values were integrated to the borda analysis (ranked integration of a set of characteristics. One difficulty in determining the genetic mechanism of schizophrenia is that it's believed to be a polygenetic disorder and previous studies have identified many different genes from different regions of the DNA and with different functions. Thus, it can be beneficial to consider proximities of genes particularly those that have strong gene-wise associations with schizophrenia. In general, studying closer gene networks linked through shorter paths may lead to better understanding if they truly play a role in schizophrenia development and how exactly the path or combination of gene expressions may lead to that. The ranks based on shorter distances have yielded results that were not all surprising; the top 100 list included familiar players such as APP and some whose connection to the disease is less clear such as ZNF165 (although, the encoded protein linked to this gene is most commonly expressed in male-related tissues and males have higher odds of developing schizophrenia [Picchioni 2007]). **Component (4): node centrality based on the PPI network.**The PPI network is among most popular molecular systems used in genetic disease research [Jia 2014]. Consistently, we thought it would be important to take into account the interactions among proteins. **Component (5): node centrality based on a tissue co-expression network.** We considered ranking based on co-expression to be important. In addition to complexities of linking genetics to schizophrenia already mentioned, the exact function of many genes is still unknown. So even if they are implicated as statistically significant for the given disease, their exact role can be hard to establish. Co-expression networks are used as an attempt to overcome this lapse in information, which is particularly important for non-coding genes. Therefore, ranking based on the node centrality in a tissue co-expression network could greatly enhance information on node interactions and prove to be important for the overall ranking.

The entire list of 1000 top ranked genes is given in the supplemented materials. The top 100 genes include several that could be strong contenders.Those include genes with a known functions related to brain and neuronal development and functioning (e.g., HIST1H2BN, LRP8) and to neurodegenerative disease such as Alzheimer's (e.g., APP) or mental disorders (e.g., FXR1).

Surprisingly, the gene list provided for this analysis did not include some better known candidates in relation to schizophrenia, most notably dopamine D2 receptor (DRD2). This gene has a long history in psychiatric research, is a target of antipsychotic drugs, was reported previously to be implicated in schizophrenia based on GWAS.

Discussion

The debate over the best way of identifying genetic pathways for complex disease such as schizophrenia is likely will continue for some time. What can be said already is that simple approaches such as genome-wide p-values on their own are insufficient for such task. Disease-focused genetic search must take into account multiple attributes of individual genes as

well as of genetic networks. Molecular networks prove valuable in this field and network-assisted analyses become the standard.

In present analysis we attempted to improve GWAS ranking for schizophrenia by tapping the additional information about genes, gene functions, and gene interactions and combining all of that knowledge into a single measure (score) that can be used to rank the genes in their assumed order of importance to the disease.

Without having a gold standard to test our results against, it is hard to judge the accuracy of the predictions on our list of ranked gene candidates. But at least some of them, as discussed above, most likely are related to schizophrenia, thus it would be important to learn more about their functions, about the genes they interact with, and about the biology of paths and clusters they belong to.

One of the biggest limitations of this analysis is that, in combining the 5 components into a single ranking score, we have assigned equal weight to each component. This may not be the best strategy. For example network-wise components (3-5) may be more important than the gene-wise components (1-2), since they take into account complexities of gene interactions. On another hand simple ranking of such measures as shortest path may lead to increased bias due to ties likely occurring for many pairs of genes. A careful consideration of the weighing schema should be done using expert knowledge from both genetic networks and psychiatric disorders.

Inclusion of psychiatric expertise is important for more than one reason. It is likely that there is an overlap in genetic networks associated with schizophrenia and other psychiatric diseases such as bipolar disorder, but the overlap may not be complete. What specific mechanisms lead to divergent paths of the two disease? Furthermore, what genetic differences, if any, exist between a wide range of closely related psychotic disorders such as schizoaffective disorder, brief psychotic disorder, and others. These additional psychiatric populations should be studied in parallel with schizophrenia.

To identify the most related genes we chose to use ranking based on the Borda count method. There are, of course, other statistical methods for unsupervised learning such as principal components and clustering. While these methods are a powerful way of combining several sources of information into a format of reduced dimensions, and they are well recognized and widely used in variety of fields, including genetics, we felt that they would be less appropriate in this setting where a ranking was desired. Another drawback of these methods is that sometimes they yield results that are hard to interpret from the clinical perspective especially with minimal prior knowledge.

A potential improvement that can be done to the ranking approach we've taken is inclusion of additional components. For example, genes previously implicated in schizophrenia can be given are better ranking than those that have not. Again, careful considerations should go into

designing such attribute to make sure it increases the power of the analysis and not introducing too much additional noise that can be either due to literature bias or due to prior false positives.

It may also be important to consider demographic factors and their association with schizophrenia. It is important that the study cohorts are diverse, since schizophrenia effect is disproportional. Males are more likely to suffer from schizophrenia compared to females and it is not understood if the factors affecting this are genetic or environmental. There is also evidence that ethnic minorities may be at the higher risk of schizophrenia, although there is a strong indication that this difference is often driven by environmental and socioeconomic factors [Picchioni 2007]. Understanding the full complexity of interactions between genetic and environmental factors could improve our understanding of the disease's development.

Methods

The code used for this analysis is submitted as a supplemental material.

References

Begcevic I, Brinc D, Drabovich AP, Batruch I, Diamandis EP. Identification of brain-enriched proteins in the cerebrospinal fluid proteome by LC-MS/MS profiling and mining of the Human Protein Atlas. *Clinical Proteomics*, 2016;13:11. doi:10.1186/s12014-016-9111-3

Bouwmans C, de Sonneville C, Mulder CL, Hakkaart-van Roijen L. Employment and the associated impact on quality of life in people diagnosed with schizophrenia. *Neuropsychiatric Disease and Treatment*. 2015;11:2125-2142. doi:10.2147/NDT.S83546.

Harrison PJ. Recent genetic findings in schizophrenia and their therapeutic relevance. *Journal of Phychopharmacology*. 2015;29(2):85-96. doi:10.1177/0269881114553647

Hellemose LAA, Laursen TM, Larsen JT, Toender A. Accidental deaths among persons with schizophrenia: A nationeide population-based cohort study. *Schizophrenia Research* 2018 (In Print). doi: 10.1016/j.schres.2018.03.031

Huang JK, Carlin DE, Yu MK, Zhang W, Kreisberg JF, Tamayo P, Ideker T. Systematical evaluation of molecular networks for discovery of disease genes. Cell Systems. 2018 Apr;6:484-495. doi:10.1016/j.cels.2018.03.001

Jia P and Zhao Z. Network-assisted analysis to prioritize GWAS results: principles, methods and perspectives. *Human Genetics*. 2014 Feb;133(2):125-138. Doi: 10.1007/s00439-013-1377-1.

Lee H, Lee K, Koo JW, Park SC. Suicide in Patients with Schizophrenia: A Review on the Findings of Recent Studies. *Korean Journal of Schizophrenia Research.* 2015 Apr;18(1):59. doi:10.16946/kjsr.2015.18.1.5

Picchioni MM, Murray RM. Schizophrenia. *BMJ*. 2007:335:91-95. doi:10.1136/bmj.39227.616447.BE

Ripke et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature Genetics*. 2013 Oct;45(10):1150-1159.

Ripke et al. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*. 2014;511:421-427. doi:10.1038/nature13595

Silverstein SM, Bellack AS. A scientific agenda for the concept of recovery as it applies to schizophrenia. *Clinical Psychology Review.* 2008;28:1108–1124.

Sullivan PF, Kendler KS, Neale MC. Schizophrenia as a complex trait evidence from a meta-analysis of twin studies. *Archives of General Psychiatry.* 2003;60(12):1187-1192. doi:10.1001/archpsyc.60.12.1187

Van Os J & Kapur S. Schizophrenia. *The Lancet.*2009 Aug; 374(9690):635-645. doi: doi:10.1016/S0140-6736(09)60995-8

Funding:
None.