



Benefits and Challenges of Pre-clustered Network-Based Pathway Analysis

Miguel Castresana-Aguirre, Dimitri Guala and Erik L. L. Sonnhammer*

Department of Biochemistry and Biophysics, Science for Life Laboratory, Stockholm University, Stockholm, Sweden

Functional analysis of gene sets derived from experiments is typically done by pathway annotation. Although many algorithms exist for analyzing the association between a gene set and a pathway, an issue which is generally ignored is that gene sets often represent multiple pathways. In such cases an association to a pathway is weakened by the presence of genes associated with other pathways. A way to counteract this is to cluster the gene set into more homogenous parts before performing pathway analysis on each module. We explored whether network-based pre-clustering of a query gene set can improve pathway analysis. The methods MCL, Infomap, and MGclus were used to cluster the gene set projected onto the FunCoup network. We characterized how well these methods are able to detect individual pathways in multi-pathway gene sets, and applied each of the clustering methods in combination with four pathway analysis methods: Gene Enrichment Analysis, BinoX, NEAT, and ANUBIX. Using benchmarks constructed from the KEGG pathway database we found that clustering can be beneficial by increasing the sensitivity of pathway analysis methods and by providing deeper insights of biological mechanisms related to the phenotype under study. However, keeping a high specificity is a challenge. For ANUBIX, clustering caused a minor loss of specificity, while for BinoX and NEAT it caused an unacceptable loss of specificity. GEA had very low sensitivity both before and after clustering. The choice of clustering method only had a minor effect on the results. We show examples of this approach and conclude that clustering can improve overall pathway annotation performance, but should only be used if the used enrichment method has a low false positive rate.

OPEN ACCESS

Edited by:

Marco Pellegrini,
Italian National Research Council, Italy

Reviewed by:

Tao Huang,
Shanghai Institute of Nutrition and
Health (CAS), China
Pier Luigi Martelli,
University of Bologna, Italy

*Correspondence:

Erik L. L. Sonnhammer
erik.sonnhammer@dbb.su.se

Specialty section:

This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: 15 January 2022

Accepted: 25 April 2022

Published: 10 May 2022

Citation:

Castresana-Aguirre M, Guala D and
Sonnhammer ELL (2022) Benefits and
Challenges of Pre-clustered Network-
Based Pathway Analysis.
Front. Genet. 13:855766.
doi: 10.3389/fgene.2022.855766

Keywords: functional association networks, network clustering, biological mechanisms, pathway enrichment analysis, sensitivity increase

INTRODUCTION

The advance in high throughput experiments has led to a huge increase in the data available for understanding biological function. However, extracting function from high-throughput experiments is often not straightforward since genes and proteins are involved in many different biological mechanisms and pathways. The quest for biological insight from high-throughput experiments has therefore prompted the invention of a large number of pathway enrichment analysis tools.

The most recent family of pathway analysis methods are the network-based tools, such as EnrichNet (Glaab et al., 2012), NEAT (Signorelli et al., 2016), NEArender (Jeggari and Alexeyenko, 2017), BinoX (Ogris et al., 2017), and ANUBIX (Castresana-Aguirre and Sonnhammer, 2020). These

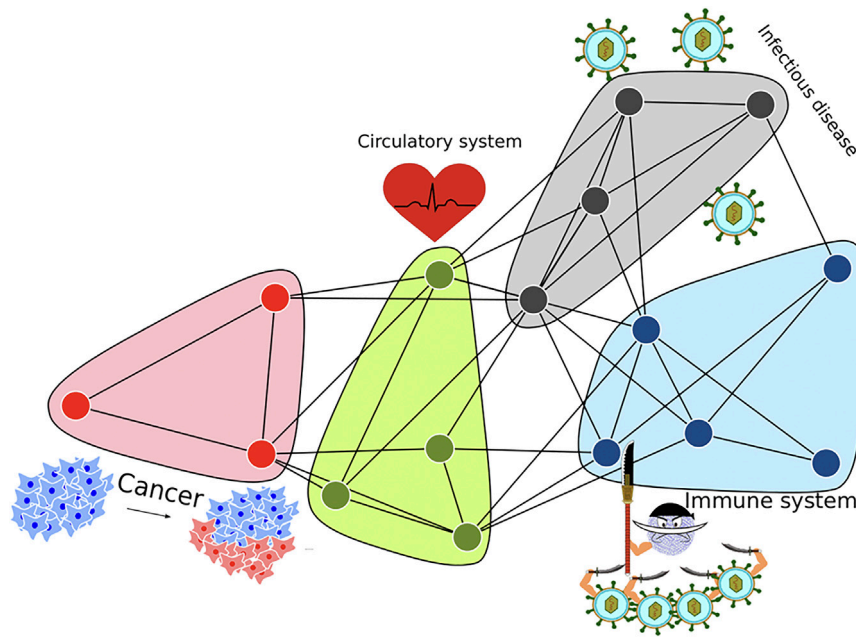


FIGURE 1 | Gene sets derived from experiments are often complex with multiple affected pathways. This illustration shows genes that belong to 4 pathways that are functionally distinct. The mixture of pathways may complicate the pathway enrichment analysis, especially for smaller pathways. By separating gene clusters prior to pathway analysis, a clearer picture of the pathway enrichment can be obtained.

methods require a functional association network, such as FunCoup (Persson et al., 2021) or STRING (Szklarczyk et al., 2021), where different types of data describing relationships between genes and/or proteins, are integrated to infer functional associations between genes. Using enrichment of network links, instead of overlap between gene sets, substantially improves the chances of detecting a relationship, as networks provide much more information (Ogris et al., 2017). Statistical significance of network-based pathway analysis methods is assessed based on the network crosstalk, i.e., links connecting the studied gene set and the pathway of interest. Methods such as BinoX rely on network randomization to obtain a null distribution, which is fit to a binomial distribution to compute the expected crosstalk. NEAT and NEArrender compute the expected crosstalk based on the node degree of the query, the pathway and the network, with the difference that NEAT fits a hypergeometric distribution and NEArrender a chi-square distribution, but their results are very similar. ANUBIX randomly samples gene sets of the same size as the original query set and fits the expected crosstalk to a beta-binomial distribution. While all these methods except ANUBIX have been shown to suffer from high false positive rates when testing random gene sets for enrichment (Castresana-Aguirre and Sonnhammer, 2020), we here included BinoX and NEAT, together with ANUBIX to study how clustering affects different methods.

Network-based methods provide the highest sensitivity of all the pathway enrichment families (Ogris et al., 2017; Castresana-Aguirre and Sonnhammer, 2020). However, experimental gene sets are often complex with multiple affected pathways, which

increases noise and leads to decreased sensitivity. An example of this would be a gene set consisting of four functional modules where each one is enriched for a specific pathway (Figure 1). A pathway analysis method would struggle to detect each module's pathway association if the genes belonging to each module is only a small fraction of all genes in the gene set. Additionally, the studied gene set could contain noise in the form of other genes not related to the main phenotypes of the gene set, which could cause false negatives, impacting the sensitivity of pathway analysis.

Due to the ubiquitous use of pathway analysis methods and reliance on their output to interpret results from diverse and important fields of research such as drug development (Jhamb et al., 2019), biomarker discovery (Chen et al., 2017) and patient diagnosis (Lu et al., 2019), it is important to ensure that these methods can cope well with complex gene sets.

One way to achieve this is to reduce the mentioned complexity by separating the mix of affected pathways. Clustering is a technique that has been used to lower complexity of data by grouping similar entities in various fields, such as pattern recognition (Baraldi and Blonda, 1999; Chen and Huang, 2003), image analysis (Chen et al., 2015; Dhanachandra et al., 2015), and analysis of biological interaction networks (Ideker et al., 2002; Opresko et al., 2004; Mitra et al., 2013). In the field of pathway analysis, clustering is used in PathFindR (Ulgen et al., 2019) and GSCluster (Yoon et al., 2019) to find subnetworks or modules in a gene set mapped to a protein-protein interaction (PPI) network, followed by gene overlap based pathway analysis. However, neither of these tools have evaluated the combination of clustering with state-of-the-art pathway analysis methods, nor

have they compared the performance of used methods with and without clustering.

The approach we take here is applying clustering to decrease complexity of the gene sets, and then apply state-of-the-art network-based pathway enrichment methods. We first investigated whether top-performing clustering methods such as MCL, Infomap, and MGclus are able to extract single pathways from pathway mixtures. The performance of clustering in combination with the network-based pathway analysis methods BinoX, NEAT, and ANUBIX, as well as classical overlap-based Gene Enrichment Analysis (GEA), was evaluated using a benchmark constructed based on the KEGG pathway database.

MATERIALS AND METHODS

Clustering is a way to group objects into different communities, where the objects within each community are more similar to each other than to objects in the other communities (Malliaros and Vazirgiannis, 2013). When clustering is used in the context of a network it involves grouping nodes with high intra-module density, i.e., that are highly connected within a network neighborhood and less connected to the nodes outside said community. There are different types of clustering, e.g., connectivity clustering, centroid clustering, density clustering, distribution clustering, network-based clustering, etc. (Emmons et al., 2016). In our study we focus on network-based clustering, since we are mapping a query gene set onto a network. Since the purpose of this study is not to benchmark the clustering methods themselves, we decided to pick three methods. These methods are MGclus, which has been shown to work well with the FunCoup network (Frings et al., 2013), Infomap (Rosvall and Bergstrom, 2008), and MCL (Van Dongen, 2008), due to their superior performances compared to other methods (Lancichinetti and Fortunato, 2009; Shemirani et al., 2021).

Clustering Methods

MGclus defines modules based on the intra- versus inter-connectivity in a module and considers shared neighbors of nodes as evidence that they belong to the same module.

Both Infomap and MCL extract modules using random walks on the underlying network. MCL performs an iterative random walk along the edges of the network to discover where the flow tends to gather. These iterative random walks are calculated using Markov chains, where the transition probability matrix changes in each run. Infomap finds the optimal set of modules that minimizes the information required to describe a random walk through a network. The description is in two levels, coding for nodes and modules (Rosvall et al., 2009). All clustering algorithms were used with their standard configurations.

Pathway Analysis Tools

GEA is an overlap-based method that tests if the overlap between two sets of genes is higher than would be expected by chance. Statistical significance is assessed using a modified Fisher's exact

test where random overlap is modeled from random samples of pairs of gene sets. This test is a conservative variation of Fisher's exact test, where 1 is subtracted from the observed overlap, as in DAVID's (Huang et al., 2009) EASE score. This means that GEA cannot determine statistical significance of overlaps smaller than 2 nodes.

BinoX assumes that the random crosstalk between two gene sets in the network is distributed according to the binomial distribution. It therefore randomizes the network and computes a distribution of pairs of randomly drawn gene sets to estimate the parameters of a binomially distributed random crosstalk. These parameters are used to determine the expected crosstalk. BinoX can assess whether a pathway is enriched or depleted for the studied gene set. A depleted pathway means that the gene set has fewer links to the pathway than expected by chance.

NEAT and NEArender use slightly different assumptions about the distribution of random crosstalk in the network. NEAT assumes a hypergeometric distribution of crosstalk while NEArender assumes a chi-square distribution. Therefore, instead of testing the observed crosstalk between the studied gene set and a pathway of interest using a sampled random distribution, they rely on the hypergeometric and chi-square test respectively to assess statistical significance. However, both methods compute the expected crosstalk in the same way, taking into account the degree of the gene set, the pathway and the network. Both methods can compute enrichment and depletion. Since NEAT and NEArender show very similar results, we only selected one of them (NEAT) for our benchmark.

ANUBIX is a novel network-based method that computes the enrichment of a gene set for a pathway of interest based on the network crosstalk. The observed crosstalk is assessed for statistical significance using a model of the null distribution of the random crosstalk in the network. This null distribution is modeled by drawing random samples of gene sets, of the same size as the studied gene set, from the genome, calculating their crosstalk with the pathway of interest and fitting the parameters of a beta-binomial distribution for the distribution of the random crosstalk. The procedure can be applied to one or multiple pathways of interest. The statistical significance of the observed crosstalk is only assessed for enrichment, where the observed crosstalk is larger than would be expected by chance.

Null Model Modification of ANUBIX

To generate a null distribution of random crosstalk, ANUBIX samples gene sets from the genome, at random. The assumptions behind this null distribution may be weak when the gene sets under study contain genes not present in the used functional association network or have node degrees that deviate from the expected degrees when drawing random genes. To make the underlying null model more accurate we used degree-aware node sampling (McCormack et al., 2013) to construct the underlying distribution. We achieved this by first grouping all network nodes into bins, one per degree if more than 100 nodes exist for a given degree, or bins representing a range of degrees if this was needed to obtain at least 100 nodes in the bin. Sampling to produce

random gene sets was done by randomly selecting nodes from bins with the same degree as the nodes in the query set.

To assess the improvement of this modification, we generated 100 random gene sets by sampling from the whole genome and another 100 random gene sets by sampling from the subset of genes present in all Chemical and Genetic interaction (CGP) gene sets in the Molecular Signatures Database (MSigDB) (Liberzon et al., 2011). Sampling was done such that the gene frequencies in the MSigDB gene sets were preserved. The size of the gene sets was fixed to 50 genes, which was the median size of all the gene sets in MSigDB.

Functional Association Network

Network-based pathway enrichment methods require a protein interaction network. In our study we used FunCoup, which is one of the most comprehensive functional association networks of genes/proteins available. FunCoup infers functional associations between genes by integrating different types of evidence using a redundancy-weighted naïve Bayesian approach, combined with orthology transfer. FunCoup's high coverage comes from the number and variety of different evidence types used, such as: mRNA and protein co-expression, co-evolution based on phylogenetic profile similarity, Protein-Protein and domain-domain interactions, sub-cellular co-localization, co-regulation via miRNA and transcription factors, as well as genetic interaction.

For this study, we used the *Homo sapiens* FunCoup 5 network. To avoid noise, we used the default link confidence cutoff of 0.8 resulting in a network of 612,276 links and 12,890 genes.

Pathway Database

For this study we use the 313 *H. sapiens* pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (v.96.0) (Kanehisa et al., 2016).

BENCHMARKS

Pathway Recovery for Each Clustering Method

Performance of clustering algorithms may vary depending on the properties of the network they are applied to, so we constructed a simple benchmark to assess this. We generated 100 gene sets by merging different KEGG pathways that had shared links, three pathways at a time. Then we applied the different clustering methods to these gene sets to produce modules. Each module was assigned to the pathway with the highest overlap, and the Jaccard index between the sets of assigned and true pathways was computed for each method. The Jaccard index distributions of the clustering methods were compared using Kruskal-Wallis and Wilcoxon tests.

True Positive Benchmark

KEGG pathways were bisected into two parts with similar number of nodes and total node degree. The overlap between the bisected parts was emulated based on the median overlap between gene sets in the MSigDB database and KEGG pathways.

KEGG pathways were ordered by size and grouped into seven bins with an equal (or as equal as possible) number of pathways in each bin. We then sampled one pathway from each bin at random and merged them into a unique gene set. To decide how many pathways to join, we performed a pathway analysis study of Chemical and Genetic interaction (CGP) MSigDB gene sets against KEGG pathways using the null model modified ANUBIX. To keep a reasonable gene set size, and to avoid merging too many pathways, we used Bonferroni correction (Abdi, 2007) and a family-wise error rate (FWER) of 5% as a cutoff. This resulted in a median number of significantly enriched pathways of seven per gene set. We therefore chose to join seven pathways for the construction of the multi-pathway gene sets. Since our sampling was constrained by the binning procedure, to avoid having too much overlap between the constructed gene sets, but still retain a statically usable number of gene sets we generated 100 gene sets and ran pathway enrichment against the other parts of the bisected pathways. Since each gene set was constructed from seven different pathways and we were aiming to recover the other half of each of those pathways, we could at most have 700 true gene set-pathway associations or True Positives (TPs).

False Positive Benchmark

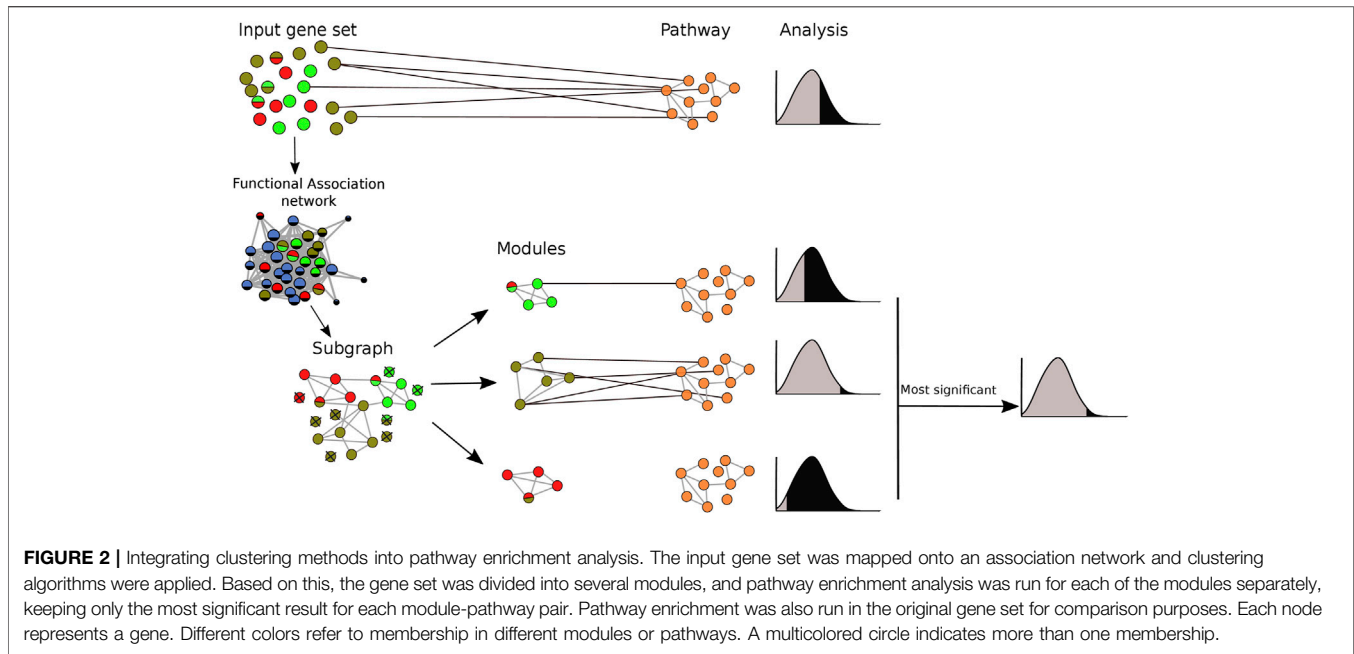
For the false positive (FP) benchmark we generated 100 random gene sets of the average size of the true positive gene sets, 280 genes. The generated gene sets were tested for enrichment against the true KEGG pathways. Considering their randomness, we did not expect to find any enriched pathways.

Performance Measures

Both the true positive and false positive benchmarks were applied with and without clustering of gene sets prior to pathway analysis. When clustering was applied, pathway enrichment was tested individually for each identified module. The pathways with the lowest *p*-value for each module were merged into a single list. The performance of each method was assessed by Receiver Operator Characteristics (ROC) curves (Bradley, 1997). For our analysis, we select only the pairs that were statistically significantly (FDR < 0.05) enriched after adjusting *p*-values using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). The pipeline of the clustering implementation in pathway enrichment analysis is shown in **Figure 2**.

Adaptive Module Size Filtering

Applying clustering to the query gene sets increases the sensitivity of the underlying analysis. However, this often comes with an increase in false positives, mainly stemming from small modules. To control for this, we devised a filtering approach for small modules prior to the pathway enrichment analysis. To calibrate it, we generated 100 random gene sets for a range of sizes between 50 and 600 genes, increasing the size by 50 genes, and ran the clustered pathway enrichment pipeline against KEGG pathways. At FDR < 0.05, we studied which minimum module size cutoff was necessary to keep the FPR below 5%. With the selected range of gene set sizes, we observed that the required module size cutoff increased linearly with the query gene set size (**Supplementary Figure S1**), suggesting that the cutoff should be adapted to



different gene set sizes. This approach only works well for methods that already control the FPR well prior to clustering, here yielding good results only for ANUBIX. The adaptive module size filtering ensures an FPR level matching the set FDR level in ANUBIX when filtering out modules whose size is below 2% of the query gene set size, hence this filter was applied to ANUBIX here. For BinoX and NEAT this was however not possible to achieve without a massive loss of sensitivity, hence the filter could not be applied to them.

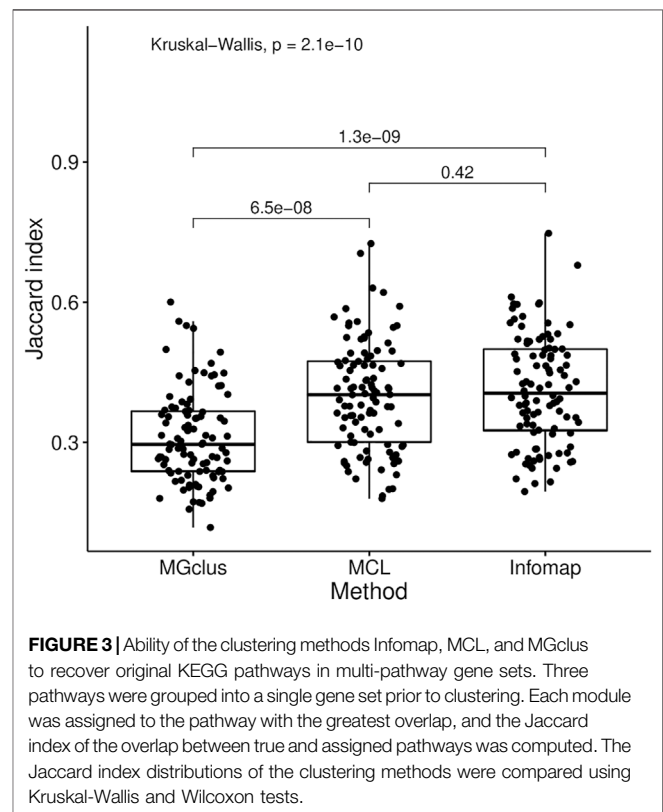
Clustered vs. Non-Clustered MSigDB Gene Sets Analysis

We ran pathway enrichment analysis against KEGG pathways for all the CGP MSigDB gene sets in two different scenarios, with and without pre-clustering the gene sets. To showcase that different gene sets are a mixture of different pathway or pathway families, for each MSigDB gene set, we studied how often a certain pathway subclass, as defined by KEGG, was targeted by the same gene set module. The KEGG database classifies pathways into 6 classes and 42 subclasses. The overlap in significantly enriched pathways between (A) with pre-clustering and (B) without pre-clustering was computed using the Jaccard Index as described in Eq. 1:

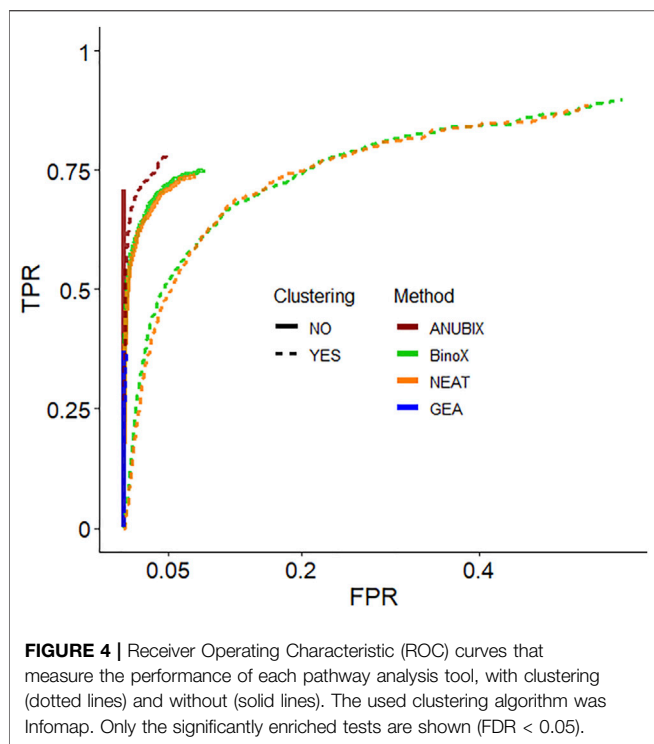
$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

RESULTS

Gene sets derived from experiments typically represent multiple affected pathways. Therefore, mapping these gene sets onto a



network such as FunCoup and applying network-based clustering algorithms to divide gene sets into more homogeneous subsets was expected to reduce noise and lead to more accurate pathway analysis. We investigated the effect of clustering on pathway



analysis using MGclus, MCL, and Infomap. To assess the clustering performance of these methods on data used in pathway analysis we applied them to gene sets constructed by joining multiple KEGG pathways. Infomap and MCL demonstrated the greatest ability to recover the original pathways with a mean Jaccard index of 41.2% for Infomap and 39.9% for MCL, followed by MGclus at 31% (Figure 3). The difference between Infomap and MCL was not significant ($p = 0.42$), however both Infomap and MCL were significantly different from MGclus, with $p = 1.3 \times 10^{-9}$ and $p = 6.5 \times 10^{-8}$, respectively.

The original null model of ANUBIX is suitable to capture non-randomness in pathways. However, it may not optimally handle biases present in the query gene set such as genes that are not in the network or genes with very high node degrees. To account for these biases and make the null model more strict we improved the random sampling step to take into account the degree distribution of the query genes. To assess the modified null model generation procedure we created two datasets of random gene sets: one by sampling from the whole genome, and another by sampling from the pool of genes present in the MSigDB CGP gene sets. For the first dataset, both the original and the null model modified ANUBIX had 0% FPR. However, for the second dataset the original ANUBIX had an FPR of 6.6%, while the FPR of the null model modified ANUBIX was only 0.2%.

We then devised a benchmark to show the effect of pre-clustering of query gene sets. The first part of the benchmark was intended to assess the ability to recover True Positive gene set-pathway pairs. Construction of the benchmark involved bisecting KEGG pathways, merging the first half of several pathways into a heterogeneous gene set and trying to detect

enrichment between this gene set and the other bisected halves. In the second part of the benchmark we simulated False Positive gene set-pathway associations by generating random gene sets of the average size of the true positive gene sets. We then assessed the performance of pathway analysis methods: ANUBIX, BinoX, NEAT, and GEA, with, and without pre-clustering on this benchmark. Figure 4 shows the results as a Receiver Operating Characteristic (ROC) curve for MCL and all pathway analysis algorithms. ROC curves when clustering by Infomap and MGclus are in Supplementary Figure S2. The ROC curves only show the statistically significant results at FDR < 0.05, and only for enrichment (i.e. not depletion).

Detailed True Positive Rate (TPR) and False Positive Rate (FPR) results are shown in Table 1. The best balanced performance prior to the application of clustering was demonstrated by ANUBIX, with a TPR of 71% and a FPR of 0%. BinoX and NEAT showed higher TPRs, of 75% and 74% respectively, but had a much higher FPR of 9% and 8%, respectively. As expected, GEA had a low TPR of only 37% due to the low coverage that overlap-based methods tend to have. However, it had a flawless specificity. A significant difference was observed between the results of ANUBIX and the other methods (McNemar's test, $p < 0.001$).

When applying clustering of the gene sets prior to pathway analysis, we observed a statistically significant (McNemar's test, $p < 0.001$) increase in TPR for all the network-based pathway enrichment methods ANUBIX, BinoX, and NEAT, but not for GEA, which decreased. The TPR for ANUBIX increased by at most 7 percentage points, when using Infomap, still maintaining an FPR not exceeding the requested FDR level of 5%. BinoX and NEAT exhibited higher increases in TPR of up to 14–15 percentage points. However, this increase came with a very high increase in FPR from 9% to 56–61% for BinoX and from 8% to 52–56% for NEAT. There is a significant difference between the results of the other methods and ANUBIX for all the clustering algorithms ($p < 0.001$).

We observed that almost all of the enrichments found without clustering were also found using pre-clustering of the query sets (Figure 5). For BinoX and NEAT the fraction of unique enrichments found without clustering were the lowest, below 2%, while for GEA they were the highest at 12–15%. Looking at enrichments only found by pre-clustering, these fractions were generally higher, 8–17%. We further noted that most of the associations, 99.6%, identified by GEA were also found by the network-based methods.

TABLE 1 | True positive rate (TPR) and false positive rate (FPR) for combinations of the clustering and pathway enrichment methods run at FDR = 0.05.

	ANUBIX		BinoX		NEAT		GEA	
	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
No clustering	0.71	0.00	0.75	0.09	0.74	0.08	0.37	0.00
MCL	0.73	0.03	0.90	0.57	0.88	0.53	0.35	0.00
MGclus	0.75	0.03	0.88	0.61	0.88	0.56	0.35	0.00
Infomap	0.78	0.05	0.90	0.56	0.88	0.52	0.36	0.00

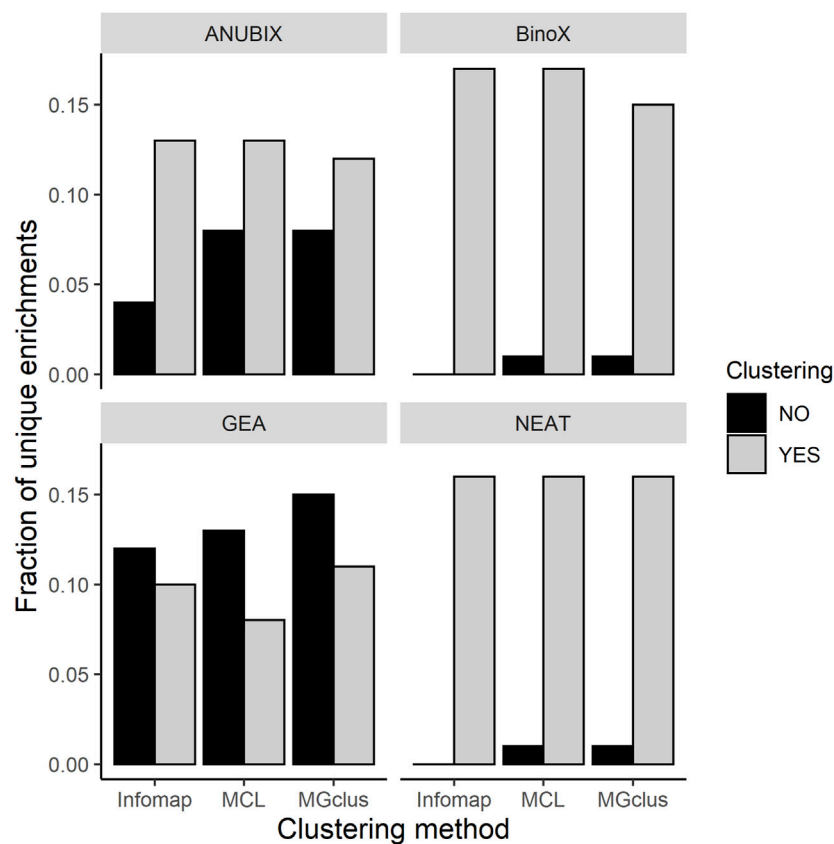


FIGURE 5 | Fractions of unique pathway enrichments found with pre-clustering relative to without pre-clustering, and vice versa, run at FDR = 0.05 for all the combinations of clustering methods and pathway enrichment tools.

Clustered Versus Non-Clustered Gene Sets Analysis

A large-scale analysis was carried out for 3302 gene sets from MSigDB/CGP against the 313 human pathways in KEGG, to observe possible benefits of applying clustering to experimental gene sets. Clustering was applied using Infomap and ANUBIX was used for the pathway enrichment analysis. Pathway enrichment analysis web server tools, such as PathBIX (Castresana-Aguirre et al., 2021) or PathwAX (Ogris et al., 2016), are implemented in a way that allows only single gene set queries. By analogy, we studied MSigDB gene sets by assuming independence between gene sets, i.e., multiple testing correction was only performed for the number of pathways each query is compared to.

Clustering of MSigDB gene sets occurred in 2703 of the 3302 gene sets. Pathway analysis without pre-clustering resulted in 129,044 significant (FDR < 0.05) crosstalks across 2,222 gene sets. Clustered analysis produced 122,819 significant crosstalks for 2,178 gene sets, of which 1,890 were shared with the non-clustering approach. The Jaccard index overlap (see Materials and Methods) of significant crosstalks between clustering and non-clustering was 52.5%, and 67.2% of the non-clustering crosstalks were found by the clustering approach while 70.6% of the clustering crosstalks were found by non-clustering.

To show that clustering helps to isolate different mechanisms within a gene set, we used the pathway subclasses as defined in the KEGG database and mapped them to the significant pathway crosstalks from the MSigDB large-scale analysis. Each pathway belongs to a KEGG subclass, and on average 95% of the significant pathways of a certain subclass had crosstalk to just one module in a gene set.

An Application of Clustered Pathway Enrichment Analysis

To illustrate the usefulness of clustering we provide an example with an MSigDB gene set, HAHTOLA_SEZARY_SYNDROM_UP (Hahtola et al., 2006). More examples can be found in **Supplementary File S1** where we provide all significant pathway enrichments found by pre-clustering using ANUBIX and Infomap but not without clustering. The selected example query set contains 99 up-regulated genes (**Supplementary Table S1**) from peripheral blood samples of Sezary syndrome patients compared to samples from healthy donors. Sezary syndrome is an aggressive form of cutaneous T-cell lymphoma (<http://ghr.nlm.nih.gov/condition/sezary-syndrome>) and is a rare disease driven by cancerous T-cells with one or several chromosomal

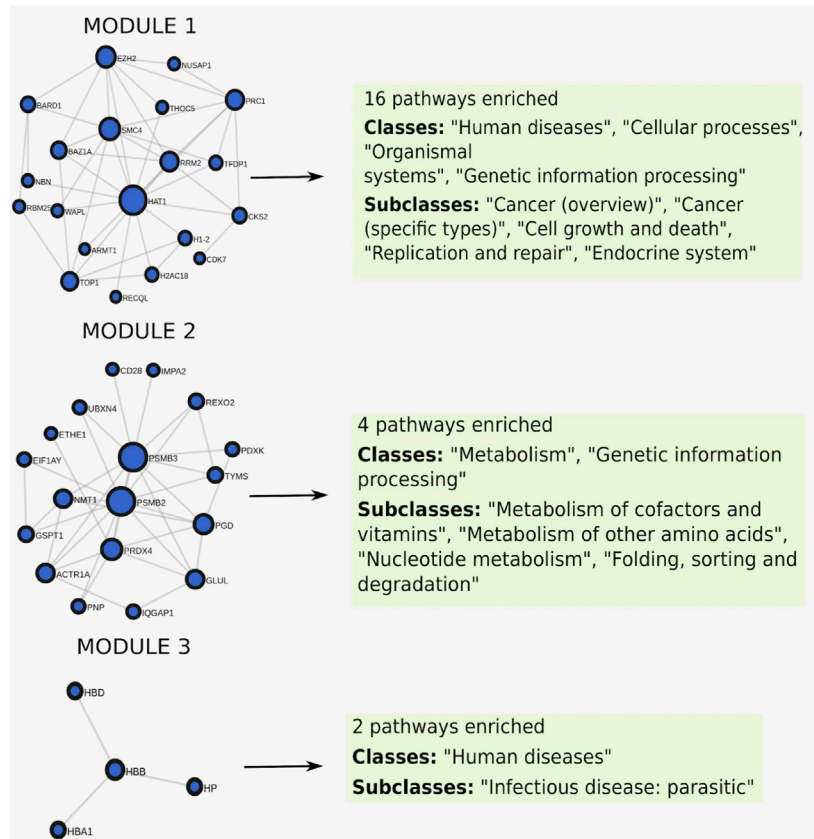


FIGURE 6 | Clustered pathway enrichment analysis of the MSigDB gene set HAHTOLA_SEZARY_SYNDROM_UP. The gene set is divided into 3 modules by applying the network clustering algorithm Infomap. Each module finds different classes of pathways.

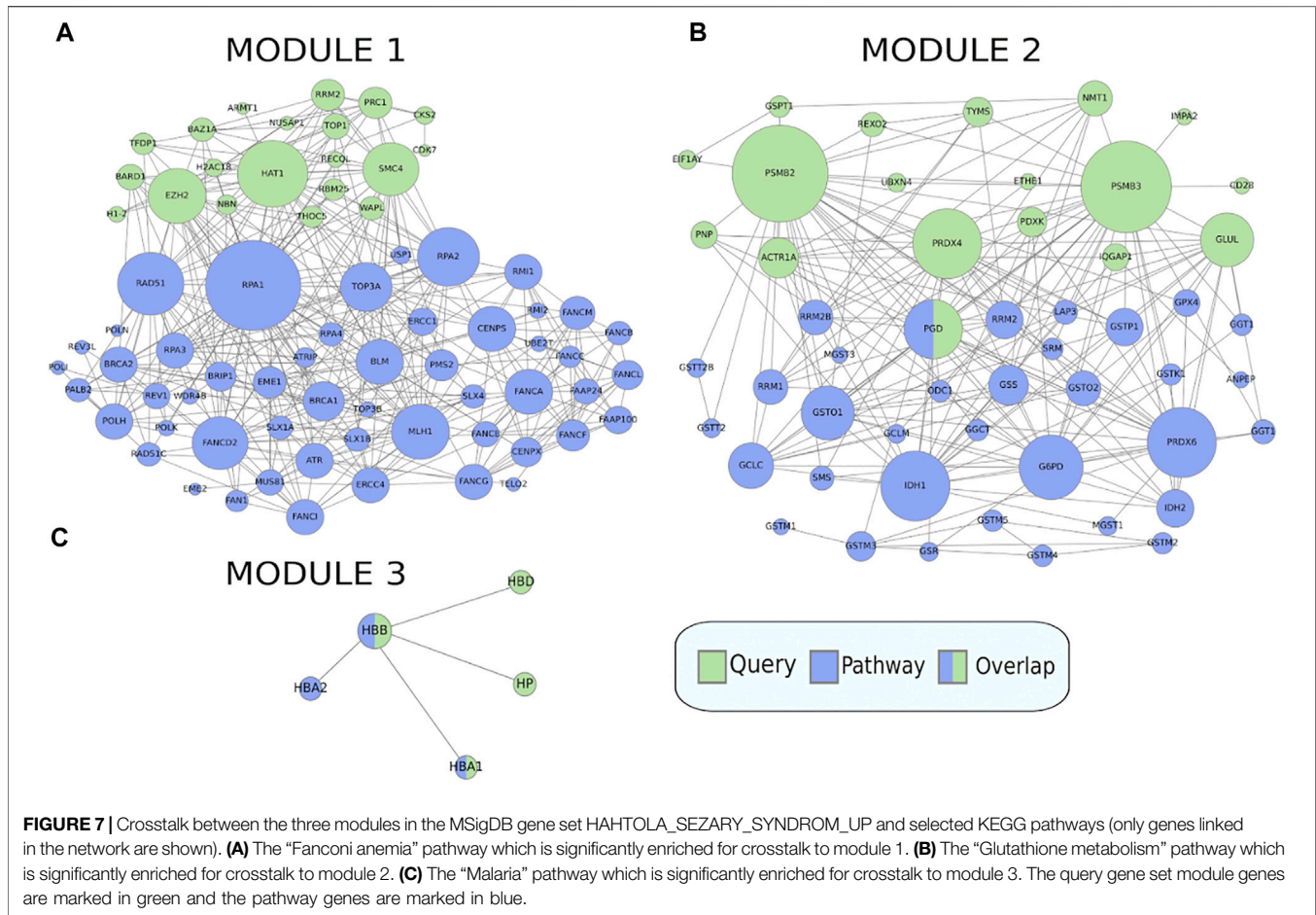
abnormalities. We used the web-server PathBIX, which provides both regular ANUBIX and clustered ANUBIX. We ran this gene set against the KEGG pathway database with a FunCoup cutoff of 0.8 and compared the results obtained from non-clustering and clustering. At $FDR < 0.05$, non-clustering finds 8 significantly enriched pathways, full results in **Supplementary Table S2**. The top seven pathways belonged to the KEGG classes of "Replication and repair" and "Cell growth and death", which are pathway classes affected by cancer. The eighth was the "Human T-cell leukemia virus 1 infection" pathway at $FDR = 0.01$. As opposed to the other seven unspecific cancer related pathways, the last one has been associated with Sezary syndrome (Pancake et al., 1995).

When clustering was applied to this gene set, it was split into three modules of size 20, 18, and 4, where each module was enriched for 16, 4, and 2 pathways respectively (**Figure 6**), full results in **Supplementary Table S3**. The first module retrieved all the enriched pathways found by the non-clustering approach, while finding additional enriched pathways belonging to the same pathway classes as the pathways found by non-clustering. Pathways relevant to cancer included "Fanconi anemia" (**Figure 7A**) at $FDR = 2.8e-3$, a bone marrow failure syndrome whose complications can result in leukemia (Cheung and Taniguchi, 2017), due to a failure in the repair of DNA interstrand crosslinks in the genome (Ceccaldi et al.,

2016). The first module was further enriched in other cancer related pathways, such as "Transcriptional misregulation in cancer" at $FDR = 1.77e-3$. Furthermore, it was enriched in the "Viral carcinogenesis" pathway ($FDR = 0.01$). This pathway includes genes targeted by the Human T-cell leukemia virus 1 (HTLV1 virus), which is thought to be the potential trigger for Sezary syndrome. This is as relevant as the HTLV1 infection pathway identified by the non-clustering approach.

The second module finds pathways belonging to the metabolism class, such as "Glutathione metabolism" (**Figure 7B**) at $FDR = 0.02$, which is reasonable as glutathione has been proven to effectively block cell death in primary T cells from Sezary patients (Kiessling et al., 2009). Other metabolism pathways like "Purine metabolism" at $FDR = 0.03$, and "One carbon pool by folate" at $FDR = 0.03$, are reasonable as purine and folate are potential therapeutic drugs for Sezary syndrome (Oka and Miyagaki, 2019).

The third module finds pathways belonging to the class of parasitic infectious diseases, with "Malaria" at $FDR = 3.79e-3$ (**Figure 7C**) and "African trypanosomiasis" at $FDR = 8.72e-4$. Biomarkers such as miRNA are used for detecting infectious diseases. In malaria, some of the most expressed miRNAs are miR451 and miR92 (Babatunde et al., 2018), where the former is significantly correlated with diagnosis and prognosis of Sezary



syndrome, and the latter is downregulated in it (Narducci et al., 2011).

DISCUSSION

This study aimed at assessing the added benefit of pre-clustering gene sets prior to conducting pathway enrichment analysis. In order to achieve this we evaluated combinations of three network clustering methods in conjunction with one overlap-based and three network-based pathway analysis algorithms. Our findings indicate that pre-clustering increases sensitivity of pathway analysis with network-based methods but observed that it comes with the challenge of risking a high false positive rate. For two of these methods, the improvement in sensitivity came with an unacceptable loss of specificity. However, ANUBIX was able to substantially increase the sensitivity while keeping a high specificity.

The large-scale application of ANUBIX with clustering to the MSigDB gene sets against all KEGG pathways resulted in a similar number of significant enrichments as when no clustering was applied, but about a third of the enrichments were unique to each approach. We further observed that each network module within a gene set tended to be enriched by a different subclass of pathways. This supports the hypothesis that experimentally

derived gene sets often represent mixtures of genes with different mechanisms, and isolating these provides a more informative analysis of the different mechanisms that are related to the condition under study. In this analysis we used Infomap for clustering as it was the best method in the benchmarks, and for the pathway enrichment analysis we used ANUBIX since it outperformed the other methods.

Before the pre-clustering analysis, we introduced a modification to the null model of ANUBIX. The new null model of ANUBIX evaluated in the study uses degree-aware sampling of genes in the network instead of randomly sampling genes from the whole genome. This null model modification resulted in a lower FPR compared to the original implementation, hence the modified version of ANUBIX was used in the rest of this study.

A previous benchmark showed that BinoX and NEAT suffer from a relatively high false positive rate (Castresana-Aguirre and Sonnhammer, 2020). To compute the crosstalk between a query gene set and a pathway, BinoX randomizes the network leading to a loss of the internal pathway structure. NEAT does not randomize the network to assess statistical significance but relies on the degrees of the query gene set, pathway, and the whole network, regardless of how that degree is distributed across the pathway. It has been demonstrated that there is a correlation between the FPs of these network-based methods and the fraction of intralinks of the

pathways (Castresana-Aguirre and Sonnhammer, 2020), meaning that the less random the pathway topology is, the more prone it is to produce FPs. The distribution of crosstalk between a random gene set and a pathway often suffers from overdispersion, i.e., when the variance is larger than the mean. When this happens, the null distributions of crosstalk assumed by the different methods, binomial (BinoX) or hypergeometric (NEAT), are not appropriate. Both the overdispersion and the high false positive rate are resolved by ANUBIX. Instead of randomizing the whole network which distorts the pathway structure, ANUBIX assesses statistical significance by sampling random gene sets of the same size as the query gene set and computing an expected crosstalk distribution for each pathway. The resulting null distribution is fitted to a beta-binomial distribution, which has been demonstrated to accurately capture overdispersion (Young-Xu and Chan, 2008), and this is used to assess the significance of an observed crosstalk. Even though ANUBIX is the best performing method in that benchmark, we wanted to include other network-based methods to study if clustering could decrease their FPR. However, this issue became even more apparent when clustering was applied. We further observed that the average degree in the unclustered ANUBIX FP gene sets was 82 while the average degree of the genes in FP modules generated from those gene sets increased significantly ($p < 0.001$) to 150, 161, and 193 for Infomap, MCL and MGclus respectively. Statistical significance was assessed using a permutation test by computing the average degree for 2,000 data sets with 100 gene sets in each.

For this benchmark, we did not include quantitative pathway analysis tools, such as GSEA (Subramanian et al., 2005), CAMERA (Wu and Smyth, 2012) or SPIA (Tarca et al., 2009). In order to work, these methods require as input the differential expression of all genes. Several limitations were described previously (Subramanian et al., 2005) when selecting subsets of genes from such a list. Thus, clustering the whole set of genes into independent subsets is unlikely to be beneficial for these methods.

We have demonstrated that the application of clustering of query gene sets prior to pathway analysis improves the sensitivity of all studied pathway enrichment methods, and helps to elucidate complex mechanisms within an experimental gene set. However, pre-clustering is recommended to be used

primarily with methods that can control the false positive rate well. The approach finds almost all associations found without clustering, while adding many new ones, and thus represents a powerful new tool in the quest for more accurate pathway analysis.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

MC-A: Conceptualization, Software, Data curation, Writing—Original Draft DG: Writing—Original Draft ES: Conceptualization, Writing—Original Draft.

FUNDING

This work was supported by the Swedish Research Council Project (2015- 05342). Open access funding provided by Stockholm University.

ACKNOWLEDGMENTS

We thank the Swedish Research Council for financial support, and Christoph Ogris and Sam de Meyer for preliminary studies of this subject.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2022.855766/full#supplementary-material>

REFERENCES

- Abdi, H. (2007). Bonferroni and Šidák Corrections for Multiple Comparisons. *Encycl. Meas. statistics* 3, 103–107.
- Babatunde, K. A., Mbagwu, S., Hernández-Castañeda, M. A., Adapa, S. R., Walch, M., Filgueira, L., et al. (2018). Malaria Infected Red Blood Cells Release Small Regulatory RNAs through Extracellular Vesicles. *Sci. Rep.* 8, 884. doi:10.1038/s41598-018-19149-9
- Baraldi, A., and Blonda, P. (1999). A Survey of Fuzzy Clustering Algorithms for Pattern Recognition. I. *IEEE Trans. Syst. Man. Cybern. B* 29, 778–785. doi:10.1109/3477.809032
- Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* 57, 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x
- Bradley, A. P. (1997). The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recognit.* 30, 1145–1159. doi:10.1016/s0031-3203(96)00142-2
- Castresana-Aguirre, M., Persson, E., and Sonnhammer, ELL. (2021). PathBIX—a Web Server for Network-Based Pathway Annotation with Adaptive Null Models. *Bioinform. Adv.* 1, vbab010. doi:10.1093/bioadv/vbab010
- Castresana-Aguirre, M., and Sonnhammer, E. L. L. (2020). Pathway-specific Model Estimation for Improved Pathway Annotation by Network Crosstalk. *Sci. Rep.* 10, 13585. doi:10.1038/s41598-020-70239-z
- Ceccaldi, R., Sarangi, P., and D'Andrea, A. D. (2016). The Fanconi Anaemia Pathway: New Players and New Functions. *Nat. Rev. Mol. Cell Biol.* 17, 337–349. doi:10.1038/nrm.2016.48
- Chen, T., He, P., Tan, Y., and Xu, D. (2017). Biomarker Identification and Pathway Analysis of Preeclampsia Based on Serum Metabolomics. *Biochem. Biophysical Res. Commun.* 485, 119–125. doi:10.1016/j.bbrc.2017.02.032
- Chen, X.-W., and Huang, T. (2003). Facial Expression Recognition: A Clustering-Based Approach. *Pattern Recognit. Lett.* 24, 1295–1302. doi:10.1016/s0167-8655(02)00371-9
- Chen, Z., Qi, Z., Meng, F., Cui, L., and Shi, Y. (2015). Image Segmentation via Improving Clustering Algorithms with Density and Distance. *Procedia Comput. Sci.* 55, 1015–1022. doi:10.1016/j.procs.2015.07.096

- Cheung, R. S., and Taniguchi, T. (2017). Recent Insights into the Molecular Basis of Fanconi Anemia: Genes, Modifiers, and Drivers. *Int. J. Hematol.* 106, 335–344. doi:10.1007/s12185-017-2283-4
- Dhanachandra, N., Manglem, K., and Chanu, Y. J. (2015). Image Segmentation Using K -means Clustering Algorithm and Subtractive Clustering Algorithm. *Procedia Comput. Sci.* 54, 764–771. doi:10.1016/j.procs.2015.06.090
- Emmons, S., Kobourov, S., Gallant, M., and Börner, K. (2016). Analysis of Network Clustering Algorithms and Cluster Quality Metrics at Scale. *PLoS One* 11, e0159161. doi:10.1371/journal.pone.0159161
- Frings, O., Alexeyenko, A., and Sonnhammer, E. L. L. (2013). MGclus: Network Clustering Employing Shared Neighbors. *Mol. Biosyst.* 9, 1670–1675. doi:10.1039/c3mb25473a
- Glaab, E., Baudot, A., Krasnogor, N., Schneider, R., and Valencia, A. (2012). EnrichNet: Network-Based Gene Set Enrichment Analysis. *Bioinformatics* 28, i451–i457. doi:10.1093/bioinformatics/bts389
- Hahtola, S., Tuomela, S., Elo, L., Häkkinen, T., Karenko, L., Nedoszytko, B., et al. (2006). Th1 Response and Cytotoxicity Genes Are Down-Regulated in Cutaneous T-Cell Lymphoma. *Clin. Cancer Res.* 12, 4812–4821. doi:10.1158/1078-0432.ccr-06-0532
- Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources. *Nat. Protoc.* 4, 44–57. doi:10.1038/nprot.2008.211
- Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A. F. (2002). Discovering Regulatory and Signalling Circuits in Molecular Interaction Networks. *Bioinformatics* 18 (Suppl. 1), S233–S240. doi:10.1093/bioinformatics/18.suppl_1.s233
- Jeggari, A., and Alexeyenko, A. (2017). NEARender: an R Package for Functional Interpretation of 'omics' Data via Network Enrichment Analysis. *BMC Bioinforma.* 18, 118. doi:10.1186/s12859-017-1534-y
- Jhamb, D., Magdi-Slav, M., Hurler, M. R., and Agarwal, P. (2019). Pathway Analysis of GWAS Loci Identifies Novel Drug Targets and Repurposing Opportunities. *Drug Discov. Today* 24, 1232–1236. doi:10.1016/j.drudis.2019.03.024
- Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a Reference Resource for Gene and Protein Annotation. *Nucleic Acids Res.* 44, D457–D462. doi:10.1093/nar/gkv1070
- Kiessling, M. K., Klemke, C. D., Kamiński, M. M., Galani, I. E., Krammer, P. H., and Gülow, K. (2009). Inhibition of Constitutively Activated Nuclear Factor- κ B Induces Reactive Oxygen Species- and Iron-dependent Cell Death in Cutaneous T-Cell Lymphoma. *Cancer Res.* 69, 2365–2374. doi:10.1158/0008-5472.can-08-3221
- Lancichinetti, A., and Fortunato, S. (2009). Community Detection Algorithms: A Comparative Analysis. *Phys. Rev. E Stat. Nonlin Soft Matter Phys.* 80, 056117. doi:10.1103/PhysRevE.80.056117
- Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular Signatures Database (MSigDB) 3.0. *Bioinformatics* 27, 1739–1740. doi:10.1093/bioinformatics/btr260
- Lu, Y., Wang, X., Dong, H., Wang, X., Yang, P., Han, L., et al. (2019). Bioinformatics Analysis of microRNA Expression between Patients with and without Latent Tuberculosis Infections. *Exp. Ther. Med.* 17, 3977–3988. doi:10.3892/etm.2019.7424
- Malliaros, F. D., and Vazirgiannis, M. (2013). Clustering and Community Detection in Directed Networks: A Survey. *Phys. Rep.* 533, 95–142. doi:10.1016/j.physrep.2013.08.002
- McCormack, T., Frings, O., Alexeyenko, A., and Sonnhammer, E. L. L. (2013). Statistical Assessment of Crosstalk Enrichment between Gene Groups in Biological Networks. *PLoS One* 8, e54945. doi:10.1371/journal.pone.0054945
- Mitra, K., Carvunis, A.-R., Ramesh, S. K., and Ideker, T. (2013). Integrative Approaches for Finding Modular Structure in Biological Networks. *Nat. Rev. Genet.* 14, 719–732. doi:10.1038/nrg3552
- Narducci, M. G., Arcelli, D., Picchio, M. C., Lazzeri, C., Pagani, E., Sampogna, F., et al. (2011). MicroRNA Profiling Reveals that miR-21, miR486 and miR-214 Are Upregulated and Involved in Cell Survival in Sézary Syndrome. *Cell Death Dis.* 2, e151. doi:10.1038/cddis.2011.32
- Ogris, C., Guala, D., Helleday, T., and Sonnhammer, E. L. L. (2017). A Novel Method for Crosstalk Analysis of Biological Networks: Improving Accuracy of Pathway Annotation. *Nucleic Acids Res.* 45, e8. doi:10.1093/nar/gkw849
- Ogris, C., Helleday, T., and Sonnhammer, E. L. L. (2016). PathwAX: a Web Server for Network Crosstalk Based Pathway Annotation. *Nucleic Acids Res.* 44, W105–W109. doi:10.1093/nar/gkw356
- Oka, T., and Miyagaki, T. (2019). Novel and Future Therapeutic Drugs for Advanced Mycosis Fungoides and Sézary Syndrome. *Front. Med.* 6, 116. doi:10.3389/fmed.2019.00116
- Opreko, L. K., Gephart, J. M., and Mann, M. B. Editors (2004). *Advances in Systems Biology*. Boston, MA: Springer Science & Business Media, 547.
- Pancake, B. A., Zucker-Franklin, D., and Coutavas, E. E. (1995). The Cutaneous T Cell Lymphoma, Mycosis Fungoides, Is a Human T Cell Lymphotropic Virus-Associated Disease. A Study of 50 Patients. *J. Clin. Investig.* 95, 547–554. doi:10.1172/jci117697
- Persson, E., Castresana-Aguirre, M., Buzzao, D., Guala, D., and Sonnhammer, E. L. L. (2021). FunCoup 5: Functional Association Networks in All Domains of Life, Supporting Directed Links and Tissue-Specificity. *J. Mol. Biol.* 433, 166835. doi:10.1016/j.jmb.2021.166835
- Rosvall, M., Axelsson, D., and Bergstrom, C. T. (2009). The Map Equation. *Eur. Phys. J. Spec. Top.* 178, 13–23. doi:10.1140/epjst/e2010-01179-1
- Rosvall, M., and Bergstrom, C. T. (2008). Maps of Random Walks on Complex Networks Reveal Community Structure. *Proc. Natl. Acad. Sci. U.S.A.* 105, 1118–1123. doi:10.1073/pnas.0706851105
- Shemirani, R., Gillian, M B, Keith, B, Kristina, L, Christy, L A, Eimear, E K, et al. (2021) Selecting Clustering Algorithms for IBD Mapping, *bioRxiv*, 29, doi:10.1101/2021.08.11.456036
- Signorelli, M., Vinciotti, V., and Wit, E. C. (2016). NEAT: an Efficient Network Enrichment Analysis Test. *BMC Bioinforma.* 17, 352. doi:10.1186/s12859-016-1203-6
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-wide Expression Profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi:10.1073/pnas.0506580102
- Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., et al. (2021). Correction to "The STRING Database in 2021: Customizable Protein-Protein Networks, and Functional Characterization of User-Uploaded Gene/measurement Sets". *Nucleic Acids Res.* 49, 10800. doi:10.1093/nar/gkab835
- Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J.-s., et al. (2009). A Novel Signaling Pathway Impact Analysis. *Bioinformatics* 25, 75–82. doi:10.1093/bioinformatics/btn577
- Ulgen, E., Ozisik, O., and Sezerman, O. U. (2019). pathfindR: An R Package for Comprehensive Identification of Enriched Pathways in Omics Data through Active Subnetworks. *Front. Genet.* 10, 858. doi:10.3389/fgene.2019.00858
- Van Dongen, S. (2008). Graph Clustering via a Discrete Uncoupling Process. *SIAM J. Matrix Anal. Appl.* 30, 121–141. doi:10.1137/040608635
- Wu, D., and Smyth, G. K. (2012). Camera: a Competitive Gene Set Test Accounting for Inter-gene Correlation. *Nucleic Acids Res.* 40, e133. doi:10.1093/nar/gks461
- Yoon, S., Kim, J., Kim, S.-K., Baik, B., Chi, S.-M., Kim, S.-Y., et al. (2019). GScluster: Network-Weighted Gene-Set Clustering Analysis. *BMC Genomics* 20, 352. doi:10.1186/s12864-019-5738-6
- Young-Xu, Y., and Chan, K. A. (2008). Pooling Overdispersed Binomial Data to Estimate Event Rate. *BMC Med. Res. Methodol.* 8, 58. doi:10.1186/1471-2288-8-58

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Castresana-Aguirre, Guala and Sonnhammer. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.