**Biologically Guided Multi-Omics Local Enrichment Analysis Protocol**

This is a new network-based multi-omics enrichment analysis workflow

This workflow combines previous methods into a biologically meaningful network diffusion process for multi-omics module identification and enrichment analysis

Ghadi, GEH, El Hasbani

Bioinformatics BS & Psychology Minor, Lebanese American University. [ghadi.elhasbani@lau.edu](mailto:ghadi.elhasbani@lau.edu)

Bilal, BWH, W. Hamdanieh

Bioinformatics BS, Lebanese American University. [bilal.hamdanieh@lau.edu](mailto:lara.issa02@lau.edu)

Frederick , FAC, Abi Chahine

Bioinformatics BS, Lebanese American University. [frederick.abichahine@lau.edu](mailto:frederick.abichahine@lau.edu)

**ABSTRACT**

Although typical Differential Expression Analysis are gene-centered, enrichment analysis has recently moved towards network-based workflows that account for gene-gene relationships. In turn, this resulted in a focus on functional, disease-related module identification for interpretable, prioritized enrichment analysis input. Single-omics approaches, although limited, have presented many advances in terms of expression data processing. Moreover, multi-omics approaches, especially those utilizing Network Diffusion, account for the global network topology across data sources or layers of information. In this paper, we propose a new protocol for multi-omics enrichment analysis that combines advantages from previous methods: interpretability, data-integration, and biologically-guided consideration of the global network topology. We then evaluate our method using coverage, enrichment, and connectivity metrics as well as a new metric measuring the cumulative decay score. In all, our protocol, awaiting further investigation, offers promising prospects for the integration of biological meaning into computational omics analyses.

CCS CONCEPTS • Gene-set enrichment analysis • Network Diffusion • multi-omics • data integration • gene expression • differential expression • functional modules

**Additional Keywords and Phrases:** Decay of gene expression, GeneSurrounder, mND, module detection, functionally enriched connectivity, network resampling, gene coverage, breast cancer, random walk with restart, cumulative decay score

**INTRODUCTION**

1. **The Effect of Genes Over their Surrounding Area**

Differential Gene Expression Analysis (DEA) is the most widely used method for detecting significant gene-associations based on their mean expression between phenotypes [1]. However, while DEA can identify specific disease associated genes, it does not take into consideration the network of interactions that govern the studied set of genes leading to limits in mechanistic insights. Hence, the analysis might miss crucial multi-gene interactions that underlie complex phenotypes. As a result. DEA can exhibit poor consensus with studies of the same conditions [2, 3]. This necessitates the emergence of new methods that take into consideration the network structure of genes.

Recently proposed methods use networks to measure the effect of genes over their surrounding area of other genes, identifying modules: a group of dysregulated genes that contribute to the disease/phenotype of study. One method, Local Enrichment Analysis (LEAN) [4] identifies dysregulated subnetworks from genome-wide omics datasets by substituting common subnetworks with local subnetwork models that consist of only the direct neighboring genes (radiuslocal\_subnetwork = 1). The method is also parameter free and exhaustive over all the genes. Another method, PathFindR [5], extends LEAN by taking advantage of user input to specify the radius of the local subnetwork to be enriched using three different algorithms of choice: Greedy Algorithm (GD), Simulated Annealing, (SA) and Genetic Algorithm (GA). Further, it has been shown that the Greedy Algorithm performs better than SA & GA; SA and GA are heuristic methods that do not make biologically-relevant assumptions on the active subnetwork model, which lets insignificant genes between two clusters of significant genes become a single connected active subnetwork, which results in high scoring active subnetworks with the remaining subnetworks becoming fewer and less informative. In short, there is a tendency towards large subnetworks which is attributed to a statistical bias prevalent in many tools [6]. Another recently developed algorithm is GeneSurrounder [7] that proposes an exhaustive method to consider the decay of DE and the sphere of influence of a gene. The optimal radius that identifies the effect of the gene on the neighbors is given from the combination of the two p-values using the Fisher method (pfisher from pdecay and psphere). However, the time complexity and implementation of GS requires further development and optimization to be favorable for common use.

Other research proposed network diffusion (ND) – also referred to as network propagation – for the development of integrative methods to analyze multiple gene-centered datasets while considering known or inferred relations between genes [8, 9, 10]. ND can be of several types: Random Walk, Random Walk with Restart (RWR), Insulated Heat Diffusion (IHD) and Diffusion Kernel. The latter two methods are characterized by differences in the normalization of the adjacency matrix which implies dissimilar behaviors of information flow, mainly in relation to network hubs: at infinite time in the RWR hubs tend to naturally gather relatively more information than IHD which is characterized by an intrinsic hub penalization. Therefore, despite RWR and IHD are conceptually similar, they may present sensibly different results, especially when applied to complex biological networks. For this reason, RWR is more common in biological analysis. One of the previous methods, DMFIND [11], uses RWR network diffusion to define network proximity and uses a smoothing index quantity that allows to jointly quantify the amount of omics information in genes and in their neighborhood (not only direct neighbors). Another method, which is of main interest, is mND [12] that, similarly to DMFIND, uses RWR network diffusion but uses a different scoring method (compared to the smoothing index) which takes into consideration the top k direct neighbors of a gene. This method classifies genes as modules (M), Linkers (L), Isolated (I), and Not-Selected (NS) and takes into consideration the global topology of the network while maintaining modules to contain only a gene and the top k direct neighbors (radius = 1).

Since the previous methods either consider the direct neighbors, global network topology or the decay of the gene expression, but not all together (Table 1), our proposed method suggested a new protocol for considering network topology and indirect-neighbors-based gene expression decay while still maintaining modules containing only a gene and its direct neighbors (radius = 1). Our protocol is presented as part of a multi-omics workflow considering it is highly dependent on mND. Nevertheless, GeneSurrounder adjustment could be applied before other ND methods such as DMFIND, NDPROD [13], and TieDIE [14] as well as single-omics expression analysis workflows where the adjusted scores are simply the output. Finally, it is also possible to use single-omics expression data with a sample per layer to apply ND methods as ND is independent of the type or source of data. This could be applied to identify a consensus of genes that in certain sample group. All our analyses, implementations, and data can be found on our GitHub repository (https://github.com/Frederick-abichahine/Computational-Biology-Final).

| **Tool** | **Has** | **Does not have** |
| --- | --- | --- |
| **LEAN** (Gwinner et al., 2017) | Interpretability, less statistical bias | Multi-omics, consideration of global topology, |
| **PathFindR** (Ulgen et al., 2018) – Greedy | Interpretability, less statistical bias, | User-defined R, consideration of global topology, single-omics |
| **PathFindR** – Simulated Annealing, Genetic Algorithm | Automated | Single-omics, interpretability, biological relevance of R |
| **GeneSurrounder** (Shah & Braun, 2019) | Automated & biologically relevant (Decay) | Speed, User-friendliness, consideration of global topology, multi-omics |
| **DMFIND** (Bersanelli et al., 2016) | Automated , multi-omics, consideration of global topology | Interpretability, less statistical bias, consideration of DE decay effect |
| **mND** (Di Nanni et al., 2020) | Interpretability, less statistical bias, consideration of global topology, multi-omics | Biologically relevant |

Table 1: Advantages and disadvantages of previous methods

**PROTOCOL GS & mND**

1. **Materials and Methods**
2. **Data Collection**

The data used was extracted from the pipeline of mND which was originally sourced from TCGA [15] using R package TCGAbiolinks [16]. It consists of two layers: somatic mutations frequency and differential expression scores from 11,796 genes expression data from matched tumor-normal samples (blood samples for SM and solid tissue samples for GE) for breast cancer patients and considering the human genome version 38 (hg38). The same adjacency network was also used from mND, which was originally obtained from STRING [17], to maintain comparable gene-gene interactions.

The reasoning behind choosing the same datasets used by mND is to allow us to assess and compare the results obtained with previous methods to validate any changes. This ensures the protocol for preprocessing, DE score calculation, and postprocessing is constant.

1. **Methodology**

The proposed method suggests using GeneSurrounder (GS) to adjust the gene expressions prior to performing mND, then examining if the suggested workflow successfully identifies genes that have potential in the network due to the expression decay effect of other genes not necessarily in direct neighboring distance of the gene (considers radius > 1). The reasoning is that genes that are individually differentially expressed have the potential to affect nearby genes to certain extent (radius). The more this pattern is observed, the more this gene should represent a functionally enriched model that systematically affects its direct and indirect neighborhood. If this adjustment is carried out, hypothesized that genes nearby to other powerful hubs will transmit such decay effect as indicated in their DE score, and they themselves will be reinforced by GS adjustment although to a lesser degree. Moreover, genes could also be penalized if they do not display a decay effect pattern to shortlist pathway-relevant genes. If this behavior is observed, GS-adjusted mND results should display superior enrichment results as well as maintain connectivity, meaningful gene classifications, and disease-relatedness.

First, GS functions and methods were downloaded from GitHub repository to allow us to measure the decay of gene expression exhaustively (radius >= 1). Only pdecay was obtained from the GS pipeline as the data/sample was not readily available in the mND package. Then, the mND package was used to perform network diffusion using RWR and scoring the genes according to the top k direct neighbors’ network diffusion scores.

1. **Pipeline Analysis**

The pipeline starts by performing GeneSurrounder, with parallelized implementation to cut down the running time of the method (ending up with an average of 4.5h/1000genes), to quantify the effect of genes on all other genes. If a gene is a source of disease-associated dysregulation, we may expect its neighbors to also exhibit a dysregulation. The quantity measured, using Kendall rank correlation, is called Decay of Differential Expression which tests whether the magnitude of the differential expression of a gene is inversely related to the distance to other genes in the interaction network (Equation 1). The latter, and random permutations, allows us to obtain a value (Equation 2) which can then be used to adjust values of expression layers prior to performing multi-network diffusion. Ideally, would be obtained by also calculating the Sphere of Influence () of a gene using data per sample.

(1)

(2)

Therefore, we adjust the Differential Gene Expression layer by multiplying its values by -log10(pdecay) to increase the effect of a gene that is highly expressed and has a high decay score over its neighbors (Equation 3), followed by multi-network diffusion using k=2 (as indicated by k optimization results; see below) and k=3. After that, we classified the genes into Modules (M), Linkers (L), Isolated (I), and Not Significant (NS) based on their differential gene expression and their neighboring information collected from network diffusion as per the mND protocol. M class is dependent on top-genes, and I and L are dependent on the choice of M. The results were then saved for comparison with mND pipeline run alone without any modifications.

(3)

**RESULTS & DISCUSSION**

1. **Classification Improvement**

The mND classification results were compared before (R1) and after (R2) applying adjustments to the differential expression scores. mND performed alone (R1) obtained 2,297 selected genes as I, L or M, while our protocol (R2) obtained fewer selected genes (2,109). The confusion matrix of the classification percentage showed that the difference in the genes is mainly in the I (76% of R1), L (69% of R1), or M (88% of R1) and not in the NS genes (96% of R1) (Table 2). For M, 5.05% previously selected were not selected after adjustment. Although this seems striking, M was the class to transition to NS the least, with 0.14% of NS genes after adjustment as compared to 1.88% and 3.11% for L and I, respectively. No I genes transitioned to class L, and only 0.27% of L transitioned to I. It seems that the highest rate of L genes transition is towards class M (8.92%). Moreover, only 0.084% of NS transitioned to M and constitute 2.55% of the latter class after adjustment. This shows that noisy changes in classification are minimal, and genes in strategic locations (L genes) are reinforced to M.. In addition, 20% of the genes previously selected as Isolated and 27% of the those previously selected as Linkers are now not being selected. This means that the opposite seems to also be true, whereby I and L genes mostly transition to NS. This might indicate a decrease in false positive genes selected, and raises the question of how the decay score leads to some genes not being selected anymore, and the reason for newly selected modules to appear. The reason could be that I and L classes are topology-dependent and could change with manipulation of the M class through score adjustment. We evaluate our protocol: is classification affected by the decay scores?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Our Protocol Classification** | | | | |
| **mND Only Protocol** | Isolated | Linkers | Modules | Not Significant |
| Isolated | 76.70 | 0.00 | 3.57 | 19.71 |
| Linkers | 0.27 | 69.73 | 2.55 | 27.4 |
| Modules | 4.69 | 1.80 | 88.44 | 5.05 |
| Not Selected | 1.70 | 1.46 | 0.084 | 96.75 |

Table 2A. Confusion Matrix Percentages Over Original mND Scores

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Our Protocol Classification** | | | | |
| **mND Only Protocol** | Isolated | Linkers | Modules | Not Significant |
| Isolated | 79.91 | 0.00 | 10.51 | 1.88 |
| Linkers | 0.34 | 84.15 | 8.92 | 3.11 |
| Modules | 1.47 | 0.55 | 78.03 | 0.14 |
| Not Selected | 18.28 | 15.29 | 2.55 | 94.87 |

Table 2B. Confusion Matrix Percentages Over Adjusted mND Scores

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Our Protocol Classification** | | | | |
| **mND Only Protocol** | Isolated | Linkers | Modules | Not Selected |
| Isolated | 6.00 | 0.00 | 0.28 | 1.54 |
| Linkers | 0.03 | 6.49 | 0.24 | 2.55 |
| Modules | 0.11 | 0.04 | 2.08 | 0.12 |
| Not Selected | 1.37 | 1.18 | 0.07 | 77.91 |

Table 2C. Confusion Matrix Percentages Over All Genes

When each class is analyzed, we notice that although we have a decrease in the number of selected genes, 7.7% of genes were additionally significant with an mND(p) value threshold of 0.05. Moreover, additional genes that appeared amongst selected genes based on classification were distributed amongst Isolated, Linkers and Modules; with 69 new modules, 32 discarded modules, 544 new Linkers and Isolated genes, and 319 discarded genes of the latter types. To assess the reasoning behind the appearance of these new genes or the discarding of others, we introduce a Cumulative Decay Score and perform significance tests.

1. **Cumulative Decay Score & Significance Test**

After obtaining the classification of every gene, we can now check the Cumulative Decay Score (CDS) on target genes. This score is used to assess the decay in relation to 1/distance from genes and the formula for the decay score for a gene is defined in Equation (4) with respect to a previously M gene *.* The procedure for calculating the cumulative version of this score is described in Algorithm (1). In order to evaluate the classification transition, we define this score as the cumulative contributory effect of the decay, normalized by distance, of genes classified as M by mND protocol alone. This score will be useful in testing the relationship between distinct M genes, classified as M after adjustment only, and previously classified M genes. The assumption is that previously classified M genes are hubs that contribute to decay effect.

(4)

To reason, we hypothesize that distinct genes would show a significant pattern of high scores as compared to the discarded genes. In other words, discarded genes should be previously of class M but not topologically-relevant enough to be affected by decay effect of other M genes as opposed to distinct genes that would appear to be differentially affected by the decay effect. Although I and L genes are not chosen on the same basis as M genes, I genes could have a high score that is contributing to the neighborhood, and M genes could be affecting them. This is because an I gene could be 2 edges away from an M genes and yet not be part of the main connected cluster and, hence, not in class L. As for L, these genes are inherently near powerful genes and their transition on the basis of decay score is important.

ALGORITHM 1: Cumulative Decay Score

for each gene i, do

if(old\_classi == Module) # If the gene is an old module then we add its effect to neighboring (dr) nodes

for each gene j in neighborhood, do # Traverse other genes in the surrounding radius and # update their cumulative decay score

if& , do

return scores

end

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart

Description automatically generatedChart, histogram

Description automatically generated

Figure 1.B: Discarded Module Genes in Green

Figure 1.D: Discarded I & L Genes in Green

Figure 1.C: Distinct I & L Genes in Green

Figure 1.A: Distinct Module Genes in Green

The density plots of the cumulative decay score (Figure 1), shows a visual difference between the distinct genes (M in Fig1.A, and I&L in Fig1.C). To assess this difference, we perform a significance test by comparing variances (F-test) and means (T-test) of distinct and discarded genes. For each subset of genes tested, the background distribution is defined as the CDSs of all genes except target genes. For distinct modules, we find a statistically significant (p-value = 5.896e-09) difference in the means whereby the distinct genes had a significantly higher mean (168.97) than that of the background distribution (126.33) with a confidence interval [29.82, 55.46]. Also, we obtain a p-value of 0.002 and ratio of 0.55 in the F-test which means that the variance of the distinctive genes that were selected as new modules was significantly lower than the background distribution for the calculated CDS with a confidence interval [0.41, 0.80]. Next, for the discarded modules, we find a statistically significant (p-value = 5.055e-06) difference in the means whereby the discarded modules had a significantly higher mean (180.41) than that of the background distribution (126.44) with a confidence interval [33.95, 74.00]. Also, we obtain a p-value of 0.083 and ratio of 0.61 in the F-test which means that the variance of the discarded M genes was not significantly lower than the old modules for the calculated CDS with a confidence interval [0.39, 1.07]. Although both distinct and discarded M genes have significantly higher CDSs, the insignificant difference in variance for discarded M genes could mean that these genes were not consistently in strategic position and their environment is variable, possibly with weaker neighboring sources of decay effect or are located at variable distances.

As for the linkers and the isolated genes, we obtain a similar pattern with the distinct I & L genes having a significant difference in mean (p-value of 0.0006) compared to the background mean of all other genes, and a significant difference in variance (p-value of 5.432e-18). On the other hand, the variance of the I&L discarded genes showed to be higher than the background but insignificantly so (p-value = 0.053) while the mean showed the same pattern as M genes with a significantly () higher mean for discarded I&L than the background. This further validates the possibility discarded groups differ from distinctive ones in that they have variable environments with respect to previously M genes.

This means that the since the same pattern is exhibited in both cases (distinct vs discarded), we realize that the variance might be an important factor in deciding the classification of genes into M or I & L. This means that another factor might be at apply, possibly the shift to and from class M. This might be due to neighbors’ decay score of M genes and network topology of I&L which requires further research and validation.

1. **Coverage & Enrichment**

We obtain the list of breast cancer (BC) related genes from three databases (eDGAR [18], DrugBank [19], & MalaCards [20]) using the AutoSeed R package [21]. Only the intersection between this list and genes included in the study was considered. Coverage was then calculated as the percentage of BC-related genes found in the full sorted gene lists at different mND(p) cutoffs.

**Chart, line chart

Description automatically generated**

As observed in Figure 2, we can see that the lowest percentage of cumulative coverage was consistently that of mND (k=3) and the highest was that of GS-adjusted mND (k=3). Besides, GS-adjusted mND (k=2) appeared to be congruent with GS-adjusted mND (k=3) for the first few mND(p) cutoffs (until ~0.005) but the difference becomes apparent as the cutoff increases. Nevertheless, the difference is evident in all cutoffs between mND (k=3) and GS-adjusted mND at both values of k. In this sense, our algorithm outperforms mND in prioritizing genes no matter the value of k for the first few cutoffs. Nonetheless, k=3 for our algorithm performs best in the task of prioritizing disease-related genes at all cutoffs of mND(p).

Figure 2

Graphical user interface, chart

Description automatically generated

Graphical user interface, chart, map

Description automatically generated Graphical user interface, chart, line chart

Description automatically generated

Figure 3: Coverage by pathway (first 4 are BC-related); DE in magenta, mND (k=3) in black, GS-adjusted mND k=3 in red and k=2 in blue

Enrichment analysis was performed using KEGG [23] with the help of the KEGGREST R package [23]. mND(p) values were used as input accompanied by the full gene list. It is important to note that none of the important relevant pathways (Figure 3) were significant for all mND (k = 3) and GS-adjusted mND (k = 2,3). All protocols also yielded around the same number of significant pathways whereby GS-adjusted mND yielded a few more. This points to the need to introduce a form of disease-specific prioritization. Nevertheless, the GS-adjusted mND protocol yielded 2 significantly dysregulated pathways related to chemical carcinogenesis in the top 40 as compared to 1 for mND alone. All methods had 1 such pathway in the top 10.

For comparative enrichment, genes per pathway were also extracted using KEGGREST and assessed for 16 pathways (Figure 3) related to cancer, 4 of which are directly BC-related (first 4). All methods consistently performed better than raw DE scores, pointing to the importance of a network-based analysis for biological, especially omics, data. Surprisingly, GS-adjusted mND (k=3) was lagging behind mND (k=3) and GS-adjusted mND (k=2). The latter two protocols are in close competition for most cutoffs. Nevertheless, for the top 50 genes, GS-adjusted mND (k=2) is almost always superior in coverage. This trend is not exclusive to the first 50 genes and occurs frequently at different cutoffs. This means that GS-adjusted mND (k=2) seems to be slightly superior to mND alone in the task of prioritizing cancer-related pathways. This s especially true for the top few genes, congruent with previous coverage results. This is in agreement with the biological relevance of the decay effect which seems to prioritize pathway structure and information flow.

1. **Enriched Connectivity**

To account for the connectivity of genes, we use a scoring method, network resampling, proposed by Besanelli et, al. (2016), and previously used for mND validation. The score checks for significantly connected, functionally enriched nodes at different cutoffs:

(5)

(6)

The score is calculated through the dot product of subset of scores and adjacency matrix corresponding to the top genes in the ranked list at any given and layer . The ranking is done through mND or DE scores which are also used for score vector . The sample procedure is done on permutated adjacency matrices whereby the vertices have been randomly resampled. As indicated in Equation (6), In our protocol, we were only able to permute over the top 1000 genes due to time constraints. This produced results different from mND reported results (Figure 3) due to the quite different sample provided for resampling of vertices in significance assessment. Regardless, using the DE scores yielded insignificant enriched connectivity for lower cutoff of sorted genes. The distribution of enriched connectivity score (Figure 4.A) showed a slight decrease in mean from our protocol with k=2, to k=3, and a decrease of around double that size between k=3 and mND protocol alone (k=3). These results need to be reassessed with permuting for more than 1,000 genes (2000 genes were also not enough). The difference realized in enriched connectivity score for original DE scores is most probably due to initial difference in the scores, as evident by the original score distributions in Figure 4.B.

Chart, histogram

Description automatically generated

Figure 4: All p-values but those for original DE scores (up until cutoff 50) are superimposed at 0

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

Figure 5.B

Figure 5.A

1. **Optimizing k**

Since results have been encouraging for k = 2 as compared to k = 3, which was reported as ideal for mND protocol, it would useful to optimize k and check for connectivity results at different values of k. As such, we use the same method described in mND available in the R package. Here, the score used for each gene is the non-decreasing (cumulative) form of the enriched connectivity score normalized across layers:

(7)

This quantity is then calculated at different values of k for different cutoffs of the gene list sorted by mND scores.

Chart, line chart

Description automatically generatedChart, line chart, scatter chart

Description automatically generated

Figure 6: Cumulative normalized enriched connectivity score at different values of k for mND protocol alone (left) and GS-adjusted mND (right)

Consistent with previous results highlighting the superior performance of GS-adjusted mND protocol with k = 2, Figure 5 shows that although there is an evident difference in enriched connectivity at around cutoff 90 between k = 1, k = 2, and k = 3 for mND alone, the GS-adjusted protocol only shows slight difference between k = 1 and all other values of k above around cutoff 90. This shows that not only did GS adjustment not necessitate consideration of indirect neighbors after mND, it also performs much better with less neighboring information, almost closing the gap for enriched connectivity at different values of k. This indicates that information is being efficiently collected and accounted for to produce functionally enriched and connected networks of disease-related genes with minimal neighbor consideration and small modules.

**LIMITATIONS & FURTHER RESEARCH**

The limitations of our work include the absence of samples data which would allow us to obtain in using GS which can be combined with using the Fisher method, yielding . GS shows can capture the decay and influence of a gene more significantly than only one of the latter p-values. Further, GS can be optimized through heuristics or Genetic Algorithm approach to cut down the running time, making it feasible for other researchers to replicate our and others’ work. Optimizing GS and structuring its implementation into a user-friendly package is of utmost importance for its academic usage and future improvements to this study. In addition, Cumulative Decay Score variance showed the potential to use Cumulative Decay Score as a score to classify genes or as a cutoff for classification into a possibly new class. The score could also be maintained as a continuous measure indicating the relative influence of the decay effect on any given gene. It would also be useful to investigate CDS by taking the intersection of M genes before and after (consistent M genes) as sources of decay effect for more reliable information. This is since it is not yet not clear whether or not new and discarded M genes Moreover, the enrichment results are not encouraging for tested methods. Perhaps GS-adjustment can be carried out only on DE genes and their surroundings as determined by R. This is in order not to reinforce naturally occurring decay effect as well as prioritizing the collection of disease-related information in the ND process. Also, enriched connectivity values require more validation and optimization as running the workflow on 1000 genes was very time-consuming. Other methods to validate the best K and calculation of could be taken into consideration in further research. Finally, comparing our results to DMFIND and other ND algorithms might reveal new insights about our protocol.

**REFERENCES**

1. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015; 43(7):e47.
2. Manoli T, Gretz N, Gröne HJ, Kenzelmann M, Eils R, Brors B. Group testing for pathway analysis improves comparability of different microarray datasets. 2006; 22(20):2500.
3. Gwinner, F., Boulday, G., Vandiedonck, C., Arnould, M., Cardoso, C., Nikolayeva, I., ... & Schwikowski, B. (2017). Network-based analysis of omics data: The LEAN method. *Bioinformatics*, *33*(5), 701-709.
4. Ulgen, E., Ozisik, O., & Sezerman, O. U. (2018). pathfindR: an R package for pathway enrichment analysis utilizing active subnetworks. *BioRxiv*, 272450.
5. Shah, S. D., & Braun, R. (2019). GeneSurrounder: network-based identification of disease genes in expression data. *BMC bioinformatics*, *20*(1), 1-12.
6. Braun R, Shah S. Network methods for pathway analysis of gene expression data. 2014. arXiv [q-bio.QM]. arXiv. http://arxiv.org/abs/1411.1993.
7. Nikolayeva, I., Guitart Pla, O., Schwikowski, B. (2018). Network module identification—a widespread theoretical bias and best practices. *Methods* 132, 19–25. doi: 10.1016/j.ymeth.2017.08.008
8. Cowen, L., Ideker, T., Raphael, B. J., & Sharan, R. (2017). Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*, *18*(9), 551-562.
9. Di Nanni, N., Bersanelli, M., Milanesi, L., & Mosca, E. (2020). Network diffusion promotes the integrative analysis of multiple omics. *Frontiers in Genetics*, *11*, 106.
10. Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G., & Milanesi, L. (2016). Methods for the integration of multi-omics data: mathematical aspects. *BMC bioinformatics*, *17*(2), 167-177.
11. Di Nanni, N., Gnocchi, M., Moscatelli, M., Milanesi, L., & Mosca, E. (2020). Gene relevance based on multiple evidences in complex networks. *Bioinformatics*, *36*(3), 865-871.
12. Bersanelli, M., Mosca, E., Remondini, D., Castellani, G., & Milanesi, L. (2016). Network diffusion-based analysis of high-throughput data for the detection of differentially enriched modules. *Scientific reports*, *6*(1), 1-12.
13. Ruffalo, M., Koyutürk, M., & Sharan, R. (2015). Network-based integration of disparate omic data to identify" silent players" in cancer. *PLoS computational biology*, *11*(12), e1004595.
14. Paull, E. O., Carlin, D. E., Niepel, M., Sorger, P. K., Haussler, D., & Stuart, J. M. (2013). Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). *Bioinformatics*, *29*(21), 2757-2764.
15. Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary oncology (Poznan, Poland)*, *19*(1A), A68–A77. https://doi.org/10.5114/wo.2014.47136
16. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot TS, Malta TM, Pagnotta SM, Castiglioni I, Ceccarelli M, Bontempi G, Noushmehr H. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 2016 May 5;44(8):e71. doi: 10.1093/nar/gkv1507. Epub 2015 Dec 23. PMID: 26704973; PMCID: PMC4856967.
17. Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., ... & Mering, C. V. (2010). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*, *39*(suppl\_1), D561-D568.
18. Babbi, G., Martelli, P. L., Profiti, G., Bovo, S., Savojardo, C., & Casadio, R. (2017). eDGAR: a database of Disease-Gene Associations with annotated Relationships among genes. *BMC genomics*, *18*(5), 25-34.
19. Rappaport, N., Nativ, N., Stelzer, G., Twik, M., Guan-Golan, Y., Iny Stein, T., ... & Lancet, D. (2013). MalaCards: an integrated compendium for diseases and their annotation. *Database*, *2013*.
20. Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., ... & Hassanali, M. (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research*, *36*(suppl\_1), D901-D906.
21. Nitta, K. R., Jolma, A., Yin, Y., Morgunova, E., Kivioja, T., Akhtar, J., ... & Taipale, J. (2015). Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *elife*, *4*, e04837.
22. Kanehisa, M. (2002). The KEGG database.
23. Tenenbaum, D., RUnit, S., Maintainer, M. B. P., Carlson, M., biocViews Annotation, P., & ThirdPartyClient, K. E. G. G. (2019). Package ‘KEGGREST’. *R Foundation for Statistical Computing: Vienna, Austria*.