Trabajo Fin de Grado

Gene co-expression network analysis in Mycobacterium tuberculosis

Análisis de redes de coexpresión génica en Mycobacterium tuberculosis

Autor

Héctor García Cebollada

Director

Yamir Moreno Vega

BIFI/Facultad de Ciencias
Año 2016

Contents

1. Abstract 1
2. Background 2
3. Objectives 4
4. Materials and methods 4
4.1. Data retrieval and ordering 4
4.2. Correlation analysis 5
4.3. Statistical significance 6
4.3.1. For gene co-expression 6
4.3.2. For gene differential co-expression 6
4.4. Analysis 7
4.4.1. For gene co-expression 7
4.4.2. Comparison between layers 8
4.4.3. For gene differential co-expression 9
4.5. Data enrichment 10
5. Results 10
5.1. Gene pairs 10
5.1.1. Correlation 11
5.1.2. Statistical significance 12
5.1.3. Relevance of each type of gene 12
5.2. Gene co-expression analysis 13
5.3. Comparison between layers 14
5.4. Differential analysis 15
6. Discussion 16
6.1. Gene pairs 16
6.2. Gene co-expression analysis 18
6.3. Comparison between layers 19
6.4. Differential analysis 21
6.5. Outlook 23
7. Conclusions 24
8. Bibliography 24
Appendix A. Codes of layers of stress and correspondence of subtypes to major types I
Appendix B. Dictionary of samples used in this work. II
Appendix C. Equivalent thresholds for different measures VIII

Appendix D. Other graphics of gene-coexpression analysis. IX

Appendix E. Full DAVID enrichment analysis results. XI

Appendix F. Genes with positive cosine similarity index XXI

Appendix G. Detailed information about the genes with positive cosine similarity index in 3 or 4 layers XXXV

Appendix H. Thresholding effect on positive interactions. XXXVI

Appendix I. Simulation of the statistical method XXXVIII

1. Abstract

Tuberculosis is a poverty-linked illness that can be lethal, especially in combination with HIV. In this work, we perform a Systems Biology approach to its causal agent, *Mycobacterium tuberculosis*, to try to unravel the regulatory mechanisms used as a response to the kinds of stress it usually finds in its intracellular niche, the endosomes. For this purpose, an already developed method for gene co-expression networks is applied to the publicly available data of high throughput transcriptome data for *Mycobacterium tuberculosis* under different conditions of stress, making a multilayer network, where each layer is a different type of stress. To find differences between the different kinds of stress, distance metrics between the conditions have been developed and an analysis of the whole multilayer network has been performed to find general and specific stress response genes. To be able to identify differences between two different types of stress, differential co-expression networks have been built. As there is no consensus on a method for this, we have developed a new method for thresholding and differential co-expression networks analysis that solves problems found in previous methods, especially those concerning loss of significant data and inability to identify genes with alternate regulatory mechanisms. To validate the biological relevance all methods, data enrichment has been performed with the obtained results, showing concordant results with already published data, such as the importance of oxidative phosphorilation for adaptation to stress, specific proteins for oxidative stress such as CysD, lower differences when comparing the stresses happening in the endosome between them than when comparing with a kind of stress from outside of the endosome (UV) and a direct correlation between ribosomal proteins and proteins responding to any stress. However, further validation and improvement is required for this method.

La tuberculosis es una enfermedad ligada a la pobreza que causa la muerte, especialmente combinada con VIH. Aquí realizamos una aproximación de Biología de Sistemas al patógeno causal, *Mycobacterium tuberculosis*, para descubrir los mecanismos reguladores desencadenados como respuesta a los diferentes tipos de estrés de su nicho intracelular, los endosomas. Para ello, aplicamos un método ya desarrollado de redes de co-expresión génica sobre los datos transcriptómicos de "alto rendimiento" públicamente disponibles en *Mycobacterium tuberculosis* sometida a diversos tipos de estrés, generando una red multicapa, donde cada capa es un tipo diferente de estrés. Para hallar diferencias entre los diferentes estreses, se han desarrollado medidas de distancia entre las condiciones y se ha estudiado la red multicapa como un conjunto, para identificar genes de respuesta a estreses genéricos y específicos de estrés. Se han construido redes de co-expresión diferencial de genes para identificar diferencias entre dos tipos concretos de estrés. Al no haber consenso en un método para ello, hemos desarrollado un nuevo método para determinar un umbral de significancia y análisis de la red que resuelve algunos problemas de los métodos anteriores, especialmente los relacionados con la pérdida de datos significativos y con la incapacidad para detectar genes con mecanismos reguladores alternativos. Para validar la relevancia biológica de los métodos, se ha realizado enriquecimiento de datos con los resultados obtenidos, mostrando resultados acordes con la literatura publicada actual, como la importancia de la fosforilación oxidativa en la adaptación al estrés, proteínas específicas para el estrés oxidativo como CysD, menores diferencias entre los estreses fisiológicos del endosoma que al comparar uno de estos estreses con otro estrés de fuera del endosoma (UV) y una correlación directa entre los niveles de proteínas ribosomales y de respuesta a cualquier estrés. Sin embargo, es necesario mejorar y validar este método aún más.

2. Background

The object of study in this work is *Mycobacterium tuberculosis*, the main causal agent of tuberculosis (TB) in humans. This illness, together with malaria and HIV, is poverty-linked and it has been one of the main areas of work in health for the Millenium Development Goals set for 2015. In 2014, TB killed 1.5 million people, from which 0.4 million were HIV-positive, and 9.6 million people are estimated to have fallen ill with TB in 2014 worldwide. From 2015, new objectives have been set with the Sustainable Development Goals and the End TB strategy. [1]

*Mycobacterium tuberculosis (*Mtb*)* is an acid fast Gram positive bacilli that shows a complex life cycle of infection, even though it doesn't undergo such morphologic changes as some parasites as *Plasmodium spp.*, the causal agent of malaria. However, it undergoes different stages during its cycle of infection [2]. Depending on how the infection occurs, several forms of the illness can be found, such as extrapulmonary disseminated or miliary forms and pulmonary form of the illness. However, the former is not usual in immunotolerant adults, being the pulmonary form the main focus of the efforts against tuberculosis [3], so that its life cycle is best known in this case, as described below.

First, Mtb is inhaled as droplets, the main way of transmission of the illness. After droplets are deposited in the distal alveoli, Mtb is ingested by various phagocytic cells, such as dendritic cells, neutrophils or macrophages, being the alveolar macrophages the best studied. Several receptors are implicated in this process [4]. Once in the cell, Mtb starts multiple adaptive mechanisms to improve their chances of survival, such as inhibiting the maturation of phagosomes, promoting necrosis and inhibiting apoptosis, or recruiting new macrophages to use as a host [2, 4]. These mechanisms improve the cell to cell transmission and survival of Mtb in cell compartments that resemble the early endosomes [2], even though new studies have found several other intracellular niches for Mtb [4].

After this stage of growth, a delayed initiation of adaptative immunity happens, believed to be caused by an inhibition of migration of dendritic cells to lymph nodes. Here, an equilibrium is reached, where the growth of bacteria is arrested but bacteria aren't killed [2]. In this phase, some lesions known as granulomas can be developed. Initially, they were believed to keep bacteria in a non-replicative form, but still alive [3]. However, new studies have shown efficient replication of Mtb in granulomas [4]. The contribution to this equilibrium is not only from the host, activating its immune system, but also from Mtb, which activates a dormancy regulon controlled by the two component system DosR-DosS, changing the antigens of Mtb and using alternative energy sources. DosR-DosS can be induced by stimuli such as local hypoxia, nitric oxide and carbon monoxide [2]. However, latent Mtb infection can be reactivated by some not very well known mechanisms, even though there is a clear correlation with immunosuppression (e.g. HIV, therapeutic neutralization of TNF). Several mechanisms have been proposed, but there is no consensus about it [2].

The last step of this cycle is transmission. Transmission usually occurs through the airborne route, after the reactivation of latent Mtb, as the latent form is not contagious. The most contagious form is cavitary TB, where destruction of the lung tissue generates open spaces containing Mtb bacilli connected to the main airways, facilitating the transmission [2].

This complex cycle makes the treatment of TB harder, as most treatments have no effect on latent TB [1, 3]. Apart from this, resistances to first and second line antiTB treatments are appearing, being multirresistant the 3.3 % of new cases and the 20% of already treated cases [1].

This increases the cost of treatments and causes a higher rate of mortality, due to the lack of new drugs. A big effort on finding new treatments and on prevention must be developed, as the currently used BCG (Bacillus Calmette-Guérin) is not efficient in preventing adult pulmonary forms of TB [1, 5]. Also, TB becomes a bigger problem for HIV-positive population [1, 3, 6], being HIV-positive 12% of new cases, with around 400000 deaths in HIV-positive by tuberculosis in 2014, so the co-epidemic of HIV and TB should be addressed together [1].

More knowledge about the mechanisms causing TB and how the immune system recognizes it is necessary to improve this development. Animal models are a widespread approach to this illness. However, each animal is appropriate only for certain stages of the illness and none of the usual animals are as good as humans in transmitting the disease [3]. Furthermore, some mechanisms are not the same between species (e.g. mouse macrophages are not activated to kill Mtb by vitamin D, while humans are [4]) and these studies are limited to the first stages of disease [3]. Also macrophages can be used, but there is a high variability between them, giving different results depending on the tissue from where they were isolated [3].

The use of high throughput techniques for the analysis of the expression of all genes in Mtb or in its host is an alternative to these methods. It can be done at a protein level [7] or at different genetic levels, such as the study of polymorphisms that confer resistance/propensity to an illness [8] or the expression of the mRNAs to find a global network [9]. All of these techniques are comprised in what is known as Systems Biology.

An important part of Systems Biology is the processing of the High Throughput data to get information. One of the most used methods to try to explain the properties coming from the interaction of all elements studied in a high throughput set of experiments are networks. A network is the set of all interactions (links, e.g. physical interaction, expression correlation, regulatory relationship) between some objects of study (nodes, e.g. genes, proteins, transcription factors, functions) [10, 11, 12]. Networks can be defined for physical interaction between proteins [7, 13], for regulatory relationships [14], or for gene coexpression [9, 15, 16, 17, 18], to name a few examples. Even though the objects of study may vary, the methods and terms used keep the same.

Biological networks usually have some common topological properties. They don't usually follow a random distribution, where most of the nodes have a number of links (node degree) approximately equal to the average of degree in the network [12]. Instead, they follow a scale-free model (also known as power law degree distribution [19]), where $P(k) \sim k^{-\gamma}$, being P(k) the probability of a node to have k links, and γ the degree exponent [11, 18, 19]. This way, there are few nodes with a lot of links (known as hubs) and many nodes with few links. Also, the shortest path between two nodes picked at random usually involves a really small number of links, in a small-world effect [11, 12, 19]. Apart from that, some modules and clusters can usually be found. Modules are groups of nodes with a higher number of connections between them (clustering) than with the rest of the network [10, 11]. This usually has biological implications (e.g. modules corresponding to hypoxia response or to responses to other kinds of stress [9]). Furthermore, it is usual to find a high robustness, that is to say, a high resistance against failure in one random node, but, if the failure happens in some hubs, it could be fatal for the system [11]. Therefore, it is important to find hubs that can be used as targets for antibiotics.

In this work, gene co-expression and differential co-expression networks are being constructed and analyzed. Gene co-expression networks use genes as nodes and they are connected if a strong correlation between both genes is found among different conditions [18]. This way, the hubs of the network are genes with correlation with many other genes, so they are probably transcription factors or genes involved in regulation of gene expression. These networks also allow for the identification of modules that are usually related to a biological function, giving the chance to predict the function of some genes. Some studies have been performed before in human tissues for getting a global co-expression landscape in humans [16] and in brain cancer and yeast cell cycle [18], to find interesting modules for prognosis and development, respectively. Also, some interesting global [9] and regulatory [14] networks for *Mycobacterium tuberculosis* have been developed.

Differential gene co-expression networks study the change between two different conditions, such as healthy and infected individuals. In this case, nodes are also genes, but links can be defined in different ways depending on the method [17]. Usually, links are calculated for both conditions and then some operations are performed, but they can also be defined as the change in correlation between both conditions for a pair of genes. Also, there is no consensus in the most appropriate methods to use for these studies, as every method has its strengths and its drawbacks [17]. The main problem is usually not identifying genes that change the sign of their correlations, using an alternative mechanism, as it happens with human genes *p53* and *Klf4*, whose sign of correlation determines the way for DNA repair or apoptosis [20]. This methods have already been used for type 2 diabetes [17], different types of leukemia [15], finding relevant genes for the diagnosis of these illnesses, and flavonoid-deficient *Arabidopsis* [15], providing better knowledge of the implications of this changes.

3. Objectives

The main objective of this work is developing and enhancing a new method to create and analyse gene co-expression and differential co-expression networks using data from High Throughput experiments, solving problems of previously developed methods [17], with special emphasis on the following points:

- Loss of significant data due to thresholding methods.
- Reducing computational requirements of methods that use all data for all the analysis.
- Identification of genes with alternate mechanisms.

To check the biological relevance of the obtained results with this method, *Mycobacterium tuberculosis* data from Gene Expression Omnibus (GEO) database will be analysed for different types of stress and network data will be compared with publicly available experimental results.

4. Materials and methods

4.1. Data retrieval and ordering

All data in this work has been retrieved from Gene Expression Omnibus (GEO) database. GEO database is organised in samples, which are the results of a single microarray experiment, usually given as base 2 logarithms of the Folding Rate (FR) of the expression. The Folding Rate is the relation between the measure of the gene in the conditions given (usually stress) and basal conditions. In that way, an FR higher than 1 means that gene is over-expressed in that stress, lower than 1 means it is repressed and equal to 1 means there are no variations in expression of that gene [21].

As not every experiment uses the same probe sets or techniques, GEO defines also platforms, which are usually lists of the probes used in an experiment, which correspond (or not) to a gene.

In that way, samples are coded by platforms. There could be more than one sample per gene and experiment. There could also be replicates of the same experiment. Apart from that, it is usual to group GEO samples in series, which contain related GEO samples (e.g. increasing time of exposure to a stress, or exposure to different conditions causing a similar stress). In few words, GEO samples are coded by platforms and grouped in series [21].

For this work, we have defined seven different kinds of stress, six of them (hypoxia, nutrient starvation, cell wall damage, ion deprivation, NO exposure and generic oxidative stress) corresponding to physiological stresses created in the endosome[2,3], where *M. tuberculosis* lives for a part of its cycle [2,3,4]. The seventh stress is DNA damage caused by UV radiation. Each stress is classified in subtypes also (See more in Appendix A). Samples related with drugs or not related with these stresses were discarded.

In order to proceed with further steps, a matrix of base-2 logarithms of Folding Rates for each kind of stress must be created containing a GEO sample in each column and a gene in each line. If a gene is reported more than once in a single sample, its average is given in this file. If a gene is not reported in a sample, "NA" is written in its corresponding cell. All samples corresponding to a kind of stress must be contained in its respective file (See samples in each stress file in Appendix B). At this point, we decided to discard the nutrient starvation stress due to a low number of samples.

4.2. Correlation analysis

In order to create a gene co-expression network, a measure of how a pair of genes varies their expression must be defined. In this case, Spearman correlation coefficient is used, for further use in Fisher's z-test, as suggested in previous papers [15, 22], due to a better robustness against outliers. Spearman correlation coefficient for genes α and β in condition i $\rho^{i}(\alpha,\beta)$ is defined by the equation below:

$$\rho^{i}(\alpha,\beta) = 1 - \frac{6\sum_{k=1}^{N_{\alpha,\beta}^{i}}\left\{r\left[G_{\alpha}^{i}(k)\right] - r\left[G_{\beta}^{i}(k)\right]\right\}^{2}}{N_{\alpha,\beta}^{i}\left({N_{\alpha,\beta}^{i}}^{2} - 1\right)} \qquad (1)$$

Where α and β are two different genes, $N_{\alpha,\beta}^{i}$ is the number of samples with expression data available for both α and β at condition i, and $r\left[G_{\alpha}^{i}(k)\right]$ is the rank of gene α expression in the kth sample when all gene expressions in the sample are ordered from higher to lower in condition i. Comparing ranks instead of expression data is the main cause of its robustness. Spearman coefficient ranges from -1 (negative correlation, one gene is overexpressed when the other is repressed) to 1 (positive correlation, both genes are overexpressed or repressed at the same time). 0 means no apparent correlation.

These results are saved in a nxn matrix for each condition i (where n is the number of genes analysed). The coefficient of genes α and β is stored in line α, column β. As $\rho^{i}(\alpha,\beta) = \rho^{i}(\beta,\alpha)$, the resulting matrix will be symmetric. These matrices will be used in further steps.

4.3. Statistical significance

4.3.1. For gene co-expression

For analysing the statistical significance of data, Fisher's z-test will be performed. First of all, the Fisher transform of Spearman coefficients will be performed, as defined below:

$$\mathrm{F}(\rho) = \frac{1}{2}\ln\frac{1+\rho}{1-\rho} \qquad (2)$$

After this transformation, data behave as a normal distribution with average $\overline{\mathrm{F}(\rho)}$ equal to 0 and variance $\sigma^2 \approx \frac{1.06}{N^i_{\alpha,\beta}-3}$, dependant only in the number of samples containing data for both genes whose correlation is being studied ($N^i_{\alpha,\beta}$) [22]. As it follows a normal distribution of known average and variance, a Z value can be assigned to each pair of genes using the equation below. Z value measures, in standard deviations, how distant a value is from the average of its distribution. Higher absolute Z values imply higher significance.

$$Z = \frac{F(\rho) - \overline{F(\rho)}}{\sigma} = \frac{F(\rho)}{\sigma} \qquad (3)$$

As it follows a normal distribution, the error function (erf) can be used to calculate the p-values associated to the data of each pair of genes, with the following formula:

$$\mathrm{p} = 1 - \mathrm{erf}\left(\frac{|\mathrm{Z}|}{\sqrt{2}}\right) \qquad (4)$$

P-values measure the probability of a point of data of being as far from the average as it is in the case of a null hypothesis. This shows a problem with multiple testing. When the number of data is in the magnitude of millions, thousands of them may exhibit p-values lower than the usual 0.05 threshold due to chance and not to a real effect behind it. For this reason, false discovery rates (FDR) are used in this work. FDR is the proportion of false positives in the total of data considered positive. It can be calculated using the distribution of p-values. There are several methods for this aim. In this work, we will use one of the most frequent methods for FDR estimation, the Benjamini-Hochberg method [23], as it demands low computational requirements while it still provides a less conservative estimation than other popular methods, such as the Bonferroni correction [24], which allows the rejection of less data.

4.3.2. For gene differential co-expression

As a measure of the change in correlation between genes α and β when experimental condition i changes to j, $\Delta\mathrm{F}^{(\mathrm{i,j})}_{(\alpha,\beta)}$is defined, following the equation below, as done in previous work [15]:

$$\Delta\mathrm{F}^{(\mathrm{i,j})}_{(\alpha,\beta)} = \mathrm{F}\left[\rho^{\mathrm{i}}(\alpha,\beta)\right] - \mathrm{F}\left[\rho^{\mathrm{j}}(\alpha,\beta)\right] \qquad (5)$$

That is to say, $\Delta\mathrm{F}^{(\mathrm{i,j})}_{(\alpha,\beta)}$is the difference between the Fisher transform calculated for genes α and β in condition i and the same for condition j. As both terms behave as a normal distribution with average 0, the result will be a distribution with average 0 and a variation equal to the sum of both variations, that is to say[15]:

$$\sigma^2 = \sigma_{\mathrm{i}}^2 + \sigma_{\mathrm{j}}^2 = \frac{1.06}{\mathrm{N_i} - 3} + \frac{1.06}{\mathrm{N_j} - 3} \qquad (6)$$

After that, Z values, p values and FDR are calculated, using the same equations as in gene co-expression, with the only difference that $\Delta\mathrm{F}^{(\mathrm{i,j})}_{(\alpha,\beta)}$ is used instead of F (ρ) in the equation for Z values.

4.4. Analysis

4.4.1. For gene co-expression

First of all, a threshold is set at FDR=0.01, which means only 1 % of positives will be false positives. From this data, we calculate the set of genes interacting with gene α in condition i with a FDR lower than the threshold set, which is expressed in this terms:

$$S(\alpha, i, \tau) = \{\beta \neq \alpha : \mathrm{FDR}[\rho^i(\alpha, \beta)] < \tau\} \qquad (7)$$

Where τ is the threshold set and $\mathrm{FDR}[\rho(G_\alpha^i, G_\beta^i)]$is the FDR calculated for genes a and b in condition i. With this set of significant interactions, several metrics are calculated. First of them is node degree, which is the number of links a node has (i.e. the number of genes in the set).

However, degree doesn't have into account how strong the correlation of the gene with the rest is. To solve this problem, we define node strength, a measure of the importance of a node's interactions in a weighted matrix. First of all, a co-expression vector is calculated, which contains the values for Fisher transforms for gene α and every other gene β_k present in the set of genes defined for α. Mathematically, this is the expression of component k of that vector:

$$C_\alpha^i(k) = F[\rho^i(\alpha, \beta_k)] \qquad (8)$$

From this vector, we can define strength as the sum of all its components (in absolute values). As Fisher transforms can be positive or negative depending on the sign of the correlation and both positive and negative correlations are important, strength is the sum of absolute values and not of the signed components, as shown below:

$$s_\alpha^i = \sum_k |C_\alpha^i(k)| \qquad (9)$$

Once strength is calculated, other metrics can be used to study how it is distributed. Gene participation coefficient is a measure of how the strength is distributed between all significant links of a gene, ranging from 0, where all the strength comes from a single link (in this study it is only possible if there is only one gene in the set of genes), to 1, where all strength is equally distributed between links. It is defined by the following formula:

$$p_\alpha^i = \left[1 - \sum_k \left(\frac{F[\rho^i(\alpha, \beta_k)]}{s_\alpha^i}\right)^2\right] \frac{n_\alpha^i}{n_\alpha^i - 1} \qquad (10)$$

Where $n_\alpha{}^i$ is the number of components of vector $\overrightarrow{C_\alpha^i}$, i.e. the number of genes in the set of genes for gene α in condition i. Also, as a measure of how many correlations are positive and negative, an asymmetry index is defined. It shows the fraction of the strength that comes from positive correlations. Mathematically, it follows this equation:

$$a_\alpha^i = \sum_k \frac{F[\rho^i(\alpha, \beta_k)] H[\rho^i(\alpha, \beta_k)]}{s_\alpha^i} \qquad (11)$$

H function is equal to 0 if the correlation is negative and equal to 1 if the correlation is positive. In other words, it only sums data if its corresponding correlation is positive, giving as a result the fraction of strength from positive correlations. Therefore, the fraction from negative correlation would be 1-a_α^i.

4.4.2. Comparison between layers

To reach a better understanding of how *M. tuberculosis* reacts to different types of stress, comparison between layers is crucial to understand which mechanisms are shared for different kinds of stress. For this purpose, distance metrics have been designed, to estimate which layers are more similar among them. Two different methods have been applied, which study distance in two different ways.

In previous works [13], the conditional probability of finding a link in layer j when already present in layer i P (j|i) was used. But this approach has some problems. The first one is that P (i|j) ≠P (j|i), so the result is an asymmetric matrix, and it was solved using only the results corresponding to the layer with the lower number of links. As the maximum number of links shared is the number of links of the smaller layer, 1 could be reached, but it doesn't distinguish two equal networks from one network fully contained in the other, as both cases would score 1. A way of solving this is using the absolute fraction of all links in any of both layers which are in both layers (F_{Sa}), so that the result depends on the size of both layers. It is defined like this:

$$F_{Sa} = \frac{\#\mathrm{links}(i \cap j)}{\#\mathrm{links}(i \cup j)} = \frac{\#\mathrm{links}(i \cap j)}{\#\mathrm{links}(i) + \#\mathrm{links}(j) - \#\mathrm{links}(i \cap j)} \quad (12)$$

Where # links(i ∩ j) is the number of links shared in both layers, # links(i U j) is the number of links present in any of both layers (or in both) and # links (i) is the number of links present in i. However, the problem of not being able to reach value 1 when the size of the layers is different appears here. To solve this, we are using the relative fraction of shared links (F_{Sr}), obtained by dividing the absolute fraction by the maximum value reachable for that layers size (i.e. when all the smaller layer is contained in the bigger one). If layer i is bigger than layer j (in its maximum, # links (i ∩ j) = # links (j)) it is defined like this:

$$F_{Sr} = \frac{F_{Sa}}{\max(F_{Sa})} = \frac{F_{Sa}}{\frac{\#\mathrm{links}(j)}{\#\mathrm{links}(i) + \#\mathrm{links}(j) - \#\mathrm{links}(j)}} = \frac{F_{Sa}}{\frac{\#\mathrm{links}(j)}{\#\mathrm{links}(i)}} = \frac{F_{Sa}\,\#\mathrm{links}(i)}{\#\mathrm{links}(j)} \quad (13)$$

In this work, we will use both absolute and relative fractions of shared links, because they can give different information. For example, if a layer is a subset of another layer, its absolute fraction would be low, as both layers are different, but their relative fraction would be high, due to one layer being contained in the other. In that way, using both fractions can explain better the differences between layers than using only one of them.

Apart from studying similarity between layers, we measure the importance of the interactions of a gene in the whole multiplex and how it is distributed. For the first one, the measure used is overlap. Overlap is the sum of the strength of a gene over all layers of the multiplex [10, 25]:

$$o(\alpha) = \sum_i s_\alpha^i \quad (14)$$

To study how genes' strength is distributed in layers, multiplex participation coefficient is used. It is similar to gene participation coefficient, but instead of studying how the strength of a gene is distributed in single links for a layer, it studies how the overlap of a gene is distributed in layers for the whole multiplex [10, 25]. It is defined as follows:

$$p_\alpha = \left[1 - \sum_i \left(\frac{s_\alpha^i}{o(\alpha)}\right)^2\right] \frac{M}{M-1} \quad (15)$$

Where M is the number of layers in the multiplex. As the gene participation coefficient, it ranges from 0, when all strength is in only one layer, to 1, when strength is homogenously distributed among all layers. A lower value of this coefficient means it is more specific of a single layer. Studying overlap and multiplex participation coefficient at the same time is useful for finding important genes (or hubs) and finding if they are stress-specific or general responses to stress. [10, 25]

4.4.3. For gene differential co-expression

In gene differential co-expression, many metrics are shared with gene co-expression, but with a change in the point of view. Here, the object of study are interactions that change significantly between one layer and the other. As not all statistically significant interactions change in a significant way on other layer and significant changes don't have to necessarily be in significant interactions for both layers, a new definition for set of genes is required:

$$\mathrm{S}(\alpha, \mathrm{i}, j, \tau) = \left\{\beta \neq \alpha\colon \mathrm{FDR}\left[\Delta \mathrm{F}^{(\mathrm{i,j})}_{(\alpha,\beta)}\right] < \tau\right\} \qquad (16)$$

Here, the measure used in thresholding is not of a single layer but of the change between two of them, as shown in equation (5). As the set of genes changes, new definitions for co-expression vectors, degree and strength are needed. Also, they will have to be calculated for both layers independently. Notation is a bit more complicated than before, as we are defining the measure of layer i when comparing layer i and j. In that way, co-expression vector for gene α in condition i when comparing conditions i and j is noted as $\overrightarrow{\mathrm{C}_\alpha^{\mathrm{i}|(\mathrm{i,j})}}$, and its component k is defined by the following equation:

$$\mathrm{C}_\alpha^{\mathrm{i}|(\mathrm{i,j})}(\mathrm{k}) = F[\rho^i(\alpha, \beta_k)] \qquad (17)$$

Where gene β^k represents the kth gene in the set of genes defined for gene α in equation (16). Strength is then defined in an analogous manner to equation (9), just changing the set of genes:

$$\mathrm{s}_\alpha^{\mathrm{i}|(\mathrm{i,j})} = \sum_{\mathrm{k}} \left|\mathrm{C}_\alpha^{\mathrm{i}|(\mathrm{i,j})}(\mathrm{k})\right| \qquad (18)$$

Once strengths are calculated for both layers, a measure of the importance of the change is needed. This measure is the strength shift, which is the difference between the strength in both conditions, normalized dividing it by the mean of both strengths, as in the formula below:

$$\mathrm{d}\left(\mathrm{s}_\alpha^{\mathrm{i}\to\mathrm{j}}\right) = 2\frac{\mathrm{s}_\alpha^{\mathrm{j}|(\mathrm{i,j})} - \mathrm{s}_\alpha^{\mathrm{i}|(\mathrm{i,j})}}{\mathrm{s}_\alpha^{\mathrm{j}|(\mathrm{i,j})} + \mathrm{s}_\alpha^{\mathrm{i}|(\mathrm{i,j})}} \qquad (19)$$

This measure ranges from -2 (all strength is in layer i) to 2 (all strength is in layer j). Strengths shifts near to 0 mean strengths in i and j are approximately the same. Strengths measure the importance of genes without having into account its sign, so that a strength shift of 0 could exist where all correlations are the same, but with altered sign. To be able to identify genes changing its mechanism of regulation, a measure for difference in signs is developed. This measure is the cosine similarity index. Mathematically, it is the cosine of the angle formed by vectors $\overrightarrow{\mathrm{C}_\alpha^{\mathrm{i}|(\mathrm{i,j})}}$ and $\overrightarrow{\mathrm{C}_\alpha^{\mathrm{j}|(\mathrm{i,j})}}$, which is calculated with the following equation:

$$\cos\left(\overrightarrow{\mathrm{C}_\alpha^{\mathrm{i}|(\mathrm{i,j})}}, \overrightarrow{\mathrm{C}_\alpha^{\mathrm{j}|(\mathrm{i,j})}}\right) = \frac{\sum_{\mathrm{k}} \mathrm{C}_\alpha^{\mathrm{i}|(\mathrm{i,j})}(\mathrm{k})\mathrm{C}_\alpha^{\mathrm{j}|(\mathrm{i,j})}(\mathrm{k})}{\left|\overrightarrow{\mathrm{C}_\alpha^{\mathrm{i}|(\mathrm{i,j})}}\right|\left|\overrightarrow{\mathrm{C}_\alpha^{\mathrm{j}|(\mathrm{i,j})}}\right|} \qquad (20)$$

Where $\left|\overrightarrow{\mathrm{C}_\alpha^{\mathrm{i}|(\mathrm{i,j})}}\right|$ is the module of the co-expression vector for gene α, calculated like this:

$$\overrightarrow{\left|C_\alpha^{i|(i,j)}\right|} = \sqrt{\sum_k \left[C_\alpha^{i|(i,j)}(k)\right]^2} \qquad (21)$$

Cosine similarity index ranges from -1 to 1, as a regular cosine. A value of 1 means correlations are all the same sign and all the components of the vector keep the same proportion with their equivalent components in the other vector. A value of -1 means all correlations have changed their sign. Therefore, studying strength shift and cosine similarity index, the way how its correlation changes can be studied.

4.5. Data enrichment

Sets of genes apparently relevant in previous analysis will be studied further to find biological relevance and possible explanations of the results obtained. Gene Ontology (GO) and KEGG Pathway databases are used to find out if relevant genes are specially implied in a concrete function and TubercuList [26] is used to retrieve further information for concrete genes.

In this work, the studied sets of genes are the 100 genes with highest strength in each layer, the 100 genes with higher overlap and genes with positive cosine similarity index. The search for terms significantly overrepresented is performed in DAVID database with its functional annotation tools [27, 28]. Thresholds considered for significant terms are p-value under 0.1 and gene count of at least 2, which are used as default.

5. Results

5.1. Gene pairs

As 3924 genes are being studied in this work, 3924^2 gene pairs are being analysed in different stress conditions, so that using representative examples is needed for showing results at gene-pair level. In this section, we are using interactions in cell wall damage and ion deprivation stresses (layers 2 and 3, respectively) as example. For choosing representative examples, we have sorted gene pair interactions into three groups, according to the number of layers in which they are statistically significant (in both, only in one or in none of the layers). In these groups, we have subdivided them in other two groups, using whether they are statistically significant in the differential co-expression network or not. According to this classification, we have selected one of each type, all of them with gene Rv0239 involved, encoding possible antitoxin VapB24, related with virulence, detoxification and adaptation. Interactions are classified in table 1.

Present in	Differential coexpression	Rv number	Coding for [†]	Function [†]
Both layers	Present	Rv0578c	PE-PGRS7	Antigenicity, modulation of immune response*
	Absent	Rv3489	Uncharacterized protein	Virulence, detoxification and adaptation
One layer	Present	Rv0017c	RodA	Peptidoglican biosynthesis, cell wall formation
	Absent	Rv0947c	Pseudogene of mycolyl transferase	Lipids metabolism
None	Present	Rv1733c	Hypothetical transmembrane protein	Cell wall and cell processes
	Absent	Rv1322	Hypothetical protein	Intermediary metabolism and respiration

Table 1. Genes interacting with Rv0239 for each described group, used as examples in this work.

[†] Data obtained from TubercuList [26] unless otherwise stated.

*Function of PE/PPE proteins reviewed in Mukhopadhyay, B. and Balaji, K. N. [29]

5.1.1. Correlation

As a graphical way of seeing correlation as Spearman coefficient, graphics with ranks of both genes whose interaction is being studied can be used, using the slope of linear regression as a visual measure of correlation. Rank is the position in which a point of data is placed when the list of expression values of that gene is ordered from higher to lower for the different samples, as described for Spearman coefficient. Rv0239 ranks are represented in axis x while their corresponding ranks of the interacting gene are in axis y. In this case, the threshold slope is around 0.54 (i.e. there is a change of more than 15 units in axis y in the change between 0 and 28 in axis x). This graphics are shown in figure 1.

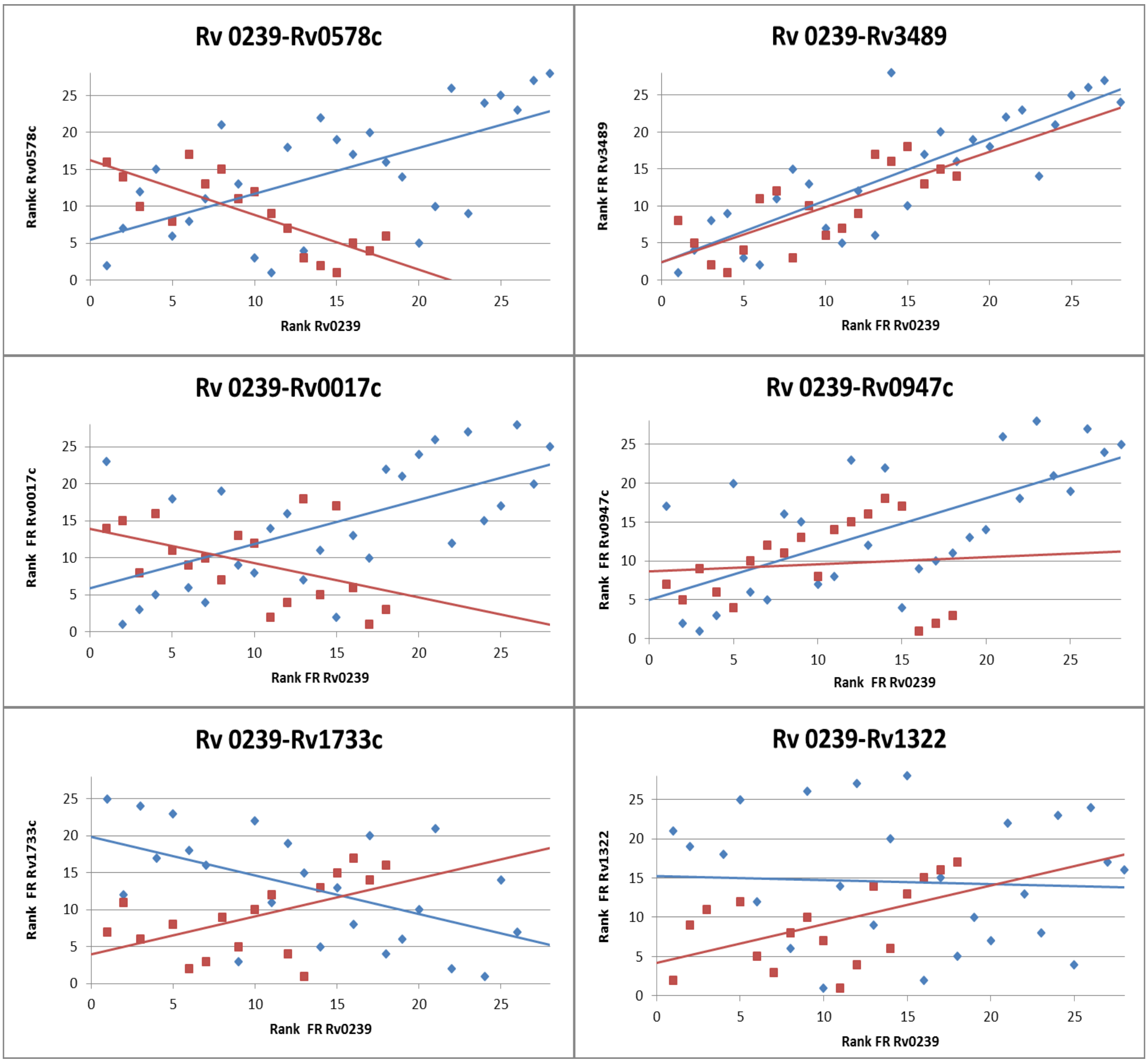
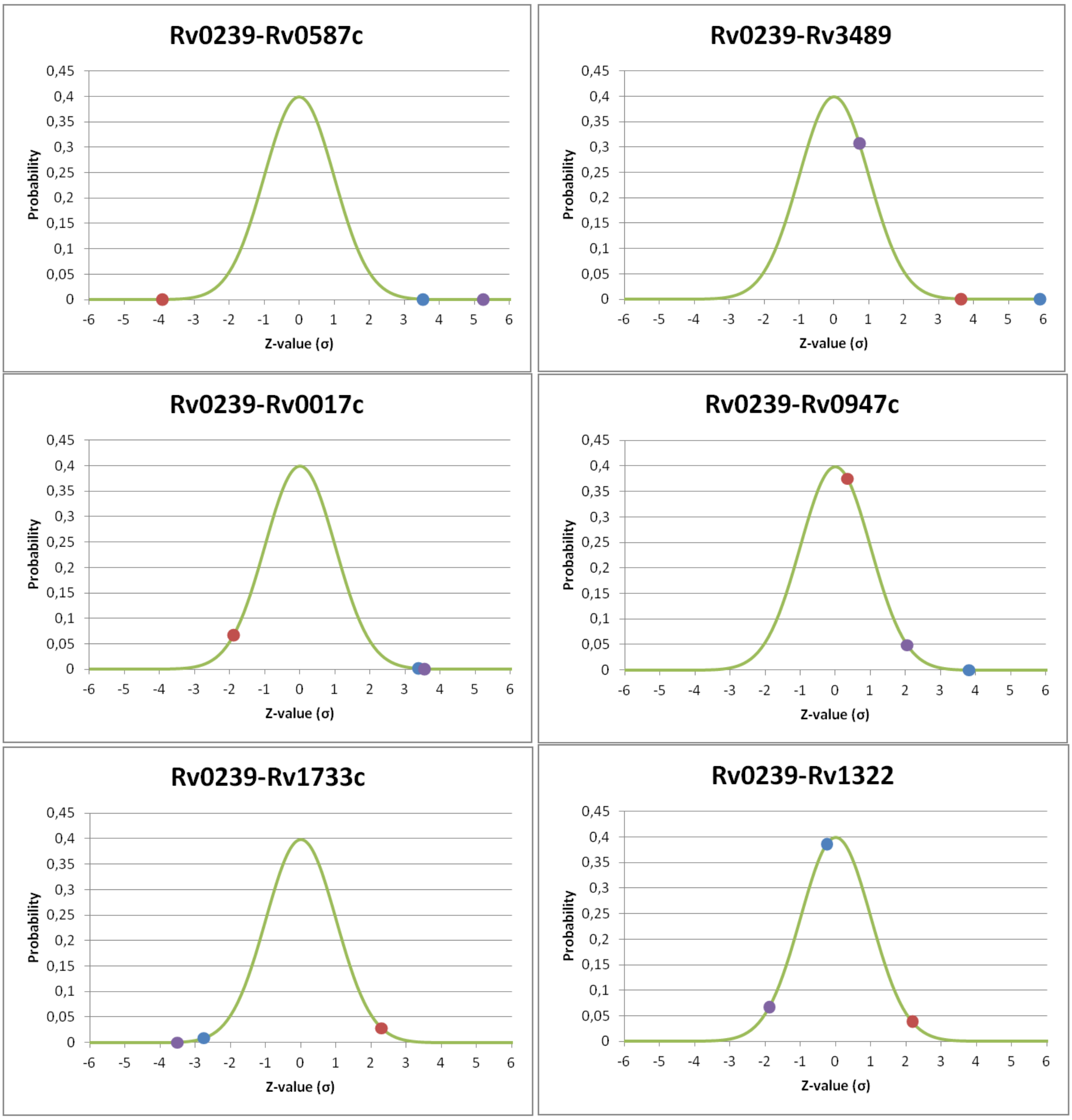


Figure 1. Correlation in six different gene pairs. Name of each gene pair is indicated at the title of each graphic. Blue dots correspond to the data of layer 2 (Cell wall damage), with its linear regression as a blue continuous line. The same applies to red, but with data for layer 3 (Ion deprivation) instead. Ranks are calculated (as explained in Spearman coefficient) from base 2 logarithms of the genes' folding rate (FR) compared to basal state, as retrieved from GEO database. The first row correspond to the pair of genes statistically significant in both layers; the second, to the pair of genes interacting only in one layer and the third, for the ones not present in any of the layers. Left column contains those interactions with significant changes in correlation (slopes are very different between both lines) and right column contains interactions without significant changes in correlation between layers.

5.1.2. Statistical significance

As a measure of statistical significance , in figure 2 we are showing z-values for all pairs of gene interactions on a normal distribution curve. P-value is twice the area under the curve from the z-value to the infinite of the same sign as the z-value. Because of this, the further a value is from 0 in this graphic, the more significant it is. We are using FDR=0.01 as threshold. As FDR is a function of the distribution of p-values, the threshold (in Z-values) is different for each co-expression or differential co-expression network (see more in Appendix C). In this case, an interaction is considered significant in layer 2 if Z-value is higher (in absolute value) than 3.257. For layer 3, this threshold is 3.513 and for the differential between them, 3.473. Results in figure 2 show also the Z-value for differential coexpression of that pair of genes.



Rv0239-Rv0587c

Rv0239-Rv3489

Rv0239-Rv0017c

Rv0239-Rv0947c

Rv0239-Rv1733c

Rv0239-Rv1322

Figure 2. Statistical significance in six different gene pairs. Name of each gene pair is indicated at the title of each graphic. In blue, data from layer 2, in red, from layer 3 and in purple, from differential between them. The green line shows the normal distribution for mean = 0 and σ = 1 (Used in Z-values). As z-values show how many standard deviations (σ) a point of data is from its mean, σ can be used as units for z-values. The position of the graphics is the same as in figure 1, to make it easier to interpret.

5.1.3. Relevance of each type of gene

Three types with two subtypes of gene pairs each have been described before according to their statistical significance. To get a global view of the distribution of the studied data in these groups, the abundance of each type of gene has been studied for each differential co-expression layer and studying all layers as a whole in figure 3.

5.2. Gene co-expression analysis

Several analysis have been performed with the metrics described before for co-expression networks. Here, only the graphics with relevant data will be shown, but the rest of them can be found in appendix D. As expected, the strength and the distribution are directly correlated with an almost linear fashion and show no relevant deviations from this behaviour (appendix D.1). Also, it can be seen that our network is quite homogenously distributed, as its participation coefficient is higher than 0.999 in all layers but one. Because of this, no additional relevant information can be obtained from the graphics of strength versus participation coefficient (appendix D.2).

Another important measure is the degree distribution, as explained in the background section. A scale-free distribution (i.e. with a high number of nodes with low degree and a low number of nodes with high degree) is a common result for biological networks. However, this only happens in the three layers with a lower number of samples, as shown in figure 4A.

Also, when studying the relation between strength and assymetry index, a symetrical pattern can be distinguished for all layers. It shows two peaks, better defined in some layers than in others, whose sum of assymetry indices is approximately one in all layers. Also, in some cases the first peak is located at a much lower assymetry index than in other layers (therefore, the higher peak is higher). Distance between peaks doesn't seem to be correlated with the number of samples of each layer and could show a different regulatory behaviour against different stress types.

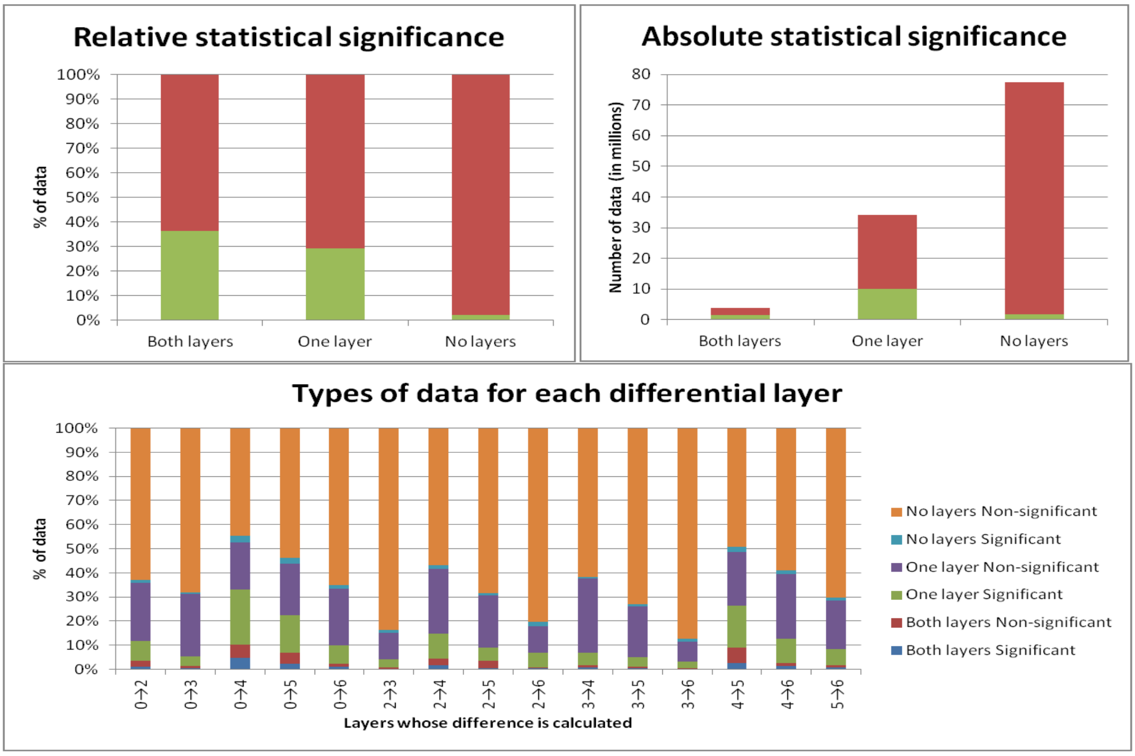


Figure 3. Abundance of each type of data. In the upper graphics, green represents significant data in differential coexpression and red represents non-significant. The upper left graphic shows the percentage of each group of genes (Present in both layers, in one or in none) which is signicative or not. The upper right graphic shows the amount of data of each group in absolute numbers (in millions), subdivided in significant and non-significant. In the lower graphic, this data is shown disaggregated for each differential layer. Color code can be found at the right part of the graphic. In this case, the percentage of data is used but, as the number of possible interactions in each layer is the same (coming all of them from the same organism with the same genes), the graphic is the same with absolute data and is not shown. Numbers in this graphic are the number codes for the different layers.

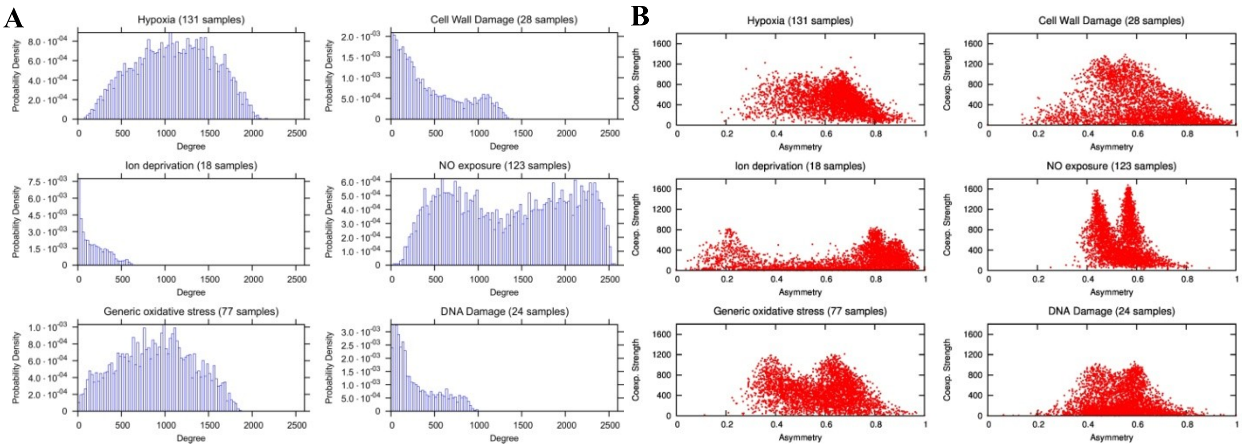


Figure 4. Relevant metrics in gene co-expression layers. Name and number of samples in each layer are indicated for each graphic. **A) Degree distribution**. Degree (x-axis) is shown in number of links and probability (y-axis) is calculated as the fraction of nodes within a degree range by the total number of nodes. **B) Assymetry index vs co-expression strength**. Assymetry index (x-axis) is shown as possitive correlated genes per unit and co-expression strength (y-axis) is shown in its arbitrary units.

Enrichment analysis shows similar functions enriched for all types of stress present in the endosome. Most of these functions are related to the obtention of energy and protein metabolism, such as aminoacid biosynthesis or translation from RNA. This suggests that a general mechanism for stresses present in the endosome may exist in *M. tuberculosis*. However, some more specific functions appear, such as photosynthesis in layer 0 (Hypoxia) or sulfate assimilation in layer 5 (Generic oxidative stress). Photosynthesis is clearly a misinterpretation generated by GO terms, which class NADH quinone oxidorreductase as related to photosynthesis, being in this case related only to the obtention of energy. Thus, it could be considered part of the general mechanism mentioned before. Sulfate assimilation genes overexpressed in layer 5 are cysN and cysD, the latter of them being associated in *E. coli* to shorter survival to oxidative stress in knock out bacteria [30]. These data confirms our analysis is showing biologically relevant results similar to previous studies. Full detailed DAVID enrichment analysis results can be found at appendix E. In layer 6, different results from the ones before are obtained. They are specially related to DNA repair, which seems a biologically coherent response to this stress not found in the endosomal environment.

5.3. Comparison between layers

The overlap and participation coefficient graphic (figure 5) shows no special patrons, but some genes with both high overlap and participation coefficient or both low overlap and participation coefficient can be found.

In the 100 genes with highest overlap, gene enrichment analysis shows the functions corresponding to the general response described before, such as translation, gene expression or protein metabollic processes. Also, having a look in which pathways are enriched, ribosomes can be found, as expected with such an importance of gene expression, but also oxydative phosphorilation is found, a term usually related to adaptation to different conditions and to slow growth rate by oxidating or reducing potentially dangerous species at many different levels and regulating the energetics of the cell, as reviewed by Cook et al [31]. When performing the analysis of the genes with participation coefficients lower than 0.8, less specific functions are found, mainly due to the kind of stress not to be found in endosomes, DNA damage, such as DNA integration. Some pathway spercific terms such as degradation of unusual substrates (limonene, pinene, anthracene, naphthalene…) and nucleotide excision repair, the latter of them being related to DNA damage.

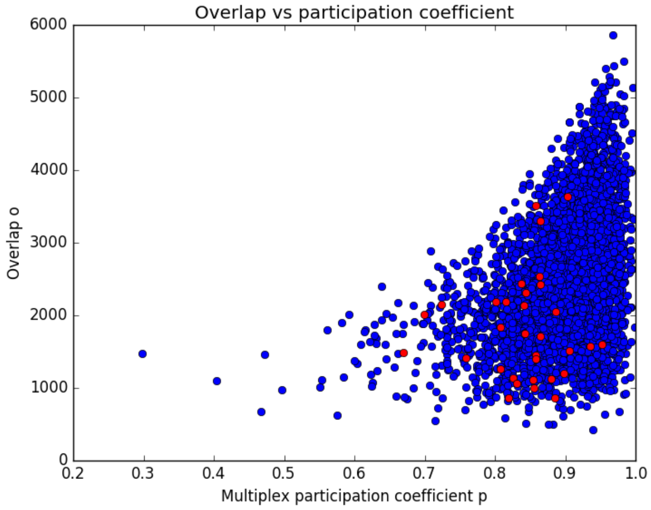


Figure 5. Overlap vs multiplex participation coefficient. On x axis, multiplex participation coefficient in a scale from 0.2 to 1. On y axis, overlap, in strength arbitrary units. Red dots represent genes with interactions only in 5 layers, while blue ones represent genes with interactions in all of them. For genes in 5 layers, a value of 0 is taken for the sixth layer and the calculation is made for six layers. No special patrons seem to appear here, but the genes with highest overlap or with multiplex participation coefficient lower than 0.8 will be analyzed for gene enrichment. Graphic made with Matplotlib [32]

Prior to differential analysis, the matrix of distances between layers has been performed with the two metrics described before. The first one gives a sense of how similar both layers are, while the second shows how much of the smaller layer is contained in the other one. When a layer is compared with itself, its result is always 100% match. These distance matrices can be seen in figure 6.

Fsa	0	2	3	4	5	6
0	100,0%	10,2%	4,4%	19,4%	16,0%	6,4%
2	10,2%	100,0%	4,2%	10,6%	11,6%	5,0%
3	4,4%	4,2%	100,0%	4,5%	4,6%	3,2%
4	19,4%	10,6%	4,5%	100,0%	18,2%	6,7%
5	16,0%	11,6%	4,6%	18,2%	100,0%	6,5%
6	6,4%	5,0%	3,2%	6,7%	6,5%	100,0%

FSr	0	2	3	4	5	6
0	100,0%	25,4%	27,8%	23,9%	19,8%	24,3%
2	25,4%	100,0%	10,7%	32,7%	23,4%	7,5%
3	27,8%	10,7%	100,0%	35,3%	23,6%	5,4%
4	23,9%	32,7%	35,3%	100,0%	27,8%	31,3%
5	19,8%	23,4%	23,6%	27,8%	100,0%	19,7%
6	24,3%	7,5%	5,4%	31,3%	19,7%	100,0%

Figure 6. Distance matrices for co-expression layers. On the left, FSa, absolute fraction of shared links, as defined in equation (12). On the right, FSr, relative fraction of shared links, as defined in equation (13). In both cases, results are expressed as a percentage. A color scale has been used to represent more visually these data, being green the most similar pairs of layers, red the less similar and yellow the intermediate ones.

5.4. Differential analysis

For differential co-expression networks analysis, the main analysis performed is strength shift and cosine similarity index (figure 7). For this graphic, a null model has been developed with random simulated data. In all cases, this model is not centered in 0,0. Instead, it is located to the side corresponding to an overexpression of the layer with the lower number of samples, as networks' interaction stregths diminish with a higher number of data samples [18]. Apart from that, cosine similarity index is mostly negative, as the difference between both correlations in the original layers is usually higher if they change their sign. Because of that, this graphics are shown with a code of colours showing the probability for each point of being represented by the null model.

As genes with a possitive cosine similarity index are unusual, further analysis is performed. From each condition, all genes with positive cosine similarity index are taken and these lists of genes are compared to find if this characteristic is replicated for a gene in several layers (This table can be found in appendix F) . Then, a gene enrichment analysis is performed for the genes appearing as positive in more layers (3 or 4, for a total of 28 genes, concretely Rv0177, Rv0315, ***Rv0396***, Rv0563, ***Rv0635***, ***Rv0642c***, Rv0700, Rv0701, Rv0733, ***Rv0957***, ***Rv1201c***, Rv1528c, Rv1613, Rv1671, ***Rv1770***, Rv1806, Rv1919c, Rv1980c, Rv2882c, Rv2929, Rv2954c, ***Rv2973c***, Rv3040c, Rv3061c, ***Rv3553***, Rv3671c, Rv3742c, Rv3752c, being in bold italics the ones appearing in 4 layers). Individual information about each gene can be found in appendix G. Also, a gene enrichment analysis is performed on all genes with positive cosine similarity index for 1 or more layers.

For the gene enrichment analysis on all genes with at least one positive correlation index, mainly terms related with translation, ribosomes and protein and nucleotide synthesis are found. This is unsurprising, as responses for several stresses need the production of proteins via translation, where the ribosomes and nucleotides are needed, so that it is not unusual to see that processes related to the response to stress correlate in a positive way with these mechanisms (Full results are shown in appendix E). The results for gene enrichment analysis for the 28 genes with positive cosine similarity index in 3 or 4 layers are shown in table 2.

As it can be seen, the results are mainly related to proteins, translation, ribosomes and purine metabolism, which is the same as in the more general group, so that these genes represent also the enrichment of the bigger group.

Term	Count	%	P-Value	F. E.
Ribosome	2	7.7	1.40E-01	11.8
Ribosome	2	7.7	1.40E-01	11.8
Ribosome	2	7.7	1.40E-01	11.5
Ribosome	2	7.7	1.40E-01	11.3
Purine metabolism	2	7.7	1.40E-01	11.1
Purine metabolism	2	7.7	1.50E-01	10.4
Purine metabolism	2	7.7	1.50E-01	10.4
Purine metabolism	2	7.7	1.60E-01	10.2

Term	Count	%	P-Value	F. E.
Protein metabolic process	5	19.2	2.00E-02	4.2
Primary metabolic process	11	42.3	2.80E-02	1.7
Translation	3	11.5	3.00E-02	10.1
Biosynthetic process	8	30.8	7.40E-02	1.8

Table 2. Data enrichment for genes with a common positive cos similarity index. On the left, results for KEGG database. On the right, results for GO terms. Due to the low number of genes, data enrichment parameters have been made less strict, using 0.2 as the threshold for significant enrichments instead of 0.1

6. Discussion

6.1. Gene pairs

As it can be seen in the examples shown in figures 1 and 2, this method allows us to find important differential coexpression links even when the correlation in both layers is not statistically signicative but its change between both of them is. This is important because, even though only a small fraction of non-significant data in both layers shows a significant change (figure 3, upper left), as the amount of non-significant data in both layers is around two thirds of the total data, the absolute number of positive changes in correlation for these data is comparable to the one for genes significantly correlated in both layers (figure 3, upper right), being 12,63% of the significative changes. Having this facts in mind, we can proceed to analyse and compare different thresholding methods used up to date, reviewed by Yu, Liu et al [17].

Up to date, three different kinds of thresholding have been used for this purpose: hard, soft and half thresholding. In this work we propose a fourth method, differential thresholding.

Hard thresholding is simple, a threshold is set for a concrete parameter, such as the correlation coefficients, p-values or FDRs. If the value for that parameter is higher (for correlation coefficients) or lower (for p-values and FDRs) than the threshold set, both genes are connected. If not, there will be no direct link in the network between that pair of genes. The problems with this method are several. First of all, the election of the threshold could significatively change the resulting network, allowing more or less interactions. Apart from this, the network usually becomes unweighted by this method and analysis have some more problems because of that, which will be discussed below. As non-significant data doesn't surpass the threshold, the significative changes in these data won't be detected by this method.

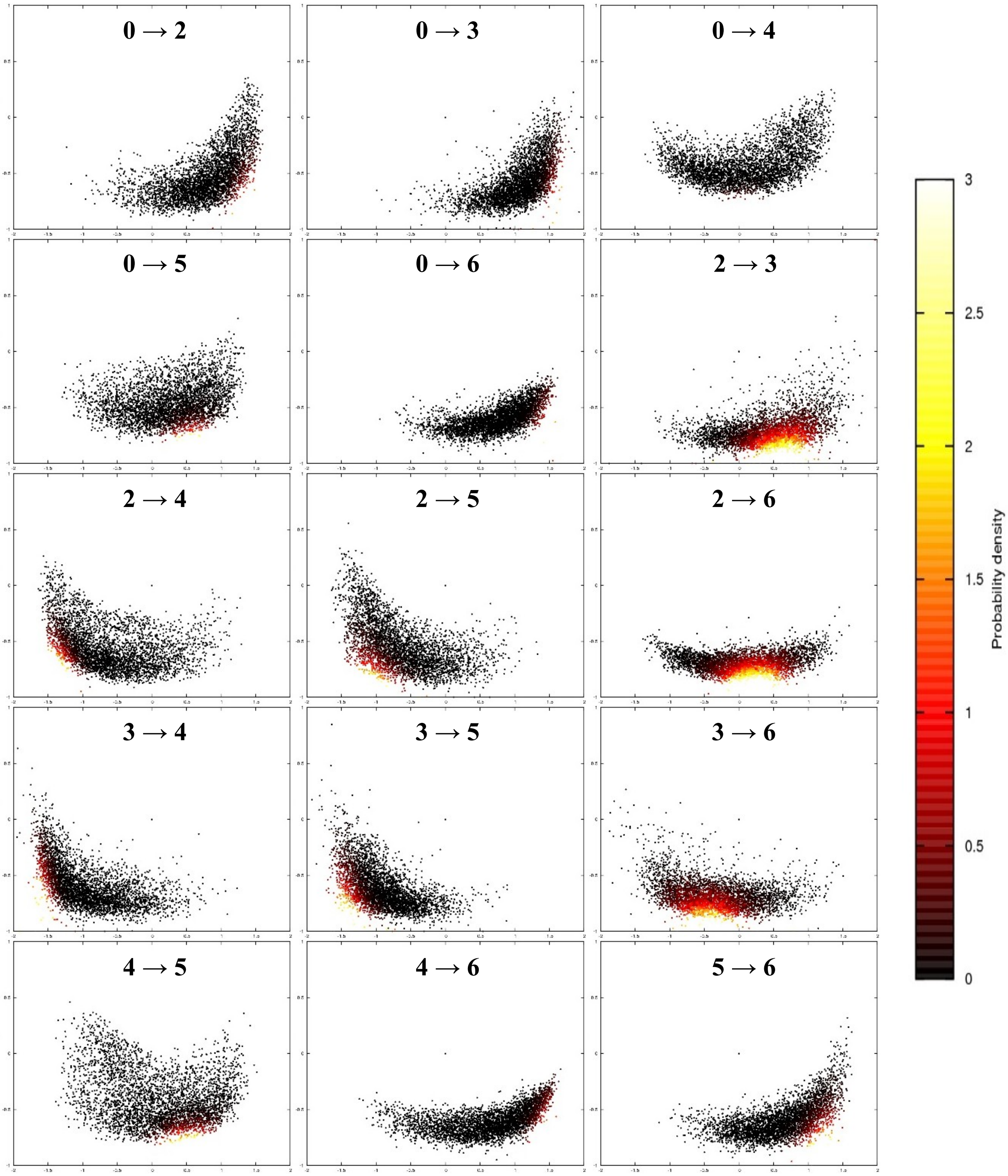


Figure 7. Cosine similarity index vs. strength shift. Strength shift on x-axis, ranging from -2 (left) to 2 (right), with partitions each 0.5 strength shift units. Cosine similarity index on y-axis, ranging from -1 (bottom) to 1 (top), with partitions each 0.5 units. The color code shown on the bar on the right represents the probability of each point of data (representing a gene) of being caused by a random distribution, where 0 (black) is a highly improbable point of data at random and 3(white/yellow) is a very probable point of data at random. This data has been obtained by comparison with a null model in which lists have been generated at random with the same number of samples as their respective layers. Numbers represent the different types of stress, using the following code: 0 (Hypoxia), 2 (Cell wall damage), 3 (Ion deprivation), 4 (NO exposure), 5 (Generic oxidative stress), 6 (DNA damage). The codes are written as i →j in formulas (19) and (20), so that a positive strength shift value indicates higher strength in the layer at the right side of the arrow.

Soft thresholding is a method for getting weighted networks. It considers all interactions exists and tries to filter the most important ones with different operations, usually raising the correlation coefficient to a power beta, with makes small values became smaller faster than higher values. Beta is usually defined in a way that the network behaves accordingly to the scale free model, but there are some values of beta that usually work well, such as six [33]. The problem with this method is that it may not make a clear distinction between important and unimportant interactions. Also, as all interactions are considered, the computational cost is higher and, as small values are made even smaller, important changes in these values will be made small too, being unable to be detected.

In view of the problems mentioned for this methods, Yu and Liu proposed half thresholding for differenttial co-expression networks [17], which is considering the interactions which are significant in at least one of the two layers which are being compared and giving them as a weight the value for the original network. This solves the problems of unweighted networks and makes a better distinction between significant and non-significant interactions. Anyway, it isn't able to detect important changes for interactions shown as non-significant for both layers, which we have seen can be important, as shown in figure 3.

Because of that, differential thresholding has been developed in this work using previously proposed definitions of significance in change [15]. This method evaluates not the value for each layer, but the change in those values, as explained in the methods section. That way, all significative changes are detected, including the ones when the correlation is non-significant for both layers. This is the main advantage over the rest of methods, but it also discards non-significant changes for some genes significant in one or both layers, which makes the set of data to analyse smaller, reducing the computational cost. Apart from this, it makes a weighted network, using the Fisher transforms for correlation coefficients as weights for interactions.

An overview of how each method interprets each type of data in a network using the same genes that in gene-pair results section is shown in figure 8, as well as different methods for its analysis. Further discussion about the different methods for differential analysis can be found below.

6.2. Gene co-expression analysis

In figure 4A, it can be observed that the behaviour of the constructed coexpression networks only follows the scale free model in half of the layers. This could be a consequence of having very dense networks. A way of making these layers less dense could be using an even lower threshold, but lower values than 1 million of interactions in the bigger layers aren't reached even with a threshold of FDR = 0.001(See appendix H) and lowering it more could result in a loss of important data.

An interesting thing in common between the layers that don't fulfill the characteristics of a scale-free model is that the number of samples is higher than 75. Even though the definition of σ^2 was empirically calculated for groups of data with a number of samples between 10 and 50 [22], simulations have been made and this formula seems to work better for higher values up to 150 samples (See appendix I). Other coexpression studies have been performed with the global coexpression network of *M. tuberculosis* without this problem [9]. The difference may reside in which conditions are being studied. While in our study, conditions are seggregated according to specific types of stress, they are using all of them as a whole. Furthermore, they are forcing a concrete mean degree of distribution with a soft-thresholding approach, and making a network of modules, which contributes to a less dense network. Some of these methods, such as the WCGNA approach for link strengths [33], could be used for further studies of modularity.

In figure 4B, an almost simetric pattern can be seen for each of the layers. To find out what happens, a network containing only the 100 genes with more strength placed in the positions they are in figure 4B is represented (figure 9A). In this figure, it can be seen that two genes with a negative correlation tend to have the same neighbours but with different sign of interaction. This, together with the fact that participation coefficients are very high (all interactions of a gene are of similar strength, which also explain the linear correlation between degree and strength of a gene) explains why two genes can have similar strength (having the same neighbours) but with the assymetry of one being one minus the assymetry of the other one (different sign of interaction with their neighbours), as shown in figure 9B.

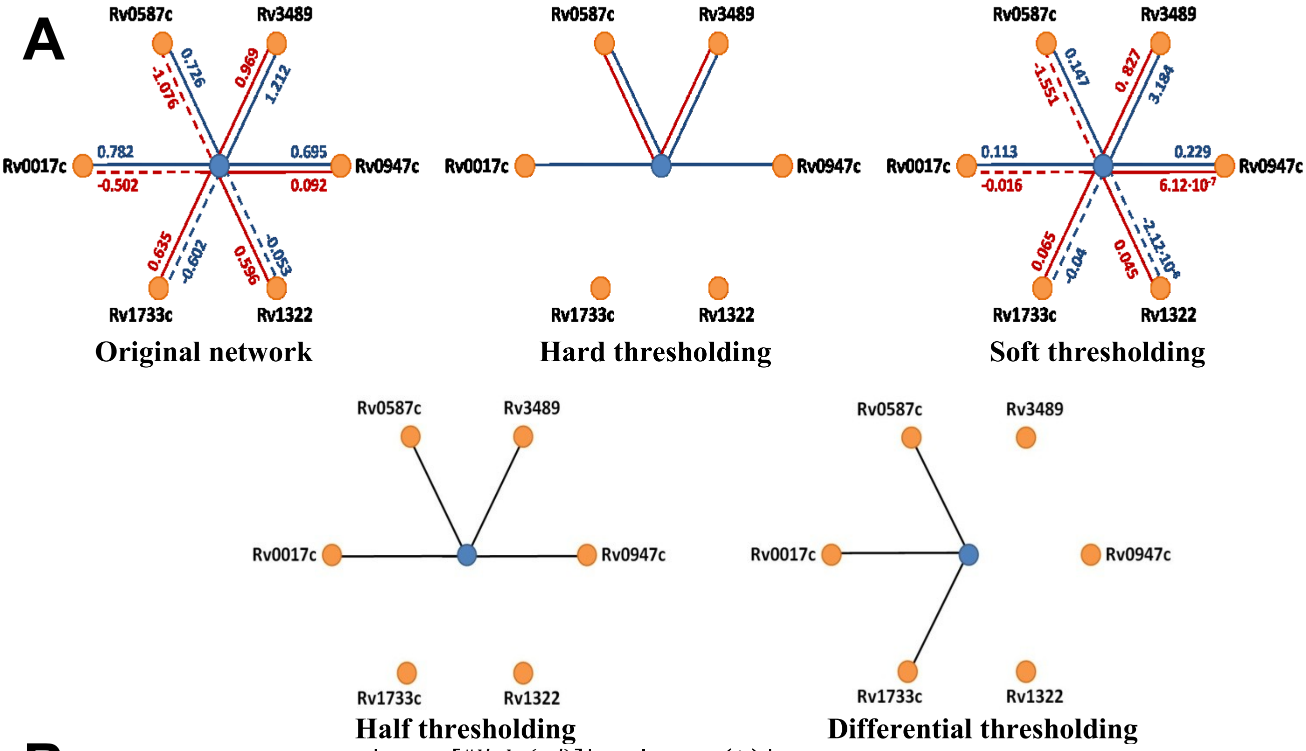
As shown in the results section, enrichment analysis leads to expectable results, such as a generic importance for all stresses related to a bigger regulation of all steps necessary for protein synthesis, including transcription and translation. It also shows some results found before in another publication, such as the increase of the importance of cysD and cysN for generic oxidative stress [30], which is not present in any other stress. This is a good sign for the network analysis procedure, as it has been able to predict something that has already been described. Anyway, the amount of data obtained from enrichment analysis is very big and some procedures for analysing this data and improving its quality should be developed. Further studies of modularity in this network could also lead to clusters of genes regulated by the same effectors, but they haven't been performed for these networks yet.

6.3. Comparison between layers

We found no outstanding genes or groups of genes in the overlap vs multiplex participation coefficient. Also, most of the genes have significant interactions in all layers, and the ones that are only in five layers seem to behave in a similar way than the others, except for the lower multiplex participation coefficient, as a consequence of not appearing in one layer. Data enrichment of the genes with higher overlap shows the terms appearing in the general response found in all layers, as expected. This shows that the results obtained are coherent within this method. Also, gene enrichment of the genes with lower multiplex participation coefficient shows specific responses, so this method also works for finding specific responses by finding genes specifically enriched in one layer and not in the rest. Anyway, due to the high density of our network, analysis is harder to perform and may show some enriched terms without real biological significance.

When looking at the distance metrics between layers, it could be seen how both of them don't give the same results as they are not measuring the same. The most informative of them is the relative fraction of shared links, as it depends less on the density of the networks. Anyway, some dependance still exists, as it can be seen in layer 4 being the closest to most of the other layers. This is because the bigger both layers are, the more probable is both of them share links by random.

Appart from this, some interesting things can be seen, such as layer 6 (DNA damage) being the most distant to the other layers in most of cases. This has biological significance because this is the only studied stress not happening in the endosome, so it may use very different mechanisms as a response and, because of that, have a big distance with the rest of stresses. Furthermore, some distant layers can be found inside the endosome stresses, such as 0 (Hypoxia) and 5(Generic oxidative stress) and 2 (Cell wall damage) and 3(Ion deprivation). Given the fact that 2 and 3 are small layers, this distance may be an artifact caused by that. For 0 and 5, as they are both big layers, this information is more reliable. Also, it seems to have biological significance, as oxidative stress is caused by an excess of oxygen reactive species and hypoxia is a lack of oxygen, so response to this stresses may go in a different way, as the distance matrix shows.



Original network

Hard thresholding

Soft thresholding

Half thresholding

Differential thresholding

B

i) $$dC(LRC) = \left|\log_{10}\left[\frac{\#links(\alpha,i)}{\#links(\alpha,j)}\right]\right| = \left|\log_{10}\left(\frac{4}{2}\right)\right| = 0.301$$

ii) $$dC(ASC) = \frac{\#specific\ links(\alpha,i) + \#specific\ links(\alpha,j)}{2} = \frac{2+0}{2} = 1$$

iii) $$dC(WGCNA) = \frac{S_\alpha^i}{\max(S^i)} - \frac{S_\alpha^j}{\max(S^j)} = \frac{0.147+3.184+0.113+0.229-0.04-2.12\cdot10^{-8}}{3.625} - \frac{-1.551+0.827-0{,}016+6.12\cdot10^{-7}+0.065+0.045}{0.827} = 1-(-0.761) = 1.761$$

iv) $$dC(DCp) = \sqrt{\frac{\sum_k^n\left(c_\alpha^i(k)-c_\alpha^j(k)\right)^2}{n}} = \sqrt{\frac{(0.726+1.076)^2+(1.212-0.969)^2+(0.782+0.502)^2+(0.695-0.092)^2}{4}} = 1.142$$

v) $$\mathrm{d}\left(\mathrm{s}_\alpha^{\mathrm{i}\to\mathrm{j}}\right) = 2\frac{\mathrm{s}_\alpha^{\mathrm{j}|(\mathrm{i,j})}-\mathrm{s}_\alpha^{\mathrm{i}|(\mathrm{i,j})}}{\mathrm{s}_\alpha^{\mathrm{j}|(\mathrm{i,j})}+\mathrm{s}_\alpha^{\mathrm{i}|(\mathrm{i,j})}} = 2\frac{(1.076+0.502+0.635)-(0.726+0.782+0.602)}{2.213+2.11} = 0.0477$$

Figure 8. Methods for thresholding and differential analysis. A) Methods for thresholding. Blue circle represents Rv0239, while orange circles represent the other genes studied in the section of results at the gene-pair level. Names of the genes are written beside the circles. Blue lines represent interactions in layer 2, while red ones are for layer 3. Continuous lines are used for positive correlations while discontinuous are used for negative correlations. The method used for thresholding is written right below the network. For the original network, the number beside the lines indicates the Fisher transform of the correlation value of the interaction of that pair of genes in the layer using the color of the number. For soft thresholding, the number is this value to a power beta of 6, which is a number that usually works well [33]. Hard thresholding only takes as data whether the interaction is over the threshold or not, so that no numerical data is used in the network. For half and differential thresholding, links to be studied are indicated with black lines. Only data for these interactions is preserved in these methods, the rest of information is ignored. The processing of the data will be performed later with the data of the original network. B) Methods for differential analysis. LRC: Logarithmic Ratio of Connections, ASC: Average Specific Connections, WGCNA: Weighted Gene Co-expression Network Analysis, DCp: Differential Co-expression profile; $\mathrm{d}\left(\mathrm{s}_\alpha^{\mathrm{i}\to\mathrm{j}}\right)$: Strength shift. Nomenclature used in these formulas is the same as in formulas 12 and 16-21. # specific links (α, i) refers to the number of links present in layer i but not in j. Each differential analysis has been performed using the networks from A correspondent to the usual thresholding for each method. LRC and ASC use hard thresholding; WGCNA, soft thresholding; DCp, half thresholding and strength shift, differential thresholding. A more detailed explanation of each method can be found below at the differential analysis discussion.

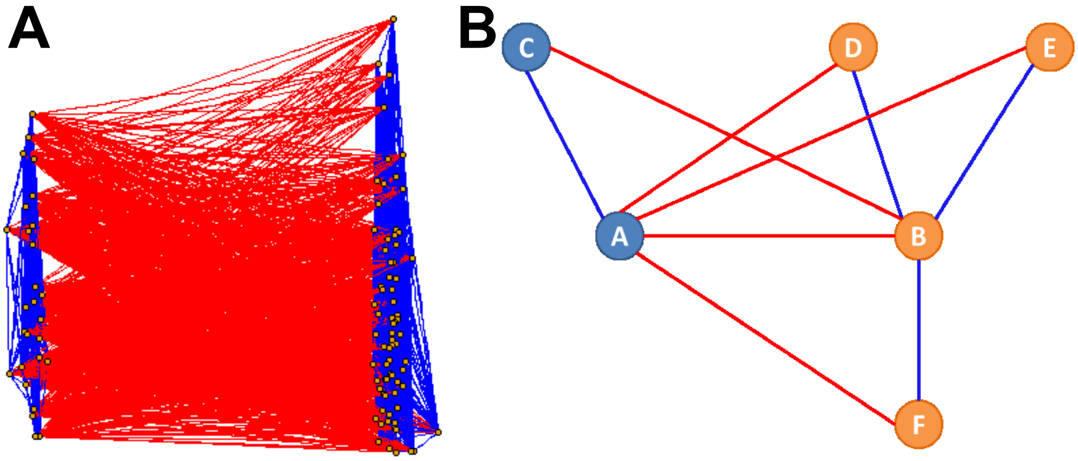


Figure 9. Analysis of the gene interaction network for layer 3 for Assymetry index vs co-expression strength graphic. A) Network for the 100 genes with higher strength. Genes are represented with orange squares and are located in the same relative position as they were in figure 4B for the NO exposure layer. Blue lines represent positive interactions, while red ones represent negative interactions. B) Simplified model of A. Only the interactions of genes A and B have been represented here. Here, genes A and B interact positively with the genes of their same colour and negatively with the ones of different colour. As A and B have the same neighbours and the interaction between them is negative, the genes whose interaction with A is positive will have a negative interaction with B and vice versa. In that way, the only interaction with the same sign for A and B is the interaction between A and B. As all interactions are of similar strength, we can suppose their strength is 1 or -1 for all interactions in this example. Having this into account, the assymetry index for A would be 1/5 and for B, 3/5. Here, the sum of both is not 1 but, if we find the formula for the sum of both as a function of neighbours per gene, the sum will be $\frac{n-1}{n}$, where n is the number of neighbours, so that when n becomes bigger, the sum tends to 1, as observed in figure 4B. This way, we can explain the particular shape of that graphic is due to a dense network with high participation coefficient for all genes.

6.4. Differential analysis

When performing differential analysis, there are some considerations to take into account for our method. The most important thing is that we are using Fisher transforms, which have a high dependance on the size of our sample [18] so, if metaanalysis is being performed and the size of the sample for both layers is different, as in this case, distributions of strength shift are going to be skewed towards the side of the layer with the smaller sample size. Apart from that, positive cosine similarity indices are rare, as there is a lower range of data for numbers of the same sign (positive cosine similarity index) than for numbers of diferent sign (negative cosine similarity index). Because of this, it is really important to know where the data is probable or not to be due to random, as it is done in figure 7 with the color code. The problem with the deviation in strength shift could probably be solved using a correction made from a previously simulated random distribution with the corresponding sizes to the case of study. However, this needs to be further adressed to find an optimal solution.

Looking at how defined distributions can be by random, three differential networks seem to be represented in a higher way than the rest: 2 → 3, 2 → 6 and 3 → 6, which are exactly the three combinations of layers with lower similarity (figure 6, right). This could be a consequence of layers 2, 3 and 6 being the three smallest layers in sample size and number of links. As discussed before, a small number of links could lead to a lower value of similarity between matrices. Also, a small number of samples makes the null model represent a bigger area in the graphics, as there are less combinations possible with a smaller sample size.

Also, the number of layers in which genes have positive cosine similarity index is low even in the best cases (4 out of 15). This can be due to our differential thresholding, which may only allow interactions with the same sign when one of them is really high and the other really low (in absolute value), so that if a gene in one layer has a positive cosine similarity index with other two layers, the interactions for that gene will have a similar level in that other two layers and then those interactions between those two other layers would be non-significant. Anyway, this would not necessarily lead to having positive cosine similarity indices in few layers. If there were some kind of alternative mechanisms for a gene, it could happen that the interactions which make its cosine similarity index positive between one layer and the rest would become non-significant when comparing the other layers but new interactions could appear to make that cosine similarity index positive. However, this doesn't seem to happen in this network, probably due to its high density of interactions.

Having a look at the data enrichment for these genes, it can be seen that there's probably a biological explanation for why this happens. These genes are mainly related to translational and transcriptional processes, which are necessary for all kind of cellular responses, as explained in the results section, so that they can vary the intensity of their correlation but they are usually positively correlated with an increase in other genes (unless they are coding for a repressor).

The main objective of this work is developing a better method of analysis for differential coexpression networks. For that, we should overcome the problems that previously developed methods show, as pointed out by Yu, Liu et al. [17].

Logarithmic Ratio of Connections (LRC) method considers the base-10 logarithm of the ratio of the number of connections between one layer and another. This method has the problem of not distinguishing properly the neigbours of a gene [17], so that a gene with two alternative mechanisms which make the gene be connected to a completely different set of genes for each layer can have a low score if both layers have a similar number of connections (even though the connections are totally different). This is solved by the Average Specific Connections (ASC) method, which uses the number of links appearing only in one of both layers and makes an average as a marker. The common problem of these two methods is the use of hard thresholding, which generates a loss of quantitative information for the links and makes impossible to detect changes of sign in correlation, apart from making the results highly dependant on the chosen threshold. [17]

Soft thresholding is then applied in Weighted Gene Coexpression Network Analysis method (WGCNA), one of the most used methods currently [17], which offers a package on R for analysis [33]. This method shows two variants, the signed and the unsigned, depending on whether the sign of correlation is conserved after the thresholding or not. Usually, signed WGCNA is used. In WGCNA, a comparison of the normalized strength of a gene in both layers is used. This gives an idea of the change of importance of the gene between the two layers, but fails to point genes which change the sign of their correlation between two layers [17]. Even in signed WGCNA, a gene can change the sign of some important interactions and keep their strength at the same level. Also, the election of the power beta in this method is highly empirical, trying to make the network fulfill a scale free model, and sometimes important interactions are difficult to distinguish from trivial ones.

To solve these problems, Yu, Liu et al propose the Diferential Coexpression profile method (DCp), in which the difference of coexpression between layers is evaluated link by link for all links of a gene. This allows the detection of changes in the pattern of coexpression of a gene [17], but it requires higher computational cost, as each link is evaluated separatedly, specially in the case of comparing more than two layers, as is the case in this work. Also, the method of half thresholding doesn't allow for significative changes which are not significative in neither of the layers to be detected.

This is solved in our method, as it uses differential thresholding. Also, strengths are used for strength shift, reducing the computational costs. The problem with this is the same than with signed WGCNA, it measures the change of importance of a gene. Because of that, the cosine similarity index is necessary for the analysis, as it is an indirect measure of the number of links that change the sign of their correlation. This way, a two dimensional measure of the importance of a gene is given with less computational requirements.

6.5. Outlook

In this metaanalisis work, we have managed to develop a multiplex network for *Mycobacterium tuberculosis* in different types of stress and analyzed it with some advantages over other methods. However, many issues remain to be adressed.

First of all, the use of an appropiate set of data is vital for gene coexpression analysis [9]. A tool for analysing the quality of the data obtained from the database could be of a great importance for this kind of metaanalysis studies, even though it would still exist dependance on how good GEO annotations are.

Another topic that hasn't been issued in this work is the use of different strains in the transcriptomic studies. For this work, no strain restriction has been made, but some strains are overrepresented for some kinds of stress, and that could induce to incorrect correlations not due to stress but to a certain strain. However, available data wasn't enough for adressing this issue.

The main problem of our analysis has been the high density of interactions of our network. However, we have shown this is not due to our method, so we think it might be due to the object of our study. Other studies performed on *M. tuberculosis* [9]don't show this problem, but there are some difference between our procedures. In other studies, networks have been depeloped with modules and not with genes, and a soft thresholding has been performed to make the network fit to a scale-free model. Also, they have generated their own data for some conditions, and they have analyzed the global network. It may happen in our network that the conditions considered for each type of stress are so similar between them that a high correlation is found for most of cases, and so a high density of interactions appear. In the global network [9], all stresses are studied together, so that correlations are usually smaller than in our work. A possible way of solving this could be using a power beta on the strength of the links to fit our network to a scale-free model after using our differential thresholding.

Another way of looking at how good our method is could be using the methods developed by Yu et al to evaluate all main data analysis methods to date [17]. The problem with that is that they use one measure while we are using two, so some kind of adaptation should be developed. Also, the same datasets of their work can be used for the same purposes of looking for biological significance of the results with a small network instead of the dense network we have developed here. Also, using their DCp method with our differential thresholding could be interesting to test if there is any improvement and/or any correlation between both of our metrics and their only result.

Further studies should include modularity studies, to search function-enriched modules, but also some studies of biological relevance can be performed, using relevant sets of genes, such as pathogenesis-related genes reviewed by Smith [3], or the results of this network can be compared with previously predicted regulatory networks, such as the one developed by Turkarslan et al [14], to check whether its findings can be confirmed with our method or not.

7. Conclusions

1. We have developed a method to adress differential co-expression that uses only data with a significant change in correlation, avoiding loss of significant data as well as reducing computational requirements.
2. Our method generates a bidimensional result that allows for identification of genes with alternate mechanisms.
3. Results obtained from the analysis of *Mycobacterium tuberculosis* data are coherent with already published literature, showing the goodness of our method.
4. Further analysis should be performed to compare this method with already developed ones [17].
5. Combination of this method with techniques from other methods can be useful for the obtention of scale-free model fitting networks.

1. Hemos desarrollado un método para coexpresión diferencial que emplea solamente los datos con un cambio significativo en correlación, evitando la pérdida de datos significativos y reduciendo los costes computacionales.
2. Nuestro método genera un resultado bidimensional que permite identificar genes con mecanismos alternativos
3. Del análisis de los datos de Mycobacterium tuberculosis se obtienen resultados coherentes con la literatura actual, mostrando la bondad de nuestro método.
4. Para comparar este método con los métodos anteriores, más análisis son necesarios [17].
5. La combinación de este método con técnicas de otros métodos puede servir para obtener redes que se ajusten al modelo de redes libres de escala.

8. Bibliography

1. World Health Organization (WHO). *Global tuberculosis report 2015*. (2015).

2. Ernst, J. D. The immunological life cycle of tuberculosis. *Nat Rev Immunol* **12,** 581–591 (2012).

3. Smith, I. Mycobacterium tuberculosis Pathogenesis and Molecular Determinants of Virulence. *Clin. Microbiol. Rev.* **16,** 463–496 (2003).

4. Philips, J. a. & Ernst, J. D. Tuberculosis Pathogenesis and Immunity. *Annu. Rev. Pathol. Mech. Dis.* **7,** 353–384 (2012).

5. Mangtani, P. *et al.* Protection by BCG vaccine against tuberculosis: A systematic review of randomized controlled trials. *Clin. Infect. Dis.* **58,** 470–480 (2014).

6. Au-Yeung, C. *et al.* Tuberculosis mortality in HIV-infected individuals: A cross-national systematic assessment. *Clin. Epidemiol.* **3,** 21–29 (2011).

7. Wang, Y. *et al.* Global Protein−Protein Interaction Network in the Human Pathogen Mycobacterium tuberculosis H37Rv. *J. Proteome Res.* **9,** 6665–6677 (2010).

8. Engle, L. J., Simpson, C. L. & Landers, J. E. Using high-throughput SNP technologies to study cancer. *Oncogene* **25,** 1594–1601

9. Jiang, J. *et al.* Construction and application of a co-expression network in Mycobacterium tuberculosis. *Sci. Rep.* **6,** 28422 (2016).

10. Battiston, F., Nicosia, V. & Latora, V. Structural measures for multiplex networks. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* **89,** 1–14 (2014).

11. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. & Hwang, D. U. Complex networks: Structure and dynamics. *Phys. Rep.* **424,** 175–308 (2006).

12. Strogatz, S. H. Exploring complex networks. *Nature* **410,** 268–276 (2001).

13. Cid, F. Characterization of context-specific networks of protein-protein interactions in Mycobacterium tuberculosis. (2014).

14. Turkarslan, S. *et al.* A comprehensive map of genome-wide gene regulation in Mycobacterium tuberculosis. *Sci. Data* **2,** 150010 (2015).

15. Fukushima, A. DiffCorr: An R package to analyze and visualize differential correlations in biological networks. *Gene* **518,** 209–214 (2013).

16. Prieto, C., Risueño, A., Fontanillo, C. & De Las Rivas, J. Human gene coexpression landscape: Confident network derived from tissue transcriptomic profiles. *PLoS One* **3,** (2008).

17. Yu, H. *et al.* Link-based quantitative methods to identify differentially coexpressed genes and gene pairs. *BMC Bioinformatics* **12,** 315 (2011).

18. Zhang, B. & Horvath, S. A General Framework for Weighted Gene Co-Expression Network Analysis. *Stat. Appl. Genet. Mol. Biol.* **4,** (2005).

19. Newman, M. E. J. The structure and function of complex networks. *SIAM Rev.* **45(2),** 167–256 (2003).

20. Zhou, Q., Hong, Y., Zhan, Q., Shen, Y. & Liu, Z. Role for Krüppel-Like Factor 4 in Determining the Outcome of p53 Response to DNA Damage. *Cancer Res.* **69,** 8284 LP-8292 (2009).

21. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30,** 207–10 (2002).

22. Fieller, E. C., Hartley, H. O. & Pearson, E. S. Trust Tests for Rank Correlation Coefficients. I. *Biometrika* **44,** 470–481 (1957).

23. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 289–300 (1995).

24 . Abdi, H. The Bonferonni and Šidák Corrections for Multiple Comparisons. *Encycl. Meas. Stat.* **1,** 103–107 (2007).

25. Guimera, R. & Nunes Amaral, L. A. Functional cartography of complex metabolic networks. *Nature* **433,** 895–900 (2005).

26. Lew, J. M., Kapopoulou, A., Jones, L. M. & Cole, S. T. TubercuList--10 years after. *Tuberculosis (Edinb).* **91,** 1–7 (2011).

27. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37,** 1–13 (2009).

28. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4,** 44–57 (2008).

29. Mukhopadhyay, S. & Balaji, K. N. The PE and PPE proteins of Mycobacterium tuberculosis. *Tuberculosis* **91,** 441–447 (2011).

30. Krisko, A., Copic, T., Gabaldón, T., Lehner, B. & Supek, F. Inferring gene function from evolutionary change in signatures of translation efficiency. *Genome Biol.* **15,** R44 (2014).

31. Cook, G. M., Hards, K., Vilchèze, C., Hartman, T. & Berney, M. Energetics of Respiration and Oxidative Phosphorylation in Mycobacteria. *Microbiol. Spectr.* **2,** (2014).

32. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9,** 90–95 (2007).

33. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9,** 1–13 (2008).