UNIVERSITAT POLITÈCNICA DE VALÈNCIA

ESCOLA TÈCNICA SUPERIOR D´ENGINYERIA AGRONÒMICA I DEL MEDI NATURAL

EVOLUTION OF TRANSCRIPTIONAL REGULATORY REGIONS IN THE MYCOBACTERIUM TUBERCULOSIS COMPLEX

TRABAJO FIN DE GRADO EN BIOTECNOLOGÍA

ALUMNO/A: MIGUEL ANGEL MORENO MOLINA

TUTOR/A: Mª ANTONIA FERRÚS PÉREZ
COTUTOR: IÑAKI COMAS ESPADAS

Curso Académico: 2015/2015

VALENCIA, 7 DE JULIO DE 2015

Título: "Evolución de las regions reguladoras de la transcripción en el complejo de Mycobacterium Tuberculosis"

Resumen: La bacteria de la tuberculosis mata a más de un millón y medio de personas al año y representa un problema de salud global. Las técnicas genómicas están permitiendo entenderla mejor de tal manera que la cantidad de datos genómicos, transcriptómicos, proteómicos y metabolómicos ha ido aumentando. Puesto que es una enfermedad global, la bacteria que la causa es también heterogénea. Recientemente, en el laboratorio se ha caracterizado a nivel genómico una colección global de cepas de tuberculosis. Esto nos ha dado un arsenal de polimorfismos presentes en las cepas circulantes de todo el mundo que pueden tener un impacto en el fenotipo (diferencias en expresión génica y proteica, epidemiología, presentación clínica). Particularmente muchos de esos polimorfismos afectan zonas reguladoras de la transcipción. Para empezar a entender ese impacto se propone este trabajo de investigación. El hecho de que dichas zonas estén o no conservadas a nivel global puede indicar su importancia para la virulencia y el desarrollo de la enfermedad. En este trabajo se plantea estudiar dichas regiones. Para ello, se han identificado diferentes elementos reguladores (sitios de unión de factores de transcripción, regiones de inicio de la transcripción y la traducción así como ARNs pequeños). Usando la información de mutaciones acumuladas en nuestra muestra genómica global clasificaremos dichas regiones como esenciales (conservadas evolutivamente) y no esenciales.

Title: 'Evolution of transcriptional regulatory regions in the Mycobacterium Tuberculosis Complex'

Summary: *Mycobacterium tuberculosis* is a bacterium that kills more than one and a half million people each year, constituting a global health issue. Modern genomic techniques are allowing to understand it better, so the amount of genomic, transcriptomic, proteomic and metabolomic data is constantly rising. As it is a global disease, bacteria which cause it are heterogeneous. Recently, a global collection of *M. tuberculosis* strains has been characterized in the lab, providing us with a vast set of polymorphisms present in the different strains which could have a great impact on their phenotype and clinical presentation or epidemiology. Particularly lots of these polymorphisms affect transcriptional regulatory regions in the genome. This work's goal is to begin understanding that impact studying those regions. For that purpose, different regulatory regions have been identified and using the information stacked in the global sample of strains, we can classify them as essential (evolutionarily conserved) or nonessential.

Keywords: tuberculosis, evolution, transcriptional regulation, polymorphism, bioinformatics.

Student: Miguel Ángel Moreno Molina
Supervisor: Mª Antonia Ferrús Pérez
Co-supervisor: Iñaki Comas Espadas

Valencia, July 7th, 2015

INDEX

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Tuberculosis remains one of the world's deadliest diseases of today, caused by a genetically related group of bacteria called the *Mycobacterium tuberculosis* complex. According to data from the World Health Organization (WHO), over 9 million people developed tuberculosis in 2013, and 1.5 million of them died from it. Nevertheless, this disease is preventable and fortunately can be treated as well, so a great effort is being made in order to fully understand the bacterial molecular mechanisms which cause it and allow *Mycobacterium tuberculosis* to persist in the host. The bacteria usually establish a long asymptomatic infection and evolve to the active disease only in some hosts, adapting to a greatly changing environment. This has frustrated the effective control of the disease by needing long complex treatments and inefficient vaccines. The main challenge when approaching these problems is the fact that the bacterium is very heterogeneous, thus a huge collection of genomic, transcriptomic, proteomic and metabolomic data coming from modern techniques needs to be analyzed before actually developing effective treatments for the disease.

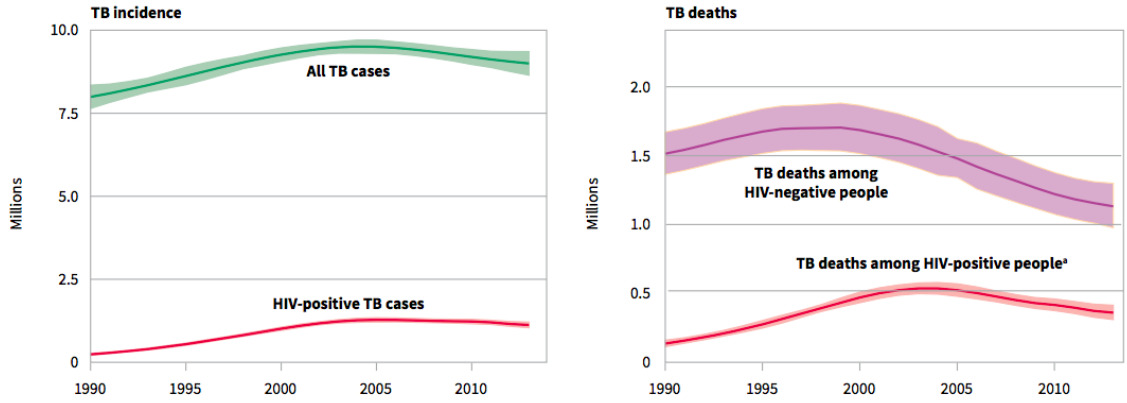

Figure 1. Estimated absolute numbers of TB cases and deaths (in millions per year), 1990–2013. This figure illustrates the growing nomber of cases and the decreasing number of deaths associated to research progress (extracted from the Global Tuberculosis Report, WHO 2104).

1.1. *M. tuberculosis* evolution, resistances and survival

Evolution through genetic variability has produced seven different *M. tuberculosis* lineages from its origin 70 thousand years ago, distributed today across the planet. It has been proved that this evolution is associated to ancient human migrations and their changes in population, concluding that modern lineages of the bacteria have a potential higher virulence in the sense of shorter latency times (Comas *et al*, 2013). These lineages have evolved from adaptation to high host densities, while older lineages show lower virulence evolved from low host densities. More than 34,000 SNPs have been identified when analyzing diversity across lineages and strains, used to construct phylogenetic relationships between them. Genetic diversity among the lineages of *M. tuberculosis* has some important clinical and epidemiological considerations, as it rises differences in the way disease progresses and the host response to it.

M. tuberculosis has high survival and resistant capabilities that difficult the global control of the disease. Susceptible *M. tuberculosis* can be treated today with strict antibiotic therapies in two phases: an intensive two-month phase using isoniazid, rifampicin, pyracinamide and ethambutol; and a continuation four-month phase with just two of the later (Espinal *et al*, 2000). Still, resistance to those treatments increases over time and some strains require the use of second-line drugs which are more toxic and less effective. The mechanisms by which resistance develops and fixates in *M. tuberculosis* are not well understood yet, and involve a number of strategies used by mycobacteria to adapt and evolve (Fonseca *et al*, 2015).

Moreover, *M. tuberculosis* ability to survive and adapt to changing microenvironments inside the host is surprisingly strong. Upon infection by aerosol spreading, the bacteria are fagocytosed by macrophages in the lung and able to replicate inside these cells by inhibiting the fusion between the phagosome and bactericidal lysosomes. Infection can persist for long periods of time without activating, surviving the dynamic host environment through a set of resources. Some of their challenges are reactive oxygen and nitrogen stress, or pronounced shifts in oxygen availability (Bartek *et al*, 2014). Figure 2 illustrates how the bacteria can adapt to anaerobiosis within a short period of time.

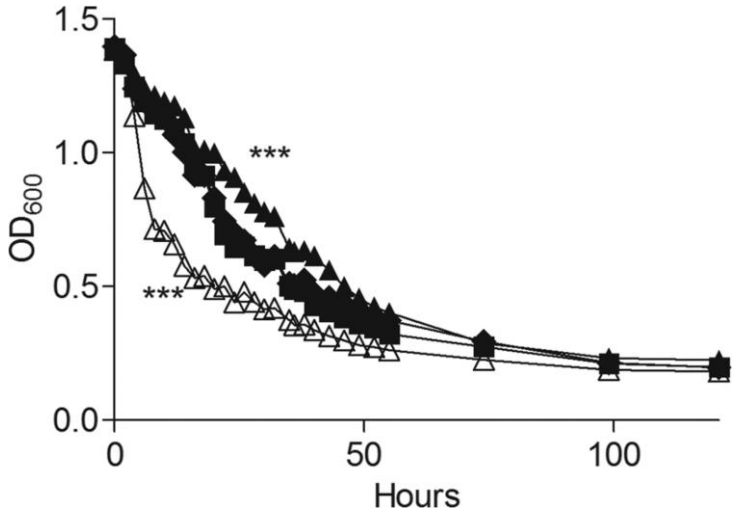

Figure 2. Tuberculosis oxygen consumption during adaptation to anaerobiosis as measured by methylene blue decolorization. It shows how bacteria shift their metabolism and adapt to hypoxia through a set of transcriptional regulations within a few days (extracted from Bartek *et al*, 2014).

These properties of *M. tuberculosis*, its latent persistence in the host combined with the easy transmission make it so difficult to control and eradicate worldwide.

1.2. Research and data collection

Collecting as much data as possible is crucial to understand *M. tuberculosis*. Studying the mechanisms by which it survives in the host and gathering a list of essential genes for the bacterium were the first actions that needed to be done.

After a century of research with rising drug resistances, studies first tried to identify genes essentially required for bacterial growth (Sassetti *et al*, 2003). The main goal was to find new targets for antimycobacterial agents through the use of transposon site

hybridization, so those genes which could not sustain a transposon insertion were considered important for growth. This first set of essential genes was classified by gene function as shown in Table 1. At the same time, experiments in vivo (Sassetti and Rubin, 2003) managed to narrow a list of 194 genes essential for growth.

Years later, the advent of RNA sequencing allowed the discovery of a considerable amount of non-coding RNA in the transcriptome of *M. tuberculosis* (Arnvig et al, 2011). It was noted that a lot of sequence reads mapped intergenic regions over annotated genes, and those sRNAs accumulated up to a high degree in the lungs of infected mice in the lab, suggesting its relation with pathogenesis. Also, a potential post-transcriptional regulatory network was thought to control the adaptation responses of the bacteria.

Table 1. Essential pathways in *M. tuberculosis* and *M. bovis*. Predicted essential genes play a central key role in metabolism. Essential steps are those that are performed by genes identified as being required for optimal growth (extracted from Sassetti *et al*, 2003).

Pathway	Essential steps/total steps
Alanine biosynthesis	2/2
Arginine biosynthesis	9/9
Asparagine biosynthesis	1/1
Aspartate biosynthesis	2/2
Chorismate biosynthesis	5/6
Cysteine biosynthesis	2/2
dTDP-rhamnose biosynthesis	3/4
Folate biosynthesis	5/9
Glutamate degradation	2/2
Glycine degradation	3/3
Haeme biosynthesis	7/10
Histidine biosynthesis	6/6
Homoserine biosynthesis	3/3
Homoserine/Methionine biosynthesis	2/3
Isoleucine biosynthesis	3/4
Leucine biosynthesis	3/4
Lysine and diaminopimelate synthesis	6/7
Mannose and GDP-mannose metabolism	2/3
Non-oxidative pentose phosphate pathway	3/4
Panthothenate and coenzyme A biosynthesis	3/5
Peptidoglycan synthesis	7/10
Proline biosynthesis	2/3
Proline utilization	2/2
PRPP biosynthesis	3/3
Purine biosynthesis	10/13
Pyridoxal 5′phosphate biosynthesis	2/2
Pyrimidine biosynthesis	3/6
Riboflavin, FMN and FAD biosynthesis	3/3
Serine biosynthesis	2/2
Sulphur-containing amino acid metabolism	3/4
Thiamine biosynthesis	4/4
Threonine biosynthesis from homoserine	2/2
Trehalose anabolism	3/3
Tryptophan biosynthesis	4/4
Valine biosynthesis	2/3

Further studies employing new methods of global phenotypic profiling carried on characterizing essential genes in *M. tuberculosis* (Griffin *et al*, 2011), consisting on high-density mutagenesis and deep sequencing of libraries of complex mutants under different conditions and pointing out the importance of sterol catabolic functions related with survival within the host. This survival, achieved by alternating between replicating and non-replicating states, was further studied through the architecture and expression of promoters (Cortes *et al*, 2013). Over 4,000 transcription start sites were identified, many of them overlapping with start codons and revealing that a part of the transcriptome in *M. tuberculosis* was leaderless. It was also noted that the leaderless transcripts did not match essential genes, but these accumulated in models of starvation *in vitro*, indicating a possible importance related to the non-replicating state of the bacteria.

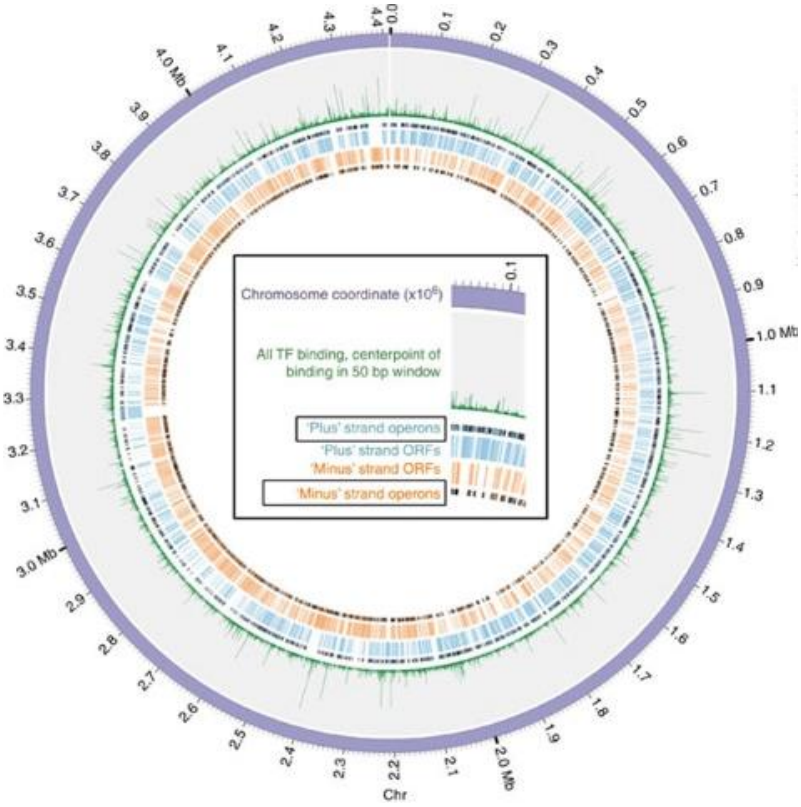

Figure 3. TF-binding sites identified by ChIP-seq. The 4.4-Mb H37Rv strain chromosome is divided into nonoverlapping 50-bp windows, and green spikes represent the total number of TF-binding events within each window (extracted from Minch *et al*, 2015).

Other projects analyzing the bacterial adaptations to hypoxia managed to map transcription factor binding sites by ChIP-Seq experiments, establishing an initial reconstruction of the transcriptional regulatory network of *M. tuberculosis* with 50 TFs (Galagan *et al*, 2013). Shortly after, this network was refined and extended to include 154 TFs (Figure 3), identifying new binding events and several binding motifs and so providing a huge set of possible binding regions to analyze (Minch *et al*, 2015).

All the previous data collected over time has helped to answer many questions and test hypothesis, and its global analysis as a whole could lead to important advances in the fight against tuberculosis.

1.3. Essential and nonessential genes in *Mycobacterium tuberculosis*

Essential genes are the ones thought to be critically important for the growth and survival of an organism. Identifying essential genes helps to understand the physiology of a species, which in turn allows for the proposal of new drug targets (Grazziotin *et al*, 2015). However, the essentiality of a gene is relative to the experimental conditions and the measured output variable like growth or colony formation *in vitro* (Fang *et al*, 2005). Here, the concept of nonessential genes can be introduced, understood as those genes which are not absolutely vital for survival under certain conditions.

In bacteria, essential genes seem to be more conserved than nonessential ones (Jordan et al, 2002), and the rate of evolution of these essential genes has traditionally been thought to be slower than the nonessential one. Analyses of evolutionary conservation have demonstrated that essential genes are generally conserved among bacteria compared to nonessential genes (Xiaodong *et al*, 2007).

In the particular case of *Mycobacterium tuberculosis*, conservation of its genes has been addressed and as expected, essential genes were more conserved than nonessential ones through sequence comparative analyses (Comas et al, 2010). The average nucleotide diversity (π) for both sets of regions was calculated in that study, illustrating that nonessential genes accumulated more polymorphisms in average than essential ones (Figure 4). This parameter measures genic variation, in other words, the degree of polymorphism of a population (Nei and Li, 1979), in this case a population of sequences. It can be seen that essential gene regions have lower diversity in average and maximum range than nonessential gene regions.

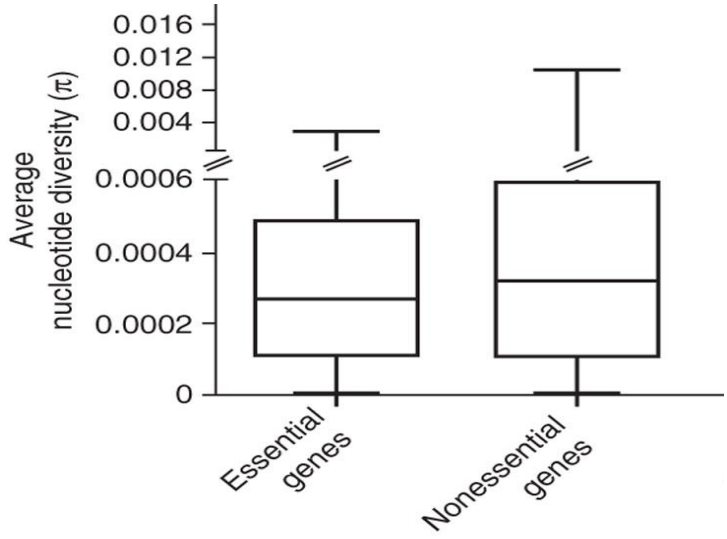

Figure 4. Average gene-by-gene nucleotide diversity across essential and nonessential genes. Box plot indicates median (horizontal line), interquartile range (box) and minimum and maximum values (whiskers) (extracted from Comas et al, 2010).

1.4. Analyzing transcription regulation: a systems biology approach

Transcriptional regulation is the way cells control their gene activity by adjusting the rate of transcription. This control allows the cell to respond to signals and elaborate a response, and it is mediated by certain genomic regions, transcription factors and other proteins. The comprehension of this complex control mechanisms requires a systems biology approach. The understanding of a cell as something more than the sum of its parts calls for the integration of high-throughput omics data to build and refine predictive models of dynamics and interactions of cellular components (Aderem *et al*, 2011). In M. tuberculosis, the pathogen-host system is especially important to model due to the complex relationships established by the bacteria and the immune system. Proteins that interact with other small molecules messengers, other proteins and DNA mediate *M. tuberculosis* adaptation inside the host. These form the bacterial transcriptional landscape by converting stimuli into responses in a coordinated fashion (Minch *et al*, 2015). However, genetic diversity is something to consider when constructing models, as it affects the outcome of predictions to search for global solutions to a disease. In this context, analyzing the transcription regulatory regions of *M. tuberculosis* means one more step towards the iterative refining of a model which goal is to identify potential new therapeutic targets.

Recently, a global collection of *M. tuberculosis* strains has been characterized in the lab (Comas *et al*, 2013), providing us with a set of more than 34,000 polymorphisms that can influence the bacterial phenotype through differences in gene and protein expression, epidemiology or clinical presentation. A high number of these polymorphisms particularly affect genomic regions which regulate transcription, and the fact that the regions are conserved or not could indicate its importance for the virulence and development of the disease. Analyzing the genetic diversity of all regions that regulate transcription across the strain collection may shed some light on the understanding of how significant their functions are for the bacteria and how evolution has affected and selected these regions among the different strains.

2. OBJECTIVE

1. To calculate the genetic diversity of essential and nonessential gene regions from a global genetic sample of 216 *Mycobacterium tuberculosis* strains as proof of concept. The values obtained will serve as reference to classify other regions as conserved (values similar to essential genes) or not conserved (values similar to nonessential genes).
2. To calculate the genetic diversity of five transcription regulatory regions from the same sample of sequences:
    a. 16S RNA
    b. Antisense RNA (asRNA)
    c. Noncoding RNA (ncRNA)
    d. Transcription factor binding sites (TFBS)
    e. tRNAs
3. To compare the former genetic diversity values to the reference ones and classify every transcription regulatory region as conserved or non-conserved based on statistical tests.

3. MATERIALS AND METHODS

3.1. Data sources

The first data source was the open tuberculosis database TubercuList (tuberculist.epfl.ch; Lew et al, 2011), from where data for 16S, ncRNA and tRNA regions was extracted. For this purpose, an advanced search was performed in the database looking for the entire annotated genome of the bacterium. The website generated a list of every annotated gene and region which was downloaded as a spreadsheet. This file containing the genomic position of every region was filtered with a custom Python script to select and keep the three sets of regions mentioned before, resulting in three individual files with the format 'start – end – annotation'.

The second data source was the MTB Network Portal database (networks.systemsbiology.net/mtb), from where data for the transcription factor binding regions was extracted. Browsing to the 'Networks' section, 'ChIP-Seq binding Dataset', a table containing the TF genomic binding locations was downloaded, and again processed with a custom Python script to achieve the format start – end – annotation. Regions selected were ±15 nucleotides from the center of the ChIP-Seq peak.

The third data source was bibliography (Arnvig *et al*, 2011; Arnvig *et al*, 2014), where experimental data for asRNA and complementary data for ncRNA regions was extracted and given the same format than the rest of the regions. Finally, the list of essential and nonessential genes was provided by Dr. I. Comas from his previous research, as well as the multiple alignment file of *M. tuberculosis* strains and a reference sequence of a predicted ancestor to those strains (Comas *et al*, 2013).

3.2. Extraction of regions

From the multiple sequence alignment file containing the genomes of 216 *Mycobacterium tuberculosis* strains, four different region datasets were extracted (Figure 5) using a combination of custom Python scripts and the 'extractalign' Emboss script (Rice *et al*, 2000). The datasets consisted on:

- Dataset 1: subalignments of each type of region, one file per region.
- Dataset 2: subalignments of each type of region, one file per region and containing an extra reference sequence from the ancestor.
- Dataset 3: concatenated subalignment of all regions and the ancestor reference sequence, one file per type of region.
- Dataset 4: concatenated subalignment of all regions and the ancestor reference sequence, one file per region.

Figure 5. Example of an essential region for each type of dataset. It can be seen that Dataset 1 contains a region for each strain, Dataset 2 contains an extra last ancestor sequence, Dataset 3 starts with ancestor and then the rest of regions concatenated for all strains, and Dataset 4 has the ancestor sequence and a concatenated for just one strain.

Each dataset was meant to provide a different type of results after analysis. Dataset 1 and 2 were intended to compare the resulting values of genetic diversity with and without the presence of the ancestor reference sequence as outgroup. Dataset 3 was used to find out the global number of mutations across the concatenated of each particular type of region. Dataset 4 paired the ancestor sequence with each separated strain concatenate to calculate its average divergence per site.

3.3. Analyses of genetic diversity

Following extraction, every file in the datasets was converted from FASTA to PHYLIP (interleaved format) by using the 'seqret' Emboss script (Rice *et al*, 2000). After that, datasets were ready to be analyzed with the software VariScan to calculate several genetic diversity population parameters. VariScan is a package for the analysis of DNA sequence polymorphisms at the whole-genome scale that performs sliding window or wavelet-transform based runs to capture information about the number and position of polymorphisms in an alignment (Vilella *et al*, 2005; Hutter *et al*, 2006). A different configuration was used for every dataset in the analyses (Table 2).

Table 2. Summary of relevant configuration parameters for running VariScan with every dataset.

Parameter	Dataset 1	Dataset 2	Dataset 3	Dataset 4
StartPos	1	1	1	1
EndPos	0	0	0	0
RefPos	0	1	1	1
Outgroup	none	last	first	first
RunMode	12	22	22	21
SlidingWindow	1	0	0	0
WidthSW	50	-	-	-

All output files from the VariScan software were loaded into a custom Python script which extracted and calculated the averages of the following diversity parameters from the sequences (some of them for each dataset): number of segregating sites (S), total number of mutations (Eta), nucleotide diversity (π), Waterson's nucleotide diversity per site (Theta), average divergence per site (K), Tajima's D, Fu and Li's D and F, and Fay and Wu's H.

3.4. Statistical tests

All statistical tests were conducted using the software package R (R-project.org). The following tests were applied to data coming from the results of VariScan analyses:

- Wilcoxon rank-sum test: this non-parametric method tests if two sets of samples originate from the same distribution. It is more efficient than a t-test on non-normal distributions. This test was performed in R through the function 'wilcox.test ( )' for testing the independence of essential versus nonessential nucleotide diversity (π) value's populations from Dataset 1.

- Kruskal-Wallis one-way analysis of variance by ranks: this non-parametric method extends the Wilcoxon test to more than two groups with the same purpose. It was performed in R through the function 'kruskal.test ( )' on the whole set of nucleotide diversity (π) values from all regions of Dataset 1 to test their independence.

- Pearson's chi-squared test: this method tests a null hypothesis that the frequency distribution of certain events observed in a sample fits the chi-squared distribution. This evaluated how likely it is that differences between the sets are by chance. It was performed on R through the function 'chisq.test ( )' on several tables with values of segregating and non-segregating sites from Dataset 3 to find out the relationships between them.

4. RESULTS AND DISCUSSION

In this section, results for the genetic diversity of each dataset will be presented along with plots to illustrate them. In addition, interpretation and validation of these results will be discussed and compared to initial hypothesis.

4.1. Dataset 1 and 2 results

Starting with the results from Dataset 1 (Table 3) and Dataset 2 (Table 4), average values for several genetic diversity parameters are shown. These two datasets had the same structure but the second one included the sequence of the ancestor. The goal was to evaluate the impact of the presence of this sequence in the analysis. Taking a quick look at the pi (π) column from both tables, it is seen that there is little difference in the results from including the ancestor sequence, so values from Table 1 were used in the following plots and calculations.

Table 3. Genetic diversity results for Dataset 1, showing the average values for different parameters. Values were calculated averaging all individual values from each type of region.

Region	S	Eta	Eta_E	Pi	Theta	Tajima_D	FuLi_D	FuLi_F
16S	9,000000	9,000000	7,000000	0,000066	0,000984	-2,100872	-4,56664	-4,25686
asRNA	3,947368	3,947368	2,473684	0,000405	0,001329	-1,219222	-2,15631	-2,09963
essential	8,441860	8,441860	5,412145	0,000256	0,001183	-1,605096	-3,21734	-3,04158
ncRNA	1,515152	1,515152	0,969697	0,000646	0,001968	-0,909566	-1,77851	-1,70353
nonessential	7,404422	7,404422	4,765646	0,000299	0,001356	-1,547591	-3,05893	-2,90256
tfactors	0,325310	0,325310	0,204407	0,000486	0,001764	-0,775153	-1,33295	-1,31046
tRNA	0,200000	0,200000	0,133333	0,000038	0,000438	-0,913855	-1,32889	-1,35424

Table 4. Genetic diversity results for Dataset 2, showing the average values for different parameters. Values were calculated averaging all individual values from each type of region.

Region	S	Eta	Eta_E	Pi	FuLi_D	FuLi_F	FayWu_H
16S_out	9,000000	9,000000	7,000000	0,000066	-4,633470	-4,453795	0,101034
asRNA_out	3,684211	3,684211	2,315789	0,000395	-2,220612	-2,230967	0,051090
essential_out	8,001292	8,001292	5,140827	0,000244	-3,236812	-3,152948	0,215645
ncRNA_out	1,363636	1,363636	0,909091	0,000600	-1,764253	-1,757162	0,050406
nonessential_out	6,512245	6,512245	4,194558	0,000266	-3,020020	-2,962339	0,178684
tfactors_out	0,292037	0,292037	0,180741	0,000455	-1,305171	-1,338023	0,009060
tRNA_out	0,200000	0,200000	0,133333	0,000038	-1,333447	-1,416456	0,002840

An initial plot of the nucleotide diversity (π) values of all individual regions from Dataset 1 was made to have a general view of what type of regions had the most diversity (Figure 6). However, to better appreciate these results, a boxplot was built where the differences in diversity ranges were far more clear. Looking at these π values,

nonessential gene regions have a wider box than the essential ones, which means that their range of diversity is higher. 16S and tRNA regions have low diversity, asRNA regions are similar to essential regions, and ncRNA and TFBS regions have higher diversity. In reference to Tajima's D, values of every region type are negative due to a high number of segregating sites being singletons (mutations occurring just in one sequence). A negative Tajimas's D can have several explanations, including purifying selection, presence of deleterious mutations or population expansion. This last one has more sense, since modern lineages of *M. tuberculosis* evolved from adaptation to high host densities (Comas *et al*, 2013).

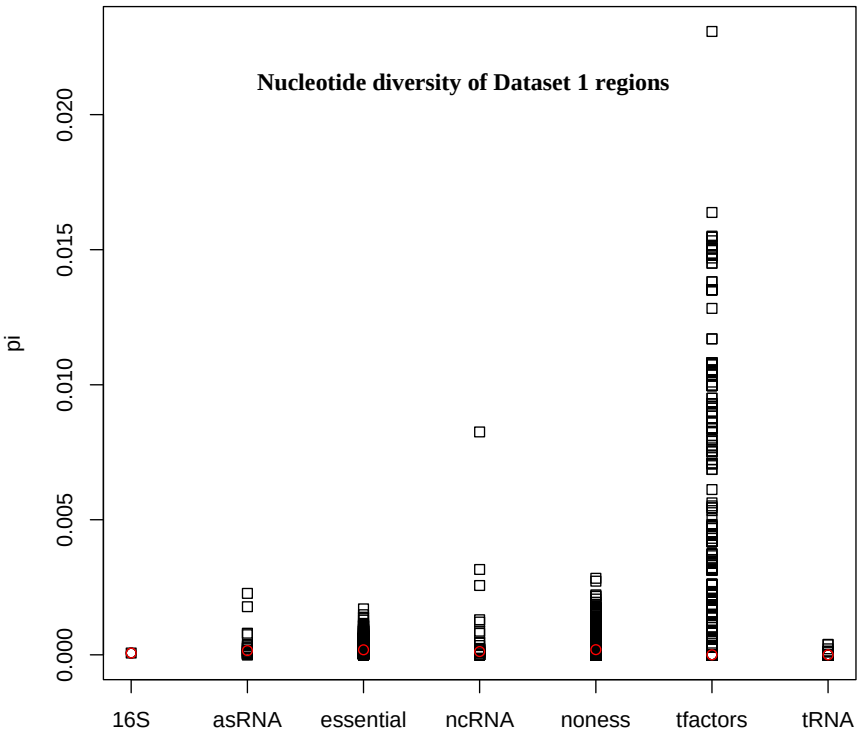

Figure 6. Scatterplot of nucleotide diversity values of every region from Dataset 1. Black squares represent every data point, and red circles the median value of each region subset.

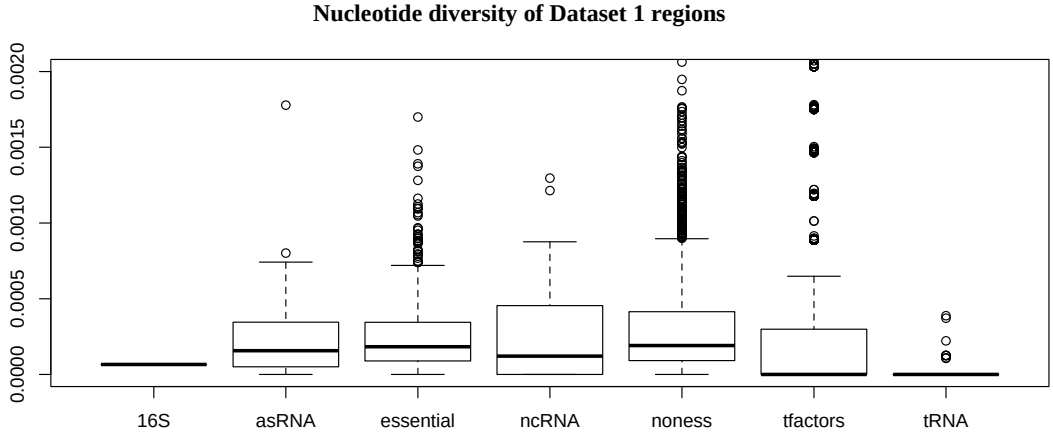

Figure 7. Boxplot of nucleotide diversity values of every region from Dataset 1. Extreme outliers of nonessential and TFBS regions are left out of the plot for better clarity.

The next step was the statistical validation of the differences between the values of nucleotide diversity across the regions. For this purpose, a Kruskal-Wallis test was performed selecting each region type as a different category and with the null hypothesis that all samples come from the same distribution. A pi-value of $2.2x10^{-16}$ was obtained in the test. As it is much lower than 0.05 (accepted α), the null hypothesis is rejected and it is assumed that the differences between region diversity are significant.

Now, in order to validate the statistical difference between essential and nonessential gene regions as proof of concept, those two categories were individually compared in a Wilcoxon test, obtaining a p-value of 0.0271. Again, this value is lower than 0.05 and the difference between both regions is now statistically verified.

4.2. Dataset 3 results

Dataset 3 consisted on a concatenate of all region sequences against the ancestor reference. One file per region type was analyzed in order to get the total number of polymorphisms of that region type (Table 5).

Table 5. Number of segregating sites (polymorphisms) and non-segregating sites. Second one was calculated as the difference between the first and the total length of the alignment.

Region	Segregating sites	Non-segregating sites
essential	6193	921746
nonessential	19146	2712669
16S	9	1528
asRNA	70	9257
ncRNA	45	4481
TFBS	1577	165823
tRNA	9	3376

Using Table 5, a Pearson's chi-squared test of independence was performed to check if the presence of polymorphisms had any relationship with the region analyzed. The result was a pi-value of $2.2x10^{-16}$, so it is statistically accurate to say that some regions have higher and lower levels of diversity than others, in other words, there are more and less genetically conserved regions.

Assuming that essential gene regions are conserved through evolution and once proved that their genetic diversity is lower than the nonessential gene regions, it is possible to perform additional chi-square tests using the values of essential regions and every of the other regions individually. In this case, values of the two regions are taken with the null hypothesis that the occurrence of their polymorphism distribution is statistically independent, so finding a pi-value lower than 0.05 means that both regions have some kind of association or relationship. This way, comparing essential gene regions to the rest can give an idea of their evolutionary conservation:

- Essential vs. nonessential, pi-value = 0.000805: the comparison between those two yields a very low pi-value, meaning once more that both types of regions do not share a common polymorphism distribution. They have a negative relationship where essential regions are more conserved than nonessential ones.

- Essential vs. 16S, pi-value = 0.8127: 16S RNA is a component of the 30S small subunit of bacterial ribosomes which rate of evolution is very slow and so it has been widely used to construct phylogenies. It is then not surprising to find a high pi-value in this test, meaning that the high conservation of the 16S region is closely similar to the essential gene regions.

- Essential vs. asRNA, pi-value = 0.3594: antisense RNAs are single-stranded RNAs complementary to messenger RNAs which function generally is to inhibit translation of that mRNA. They are encoded inside the same gene region than the mRNA they regulate, and so it is expected for them to have the same diversity than the gene they belong to. This relatively high pi-value means that they are well-conserved, as they are subjected to a double evolutionary pressure. A first pressure is the one that they gene they are encoded into suffers, and then the second is relative to their true regulatory function (Arnvig *et al*, 2014).

- Essential vs. ncRNA, pi-value = 0.0093: non-coding RNAs are functional RNA molecules that are not translated into proteins but have a broad set of regulatory functions. This pi-value points out to think that they are less conserved than essential gene regions, but it must be taken into account that inside the variety of regulatory roles they play, some may be very important and then conserved, and many more may be not subjected to evolutionary pressure and then much more free to accumulate mutations. So the individual conservation of this type of regions relies to a great extent whether they play an important role or not (Arnvig *et al*, 2014).

- Essential vs. TFBS, pi-value = $2.2x10^{-16}$: data for transcription factor binding sites was extracted from ChIP-Seq experiments, and it is sure that lots of the peaks originate from random affinity binding to non-regulatory DNA sequences, so many of the analyzed regions may be inaccurate. The pi-value obtained indicates there is no conservation in relation to essential genes, but keeping in mind the discussed experimental bias, it is needed to further study this regions in the lab trying to prove if there is expression associated to every TFBS analyzed.

- Essential vs. tRNA, pi-value = 0.00576: transfer RNAs are adaptor molecules that carry amino acids to the ribosomal translation machinery. One of the ends of a tRNA matches the genetic code in a sequence of three nucleotides called the anticodon. The pi-value shows that they are less conserved than essential gene regions, which could be explained by the fact that as the genetic code is degenerate, more than one anticodon can carry a given amino acid, so evolutionary pressure diminishes due to this phenomenon. The difference in codon usage of the different strains also support this theory (Andersson and Sharp, 1996; Cole *et al*, 1998).

In addition to comparing essential gene regions to the rest, every region was compared in the same way (chi-squared independence test) to the others, getting a matrix of values represented in Figure 8 as a heatmap. A multiple test correction using Bonferroni and Hochberg corrections was made but it yielded no significant changes in the p-values. This figure tries to represent all relationships between categories in a quick graphical way. Green squares mean that both regions share common levels of diversity, while orange tones indicate the opposite.

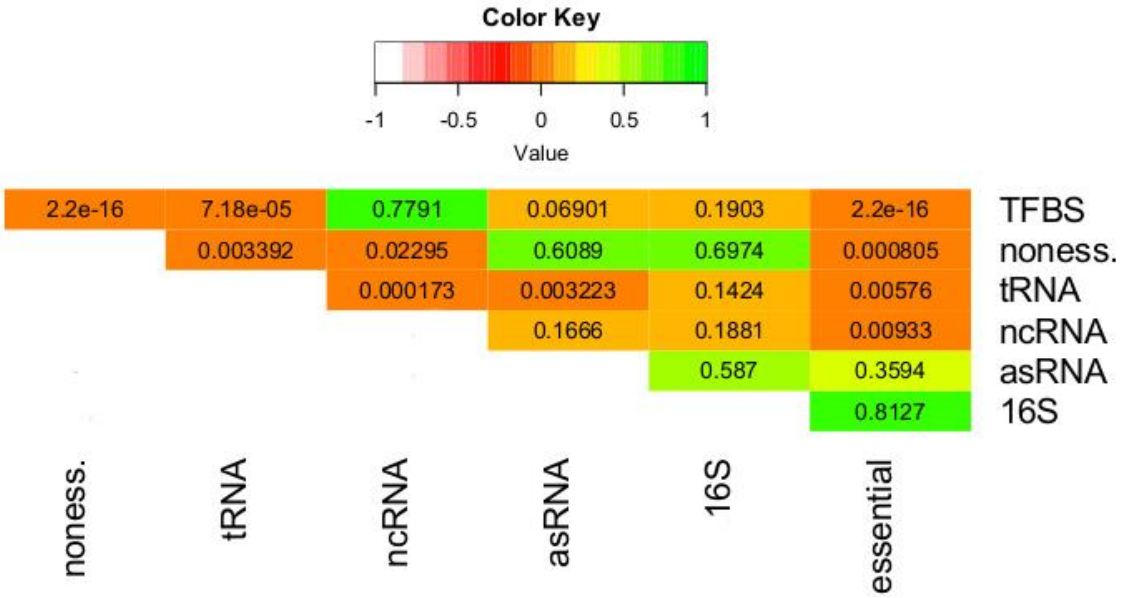

Figure 8. Heatmap of the chi-square test comparisons between every pair of region types. Color key establishes the degree of relationship between two regions according to the pi-value.

4.3. Dataset 4 results

Dataset 4 was constructed by concatenating every region of a certain type for a given strain and comparing the resulting concatenate to the ancestor reference sequence. This way, values of divergence (K) from the ancestor for each strain and region could be obtained. Genetic divergence is the process of independently accumulating mutations over time, so analyzing theses values gives an idea of how evolutionarily distant is a strain from the ancestor.

All values for each region type were grouped by lineage (from 1 to 7) and averaged to obtain the average divergence from the ancestor of every lineage and for each region. The resulting values were represented by bar plots (Figure 9).

Looking at the essential and nonessential regions graphs, both share a common shape. lineages 2, 3 and 4 (the modern ones) appear to be more conserved than the rest for these regions. The rest of lineages show some more diversity, which could mean they have undergone more purifying selection. For asRNA, lineages 4 and 7 have the most

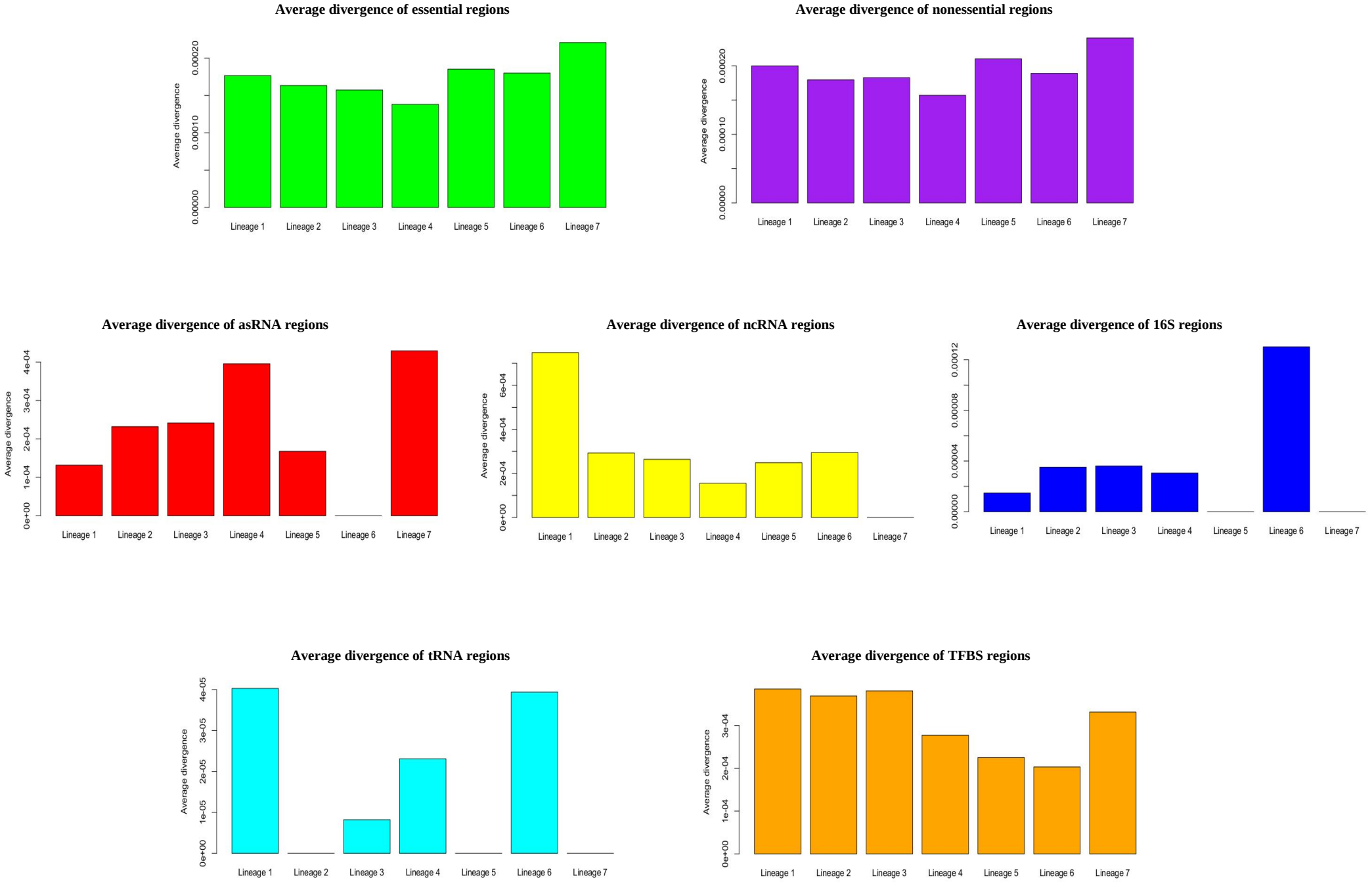

Figure 9. Average divergence of each *M. tuberculosis* lineage for every region type.

diversity, the rest are more conserved, especially lineage 6. For ncRNA, lineage 1 stands out as the most divergent, while the others show little diversity, especially lineage 7. For the 16S region, the one which stands out is lineage 6, and the rest appear conserved, particularly lineages 5 and 7. For tRNA, lineages 1 and 6 have the highest diversity, while lineages 2, 5 and 7 are very conserved. Finally, every lineage shows high divergence in TFBS, the lowest ones being 4, 5 and 6. This is an exploratory analysis, and raises a lot of questions about why some regions are so conserved in certain lineages. This disparity could translate into differences in clinical manifestation of the strains of a lineage, something that should be further studies in the future.

5. CONCLUSION

This report presents an initial step in the understanding of how important transcriptional regulatory regions are in the *Mycobacterium tuberculosis* complex. It tries to address their levels of genetic diversity and conservation by comparison to essential and nonessential gene regions.

From a global sample of 216 *M. tuberculosis* strains, nucleotide diversity, polymorphic sites and divergence of all the aforementioned regions were calculated using the software package VariScan. After it was statistically validated that essential regions were more conserved than nonessential regions, comparisons to every other region were assessed to check for relationships. Chi-square tests revealed that 16S and antisense RNA regions were conserved showing a diversity degree similar to essential regions. As expected, the 16S gene had few polymorphisms across the sample collection, thus presenting a low value of nucleotide diversity. Antisense RNA regions similarity to essential regions was explained by their double evolutionary pressure which difficults polymorphism fixation. Non-coding RNA regions and transcription factor binding sites appeared to be poorly conserved, but their complete dataset contained both many regions with high variability and many with zero variability. In the case of non-coding RNA, this fact suggests that regions with important regulatory roles in the cell do not allow diversity while other less essential regions are free from this pressure. Regarding the binding sites of transcription factors, the fact that the dataset comes from ChIP-Seq experiments where random affinity bindings can happen, could explain why the results are so polarized. Further experiments need to be done to check for active transcription associated to every predicted binding site. This way, new analyses could be performed to precisely determine which transcription factors are more essential for the cell and why. Finally, transfer RNA regions showed less conservation than essential genes, hinting that the particular codon usage of the bacteria, together with codon degeneracy, allows for some diversity within these regions.

Future projects following this line of study could further analyze single regions from the datasets and perform lab experiments involving the most conserved regulatory regions to see if they could constitute a therapeutic target. Developing a global regulatory network model of *M. tuberculosis* which can predict this kind of outcomes is also a critical milestone in the fight against the pathogen. The results in this report could help in the refining of such complex models which need all the possible data of the different strains from around the world to better understand bacterial adaptation and intracellular persistence, and develop strain-specific therapeutics in the future.

6. REFERENCES

ARNVIG, K. B., COMAS, I., THOMSON, N. R., HOUGHTON, J., BOSHOFF, H. I., CROUCHER, N. J. YOUNG, D. B. (2011). Sequence-based analysis uncovers an abundance of non-coding RNA in the total transcriptome of Mycobacterium tuberculosis. *PLoS Pathogens*, *7*(11).

ARNVIG, K. B., YOUNG, D. B., & CORTES, T. (2014). Noncoding RNA in Mycobacteria. *Microbiology Spectrum*, *2*(2).

BARTEK, I. L., WOOLHISER, L. K., BAUGHN, A. D., BASARABA, R. J., JACOBS, W. R., LENAERTS, A. J., & VOSKUIL, M. I. (2014). Mycobacterium tuberculosis Lsr2 Is a Global Transcriptional Regulator. *Mbio*, *5*(3).

COLE, S. T., BROSCH, R., PARKHILL, J., GARNIER, T., CHURCHER, C., HARRIS, D, BARRELL, B. G. (1998). Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*, *393*(6685).

COMAS, I., CHAKRAVARTTI, J., SMALL, P. M., GALAGAN, J., NIEMANN, S., KREMER, K., GAGNEUX, S. (2010). Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved. *Nature Genetics*, *42*(6).

COMAS, I., COSCOLLA, M., LUO, T., BORRELL, S., HOLT, K. E., PARKHILL, J., THWAITES, G. (2013). Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. Nat Genet. Oct;45(10).

CORTES, T., SCHUBERT, O. T., ROSE, G., ARNVIG, K. B., COMAS, I., AEBERSOLD, R., & YOUNG, D. B. (2013). Genome-wide Mapping of Transcriptional Start Sites Defines an Extensive Leaderless Transcriptome in Mycobacterium tuberculosis. *Cell Reports*, *5*(4).

FANG, G., ROCHA, E., & DANCHIN, A. (2005). How essential are nonessential genes? . *Molecular Biology and Evolution*, *22*(11).

FONSECA, J. D., KNIGHT, G. M., & MCHUGH, T. D. (2015). The complex evolution of antibiotic resistance in Mycobacterium tuberculosis. *International Journal of Infectious Diseases*, *32*, 94–100.

GALAGAN, J. E., MINCH, K., PETERSON, M., LYUBETSKAYA, A., AZIZI, E., SWEET, L., SCHOOLNIK, G. K. (2013). The Mycobacterium tuberculosis regulatory network and hypoxia. *Nature*, *499*(7457).

GRIFFIN, J. E., GAWRONSKI, J. D., DEJESUS, M. A., IOERGER, T. R., AKERLEY, B. J., & SASSETTI, C. M. (2011). High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism. *PLoS Pathogens*, *7*(9), 1–9.

HUTTER, S., VILELLA, A. J., & ROZAS, J. (2006). Genome-wide DNA polymorphism analyses using VariScan. *BMC Bioinformatics*, *7*, 409.

JORDAN, I. K., ROGOZIN, I. B., WOLF, Y. I., & KOONIN, E. V. (2002). Essential Genes Are More Evolutionarily Conserved Than Are Nonessential Genes in Bacteria. *Genome Research*, 962–968.

MCDERMOTT, J. G., PROLL, S. C., ROSENBERGER, C., SCHOOLNIK, G., & KATZE, M. G. (2011). A Systems Biology Approach to Infectious Disease Research : Innovating the Pathogen-Host Research Paradigm. *MBio*, *2*(1), 1–4.

MINCH, K. J., RUSTAD, T. R., PETERSON, E. J. R., WINKLER, J., REISS, D. J., MA, S., SHERMAN, D. R. (2015). The DNA-binding network of Mycobacterium tuberculosis. *Nature Communications*, *6*, 1–10.

NEI, M., & LI, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, *76*(10).

PETERSON, E. J. R., REISS, D. J., TURKARSLAN, S., MINCH, K. J., RUSTAD, T., PLAISIER, C. L., BALIGA, N. S. (2014). A high-resolution network model for global gene regulation in Mycobacterium tuberculosis. *Nucleic Acids Research*, *42*(18), 11291–11303

REDDY, T. B. K., RILEY, R., WYMORE, F., MONTGOMERY, P., DECAPRIO, D., ENGELS, R., SCHOOLNIK, G. K. (2009). TB database: An integrated platform for tuberculosis research. *Nucleic Acids Research*, *37*(SUPPL. 1), 499–508.

SASSETTI, C. M., BOYD, D. H., & RUBIN, E. J. (2003). Genes required for mycobacterial growth defined by high density mutagenesis. *Molecular Microbiology*, *48*(1), 77–84.

SASSETTI, C. M., & RUBIN, E. J. (2003). Genetic requirements for mycobacterial survival during infection. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(22).

VILELLA, A. J., BLANCO-GARCIA, A., HUTTER, S., & ROZAS, J. (2005). VariScan: Analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. *Bioinformatics*, *21*(11).