Report

Abstract

GSEA and ORA are both methods for characterizing expression changes associated with a phenotype of interest. In this project, different tools capable of ORA or GSEA were compared in the hopes of creating a wrapper that can be run through R. In addition, two datasets comparing modes in which samples are transported through a hospital were analysed as a demonstration of the effectiveness of the tools used.

Introduction

Overrepresentation analysis and gene set enrichment analysis are both methods of characterizing wide expression changes in any sample that goes through a mass spectrometer. They are necessary because it is impossible for a human to comprehensively go through the vast amount of data produced by a mass spectrometer. Luckily, there are large databases detailing the functions of almost all known genes via a controlled vocabulary, in the form of GO:terms, as well as through pathways recorded in pathway databases such as KEGG and Reactome.

The way these methods work is by testing either if the number of genes/proteins associated with a certain term is higher above some cutoff than what would be expected compared to the background in the case of overrepresentation analysis, and to check for coordinated changes throughout the output in the case of gene set enrichment analysis.

Overrepresentation analysis (ORA) works through Fisher's exact tests or the hypergeometric test, where one measures the likelihood of obtaining the observed number of genes associated with the gene set of interest above the defined cutoff. The genes above this cutoff are what is referred to as the foreground and the entire list of genes used in the test to calculate what is being tested for is what is referred to as the background. Overrepresentation analysis has a few weaknesses: it assumes that genes are independent from each other, ignoring the possibility of proteins working together, relies on an arbitrary p-value cutoff and considers all genes in the list equal, regardless of the magnitude of their fold change or p-value of their fold-change.

GSEA doesn't rely on an arbitrary p-value cutoff for its gene-lists. Instead, it takes a ranking, whether it is just a pre-ranked list or a list with an associated ranking metric, and checks if there are coordinated changes within this list, rather than comparing it to a background list. GSEA is largely based on the calculation of enrichment scores (ES) based on the fraction of genes in a set compared to all genes in the list [1] The statistical significance of a gene set is determined through a walk down the ranked list, where the „running sum" statistic is increased when a gene belonging to the set is encountered and decreased when a gene not belonging to it is encountered. The thus obtained enrichment score (ES) refers to the maximum deviation from zero achieved during his walk.The significance of the ES is calculated using permutation of the genes in the list. However, even GSEA has its weaknesses. Even though it is threshold-free and able to account for modest, coordinated changes, gene sets are still treated independently.

To address the weaknesses inherent to different approaches to ORA and GSEA, the aim of this project was to begin the creation of a wrapper for multiple GSEA/ORA tools after comparing the usefulness of them, in the hopes that combining the information returned by them one can obtain more nuanced information about what genes are associated with the phenotype of interest. This is especially interesting, since GSEA and ORA are not exact sciences yet, and there are differences in the ways that different tools obtain and correct their results.

Methods

Datasets

In the first stage of the project, I used various methods capable of overrepresentation analysis on 5 different data-sets. These datasets were the experimental results of a study using acute myeloid leukaemia CD4+ cells (available with GSEABenchmarker), two datasets concerning Dengue infected Huh7-liver cells, and two datasets which had been used to compare the impact of a human carrier and a pneumatic tube system on samples in the Kinderklinik in Bern, which were provided by the Mass Spectrometry Core Facility of the Inselspital.

The samples from the Kinderklinik were a dataset concerning a differential expression test of platelet-free-plasma from one donor between samples transported by a human carrier of blood samples in the hospital and the pneumatic tube system used to transport such samples automatically across the hospital, (hence referred to as the v´carrier vs tube system dataset), and the results of an ANOVA test across all platelet-free-plasma blood samples carried by the human sample carrier. Thus, this project is partly about evaluating the impact of the choice of transportation between the tube system and a human carrier of blood samples.

The datasets from the Kinderklinik and the Huh7 cells were analysed in more detail than the CD4 acute myeloid leukemia dataset. This is because the Huh7 cell datasets have already been analysed in some detail in the past, and are quite easy to work with, while the datasets from the Kinderklinik were datasets with as of yet unknown GSEA / overrepresentation analysis results.

Overrepresentation

Overrepresentation analysis relies on a background, which provides the reference as to what it is that one should expect. In this project, the background set was all genes detected in the experiment rather than the whole genome. The reason for this is that using the whole genome means that proteins that weren't detected during the experiment are also tested for, rather than the proteins that were detected. This will inevitably lead to inaccurate results as well as the detection of numerous terms that will be associated with the specific tissue that is being considered, rather than the changes caused by whatever disease, medication etc. the researcher is testing for. Nonetheless, almost all tools give the researcher the option to also use the entire human genome as the background. A p-value cutoff of 0.05 was used in each case to determine the foreground gene list.

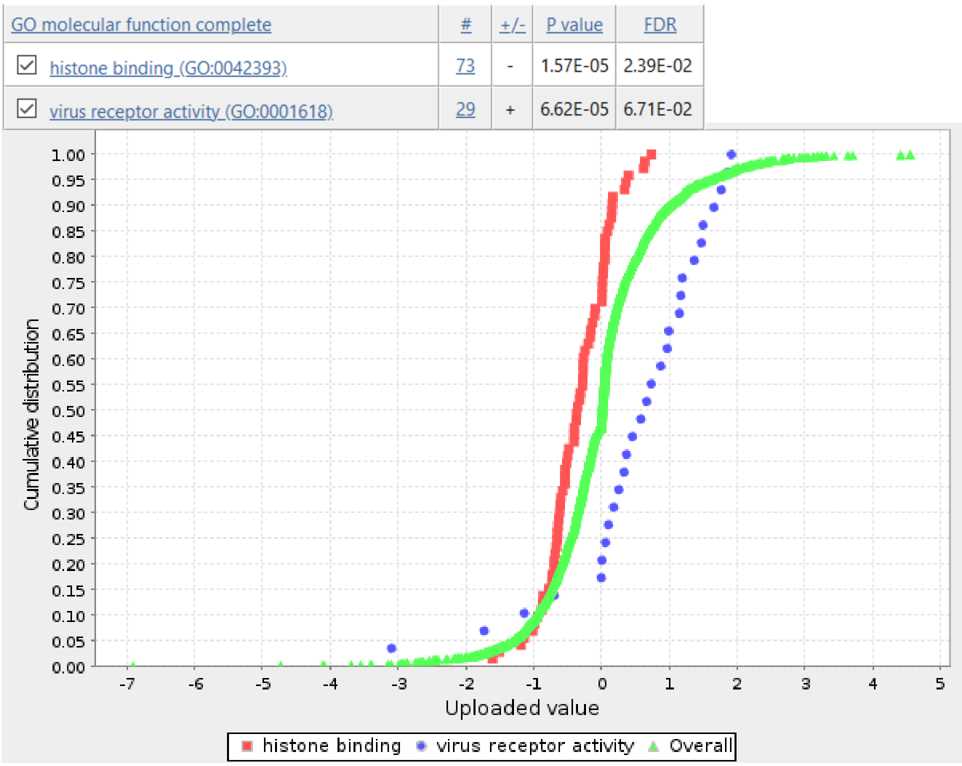
The tools used to do overrepresentation analysis on these datasets were the web-tools DAVID, PANTHER and gProfiler (of which gProfiler was queried through R), and SetRank.

Packages

The GSEA Benchmerker package contains 75 expression datasets for 42 well-characterized human diseases with a dedicated KEGG pathway associated with them. The creators of this package hope that this will allow researchers to use a gold standard for testing new methods for GSEA. [2]

SetRank is a package developed by Cedric Simillion with the aim of tackling problems like the arbitrariness of p-value cutoffs and the amount of false positives due to bias and pathway overlap. It uses a hypergeometric test, also called a one-sided Fisher's test. Its output includes a ranked list of pathways (terms) it found significant in the dataset, the membership of genes in each pathway, and an xls file that can be visualized in cytoscape to allow the visualization of the relationships between identified gene sets. When used in ranked mode, SetRank does not require a ranking metric, but instead a preranked list of genes. It also always requires a collection item, which refers to an object that is a compilation of the different annotation and pathway databases that SetRank is to work with [3].

PANTHER is an online tool widely used for data analysis. It uses a Fisher’s exact test to determine enrichment and Benjamini-Hochberg FDR correction when correcting for multiple testing. It can do both gene set enrichment and overrepresentation analysis. PANTHER is configurable, allowing for both Bonferroni and Fisher’s exact test to be performed, but for the purposes of this project, only Fisher’s exact test was used. It is able to display parent-child relationships between its results, meaning that terms that are more specific will be displayed under the more general term. This helps in the visualization of cases where genes map to multiple terms that are related, causing both to be significant. [4] In figure 1, it is demonstrated how panther can measure both over- and underrepresentation. The graphical representation allows the user to view the distributions of genes associated with terms compared to the overall distribution. One can see that the terms reach their peaks before the overall distribution, which indicates their significance. The nuances caused by over- and underrepresentation were not considered during this project. For ranked analysis, PANTHER needs a ranking metric of the users’ definition.



GO molecular function complete	#	+/-	P value	FDR
☑ histone binding (GO:0042393)	73	-	1.57E-05	2.39E-02
☑ virus receptor activity (GO:0001618)	29	+	6.62E-05	6.71E-02

Figure 1: cumulative distribution produced by PANTHER for two significant GOMF terms on the Huh7 cellular proteome, “histone binding” and “virus receptor activity”, compared to overall distribution.

DAVID is an older tool, which is only capable of doing different types of overrepresentation-tests. I used the “functional annotation chart” method of DAVID. For the majority of the project, the web-tool for DAVID was used instead of querying it through the R-package due to the fact that the documentation of the package was not clear. The package requires a registered email account as well as an up-to-date url for the david website to work, and will take the desired foreground and background gene-lists as objects and combine them into a david-object used in the query. It is highly configurable, but in this project I used fishers’ for correction, only looked for GO terms and only used the functional annotation chart [5].

gProfiler is a collection of tools rather than a single tool. In this project, I used g:GOSt, which can perform functional enrichment analysis. Unlike the other tools, g:GOSt by default uses a multiple correction method called g:SCS, which is, in terms of conservativeness, somewhere between the Benjamini-Hochberg FDR and the Bonferroni correction, and also uses the hypergeometric test instead of the Fisher's exact test. GOST is primarily an overrepresentation tool, but also allows for ranked lists to be used if its argument ordered_query is set to TRUE. However, whether or not it should be used for GSEA will be discussed later in this report. That being said, GOST takes a preranked list of genes when used in ranked analysis, so it does not need a ranking metric. [6]

Fgsea is an acronym for fast gene set enrichment analysis, and the name comes from how FGSEA calculates cumulative gene set enrichment values. This is where the "fast" in its name comes from, since the more traditional GSEA calculates an empirical null distribution, which takes time and requires a large amount of sampling, which makes it slow. The FGSEA algorithm works by incrementally adding genes to its calculation according to their rank, as explained in the introduction. FGSEA allows the visualization of this process, as shown in figure 2. Fgsea needs a ranking metric in order to work [7].

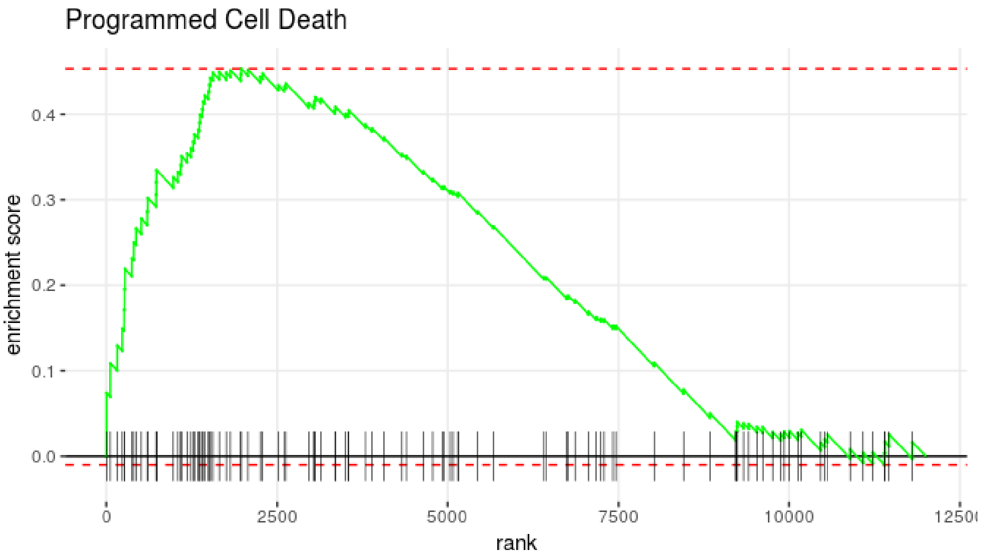


Figure 2: development of enrichment score for the term programmed cell death troughout fgsea analysis, as shown on the fgsea vignette.

The packages for PADOG, CAMERA and ORA were also used to compare their respective results on the GSEABenchmarker datasets, since Cedric Simillion had previously identified these packages as having good sensitivity and specificity [3], but for reasons which will be discussed later I will not explain how they work in detail. That being said, it is important to make the distinction between ORA the package, and ORA the unranked gene set analysis method. They are both acronyms for overrepresentation analysis, but the package will only be mentioned alongside CAMERA nad PADOG.

GSEA

In the second part of the project, I compared tools capable of ranked anaylsis. This includes the previously mentioned g:Profiler, PANTHER and SetRank, since, as mentioned, these also allow the user to use rankings, but also fgsea and briefly PADOG, ORA and CAMERA.

At this stage the decicion was made to move away from using GO terms for two reasons: firstly, working with GO terms takes a long time in terms of processing with the available hardware, especially when building collections fro SetRank. Thirdly, the fgsea package uses the reactome database and analyses pathway enrichment, not allowing the usage of GO terms unless it is provided an annotation for ENTREZ ids with GO terms. Thus, to make the packages more comparable, it was more reasonable to compare the pathways found significant between them.

The usage of PADOG, CAMERA and ORA on the CD4 dataset was justified in the sense that they can serve as a point of reference as to how much overlap can be expected between different tools. If they resulted in very high overlap, it would have indicated that that was what one should expect to be the result when comparing tools.

Results

The differences between the GO terms returned by each tool were analysed. For the overrepresentation tests, there was a lot of variability between the tools. SetRank and PANTHER of didn’t find any significant terms in certain datasets (ANOVA dataset of human carrier samples, Huh7 secretome), indicating that especially SetRank and PANTHER were much more stringent about when a term was significant. In figure three, the results from the overrepresentation analysis from DAVID, gProfiler, PANTHER and SetRank are shown for the cellular proteome and secretome of Dengue infected Huh7 cells.

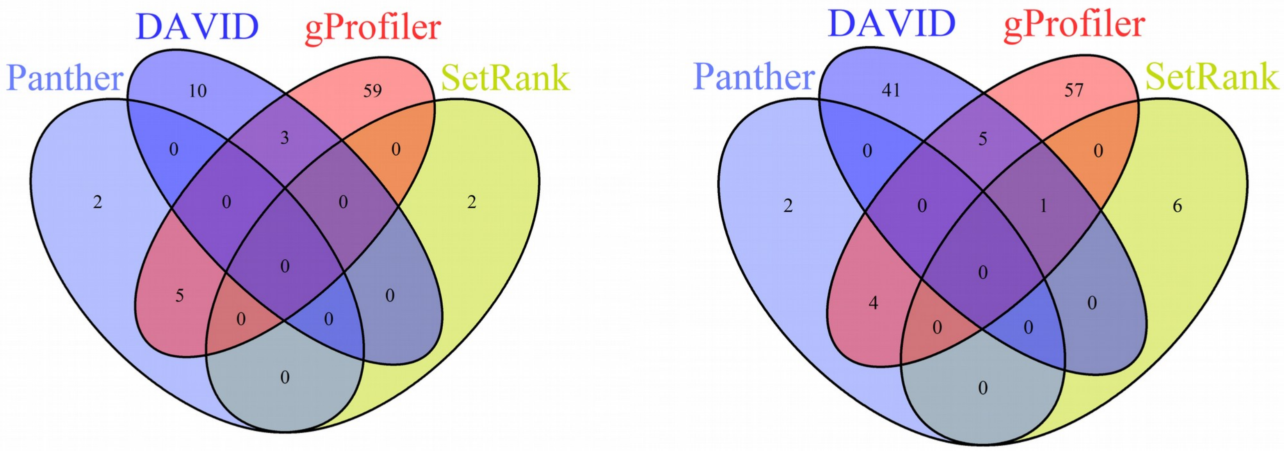


Figure 3: These Venn diagrams represent the overlap of GO terms produced by different tools doing overrepresentation analysis, using the secretome (left) and the cellular proteome (right) of Dengue-virus infected Huh7 liver cells.

The results in figure 3 are the ones highlighted because they were the only ones where each test returned some significant terms. This illustrates two things: firstly, there was surprisingly little overlap between the terms, and secondly, gProfiler and DAVID returned significantly more results than SetRank and PANTHER. This might be because overrepresentation tests are less sensitive to coordinated changes in gene expression, but it is strange that even though gProfiler and DAVID returned so many results, only very few overlapped with SetRank and PANTHER, since one could expect the results for the more stringent terms to be also found in the output for the less stringent ones.

Next, I was looking at gene set enrichment analysis using ranked gene lists.
First, I tried to get PADOG, ORA and CAMERA to work on the Huh7 datasets, since they are silently installed with the GSEABenchmarker package. However, the packages were not designed for datasets provided by the user, as they make specific transformations to the data necessary. In figure 4, the overlap of GO terms between ora, padog and CAMERA on the CD4 acute myeloid leukemia dataset, numbered 27 in the GSEABenchmarker package, is shown.

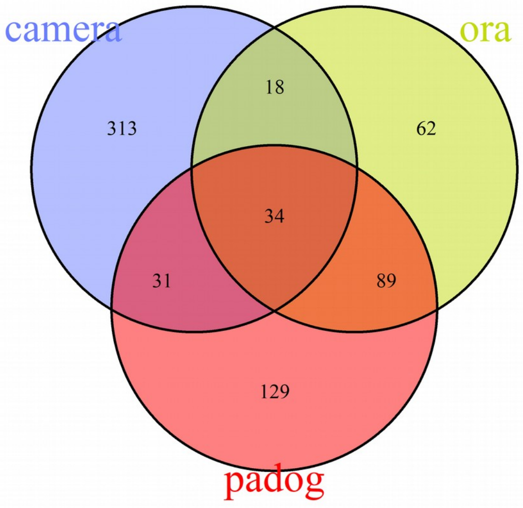
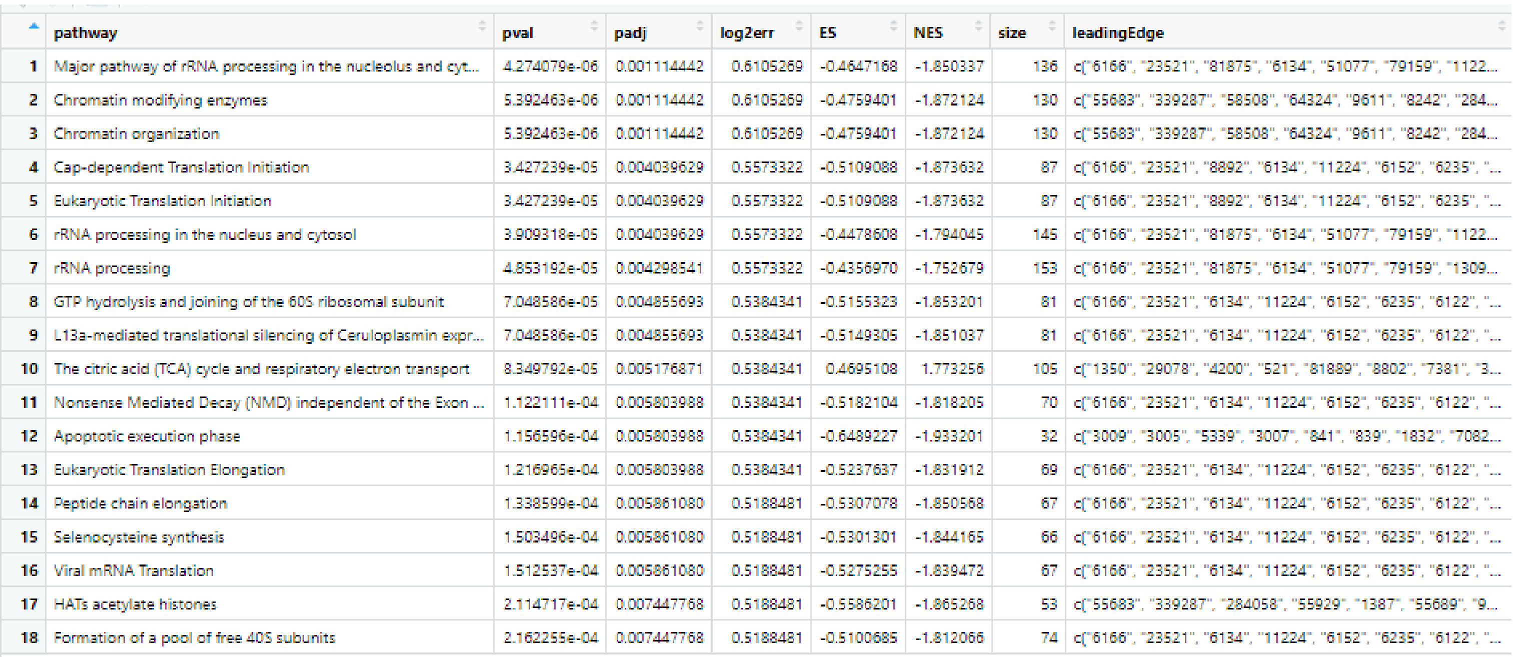


Figure 4: overlap of GO terms detected as significant by PADOG, CAMERA and ORA when using the CD4 dataset (number 27 in the GSEABenchmarker package) fir acute myeloid leukaemia

PADOG, CAMERA and ORA are all methods with good accuracy and sensitivity, as described by Cedric Simillion [3]. Even so, the overlap in the identifiers they returned was by no means large. This might just be the state of the current tools, meaning that due to differences in correction and calculation of initial p-values, different ID:s are selected as enriched.

Fgsea needs a special .rnk file with ranked gene names and a ranking metric, and will use the reactome pathway database as it’s source. The first two datasets that fgsea was used on were the Huh7 liver-cell datasets. For the ranking metric, I used the log10 p-value divided by the sign of the fold change, which gives, rather than a statistical meaning, a biological meaning through the ranking, as the ranking metric reflects the intensity of the increase or decrease.



	pathway	pval	padj	log2err	ES	NES	size	leadingEdge
1	Major pathway of rRNA processing in the nucleolus and cyt...	4.274079e-06	0.001114442	0.6105269	-0.4647168	-1.850337	136	c("6166", "23521", "81875", "6134", "51077", "79159", "1122...
2	Chromatin modifying enzymes	5.392463e-06	0.001114442	0.6105269	-0.4759401	-1.872124	130	c("55683", "339287", "58508", "64324", "9611", "8242", "284...
3	Chromatin organization	5.392463e-06	0.001114442	0.6105269	-0.4759401	-1.872124	130	c("55683", "339287", "58508", "64324", "9611", "8242", "284...
4	Cap-dependent Translation Initiation	3.427239e-05	0.004039629	0.5573322	-0.5109088	-1.873632	87	c("6166", "23521", "8892", "6134", "11224", "6152", "6235", "...
5	Eukaryotic Translation Initiation	3.427239e-05	0.004039629	0.5573322	-0.5109088	-1.873632	87	c("6166", "23521", "8892", "6134", "11224", "6152", "6235", "...
6	rRNA processing in the nucleus and cytosol	3.909318e-05	0.004039629	0.5573322	-0.4478608	-1.794045	145	c("6166", "23521", "81875", "6134", "51077", "79159", "1122...
7	rRNA processing	4.853192e-05	0.004298541	0.5573322	-0.4356970	-1.752679	153	c("6166", "23521", "81875", "6134", "51077", "79159", "1309...
8	GTP hydrolysis and joining of the 60S ribosomal subunit	7.048586e-05	0.004855693	0.5384341	-0.5155323	-1.853201	81	c("6166", "23521", "6134", "11224", "6152", "6235", "6122", "...
9	L13a-mediated translational silencing of Ceruloplasmin expr...	7.048586e-05	0.004855693	0.5384341	-0.5149305	-1.851037	81	c("6166", "23521", "6134", "11224", "6152", "6235", "6122", "...
10	The citric acid (TCA) cycle and respiratory electron transport	8.349792e-05	0.005176871	0.5384341	0.4695108	1.773256	105	c("1350", "29078", "4200", "521", "81889", "8802", "7381", "3...
11	Nonsense Mediated Decay (NMD) independent of the Exon ...	1.122111e-04	0.005803988	0.5384341	-0.5182104	-1.818205	70	c("6166", "23521", "6134", "11224", "6152", "6235", "6122", "...
12	Apoptotic execution phase	1.156596e-04	0.005803988	0.5384341	-0.6489227	-1.933201	32	c("3009", "3005", "5339", "3007", "841", "839", "1832", "7082...
13	Eukaryotic Translation Elongation	1.216965e-04	0.005803988	0.5384341	-0.5237637	-1.831912	69	c("6166", "23521", "6134", "11224", "6152", "6235", "6122", "...
14	Peptide chain elongation	1.338599e-04	0.005861080	0.5188481	-0.5307078	-1.850568	67	c("6166", "23521", "6134", "11224", "6152", "6235", "6122", "...
15	Selenocysteine synthesis	1.503496e-04	0.005861080	0.5188481	-0.5301301	-1.844165	66	c("6166", "23521", "6134", "11224", "6152", "6235", "6122", "...
16	Viral mRNA Translation	1.512537e-04	0.005861080	0.5188481	-0.5275255	-1.839472	67	c("6166", "23521", "6134", "11224", "6152", "6235", "6122", "...
17	HATs acetylate histones	2.114717e-04	0.007447768	0.5188481	-0.5586201	-1.865268	53	c("55683", "339287", "284058", "55929", "1387", "55689", "9...
18	Formation of a pool of free 40S subunits	2.162255e-04	0.007447768	0.5188481	-0.5100685	-1.812066	74	c("6166", "23521", "6134", "11224", "6152", "6235", "6122", "...

Table 1: Fgsea most significant pathways of cellular proteome

The pathways detected by fgsea were largely the same (at least in the case of the Huh7 cellular proteome) as what was produced by PANTHER. The pathways were also very descriptive of what was happening in the sample, including such pathways like “viral mRNA synthesis” and “major pathway of rRNA processing in the nucleolus and cytosol”, which were both also found by PANTHER.

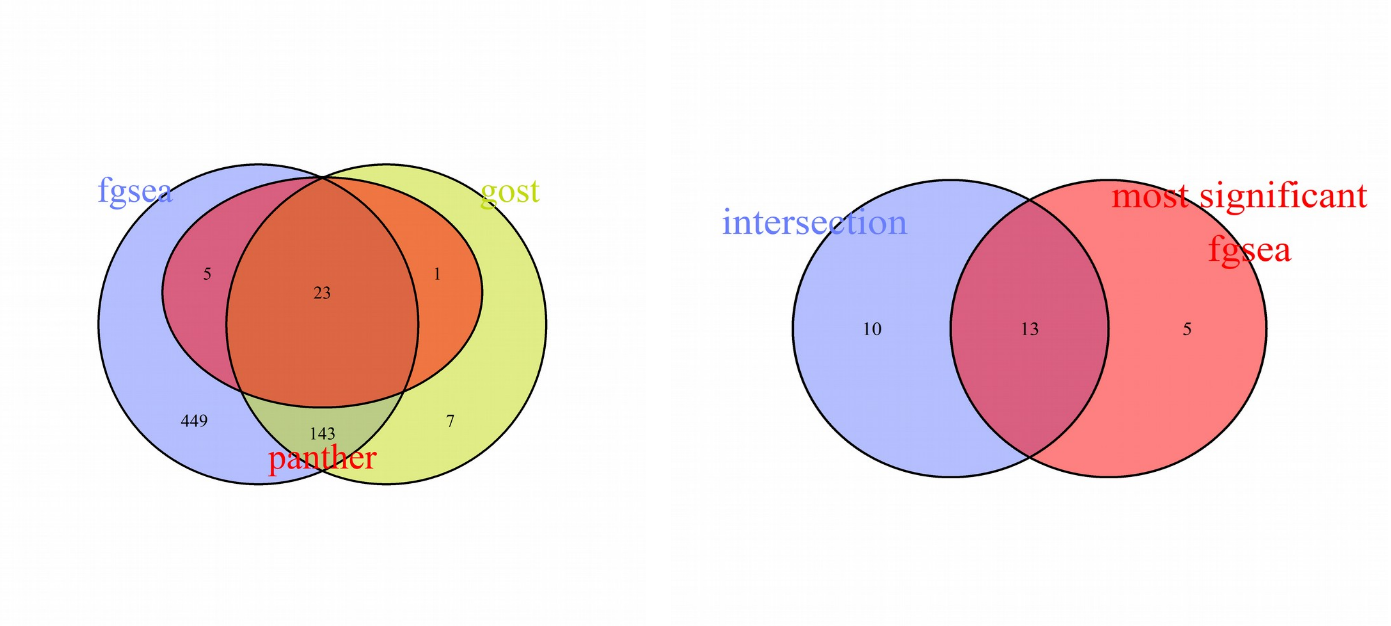


Figure 4: pathways found by PANTHER, FGSEA and gProfiler for the Dengue-infected Huh7-proteome (left), and overlap of the intersection of these three tools and the most significant results from FGSEA (adjusted p-value 0.01)

In figure 4, it is shown how, firstly, all reactome pathways found by PANTHER are also found by either gProfiler and FGSEA, and secondly, how thirteen out of the eighteen most significant results (which are also shown in table 1) are found in the intersection of all three tools. This creates a stark contrast to what was found when doing overrepresentation analysis, where there was great variation and precious little overlap. This goes to show the reliability of fgsea in finding terms that multiple tools can agree on, regardless of their differences, making it an interesting tool for the wrapper.

If the results for SetRank are also added to the analysis, the two pathways that all 4 tools agree on are "Major pathway of rRNA processing in the nucleolus and cytosol" and "Cellular responses to stress". This probably means that these are pathways that are definitively affected, since they were found regardless of correction method. This conclusion is supported by the fact that "Major pathway of rRNA processing in the nucleolus and cytosol" is also the most significant pathway found by fgsea.

When analysing the dataset for the ANOVA of the sample carrier, similarly to what was observed in the Huh7 secretome, fgsea didn’t find any significant results below p=0.01 for the tube carrier, and PANTHER found no significant pathways. However, the analysis for the dataset of the carrier vs pneumatic tube system was more informative. The overlap is presented in figure 5.

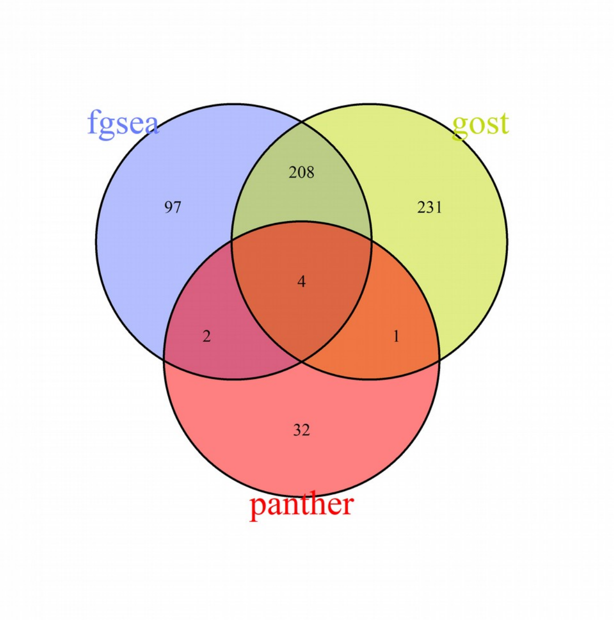


Figure 5: overlap of pathway analysis for the dataset for the expression change between carrier and pneumatic tube system. Additionally, there was no overlap between the intersection and the most significant terms found by fgsea.

One could say that there were some coordinated changes. While, upon the addition of SetRank to the analysis, the combination of all tools doesn’t yield any pathways that they would all agree on, they still found similar pathways to be significant. Pathways related to RHO-GTPases, which are proteins regulating intracellular actin dynamics, keratinization and chromatin remodelling are detected by multiple tools, but with different names. The rankings of the pathways returned by different tools also vary greatly.

This might mean that even though it would be tempting to just combine the output from multiple tools and get an intersect for the results, there can easily be enough variation to cause the elimination of all results.

In order to allow fgsea to work with GO terms, I attempted to create the appropriate annotation for ENTREZ-identifiers using the wrapper for ID conversion provided by SetRank. The full human gene ontology annotation was used for this. However, when the resulting table with ENTREZ ids mapped to GO terms was examined, a large portion of the uniprot identifiers had not been converted to ENTREZ ids. 58.7% of all unique uniprot ids had not been mapped, and when randomly sampling the unconverted ids, a number of the sampled ids had existing ENTREZ ids that had not been found by the converter.

To adress this problem, the package UniProt.ws [8] was employed. This package has an up-to-date ID converter. However, the respective function is prone to crashing if more than 100 ids are queried at the same time, and will encounter problems if it is unable to map ids. Due to these problems and in the interest of time, fgsea is not yet used in the final wrapper, until a more reliable way to convert uniprot ids into ENTREZ ids has been tested. That being said, even though UniProt.ws seemed unreliable, it was still able to map ids that the wrapper from SetRank was not able to map.

Systematic analysis

An attempt was made to reproduce the graph showing the SetRank network for the CD4 acute myeloid leukaemia, visualized in cytoscape, presented in the paper by Mr. Simillion [3]. However, SetRank returned a far richer network (figure 7) than what was shown in the paper by Simillion. This might suggest that the annotation might have changed since then, or the original data may have undergone some additional trimming before analysis. The cytoscape view of the output shows gene sets as nodes and the edges show the interactions between them.

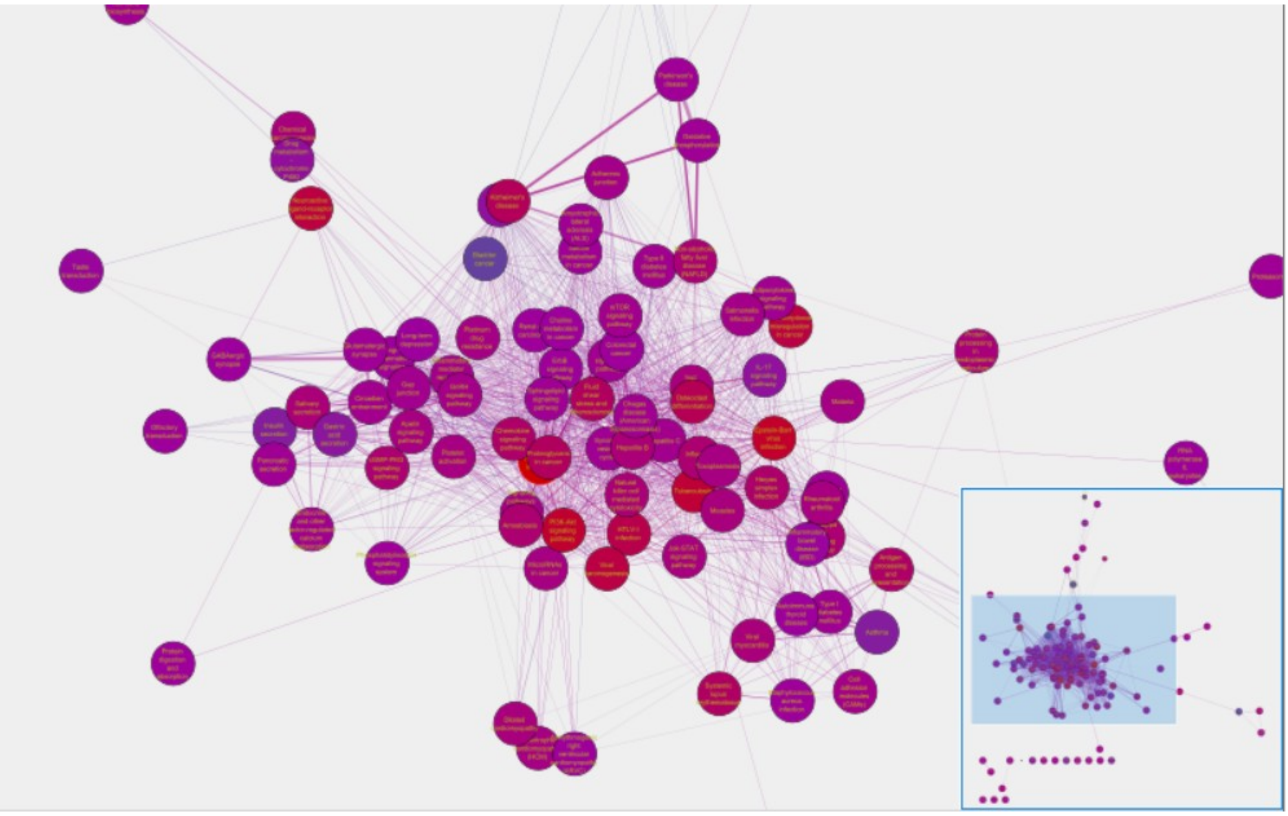
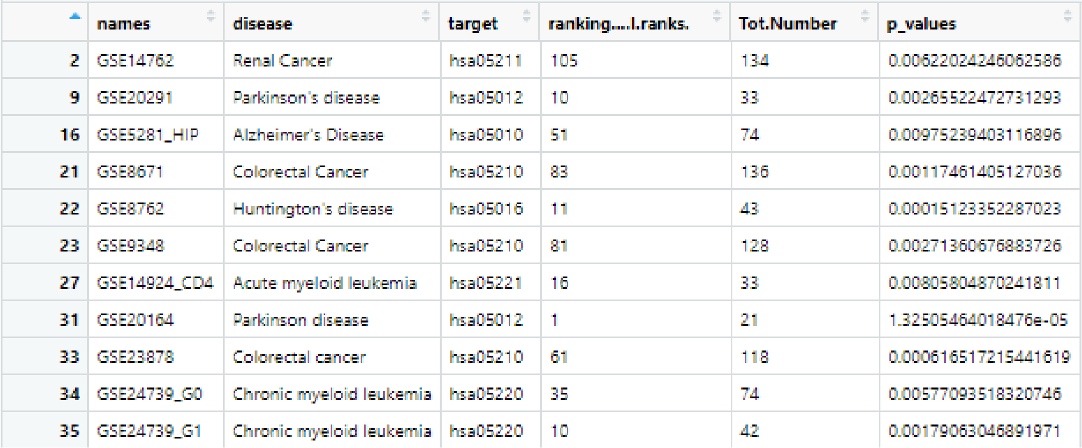


Figure 7: Cytoscape network view of the SetRank output for the CD4 acute myeloid leukaemia dataset, which was also used in the paper by Cedric Simillion

SetRank was also looped through all GSEABenchmarker datasets. Each of the datasets were for a defined disease, all of which (except for one) had a defined KEGG pathway associated with them. The rank of the pathway in question was recorded, as well as the associated p-value, if it was present. This search was done using the KEGG pathway identifiers.

Out of 42 datasets, SetRank found the specific disease pathway for 11 datasets, as shown in table 2.



	names	disease	target	ranking....l.ranks.	Tot.Number	p_values
2	GSE14762	Renal Cancer	hsa05211	105	134	0.00622024246062586
9	GSE20291	Parkinson's disease	hsa05012	10	33	0.00265522472731293
16	GSE5281_HIP	Alzheimer's Disease	hsa05010	51	74	0.00975239403116896
21	GSE8671	Colorectal Cancer	hsa05210	83	136	0.00117461405127036
22	GSE8762	Huntington's disease	hsa05016	11	43	0.00015123352287023
23	GSE9348	Colorectal Cancer	hsa05210	81	128	0.00271360676883726
27	GSE14924_CD4	Acute myeloid leukemia	hsa05221	16	33	0.00805804870241811
31	GSE20164	Parkinson disease	hsa05012	1	21	1.32505464018476e-05
33	GSE23878	Colorectal cancer	hsa05210	61	118	0.000616517215441619
34	GSE24739_G0	Chronic myeloid leukemia	hsa05220	35	74	0.00577093518320746
35	GSE24739_G1	Chronic myeloid leukemia	hsa05220	10	42	0.00179063046891971

Table 2: The table presenting the results for the SetRank loop through the datasets present in the GSEABenchmarker R-package. The colums are such: index of the dataset in the package, name of the GSEA dataset, name of the specific disease, the ID of the target pathway, rank of the found ID in the SetRank output, the total number of KEGG pathways identified, and the p-value of the rank.

This test was mostly to test for the ability of SetRank to definitively find the pathway best describing the disease found in each dataset.

Framework for the wrapper:

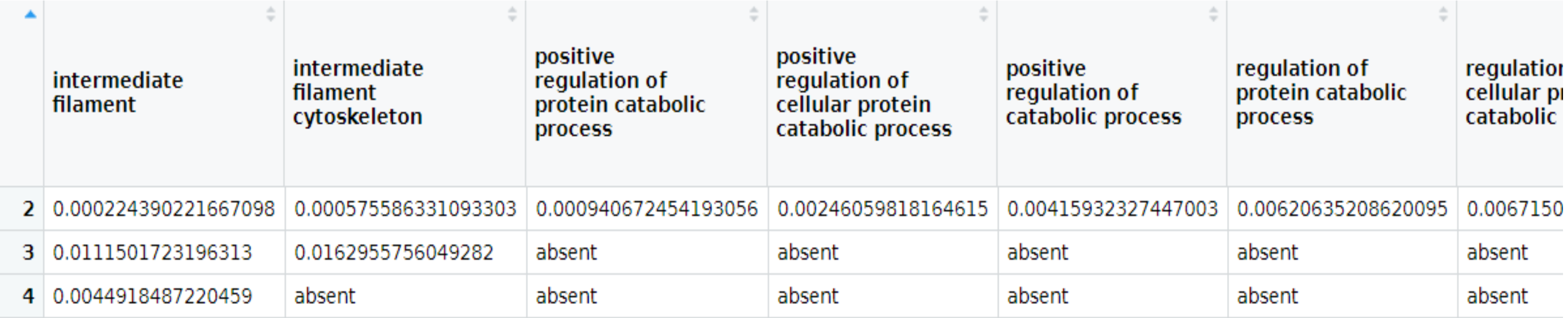
Ultimately, due to the differences in the number of significant results returned by different tools, it was necessary to limit the results of the analysis. This was due to the fact that the output of g:Profiler, which ended up being one of only two tools used in the ranked analysis in the wrapper, can be so much larger than that of SetRank. Once the number of columns exceeds a certain size, a user can’t properly view the results in R. However, this cutoff is left to the users discretion.

Due to time restrictions, fgsea and PANTHER couldn’t be included in the final wrapper. Fgsea because the conversion of the uniprot annotation into ENTREZ couldn’t be easily completed, and PANTHER because the algorithm wasn’t easily reproducible.

The necessary libraries/ to run this framework are: dplyr, tidyr, annotate, RdavidWebservice, EnrichmentBrowser, gprofiler2, openxlsx, SetRank, reactome.db, GeneSets.Homo.sapiens and pRoloc. The first three to be mentioned are mostly for data processing before the analysis, openxlsx is used so sheets can be imported into R, reactome.db is also mentioned, because the code also has parts that could be adapted to incorporate fgsea. pRoloc deserves special mention, as it hasn’t been mentioned before. It is a package that allows GO ids be converted into the term names, which will make the output more human-readable.

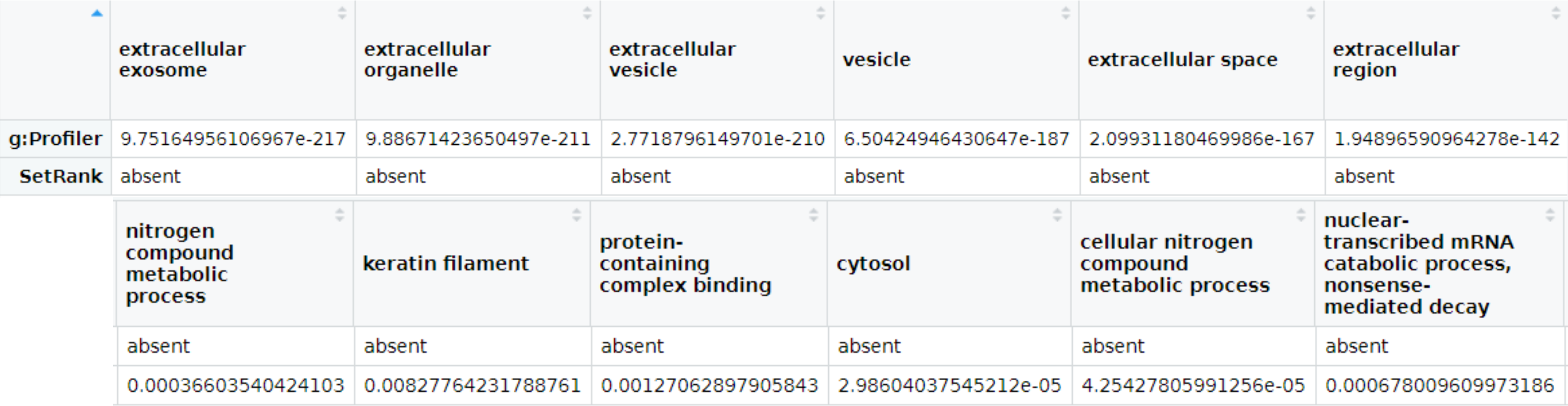
This first version of the wrapper takes, as arguments, a ranking metric (which is the p-value for unranked analysis, which is used for the cutoff in the creation of the foreground), whether the analysis is ranked or not, the list of identifiers (in the ENTREZ format, the conversion from uniprot can be done by the wrapper), a desired cutoff for the gene list for unranked analysis, as well as a collection for SetRank.

The wrapper returns a table with identifiers as columns and the tools as rows. The rows contain the relevant p-value for that term returned by the respective tool, or in the case that the term is not present in the output of that tool, „absent“. The results for the tube-system, ranked and unranked, are presented in the tables 3 and 4.



	intermediate filament	intermediate filament cytoskeleton	positive regulation of protein catabolic process	positive regulation of cellular protein catabolic process	positive regulation of catabolic process	regulation of protein catabolic process	regulatio cellular p catabolic
2	0.000224390221667098	0.000575586331093303	0.000940672454193056	0.00246059818164615	0.00415932327447003	0.00620635208620095	0.0067150
3	0.0111501723196313	0.0162955756049282	absent	absent	absent	absent	absent
4	0.0044918487220459	absent	absent	absent	absent	absent	absent

Table 3: excerpt from wrapper output in unranked mode. Notably there is overlap between the different tools.



	extracellular exosome	extracellular organelle	extracellular vesicle	vesicle	extracellular space	extracellular region
g:Profiler	9.75164956106967e-217	9.88671423650497e-211	2.7718796149701e-210	6.50424946430647e-187	2.09931180469986e-167	1.94896590964278e-142
SetRank	absent	absent	absent	absent	absent	absent

nitrogen compound metabolic process	keratin filament	protein-containing complex binding	cytosol	cellular nitrogen compound metabolic process	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
absent	absent	absent	absent	absent	absent
0.00036603540424103	0.00827764231788761	0.00127062897905843	2.98604037545212e-05	4.25427805991256e-05	0.000678009609973186

Table 3: Two excerpt from ranked wrapper output. Notably there is little overlap, probably due to the fact that the verbose g:Profiler output had to be trimmed to accommodate the format.

Conclusions

The aim of this project was to try different tools capable of ORA and GSEA and combine them in a wrapper, and this has been sufficiently achieved. Further work would be needed to include the PANTHER algorithm and allow fgsea to perform GO enrichment analysis.

It would seem that there are quite a few significant terms found when investigating the carrier vs tube-system dataset, but less so with the ANOVA for the carrier. This indicates that there is less variability between the samples from the human carrier, while the difference between carrier and tube-system is more significant. The results for the fgsea and SetRank ranked analysis are included as an appendix.

To first address the results during the ranked analysis before the creation of the wrapper: considering that the samples are from platelet-free plasma, there was a surprising number of terms associated with intracellular processes such as „epigenetic regulation of gene expression“ and „gene silencing by RNA“, as well as terms associated with the structural integrity of the cell, like „keratinization“. The most significant pathway, „B-WICH regulates rRNA expression“, is also a pathway related to chromatin remodelling, and certainly not something one would expect to find in platelet free plasma. Especially telling is the reactome pathway „G2/M DNA damage checkpoint“, which directly indicates cellular damage.

The reactome pathways found during the ranked analysis for the ANOVA for the carrier samples were less significant than what was found for the carrier vs tube system dataset, and with an adjusted p-value of 0.01 as the threshold, nothing is found. When the results are examined, there is indeed nothing that has an adjusted p-value of below 0.05 either. The 15 most significant results are included in the appendix.

Furthermore, the appendices 3 and 4 include the SetRank results for the two datasets as well. Looking at these, it would seem that some of the pathways found are shared between the datasets. More specifically, pathways associated with the innate immune system, like „Neutrophil degranulation“ and „platalet activation“ are shared, and can thus probably be ignored when analysing the differences in the datasets. What is not shown due to problems with the formatting, is that the p-values for the detected terms are orders of magnitude lower in the carrier vs tube system dataset. This might mean that the detected pathways are that much more impacted in this dataset.

In tables 3 and 4, part of the output for the wrapper is shown to give the reader a general idea of what it looks like. This data was produced with the carrier vs tube system dataset. Since it was in this case not possible to include PANTHER (since the algorithm couldn’t yet be reproduced in R) and fgsea (due to incomplete conversion of the UniProt annotation for all human proteins), only SetRank and g:Profiler were used in the ranked analysis. This is a pity, because during the preliminary analysis of the packages, SetRank and g:Profiler were possibly the most different. G:Profiler usually returns a very large number of hits during GSEA, while SetRank is one of, if not the most stringent of the tools tested in this project. I made the decision to limit the output of g:Profiler in the wrapper, because having a very large list from one tool doesn’t make sense for the construction of a readable table.However, this is left to the user in the form of the argument max_nr, which indicates how many of the most significant results will be considered. Should fgsea and PANTHER be eventually incorporated, it would definitely be necessary to have the output for fgsea be affected by this parameter as well, since it also really suffered from returning a huge number of results.

What is shown in table 4 precisely demonstrates the problem with g:Profiler. The top results are related terms, but were picked up by the tool over and over again. By restricting the output to 20 results, there is no overlap between SetRank and g:Profiler. Thus, g:Profiler might not be suitable for the ranked analysis of the wrapper. However, as previously stated, that is an aspect of this project that needs more work.

Looking at the data, a part of which is displayed in table 3, there is yet again overlap in terms like „Keratinization“ and „Intermediate filament“ (the latter of which is something all 3 agree on), this provides further support in the favor of there being a significant difference in the choice of transportation for samples, as the tube system seems to potentially cause damage to the samples.

In conclusion, this project aimed to create the framework for a wrapper of multiple useful tools for overrepresentation analysis and gene set enrichment analysis. Due to a lack of time, it was not possible to include two of the more useful tools, fgsea and PANTHER, in the framework.

Over the course of this project, a number of datasets were analysed. Even though the wrapper isn’t as useful as I had hoped yet, the dataset for the expression change between the samples from the pneumatic tube system and carrier as well as the ANOVA for the samples from the carrier had some light shed on them, especially during the ranked analysis. It would seem that, while few terms were found to be significant for the dataset for the sample carrier, many were found to be significant for the carrier vs pneumatic tube system -dataset. Interestingly, a lot of these had to do with intracellular processes, structural proteins of the cell as well as even DNA damage. A similar effect was shown by the overrepresentation analysis of the wrapper, where intermediate filaments, keratinization and mitochondrial membrane organization are something the tools agree on. This means that there was little evidence of the variability between carrier samples to be the reason for the observed changes, and that there almost certainly is a difference between the modes of transportation.

References

1: Jing Shi, Michael G Walker: Gene Set Enrichment Analysis (GSEA) for Interpreting Gene Expression Profiles, (May 2007, Current Bioinformatics)

2: Ludwig Geistlinger, Gergely Csaba, Mara Santarelli, Marcel Ramos, Lucas Schiffer, Nitesh Turaga, Charity Law, Sean Davis, Vincent Carey, Martin Morgan, Ralf Zimmer, Levi Waldron: Toward a gold standard for benchmarking gene set enrichment analysis (06 February 2020)

3: Cedric Simillion, Robin Liechti, Heidi E.L. Lischer, Vassilios Ioannidis, and Rémy Bruggmann: Avoiding the pitfalls of gene set enrichment analysis with SetRank (BMC Bioinformatics (2017) 18:151)

4: Huaiyu Mi, Anushya Muruganujan, Xiaosong Huang, Dustin Ebert, Caitlin Mills, Xinyu Guo and Paul D. Thomas1: Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0) (2019)

5: Cristóbal Fresno, Elmer A. Fernández: RDAVIDWebService: a versatile R interface to DAVID (*Bioinformatics*, Volume 29, Issue 21, 1 November 2013)

6: Jüri Reimand, Meelis Kull, Hedi Peterson, Jaanus Hansen and Jaak Vilo: g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments (May 3, 2007)

7: Alexey A. Sergushichev: An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation (June 20, 2016.)

8: Marc Carlson: UniProt.ws: A package for retrieving datafrom the UniProt web service, (April 27 2020)

Appendix 1:

Fgsea results for tube system vs carrier

rank	name	Adjusted p-value
1	B-WICH complex positively regulates rRNA expression	7.775E-09
2	Epigenetic regulation of gene expression	7.775E-09
3	Gene Silencing by RNA	7.775E-09
4	Positive epigenetic regulation of rRNA expression	7.775E-09
5	Pre-NOTCH Expression and Processing	1.45136008111782E-07
6	RUNX1 regulates genes involved in megakaryocyte differentiation and platelet function	1.40271833260836E-06
7	Mitotic Prophase	2.31689169975015E-06
8	Keratinization	8.55059029255378E-06
9	Signaling by Nuclear Receptors	9.6953221949119E-06
10	Estrogen-dependent gene expression	1.76484369848363E-05
11	ESR-mediated signaling	3.94531678382814E-05
12	Formation of the cornified envelope	0.000116434758918
13	G2/M DNA damage checkpoint	0.000441367064971
14	RHO GTPase Effectors	0.002087948952105
15	Signaling by Rho GTPases	0.002311956376735
16	RHO GTPases activate PKNs	0.002311956376735
17	Amyloid fiber formation	0.005858587602562

Appendix 2:

Reactome pathways for sample carrier

rank	name	Adjusted p-value
1	Hemostasis	0.06375817804755
2	Platelet activation signaling and aggregation	0.08024634132444
3	Immune System	0.171713389318271
4	Keratinization	0.171713389318271
5	Glycolysis	0.171713389318271
6	Glucose metabolism	0.171713389318271
7	Cytokine Signaling in Immune system	0.18326491146924
8	Signaling by Interleukins	0.233589334418107
9	Platelet degranulation	0.238550310205155
10	Neutrophil degranulation	0.238550310205155
11	Formation of the cornified envelope	0.238550310205155
12	Response to elevated platelet cytosolic Ca2+	0.238550310205155
13	Gene and protein expression by JAK-STAT signaling after Interleukin-12 stimulation	0.288606538622791
14	Interferon Signaling	0.340733333333333
15	Interleukin-12 family signaling	0.340733333333333

Appendix 3: SetRank for carrier vs tube system

description
Hemostasis
Integrin cell surface interactions
Platelet degranulation
Regulation of Complement cascade
Platelet Aggregation (Plug Formation)
Cell surface interactions at the vascular wall
Neutrophil degranulation
Classical antibody-mediated complement activation
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell
Regulation of TLR by endogenous ligand
Cell junction organization
Cell-extracellular matrix interactions
Platelet activation
FCGR activation
Role of phospholipids in phagocytosis
Regulation of cytoskeletal remodeling and cell spreading by IPP complex components
Fibronectin matrix formation
Innate Immune System
Diseases associated with glycosaminoglycan metabolism

Type I hemidesmosome assembly

Other semaphorin interactions

Activation of C3 and C5

Platelet Adhesion to exposed collagen

Axon guidance

Scavenging of heme from plasma

Chk1/Chk2(Cds1) mediated inactivation of Cyclin B:Cdk1 complex

VLDL biosynthesis

Amyloid fiber formation

G2/M DNA damage checkpoint

Retinoid metabolism and transport

Meiotic synapsis

HDL-mediated lipid transport

Lipoprotein metabolism

Keratinization

Smooth Muscle Contraction

Formation of tubulin folding intermediates by CCT/TriC

Bicarbonate transporters

Detoxification of Reactive Oxygen Species

Transport of glucose and other sugars

Facilitative Na+-independent glucose transporters

SLC-mediated transmembrane transport

Cross-presentation of particulate exogenous antigens (phagosomes)

Chemokine receptors bind chemokines

Glycolysis

Appendix 4: SetRank results for ANOVA of carrier samples

Platelet activation

Neutrophil degranulation

Interleukin-4 and 13 signaling

Innate Immune System

Metal sequestration by antimicrobial proteins

Regulation of Complement cascade

Cell surface interactions at the vascular wall

RHO GTPases Activate WASPs and WAVEs

Antimicrobial peptides

Glycolysis

Gluconeogenesis

L1CAM interactions

Cross-presentation of particulate exogenous antigens (phagosomes)

Diseases of Immune System

Nef-mediates down modulation of cell surface receptors by recruiting them to clathrin adapters

HSF1 activation

Smooth Muscle Contraction

Purine catabolism

ZBP1(DAI) mediated induction of type I IFNs

Biological oxidations

Folding of actin by CCT/TriC

Chk1/Chk2(Cds1) mediated inactivation of Cyclin B:Cdk1 complex

Insulin effects increased synthesis of Xylulose-5-Phosphate

Appendix 5: PANTHER output for carrier vs tube system

HATs acetylate histones (R-HSA-3214847)

Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Prot
HSA-381426)
HCMV Early Events (R-HSA-9609690)

HCMV Infection (R-HSA-9609646)

Post-translational protein phosphorylation (R-HSA-8957275)

Chromatin modifying enzymes (R-HSA-3247509)

Chromatin organization (R-HSA-4839726)

RUNX1 regulates genes involved in megakaryocyte differentiation and platelet function (R-HSA-8936459)

TCF dependent signaling in response to WNT (R-HSA-201681)

IKK complex recruitment mediated by RIP1 (R-HSA-937041)

PRC2 methylates histones and DNA (R-HSA-212300)

RNA Polymerase I Promoter Escape (R-HSA-73772)

Formation of the beta-catenin:TCF transactivating complex (R-HSA-201722)

RNA Polymerase I Promoter Opening (R-HSA-73728)

DNA methylation (R-HSA-5334118)

Meiosis (R-HSA-1500620)

Meiotic synapsis (R-HSA-1221632)

Condensation of Prophase Chromosomes (R-HSA-2299718)

Negative epigenetic regulation of rRNA expression (R-HSA-5250941)

Deposition of new CENPA-containing nucleosomes at the centromere (R-HSA-606279)

Base-Excision Repair

Depyrimidination (R-HSA-73928)

Depurination (R-HSA-73927)

Nucleosome assembly (R-HSA-774815)

NoRC negatively regulates rRNA expression (R-HSA-427413)

HDACs deacetylate histones (R-HSA-3214815)

Meiotic recombination (R-HSA-912446)

Activated PKN1 stimulates transcription of AR (androgen receptor) regulated genes KLK2 and KLK3 (R-HSA-5625886)
HCMV Late Events (R-HSA-9610379)

Packaging Of Telomere Ends (R-HSA-171306)

Cleavage of the damaged purine (R-HSA-110331)

Recognition and association of DNA glycosylase with site containing an affected purine (R-HSA-110330)
Cleavage of the damaged pyrimidine (R-HSA-110329)

Recognition and association of DNA glycosylase with site containing an affected pyrimidine (R-HSA-110328)
RNA Polymerase I Transcription (R-HSA-73864)

RNA Polymerase I Promoter Clearance (R-HSA-73854)

ERCC6 (CSB) and EHMT2 (G9a) positively regulate rRNA expression (R-HSA-427389)

SIRT1 negatively regulates rRNA expression (R-HSA-427359)

RHO GTPases activate PKNs (R-HSA-5625740)

Cell Cycle Checkpoints (R-HSA-69620)