DOMINO: a novel network-based module detection algorithm with reduced rate of false calls

Hagai Levi[1], Ran Elkon[2,3,*] and Ron Shamir[1,*]

[1]The Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel. [2]Department of Human Molecular Genetics and Biochemistry, Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel. [3]Sagol School of Neuroscience, Tel Aviv University, Tel Aviv, Israel. *equal contribution.

Abstract

Network-based module discovery (NBMD) methods have taken a central role in integrative analyses of omics data in modern bioinformatics. NBMD algorithms receive a gene network and nodes' activity scores as input and report sub-networks (modules) that are putatively biologically meaningful in the context of the activity data. Although NBMD methods exist for almost two decades, only a handful of studies attempted to compare the biological signals captured by different methods. Here, we first set to systematically evaluate six popular NBMD methods on gene expression (GE) data and Gene-Wide-Association Studies (GWAS). Notably, testing Gene Ontology (GO) enrichment of modules obtained by these methods, we observed that GO terms enriched on modules detected on the real data were often also enriched after randomly permuting the input data. To tackle this bias, we designed the EMpirical Pipeline (EMP), a method that infers the empirical significance of GO enrichment scores of an NBMD solution by computing, for each term, a background distribution of scores on permuted data. We used the EMP to fashion five novel performance evaluation criteria for NBMD methods. Last, we developed DOMINO (Discovery of Modules In Networks using Omics) - a novel NBMD algorithm. In extensive testing on gene expression and genome-wide association study data it outperformed the other six algorithms. As it produces solutions with only a few non-specific GO terms, DOMINO can be used without empirical validation. EMP and DOMINO are available at https://github.com/Shamir-Lab/.

Introduction

The maturation of high-throughput technologies has led to an unprecedented abundance of omics studies. With the ever-increasing availability of genomic, transcriptomic and proteomic data (McLendon *et al*, 2008; VanderSluis *et al*, 2018; Zhang *et al*, 2011), a main challenge remains to uncover biological and biomedical insights by examination of these datasets as a whole (Chen *et al*, 2019). A leading approach to this challenge relies on biological networks(Aittokallio & Schwikowski, 2006), simplified yet solid mathematical abstractions of complex intra-cellular systems. In these networks, each node represents a cellular subunit (e.g. a protein) and each edge represents a relationship between two subunits (Szklarczyk *et al*, 2017; Wu *et al*, 2014; Xenarios *et al*, 2002) (e.g. a physical interaction between two proteins). Among the many bioinformatics tasks that are tackled by network-based methods, including gene function and drug target predictions (Emig *et al*, 2013; Warde-Farley *et al*, 2010), one of the most popular is the discovery of "active" modules in data. Given an omics dataset, network-based module discovery (NBMD) aims to detect subnetworks (i.e. modules) that are functionally relevant ("active") in the probed biological condition. The core of the NBMD task is to pinpoint highly scoring sets of interacting nodes, where the score of each node (i.e. the activity score) is derived from the data (e.g. $\log_2(fold-change\ of\ expression)$). As this problem has been proven to be NP-hard (Ideker *et al*, 2002), many heuristics were suggested to detect active modules (Mitra *et al*, 2013; Creixell *et al*, 2015).

An NBMD solution is composed of a set of modules that are enriched for the activity signal.

Typically, with a solution at hand, each module is subjected to functional analysis (Eden *et al*, 2009; Subramanian *et al*, 2005) wherein the GO (The Gene Ontology Consortium, 2019) functions of the modules genes are assessed. The most popular approach for biologically interpreting a gene set is the Hypergeometric (HG) test, where the proportion of genes annotated for a certain property (GO functional category) in the set is compared to a background set of genes. Applying the HG test to the entire set of the responsive genes in a dataset ignores the modular organization of the response and might miss the more elusive biological signals. In constrast, an NBMD solution can provide a finer understanding of the examined biological conditions and cellular responses by delineating subnetworks that carry out distinct biological endpoints (Leiserson *et al*, 2015; Cerami *et al*, 2010). For example, biological responses to stress often comprise the concurrent activation and repression of multiple biological processes, each mediated by a single or a few dedicated signaling pathways (Kyriakis & Avruch, 2012; Ashcroft *et al*, 2000). NBMD would ideally dissect such a complex response into distinct sub-networks, each representing a certain functional module.

Another key utility of NBMD methods is the amplification of weak signals, where an active module comprises multiple nodes that individually have only marginal scores, but collectively score

significantly higher. This ability of NBMD methods is especially critical for the functional interpretation of Genome-Wide Association Studies (GWASs) (Visscher *et al*, 2012). Numerous GWASs conducted over the last decade have demonstrated that the genetic component of complex diseases is highly polygenic (Khera *et al*, 2018; Musunuru & Kathiresan, 2019; Sullivan & Geschwind, 2019), affected by hundreds or thousands of genetic variants, the vast majority of which have only a very subtle effect. Therefore, most of the "risk SNPs" do not pass statistical significance when tested individually after correcting for multiple testing (Stringer *et al*, 2011; Boyle *et al*, 2017). This stresses the need for integrative approaches that consider multiple related nodes together, and NBMD methods are among the most effective for fulfilling this task (Marbach *et al*, 2016; Barrenas *et al*, 2009; Cowen *et al*, 2017).

Evaluation of NBMD solutions based on GO terms enrichment suffers from a substantial drawback: the lack of ground-truth annotations. The manual cherry-picking (Geistlinger *et al*, 2019) of the right terms out of those identified, is inevitably subject to researcher bias. Moreover, as different NBMD algorithms tend to capture different biological signals, the underlying functional processes cannot conclusively be determined.

In this study, we first aimed to systematically evaluate popular NBMD algorithms across multiple gene expression (GE) and GWAS datasets based on the enrichment of the called modules for functional GO categories. Unexpectedly, our analysis revealed that algorithms often obtained modules enriched for a high number of GO terms even when run on permuted datasets. Moreover, some of the GO terms that were recurrently enriched on permuted datasets, were also enriched on the original dataset, indicating that NBMD solutions commonly suffer from a high rate of false calls. We therefore designed a procedure for validating the functional analysis of an NBMD solution by comparing it to null distributions obtained on permuted datasets. We used the empirically validated set of GO terms to define novel metrics for evaluation of NBMD algorithms. Finally, we developed DOMINO (Discovery of Modules In Networks using Omics) – a novel NBMD method, and demonstrated that its solutions outperform extant methods in terms of the novel metrics and are typically characterized by a high rate of validated GO terms.

Results

NBMD algorithms suffer from a high rate of non-specific GO term enrichments

We set out to evaluate the performance of leading NBMD algorithms. Our analysis included six algorithms – jActiveModules (Ideker *et al*, 2002) in two strategies: greedy and simulated annealing (abbreviated JAM_greedy and jAM_SA, respectively), BioNet (Beisser *et al*, 2010), HotNet2 (Leiserson *et al*, 2015), NetBox (Cerami *et al*, 2010) and KeyPathwayMiner (Baumbach *et al*, 2012) (abbreviated KPM). These algorithms were chosen based on their popularity, computational methodology and diversity of original application (e.g., gene expression data, somatic mutations) (**Table S1**). As we wished to test these algorithms extensively, we focused on those that had a working tool/codebase that can be executed in a stand-alone manner, have reasonable runtime and could be applied to different data types. Details on the execution procedure of each algorithm are available in the Appendix. We applied these algorithms to two types of data: (1) a set of ten gene-expression (GE) datasets of diverse biological physiology (**Table S2**) where gene activity scores correspond to differential expression between test and control conditions, and (2) a set of ten GWAS datasets of diverse pathological conditions (**Table S3**) where gene activity scores correspond to association with the trait (Methods). In our analysis, we used the Database of Interacting Proteins (DIP (Xenarios *et al*, 2002)) as the underlying global network. Although the DIP network is relatively small - comprising about 3000 nodes and 5000 edges, in a recent benchmark analysis (Huang *et al*, 2018) it got the best normalized score on recovering literature-curated disease gene sets, making it ideal for multiple systematic executions.

First, applying the algorithms to the GE and GWAS datasets we observed that their solutions showed high variability in the number and size of modules they detected (**Figure S1** and **Figure S2**). On the GE datasets, jAM_SA tended to report a small number of very large modules while HotNet2 usually reported a high number of small modules (**Figure S1**). jAM_SA tended to report large modules also on the GWAS datasets (**Figure S2**). Next, we used the hypergeometric (HG) GO enrichment test to functionally characterize the solutions obtained by the algorithms. As part of our evaluation analysis, we applied the algorithms also on random datasets that we generated by permuting the original activity scores. Importantly, we observed that modules detected on the permuted datasets were frequently enriched for GO terms (**Figure 1A**). Moreover, different algorithms showed varying degree of overlap between the enriched terms obtained on real and permuted datasets (**Figure 1B**). These findings imply that some - or even many of terms reported by NBMD algorithms do not stem from the specific biological condition that was assayed in each dataset, but rather from other non-specific factors that bias the solution, such as the structure of the network, the methodology of the algorithm and the distribution of the activity scores.

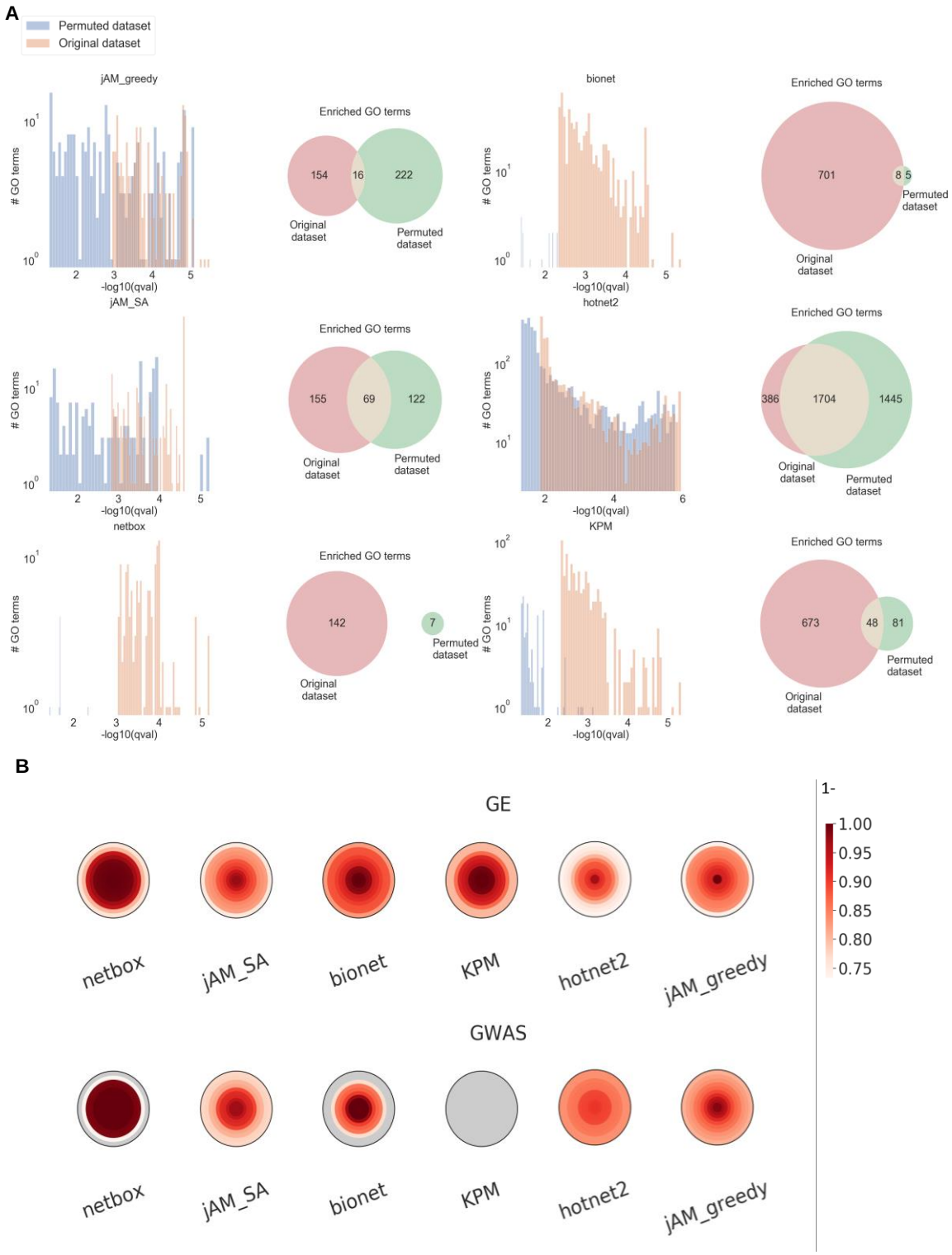


Figure 1. A. Comparison of GO enrichment results obtained on the original CBX GE dataset and on its permuted datasets. The histograms show the distributions of GO enrichment scores obtained for the modules detected on both datasets. The Venn diagrams show the overlap between the GO terms detected in the two solutions. **B**. Comparison of GO terms reported on the original and permuted GE and GWAS datasets. We used 1 minus the Jaccard score to measure the dissimilarity between the GO terms. Values close to 1 indicate low similarity between the results on the real and permuted data. Each circle shows, per algorithm, this measure (averaged over ten random permutations) over the 10 datasets. For each algorithm, the datasets are ordered such that higher scores are closer to the center. The gray color represents empty solutions. The results are shown separately for the GE and GWAS datasets.

A permutation-based method for filtering false GO terms

The high overlap between sets of enriched GO terms obtained on real and permuted datasets indicates that the results of most NBMD algorithms tested are highly susceptible to false calls that might lead to functional misinterpretation of the data. We looked for a way to filter out such non-specific terms while preserving the ones that are biologically meaningful in the context of the analyzed dataset. For this purpose, we developed a procedure called the EMpirical Pipeline (EMP). It works as follows: Given an NBMD algorithm and a dataset, EMP permutes the genes in the dataset and executes the algorithm. For each module reported by the algorithm, it performs GO enrichment analysis. The overall reported enrichment score for each GO term is its maximal score over all modules (**Figure 2A**). The process is repeated many times (typically, in our analysis, 5,000 times), generating a background distribution per GO term (**Figure 2B**). Next, the algorithm and the enrichment analysis are run on the real (i.e. non-permuted) dataset (**Figure 2C**). Denoting the background CDF obtained for GO term t by F_t, the empirical significance of t with enrichment score s is $e(t) = 1 - F_t(s)$. EMP reports only terms t that passed the HG test (q-value ≤ 0.05 on the original data) and had empirical significance $e(t) \leq 0.05$ (**Figure 2D**). We call such terms *empirically validated GO terms (EV terms)*. In addition, for each NBMD algorithm solution, we define the *Empirical-to-Hypergeometric Ratio* (EHR) as the fraction of EV terms out of all GO terms that passed the HG test (**Figure 2E,F**).

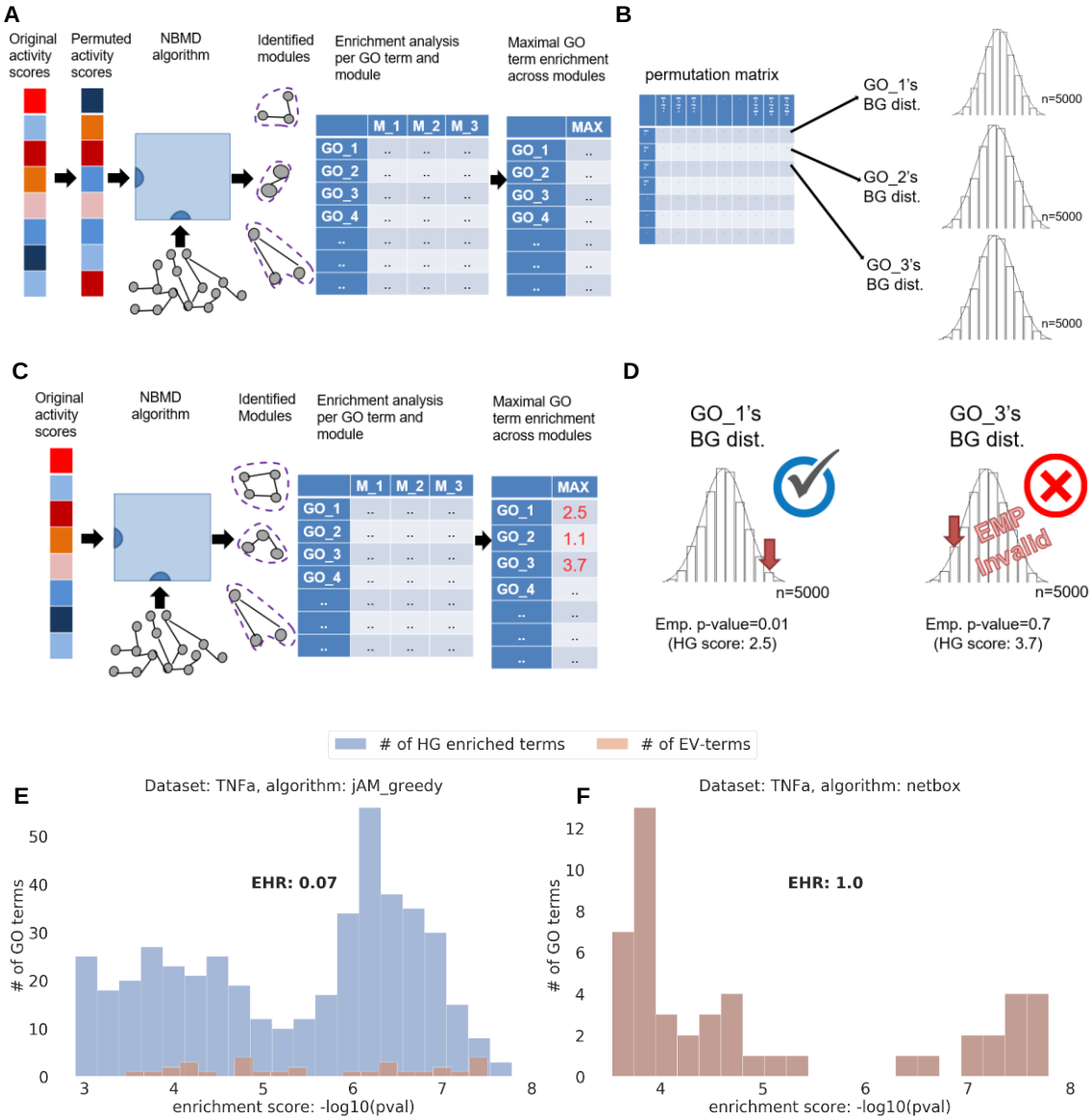


Figure 2. Overview of the EMpirical Pipeline (EMP) procedure. **A.** The NBMD algorithm and the GO enrichment analysis are applied on many instances (typically, n=5000) with permuted activity scores. **B.** A null distribution of enrichment scores is produced per GO term. **C.** The NBMD algorithm is applied to the original (un-permuted) activity scores, to calculate the real enrichment scores. **D.** For each GO term, the real enrichment scores is corrected according to its corresponding empirical distribution. In this example, GO_3 passed the HG test, but failed the empirical test and thus was filtered out. **E, F**. Distributions of HG enrichment scores for all the GO terms that passed the HG test and for the subset of the EV terms obtained on the TNFa expression dataset by jActiveModules with greedy strategy (E) and NetBox (F). The EHR measures the ratio between the number of EV terms and the number of GO terms that passed the HG test. The EHR scores summarize the advantage of NetBox in avoiding false reported terms.

The DOMINO algorithm

While the EMP method is a potent way for filtering out false GO term calls from NBMD solutions, this procedure is computationally demanding, as it requires several thousands of permutation runs. In our analyses, using a 44-cores server, EMP runs typically took several days to complete, depending

on the algorithm and the dataset. In order to provide a more frugal alternative that can be used on a desktop computer, we developed a novel NBMD algorithm called DOMINO (*Discovery of Modules In Networks using Omics*), with the goal of producing confident modules that also lead to high EHR values.

DOMINO receives as input a set of genes flagged as the *active genes* in a dataset (e.g., the set of genes that passed a differential expression test) and a network of gene interactions, aiming to find disjoint connected subnetworks in which the active genes are enriched. It has four main steps:

0. Dissect the network into disjoint, highly connected subnetworks (*slices).*
1. Detect *relevant slices* where active genes are enriched
2. For each relevant slice S
    a. Refine S to a sub-slice S'
    b. Repartition S' into putative modules
3. Report as final modules those that are enriched for active genes.

Step 0 - Dissecting the network into slices: This pre-processing step is done once per network (and reused for any analyzed datasets). In this step, the network is split into disjoint subnetworks called slices. Splitting is done using a variant of the Newman-Girvan modularity detection algorithm (Girvan & Newman, 2002) (Methods). Each connected component in the final network that has more than three nodes is defined as a *slice* (**Figure 3A**).

Step 1 - Detecting relevant slices: each slice that contains more active nodes than a certain threshold (see Methods) is tested for enrichment for active nodes using the Hypergeometric (HG) test, correcting the p-values for multiple testing using FDR(Benjamini & Hochberg, 1995). Slices with q-values < 0.3 are accepted as *relevant slices* (**Figure 3B**).

Step 2a - Refining the relevant slices into sub-slices: From each slice**,** the algorithm extracts a single connected component that captures most of the activity signal. The single component is obtained by solving the Prize Collecting Steiner Tree (PCST) problem(Johnson *et al*, 2000) (Methods). The resulting subgraph is called a *sub-slice* (**Figure 3C**).

Step 2b - Partitioning sub-slices into putative modules: Each sub-slice that is not enriched for active nodes and has more than 10 nodes is partitioned using the Newman-Girvan algorithm (Methods). The resulting parts, and the sub-slices of $\leq$ 10 nodes, are called *putative modules* (**Figure 3D**).

Step 3 - Identifying the final modules: Each putative module is tested for enrichment for active nodes using the HG test. In this step, we correct for multiple testing using the more stringent Bonferroni correction. Those with q-value < 0.05 are reported as the final modules (**Figure 3E**).

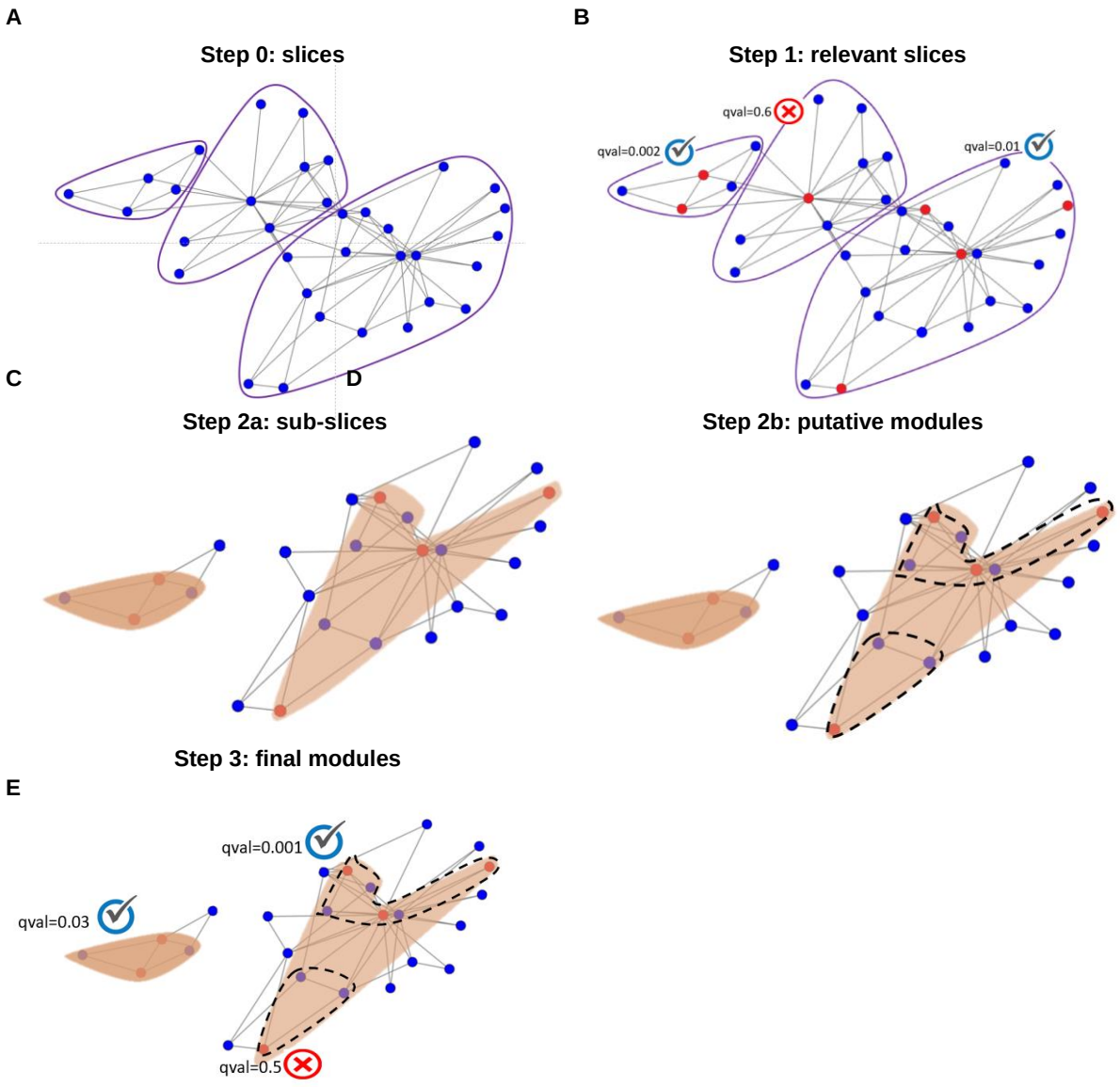


Figure 3. Schematic illustration of DOMINO. **A.** The global network is dissected by the Newman-Girvan (NG) modularity algorithm into slices (encompassed in purple line). **B.** A slice is considered relevant if it passes a moderate HG test for enrichment for active nodes ($FDR\ q \leq 0.3$). **C.** For each relevant slice the most active sub-slice is identified using PCST (red areas). **D.** Sub-slices are dissected further into putative modules using the NG algorithm. **E.** Each putative module that passes a strict enrichment test for active nodes ($Bonferroni\ qval \leq 0.05$) is reported.

Systematic evaluation of NBMD algorithms on gene-expression and GWAS datasets

We next carried out a comparative evaluation of DOMINO and the six NBMD algorithms described above (**Table S1**) over the same ten GE and ten GWAS datasets (**Tables S2,S3**). This evaluation task is challenging as there are no "gold-standard" solutions to benchmark against. To address this

difficulty, we introduce five novel scores for the systematic evaluation of NBMD algorithms. These scores are based on our EMP method and the GO terms that pass this empirical validation procedure. The scores are described in Methods and the results on all algorithms are summarized in **Figures 4-6**.

(a) EHR (*Empirical-to-Hypergeometric Ratio*). EHR summarizes the tendency of an algorithm to capture biological signals that are specific to the analyzed data, i.e. GO terms that are enriched in modules found on the real but not on permuted data. EHR has values between 0 to 1, with higher values indicating better performance. In our evaluation, DOMINO and NetBox scored highest on EHR. In both GE and GWAS datasets, DOMINO performed best with an average above 0.8. (**Figure 4A,B**). Importantly, these high EHR levels were not a result of reporting few terms: DOMINO reported a high number of enriched GO terms with only NetBox and jAM_greedy on GWAS reporting more (**Figure 4C,D**). Since HotNet2, originally developed for analysis of somatic mutation data, yielded poor results on both GE and GWAS datasets we excluded it from subsequent evaluations. For the same reason we included KPM only in the subsequent evaluations of GE datasets.

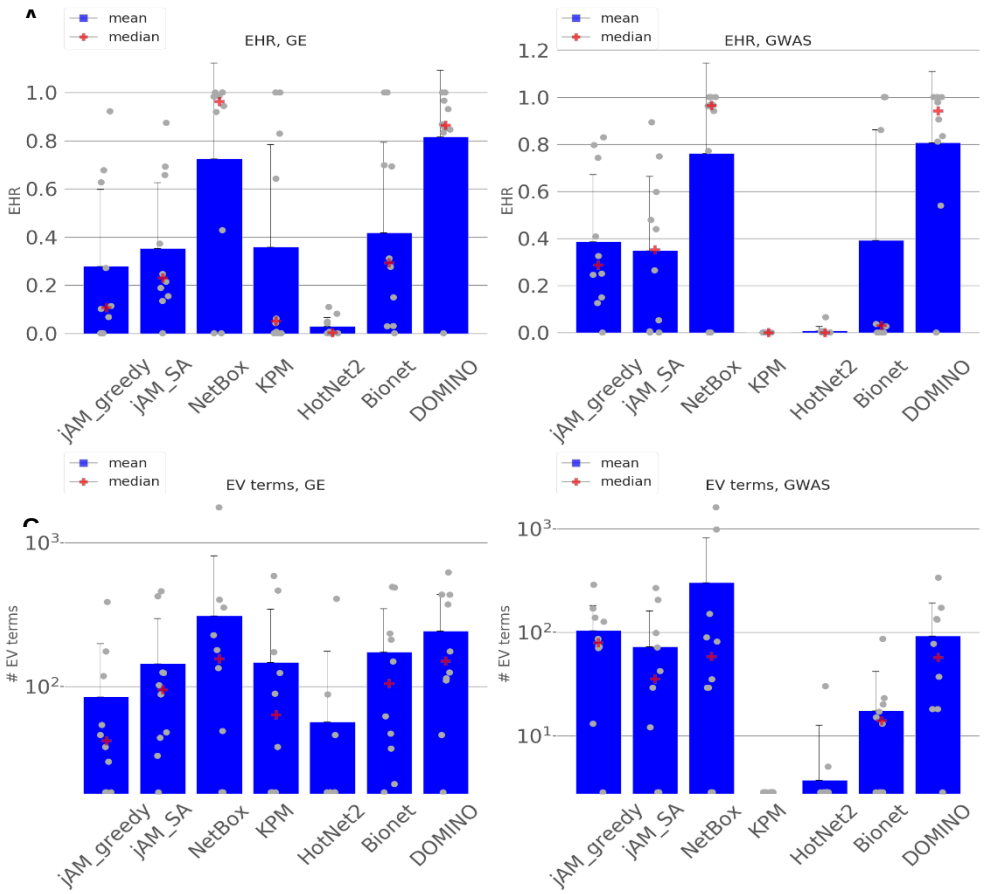


Figure 4. EHR and number of reported terms. **A**. EHR for the GE datasets. **B.** EHR for the GWAS datasets. **C**. The number of EV terms reported for the GE datasets. **D**. The number of EV terms reported for the GWAS datasets. The dots indicate results for each dataset. Error bars indicate the SD across datasets.

(b) **Module-level EHR (mEHR).** While the EHR characterizes a solution as a whole by considering the union of GO terms enriched on any module, biological insights are often obtained by functionally characterizing each module individually. We therefore next evaluated the EHR of each module separately. Specifically, for each module, we calculated the fraction of its EV terms out of the HG terms detected on it (Methods). The results are summarized in **Figure 5A**. Notably, solutions can have a broad range of mEHR scores (see, for example, NetBox solution on the IEM dataset, where the best module has an mEHR above 0.9 while the poorest has an mEHR below 0.2). To summarize the results over multiple modules, we averaged the k top scoring modules (using k=1 to 20; **Figure 6A**). In this criterion, DOMINO got highest mEHR scores for most values of k, followed by NetBox. The results for GWAS datasets are shown in **Figure S3A** and **Figure S4A**.

Furthermore, the EMP procedure enhances the functional interpretation of each module by distinguishing between its enriched GO terms that are specific to the real data (i.e., the EV terms) and those that are recurrently enriched also on the permuted ones.This utility of EMP is demonstrated, as one example, on a module detected by jAM_greedy on the TNFa GE dataset (**Figure 5B**). TNFa is a potent inducer of immune reponses largely mediated by the NFκB transcription factors. This biological process is well captured by the GO terms that passed EPM validation (e.g., "NIK/NF-kappaB signaling") (Hayden & Ghosh, 2014). In contrast, GO terms that failed passing this validation procedure represent less specific processes (e.g., "regulation of RNA biosynthetic process"). Similarly, EV terms of a module detected by DOMINO on the schizophrenia GWAS data are highly relevant for this trait (e.g., "neurotransmitter metabolic process", "regulation of neurogenesis" and "learning") (Ripke *et al*, 2014) while GO terms that did not pass validation are either generally less specific (e.g., "system development" and "regulation of localization") or seem less relevant biologically (e.g., "regulation of apoptosis") **(Figure S3B)**.

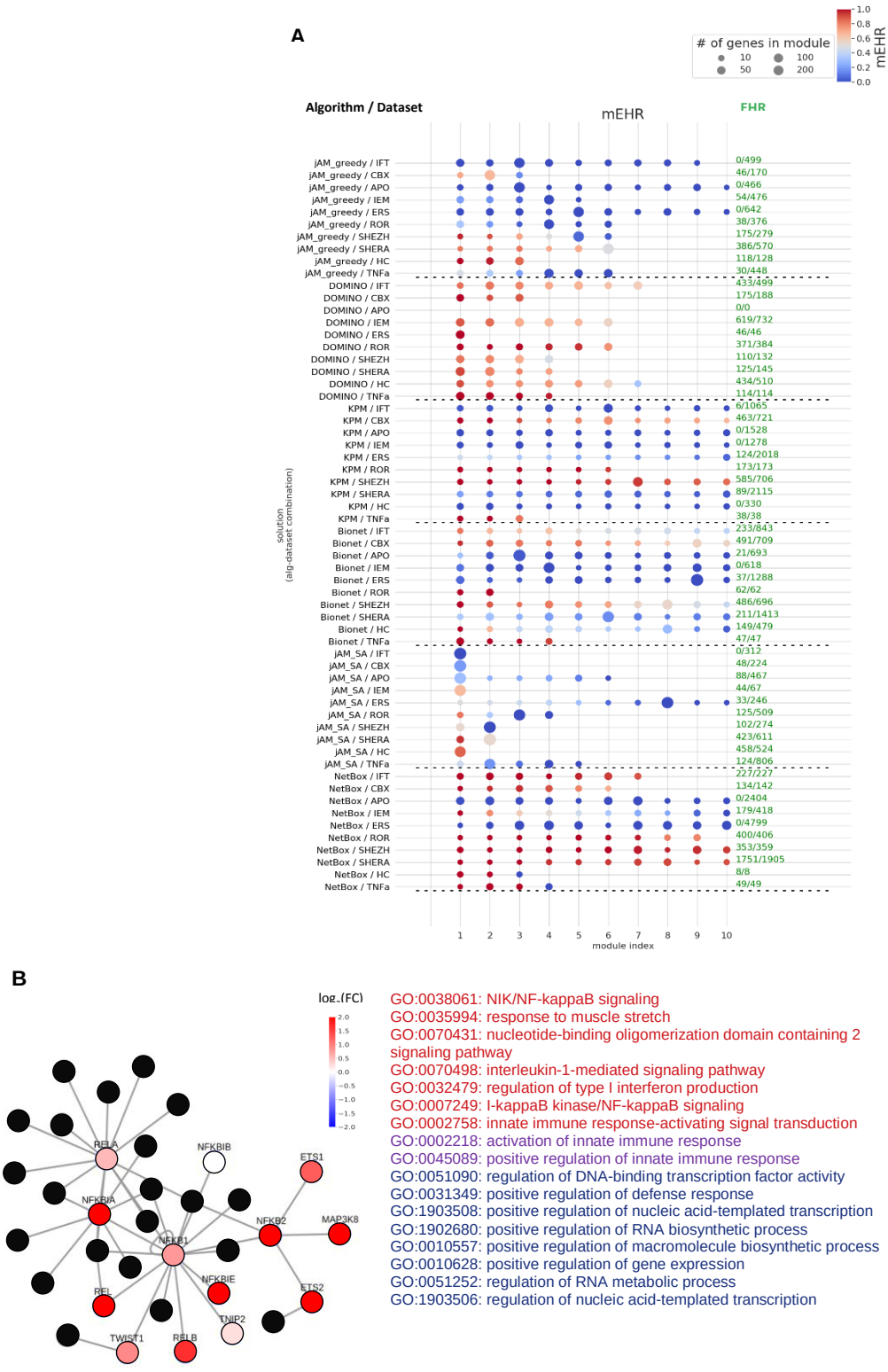


Figure 5. Performance measured using the module-level EHR (mEHR) criterion on GE datasets. **A**. mEHR scores for each algorithm and dataset. Up to ten top k modules are shown per datasets, ranked by their mEHR. **B.** An example of a module from the solution reported by jAM_greedy on the TNFa dataset (mEHR=0.35). The nodes' color indicates the logarithm of their fold change in the dataset. The black nodes are the neighbors of the module's nodes in the network. Right: The EV terms for this module are shown in red and those that did not pass the empirical validation in blue. GO terms with borderline EV score ($0.05 < \text{q-val} < 0.1$) are colored in purple).

(c) **Biological richness.** The next criterion aims to measure the diversity of biological processes captured by a solution. Our underlying assumption here is that the biological systems are complex and their responses to triggers involve the concurrent modulation of a diversity of biological processes. For example, genotoxic stress concurrently activates DNA damage repair mechanisms and apoptotic pathways and suppresses cell-cycle progression. However, merely counting the number of EV terms of a solution would not faithfully reflect its biological richness because of the high redundancy between GO terms. This redundancy stems from overlaps between sets of genes assigned to different GO terms, mainly due to the hierarchical structure of the ontology. We therefore used REVIGO(Supek *et al*, 2011) to derive a non-redundant set of GO terms based on semantic similarity scores(Lord *et al*, 2003)(Resnik, 1999). We defined the *biological richness score* of a solution as the number of its non-redundant EV terms (Methods). The results in **Figure 6B** show that on the GE datasets, DOMINO and NetBox performed best. On the GWAS datasets, jAM_greedy performed best (**Figure S4B**).

(d) **Intra-module homogeneity**. While high biological diversity (richness) is desirable at the solution level, a single module should ideally capture only a few related biological processes. Solutions in which the response is dissected into modules where each represents a distinct biological endpoint are easier to interpret biologically and are preferred over solutions with larger modules, where each represents several composite processes. To reflect this preference, we introduced the *intra-module homogeneity score,* which quantifies how functionally homogeneous the EV terms captured by each module are (Methods). For each solution, we take the average score of its modules. On the GE datasets, BioNet and NetBox performed best in this criterion for the lower similarity cutoffs while KPM scored the highest for the higher cutoffs (**Figure 6C**). On the GWAS datasets, NetBox, DOMINO, and jAM_greedy scored higher than jAM_SA and BioNet (**Figure S4C**).

(e) **Robustness**. This criterion measures how robust an algorithm's results are to subsampling of the data. It compares the EV-terms obtained on the original dataset with those obtained on randomly subsampled datasets. Running 100 subsampling iterations and using the EV terms found on the original dataset as the gold-standard GO terms, we compute AUPR and average F1 scores for each solution (Methods). DOMINO's solutions showed the highest robustness on the GE datasets, followed by NetBox (**Figure 6D,E**). It also performed best on the GWAS datasets, showing markedly higher robustness than all other algorithms (**Figure S4D,E**).

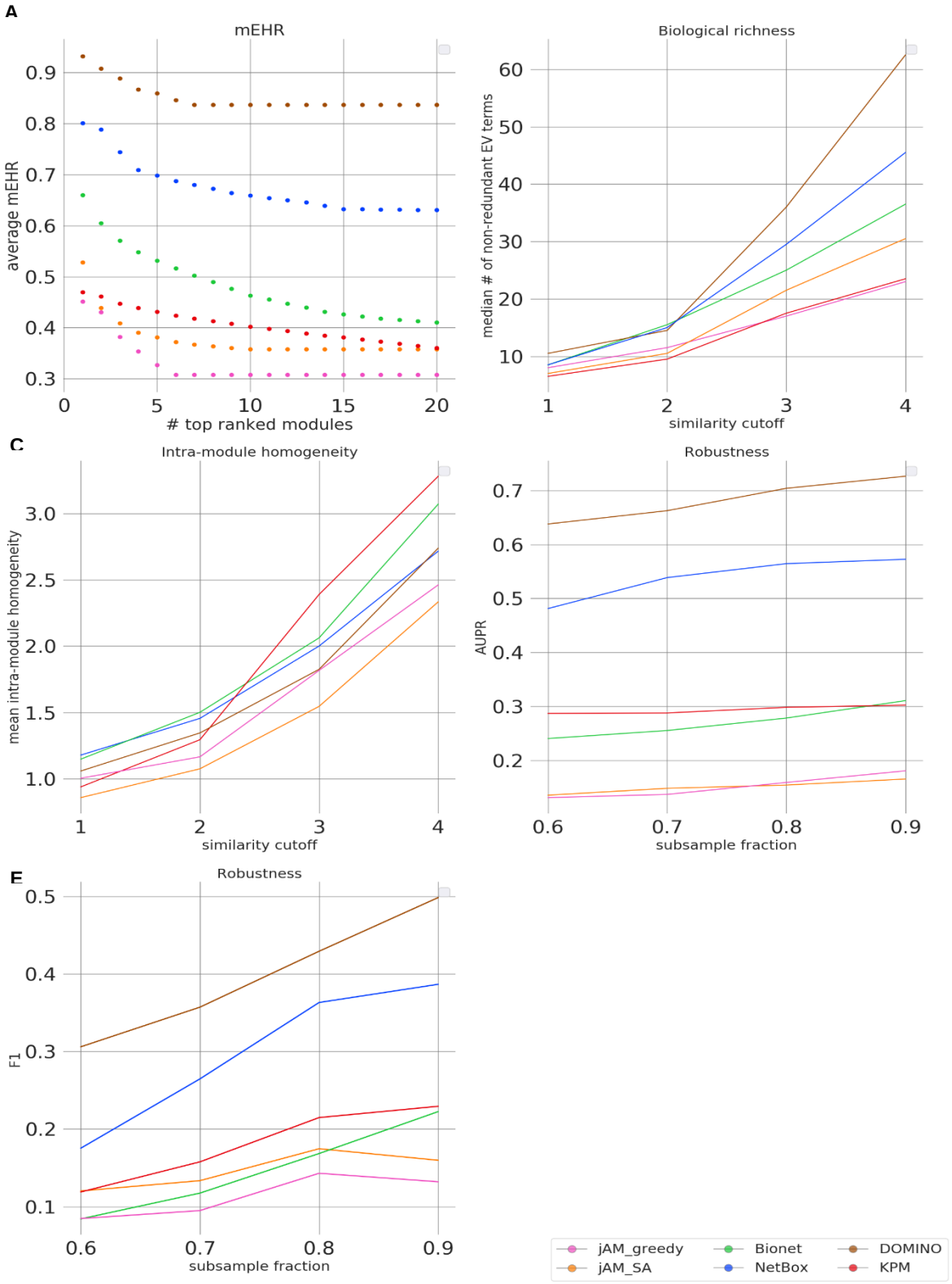


Figure 6. Evaluation results for the GE datasets. **A.** Module-level EHR scores. The plots show the average mEHR score in the k top modules, as a function of k in each dataset. Modules were ranked by their mEHR scores. **B.** Biological richness. The plots show the median number of non-redundant terms (richness score) as a function of the Resnik similarity cutoff. **C.** Intra-module homogeneity scores as a function of the similarity cutoff. **D.** Robustness measured by the average AUPR over the datasets, shown as a function of the subsampling fraction. E. Robustness measured by the average F1 over the datasets shown as a function of the subsample fraction. For each dataset and subsampling fraction 100 samples were drawn and averaged.

Table 1 (and **Table S4**) summarizes the results of the benchmark on GE and GWAS datasets. For the GE datasets, DOMINO performed best in five of the six criteria, while KPM scored highest in intra-module homogeneity. On the GWAS datasets, DOMINO scored best in four criteria, while jAMgreedy had the highest biological richness and NetBox had the highest intra-module homogeneity. Overall, these results demonstrate the high performance of DOMINO in multiple solution facets in both GE and GWAS datasets. NetBox tended to give the second-best results overall.

Discussion

The fundamental task of network-based module discovery (NBMD) algorithms is to identify active modules in an underlying network based on genes activity profiles. The comparison of such algorithms is challenging due to the complex nature of the solutions produced. Algorithms differ dramatically in the number, size, and properties of the modules they detect. Although NBMD algorithms have been extensively used for some two decades, there is no accepted community benchmark and no consensus evaluation criteria have emerged. Since modules are often used to characterize the biological processes that are activated/repressed in the probed biological conditions, we analyzed the solutions produced by the algorithms from the perspective of functional enrichment. Early on, we observed that many enriched GO terms also appear on permuted datasets, suggesting that such enrichments stem from some proprieties of the algorithms or the data that bias the results. Following this observation, we developed the EMP procedure, which empirically calibrates the enrichment scores and filters out non-specific terms.

Our analysis highlighted the need for improved NBMD algorithms and better benchmark methodology. We developed the DOMINO algorithm and defined five novel evaluation criteria to allow systematic comparison of NBMD algorithms. Each of these criteria emphasizes a different aspect of the solution (**Figure 7**). We used these criteria to evaluate the performance of six popular NBMD algorithms and our DOMINO algorithm on a set of ten GE and ten GWAS datasets that collectively cover a very wide spectrum of biological conditions. Overall, DOMINO performed best, indicating its ability to produce "clean", stable and concise modules. NetBox also scored high in our evaluation analysis. Interestingly, both DOMINO and NetBox handle the activity scores as binary ones. Intuitively one may expect that such a step could lead to a loss of important biological signals. However, the high performance of these algorithms suggests that at least on our benchmark binarizing the data helped in reducing noise. Further study of this observation is needed.

Notably, the algorithms that we tested differ substantially in their empirical validation rates (i.e., EHR). Some algorithms produced solutions with very low EHR (<0.5), and therefore running the EMP on them is critical. While empirical correction is desirable and adds confidence to the reported

results, it is computationally highly demanding even with a relatively small network (DIP). Using larger networks, of course, makes this procedure even slower. A notable advantage of DOMINO is the high validation rates it consistently obtained: its average EHR and average mEHR were above 0.8. This indicates that DOMINO can be confidently run without EMP when computational resources are limited. The EMP and DOMINO software and codebases are freely available to the community at https://github.com/Shamir-Lab/.

One shortcoming of EMP is that it does not lend itself to provide module-based correction for enrichment scores, since each randomized run can produce a different number of modules of different sizes. Ideally, one would like to validate the GO terms on the module level. Nevertheless, we do provide means for validation of terms on the module-level by the mEHR index, which calculates the proportion of enriched GO terms that passed the EMP filter in each module. Another limitation is the speed, which also limits the size of networks one can use.

An additional future task is to understand better the sources of the bias that causes over-reporting of enriched GO terms. The sources may be the activity score distribution, network structure, algorithm strategy, etc. Obtaining such understanding could lead to improved module discovery and shorter runtimes of EMP. It could also enable tuning of each algorithms' hyper-parameters, which is another open issue in our analysis.

In summary, in this study we (1) report on a highly prevalent bias in popular NBMD algorithms that leads to non-specific calls of enriched GO terms, (2) implemented a procedure to allow for the correction of this bias, (3) introduced novel evaluation criteria of solutions and (4) developed DOMINO – a novel NBMD algorithm with low rate of non-specific calls and better performance across most of the criteria.

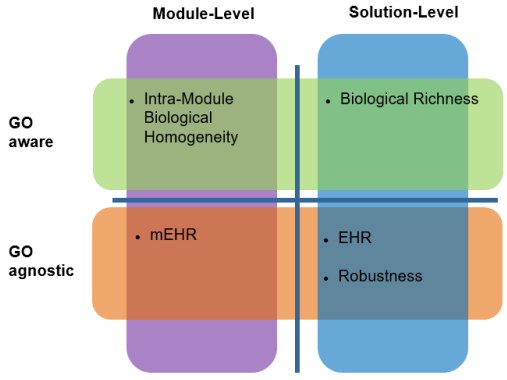


Figure 7. A breakdown of the evaluation criteria by their properties. Richness, EHR and robustness score solutions based only on the whole set of the reported GO terms, without taking into account the results for individual modules. In contrast, mEHR and intra-module homogeneity score solutions in a module-aware fashion. From another perspective, biological richness and intra-module homogeneity consider the functional relations among the reported GO terms, while EHR, mEHR, and robustness do not. Colors highlight the different facets considered by each group of scores.

Table 1. Summary of the benchmark results.

	EHR	mEHR*	Robustness (F1)	Robustness (AUPR)	Biological Richness#	Intra-Module Homomgeneity#
	GE					
NetBox	0.73	0.66	0.36	0.54	29.5	2.00
jAM_sa	0.35	0.36	0.17	0.15	21.5	1.55
Bionet	0.42	0.46	0.17	0.26	25	2.06
KPM	0.36	0.40	0.21	0.29	17.5	**2.39**
DOMINO	**0.82**	**0.84**	**0.43**	**0.66**	**36**	1.83
jAM_greedy	0.28	0.31	0.14	0.14	17	1.82
	GWAS					
NetBox	0.76	0.78	0.39	0.45	13	**1.81**
jAM_sa	0.35	0.38	0.12	0.14	10	0.93
Bionet	0.39	0.43	0.33	0.33	5	1.06
DOMINO	**0.81**	**0.81**	**0.77**	**0.77**	15	1.56
jAM_greedy	0.39	0.46	0.28	0.24	**18**	1.72

Per algorithm, the average over the ten datasets is shown
*Results are average over the top 10 modules
#Results shown for Resnik cutoff=3

Methods:

1. The Newman-Girvan algorithm in DOMINO

The Newman-Girvan (NG) algorithm is a community detection method(Girvan & Newman, 2002). This method iteratively removes edges using the Betweenness-centrality metric for edges and recomputes the modularity score for each intermediate graph. Let M_i be the modularity score for the graph in iteration i. The process continues until a stopping criterion is met. The stopping criterion we used in DOMINO's step (1) is that $M_{i+1} \leq M_i$. For step (2b), the stopping criterion is $\frac{log(\#\ of\ nodes\ in\ sub-slice)}{log(\#\ of\ nodes\ in\ network)} \leq M_i$. For more details see Appendix.

2. Threshold for testing relevant slices

Slices that contain only a few active nodes are unlikely to be relevant. Testing multiple such slices would diminish the significance of the actual relevant slices. Therefore, we test for relevance only slices that satisfy either

$$\frac{\#active\ nodes\ in\ slice}{\#active\ nodes\ in\ network} \geq 0.2\ .$$

or

$$\#active\ nodes\ in\ slice > log_2(\#active\ nodes\ \text{in network}).$$

3. The PCST application in DOMINO

In PCST (Johnson *et al*, 2000), nodes have values called prizes, and edges have values called penalties. All values are non-negative. The goal is to find a subtree T that maximizes the sum of the prizes of nodes in T minus the sum penalties of the edges in it, i.e., $\sum_{v\in T} p(v) - \sum_{e\in T} c(e)$ where $p(v)$ is the prize of node v, and $c(e)$ is the cost of edge e.

The node prizes are computed by diffusing the activity of the nodes using influence propagation with the linear threshold model(Kempe *et al*, 2015). The process is iterative: Initially, the set of active nodes is as defined by the input. In each iteration, an inactive node is activated if the sum of the influence of its active neighbors exceeds $\theta = 0.5$. The influence of a node that has k neighbors on each neighbor is $\frac{1}{k}$. Activated nodes remain so in all subsequent iterations. The process ends when no new node is activated. If v became active in iteration l then $p(v) = 0.7^l$. We define the penalty of edge e as $c(e) = 0$ if it is connected to an active node, and $c(e) = 1 - \epsilon$ otherwise (we used $\epsilon = 10^{-4}$).

PCST is NP-hard but good heuristics are available. In DOMINO we used FAST-PCST (Hegde *et al*, 2014). The resulting subgraph obtained by solving PCST on each slice is called its sub-slice. See **Figure 3C**.

4. Derivation of p-values and q-values for the GE and GWAS datasets

For the GE datasets, we calculated p-values for differential expression between test and control conditions using edgeR (Robinson *et al*, 2010) for RNAseq and student t-test for microarray datasets. We computed q-values using Benjamini-Hochberg FDR method (Benjamini & Hochberg, 1995). For GWAS we took the p-values of each SNP for the significance of its association with the analyzed trait and summarized them to gene-level p-values with PASCAL (Lamparter *et al*, 2016), using the sum chi-square option and flanks of 50k bps around genes. We computed q-values using Benjamini-Hochberg FDR method (Benjamini & Hochberg, 1995).

5. NBMD tools - execution details

See the Appendix for details on the execution of each of the six algorithms benchmarked.

6. Criteria for evaluating NBMD solutions

We defined five novel criteria to allow systematic evaluation of solutions provided by NBMD algorithms. For a specific solution, we considered the list of GO terms that passed the HG enrichment test (HG terms) and the terms that passed the EMP validation procedure (EV terms).

Solution-Level Criteria

(1) **Empirical to Hypergeometric Ratio (EHR)**. We define the *Empirical-to-Hypergeometric Ratio* (EHR) as the ratio between the number of reported HG terms and EV-terms. EHR summarizes the tendency of an algorithm to over-report GO terms, with values close to 1.0 indicating good solutions while values close to 0 indicating poor ones. EHR reflects the precision (true positive rate) of a solution.

(2) **Biological Richness.** This criterion quantifies the biological information collectively captured by the EV-terms. As often there is high redundancy among enriched GO terms - mainly due to the hierarchical structure of the GO ontology - we use the method implemented in REVIGO(Supek *et al*, 2011) to derive a non-redundant set of EV terms, that is, a measurement of the biological diversity of the solution. This method is based on a similarity matrix of GO terms, which is generated using Resnik similarity score (Resnik, 1999). The *biological richness score* is defined as the number of non-redundant EV terms in a solution. We calculated this measure using different similarity cutoffs (1.0 to 4.0 in REVIGO).

(3) **Solution Robustness.** This criterion evaluates the robustness of a solution to incomplete gene activity data. It compares the EV-terms obtained on the original dataset with those obtained on randomly subsampled datasets, where non-sampled gene levels are treated as missing. We repeated this procedure for subsampling fractions 0.6, 0.7, 0.8, and 0.9, iterating each fraction 100 times. Using the EV terms of the full dataset as the truth, we then computed average precision, recall and F1 scores across these iterations. Another perspective is provided by the examination of the frequency by which GO terms are detected in the subsampled datasets: higher frequency for a specific EV-term implies higher robustness. We measured this robustness aspect of a solution using AUPR, in which EV terms are ranked according to their frequency across iterations, and EV terms detected on the full dataset are used as the positive instances). Note that cases in which an algorithm results in many empty solutions (that is, solutions with no enriched

GO terms) and a few non-empty ones that are enriched for true EV terms can yield a high but misleading AUPR score. Therefore we validated that the fraction of non-empty solutions obtained by the algorithms on the subsampled runs is high: all the algorithms achieved around 70% or more non-empty solutions on GE data (**Figure S6**).

Module-Level Criteria

(1) Module-Level EHR (mEHR). This criterion calculates a single module's EHR. We define the module-level EHR (*mEHR*), as the ratio between the number of a module's EV terms and HG terms (**Figure S5A**). We score each solution by averaging the mEHR of its k top-ranked modules (k values ranging from 1-20).

(4) **Intra-Module Homogeneity.** This index measures the homogeneity of the biological signal that is captured by each module compared to the biological signal in the entire solution. For its calculation, we build a (complete) graph for the solution's EV terms (GO graph) in which nodes represent the EV-terms and the weights on the edges are the pairwise Resnik similarity score (**Figure S5B**). Next, edges whose weight is below a cutoff are removed. The *intra-module homogeneity* is defined as the module's relative edge-density:

$$\frac{\left(\frac{\#\ of\ edges\ in\ module}{\#\ of\ edges\ in\ a\ complete\ module\ of\ the\ size}\right)}{\left(\frac{\#\ of\ edges\ in\ graph}{\#\ of\ edges\ in\ a\ complete\ graph\ of\ the\ same\ size}\right)}$$

We calculate the intra-module homogeneity score for a solution by averaging its modules' scores (**Figure S5B**). We repeat this test for a range of similarity cutoffs – from 1.0 to 4.0. This criterion provides a complementary view on top of the one captured by the biological richness criterion, by characterizing its the biological coherence of the reported modules.

Funding:

Study supported in part by German-Israeli Project DFG RE 4193/1-1 (to RS and RE), by the Israel Science Foundation grants No. 1339/18 (to RS) and 2118/19 (to RE), by Len Blavatnik and the Blavatnik Family foundation (to RS) and the Koret-UC Berkeley-Tel Aviv University Initiative in Computational Biology and Bioinformatics (to R.E.). HL was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University. R.E. is a Faculty Fellow of the Edmond J. Safra Center for Bioinformatics at Tel Aviv University.

Table S1. NBMD algorithms included in our analysis.

Method name	**Published on**	**Designed for**	**Algorithmic Approach**	**Code language**	**# citations (updated to 11/2019)**
jActiveModules (Ideker *et al*, 2002)	2002	GE	Seek high scoring subnetworks either by simulated annealing (jAM_SA) or by a greedy search (jAM_greedy)	Java	1207
NetBox (Cerami *et al*, 2010)	2010	Somatic Mutations	Enrichment of Perturbed neighbors, Newman-Girvan (NG) modularity score	Java, Python	304
BioNet (Beisser *et al*, 2010)	2010	GE	Prize collecting Steiner tree	R	218
HotNet2 (Leiserson *et al*, 2015)	2015	Somatic Mutations	Heat diffusion	Python	460
KeyPathwayMiner (Baumbach *et al*, 2012)	2012	GE	Choose modules with at most K non-perturbed genes	Java	41

Table S2. The ten gene expression datasets used in our benchmark analysis.

Datasets name (acronym)	access to data	Technology	General description
TNFa (Schmidt *et al*, 2015)	GSE64233	RNA-seq	TNFa, a potent inducer of immune responses
HC (Elkon *et al*, 2015)	GSE67478	RNA-seq	Hair cell from the cochlea and vestibular system, compared to non-hair cell from these inner-ear organs.
SHERA (Miano *et al*, 2018)	GSE108693	RNA-seq	Luminal lncRNAs regulation by ERα-controlled enhancers in a ligand-independent manner in breast cancer cells. Comparison was made between ER siRNA to control siRNA
SHEZH (Ito *et al*, 2018)	GSE109064	RNA-seq	Downregulation of EZH2 leads to cellular senescence with features of SASP. Comparison between control to 4d samples.
ERS (Kroeger *et al*, 2018)	GSE106847	RNA-seq	ATF6 encodes a transcription factor that is activated during the Unfolded Protein Response to protect cells from ER stress. Comparison was made between ATF6-activated and control cells.
IEM (Hertzano *et al*, 2011)	---	Microarray	Comparison between 2 different cell types in the inner-ear: blood cells and mesenchymal cells.
ROR (Bayerlová *et al*, 2017)	GSE74383	RNA-seq	RNA-Seq profiling of estrogen-receptor-positive MCF-7 cell lines with different perturbations of non-canonical WNT signaling. Comparison was made between ROR2-overexpression and control conditions.
APO (Pulikkan *et al*, 2018)	GSE101788	RNA-seq	Comparison between ME-1 cells (a human leukemia cell line) treated with either the AI-10-49 drug (which induces apoptosis) or DMSO (control).
CBX (Connelly *et al*, 2019)	GSE123689	RNA-seq	CBX8 is a subunit of the polycomb repressive complex 1 (PRC1). This RNA-seq experiment compared CBX8-KO and control cells.
IFT (Forbes *et al*, 2018)	GSE107230	RNA-seq	IFT140 is involved in the formation and maintenance of cilia. This RNA-seq experiment compared uncorrected (IFT140 compound heterozygous) and gene-corrected (IFT140 heterozygous) epithelial cells isolated from patient's iPSC that were derived from kidney organoids.

Table S3. The ten GWAS datasets used in our benchmark analysis.

Datasets name (acronym)	**Trait**
BC (Michailidou *et al*, 2017)	Breast Cancer
CD (De Lange *et al*, 2017)	Crohn's Disease
SCZ (Ripke *et al*, 2014)	Schizophrenia
TRI (Teslovich *et al*, 2010)	Triglycerides
T2D (Mahajan *et al*, 2018)	Type 2 Diabetes
CAD (Nelson *et al*, 2017)	Coronary Artery Disease
BMD (Kemp *et al*, 2017)	Bone Mineral Density
Height (Allen *et al*, 2010)	Height
AF (Nielsen *et al*, 2018)	Arterial Fibrillation
AMD (Fritsche *et al*, 2016)	Age Related Macular Degeneration

Table S4. Summary of the standard deviation of the results in Table 1

	EHR	mEHR	Robustness (F1)	Robustness (AUPR)	Biological Richness	Intra-Module Homomgeneity
	GE					
NetBox	4.20E-01	1.17E-16	2.25E-01	3.31E-01	3.41E+01	1.58E+00
jAM_sa	2.89E-01	5.85E-17	1.70E-01	1.80E-01	1.65E+01	1.55E+00
Bionet	3.96E-01	0.00E+00	1.33E-01	2.70E-01	2.50E+01	1.66E+00
KPM	4.50E-01	5.85E-17	2.55E-01	3.63E-01	2.74E+01	2.59E+00
DOMINO	2.94E-01	0.00E+00	1.96E-01	2.53E-01	2.44E+01	8.68E-01
jAM_greedy	3.38E-01	5.85E-17	1.80E-01	1.93E-01	1.37E+01	2.79E+00
	GWAS					
NetBox	4.07E-01	1.17E-16	4.25E-01	4.13E-01	3.52E+01	1.78E+00
jAM_sa	3.33E-01	5.85E-17	1.69E-01	2.08E-01	1.14E+01	6.70E-01
Bionet	4.95E-01	0.00E+00	4.29E-01	4.36E-01	6.69E+00	1.50E+00
DOMINO	3.18E-01	1.17E-16	2.90E-01	3.12E-01	1.70E+01	1.58E+00
jAM_greedy	2.99E-01	5.85E-17	2.11E-01	2.12E-01	1.05E+01	1.48E+00

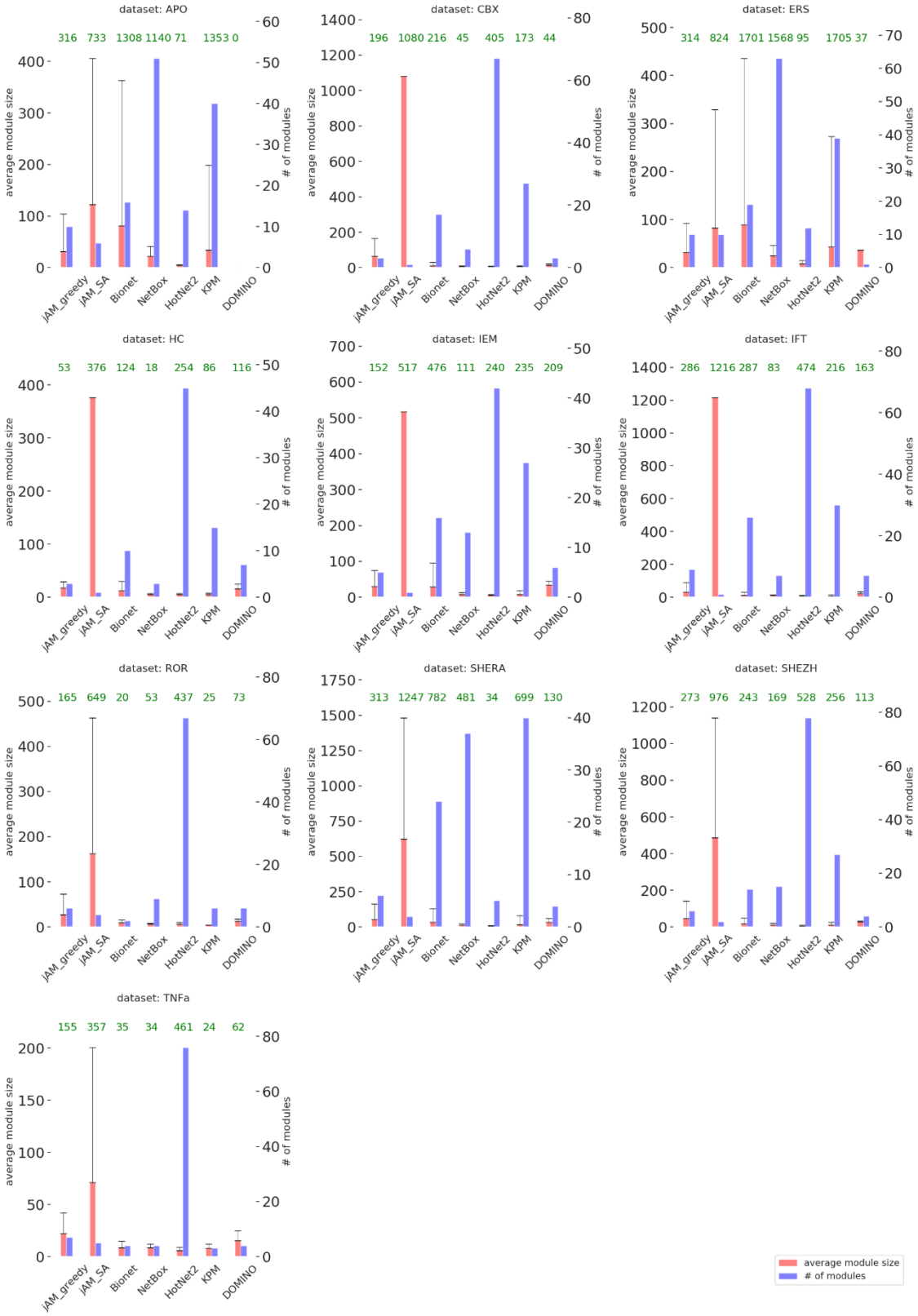


Figure S1. Summary statistics of the solutions obtained on the GE datasets. For each dataset, the number of modules detected by each NBMD algorithm and their sizes are indicated. (Error bars represent 1 SD of the number of genes in modules). The numbers in green are the total number of genes in the union of all modules in the solution.

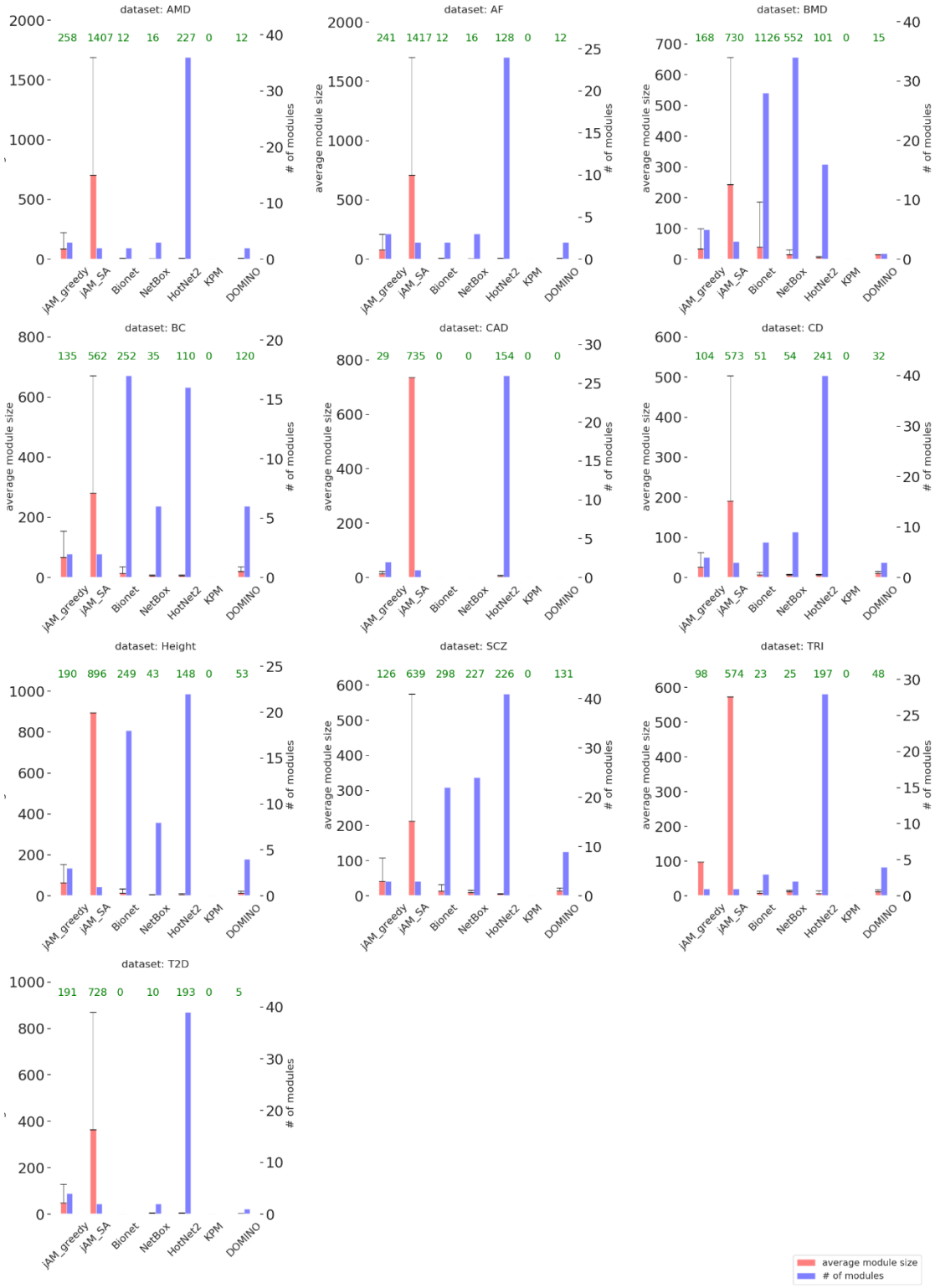
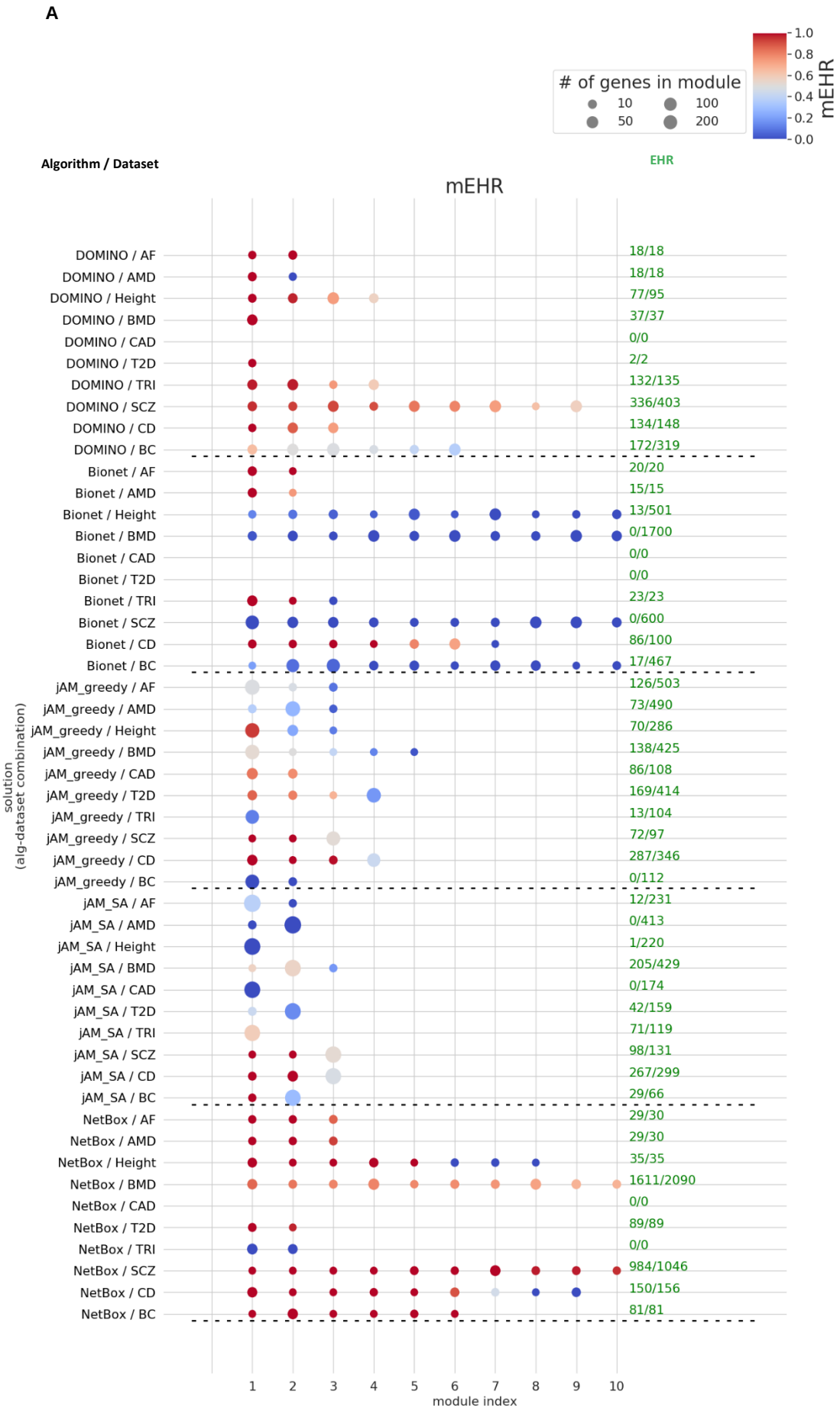
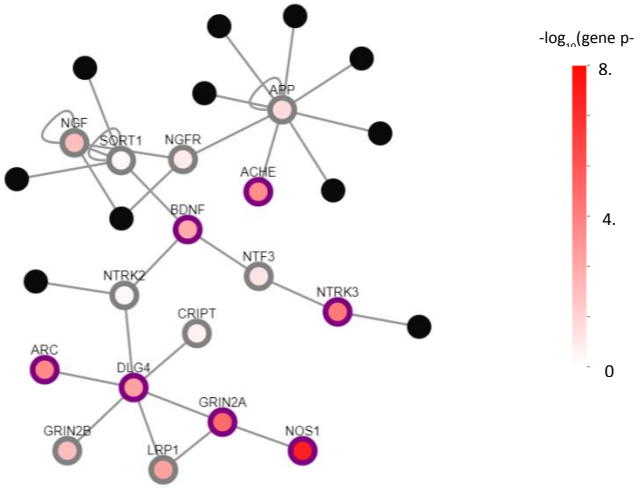


Figure S2. Summary statistics of the solutions obtained on the GWAS datasets. For each dataset, the number of modules detected by each NBMD algorithm and their sizes are indicated. (Error bars represent 1 SD of the number of genes in modules). We excluded empty solutions. Green numbers are the total number of genes of the union of all modules in the solution.



B



Go ID : GO Name

GO:0042133: neurotransmitter metabolic process
GO:0042737: drug catabolic process
GO:0050806: positive regulation of synaptic transmission
GO:0035235: ionotropic glutamate receptor signaling pathway
GO:0007215: glutamate receptor signaling pathway
GO:0060191: regulation of lipase activity
GO:0050807: regulation of synapse organization
GO:0042136: neurotransmitter biosynthetic process
GO:0060291: long-term synaptic potentiation
GO:0050772: positive regulation of axonogenesis
GO:0051963: regulation of synapse assembly
GO:0007612: learning
GO:0098976: excitatory chemical synaptic transmission
GO:0099175: regulation of postsynapse organization
GO:0045471: response to ethanol
GO:0099601: regulation of neurotransmitter receptor activity
GO:1900449: regulation of glutamate receptor signaling pathway
GO:0050905: neuromuscular process
GO:0120035: regulation of plasma membrane bounded cell projection organization
GO:0051965: positive regulation of synapse assembly
GO:0038179: neurotrophin signaling pathway
GO:0031344: regulation of cell projection organization
GO:0050804: modulation of chemical synaptic transmission
GO:0050767: regulation of neurogenesis
GO:0038180: nerve growth factor signaling pathway
GO:0048167: regulation of synaptic plasticity
GO:0007422: peripheral nervous system development
GO:0010975: regulation of neuron projection development
GO:0021675: nerve development
GO:2000310: regulation of NMDA receptor activity
GO:1905606: regulation of presynapse assembly
GO:0048812: neuron projection morphogenesis
GO:0032501: multicellular organismal process
GO:1903539: protein localization to postsynaptic membrane
GO:0043524: negative regulation of neuron apoptotic process
GO:0045664: regulation of neuron differentiation
GO:0048858: cell projection morphogenesis
GO:0120039: plasma membrane bounded cell projection morphogenesis
GO:0032990: cell part morphogenesis
GO:0045595: regulation of cell differentiation
GO:1900271: regulation of long-term synaptic potentiation
GO:0048672: positive regulation of collateral sprouting
GO:1900273: positive regulation of long-term synaptic potentiation
GO:0060041: retina development in camera-type eye
GO:0050885: neuromuscular process controlling balance
GO:0048011: neurotrophin TRK receptor signaling pathway
GO:0007611: learning or memory
GO:0022604: regulation of cell morphogenesis
GO:0048168: regulation of neuronal synaptic plasticity
GO:0007416: synapse assembly

Go ID : GO Name

GO:0042391: regulation of membrane potential
GO:0014068: positive regulation of phosphatidylinositol 3-kinase signaling
GO:0010469: regulation of signaling receptor activity
GO:0050731: positive regulation of peptidyl-tyrosine phosphorylation
GO:0048731: system development
GO:2000116: regulation of cysteine-type endopeptidase activity
GO:0007154: cell communication
GO:0050730: regulation of peptidyl-tyrosine phosphorylation
GO:0045597: positive regulation of cell differentiation
GO:0010942: positive regulation of cell death
GO:0010941: regulation of cell death
GO:0032879: regulation of localization
GO:0032268: regulation of cellular protein metabolic process
GO:0007169: transmembrane receptor protein tyrosine kinase signaling pathway
GO:0051246: regulation of protein metabolic process
GO:0045860: positive regulation of protein kinase activity
GO:0033674: positive regulation of kinase activity
GO:0045859: regulation of protein kinase activity
GO:0023051: regulation of signaling
GO:0042981: regulation of apoptotic process
GO:0007167: enzyme linked receptor protein signaling pathway
GO:0045937: positive regulation of phosphate metabolic process
GO:0009967: positive regulation of signal transduction

Figure S3. Module-level EHR (mEHR) scores on the GWAS datasets. **A**. mEHR scores for each algorithm and GWAS dataset. **B.** An example of a module from the solution reported by DOMINO on the Schizophrenia dataset (mEHR=0.8), and its enriched GO terms. The nodes are color coded by their gene scores as calculated by PASCAL (Lamparter *et al*, 2016) and -$\log_{10}$ transformed. Black nodes are the neighbors of the module's nodes in the network. Nodes with purple border are active nodes ($qval < 0.05$). Red: top 50 EV terms. Blue: enriched HG terms that failed the empirical test.

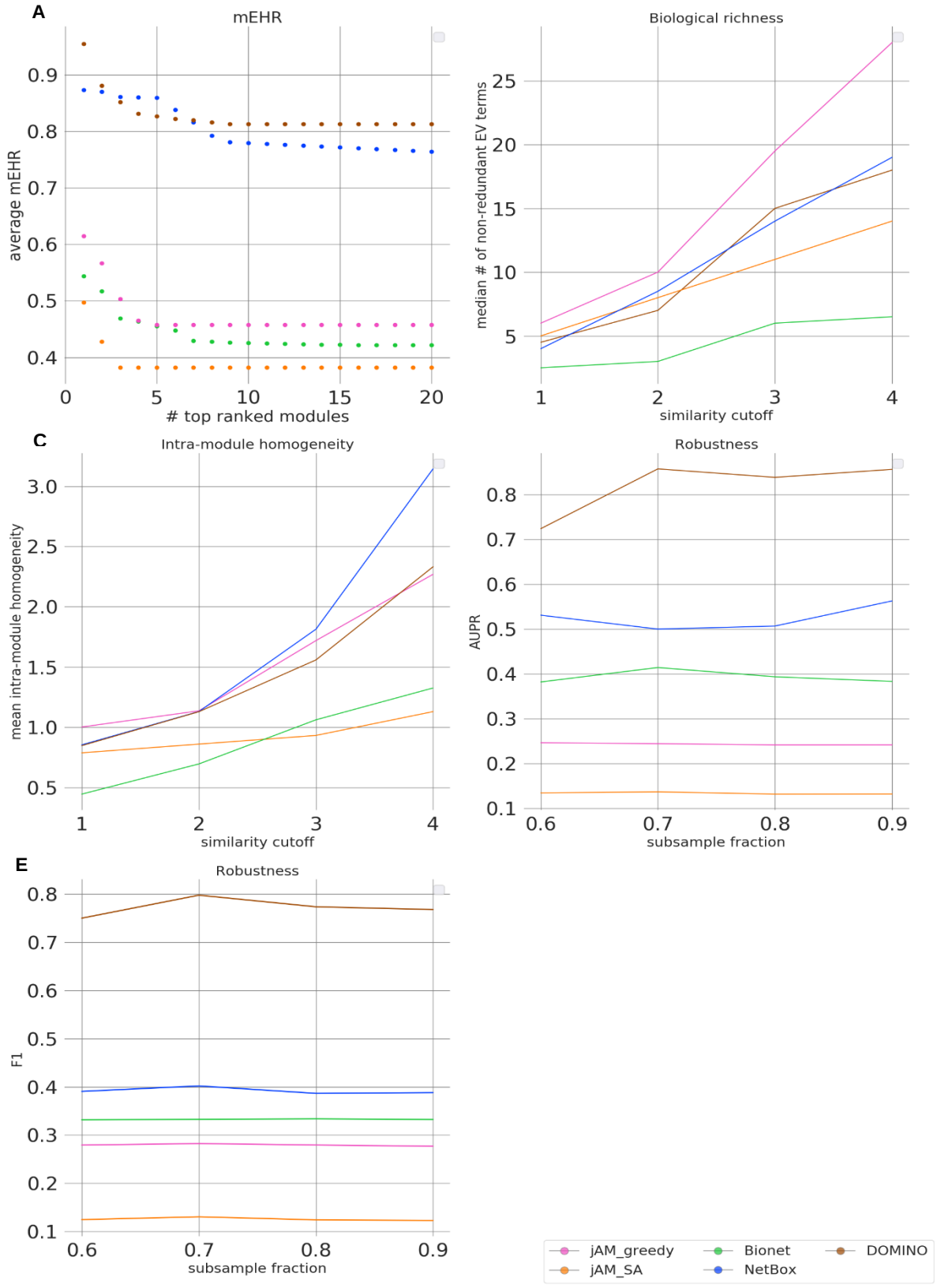


Figure S4. Evaluation results for the GWAS datasets. **A.** Module-level EHR scores. The plots show average mEHR score in the k top modules, as a function of k. Modules are ranked by their mEHR scores. **B.** Biological richness. The plots show the median number of non-redundant terms (richness score) as a function of the Resnik similarity cutoff. **C.** Intra-module homogeneity scores as a function of Resnik similarity cutoff. **D.** Robustness measured by the average AUPR over the datasets, shown as a function of the subsampling fraction. **E.** Robustness measured by the average F1 over the datasets shown as a function of subsample fraction (results for each dataset and fraction were averaged over 100 subsampling).

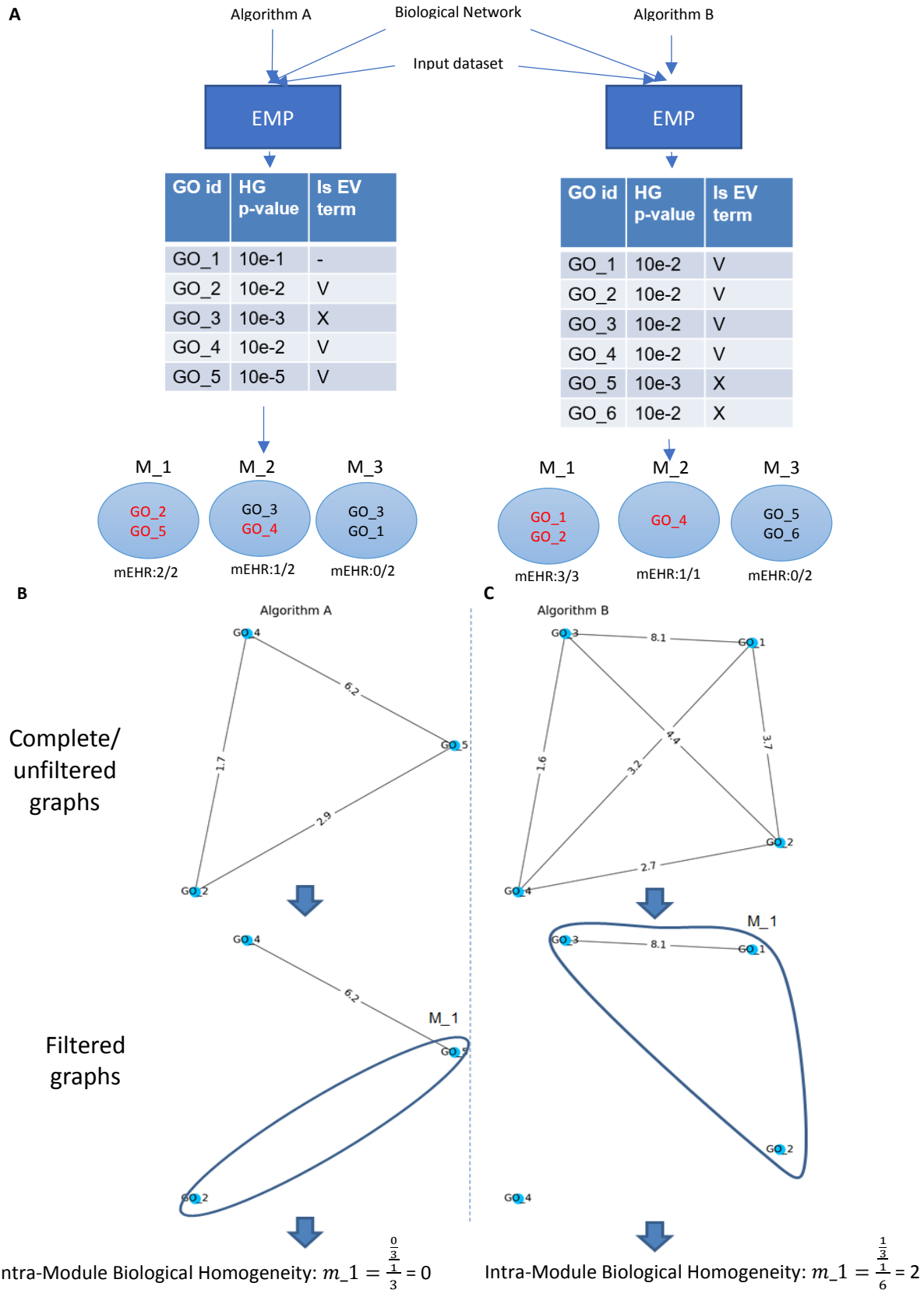


Figure S5. Module-level evaluation criteria. **A. mEHR**. Enriched GO terms in each module are examined by the EMP procedure (EV terms are colored in red) and mEHR is calculated for each module in the solution. **B. Intra-module homogeneity**. A GO graph is first built for the union of all the EV terms in a solution using Resnik similarity scores. Then, a certain cut-off is applied (here, 4.0) for filtering low scoring edges. Last, the intra-module homogeneity score is calculated as the density ratio between the EV terms that are enriched in the module and the entire GO graph

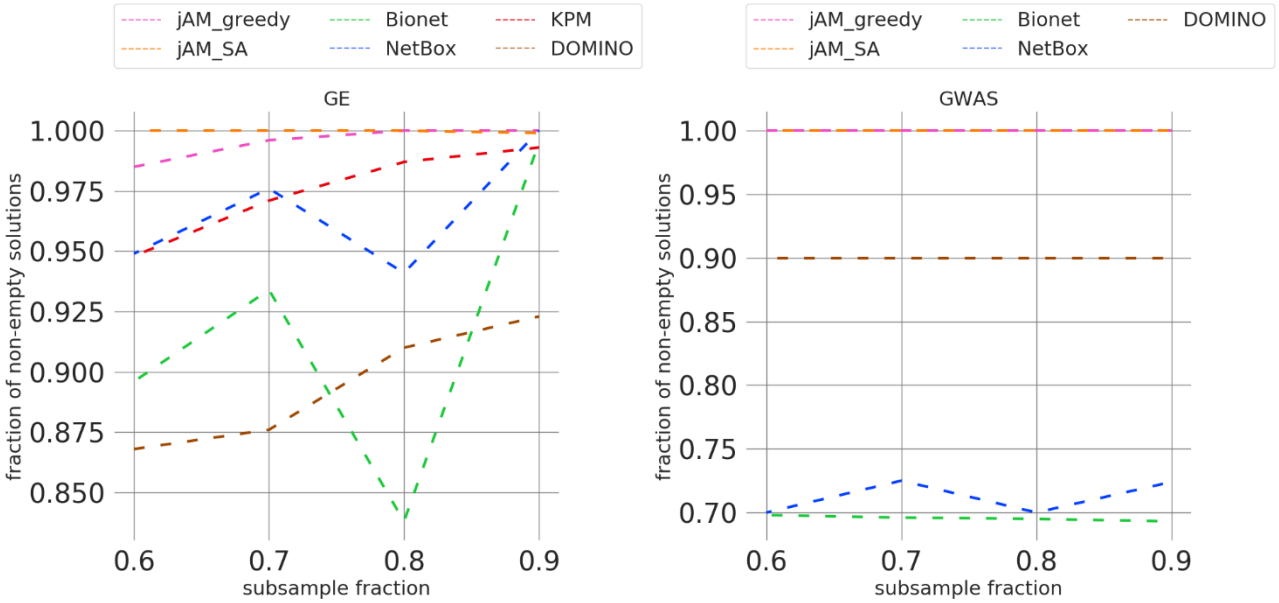


Figure S6. The fraction of non-empty solutions as a function of the subsampling fraction. For each algorithm and subsampling fraction we report the average over the datasets.

Appendix 1: NBMD tools - execution details

The NBDM algorithms that we tested differ in preprocessing, input and output. We describe below the specific execution details for each algorithm.

jActiveModules (Ideker *et al*, 2002). jActiveModules was written as a plugin for Cytoscape (Shannon *et al*, 2003), a powerful platfrom for network analysis of biological data. We modified the codebase of jActiveModules so we could run it independently of Cytoscape. jActiveModules expects a list of genes and their p-values as the gene activity scores. We increased the default number of requested modules (from n=5 to n=50) to retrieve more modules and required that reported modules would be mutually exclusive. The algorithm typically producing no more than 10 modules with more than 3 genes.

NetBox (Cerami *et al*, 2010). We modified NetBox codebase so we can choose the networks it uses. NetBox gets as an input a list of mutated genes, that is, binary gene activity scores. We used the genes' q-values and set the gene score to 1 if its q-value was < 0.05, and 0 otherwise.

BioNet (Beisser *et al*, 2010)**.** BioNet is designed to retrieve only one module. To retrieve multiple mutually exclusive modules we executed BioNet iteratively, removing the genes in the identified module in each iteration. We stopped these iterations after retrieving modules smaller than four genes in five consequtive runs.

HotNet2 (Leiserson *et al*, 2015)**.** HotNet2 expects gene activity scores that are calculated by mutation rate p-values (e.g., using MutSig). We transformed the q-values calculated from our datasets into $-log10(q\_value)$ scale and used them as the input activity scores. We considered all the reported modules, ignoring their scores reported by HotNet2.

KeyPathwayMiner (Baumbach *et al*, 2012)**.** We used the version of KPM with the greedy strategy. It expects binary gene activity scores: 1 marks a gene as active and 0 otherwise. We used the genes' q-values and scored a gene with 1 if its q-value was < 0.05, and 0 otherwise. As the reported modules considerably overlap each other, we executed the algorithm iteratively, removing in each iteration the genes in the identified module.

DOMINO. DOMINO gets as an input a set of active genes, that is, binary gene activity scores. We used the genes' q-values and set the gene score to 1 if its q-value was < 0.05, and 0 otherwise.

The Newman-Girvan (NG) algorithm

The Newman-Girvan method (Girvan & Newman, 2002) iteratively removes edges using the Betweenness-centrality metric for edges. This method iteratively removes edges using the Betweenness-centrality metric for edges. Betweenness-centrality scores each edge according to its frequency in shortest paths between all node pairs. For each node pair, a shortest path is calculated and a score of 1 is added to each edge that appears in the path. For node pairs with multiple shortest paths the score is split evenly among the different paths (e.g. a node pair with two shortest paths will add 0.5 to the score of an edge for of each appearance of the edge in any of the paths). The highest scoring edge is thereafter removed from the graph and the process repeats. In some iterations, the process breaks connected components into smaller ones. The overall solution at iteration i is given a "modularity score", which measures how well-connected are the nodes inside each CC, while being disconnected from nodes in other CCs. M_i is calculated as follows:

$$M_i = \sum_{s=1}^{N_M} \frac{l_s}{L} - \left(\frac{d_s}{2L}\right)^2$$

Where N_M is the number of modules (connected components in the current graph), l_s is the number of edges within modules, L is the total number of edges in the network, and d_s is the sum of the degrees of all nodes within modules. Originally, the algorithm reports the partition that is associated with the highest modularity score. In DOMINO, the process continues until a stopping criterion is met. The stopping criteria we use in DOMINO are:

In step (1) : $M_{i+1} \leq M_i$.

In step (2b): $\frac{log(\#\ of\ nodes\ in\ sub-slice)}{log(\#\ of\ nodes\ in\ network)} \leq M_i$.

References

Aittokallio T & Schwikowski B (2006) Graph-based methods for analysing networks in cell biology. *Brief. Bioinform.* **7:** 243–255

Allen HL, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S, Ferreira T, Wood AR, Weyant RJ, Segrè A V., Speliotes EK, Wheeler E, Soranzo N, Park JH, Yang J, Gudbjartsson D, et al (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467:** 832–838

Ashcroft M, Taya Y & Vousden KH (2000) Stress signals utilize multiple pathways to stabilize p53. *Mol. Cell. Biol.* **20:** 3224–3233

Barrenas F, Chavali S, Holme P, Mobini R & Benson M (2009) Network properties of complex human disease genes identified through genome-wide association studies. *PLoS One* **4:** e8090

Baumbach J, Friedrich T, Kötzing T, Krohmer A, Müller J & Pauling J (2012) Efficient algorithms for extracting biological key pathways with global constraints. In *Proceedings of the genetic and evolutionary computation conference, GECCO 2012* pp 169–176.

Bayerlová M, Menck K, Klemm F, Wolff A, Pukrop T, Binder C, Beißbarth T & Bleckmann A (2017) Ror2 signaling and its relevance in breast cancer progression. *Front. Oncol.* **7:** 135

Beisser D, Klau GW, Dandekar T, Müller T & Dittrich MT (2010) BioNet: An R-Package for the functional analysis of biological networks. *Bioinformatics* **26:** 1129–1130

Benjamini Y & Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57:** 289–300

Boyle EA, Li YI & Pritchard JK (2017) An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169:** 1177–1186

Cerami E, Demir E, Schultz N, Taylor BS & Sander C (2010) Automated network analysis identifies core pathways in glioblastoma. *PLoS One* **5:** e8918

Chen M, Hofestädt R & Taubert J (2019) Integrative bioinformatics: History and future. *J. Integr. Bioinform.* **16:** 20192001

Connelly KE, Weaver TM, Alpsoy A, Gu BX, Musselman CA & Dykhuizen EC (2019) Engagement of DNA and H3K27me3 by the CBX8 chromodomain drives chromatin association. *Nucleic Acids Res.* **47:** 2289–2305

Cowen L, Ideker T, Raphael BJ & Sharan R (2017) Network propagation: A universal amplifier of genetic associations. *Nat. Rev. Genet.***:** 551–562

Creixell P, Reimand J, Haider S, Wu G, Shibata T, Vazquez M, Mustonen V, Gonzalez-Perez A, Pearson J, Sander C, Raphael BJ, Marks DS, Ouellette BFFF, Valencia A, Bader GD, Boutros PC, Stuart JM, Linding R, Lopez-Bigas N, Stein LD, et al (2015) Pathway and network

analysis of cancer genomes. *Nat. Methods* **12:** 615–621

Eden E, Navon R, Steinfeld I, Lipson D & Yakhini Z (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10:** 48

Elkon R, Milon B, Morrison L, Shah M, Vijayakumar S, Racherla M, Leitch CC, Silipino L, Hadi S, Weiss-Gayet M, Barras E, Schmid CD, Ait-Lounis A, Barnes A, Song Y, Eisenman DJ, Eliyahu E, Frolenkov GI, Strome SE, Durand B, et al (2015) RFX transcription factors are essential for hearing in mice. *Nat. Commun.* **6:** 593

Emig D, Ivliev A, Pustovalova O, Lancashire L, Bureeva S, Nikolsky Y & Bessarabova M (2013) Drug target prediction and repositioning using an integrated network-based approach. *PLoS One* **8:** e60618

Forbes TA, Howden SE, Lawlor K, Phipson B, Maksimovic J, Hale L, Wilson S, Quinlan C, Ho G, Holman K, Bennetts B, Crawford J, Trnka P, Oshlack A, Patel C, Mallett A, Simons C & Little MH (2018) Patient-iPSC-derived kidney organoids show functional validation of a ciliopathic renal phenotype and reveal underlying pathogenetic mechanisms. *Am. J. Hum. Genet.* **102:** 816–831

Fritsche LG, Igl W, Bailey JNC, Grassmann F, Sengupta S, Bragg-Gresham JL, Burdon KP, Hebbring SJ, Wen C, Gorski M, Kim IK, Cho D, Zack D, Souied E, Scholl HPN, Bala E, ELee K, Hunter DJ, Sardell RJ, Mitchell P, et al (2016) A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nat. Genet.* **48:** 134–143

Geistlinger L, Csaba G, Santarelli M, Ramos M, Schiffer L, Law C, Turaga N, Davis S, Carey V, Morgan M, Zimmer R & Waldron L (2019) Towards a gold standard for benchmarking gene set enrichment analysis. *bioRxiv***:** 674267

Girvan M & Newman MEJ (2002) Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U. S. A.* **99:** 7821–7826

Hayden MS & Ghosh S (2014) Regulation of NF-κB by TNF family cytokines. *Semin. Immunol.* **26:** 253–266

Hegde C, Indyk P & Schmidt L (2014) A fast, adaptive variant of the Goemans-Williamson scheme for the prize-collecting Steiner tree problem. *Work. 11th DIMACS Implement. Chall.*

Hertzano R, Elkon R, Kurima K, Morrisson A, Chan SL, Sallin M, Biedlingmaier A, Darling DS, Griffith AJ, Eisenman DJ & Strome SE (2011) Cell type-specific transcriptome analysis reveals a major role for Zeb1 and miR-200b in mouse inner ear morphogenesis. *PLoS Genet.* **7:** e1002309

Huang JK, Carlin DE, Yu MK, Zhang W, Kreisberg JF, Tamayo P & Ideker T (2018) Systematic evaluation of molecular networks for discovery of disease genes. *Cell Syst.* **6:** 484-495.e5

Ideker T, Ozier O, Schwikowski B & Siegel AF (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* **18:** S233–S240

Ito T, Teo YV, Evans SA, Neretti N & Sedivy JM (2018) Regulation of cellular senescence by polycomb chromatin modifiers through distinct DNA damage- and histone methylation-dependent pathways. *Cell Rep.* **22:** 3480–3492

Johnson DS, Minkoo M & Phillips S (2000) The prize collecting Steiner tree problem: Theory and practice. *SODA '00 Proc. Elev. Annu. ACM-SIAM Symp. Discret. algorithms***:** 760–769

Kemp JP, Morris JA, Medina-Gomez C, Forgetta V, Warrington NM, Youlten SE, Zheng J, Gregson CL, Grundberg E, Trajanoska K, Logan JG, Pollard AS, Sparkes PC, Ghirardello EJ, Allen R, Leitch VD, Butterfield NC, Komla-Ebri D, Adoum AT, Curry KF, et al (2017) Identification of 153 new loci associated with heel bone mineral density and functional involvement of GPC6 in osteoporosis. *Nat. Genet.* **49:** 1468–1475

Kempe D, Kleinberg J, Tardos É & Tardos E (2015) Maximizing the spread of influence through a social network. *Theory Comput.* **11:** 105–147

Khera A V., Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, Natarajan P, Lander ES, Lubitz SA, Ellinor PT & Kathiresan S (2018) Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50:** 1219–1224

Kroeger H, Grimsey N, Paxman R, Chiang WC, Plate L, Jones Y, Shaw PX, Trejo JA, Tsang SH, Powers E, Kelly JW, Luke Wiseman R & Lin JH (2018) The unfolded protein response regulator ATF6 promotes mesodermal differentiation. *Sci. Signal.* **11:** eaan5785

Kyriakis JM & Avruch J (2012) Mammalian MAPK signal transduction pathways activated by stress and inflammation: A 10-year update. *Physiol. Rev.* **92:** 689–737

Lamparter D, Marbach D, Rueedi R, Kutalik Z & Bergmann S (2016) Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. *PLOS Comput. Biol.* **12:** e1004714

De Lange KM, Moutsianas L, Lee JC, Lamb CA, Luo Y, Kennedy NA, Jostins L, Rice DL, Gutierrez-Achury J, Ji SG, Heap G, Nimmo ER, Edwards C, Henderson P, Mowat C, Sanderson J, Satsangi J, Simmons A, Wilson DC, Tremelling M, et al (2017) Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49:** 256–261

Leiserson MDM, Vandin F, Wu H-T, Dobson JR, Eldridge J V, Thomas JL, Papoutsaki A, Kim Y, Niu B, McLellan M, Lawrence MS, Gonzalez-Perez A, Tamborero D, Cheng Y, Ryslik GA, Lopez-Bigas N, Getz G, Ding L & Raphael BJ (2015) Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **47:** 106–14

Lord PW, Stevens RD, Brass A & Goble CA (2003) Semantic similarity measures as tools for exploring the gene ontology. *Pac. Symp. Biocomput.***:** 601–612

Mahajan A, Taliun D, Thurner M, Robertson NR, Torres JM, Rayner NW, Payne AJ, Steinthorsdottir V, Scott RA, Grarup N, Cook JP, Schmidt EM, Wuttke M, Sarnowski C, Mägi R, Nano J, Gieger C, Trompet S, Lecoeur C, Preuss MH, et al (2018) Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50:** 1505–1513

Marbach D, Lamparter D, Quon G, Kellis M, Kutalik Z & Bergmann S (2016) Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat. Methods* **13:** 366–370

McLendon R, Friedman A, Bigner D, Van Meir EG, Brat DJ, Mastrogianakis GM, Olson JJ, Mikkelsen T, Lehman N, Aldape K, Yung WKA, Bogler O, Weinstein JN, VandenBerg S, Berger M, Prados M, Muzny D, Morgan M, Scherer S, Sabo A, et al (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455:** 1061–1068

Miano V, Ferrero G, Rosti V, Manitta E, Elhasnaoui J, Basile G & De Bortoli M (2018) Luminal lncRNAs regulation by ERα-controlled enhancers in a ligand-independent manner in breast cancer cells. *Int. J. Mol. Sci.* **19:**

Michailidou K, Lindström S, Dennis J, Beesley J, Hui S, Kar S, Lemaçon A, Soucy P, Glubb D, Rostamianfar A, Bolla MK, Wang Q, Tyrer J, Dicks E, Lee A, Wang Z, Allen J, Keeman R, Eilber U, French JD, et al (2017) Association analysis identifies 65 new breast cancer risk loci. *Nature* **551:** 92–94

Mitra K, Carvunis A-RR, Ramesh SK & Ideker T (2013) Integrative approaches for finding modular structure in biological networks. *Nat. Rev. Genet.* **14:** 719–732

Musunuru K & Kathiresan S (2019) Genetics of common, complex coronary artery disease. *Cell* **177:** 132–145

Nelson CP, Goel A, Butterworth AS, Kanoni S, Webb TR, Marouli E, Zeng L, Ntalla I, Lai FY, Hopewell JC, Giannakopoulou O, Jiang T, Hamby SE, Di Angelantonio E, Assimes TL, Bottinger EP, Chambers JC, Clarke R, Palmer CNA, Cubbon RM, et al (2017) Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nat. Genet.* **49:** 1385–1391

Nielsen JB, Thorolfsdottir RB, Fritsche LG, Zhou W, Skov MW, Graham SE, Herron TJ, McCarthy S, Schmidt EM, Sveinbjornsson G, Surakka I, Mathis MR, Yamazaki M, Crawford RD, Gabrielsen ME, Skogholt AH, Holmen OL, Lin M, Wolford BN, Dey R, et al (2018) Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nat. Genet.* **50:**

1234–1239

Pulikkan JA, Hegde M, Ahmad HM, Belaghzal H, Illendula A, Yu J, O'Hagan K, Ou J, Muller-Tidow C, Wolfe SA, Zhu LJ, Dekker J, Bushweller JH & Castilla LH (2018) CBFβ-SMMHC inhibition triggers apoptosis by disrupting MYC chromatin dynamics in acute myeloid leukemia. *Cell* **174:** 172-186.e21

Resnik P (1999) Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *J. Artiicial Intell. Res.* **11:** 95–130

Ripke S, Neale BM, Corvin A, Walters JTR, Farh KH, Holmans PA, Lee P, Bulik-Sullivan B, Collier DA, Huang H, Pers TH, Agartz I, Agerbo E, Albus M, Alexander M, Amin F, Bacanu SA, Begemann M, Belliveau RA, Bene J, et al (2014) Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511:** 421–427

Robinson MD, McCarthy DJ & Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26:** 139–40

Schmidt SF, Larsen BD, Loft A, Nielsen R, Madsen JGS & Mandrup S (2015) Acute TNF-induced repression of cell identity genes is mediated by NFκB-directed redistribution of cofactors from super-enhancers. *Genome Res.* **25:** 1281–1294

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B & Ideker T (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13:** 2498–2504

Stringer S, Wray NR, Kahn RS & Derks EM (2011) Underestimated effect sizes in GWAS: Fundamental limitations of single SNP analysis for dichotomous phenotypes. *PLoS One* **6:** e27964

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES & Mesirov JP (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102:** 15545–50

Sullivan PF & Geschwind DH (2019) Defining the genetic, genomic, cellular, and diagnostic architectures of psychiatric disorders. *Cell* **177:** 162–183

Supek F, Bošnjak M, Škunca N & Šmuc T (2011) REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* **6:** e21800

Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P, Jensen LJ & von Mering C (2017) The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.* **45:** D362–D368

Teslovich TM, Musunuru K, Smith A V., Edmondson AC, Stylianou IM, Koseki M, Pirruccello JP,

Ripatti S, Chasman DI, Willer CJ, Johansen CT, Fouchier SW, Isaacs A, Peloso GM, Barbalic M, Ricketts SL, Bis JC, Aulchenko YS, Thorleifsson G, Feitosa MF, et al (2010) Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466:** 707–713

The Gene Ontology Consortium (2019) The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47:** D330–D338

VanderSluis B, Costanzo M, Billmann M, Ward HN, Myers CL, Andrews BJ & Boone C (2018) Integrating genetic and protein–protein interaction networks maps a functional wiring diagram of a cell. *Curr. Opin. Microbiol.* **45:** 170–179

Visscher PM, Brown MA, McCarthy MI & Yang J (2012) Five years of GWAS discovery. *Am. J. Hum. Genet.* **90:** 7–24

Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT, Maitland A, Mostafavi S, Montojo J, Shao Q, Wright G, Bader GD & Morris Q (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* **38:** W214–W220

Wu G, Dawson E, Duong A, Haw R & Stein L (2014) ReactomeFIViz: a Cytoscape app for pathway and network-based data analysis. *F1000Research* **3:** 146

Xenarios I, Salwínski L, Duan XJ, Higney P, Kim S-M & Eisenberg D (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* **30:** 303–305

Zhang J, Baran J, Cros A, Guberman JM, Haider S, Hsu J, Liang Y, Rivkin E, Wang J, Whitty B, Wong-Erasmus M, Yao L & Kasprzyk A (2011) International cancer genome consortium data portal-a one-stop shop for cancer genomics data. *Database* **2011:** bar026