10

Analysis of Duplicate Gene Families in Microbial Genomes and Application to the Study of Gene Duplication in *M. tuberculosis*

Venu Vuppu and Nicola Mulder
Computational Biology Group, Department of Clinical Laboratory Sciences
Institute of Infectious Diseases and Molecular Medicine, Health Science Faculty
University of Cape Town
South Africa

1. Introduction

Though considerable sequence information from different organisms was available prior to the recent advances in genome sequencing technology, the foundation for our current understanding of the mechanisms of bacterial pathogenesis was laid by the release of the first complete genome sequence of *Heamophilus influenza* in 1995 (Fraser-Liggett, 2005). Ever since, significant progress in the availability of data for different genomes has been possible due to the contribution of various genome sequencing projects (Koonin & Wolf, 2008). Despite the complete genome sequences of many pathogenic organisms being available, the mortality rates due to these infectious agents still remains a problem, highlighting the need to decipher the complex molecular mechanisms responsible for survival of the bacteria. The wealth of complete genome information for pathogens can be effectively explored using comparative genomic tools for the identification of common and unique sets of genes involved in the propagation of virulence. Sequence comparison tools have been developed to identify homologous genes from the complete genomes of microorganisms. Homologous genes which arise from speciation tend to maintain functions similar to that of their ancestral molecule and are known as orthologs, while the genes originating from duplication events often evolve new functions and are defined as paralogs (Tatusov *et al.*, 1997).

The world of microbes is highly diverse with genome complexity differing across a wide range of microorganisms. In general, the difference in the complexity of genomes is dictated by the life style and environment of the organism (Cordero & Hogeweg, 2009). Life style plays an important role in regulating the genome dynamics of an organism, and functional novelty provided by gene duplication is thought to enhance the adaptation capability of the organism. In addition, horizontal transfer of operons or functional units of genes from external sources may provide an immediate functional benefit to the organism, thereby adding to the functional complexity of the genomes. The availability of complete genome sequences of important mycobacteria such as *Mycobacterium tuberculosis, Mycobacterium ulcerans, Mycobacterium bovis, Mycobacterium leprae, Mycobacterium paratuberculosis, Mycobacterium avium* and others, can be used to gain deeper insights into possible

mechanisms of prokaryotic gene innovation. Genome data has been used in recent studies to compare different species of mycobacteria, as well as different strains, to understand the evolution and pathogenesis of *M. tuberculosis* (Marri *et al.,* 2006). In our study, duplicate gene sets from different mycobacteria were investigated to identify the distribution of important functional classes of protein families, and the evolution of these functional classes was further analyzed by comparing their genetic divergence following duplication.

The importance of gene duplication in prokaryotic gene innovation is well established and comparative analysis of duplicate genes with basic characteristic features of genomes like GC content and genome size may aid in deciphering their contributions. In contrast to eukaryotes, GC content varies widely across different bacterial genomes (Mann & Chen, 2010), and analysis of GC variations between related bacteria could be useful in establishing evolutionary relationships (Mann *et al.,* 2010). The focus of the majority of earlier studies was on deciphering the role of GC composition in HGT (Nelson *et al.,* 1999; Hamady *et al.,* 2006), transcription start and stop sites (Zhang *et al.,* 2004), nucleotide substitution rates (DeRose-Wilson & Gaut, 2007), optimal growth temperature (Basak & Ghosh, 2005; Musto, 2006) and metabolic characteristics (Naya *et al.,* 2002). Furthermore, genome size has been reported to increase with an increase in number of genes in duplicate gene families (Snel *et al.,* 2002; Pushker *et al.,* 2004). In this study we analyzed the genomes of 56 pathogenic and 20 non-pathogenic microorganisms to identify and characterize the expanded gene families across these organisms. In addition to the GC content, we investigated the relationship between genome size and duplicate gene percentage. On finding sufficient evidence for a correlation between genome size and extent of gene duplication, we further investigated the significance of duplicate genes in enhancing genome complexity. He and Zhang (2005) previously reported the importance of gene duplication in enhancing genome and organism complexity in eukaryotes. However, due to the difference in the selective pressures operating on prokaryotic and eukaryotic genomes, we used the duplicate and single copy genes to investigate the influence of protein lengths on genome and organism complexity of prokaryotic organisms, with a specific focus on investigating the role of duplicate genes in enhancing the genome complexity of *M. tuberculosis*.

2. Materials and methods

2.1 Data selection and identification of homologous sequences

Comparative sequence analysis of different genomes is the most common approach for identifying orthologs and paralogs. However, here we used both the sequence and protein signature data as the latter could substantiate the former, and enables identification of more distantly related members of a protein family. We collected non-redundant protein sets for 76 microorganisms, including pathogens and non-pathogens, to identify expanded gene families in these organisms. The selection of the non-pathogenic bacteria in this study is of value, since many of these may also contain virulent genes which could act as barriers conferring protection against the defense mechanisms of the host, thus enhancing the survival capabilities and adapting the organism to intracellular conditions. In addition, acquisition of specific virulent gene clusters can transform these non-pathogenic agents to pathogenic microorganisms.

For the selected organisms, approximately 1,91,497 protein signatures, 2,47,858 protein sequences, Genome size and G+C composition data were retrieved from the InterPro (http://www.ebi.ac.uk/interpro) (Apweiler *et al.,* 2001; Mulder *et al.,* 2007) and Integr8 (http://www.ebi.ac.uk/integr8) (Kersey *et al.,* 2005) databases respectively. The protein

signature data from InterPro enabled the identification of approximately 27,827 proteins which exhibited complete domain identity (same InterPro matches) over their entire length to one or more proteins in *M. tuberculosis* strain H37Rv. Within each organism and across all organisms, the proteins showing complete domain identity were grouped together as duplicate gene sets or ortholog and paralog sets, respectively, and those with no common signature matches were considered to be single copies. In addition to the identification of expanded families using InterPro data, homologous sequences were clustered using BlastClust in two separate clustering procedures:

a. **Independent Genome Clustering**: This involves within genome clustering to generate clusters of paralogs or protein families for each genome. BlastClust was executed at a wide range of percentage identities over varying lengths of the sequence to select the optimum parameters. Amongst the tested parameters, a 30% similarity over 60% sequence length cut-off was chosen, as it generated a suitable number of clusters (in line with previously reported numbers of duplicated families for *M. tuberculosis*).
b. **Multiple Genome Clustering**: In this, all of the 76 genomes were appended together for the clustering of related proteins (orthologs and paralogs). In addition, the clustering of six of the mycobacterial species was performed separately for the evolutionary analysis of expanded gene families in *M. tuberculosis*.

2.2 Evolutionary analysis

For evolutionary studies, in addition to 66 paralogous gene clusters, 116 multiple genome clusters from the phylogenetic matrix of six of the closely related mycobacterial genomes that showed gene family expansions in both *M. tuberculosis* and *M. leprae*, as well as other mycobacteria, were selected. The proteins in each of the clusters were aligned with T-coffee (Notredame *et al.*, 2000), and poorly aligned regions were edited using the Gblocks program (Castresana**,** 2000) with adjustments in the default settings for the generation of optimal sequence alignments. For each of these protein alignments, selection of the best-fitting amino acid substitution model was performed according to the Akaike informational criterion, and the gamma correction factor (alpha), the proportion of invariable sites (I), and observed amino acid frequencies (F) were estimated and selected for subsequent phylogenetic analysis using ProtTest (Abascal *et al.*, 2005). Since, PhyML is a maximum likelihood method with the ability to incorporate the estimated values of alpha, proportion of invariable sites, and observed frequencies, the tree topologies for the gene sets in the identified clusters were constructed using this program (Guindon & Gascuel, 2003). The genetic distance measures from each of the estimated tree topologies were used to compute average and maximum genetic distance using Perl scripts.

3. Results and discussion

3.1 Identification of expanded gene families and relation to GC content and genome size

We used sequence clustering and protein signature data to identify expanded genes families within and across several different microbial genomes. The across-genome clustering of protein sequence data yielded 1,984 expanded genes in 441 clusters for *M. tuberculosis* H37Rv. The protein signature method allowed us to group 30,885 proteins into 2238 clusters from all the organisms. InterPro signatures usually match between 50% and 80% of a genome, so data is not available for every protein. Since signature data enables identification

of more distantly related members of a cluster, but loses data where proteins do not match InterPro, the sequence and signature-based cluster data was merged, and used to generate a phylogenetic profile, which reflected the number of copies in each expanded family for each organism. From this, we identified 2011 duplicate/expanded genes in 461 clusters for *M. tuberculosis*, confirming previous reports that the duplicate genes make up approximately half of the *M. tuberculosis* genome (Tekaia *et al.,* 1999). The percentages derived from the 2 methods are shown in Table 1 for the 6 mycobacteria studied. The 461 clusters in the merged data include gene families that are also expanded in different organisms.

S.No	Organism	Sequence	Signature	Union
1	*M. tuberculosis*	31.47%	38%	50.96%
2	*M. bovis*	30.28%	42%	48.69%
3	*M. paratuberculosis*	39.75%	49%	56.46%
4	*M. avium*	42.06%	49%	55.19%
5	*M. ulcerans*	36.82%	46%	53.51%
6	*M. leprae*	12.03%	20%	30.44%

Table 1. Percentage of the genome belonging to expanded gene families for the mycobacteria. Data was generated using sequence clustering, protein signatures and a combination of the two (union).

Next, we investigated the GC composition of different bacteria in relation to the duplicate gene percentages (Figure 1) to understand the characteristic features of genomes maintaining high percentages of duplicate genes. A statistical analysis of the data using Pearson's correlation, revealed a moderate correlation between the GC content and estimated duplicate gene percentages. However, the analysis of the trend lines of the scatter plot in figure 1 reveals the presence of three different kinds of relationships in the data: i) an initial increase of the trend line, ii) the initial increase is followed by a phase of neutrality, iii) and a steady increase of the trend line follows the phase of neutrality. We analyzed histograms of GC content and duplicate gene percentages and observed differences in the modality of the data distribution; GC percentages followed a trimodal distribution, while the duplicate gene percentages followed a unimodal distribution. Thus, although a positive correlation could be inferred from the analysis of the scatter plot, the correlation coefficient could be subdued by the differences in the modality of the data distributions. Hence, based on the analysis of the scatter plot and trimodal distribution of the GC percentage histogram, the organisms were grouped into three categories based on their GC compositions:

Group 1: Organisms having GC content greater than 54 percent.

Group 2: Organisms having GC content greater than 44 percent and less than 54 percent.

Group 3: Organisms having GC content less than 44 percent.

We then performed a one-way ANOVA on the data (Table 2). Taking into consideration the mean square values (Mean Sq) and the calculated p-value of 2×10^{-16}, we predicted that the mean variance between groups is significant compared to the within sample variance. These results indicate the existence of differences in the means of the three groups of organisms, and hence, we reject the null hypothesis and accept the alternative. Further, to estimate how significantly different the means of each group are compared to one another, a Tukey's Honest Significant Difference (Tukey's HSD) test was performed, and significant differences in the mean values of group2 and group1, group3 and group1, and group3 and group2 were found. From the results table of the Tukey's multiple comparison test (Table 3), it can be

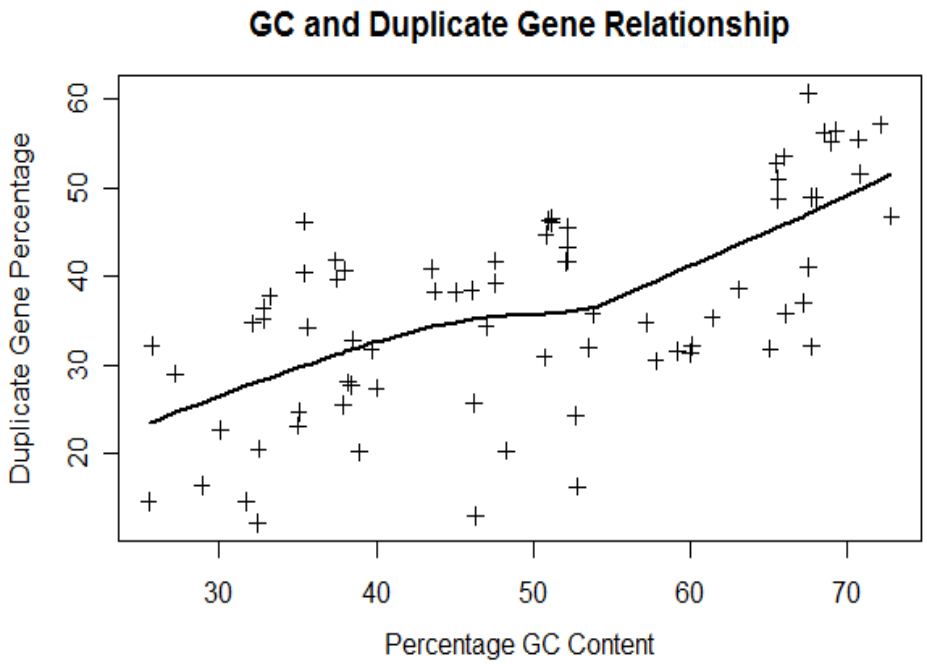


Fig. 1. Scatter plot analysis of the relationship between GC content and duplicate gene percentages of the selected organisms. The percent GC content is plotted on the X-axis and duplicate gene percentage on the Y-axis. The graph suggests that a positive correlation between GC content and duplicate gene percentages exists for the majority of the investigated organisms.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Groups	2	13174.8	6587.4	417.28	<2.2e-16***
Residuals	73				
Signif Codes	0 '***' 0.001 "**' 0.01 '*' 0.05 '.'0.1 ' ' 1				

Table 2. One-Way ANOVA Results. The columns of the table display the degrees of freedom (df), sum square values (Sum Sq), Mean square values (Mean Sq), F value and p-value (Pr(>F)) reported by One-Way ANOVA for the data.

Tukey's Multiple Comparison of Means			
Groups	Diff	Lwr	Upr
G2-G1	-16.36	-19.08	-13.64
G3-G1	-31.54	-34.15	-28.92
G3-G2	-15.18	-17.88	-12.48

Table 3. The table displays the differences between the mean values of the groups. The Groups column represents the investigated groups: group2 and group1 (G2-G1), group3-group1 (G3-G1), and group3-group2 (G3-G2). The differences in the means of the groups are given by the difference (diff) column, and the lower (lwr) and upper (upr) columns represent the lower and upper boundaries for the estimated mean difference between the groups.

inferred that both groups, G2-G1 and G3-G2 exhibit similar mean differences (-16.36 and -15.18). However, the mean differences of both these groups are higher than the mean difference (-31.54) of group G3-G1. Therefore, group2 organisms, which have higher mean differences compared with group1 and group3, could be responsible for the reduced correlation coefficient values. Hence, their elimination from the list of investigated organisms could result in the prediction of strong positive correlation between the GC composition and duplicate gene percentages of group1 and group3 organisms. Thus, we suggest that gene duplication events may be a characteristic feature of GC rich bacterial genomes. Since all of the selected mycobacterial species in the present study are representatives of group1, the phenomenon of gene duplication in this genus could be attributed to its high GC content.

In addition to GC compositions, we analyzed the influence of duplicate genes on the physical expansion of the genomes (Figure 2). An observed correlation coefficient value of 0.84 at a p-value of 2×10^{-16} between the genome size and duplicate genes provides sufficient evidence to prove the contribution of duplicate genes to genome expansion of these organisms. This is not surprising, since the addition of genes through gene duplication will obviously increase genome size unless some genes are lost in the process.

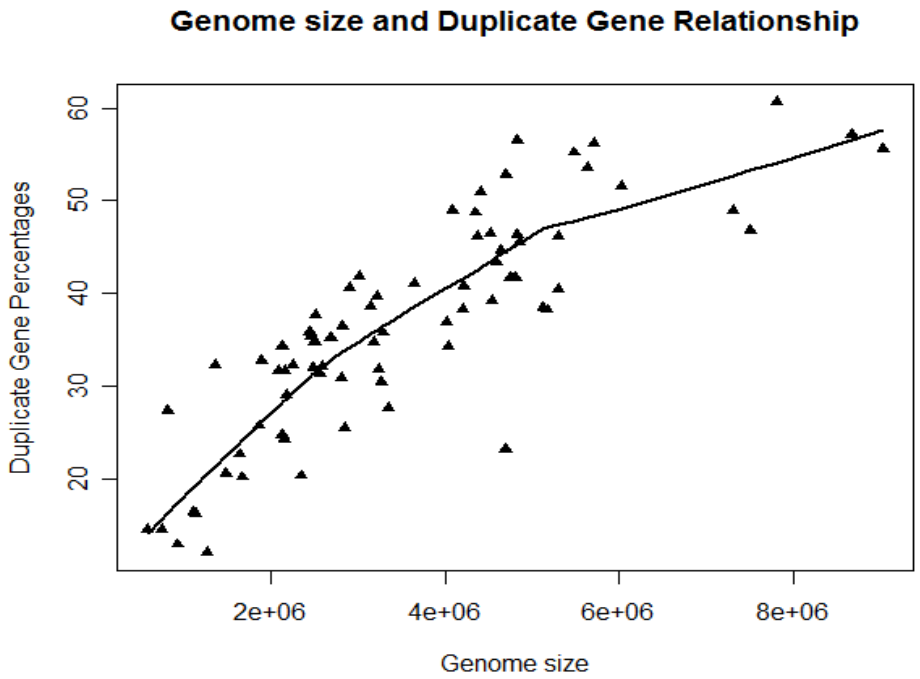


Fig. 2. The graph displays the relationship between duplicate gene percentage and genome size for the selected organisms. The identified duplicate gene percentages are plotted on the X-axis and genome size on the Y-axis. From the graph, a positive correlation can be observed between duplicate gene percentages and genome size.

3.2 Investigation of functional complexity of the duplicate and single copy genes

In eukaryotes, single copy genes were reported to be shorter and to contain fewer domains than duplicate genes (He & Zhang, 2005). Here, we investigated the gene lengths and complexity (using domain number) of both pathogenic and non-pathogenic bacteria to determine the role of gene duplication in enhancing genome complexity in prokaryotes. A preliminary determination of the functional complexity of the expanded genes in the selected organisms was graphically analyzed by plotting the mean gene lengths of both the duplicate and single copy genes. Figure 3 shows that the average gene length in the majority of the organisms is comparatively higher for duplicate genes than for single copy genes.

The difference in the mean gene lengths of duplicate and single copy genes was statistically analyzed using the Mann-Whitney U test. An observed W value of 4880 at a p-value of 2×10^{-13} estimated from the Mann-Whitney U test confirms that the mean gene lengths of the duplicate genes are significantly higher than that of the single copy genes.

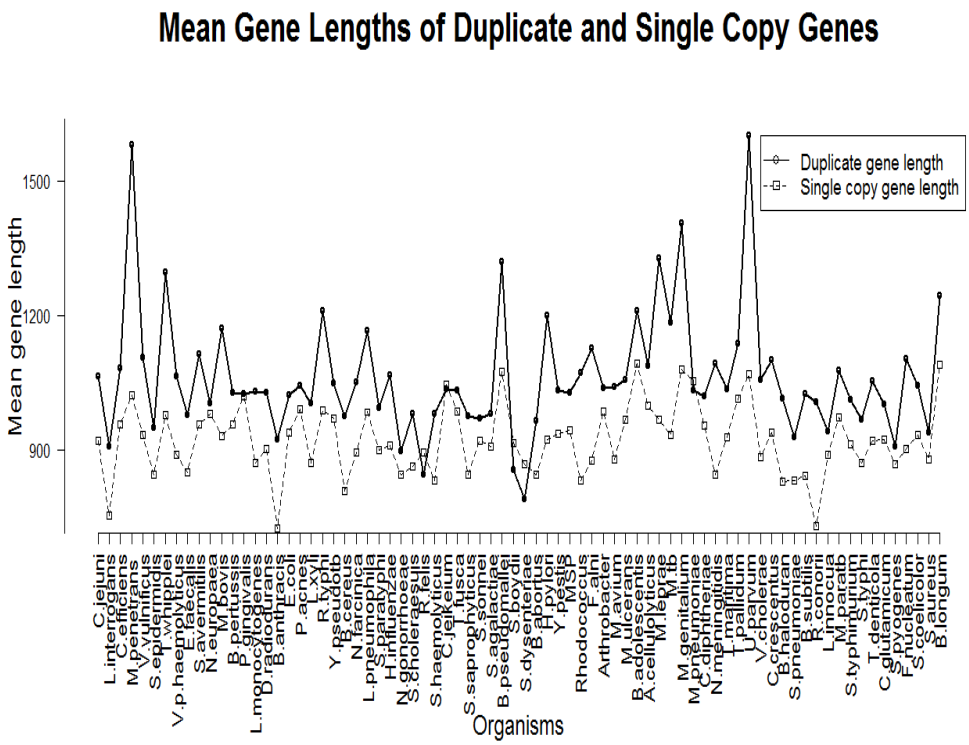


Fig. 3. Comparison of Functional Complexity of the Expanded and Single Copy Gene Families Based on Nucleotide Sequence Data. The graph displays the mean gene lengths of duplicate and single copy genes in the investigated organisms. The organisms are plotted on the X-axis and the corresponding mean gene length on the Y-axis.

We went on to investigate the domain complexity of single and duplicate copy genes to further enhance our understanding of the functional complexity of these organisms. As a measure of domain complexity, the number of domains present in each of the

corresponding proteins of the genes was computed from InterPro data, and the mean for the total number of domains was estimated for the duplicate and single copy genes using Perl scripts. From figure 4, we can see that the mean number of domains per duplicate gene is lower than that of the single copy genes. This suggests that single copy genes should be more complex due to the presence of more domains. We further analyzed the results, by statistically comparing the difference in the mean domain numbers of duplicate and single copy genes using the Mann-Whitney U test. From the resulting W value of 5717 at a p-value of 2×10^{-16}, we inferred that the mean number of domains per single copy genes is significantly higher than that of duplicate genes. Thus, these studies suggest that the single copy genes are functionally more complex than duplicate genes. This was a surprising result, given that the mean length of duplicate genes was found to be higher than that of the single copy genes. Therefore, we specifically investigated the influence of gene lengths on the domain complexity of *M. tuberculosis*, and compared this statistic in two other organisms, *Leptospira interrogans* and the model organism, *Escherichia coli.*

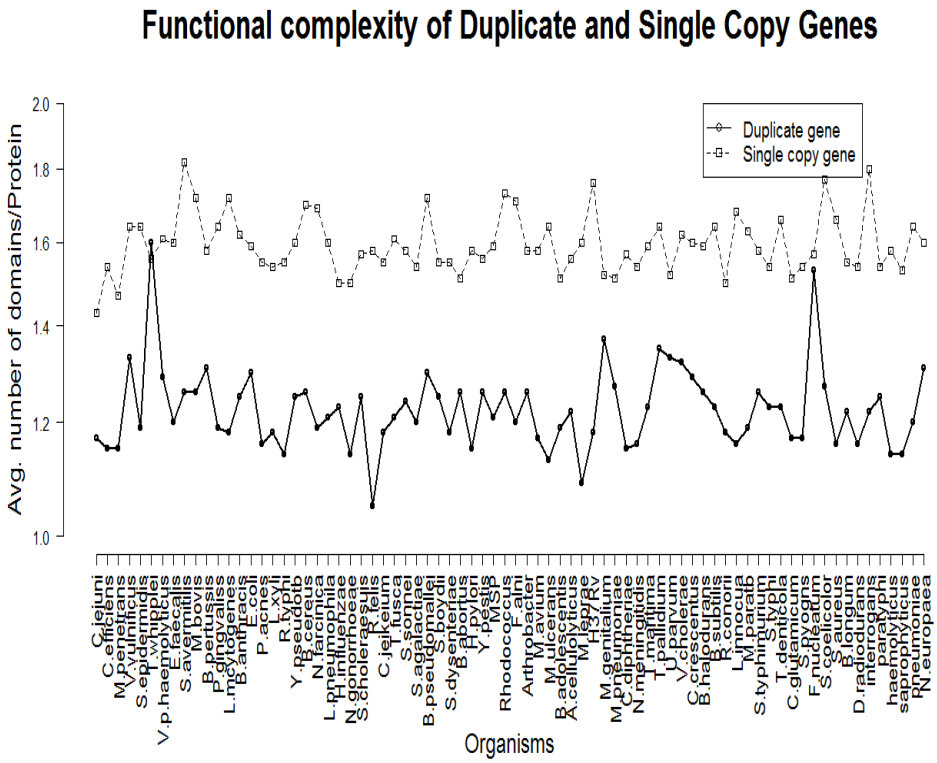


Fig. 4. Comparison of Functional Complexity of the Expanded and Single Copy Genes Based on InterPro Signature Data in Selected organisms. The graph displays the mean number of domains per protein in the duplicate and single copy genes of each organism. The organisms investigated are plotted on the X-axis and the corresponding mean number of domains per protein for each organism on the Y-axis.

For each of these three organisms, the total number of genes in the genome was retrieved, and for every gene, we estimated the total number of domains to determine the relationship between gene length and domain number (Figure 5 -Whole Genome Analysis). In addition, the number of domains for each of the duplicate (Figure 6) and single copy genes (Figure 7) were also estimated. The preliminary analysis of the relationships using scatter plots suggested that the number of domains per gene does not necessarily increase with an increase in the gene length. Further, correlation coefficient values estimated from the Pearson moment correlation were used for statistical confirmation of the relationships.

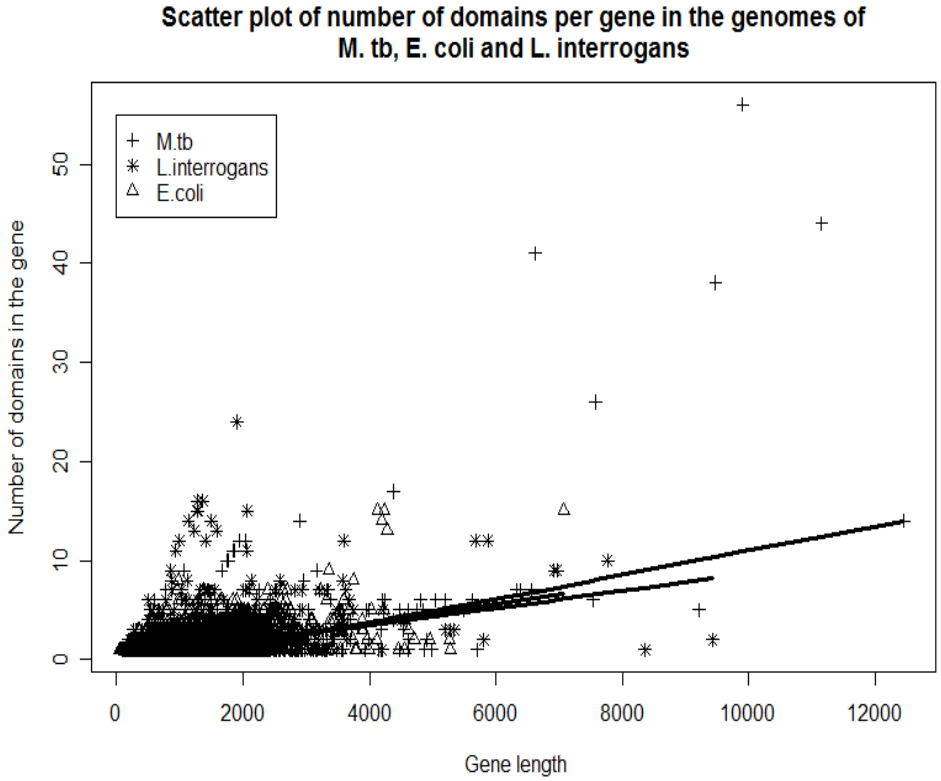


Fig. 5. Investigation of number of domains per gene in the *M. tuberculosis* H37Rv, *E. coli* and *L. interrogans* genomes (Whole Genome Analysis). The graph displays the relationships between gene length and number of domains. The sequence lengths of each gene are plotted on the X-axis and the corresponding number of domains per gene on the Y-axis. From the graph, it can be inferred that the gene length is not necessarily dependent on domain number.

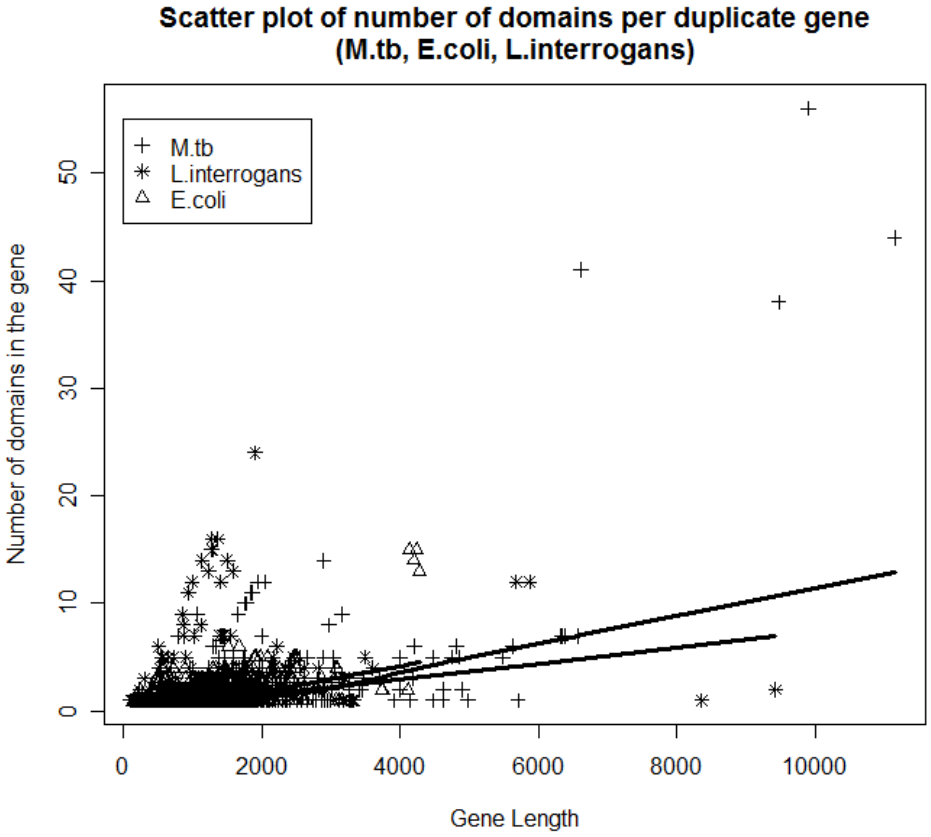


Fig. 6. Investigation of number of domains per duplicate gene in the *M. tuberculosis* H37Rv, *E. coli* and *L. interrogans* genomes (Duplicate Gene Analysis). The graph displays the relationship between sequence length (X-axis) and number of domains (Y-axis) of the duplicate genes.

The correlation coefficient values for the whole genome, duplicate gene and single copy gene analysis in *E. coli* were 0.48, 0.47, and 0.49, respectively, while the values of 0.39, 0.25, and 0.53 were reported for *L. interrogans* (Table 4). The results from these two organisms suggest that the number of domains does not increase significantly with the increase in gene length and hence, domain complexity may be independent of the gene length or vice versa. Although the reported correlation coefficient values of 0.58, 0.62 and 0.59 corresponding to the whole genome, duplicate and single copy gene analyses, respectively, in *M. tuberculosis* are higher than those for *E. coli* and *L. interrogans*, these correlation coefficient values still do not suggest significant positive correlation between domain complexity and gene lengths. Thus, the specific protein complexity studies of these three genomes show that it is not necessarily surprising that while the duplicate genes are generally longer than single copy genes, they tend to contain fewer domains.

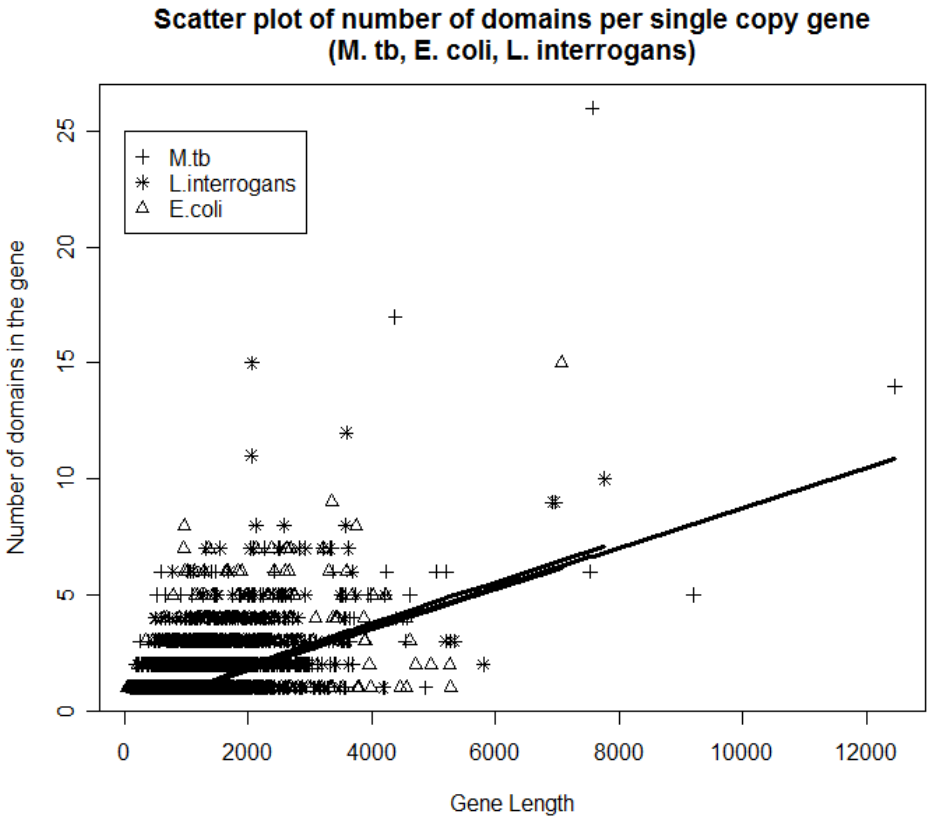


Fig. 7. Investigation of number of domains per single copy gene in the *M. tuberculosis* H37Rv, *E. coli* and *L. interrogans* genomes (Single Copy Gene Analysis). The graph displays the relationship between sequence length and number of domains of the single copy genes in these genomes.

Pearson's product-moment correlation						
Organism and P-value	E. coli	P-value	H37Rv	P-Value	L. interrogans	P-Value
All Proteins	0.48	2.2e-16	0.58	2.2e-16	0.39	2.2e-16
Duplicate	0.47	2.2e-16	0.62	2.2e-16	0.25	2.64e-10
Single	0.49	2.2e-16	0.59	2.2e-16	0.53	2.2e-16

Table 4. Correlation Coefficient and P-values of the whole genome, duplicate and single copy gene analysis in *E. coli, M. tuberculosis* H37Rv and *L. interrogans*. The table displays the results of Pearson's product-moment correlation.

3.3 Functional and evolutionary analysis of expanded genes in *M. tuberculosis*

The protein sequence and signature data for the 76 genomes were clustered into related sets of duplicate genes for the study of relationships between percentage duplication and the GC content, genome size and gene complexity described above. However, since the comparison of closely related organisms is better for inferring evolutionary relationships, we separately clustered six of the closely related mycobacterial genomes and identified 390 duplicate gene clusters in *M. tuberculosis*. The results were represented as a phylogenetic profile and a summary is shown in Table 5.

Organism	Total Genes	TGC	SCG	DGC	Total duplicate Genes	Estimated Duplicate Gene Percentages
M. tuberculosis	3947	2815	2425	390	1521	38.53
M. bovis	3910	2817	2439	378	1471	37.62
M. paratuberculosis	4316	2807	2343	464	1973	45.71
M. avium	5040	3199	2679	520	2361	46.84
M. ulcerans	4206	2755	2359	396	1847	43.91
M. leprae	1036	1603	1261	119	342	21.33

Table 5. Protein sequence clustering of the mycobacterial group. The columns of the table represent the selected organisms, total number of genes in the genome, total number of identified gene clusters (TGC), total number of single copy genes (SCG), total number of duplicate gene clusters (DGC) in the organism, total number of duplicate genes in the duplicate gene clusters identified, and the percentage of duplicate genes estimated for each organism.

The biggest expanded family in *M. tuberculosis* was the PE/PPE/PGRS family with 164 members, followed by a family of alcohol dehydrogenases and oxidoreductases with 44 members, the fatty-acid-CoA ligase family with 33 members, then acyl-CoA dehydrogenase with 27 members. A manual assignment of high-level functional classes was done previously in the laboratory for all *M. tuberculosis* proteins. This was used here to determine the functional distribution of all 390 expanded families in this organism. Figure 8 shows the number of families and number of proteins belonging to each of the functional classes. The biggest class is made up of enzymes or proteins involved in metabolism, followed by proteins of unknown function. From the data, we selected 116 gene clusters which showed gene family expansions in *M. tuberculosis* and *M. leprae*, as well as other mycobacteria. We are interested in expansion in *M. leprae*, as this is a highly reduced mycobacterial genome, so expanded genes that have been maintained are likely to be important. When considering only the 116 families that are also expanded in *M. leprae*, the distribution of functional classes is similar, except for a large reduction in the number of unknown protein families.

For each of the 116 clusters of interest, we calculated the genetic distance between family members and investigated the relationship between genetic distance and gene family size. Our results suggest that the genetic distance between two of the most distant proteins in the clusters increases with an increase in cluster size. In addition, the correlation coefficient value of 0.87 at a p-value of 2.2×10^{-16} is indicative of strong positive correlation between these factors. These sets of duplicate copies of proteins are clustered from different mycobacterial genomes, and since the estimated maximum genetic distance between the 2

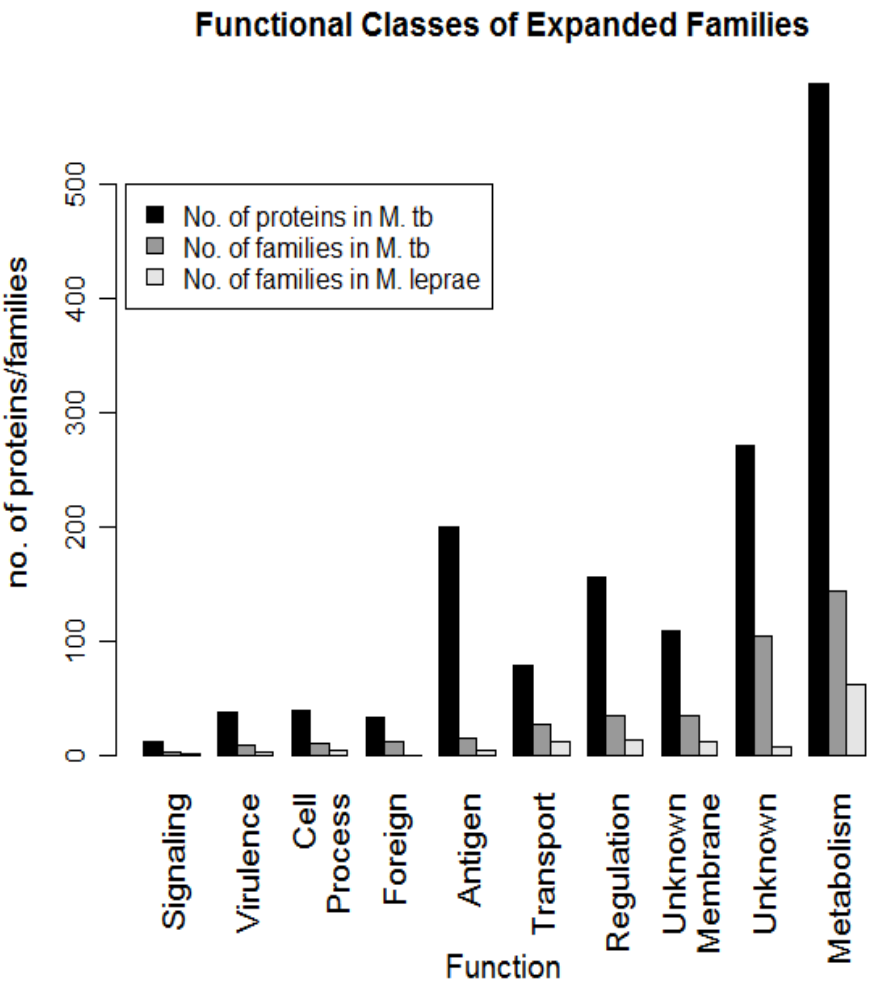


Fig. 8. Distribution of functions in *M. tuberculosis* (all 390 clusters) and *M. leprae*-shared (116 clusters) expanded families.

most distant proteins in each of these sets increases with an increase in cluster size, it was inferred that some of the duplicate copies show a tendency to diverge from the original ancestral functions after multiple duplication events in bigger families. To investigate the average divergence of proteins in these clusters, the relationship between average genetic distance and cluster size was determined. The results suggest that the average genetic distance between the gene families does not increase with the cluster size, except perhaps for the few larger families. In order to statistically verify the results, correlation coefficient values were estimated using the Pearson's product-moment correlation. The correlation coefficient value of 0.43 at a p-value of 1.17×10^{-6} indicates the presence of moderate

correlation between the average genetic distance and cluster size. It suggests that the majority of homologous gene families identified from these mycobacterial species have not undergone significant functional divergence and still show close evolutionary relatedness, but these results may be skewed by the fact that the clusters contain orthologs and paralogs. Orthologs are generally predicted to maintain similar functions, while paralogs are known to have diverged functions.

The relationship between cluster size and genetic distance was also studied for 66 paralogous families only (within genome clustering). Within each of the selected mycobacterial species, the estimated tree topologies were used to investigate the genetic divergence of the identified paralogous gene families by computing the maximum genetic distance between two of the most distant paralogs in each of the clusters. A scatter plot analysis of the computed maximum genetic distances and cluster sizes was performed (Figure 9), and correlation coefficient values were estimated for studying the genetic divergence of these paralog gene families. From the analysis of the scatter plot (Figure 9) and correlation coefficient values (Table 6), we inferred that the genetic distance between the two most distant proteins increases with the cluster size.

Organism	df	Pearson's correlation	P-value
M. tuberculosis	64	0.88	2.20e-16
M. bovis	62	0.81	4.4e-16
M. paratuberculosis	59	0.93	2.20e-16
M. avium	56	0.95	2.20e-16
M. ulcerans	59	0.93	2.20e-16
M. leprae	36	0.66	5.94e-06

Table 6. Results of the correlation calculations for maximum genetic distance versus cluster size, including degrees of freedom (df), Pearson's correlation coefficient values, and the corresponding p-values.

In addition to maximum genetic distance, the average genetic distance for each of the paralog gene families was computed to investigate the evolutionary relationships between the members within the selected mycobacterial genomes. To provide statistical significance for the scatter plot observations, correlation coefficient values were estimated using the Pearson's product-moment correlation (Table 7). The scatter plot (Figure 10) and correlation coefficient values (Table 7), suggest a moderate negative correlation between the average genetic distance and cluster size of the paralogous gene families.

3.4 Further analysis of one example expanded gene family in *M. tuberculosis*

While we have evolutionary data for all the orthologous and paralogous families of *M. tuberculosis*, we cannot show all the results, so we have selected an important class of regulatory proteins as an example. The adaptability of *M. tuberculosis* to enable successful survival of the stressful conditions in the host during infection is attributed to the existence of a diverse class of sigma factors in the organism (Fontan, 2009). The organism is suggested to contain numerous sigma factors that bind to the core subunit of RNA polymerase to provide promoter specificity (Fontan, 2008). To investigate the phylogenetic diversification of sigma factors in *M. tuberculosis* and other mycobacteria, we studied the sigma factors

Organism	df	Pearson's correlation	P-value
M. tuberculosis	64	-0.44	0.0001966
M. bovis	62	-0.5	2.38e-05
M. paratuberculosis	59	-0.41	0.0007649
M. avium	56	-0.43	0.0007819
M. ulcerans	59	-0.43	0.000633
M. leprae	36	-0.44	0.005978

Table 7. Pearson's correlation coefficient results for the relationship between the average genetic distance and cluster size. The columns include degrees of freedom (df), Pearson's correlation coefficient values, and the corresponding P-values.

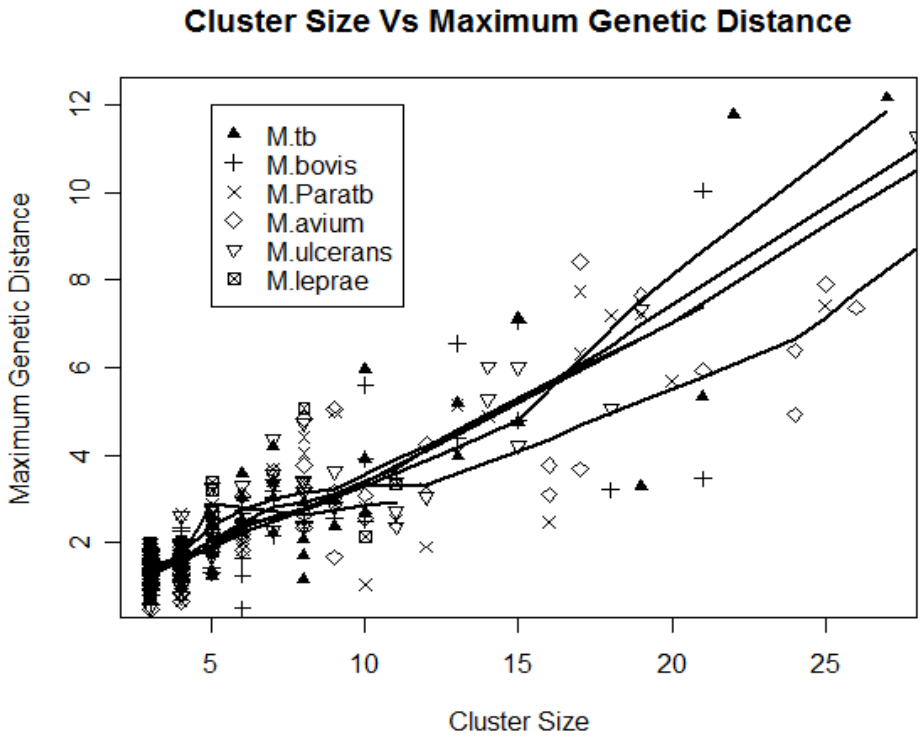


Fig. 9. Relationship between maximum genetic distance and cluster size for families of *M. tuberculosis* H37Rv, *M. bovis, M. paratuberculosis, M. avium, M. ulcerans* and *M. leprae.* The X-axis represents the cluster size (total proteins in each cluster) and Y-axis shows the genetic distance between the two most distant proteins in the clusters of each organism. The genetic distance appears to increase with the cluster size, suggesting a correlation between them.

identified by our ortholog and paralog clustering methods. From the analysis of the sigma factor phylogenetic trees of *M. tuberculosis* (Figure 11), we infer that gene duplication events followed by divergence could have resulted in the bifurcation of the sigma factor class of proteins into two subfamilies (marked as A and B in the Figure).

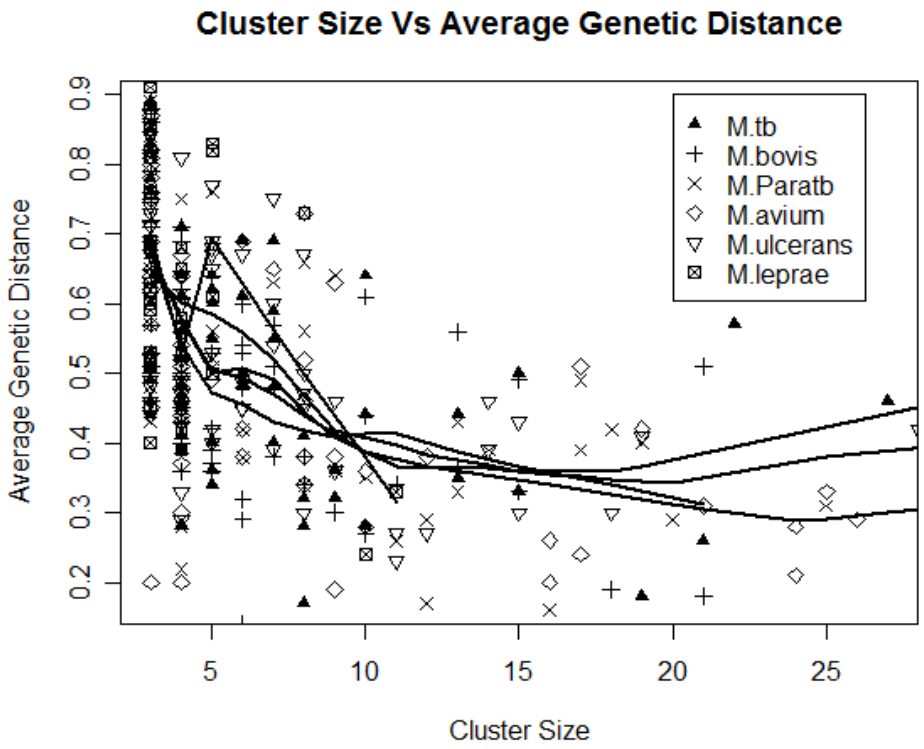


Fig. 10. Relationship between average genetic distance and cluster size for duplicate gene families of *M. tuberculosis* H37Rv, *M. bovis, M. paratuberculosis, M. avium, M. ulcerans and M. leprae*. The X-axis represents the cluster size (total proteins in each cluster) and Y-axis shows the average genetic distance between the identified gene families of each organism. The average genetic distance appears to decrease with the cluster size, suggesting a negative correlation between these two factors.

3.4.1 Analysis of sigma factor proteins in subfamily A

Following duplication, the proteins of this subfamily have diverged into 2 groups: SigE and SigM (Figure 11). The 2 proteins in the SigE group have further diverged following duplication and divergence. However, one of the proteins in the sigE group was identified to have no orthologs in *M. leprae* (Figure 12), and loss of various sigma factors is suggested to be the reason for *M. leprae* reductive genome evolution (Babu, 2003). Interestingly, all the paralogs of *M. tuberculosis* appear to have orthologs in *M. bovis*, but the absence of sigM proteins in *M. bovis*, and the large divergence of this protein group in *M. tuberculosis* compared to other mycobacteria enables us to speculate on its significance in *M. tuberculosis* evolution. Though an error in available *M. bovis* sequences could have resulted in incorrect annotation of the sigM locus as a psuedogene (Manganelli *et al.,* 2004), the extent of divergence of this protein in *M tuberculosis* compared to other mycobacteria prompts further investigation into its possible paths of pseudogenization or neofunctionalization.

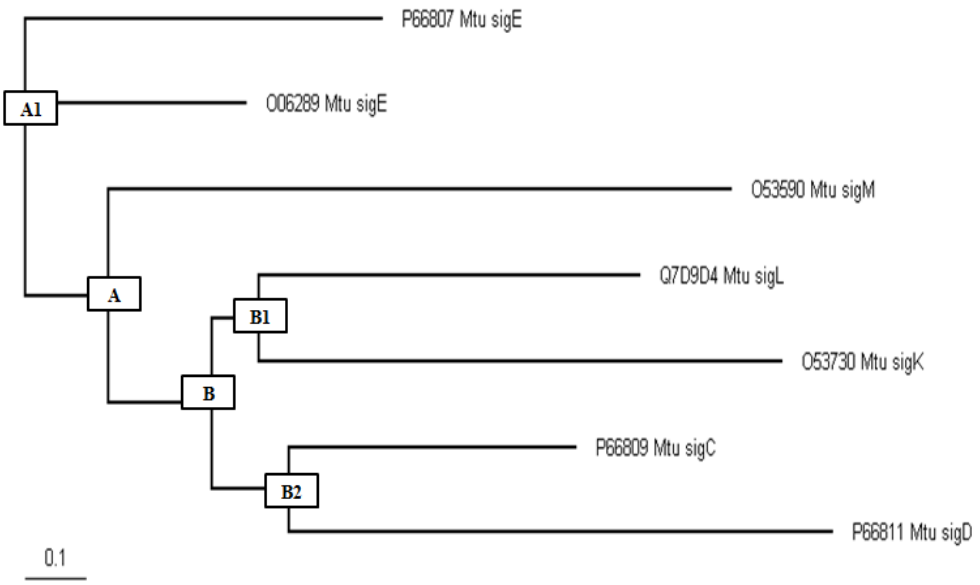


Fig. 11. Phylogenetic tree of the Sigma factor paralog cluster inferred by the maximum likelihood method. The duplication events are marked by A's and B's. The figure displays the phylogenetic diversification of sigma factors (sigE, sigM, sigL, sigK, sigC and sigD).

3.4.2 Analysis of the sigma factor proteins in subfamily B

From the analysis of the sigma factor class of paralogs in GroupB (Figure 11), we infer that duplication events followed by divergence resulted in the 2 groups of sigma factor subfamilies (marked as B1 and B2 in Figure 11). The sigma factor proteins in each of the subfamilies have significantly diverged after gene duplication. For the two sigma proteins (sigK and sigL) in one of the subfamilies (Figure 12), sigK was identified to have no orthologs in *M, avium, M paratuberculosis* or *M. leprae*, and sigL was noted to have no orthologs in *M. leprae*. These results are inconsistent with the published reports of Manganelli *et al*, 2003. For the other subfamily of sigma proteins (sigC and sigD) on the phylogenetic tree (Figure 12), we did not identify orthologs for sigC in *M. paratuberculosis* or sigD in *M. leprae*.

4. Discussion

The availability of complete genome sequences of many bacteria and significant progress in the development of modern computational biology methods has resulted in the evolution of a powerful platform for the comparative investigation of genome diversity across different organisms. Here, we make use of the wealth of genome information and bioinformatics tools to understand the significance of gene duplication in *M. tuberculosis* evolution. The investigation of relationships between the GC composition and duplicate gene percentages identified from the sequence and InterPro domain data provides sufficient evidence to suggest a positive correlation between them for group1 and group3 organisms. Here, the mycobacterial species are part of the group1 organisms, so the maintenance of

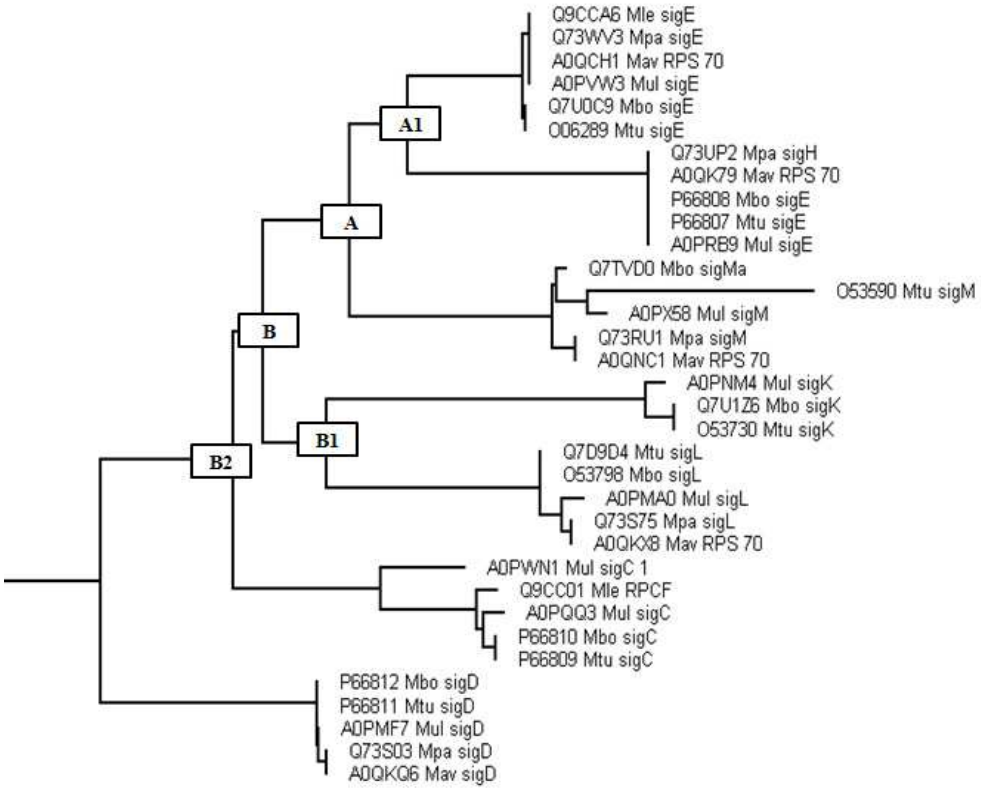


Fig. 12. Phylogenetic tree of the Sigma factor ortholog and paralog cluster inferred by the maximum likelihood method. The duplication events are marked by A's and B's. The labels Mtu, Mbo, Mav, Mul, Mpa and Mle represent proteins from *M. tuberculosis, M. bovis, M. avium, M. ulcerans, M. paratuberculosis and M. leprae* respectively.

high duplicate gene percentages in these species, with the exception of *M. leprae,* could be attributed to the high GC composition of their genomes. Unsurprisingly, the study has also shown a correlation between duplicate gene percentage and genome size, suggesting that gene duplication increases genome size. Further, our investigations on protein complexity provide deeper insights into the general trend in gene length and domain number in duplicate genes in these organisms.

He and Zhang (2005), investigating *Saccharomyces cerevisiae*, showed duplicate genes to be complex molecules with longer sequences containing more functional domains. From the investigations of the mean gene lengths of 76 pathogenic and non-pathogenic organisms, it is evident that the average length of the duplicate genes is comparatively higher than that of single copy genes. However, the analysis of mean number of domains in the duplicate and single copy genes reveals the presence of a higher number of domains in the single copy genes compared to the duplicate genes. According to Stoltfus (1999), partial loss-of-function

mutations lead to the preservation of duplicate genes with single functions (Stoltfus, 1999; Lynch & Force, 2000). Moreover, it has been suggested that duplicate genes lose one of the domains that were originally present in the ancestral molecule, and by complementation of the lost domains, both the daughter copies are reported to reflect the original ancestral function. Thus, gene complexity is suggested to be reduced after subfunctionalization of duplicate genes (He & Zhang, 2005). Further, the complementary loss of subfunctions is considered to facilitate the preservation of duplicate gene pairs, and due to relaxed evolutionary constraints following subfunctionalization, the chances of long-term evolution of new functions is enhanced (Force *et al.,* 1999). However, since deleterious mutations are more common than beneficial mutations (Cun, 2010), evolution of new and essential protein functions is considered to be a rare event (Nadeau & Sankoff, 1997; Force *et al.,* 1999). According to the predominant argument, the evolution of new domains would be favored only if they can perform a function different to that of preexisting domains or domain combinations (Lagomarsino *et al.,* 2009). Further, the majority of duplicate genes are predicted to develop new functions from the already existing ancestral gene functions, and if new functions evolve by mutation from prior domains, it is less likely that all of the domains would evolve into new domains due to the mutational bridge for new domain evolution being too far from the ancestral molecule (Lagomarsino *et al.,* 2009). Hence, evolution of new functions following subfunctionalization could be a rare event. Therefore, the presence of fewer domains in the duplicate genes compared to the single copy genes could be due to evolution of duplicate genes by subfunctionalization, where complementary loss of subfunctions is viewed to primarily facilitate preservation of the duplicate gene. Alternatively, the addition of new domains into a bacterial genome could be due to acquisition by HGT. Indeed, acquisition of one or more domains by HGT in 30 to 50 percent of bacteria has been reported (Choi & Kim, 2007). The acquired gene or gene segment is known to be beneficial only if it has some properties different to that of recipient genome (Kinsella *et al.,* 2003). Since the selection of a new domain would depend upon its ability to perform a biological function that is not covered by pre-existing domains, addition of such rare domains by HGT could be an uncommon phenomenon (Lagomarsino *et al.,* 2009).

Adaptation of bacteria to new environments requires evolution of new functions (Hooper & Berg, 2003), and gene duplication is viewed to be the general mechanism of adaptation to different environmental conditions (Kondrashov, 2002). However, a recent study suggests HGT to be a far more important route to adaptation compared to gene duplication (Koonin & Wolf, 2009). Further, duplication of horizontally transferred genes with weak or no functions is suggested to accelerate the evolutionary process of gene innovation. Since both gene duplication and HGT are considered to be important routes of bacterial adaptation to changing environments, and amplification of weak ancillary functions is considered to be the easiest route to gene innovation, quick adaptation of bacteria to changing environments could be due to amplification of weak ancillary functions. Thus, reduced functional complexity of the investigated duplicate genes compared to single copy genes could be due to preservation of the majority of the paralogs by subfunctionalization, and the rare event of neofunctionalization could have been either due to divergence of subfunctions over an evolutionary period of time following preservation of subfunctionalized paralogs, or mostly due to rapid amplification of weak ancillary functions after gene duplication. To gain deeper insights into the functional complexity of duplicate genes in *M. tuberculosis*, we focussed on the evolutionary analysis of the duplicate genes in six of the closely related mycobacterial species.

From the analysis of maximum genetic distance between the two most distant proteins of the mycobacterial multiple genome clusters, we suggest that the divergence of at least one of the duplicate gene copies from the ancestral gene increases with the increase in cluster size. These homologous gene families consist of orthologs and paralogs. The lack of a strong correlation between the average genetic distance and cluster size of the duplicate gene copies in the multiple genome clusters indicates that the homologous gene families including proteins from different mycobacterial species have not undergone complete functional divergence. This is to be expected for orthologs, which tend to maintain their functions.

The average genetic distance estimated for single genome paralogous gene clusters, on the other hand, decreases with the increase in cluster size, suggesting that, on average, smaller families tend to diverge more rapidly than the larger families. This is apart from some members of the larger families, which have obviously diverged further as they are contributing to the increased maximum genetic distance with cluster size. Though gene duplication is considered to be an important mechanism for acquiring new genes, and creating evolutionary novelty (Torgerson and Singh, 2004), horizontal gene transfer (HGT) is also known to be a wide spread phenomenon, and a significant proportion of genes in bacteria are accepted to have been acquired by HGT (Price *et al.*, 2007). The genome of *M. tuberculosis* is known to contain 19 genes of eukaryotic origin, and it is speculated that the organism may have also acquired genes from other prokaryotes by HGT (Kinsella *et al.*, 2003). In addition, the occurrence of many intraspecies HGT events in the progenitor of *M. tuberculosis* has been reported (Rosas-Magallanes *et al.*, 2006). The ability of HGT to incorporate a new gene which is homologous to an existing gene family member is well recognized (Ochman, 2001; Kinsella *et al.*, 2003; Krzywinska, 2004), and in comparison to its gene family members, the newly introduced gene may be more divergent in sequence and function (Pushker *et al.*, 2004). Following duplication, such laterally transferred genes with already divergent functions may further diversify in the process of evolving new functions, and this could result in an increase in genetic distance between the laterally transferred duplicate gene and its paralog gene family members. The chance of this should increase with the number of members.

Phylogenetic analysis of the sigma factors in *M. tuberculosis* suggested that most of the sigma factors have orthologs in other mycobacteria. However, we could not observe orthologs in *M. leprae* for a few of the subfamilies, and this could have been due to the extensive loss of sigma factors during its reductive genome evolution. Agarwal *et al.,* 2007 reports that sigM proteins control only a small subset of genes, and their loss would not influence *M. tuberculosis* virulence (Agarwal *et al.,* 2007). The difference in the divergence of sigM in *M. tuberculosis* compared to other mycobacteria, and absence of its ortholog in *M. bovis* should be considered further to study the importance of the sigM factor in *M. tuberculosis* virulence.

5. Conclusions and future work

The estimated duplicate gene percentages for *M. tuberculosis* from independent genome clustering (31%), InterPro signature methods (38%), across genome clustering (49%) and a union of the methods (51%) were all relatively high, showing the significance of gene duplication in *M. tuberculosis* genome evolution. The investigation of relationships between the GC composition and duplicate gene percentages identified from the sequence and InterPro domain data provides sufficient evidence to suggest that for the mycobacterial species, with the exception of *M. leprae*, the maintenance of high duplicate gene percentages

could be attributed to the high GC composition of their genomes. The study has also shown a correlation between duplicate gene percentage and genome size, suggesting that gene duplication increases genome size, which is a logical result. Interestingly, our functional complexity results were in contrast to recent finding in eukaryotes, and we show that, on average, duplicate genes have longer sequences but fewer domains than single copy genes in the investigated organisms. The reduced functional complexities of duplicate genes could be due to their evolution by subfunctionalization following duplication.

We also show that duplicate gene families of mycobacterial multiple genome clusters have not undergone complete functional divergence following gene duplication and still tend to maintain their functions. Our maximum genetic distance results suggest that multiple duplication events in a few of the duplicate copies of bigger families may result in their functional divergence from the original ancestral functions. Our paralog maximum genetic distance results suggest that the increase in genetic distance between the two most distant proteins with the size of the gene family may be due to duplication followed by paralog evolution of some of the distant genes that already have divergent functions compared to its paralog members. From the study of average genetic distance of paralogs, we suggest that slow evolution of paralogs of large families in *M. tuberculosis* could be due to preservation of original ancestral functions by the mechanism of subfunctionalization.

For future studies, we are investigating selection pressure and comparison between smaller and larger gene family evolution to shed light on the evolutionary fate of duplicate genes and functional innovation in *M. tuberculosis*. In addition, since the functional constraints on amino acid residues are known to differ due to the potential changes in protein function, we have studied site specific rate differences between the amino acids of closely related mycobacterial species to aid in deciphering specific subfamily evolutionary divergence following gene duplication. Such predicted critical amino acids when mapped on to protein secondary structure could help in evaluation of important structural locations in functional diversification. In addition to functional divergence, gene expression data from different experimental conditions is of use to understand the degree of expression divergence of the genes following duplication events. Overall, we are working on further investigation of *M tuberculosis* duplicate genes with the integration of phylogeny-sequence-structure-function-expression information, which will be valuable for understanding the functional and evolutionary fate of genes following gene duplication in *M. tuberculosis*.

6. Acknowledgement

We thank the National Bioinformatics Network and Computational Biology Group, University of Cape Town, South Africa for supporting this work.

7. References

Abascal, F.; Zardoya, R. & Posada, D. (2005). ProtTest: selection of best-fit models of protein evolution. *Bioinformatics Applications Note,* Vol. 21, No. 9, pp. 2104–2105.

Agarwal, N.; Woolwine SC, Tyagi S, Bishai WR. (2007). Characterization of the *Mycobacterium tuberculosis* Sigma Factor SigM by Assessment of Virulence and Identification of SigM-Dependent Genes. *Infection and Immunity,* Vol. 75, No 1, pp. 452–461.

Apweiler, R.; Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD, Durbin R, Falquet L, Fleischmann W, Gouzy J, Hermjakob H, Hulo N, Jonassen I, Kahn D, Kanapin A, Karavidopoulou Y, Lopez R, Marx B, Mulder NJ, Oinn TM, Pagni M, Servant F, Sigrist CJ, Zdobnov EM. (2001). The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Research*, Vol. 29, No. 1, pp. 37-40.

Babu, MM. (2003). Did the loss of sigma factors initiate pseudogene accumulation in *M. leprae*? *Trends in Microbiology,* Volume 11, No. 2, pp. 59–61.

Basak, S. & Ghosh, TC. (2005). On the origin of genomic adaptation at high temperature for prokaryotic organisms. *Biochemical and Biophysical Research Communications,* Vol. 330, No. 3, pp. 629-632.

Castresana, J. (2000). Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Molecular Biology and Evolution,* Vol. 17, Vol. 4, pp. 540–552.

Choi, I. & Kim, S. (2007). Global extent of horizontal gene transfer. *Proceedings of National Academy of Science*, Vol. 104, No. 11, pp. 4489–4494.

Cordero, OX. & Hogeweg, P. (2009). The impact of long-distance horizontal gene transfer on prokaryotic genome size. *Proceedings of National Academy of Science*, Vol. 106, No. 51, pp. 21748–21753.

Cun Y. (2010). The Evolutionary Dynamics of Mutant Allele at Duplicate Loci *arXiv*:1007.0333v1.

DeRose-Wilson, LJ. & Gaut, BS. (2007). Transcription-related mtations and GC content drive variation in nucleotide substitution rates across the genomes of *Arabidopsis thaliana* and *Arabidopsis lyrata. BMC Evolutionary Biology*, Vol. 7, No. 66.

Fontan, PA.; Voskuil MI, Gomez M, Tan D, Pardini M, Manganelli R, Fattorini L, Schoolnik GK, Smith I. (2009). The Mycobacterium tuberculosis Sigma Factor B Is Required for Full Response to Cell Envelope Stress and Hypoxia In Vitro, but It Is Dispensable for In Vivo Growth. *Journal of Bacteriology*, Volume 191, No. 18, pp. 5628–5633.

Fontan, PA.; Aris V, Alvarez ME, Ghanny S, Cheng J, Soteropoulos P, Trevani A, Pine R, and Smith I. (2008). Mycobacterium tuberculosis Sigma Factor E Regulon Modulates the Host Inflammatory Response. Vol. 198, pp. 877-85.

Force, A.; Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics,* Vol. 151, No. 4, pp. 1531-1545.

Fraser-Liggett, CM. (2005). Insights on biology and evolution from microbial genome sequencing. *Genome Research,* Vol. 15, No. 12, pp.1603–1610.

Guindon, S. & Gascuel, O. (2003). A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology,* Vol. 52, No. 5, pp. 696–704.

Hamady, M; Betterton, MD., & Knight, R. (2006). Using the nucleotide substitution rate matrix to detect horizontal gene transfer. *BMC Bioinformatics*, Vol. 7, No. 476.

He, X. & Zhang, J. (2005). Gene Complexity and Gene Duplicability. *Current Biology*. Vol. 15, No. 11, pp. 1016-1021.

Hooper, DS. & Berg, GO. (2003). On the Nature of Gene Innovation: Duplication patterns in Microbial Genomes. *Molecular Biology and Evolution*, Volume 20, No. 6, pp. 945-954.

Kersey, P.; Bower L, Morris L, Horne A, Petryszak R, Kanz C, Kanapin A, Das U, Michoud K, Phan I, Gattiker A, Kulikova T, Faruque N, Duggan K, Mclaren P, Reimholz B, Duret L, Penel S, Reuter I and Apweiler R. (2005). Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. *Nucleic Acids Research,* Vol. 33, Database issue, D297–D302.

Kinsella, RJ.; Fitzpatrick DA, Creevey CJ, McInerney JO. (2003). Fatty acid biosynthesis in Mycobacterium tuberculosis: Lateral gene transfer, adaptive evolution, and gene duplication. *Proceedings of National Academy of Science*, Vol. 100, No. 18, pp. 10320–10325.

Kondrashov, FA.; Rogozin IB, Wolf YI, Koonin EV. (2002). Selection in the evolution of gene duplications. *Genome Biology,* Vol. 3, No. 2.

Koonin, EV. & Wolf YI. (2008). Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Research*, Vol. 36, No. 21.

Koonin, EV. & Wolf, YI. (2009). Is evolution Darwinian or/and Lamarckian? Biology Direct, *Vol. 4, No. 42.*

Krzywinska, E.; Krzywinski J, & Schorey JS. 2004. Naturally occurring horizontal gene transfer and homologous recombination in *Mycobacterium*. *Microbiology,* Vol. 150, No. Pt 6, pp. 1707-1712.

Lagomarsino, MC. (2009). Universal features in the genome-level evolution of protein domains. Genome Biology, *Vol. 10, No. 1.*

Manganelli, R.; Provvedi R, Rodrigue S, Beaucher J, Gaudreau L, Smith I. (2004). σ Factors and Global Gene Regulation in *Mycobacterium tuberculosis. Journal of Bacteriology*, Volume 186, No. 4, pp. 895–902.

Marri, PR.; Bannantine, JP. & Golding, GB. (2006). Comparative genomics of metabolic pathways in mycobacterium species: gene duplication, gene decay and lateral gene transfer *FEMS. Microbiology* Review, Vol. 30, No. 6, pp. 906-925.

Mulder, NJ; Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Buillard V, Cerutti L, Copley R, Courcelle E, Das U, Daugherty L, Dibley M, Finn R, Fleischmann W, Gough J, Haft D, Hulo N, Hunter S, Kahn D, Kanapin A, Kejariwal A, Labarga A, Langendijk-Genevaux PS, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Nikolskaya AN, Orchard S, Orengo C, Petryszak R, Selengut JD, Sigrist CJ, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C. (2007). New developments in the InterPro database. *Nucleic Acids Research*. Vol. *35*, No. D224-D228.

Musto, H.; Naya H, Zavala A, Romero H, Alvarez-Valín F, and Bernardi G. (2006). Genomic GC level, optimal growth temperature, and genome size in prokaryotes. *Biochemical and Biophysical Research Communications,* Vol. 347, No. 1, pp. 1-3.

Mann, S. & Chen, YP. (2010). Bacterial genomic G + C composition-eliciting environmental adaptation. *Genomics*, Vol. 95, No. 1, pp. 7-15.

Mann, S.; Li, J. & Chen YP. (2010). Insights into Bacterial Genome Composition through Variable Target GC Content Profiling. *Journal of Computational Biology*, Vol. 17, No. 1, Pages 79-96.

Nadeau, JH. & Sankoff, D. (1997). Comparable Rates of Gene LOSS and Functional Divergence After Genome Duplications Early in Vertebrate Evolution. *Genetics,* Vol. 147, No. 3, pp. 1259-1266.

Naya, H.; Romero H, Zavala A, Alvarez B, Musto H. (2002). Aerobiosis Increases the Genomic Guanine Plus Cytosine Content (GC%) in Prokaryotes. *Journal of Molecular Evolution,* Vol. 55, No. 3, pp. 260-264.

Nelson, KE.; Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, McDonald L, Utterback TR, Malek JA, Linher KD, Garrett MM, Stewart AM, Cotton MD, Pratt MS, Phillips CA, Richardson D, Heidelberg J, Sutton GG, Fleischmann RD, Eisen JA, White O, Salzberg SL, Smith HO, Venter JC, Fraser CM. (1999). Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritime.* Nature, Vol. 399, pp. 323-329.

Notredame, C.; Higgins, DG. & Heringa J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. Journal of Molecular Biology, Vol. 302, No. 1, pp. 205-217.

Ochman, H. 2001. Lateral and oblique gene transfer. Current Opinion in Genetics & Development, Vol. 11, No. 6, pp. 616–619.

Price, MN.; Dehal PS, & Arkin AP. 2007. Orthologous transcription factors in bacteria have different functions and regulate different genes. *PLoS Computational Biology,* Vol. 3, No. 9, pp. 1739-1750.

Pushker, R.; Mira A, & Rodriguez-Valera F. (2004). Comparative genomics of gene-family size in closely related bacteria. *Genome Biology*, Vol. 5, No. 4.

Rosas-Magallanes, V.; Deschavanne P, Quintana-Murci L, Brosch R, Gicquel B, Neyrolles O. 2006. Horizontal Transfer of a Virulence Operon to the Ancestor of Mycobacterium tuberculosis. *Molecular Biology and Evolution,* Vol. 23, No. 6, pp. 1129–1135.

Snel. B.; Bork, P. & Huynen MA. (2001). Genomes in Flux: The Evolution of Archaeal and Proteobacterial. *Gene Content Genome Res*earch, *Vol 12*, No. 1, pp.17–25.

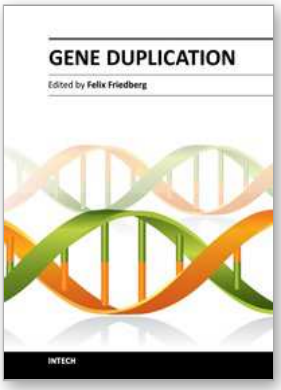
Stoltzfus, A. (1999). On the possibility of constructive neutral evolution. *Journal of Molecular Biology*, *Vol 49*, No. 2, pp. 169–181.

Tatusov, RL.; Koonin, EV. & Lipman DJ. (1997). A Genomic Perspective on Protein Families. *Science,* Vol. 278, No. 631.

Tekaia, F.; Gordon SV, Garnier T, Brosch R, Barrell BG, Cole ST. (1999). Analysis of proteome of mycobacterium tuberculosis in silico. *Tubercle and Lung Diseases*, Vol. 79, No. 6, pp. 329-342.

Torgerson, DG. & Singh RS. 2004. Rapid Evolution Through Gene Duplication and Subfunctionalization of the Testes-Specific α4 Proteasome Subunits in Drosophila. *Genetics,* Vol. 168, No. 3, pp. 1421–1432.

Zhang, L.; Kasif S, Cantor CR, Broude NE. (2004). GC/AT-content spikes as genomic punctuation marks. *Proceedings of National Academy of Science,* Vol. 101, No. 48, pp. 16855–16860.



Gene Duplication

Edited by Prof. Felix Friedberg

ISBN 978-953-307-387-3
Hard cover, 400 pages
Publisher InTech
Published online 17, October, 2011
Published in print edition October, 2011

The book Gene Duplication consists of 21 chapters divided in 3 parts: General Aspects, A Look at Some Gene Families and Examining Bundles of Genes. The importance of the study of Gene Duplication stems from the realization that the dynamic process of duplication is the "sine qua non" underlying the evolution of all living matter. Genes may be altered before or after the duplication process thereby undergoing neofunctionalization, thus creating in time new organisms which populate the Earth.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Venu Vuppu and Nicola Mulder (2011). Analysis of Duplicate Gene Families in Microbial Genomes and Application to the Study of Gene Duplication in M. tuberculosis, Gene Duplication, Prof. Felix Friedberg (Ed.), ISBN: 978-953-307-387-3, InTech, Available from: http://www.intechopen.com/books/gene-duplication/analysis-of-duplicate-gene-families-in-microbial-genomes-and-application-to-the-study-of-gene-duplic

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821