bioRxiv preprint doi: https://doi.org/10.1101/272450; this version posted March 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

pathfindR: An R Package for Pathway Enrichment Analysis Utilizing Active Subnetworks

Ege Ulgen[1]*, Ozan Ozisik[2], Osman Ugur Sezerman[1]

[1]Department of Biostatistics and Medical Informatics, School of Medicine, Acibadem Mehmet Ali Aydinlar University, Istanbul, Turkey.
[2]Department of Computer Engineering, Electrical & Electronics Faculty, Yildiz Technical University, Istanbul, Turkey.

*Correspondence to:
Ege Ulgen
Address: Department of Biostatistics and Medical Informatics, School of Medicine, Acibadem Mehmet Ali Aydinlar University, Istanbul, Turkey
E-mail: egeulgen@gmail.com

bioRxiv preprint doi: https://doi.org/10.1101/272450; this version posted March 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Summary: PathfindR is a tool for pathway enrichment analysis utilizing active subnetworks. It identifies gene sets that form active subnetworks in a protein-protein interaction network using a list of genes provided by the user. It then performs pathway enrichment analyses on the identified gene sets. Further, using the R package pathview, it maps the user data on the enriched pathways and renders pathway diagrams with the mapped genes. Because many of the enriched pathways are usually biologically related, pathfindR also offers functionality to cluster these pathways and identify representative pathways in the clusters. PathfindR is built as a stand-alone package but it can easily be integrated with other tools, such as differential expression/methylation analysis tools, for building fully automated pipelines. In this article, an overview of pathfindR is provided and an example application on a rheumatoid arthritis dataset is presented and discussed.

Availability: The package is freely available under MIT license at: https://github.com/egeulgen/pathfindR

bioRxiv preprint doi: https://doi.org/10.1101/272450; this version posted March 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1. Introduction

High-throughput technologies have revolutionized biomedical research by enabling comprehensive characterization of biological systems. These technologies allow researchers to identify a list of differentially expressed genes/proteins or differentially methylated genes, which most likely play a role in the formation of the phenotype. However, this list often falls short of providing mechanistic insights into the underlying biology of the disease being studied[1]. Therefore, we face a challenge posed by high-throughput experiments: extracting relevant information that allows us to understand the underlying mechanisms from a long list of genes or shortly, finding a needle in the haystack.

One approach, which reduces the complexity of analysis while simultaneously providing great explanatory power, is identifying groups of genes that function in the same pathways, i.e. pathway analysis[1]. Pathway analysis has been successfully and repeatedly applied to gene expression[2,3], proteomics[4] and DNA methylation data[5], in addition to various other applications[6-10].

However, there are drawbacks to pathway analysis. Most importantly, the statistics used by pathway analysis approaches usually consider the number of genes in a list alone and are independent of the values associated with genes, such as fold-changes or p values. By treating each gene equally, they also assume that each gene is independent of the other genes. Because they ignore information on interactions of genes, directly performing pathway analysis on a gene set is not completely informative.

For a given list of significant genes, an active subnetwork is defined as a group of interconnected genes in a protein-protein interaction network (PIN) that mostly consists of significant genes. In short, active subnetworks define distinct disease-associated sets of interacting genes. For the identification of active subnetworks, various algorithms have been proposed, such as greedy algorithms[11-19], simulated annealing[20-21], genetic algorithms[22-26] and mathematical programming-based methods[27-31].

With pathfindR, we propose to leverage interaction information from active subnetworks to extract the most relevant pathways, utilizing both the p values of individual genes and information from a PIN. In the pathfindR approach, information from four resources are integrated to determine the mechanisms underlying the disease: (i) differential expression/methylation information obtained through omics analyses, (ii) interaction information through the protein-protein interaction network, (iii) Kyoto Encyclopedia of Genes and Genomes (KEGG)[32,33] pathways, (iv) clustering of related pathways and establishment of representative pathways.

The pathfindR package was developed based on a previous approach developed by our group for genome-wide association studies (GWASes): Pathway and Network-Oriented GWAS Analysis (PANOGA)[34]. PANOGA was successfully applied to uncover the underlying mechanisms in GWASes of various diseases, such as rheumatoid arthritis[35], intracranial aneurysm[36], epilepsy[37] and Behcet's disease[38]. pathfindR applies an

bioRxiv preprint doi: https://doi.org/10.1101/272450; this version posted March 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

approach similar to PANOGA to “omics” experiments with additional functionality, further described below.
In this article, we present the details on pathfindR along with an example application on rheumatoid arthritis (RA) differential expression data.

2. Methods

2.1. The pathfindR Case Study - Analysis on RA Data

The dataset GSE15573 was obtained from the National Center for Biotechnology Information (NCBI) - Gene Expression Omnibus (GEO). This dataset aimed to characterize gene expression profiles in the peripheral blood mononuclear cells of 18 RA patients versus 15 healthy subjects. We performed differential expression analysis between these two groups using the R[39] package limma[40]. The differentially-expressed genes (DEGs) with adjusted p values ≤ 0.05 (n = 571) were used to create the example input dataset *RA_input*. This dataset includes the gene symbols, log-fold-change values and adjusted p values for the DEGs.
Active subnetwork search and enrichment analysis of the RA differential-expression data was performed with pathfindR using the greedy active subnetwork search algorithm and the Biogrid PIN.
Next, the enriched pathways were clustered and representative pathways were obtained.
The analysis approach is explained in detail below.

2.2. Protein-protein Interaction Networks

The user can choose between the protein-protein interaction data of KEGG, Biogrid[41,42], GeneMANIA[43] and InTact[44].
The KEGG PIN was created by an in-house script using the KEGG pathways on December 31, 2017. The relations among genes were added to the PIN as undirected links, removing any duplicate interactions.
For the GeneMania PIN, only interactions with weights ≥ 0.0006 were kept, allowing only strong interactions.
All PINs were formatted as simple interaction files (SIFs) for use in analyses.
The user can also use a PIN of their choice by supplying the path of the SIF to the wrapper function *run_pathfindR*.

2.3. Scoring of Subnetworks

In pathfindR we followed the scoring scheme that was proposed by Ideker et al.[20]. p value of each gene is converted to z-score using Eq. 1 and the score of a subnetwork is calculated using Eq. 2. In Eq. 2, A is the set of genes in the subnetwork and k is its cardinality.

$$z_i = \Phi^{-1}(1 - p_i) \quad (1)$$

bioRxiv preprint doi: https://doi.org/10.1101/272450; this version posted March 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$z_A = \frac{1}{\sqrt{k}} \sum_{i \in A} z_i \quad (2)$$

In the same scoring scheme, there is also a Monte Carlo approach for the calibration of the scores of subnetworks against background distribution. Using randomly selected genes, 2000 subnetworks of each possible size are constructed and for each possible size, the mean and standard deviation of the score is calculated. These values are used to calibrate subnetwork score using Eq. 3.

$$s_A = \frac{(z_A - \mu_k)}{\sigma_k} \quad (3)$$

2.4. Active Subnetwork Search Algorithms

Currently, there are three algorithms implemented in the pathfindR package for active subnetwork search: greedy algorithm, simulated annealing algorithm and genetic algorithm.

2.4.1. Greedy Algorithm

Greedy algorithm is the problem-solving/optimization concept that chooses locally the best option in each stage with the hope of reaching the global optimum. In active subnetwork search, this is generally applied by starting with a significant seed node and considering addition of a neighbor in each step to maximize the subnetwork score. In pathfindR, we used the approach in Chuang et al.[13]: This algorithm considers addition of a node within a specified distance d to the current subnetwork. In our method maximum depth from the seed can also be set. With the default parameters, our greedy method considers addition of direct neighbors (d=1) and forms a subnetwork with a maximum depth of 1 for each seed. Because the expansion process runs for each significant seed node, several overlapping subnetworks emerge. In pathfindR, overlapping subnetworks are handled by discarding a subnetwork that overlaps with a higher scoring subnetwork more than a threshold, which is set to 0.5 by default.

2.4.2. Simulated Annealing Algorithm

Simulated annealing improves the greedy search by accepting non-optimal actions to increase exploration in the search space. The probability of accepting a non-optimal action decreases in each iteration. In active subnetwork search context, the search begins with a set of randomly chosen genes (that will be referred to as genes in “on” state), connected components in this candidate solution are found and the scores are calculated. In each iteration the state of a random node is changed from on to off, vice versa, connected components are found in the new solution and their scores are calculated. If the score improves, the change is accepted, if the score decreases, the change is accepted with a probability proportional to the temperature parameter that decreases in each step.

bioRxiv preprint doi: https://doi.org/10.1101/272450; this version posted March 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.4.3. Genetic Algorithm

Genetic Algorithm is a bio-inspired algorithm that mimics natural selection by implementing fitness-based parent selection, crossover of genes and mutation. In our genetic algorithm implementation, candidate solutions represent on/off state of each gene. In the algorithm, we used rank selection and uniform crossover. In each iteration, the fittest solution of the previous population is preserved if the highest score of the current population is less than the previous population's score. In every ten iterations, the worst scoring 10% of the population is changed with random solutions. Because uniform cross-over and addition of random solutions make adequate contribution to exploration of the search space, mutation is off in default settings.

2.5. Active Subnetwork-Oriented Pathway Enrichment Analysis

Our active subnetwork-oriented pathway enrichment is implemented as the wrapper function *run_pathfindR*. The overview of the approach is presented in Figure 1.

Initially, the input is filtered so that all p values are less than or equal to the given threshold (default is 0.05). Next, gene symbols that are not found in the PIN are identified. If aliases of these gene symbols, obtained through the R package org.Hs.eg.db[45], are found in the PIN, the symbols are converted to the corresponding aliases.

The processed data is used for active subnetwork search. The identified active subnetworks are then filtered via the following criteria: (i) has a score larger than the given threshold (default is 3) and (ii) contains at least a specified number of DEGs (default is 2).

Using the genes in each of the remaining subnetworks, pathway enrichment analyses are performed via one-sided hypergeometric testing. The enrichment tests use the genes in the PIN as the gene pool. Using the genes in the PIN instead of the whole genome provides more statistical strength because active subnetworks are identified using only the genes in the PIN. The p values obtained from the enrichment tests are adjusted using the Bonferroni method.

Pathways with adjusted p values larger than the given threshold (default is 0.05) are discarded. This process of active subnetwork search and enrichment analysis is repeated for a selected number of iterations (default is 10 iterations for greedy and simulated annealing algorithms, 1 for genetic algorithm). These iterations are executed in parallel via the R package foreach[46].

Finally, the lowest and the highest adjusted p values, the number of occurrences over all iterations and up-regulated and down-regulated DEGs in each enriched pathway are returned as a data frame. Additionally, Hypertext Markup Language (HTML) format reports with the pathfindR enrichment results, linked to the visualizations of the pathways, as well as the table of converted gene symbols are created. The pathway diagrams are created using the R package pathview[47]. These diagrams display the involved genes colored by change values on a KEGG pathway graph.

bioRxiv preprint doi: https://doi.org/10.1101/272450; this version posted March 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.6. Pathway Clustering and Partitioning

Enrichment analysis usually yields a large number of related pathways. In order to establish representative pathways among similar groups of pathways, we propose that clustering can be performed via an approach based on a method described previously by Chen et al[48]. This approach is described below:

Firstly, an overlap index matrix *OI* containing overlap indices between all pairs of pathways is calculated. For each pathway P_i in the dataset, let G_i be the set of all genes in P_i. For a pair of pathways P_i and P_j, $OI_{i,j}$ is defined in Eq. 4.

$$OI_{i,j} = \frac{|G_i \cap G_j|}{min(|G_i|, |G_j|)} \quad (4)$$

Afterwards, defining each row o_i of the matrix *OI* as the gene overlap profile of pathway P_i, the Pearson correlation coefficients $R_{i,j}$ are calculated for each pair of o_i and o_j. These are then transformed into pairwise distances $PD_{i,j} = 1 - R_{i,j}$. This distance calculation approach is implemented in pathfindR as the function *cluster_pathways*.

Using this distance metric *PD*, pathways are clustered via hierarchical clustering with the desired agglomeration method. Via a shiny[49] application, the hierarchical clustering dendrogram is visualized. In this application, the user can select the agglomeration method and the distance value at which to partition the tree. The representative pathway for each cluster is chosen as the pathway with the smallest "lowest p" value. The dendrogram with the cut-off value marked with a red line is dynamically visualized and the resulting cluster assignments of the pathways and annotation of representative pathways are presented as a table. This table can be saved as a comma-separated values (CSV) file.

This clustering and portioning method is implemented as the wrapper function *choose_clusters* in the pathfindR package.

3. Results of Analysis on RA Data

In the analysis of the RA differential expression data, pathfindR identified 36 KEGG pathways to be enriched (Table S1). Upon examination of these pathways, some appeared to be biologically related, such as various signaling pathways. Therefore, clustering of these 36 KEGG pathways was performed. Upon manual inspection, the clustering dendrogram was cut at a distance of 0.66 (Figure 2), and 13 representative pathways were obtained (Table 1). Below, we discuss the functional relevance of the identified representative pathways to the pathogenesis of RA.

The most significantly enriched pathway was "Spliceosome". Autoimmune response to the spliceosome was previously reported in numerous autoimmune diseases, including RA[50]. Moreover, a recent study revealed there is a significant alteration of spliceosome components in RA patients[51]. This study suggested that alterations in the spliceosome

bioRxiv preprint doi: https://doi.org/10.1101/272450; this version posted March 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

could be associated with the development of RA and could also drive development of cardiovascular disease by altering the atherothrombotic profile in patients.

"Pathogenic Escherichia coli infection" was found to be the representative pathway of the cluster, which also included the pathways, "Bacterial invasion of epithelial cells", "Shigellosis" and "Salmonella infection". The association between these pathways, suggesting a response to infection, and RA development is not certain. However, in a review on microbial infections and RA, it was reported that infections play an important role in the initiation and advancement of RA[52]. The review also discusses potential mechanisms whereby infection may promote the development of RA, such as generation of neo-autoantigens, molecular mimicry, and bystander activation of the immune system.

There is currently no study that explains the association of "RNA transport" with RA. However, a recent study that analyzed dysregulated genes in RA also found that DEGs were enriched in "RNA transport" among other pathways[53]. This implies that dysregulation of "RNA transport" may play an important role in RA.

The association between the "Neurotrophin signaling pathway" and RA is well supported by literature. A 2005 study compared nerve growth factor (NGF), brain derived neurotrophic factor (BDNF), neurotrophin 3 (NT-3), and neurotrophin 4 (NT-4) concentrations in the serum of spondyloarthritis (SpA), rheumatoid arthritis (RA) and osteoarthritis (OA) patients, and healthy subjects[54]. Significantly higher concentrations of NT-4 and lower concentrations of BDNF were reported in disease group compared to healthy controls. Another study investigated the mRNA expression of BDNF and NGF in synovial fluid cells of RA, SpA and OA patients[55]. It was detected that NGF was expressed at significantly higher levels in RA and SpA patients than in the OA group. A recent study that investigated the methylation patterns affecting the pathogenesis of RA identified that the differentially methylated genes participated in the "Neurotrophin signaling pathway"[56] among others. This provides a further level of evidence on the involvement of this pathway in RA pathogenesis.

The "NF-kappa B signaling pathway" is known to play a key role in RA pathology. The transcription factor nuclear factor kappa B (NF-κB) is accepted as a pivotal regulator of inflammation in RA along with other aspects of RA pathology[57]. Studies in animal models of RA demonstrated the efficacy of inhibitors of this pathway. Therefore, the "NF-kappa B signaling pathway" is also considered a therapeutic target in RA[58].

We identified the "Parkinson's disease" pathway as the representative pathway in the cluster which also included "Huntington's disease". The association between RA and neurodegenerative diseases is not entirely clear. However, a recent study investigated genome-wide pleiotropy between Parkinson's disease (PD) and autoimmune diseases and found a genetic link between PD and RA[59]. This study identified 4 loci with genetic risk variants conveying risk of both PD and RA. This genetic evidence supports our transcriptomic finding and suggests that PD and RA are affected by or induce similar

bioRxiv preprint doi: https://doi.org/10.1101/272450; this version posted March 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

biological processes. One of such common processes is likely inflammation: Known to be a key process in RA, inflammation is also reported to be etiologically involved in PD[60].

"cGMP-PKG signaling pathway" enrichment is supported by two studies, which showed that RA shared epitope (an HLA-DRB1-encoded 5-amino acid sequence motif carried by most of RA patients) acted as a signal transduction ligand that interacted with cell surface calreticulin, triggered nitric oxide-mediated signaling events in opposite cells, and affected cGMP levels[61,62].

The representative pathway "Mismatch repair" (MMR) consisted of only genes downregulated in RA. Supporting our finding, Lee et al. also identified suppressed MMR enzyme expression in RA[63]. This study also observed abundant microsatellite instability in RA synovium most likely due to MMR deficiency.

We also identified the "Citrate cycle (TCA cycle)" as one of the representative pathways. The cluster of this representative pathway also included "Pyruvate metabolism" and "Glycolysis – Gluconeogenesis", suggesting a dysregulation in energy metabolism. This finding is supported by Yang et al. who screened different proteins and metabolites in the synovial fluid samples of 25 RA patients and 10 normal subjects to explore the pathogenesis of RA[64]. Ultimately, they identified energy metabolism disorder as a contributing factor of RA.

4. Conclusion

PathfindR is an R package that enables active subnetwork-oriented pathway analysis, complementing the gene-phenotype associations identified through differential expression/methylation analysis. Initially identifying active subnetworks in a list of significant genes and then performing pathway enrichment analysis of these active subnetworks makes the best use of interaction information between the genes. This, in turn, helps uncover novel in addition to known mechanisms underlying the disease, as demonstrated in the RA example.

As stated above, the pathfindR approach is based on PANOGA. This package extends the use of the active subnetwork-oriented pathway analysis approach to omics data. Additionally, pathfindR provides numerous improvements and useful new features, listed in detail below.

To overcome inconsistent annotation issues, pathfindR converts gene symbols that are not in the PIN to alias symbols that are in the PIN. This ensures that the majority of genes from the experiment can be mapped to the PIN and the user can make the best use of the data at hand.

The package provides three active subnetwork search algorithms. The user is therefore able to choose between the different algorithms to obtain the optimal results.

For the greedy and simulated annealing active subnetwork search algorithms, the search and enrichment processes are executed several times. By summarizing results

bioRxiv preprint doi: https://doi.org/10.1101/272450; this version posted March 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

over the iterations and identifying consistently enriched pathways, the stochasticity of these algorithms is overcome. Because the genetic algorithm is time-exhaustive, it is executed only once.

In addition to the data frame object, the package provides an HTML report with links to a table of the active subnetwork-oriented pathway enrichment results and a table of converted gene symbols. The table of enrichment results contains links to the pathway diagrams of individual pathways. These diagrams display the involved genes colored by change values.

pathfindR also allows for clustering of related pathways. This allows for further abstraction of the data and reduces the complexity of analysis.

All features in pathfindR work together to enable identification of dysregulated pathways that potentially reflect the underlying pathological mechanisms. We believe that this approach will allow researchers to better answer their research questions and discover novel mechanisms.

The pathfindR package is available on: https://github.com/egeulgen/pathfindR

5. References

1. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012;8(2):e1002375.
2. Emmert-streib F, Glazko GV. Pathway analysis of expression data: deciphering functional building blocks of complex diseases. PLoS Comput Biol. 2011;7(5):e1002053.
3. Werner T. Bioinformatics applications for pathway analysis of microarray data. Curr Opin Biotechnol. 2008;19(1):50-4.
4. Wu X, Hasan MA, Chen JY. Pathway and network analysis in proteomics. J Theor Biol. 2014;362:44-52.
5. Wang XX, Xiao FH, Li QG, Liu J, He YH, Kong QP. Large-scale DNA methylation expression analysis across 12 solid cancers reveals hypermethylation in the calcium-signaling pathway. Oncotarget. 2017;8(7):11868-11876.
6. Schilling CH, Schuster S, Palsson BO, Heinrich R. Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. Biotechnol Prog. 1999;15(3):296-303.
7. Xia J, Wishart DS. MetPA: a web-based metabolomics tool for pathway analysis and visualization. Bioinformatics. 2010;26(18):2342-4.
8. Welch RP, Lee C, Imbriano PM, et al. ChIP-Enrich: gene set enrichment testing for ChIP-seq data. Nucleic Acids Res. 2014;42(13):e105.
9. Ganter B, Giroux CN. Emerging applications of network and pathway analysis in drug discovery and development. Curr Opin Drug Discov Devel. 2008;11(1):86-94.

bioRxiv preprint doi: https://doi.org/10.1101/272450; this version posted March 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

10. Zheng W, Zhang Z, Liu C, et al. Metagenomic sequencing reveals altered metabolic pathways in the oral microbiota of sailors during a long sea voyage. Sci Rep. 2015;5:9131.
11. Sohler F, Hanisch D, Zimmer R. New methods for joint analysis of biological networks and expression data. Bioinformatics. 2004;20(10):1517-1521.
12. Breitling R, Amtmann A, Herzyk P. Graph-based iterative Group Analysis enhances microarray interpretation. BMC Bioinformatics. 2004;5:100.
13. Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol. 2007;3:140.
14. Nacu S, Critchley-Thorne R, Lee P, Holmes S. Gene expression network analysis and applications to immunology. Bioinformatics. 2007;23(7):850-858.
15. Ulitsky I, Shamir R. Identification of functional modules using network topology and high-throughput data. BMC Syst Biol. 2007;1:8.
16. Ulitsky I, Shamir R. Identifying functional modules using expression profiles and confidence-scored protein interactions. Bioinformatics. 2009;25(9): 1158-1164.
17. Karni S, Soreq H, Sharan R. A network-based method for predicting disease-causing genes. J Comput Biol. 2009;16(2):181-189.
18. Fortney K, Kotlyar M, Jurisica I. Inferring the functions of longevity genes with modular subnetwork biomarkers of Caenorhabditis elegans aging. Genome Biol. 2010;11(2):R13.
19. Doungpan N, Engchuan W, Chan JH, Meechai A. GSNFS: Gene subnetwork biomarker identification of lung cancer expression data. BMC Medical Genomics. 2016;9(Suppl 3):70.
20. Ideker T, Ozier O, Schwikowski B, Siegel AF. Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics. 2002;18 Suppl 1:S233-240.
21. Guo Z, Wang L, Li Y, Gong X, Yao C, Ma W, Wang D, Li Y, Zhu J, Zhang M, Yang D, Rao S, Wang J. Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network. Bioinformatics. 2007;23(16):2121-2128.
22. Klammer M, Godl K, Tebbe A, Schaab C. Identifying differentially regulated subnetworks from phosphoproteomic data. BMC Bioinformatics. 2010;11:351.
23. Ma H, Schadt EE, Kaplan LM, Zhao H. COSINE: COndition-SpecIfic sub-NEwork identification using a global optimization method. Bioinformatics. 2011;27(9):1290-1298.
24. Wu J, Gan M, Jiang R. A genetic algorithm for optimizing subnetwork markers for the study of breast cancer metastasis. Proceedings of Seventh International Conference on Natural Computation (ICNC); 2011 26-28 July; Shanghai, China.
25. Amgalan B, Lee H. WMAXC: a weighted maximum clique method for identifying condition-specific sub-network. PLoS ONE. 2014;9(8):104993.

bioRxiv preprint doi: https://doi.org/10.1101/272450; this version posted March 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

26. Ozisik O, Bakir-Gungor B, Diri B, Sezerman OU. Active Subnetwork GA: A Two Stage Genetic Algorithm Approach to Active Subnetwork Search. Current Bioinformatics. 2017;12(4):320-328.
27. Dittrich MT, Klau GW, Rosenwald A, Dandekar T, Muller T. Identifying functional modules in protein-protein interaction networks: an integrated exact approach. Bioinformatics. 2008;24(13):i223-231.
28. Zhao XM, Wang RS, Chen L, Aihara K. Uncovering signal transduction networks from high-throughput data by integer linear programming. Nucleic Acids Res. 2008;36(9):e48.
29. Qiu YQ, Zhang S, Zhang XS, Chen L. Identifying differentially expressed pathways via a mixed integer linear programming model. IET Syst Biol. 2009;3(6):475-486.
30. Backes C, Rurainski A, Klau GW, Müller O, Stöckel D, Gerasch A, Küntzer J, Maisel D, Ludwig N, Hein M, Keller A, Burtscher H, Kaufmann M, Meese E, Lenhof HP. An integer linear programming approach for finding deregulated subgraphs in regulatory networks. Nucleic Acids Res. 2012;40(6):e43.
31. Beisser D, Brunkhorst S, Dandekar T, Klau GW, Dittrich MT, Muller T. Robustness and accuracy of functional modules in integrated network analysis. Bioinformatics 2012; 28(14):1887-1894.
32. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000;28(1):27-30.
33. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45(D1):D353-D361.
34. Bakir-gungor B, Egemen E, Sezerman OU. PANOGA: a web server for identification of SNP-targeted pathways from genome-wide association study data. Bioinformatics. 2014;30(9):1287-9.
35. Bakir-gungor B, Sezerman OU. A new methodology to associate SNPs with human diseases according to their pathway related context. PLoS ONE. 2011;6(10):e26277.
36. Bakir-gungor B, Sezerman OU. The identification of pathway markers in intracranial aneurysm using genome-wide association data from two different populations. PLoS ONE. 2013;8(3):e57022.
37. Bakir-gungor B, Baykan B, Ugur İseri S, Tuncer FN, Sezerman OU. Identifying SNP targeted pathways in partial epilepsies with genome-wide association study data. Epilepsy Res. 2013;105(1-2):92-102.
38. Bakir-gungor B, Remmers EF, Meguro A, et al. Identification of possible pathogenic pathways in Behçet's disease using genome-wide association study data from two different populations. Eur J Hum Genet. 2015;23(5):678-87.

bioRxiv preprint doi: https://doi.org/10.1101/272450; this version posted March 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

39. R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/ [last accessed: March 2, 2018].
40. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
41. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 2006;34(Database issue):D535-9.
42. Chatr-aryamontri A, Oughtred R, Boucher L, et al. The BioGRID interaction database: 2017 update. Nucleic Acids Res. 2017;45(D1):D369-D379.
43. Warde-farley D, Donaldson SL, Comes O, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res. 2010;38(Web Server issue):W214-20.
44. Hermjakob H, Montecchi-palazzi L, Lewington C, et al. IntAct: an open source molecular interaction database. Nucleic Acids Res. 2004;32(Database issue):D452-5.
45. Marc Carlson (2017). org.Hs.eg.db: Genome wide annotation for Human. R package version 3.5.0.
46. Microsoft and Steve Weston (2017). foreach: Provides Foreach Looping Construct for R. R package version 1.4.4. https://CRAN.R-project.org/package=foreach [last accessed: March 2, 2018].
47. Luo W, Brouwer C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. Bioinformatics. 2013;29(14):1830-1.
48. Chen YA, Tripathi LP, Dessailly BH, Nyström-persson J, Ahmad S, Mizuguchi K. Integrated pathway clusters with coherent biological themes for target prioritisation. PLoS ONE. 2014;9(6):e99030.
49. Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson (2017). shiny: Web Application Framework for R. R package version 1.0.5. https://CRAN.R-project.org/package=shiny [last accessed: March 2, 2018].
50. Hassfeld W, Steiner G, Studnicka-benke A, et al. Autoimmune response to the spliceosome. An immunologic link between rheumatoid arthritis, mixed connective tissue disease, and systemic lupus erythematosus. Arthritis Rheum. 1995;38(6):777-85.
51. Ruiz-Limon P, Perez-Sanchez C, Ortega-Castro R, et al. AB0104 Alterations of spliceosome components in leukocytes from patients with rheumatoid arthritis influence their autoimmune and inflammatory profile, and the development of cardiovascular disease. Ann Rheum Dis. 2017;76(Suppl 2):1082.3-1082.
52. Li S, Yu Y, Yue Y, Zhang Z, Su K. Microbial Infection and Rheumatoid Arthritis. J Clin Cell Immunol. 2013;4(6)

bioRxiv preprint doi: https://doi.org/10.1101/272450; this version posted March 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

53. Hao R, Du H, Guo L, et al. Identification of dysregulated genes in rheumatoid arthritis based on bioinformatics analysis. PeerJ. 2017;5:e3078.
54. Rihl M, Kruithof E, Barthel C, et al. Involvement of neurotrophins and their receptors in spondyloarthritis synovitis: relation to inflammation and response to treatment. Ann Rheum Dis. 2005;64(11):1542-9.
55. Barthel C, Yeremenko N, Jacobs R, et al. Nerve growth factor and receptor expression in rheumatoid arthritis and spondyloarthritis. Arthritis Res Ther. 2009;11(3):R82.
56. Lin Y, Luo Z. Aberrant methylation patterns affect the molecular pathogenesis of rheumatoid arthritis. Int Immunopharmacol. 2017;46:141-145.
57. Makarov SS. NF-kappa B in rheumatoid arthritis: a pivotal regulator of inflammation, hyperplasia, and tissue destruction. Arthritis Res. 2001;3(4):200-6.
58. Jue DM, Jeon KI, Jeong JY. Nuclear factor kappaB (NF-kappaB) pathway as a therapeutic target in rheumatoid arthritis. J Korean Med Sci. 1999;14(3):231-8.
59. Witoelar A, Jansen IE, Wang Y, et al. Genome-wide Pleiotropy Between Parkinson Disease and Autoimmune Diseases. JAMA Neurol. 2017;74(7):780-792.
60. Whitton PS. Inflammation as a causative factor in the aetiology of Parkinson's disease. Br J Pharmacol. 2007;150(8):963-76.
61. Ling S, Lai A, Borschukova O, Pumpens P, Holoshitz J. Activation of nitric oxide signaling by the rheumatoid arthritis shared epitope. Arthritis Rheum. 2006;54(11):3423-32.
62. De almeida DE, Ling S, Holoshitz J. New insights into the functional role of the rheumatoid arthritis shared epitope. FEBS Lett. 2011;585(23):3619-26.
63. Lee SH, Chang DK, Goel A, et al. Microsatellite instability and suppressed DNA repair enzyme expression in rheumatoid arthritis. J Immunol. 2003;170(4):2214-20.
64. Yang XY, Zheng KD, Lin K, et al. Energy Metabolism Disorder as a Contributing Factor of Rheumatoid Arthritis: A Comparative Proteomic and Metabolomic Study. PLoS ONE. 2015;10(7):e0132695.

bioRxiv preprint doi: https://doi.org/10.1101/272450; this version posted March 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1: Representative pathways that were enriched in the RA differential-expression data. Pathway Description indicates the description of the given KEGG pathway. Occ. indicates the occurrence, i.e., the number of times the pathway was identified to be enriched over 10 iterations. Lowest p and Highest p indicate the lowest and highest p values calculated for the given pathway over all iterations. Up-regulated and Down-regulated indicate the up- and down-regulated DEGs that are involved in the given pathway.

Pathway Description	Occ.	Lowest p	Highest p	Up-regulated	Down-regulated
Spliceosome[50,51]	10	1.10E-06	1.50E-06	SF3B6, LSM3, BUD31	SNRPB, SF3B2, U2AF2, PUF60, HNRNPA1, PCBP1, SRSF5, SRSF8, SNU13, DDX23, EIF4A3
Pathogenic Escherichia coli infection[52]	10	1.90E-05	2.50E-03	LY96, TLR5	ABL1, ITGB1, TUBB, ACTB, ACTG1
RNA transport[53]	10	3.30E-05	2.20E-03	NUP214	GEMIN4, EIF4A3, RNPS1, SRRM1, NUP62, NUP93, UBE2I, RANGAP1, SUMO3, EIF2S3, EIF2B1
Neurotrophin signaling pathway[54-56]	4	5.20E-05	6.80E-04		CRKL, FASLG, SH2B3, ABL1, MAGED1, IRAK2, IKBKB, CALM1, CALM3
NF-kappa B signaling pathway[57,58]	1	8.50E-04	8.50E-04	LY96	IKBKB, PRKCQ, CARD11, TICAM1, PARP1, UBE2I
Parkinson's disease[59,60]	2	2.90E-03	2.90E-03	NDUFA1, NDUFB3, UQCRQ, COX6A1, COX7A2, COX7C, ATP5E, ATP5J	ATP5G2, SLC25A5, VDAC1, UBE2G1
cGMP-PKG signaling pathway[61,62]	2	4.30E-03	4.30E-03		NFATC3, SRF, ATP2A2, CREB1, ADCY7, SLC25A5, VDAC1, CALM1, CALM3
Mismatch repair[63]	2	7.70E-03	7.70E-03		MLH1, POLD2, RPA1
Citrate cycle (TCA cycle)[64]	5	7.90E-03	7.90E-03		MDH2, PDHA1, PDHB
SNARE interactions in vesicular transport	3	1.70E-02	1.70E-02	STX10, STX6	BET1L, SNAP23, STX2

bioRxiv preprint doi: https://doi.org/10.1101/272450; this version posted March 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Proteasome	5	1.80E-02	1.80E-02		PSMD7, PSMB10
Vibrio cholerae infection	8	2.10E-02	3.70E-02	ATP6V0E1, ATP6V1D	ARF1, ATP6V0E2, ACTB, ACTG1, PDIA4
Cysteine and methionine metabolism	4	3.00E-02	3.00E-02		MRI1, MAT2B, AHCYL2, DNMT1, GOT1, MDH2

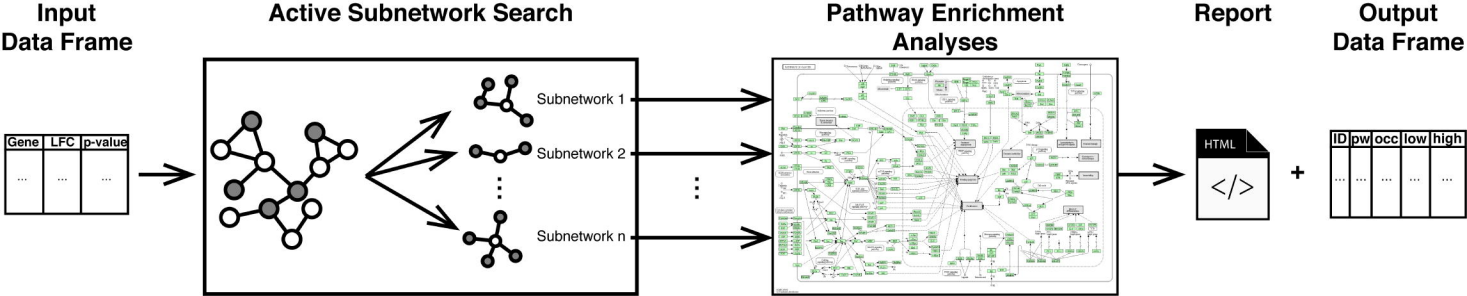
Supplementary Table 1: Table of enriched pathways identified in the analysis of the RA differential-expression data with pathfindR. KEGG ID indicates the KEGG ID of the pathway. Pathway indicates the description of the pathway. Occ indicates the occurrence, i.e., the number of times the pathway was identified to be enriched over 10 iterations. Lowest p and Highest p indicate the lowest and highest p values calculated for the pathway over all iterations. Up_regulated and Down_regulated indicate the up- and down-regulated DEGs that are involved in the given pathway. Cluster indicates the cluster the pathway is assigned to upon clustering of the pathways. Status indicates whether the pathway is the representative pathway or a regular member in its cluster.

bioRxiv preprint doi: https://doi.org/10.1101/272450; this version posted March 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure Legends

Figure 1: Flow diagram of the pathfindR active subnetwork-oriented pathway enrichment analysis approach

Figure 2: Clustering dendrogram of enriched pathways identified in the RA differential expression dataset. Vertical axis indicates the pairwise distance. The horizontal red line indicates the height at which the dendrogram is cut. The representative pathways, i.e. the pathways with the lowest p value in each cluster, are indicated as bold text.


Input
Data Frame
Gene
LFC
p-value
Active Subnetwork Search
Subnetwork 1
Subnetwork 2
Subnetwork n
Pathway Enrichment
Analyses
Report
HTML
Output
Data Frame
ID
pw
occ
low
high

bioRxiv preprint doi: https://doi.org/10.1101/272450; this version posted March 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.