**Subtracting regulatory genes in *Mycobacterium tuberculosis* based on three network inference methods**

Jeroen Lodewijk - 930 203 525 030

Thierry Haddad - 911 031 296 110

# Introduction

Tuberculosis (TB), a widespread infectious disease caused by the pathogenic host *Mycobacterium tuberculosis*, is affecting millions of people world-wide in either its latent or active state. A majority of the infections and deaths related to TB occur in developing countries, where aid can be scarce and infected people, especially with latent TB, hard to track. Solidifying our understanding of active and latent TB is required in order to save people from TB. Therefore, in this concise research we attempt to identify genes that show characteristics that could prove fruitful for further research, by making use of regulatory network inference methods on a TB dataset aggregated by Van Dam *et al*. [1]. By making use of methods that infer gene relations based on statistical, context-related significance, we hope to find genes that appear to be co-regulated and could form a basis for further research on the detection and identification of active and latent tuberculosis.

# Materials & methods

The dataset was being processed using R through various steps. These steps included: data processing, replacing NAs with KNN, plotting, PCA and network inference. NAs were replaced using a k-nearest neighbour solution (neighbours = 3). Afterwards, networks were visualized in Cytoscape and analysed for GO-terms with BiNGO. With BiNGO, we used all three GO fields to obtain a side-by-side comparison of the results for all three network methods.

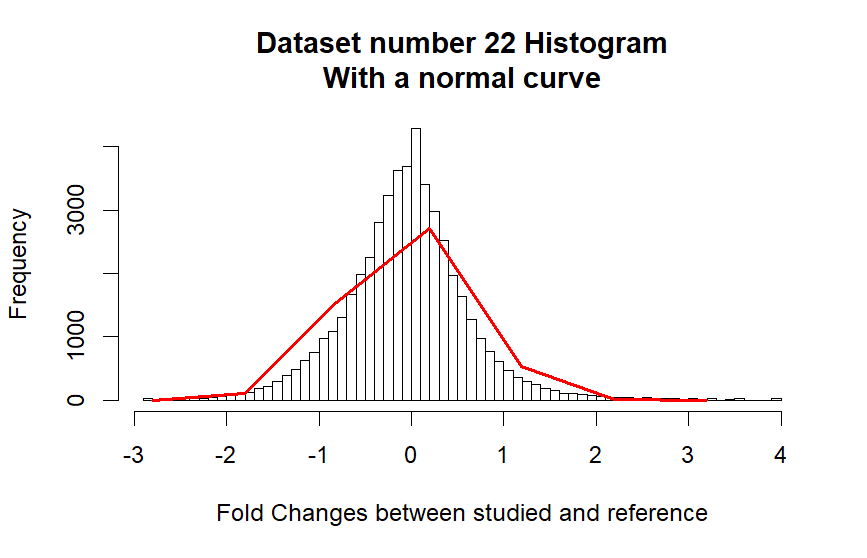
During the creation of the inference network of the expression dataset three methods were used. These are Maximum Relevance/minimum redundancy Network (MRnet), Context Likelihood of Relatedness (CLR) and Accurate Reconstruction of Accurate Cellular Networks (ARACNE). In short, MRnet creates a mutual information (MI) matrix for which it will infer a network based on maximum relevance/minimum redundancy feature selection [1], CLR also creates a MI matrix, but will additionally perform a network context check for each MI-value to identify false-positive correlations [2] and lastly ARACNE also creates a MI matrix, but makes use of the data inequality paradigm to detect indirect relations and removes these [3].

Initially it was decided to not use any threshold value, but this lead into a lot of nodes and edges. The sheer number of edges and nodes led into issues with finding GO terms using Bingo. Therefore it was decided to put the threshold value at 0.25.

# Results

**Data Validation**

The first step is checking the data for any missing values. This is done using R version 3.5.0. R managed to count eight missing values over two genes. These genes were fixed using a K-nearest-neighbour imputation implementation.

****

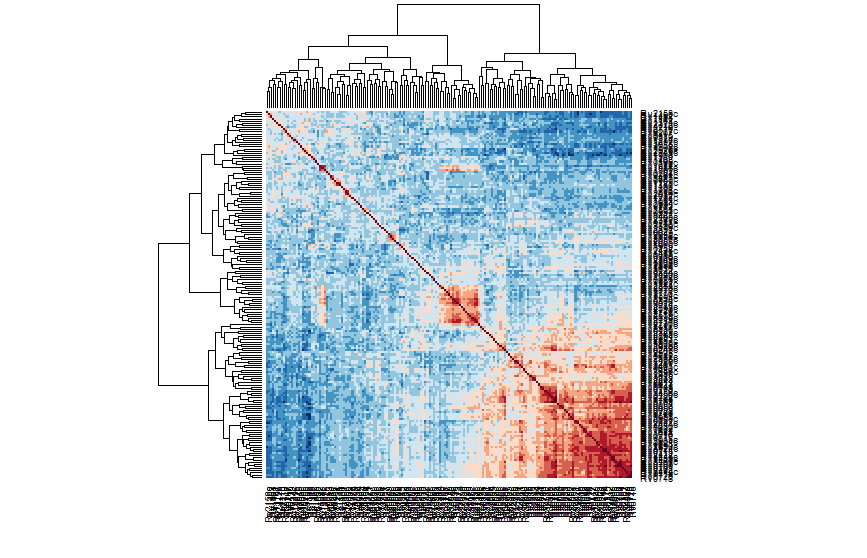
*Figure 1: Shows the data being plotted as a histogram, the x-axis contains the fold change*

*between studied condition and reference condition. The red line is the normal curve based on*

*the same data.*

Next step in the data validation was checking if the data is being normalized and normally distributed. This was done by creating a histogram, within it a line was fitted to follow the distribution. It appears that it can be assumed that the data is log2-transformation normalized and reasonably normally distributed.

**Heatmap correlation**

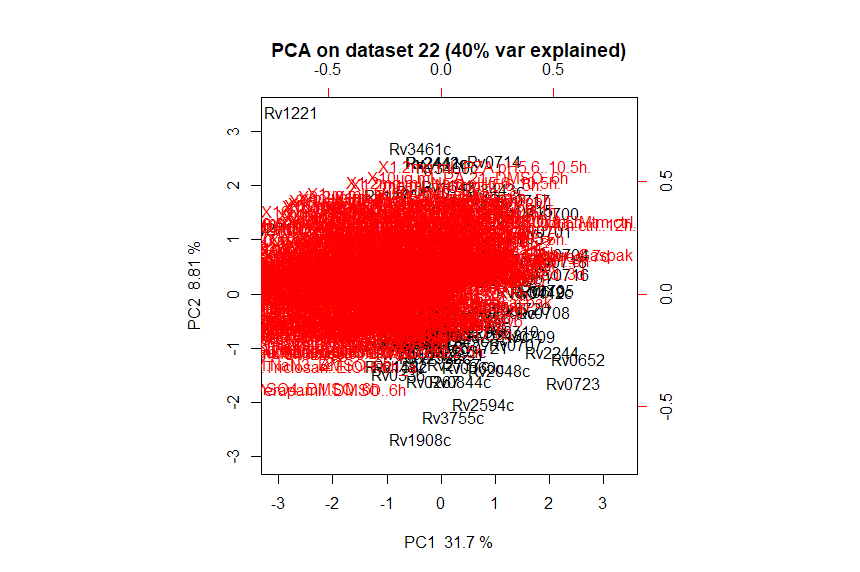
****

*Figure 2: Heatmap that shows the correlation between the different genes. Colors range from blue (negative correlation) to red (positive correlation). Heatmap also has a dendrogram.*

Heatmap (figure 2) shows that there is correlation between different genes. It appears that there are a few similar correlation clusters within the heatmap, like Rv1163 and Rv1164. Genes that shows some correlation were taken in consideration for gene ontology step.

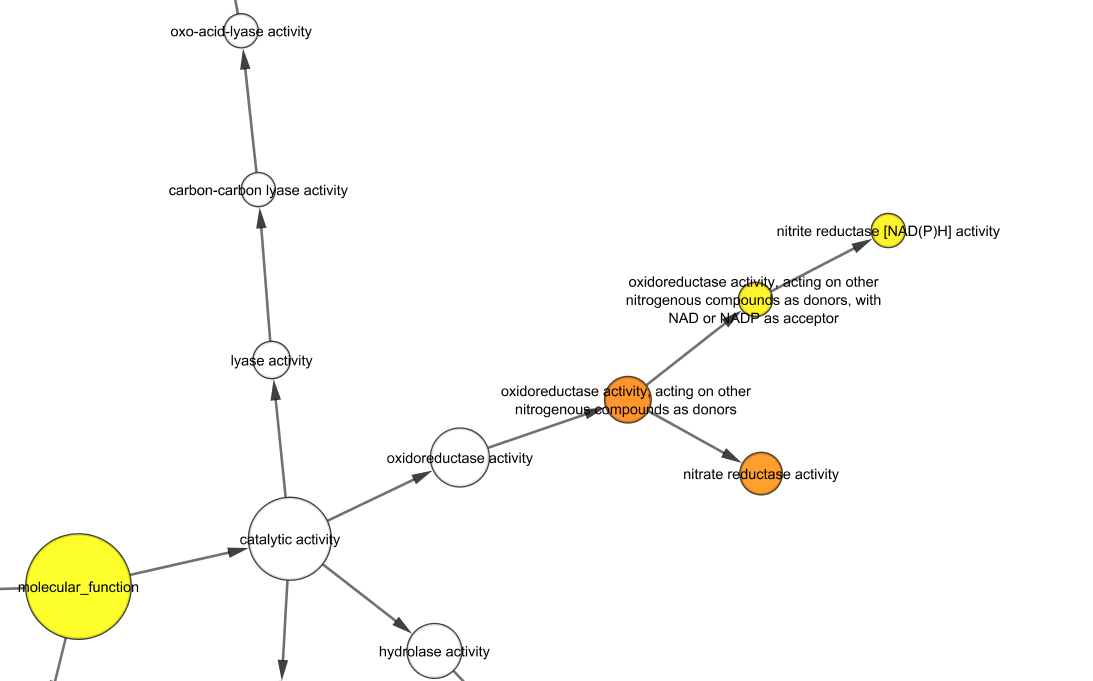
**Principal components analysis**

A principal component analysis (PCA) was performed on the log2 fold-change RNA-seq dataset. However, as can be seen in Figure 3, the PCA contains little explanative value due to the nature of the dataset. In typical use-cases the PCA is performed on datasets with fewer conditions or groups, to attempt to visualize the difference between and inside groups. Unfortunately, our dataset has a large amount of conditions (p=287) and lacks a reference-condition distinction, as we are working with the log2-transformed dataset, not the raw expression values for reference and conditions. This makes defining distinct groups in the PCA hard and the PCA too cluttered to be informative.

****

*Figure 3: Displays the results of the PCA biplot. The words colored in red are the conditions and the black colored words are the gene names. Each red line indicated its own principal component in the biplot.*

**Network**



*Figure 4: Showing a Bingo results of the CLR network. A part of the network is shown to highlight nitrite and nitrate related genes.*

Bingo is used to perform gene ontology on the all the inferred networks (figure 4). In the CLR inference network Bingo results a set of genes are found related to nitrate assimilation (table 1). These genes occur only in tuberculosis Mycobacterium species, non-tuberculosis Mycobacterium lack these genes.

| **Compound** | **Genes found** |
| --- | --- |
| Nitrate | RV1737C, RV1161, RV1162, RV1164, RV1163, RV0252, RV0253, RV1736C |
| Nitrite | RV0253 RV0252 |

*Table 1: The two compounds the GO-terms relate to and their respective genes. Note the overlap in Nitrite and Nitrate genes, which can be explained by the fact that Nitrate is reduced to Nitrite by Mycobacterium tuberculosis.*

# Conclusion and discussion

Initial data visualization shows that the data is likely log2-transformed and normally distributed (figure 1). The correlations between different genes was checked using heatmap (figure 2). This heatmap did visualize the correlation between genes, yet the labeling proved to be hard to read. So it was decided to take a few genes that showed some form of correlation with each other and examine their Bingo results. This method did not prove to be useful, since genes that belong to a particular metabolic process (nitrate assimilation) showed various levels of correlation with each other (0.05 to 0.48). The PCA proved to be hard to interpret due to the large amount of conditions.

The dataset contains a set of genes that are related to handling of nitrate and nitrite (Table 1), two compounds used by *Mycobacterium tuberculosis*. It is speculated these aid the organism in survival in low-oxygen environments [5], where Nitrate reduction changes Nitrate to Nitrite, of which the intake is regulated by NarK2 which was also found using Mycobrowser [6] when searching for RV1737C. When transforming the list of genes to a non-redundant set, we obtain 'RV1737C', 'RV1161', 'RV1162', 'RV1164', 'RV1163', 'RV0252', 'RV0253', and 'RV1736C'. These genes only occur in Mycobacterium species that cause tuberculosis. Therefore it is expected that these genes were found using Bingo, as they appear to be differentially expressed in reference vs condition and form a regulated network.   
Currently only CLR was able to detect these genes. This might be caused by setting a too high of a threshold during the network inference steps, but might also be caused by the difference in methodology these network inference methods adhere to. A potential solution for this issue could be using a more relaxed threshold. Unfortunately Bingo proved prone to crashing with more relaxed thresholds, which initially lead us to using our current threshold of 0.25. Additionally, what could be a useful endeavour would be adding non-tuberculosis Mycobacterium species in the dataset, thereby giving a contrast between reference, *Mycobacterium tuberculosis* and other Mycobacterium species. A last recommendation might be to increase the overall size of the dataset to somewhat lessen the statistical problems a small-n, large-p dataset might induce.

# References

1: van Dam, J. C., Schaap, P. J., dos Santos, V. A. M., & Suárez-Diez, M. (2014). Integration of heterogeneous molecular networks to unravel gene-regulation in Mycobacterium tuberculosis. *BMC systems biology*, *8*(1), 111.

2: Meyer, P. E., Kontos, K., Lafitte, F., & Bontempi, G. (2007). Information-theoretic inference of large transcriptional regulatory networks. *EURASIP journal on bioinformatics and systems biology*, *2007*, 8-8.

3: Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., ... & Gardner, T. S. (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS biology*, *5*(1), e8.

4: Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., & Califano, A. (2006, March). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In *BMC bioinformatics* (Vol. 7, No. 1, p. S7). BioMed Central.

5: Khan, A., & Sarkar, D. (2012). Nitrate reduction pathways in mycobacteria and their implications during latency. *Microbiology*, *158*(2), 301-307.

6: Kapopoulou, A., Lew, J. M., & Cole, S. T. (2011). The MycoBrowser portal: a comprehensive and manually annotated resource for mycobacterial genomes. *Tuberculosis*, *91*(1), 8-13.