A web-based pathway enrichment analysis module for the PharMeBINet database

Ein webbasiertes Pathway-Enrichment-Analysemodul für die PharMeBINet-Datenbank

Cassandra Königs[1]
Theresa Dietrich[1]

1 Bielefeld University, Faculty of Technology, Bioinformatics/Medical Informatics Department, Bielefeld, Germany

Abstract

In modern molecular biology, the quantification of proteins, RNA, and DNA is a standard procedure. Resulting in the generation of large data, researchers need appropriate tools for interpretation. A common method for interpreting gene expression data is pathway enrichment analysis.
The heterogeneous pharmacological medical biochemical network (PharMeBINet) is a Neo4j database in combination with a website for browsing and analysis available at https://pharmebi.net. Here we present a new analysis module for the website developed in JavaScript that applies enrichment analyses to gene expression data. The analysis has two methods for enrichment: Fisher's exact test and modified Fisher's exact test. The modified test considers the order of the gene expression data. Additionally, Bonferroni-correction, Dunn-Šidák correction, Holm-Bonferroni method, and Benjamin-Hochberg method are implemented to reduce the false positive pathways. The analysis was tested with the gene expression data of the first cluster described by the analysis of Shin et al.
The result of Fisher's exact test with corrected p-values ($p<0.01$) was 68 pathways. In contrast, the result of the modified Fisher's exact test was 104 different pathways. The pathway with the best p-value is "Generation of second messenger molecules". The results are presented in multiple forms. The first is a table ordered by p-values. Secondly, a bar plot with the log10(p-value) for all pathways provides a general impression of the resulting pathways. Thirdly, a combination of heat map and bar plot for all pathway gene combinations shows an overview of how the genes are connected to the relevant pathways and with the p-values beside it. Further, the input data was analyzed. The results are presented as a pie chart and bar plot. The pie chart shows how many of the input genes have a connection to pathways and how many do not. The bar plot displays the number of enriched pathways the genes appear in. The resulting pathways are well-fitting results for the gene expression data. This analysis module returns similar results compared to other enrichment tools.

Keywords: pathway enrichment analysis, heterogeneous database

Zusammenfassung

In der modernen molekularen Biologie ist die Quantifizierung von Proteinen, RNA und DNA ein Standardverfahren. Für die daraus resultierenden großen Datenmengen benötigen Forscher geeignete Instrumente für die weiterführende Interpretation. Ein üblicher Ansatz zur Reduzierung einer großen Genliste mit zugehörigen Expressionsdaten ist die sogenannte Enrichment Analyse.
Das Heterogeneous Pharmacological Medical Biochemical Network (PharMeBINet) ist eine Neo4j-Datenbank in Kombination mit einer Website zur Betrachtung und Analyse der Datenbank und unter

https://pharmabi.net verfügbar. Hier präsentieren wir ein neues Analysemodul für die Website, welches Enrichment-Analysen auf Genexpressionsdaten anwenden kann und mit JavaScript programmiert wurde. Um signifikante Pathways zu berechnen, werden der exakte Test nach Fisher und eine modifizierte Variante des Tests verwendet. Der modifizierte Test berücksichtigt die Reihenfolge der Genexpressionsdaten. Außerdem werden die Bonferroni-Korrektur, Dunn-Šidák-Korrektur, Holm-Bonferroni-Methode und die Benjamini-Hochberg-Methode verwendet, um die Anzahl falsch positiver Pathways zu reduzieren. Als Beispiel für einen Anwendungsfall wurden die Genexpressionsdaten des Clusters 1, der Analyse von Shin et al., für die Enrichment-Analyse verwendet.
Infolgedessen werden 68 verschiedenen Pathways für den exakten Fisher-Test sowie 104 für den modifizierten Test mit dem angepassten p-Wert ($p<0,01$) gefunden. Der Pathway mit dem besten p-Wert ist „Generation of second messenger molecules“. Die Ergebnisse werden als Tabelle sortiert nach p-Wert dargestellt. Zweitens vermittelt ein Balkendiagramm mit dem log10(p-Wert) für alle Pathways einen allgemeinen Eindruck der resultierenden Pathways. Drittens gibt eine Heat-Map für alle Pathway-Genkombinationen einen Überblick darüber, wie die Gene mit den relevanten Pathways verbunden sind. Außerdem wird eine Analyse der Eingabedaten in einem Kreisdiagramm und einem Balkendiagramm angezeigt. Das Kreisdiagramm zeigt, wie viele der Input-Gene eine Verbindung zu Pathways haben oder nicht. Das Balkendiagramm zeigt die Anzahl der relevanten Pathways, in denen die Gene erscheinen. Die resultierenden Pathways passen gut zu den eingegebenen Genen. Auch im Vergleich zu anderen Enrichment-Tools haben sie ähnliche Ergebnisse.

Schlüsselwörter: Pathway-Enrichment-Analyse, heterogene Datenbank

Introduction

In recent years, the improvements of high-throughput technologies for Omics experiments produce more data sets for analysis [1], [2], [3]. Consequently, the challenge is to interpret these data for a better understanding of diseases or a given phenotype [1], [4], [5], [6]. The most common outputs are gene expression profiles which are compared to examine the changes in gene expression between different test groups [5]. A standard analysis method on such data is the statistical enrichment with pathways, molecular functions, or biological processes [7]. There are various algorithms to compute enrichments which can be divided into over-representation analysis (ORA) [8], functional class sorting, and topology-based methods [9]. Unfortunately, no standard method or preferred category exists as such but ORA and functional class sorting methods are utilized in many publications [1], [2], [9].
Currently, a lot of different pathway databases exist like Kyoto Encyclopedia of Genes and Genomes (KEGG) [10], Reactome [11], WikiPathways [12], and PathBank [13]. Every pathway database has its focus and applications. The choice of pathway databases influences the results of enrichment [14]. That is the reason why recently the usage of multiple pathway databases or a merged pathway database became more prevalent [14], [15], [16], [17].
Here, a new tool of PharMeBINet's web application is introduced allowing users to perform pathway enrichment analysis and investigate results in multiple ways. In PharMeBINet multiple pathway databases are connected and used for enrichment. For the enrichment analysis, two methods of the Fisher's exact test are used, the standard Fisher's exact test and a modified version of the Fisher's exact test, which considers the ranking of the gene set. Four correction approaches for multiple testing are available to the user controlling the rate of false positives. The results are visualized as bar plots, heat map, and pie chart. The tool is available at https://pharmebi.net/#/analysis/enrichmentanalysis.

Methods

In the following, the development of the website is described as well as the databases utilized for the enrichment methods. The implemented enrichment and correction methods are explained. Afterward, a short explanation of how to use the website and analysis module is described as well as the use-case for comparison.

Implementation of PharMeBINet

The web application is divided into frontend and backend. Vue.js 2.6.11 [18], an open-source JavaScript framework for user interfaces, generates the frontend. For the design of Vue, the library Vuetify 2.6.2 [19] is used intending to create intuitive and reliable user interfaces based on material design. Echarts 5.3.2 [20] generates different graphics and charts on the webpage. Node.js v14 LTS [21] is an open-source JavaScript framework which forms the backend. It communicates with the Neo4j database utilizing the library Neode 0.4.8 [22], a Neo4j object graph mapping (OGM) which takes care of creating, reading, updating, and deleting (CRUD) operations on data in the database. The communication between frontend and backend is implemented using Axios 0.27.2 [23], an open-source JavaScript library for HTTP requests.

Structure of the PharMeBINet database

The backend utilized the heterogeneous Neo4j database PharMeBINet [24]. Pathway information is derived from Comparative Toxicogenomics Database (CTD) (version 2022-04) [25], Pathway Commons (version 12) [26], Reactome (version 2022-03-31) [11], and WikiPathways (version 2022-04-10) [12].

The databases Pathway Commons and WikiPathways provide gene-pathway relationships. However, Pathway Commons is filtered for information with compatible licenses before integration. Reactome's pathway information is mapped via Reactome identifier and name. Reactome pathways that do not map to Pathway Commons or WikiPathways are added as new nodes. CTD is mainly used to fill up additional gene-pathway relationships. Gene information is derived from Entrez Gene [27]. The resulting database contains 3,689 different pathway nodes extracted from four databases.

Implementation of pathway enrichment analysis into PharMeBINet

Fisher's exact test is a common ORA method utilized for the pathway enrichment analysis [1]. A list of candidate genes is tested for over-representation in a list of genes connected to a given pathway. Fisher's exact test calculates the statistical significance between the genes of interest and the genes of a pathway. The null hypothesis for each pathway is that a special set of genes is not over-represented in comparison to all genes of the pathway. For this, a contingency table (Table 1) is utilized where **N** is the number of genes in the input list, **n** is the number of genes connected to the pathway, and **M** is the number of all genes.

The p-value is computed by the hypergeometric distribution:

$$Pr(x = k) = \frac{\binom{n}{k}\binom{M-n}{N-k}}{\binom{M}{N}}$$

For the right-tailed test, the p-value is computed with $Pr(x \leq k)$. The input list is significantly over-represented in a given pathway if the p-value falls below a threshold, typically set to $p<0.05$. The repeating calculation of Fisher's exact test on different pathways increases the possibility for false-positive results [28], [29]. For this problem, multiple methods to control the type I error have been developed like Bonferroni correction (BC), Dunn-Šidák correction (DS-correction), Holm-Bonferroni method (HB-correction), and Benjamini-Hochberg procedure (FDR-correction) [28], [29]. Dunn-Šidák and Bonferroni correction adjust the threshold with a number of tests [29]. In contrast, for the Benjamini-Hochberg procedure and the Holm-Bonferroni method the p-values are ordered and assigned a rank. Consequently, the threshold is changed by the number of tests and depending on their rank [28], [29].

Usually Fisher's exact test does not take the ranking of genes into account [15]. However, g:Profiler [15] developed an algorithm applying Fisher's exact test on an ordered gene list. First, the gene list is ordered, for example, by fold-change. For each prefix sublist of the ordered gene list with length $\leq$**N** the algorithm computes the intersection **x** to the genes of a pathway as well as the p-value. Afterward, the minimal p-value and respective prefix are returned.

Additionally, the size of expected overlap by chance is calculated by considering the amount of genes from a given pathway compared to the background genes and multiplied by the number of genes in the input list. The fold-enrichment value is computed by dividing the overlap of genes between the pathway and the input list through the expected overlap by chance.

The general usage of enrichment analysis

Not only genes can be utilized as input in the new PharMeBINet analysis module, but also diseases and chemicals (see Figure 1). The tool can perform enrichment analyses for pathways, diseases, chemicals, biological processes, cellular components, molecular functions, and genes. First, the threshold for rejecting the null hypothesis can be set manually. Secondly, it can be selected if the threshold is used on the raw p-value or the corrected p-value. Thirdly, four different correction methods can be selected. It is possible to use multiple corrections at the same time. A pathway passes the threshold if at least one of the selected correction methods is below the threshold. Lastly, it is possible to select the background value from the database or to define a custom value. The background from the database is defined as the number of nodes from the input label connected to the enrichment label.

The right side of Figure 1 shows the possibilities for input data. Genes, diseases, and chemicals can either be put in using a text field or by uploading a tab-separated values (TSV) file with or without a header. For both methods, the

Table 1: The contingency table for Fisher's exact test

	Entries in input list	Entries not in input list	Total
Connected with pathway	x	n – x	n
Not connected with pathway	N – x	M – (n + N) + x	M – n
Total	N	M - N	M

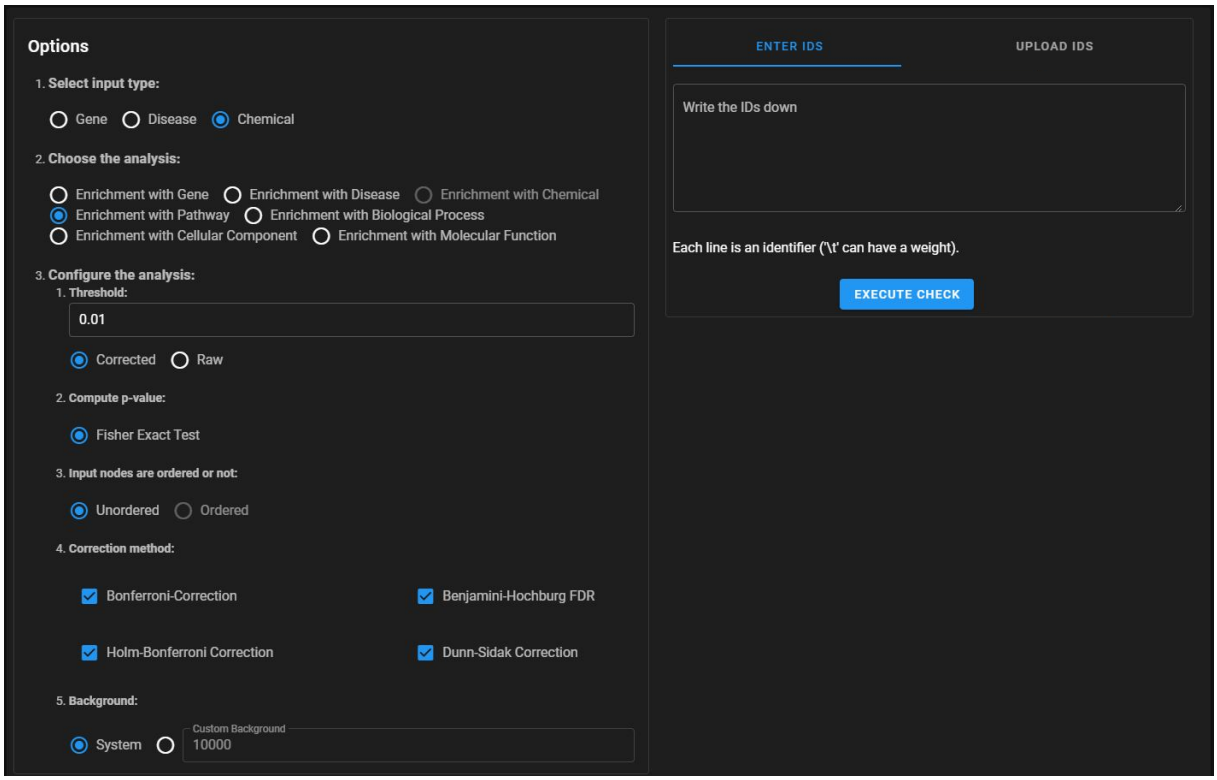

Figure 1: The graphical user interface on the website of PharMeBINet shows the general overview of all options.

first column represents an identifier field and the second column has some kind of weight. For genes, these weights could, for example, represent the fold-change.

Use-Case for pathway enrichment analysis of PharMeBINet

Shin et al. [30] took 63 samples from 62 patients with cutaneous T-cell lymphomas (CTCL) to analyze the gene expression profile. CTCL represents a subset of the non-Hodgkin lymphomas, with T-cell lymphomas present in the skin [31], [32]. They analyzed samples from patients with cancer stages IA, IB, IIB, and III. The expression results are clustered using hierarchical clustering and self-organizing maps, from which three unique clusters are extracted. A unique selling point is that the first cluster contains patients from all stages of cancer. Additionally, the authors performed pathway enrichment on the first cluster. This is why upregulated genes of this cluster 1 (see Table 2) are chosen for testing the pathway enrichment analysis presented here.

Results

For the given gene expression data of upregulated genes of cluster 1, PharMeBINet finds 68 different pathways (Attachment 1) of which the top fifteen are shown in Table 3. The most significant pathway is "Generation of second messenger molecules" with an adjusted p-value of 2.858e-13 and a fold-enrichment of 57.25. Conditions of the previous study are applied to this data set in which only pathways are considered which contain at least 5 genes [30], reducing the result to 45 different pathways. The enrichment with the modified Fisher's exact test for ranked genes returns 104 different pathways shown in Attachment 2. The most significant pathway is "Generation of second messenger molecules" too, but the adjusted p-value is better with 9.526e-14 and fold-enrichment of 61.57. The better values are because only the first 53 inputs were considered. Additionally, the top fifteen of the Fisher's exact and the modified version are the same but some are ranked differently. Moreover, with the condition to have at least five genes in overlap with the input list only 57 pathways are left.

In addition to the tabular demonstration of the pathway enrichment, there is a graphical visualization. A bar plot (Figure 2 and Attachment 3) shows the

Table 2: Gene cluster 1 is generated by hierarchical clustering and self-organizing maps. This cluster contains patients from all stages and is used for producing pathway enrichment [30].

# Gene symbol	Fold change	Gene id	# Gene symbol	Fold change	Gene id
1 IGLC1	8.63	3537	32 ST8SIA1	3.01	6489
2 IL26	8.44	55801	33 IL2RG	3.01	3561
3 IGHM	7.16	3507	34 LILRB4	2.95	11006
4 POU2AF1	5.93	5450	35 CCR4	2.91	1233
5 TNFRSF17	5.78	608	36 ITGAL	2.91	3683
6 T3JAM	4.73	80342	37 CD3E	2.89	916
7 LEF1	4.66	51176	38 CD3Z	2.88	919
8 SH2D1A	4.61	4068	39 PRF1	2.86	5551
9 GNLY	4.56	10578	40 CCR5	2.84	1234
10 PTPN7	4.45	5778	41 LTB	2.82	4050
11 FYB	4.44	2533	42 CCR7	2.72	1236
12 TNFSF14	4.16	8740	43 BIRC3	2.71	330
13 LCK	4.11	3932	44 MAP4K1	2.68	11184
14 RAC2	4.10	5880	45 IL27RA	2.67	9466
15 IL21R	4.09	50615	46 CD69	2.67	969
16 TCRIM	4.01	50852	47 IFNG	2.66	3458
17 ZAP70	3.89	7535	48 CXCR4	2.66	7852
18 TNFSF4	3.86	7292	49 CD3D	2.58	915
19 CCR8	3.65	1237	50 ST8SIA4	2.57	7903
20 FUT7	3.64	2529	51 MYB	2.55	4602
21 SELL	3.64	6402	52 ITGB7	2.50	3695
22 CD3G	3.64	917	53 LPXN	2.40	9404
23 CCL4	3.62	6351	54 PTPRC	2.27	5788
24 ITK	3.61	3702	55 TRA@	2.25	6955
25 CXCL13	3.60	10563	56 EVI2B	2.19	2124
26 IL2RA	3.49	3559	57 TNF	1.90	7124
27 TNIK	3.40	23043	58 TNFAIP8	1.87	25816
28 IL2RB	3.29	3560	59 WIF1	4.93	11197
29 TNFRSF7	3.23	939	60 IL1F7	4.17	27178
30 TRBC1	3.20	28639	61 C1orf68	3.71	100129271
31 LAT	3.16	27040	62 PSORS1C2	3.06	170680

−log10(p-value) for the raw p-values and the different corrected (BC, FDR-, DS-, HB-correction) p-values for each pathway. It is possible to zoom in and out with the slide bar. The bar plot demonstrates that the BC is the strictest correction. In contrast, the FDR-correction is the least strict. However, the p-value is strongly reduced by each correction method. The FDR-correction demonstrates that with increasing rank the distance to raw p-values is decreased. Also, the black dashed line defines the threshold. The last ten pathways are only considered due to the FDR-correction.

Another visualization is a combination of a heat map and bar plot as demonstrated in Figure 3 and Attachment 4. Each line represents a pathway and the heat map highlights the gene from the input list which is connected to this pathway. The bar shows the −log10(p-value) of the pathway. A slide bar makes it possible to zoom in and out of the heat map. The pathway "Immune System" includes the most genes from the input list but does not have the best p-value.

Further, a bar plot visualizes the input genes and their count of enriched pathways as seen in Figure 4 and Attachment 5. It demonstrates that some genes do not appear in any of the enriched pathways. In contrast, the gene *LCK* appears in 37 and in the modified Fisher's exact test in 52 different enriched pathways.

Another statistic is shown in Figure 5, demonstrating how many genes are in the input and how many have at least one connection to any pathway in PharMeBINet. The input lists consist of 62 different genes for this example and 5 have no connection to any pathway connection. The genes without any connection are *POU2AF1*, *C1orf68*, *PSORS1C2*, *TRA*@, and *EVI2B*.

Discussion

In the following, the results are compared and discussed. First, the results of the PharMeBINet enrichment analysis for the use-case are compared with the published results as well as the resulting pathways validated for CTCL. The results are compared between the standard and the modified Fisher's exact test and also between the different correction methods. Afterward, the results are compared with results from other available enrichment tools as well as their provided result visualizations.

Table 3: The top fifteen pathways from pathway enrichment for the gene cluster one for Shin et al. [30].

pathway	pathway ID	p-value	BC	FDR correction	DS correction	HB correction	rank	annotated genes	all genes considered input nodes	total number of considered input genes	number of nodes expected by chance	fold-enrich-ment
Generation of second messenger molecules	PC12_685	6.660e-16	2.860e-13	2.860e-13	2.860e-13	2.860e-13	1	10	40	57	0.1747	57.25
Th17 cell differentiation pathway	PC12_4425	8.440e-15	3.620e-12	1.810e-12	3.620e-12	3.620e-12	2	11	71	57	0.3101	35.48
Modulators of TCR signaling and T cell activation	PC12_4234	8.370e-14	3.590e-11	1.030e-11	3.590e-11	3.570e-11	3	10	61	57	0.2664	37.54
Immune System	PC12_5506	1.230e-13	5.280e-11	1.030e-11	5.280e-11	5.240e-11	4	34	2118	57	9.2496	3.68
TCR signaling	PC12_3341	1.270e-13	5.460e-11	1.030e-11	5.460e-11	5.410e-11	5	12	122	57	0.5328	22.52
T-cell receptor signaling pathway	PC12_4289	1.440e-13	6.180e-11	1.030e-11	6.180e-11	6.110e-11	6	11	91	57	0.3974	27.68
T Cell Receptor Signaling Pathway	PC12_2849	3.577e-13	1.535e-10	2.192e-11	1.513e-10	1.535e-10	7	47	9	57	0.2053	43.85
Translocation of ZAP-70 to Immunological synapse	PC12_684	1.280e-11	5.480e-9	6.840e-10	5.480e-9	5.390e-9	8	7	26	57	0.1135	61.65
Phosphorylation of CD3 and TCR zeta chains	PC12_683	3.000e-11	1.290e-8	1.430e-9	1.290e-8	1.260e-8	9	7	29	57	0.1266	55.27
TCR	PC12_3341	6.470e-11	2.780e-8	2.780e-9	2.780e-8	2.720e-8	10	13	260	57	1.1355	11.45
T-cell activation SARS-CoV-2	PC12_4679	1.180e-10	5.060e-8	4.600e-9	5.060e-8	4.940e-8	11	9	87	57	0.3799	23.69
T cell activation	PC12_136	2.240e-10	9.600e-8	8.000e-9	9.600e-8	9.350e-8	12	8	62	57	0.2708	29.55
T-cell antigen receptor (TCR) pathway during Staphylococcus aureus infection	PC12_4013	2.550e-10	1.100e-7	8.430e-9	1.100e-7	1.070e-7	13	8	63	57	0.2751	29.08
Cancer immunotherapy by PD-1 blockade	PC12_3991	5.04e-10	2.16e-7	1.55e-8	2.16e-7	2.100e-7	14	6	23	57	0.1004	59.73
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	PC12_660	7.85e-10	3.37e-7	2.25e-8	3.37e-7	3.260e-7	15	11	199	57	0.8691	12.66

BC: Bonferroni correction, FDR: false discovery rate (Benjamini-Hochberg correction), DS: Dunn-Šidák correction, HB: Holm-Bonferroni method

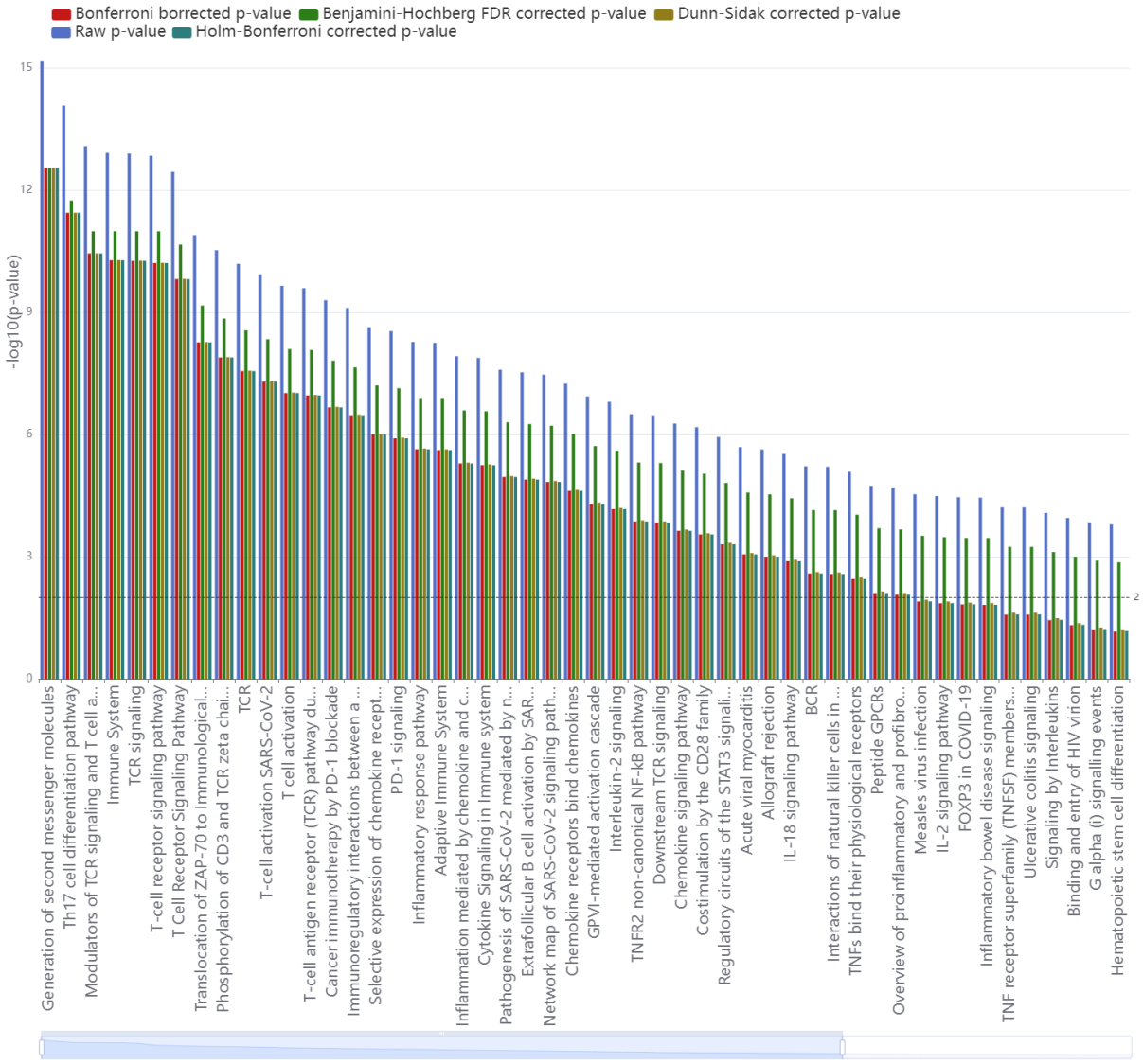

Figure 2: The figure demonstrates the p-values of the different enrichment pathways with –log10 of the p-value from the Fisher's exact method.

Comparison to the pathway enrichment analysis of Shin et al.

PharMeBINet's pathway enrichment analysis finds 68 relevant pathways for gene cluster 1. The 15 best hits mainly fit the "T cell receptor signaling pathway" (rank 5, 6, 10, and 13) and "T cell activation" (rank 11 and 12) comparable to the analysis of Shin et al. [30]. According to these pathways, "generation of second messenger molecules" (rank 1) is detected which is a specific part of the TCR-signaling in *H. sapiens*. Not only specific pathways are found but also the "immune system pathway" (rank 4) and "inflammatory response pathway" (rank 19), which represent both more general pathways and both contribute to the phenotype of CTCL. After TCR activation, naive CD4- T-cells differentiate into one of various T helper cells like Th17 cells [33]. In the Th17 cell differentiation pathway (rank 2), Th17 cells are required for immune responses to various extracellular bacteria and fungi. In CTCL, mycosis fungoides is the most common type. So, this suits genes that are highly expressed in this tissue. These cells produce Interleukin (IL)-17 among others, which describes a pathway Shin et al. found in their pathway-enrichment analysis. Not only similar results are found in terms of content, but also much more relevant pathways than in the analysis of Shin et al., which is due to the higher number and specificity of pathways. In PharMeBINet 3,689 pathways are included from three databases, whereas Shin et al. considered six databases with 408 pathways.

Some of the enriched pathways are opposites of each other like "PD-1 signaling" and "Cancer immunotherapy by PD-1 blockade'". The PD-1 blockade pathway is the pharmacological treatment of cancer by immunotherapeutic inhibition (such as Cemiplimab) of reactions in the "PD-1 signaling" pathway.

Multiple enriched pathways are part of inflammatory processes of the immune system. As T-cells play an im-

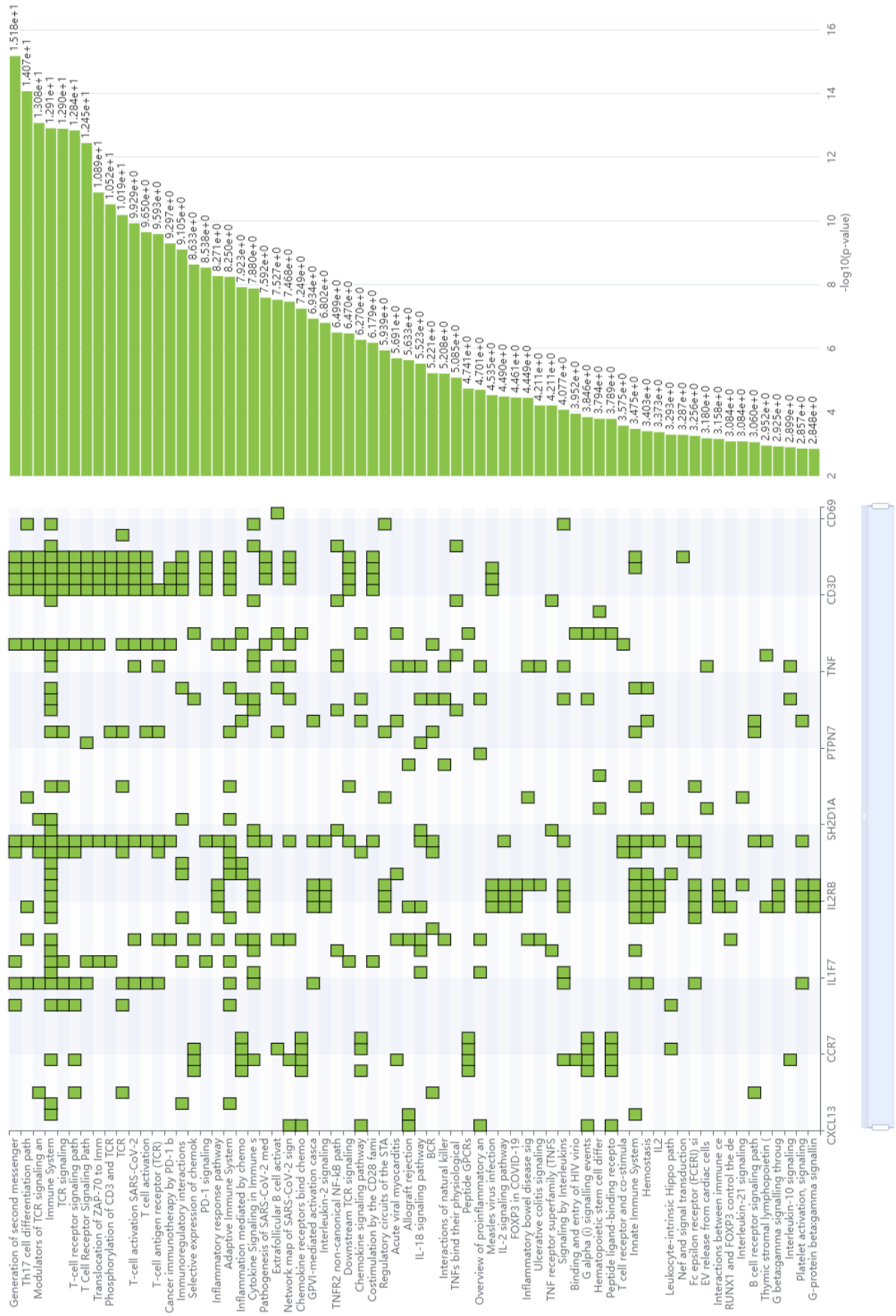

Figure 3: The heat map component of the visualization represents which genes in the input list are connected to which of the enriched pathways. A bar plot on the right shows the –log10(raw p-values) for each pathway. Here, the result of Fisher's exact test is shown.

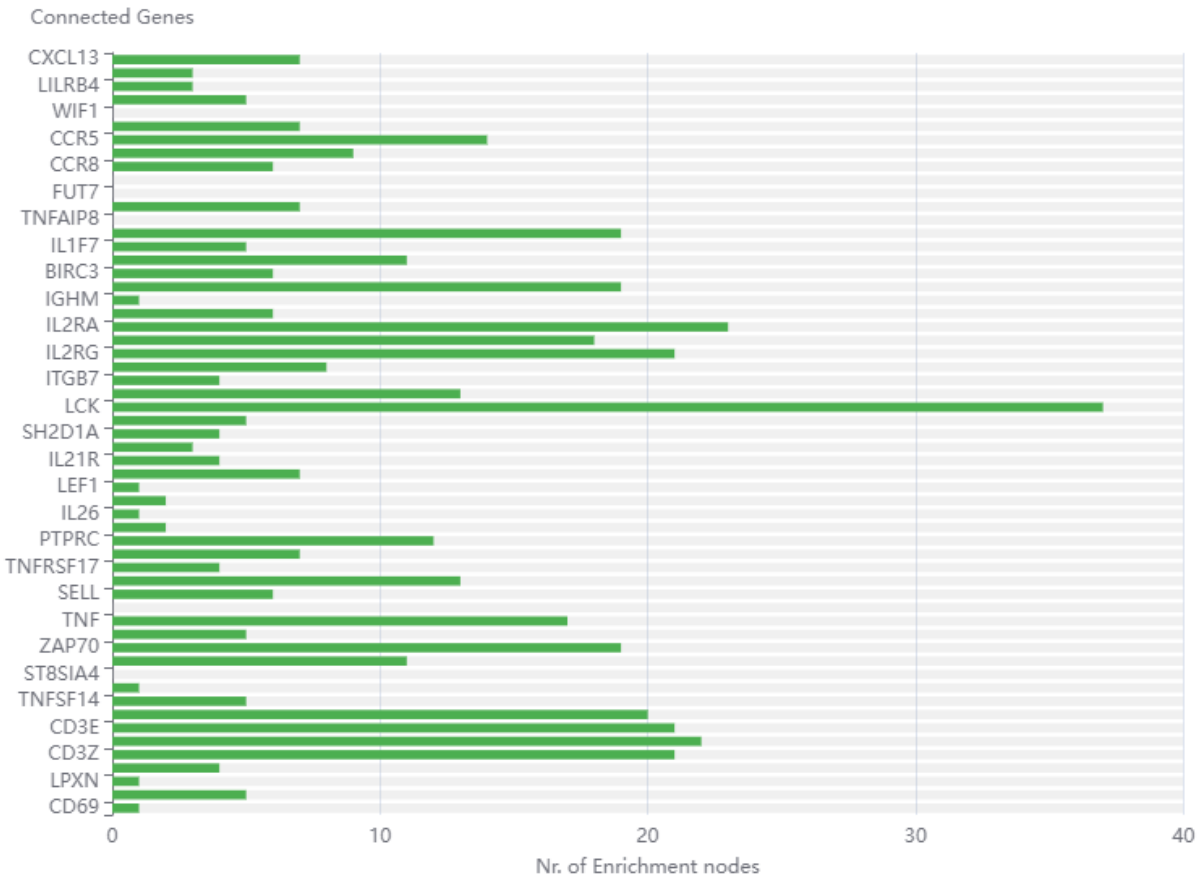

Figure 4: Bar plot for each input gene the number of enriched pathways in which the gene appears. Here, the result of Fisher's exact test is shown.

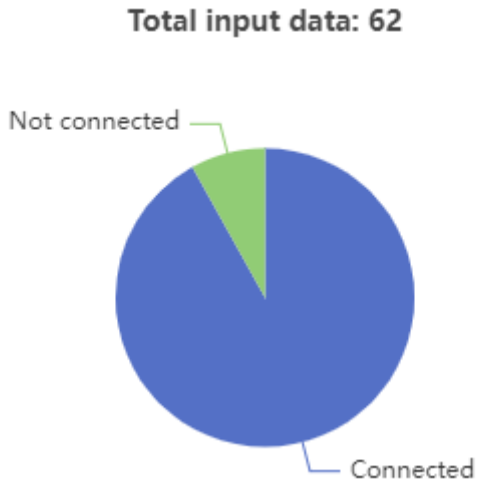

Figure 5: Pie chart of the gene input list. The total number is shown on the top. Five genes of the input list are not connected to pathways (green part). 57 have at least a connection to a pathway (blue part).

portant role in the immune system response, the inclusion of pathways involving T-cells like "T-cell activation SARS-CoV-2" and pathways involving T-cell receptor CD4 like "Binding and entry of HIV virion" in the results are to be expected.

Comparing Fisher's exact test and modified Fisher's exact test

The three best pathways of both results are equal, however, the p-values, fold enrichments and the numbers of nodes expected by chance are different. This difference occurs due to the use of a prefix sublist in the modified

test changing the value for **N**. However, the top 15 pathways of both enrichment tests are the same although their order, p-values, the number of nodes expected by chance, and the fold enrichments are different (see Table 3 and Attachment 2). This is caused by the different prefixes of the input list by the modified Fisher's exact test to get the best p-value and results in some genes of the input list not being considered even though they appear in the pathway. An example of this is the pathway "Innate Immune System" PC12 4895. In Fisher's exact test it has an overlap of 15 genes but in the modified version the overlap is only 14. In the modified version only the first 47 genes are considered so the gene PTPRC from position 54 is not considered (see Attachment 1, Attachment 2, Attachment 4, and Figure 3).
The pathway "Wnt signalling" is only part of the modified enrichment result. In the paper Shin et al. the association between the input gene and the pathway is described. This underlines that the modified Fisher's exact test returns additional useful pathways.
The number of results for Fisher's exact test and modified Fisher's exact test would be nearly the same if the FDR-correction is not used. The number would be 40 for Fisher's exact test and 43 for the modified test. This demonstrates that the other corrections reduce the number of results drastically for both enrichment methods.

Comparison of different correction methods

For the enrichment analysis, four different correction methods are implemented: BC, DS, HB, and FDR. BC is the strictest correction of the four (see Figure 2 and Attachment 3). This is caused by the fact that the family-wise error is below the given threshold [34]. In contrast, it increases the type II error filtering out significant pathways. The DS- and HB-correction are less strict than BC; however, for this example data set all three surpass the threshold for the same pathways (see Attachment 1, Attachment 2, and Attachment 3). The loosest correction is the FDR-correction which results in 28 additional pathways for the Fisher's exact test and 61 for the modified test.
"Measles virus infection" is the first pathway that is only included because of the FDR-correction. The paper Künzi et al. [35] shows that the use of the measles virus is a possibility to treat cutaneous T-cell lymphoma. This demonstrates that the pathway "Measles virus infection" is associated with cutaneous T-cell lymphoma and associated with the input list.
"T cell receptor and co-stimulatory signalling" is the first pathway of the modified Fisher's exact test result where only the FDR-correction value is under the threshold. This pathway is similar to the pathway "T cell receptor signaling pathway", which belongs to the top enriched pathways and is a good significant pathway for the input list.
Another pathway that belongs to both results but is only included because of the FDR-correction is "IL2". IL-2 plays a role in cutaneous T-cell lymphoma [32]. This demonstrates that many of the enriched pathways are only accepted because the FDR-correction is still significant.

Comparison to other pathway enrichment tools

For comparing the results of PharMeBINet with other enrichment tools the same gene list is submitted to g:Profiler [15], ToppGene [36], DecoPath [16], Panther [37], and Enrichr [38]. If possible the threshold is set to 0.01 or manually filtered to have similar start conditions. Table 4 shows the number of enriched pathways returned for which enrichment method, database, and correction method.
Panther returns the lowest number of pathways for the input list. It is tested with Fisher's exact test and binomial test with the FDR-correction. Both combinations return only three different pathways which are included in both PharMeBINet enrichments, as well (see Table 4).
The next tool is g:Profiler which uses three different databases KEGG, Reactome, and WikiPathways with modified Fisher's exact test and a custom correction method g:SCS threshold [15]. Each database returns its own pathways. The enrichment pathways from Reactome and WikiPathways are in the enrichment results of PharMeBINet. In contrast, four of the KEGG enrichment pathways are not in the PharMeBINet results (see Table 4).
Enrichr uses Fisher's exact test and the FDR-correction for multiple databases. Only the latest databases are utilized for the comparison: WikiPathways 2021 and KEGG 2021. The number of enriched pathways is higher than the number of g:Profiler. This may be because of the different database versions and/or because of the different correction methods. Not all WikiPathways 2021 results are included in the PharMeBINet results which may also be caused by the different database versions. The overlap between KEGG 2021 and PharMeBINet is low as seen in Table 4.
DecoPath uses Fisher's exact test and FDR-correction in combination with the databases KEGG, Reactome, WikiPathways, PathBank, and their own DecoPath database. The returned enrichment pathways are grouped by their respective database as in the other tools. This causes some pathways to appear multiple times from different databases and the reason why in Table 4 the number of common pathways is split into common pathways from PharMeBINet and common pathways from other databases such as DecoPath. More than half of the pathways of DecoPath appear in the PharMeBINet results and the modified Fisher's exact test has a greater overlap than the Fisher's exact test.
The last tool, ToppGene uses Fisher's exact test, FDR-correction, and multiple databases KEGG, MSigDB C2 BIOCARTA, BioCyc, Reactome, GenMAPP, Pathway Interaction Database, PantherDB, Pathway Ontology, and SMPDB. Also, like in DecoPath, all database results are in a single table increasing the number and causing the appearance of duplicated pathways. The overlap from

Table 4: This table demonstrates the number of results of the different enrichment tools and compares different online enrichment tools.

Tool	Database	Test – correction method	Number of enriched pathways	Fisher's exact		68	Fisher's exact modified		104
				common PharmMeBINet/ other data source	only in PharMeBINet	only in other	common PharmMeBINet/ other data source	only in PharMeBINet	only in other
g:Profiler	KEGG	modified Fisher's exact – g:SCS	14	10/10	58	4	10/10	94	4
	Reactome	modified Fisher's exact – g:SCS	15	15/15	53	0	15/15	89	0
	WikiPathways	modified Fisher's exact – g:SCS	17	17/17	51	0	1//17	89	0
DecoPath	KEGG, Reactome, WikiPathways, PathBank, DecoPath	Fisher's exact – FDR-correction	112	57/61	11	51	66/69	35	43
ToppGene	KEGG, MSigDB C2 BIOCARTA, BioCyc, Reactome, GenMAPP, Pathway Interaction Database, PantherDB, Pathway Ontology, SMPDB	Fisher's exact – FDR-correction	169	67/74	1	95	73/80	31	89
Panther	Panther	Fisher's exact – FDR-correction	3	3/3	65	0	3/3	101	0
	Panther	Binomial – FDR-correction	3	3/3	65	0	3/3	101	0
Enrichr	WikiPathways 2021	Fisher's exact – FDR-correction	29	23/23	45	6	26/26	78	3
	KEGG 2021	Fisher's exact – FDR-correction	34	9/9	59	25	9/9	95	25

ToppGene to PharMeBINet Fisher's exact test is high and nearly all pathways are included in PharMeBINet results. The overlap to the modified Fisher's exact test is even higher as the test returns more pathways.
In total, the results of PharMeBINet enrichment are on par with other tools. It does not yet include all pathways so a logical next step is to add more pathway databases such as KEGG.

Comparison of result presentation

Panther and ToppGene only provide table representations for the enrichment results. These tables are similar to the table of PharMeBINet.
Enrichr provides a tabular representation as well but also a log bar plot of the p-values/combined score rank/rank value for the top ten results. Additionally, it has a heat map for the top 30 enriched pathways with an additional bar plot in the graphic with log p-values/combined score rank/rank value. It is similar to the heat map and bar plot illustration of PharMeBINet. In contrast, in PharMeBINet all pathways are shown. However, Enrichr has additional plots on Appyter.
DecoPath also provides a tabular representation of the resulting pathways. Additionally, a pie chart shows the number of pathways which are concordant (multiple databases say it is significantly enriched), no mapping (pathways that are not mapped), and discordant (multiple databases say opposites). An additional table shows the pie chart's information. This representation is unique. The last plot of DecoPath is a pathway hierarchy that highlights the significant pathways. This is a unique plot as well.
g:Profiler provides a Manhattan plot and a combination of table, log bar plot, and heat map. This tool has some plots in common with PharMeBINet but includes some unique plots, too.
However, none of them shows a graphical representation of how many of the genes are used or not. They only represent this in text form. Additionally, the last bar plot in PharMeBINet is unique, showing the number of enriched pathways the genes appear in.

Conclusion

Here, we present a new enrichment analysis module for the PharMeBINet website. It allows setting different parameters manually. The pathway enrichment for the use-case retrieved good results for Fisher's exact test and modified Fisher's exact test. Multiple figures visualize the results of the enrichment analysis. Not only the ranks of the pathways are shown but also the overlap of genes. Additionally, a diagram shows how many genes have no connection to a pathway. Furthermore, a bar plot demonstrates the number of enriched pathways each gene appears in. The enrichment already returned good results and got similar results in comparison to other online tools. However, the enrichment analysis could be improved if more pathway databases are considered. Further, it would be a good idea to add other enrichment algorithms than Fisher's exact test as other methods are often used in research analyses as well.

Notes

Author's ORCID

- Cassandra Königs: 0000-0001-8991-9272

Competing interests

The authors declare that they have no competing interests.

Attachments

Available from https://doi.org/10.3205/mibe000243

1. Attachment1_mibe000243.csv (14 KB)
   Contains all enriched pathways for the Fisher's exact test.
2. Attachment2_mibe000243.csv (20 KB)
   Contains all enriched pathways for the modified Fisher's exact test.
3. Attachment3_mibe000243.png (227 KB)
   Demonstrates the p-values of the different enrichment pathways with –log10 as the bar plot of the p-value from the modified Fisher's exact method.
4. Attachment4_mibe000243.png (380 KB)
   Presents for each enriched pathway the genes of the input list which take place in the pathway in the heat map for each enriched pathway. Additionally, at the end of the line the –log10 (raw p-value) is demonstrated as a bar plot. These are the results of the modified Fisher's exact test.
5. Attachment5_mibe000243.png (42 KB)
   Bar plot for each input gene the number of enriched pathways in which the gene appears. This is the result of the modified Fisher's exact test.

References

1. Liu L, Wei J, Ruan J. Pathway Enrichment Analysis with Networks. Genes (Basel). 2017 Sep;8(10):246. DOI: 10.3390/genes8100246
2. Reimand J, Isserlin R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A, Wadi L, Meyer M, Wong J, Xu C, Merico D, Bader GD. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nat Protoc. 2019 Feb;14(2):482-517. DOI: 10.1038/s41596-018-0103-9

3. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloë D, Le Gall C, Schaëffer B, Le Crom S, Guedj M, Jaffrézic F; French StatOmique Consortium. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform. 2013 Nov;14(6):671-83. DOI: 10.1093/bib/bbs046
4. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005 Oct;102(43):15545-50. DOI: 10.1073/pnas.0506580102
5. Mathur R, Rotroff D, Ma J, Shojaie A, Motsinger-Reif A. Gene set analysis methods: a systematic comparison. BioData Min. 2018;11:8. DOI: 10.1186/s13040-018-0166-8
6. Nguyen TM, Shafi A, Nguyen T, Draghici S. Identifying significantly impacted pathways: a comprehensive review and assessment. Genome Biol. 2019 Oct;20(1):203. DOI: 10.1186/s13059-019-1790-4
7. Paczkowska M, Barenboim J, Sintupisut N, Fox NS, Zhu H, Abd-Rabbo D, Mee MW, Boutros PC; PCAWG Drivers and Functional Interpretation Working GroupReimand J; PCAWG Consortium. Integrative pathway enrichment analysis of multivariate omics data. Nat Commun. 2020 Feb;11(1):735. DOI: 10.1038/s41467-019-13983-9
8. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. Nat Genet. 1999 Jul;22(3):281-5. DOI: 10.1038/10343
9. Zyla J, Marczyk M, Domaszewska T, Kaufmann SHE, Polanska J, Weiner J. Gene set enrichment for reproducible science: comparison of CERNO and eight other algorithms. Bioinformatics. 2019 Dec;35(24):5146-54. DOI: 10.1093/bioinformatics/btz447
10. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. Nucleic Acids Res. 2021 Jan;49(D1):D545-D551. DOI: 10.1093/nar/gkaa970
11. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, Sidiropoulos K, Cook J, Gillespie M, Haw R, Loney F, May B, Milacic M, Rothfels K, Sevilla C, Shamovsky V, Shorser S, Varusai T, Weiser J, Wu G, Stein L, Hermjakob H, D'Eustachio P. The reactome pathway knowledgebase. Nucleic Acids Res. 2020 Jan;48(D1):D498-D503. DOI: 10.1093/nar/gkz1031
12. Martens M, Ammar A, Riutta A, Waagmeester A, Slenter DN, Hanspers K, A Miller R, Digles D, Lopes EN, Ehrhart F, Dupuis LJ, Winckers LA, Coort SL, Willighagen EL, Evelo CT, Pico AR, Kutmon M. WikiPathways: connecting communities. Nucleic Acids Res. 2021 Jan;49(D1):D613-D621. DOI: 10.1093/nar/gkaa1024
13. Wishart DS, Li C, Marcu A, Badran H, Pon A, Budinski Z, Patron J, Lipton D, Cao X, Oler E, Li K, Paccoud M, Hong C, Guo AC, Chan C, Wei W, Ramirez-Gaona M. PathBank: a comprehensive pathway database for model organisms. Nucleic Acids Res. 2020 Jan;48(D1):D470-D478. DOI: 10.1093/nar/gkz861
14. Mubeen S, Hoyt CT, Gemünd A, Hofmann-Apitius M, Fröhlich H, Domingo-Fernández D. The Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling. Front Genet. 2019;10:1203. DOI: 10.3389/fgene.2019.01203
15. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, Vilo J. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. 2019 Jul;47(W1):W191-W198. DOI: 10.1093/nar/gkz369
16. Mubeen S, Bharadhwaj VS, Gadiya Y, Hofmann-Apitius M, Kodamullil AT, Domingo-Fernández D. DecoPath: a web application for decoding pathway enrichment analysis. NAR Genom Bioinform. 2021 Sep;3(3):lqab087. DOI: 10.1093/nargab/lqab087
17. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, Benner C, Chanda SK. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun. 2019 Apr;10(1):1523. DOI: 10.1038/s41467-019-09234-6
18. Vue.js. Available from: https://vuejs.org
19. Vuetify. Available from: https://vuetifyjs.com
20. Apache ECharts. Available from: https://echarts.apache.org
21. Node.js. Available from: https://nodejs.org
22. Cowley A. Neode. Available from: https://github.com/adam-cowley/neode
23. vue-axios. Available from: https://npmjs.com/package/vue-axios
24. Königs C, Friedrichs M, Dietrich T. The heterogeneous pharmacological medical biochemical network PharMeBINet. Sci Data. 2022 Jul;9(1):393. DOI: 10.1038/s41597-022-01510-3
25. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, Wiegers J, Wiegers TC, Mattingly CJ. Comparative Toxicogenomics Database (CTD): update 2021. Nucleic Acids Res. 2021 Jan;49(D1):D1138-D1143. DOI: 10.1093/nar/gkaa891
26. Rodchenkov I, Babur O, Luna A, Aksoy BA, Wong JV, Fong D, Franz M, Siper MC, Cheung M, Wrana M, Mistry H, Mosier L, Dlin J, Wen Q, O'Callaghan C, Li W, Elder G, Smith PT, Dallago C, Cerami E, Gross B, Dogrusoz U, Demir E, Bader GD, Sander C. Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. Nucleic Acids Res. 2020 Jan;48(D1):D489-D497. DOI: 10.1093/nar/gkz946
27. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res. 2011 Jan;39(Database issue):D52-7. DOI: 10.1093/nar/gkq1237
28. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological). 1995 Jan;57(1):289-300. DOI: 10.1111/j.2517-6161.1995.tb02031.x
29. Noble WS. How does multiple testing correction work? Nat Biotechnol. 2009 Dec;27(12):1135-7. DOI: 10.1038/nbt1209-1135
30. Shin J, Monti S, Aires DJ, Duvic M, Golub T, Jones DA, Kupper TS. Lesional gene expression profiling in cutaneous T-cell lymphoma reveals natural clusters associated with disease outcome. Blood. 2007 Oct;110(8):3015-27. DOI: 10.1182/blood-2006-12-061507
31. Willemze R, Cerroni L, Kempf W, Berti E, Facchetti F, Swerdlow SH, Jaffe ES. The 2018 update of the WHO-EORTC classification for primary cutaneous lymphomas. Blood. 2019 Apr;133(16):1703-14. DOI: 10.1182/blood-2018-11-881268
32. Dummer R, Vermeer MH, Scarisbrick JJ, Kim YH, Stonesifer C, Tensen CP, Geskin LJ, Quaglino P, Ramelyte E. Cutaneous T cell lymphoma. Nat Rev Dis Primers. 2021 Aug;7(1):61. DOI: 10.1038/s41572-021-00296-9
33. O'Shea JJ, Steward-Tharp SM, Laurence A, Watford WT, Wei L, Adamson AS, Fan S. Signal transduction and Th17 cell differentiation. Microbes Infect. 2009 Apr;11(5):599-611. DOI: 10.1016/j.micinf.2009.04.007
34. Harris RJ. A primer of multivariate statistics. 3rd ed. New York, NY: Psychology Press; 2001. DOI: 10.4324/9781410600455
35. Künzi V, Oberholzer PA, Heinzerling L, Dummer R, Naim HY. Recombinant measles virus induces cytolysis of cutaneous T-cell lymphoma in vitro and in vivo. J Invest Dermatol. 2006 Nov;126(11):2525-32. DOI: 10.1038/sj.jid.5700529

36. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W305-11. DOI: 10.1093/nar/gkp427

37. Mi H, Ebert D, Muruganujan A, Mills C, Albou LP, Mushayamaha T, Thomas PD. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. Nucleic Acids Res. 2021 Jan;49(D1):D394-D403. DOI: 10.1093/nar/gkaa1106

38. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, McDermott MG, Monteiro CD, Gundersen GW, Ma'ayan A. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016 Jul;44(W1):W90-7. DOI: 10.1093/nar/gkw377

Corresponding author:
Cassandra Königs
Universität Bielefeld, Technische Fakultät, Arbeitsgruppe Bioinformatik/Medizininformatik, Postfach 10 01 31, 33501 Bielefeld, Germany
c.koenigs@uni-bielefeld.de

Please cite as
Königs C, Dietrich T. A web-based pathway enrichment analysis module for the PharMeBINet database. GMS Med Inform Biom Epidemiol. 2023;19:Doc04.
DOI: 10.3205/mibe000243, URN: urn:nbn:de:0183-mibe0002434

This article is freely available from
https://doi.org/10.3205/mibe000243

Published: *2023-07-04*