Title

Strategies for Network GWAS Evaluated Using Classroom Crowd Science.

Journal

Cell systems, 8(4)

ISSN

2405-4712

Authors

Fong, Samson H
Carlin, Daniel E
Ozturk, Kivilcim
et al.

Publication Date

2019-04-01

DOI

10.1016/j.cels.2019.03.013

Peer reviewed

HHS Public Access

Author manuscript

Cell Syst. Author manuscript; available in PMC 2019 September 27.

Published in final edited form as:

Cell Syst. 2019 April 24; 8(4): 275–280. doi:10.1016/j.cels.2019.03.013.

Strategies for Network GWAS Evaluated Using Classroom Crowd Science

Samson H. Fong[1,2], **Daniel E. Carlin**[1], **2018 UCSD Network Biology Class**[3], **Trey Ideker**[1,2,3,4,*]

[1]Department of Medicine, University of California San Diego, La Jolla, CA 92093, USA

[2]Department of Bioengineering, University of California San Diego, La Jolla, CA 92093, USA

[3]Program in Bioinformatics, University of California San Diego, La Jolla, CA 92093, USA

[4]Lead Contact

Abstract

Biological networks can substantially boost power to identify disease genes in genome-wide association studies. To explore different network GWAS methods, we challenged students of a UC San Diego graduate level bioinformatics course, Network Biology and Biomedicine, to explore and improve such algorithms during a four-week-long classroom competition. Here, we report the many creative solutions and share our experiences in conducting classroom crowd science as both a research and pedagogical tool.

Introduction

Genome-wide association studies (GWAS) analyze genotypes from a population of individuals to identify genetic variants associated with a disease or other phenotypic trait. Such variants implicate genes that are potentially causal for the phenotype, elucidating disease mechanisms and suggesting new routes for disease diagnosis or therapy. Thus far, GWAS has identified numerous loci relevant to a wide range of diseases such as coronary artery disease (Lu et al., 2012), diabetes (Zhao et al., 2010), and cancer (Chang et al., 2014). A long-standing challenge of GWAS is that the statistical power to find gene-disease associations can be greatly limited by the millions of genetic loci tested in a genome-wide study (Sham and Purcell 2014). This large number of loci requires that each individual locus test passes a very strict significance threshold (typically, $p < 5 \times 10^{-8}$), resulting in potentially many disease genes that are missed (Lander and Kruglyak 1995). In addition, causal variants may act in genetic epistasis with other variants, and the linear models typically used in GWAS do not capture these nonlinear interactions (Visscher et al., 2017). Moreover, subsequent interpretation of the significant variants is made difficult when the associated variants lie in non-coding regions of the genome, as is typically the case for common variants (Hou and Zhao 2013). These factors often lead to an incomplete set of candidate variants, obscuring the ability to infer the mechanistic basis of a trait.

*Correspondence: tideker@ucsd.edu.

Networks of known or suspected gene-gene and protein-protein interactions, captured in a growing community of molecular network databases, can be used to at least partially address these problems. For instance, the methods of genome-wide association boosting (GWAB) (Lee et al., 2011) and network-wide association studies (NetWAS) (Greene et al., 2015) both use networks to analyze GWAS summary statistics to prioritize the top genetic variants associated with a disease. GWAB aims to detect weakly implicated disease-associated genes by their proximity to other strongly implicated genes in a molecular network using a naive Bayes guilt-by-association approach. NetWAS first constructs functional tissue-specific networks from correlations in mRNA expression, where edges are weighted based on a tissue-specific Bayesian method. These edge weights are then used as features in a support vector machine (SVM) classifier for which positive and negative sets are chosen based on their association to a disease using GWAS data. In addition to molecular networks, gene sets provide another way to summarize higher order pathway effects. For instance, MAGENTA (meta-analysis gene-set enrichment of variant associations) detects enrichment of functionally related gene sets when performing a meta-analysis of multiple GWAS data collected for the same disease (Segrè et al., 2010). Conversely, predicted gene functions can help select which genes at an associated locus are causal, as demonstrated by DEPICT (data-driven expression prioritized integration for complex traits) (Pers et al., 2015).

Given this nascent field of network GWAS approaches, we considered that a very timely research activity would be to survey and synthesize the recent developments and to explore methodological alternatives. Here, we describe such a research project, conducted not in any individual laboratory but in the context of a university classroom.

The Class Competition

In spring 2018, we conducted a biological networks class at the University of California San Diego, entitled Network Biology and Biomedicine (BNFO286/MED283), intended to introduce network concepts to graduate students from various backgrounds. The course taught an understanding of the types, roles, and uses of networks in the biomedical sciences by covering theory and practice of network analysis, and it culminated in a final class competition that required students to apply network fundamentals to an ongoing area of biomedical research. In particular, students were presented with the general problem of deriving a ranked list of 100 disease-relevant genes from a schizophrenia GWAS (Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium, 2011). This GWAS consisted of 51,695 individuals; the original study had reported seven genomic loci associated with schizophrenia, five of which were novel at the time. Students were asked to self-assemble into small teams of up to three people to develop methods to improve a default workflow that utilizes networks to prioritize disease genes. This process led to creation of ten student teams in total, which we henceforth de-identify and reference as teams A–J (Table 1).

The default workflow involved three major steps: mapping variants to genes, selecting an appropriate network, and analyzing variants on the network (Figure 1). For each step, we either introduced a method or reminded students of relevant lecture topics while also making it clear that students should create and explore new alternatives. The first step of the default

workflow was to map genetic variants to genes. This step addressed a fundamental challenge in GWAS: how should the set of variants in and around a gene be used to assign a single score for prioritizing that gene's likelihood of association with phenotype? As the default, we presented a simple approach of assigning each genetic variant to the nearest coding gene based on chromosomal distance (Lee et al., 2011). In the second step, students needed to select the appropriate network for the analysis. The fundamental challenge posed by this step was to select a network enriched for interactions that are highly relevant to the disease context. As a starting point, we reminded the students of a list of publicly available gene and protein networks discussed in class as well as the many other networks available on NDEx (Pratt et al., 2015), an online repository of networks. In the final step, students had to decide on how best to apply their chosen network to analyze the gene scores derived in step 1. The challenge of this step was to select a method that best leverages network knowledge to boost the priority of causal disease genes. We presented two default options based on network propagation methods discussed in class: random walk and heat diffusion. Both methods had been used with success in previous genomic tasks such as stratifying cancer patients (Hofree et al., 2013), identifying cancer relevant pathways (Leiserson et al., 2015; Paull et al., 2013), and evaluating network quality (Huang et al., 2018) but to our knowledge had not been applied to analyze GWAS results. Students were also free to explore any existing or novel algorithms that utilize networks to derive a ranked list of schizophrenia disease genes. Once again, we presented this workflow as a starting point only and challenged students to improve upon each of the above steps as they saw fit.

Submitted Approaches

Given this default template (Figure 1), students creatively altered and extended the workflow to arrive at a range of submitted approaches, often choosing to experiment on a specific point in the workflow (Table 1). In what follows, we survey methodologies implemented for each step by the student teams, grouping similar methodologies where appropriate.

The first step of the workflow saw the least innovation, with most teams choosing the default method of assigning variants to genes based on nearest genomic (chromosomal) distance. Team H deviated from this method, instead choosing to incorporate physical distances measured by high-resolution three-dimensional DNA contact maps of chromatin (Won et al., 2016). In particular, genetic variants were mapped to the nearest gene within each topologically associated domain (TAD) identified by the contact map. TADs are segments of the genome that are highly enriched for DNA-DNA contacts within the segment; owing to this topological constraint, TADs are also thought to group variants with the genes they likely regulate. Any variants not associated with a TAD were mapped using the nearest genomic distance method.

For step 2, groups generally used one of two gene networks for their analyses (Table 1). These two networks were presented in early class lectures alongside a number of alternatives: the so-called parsimonious composite network (PCNet) (Huang et al., 2018) and the brain network from GIANT (genome-scale integrated analysis of gene networks in tissues) (Greene et al., 2015). PCNet is an amalgamation of 21 different networks where each edge is supported by more than 1 network. While this network is not tissue specific, it

had shown excellent performance in recovering literature gene sets of various diseases given only a subset of these disease genes, while being substantially smaller than most other publicly available networks (Huang et al., 2018). Other groups used the brain network from GIANT, arguing for the importance of tissue context in recovering disease-specific genes. PCNet and GIANT represent two very different approaches to network construction. On the one hand, PCNet is a composite of a broad range of networks, agnostic to cellular context, but it takes only those that are verified by multiple sources yielding a very compact network (19,781 nodes, 2.7 million edges, with density of 0.014). On the other hand, GIANT utilizes tissue-specific information across many different sources resulting in a substantially larger network (25,689 nodes, 42 million edges, with density of 0.126). Team F explored a third route, opting for a protein-protein interaction network and a co-expression network which were combined into a single interaction database.

In step 3, the network analysis step of the workflow, we saw the most diversity of approaches. Many teams (teams A, B, C, D, G, H and I) implemented and tested different variants of the two default network propagation methods, but teams explored different ways to place the initial gene scores on the network as well as to tune the restart probability parameter in random walk with restart. Three teams implemented network analysis methodologies very different from the above. Of these, team F decided to prioritize disease genes based on the sum of ranks across five different scoring schemes: gene-level p values, average protein levels in brain sample tissue derived from the Human Protein Atlas (Uhlén et al., 2015), network proximity to known GWAS genes, node centrality based on a protein-protein interaction network, and node centrality based on a tissue co-expression network. This consensus ranking method thus integrated multiple networks and measurements and prioritized genes relevant across different contexts.

Two teams implemented deep learning methods, based on recurrent neural networks (team E) or word embedding (team J). Team E used a recurrent neural network supervised to recover a hold-out set of significant genes implicated by the GWAS summary statistics. They argued that each iteration of random walk can be approximated using two fully connected layers. The input of the recurrent neural network is identical to that of random walk with restart, a vector of initial node weights and the network's adjacency matrix. The recurrent neural network attempts to create a more flexible model by allowing the model to learn a nonlinear representation of the input vectors and aggregate the information using a nonlinear function.

Team J devised a novel GWAS analysis using node embedding, network motifs, and network propagation. First, a node embedding algorithm (the skip-gram model) (Mikolov et al., 2013) is used to learn a vector representation of each network node. Features used for this learning are a random collection of paths through that node, the number of network motifs in which that node participates, and its gene score. Second, disease nodes are prioritized by random walk over a graph of node similarities in the embedded space. Notably, both Teams E and J incorporated the topology of networks and attempted to derive associations in nonlinear ways. These deep learning methods were not introduced in class, but due to the freedom provided by the competition, students were able to creatively explore a wide range of alternatives.

Evaluation Phase

We evaluated the performance of each team's algorithm in two ways, based on comparison of the resulting network-prioritized gene list to (1) a literature-curated schizophrenia gene set or (2) genes validated by additional GWAS studies. For both evaluation methods, ranked gene sets from all teams were evaluated relative to each other as well as to a baseline ranking derived from the discovery schizophrenia GWAS prior to network analysis (baseline or BL). For this baseline ranking, we mapped genetic variants from the discovery GWAS to genes using a ±10 kb window and ranked the genes in increasing order of p value. These baseline results represented the "null" method where the GWAS results were returned without any transformation. The computational code used to evaluate all results is available online (https://github.com/shfong/2018NetBioEval).

The first method of evaluation considered each method's ability to recover a previously documented set of 1,147 schizophrenia disease genes (Shim et al., 2017; Allen et al., 2008). We determined the number of these documented disease genes in each team's top 100 ranking and verified the statistical significance by a hypergeometric test. The winning entry according to this evaluation metric, by team A, employed random walk with restart to propagate gene association scores over the PCNet molecular network (Figure 2A). The team placed initial scores according to the negative log of the gene's assigned p value of association using the 10 kb window described above. This simple approach outperformed all other methods and discovered 33 documented schizophrenia genes in the top 100 genes reported overall (hypergeometric test $p < 10^{-27}$) (Figure 2A). This result represented a significant improvement over the baseline gene list, which recovered 9 schizophrenia genes (hypergeometric test $p < 0.06$). Notably, nearly all teams out-performed this baseline and yielded significant enrichment in their top 100 ranked gene lists (Figure 2A).

The second evaluation experiment tested the reproducibility of the results in comparison to another large schizophrenia GWAS published several years later (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). In addition, as schizophrenia and bipolar disorder are related diseases (Craddock et al., 2005), we expected that some genes linked to schizophrenia might also show linkage to bipolar disorder. Thus, we also tested how well each team's algorithm could discover significantly altered genes from the bipolar study. For these GWAS validation sets, we evaluated a team's algorithm by the number of genes in the top 100 ranking that were within 10 kb of a variant with significant association to the disease, with the significance threshold adjusted using Bonferroni correction for 100 hypotheses ($p < 5\times10^{-4}$).

Unlike the first task, where almost all teams performed better than baseline, no team was able to recover more significant associations from the validation schizophrenia GWAS than the baseline (71 genes). Teams I and C came very close, with 70 and 69 genes, respectively, both of which had used very similar methodologies to team A (Figure 2B). Most of the teams identified genes which also had significant associations to bipolar disorder. None of the teams were able to beat the baseline recovery of bipolar genes, however, although teams I, F, and E came very close to this baseline.

Based on the superior performance of team A in the first evaluation method, the team members were invited to work with the instructors to describe this winning method in a companion publication (Carlin et al., 2019).

General Conclusions across Methods

The evaluation results highlight several broad trends. First, random walk performs very well in prioritizing GWAS associations, as the top-performing methods in recovering literature curated schizophrenia genes and significantly associated GWAS genes both used a version of the algorithm. The random walk algorithm is simple to implement, and with only a single parameter, the restart probability, it performs well when considering the data in the form of summary statistics, perhaps because the simplicity of the approach matches that of the data. In contrast, deep learning methods performed less well. We speculate that the greater flexibility afforded by deep learning actually hurts performance here due to overtraining. In the future, and especially as network GWAS methods are applied at the level of an individual patient's variants, we expect that more complex interactions between variants will be captured by more expressive models.

Second, despite the good performance of random walk generally, it shows significant variance in performance, ranging from recovering 33 literature genes (team A) to 9 literature genes (team I). These methods differed in how each gene in the network is assigned an initial score. The winning team A scaled the initial score proportionally to the gene's significance by using a negative logarithm of the p value, while others binarized the significance using a threshold. The success of the former technique may owe to the fact that it transfers more information from the GWAS than the latter technique, by weighting more confident genes more heavily.

The baseline performed significantly better than network methods in recovering the validation GWAS; incorporating networks in this analysis thus appears to only degrade the signal (Figure 2B). One explanation is that the validation and discovery GWAS had some overlap in information; all three GWAS (the discovery and validation schizophrenia studies and the bipolar disorder study) used some of the same control individuals. Therefore, analysis of the second (validation) schizophrenia cohort was complicated by the overlapping control samples with the discovery set. Performance in recovering schizophrenia genes from the literature showed little correlation with ability to validate genes in the second schizophrenia GWAS (Spearman correlation of −0.28, $p = 0.40$) (Figure 2C). Rather, performance on the validation schizophrenia GWAS was far more correlated with the bipolar GWAS (Figure 2D), further highlighting the similarity between the two tasks (Spearman correlation of 0.91, $p = 0.00$), perhaps due to the use of the same control individuals.

Nevertheless, the additional case samples provided a valuable hold-out validation set to carry out a comparative evaluation of the methods. Notably, the top two teams in recovering significant associations in the validation GWAS, teams I and C, both thresholded the gene scores instead of using a quantitative transformation. However, in recovering schizophrenia genes from the literature, these two teams performed poorly, emerging among the bottom three teams of the class. This stark difference in performance caused by a seemingly small

change in input scores suggests that the manner in which these scores are handled can be quite consequential.

Finally, the high performance of network methods in the first evaluation task suggests that the underlying networks contain significant numbers of relevant interactions to schizophrenia. Genes implicated in schizophrenia tend to be well-studied, making interactions among these genes more likely to be identified in networks and subsequently discovered by network methods. Gene study bias may thus skew performance in favor of network methods. At the same time, network methods represent a systematic approach to incorporate the substantial prior knowledge present in multiple sources of evidence, including independent sample cohorts and gene expression.

Lessons Learned for Instructors

Since the fundamental idea of a class competition is to involve students in the scientific process, both instructors and students benefit from treating the competition with the same rigor needed to craft a typical scientific research project. This rigor includes baseline controls and quantifiable results to enable appropriate comparisons across different student methods. Due to the very strict time constraints of a class in comparison to an open-ended research study, the project topic needs to be limited in scope and the deliverables very specific. Students should also be encouraged to employ or develop algorithms that run quickly so they can be iteratively improved throughout the competition.

This time limitation creates extraordinary pressure on students to simultaneously learn the course materials and accomplish a large research project. Such pressure can be mitigated by providing students with a complete baseline model which they are required to improve over the course of the project. Use of a baseline model ensures (and assures) that the task can in fact be completed as planned, and it outlines the necessary data-processing steps. On the other hand, while a default model can be quite beneficial in orienting students to a challenge, it may limit the diversity of methods that the students produce. Such limitation can be offset by early checkpoints and active engagement with students to guide student teams toward novel solutions.

Finally, for classroom competitions involving GWAS, we note that some students may request information on each individual’s genetic variants, on the premise that GWAS data are best integrated with molecular networks patient-by-patient. However, patient-level data typically requires privileged access, which is not easy to provide in a classroom environment. This point extends beyond GWAS to any challenge involving clinical data.

Coda

In this class competition, we took an unorthodox approach to bioinformatics research in which we invited an entire classroom to collaboratively advance a common research topic. This process produced a wide range of different methods. Class competition prompts students to think critically about the course materials, encouraging them to gain mastery over key concepts they will need to best compete. Rather than accepting algorithms taught in class as absolutes, they begin to think about the relative strengths and weaknesses of these algorithms and are motivated to improve upon them. More generally, engaging many

different teams to explore a spectrum of methods to solve a common problem is the epitome of crowd science. Carefully studying how methods succeed and fail can yield surprising insights into algorithms in ways not readily achievable through more conventional means.

Consortia

The members of the 2018 UCSD Network Biology Class are Nadia Arang, Bokan Bao, Hunter Bennett, Xiaochun Cai, Kevin Chau, Bethany Fixsen, Edahi Gonzalez-Avalos, Alexander Hakansson, Vincent Hu, Arya Kaul, Irina Kufareva, Duong Nguyen, Elly Poretsky, Yue Qin, David Rideout, Isaac Shamie, Alex Sharp, Erica Silva, James Sorrentino, Anya Umlauf, Chao Zhang, and Jessica Zhou.

ACKNOWLEDGMENTS

We gratefully acknowledge support for these studies from the University of California as well as grants from the National Institutes of Health (R01HG009979, P41GM103504, U24CA184427).

REFERENCES

Allen NC, Bagade S, McQueen MB, Ioannidis JPA, Kavvoura FK, Khoury MJ, Tanzi RE, and Bertram L (2008). Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database. Nat. Genet. 40, 827–834. [PubMed: 18583979]

Carlin D, Fong SH, Qin Y, Jia I, Huang JK, Bao B, Zhang C, and Ideker T (2019). A fast and flexible framework for network assisted genomic association. iScience 15.

Chang CQ, Yesupriya A, Rowell JL, Pimentel CB, Clyne M, Gwinn M, Khoury MJ, Wulf A, and Schully SD (2014). A systematic review of cancer GWAS and candidate gene meta-analyses reveals limited overlap but similar effect sizes. Eur. J. Hum. Genet. 22, 402–408. [PubMed: 23881057]

Craddock N, O'Donovan MC, and Owen MJ (2005). The genetics of schizophrenia and bipolar disorder: dissecting psychosis. J. Med. Genet. 42, 193–204. [PubMed: 15744031]

Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, Zhang R, Hartmann BM, Zaslavsky E, Sealfon SC, et al. (2015). Understanding multicellular function and disease with human tissue-specific networks. Nat. Genet. 47, 569–576. [PubMed: 25915600]

Hofree M, Shen JP, Carter H, Gross A, and Ideker T (2013). Network-based stratification of tumor mutations. Nat. Methods 10, 1108–1115. [PubMed: 24037242]

Hou L, and Zhao H (2013). A review of post-GWAS prioritization approaches. Front. Genet. 4, 280. [PubMed: 24367376]

Huang JK, Carlin DE, Yu MK, Zhang W, Kreisberg JF, Tamayo P, and Ideker T (2018). Systematic Evaluation of Molecular Networks for Discovery of Disease Genes. Cell Syst. 6, 484–495.e5. [PubMed: 29605183]

Lander E, and Kruglyak L (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat. Genet. 11, 241–247. [PubMed: 7581446]

Lee I, Blom UM, Wang PI, Shim JE, and Marcotte EM (2011). Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res. 21, 1109–1121. [PubMed: 21536720]

Leiserson MDM, Vandin F, Wu H-T, Dobson JR, Eldridge JV, Thomas JL, Papoutsaki A, Kim Y, Niu B, McLellan M, et al. (2015). Pancancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat. Genet. 47, 106–114. [PubMed: 25501392]

Lu X, Wang L, Chen S, He L, Yang X, Shi Y, Cheng J, Zhang L, Gu CC, Huang J, et al.; Coronary ARtery DIsease Genome-Wide Replication And Meta-Analysis (CARDIoGRAM) Consortium (2012). Genome-wide association study in Han Chinese identifies four new susceptibility loci for coronary artery disease. Nat. Genet. 44, 890–894. [PubMed: 22751097]

Mikolov T, Sutskever I, Chen K, Corrado GS, and Dean J (2013). Distributed Representations of Words and Phrases and Their Compositionality In Advances in Neural Information Processing Systems, Volume 26, Burges CJC, Bottou L, Welling M, Ghahramani Z, and Weinberger KQ, eds. (Curran Associates, Inc), pp. 3111–3119.

Paull EO, Carlin DE, Niepel M, Sorger PK, Haussler D, and Stuart JM (2013). Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). Bioinformatics 29, 2757–2764. [PubMed: 23986566]

Pers TH, Karjalainen JM, Chan Y, Westra H-J, Wood AR, Yang J, Lui JC, Vedantam S, Gustafsson S, Esko T, et al.; Genetic Investigation of ANthropometric Traits (GIANT) Consortium (2015). Biological interpretation of genome-wide association studies using predicted gene functions. Nat. Commun. 6, 5890. [PubMed: 25597830]

Pratt D, Chen J, Welker D, Rivas R, Pillich R, Rynkov V, Ono K, Miello C, Hicks L, Szalma S, et al. (2015). NDEx, the Network Data Exchange. Cell Syst. 1, 302–305. [PubMed: 26594663]

Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium (2011). Genome-wide association study identifies five new schizophrenia loci. Nat. Genet. 43, 969–976. [PubMed: 21926974]

Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature 511, 421–427. [PubMed: 25056061]

Segrè AV, Groop L, Mootha VK, Daly MJ, and Altshuler D; DIAGRAM Consortium; MAGIC investigators (2010). Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. PLoS Genet. 6, e1001058, 10.1371/journal.pgen.1001058. [PubMed: 20714348]

Sham PC, and Purcell SM (2014). Statistical power and significance testing in large-scale genetic studies. Nat. Rev. Genet. 15, 335–346. [PubMed: 24739678]

Shim JE, Bang C, Yang S, Lee T, Hwang S, Kim CY, Singh-Blom UM, Marcotte EM, and Lee I (2017). GWAB: a web server for the network-based boosting of human genome-wide association data. Nucleic Acids Res. 45 (W1), W154–W161. [PubMed: 28449091]

Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A, et al. (2015). Proteomics. Tissue-based map of the human proteome. Science 347, 1260419. [PubMed: 25613900]

Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, and Yang J (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet. 101, 5–22. [PubMed: 28686856]

Won H, de la Torre-Ubieta L, Stein JL, Parikshak NN, Huang J, Opland CK, Gandal MJ, Sutton GJ, Hormozdiari F, Lu D, et al. (2016). Chromosome conformation elucidates regulatory relationships in developing human brain. Nature 538, 523–527. [PubMed: 27760116]

Zhao J, Bradfield JP, Zhang H, Annaiah K, Wang K, Kim CE, Glessner JT, Frackelton EC, Otieno FG, Doran J, et al. (2010). Examination of all type 2 diabetes GWAS loci reveals HHEX-IDE as a locus influencing pediatric BMI. Diabetes 59, 751–755. [PubMed: 19933996]

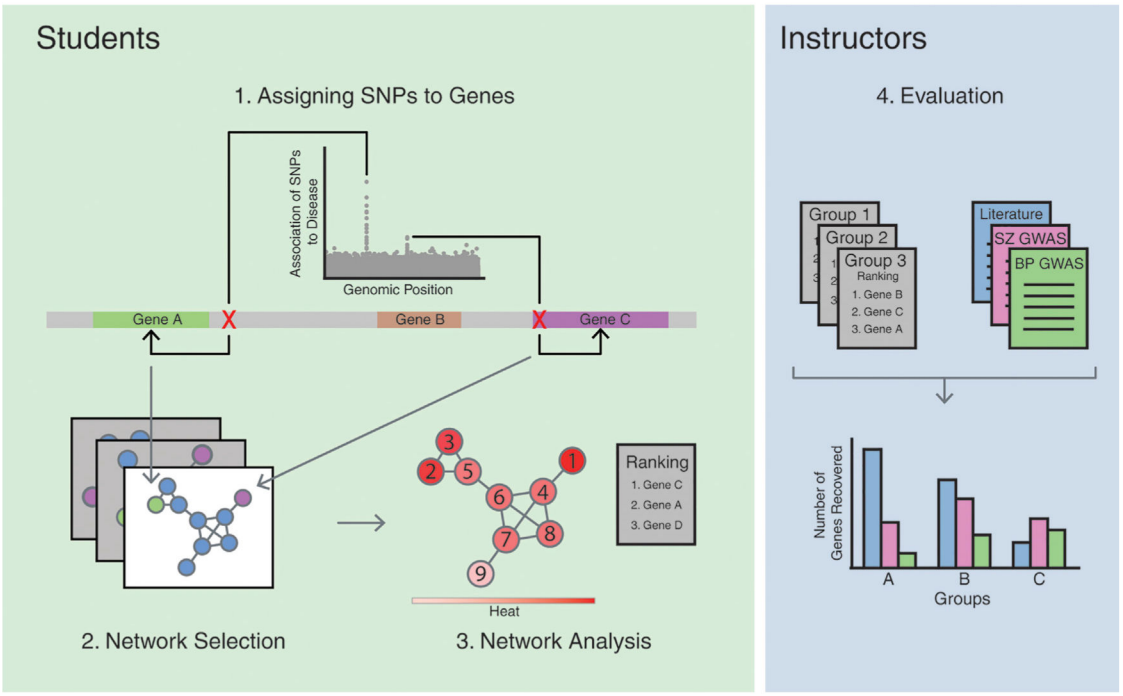

Figure 1. Workflow of the Network GWAS Challenge

This figure describes major steps in the class challenge. Student teams are responsible for developing the algorithm to analyze GWAS associations (first three steps) while the instructors evaluate these algorithms (final step), as follows: (1), GWAS summary statistics of associations between phenotype and genetic variants are assigned to genes. (2) Students select the appropriate network for the analysis. (3) Each gene's assigned associations are projected to scores on nodes in a molecular network and analyzed using any algorithm the students see fit. Each group produces a ranking of top 100 genes most associated with the target disease, schizophrenia. (4) Instructors evaluate these rankings in comparison to literature-derived schizophrenia genes (Literature), another schizophrenia GWAS acting as a validation set (SZ GWAS), and a bipolar GWAS (BP GWAS).

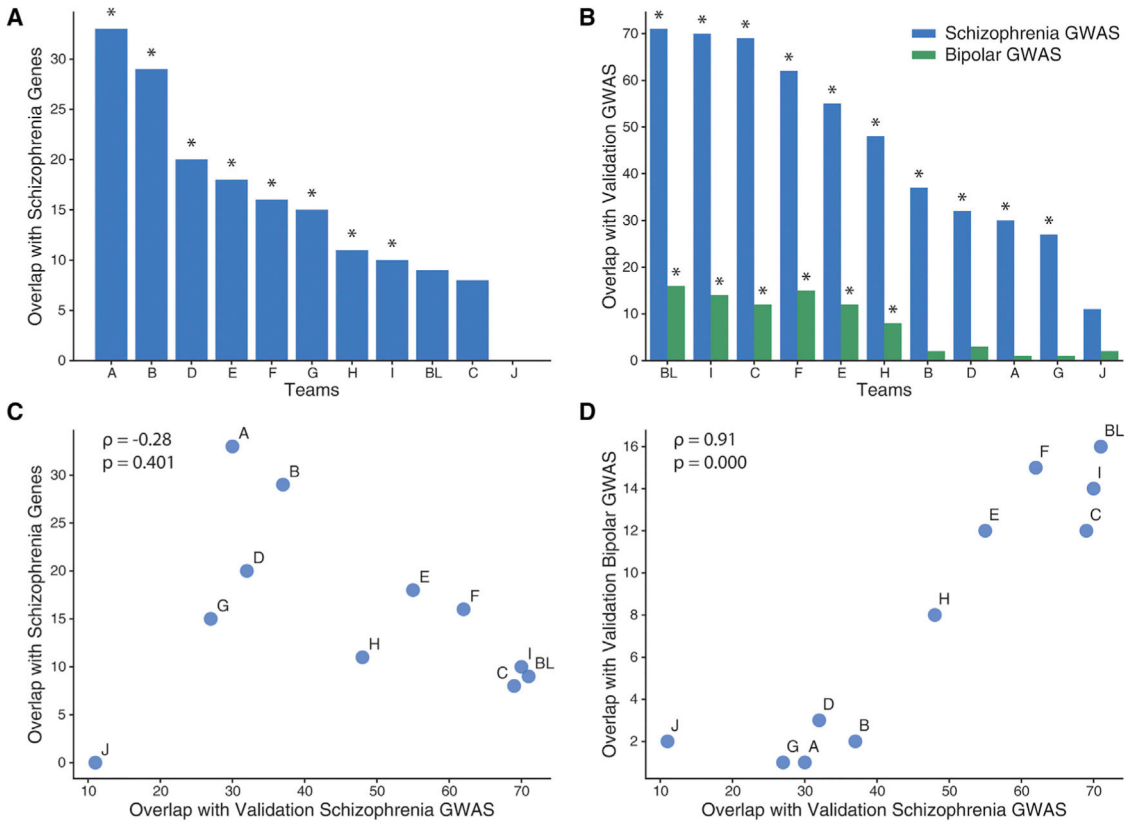

Figure 2. Comparison of Algorithm Performance

(A) The performance of each team (A–J) in recovering the schizophrenia literature curated gene set (1,147 genes total) among their top 100 prioritized genes, with significance evaluated using the hypergeometric test. Stars denote significant p values less than 0.05. BL refers to the baseline method which maps variant significance to the nearest coding gene within a 10kb window and ranks genes in increasing p value.

(B) Each team's performance in the validation schizophrenia and bipolar studies, again using the top 100 prioritized genes as in (A).

(C) Scatterplot of performance in GWAS validation (x axis) versus performance in recovery of literature-derived disease genes (y axis), shown for schizophrenia. The plot shows no significant correlation between these two validation tasks.

(D) Scatterplot of schizophrenia validation performance (x axis) versus bipolar disorder validation performance (y axis). Good performance on the schizophrenia GWAS correlates strongly with good performance on the bipolar GWAS (rho = 0.77, p value = 0.006).

Table 1.

Methodologies for Each Workflow Step Proposed by Class Teams

		Teams										
		A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**BL**[a]
SNP to gene assignments	Nearest genomic distance	x	x	x	x	x	x	x		x	x	x
	Nearest chromatin distance								x			
Networks	PCNet	x		x		x				x	x	
	Tissue Specific Network (GIANT)		x		x			x	x			
Propagation methods	Random walk with restart	x	x	x				x		x		
	Heat diffusion				x				x			
Deep learning	Recurrent neural network					x						
	Word embedding										x	
Other	Consensus network metrics							x				

[a]BL refers to the baseline method of using the untransformed GWAS results.