

Chapter 3

The Seqword Genome Browser

A large part of the work in this chapter was published as:

Ganesan H., Rakitianskaia AS., Davenport CF., Tümmler B., Reva ON. (2008) The Seq-Word Genome Browser: an online tool for the identification and visualization of atypical regions of bacterial genomes through oligonucleotide usage. *BMC Bioinformatics* **7**, p333.

3.1 Introduction

The greater FunGIM system includes the Seqword Genome Browser (SWGB) which is a sub-project of the Seqword research team. The SWGB aims to visualize whole bacterial genomes on the level of oligonucleotide usage (OU) statistics thus providing a novel method of genome visualization. The specifics of OU statistics and the SWGB will be discussed.

3.2 Background

The study of genome OU signatures has a long history dating back to early publications by Karlin *et al.*, who focused mainly on dinucleotide compositional biases and their evolutionary implications (Karlin & Burge, 1995; Karlin, 1995; Pride *et al.*, 2003). Statistical approaches of OU comparison were further advanced by Deschavanne *et al.* (1999) who applied chaos game algorithms and by Pride *et al.* (2003) who extended the analysis to tetranucleotides using Markov Chain Model simulations. Later, a number of practical tools for phylogenetic comparison of bacterial genomes (Deschavanne *et al.*, 1999; Coenye & Vandamme, 2004; van Passel *et al.*, 2006), identification of horizontally transferred genomic islands (Mrazek & Karlin, 1999; ; Becq *et al.*, 2007; Dufraigne *et al.*, 2005; Nakamura *et al.*, 2004; Pride & Blaser, 2002) and assignment of unknown genomic sequences (Abe *et al.*, 2003; Teelin *et al.*, 2004) based on OU statistics became

publicly available. These approaches exploited the notion that genomic OU composition was less variable within genomes rather than between them, regardless of which genomic regions had been taken into consideration (Jernigan & Baran, 2002). A general belief was that if a significant compositional difference was discovered in genomic fragments relative to the core genome, these loci most likely can be assigned to horizontally transferred genetic elements (transposons, prophages or integrated plasmids). This approach was criticized by several researchers (Koski *et al.*, 2001; Wang, 2001) who pointed out that codon bias and base composition are poor indicators of horizontal gene transfer. Therefore, there is a need for more informative parameters which also take into account higher order DNA variation. An overview of the current OU statistical methods based on di-, tetra- and hexanucleotides has been published recently. The conclusion of the review was that all methods were context dependent and, though being efficient and powerful, none of them were superior in all applications (Bohlin, 2008). Thus, the major motivation in this work was to develop more flexible and informative algorithms seamlessly integrating di- to heptanucleotides OU analysis for reliable identification of divergent genomic regions.

Recently the concept of OU patterns was introduced into the literature (Reva & Tummler, 2004). Each OU pattern is characterized by a number of OU statistical parameters namely, local pattern deviation (D), pattern skew (PS), relative variance (RV) and others (see Methods section). Novelty of the developed algorithms relative to other existing methods include the following: i) distances between patterns of different word length (from di- through to heptanucleotides) calculated for the same sequences are comparable; i.e. one may use longer word patterns to perform a large scale analysis and then switch to shorter word patterns for a more detailed view; ii) OU patterns calculated for sequences of different lengths are comparable provided that the length of the sequence is longer than the corresponding thresholds (specified in the Methods section); iii) alterations of OU patterns may be analyzed by different non-redundant parameters (D, PS and RV with different schemes of normalization by frequencies of shorter constituent words). Superimposition of these OU characteristics allows better discrimination of divergent genomic regions relative to other contemporary approaches (Reva & Tummler, 2005). This is described by:

$$\Delta_{[\xi_1 \dots \xi_N]} = (C_{[\xi_1 \dots \xi_N]_{obs}} - C_{[\xi_1 \dots \xi_N]_e}) / C_{[\xi_1 \dots \xi_N]_0}$$

where ξ_n is any nucleotide A, T, G or C in the N-long word; $C_{[\xi_1 \dots \xi_N]_{obs}}$ is the observed count of the word $[\xi_1 \dots \xi_N]$; $C_{[\xi_1 \dots \xi_N]_e}$ is the expected count and $C_{[\xi_1 \dots \xi_N]_0}$ is a standard count estimated from the assumption of an equal distribution of words in the sequence: ($C_{[\xi_1 \dots \xi_N]_0} = L_{seq} \times 4^{-N}$).

Expected counts of words $C_{[\xi_1 \dots \xi_N]_e}$ were calculated in accordance with the applied normalization scheme. Thus, $C_{[\xi_1 \dots \xi_N]_e} = C_{[\xi_1 \dots \xi_N]_0}$ if OU is not normalized, or $C_{[\xi_1 \dots \xi_N]_e} = C_{[\xi_1 \dots \xi_N]_n}$ if OU is normalized by empirical frequencies of all shorter words of the length n. The expected count of a word $C_{[\xi_1 \dots \xi_N]_e}$ of length N in a L_{seq} long sequence normalized by frequencies of n-mers ($n < N$) was calculated as follows:

$$C_{[\xi_1 \dots \xi_w]n} = L_{seq} \times F_{[\xi_1 \dots \xi_n]} \times \prod_{i=2}^{N-n+1} \left(\frac{F_{[\xi_i \dots \xi_{i+n-1}]\xi_{i+n}}}{\sum_{\xi \in \{A, T, G, C\}} F_{[\xi_i \dots \xi_{i+n}]\xi}} \right)$$

where the $F_{[\xi_1 \dots \xi_N]}$ values are the observed frequencies of the particular word of length n in the sequence and ξ is any nucleotide A, T, G or C. For example, expected count of a word ATGC in a sequence of L_{seq} nucleotides normalized by frequencies of trinucleotides is:

$$C_{ATGC} = L_{seq} \times F_{ATG} \times \frac{F_{TGC}}{F_{TGA} + F_{TGT} + F_{TGG} + F_{TGC}}$$

Two approaches of normalization have been exploited where the F values were calculated for the complete sequence of a chromosome, plasmid, etc (generalized normalization) or for a given sliding window (local normalization). The normalization by equation 2 allows identification of words, frequencies of which cannot be predicted exactly by frequencies of shorter constituent words.

The distance D between two patterns was calculated as the sum of absolute distances between ranks of identical words (w , in a total 4^N different words) after ordering of words by $\Delta_{[\xi_1 \dots \xi_N]}$ values (see equation 1) in patterns i and j as follows:

$$D(\%) = 100 \times \frac{\sum_{w=1}^{4^N} |rank_{w,i} - rank_{w,j}| - D_{min}}{D_{max} - D_{min}}$$

Application of ranks instead of relative oligonucleotide frequency statistics made the comparison of OU patterns less biased to the sequence length provided that the sequences are longer than the limits of 0.3, 1.2, 5, 18.5, 74 and 295 kbp for di-, tri-, tetra-, penta-, hexa- and heptanucleotides, respectively (Reva & Tummler, 2004)

PS is a particular case of D where patterns i and j were calculated for the same DNA but for direct and reversed strands, respectively. $D_{max} = 4^N \times (4^N - 1)/2$ and $D_{min} = 0$ when calculating a D or, in a case of PS calculation, $D_{min} = 4^N$ if N is an odd number or $D_{min} = 4^N - 2^N$ if N is an even number due to presence of palindromic words (Reva & Tummler, 2004) Normalization of D-values by Dmax ensures that the distances between two sequences are comparable regardless of the word length of OU patterns.

Relative variance of an OU pattern was calculated by the following equation:

$$RV = \frac{\sum_{w=1}^{4^N} \Delta_w^2}{\left(4^N - 1\right) \sigma_0^2}$$

where N is word length; Δ_w^2 is the square of a word w count deviation (see equation 1); and σ_0^2 is the expected variance of the word distribution in a randomly generated sequence that depends on the sequence length and the word length:

$$\sigma_0^2 = 0.14 + \frac{4^N}{L_{seq}}$$

where L_{seq} is sequence length, and N is word length. Normalization of OU pattern variance by σ_0 makes the variances comparable regardless of the word length of OU patterns and the sequence length. The regression equation was tested on 300 randomly generated sequences with an equiprobable occurrence of all 4 nucleotides by the DataFit 7.1.44 software. The SWGB is coded in Java to be used as an applet in a Web-browser either on the Internet or locally (the programs OligoWords in Python and SeqWord_Viewer, which respectively calculate and visualize the OU patterns for DNA sequences, are available for download from the SWGB website). SWGB should run on any platform with a Java 1.5.x runtime environment or newer. The pre-calculated data-sets are saved in a MySQL Server 5.0 database. The size of the sliding window and the OU pattern type were applied according to the sequence length (Table 3.1) At the time of writing, the SeqWord database contained OU patterns pre-calculated for the sequences of 682 bacterial chromosomes belonging to 637 different organisms (strains and species), 412 plasmids, 100 bacteriophages and 39 other viruses, which were downloaded from the NCBI (14).

Table 3.1: Sliding window size and OU pattern types (oligomer lengths) selected for sequences of different length present in the SeqWord database.

Sequence length	Sliding window	Step	OU pattern type
> 2 Mbp	8 kbp	2 kbp	4 mer
from 1 mbp to 2 Mbp	5 kbp	0.5 kbp	4 mer
from 0.5 mbp to 1 Mbp	3 kbp	0.3 kbp	3 mer
< 0.5 Mbp	1.5 kbp	0.15 kbp	3 mer

3.3 Results

User familiarity with the abbreviations of the various OU statistical parameters is important. Different types of OU patterns were abbreviated as type_Nmer. Types might be "n0" for non-normalized, or "n1" for normalized by mononucleotide frequencies. For example, the non-normalized tetranucleotide usage pattern is denoted as n0_4mer; tetranucleotide usage pattern normalized by mononucleotide content is n1_4mer etc. The genomes in the SWGB database were

analyzed by the following statistical parameters: D – distance between two patterns of the same type (in this work we used distances (D) between local patterns calculated for overlapping genome fragments and the global genome patterns calculated for the complete sequence – the local pattern deviation); PS – pattern skew, distance between the two patterns of the direct and reverse strands of the same DNA sequence; RV and GRV – oligonucleotide usage variances normalized locally and globally, respectively, and reduced to the OU variance expected for a randomly generated sequence (see Background section); GC-content (GC) and GC-skew (GCS) in DNA fragments. The SeqWord Genome Browser (SWGB) applet is available via the Internet through mirror sites (University of Pretoria, South Africa [<http://www.bi.up.ac.za/SeqWord/mhhapplet.php>]; Hannover Medical School, Germany [<http://genomics1.mh-hannover.de/seqword/genomebrowser/mhhapplet.php>]; Penn State University, USA [<http://seqword.bx.psu.edu/mhhapplet.php>]) and is mouse and menu driven. The Web-based applet is used to visualize DNA compositional variations in bacterial and viral genomes stored in the SeqWord database. Every genome in the database is represented by a set of statistical OU parameters (D, PS, GV, GRV, GC and GCS) calculated for genomic fragments, which were selected by a sliding window (sliding window length and step were set according to the total length of the sequence as demonstrated in Table 3.1). While in 70 to 99% of genomic fragments the OU compositional bias is similar to the complete genome OU pattern, some regions with atypical OU composition, however, are always present. Superimposition of different OU parameters allows discrimination of divergent genomic regions, as was published previously (Reva & Tummler, 2005). Briefly: rRNA operons are characterized by extremely high PS and low RV; giant genes with multiple repeated elements have high or moderate PS and high RV; horizontally transferred genetic elements are characterized by increased divergence between RV and GRV accompanied by high D; and genes for ribosomal proteins show a moderate increase of D, PS and RV above genomic averages. Having analyzed 1243 sequences of different microorganisms including viruses and plasmids in the SeqWord database, it was confirmed that the approaches developed and tested previously (Reva & Tummler, 2008) (mainly on *Pseudomonas putida* KT2440 chromosomal DNA) are appropriate and useful for analysis of genomic sequences of other microorganisms and viruses. In an open applet window, the user has the ability to choose from an ever growing list of available sequences (Figure 3.1) The user also has the option of restricting the list to display only bacterial chromosomes, plasmids, phages, viruses or all sequences by selecting the corresponding filter button. Users have to select a genome in the list and click the 'Display in the Applet' button to retrieve the pre-calculated data. All OU parameters calculated for a given genome may be exported to a local text file by using the 'Export' function from the applet's 'File' menu. Later, instead of again having to connect to the database, users may open and view their local files (previously exported from the applet or calculated by the OligoWords program, see below) via the 'Open' function in the 'File' menu.

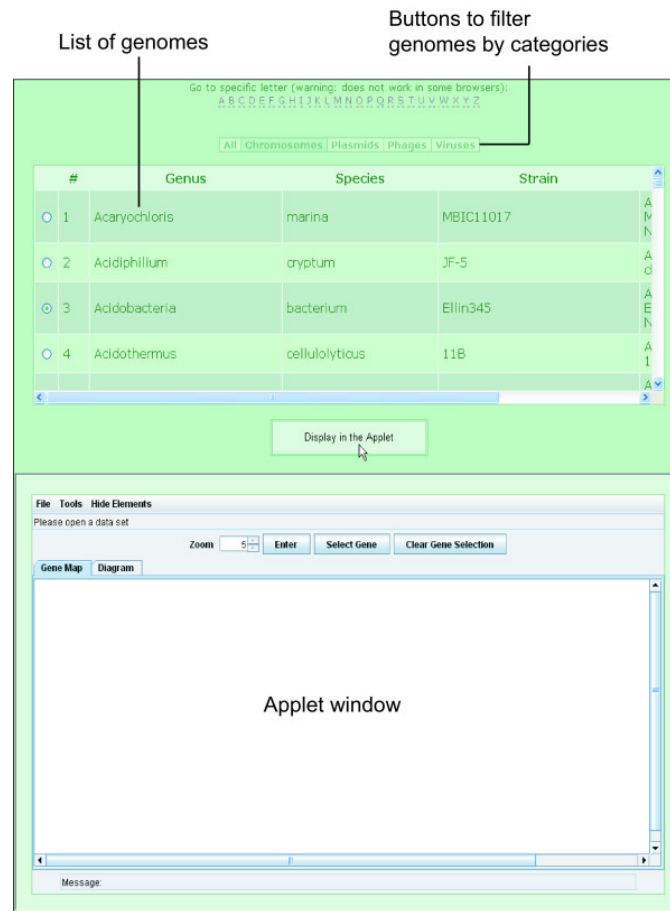


Figure 3.1: General view of the web-based SWGB with a list of genomes present in the database and an enclosed Java applet for data visualization. To show OU statistical parameters for a selected genome, click the 'Display in the Applet' button. Click a filter button to order genomes by the corresponding category and use the interactive letters at the top to scroll the list to a sequence of interest.

The SWGB is basically comprised of two views, denoted by the 'Gene Map' and 'Diagram' tabs. The applet is instrumental for visualization of natural variation in DNA sequences by the interactive diagrams on the 'Gene Map' and 'Diagram' tabs. Users may save the current diagram in JPG format by using the 'Save picture' function in the 'File' menu. The 'Gene Map' tab offers a simple view of an entire genome at a glance and gives users access to a number of important pre-calculated OU statistics superimposed on the gene map (Figure 3.2) Displays for each of the statistical parameters can be toggled on/off by checking items in the 'Hide Elements' menu. By merely mousing over any region on the plot, a message displaying detailed information for the pointed curve will be shown in the 'Message' bar. Clicking a gene on the map displays a dialog with the annotation details (Figure 3.2).

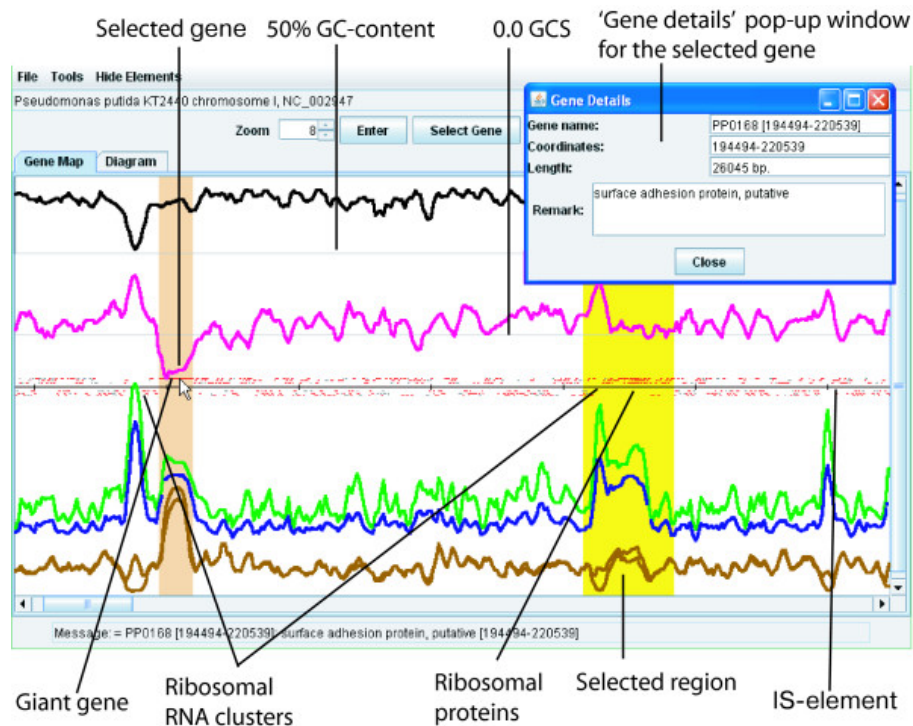


Figure 3.2: Identification of divergent genomic regions on the 'Gene Map' view. Superimposition of different OU parameters such as GC (black line), GCS (pink), PS (green), D (blue), GRV (upper brown line) and RV (lower brown line) allows discrimination of divergent genomic regions. In this example a part of the chromosome of *Pseudomonas putida* KT2440 (127–774 kbp) is displayed in the applet window. A genomic fragment was highlighted using the function 'Select region' and a giant gene, PP0168, was selected by 'Select gene'. A pop-up window 'Gene Details' was opened by double-clicking the gene on the map. Genes are indicated by red and grey (for hypotheticals) bars. The black horizontal line separates genes by their direction of translation.

The 'Zoom' function is straight-forward and allows users to control the amount of data viewed in the plot area. Clicking the 'Enter' button after setting the desired zoom value will then redraw the map. A 'Zoom into region' function under the 'Tools' drop-down menu allows users to zoom into exact genomic regions by merely entering their desired co-ordinates into the pop-up dialog box. The 'Tools' → 'Select region' menu item allows highlighting of selected regions without zooming. Use the option 'Clear ...' in the 'Tools' menu to undo zooming or highlighting. To locate a genomic region by gene, click the button 'Select Gene'. In the pop-up dialog box one may order the gene list by gene names, functionality or coordinates, then select a gene in the list and click 'OK'. When a gene annotation is not available, the values of the locus coordinates are used as a gene name. The applet window will be scrolled to the selected gene highlighted on the map (see Figure 3.2).

The 'Diagram' tab allows flexible filtering of the underlying data based on the criteria chosen by users. Although the underlying data is pre-calculated, the user may, by simply changing selected parameters, generate very different images which give different insights into the natural genomic variation. To start with, the 'Diagram' view offers a bar chart or a dot-plot presentation of the pre-calculated data. To view a bar chart of the distribution statistics for a given OU parameter, select the desired parameters from the X or Y-axis drop-downs and click 'Enter'. The number of bars displayed can be adjusted using the '# Bars' selector.

On the dot-plot diagram, each genomic fragment (selected by the sliding window) is represented by a dot with X and Y coordinates that correspond to values of OU parameters chosen from X and Y drop-down lists, respectively. The Z axis parameter may be set as well. In this case, the dots are colored by values of OU parameters selected for the Z axis, and the color range is displayed on the vertical color bar on the left of the plot area (Figure 3.3).

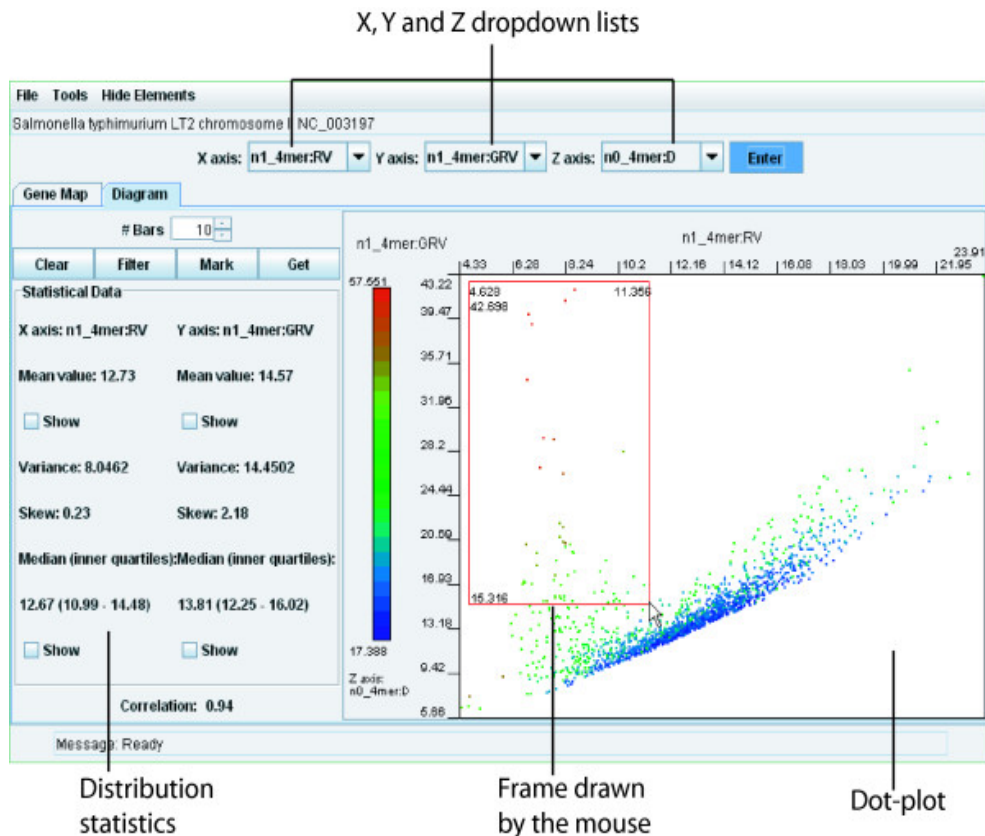


Figure 3.3: The 'Diagram' view. To draw a diagram, first select corresponding OU parameters using the dropdown lists and click the 'Enter' button. In this example n1_4mer:RV, n1_4mer:GRV and n0_4mer:D were selected for the X, Y and Z axes, respectively. Every dot on the dot-plot corresponds to a genomic fragment selected by the sliding window. Dots are spread and colored in accordance with their values of the selected statistical OU parameters. Information for each dot may be found by one of the following methods: i) information for a dot under the mouse pointed by the mouse is shown in the 'Message' bar; ii) double clicking a dot returns us to the 'Gene map' tab with the corresponding genomic fragment highlighted; iii) framing the dots and clicking the 'Get' button opens a new applet window with the information about all selected regions. In this example the genomic regions of *Salmonella typhimurium* LT2 (NC_003197) that correspond to horizontally transferred genetic elements were selected (see discussion in the text).

Having set up the dot-plot, users will be able to identify divergent genomic regions (see next section). To retrieve annotations of genomic fragments corresponding to a group of dots, frame the dots of interest by clicking and dragging over the desired area. A selector frame then appears around the dots (Figure 3.3). Clicking the 'Get' button displays the selected genomic fragments with their coordinates and gene annotations. Furthermore, identification and isolation of specific genomic regions may be improved significantly by filtering dots by OU parameters. The simplest way of filtering is by the third (Z axis) parameter. One may select an area on the color bar to exclude all dots from the plot lying outside of the selected color range (see an example in help

files on-line). The hidden dots will not be selected by the 'Get' button. A more sophisticated way to filter genomic regions is provided by the 'Filter' button. An example will be discussed below. The 'Mark' button enables genomic fragments to be selected by their coordinates and highlighted on the dot-plot. Click the 'Mark' button to open a dialog and enter coordinates of one or multiple fragments (Figure 3.4). Co-ordinates of each fragment must be added to the list by clicking the 'Add' button. Close the dialog by clicking 'OK'. The corresponding dots on the dot-plot will be highlighted as shown in Figure.3.4.€

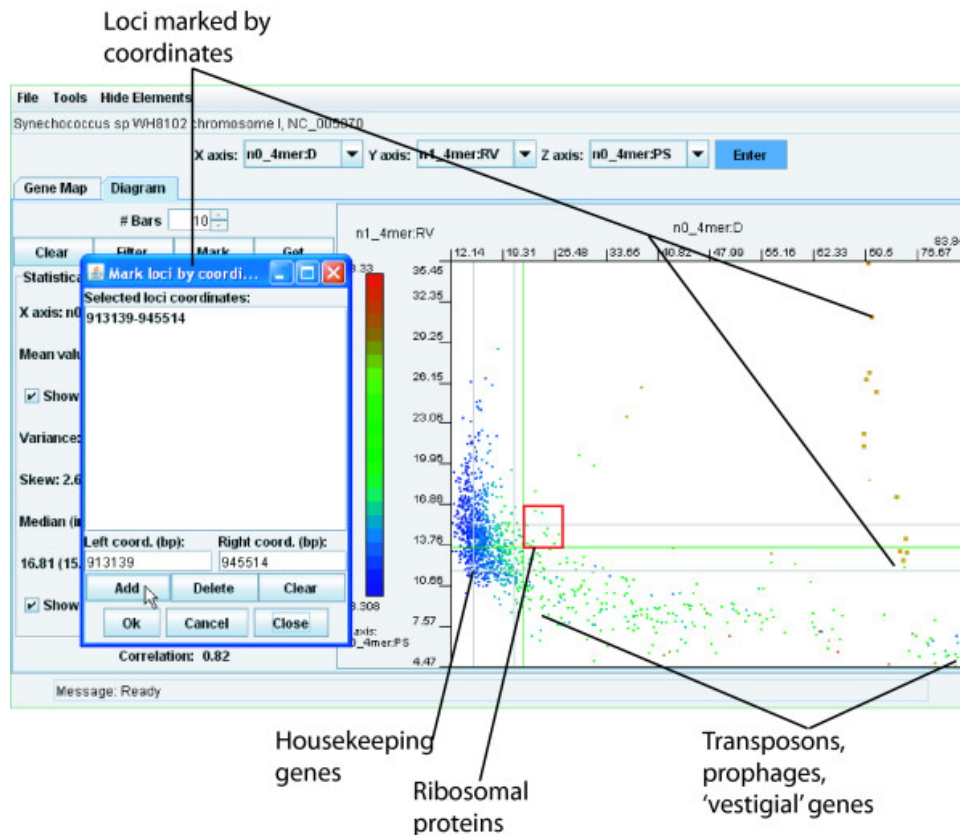


Figure 3.4: Identification of divergent genomic regions by plotting and highlighting. In this example the genome of *Synechococcus* sp. WH8102 was analyzed. The parameters n0_4mer:D, n1_4mer:RV and n0_4mer:PS were selected for the X, Y and Z axes, respectively. The genomic regions covering the giant gene for the surface protein SwmB (Reva & Tummeler, 2008) were highlighted by entering the coordinates of this gene into the 'Mark loci by coordinates' dialog. The genomic regions enriched with i) housekeeping genes; ii) genes for ribosomal proteins; iii) vestigial genetic elements (comprising pseudogenes, transposons, prophages and IS-elements) are indicated.

3.4 Identification of divergent genomic islands

Several routines have been developed to identify the horizontally transferred genomic islands, genes for ribosomal RNA and proteins, non-functional pseudogenes and genes of other functional categories. All these routines are described in detail with illustrations in supplementary web-pages (use the 'Help' link in the applet window). The approach to identify inserts of foreign genomic elements by OU statistical parameters have been described recently (Reva & Tummler, 2005). While several algorithms allow identification of horizontally transferred genomic islands (Mrazek & Karlin, 1999; Azad & Lawrence, 2005; Becq *et al.*, 2008; Dufraigne *et al.*, 2005; Nakamura *et al.*, 2004; Pride & Blaser, 2002), the multiple oligomer parameters used in the SWGB even allows tentative attribution of genomic fragments (and, given the right scale, genes or gene clusters) to different functional classes using only a FASTA sequence as input. However, the emphasis of the SWGB is not primarily its annotation capability, but its ability to display the natural internal variability of genome sequences. *Pseudomonas putida* KT2440 was used, which is a known mosaic genome with 105 genomic islands above 4000 bp in length (Weinel *et al.*, 2004) as an example. Many of these features can be visualized at a glance using the SWGB without any in depth analysis (see Figure 3.2). On the 'Diagram' view the parameters `n1_4mer:RV`, `n1_4mer:GRV` and `n0_4mer:D` were selected for the X, Y and Z axes, respectively, as shown previously (see Figure 3.3). Plotting local relative oligomer variance (RV) against global relative variance (GRV) basically shows the effect of normalization by global mononucleotide content. The core genome is then represented on the dot plot as the positive linear correlation line where $RV \approx GRV$ (Figure 3.3). In other words, these fragments exhibit such compositional closeness to the core genome that normalizing by local mononucleotide content does not have a different effect compared to normalizing by global content. These genomic fragments also exhibit a low distance from the genomic average; and are therefore colored blue. Scattered dots lying peripheral to the expected strong linear correlation do not belong to the core genome and also have a higher distance from the genomic average and are hence colored green. Using the filter settings recommended in Figure 3.5, twenty one fragments were found to be genomic islands (note that while border values of OU parameters are not the same for different genomes, the grading notches of the sliders represent relative values that allows identification of homologous regions in many different genomes). For a number of reasons, many more islands were found in a similar analysis by Weinel *et al.*, (2004) Firstly, the sliding window size of 8 kbp means many of the 4 kbp features from their analysis were not identified automatically. Furthermore, they were looking for all compositionally atypical regions, whereas here, restriction was made to horizontally transferred regions.

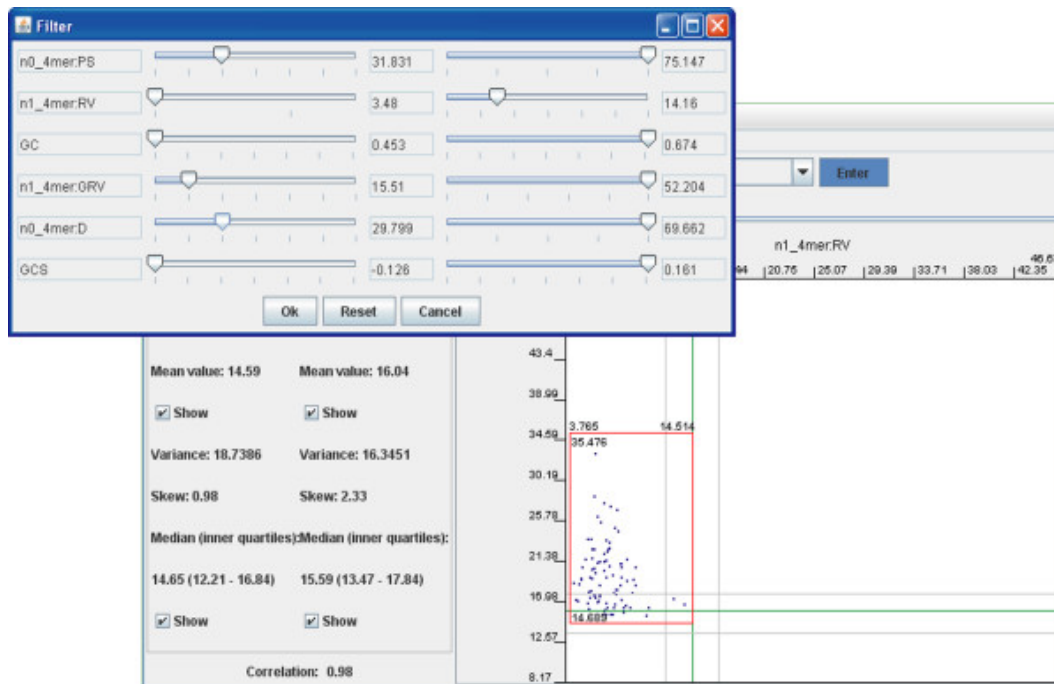


Figure 3.5: Filtering genomic regions by multiple parameters. Click the 'Filter' button to open a dialog as shown in the figure. Setting up border values of multiple OU statistical parameters allows more precise localization of regions of interest.

A known 40 kbp bacteriophage insertion [2586000–2626000] is, surprisingly, not among the genomic fragments selected in the SWGB using this filter. Although the prophage is still perceptible on the 'Gene Map' view (see a figure in the supplementary help web-pages), the OU parameters of the region do not differ markedly enough from the core sequence to be isolated automatically as a horizontally transferred region.

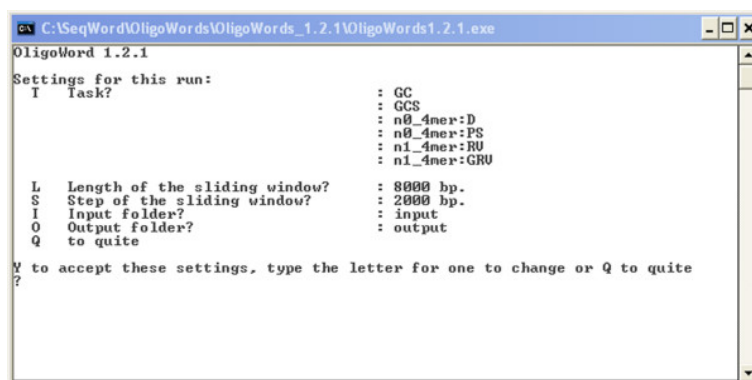
As the SWGB uses parameters that are based on comparison of local fragments to the global genomic average, strains with abundant insertions of homogenous DNA can confound this form of analysis. One example is the *Methanosarcina acetivorans* C2A genome which is composed of an estimated 25% of putatively horizontally acquired DNA, one of the highest amounts discovered to date (Dufraigne *et al.*, 2005). As a result of these insertions, the genomic signature has been strongly influenced, resulting in a large amount of scatter and a poorly defined core genome on the plots. On the other hand, this type of analysis allows estimation of genome stability in a simple, multi parameter view (see the *Vibrio cholerae* N16961-O1-eltor example in the online help files). To conclude, filtering provides a convenient way to automatically isolate divergent genomic regions of interest. However, some regions may erroneously remain undetected due to possible amelioration of older inserts (Lawrence & Ochman, 1997) or a higher level of noise in unstable genomes. However, many problematic genomic fragments can in some cases be easily attributed to functional gene categories using the SWGB 'Diagram' window (see Figure 3.2).

Methodologies for discovering long modular genes have already been discussed in a previous publication (Reva & Tummler, 2008). Briefly, long genes display a particular tetranucleotide usage and can be discovered by plotting $n0_4mer:D$ (X axis) versus $n1_4mer:RV$ (Y axis). The positively linear correlated outlier fragments (towards the top right of the image) are often fragments of long genes with their characteristic repeats. An example using the gene encoding the 1.12 megadalton cell surface protein of *Synechococcus sp.* WH8102 (McCarren & Brahamsha, 2007) marked on the dot-plot is shown in Figure. 3.4. Ribosomal RNA operons (but not genes for ribosomal proteins) are characterized by extremely high pattern skew and a large distance from the core genome (Figure 3.2). Thus, there is a tendency to find many genomic fragments containing rRNA genes colored dark brown to red in the bottom right section of the 'Diagram' tab. The annotation for rRNA operons is not present in the database; therefore, these are seen in the 'Gene Map' tab as un-annotated areas with high pattern skew (Figure 3.2). Ribosomal proteins tend to be increasingly present at a slightly greater than average RV and above average D (see Figure 3.2), which is in agreement with observations that highly expressed genes for ribosomal proteins have a highly specific codon usage compared to housekeeping genes of the organism (Puigbo *et al.*, 2008). The majority of genomic fragments form a cluster characterized by average and higher than average RV, stable OU patterns (low D) and low PS. These tend to be the core, or bulk genes and genomic regions with their typical tetranucleotide usage. Some other core sequence fragments spread from this area toward lower RV and less specific OU patterns (higher D and PS) – these are all characteristics of an unstable or randomly generated sequence (Reva & Tummler, 2004). These regions were found to be enriched with many hypothetical genes, prophages and transposons (the data is not shown but is easily verified with any genome using the 'Get' button. Consider, for example, this area in the pseudogene rich *Mycobacterium leprae* TN or *Methanosarcina acetivorans* C2A genomes (Dufraigne *et al.*, 2005; Klockgether *et al.*, 2006) and the relatively homogenous *Alcanivorax borkumensis* SK2 genome (Reva *et al.*, 2008). These regions were thus categorized as rich in 'vestigial' genes in contrast to the core genome regions rich in housekeeping genes (Figure 3.4).

It must be stressed that with an average length of genes being around 1 kbp and overlapping sliding windows of 8 kbp, one cannot expect precise separation of housekeeping and vestigial genes by the method described above. However, when analyzing an unknown DNA sequence prior to annotation, it may be helpful to identify genomic regions enriched with a higher proportion of these so called housekeeping genes and other regions rich in vestigial genes. These tentative results should be verified with other complementary algorithms such as BLAST, gene finding and annotation techniques.

The most important feature of the supplemented software available from the SWGB web-server for download is the ability to quickly and easily analyze a novel sequence on a local computer. The command-line Python program OligoWords is first used to analyze FASTA or GenBank formatted sequences. The program is available for download in several packages as precompiled executable files and as Python source code. The command-line interface of the

OligoWords program is shown in Figure 3.6. Parameters such as oligomer length and window size can all be set depending on the sequence length and desired resolution (see Table 3.2 for suggestions). Since the SWGB is implemented as a Java applet, it can be run within a web browser locally. The HTML-embedded applet is available for download from the same FTP site (15) (select SeqWord_Viewer.zip). The output file from OligoWords is read into the SWGB via the 'Open' function of the 'File' menu, and the complete functionality of the online system is then available. For example, a new sequence can be analyzed for ribosomal gene clusters, putative horizontally transferred elements or other regions of atypical DNA structure prior to the lengthy annotation step. A complete description of how to run the SWGB and OligoWords locally is presented in the online help files.



```

C:\SeqWord\OligoWords\OligoWords_1.2.1\OligoWords1.2.1.exe
OligoWord 1.2.1
Settings for this run:
I Task?                : GC
                      : GCS
                      : n0_4mer:D
                      : n0_4mer:PC
                      : n1_4mer:RU
                      : n1_4mer:GRU

L Length of the sliding window? : 8000 bp.
S Step of the sliding window?   : 2000 bp.
I Input folder?                 : input
O Output folder?                : output
Q to quite

Y to accept these settings, type the letter for one to change or Q to quite
?

```

Figure 3.6: Command-line interface of the OligoWords program. To change the setting for the current run, type the option's letter and enter a new value as prompted. Users may change: T) the set of statistical OU parameters to be calculated for every local pattern; L) length of the sliding window; S) step of the sliding window; I) the name of the input folder that contains FASTA and/or GenBank files with source DNA sequences; and O) the name of the output folder where the result files will be stored.

3.5 Scientific Investigation – Application to mycobacteria

In this section, the results of a comparative study of mycobacterial genomes by using SWGB routines described above will be shown and described. The aim of this study was to identify a few of the main sub-genomic components contributing toward genetic variation seen among mycobacteria. With the help of literature, we also aim to gauge these identified components contribution toward virulence of the organism.

Firstly, few mycobacterial genomes were analysed for the presence of horizontally transferred genetic elements (HTGE). RV (X-axis), GRV (Y-axis) and D (Z-axis) gene diagram plots were generated for *M. tb* H37Rv (Figure 3.7). The expected area on the plot where fragments of horizontally transferred genetic elements were expected to appear was empty (see Figure 3.7 below). Similar plots lacking traces of HTGE were obtained for all other *M. tb* and *M. bovis*

genomes.

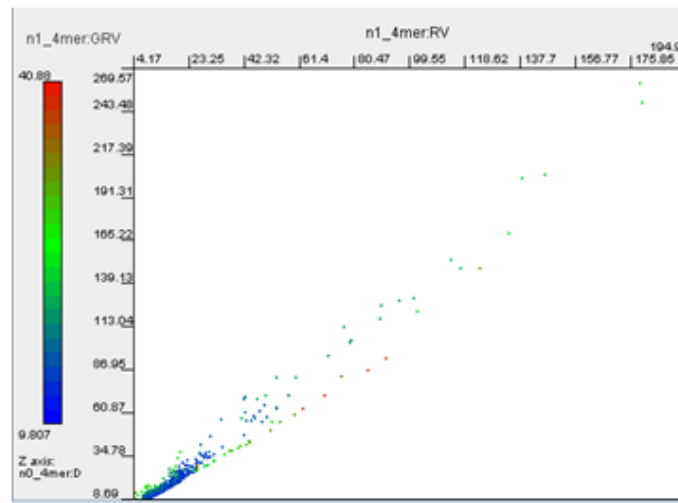


Figure 3.7: RV, GRV and D gene diagram plot for *M. tb* H37Rv.

On the contrary, the SWGB plot for *Mycobacterium avium* K10 (NC_002944) revealed many genomic regions of putatively lateral origin (Figure 3.8). The coordinates and annotation of these identified regions are shown in Table 3.2.

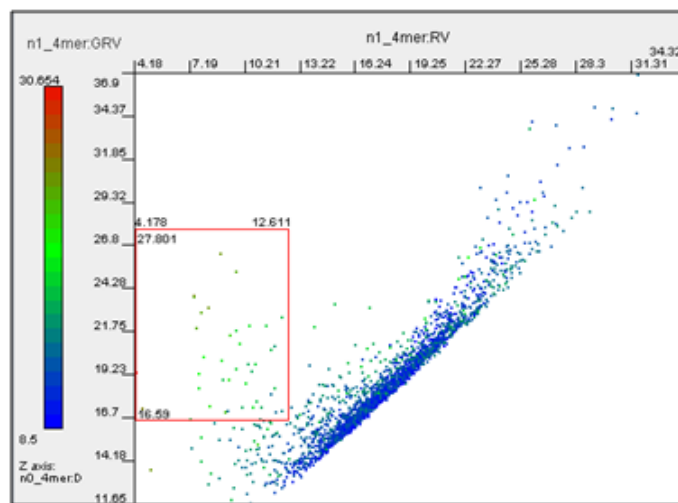


Figure 3.8: RV, GRV and D dot-plot generated for *Mycobacterium avium* K10.

The area of the distribution of horizontally transferred genomic elements is outlined in the figure above (Figure 3.8) and the coordinates of the outlined genes are shown in the table below.

Table 3.2: Coordinates and annotations of the gene islands in the genome of *M. avium* K10.

Left	Right	n1_4mer:RV	n1_4mer:GRV	n0_4mer:D	Annotations
78000	86000	11.6606	18.0853	13.7828	dnaB; mmpL4_1 and 4 genes for hypothetical proteins
870000	892000	8.6089	22.8127	22.0832	nramp and 20 hypothetical proteins
1290000	1304000	8.7803	20.6197	21.1895	lipL and a hypothetical gene

The RV, GRV, D plot above highlights several areas of possible horizontally transferred origin. The first genomic island (as outlined in the table above) is in the region 78000–86000 bp. In this region dnaB, mmpL4 and 4 genes of hypothetical proteins were found. Based on the RV, GRV distribution for this region, the 4 hypothetical proteins indeed exhibit atypical oligonucleotide usage relative to the rest of the genome. dnaB however, is found very proximal to this genomic island but is not necessarily horizontally transferred. This is the same situation with the next genomic island (870000 – 892000 bp) and nramp, where nramp lies proximal to the terminal regions of the genomic islands. If OU parameters are examined for this genomic region the following is seen.

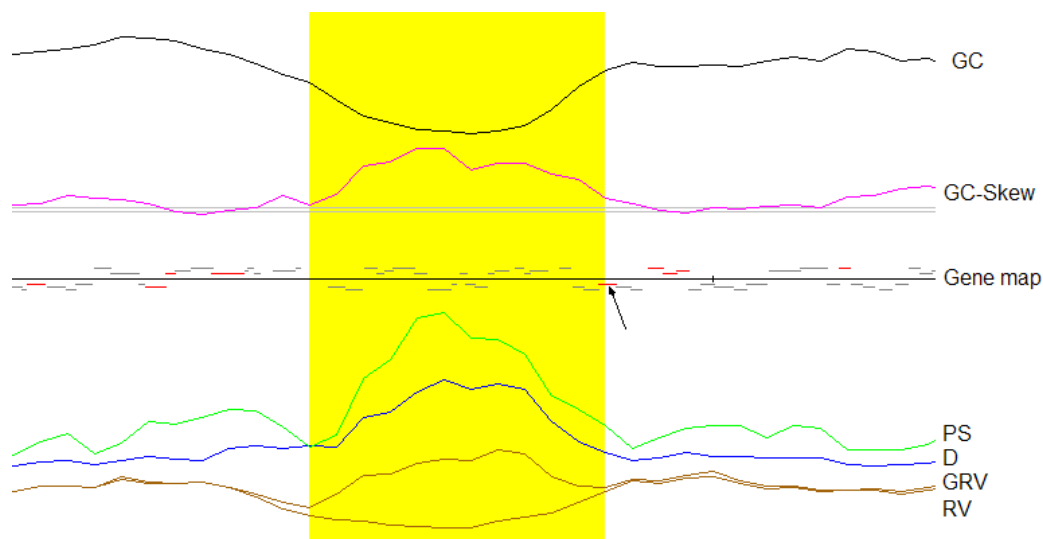


Figure 3.9: SWGB view for genomic region 87000-892000 (highlighted). An arrow marks nramp (in red) on the border of the highlighted region.

The above view of the genomic region reveals typical attributes for a horizontally transferred region. RV (the bottom most line) represents oligonucleotide usage normalized by mononucleotide content for the local pattern. GRV (above RV) represents oligonucleotide usage normalized by mononucleotide content for the whole genome. When regions (such as this) contain

atypical oligonucleotide usage patterns, GRV and RV diverge from each other as evident in the above figure. Nramp lies on the border of the genomic island is its horizontal origin should be checked separately.

Using different plot parameters (RV, PS and GC) on *M. tb* H37Rv revealed some regions of atypical and quite unusual OU.

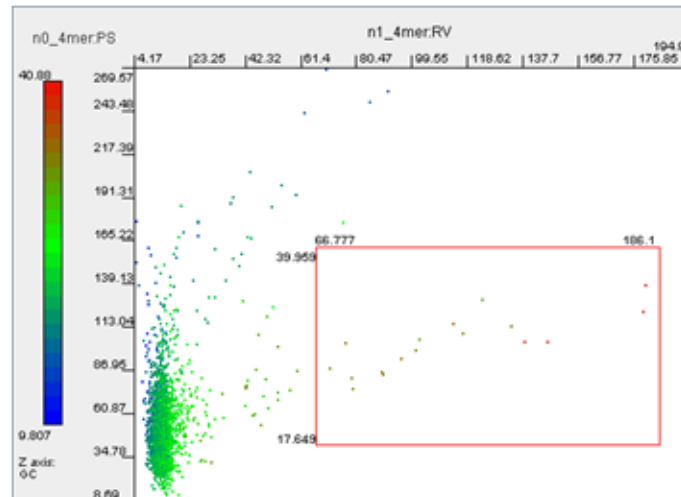


Figure 3.10: RV, PS and GC gene diagram plot for *M. tb* H37Rv.

Note that the core genomic elements form a dense cloud of dots with two jets of outliers directed rightward and upward relative to the core genome sequence. RV increase in local OU patterns shows an increased compositional bias in oligonucleotide frequencies that may correlate with the codon usage bias in coding sequences. A simultaneous increase of PS and RV parameters usually imply multiple tandem repeats in these genomic regions (Reva and Tummler 2008)

Annotations for the outlined genomic fragments for the *M. tb* H37Rv plot above (Figure 3.10) is shown in Table 3.3 below.

Table 3.3: Coordinates and annotations of the gene islands in the genome of *M. tb* H37Rv.

LEFT	RIGHT	n1_4mer:RV	n0_4mer:PS	GC	ANNOTATION
332000	342000	75.5	25.7	73%	PE-PGRS and PPE family proteins
1630000	1638000	77.1	29.1	75%	PE-PGRS and PPE family proteins
3734000	3746000	120.5	31.5	73%	PE-PGRS and PPE family proteins
3924000	3954000	121.6	29.1	76%	PE-PGRS and PPE FAMILY PROTEINS; acyl-CoA synthase; acyl-CoA dehydrogenase; acyl-CoA lygase FADD18; enoyl-CoA hydratase;thiamine-pyrophosphate requiring enzyme and many hypotheticals

This fragment selection from the RV, PS and GC SWGB dot-plot screen (Figure 3.10) allows identification of PE-PGRS and PPE family proteins which in *M. tb* are indirectly associated with virulence (Zheng *et al.*, 2008).

These PE-PGRS genes acquired highly peculiar characteristics in the *M. tb*/*M. bovis* genomes as compared to *M. avium* which contains homologous genes that can not be distinguished from the core genome sequences by the method described above. The genomes of *M. ulcerans* and *M. marinum* on the other hand, show intermediate states of evolution from the pattern of *M. avium* toward that of *M. tb* (Figure 3.11)

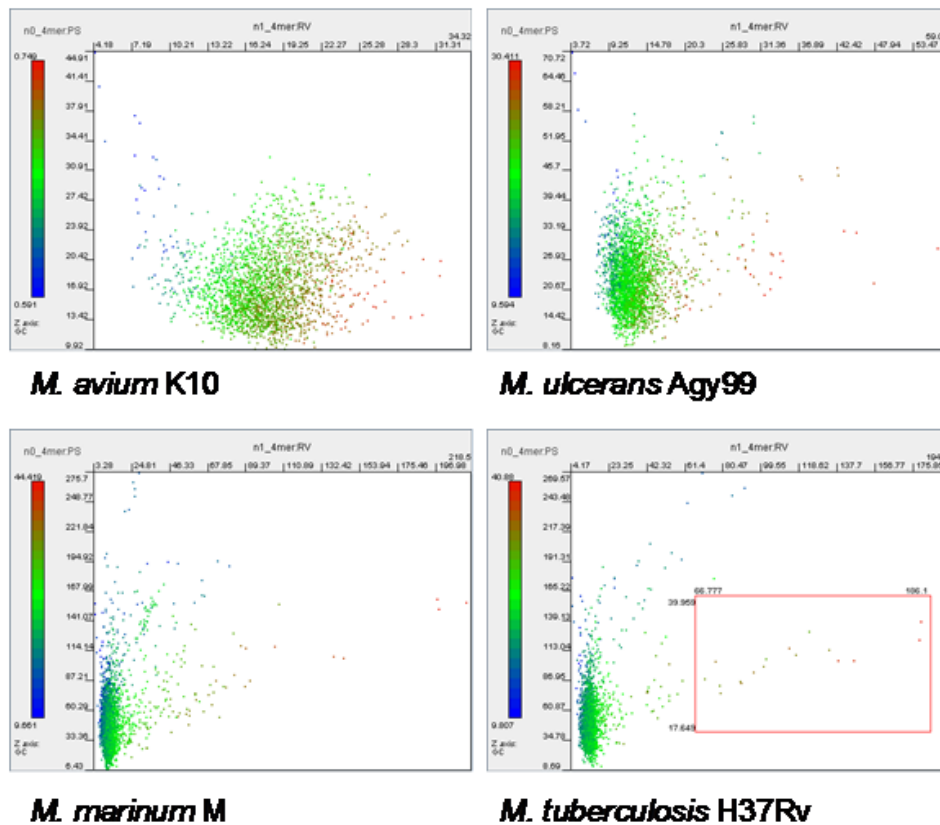


Figure 3.11: Global evolutionary changes in mycobacterial genomes as revealed by SWGB dot plots. Each dot corresponds to the calculated oligonucleotide usage pattern for an 8kb sliding window of step size 2kb.

Taking into account that the differences between 16S rRNA sequences of these genomes are less than 3%, the rate of evolutionary changes of the genes of PE_PGRS and PPE families is fascinating and yet poorly-studied (Harmsen *et al.*, 2003).

The rate of mutations in these genomic loci are several fold higher than the average rate of mutations per genome. In Figure 3.12 below, single nucleotide polymorphisms (SNPs) between the two closely related *M. tb* strains H37Rv and H37Ra is shown. Note the frequency of SNPs accompanied with genome rearrangements in a hypervariable locus toward the right of the figure.

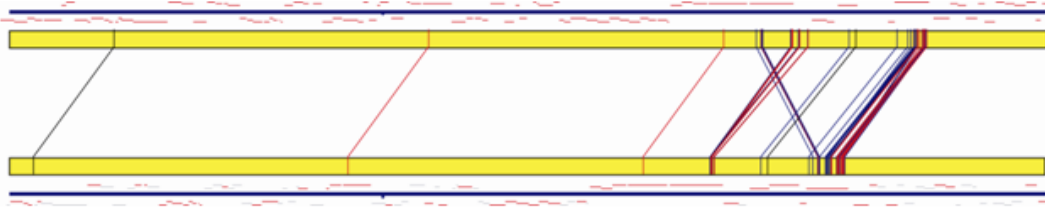


Figure 3.12: SNP distribution in homologous loci of *M. tb* H37Ra [3,800,000-4,000,000] (above) and *M. tb* H37Rv [3,850,000-4,000,000] (below) genomes. Transitions, transversions and deletions are depicted by blue, red and black connecting lines, respectively. Genes are depicted by red and grey (hypothetical) bars above (forward) and below (reverse) the blue lines according to the direction of their transcription.

Figure 3.11 above emphasizes the differences of these organisms by displaying the genomic fragments that have experienced evolutionary changes from *M. avium* lineage to that of *M. tb*. For *M. avium* the variability of the calculated parameters RV is found to be in the range 5-35 while PS found in the range 10-45. These ranges are considered homogenous as they are consistent among many other bacterial strains including *E. coli* and *B. subtilis*. On the contrary, some loci in *M. tb* exhibits extreme divergence for these calculated parameters. For example, having a RV up to 200 and PS up to 270. This signifies local mutational hotspots which have not been well studied thus far. In the following chapter the mycobacterial comparison project (MCP) will be used to further elucidate these regions of interest.

3.6 Discussion

In this chapter, the SWGB and its functionality was used to reveal differences among mycobacterial genomes that would have otherwise gone un-noticed. Based on sequence studies thus far, it is known that mycobacterial genomes are largely similar on the nucleotide level (Fleischmann *et al.*, 2002). Using conserved genes such as 16s RNA, dnaB and gyrB genes, it is possible to sometimes differentiate between various strains but not to a high level. Using our novel algorithmic approaches to the genomic analyses and novel display techniques (Ganesan *et al.*, 2008), however, we were able to not only identify regions of gross differences between some mycobacterial strains, but also identity specific genes and genomic islands that account for much of the sequence differences exhibited in these organisms. Several interesting genomic islands containing important genes were identified.

Several gene islands which had most likely been acquired by the lateral transfer but had been significantly ameliorated towards the OU pattern of the core genome sequence were identified in the *M. tb* H37Rv genome by an in-house program Gene Island Sniffer (12). Many virulence associated proteins were found in these former gene islands. The *M. tb* genome contains 13 genes encoding RND (resistance, nodulation & cell division) proteins designated, MmpL (Mycobacterial membrane protein Large). RND proteins are a family of multi-drug resistance pumps that

function to recognize and mediate the transport of a wide variety of cationic, anionic or neutral compounds such as various drugs, fatty acids, bile salts etc. (Domenech *et al.*, 2005). Although MmpL proteins play a role in drug resistance in certain bacteria, it was found not to be the case within mycobacteria however, MmpL4 mutants showed a decreased level of virulence in a low-dose aerosol murine model of infection. The study supports the concept that MmpL-mediated lipid secretion affects both the pathogens ability to survive intracellularly as well as host-pathogen dialogue which determines the ultimate outcome of infection (Domenech *et al.*, 2005; Rodriguez *et al.*, 2002). lipL was one of the other elements identified. lipL is a gene belonging to the hormone-sensitive lipase family and is responsible for fatty acid metabolism during an adverse nutrient climate (Deb *et al.*, 2006). This activity may account for the organisms utilization of stored triacylglycerols during dormancy and its subsequent reactivation.

Also identified were genomic islands in *M. tb* H37Rv rich in PE-PGRS genes, perhaps representing a gene cluster of some sort. Studies in the past have shown that virulent mycobacterial genes sometimes do appear clustered together in specific loci. Camacho *et al.*, (1999) identified a 50kb chromosomal region in *M. tb* which contained several virulence genes. The group created a library of signature-tagged transposon mutants of *M. tb* and then screened for those which had their ability to replicate within lung of mice negatively affected. The insertions for those mutants which had their virulence inhibited were then mapped onto the *M. tb* genome. Apart from the identification of the 'pathogenicity island' the group also noticed that most of the mutated loci seemed to be involved in lipid metabolism and transport across the membrane (Camacho *et al.*, 1999). Similarly, Danelishvili *et al.*, (2007) also conducted studies with some transposon mutant mycobacteria this time with the intention of identifying *M. avium* genes and host cell pathways involved in their uptake by macrophages. In the clones with impaired macrophage uptake, they revealed that 4 of the six genes examined, all lie within the same region of the chromosome. Analysis of this chromosome region revealed a pathogenicity island of 58% GC content (compared to 69% for the genome) inserted between two tRNA sequences. This region was also found to be unique to *M. avium* and absent in *M. tb* and *M. tb paratuberculosis*. Gene islands indeed play a role in the life-cycle and virulence of mycobacteria thus it is imperative that these regions can be accurately and efficiently identified.

In terms of the genomic islands (loci with atypical OU) found within *M. tb* H37Rv (Figure 3.10), it is seen that the PE-PGRS, PPE gene family are a dominant feature. What are these genes and what role do they play? Approximately 8% of the potential coding capacity of *M. tb* H37Rv was found to be accounted for by two unrelated gene families encoding the PE and PPE proteins (Banu *et al.*, 2002). The PE/PPE names are derived from the motifs Pro-Glu/Pro-Pro-Glu which in most cases are found near the N-terminus of these glycine and alanine rich proteins. The PE and PPE family comprises of about 100 and 68 members, respectively. The largest class of the PE family, having 67 members in *M. tb* H37Rv is referred to as the PE-PGRS sub-family. These proteins consist of the PE domain followed by C-terminal extension with multiple tandem repetitions of Gly-Gly-Ala or Gly-Gly-Asn encoded by the PGRS (polymorphic GC-rich repeti-

tive sequence) motif. PE-PGRS proteins may contain up to 1900 amino acids (based on predictive models), up to 50% of these can be glycine (Banu *et al.*, 2002). Implicit in the name, PGRS genes are GC rich and may be a major source of polymorphism, this then lead to the question of whether PE-PGRS proteins of *M. tb* variable surface antigens? Banu and group tested this hypothesis by raising antibodies in mice against 5 PE-PGRS proteins. These antibodies detected single proteins when the original plasmid constructs (used for immunization) were expressed in epithelial and reticulocyte extracts, thus confirming the proteins antigenicity. Furthermore, the antibodies cross reacted with several PE-PGRS proteins suggesting that different proteins share common epitopes. The group then went on to perform sub-cellular fractionation studies and immunoelectron microscopy which localized many PE-PGRS proteins in the cell wall and membrane of *M. tb*. Their findings further suggested that PE-PGRS proteins play a role in antigenic variability on the cell surface (Banu *et al.*, 2002; Okkels *et al.*, 2003). Similarly, in a comparative study of gene products of key metabolic pathways among 5 mycobacterial genomes (*M. tb*, *M. leprae*, *M. avium*, *M. bovis* and *M. avium ssp. Paratuberculosis* K10), it was shown that the major differences between these species is accounted for by gene products constituting the cell wall and gene families encoding the PE/PPE/PGRS proteins. What is interesting is that *M. avium ssp. Paratuberculosis* lacks PE-PGRS genes. This gene set is the very likely set of genes responsible for the survival of *M. tb* in macrophages, which then leads to the idea that *M. tb* and *M. avium ssp. Paratuberculosis* exhibit differences in the survival mechanisms of these species within macrophages (Marri *et al.*, 2006). There is also evidence that PE-PGRS may be important for bacterial survival during early stages of infection indirectly through alternative sigma factor (SigD) (Raman *et al.*, 2004) as well as provide resilience for a changing host micro-environment through differential expression (Voskuil *et al.*, 2007)

It has been shown how easily gross differences between mycobacterial genomes can be detected using our SWGB applet. Several gene islands have been identified and along with genes that account for the differences between the various mycobacterial species. In the following chapter, a closer look will be taken at these genes and gene families using the mycobacterial comparison project (MCP) system in an attempt to provide deeper understanding of these differences and the role they play in the evolution and virulence of mycobacteria.