Data Analysis of the Gene Expression Profiles of IFN - β and IFN - γ - Conditioned Human Macrophages

Plazzotta Giovanni, MAT. 232312 **12/6/2023**

Abstract

The paper examines the effect of Interferon gamma (IFN - γ) and Interferon beta (IFN - β) on the expression profiles of non - activated (naive) human macrophages, with the goal of highlighting the cellular processes impacted by each of these chemical mediators. After data normalization and filtering, the presence of latent structures or patterns in the data was investigated with a set of unsupervised learning methods, which failed to discriminate between some of the groups considered. After a feature selection step, the ability of a set of supervised classification methods to correctly distinguish between naive and conditioned samples was tested: all the techniques considered showed 100% accuracy. Finally, functional and network - based analyses were performed: a significant overlap was detected between the cellular processes impacted by each interferon.

1 Introduction

Macrophages are white blood cells that, together with granulocytes, mast cells, natural killer cells and dendritic cells, constitute the subset of leukocytes responsible for the innate immune system. Tissue - resident macrophages serve multiple roles. First of all, these are able to detect the presence of necrotic cells fragments and foreign microorganism by means of their surface pattern recognition receptors (PRRs). The detected entities are then phagocyted, their protein content fragmented and exposed on class II MHC surface receptors. PRRs are also responsible for the initialization of a signal cascade that culminates in the induction of cytokine - genes transcription: the chemical mediators produced this way are able to activate the endothelium and recruit additional leukocytes, amplifying the pro - inflammatory response. However, in non - activated macrophages, these receptors often fails to initialize a signalling cascade: consequently, most of the phagocytosis that occurs on a daily basis happens without the intervention of additional immune cells.

As members of the Antigen - Presenting Cell (APC) family, antigen - carrying macrophages are able to interact with the mature lymphocytic population, triggering their activation and differentiation towards the T $CD4^+$ helper phenotype. However, Dendritic Cells are often recognized as the primary activators of the naive T population, given that macrophages are located mainly in non - lymphoid tissues [1].

In addition to producing chemical signals able to alert the immune system of the presence of foreign microorganism or local cellular debris, macrophages themselves can respond to the endogenous stimuli generated by the cellular component of the innate immunity as well as by activated lymphocytes, stimuli that exert a marked effect on the physiology of macrophages. To complicate the matter, macrophages themselves can produce several factors that influence their own physiology. While these chemical signals are known to induce a spectrum of different phenotypes in this population, the literature usually distinguishes between *classically* and *alternatively* activated macrophages as the two extrema of this range [2].

The classically activated phenotype, also labelled "M1", is the result of the combination of the following signals, namely interferon γ (IFN - γ), mainly produced by activated T lymphocytes and Natural Killer cells, tumor necrosis factor (TNF), secreted by APCs, and pathogen - associated lipopolysaccharides (LPS). It is characterized by an enhanced microbicidal and tumoricidal capacity and, unlike the non - activated phenotype, is able to secrete high levels of pro - inflammatory cytokines. For these reasons, M1 macrophages are the protagonists of the initiation and up-keeping of the acute inflammation state.

The alternatively activated "M2" phenotype arises instead due to chemical stimulation via interleukin 4 (IL-4), IL-13 and IL-6. Characterized by a pro - resolving behaviour, M2 macrophages secrete anti - inflammatory cytokines such as IL-10 and TGF - β and contribute to tissue repair, remodeling and to the initiation of vasculogenesis [3].

In addition to the aforementioned chemical mediators, interferon β (IFN - β) has also been shown to play a substantial role in regulating the activity of macrophages. However, contrasting evidence exists on the outcome of this interaction: while in some studies the activity of IFN - β resulted in an anti - inflammatory state that is in contrast to the pro - inflammatory role of IFN - γ, TNF and LPS, others have shown how IFN - β enhances antigen presentation and participates in signaling cross - talk with other cytokines, such as the pro - inflammatory TNF [4].

In order to elucidate the role of IFN - β in the context of the phenotypic fine - tuning of human macrophages, it was decided to compare the expression profiles of non activated macrophages with the one displayed by, respectively, IFN - γ and IFN - β stimulated cells, with the goal of highlighting eventual overlapping effects on the trascriptome. The latter could suggest a similar functional outcome for the macrophages, after exposure to these chemical mediators. On the other hand, failure to identify any similarity in the pathways affected by the two signals would mean that their effect on the macrophages' phenotype is not expected to be comparable: in this situation, the analysis of the impacted cellular processes could shed some light on the specific effect of IFN - β.

2 Data inspection and preprocessing

The dataset under analysis was provided by the department of Microbiology, Immunology, and Molecular Genetics of the University of California, and contains the RNA sequencing data relative to 63856 transcripts across 133 subject. The macrophages of each patient were isolated form peripheral blood. Of the resulting 133 samples, 59 were not conditioned ("naive" group), 33 were conditioned with IFN - β and 41

with IFN - γ.

After converting the raw transcript coverage counts into read counts and labelling each sample with its own group of membership, a filtering step was considered: in particular, all the rows such that more than half the entries were zero were removed. This resulted in a final dataset comprised of 24574 rows and 133 columns. After a *log* transformation, a boxplot of each sample was produced, that revealed how the data were in clear need of normalization. Normalization was thus performed: the Reads per Kilobase (RPK) were computed for each transcript, exploiting the transcript length information contained in the metadata, and fed to the `EdgeR` package, which yield the normalized counts. After a second *log* transformation another set of boxplots was produced, which showed normalized data (Figure 1).

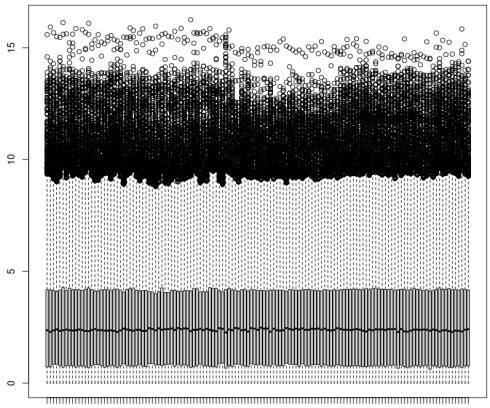


Figure 1: distribution of the gene expression profiles after the normalization step.

3 Exploratory data analysis

In order to investigate the presence of latent structures or patterns in the data, three different unsupervised method were considered: Principal Component Analysis (PCA), K - Means and Hierarchical Clustering.

3.1 Principal Component Analysis

PCA was performed on the whole dataset. The first two principal components explained respectively 26% and 14% of the total variance (Figure 2), with the first three reaching a cumulative value of 50%.

The plot obtained by projecting the 133 samples in the two dimensional space identified by the first two principal components showed two main clusters of datapoints (Figure 3). By colouring the latter according to the labels provided in the metadata, it was noticed that the green cluster on the right was comprised by all and only the samples conditioned with IFN - γ, while the cluster on the left showed a mixed composition of naive (in red) and IFN - β - exposed samples (in blue).

The PCA seems to suggest that, at least according to the first two principal components, the expression profiles of the samples treated with IFN - γ are remarkably different than the ones that characterize naive and IFN - β - conditioned samples. Additionally, these latter seems to be similar enough to at least partially overlap in the 2D PCA plot. Obviously, additional analyses are required to draw any significant conclusion.

3.2 K - Means

A first clustering instance was performed with the K - Means method, by specifying $K = 3$. To allow for the visualization

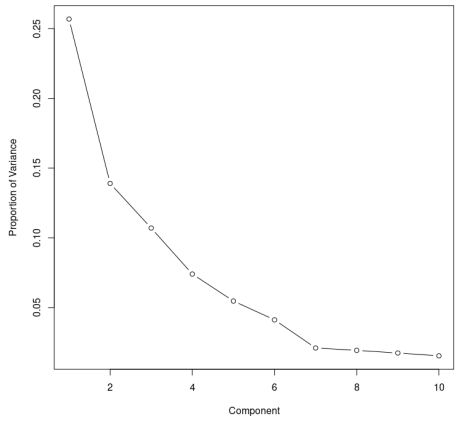


Figure 2: scree plot of the proportion of the total variance explained by each component.

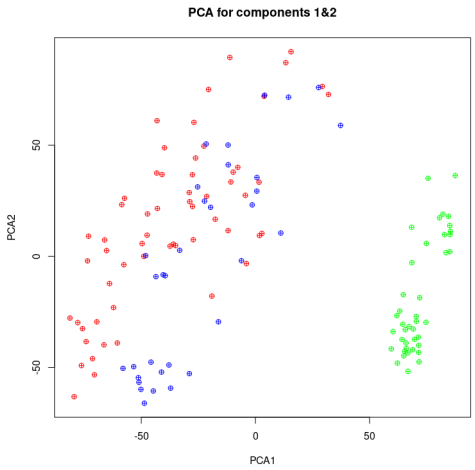


Figure 3: projection of the 133 datapoints on the space defined by the first two principal components.

of the identified clusters, a second PCA was performed: the results of the projection of the samples, coloured by cluster membership, into the space identified by the first two principal components is reported in Figure 4. It is evident that, while K - Means is able to correctly identify the cluster comprised of IFN - γ - conditioned samples, it struggles to deal with the other two groups. By comparing the clusters identified by K - Means with the PCA plot reported in Figure 3, one can seen how the first hallucinates a boundary between the naive and the IFN - β - treated samples, boundary that is not so clear nor positioned in the same location in the original PCA plot.

The extent of agreement between the groups identified by the labels and the ones identified by K - Means was quantified with the Adjusted Rand Index (ARI), which provided a similarity score of $ARI = 0.475$. This denotes an intermediate goodness for the clustering, between perfect agreement ($ARI = 1$) and a clustering equivalent to the one generated by a random assignment of labels ($ARI = 0$).

3.3 Hierarchical Clustering

A second clustering instance was performed, this time with the Hierarchical Clustering method. The resulting dendrogram was cut so to generate three clusters. In order to compare the results with the original labels, the terminal nodes

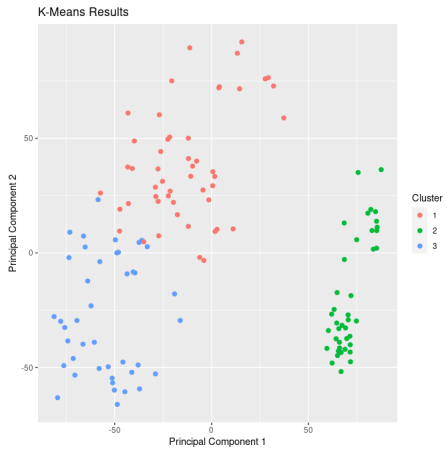


Figure 4: result of the K - Means clustering, projected in the space of the two principal components.

of the dendrogram were renamed according to the samples' metadata, recoded as "1" for the naive samples, "2" for the samples exposed to IFN - β, and "3" for the ones conditioned with IFN - γ (Figure 5). By inspecting the terminal nodes, it is possible to conclude that the Hierarchical Clustering was able to group all and only the samples exposed to IFN - γ in the cluster represented by the top red box. However, similar separation was not achieved for the other two groups: this is demonstrated by the mixed composition of the terminal node's names displayed by the red box in the center and at the bottom. These results are in line with what already demonstrated by K - Means.

The similarity between the grouping identified by Hierarchical Clustering and the one determined by the labels was again quantified using the ARI, which provided a score of $ARI = 0.485$, similar to the one obtained with K - Means.

4 Feature selection

Ahead of the subsequent analyses, the original data was split into two datasets: the first was comprised of the expression profiles of the naive and IFN - β - conditioned samples, and was referred to as the "naive - IFN - β" set. The second was built from the transcriptomic data of the naive and IFN - γ - exposed samples, and was named the "naive - IFN - γ" set. A row - wise t - test was performed on both of these, in order to select the transcripts whose expression levels showed a significant variability across samples belonging to the different treatment groups. After rescaling the $p - value$ with the Holm correction (this was deemed necessary given the large number of comparisons), different $p - value$ thresholds were considered, with the goal of maximizing the significance of the selected features, while still retaining a number of them large enough not to hinder the activity of the subsequent statistical methods. A threshold of 10^{-9} was eventually chosen. This process led to the construction of two datasets comprised, respectively, of 1049 rows and 92 columns and of 3404 rows and 100 columns.

5 Classification methods

To investigate the capability of statistical learning methods to correctly discriminate between naive and IFN - β and between naive and IFN - γ conditioned samples, four supervised classification techniques, namely Random Forest, Linear Discriminant Analysis, LASSO and SCUDO, were tested.

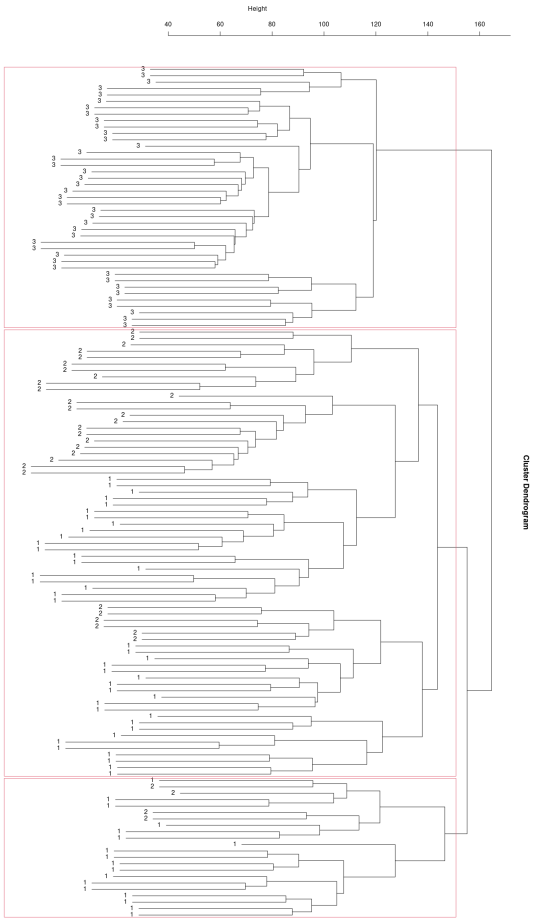


Figure 5: result of the Hierarchical Clustering. The red boxes represents the three clusters generated by cutting the dendrogram.

5.1 Random Forest

A random forest model was fit to both naive - IFN - β and naive - IFN - γ datasets. The parameter `ntree` was set to 50 after observing that, in both cases, the classification error dropped to 0 with values of `ntree` larger than 30. The plot reporting the classification error in function of this parameters is shown below, in Figure 6. The parameter `mtry`, representing the number of variables randomly sampled as candidates at each split, was instead left to the default value of $\sqrt{p}$, where p is the number of features in the specific dataset.

In order to accurately estimate the performance of the two models, a repeated 10 fold cross validation (repeated CV) was then performed. Given that each cross validation instance produces a cross validation error as the mean of the classifications errors computed on each fold, repeating the CV multiple times allows to collect a series of performance estimates that can be used to compute and plot some statistics about the CV errors, such as mean and standard deviation. The repeated cross validation was performed with $K = 10$ and $r = 10$, where K is the number of folds and r is the number of repetitions. Given that each CV instance loops across different values of the sole `mtry` parameter, the number of trees was set to the aforementioned value of 50. In both cases, the repeated CV showed a 100% accuracy across every fold of each repetition.

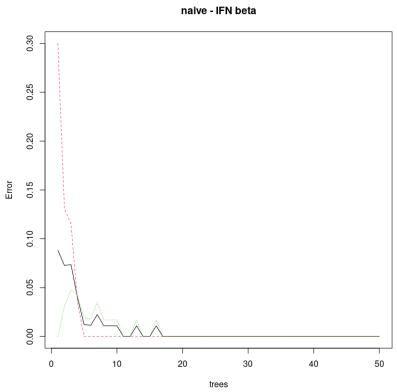


Figure 6: plot of the error rate in function of the number of trees considered when fitting the random forest on the naive - IFN - β dataset. The same plot, relative to the naive - IFN - γ dataset, is reported in the Additional Files section.

5.2 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) was then considered. A LDA model was fit to both naive - IFN - β and naive - IFN - γ datasets. The projections of the datapoints belonging to each of the two groups on the direction perpendicular to the vector that maximizes the between class distances while minimizing the within class ones were reported in Figure 7.

Similarly to what was performed with Random Forest, the current method was benchmarked in a repeated cross validation fashion, which reported 100% accuracy across every fold of each repetition.

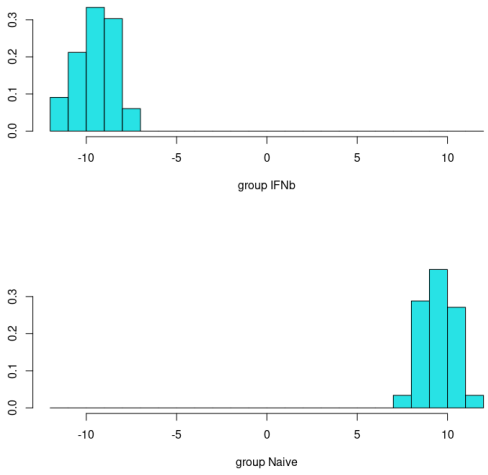


Figure 7: projections of the datapoints of the naive - IFN - β dataset along the direction of optimal class separability. The same plot, relative to the naive - IFN - γ dataset, shows the same optimal separation and is attached in the Additional Files section.

5.3 LASSO

A logistic regression model equipped with LASSO regularization was then considered. In principle, LASSO models do not require a prior feature selection step, given that they themselves are able to identify the most informative covariates. However, some implementations of this constrained optimization algorithm will rise a protection stack overflow error if challenged with a high dimensional dataframe. For this reason, it was decided to fed the LASSO algorithm with the naive - IFN - β and naive - IFN - γ datasets resulting from the aforementioned feature selection step. The model is characterized by two parameters: the first, α, quantifies the blend between LASSO and Ridge regression on the underlying generalized regression model. In this case, α was set to 0 to ensure a purely LASSO - based logistic regression. The second parameter, λ, quantifies the weight to be put on the constrained section of the LASSO optimization problem, and was selected, for both datasets, via 10 fold cross validation (Figure 8), using the default search grid.

The constrained logistic regression model was then evaluated via repeated CV, which reported, once again, a 100% accuracy score across every fold of each repetition.

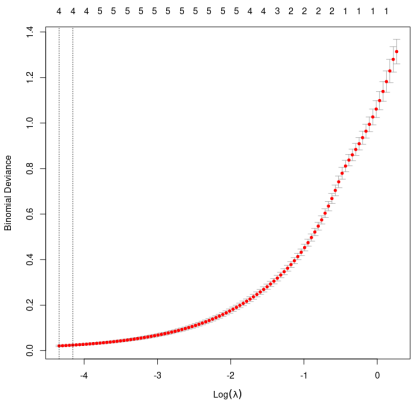


Figure 8: plot of the cross validation errors (binomial deviances) in function of different values of the parameter λ, in the context of the naive - IFN - β dataset. The best value of λ was found as $\lambda = 0.01303247$. The same plot, for the naive - IFN - γ dataset, showed a similar behaviour and is attached in the Additional Files section.

5.4 SCUDO

A rank based strategy was then explored, given the robustness displayed by these classification algorithms when challenged with data sourced from different laboratories. The following steps were performed in the context of both datasets: first, a 10 fold CV was exploited to select the number of features to retain form the top and the bottom of the sorted list of transcript IDs, across all samples. In both evaluations, the best value for the `nTop` and `nBottom` parameters was found at `nTop` $=$ `nBottom` $= 250$. After that, a SCUDO model equipped with this optimal parameter set was fit to each dataset, and used to build graphs in which the nodes are the signatures identified by the method, while the edges represent the pairwise distance between these. Only the edges relative to the smallest 50% distances were retained. Smaller values of this threshold led, in both cases, to unconnected components. Clustering was then performed on the resulting graphs, using a spin - glass model. The results of the SCUDO classification can be appreciated in Figure 9 and Figure 10.

While the separation between the two classes seems clear, at least in the context of the naive - IFN - β dataset, from this qualitative overview it is impossible to tell whether all the samples were actually classified in their respective group. In order to obtain a quantitative estimate of the performance of the SCUDO method, it was decided to proceed with a repeated cross validation. Here too, an accuracy score of 100% was recorded across every fold of every repetition.

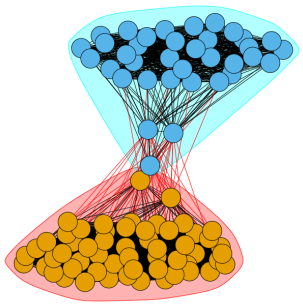


Figure 9: communities identified by the spin - glass clustering method on the graph build by SCUDO, in the context of the naive - IFN - β dataset.

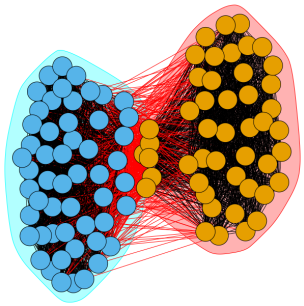


Figure 10: communities identified by the spin - glass algorithm on the SCUDO graph built from the naive - IFN - γ dataset.

5.5 Comparison of the classification methods

As stated above, repeating the 10 fold cross validation multiple times allow to gather data that can then be used to compute some statistics for the cross validation errors. Having performed the repeated CV for all four methods considered, one can find the best performing one by comparing these statistics: Figure 11 reports mean and standard deviation for the CV errors across all techniques considered. It can be concluded that all the methods performed equally perfectly on these datasets.

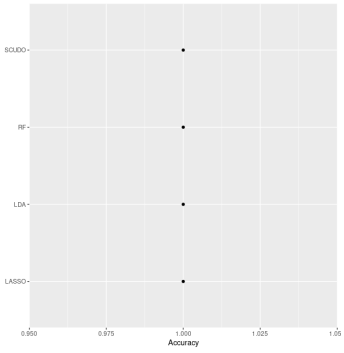


Figure 11: distribution of the CV errors across the four classification methods considered, for the naive - IFN - β dataset. The same behaviour was observed by testing these algorithms on the naive - IFN - γ dataset.

6 Functional Analysis

The latter steps of the current analysis were dedicated to uncovering the biological role of the most significant transcripts in the two databases. It was decided not to rely on the t - test's $p-values$ to detect these most important features, given that, firstly, the t - test makes assumptions on the distribution of the data, and secondly, given the large numbers of features, the resulting $p - values$ are not always trustworthy, even when adjusted. Instead, it was decided to use, as a measure of significance, the `importance` score assigned by random forest to each of the transcripts, a numerical index representing the mean decrease in accuracy of the predictions, across all trees, when each particular predictor is removed from the model. The transcripts of both datasets were then sorted according to the `importance` score. Two plots, reporting the `importance` values as a function of the indexes, were generated and attached in Figure 12.

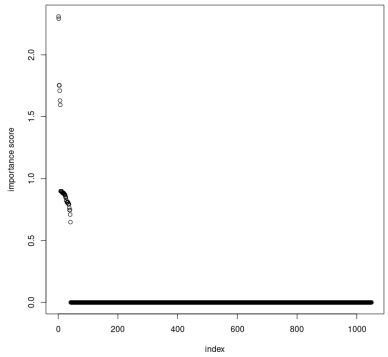
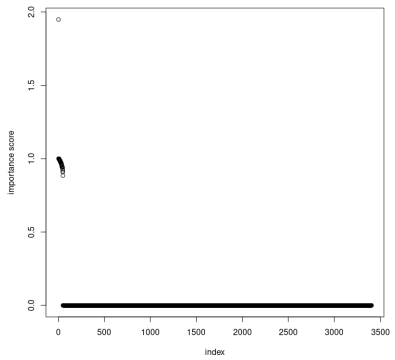


Figure 12: importance score in function of the transcript index, for the naive - IFN - β dataset (top) and for the naive - IFN - γ dataset (bottom).



After experimenting with different thresholds, it was decided to perform the subsequent analyses with the first 400 most important genes of each dataset, whose IDs were extracted form the importance plots just shown.

6.1 DAVID

In the tables below, the top 10 results provided by the DAVID tool for functional analysis were reported. From the table relative to the naive - IFN - β dataset (Figure 13), it can be appreciated how terms referring to the mitotic phase of the cell cycle, to the mitotic spindle organization and to the ribosome biogenesis were the ones associated with lower $p-value$ (in this case, computed with the Benjamini correction) for the Fisher test.

The DAVID results for the naive - IFN - γ dataset (Figure 14) were even richer in terms referring to specific events of the mitotic process, such as the assembly of the mitotic spindle, the amplification of the anti - mitotic signals coming from unconnected kinetochores and the separation of sister chromatids.

	Term	Benjamini
	<chr>	<chr>
1	GO:0051301~cell division	2.68195460809226E-15
2	R-HSA-1640170~Cell Cycle	5.95155759656958E-10
3	GO:0003735~structural constituent of ribosome	4.68853123582137E-08
4	R-HSA-69278~Cell Cycle, Mitotic	3.77610018965073E-08
5	R-HSA-156842~Eukaryotic Translation Elongation	3.77610018965073E-08
6	GO:0002181~cytoplasmic translation	1.01298566040949E-07
7	GO:0007052~mitotic spindle organization	1.01298566040949E-07
8	GO:0022626~cytosolic ribosome	9.13684820283354E-08
9	R-HSA-72689~Formation of a pool of free 40S subunits	7.12564092285052E-08
10	R-HSA-9633012~Response of EIF2AK4 (GCN2) to amino acid deficiency	7.12564092285052E-08

Figure 13: top 10 DAVID results in the context of the naive - IFN - β dataset, sorted for Benjamini score.

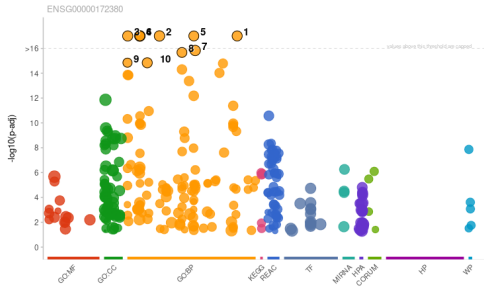
	Term	Benjamini
	<chr>	<chr>
1	GO:0070062~extracellular exosome	2.50899716604722E-07
2	R-HSA-141424~Amplification of signal from the kinetochores	3.45936355484844E-06
3	R-HSA-5663220~RHO GTPases Activate Formins	4.21457970118375E-06
4	R-HSA-69618~Mitotic Spindle Checkpoint	1.27733187107837E-05
5	R-HSA-9648025~EML4 and NUDC in mitotic spindle formation	1.55233202186765E-05
6	GO:0007052~mitotic spindle organization	0.000170169153722514
7	R-HSA-195258~RHO GTPase Effectors	2.46776362656459E-05
8	R-HSA-2500257~Resolution of Sister Chromatid Cohesion	2.67478595347843E-05
9	GO:0051301~cell division	0.00227919327727089
10	GO:0005515~protein binding	0.00285504949602747

Figure 14: top 10 DAVID results in the context of the naive - IFN - γ dataset, sorted for Benjamini score.

6.2 gProfiler

The gProfiler tool was then considered. The results of the functional analysis can be appreciated in the plots below: in the context of the naive - IFN - β dataset (Figure 15), the most significant terms detected referred to the mitotic phase of the cell cycle. A focus was also put on the separation of sister chromatids.

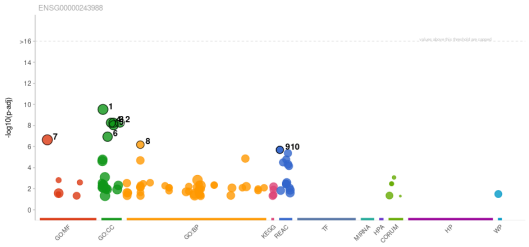
For what concerns the second dataset, the smallest $p - values$ were associated with terms referring to the cytoplasm and to different extracellular locations. More interestingly, some of the mitotic - related processes listed by DAVID, such as the mitotic spindle organization and the amplification of signal from the kinetochores, were also present in this list, fact that highlights their significance.



id	source	term_id	term_name	term_size	p_value
1	GO:BP	GO:1903047	mitotic cell cycle process	766	5.6e-23
2	GO:BP	GO:0022402	cell cycle process	1281	1.1e-22
3	GO:BP	GO:0000278	mitotic cell cycle	922	2.0e-22
4	GO:BP	GO:0007049	cell cycle	1825	5.5e-20
5	GO:BP	GO:0051301	cell division	661	7.7e-19
6	GO:BP	GO:0007059	chromosome segregation	423	6.3e-18
7	GO:BP	GO:0051726	regulation of cell cycle	1130	1.5e-16
8	GO:BP	GO:0044770	cell cycle phase transition	560	2.1e-16
9	GO:BP	GO:0000070	mitotic sister chromatid segregation	190	1.4e-15
10	GO:BP	GO:0010564	regulation of cell cycle process	726	1.4e-15

g:Profiler (bit.cs.ut.ee/gprofiler)

Figure 15: gProfiler result for the naive - IFN - β dataset (in this column) and for the naive - IFN - γ one (in the next column) ranked by $p - values$ and distributed according to the source of the terms.



id	source	term_id	term_name	term_size	p_value
1	GO:CC	GO:0005737	cytoplasm	12365	3.0e-10
2	GO:CC	GO:1903561	extracellular vesicle	2133	5.4e-09
3	GO:CC	GO:0065010	extracellular membrane-bounded organelle	2134	5.5e-09
4	GO:CC	GO:0043230	extracellular organelle	2134	5.5e-09
5	GO:CC	GO:0070062	extracellular exosome	2109	9.6e-09
6	GO:CC	GO:0031982	vesicle	3998	1.2e-07
7	GO:MF	GO:0005515	protein binding	14881	2.3e-07
8	GO:BP	GO:0007052	mitotic spindle organization	136	6.6e-07
9	REAC	REAC:R-HSA-141424	Amplification of signal from the kinetochores	92	2.0e-06
10	REAC	REAC:R-HSA-141444	Amplification of signal from unattached kinetochores via a MAD2 inhibitory signal	92	2.0e-06

g:Profiler (bit.cs.ut.ee/gprofiler)

7 Network - Based Analysis

Network - based analysis capitalizes on the information regarding the closest interactors of the selected genes to highlight pathways and processes characterized by a significantly high density of these latter. In the current work, three different network - based methods, namely PathfindR, STRING and EnrichNet, were considered. These were each given as input the 400 genes (per dataset) used for the previously performed functional analyses.

7.1 PathfindR

The results of the network based analysis performed with PathfindR were only partially in line with the ones derived form the previous methods. While, for the naive - IFN - β gene set, the impact of the genes under analysis on the ribosome biogenesis and on the cell cycle was once again confirmed, other processes, such as the activation of the complement cascade and the pro - inflammatory antiviral response, were also reported as significant. This can be seen as a consequence of the gene enrichment performed by PathfindR.

In the context of the naive - IFN - γ gene set, on the other hand, the $p - values$ of the reported terms were in general many order of magnitude higher than the ones reported for the terms extracted after enriching the first gene set. This behaviour can also be observed in the results of DAVID and gProfiler. The most significant terms were, also in this case, connected to the ribosome biogenesis and the activation of the antiviral response. An intuitive representation of the most statistical relevant terms found by PathfindR is presented below, in Figure 16. Given the non - deterministic nature of this tool, it was chosen to report as output the average result of 5 iterations of the latter.

7.2 STRING

The result of the network based analysis performed with the STRING web tool are reported below. Compared to PathfindR's, the top 10 STRING results (according to the false discovery rate statistics) were more in line with the ones reported by the previously explored methods. In the context of the naive - IFN - β set, the most significant terms were associated with the mitotic cell cycle and with the segregation of sister chromatids.

For what concerns the naive - IFN - γ set, STRING detected as statistically overrepresented multiple terms referring to the mitotic spindle checkpoint and assembly, and to the amplification of the signal from the kinetochores. All of these results are, in broad outline, in agreement with what previously reported by DAVID and gProfile. The tables of the 10 most significative terms for the two datasets are reported below, in Figure 17 and Figure 18.

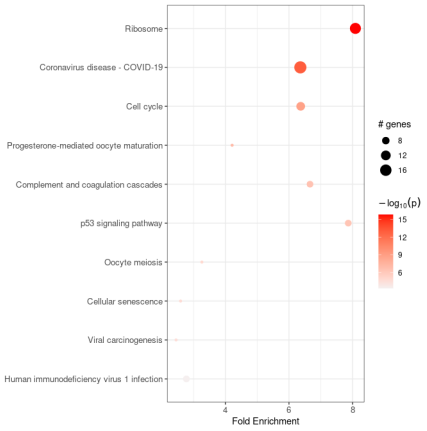
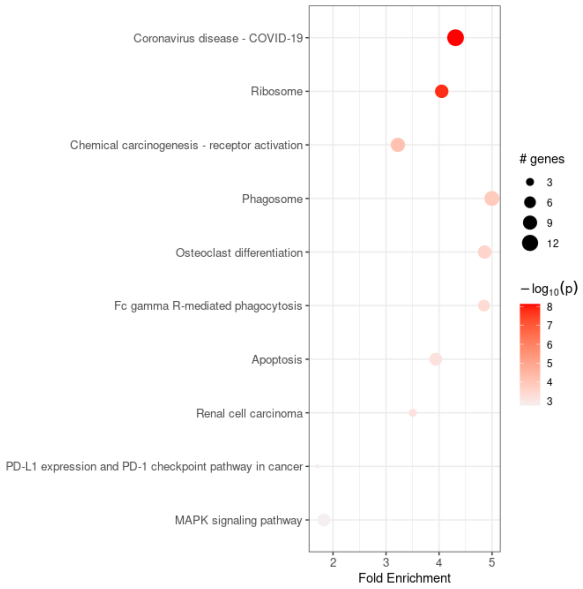


Figure 16: most significant terms according to PathfindR's $p-value$. The size of each circle is proportional to the number of genes in the enriched set found as involved in each particular process or pathway, while the colour encodes the significance in term of $p-value$. The top image refers to the naive - IFN - β gene set, while the bottom one refers to the naive - IFN - γ gene set.



7.3 EnrichNet

For what concerns the naive - IFN - β set, the results of the analysis performed with the EnrichNet tool were in accordance with the output of DAVID, gProfiler and STRING, as well as with the conclusions provided by PathfindR: different terms regarding the mitotic cell cycle, the mitotic spindle organization, the segregation of the sister chromatids, as well as terms regarding the biogenesis of the ribosomes and the triggering of the complement system (previously highlighted by PathfindR) were reported as statistically significant.

The same accordance was observed in the context of the second database, where Enrichnet reported as significative terms linked to the biogenesis of the ribosomes and to the mitotic spindle organization. Additional entries, referring to pro - inflammatory responses (respiratory burst, positive regulation of mast cell degranulation, peroxisome organization and acute inflammatory response) were also present. A portion of the EnrichNet output is reported in Figure 19 and Figure 20.

	term.description <chr>	false.discovery.rate <chr>
1	Mitotic cell cycle	3.25e-25
2	Cell cycle	3.25e-25
3	Cell cycle process	3.25e-25
4	Mitotic cell cycle process	3.25e-25
5	Cell division	7.14e-19
6	Mitotic nuclear division	3.76e-16
7	Sister chromatid segregation	1.52e-14
8	Nuclear division	1.90e-14
9	Chromosome segregation	1.90e-14
10	Mitotic sister chromatid segregation	2.73e-14

Figure 17: top 10 STRING results in the context of the naive - IFN - β gene set, sorted for the false discovery rate.

	term.description <chr>	false.discovery.rate <chr>
1	Mitotic Spindle Checkpoint, and mitotic nuclear division	0.00000000761
2	Mitotic Spindle Checkpoint	0.0000000202
3	Amplification of signal from unattached kinetochores	0.00000011
4	RHO GTPases Activate Formins	0.000000132
5	Organelle	0.000000159
6	RHO GTPase Effectors	0.000000298
7	EML4 and NUDC in mitotic spindle formation	0.000000483
8	Mixed, incl. condensed chromosome and centromeric region.	0.000000562
9	Resolution of Sister Chromatid Cohesion	0.00000107
10	Mitotic Anaphase	0.00000277

Figure 18: top 10 STRING results in the context of the naive - IFN - γ gene set, sorted for the false discovery rate.

	Annotation..pathway.process. <chr>	XD.score <dbl>
1	GO:0000085~G2 phase of mitotic cell cycle	2.548356
2	GO:0007091~mitotic metaphase/anaphase transition	2.098356
3	GO:0007052~mitotic spindle organization	2.098356
4	GO:0000070~mitotic sister chromatid segregation	1.966003
5	GO:0007076~mitotic chromosome condensation	1.925279
6	INITIAL TRIGGERING OF COMPLEMENT	1.833561
7	CYCLIN A1 ASSOCIATED EVENTS DURING G2 M TRANSITION	1.833561
8	GO:0042026~protein refolding	1.648356
9	GO:0042273~ribosomal large subunit biogenesis	1.484719
10	GO:0009607~response to biotic stimulus	1.484719

Figure 19: top 10 EnrichNet terms for the naive - IFN - β gene set, sorted for the XD score.

	Annotation..pathway.process. <chr>	XD.score <dbl>
1	GO:0045730~respiratory burst	1.644948
2	GO:0043306~positive regulation of mast cell degranulation	1.644948
3	GO:0007052~mitotic spindle organization	1.532448
4	GO:0042273~ribosomal large subunit biogenesis	1.481311
5	GO:0016050~vesicle organization	1.481311
6	GO:0050673~epithelial cell proliferation	1.481311
7	GO:0005978~glycogen biosynthetic process	1.433183
8	GO:0043039~tRNA aminoacylation	1.344948
9	GO:0007031~peroxisome organization	1.266000
10	GO:0002526~acute inflammatory response	1.229563

Figure 20: top 10 EnrichNet terms for the naive - IFN - γ gene set, sorted for the XD score.

8 Conclusions

The results of the functional and network - based analysis tools were compared. It was observed that the gene set selected form the naive - IFN - β dataset and the one extracted form the naive - IFN - γ dataset were both strongly involved in the regulation of the cell cycle, and in particular in the management of events linked to the mitotic process, such as the assembly of the mitotic spindle, the segregation of sister chromatids and the amplification of the anti - mitotic signals generated by unattached kinetochores. These findings suggest the presence of a significant overlap between the pathways that are impacted by the conditioning of naive macrophages with IFN - β, and the pathways that are affected by the exposure to IFN - γ. The two chemical mediators also seemed to play a role in controlling the ribosome's biogenesis and the modulation of the antiviral response. However, it is important to stress that this overlap does not necessarily imply an equivalence of effect (this is a way stronger statement), for two reasons. The first is that, while the cellular processes involved may be similar, the particular genes affected in the context of these pathways could be different, depending on whether the naive cells were conditioned with IFN - β or γ.

Two examples were chosen to demonstrate this problem: in Figure 21 and Figure 22 one can appreciate the EnrichNet graph outputs for the ribosomal large subunit biogenesis, term that was deemed as significant for both the naive - IFN - β gene set and the naive - IFN - γ one. From the plots it is possible to observe that, in the context of this pathway (genes in red), the genes whose transcription level was impacted by the treatment with IFN - β (in green in Figure 21) and the genes whose activity was modulated following exposure to IFN - γ (in green in Figure 22) were exactly the same. So in this case, the treatment with the two different chemical mediators impacted not only the same pathway, but also the same genes in the context of that pathway.

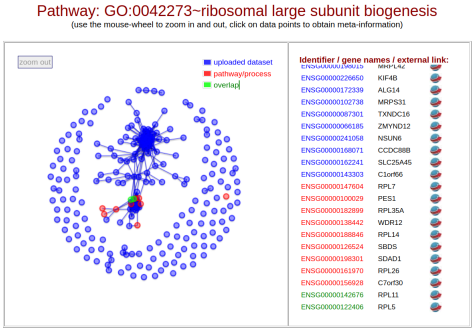


Figure 21: EnrichNet graph output for the ribosomal large subunit biogenesis pathway, in the context of the naive - IFN - β gene set.

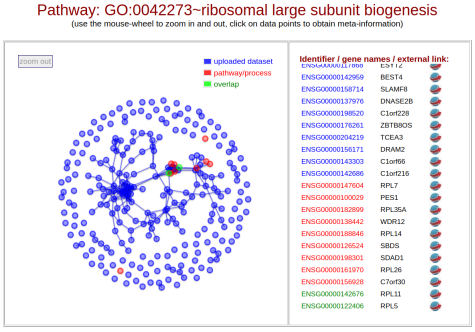


Figure 22: EnrichNet graph output for the ribosomal large subunit biogenesis pathway, in the context of the naive - IFN - γ gene set.

As a counter example, the EnrichNet graph outputs for the mitotic spindle organization was considered. Similarly to what observed for the ribosomal biogenesis pathway, this process was labelled as significant for both the naive - IFN - β gene set and the naive - IFN - γ one. From Figure 23 and Figure 24, one can notice how the overlapped genes are different across the two EnrichNet instances, even if the complete list of red and green terms is exactly the same.

The second reason for which the overlap of the results provided by the different network - based methods does not necessarily imply an equivalence of effect between IFN - β and IFN - γ is that, even in the situation in which the genes at play are found to be exactly the same (such as in the ribosomal large subunit biogenesis pathway), the analyses performed so far do not allow to discriminate whether the expression of these features was down - regulated or up - regulated as a result of the separate conditioning with the two chemical mediators. Additional evaluations are thus necessary to prove the aforementioned stronger statement.

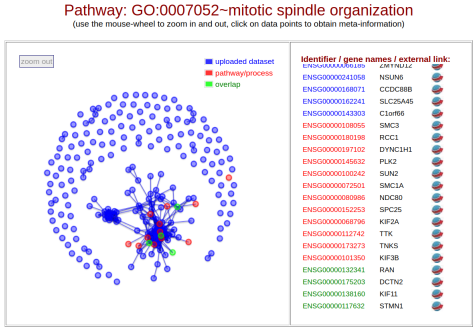


Figure 23: EnrichNet graph output for the mitotic spindle organization process, in the context of the naive - IFN - β gene set.

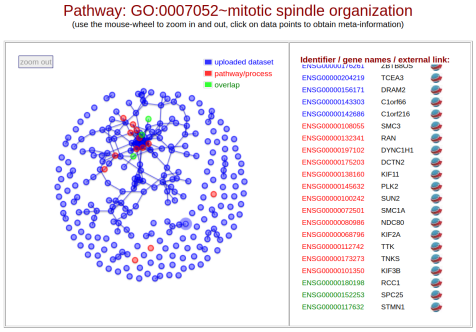


Figure 24: EnrichNet graph output for the mitotic spindle organization process, in the context of the naive - IFN - γ gene set.

9 References

[1] Antigen presentation to naive CD4 T cells in the lymph node, Itano et al., Nature Immunology, 2003

[2] Exploring the full spectrum of macrophage activation, Mosser et al., Nature Reviews Immunology, 2009 Aug 11

[3] Macrophage Polarization and Reprogramming in Acute Inflammation: A Redox Perspective, Perez et al., Antioxidants, 2022 Jul

[4] IFN - β / TNF - α synergism induces a non-canonical STAT2/IRF9-dependent pathway triggering a novel DUOX2 NADPH Oxidase-mediated airway antiviral response, Fink et al., Cell Research, 2013 May

10 Additional Files

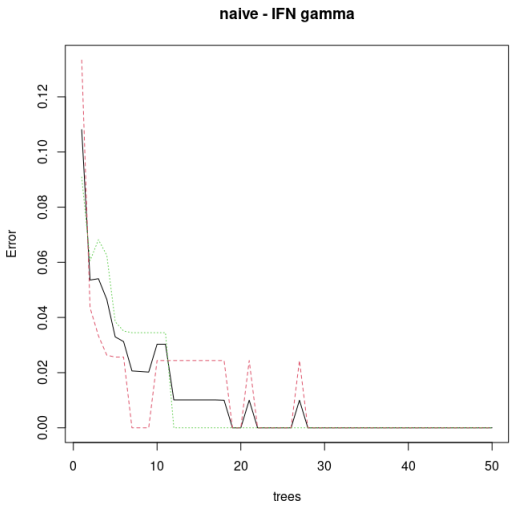


Figure 25: plot of the error rate in function of the number of trees concerning the naive - IFN - γ dataset.

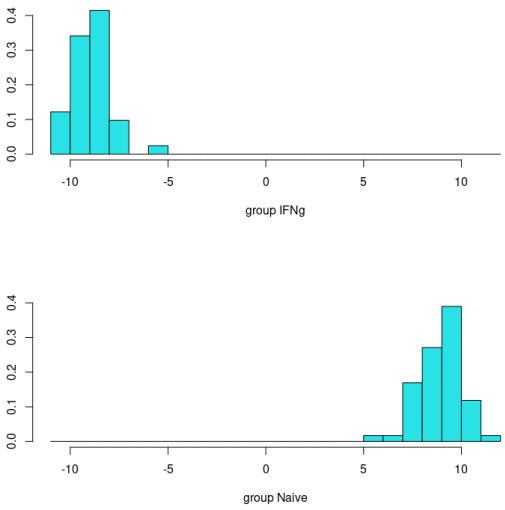


Figure 26: projections of the datapoints of the naive - IFN - γ dataset along the direction of optimal class separability.

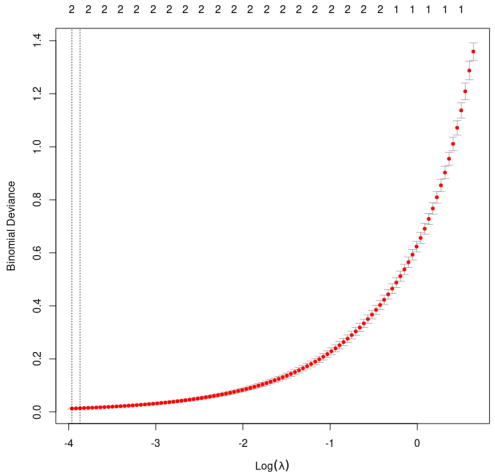


Figure 27: plot of the CV error in function of different values of the parameter λ for the naive - IFN - γ dataset. The optimal value of λ was found at $\lambda = 0.01903352$.