Bioinformatics Advance Access published August 19, 2016

Subject Section

STAMS: STRING-Assisted Module Search for Genome Wide Association Studies and Application to Autism

Sara Hillenmeyer[1], Lea K. Davis[2,3], Eric R. Gamazon[3,4], Edwin H. Cook[5], Nancy J. Cox[2,3], and Russ B. Altman[6,*]

[1]Biomedical Informatics Training Program, Stanford University, Stanford CA, [2]Vanderbilt Genetics Institute, Vanderbilt University, Nashville, TN, [3]Division of Genetic Medicine, Department of Medicine, Vanderbilt University, TN, [4]Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands, [5]Institute for Juvenile Research, Department of Psychiatry, University of Illinois at Chicago, Chicago, IL, [6]Departments of Bioengineering and Genetics, Stanford University, Stanford, CA

*To whom correspondence should be addressed.

Associate Editor: Prof. Alfonso Valencia

Abstract

Motivation: Analyzing genome wide association data in the context of biological pathways helps us understand how genetic variation influences phenotype and increases power to find associations. However, the utility of pathway-based analysis tools is hampered by undercuration and reliance on a distribution of signal across all of the genes in a pathway. Methods that combine genome wide association results with genetic networks to infer the key phenotype-modulating subnetworks combat these issues, but have primarily been limited to network definitions with yes/no labels for gene-gene interactions. A recent method (EW_dmGWAS) incorporates a biological network with weighted edge probability by requiring a secondary phenotype-specific expression dataset. In this paper, we combine an algorithm for weighted-edge module searching and a probabilistic interaction network in order to develop a method, STAMS, for recovering modules of genes with strong associations to the phenotype and probable biologic coherence. Our method builds on EW_dmGWAS but does not require a secondary expression dataset and performs better in six test cases.

Results: We show that our algorithm improves over EW_dmGWAS and standard gene-based analysis by measuring precision and recall of each method on separately identified associations. In the Wellcome Trust Rheumatoid Arthritis study, STAMS-identified modules were more enriched for separately identified associations than EW_dmGWAS (STAMS p-value 3.0x10-4; EW_dmGWAS- p-value=0.8). We demonstrate that the area under the Precision-Recall curve is 5.9 times higher with STAMS than EW_dmGWAS run on the Wellcome Trust Type 1 Diabetes data.

Availability: STAMS is implemented as an R package and is freely available at https://simtk.org/projects/stams.

Contact: rbaltman@stanford.edu

1 Introduction

Genome Wide Association Studies (GWAS) are chronically underpowered because they interrogate millions of positions in the genome. In order to overcome the multiple testing burden, such analyses require either very large cohorts or a reduced number of tests performed. One way to reduce the number of tests is to aggregate the genetic information from the Single Nucleotide Polymorphism (SNP) level to the gene level, reducing the number of tests from ~1,000,000 to ~20,000. However, in cases when it is too expensive or impossible to collect a large sample (e.g. a very rare phenotype), aggregation to the gene level may not be enough. Our group (Daneshjou *et al.*, 2014) and others (Compared in (Fehringer *et al.*, 2012)) have shown that aggregating GWAS to the pathway level is useful for analyzing underpowered GWAS and finding associations between groups of genes and a phenotype. These pathway methods provide insights into biological function.

The pathway approach has two major shortcomings. First, our present curation of genes into pathways covers only a small fraction of the genes that we measure. Second, most of the curated pathways are not designed to analyze GWAS. For instance, a metabolism pathway may contain mostly genes with no SNPs (due to stabilizing selection) but also one or two genes with significant variation. All of the genes are important to the metabolism, and so are considered part of the "pathway," but only a couple have GWAS signal. In cases like this, traditional pathway analysis cannot identify the entire pathway as significantly associated, because the noise dilutes the few genes with signal.

"Dense module searching", originally proposed for expression studies in 2002 (Ideker *et al.*, 2002), has become a popular analysis method for analyzing genome-wide measurements. By overlaying gene-based p-values on a graph of known biological interactions and identifying clusters of connected genes that have an enrichment of signal, dense module searching isolates clusters of concentrated signal within biologically related genes. These clusters range in size from two genes to complete pathway-level associations without requiring predefined pathway lists. Like traditional pathway analysis techniques, dense module searching leverages known biological relationships to aggregate measured information. Unlike traditional pathway analysis techniques, such as GSEA (Subramanian *et al.*, 2005; K. Wang *et al.*, 2007), or GO enrichment (Ashburner *et al.*, 2000; Beissbarth and Speed, 2004), dense module searching identifies clusters of biologically related genes that have the most enrichment for signal without requiring distribution of signal across all of the genes in a predefined list.

Several researchers have published improvements to Ideker's original algorithm. Nacu et al (Nacu *et al.*, 2007) proposed a different search technique, and made improvements to module scoring including control for multiple testing. Chuang et al (Chuang *et al.*, 2007) presented DMS, a method that uses a greedy search within a local neighborhood and incorporates three different kinds of significance testing to output modules. They demonstrate that significant modules in breast cancer gene expression data are better predictors of metastasis than individual markers. Based on DMS, Jia et al (Jia *et al.*, 2010; 2012) built dmGWAS, the first GWAS-specific R-based tool that allows users to easily incorporate dense module searching into their GWAS analysis workflow. dmGWAS uses a greedy search heuristic that iteratively adds nearby genes to a module if the p-value of the considered gene improves the aggregate score by an appreciable amount (r) and is described further in sections 2.1.4 and 2.1.5. However, all of these methods are limited to searching on a network with yes/no edge labels, and therefore prone to error due to false negative or false positive edges.

Recently published Edge-Weighted_dmGWAS (EW_dmGWAS) (Q. Wang *et al.*, 2015) allows users to input a co-expression matrix in order to weight edges in the input network before module searching. This is the first time that edge weights have been incorporated into dense module searching and allows the user to include information about a) the strength of evidence that two genes interact and/or b) the strength of the interaction itself. Wang et. al. demonstrate that the addition of edge weight information is valuable by comparing EW_dmGWAS performance on the GAIN schizophrenia GWAS to both standard dmGWAS (Jia *et al.*, 2010), and GiGa (Breitling *et al.*, 2004). Although phenotype-specific co-expression measurements may not be practical in every study, and many types of biological relationships are missed in co-expression analysis of eukaryotes, the expansion of the algorithm to incorporate edge weights improved performance of their algorithm, and creates an opportunity for exploring different kinds of gene-gene interaction weighting schemes.

The STRING database (Franceschini *et al.*, 2012) aggregates gene-gene and protein-protein interactions into a network, with edges scored for their confidence. STRING combines information from co-expression analysis, databases such as MIPS (Pagel *et al.*, 2005), high throughput experiments such as Chromatin Immunoprecipitation, phylogenetic co-occurrence, conservation of the genetic neighborhood (in prokaryotes), and literature co-occurrence. Associations are assigned a confidence score based on benchmarking groups of associations against KEGG (Kanehisa and Goto, 1999) pathways and generally correspond to the probability of finding the linked proteins within the same KEGG path. Although many other networks are available, STRING is notable because it consolidates many important interaction networks, has edge confidence scores based on a probabilistic framework, is very popular with more than 3,000 citations, and is easily usable with R.

In this paper, we present a method, STAMS, which incorporates general edge weights such as those provided by STRING, the search tool from EW_dmGWAS, and a new significance ranking method. STAMS allows users to use the edge-weighted searching of EW_dmGWAS without needing co-expression data. The resulting modules have high confidence biological relatedness and strong associations with phenotype. By leveraging the biological relationships in STRING, STAMS identifies modules that are more highly replicated in independent studies than standard analysis with a gene-based p-value or EW_dmGWAS with a phenotype-specific co-expression dataset.

2 Methods

2.1 STAMS

As shown in Figure 1, STAMS uses two kinds of input data: an interaction network with edge confidence scores, and gene-based p-values from a GWAS or other genome-wide experiment. These two data sources are integrated into a single graph with both node and edge weights. STAMS searches for high scoring modules using a greedy search, and then ranks the modules using a score normalization procedure.

2.1.1 STAMS: Underlying STRING network data

STAMS uses the human STRING database, version 9.1, and the STRINGdb R-plugin (Franceschini et al., 2012) as the underlying network. For each run of STAMS, we choose one of the eight edge types published by STRING. The confidence scores from each of the eight types are on the interval [0-1000], and generally correlate with the probability of finding two genes with that edge type within the same KEGG pathway. The shape of the distribution varies widely by edge type.

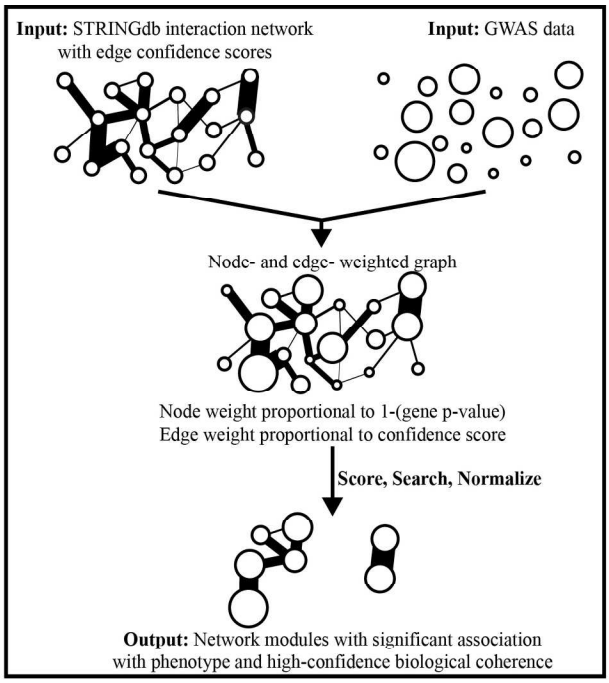

Fig. 1. The workflow overview of STAMS. STAMS overlays GWAS gene-based p-values on a graph of gene-gene interaction confidence scores from the STRING database. In the resulting graph, circular nodes represent genes with size proportional to 1-(p-value) of the gene's individual association with the phenotype. Graph edges are edges from the STRING database, weighted with confidence scores calculated by STRING. Using a search based on EW_dmGWAS, STAMS identifies modules of genes that have high biological coherence and an enrichment of GWAS signal.

(1) Neighborhood (31,550 edges): A confidence level based on how often the two genes are repeatedly in close neighborhood in prokaryotic genomes. The median confidence is 169; mean confidence is 181.

(2) Fusion (1,070 edges): A confidence level based on the number of times that the genes fuse in different species. Median confidence is 6; mean confidence is 77.5.

(3) Co-ocurrence: (11,654 edges): A confidence level based on the number of times the two proteins co-occur in different species. Median confidence is 152; mean confidence is 159.

(4) Coexpression (534,720 edges): A confidence level based on how much the genes are co-expressed across species. Median confidence is 185; mean confidence is 231.

(5) Experimental (197,709 edges): A confidence level based on protein interaction data from protein-protein interaction databases, e.g. MIPS. Median confidence is 266; mean confidence is 349.

(6) Database (126,374 edges): A confidence level based on protein interaction groups from curated databases, e.g. Reactome (Croft et al., 2010). Median confidence is 899; mean confidence is 867.

(7) Textmining (text) (1,150,456 edges): A confidence level based on protein interaction groups extracted from abstracts and open source full texts of scientific literature. Median confidence is 201; mean confidence is 244.

(8) Combined Score (CS) (1,684,531 edges): These scores are combined as if they are independent data sources:

$$Confidence(e) = 1000 * (1 - \prod_{i=1}^{7} (1 - c(e_i)/1000))$$

where $c_i(e)$ is the confidence level of the edge according to data type i. Median confidence is 229; mean confidence is 319.

2.1.2 STAMS specific: Edge weights

We transform the confidence levels from STRING into edge weights for the graph. To easily utilize the EW_dmGWAS search function, we defined edgeweight(e)= φ^{-1}(confidence(e) /1000) where φ^{-1} is the inverse CDF (quantile function) of the normal distribution, resulting in an approximately standard normal distribution of edge weights.

2.1.3 STAMS specific: Node weights

We transform the gene-level p-values into node weights. In order to make direct comparisons to EW_dmGWAS, and to directly apply the search function from EW_dmGWAS, we defined nodeweight as nodeweight(v) = φ^{-1}(1-p) where p denotes the gene-based p-value of the node v.

2.1.4 STAMS: Scoring

With the node weights and edge weights as described above, the searching and scoring functions from EW_dmGWAS were used without alteration. Briefly, the module score is defined by:

$$S = \lambda \frac{\sum_{e \in E} \text{edgeweight}(e)}{\sqrt{\text{No. of E}}} + (1-\lambda) \frac{\sum_{v \in V} \text{nodeweight}(v)}{\sqrt{\text{No. of V}}}$$

where E and V represent the edges and nodes of the module, and λ is a parameter between 0 and 1 that balances edge-weight and GWAS signals. We used the default permutation-based approach from EW_dmGWAS for choosing λ.

2.1.5 STAMS: Searching

We used the greedy searching function from EW_dmGWAS without alteration. Briefly:

(1) Assign a seed module M and calculate the module score S(M) of M. Initially, the seed module is a single gene.

(2) Examine all of the first-order neighbors of M, and identify the neighbor node N_{max} that generals the maximum increment of the module score.

(3) Add N_{max} to the current module M if the score increment is greater than S(M) x r, where r is a parameter that decides the magnitude of the increment. We used r = 0.1.

(4) Repeat steps 1-3 until no more neighbors can be added.

2.1.6 STAMS specific: Normalization of module score

For each module, we calculate a background distribution of 100,000 randomly generated modules by permuting the node weights. Lambda and the edge weights remain the same as the observed module. We calculate the mean μ and standard deviation σ of the 100,000 scores. For a candidate module with score S(M), the normalized score is $S_N(M)=(S(M)-\mu)/\sigma$. Modules are ranked by S_N. An empirical p-value is calculated as the fraction of permuted scores that meet or exceed the observed score.

2.2 Validation of STAMS

We validated STAMS on six GWAS from the Wellcome Trust Case Control Consortium (WTCCC) (Burton *et al.*, 2007), and the GAIN schizophrenia data set.

2.2.1 Calculating gene-level GWAS p-values

SNP p-values for six of the WTCCC GWAS data sets were calculated by using PLINK (Purcell *et al.*, 2007) to count alleles and then calculating p-values with built in chi-squared test from R. For the GAIN schizophrenia data, SNP p-values from European Americans were downloaded from dbGAP. Gene-based p-values for all seven data sets were calculated with VEGAS (Liu *et al.*, 2010) using default settings. VEGAS p-values are simulation-based, so in cases where VEGAS reported p=0, we substituted p=1/(number of simulations) which may overestimate some p-values. The p-values from VEGAS were approximately uniform on the interval (0,1].

2.2.2 Replication in subsequent studies using *knowngenes* list

To demonstrate that STAMS identifies genes that have true biological associations with the phenotype, we tested whether the identified modules were enriched for genes with independently identified associations. For each phenotype, we downloaded all of the reported phenotype-specific gene associations in the genome.gov GWAS catalog, except those found in the discovery datasets, into a list, denoted as *knowngenes*. The catalog contains hits with p-value < 1.0×10^{-5} in the initial + replication population from English language publications of new GWAS data measuring at least 100,000 SNPs. Fisher's exact test was used to measure enrichment of the genes in the top 1% of STAMS-identified modules for genes on the *knowngenes* list. We calculated Precision (true positives ÷ number of STAMS-identified genes) and Recall (true positives ÷ number of *knowngenes*) of STAMS alongside a standard VEGAS + Bonferroni and VEGAS + FDR analyses with corrected p-value ≤0.05.

2.2.3 Comparison of STAMS and EW_dmGWAS on six WTCCC datasets and the GAIN schizophrenia data set

In order to demonstrate the performance improvements in STAMS over EW_dmGWAS, we used both methods to analyze seven datasets. We performed standard EW_dmGWAS of (Q. Wang *et al.*, 2015), software version 3.0, dated 10/4/2014) with phenotype-specific expression datasets for type 1 diabetes (T1D), bipolar disorder (BD), rheumatoid arthritis (RA), coronary artery disease (CAD), type 2 diabetes (T2D), hypertension (HT), and the GAIN schizophrenia data (SCZ). See Supplemental Methods.

2.2.4 Precision and Recall of STAMS and EW_dmGWAS in reduced datasets

To demonstrate the performance of STAMS with under-powered datasets, we created datasets using the full control samples and varied the percentage of case patients by randomly selecting 20-100% of the cases each time in 20% increments (non-progressively). We compared the Precision/Recall curves of STAMS performed on these lower-powered datasets to the curve generated by using the full data set. The gold standard in this analysis was the list of genes that met genome-wide significance in the full dataset.

2.3 Application of STAMS to autism fGWAS

We analyzed an autism GWAS with STAMS. The cohort consisted of 654 probands from the Autism Genetic Resource Exchange (AGRE) and 1,593 unselected controls from the iControl data set, both genotyped on the Illumina HumanHap550v3. (The data selection criteria are described further in (Geschwind *et al.*, 2001)). The data (AGRE/iControl) were restricted to 1,945 individuals of western European decent based on results from multidimensional scaling conducted using PLINK. SNPs with Minor Allele Frequency (MAF) < 0.01, Hardy-Weinberg Equilibrium (HWE) p-values < 0.001, genotype missingness > 0.02, and differential missingness between cases and controls > 0.01 were excluded. After these quality control measures, 511,483 SNPs remained. We imputed with HapMAP3 R2 Build 36 and MACHv1, resulting in a total of 1.38M SNPs.

For each SNP, we did a logistic regression with the genotype and the top four principal components as covariates. We constructed gene-level annotations that included expression quantitative trait loci (eQTLs) previously identified in the parietal cortex (GSE35978) and the cerebellum (GSE35978) as well as coding variants (nonsense, frameshift, and missense) within the gene. The p-values for each SNP from its regression were filtered based on their function—eQTLs, nonsense, missense, and frameshift SNPs were included and all other SNPs were dropped from the model. For each gene, the test statistic Y was calculated as:

$$Y = \sum_{i=1}^{L} \log(p_i)$$

where L is number of SNPs that are annotated to the gene. To control for Linkage Disequilibrium, we permuted the phenotype labels in order to generate an empirical null distribution for the test statistic. Empirical gene-level p-values were then obtained by calculating the proportion of test statistics from the null distribution that met or exceeded the test statistic from the associated set of results. These p-values were the input to the STAMS algorithm.

3 Results

Figure 2 demonstrates that STAMS identifies modules with better enrichment for *knowngenes* than EW_dmGWAS. We analyzed six Wellcome Trust datasets and the GAIN schizophrenia dataset with STAMS and EW_dmGWAS. Figure 2 compares the negative log of the p-value for Fisher's exact tests of modules for enrichment with independently identified *knowngenes* for the six phenotypes with signal. Neither STAMS nor EW_dmGWAS identified modules with enrichment for *knowngenes* in the HT dataset. Since Wang et. al. (Q. Wang *et al.*, 2015) suggest using the top 1% of modules returned by EW_dmGWAS for further study, we compared the top 1% of modules returned by each method. In all six phenotypes with signal, the Textmining edge set with STAMS gave the best enrichment, followed by the Combined Score edge set with STAMS. The other six edge modalities in STRING did not perform as well, and are compared in Figure 5. We also ran EW_dmGWAS using the PINA interaction network on the schizophrenia data as described in Wang et. al., and the results were substantially consistent with the EW_dmGWAS + STRING results presented here.

Figure 3 plots precision and recall of STAMS with the Textmining edge set against *knowngenes* alongside EW_dmGWAS, a standard gene-based test (VEGAS) with Bonferroni correction, and VEGAS with an FDR correction. STAMS and EW_dmGWAS parameters (number of considered modules) were set such that their precision roughly matches

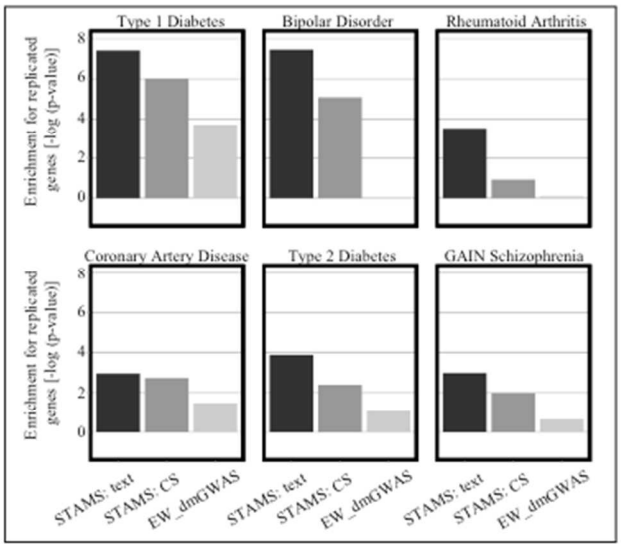

Fig. 2. Performance comparison of STAMS with EW_dmGWAS for six phenotypes. We pooled the genes in the top 1% of modules from each analysis and measured enrichment for genes reported in independent GWAS for each phenotype (*knowngenes*) using Fisher's exact test. We plot the $-\log_{10}$ of the p-value so that taller bars indicate better enrichment.

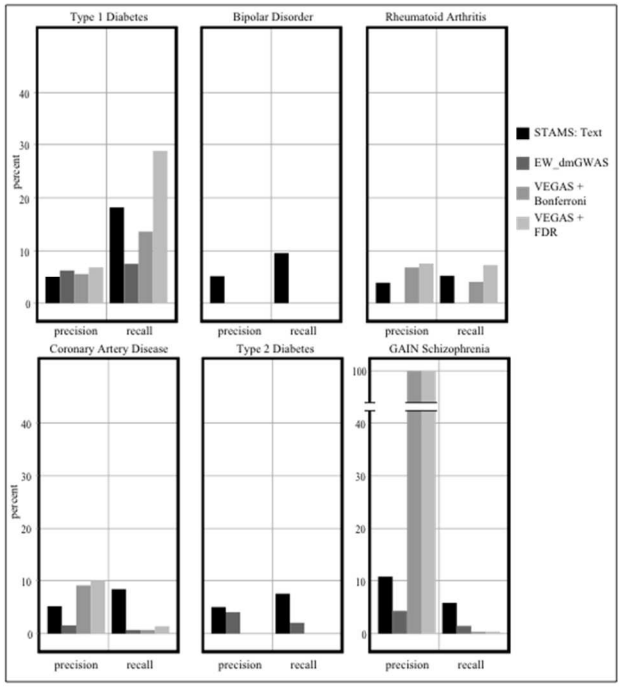

Fig. 3. Comparison of precision and recall of STAMS, EW_dmGWAS, and standard gene-based methods. After selecting a number of considered modules in order to approximately match the precision of the standard analysis in RA and T1D, we plot the precision and recall of each method on predicting genes in the *knowngenes* list. Standard gene-based p-values were calculated using VEGAS and corrected using Bonferroni and FDR.

that of the gene-based tests of T1D. STAMS is plotted with 75 top modules; EW_dmGWAS is plotted with 25 top modules. STAMS with Textmining edges has universally better performance than EW_dmGWAS, and in two phenotypes has better performance than standard gene-based analyses. We present the Precision/Recall for a corrected p-value of 0.05 because it is a commonly used cutoff in the literature. See Supplemental Figure 2 for the full Precision/Recall curves.

In BD with 75 modules, STAMS identified 12 *knowngenes* that were not individually significant. These genes include two small clusters of STRING-connected genes (GLT8D1, SPCS1, NDUFAB1; ITIH3, ITIH1, HNRNPC, NEK4, CACNA1C, PBRM1). The remaining genes were not connected to other *knowngenes*. In T2D, STAMS identified 11 *knowngenes* that were not individually significant. Nine of them were connected to each other in STRING (ZFAND6, ZBED3, IGF2BP2, HHEX, FTO, TSPAN8, CSKN2B, CDKAL1, HNF1B).

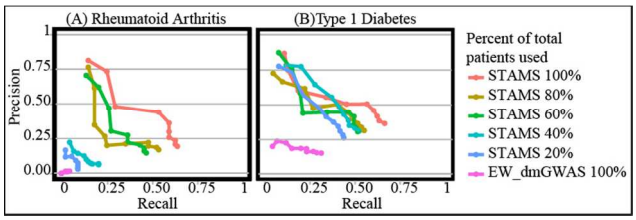

Fig. 4. Precision-Recall curves of STAMS performance across varying numbers of modules and sample sizes. We used the list of genes identified by VEGAS with a Bonferroni correction in the full dataset as a gold standard and calculated precision-recall curves for our ability to recover these genes with a reduced sample size by varying the number of top modules considered. We randomly selected subpopulations of patients to include and show that performance of STAMS with Combined Score edges decreases when fewer patients are included. We also ran EW_dmGWAS with expression data for edge weights, and show its performance. (A) shows results for Rheumatoid Arthritis (RA) and (B) shows results for Type 1 Diabetes (T1D).

Figure 4 demonstrates that STAMS with Textmining edges recovers more of the genome-wide significant genes in the original dataset than EW_dmGWAS, even when using only 20% of the patient's data. It also shows that STAMS performance is maintained in a reduced sample size. Performance of STAMS on T1D is better than on RA due to the strong edge weights between genes in the HLA region, many of which have known T1D associations.

Figure 5 shows the variability in performance of STAMS across six WTCCC disorders and different types of edge sets from STRING. The Experimental, Database, Textmining, and Combined Score edge sets are heavily enriched for *knowngenes* while the Fusion edges, which are based on the number of times that two genes fuse across species, are not. Many edge types perform well on T1D, while none perform well for Hypertension.

Figure 6 shows a top-scoring module from the AGRE autism fGWAS. The module includes an individually *non*-significant gene, CTTNBP2 (gene-based p-value=0.2) with known rare-variant autism associations. The other genes in this module are part of the STRIPAK complex. CTTNBP2 and STRIPAK interact to regulate dendritic spinogenesis. This module was chosen for discussion because it is the best scoring module (rank = 91) that contains a gene from the high confidence rare-variant associations in (DeRubeis et al 2014). Its empirical p-value based on 1 million node-label permutations is 0.001905.

Normalization improves module selection

The normalization function described in Methods 2.1.6 improves module detection. As shown in Supplemental Figure 1, the STAMS normalization function identifies modules that are more enriched for *knowngenes* than either the EW_dmGWAS normalization function, or ranking the output modules by the raw scores.

Fig 5. STAMS performance varies over different edge data sources and disorders. We ran STAMS on the subsets of edges in STRING that were curated in each edge-set modality as summarized in Methods. We pooled the genes from the top 1% of modules returned and measured enrichment for *knowngenes* with Fisher's exact test. The $-\log_{10}$ of the p-value of is plotted.

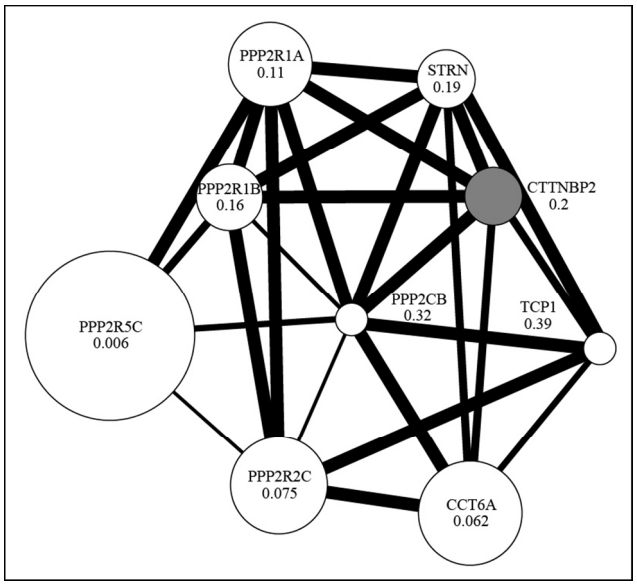

Fig 6. STAMS-identified module from autism GWAS. A high-scoring autism module from AGRE fGWAS is plotted with input gene-based p-values listed in the nodes. Line width corresponds to Combined Score edge confidence, but are all very high confidence (z-scores range from 1.43 to 3.09). The module contains CTTNBP2; rare loss of function mutations in CTTNBP2 have been associated with autism. The other genes in the module are members of the STRIPAK complex. CTTNBP2 interacts with STRIPAK to regulate dendritic spinogensis, a proposed mechanism for autism.

4 Discussion

In this paper, we show that using the multimodal edge weights from the STRING database improves the performance of EW_dmGWAS while eliminating the need for a secondary phenotype-specific expression dataset. We demonstrate this improved performance in several ways. We compare STAMS performance to EW_dmGWAS on two gold standards: genes that were identified in independent studies (*knowngenes*) and genes that were individually significant in the full dataset. We show that STAMS demonstrates high precision and recall on *knowngenes* and recovers more of the disease genes even with only 20% of the patients as input data.

The results from all GWAS analysis techniques are dataset dependent. We presented results on seven datasets in order to demonstrate the range of STAMS performance across varying levels of signal. In all datasets (excluding low-signal Hypertension), STAMS performs better than EW_dmGWAS. In two out of five WTCCC GWAS, STAMS has notably better performance than a gene-based test with a Bonferroni correction. For T1D, the recall of STAMS is better than the recall of VEGAS with a Bonferroni correction, but lower than recall of VEGAS with an FDR correction. In Bipolar Disorder and Type 2 Diabetes, STAMS has dramatically better precision and recall than either gene-based test. In RA and Coronary Artery Disease, the performance of STAMS and the gene-based tests are comparable.

In BD and T2D, where STAMS identified genes that standard analysis techniques missed, the genes were found clustered together. The improvement of recall by STAMS shows that the addition of interaction data identifies genes that would otherwise be missed in an underpowered study, even though they are true associations. The improvement of precision demonstrates that the genes that belong to clusters of biologically related high-scoring genes are more likely to be validated in independent studies than genes that have individual genome-wide significance in a single study.

When we examine STAMS and EW_dmGWAS precision and recall on the genes that were individually genome-wide significant in the full dataset, we predictably see that both methods fail to recover some of the individually significant hits. Both methods rely on strong gene-gene interactions to identify modules, so neither method detects genes that have strong individual signal but no neighbors with good p-values, which are detectable by single-gene based tests. However, STAMS maintains nearly full levels of precision and recall even with fewer patients included. As the p-values for genes (based on patient data) become less significant, the edge weights from STRING remain constant, stabilizing STAMS performance in the reduced power datasets. In Type 1 Diabetes, the strong edge weights between members of the HLA region allows STAMS to perform well on only 20% of the patient data. Performance of STAMS on the reduced-power datasets from Rheumatoid Arthritis shows nearly full precision and recall using 60% (n=1116 cases) or more of the patient data. These data indicate the utility of STAMS to rescue important gene associations that would otherwise be missed due to lack of power. Finding ways to increase power without requiring larger sample sizes is mandatory in some cases. For instance, in some rare drug events and populations, researchers have collected every known example, and are still underpowered to find associations. By demonstrating that STAMS has high precision and recall of significant genes even as less and less input data are used, we have demonstrated its utility for identifying candidate genes in a small GWAS.

When comparing the performance of STAMS using different types of edges, we find that edges created from Textmining give the best performance on the *knowngenes* in four out of the six WTCCC phenotypes. However, STAMS on the Combined Score edges outperforms the Textmining edges on Precision/Recall of individually significant genes (Combined Score shown in Figure 4, Textmining data not shown). We suspect that this is due to the greater number of edges in the Combined Score set than the Textmining set, which allows more genes to be identified. Unlike the *knowngenes* validation list, the individually significant genes may or may not be replicated in other studies. Many of the edge modalities (Fusion, Co-expression, Co-occurance, Neighborhood) had

poor performance across phenotypes. We were not surprised to learn that these edges, based on prokaryotic features or cross-species features, were less useful than the edge-sets created from eukaryotic protein-protein interaction databases (Experimental, Database). We tried combining some of the top-performing edge modalities into new edge scores, but the results were not better than Textmining alone. We recommend that researchers use the Textmining and/or Combined Score edge modalities.

Notably, the STAMS performance correlates with number of edges included in a modality. As shown in Figure 5, the best performing edge modalities are the ones that contain the most edges. In Supplemental Figure 4, we illustrate STAMS performance as edges are randomly dropped or added to the input graph. Although performance drops as edges are removed, STAMS performance is not as sensitive to the added edges, indicating that it is robust to false-positive gene-gene links.

The utility of the Textmining based edges to find replicable results demonstrates the power of Natural Language Processing (NLP) for curating biomedical knowledge. The edges currently in STRING are based on two genes or proteins co-occurring in PubMed abstracts or open- source articles. As NLP techniques improve to include sentence parsing and modeling of more complicated gene-gene relationships, this edge set will get less noisy and identifying modules based on it will improve.

We should acknowledge that the publications mined could include those about the GWAS that we used for validation or candidate gene studies based on the initial WTCCC results. These papers may have listed the GWAS hits or candidate genes listed in the abstract, thus creating links between the hits in the STRING database. With those links, the modules containing the GWAS hits are easier to find. However, given the large corpus of PubMed abstracts, and the normalization calculations in STRING, we suspect that the contamination by publications about our validation set is very small.

We improved on the significance testing and module-ranking algorithm of EW_dmGWAS by considering permutations of node weights over a fixed network topology. With randomly selected network topology and connectivity used in permutations as EW_dmGWAS does, the background scores are almost certainly underestimated since they may not have connecting edges at all. By keeping the edges fixed, we focus on whether the weights of the nodes in the module are significant. Although we ranked the modules differently, we agree with the authors of EW_dmGWAS that selecting the top 1% of modules for further investigation, although seemingly arbitrary, works well in practice. This was demonstrated by our validation against independently associated GWAS hits. However, because there are several ways to measure significance for dense modules, we do not rely heavily on the calculated p-value. With only summary statistics rather than genotype data, the p-values calculated in neither STAMS nor EW_dmGWAS measure how associated the module is with the phenotype. Instead, the current implementations measure how densely clustered the high-scoring genes are in the network. In order to get a more complete picture of significance, one would have to permute the case/control labels on the GWAS data and redo the searching under those permutations to create a null distribution. We suggest that users use the p-value to rank modules and identify candidate modules for follow up study through permutation of case/control labels, or orthogonal biological methods.

In our application to autism, we show that analyzing a common-variant GWAS with STAMS identifies a module with a known rare-variant association, demonstrating the power of STAMS to converge on true biological associations. Rare loss of function mutations in CTTNBP2, also known as CORTBP2, have been associated with autism (Cheung *et al.*, 2001; Sanders *et al.*, 2015) but have not been reported in common-variant autism GWAS to date. The module also has overlapping genes with other gene groups presented in the autism literature. Six of the genes overlap with neocortical development coexpression modules from Parikshak et. al. (PPP2R5C, PPP2CB are M4; CTTNBP2 is M2; TCP1 and CCT6A are in M14; PPP2R2C is in M16) (Parikshak *et al.*, 2013). Additionally, PP2R1A was also classified by Hormozdiari et. al. and as being part of two modules which are enriched for de novo mRNA mutations in autistic probands, Epilepsy and "M3: Extended autism spectrum disorder and intellectual disability" (Hormozdiari *et al.*, 2010). None of the genes in the module overlap with the copy number variation modules identified by the NETBAG method of Gilman et. al. (Gilman *et al.*, 2011).

Many of the edges in the identified autism module come from the STRIPAK complex described by Goudreault et al (Goudreault *et al.*, 2008). They used an iterative affinity purification/mass spectrometry approach to characterize the diversity of protein phosphotase 2A (PP2A) complexes. The protein products of PPP2R1A, PPP2R1B, STRN, PPP2R5C, PPP2CB, PPP2R2C, CCT6A, and TCP1 are all components of the STRIPAK assembly. STRIPAK interacts with CTTNBP2, which targets the assembly to dendritic spines and where it regulates dendritic spinogenesis. Since CTTNBP2's association with autism may be explained by problems in the regulation of dendritic spinogenesis (Chen *et al.*, 2012), it follows that mutations in members of the STRIPAK complex, which is also part of this process, may be associated with autism through the same mechanism. STAMS reveals meaningful association of the STRIPAK complex and this possible mechanism for autism by analyzing the GWAS data within its biological context.

STAMS is limited by the underlying search mechanisms presented in EW_dmGWAS. Notably, the search is entirely greedy, which means that genes with small contributions to the score have no chance of being added to a module, even if they provide a link to a higher-scoring gene. Several other dense module-searching methods exist which incorporate non-greedy searching (Nacu *et al.*, 2007; Ideker *et al.*, 2002; Q. Wang *et al.*, 2015) and the expansion of a non-greedy search method to include edge weighting would likely outperform STAMS in some cases.

STAMS, like other gene-based methods, is also limited by including only SNPs that are mapped to genes. In our application to autism, we incorporated SNPs that are known eQTLs for genes into the gene-based score, but there are still many measured SNPs for which data are ignored. Expanding gene definitions to include more intragenic SNPs may help. However, some individual SNPs and genes will remain undiscoverable by STAMS, so we recommend running STAMS as a complement to SNP-based and gene-based analysis techniques.

Our validation method, which uses genes from independent GWAS that have SNP that meets a cutoff of p-value $< 1.0 \times 10^{-5}$ in the initial + replication population is more relaxed than most genome-wide SNP significance criteria. While 1.0×10^{-5} is a reasonable cutoff for a gene-based test, the results might change with a stricter validation list based on a p-value $< 1.0 \times 10^{-8}$ cutoff, which is more typical of a SNP study.

There may be ways to improve these edge sets by making them more phenotype or tissue specific. For instance, augmenting the base STRING edge set with extra edges from phenotype-specific co-expression data, or up-weighting edges in STRING that have phenotype- or tissue-specific expression may increase performance. With the evaluation framework that we present in this paper, one could determine whether this kind of added information is useful.

Understanding GWAS hits in the context of their biological interactions gives insight into mechanism of action. However, most dense mod-

ule searching approaches rely on yes/no edge classification, which creates errors due to arbitrary cutoffs for edge scores. As shown by comparison to their older binary edge algorithm, the inclusion of edge weights into the search algorithm of EW_dmGWAS eliminates these errors and makes the algorithm more robust. We have shown that using the STRING confidence scores, along with appropriate changes to score normalization, improves the performance of dmGWAS even further. We provide an R-package, STAMS, that integrates the STRINGdb package, the EW_dmGWAS search function, and the STAMS-specific score normalization and allows users to easily perform this analysis on their gene-based p-values with the edge types of their choosing. We are also happy to share our reusable validation and evaluation suite. This suite represents the first systematic way to evaluate and compare dense module searching methods and parameter choices across a range of GWAS.

Acknowledgements

The authors wish to thank Susan Holmes for her general advice on the project, Ruth Schneider for comments on the manuscript, and Lichy Han for testing the R package.

Funding

This work was supported by The National Institutes of Health [MH094267, GM102365, GM61374]; The National Library of Medicine [LM07033]; and gifts from Oracle, Microsoft.

Conflict of Interest: none declared.

References

Ashburner,M. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat Genet*, **25**, 25–29.

Beissbarth,T. and Speed,T.P. (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics (Oxford, England)*, **20**, 1464–1465.

Breitling,R. *et al.* (2004) Graph-based iterative Group Analysis enhances microarray interpretation. *BMC Bioinformatics*, **5**, 100.

Burton,P.R. *et al.* (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.

Chen,Y.K. *et al.* (2012) CTTNBP2, but not CTTNBP2NL, regulates dendritic spinogenesis and synaptic distribution of the striatin-PP2A complex. *Molecular Biology of the Cell*, **23**, 4383–4392.

Cheung,J. *et al.* (2001) Identification of the Human Cortactin-Binding Protein-2 Gene from the Autism Candidate Region at 7q31. *Genomics*, **78**, 7–11.

Chuang,H.-Y. *et al.* (2007) Network-based classification of breast cancer metastasis. *Mol Syst Biol*, **3**.

Croft,D. *et al.* (2010) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research*, **39**, D691–D697.

Daneshjou,R. *et al.* (2014) Genetic variant in folate homeostasis is associated with lower warfarin dose in African Americans. *Blood*, **124**, 2298–2305.

Fehringer,G. *et al.* (2012) Comparison of Pathway Analysis Approaches Using Lung Cancer GWAS Data Sets. *PLoS ONE*, **7**, e31816.

Franceschini,A. *et al.* (2012) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, **41**, D808–D815.

Geschwind,D.H. *et al.* (2001) The Autism Genetic Resource Exchange: A Resource for the Study of Autism and Related Neuropsychiatric Conditions. *The American Journal of Human Genetics*, **69**, 463–466.

Gilman,S.R. *et al.* (2011) Rare De Novo Variants Associated with Autism Implicate a Large Functional Network of Genes Involved in Formation and Function of Synapses. *Neuron*, **70**, 898–907.

Goudreault,M. *et al.* (2008) A PP2A Phosphatase High Density Interaction Network Identifies a Novel Striatin-interacting Phosphatase and Kinase Complex Linked to the Cerebral Cavernous Malformation 3 (CCM3) Protein*. *Molecular and Cellular Proteomics*, 1–15.

Hormozdiari,F. *et al.* (2010) Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics (Oxford, England)*, **26**, i350–i357.

Ideker,T. *et al.* (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics (Oxford, England)*, **18**, S233–S240.

Jia,P. *et al.* (2010) dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks. *Bioinformatics (Oxford, England)*, 1–8.

Jia,P. *et al.* (2012) Network-Assisted Investigation of Combined Causal Signals from Genome-Wide Association Studies in Schizophrenia. *PLoS Comput Biol*, **8**, e1002587.

Kanehisa,M. and Goto,S. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 1–4.

Liu,J.Z. *et al.* (2010) A Versatile Gene-Based Test for Genome-wide Association Studies. *The American Journal of Human Genetics*, **87**, 139–145.

Nacu,S. *et al.* (2007) Gene expression network analysis and applications to immunology. *Bioinformatics (Oxford, England)*, **23**, 850–858.

Pagel,P. *et al.* (2005) The MIPS mammalian protein-protein interaction database. *Bioinformatics (Oxford, England)*, **21**, 832–834.

Parikshak,N.N. *et al.* (2013) Integrative Functional Genomic Analyses Implicate Specific Molecular Pathways and Circuits in Autism. *Cell*, **155**, 1008–1021.

Purcell,S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, **81**, 559–575.

Sanders,S.J. *et al.* (2015) Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron*, **87**, 1215–1233.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: A knowledge-basedapproach for interpreting genome-wideexpression profiles. *PNAS*, 1–6.

Wang,K. *et al.* (2007) Pathway-Based Approaches for Analysis of Genomewide Association Studies. *The American Journal of Human Genetics*, **81**, 1278–1283.

Wang,Q. *et al.* (2015) EW_dmGWAS: edge-weighted dense module search for genome-wide association studies and gene expression profiles. *Bioinformatics (Oxford, England)*, 1–4.