Development of new computational approaches for analysis and visualization of fluxes of genomic islands through bacterial species.

Louis Cronje in partial fulfillment of the degree (MSc Bioinformatics)

2015

Department of Biochemistry, School of Biological Sciences,

Faculty of Natural and Agricultural Sciences, University of Pretoria

Contents

List of Figures

List of Tables

Chapter 1. Background

1.1 Literature Review

1.1.1 Introduction

Oscillations of gene exchange in the prokaryotic world have enormous impacts on bacterial evolution and outbreaks of new diseases (Smets & Barkay, 2005; Seiffert, et al., 2013; Goldsmith , et al., 2013). Yet we have limited information on the dynamics of the complex system. Mobile genetic elements (MGEs) are shared between bacterial taxonomic ranks by means of horizontal gene transfer (HGT). Genetic exchange rates are largely determined by sequence and ecological similarities among organisms and we expect closer related species to show increased HGT as a result. Successful HGT is characterized by compatible architecture and a physiological kinship to recipient organisms, provided the novel gene exerts advantageous influence for the recipient within its niche. Furthermore, the acquisition of genomic islands (GIs) enables environmental and clinical strains to transform into threatening forces by providing both “offensive” and “defensive” virulence encoding factors. It is therefore imperative that we study the mechanisms of HGT as well as the underlying foundation of gene sharing relationships in the prokaryotic realm. Current research on the detection of GIs allows for the estimation of relative insertion times as well as transfer direction indices (Bezuidt, et al., 2011; Popa, et al., 2011). In turn, these indicators are proposed to reveal recurring HGT events to ultimately infer donor-recipient relationships. This raises some serious questions regarding the regulation of HGT between microbial organisms and our ability to predict the genetic and environmental signals that trigger them. Given the tremendous amount of GI markers already identified (Langille & Brinkman, 2009), we are able to venture into gene sharing arrangements between a multitude of species and we may simulate gene sharing networks even among diverse bacterial communities. A collection of gene flux networks will allow us to monitor new and most active gene exchange hotspots and assist in formulating accurate expectations regarding gene spread research, especially among those of pathogenic nature. However, due diligence should be given in the implementation of such a system to ensure that all necessary aspects are taken into account.

1.1.2 The influence of gene sharing among prokaryotes

In the prokaryotic domain, genomic island (GI) exchange routines have extensively controlled ecological niche stability and susceptibility. With the availability of new properties provided by the readily integrated genetic segments, HGT has greatly influenced the adaptive ability of bacterial commensalism, symbiotic relationships and environmental interaction (Ulrich, et al., 2004; Jane & Frederick, 2011; Zongfu, et al., 2013; Ho Sui, et al., 2009). Our attention however, has only recently been focused on the detection of GIs with the realization of the association between significant virulence gene frequencies and the occurrence of GIs. ***Figure 1.1*** shows the proportions of virulence factors inside and outside GIs predicted by three different IslandViewer methods (Langille & Brinkman, 2009). It is not only the offensive elements we need to concern ourselves with; defensive properties increasing an organism's resilience, for example by the addition of antibiotic resistance genes, further enable the organisms evasive or defensive strategies to overcome threatening evolutionary pressures. Makarova et al. (2013), evaluates the various known defense systems provided by genes typically associated with GIs and showed that the distribution of different defense mechanisms among archaea and bacteria are not uniform with respect to overall abundance and usage but rather cluster into distinguishable groups. This indicates different preferences for microbe resistance response and accentuates the need for unique approaches in dealing with each resistance class. Makarova et al. (2013), categorized the defense systems into four distinct subgroups. The various restriction-modification gene sets (R-M), Clustered Regularly Interspaced Short Palindromic Repeats with CAS protein gene sets (CRISPR), toxin with antitoxin (TA) systems as well as the abortive infection (ABI) gene sets. ***Figure 1.2*** reveals the fraction of these gene found within GIs as well as the ratios of each class we may expect to find relative to the total number of genes within an organism. ***Figure 1.3*** shows the various preferences of defense systems among major prokaryotic taxa.

The implications of these studies are amplified in the hospital setting where active surveillance and strict hygienic procedures are implemented to prevent cross infections and outbreaks (Sabino, et al., 2011). Resistant strains quickly become dormant as resistances accumulate via a wealth of locally available gene pools (Baldan, et al., 2012).

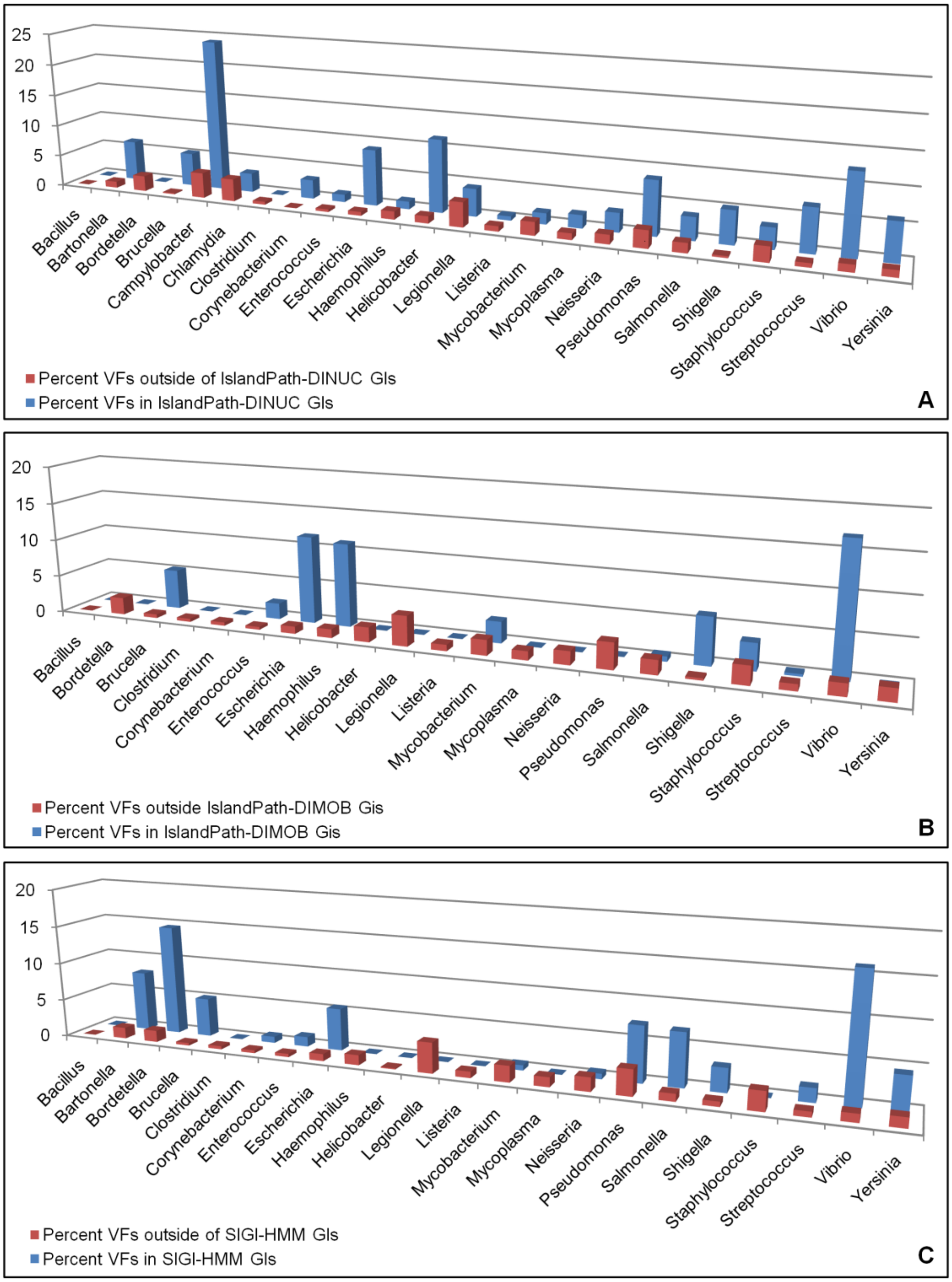


Figure 1. 1 Proportion of genes (%) that are virulence factors (VFs) inside versus outside of genomic islands predicted by (A) IslandPath-DINUC, (B) IslandPath-DIMOB, and (C) SIGI-HMM GIs. Pathogens having GI predictions are grouped by genus. Adapted from Ho Sui, et al. (2009).

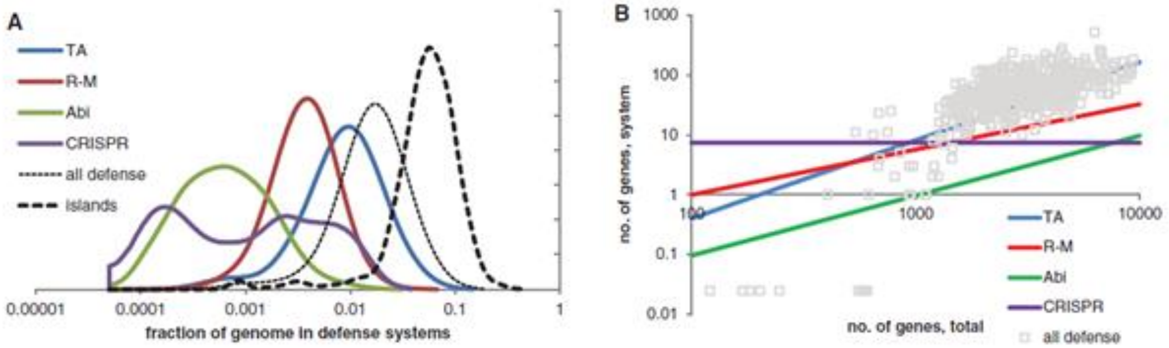


Figure 1. 2 Four of the most influential types of defense systems in 1516 bacterial and archaeal genomes (A) per occupying fraction of the genome (probability density function) and (B) per defense gene set ratio relative to the total number of genes. Adapted from Makarova, et al. (2013).

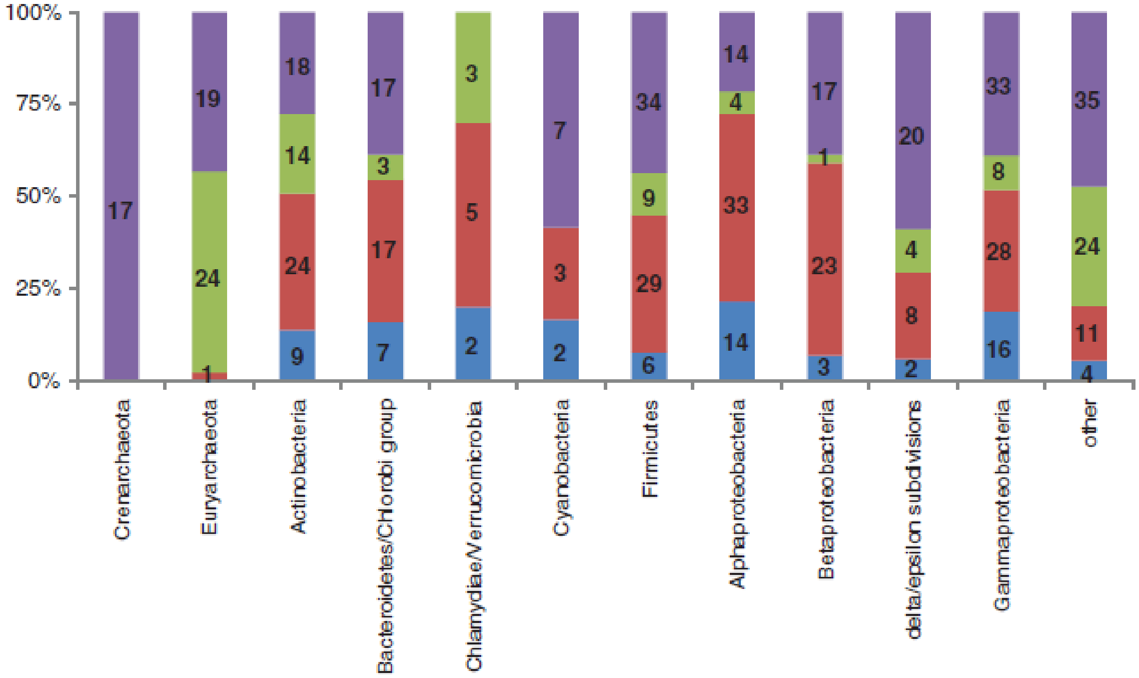
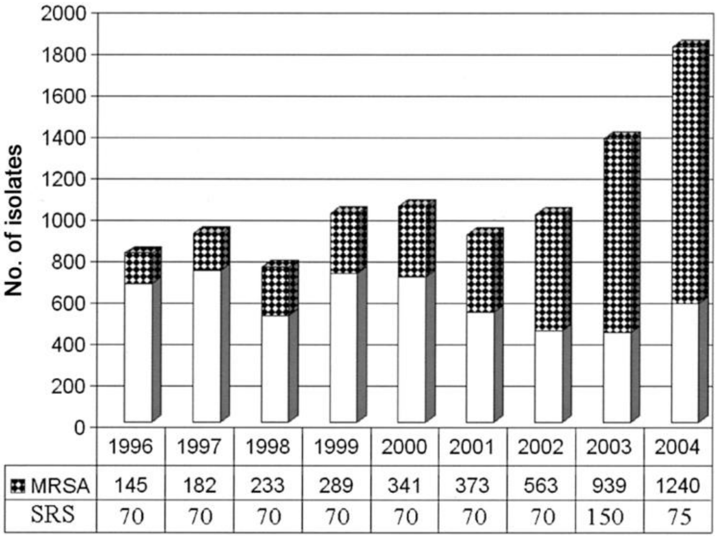


Figure 1. 3 Distribution of the defense strategies among major prokaryotic taxa. The number of analyzed genomes for each taxon is indicated inside the respective bar. The colors represent the strategies as coded in Figure 1.2. Adapted from Makarova, et al. (2013).

An Diep *et al.* (2006) associates HGT with the emergence of new Methicillin-Resistant strains of *Staphylococcus aureus* (MRSA) within the Community Health Network of San Francisco and shows an increased detection of MRSA isolates over a timespan of only 8 years. ***Figure 1.4*** reveals the steady increase of the resistant strains while the susceptible strain isolates decline.

It is of grave importance that we formulate an understanding as to where and when threating transfer events do occur. Indeed, different GI detection strategies have been employed with sufficient benchmarking to determine the accuracy and completeness of each method (Becq, et al., 2010; Karlin, 2001). However, limited work on the origination of gene flux have been done in the community setting, with studies directed at certain taxonomic involvements or with predetermined gene prominence (Bezuidt, et al., 2011; Encinas, et al., 2014). Discoveries relating to GI flows are interest-driven and we lack a comprehensive view of co-existing GI highways.



	1996	1997	1998	1999	2000	2001	2002	2003	2004
▦ MRSA	145	182	233	289	341	373	563	939	1240
SRS	70	70	70	70	70	70	70	150	75

Figure 1. 4 Secular trends of methicillin-susceptible (white bars) and methicillin-resistant (shaded bars) Staphylococcus aureus (MRSA) infections in unique patients, identified during laboratory-based surveillance for the Community Health Network of San Francisco. The total no. of MRSA isolates is listed for each year. Adapted from An Diep, et al. (2006).

1.1.3 Genomic island evolution

In an attempt to understand the multifaceted aspects of GI transfer and integration events, we have explored the evolutionary basis for the provision of HGT mechanisms. However, we are left to speculate. It has been proposed that HGT was derived from integrating plasmids or phages which had lost their abilities of self-replication and subsequently transferred themselves for more stable succession lines (Ulrich, et al., 2004). This adaptation has caused great ripples through the evolution of prokaryotes, successfully driving the creation of countless new bacterial species (Jane & Frederick, 2011). ***Figure 1.5.1* and *Figure 1.5.2***, are representations of the implications of HGT motivating genetic variation among microbial tree of life. Ulrich *et al. (2004)* explains that bacteria which are present in communities which also contain other species, have access to a greater collection of mobile gene pools. HGT then, is more likely to occur within ecological niches, which are colonized by diverse microbes than of those of sparely populated environments. ***Figure 1.6*** reveals a global network of coexisting microbial lineages as a demonstration of the multitude of bacterial communities (Chaffron, *et al*., 2010).

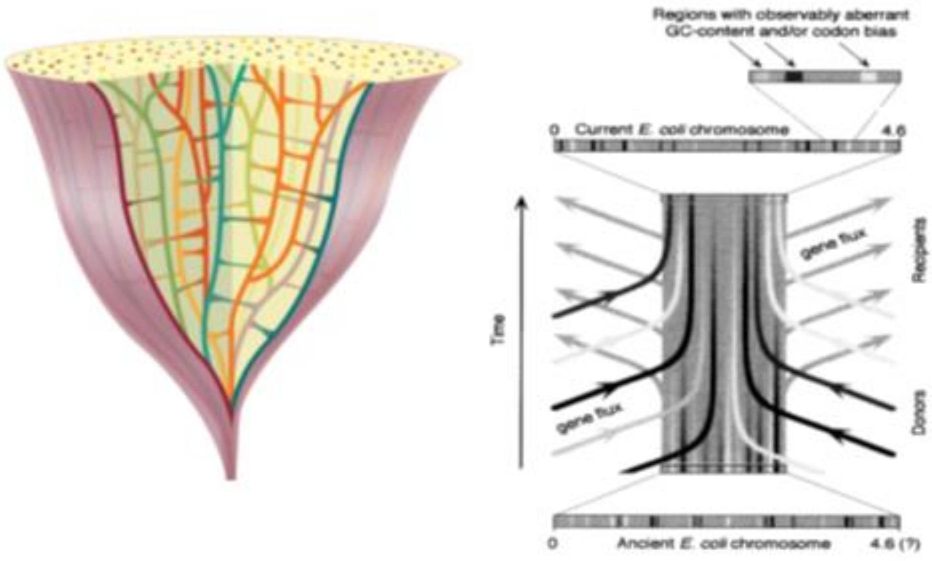

Regions with observably aberrant
GC-content and/or codon bias
0
Current E. coli chromosome
4.6
Time
gene flux
gene flux
Recipients
Donors
0
Ancient E. coli chromosome
4.6 (?)

Figure 1. 5. 1 Phylogenetic tree of microbial genomes. Tree branches correspond to genomes and branch colors represent different lineages. Horizontal connections between the branches correspond to HGT events. Adapted from Popa & Dagan (2011) Figure 1.5.2 (right) Schematic representation of gene flow through the E. coli chromosome through time. Adapted from Martin (1999).

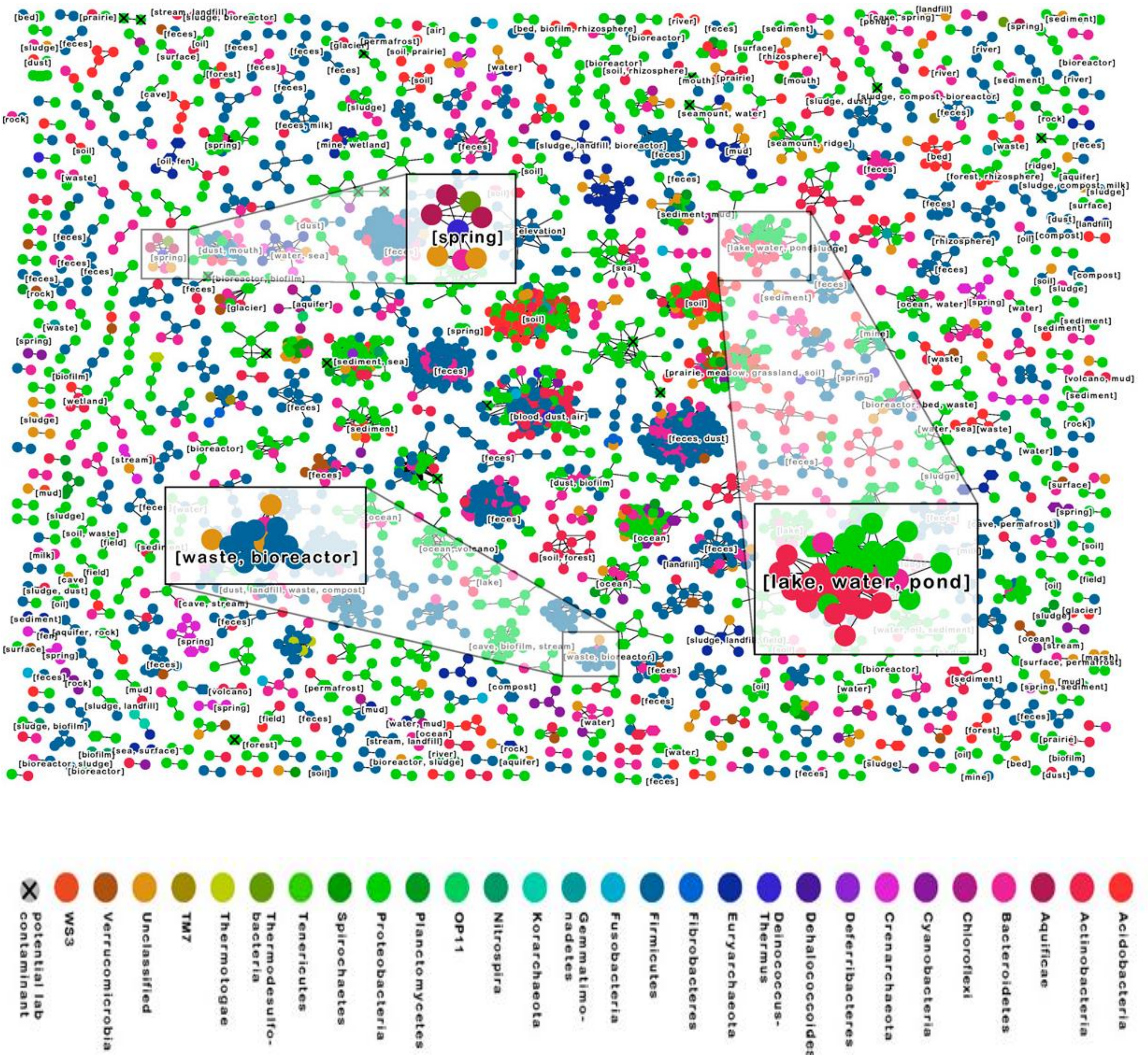


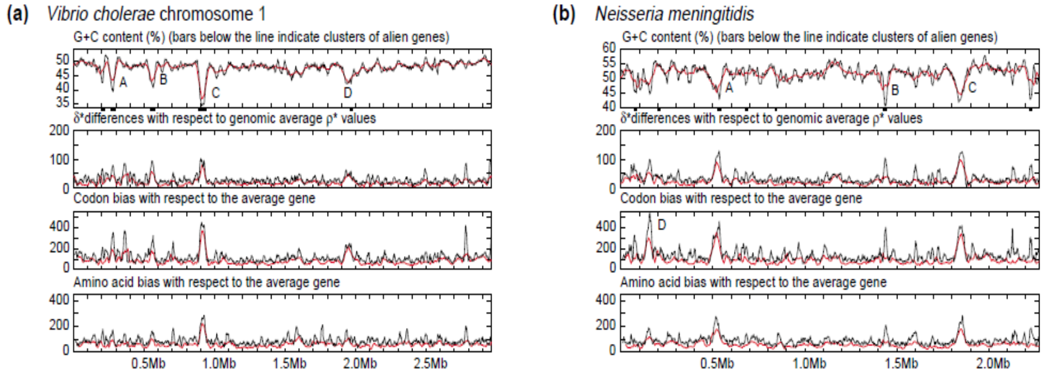
Figure 1. 6 Node color denotes taxonomic classification at the phylum level. Each node denotes a microbial lineage, and each line is a significant co-occurrence relationship partitioned using unsupervised Markov clustering, to reveal modules (clusters) of co- occurring lineages. Adapted from Chaffron, et al. (2010).

The process for the assimilation of genes obtained through HGT can be divided into stages, namely the transference of mobile DNA into the cytoplasm, the integration of the acquired DNA into the site specific regions of the genome and the clonal inheritance to following generations (Popa, et al., 2011). From the fully sequenced genomes currently available, we observe a substantial fraction of open reading frames (ORFs), have been involved in HGT events (Popa, et al., 2011; Jane & Frederick, 2011). Speculatively, almost all genes are subject to HGT with only a few resistant to it (Park & Zhang, 2012). Gene exchange events may be done through a number of mechanisms of which the most prominent are transformation, conjugation, transduction and gene transfer agents (Popa & Dagan, 2011).

It has been shown that all GIs contain a recombination mechanism consisting of either an integrase or in some cases, a recombination directionality factor (Boyd, et al., 2009). Integrases are specific to the sites they bind to and it is well established that such binding usually occur at tRNA loci. However, the number of specific tRNA genes used to integrate over the entire genome are limited to a minority (Williams, 2002). Other suggestions of integration occurring at specific sites are evident in instances such as for *Streptococcus thermophilus* CNRZ368, where a significant number of genomic islands are recombined at the same position in seven other strains of *S. thermophiles* (Pavlovic, et al., 2004). Soon after the integration of foreign elements into the genome, these regions are brought into functional questioning by the host organism. Transferred genes tend to be deleted quickly if they prove to provide no or unfavorable utility (Popa, et al., 2011). This means that although large genic regions may have been obtained through HGT, only snippets may be retained within the host genome. Furthermore, it has been suggested that the translocation of genes takes place preferentially in recently acquired genes (Hao & Golding, 2009). Upon obtaining new genetic material, a host bacterium frequently counterbalances the acquisition with a loss of native genes (Ulrich, et al., 2004; Lefeuvre, et al., 2013). Whether these steps would prove to be advantageous remains to be decided by selection.

1.1.4 Genomic island detection

The detection of GIs has received a great deal of attention. Several methods encapsulating the discovery of genomic regions foreign to the host organism exists. The first group of methods implements comparative genomics to identify genomic regions unique to one organism by comparing multiple related strains. HGT events, even when occurring infrequently, have major impacts on phylogenetic inferences. Therefore the phylogenetic incongruence method is one of the most accurate GI detection methods to date (Langille, et al., 2010). However for purposes of computational expense, such approaches are severely limited due to the method being completely dependent on the breadth and depth of the query sequence database. The detection of orphan genes of foreign origin, for example, is especially evaded by these approaches (Becq, et al., 2010). Moreover, methods based on phylogenetic differences are not well suited for the exhaustive search of HGT as a sufficient number of orthologs for each gene under inspection is required to produce any significant results. A second group of methods infers HGT when encountering atypical sequence composition from a gene or gene set when compared with the rest of the genome. Differences in GC content along with dinucleotide abundance values, codon biases and oligonucleotide frequencies has been used to great extent. Overall, methods based on oligonucleotide frequencies share the greatest accuracy with sufficient sensitivity (Becq, et al., 2010; Langille, et al., 2010; Karlin, 2001; Roos & Mark , 2011).



A: LPS biosynthesis genes (*rfb* operon); B: ORFs and a hemolysin secretion protein homolog; C: toxin-coregulated pilus biosynthesis (TCP) operon and ORFs; D: ~50 kb region contains 13 alien ORFs (unknown function)

A: ORFs and two hemagglutinin/hemolysin related proteins; B: five *frpC* operon genes (similar to cytotoxin, involved in pathogenesis), two transposases, ORFs; C: ORFs, two hemagglutinin/hemolysin-related proteins, hemolysin activator *hecB*, one transposase; D: major cluster of ribosomal protein genes

Figure 1. 7 Examples of regions in (a) Vibrio cholerae and (b) Neisseria identified as having atypical sequence properties at line peaks; Top bar: GC % content; Second bar: dinucleotide signatures; Third bar: Codon bias; Fourth bar: Amino acid bias. Adapted from Karlin (2001).

Some pitfalls associated with sequence composition methods do exist. Highly expressed genes, such as the genes in ribosomal operons, have a sequence composition that is dramatically different from the host genome and would therefore be classified as a GI. Additionally, amelioration, an evolutionary process described as mutational pressure that slowly causes genes which were integrated to adapt to the host genome signature over time, also limits the ability of sequence composition methods to detect more ancient GIs (Wiedenbeck & Frederick, 2011). Other approaches for the detection of foreign regions integrates the identification of genes associated with GIs by searching through annotation data and finding BLAST sequence similarities. In retrospect, GIs which were obtained from recent acquisitions still reflect the sequence composition patterns of their previous hosts. It is there for possible to re-trace donor organisms where sufficient matching is found. Roos & Mark *et al*. (2011), state that when multiple GIs are identified in a single organism, they may share resemblances in their genomic composition and we may expect that a common donor was responsible for those transfers. Through the use of clustering techniques, it is possible to bundle these sequences together and we may only need to account for a single common donor for each cluster. The authors evaluated 1787 GIs from 246 genome sequences in 88 species by allocating GIs in individual organisms which show similar dinucleotide relative abundance values into groups. ***Figures 1.8*** shows the calculated number of GIs per analyzed genome and number of GIs within clusters after grouping. These results show that even in lack of sequence alignment, the regions detected with atypical sequence composition can be associated with each other individually. Clustering related sequences in this way provides a reference for analyzing the roles and interactions between the donor and recipient organisms because they may share common regulatory modules or mobilizing capacities.

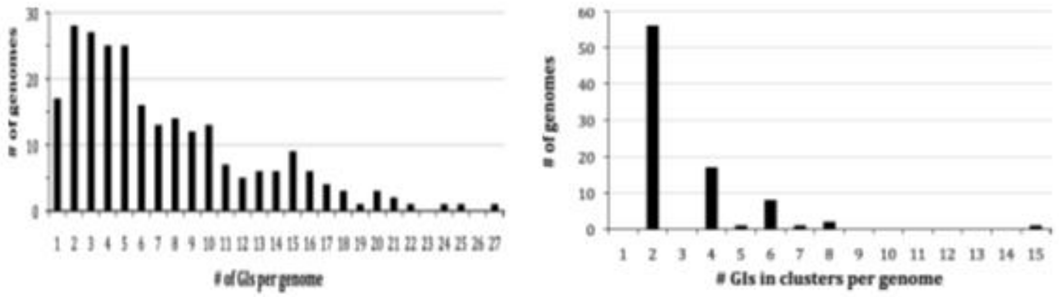


Figure 1. 8 The number of Genomic Islands per genome for the 246 genomes tested (left). Adapted from (Roos & van Passel, 2011); the number of clustered GIs per genome (right). Adapted from Roos & van Passel, 2011)

1.1.5 Barriers to horizontal gene transfer

HGT may bring genes from taxonomically unrelated species into a considered genome. Unlike point mutations, genetic rearrangements, gene loss and duplications, the speed at which microbes are able to utilize new functionality through the use of HGT is unmatchable. The current perspective regarding HGT exchange routines is that they are merely random dealings due to the promiscuity of microbes (Wiezera & Merklb, 2005). However, the evolutionary driving force for HGT is faced with transfer barriers (Thomas & Nielsen, 2005). Donor–recipient genome sequence composition similarities have been shown to correlate positively with HGT event frequencies. While gene acquisitions from distantly related species have been reported in literature, such events are proposed to evade the sequence similarity complications by integrating near or with a recognized promoter or by advances such as the nonhomologous end-joining DNA doublestrand break repair mechanisms (Popa, et al., 2011; Shuman & Glickman, 2007). Another requirement for the successful integration of foreign DNA is functional utility. Following the integration, nonfunctional or unnecessary DNA is usually discarded. Physical distance between HGT parties serves as an immediate barrier, evident in the formation of habitats (here transduction is considered as the longest range mechanism due to the phage mobility) (Popa & Dagan, 2011). The availability of nutrients can also serve to isolate geographic reproduction and may therefore also be included in the latter as a mechanism of HGT maintenance. However, it has been reported that multiple antibiotic resistance genes were shared among very diverse pathogens which proliferated in an agricultural context (Gatica & Cytryn, 2013). Furthermore, this phenomenon has also been shown to occur between pathogenic species confined to a single ecological niche (Djordjevic, et al., 2013). Therefore, a means enabling the travel of microbial genes from one part of the biosphere to another must exist with probable origins from completely separate areas (Wiezera & Merklb, 2005). Given the assumption that above mentioned restrictions do have significant impacts on the directive flow of GIs, it remains difficult to classify to what extent gene pools are available or transferable between microbial populations. To attempt such analyses, patterns of gene movements with sufficient information regarding sequence similarities and functional associations needs to be evaluated in order to maintain an understanding of the impacts of the gene sharing relations.

Lefeuvre, et al. (2013), evaluate the genomic plasticity of *R. solanacearum* to determine gene exchange flows accessible to the species. The authors suggested that through analyses of gene movement patterns, factors such as gene function and ecology reveal possible HGT highways. The study used 72 *R.*

solanacearum strains as representatives of all the phylotypes of *R. solanacearum* through the hybridization of metagenomic data using microarrays. Probes were chosen to be representative of all the coding regions of six full *R. solanacearum* genome sequences available at that time. Hybridization signals were filtered and analyzed to obtain a matrix of positive and negative probe signals. The presence or absence of a signal for each probe were used to reconstruct the phylogeny of *R. solanacearum* with frequencies of gene gain and gene loss, estimated independently. ***Figure 1.9*** reveals estimated HGT pathways on a circular phylogenetic tree for the species and is representative of the existence of transfer restrictions from certain clades.

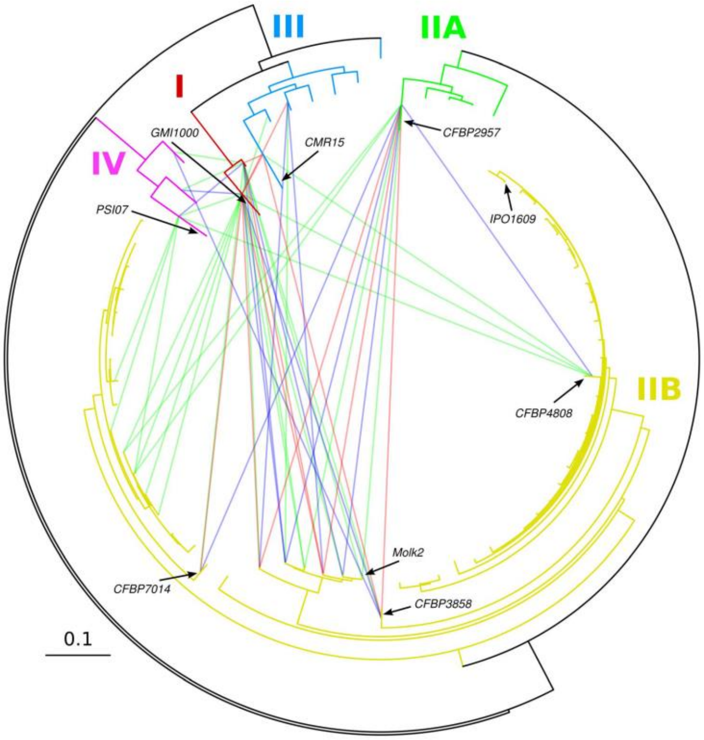


Figure 1. 9 Circular representation of the phylogenetic tree based on the hybridization of 7,055 probes from the six fully sequenced strains. Putative horizontal gene transfer events are represented using lines between tips of the tree. The lines are colored according to the significance level of having more gene sharing than expected by chance with green, blue and red for p-values superior to 0.99, 0.999 and 0.9999

While it has been shown that HGT is also more common among closer-related species sharing greater sequence composition similarity, examples of what may be thought of as improbable HGT events, have been found (Popa & Dagan, 2011). Gene translocations and nucleotide substitution rates increased for recently transferred segments as compared to those of conserved or ancient genes (Hao & Golding, 2009). Collectively these points suggest that even though HGT plays a huge part in the evolution of prokaryotic organisms, some exchange limiting influences exist. These boundaries represent active gene flux preferences forming underlying platforms which governs GI exchange routines by forces such as physical distance, sequence composition, DNA integration mechanisms and micro-flora ecology (Popa & Dagan, 2011; Lefeuvre, et al., 2013). The extent and magnitude of these influences are still debated. However, they may assist in identifying the roles of HGT between taxonomically closer species and imports from distant origins. Moreover, we may use the information shown in these studies to approach the implementation of a system for the simulation of gene sharing networks.

1.1.6 Donor-recipient relations and graph theory

The process of HGT is becoming clearer as additional research brings forth new information surrounding the mechanisms behind it. It is evident from literature that GIs are generally obtained at similar insertion sites and the integrated regions become fragmented into a mosaic of foreign regions scattered across the genome in a non-uniform fashion due to the multitude of increased evolutionary forces acting upon them. Taking the information in the above mentioned studies into account, it then becomes increasingly difficult to examine single GI markers with attempts at relational analysis regarding potential donors and recipients. Roos et al. (2011), indicated that a significant proportion of GI in individual bacterial organisms shared similar dinucleotide frequencies. As explained, these are proposed as recurring acquisitions from the same donor to the same acceptor. Another explanation may be that they are fragments of the same transfer which were subject to translocation. To prevent the loss of information in this way, attempts at relational analyses should be considered from a perspective of clustered GIs. GIs in individual organisms should be grouped with reasonable technique to counteract the fragmentation and to represent independent transfer events. In addition, the magnitude of GI data already generated poses a challenge in terms of extracting sensible information for a myriad of species.

The use of networks represented as graphs integrated with GIs research has been proposed in previous literature. Popa et al. (2011), evaluated a total of 657 sequenced prokaryotic genomes by exemplifying HGT donor-recipient events with a directed network. The authors identified 446,854 protein-coding genes using atypical GC content values which deviated from sequential sequence windows over the genome compared to the GC content of the entire genome. Subsequently these genes were scanned for most likely donors by searching for orthologs with the highest sequence similarity to the acquired gene but excluding orthologs that share a common acquisition event with the specific gene. The method revealed donors for a subset of 32,028 of the identified genes. The subset of genes was then embodied into polarized HGT events as input to a directed network. The nodes were connected by directed edges pointing from the donor node to the recipient nodes. For additional calculations on the distribution of network connections, edge weights were quantified as the number of genes that were transferred from unique donor to unique recipient. ***Figure 1.10*** shows the network graph produced by using a force-directed layout. ***Figure 1.11*** reveals the observed number of connections for a typical node along with the edge weight distribution.

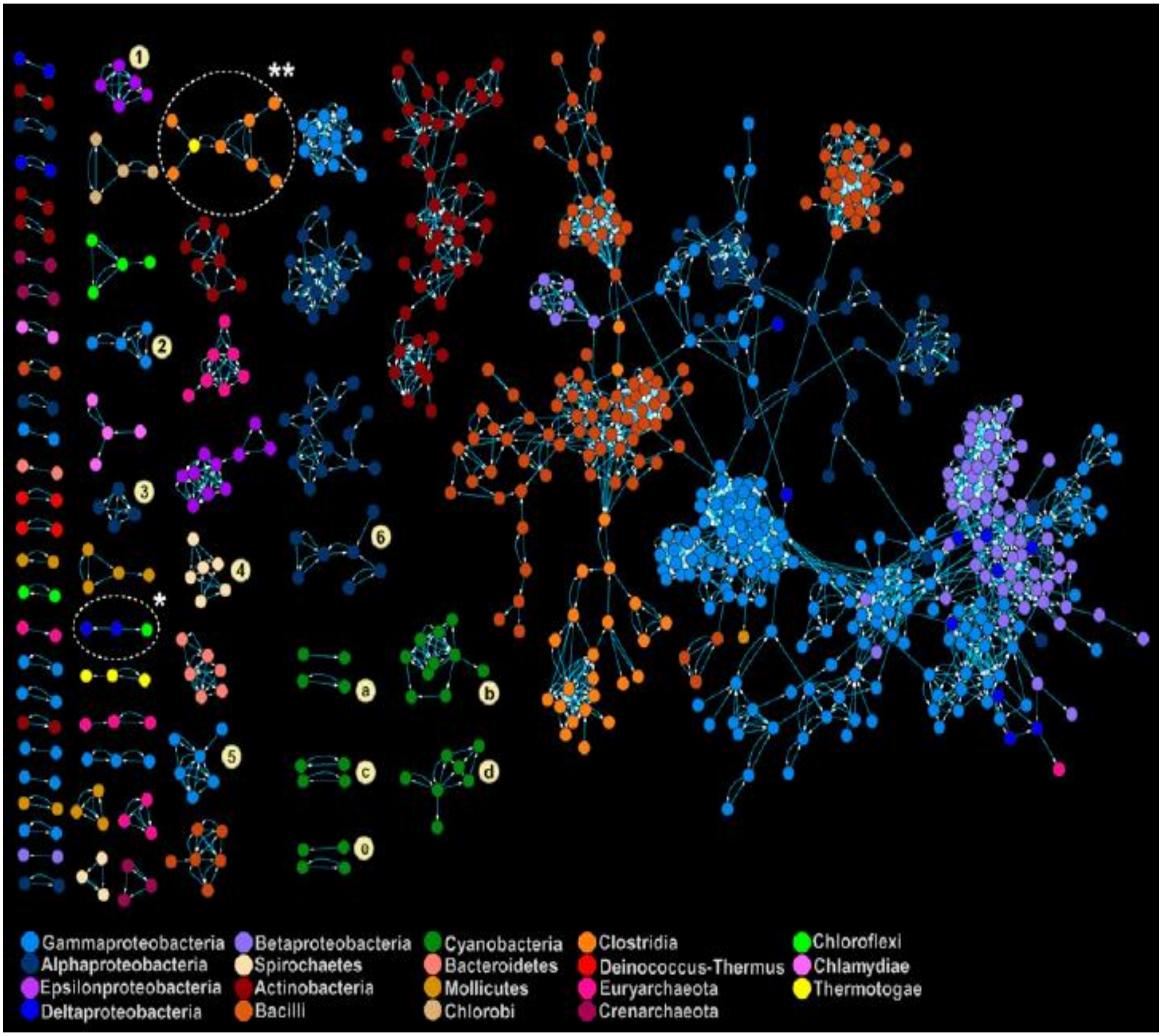


Figure 1. 10 The directed network of recent lateral gene transfers. Node color corresponds to the taxonomic group of donors and recipients listed at the bottom. Connected components of endosymbionts are marked with numbers: (1) Helicobacter, (2) Coxiella, (3) Bartonella, (4) Leptospira, (5) Legionella and (6) Ehrlichia. Clusters of cyanobacteria are marked with letters: (a) high-light adapted Prochlorococcus, (b) low-light adapted Prochlorococcus, (c) marine Synechococcus, (d) other Synechococcus, (e) Nostocales and Chroococcales. Adapted from the study Popa et al., (2011).

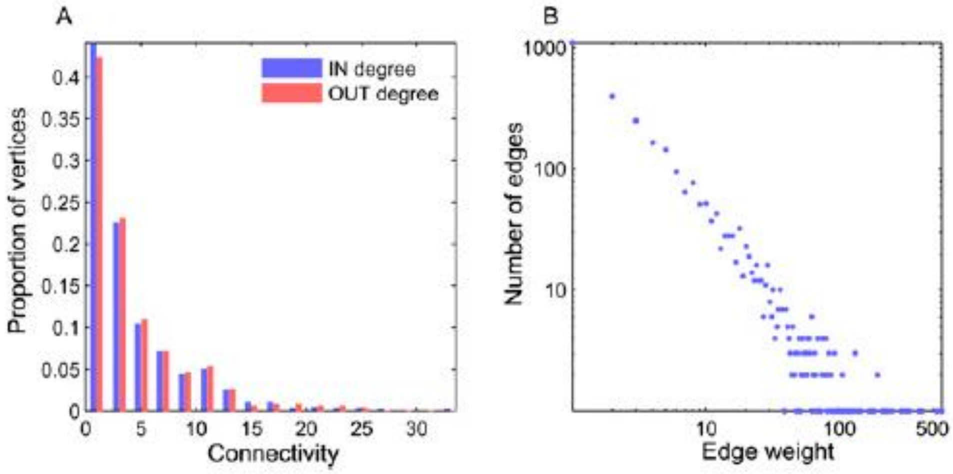


Figure 1. 11 Distribution of connectivity and edge weight in the directed network. Adapted from Popa et al., (2011).

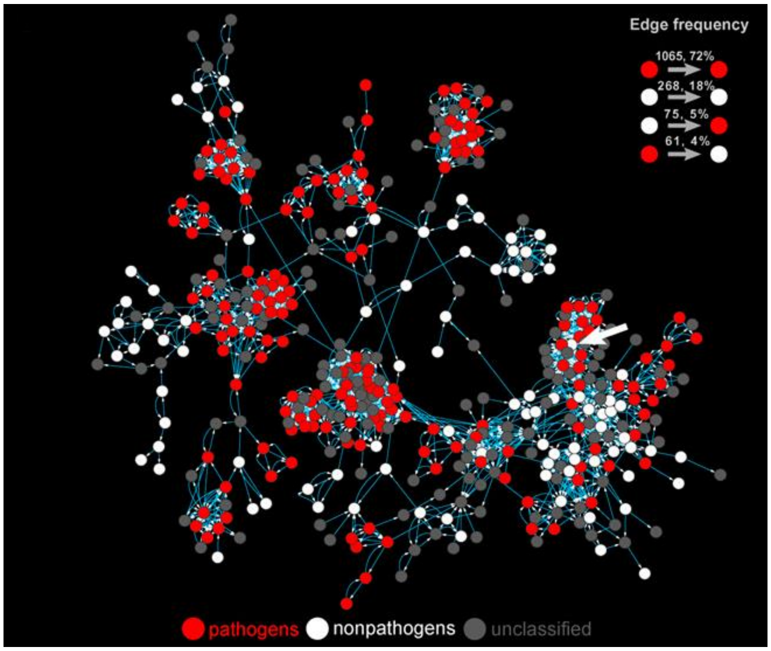


Figure 1. 12 Pathogens in the largest connected component of the network. The white arrow marks a non-pathogen (Bukholderia thailandensis) within a pathogenic community. Adapted from Popa et al., (2011).

We observe several communities forming from densely connected regions. These communities are mostly shared by members of the closely related or similar taxa. From the network, the authors conclude that most HGT occurs between donors and recipients within the same taxonomic group. They propose this is due to closely related species having similar genome sequence compositions. The usefulness of the network is further extrapolated by selecting a subset of the graph to illustrate interactions between pathogenic and non-pathogenic organisms as per ***Figure 1.12***.

For determination of GI flow, closer related species were expected to share gene exchange events more frequently and we would therefore also expect such species would harbor genes with parallel function or ontology. However, BLAST sequence similarity implies a functional conservation rather than a phylogenetic kinship. Single genes comparisons hold very little information in the context of the entire GI transfer relations. Since one GI may share the same single gene to a completely dissimilar GI in another microbe, an attempt to imply a HGT relationship between these GIs is erroneous and would give a false impression as to associations of specific donor-recipient interactions. Moreover, the exclusion of certain genera may also obscure the resulting network and is indicative of a requirement to provide specific suggestions of organisms to formulate more accurate visualizations.

1.1.7 Functional analyses of genomic islands

We have rarely ventured into correlations of co-occurring gene products found in horizontally transferred regions, which has resulted in a lacking perspective of associations between bacterial lifestyle and GI composition. Analyses of GI content and functions are challenging due to the lack of standardized annotation and a large amount of hypothetical genes. Several solutions have been proposed to overcome this issue. The COG database allows for the classification of proteins on the basis of the orthology concept and is commonly used in approaches of sorting gene products. The database of Clusters of Orthologous Groups (COG) provides an interface for the phylogenetic classifications of proteins encoded in 21 complete genomes of bacteria, archaea and from the yeast *Saccharomyces cerevisiae.* The database comprises of 2091 COGs that includes gene products of up to 83% of the bacterial and archaeal genomes and up to 35% of Saccharomyces *cerevisiae* (Tatusov, et al., 2000). The resulting gene products were divided into 25 unique functional classes. Using these functional classifications we may group genes, even from distantly related organisms, in a standardized measure. Another approach for the classification of gene products is done using the GeneOntology database (GO). The GO database is a project developed for describing gene products and characteristics using controlled vocabulary terms. The GO database has developed formal ontologies that represent over 40 000 biological concepts which are updated regularly (Ashburner, et al., 200). Proteins shared in two or more organisms provide strong evidence for the role of these proteins in all organisms, if the function or products is known in at least one. The authors describes the role of proteins to three main categories: Biological process, Molecular functions and Cellular components (Ashburner, et al., 200).

Fernandez-Gomez et al. (2012), studied the gene contents of GIs in 70 marine bacteria to reveal functional trends significant to the ecology. The author used the GO database for the classification of gene functions and identified the highest proportion of gene functions were related to DNA integration. Other categories also over-represented were found among cell membrane development and transposition. ***Figure 1.13*** shows the distribution of genes according to GeneOntology (GO).

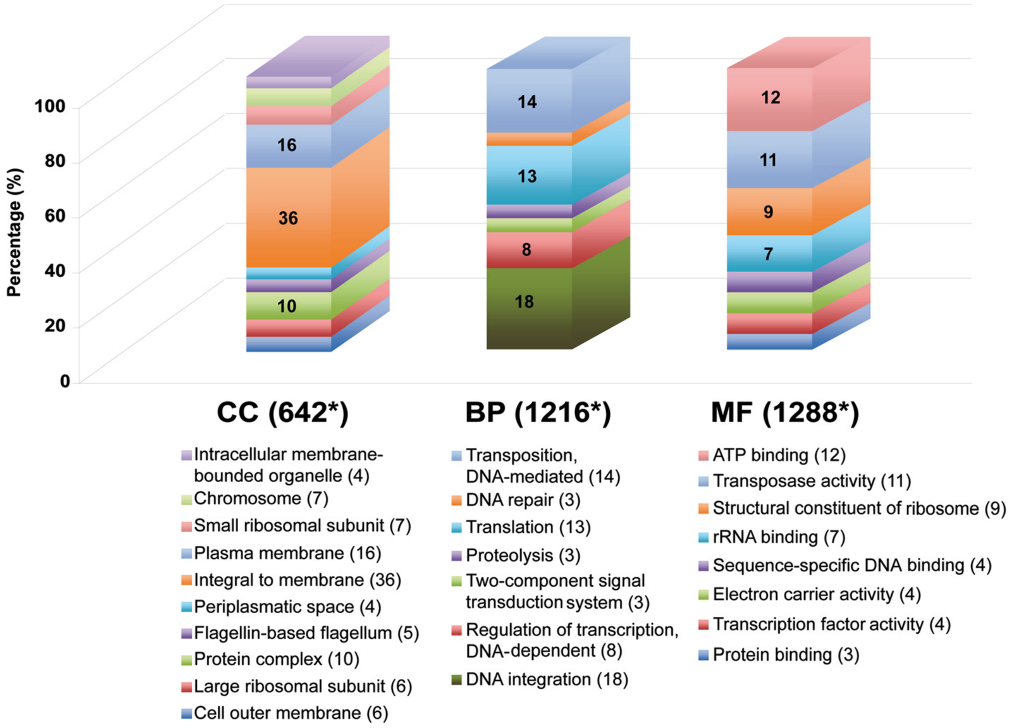


Figure 1. 13 Distribution of annotated genes within the GIs according to GO classification. Functional categories were split in three: Cellular Components (CC), Biological Processes (BP) and Molecular Functions (MF). Asterisk means the number of annotated genes to each main category. Numbers between parentheses indicate the percentage of appearance. Adapted from Fernandez-Gomez et al. (2012).

Merkl et al. (2006), evaluated the GIs of 63 bacterial organisms and attributed COG functions to the genes found within. The authors assigned the organisms to separate groups based on taxon and habitat to find significant functional class usage differences. In all groups, genes related to replication recombination and repair were always significantly over-represented, with genes related to intracellular trafficking, cell motility and defense mechanisms among the most predominant classes. Genes related to information storage and processing tracked among the least. Popa et al. (2011), evaluates the types of genes found in the GIs of 657 sequenced prokaryotes as an example of the use of the COG scheme to illustrate over-and-underrepresented gene functions among horizontally transferred genes. ***Figure 1.14*** reveals the classification of gene types per frequency of COG class.

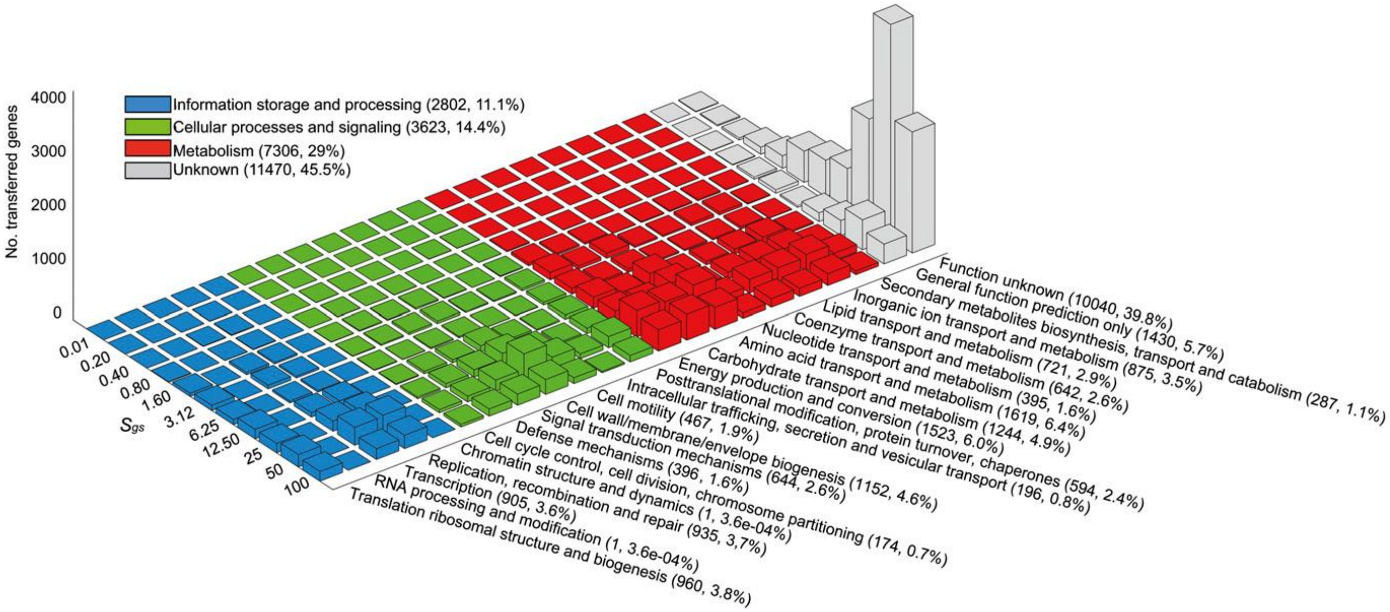


Figure 1. 14 The frequency of transferred genes by functional category and a genome sequence similarity index (Sgs) Adapted from Popa et al. (2011).

The results from these studies indicated that a large proportion of genes within the analyzed organisms were dedicated to the actual transfer and integration of the regions themselves. Given the importance of these integration genes, it is unlikely to encounter an underrepresentation of these classes in frequent and successful GI transfers. However, the exclusion of these classes may reveal interesting results among the remaining content, especially with regards to the co-occurrence of certain gene classes with others. The investigation of functional classes in a system for the monitoring of HGT will allow for the separation of GIs on a functional level and may furthermore assist in describing the roles of HGT among various organisms based on common gene content.

1.1.8 Summary

GI exchange routines in prokaryotic kingdom have been shown to exert great influence for the survivability and adaptations of microbes. Virulent strains have become prominent and widespread as pathogenicity islands become more accessible via HGT. Fortunately, we have identified some HGT barriers which require additional mechanisms for species which do not inherently share similar sequence compositions. In this regard, the promiscuity of microbes are countered to a debatable extent and we may use this information to assist in our investigations of gene movement assessments. To enable the formulation of expectations related to the outbreak of new or recurring diseases, gene movement patterns needs to be evaluated from a unified perspective. Therefore, we need an understanding of the mechanisms of HGT and suitable representations of unique HGT events. Our approaches to the detection of GIs in a bacterial host genome has reached a stage of maturity to the extent were we now have sufficient information to simulate gene fluxes even among diverse organisms. We may find foreign genetic regions in different parts of a host genome which share a likely common origin as a result of successive HGT between two organisms, or gene translocations of specific horizontally transferred regions. Therefore, GIs which have been identified as sharing compositional similarities in a single bacteria should subsequently be grouped to unify all the genes which formed part of a common HGT event. It is challenging to attempt relational analyses from a fragmented point of view, as the clustering of GIs in this way results in a minor partitioning of sources, and therefore indistinguishable origins. To visualize unique HGT events we may employ graph theory with clustering algorithms to reveal the underlying HGT flux structures. The resulting information may be used to derive preliminary hypotheses required for subsequent gene spread predictions. Furthermore, it is necessary to categorize the coding regions in GIs in a standardized format so that we may formulate possible association between the content and the role of HGT in a controlled fashion.

1.2 Methodology

1.2.1 SeqWord Project and the Pre_GI database

The earlier studies by Karlin et al. (2001) showed the evolutionary implications of dinucleotide and codon biases, with extended statistical approaches by various authors (Deschavanne, et al., 1999; Pride, et al., 2003) to finally propose methods to detect genomic regions harboring oligonucleotide sequence usages significantly different from the rest of the genome.

The SeqWord project (http://www.bi.up.ac.za/SeqWord/) incorporated tertranucleotide usage patterns as a matrix of deviations of observed versus expected oligonucleotide usage (OU) counts. These oligonucleotides or “words” are distributed logarithmically in sequences and deviations from the expected frequencies are quantified by ***Equation 1.1***.

$$\Delta_w = \Delta_{[\xi_1\ldots\xi_N]} = 6\times\frac{\ln\left(\dfrac{C^2_{[\xi_1\ldots\xi_N]|obs}\sqrt{C^2_{[\xi_1\ldots\xi_N]|e}+C^2_{[\xi_1\ldots\xi_N]|0}}}{C^2_{[\xi_1\ldots\xi_N]|e}\sqrt{C^2_{[\xi_1\ldots\xi_N]|obs}+C^2_{[\xi_1\ldots\xi_N]|0}}}\right)}{\sqrt{\ln\left(\left[C^2_{[\xi_1\ldots\xi_N]|0}\middle/C^2_{[\xi_1\ldots\xi_N]|e}\right]+1\right)}}$$

Equation 1.1

where ξ_n is any nucleotide A, T, G or C in a N-long word; $C_{[\xi_1\ldots\xi_N]|obs}$ is the observed count of a word $[\xi_1\ldots\xi_N]$; $C_{[\xi_1\ldots\xi_N]|e}$ is its expected count and $C_{[\xi_1\ldots\xi_N]|0}$ is a standard count estimated from the assumption of an equal distribution of words in the sequence: ($C_{[\xi_1\ldots\xi_N]|0} = L_{seq} \times 4^{-N}$). Expected counts of words $C_{[\xi_1\ldots\xi_N]|e}$ were calculated in accordance to the applied normalization scheme. For instance, $C_{[\xi_1\ldots\xi_N]|e} = C_{[\xi_1\ldots\xi_N]|0}$ if OU is not normalized, and $C_{[\xi_1\ldots\xi_N]|e} = C_{[\xi_1\ldots\xi_N]|1}$ if OU is normalized by the GC-content (Reva & Tümmler, 2005). These oligonucleotide usage patterns (OUP) represented the normalized measures of the sequence pattern similarity between two GIs. Recent acquisitions are known to constitute oligonucleotide usage patterns of former hosts which allow the possibility to trace down their distribution patterns and identify their putative donors (Roos & Mark , 2011).

Additionally, a distance percentage (D) was calculated between the host genome pattern and its relevant putative GI sequence pattern as the sum of absolute values of the subtractions between ranks of identical oligonucleotides or words (w, in a total 4^N different words) after the ordering of words by $\Delta_{[\xi_1 \ldots \xi_N]}$ values in patterns i and j by ***Equation 1.2***:

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

Equation 1.2

This distance percentage (D) was in turn used to classify how divergent the OUP of the GIs were from their host genome OUP. Consequently, D might be interpreted as the amount of relative evolutionary time the process of amelioration was in effect. Therefore, this measure served as a representation of the relative time of insertion (Reva & Tümmler, 2005). Hereinafter in the text percentage in D values will be omitted.

A large number of bacterial organisms and plasmids were scanned with the above mentioned procedures from SeqWord and putative genomic islands saved to a MYSQL database called Pre_GI (http://pregi.bi.up.ac.za/). The database to date consists of **26,744** predicted GIs from a total of **2,407** analyzed bacterial replicons downloaded from the NCBI (http://www.ncbi.nlm.nih.gov/) in the Genbank accession format. The Pre_GI database is focused around creating a research environment for the mining of bacterial DNA sequences though the use of genome linguistics. The database enables users to deconstruct the ontological relationships between bacterial mobile genetic elements and to examine the global spread of genetic vectors through taxonomic barriers. The Pre_GI database contained GI composition similarity scores computed by comparisons of sequence pattern similarities (OUP) for all GIs in the database. Sequence pattern similarity comparisons among GIs produced 69,176,627 matching scores, indicating the strength of OUP resemblances to every other respective GI.

1.2.2 Markov Clustering Algorithm

The use of the Markov Clustering Algorithm (MCL) is not uncommon in Bioinformatic studies and is generally regarded as well suited for high-throughput data. MCL is mostly known for its applications in protein interaction networks within the field of Biology. However, because the intended use of MCL is not to find clusters but rather to identify the underlying cluster structure, it may be used regardless of the number of nodes and without making any assumptions of the data a *priori*. MCL is also appropriate for representations of graphs due to its intuitive modelling of node and edge implementation as community interactions. MCL is indicative of a superior technique with regards to execution time and processing power requirements as it has been shown to outperform other clustering algorithms (Brandes, et al., 2008; Foggia, et al., 2008; Van Dogen, 2000; Guzzi & Cannataro, 2012).

With regards to the implementation of MCL in the analysis of HGT, GIs may be presented as nodes and high sequence composition pattern similarities as edges between the GIs. We may also provide raw BLAST sequence similarity scores as edge attributes, although base sequence similarity allows us to infer a functional resemblance of gene products, rather that of a common origin.

The MCL algorithm clusters several nodes into groups by simulating stochastic flows between edges and analyzing the distribution of the flow spread (Van Dogen, 2000). By observing the walks of a network we can derive certain behaviors of the elements of the network. For each walk the probability of taking a next route may be described as the probability that the particular route will be chosen next, based on the number of surrounding connections. Thus, highly connected regions are more likely to cluster together and the probability of diverting to a distant node from another cluster only becomes prevalent once closer related nodes have been routed. MCL also considers areas of weak and strong connections to iteratively decease the flow where it is weak and increase the flow where it is strong. Formally, MCL represents a matrix of a graph M. An associated Markov matrix MxM of the normalized values of M then relates to the inclination of a node to be attracted to each of its neighboring nodes (Guzzi & Cannataro, 2012). However, since all nodes in MxM have a non-zero attraction to each of its neighbors at this stage, additional computation is required to distinguish between nodes which are strongly connected to each other but not in the same region in the matrix. The consecutive powers of MxM is then used to calculate the increasing probability of attractions for distant but allured nodes. The algorithm uses a parameter, called the inflation value, to change node attraction probabilities. Thus based on the inflation parameter, the structure of a cluster network changes so that for inflation values greater than one, an increasing amount of clusters are obtained (Van Dogen, 2000). ***Figure 1.15*** below is representation of the stages of MCL execution (Van Dogen, 2000).

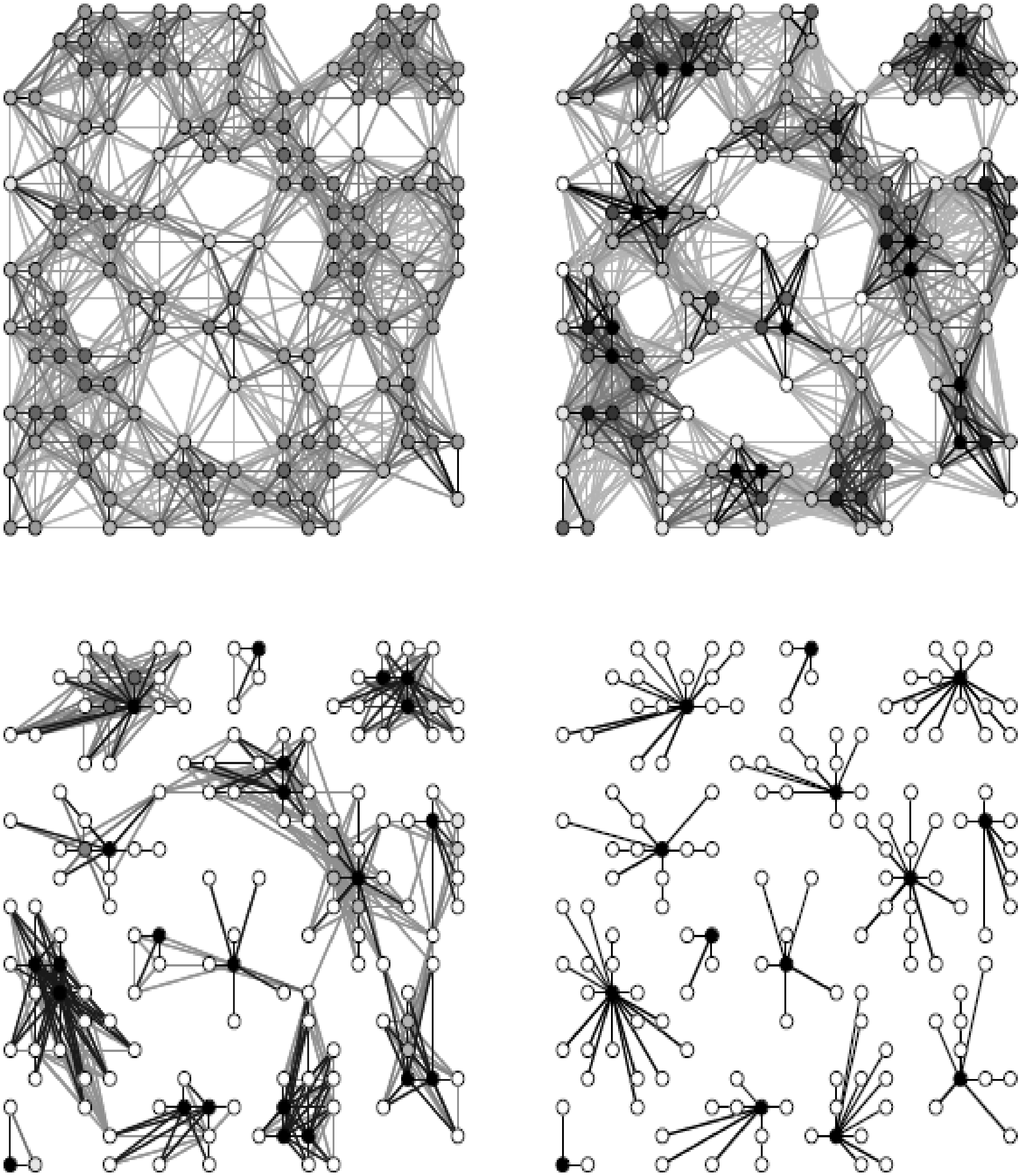


Figure 1. 15 Successive stages of MCL flow simulation. Adapted from (Van Dogen, 2000)

1.2.3 Cytoscape Freeware

Cytoscape is an open source software platform used for the modelling of interaction networks through visualization. The program was initially designed for molecular interaction networks in biological research, with integrated annotation and gene expression profiles as a standard. The program also allows for the addition of external applications *via* designed plug-ins. These plug-ins enables the program to perform multiple functions on a single input of data. ClusterMaker is a Cytoscape plugin which incorporates different clustering techniques including MCL (Morris, et al., 2011; Cline, 2007).

Cytoscape allows for the addition of external data to attribute edges and nodes with enhancements to the graph at each import. We may therefore initiate the MCL procedure with composition similarity scores as weights and provide more information to the Cytoscape server for the consumption of attributes related to functional annotation, BLAST similarity scores and taxonomic classifications. ***Figure 1.16*** shows an example usage and output of Cytoscape with the clusterMaker plugin. Here the separation of bacterial and archaeal BLAST sequence similarities are illustrated after the application of MCL and the force-directed layout. MCL detects a difference in the common genes in between the kingdoms of *Archaea* and *Bacteria* as shown in the two separate clusters.

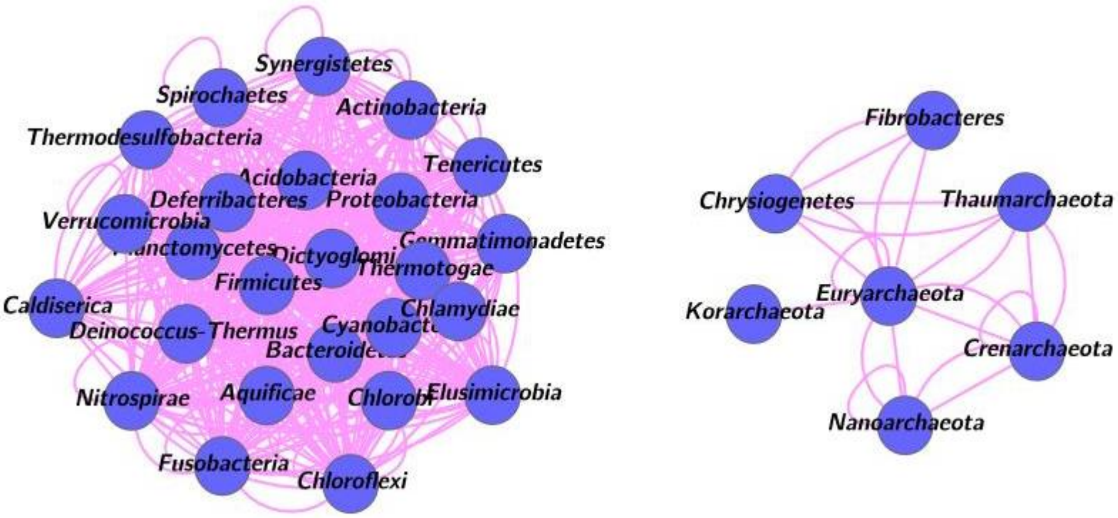


Figure 1. 16 MCL Clusters of phyla among gene networks reveal different common genes among Bacteria and Archaea.

1.3 Project Overview

The objective of the project is to develop a system (Flux Visualizer) for the graphical monitoring and large scale analysis of MGEs identified in the Pre_GI database, in order evaluate and determine non-random HGT event associations, and to highlight GI exchange pathways between various taxonomic ranks by integrating GI sequence composition patterns and relative insertion time similarities. The visualization of the donor-recipient relationships should be implemented through the Cytoscape freeware (Cline, 2007) to ensure applied procedures are accurate and up to date. Additionally, the Markov clustering algorithm will be employed with modeled layouts from Cytoscape to present unique elements revealing probable donor-recipient relationship by means of graph construction. Studies surrounding gene flux relations are usually subject to certain taxonomic involvements and would therefore also require selection criteria's to identify only the most prominent inclusions. The system will enable the user to select organisms of interest by entering Genbank accession numbers and taxonomic keywords as well as to provide suggestions from organisms identified to share significant BLAST hits among their GIs. In this regard, networks of successive BLAST matches will allow users to admission non-related inclusions based on common gene functions. Furthermore, all GI genes will be assigned to functional categories most accurately portraying their roles in individual GIs to assist in describing the characteristics of HGT between the targeted species.

The project had four main goals to achieve:

1. Investigate the level of fragmentation of genomic islands (GIs) in bacterial genomes. The working hypothesis is that the number of events of horizontal gene transfer was overestimated in the Pre_GI database due to fragmentation of GIs. To be able to predict fluxes of GIs through bacterial species, it is important to reconstruct entities of mobile genetic elements. This aim will be addressed by grouping GIs through compositional similarity (Chapter 2).

2. Construct a system for visualization of HGT events based on oligonucleotide usage similarities among the reconstructed entities of mobile genetic elements. To enable interpretation of the gene spread between organisms selected by users of the system, MCL and graph theory approaches will be applied to simulate and visualize fluxes of GIs through various bacterial

species. The system will be implemented on a web-based platform integrated with a server instance of Cytoscape to ensure appropriate procedures are applied (Chapter 3).

3. Evaluate the distributions of co-occurring functional class and genetic contents among the reconstructed entities of mobile genetic elements. We hypothesize that it is possible to separate and cluster groups of GIs based on functional preferences. To be able to describe the functional aspects of fluxes of GIs visualized with the system, the COG database will be used to assign functional roles to all of the coding regions in each GI. This aim will assist in describing the practical roles of HGT among selected organisms (Chapter 4).

4. Validate the system functionality by investigating associations of BLAST hit matches among groups of GIs as a case study. To illustrate the use of the system, organisms which show significant BLAST hits matches between their groups of GIs, will be used as example inputs to the system. This aim will be addressed by visualizing and describing ontological HGT links between the selected organisms (Chapter 5).

Chapter 2. Reconstruction of GI entities by compositional clustering of identified inserts of foreign DNA in bacterial genomes

2.1 Grouping of GIs

It is known from literature that GIs are generally exposed to accelerated substitution rates as well as increased gene translocations soon after integration (Hao & Golding, 2009). Consequently, newly integrated regions are frequently fragmented and these fragments may incorrectly be perceived as separate HGT events. Furthermore, unrelated GIs may be expected to contain homologous genes by chance and this would give a false impression of gene exchange between the hosting organisms. To obtain an accurate representation of the HGT relationships between microorganisms in the Pre_GI database (http://pregi.bi.up.ac.za/index.php), HGT events should be clearly distinguished from consequent GI fragmentation and rearrangement. Donor-recipient links formulated in such a way, will serve to identify unique GI origins among numerous equivalent matches, since we will be able to recognize which GIs were fragmented or formed part of recurring transfers from a common origin. To determine which GI inserts resulted from the fragmentation of longer entities, every GI predicted in 2,407 replicons and stored in Pre_GI database were subsequently compared against all of the other GIs in their respective host organisms. The grouping of GIs was based on oligonucleotide usage pattern (OUP) similarities (100 – GI-to-GI-distance-value) and the distance-to-host values (***Equation 1.2***). Similarity in OUP allowed for the inference of a common origin of GIs, and similar differences in OUP distance-to-host values lead to the conclusion that GIs were acquired at the same time. The rationale for the latter assumption was that foreign DNA inserts in bacterial genomes experience the DNA amelioration pressure smoothing out the difference in OUP patterns between GIs and the host organism (Wiedenbeck & Frederick, 2011). Both of these measures were extracted from the Pre_GI database and analyzed using Python scripts.

It was assumed that the similarity of randomly generated DNA sequences may be expected to be around 50 (see ***Equation 1.2.***) and it was empirically proved that pairs of GIs, which showed similarities from 50 to 70, were just random combinations (Bezuidt, et al., 2011). In this paper it was shown that for each OUP similarity rank the number of GI pairs producing significant BLASTN hits became prevalent once

OUP similarity reached above 75, and was therefore used as a lower threshold of meaningful compositional similarity (***Figure 2.1***). Therefore, in this study GIs sharing 75 or higher OUP similarity were considered as possibly originating from the same source (a donor organism or a pool of genetic vectors shared by a microbial community).

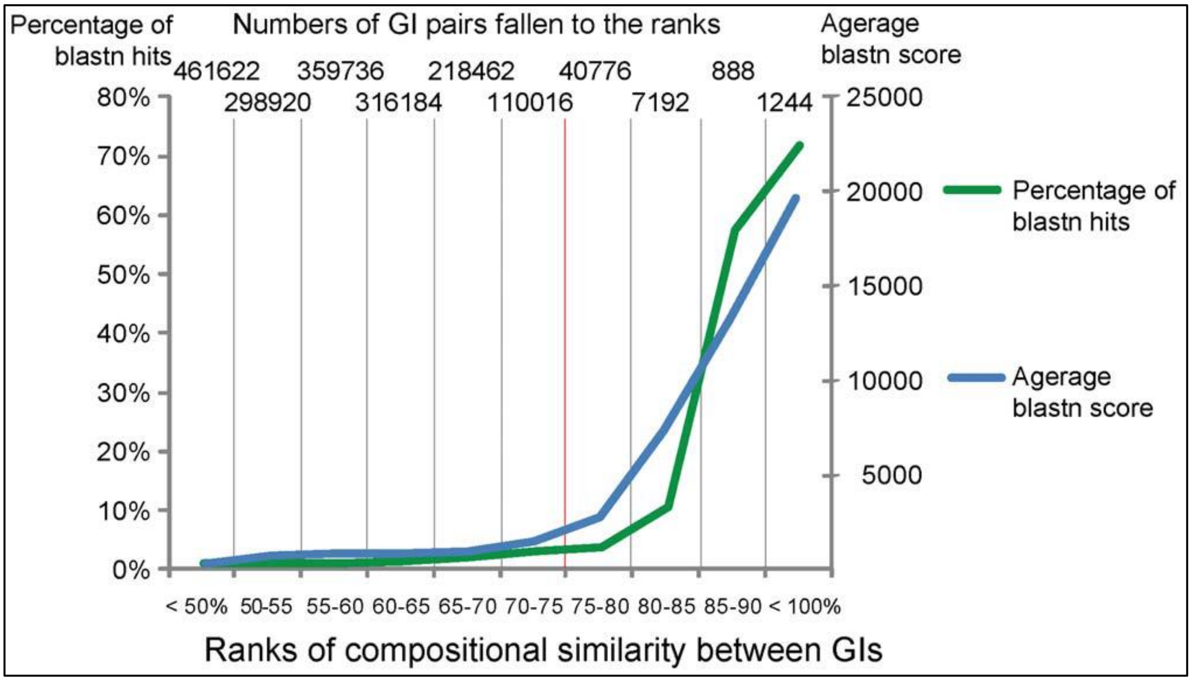


Figure 2. 1 Percentage of BLASTN hits between pairs of GIs shared ranked compositional similarity. The threshold value used for clustering of GIs is depicted by a vertical red line. Adapted from Bezuidt, et al. (2011).

Additionally, we considered using several threshold values for the absolute difference in pattern distance- to-host (DID) measures among any two GIs. GIs were only considered for grouping if they showed OUP similarities above 75 and also similar DID values supposing the same time of acquisition by the host organism. DID values were used as a representation of the timespan between each GI integration. This measure was formulated by the difference of each GI's distance-to-host value (***Equation 1.2***). A minimal threshold value of 15 GI distance-to-host value was used as a cutoff point to exclude old significantly ameliorated GIs, as their origins could not be reliable predicted. To identify suitable pattern distance-to-host differences between GIs, the clustering of GIs was performed five times iteratively with 5, 10, 15, 20 and 25 DID cutoff values. Remarkably, the results of the clustering with all of these cutoff values were quite consistent (**Figure 2.2**). This indicates that HGT events occurs infrequently and that fragments of one horizontally acquired insert may be distinguished from fragments of other successive inserts even if all of them came from the same origin.

Figure 2.2 shows how many groups of GIs were identified in the genomes of different organisms when different DID cutoff values were applied (denoted as G1 – all GIs were grouped into a single cluster; G2 – GIs were separated to two clusters; and so on). Note that the OUP similarity cutoff was 75 for all iterations of clustering. The boxplots denote clustering with 10, 15 and 20 cutoffs – bottom, middle and top lines of the box, respectively. Outliers for cutoff of 25 and 5 were depicted by short bars and open cycles. Clustering with a DID threshold of 5 may presumably lead to an overestimated grouping, while results with all other thresholds were consistent. Taking into account that the standard deviation of all the distance-to-host values (***Equation 1.2***) for the all GIs in the Pre_GI database was 8.27, the DID of 10 was selected as the optimal DID cutoff value.

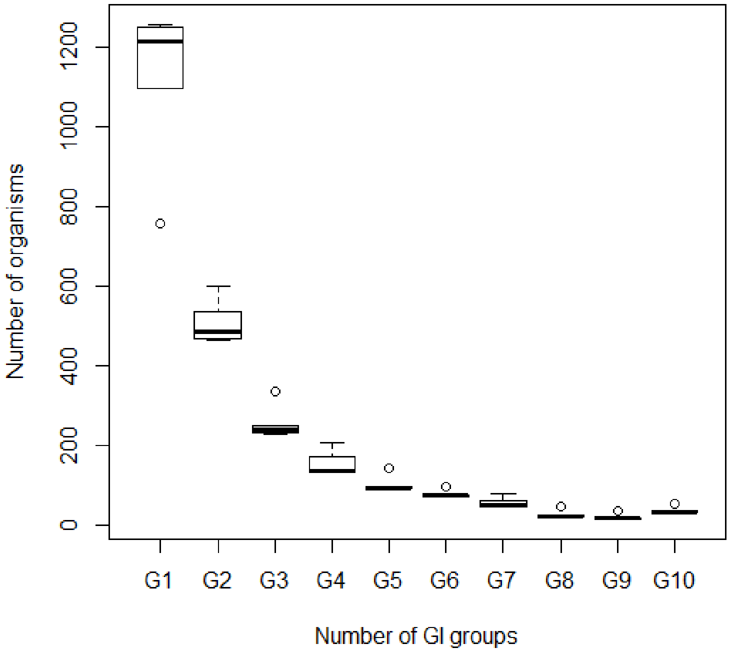


Figure 2. 2 The boxplot frequencies of organisms assigned with GI groups for all DID thresholds of 5, 10, 15, 20 and 20 combined.

Figure 2.3 shows the process diagram for the pseudo code of grouping of GIs in each host organism and the procedural implementation of the algorithm. The Pre_GI database was used as the data source and every GI with DID value (***Equation 1.2***) greater or equal than 15 was extracted and evaluated. For each organism, the first group of GIs was created for all of the GIs which were found to share 75 or higher OUP similarity as well as less than 10 DID-difference between each other. Sequential groups of GIs were only created if one or more of the GIs did not share sufficient OUP similarity, or had a higher DID-difference to members of the G1 group or other existing groups. Thus, G1 was mostly the biggest group of GIs in a given organism, and the remaining GIs were clustered into smaller groups G2, G3 and so on. It was possible for certain GIs to fall into more than one group given the GI had OUP matches to one or more GIs allocated in different groups. In such a case the GI was allocated to one of the groups, based on the maximum measure of OUP similarity and the highest amount of OUP matches to every other GI in each separate group. In many bacterial genomes the identified GIs showed a significant compositional polymorphism that was most likely associated with the acquiring of these GIs from different sources or several sequential acquisitions. Those GIs found in the same bacterial genome, which shared OUP and DID similarity, most likely resulted from the fragmentation of a longer insert.

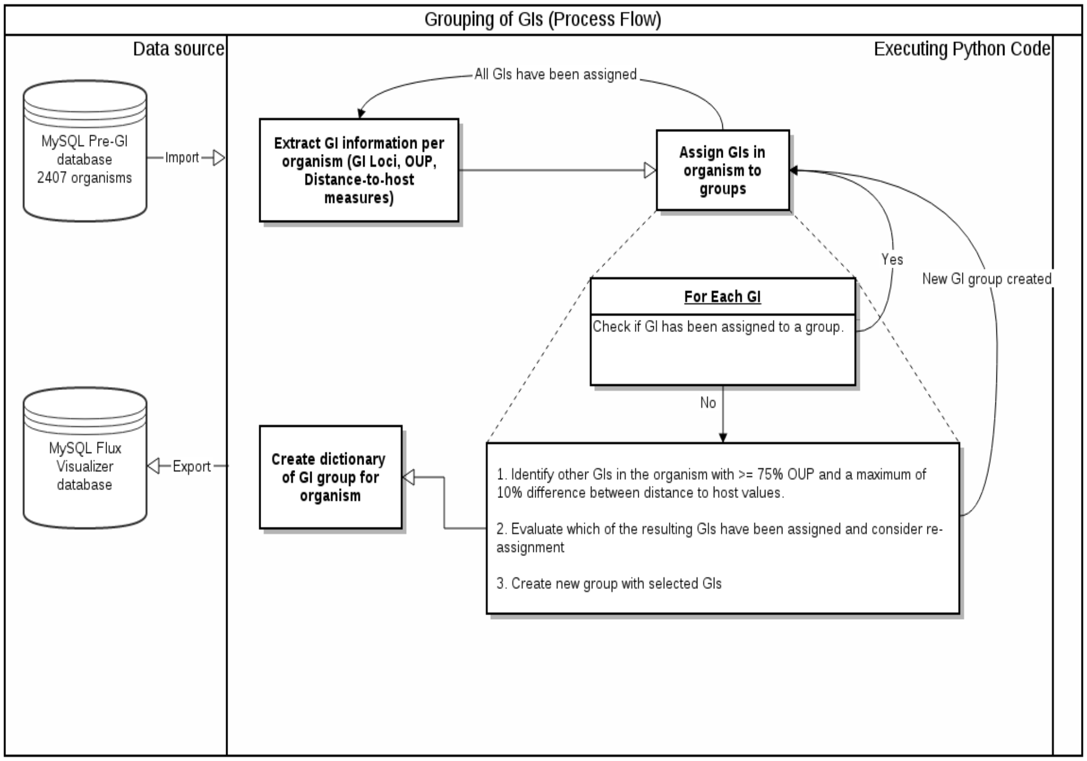

Grouping of GIs (Process Flow)
Data source
Executing Python Code
MySQL Pre-GI database 2407 organisms
Import
All GIs have been assigned
Extract GI information per organism (GI Loci, OUP, Distance-to-host measures)
Assign GIs in organism to groups
Yes
New GI group created
For Each GI
Check if GI has been assigned to a group.
No
MySQL Flux Visualizer database
Export
Create dictionary of GI group for organism
1. Identify other GIs in the organism with >= 75% OUP and a maximum of 10% difference between distance to host values.
2. Evaluate which of the resulting GIs have been assigned and consider re-assignment
3. Create new group with selected GIs

Figure 2. 3 Pseudo code for grouping GIs in each organism.

Figure 2.4 shows frequency of GI groups in different organisms after completion of the clustering procedure with 75 OUP similarity and 10 DID difference cutoffs. Combining GIs in this way produced one or more groups of GIs in each organism. Each GI group was denoted as "G" along with a number indicating the number of splits in each organism's GI combinations. For instance, an organism with three distinct groups of GIs would have been assigned G1, G2 and G3. A significant proportion of organisms was found to contain a single group comprising all of their GIs. In contrast, several organisms with up to 10 GI groups were observed. These organisms belonged to following genera: *Paenibacillus*, *Bacteroides*, *Hahella*, *Teredinibacter*, *Denitrovibrio*, *Geobacter*, *Nitrosococcus*, *Cellvibrio*, *Spirochaeta*, *Pyrobaculum*, *Parabacteroides*, *Desulfovibrio*, *Corynebacterium* and *Xenorhabdus*.

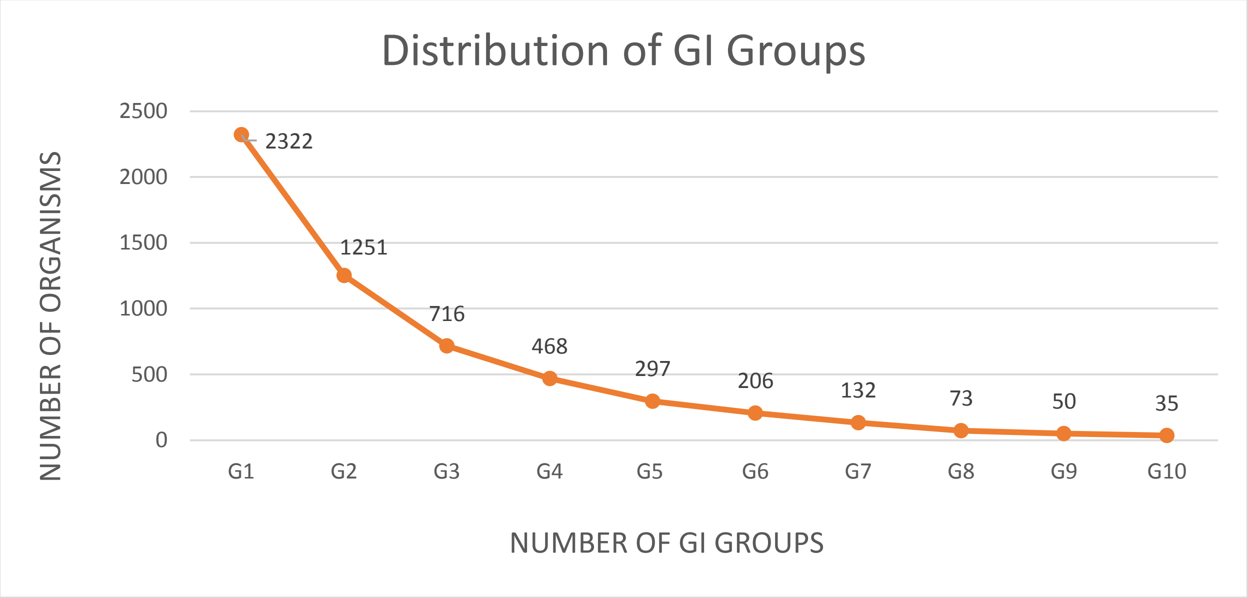


Figure 2. 4 Frequency of bacterial organisms with different numbers of GI groups.

Figure 2.5 shows a boxplot for numbers of GIs distributed in every group. The first group of GIs contains the most GIs because it was used as the initial container for the first GI comparisons per organism. Thereafter, GIs which did not satisfy the OUP and DID criteria of the initial group were combined into a second group. In the same manner, additional third, fourth, fifth and so on groups were created until every relevant GI was assigned to a group. The steady decline in the number of GIs in each successive group translates to fewer OUP similarities between the GIs within the succeeding groups.

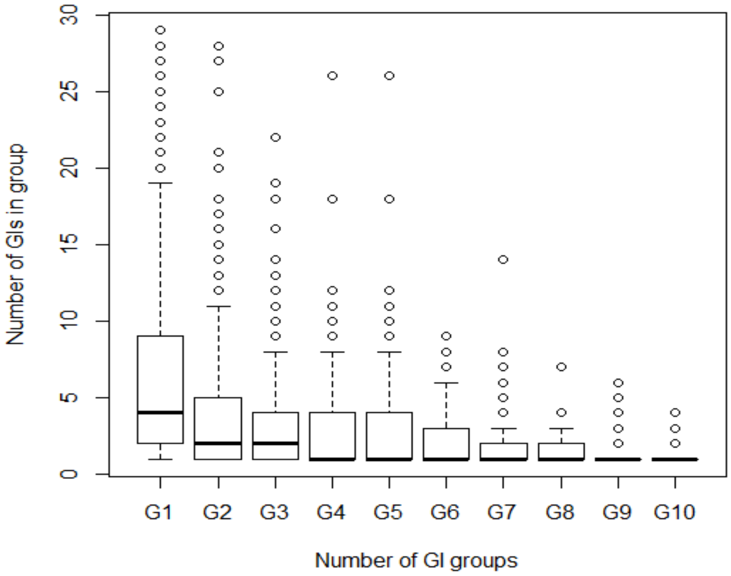


Figure 2. 5 Distribution of GI counts in each group for DID threshold of 10.

To check the appropriateness of the GI clustering, the relationships between groups were visualized using the LingvoCom toolset developed previously, which is available from the SeqWord project web site (http://www.bi.up.ac.za/SeqWord/lingvocom/index.html). ***Figure 2.6*** shows 3D projection of GIs from *Nitrosomonas europaea ATCC 19718* (NC_004757) by comparison OUP similarity values between each GI, the host organism chromosome and chromosomes of three distant organisms used as outgroups: *Clostridium thermocellum ATCC 27405*; *Salmonella enterica Typhi Ty2* and *Acidovorax ebreus TPSY*. Whole genome and GI patterns are depicted on the 3D-plot by squares and circles, respectively. The GIs of *Nitrosomonas europaea ATCC 19718* which were assigned to the same groups were marked with "G" indicators and color coded (green, purple, black and red for G1, G2, G3 and G4, respectively). In this example we observe a significant separation of GIs presumably acquired by the host organism from different sources. However, GIs within each group showed consistency in OUP patterns. Interestingly, we can see that the GIs in the same group are also spatially bundled together. In particular, GIG3 seems much more similar to the sequence patterns of the *A. ebreus TPSY* genome while GIs of GIG1 showed some similarity to Clostridial origin.

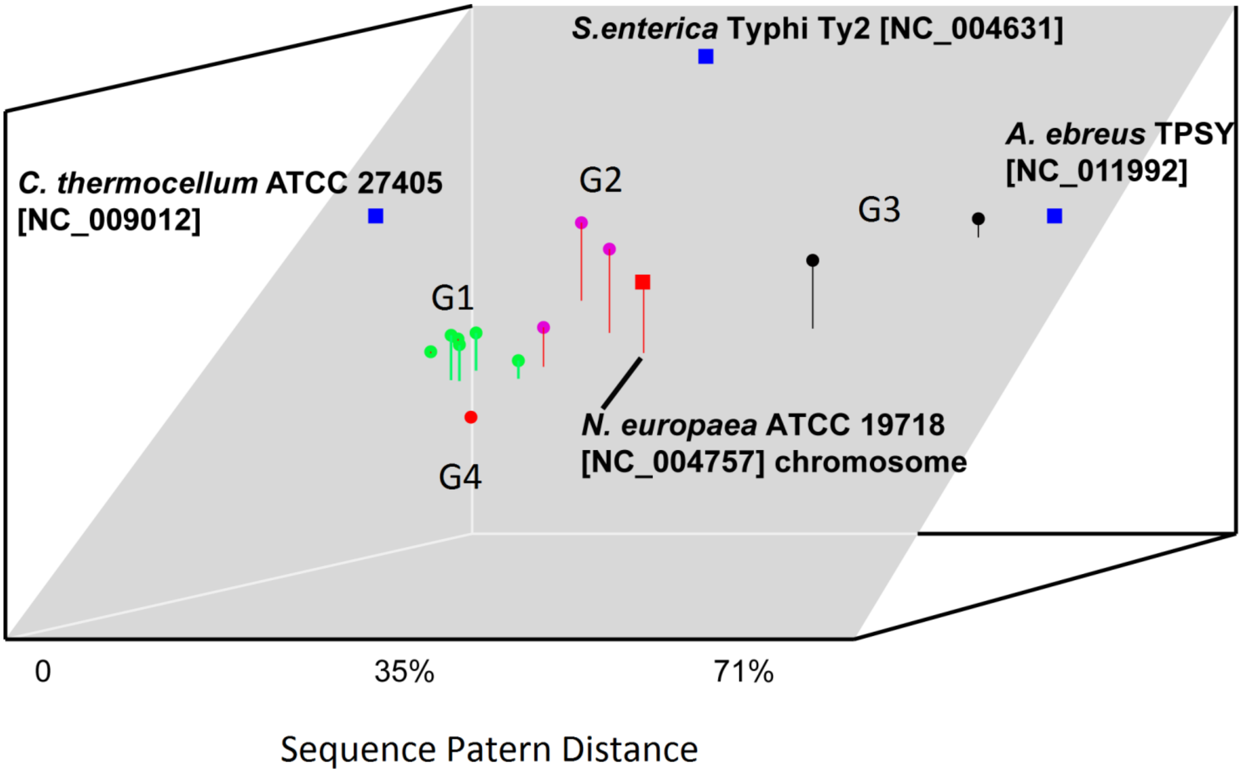


Figure 2. 6 Example of grouping of the GIs of Nitrosomonas europaea ATCC 19718 in 3D space by LingvoCom. Each GIG group is color coded to clearly observe the separations of their GIs.

A total of 5577 groups were created from 24,858 GIs accessible in Pre_GI, which were found in 2,407 different bacterial replicons, with an average of 5 GIs per group. The grouping results were saved to a MySQL database for subsequent analysis. In 27 organisms we found extreme counts of ≥ 30 predicted GIs assembled into a single group, meaning a single acquisition event followed by significant fragmentation. ***Table 2.1*** lists 27 organisms where multiple GIs were clustered into single groups.

Species	GI Count
Clostridium botulinum	30
Delftia acidovorans	30
Sebaldella termitidis	30
Clostridium botulinum	30
Bacillus cereus	30
Bradyrhizobium japonicum	30
Bacillus thuringiensis	31
Clostridium beijerinckii	31
Streptomyces cattleya	32
Clostridium botulinum	32
Clostridium cellulolyticum	32
Haliangium ochraceum	32
Stackebrandtia nassauensis	32
Clostridium botulinum	32
Bacillus cereus	33
Verminephrobacter eiseniae	33
Corallococcus coralloides	33
Bacillus anthracis	34
Clostridium phytofermentans	37
Myxococcus xanthus	44
Clostridium ljungdahlii	44
Myxococcus stipitatus	47
Trichodesmium erythraeum	49
Clostridium pasteurianum	52
Clostridium saccharoperbutylacetonicum	54
Methylobacterium nodulans	55
Stigmatella aurantiaca	64

Table 2. 1 Organisms with extreme number of GIs grouped into single G1 groups that indicated a single HGT event followed by severe fragmentation.

2.2 Calculating of OUP similarity values between groups of GIs

Since all the considered GIs had been grouped, new similarity scores were necessary to enable donor-recipient inferences between GI groups instead of using of OUP similarity values calculated for individual GIs. Two possible scores to measure similarity between groups of GIs were considered. The first score was calculated by extracting the maximum OUP similarity values between any two GIs in each separate group. This score represented the convergent OUP similarities between the two groups and was necessary to mimic a single-linkage state of GI groups. In a single-linkage state, any two observations (or GI groups) are only separated by the shortest distance between them (Zhang, et al., 2012). The second score was calculated as the average OUP similarity shared among all GIs in both groups. This score represented the centroid-linkage state of GI groups. In a centroid-linkage state, the distance between two groups is the distance between the means of the observations in the group. Since the process of amelioration has such a drastic effect on the composition of GIs in the hosting organisms, we hypothesized that the single-linkage state of the groups of GIs would produce more sensitive GI group similarity scores as an input to MCL algorithm. Both the maximum and average group scores were saved to a MySQL database to serve as the potential lookup values for MCL analyses. GI group weights generated in this way enabled subsequent analyses to initiate from a perspective of unique HGT events.

First, we analyzed the distribution of OUP similarity values between individual GIs. ***Figure 2.8*** shows frequencies of OUP similarity values in all pairs of GIs from Pre_GI database in three ranks (pairs with OUP similarity < 75 were ignored). Observation of only a few cases when OUP similarity was > 90 corresponded to our expectation counting for the effect of amelioration and other mutational pressures acting upon newly integrated genomic loci. The bulk of OUP scores fell into category 75 – 85 OUP similarity.

In the next step we were interested to see whether the replacement of OUP similarity values between individual GIs with those calculated for GI groups will change the distribution shown in ***Figure 2.8*** or not. Expectedly enough, applying of a centroid-linkage approach significantly decreased the number of pairs of GI groups linked by a significant OUP similarity (***Figure 2.9***). It was concluded that the maximum linkage approach was better suited for this study, as the use of average values may conceal too many important relations between groups of GIs due to significant pressure of genome amelioration processes.

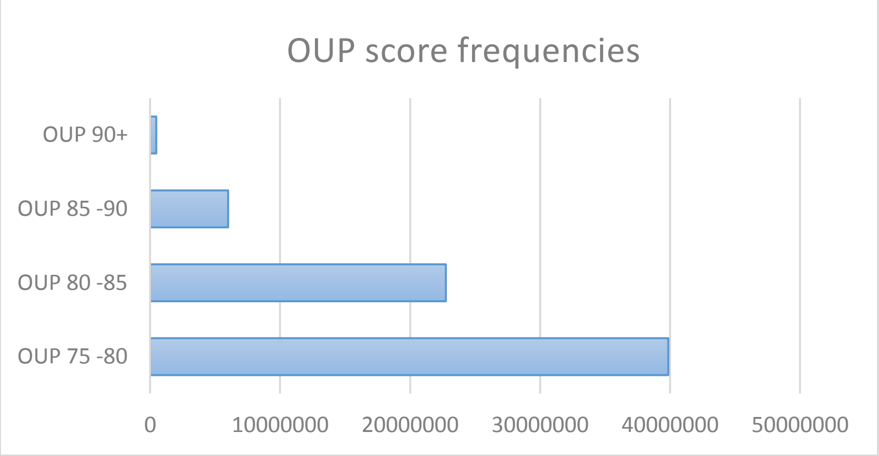


Figure 2. 7 Frequencies of pair of GI OUP scores as per the Pre_GI database.

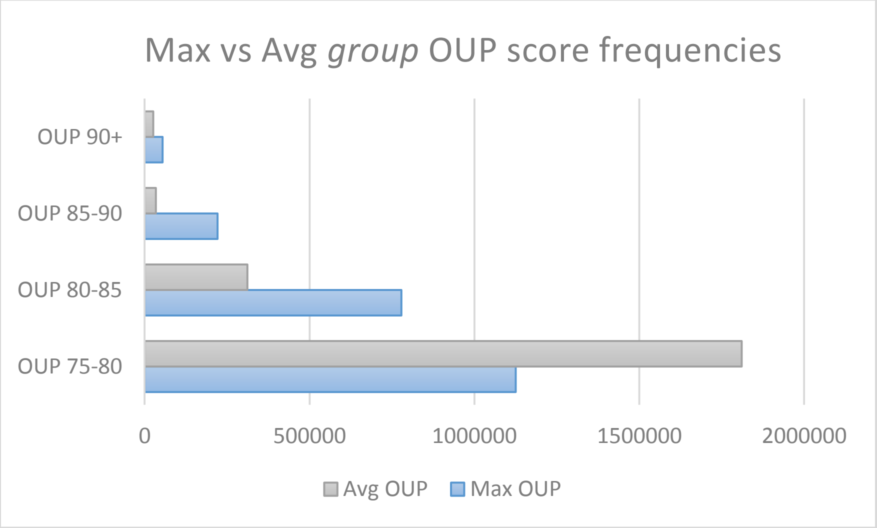
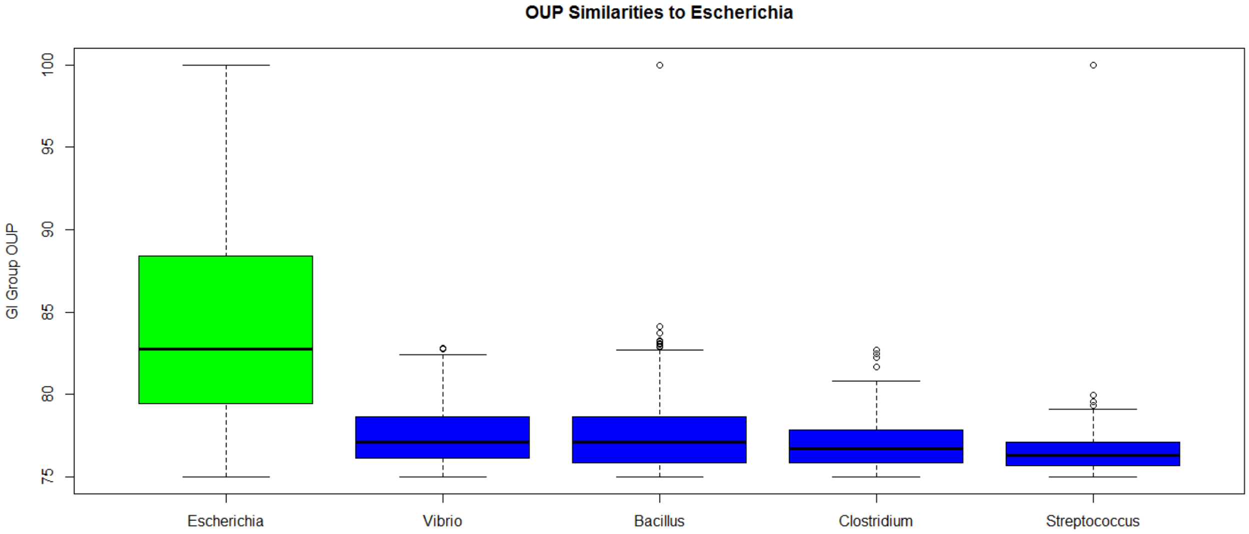


Figure 2. 8 Distribution of average and maximum OUP similarity links between groups of GIs.

To validate the precision of the newly generated OUP similarity scores and to confirm preference of the maximal single-linkage approach, we analyzed the distribution of similarity scores between different taxonomic units. The hypothesis was that groups of GIs harbored by organisms belonging to the same taxonomic unit should show a significantly higher level of OUP similarity than those from different taxonomic units. Plotting of both the single-linkage and centroid-linkage approaches would indicate which of them produced more sensitive output. To illustrate this validation, we performed analyses on a subset of organisms belonging to diverse genera counting for a more or less equal abundance of GIs from these genera in Pre_GI database. Calculated OUP similarities between groups of GIs within a selected reference taxonomic unit were compared to distances calculated between groups from other units to the reference one. In a series of experiments, the genera of *Escherichia, Bacillus* and *Streptococcus* were used as references. The results of comparisons for maximal single-linkage and centroid-linkage approaches are shown in ***Figure 2.10*** (A, B, C) and ***Figure 2.11*** (A, B, C) respectively.

A



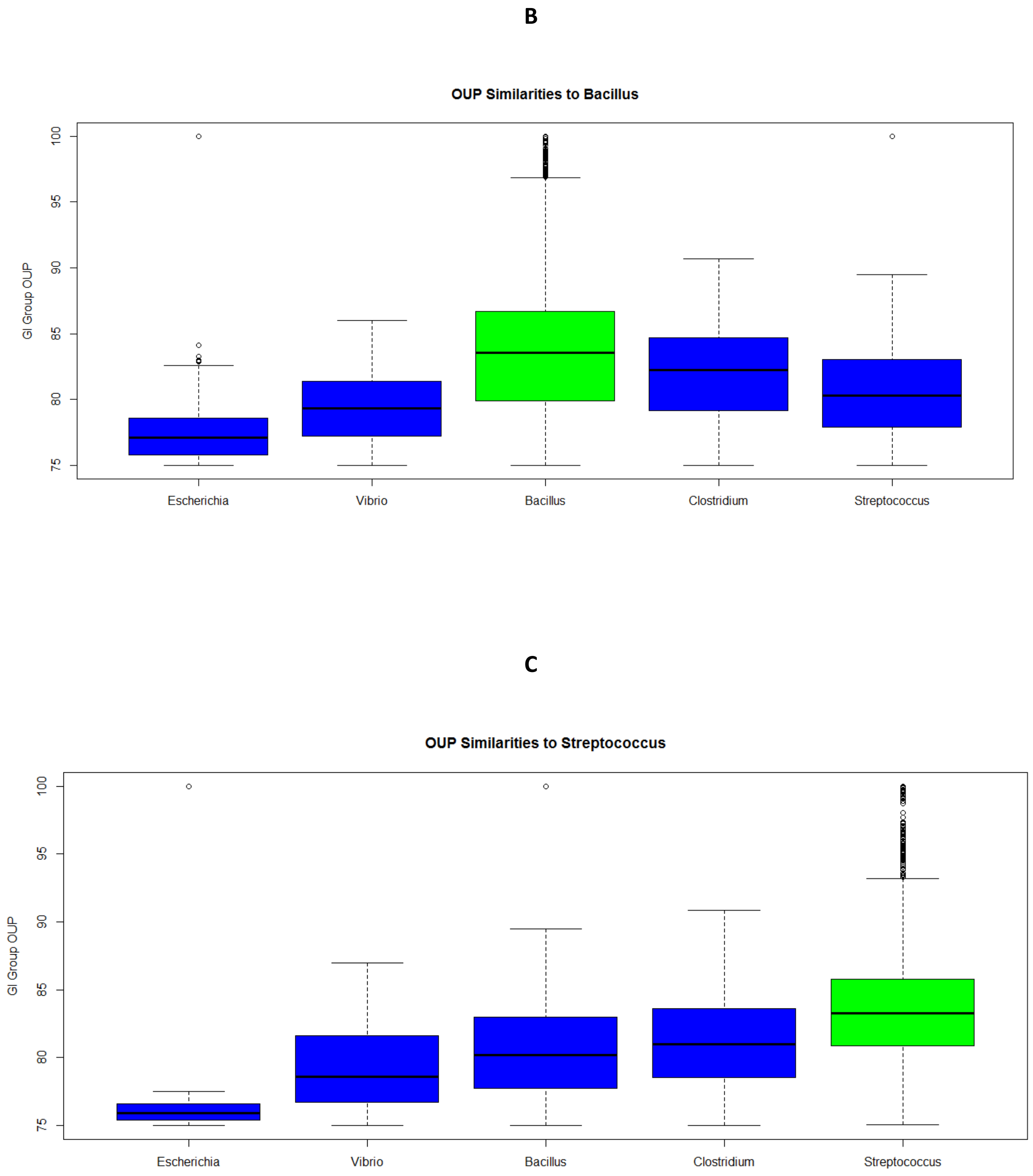
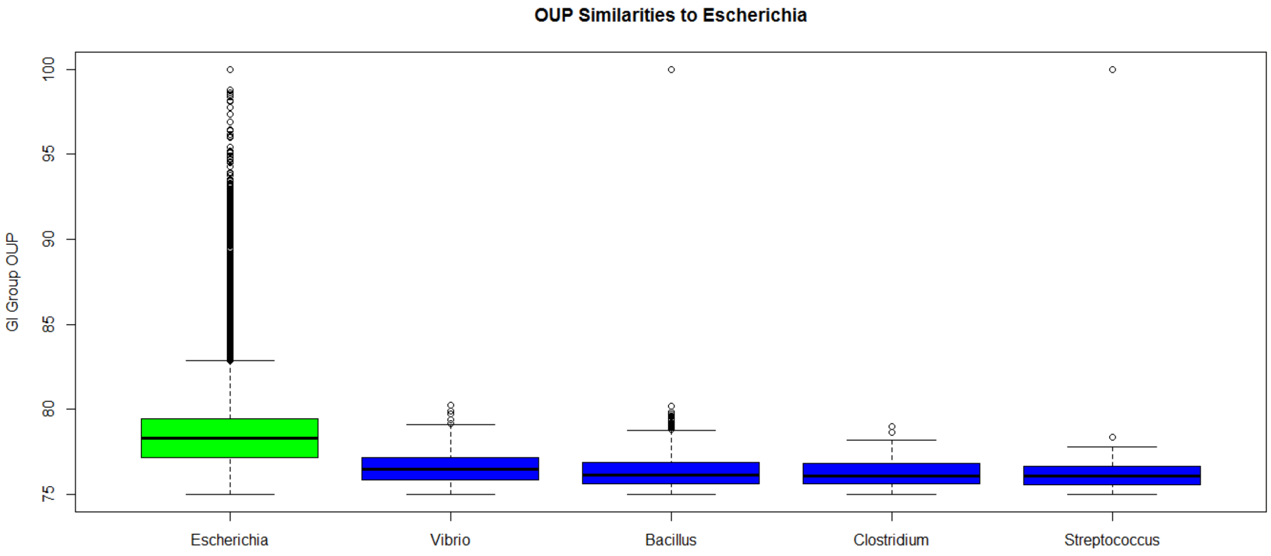
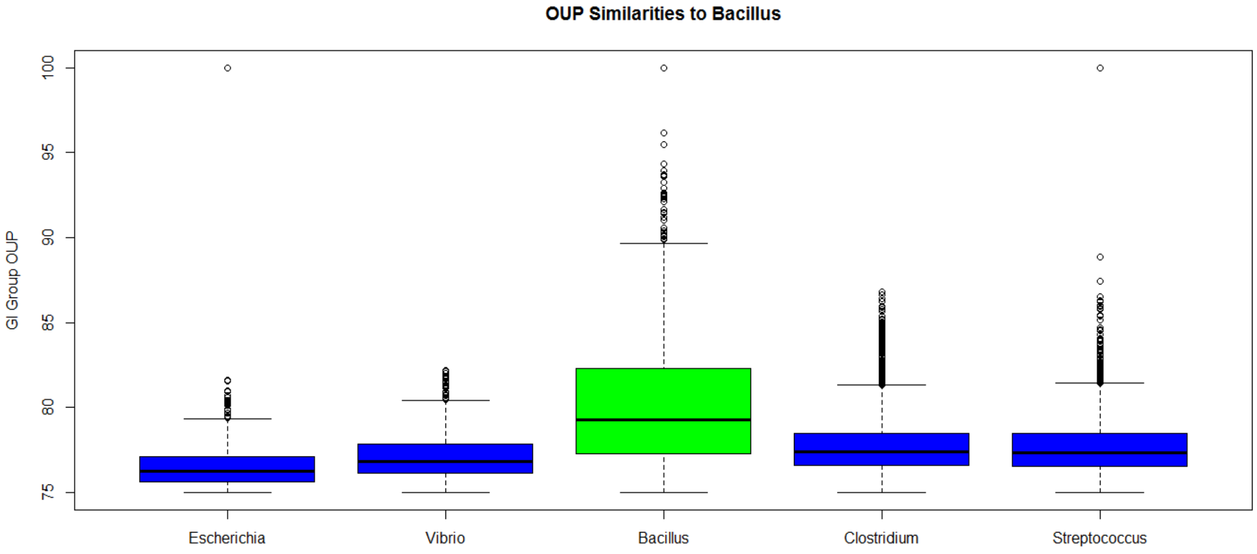


Figure 2. 9 Maximal single-linkage comparison of intra-generic and extra-generic OUP similarities with reference genera set for A) Escherichia; B) Bacillus; C) Streptococcus.

A


OUP Similarities to Escherichia
GI Group OUP
100
95
90
85
80
75
Escherichia
Vibrio
Bacillus
Clostridium
Streptococcus

B


OUP Similarities to Bacillus
GI Group OUP
100
95
90
85
80
75
Escherichia
Vibrio
Bacillus
Clostridium
Streptococcus

C

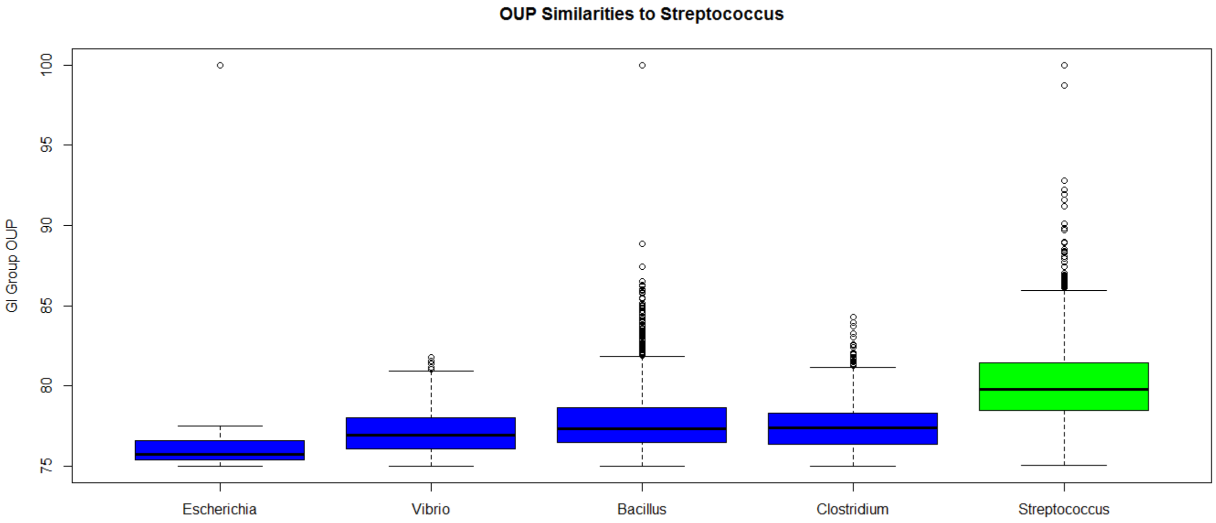


Figure 2. 11 Centroid-linkage comparison of intra-generic and extra-generic OUP similarities with reference genera set for A) Escherichia; B) Bacillus; C) Streptococcus.

It was seen that the average group linkage produced more condensed distributions compared against the maximum group linkage, which may hinder appropriate identification of ontological relations between groups of GIs found in different organisms. These results demonstrated also that exchange of genetic materials between organisms of the same genus occurs more frequently than between genera. This conclusion is probably true and for higher taxonomic levels. It is seen in ***Figures 2.10*** and ***2.11*** that Firmicutes organisms (*Bacillus*, *Clostridium* and *Streptococcus*) share a higher level of similarity than those of the Enterobacterial species of *Escherichia* and *Vibrio*.

2.3 Discussion

In order to recognize unique HGT events from a perspective of GI flux relations, GIs detected by compositional approaches has to be evaluated in a unified scheme. Fragments of GIs are often scattered across the hosting genome due subsequent evolutionary events, i.e. fragmentation, rearrangement and gene loss, acting upon them. It greatly complicates identification of donor-recipient and other ontological links between multiple fragments of GIs. Consequently, GIs from the Pre_GI database were analyzed and grouped according to sequence pattern similarities and congruent amelioration profiles. Different thresholds for both of these measures were evaluated to determine which cutoff values produced meaningful and sensible GI clusters. The grouping produced 5577 groups from 24,858 GIs among the considered replicons, with up to 10 groups of GIs separations in a minority of the organisms. From this perspective, the transferred regions were significantly larger and we could evaluate the entire summary of the genes obtained by a recipient organism instead of comparison of many disjointed regions. Integration of all available GIs into a single group (only one GI group per organism) was observed in 1098 bacterial genomes. These observations are indicative of unique sources and reduced promiscuity for these organisms. This is probably due to an inability to overcome taxonomic barriers or confinement to a specific ecology. The sequence pattern similarity scores produced among groups of GIs in different organisms were used to check the applicability of GI group similarity scoring based on the single and centroid linkage approaches. We found that the use of either of these approaches would influence the structure of the resulting MCL cluster outputs. As per our expectation, the maximal single-linkage approach allowed for more accurate clustering performance. The reconstruction of complete GI entities from fragmented GI regions showed an over representation in the number of HGT from the Pre_GI database and produced a framework of unified mobile genetic elements necessary to predict fluxes of GIs through bacterial species. This framework may be used to simulate and visualize fluxes of GIs on a software platform which will serve as the primary tool for the analysis of GI flux relations.

Chapter 3. Development of the web based platform for the visualization of fluxes of GIs

3.1 System platform

Using the constructed groups of GIs and their corresponding group weights developed for this project in section *2.1.1* and *2.1.2*, we could simulate HGT flows through the implementation of Markov Clustering (MCL). The design of the system was encapsulated in a PHP web platform, incorporating connections to MySQL and Cytoscape server instances. The MySQL database served as the primary source of GI group information. ***Figure 3.1*** shows the process diagram of the system execution procedure. The user initiates interaction with the system by providing organisms of interest to the system. Moreover, the user may opt to select additional inputs by specifying which gene description keywords they intend to study. The system will extract the necessary information from the MySQL database for each of the user's selections. This information will be used to start graph construction in Cytoscape. Cytoscape allows for remote procedure calls (RPCs) to mimic manual operation which was implemented using the *Python 2.7* scripting language. Each of the GI groups is then implemented as nodes and the GI group weights as edges. Once the graph has been constructed, the MCL algorithm will be applied to the structure. The resulting visualization is then attributed with custom settings for a variety of descriptors. Node colors and sizes are assigned according to the various species and total sequence lengths, respectively. Additionally, confirmed BLAST matches between any two GI groups are indexed and depicted by a green background. The results from the Python scripts may be retrieved via PHP handlers and displayed on a PHP webpage.

The address for the Flux Visualizer system is http://flux.bi.up.ac.za*.* The landing page of the website provides information on the project with descriptions and usage instructions. Additionally, two links from the landing page are shown to enable navigation to the separate interfaces. The first interface was designed to explore the grouping of GIs per organism, and the second interface to visualize donor-recipient relationships based on the user's selections.

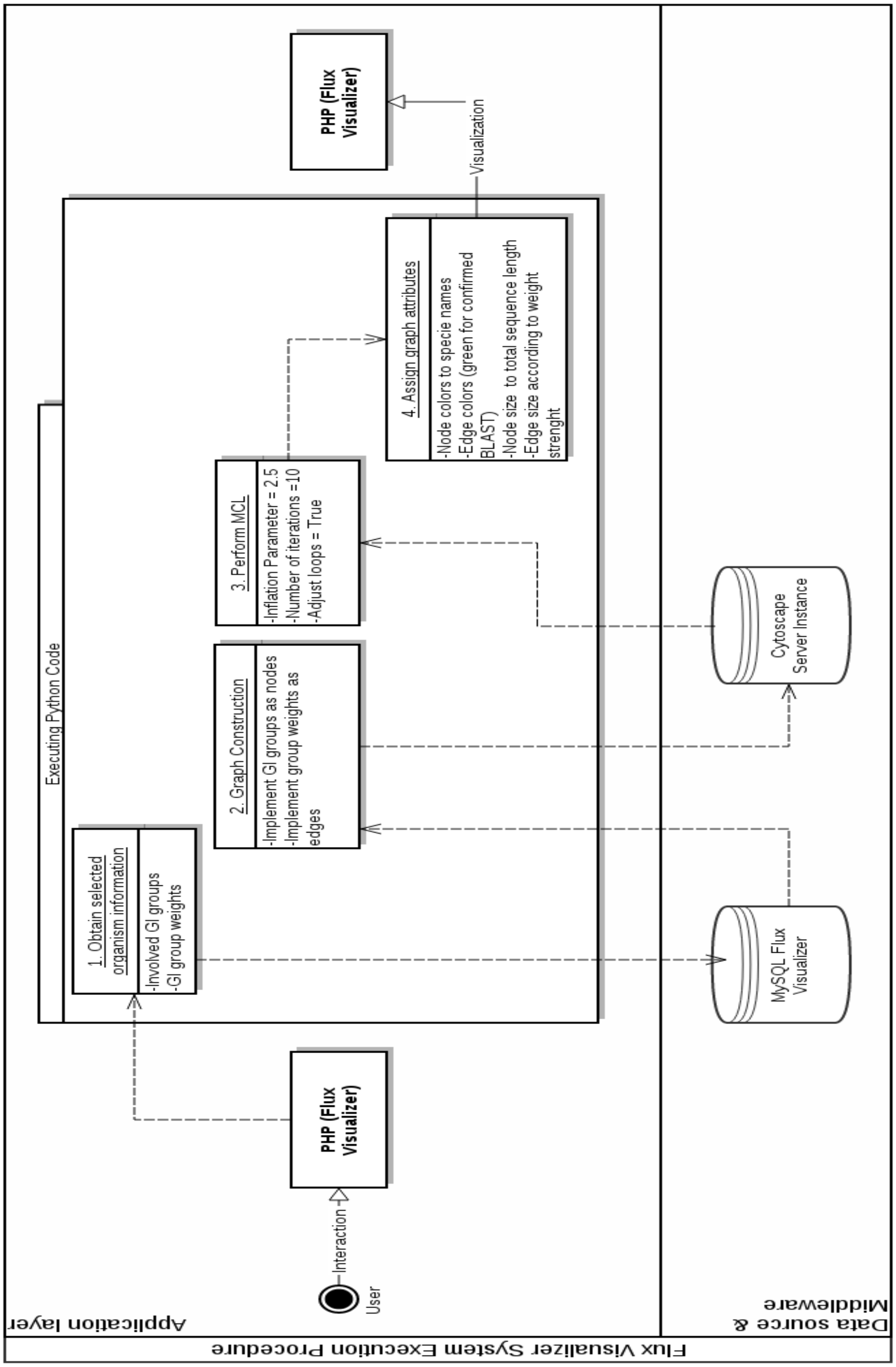
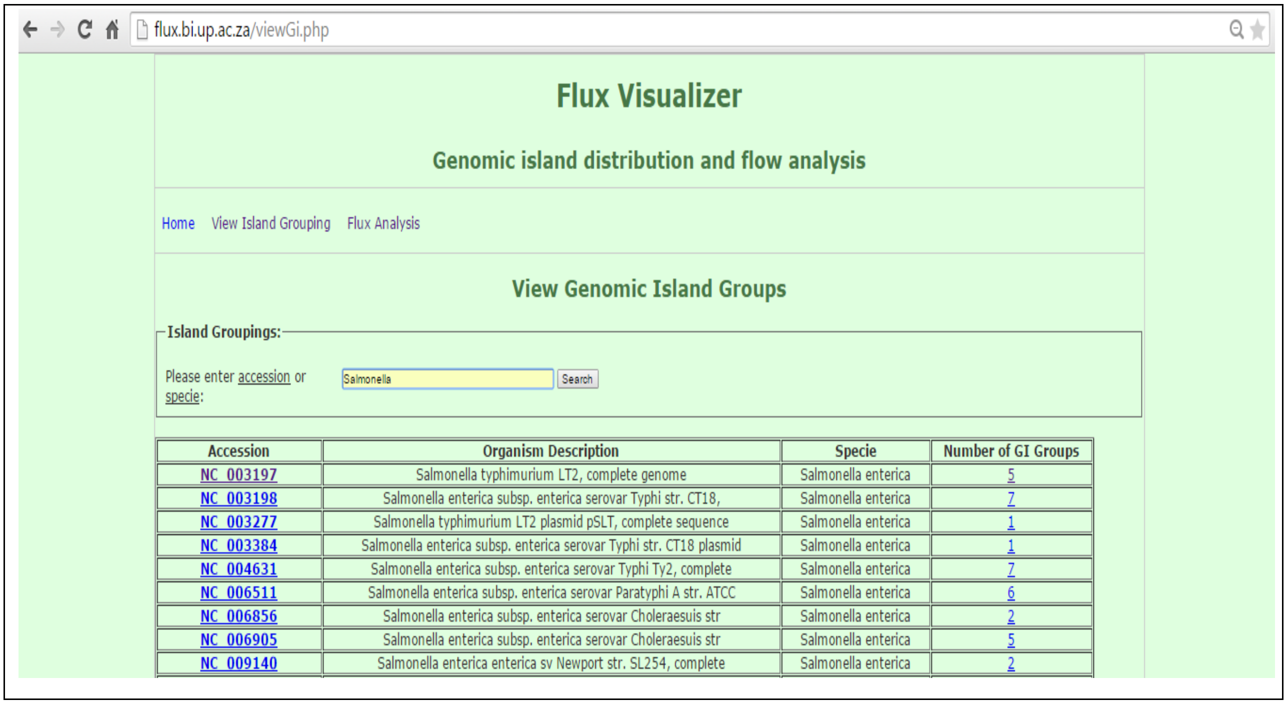

Flux Visualizer System Execution Procedure
Application layer
Data source & Middleware
User
Interaction
PHP (Flux Visualizer)
Executing Python Code
1. Obtain selected organism information
-Involved GI groups
-GI group weights
2. Graph Construction
-Implement GI groups as nodes
-Implement group weights as edges
3. Perform MCL
-Inflation Parameter = 2.5
-Number of iterations =10
-Adjust loops = True
4. Assign graph attributes
-Node colors to specie names
-Edge colors (green for confirmed BLAST)
-Node size to total sequence length
-Edge size according to weight strenght
Visualization
PHP (Flux Visualizer)
MySQL Flux Visualizer
Cytoscape Server Instance

Figure 3. 1 Process flow diagram of the Flux Visualizer core operation.

3.2 System interface

The Flux Visualizer system provides several functionalities to the users of the site. The first function (View GI groupings navigation link) allows the user to investigate the grouping and functional mapping of GIs (as per Chapter 4) by entering and searching either species names for multiple result matches or Genbank accession numbers for specific organism retrievals. The results are formatted into a table with navigational links ascribed to each individual organism presented. ***Figure 3.2*** shows the usage of this interface with the *Salmonella* species name as an example reference. ***Figure 3.3*** shows the grouping of GIs with the sequence starting positions of each GI contained within the group. In turn each GI start position may be used as a navigation link which integrates with the Pre_GI website (http://pregi.bi.up.ac.za/) to display a visual representation of the GI locations and Genbank coding region annotations.



Flux Visualizer

Genomic island distribution and flow analysis

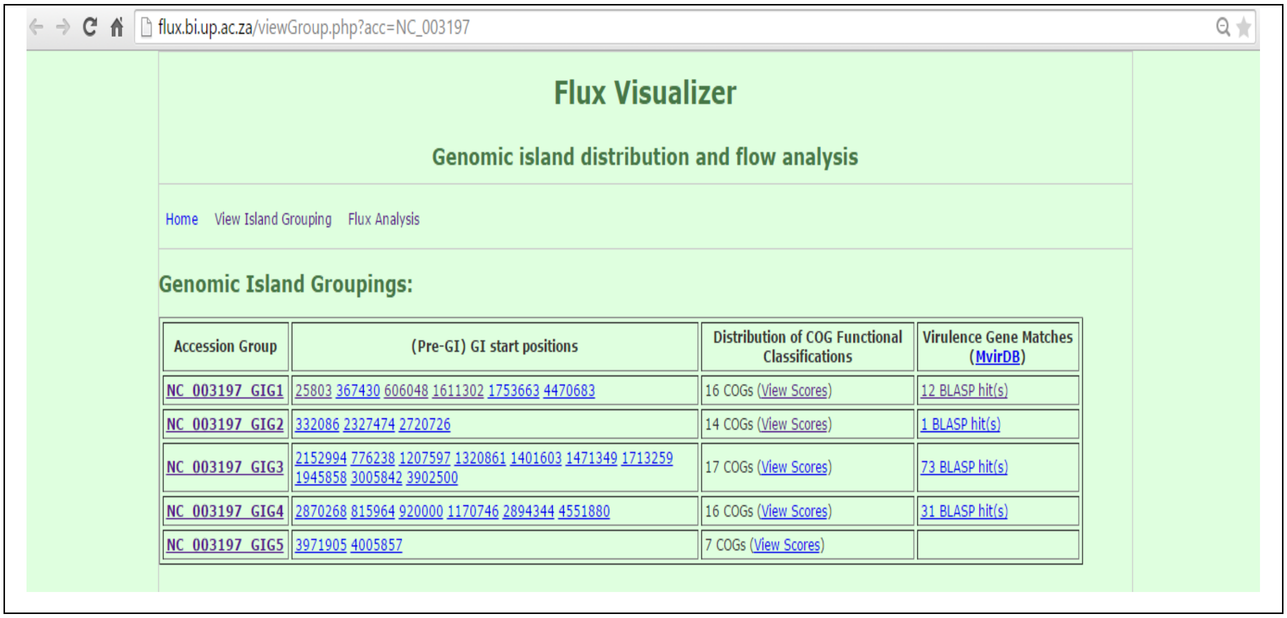
View Genomic Island Groups

Island Groupings:

Please enter accession or specie: Salmonella

Accession	Organism Description	Specie	Number of GI Groups
NC_003197	Salmonella typhimurium LT2, complete genome	Salmonella enterica	5
NC_003198	Salmonella enterica subsp. enterica serovar Typhi str. CT18,	Salmonella enterica	7
NC_003277	Salmonella typhimurium LT2 plasmid pSLT, complete sequence	Salmonella enterica	1
NC_003384	Salmonella enterica subsp. enterica serovar Typhi str. CT18 plasmid	Salmonella enterica	1
NC_004631	Salmonella enterica subsp. enterica serovar Typhi Ty2, complete	Salmonella enterica	7
NC_006511	Salmonella enterica subsp. enterica serovar Paratyphi A str. ATCC	Salmonella enterica	6
NC_006856	Salmonella enterica subsp. enterica serovar Choleraesuis str	Salmonella enterica	2
NC_006905	Salmonella enterica subsp. enterica serovar Choleraesuis str	Salmonella enterica	5
NC_009140	Salmonella enterica enterica sv Newport str. SL254, complete	Salmonella enterica	2

Figure 3. 2 View GI Grouping Interface on the Flux Visualizer page. The search functionality is used to display relevant entries for the Salmonella species keyword.



Flux Visualizer

Genomic island distribution and flow analysis

Genomic Island Groupings:

Accession Group	(Pre-GI) GI start positions	Distribution of COG Functional Classifications	Virulence Gene Matches (MvirDB)
NC_003197_GIG1	25803 367430 606048 1611302 1753663 4470683	16 COGs (View Scores)	12 BLASP hit(s)
NC_003197_GIG2	332086 2327474 2720726	14 COGs (View Scores)	1 BLASP hit(s)
NC_003197_GIG3	2152994 776238 1207597 1320861 1401603 1471349 1713259 1945858 3005842 3902500	17 COGs (View Scores)	73 BLASP hit(s)
NC_003197_GIG4	2870268 815964 920000 1170746 2894344 4551880	16 COGs (View Scores)	31 BLASP hit(s)
NC_003197_GIG5	3971905 4005857	7 COGs (View Scores)	

Figure 3. 3 The GI groupings of Salmonella enterica subsp. enterica serovar Typhimurium str. LT2 (NC_003197). The starting positions of the GIs in each group as well as the organism accession group are hyperlinked to the Pre_GI website for additional information.

The second function (Flux Analysis navigation link) depicts the core functionality of the system. Users are here able to search for organisms of interest by inputs of Genbank accessions or provided suggestions from related COG functional classes, and input gene description keywords. The user's selections are then used to create a visual interpretation of the results from the application of the MCL algorithm. ***Figure 3.4*** and ***3.5*** shows the usage of the Flux Analysis page. The user is able to facilitate one of two search functions, either through the use of suggested links or custom links.

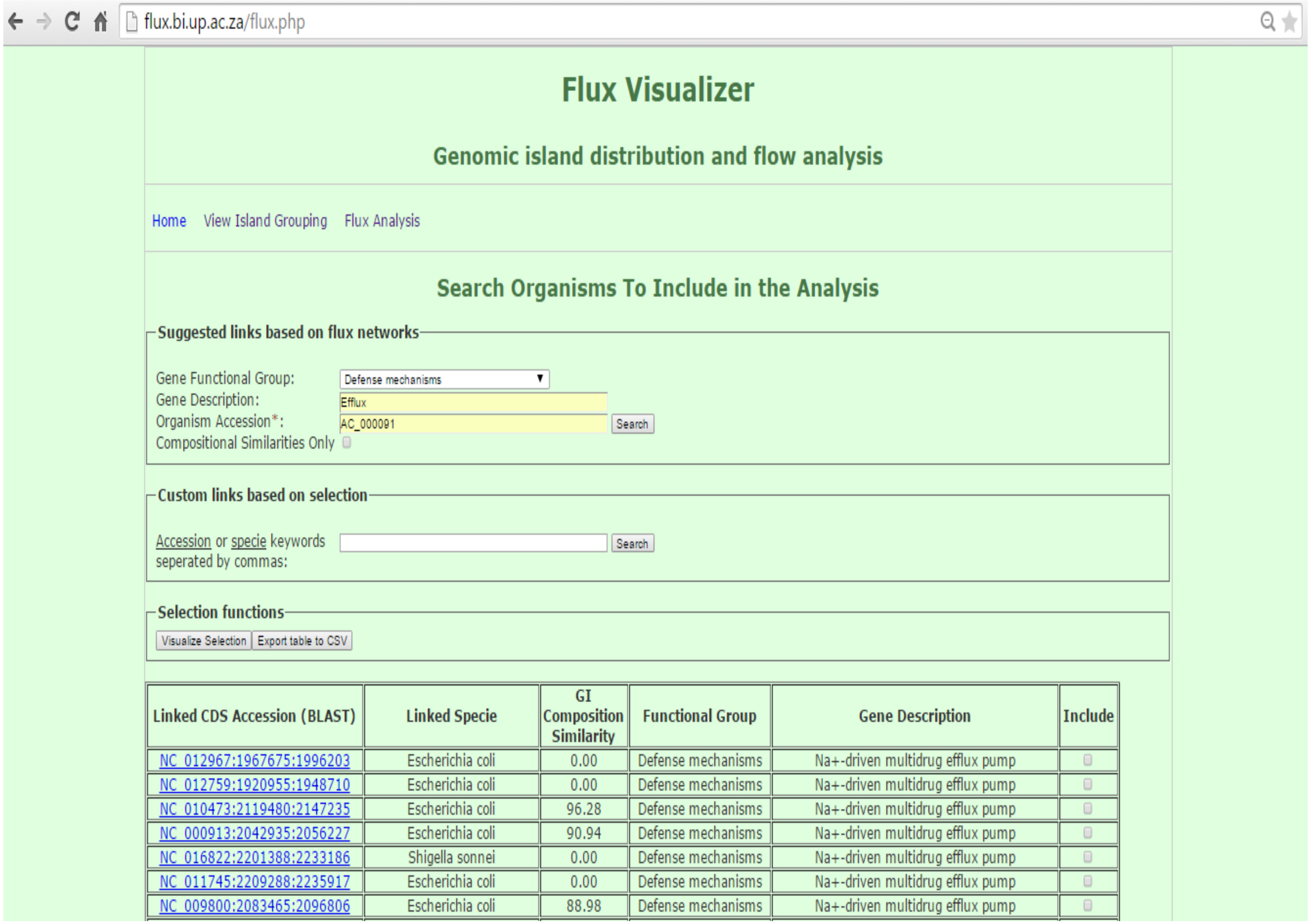
Suggested links integrate with the constructed gene networks and additional COG functional class filters (as described in Chapter 4) to enable generation of flux clusters based on organisms sharing specific gene or functional utilities. The available selections are displayed in the table once the search is started. Moreover, the selections from suggested links may be further filtered to show only genes which form part of GIs identified as sharing OUP similarities by selecting the checkbox "Compositional Similarities Only".

As per ***Figure 3.4***, the table results were extracted for the example organism *Escherichia coli str. K-12 substr. W3110* accession AC_000091, with the "Efflux" gene description keyword. Here the selection results provided information regarding the organisms found to share BLASTP hits to the genes within AC_000091 and matching gene descriptions of "Efflux". Furthermore, the selection results could also be filtered to show only organisms that have been identified as sharing OUP similarity above 75 on a GI level.

Research studies are often restricted to certain organism involvements which may require a small number of inputs. Custom links provide the ability to enter consecutive accession numbers and species names to show only relevant resulting retrievals. Accession numbers and species names may be combined into a single text input line separated by commas. As per ***Figure 3.5***, the example input specified is the following bacteria and plasmids references:

NC_002655	*Escherichia coli O157:H7 EDL933*
NC_017319	*Shigella flexneri 2002017 plasmid pSFxv_1*
NC_004851	*Shigella flexneri 2a str. 301 plasmid pCP301*
NC_016445	*Vibrio cholerae O1 str. 2010EL-1786 chromosome 1*
NC_011184	*Vibrio fischeri MJ11 chromosome I*
NC_014966	*Vibrio vulnificus MO6-24/O chromosome II*

The user can subsequently use the "Include" checkboxes to specify which of the organisms are required to be included for MCL analysis and visualization. Two selection options are provided to the user once selection results have been retrieved. The first option (Visualize Selection) will perform MCL clustering on the organisms included, and divert to a new page with the visualization of flux clusters. The second option enables the user to download the complete table for personal analysis in a CSV file format. ***Figure 3.6*** shows a demonstration of the resulting clustering of the reference input specified as per ***Figure 3.5***



Flux Visualizer

Genomic island distribution and flow analysis

Search Organisms To Include in the Analysis

Suggested links based on flux networks

Gene Functional Group: Defense mechanisms ▼
Gene Description: Efflux
Organism Accession*: AC_000091 Search
Compositional Similarities Only □

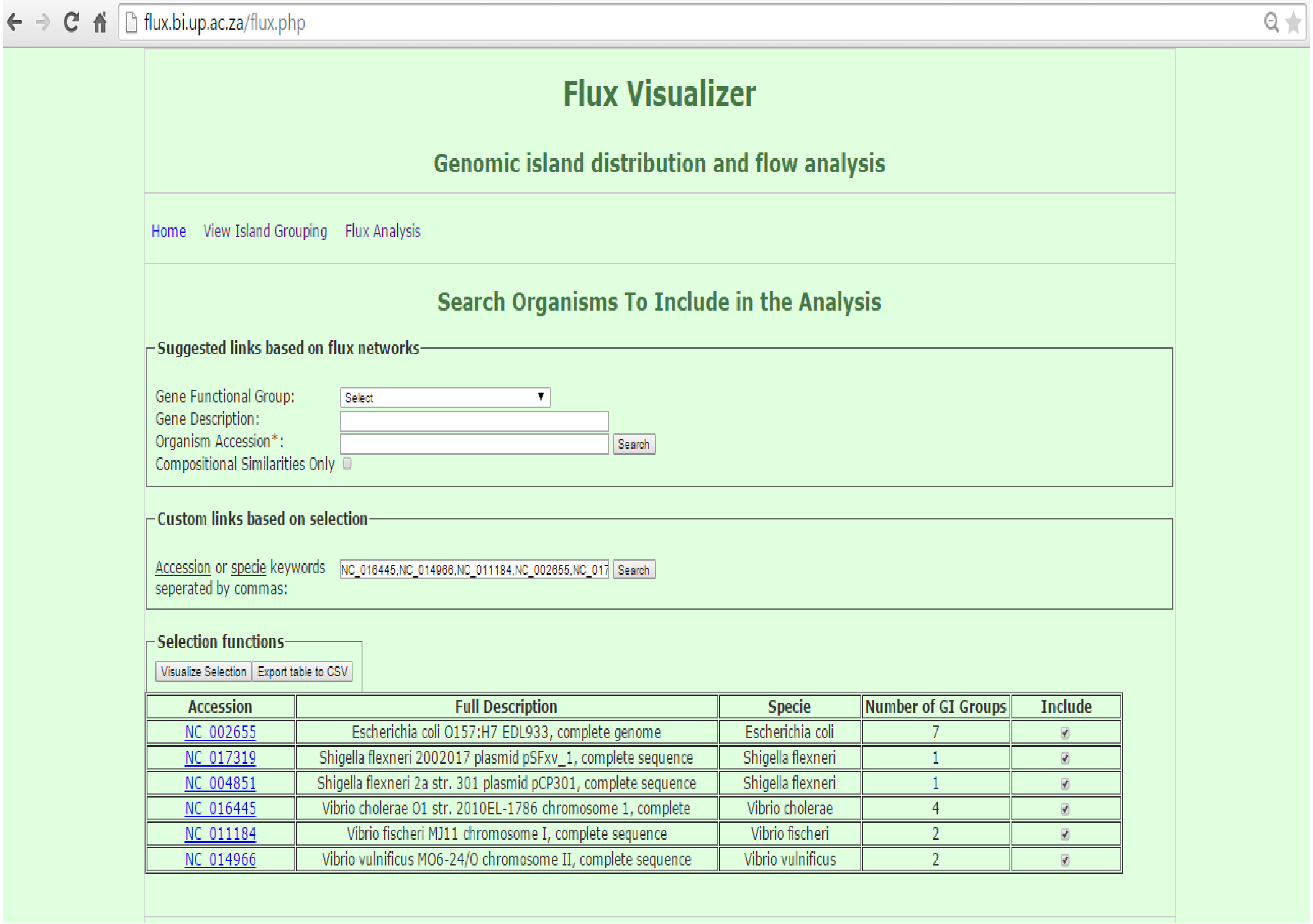
Custom links based on selection

Accession or specie keywords seperated by commas: Search

Selection functions

Linked CDS Accession (BLAST)	Linked Specie	GI Composition Similarity	Functional Group	Gene Description	Include
NC_012967:1967675:1996203	Escherichia coli	0.00	Defense mechanisms	Na+-driven multidrug efflux pump	□
NC_012759:1920955:1948710	Escherichia coli	0.00	Defense mechanisms	Na+-driven multidrug efflux pump	□
NC_010473:2119480:2147235	Escherichia coli	96.28	Defense mechanisms	Na+-driven multidrug efflux pump	□
NC_000913:2042935:2056227	Escherichia coli	90.94	Defense mechanisms	Na+-driven multidrug efflux pump	□
NC_016822:2201388:2233186	Shigella sonnei	0.00	Defense mechanisms	Na+-driven multidrug efflux pump	□
NC_011745:2209288:2235917	Escherichia coli	0.00	Defense mechanisms	Na+-driven multidrug efflux pump	□
NC_009800:2083465:2096806	Escherichia coli	88.98	Defense mechanisms	Na+-driven multidrug efflux pump	□

Figure 3. 4 Example usage of the Flux Analysis navigation link with suggested selections using Escherichia coli str. K-12 substr. W3110 accession AC_000091 and "Efflux" gene keyword as reference.



Flux Visualizer

Genomic island distribution and flow analysis

Search Organisms To Include in the Analysis

Suggested links based on flux networks

Gene Functional Group: Select ▾
Gene Description:
Organism Accession*:
Compositional Similarities Only □

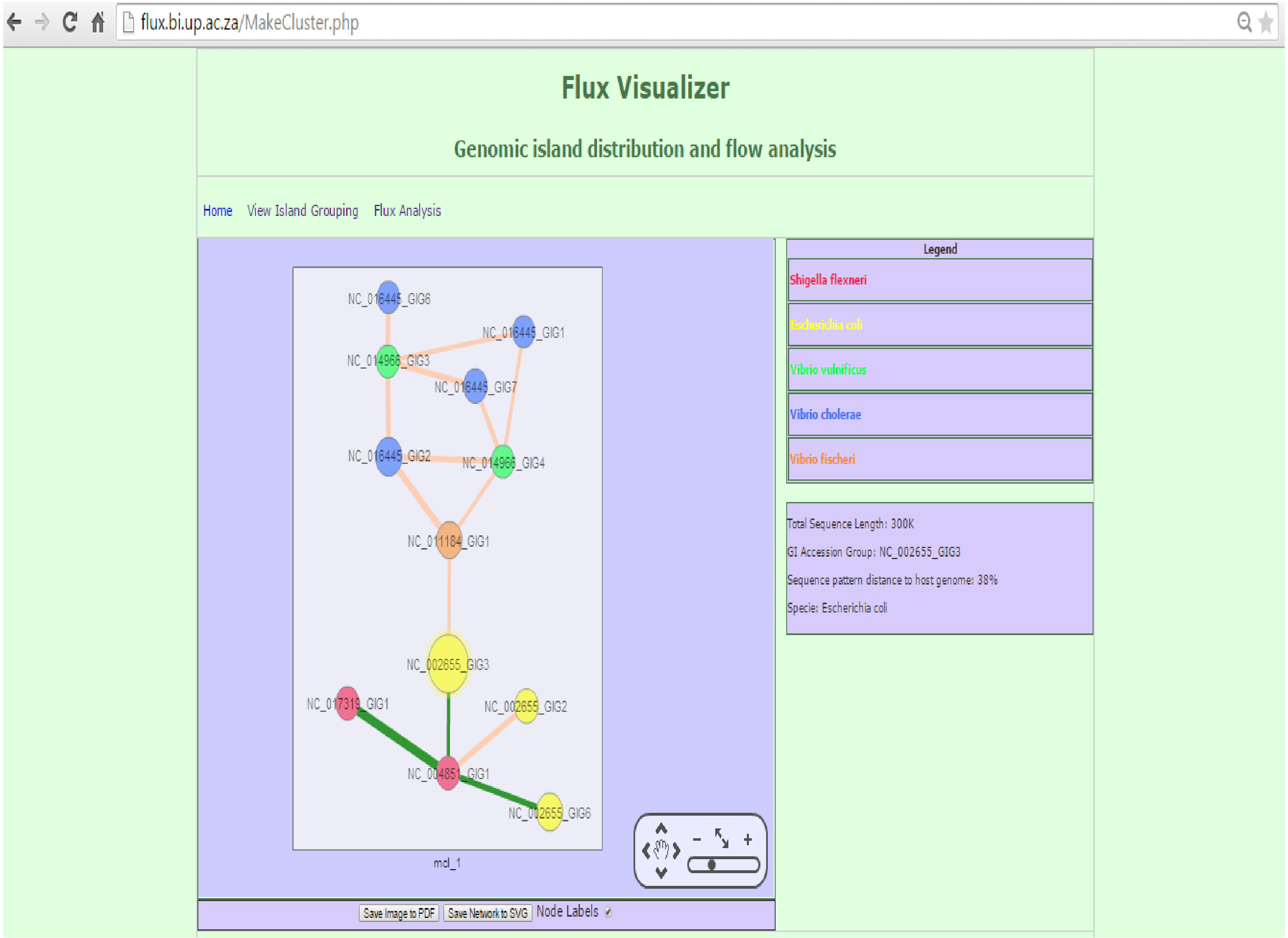
Custom links based on selection

Accession or specie keywords seperated by commas: NC_016445,NC_014966,NC_011184,NC_002655,NC_017

Selection functions

Accession	Full Description	Specie	Number of GI Groups	Include
NC_002655	Escherichia coli O157:H7 EDL933, complete genome	Escherichia coli	7	☑
NC_017319	Shigella flexneri 2002017 plasmid pSFxv_1, complete sequence	Shigella flexneri	1	☑
NC_004851	Shigella flexneri 2a str. 301 plasmid pCP301, complete sequence	Shigella flexneri	1	☑
NC_016445	Vibrio cholerae O1 str. 2010EL-1786 chromosome 1, complete	Vibrio cholerae	4	☑
NC_011184	Vibrio fischeri MJ11 chromosome I, complete sequence	Vibrio fischeri	2	☑
NC_014966	Vibrio vulnificus MO6-24/O chromosome II, complete sequence	Vibrio vulnificus	2	☑

Figure 3. 5 Example usage of the Flux Analysis navigation link with custom selections based on accession numbers.



flux.bi.up.ac.za/MakeCluster.php

Flux Visualizer

Genomic island distribution and flow analysis

Home View Island Grouping Flux Analysis

Legend

Shigella flexneri

Escherichia coli

Vibrio vulnificus

Vibrio cholerae

Vibrio fischeri

Total Sequence Length: 300K

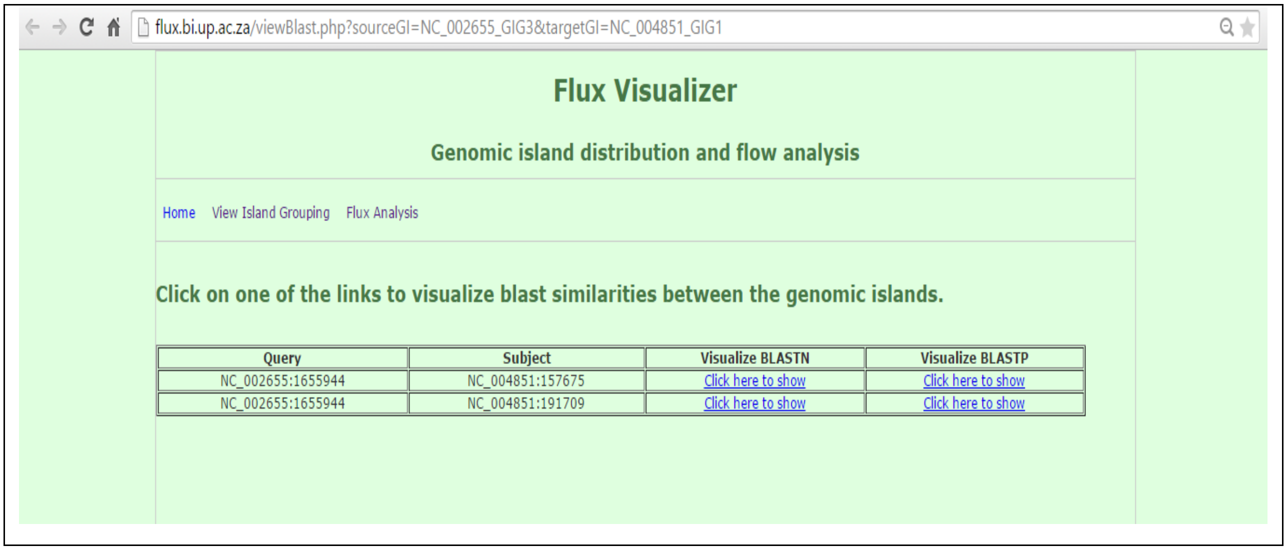
GI Accession Group: NC_002655_GIG3

Sequence pattern distance to host genome: 38%

Specie: Escherichia coli

Figure 3. 6 Example visualization of flux links among the organisms selected in Figure 3.5.

The resulting graph in ***Figure 3.6*** shows one MCL cluster with the selected organisms encapsulated within. This cluster represents the fluxes of GIs between the organisms with each GI group represented as a node and each HGT event presented as an edge. The nodes are color coded to the species legend scheme for the identification of different organisms. Tangerine colored links presents high GI group weight scores between the specified GI groups while green links specify confirmed BLAST similarities among at least one gene contained in each group. While the size of the node is directly related to the size of the GI group, the thickness of the edge is related to the strength of the OUP score. Each node and edge is clickable and reveals additional information regarding the item clicked in the panel below the legend. In ***Figure 3.6*** the node for *Escherichia coli O157:H7 EDL933,* GI group 3 is clicked for reference. The right hand side panel reveals that the GI group is approximately 300,000 base pairs in total sequence length, and has an average pattern distance-to-host values of 38% across all of its respective GIs. Moreover, each green BLAST link can be selected to visualize the gene BLAST matches through the Pre_GI database functionality, as shown between the *Shigella flexneri 2a str. 301 plasmid pCP301* (NC_004851) and *Escherichia coli O157:H7 EDL933* (NC_002655) in ***Figure 3.7*** and ***Figure 3.8***.



flux.bi.up.ac.za/viewBlast.php?sourceGI=NC_002655_GIG3&targetGI=NC_004851_GIG1

Flux Visualizer

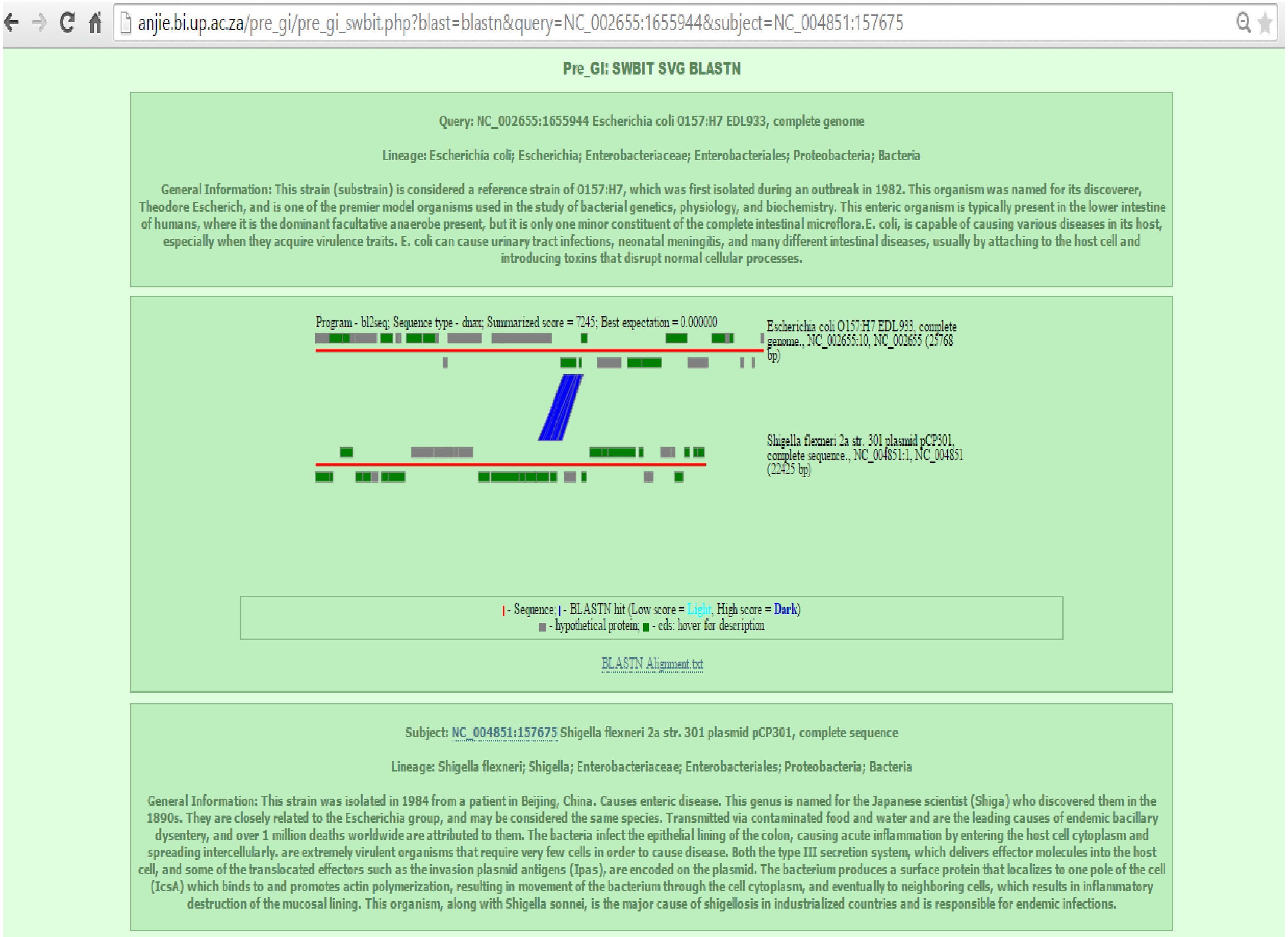
Genomic island distribution and flow analysis

Home
View Island Grouping
Flux Analysis

Click on one of the links to visualize blast similarities between the genomic islands.

Query	Subject	Visualize BLASTN	Visualize BLASTP
NC_002655:1655944	NC_004851:157675	Click here to show	Click here to show
NC_002655:1655944	NC_004851:191709	Click here to show	Click here to show

Figure 3. 7 BLAST visualization options for GIs with common genes for Shigella flexneri 2a str. 301 plasmid pCP301 (NC_004851) and Escherichia coli O157:H7 EDL933 (NC_002655).



anjie.bi.up.ac.za/pre_gi/pre_gi_swbit.php?blast=blastn&query=NC_002655:1655944&subject=NC_004851:157675

Pre_GI: SWBIT SVG BLASTN

Query: NC_002655:1655944 Escherichia coli O157:H7 EDL933, complete genome

Lineage: Escherichia coli; Escherichia; Enterobacteriaceae; Enterobacteriales; Proteobacteria; Bacteria

General Information: This strain (substrain) is considered a reference strain of O157:H7, which was first isolated during an outbreak in 1982. This organism was named for its discoverer, Theodore Escherich, and is one of the premier model organisms used in the study of bacterial genetics, physiology, and biochemistry. This enteric organism is typically present in the lower intestine of humans, where it is the dominant facultative anaerobe present, but it is only one minor constituent of the complete intestinal microflora.E. coli, is capable of causing various diseases in its host, especially when they acquire virulence traits. E. coli can cause urinary tract infections, neonatal meningitis, and many different intestinal diseases, usually by attaching to the host cell and introducing toxins that disrupt normal cellular processes.

Program - bl2seq; Sequence type - dnax; Summarized score = 7245; Best expectation = 0.000000

Escherichia coli O157:H7 EDL933, complete genome., NC_002655:10, NC_002655 (25768 bp)

Shigella flexneri 2a str. 301 plasmid pCP301, complete sequence., NC_004851:1, NC_004851 (22425 bp)

| - Sequence; | - BLASTN hit (Low score = Light, High score = Dark)
■ - hypothetical protein; ■ - cds: hover for description

BLASTN Alignment.txt

Subject: NC_004851:157675 Shigella flexneri 2a str. 301 plasmid pCP301, complete sequence

Lineage: Shigella flexneri; Shigella; Enterobacteriaceae; Enterobacteriales; Proteobacteria; Bacteria

General Information: This strain was isolated in 1984 from a patient in Beijing, China. Causes enteric disease. This genus is named for the Japanese scientist (Shiga) who discovered them in the 1890s. They are closely related to the Escherichia group, and may be considered the same species. Transmitted via contaminated food and water and are the leading causes of endemic bacillary dysentery, and over 1 million deaths worldwide are attributed to them. The bacteria infect the epithelial lining of the colon, causing acute inflammation by entering the host cell cytoplasm and spreading intercellularly. are extremely virulent organisms that require very few cells in order to cause disease. Both the type III secretion system, which delivers effector molecules into the host cell, and some of the translocated effectors such as the invasion plasmid antigens (Ipas), are encoded on the plasmid. The bacterium produces a surface protein that localizes to one pole of the cell (IcsA) which binds to and promotes actin polymerization, resulting in movement of the bacterium through the cell cytoplasm, and eventually to neighboring cells, which results in inflammatory destruction of the mucosal lining. This organism, along with Shigella sonnei, is the major cause of shigellosis in industrialized countries and is responsible for endemic infections.

Figure 3. 8 Pre_GI functionality for visualizing BLASTN similarity matches between Shigella flexneri 2a str. 301 plasmid pCP301 (NC_004851) and Escherichia coli O157:H7 EDL933 (NC_002655).

3.3 Discussion

The Flux Visualizer system was developed for the analyses of grouped GIs to envisage HGT interactions from a relational perspective. The implementation was done through a pipeline process incorporating *Python* scripts to execute procedural calls to server instances of Cytoscape and MySQL. Groups of GIs are clustered together using MCL to identify which organisms show the most probable HGT interactions and how the mobile genes were exchanged between the selected organisms. The system provides multiple ways for the user to input organisms of interest. Suggested links may prove valuable in obtaining examples of HGT partners which have been identified as sharing some form of functional commonality. Moreover, the system also provides researchers with the ability to specify explicit organism involvements and to include only the necessary. The results of the clustering analysis are displayed as a graph, generated from the Cytoscape software package. The elements of the graph are further attributed with unique colors and sizes to provide clear distinctions between different species, total sequence lengths, GI group weight scores and confirmed BLAST hits. Furthermore, the system allows the user to explore the contents of each GI group from a functional perspective as well as provide the starting genetic position of each GI in a particular group. Here the GI starting positions integrates with the Pre_GI website (*http://pregi.bi.up.ac.za/*) through hyperlinks and the user may explore each GI in more detail there. The clustering of GI groups in this way serves to aid research in providing a visual illustration of the HGT relationships between the involved organisms. To further enable specific and interest driven selections for the users of the website, we will need to investigate associations of gene functions and functional types between groups of GIs. Through the functional attribution of GI coding regions we may attempt to separate organisms which show functional similarities and in this way provide suggestions of organisms to include in the users analysis.

Chapter 4. Functional categorization of genomic islands groups

4.1 Genetic functional attribution

4.1.1 COG classifications

The functional attribution of GI coding regions were implemented to investigate relations of shared gene contents among groups of GIs to ultimately enable the separation of GI groups based on genetic utility. We hypothesized that is was possible to cluster groups of GIs based on specific functional types of HGT events. In turn, the separation of GI groups in this way would assist in creating categories of HGT and allow for additional analyses based on specific types of functional transfers. We considered allocating genes to functional classes from the evaluation of Genbank annotation data. However, due to the lack of comprehensive descriptions and standardized cataloging, it was extremely challenging to formulate accurate results based on text matching alone. Consequently, the COG database (Tatusov, et al., 2000) was used to infer phylogenetic classifications for the assignment of biochemical functions and roles to coding regions. We performed PSI-BLASTs of the 392,622 coding regions against the COG database to obtain classification hits for each GI group. Because we were not looking for strict functional conservations but rather similar functions of homologous proteins, we used PSI-BLAST to obtain more sensitive results with regards to distant functional resemblances. In this case PSI-BLAST performed better than the standard BLASTP. A bit score ≥ 100 an e-value of 0.004 was used as a homology threshold to protein sequences in the COG database. This relaxed bit score was considered to ensure that the majority of CDS were assigned to one of the functionally known categories. The COG database consists of 25 categories of functional proteins including poorly characterized, unknown or predicted only function. All of the COG functional categories are listed in ***Table 4.1***. Each CDS in GIs was assigned to one or more of these categories, depending on mapping to COG protein descriptions with the highest bit scores over the threshold value obtained by PSI-BLAST.

Representative Letter	Functional Description
J	Translation, ribosomal structure and biogenesis
A	RNA processing and modification
K	Transcription
L	Replication, recombination and repair
B	Chromatin structure and dynamics
D	Cell cycle control, cell division, chromosome partitioning
Y	Nuclear structure
V	Defense mechanisms
T	Signal transduction mechanisms
M	Cell wall/membrane/envelope biogenesis
N	Cell motility
Z	Cytoskeleton
W	Extracellular structures
U	Intracellular trafficking, secretion, and vesicular transport
O	Posttranslational modification, protein turnover, chaperones
C	Energy production and conversion
G	Carbohydrate transport and metabolism
E	Amino acid transport and metabolism
F	Nucleotide transport and metabolism
H	Coenzyme transport and metabolism
I	Lipid transport and metabolism
P	Inorganic ion transport and metabolism
Q	Secondary metabolites biosynthesis, transport and catabolism
R	General function prediction only
S	Function unknown

Table 4. 1 COG database functional classifications with representative letters of functional categories.

The distribution of COG functional categories from the CDS in predicted GIs were analyzed to obtain a consolidated view of over- and under-saturated classes. ***Figure 4.1*** shows the frequencies of CDS in all 26,744 GIs from the Pre_GI database per COG category. We observed significantly underrepresented transfer frequencies of informational and housekeeping genes denoted in COG by categories A and B that was consistent with data from literature (Popa, et al., 2011). Contrary, to the genes belonging to categories "Amino acid transport", "Cell membrane biogenesis" and "Transcription/Translation" denoted in COG as E, M, J and K, where most frequently mobilized. However, a significant proportion of the genes in GIs were classified either as "Function Unknown" or showed no matches at all to the reference proteins in the COG database.

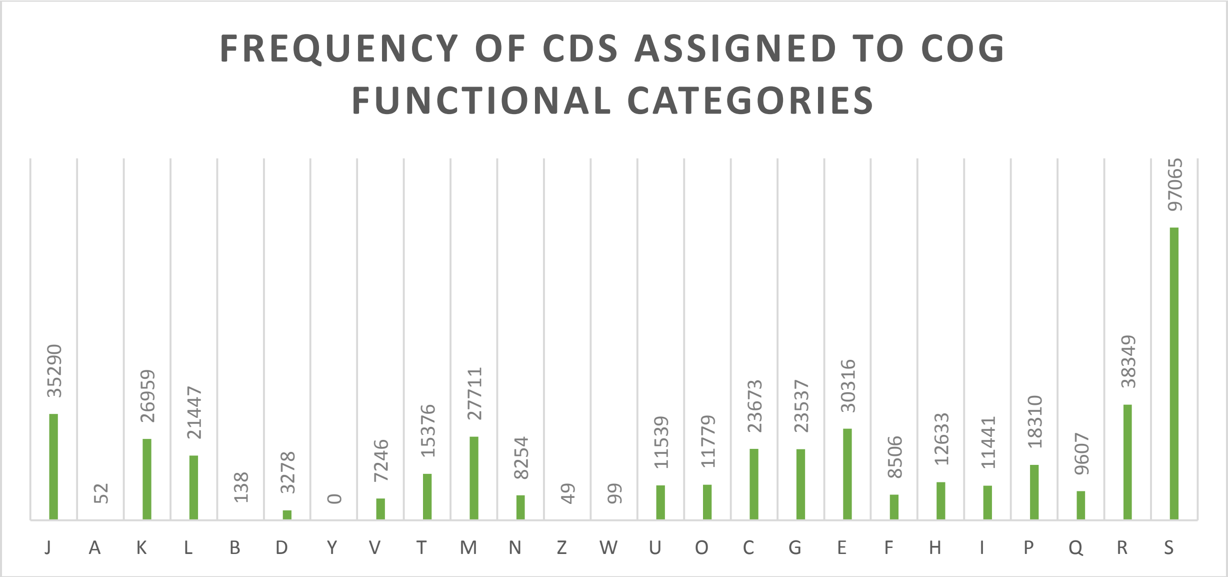


Figure 4. 1 Frequencies of GI's genes by COG functional category.

Since each CDS was allocated to one or more of the COG categories, groups of GIs could be evaluated from a perspective of functional gene products. The functional classifications for each of the individual GI groups (G1, G2, etc.) were used to evaluate the proportions of COG categories among the separate groups of GIs in all considered organisms. Since G1 was generally the most abundant group of GIs per genome, we anticipated that this group would show increased representations of COG categories compared to the other groups. ***Figure 4.2*** shows the distribution of COG category proportions among all the groups of GIs. The distributions of COG functional categories remained remarkably similar in all GI groups ordered by abundance from G1 to G10 (see discussion in Chapter 3) with a strong resemblance to the overall distribution of the CDS frequencies in ***Figure 4.1.*** It may therefore be concluded that the distribution of the genetic content of the most abundant (and/or fragmented) GIs in bacterial genomes is similar to the distribution in singletons with the exception of unknown and functionally undefined genes categories.

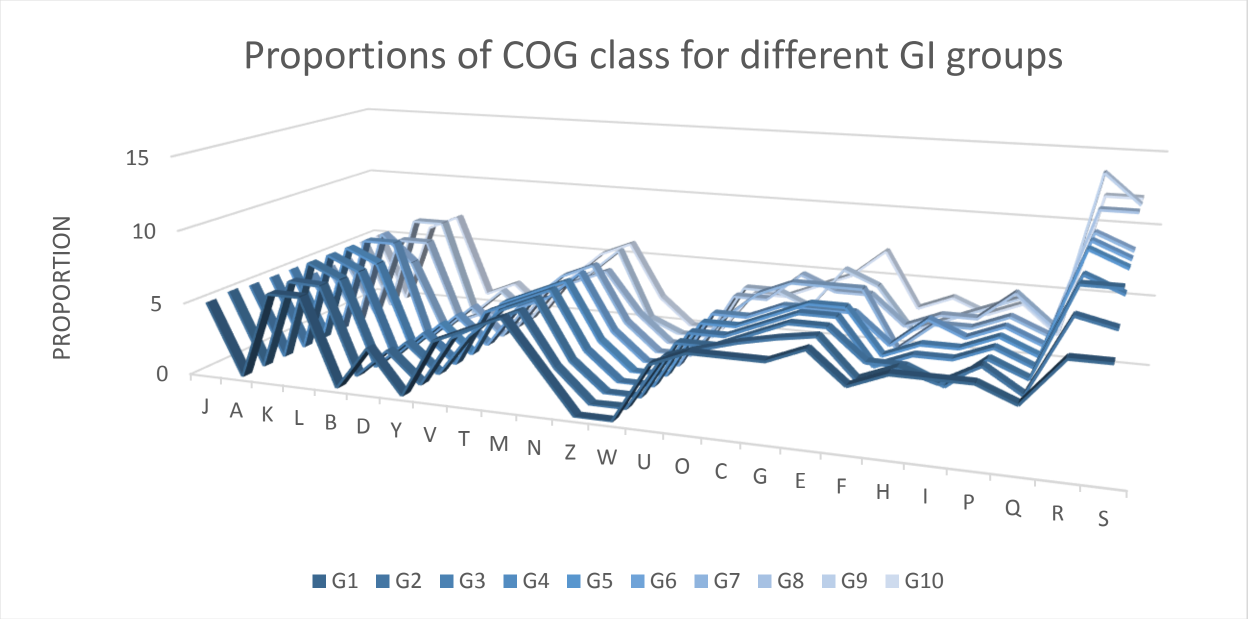
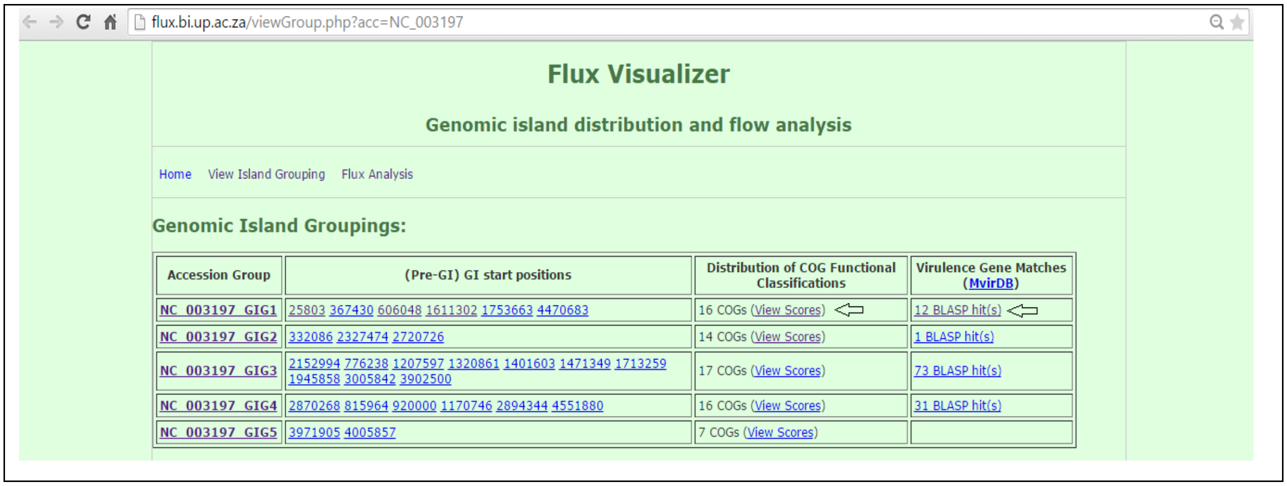


Figure 4. 2 Proportions of COG class distributions per GI group.

To allow the exploration of functional class assignments in GIs, the “View COGs” navigation system was implemented in the Flux Visualizer Web interface as shown in ***Figure 4.3*** (A) and (B). Assignments of genes to all COG functional categories and their proportions per groups of GIs may be viewed. The poorly characterized categories R (Prediction only) and S (Unknown functions) were excluded in the system.

A



flux.bi.up.ac.za/viewGroup.php?acc=NC_003197

Flux Visualizer

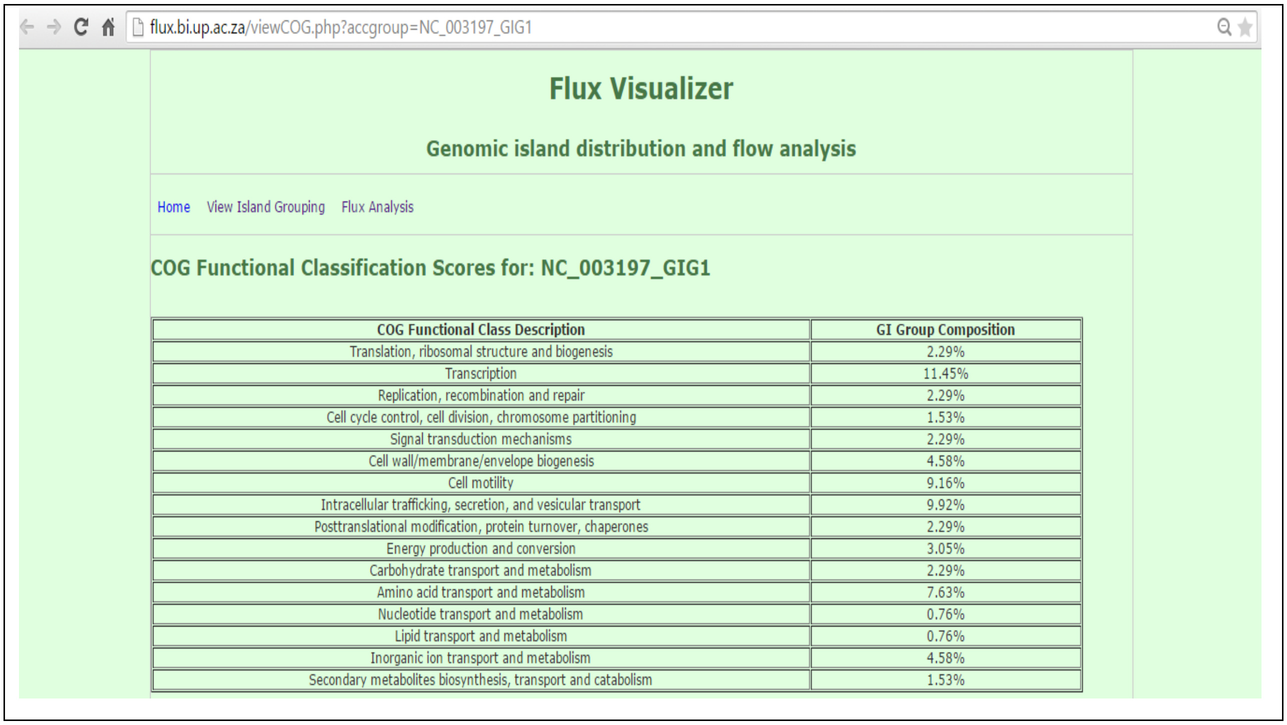
Genomic island distribution and flow analysis

Home View Island Grouping Flux Analysis

Genomic Island Groupings:

Accession Group	(Pre-GI) GI start positions	Distribution of COG Functional Classifications	Virulence Gene Matches (MvirDB)
NC_003197_GIG1	25803 367430 606048 1611302 1753663 4470683	16 COGs (View Scores) ⇦	12 BLASP hit(s) ⇦
NC_003197_GIG2	332086 2327474 2720726	14 COGs (View Scores)	1 BLASP hit(s)
NC_003197_GIG3	2152994 776238 1207597 1320861 1401603 1471349 1713259 1945858 3005842 3902500	17 COGs (View Scores)	73 BLASP hit(s)
NC_003197_GIG4	2870268 815964 920000 1170746 2894344 4551880	16 COGs (View Scores)	31 BLASP hit(s)
NC_003197_GIG5	3971905 4005857	7 COGs (View Scores)	

B



flux.bi.up.ac.za/viewCOG.php?accgroup=NC_003197_GIG1

Flux Visualizer

Genomic island distribution and flow analysis

Home View Island Grouping Flux Analysis

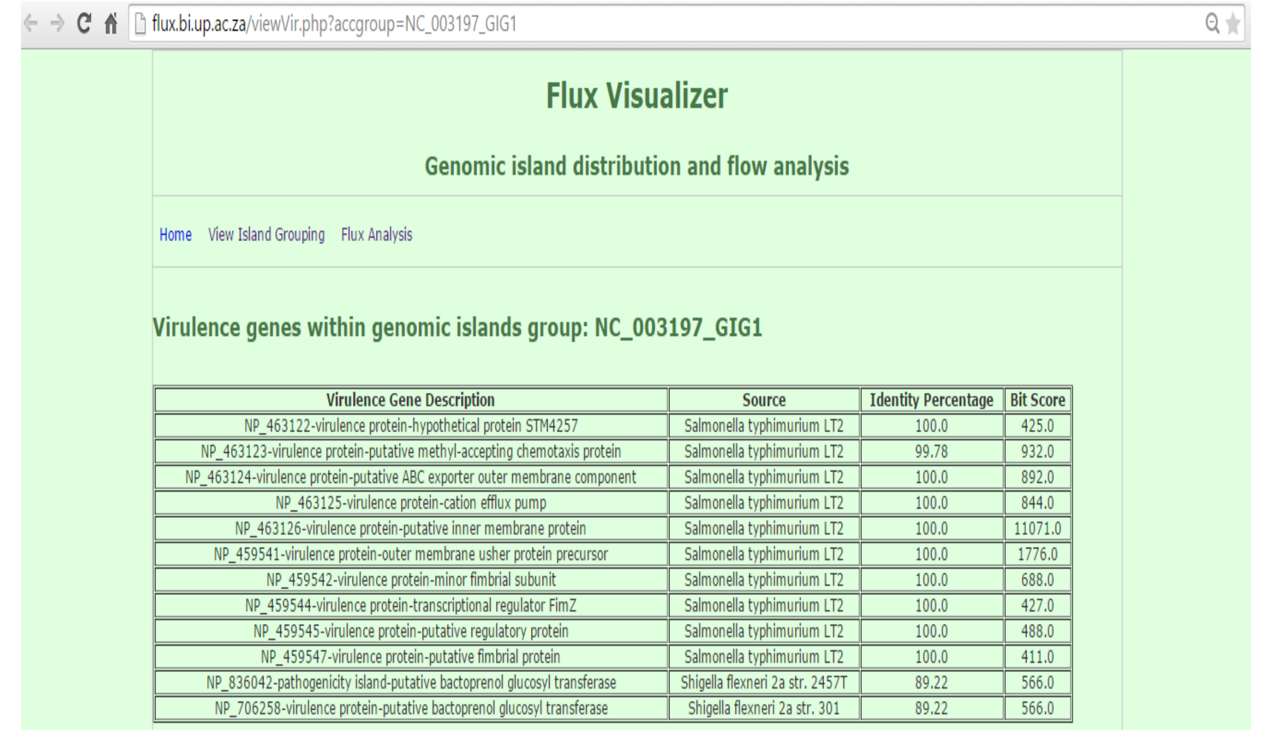
COG Functional Classification Scores for: NC_003197_GIG1

COG Functional Class Description	GI Group Composition
Translation, ribosomal structure and biogenesis	2.29%
Transcription	11.45%
Replication, recombination and repair	2.29%
Cell cycle control, cell division, chromosome partitioning	1.53%
Signal transduction mechanisms	2.29%
Cell wall/membrane/envelope biogenesis	4.58%
Cell motility	9.16%
Intracellular trafficking, secretion, and vesicular transport	9.92%
Posttranslational modification, protein turnover, chaperones	2.29%
Energy production and conversion	3.05%
Carbohydrate transport and metabolism	2.29%
Amino acid transport and metabolism	7.63%
Nucleotide transport and metabolism	0.76%
Lipid transport and metabolism	0.76%
Inorganic ion transport and metabolism	4.58%
Secondary metabolites biosynthesis, transport and catabolism	1.53%

Figure 4. 3 (A) Arrows depicting the clickable navigation links. (B) Functional class proportions (Excluding R and S classifications) for Salmonella enterica subsp. enterica serovar Typhimurium str. LT2 (NC_003197) GI group 1.

4.1.2 Virulence factor assignments

To enable identification of virulence-associated genes in groups of GIs, the MvirDB database (Zhou, et al., 2007) was used as a reference for mapping of genes by BLAST similarity. MvirDB is a collection of sequences representing known toxins, virulence factors and antibiotic resistance genes. The interface allows a user to search or match queries to database records by sequence similarity BLAST search. In total 392,622 coding sequences found in GIs of the Pre_GI database were searched by BLASTP for matches against the MvirDB database with e-value stringency set to 0.004 and a minimum bit score of at least 400. Here we considered a stricter bit score to ensure that the hits against the MvirDB are closely related matches. The results revealed 27,995 hits against the MvirDB database distributed among 3347 groups of GIs in 2002 replicons. The results of identification of virulence factors by search through MvirDB were added to our database and may be explored through the Flux Visualizer system by the virulence factor links as per ***Figure 4.4***.



flux.bi.up.ac.za/viewVir.php?accgroup=NC_003197_GIG1

Flux Visualizer

Genomic island distribution and flow analysis

Virulence genes within genomic islands group: NC_003197_GIG1

Virulence Gene Description	Source	Identity Percentage	Bit Score
NP_463122-virulence protein-hypothetical protein STM4257	Salmonella typhimurium LT2	100.0	425.0
NP_463123-virulence protein-putative methyl-accepting chemotaxis protein	Salmonella typhimurium LT2	99.78	932.0
NP_463124-virulence protein-putative ABC exporter outer membrane component	Salmonella typhimurium LT2	100.0	892.0
NP_463125-virulence protein-cation efflux pump	Salmonella typhimurium LT2	100.0	844.0
NP_463126-virulence protein-putative inner membrane protein	Salmonella typhimurium LT2	100.0	11071.0
NP_459541-virulence protein-outer membrane usher protein precursor	Salmonella typhimurium LT2	100.0	1776.0
NP_459542-virulence protein-minor fimbrial subunit	Salmonella typhimurium LT2	100.0	688.0
NP_459544-virulence protein-transcriptional regulator FimZ	Salmonella typhimurium LT2	100.0	427.0
NP_459545-virulence protein-putative regulatory protein	Salmonella typhimurium LT2	100.0	488.0
NP_459547-virulence protein-putative fimbrial protein	Salmonella typhimurium LT2	100.0	411.0
NP_836042-pathogenicity island-putative bactoprenol glucosyl transferase	Shigella flexneri 2a str. 2457T	89.22	566.0
NP_706258-virulence protein-putative bactoprenol glucosyl transferase	Shigella flexneri 2a str. 301	89.22	566.0

Figure 4. 4 BLASTP hits for Salmonella enterica subsp. enterica serovar Typhimurium str. LT2 (NC_003197) GI group 1 against the MvirDB.

4.1.3 Gene networks

Another functionality of the Flux Visualizer system is to reconstruct possible horizontal gene fluxes based on the sharing of similar genes in GIs by different microorganisms. The Pre_GI database contains records of significant BLASTP matches among CDS in GIs, which were subsequently used to find homologous pairs of genes. These BLASTP matches served for an initial outlining of organisms, which might exchange their genes by horizontal gene transfer. For the purpose of exploring mutual functional traits, we evaluated the BLASTP hits among 392,622 CDS for all of the replicons in Pre_GI database. Genes related to integration or transfer of mobile genetic vectors were excluded to avoid finding links between organisms sharing these selfish functions or with unknown functions, which are abundant in GIs. In the first step, gene annotations were scanned for keywords among the following: “Integrase”, “Transposase”, “Phage”, “Plasmid”, “Hypothetical” and “Unknown”. Any genes containing one of these keywords were omitted from subsequent analysis. Only those CDS which shared a bit score of at least 400 to any other CDS in GIs, were deemed as a connection in a network. Then by the recursive analysis of all the CDS matches for the considered genes, whole sets of connections were used to establish each gene network. A total of 44,667 unrelated networks were constructed and comprised of 1,104,771 BLAST hits, reflecting gene homology interlinks between all GIs. The gene networks were saved to the MySQL database to serve as a reference of common function indicators. ***Figure 4.5*** shows a heatplot to visualize densities of shared homologous genes in GIs for several of the most represented genera in the Pre_GI database. The plot is based on the proportion of organisms found to share common genes. Organisms from *Bacillus* and *Clostridium* showed the highest proportions of BLAST hits to genes in other genera which may indicate increased promiscuity among those organisms. Furthermore, *Vibrio* showed particularly high proportions of common genes functions to *Bacillus* and *Escherichia* with more than 75% of organisms classified under the *Vibrio* genus, showing significant BLAST references to both *Bacillus* and *Escherichia*. Other examples of organisms found to show increased common gene functions were among *Rickettsia* and *Clostridium* as well as *Mycoplasma* to both *Bacillus* and *Clostridium*.

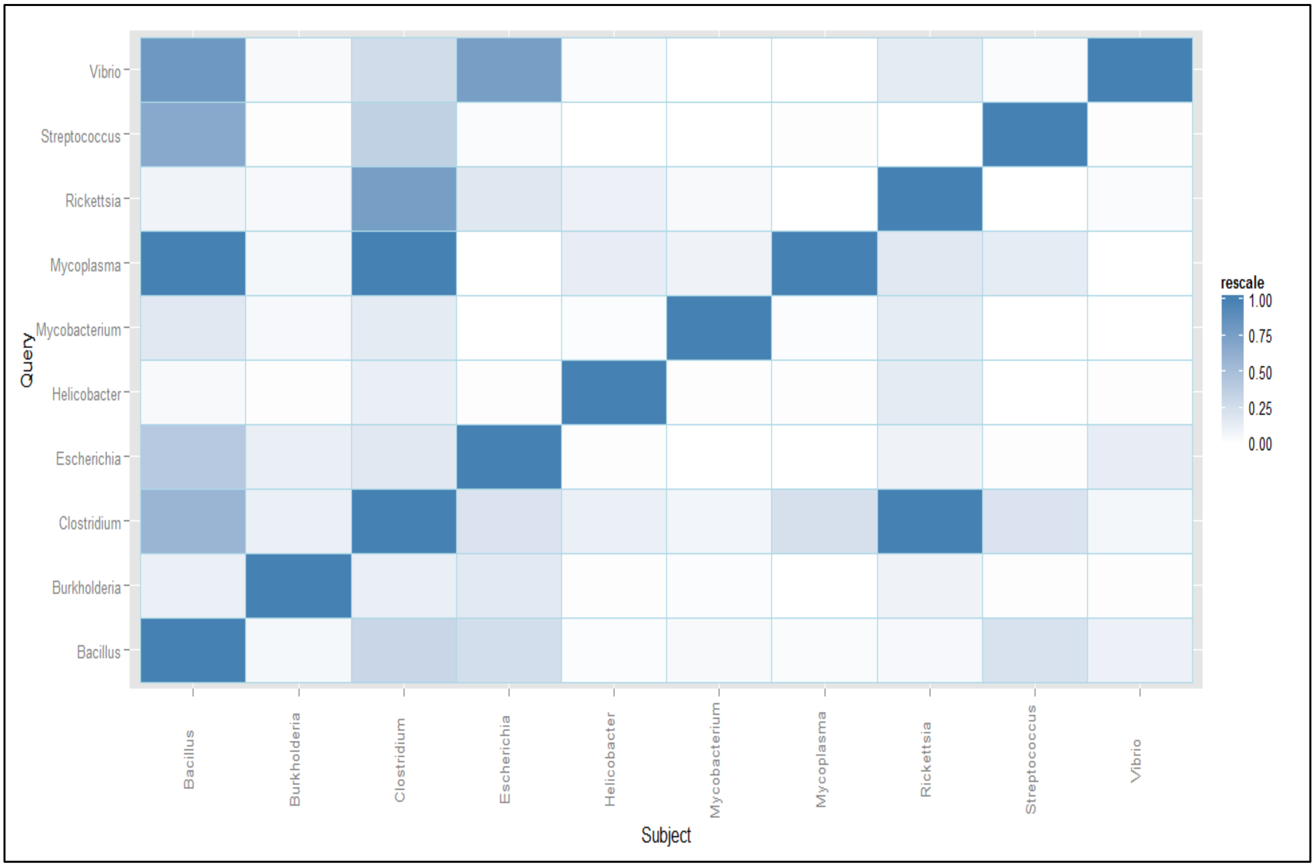


Figure 4. 5 Proportions of organisms with common genes in the networks among major genera.

Figure 4.6 shows the proportions of the number of BLAST subjects in all of the networks. Almost 50% of the organisms identified as sharing some form of common functionality, were only connected by a maximum of 5 other subjects. Likewise, 90% of the organisms were connected by a maximum of 40 other subjects. Conversely, we observed a minority of networks with up to 418 significant BLAST hits between different organisms. The functional categories of these genes which seemed extremely popular, were exclusively related to transcription and translation with the majority of annotation descriptions among the following: “DNA-dependent RNA polymerase”, “Translation elongation factor”, “Carbamoyl phosphate”.

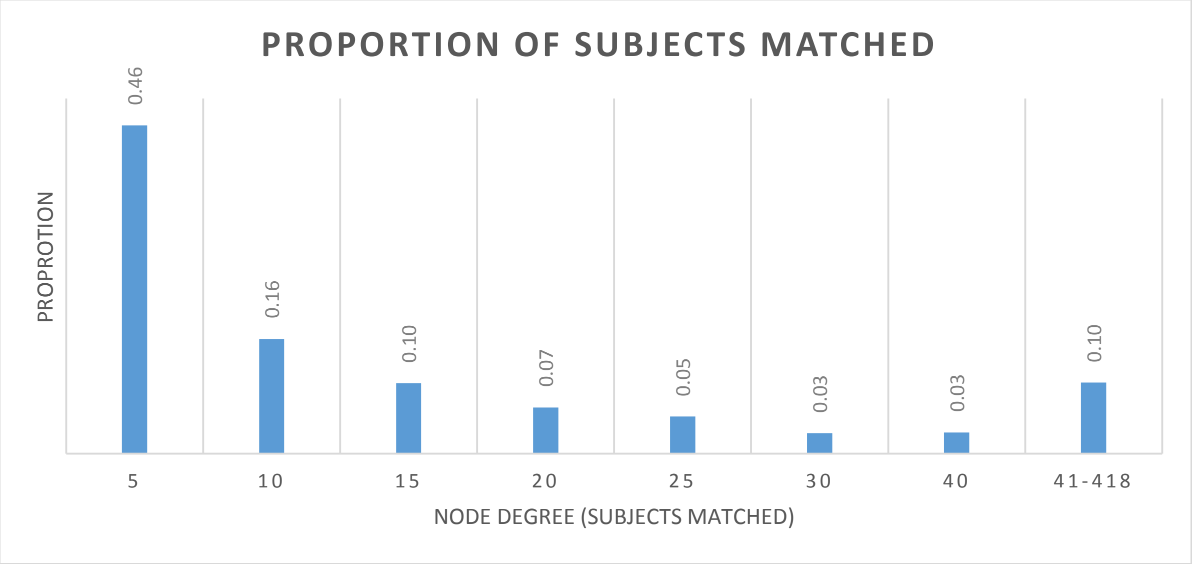


Figure 4. 6 Proportions of the of node degree ranges in the gene networks.

To evaluate associations between the identified BLAST hits and their sequence compositional similarities, homologous gene pairs which also showed OUP similarities between their respective GIs were indexed. The occurrence of both measures would propose strong evidence for HGT among the formulated networks. In contrast, lower frequencies of co-occurring BLAST and OUP similarity instances may indicate that individually, BLAST similarities are not well suited for the identification of HGT relations due to a higher rate of mutations in GIs and they merely serve to infer functional resemblances among GIs. We evaluated the GIs of the organisms which showed BLAST hits between their CDS in ***Figure 4.5*** and found that on average, 72% of these genes also shared OUP similarities between their respective GIs**.** This indicated that some correlation exists between BLAST revealed sequence similarity and the level of compositional OUP similarity among the considered GIs. ***Figure 4.7*** shows the proportion of the GIs in organisms sharing common genes as well as OUP similarities among *Escherichia*, *Burkholderia* and *Mycoplasma* in their respective genera.

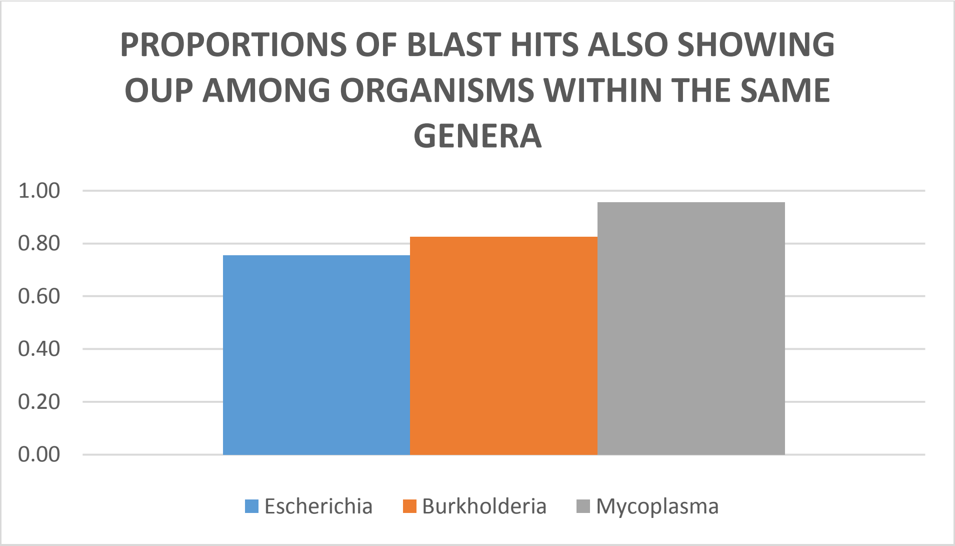
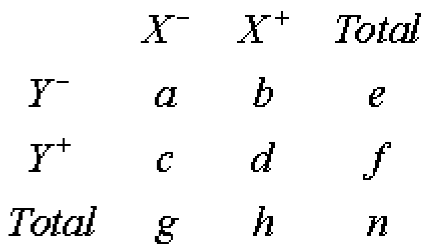


Figure 4. 7 Frequency of OUP links > 75% among groups of GI sharing BLASTP sequence hits

In retrospect, the creation of the gene networks assists in the diversification of suggestions to the users of the system as per ***Figure 3. 9*** in Chapter 3. Users are able to find organisms which share specific genes through the use of genetic keywords which are filtered against the constructed gene networks. Interlinks of homologous genes may reveal indirect donor-recipient relationships between organisms which do not inherently share strong sequence similarities but are related to each other through intermediate organisms.

4.2 Co-occurrence of functional categories

To better understand the role of HGT between microorganisms, especially among diverse species, it is important to analyze the functional categories of genes which are more likely to be transferred together within one GI. All identified GI groups were compared against each other on a basis of shared gene categories to primarily attempt the identification of any possible ontological links between the groups of GIs, COG categories and associated virulence factor indicators. A unique list of alphabetically arranged COG categories was extracted for each GI group to produce a list of all the COG combinations among the considered GI groups. This list served as a measure of COG category commonality and was used to formulate associations of co-occurring COG categories. We calculated the Phi correlation coefficients (ϕ) of individual COG categories based on their co-occurrence in a distinct list of COG combinations. Phi correlations are used to assess correlations between two variables where both variables are dichotomous. If we have a 2 × 2 table for two random variables X and Y as per ***Table 4.2*** with a, b, c and d as non-negative counts of the number of observations that sum to n, the total number of observations, and X^- and X^+ representing the absence and presence of X respectively, then the two variables are considered positively associated if most of the data falls along the diagonal cells (i.e., a and d are larger than b and c). In contrast, two binary variables are considered negatively associated if most of the data falls off the diagonal. The Phi coefficient (ϕ) that describes the association of X and Y is calculated by ***Equation 4.1***. Phi compares the product of the diagonal cells (a x d) to the product of the off-diagonal cells (b x c). The denominator is an adjustment that ensures that Phi is always between -1 and +1.



	X^-	X^+	*Total*
Y^-	a	b	e
Y^+	c	d	f
Total	g	h	n

Table 4. 2 Matrix of two random variables, X and Y, with a, b, c and d as non-negative counts of the number of observations.

$$\phi = \frac{ad - bc}{\sqrt{efgh}}$$

Equation 4. 1 Phi Correlation Coefficient

Figure 4.8 shows a heat map of Phi correlation coefficients for co-occurring COG class frequencies. The correlation coefficients produced by this matrix were all positive with a range between 0.05 to 0.51. This result showed that no COG category was strongly associated with any other category and because we did also not observe any negative coefficients we could not make any assertions regarding the absence of individual categories. The strongest positive associations were between the COG category combinations of (C-I), (E-G), (K-T), (Q-I), (O-C), and (F-J). In contrast, the COG categories, D and N showed the lowest Phi coefficients (little or no association) to every other category.

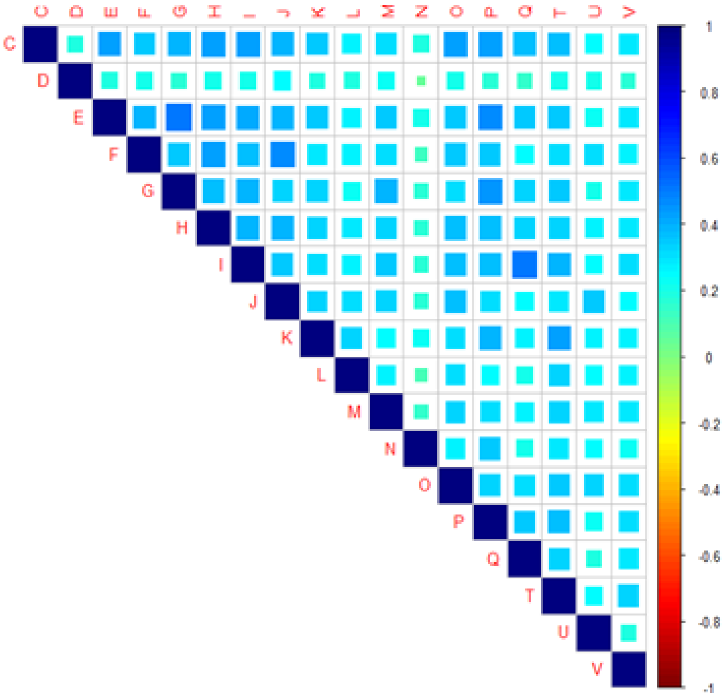


Figure 4. 8 Heatmap of the correlation matrix encompassing the frequency of co-occurrence of the COG categories.

Moreover, we investigated the associations of virulence factor frequencies to each of the COG categories by using the number of genes classified by MvirDB in each GI group (*Section 4.1.3),* to estimate which of the COG categories were more likely to be associated with the presence of virulence factors. For every COG category the frequencies of virulence factors F1 and F2 were calculated if the specific COG category was present in a GI group's COG combination (for positive associations), and also if the COG category was absent (for negative associations). The linkage association was calculated as:

$$L = \frac{F1/n - F2/(N-n)}{MAX(F1/n, F2/(N-n))}$$

Equation 4. 2 Virulence factor linkage association

Where N is the total number of GI groups, and n – the number of GI groups containing genes of the current COG category.

Table 4.3 shows the results of the virulence factor linkage associations for each of the COG categories. Positive associations to virulence factors were identified among the COG categories of N, Q and U. Microbial secondary metabolites (COG category Q) include among others, antibiotics, toxins, effectors of ecological competition and symbiosis, enzyme inhibitors and immunomodulating agent utilities. We may therefore expect this category to reflect a high degree of associated virulence. COG category N comprises cell motility proteins including bacterial pili, which are regarded as important virulence factors. Pili are recognized as crucial role-players in many virulence associated processes such as adhesion, biofilm formation and virulence factor secretion (Johanna & Westerlund-Wikström, 2013). In combination with COG classes "Intracellular trafficking and secretion" (COG category U), we may also expect genes classified under categories Q and N to show alleviated virulence factor linkage associations purely based on the combined attributes that are gained from each of these categories conjunction with COG category U. In other words, a combination of genes from each of these categories may serve to provide greater virulent utilities than genes from each separate category.

Linkage of virulence factors to COG categories								
C	D	E	F	G	H	I	J	K
-0.34993	-0.15591	-0.30076	-0.37689	-0.30935	-0.20287	-0.28937	-0.37952	-0.24752
L	M	N	O	P	Q	T	U	V
-0.06251	-0.08723	0.33333	-0.29051	-0.17532	0.312128	-0.21457	0.258938	0

Table 4. 3 Virulence factor linkage associations for each of the COG categories.

Conversely, negative associations were found among COG categories C, F and J. The COG categories of "Energy production", "Nucleotide transport" and "Translation" represent common and genetic functions which we may expect to be present in most horizontally transferred regions. It is therefore no surprise that these classes are disassociated with the identified virulence factors because they are more related to housekeeping and informational roles, necessary for general cell functioning.

To enable the separation of groups of GIs, firstly by their characterization of COG category combinations, and secondly by the frequency of virulence factor linkage, it was necessary to calculate a new similarity measure for the MCL clustering of GI groups by functional profile. The functional similarity of groups of GIs were calculated as follows:

$$Fsim = I \times \frac{(a+b)}{(a \times L_a \ + \ b \times L_b)}$$

Equation 4. 3 GI group functional similarity measure

where I is intersection, the number of COG categories shared by a query and subject GI group; a and b is the total number of groups of GI represented by query and subject COG category combinations, respectively; L_a is number of COG categories in the query GI group and L_b, the number of COG categories in the subject GI group. All GI groups pairs were compared against each other using this formula to produce GI group functional similarity scores. In turn, these functional similarity scores were used to evaluate the MCL clustering of GI groups with different similarity cutoffs. The execution of the MCL algorithm with a similarity cutoff of 0.97 and inflation value of 350 produced the result in ***Figure 4.9***. The clusters were named by their central nodes with singletons and pairs of nodes removed. The red nodes are the top 10% of GI groups with the highest frequency of virulence factors. From ***Figure 4.9*** we observe several small clusters with high virulence factors frequencies. In particular, clusters with central nodes of NQ, NU, NV, INQ and MNO present strong evidence for virulence associations and we may expect GI groups harboring these COG combinations to exert some degree of virulence. In contrast, the densest cluster (top left) presents the combination of the most versatile COG categories among the groups of GIs, where virulence factors are sparsely distributed.

The second densest cluster, characterized by the central node KMNTU, also presents a common COG combination which is indicative of increased virulence association. Here we might conclude on a similar premise, that organisms which engage in HGT and exchange genes which represent this category combination, may also be expected to show some degree of associated virulence.

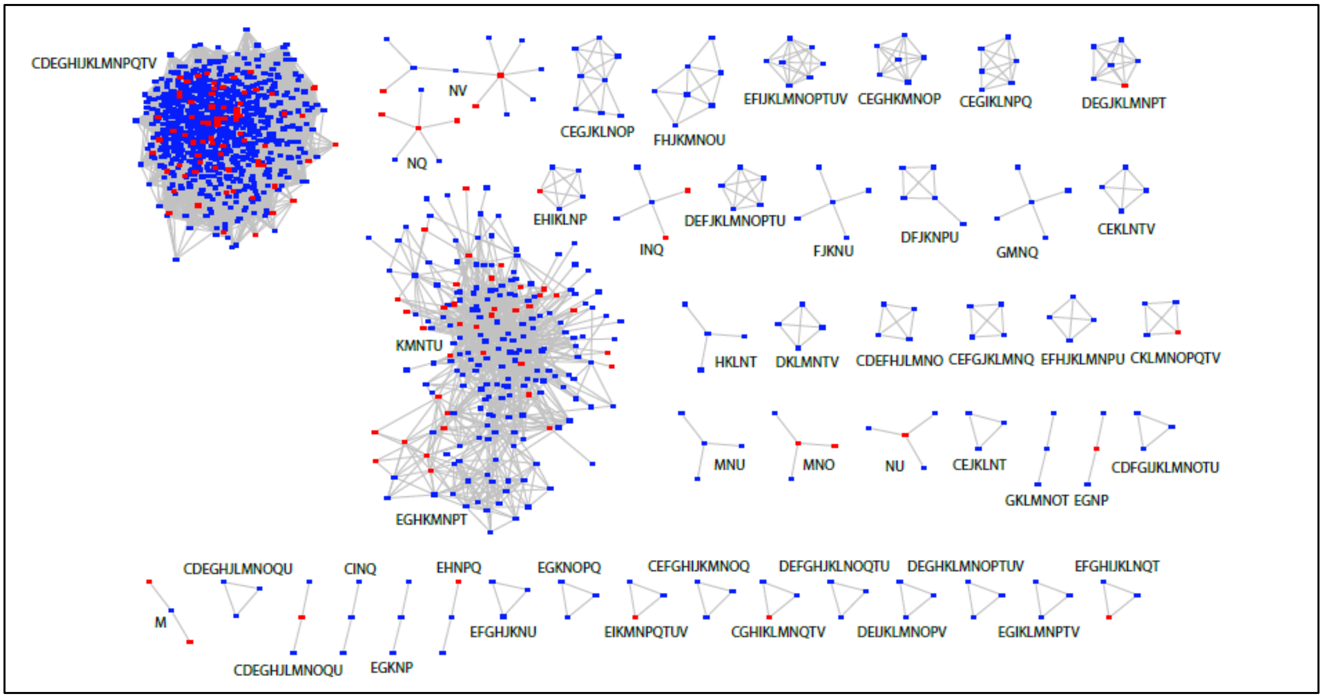


Figure 4. 9 Separation of GI groups sharing COG category combinations. The red nodes are the top 10% of groups of GIs with the highest frequency of virulence factors. The clusters are titled by their central nodes.

4.3 Associations of shared gene matches among groups of GIs

The assessment of common functional traits in HGT events among the replicons listed in the Pre_GI database, provided a foundation for the analyses of donor-recipient relations between a myriad of species. To identify a set of candidates which may be used to validate the functionality of the system, frequencies of BLAST sequence similarities found among the constructed groups of GIs were evaluated. Organisms identified as sharing a significant number of similar genes between their GI groups were perceived as prime candidates for a case study as they would most likely contain common genes acquired by true HGT events, as opposed to containing similar genes by chance. GI groups were evaluated to find similar genes in other GI groups based on BLAST similarities captured in the generated gene networks. Each of the 44,667 gene networks (as described in *Section 4.1.3*) were used to assess the number of similar genes found among each pair of GI groups.

A total of 452,259 BLAST hit references were identified between all considered GI group combinations. We found that 67% of all the GI group combinations shared only one significant BLAST match. As the number of common genes found between the GI groups combinations increased, the frequency of such occurrences diminished rapidly. ***Table 4.4*** shows the proportions of BLAST hit frequencies between the considered GI groups. This measure served as an indicator for the identification of significant BLAST match frequencies between groups of GIs and we postulated that GI group combinations containing only one gene match did not indicate sufficient evidence for subsequent analysis. ***Figure 4.10*** reveals the COG class descriptions among GI groups sharing only one BLAST hit. These genes were distributed mainly among the functional classifications of transcription and translation as well as the nucleotide and amino acid transport classes. We may expect these over-represented genes to form part of a backbone of genes necessary for the process of successful HGT integration and speculatively, expect most GI groups to harbor at least some of them.

Number of BLAST similarities	Frequency	Proportion
1	301733	66.72%
2	84087	18.59%
3	34980	7.73%
4	9813	2.17%
≥5	5176	4.79%

Table 4. 4 Proportion of BLAST similarities between GI groups.

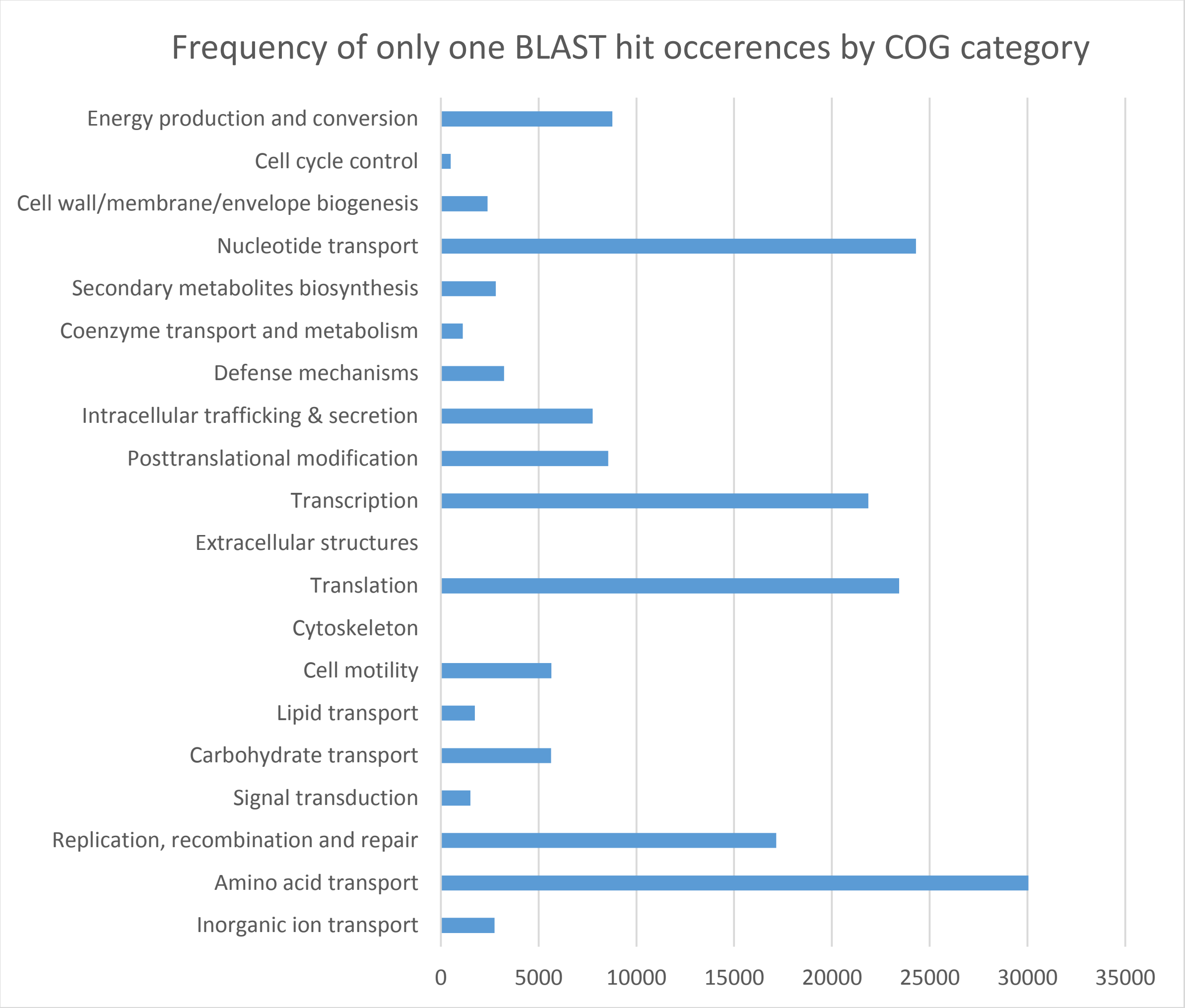


Figure 4. 10 Frequency of single BLAST matches among GI groups per COG class.

In contrast, we may expect significant frequencies of GI groups containing sequence similarities between related strains or organisms within the same species. Consequently, we considered filtering the GI group combinations to show only sequence similarities between organisms of diverse taxonomic ranks. As per ***Table 4.5*** and ***4.6***, the frequency and proportions of GI groups sharing more than one gene further decreased when we considered only sequence similarities between GI groups belonging to different species, orders and phyla.

	Proportion			
Number of BLAST similarities	Same Species	Different Species	Different Orders	Different Phyla
1	66.72%	65.56%	54.75%	36.37%
2	18.59%	17.61%	14.15%	9.47%
3	7.73%	7.04%	5.55%	3.19%
4	2.17%	1.54%	0.76%	0.34%

Table 4. 5 The proportion of sequence similarities between groups of GIs from different taxonomic ranks.

	Frequency			
Number of BLAST similarities	Same Species	Different Species	Different Orders	Different Phyla
1	301733	296501	247627	164497
2	84087	79660	64001	42846
3	34980	31829	25114	14439
4	9813	6980	3423	1518
5	5176	2993	973	409
6	4082	2241	603	258
7	2391	1059	170	72
8	1716	640	141	73
9	1283	461	38	26
10	1133	413	91	58

Table 4. 6 The frequency of sequence similarities between groups of GIs from different taxonomic ranks.

We subsequently assessed groups of GIs sharing either the most BLAST similarities between their GI groups, regardless of taxonomic classification, or multiple BLAST matches from different phylogenetic ranks. Therefore, GI group combinations which showed the highest degree of common genes were not the only sets of organisms considered. We also considered a set of diverse organisms to compensate for the expectation of organisms within the same species sharing substantial gene BLAST indexes. We used inter- and intra-species combinations of organisms showing multiple BLAST matches between their GI groups and observed a significant number of such occurrences between two sets of organisms. Firstly, two *Bacillus anthracis* strains (*Bacillus anthracis str. Sterne* NC_005945; *Bacillus anthracis str. A0248* NC_012659) were found to share the most BLAST similarities between their GI groups and secondly, various *Mycobacterium* and *Streptomyces* species were identified to share the most BLAST similarities between the groups of GIs from different taxonomic background. The GI group combinations from these organisms were selected for a case study using the Flux Visualizer system as described in Chapter 5.

4.4 Discussion

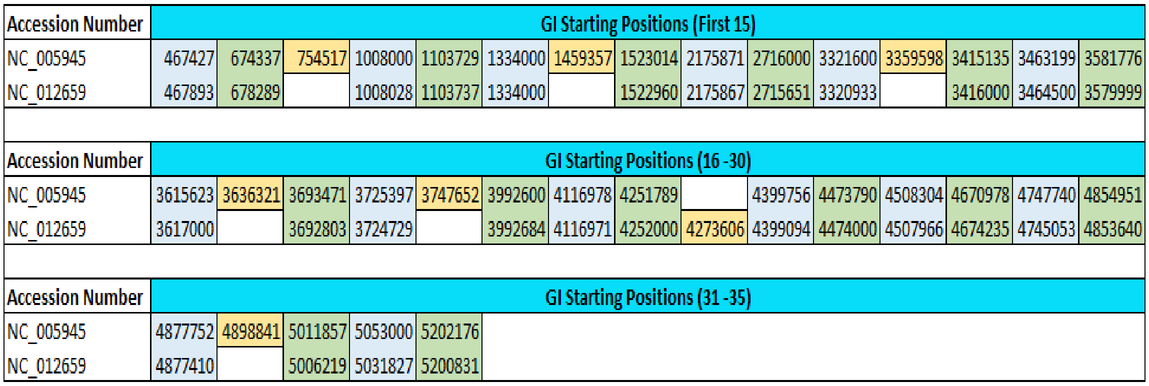
The distribution of COG categories among groups of GIs is in agreement with the results obtained by Popa et al., (2011). It is no surprise that genes relating to transcription, translation and recombination were among the highest ranking functional classes as these are necessary proteins involved in DNA integration and strongly linked to the process of HGT. However, we also observed other COG classes such as “Cell wall/membrane/envelope biogenesis” and “Amino acid transport and metabolism” as highly regarded gene representatives. This indicates that even beyond the genes required for the process of HGT, certain functional classes are favored, probably because they provide host organisms with more advantageous abilities. The COG categories most frequently identified (“Carbohydrate transport and metabolism” and “Amino acid transport and metabolism”) among groups of GIs, represent the utilities which are most common among transfer events, apart from genes related to the actual integration itself. Additionally, we observed similar distributions of COG functions in each separate GI group (***Figure 4.2***). This is interesting as it proposes that there might be a template of COG category combinations which can be associated with the abundant spread of HGT, even if the genes provide diverse utilities. Several positive co-occurring COG category combinations were found. We may expect these categories to complement each other, for example COG category combinations of E and G, which are typically associated with uptake of nutrients or the excretion of metabolic waste products, where the transport system corresponds to a sensor for external stimuli and is tightly linked to common controls between the carbohydrate metabolic pathways. COG categories N,U and Q seemed to not only show the strongest virulence factor linkage associations but combinations of these categories are trademarks for HGT characterized by virulence gene exchanges, as identified by the clustering of GI group functional similarities. Furthermore, many organisms were found to share common functionalities through horizontal gene transfer based on significant BLAST matches. The indexes of genes shared in this way, were used to generate gene networks. The gene networks enabled for the provision of suggestions of organisms to the users of the system based on common functional traits. Lastly, to identify sets of organisms which may be evaluated using Flux Visualizer system, frequencies of BLAST sequence similarities found among pairs of GIs groups were assessed and two sets or candidates were selected for flux analyses in Chapter 5.

Chapter 5 Investigation of HGT relationships using the Flux Visualizer

In this chapter we considered several case studies as examples of practical applications of the Flux Visualizer. For these case studies, the following organisms were selected: *Bacillus anthracis str. Sterne* NC_005945 and *Bacillus anthracis str. A0248* NC_012659 as representatives of organisms which showed the most BLAST matches between their GI groups from similar taxonomic backgrounds, and various organisms from *Mycobacterium* and *Streptomyces* as they shared the most BLAST similarities between the groups of GIs from different taxonomic backgrounds.

5.1 Case study for Bacillus anthracis

The maximum number of sequence similarities found between all considered GI groups was 91 genes. These matches occurred specifically for HGT between *Bacillus anthracis str. Sterne* (NC_005945_GIG1) and *Bacillus anthracis str. A0248* (NC_012659_GIG1). The virulence of most *B. anthracis* strains are associated with two plasmids namely, pXO1 and pXO2. The plasmid pXO1 is required for synthesis of the anthrax toxin proteins. Plasmid pXO2 harbors genes required for the synthesis of an antiphagocytic capsule. The Sterne strain of *Bacillus anthracis* lacks the plasmid pXO2 which results in an avirulent phenotype (Okinaka, et al., 1999). Although the pXO2 genes are not self-transmissible, they may be transferred via conjunctive plasmids originating in *B. thuringiensis* (Koehler, 2002). In both organisms we observed grouping of each organism's respective GIs into a single large group (GIG1). As per ***Figure 5.1***, the GI genetic positions between these strains are very similar. This proposes that both strains either share a common donor and comparable recombination hotspots, or they descended from the same ancestral organism, which possessed already all these GIs. It is then not surprising that we observed a large amount of common genes as per the two example GI (positions 467KB & 1008KB) BLASTP visualizations in ***Figure 5.2*** (A & B).



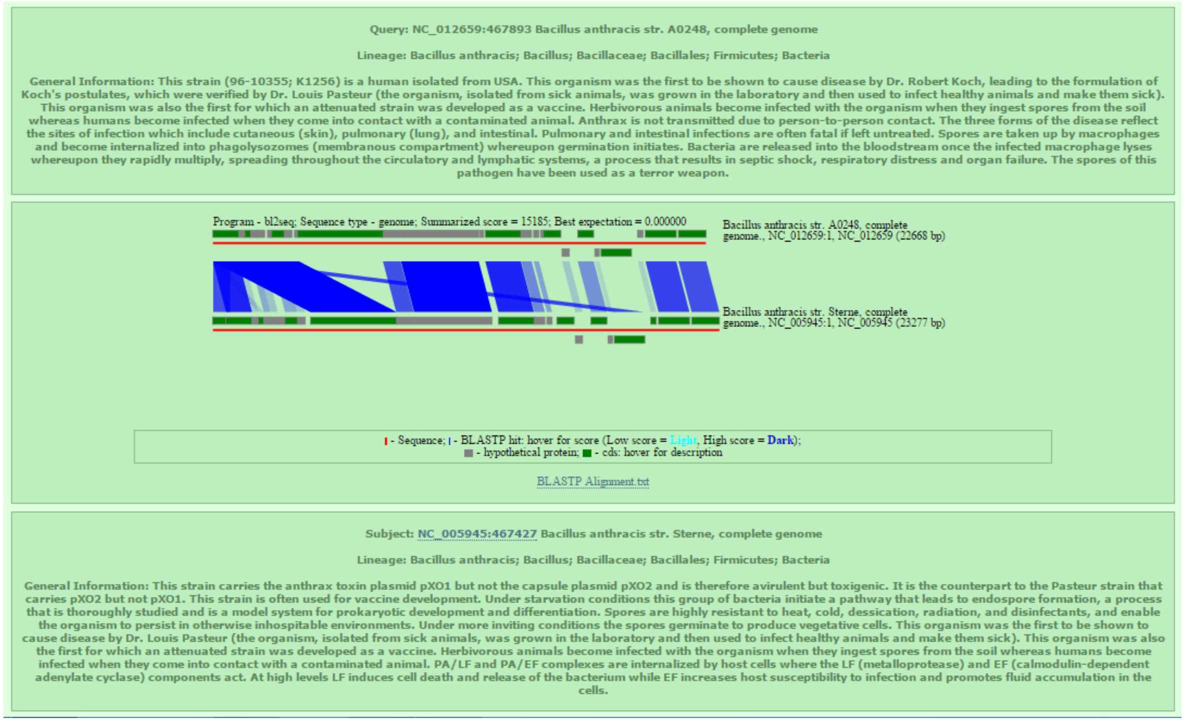
Accession Number	GI Starting Positions (First 15)														
NC_005945	467427	674337	754517	1008000	1103729	1334000	1459357	1523014	2175871	2716000	3321600	3359598	3415135	3463199	3581776
NC_012659	467893	678289		1008028	1103737	1334000		1522960	2175867	2715651	3320933		3416000	3464500	3579999

Accession Number	GI Starting Positions (16 -30)														
NC_005945	3615623	3636321	3693471	3725397	3747652	3992600	4116978	4251789		4399756	4473790	4508304	4670978	4747740	4854951
NC_012659	3617000		3692803	3724729		3992684	4116971	4252000	4273606	4399094	4474000	4507966	4674235	4745053	4853640

Accession Number	GI Starting Positions (31 -35)				
NC_005945	4877752	4898841	5011857	5053000	5202176
NC_012659	4877410		5006219	5031827	5200831

Figure 5. 1 Genetic positions of GIs in two closely related Bacillus anthracis strains.

A



Query: NC_012659:467893 Bacillus anthracis str. A0248, complete genome

Lineage: Bacillus anthracis; Bacillus; Bacillaceae; Bacillales; Firmicutes; Bacteria

General Information: This strain (96-10355; K1256) is a human isolated from USA. This organism was the first to be shown to cause disease by Dr. Robert Koch, leading to the formulation of Koch's postulates, which were verified by Dr. Louis Pasteur (the organism, isolated from sick animals, was grown in the laboratory and then used to infect healthy animals and make them sick). This organism was also the first for which an attenuated strain was developed as a vaccine. Herbivorous animals become infected with the organism when they ingest spores from the soil whereas humans become infected when they come into contact with a contaminated animal. Anthrax is not transmitted due to person-to-person contact. The three forms of the disease reflect the sites of infection which include cutaneous (skin), pulmonary (lung), and intestinal. Pulmonary and intestinal infections are often fatal if left untreated. Spores are taken up by macrophages and become internalized into phagolysozomes (membranous compartment) whereupon germination initiates. Bacteria are released into the bloodstream once the infected macrophage lyses whereupon they rapidly multiply, spreading throughout the circulatory and lymphatic systems, a process that results in septic shock, respiratory distress and organ failure. The spores of this pathogen have been used as a terror weapon.

Program - bl2seq; Sequence type - genome; Summarized score = 15185; Best expectation = 0.000000

Bacillus anthracis str. A0248, complete genome., NC_012659:1, NC_012659 (22668 bp)
Bacillus anthracis str. Sterne, complete genome., NC_005945:1, NC_005945 (23277 bp)

🟥 - Sequence; 🟦 - BLASTP hit: hover for score (Low score = Light, High score = Dark);
▨ - hypothetical protein; 🟩 - cds: hover for description

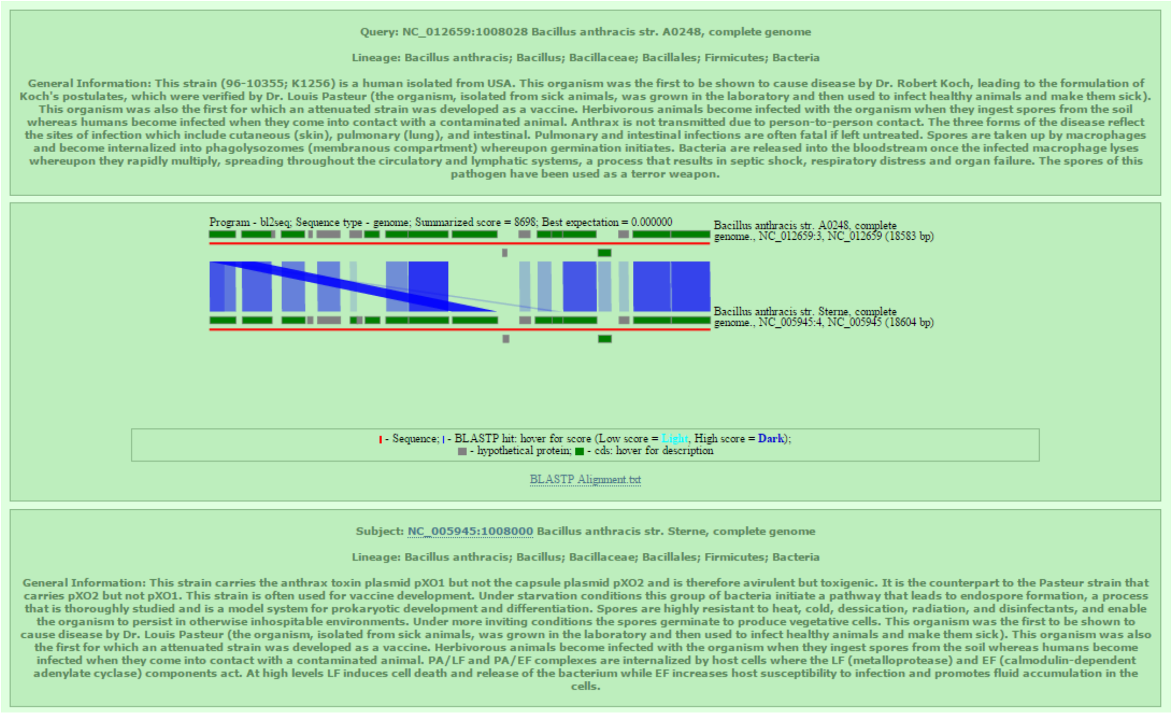
BLASTP Alignment.txt

Subject: NC_005945:467427 Bacillus anthracis str. Sterne, complete genome

Lineage: Bacillus anthracis; Bacillus; Bacillaceae; Bacillales; Firmicutes; Bacteria

General Information: This strain carries the anthrax toxin plasmid pXO1 but not the capsule plasmid pXO2 and is therefore avirulent but toxigenic. It is the counterpart to the Pasteur strain that carries pXO2 but not pXO1. This strain is often used for vaccine development. Under starvation conditions this group of bacteria initiate a pathway that leads to endospore formation, a process that is thoroughly studied and is a model system for prokaryotic development and differentiation. Spores are highly resistant to heat, cold, dessication, radiation, and disinfectants, and enable the organism to persist in otherwise inhospitable environments. Under more inviting conditions the spores germinate to produce vegetative cells. This organism was the first to be shown to cause disease by Dr. Louis Pasteur (the organism, isolated from sick animals, was grown in the laboratory and then used to infect healthy animals and make them sick). This organism was also the first for which an attenuated strain was developed as a vaccine. Herbivorous animals become infected with the organism when they ingest spores from the soil whereas humans become infected when they come into contact with a contaminated animal. PA/LF and PA/EF complexes are internalized by host cells where the LF (metalloprotease) and EF (calmodulin-dependent adenylate cyclase) components act. At high levels LF induces cell death and release of the bacterium while EF increases host susceptibility to infection and promotes fluid accumulation in the cells.

B



Query: NC_012659:1008028 Bacillus anthracis str. A0248, complete genome

Lineage: Bacillus anthracis; Bacillus; Bacillaceae; Bacillales; Firmicutes; Bacteria

General Information: This strain (96-10355; K1256) is a human isolated from USA. This organism was the first to be shown to cause disease by Dr. Robert Koch, leading to the formulation of Koch's postulates, which were verified by Dr. Louis Pasteur (the organism, isolated from sick animals, was grown in the laboratory and then used to infect healthy animals and make them sick). This organism was also the first for which an attenuated strain was developed as a vaccine. Herbivorous animals become infected with the organism when they ingest spores from the soil whereas humans become infected when they come into contact with a contaminated animal. Anthrax is not transmitted due to person-to-person contact. The three forms of the disease reflect the sites of infection which include cutaneous (skin), pulmonary (lung), and intestinal. Pulmonary and intestinal infections are often fatal if left untreated. Spores are taken up by macrophages and become internalized into phagolysozomes (membranous compartment) whereupon germination initiates. Bacteria are released into the bloodstream once the infected macrophage lyses whereupon they rapidly multiply, spreading throughout the circulatory and lymphatic systems, a process that results in septic shock, respiratory distress and organ failure. The spores of this pathogen have been used as a terror weapon.

Program - bl2seq; Sequence type - genome; Summarized score = 8698; Best expectation = 0.000000

Bacillus anthracis str. A0248, complete genome., NC_012659:3, NC_012659 (18583 bp)
Bacillus anthracis str. Sterne, complete genome., NC_005945:4, NC_005945 (18604 bp)

🟥 - Sequence; 🟦 - BLASTP hit: hover for score (Low score = Light, High score = Dark);
▨ - hypothetical protein; 🟩 - cds: hover for description

BLASTP Alignment.txt

Subject: NC_005945:1008000 Bacillus anthracis str. Sterne, complete genome

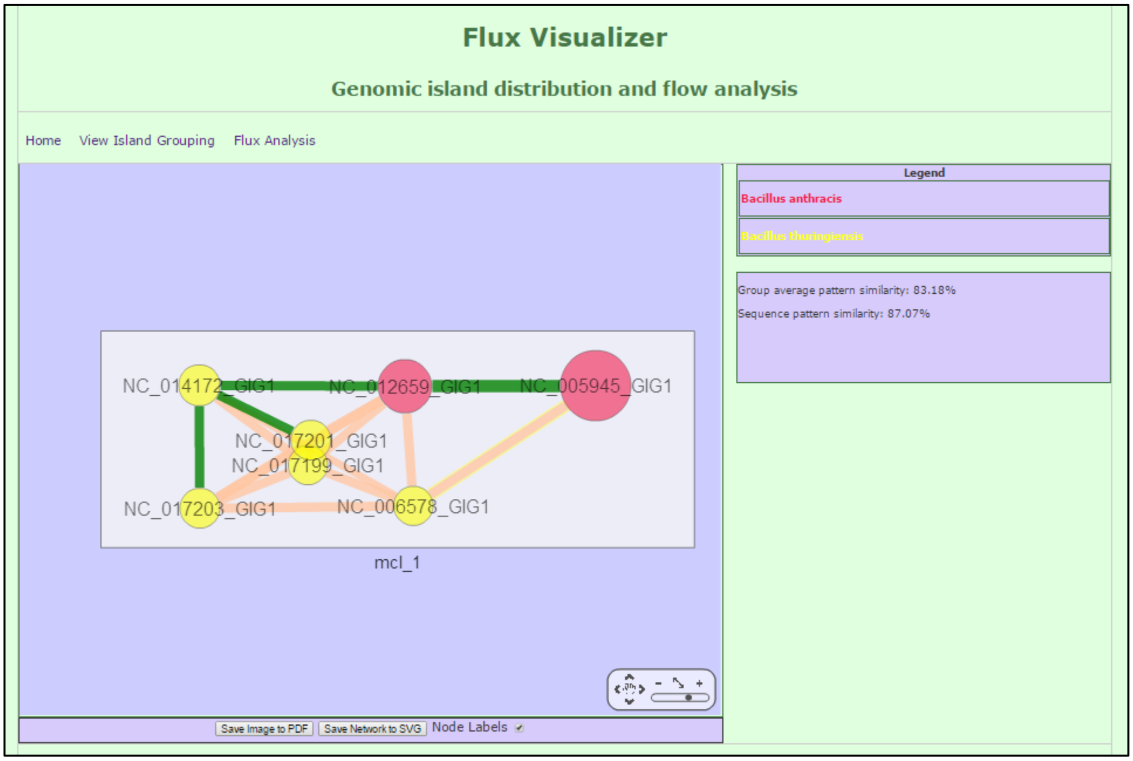
Lineage: Bacillus anthracis; Bacillus; Bacillaceae; Bacillales; Firmicutes; Bacteria

General Information: This strain carries the anthrax toxin plasmid pXO1 but not the capsule plasmid pXO2 and is therefore avirulent but toxigenic. It is the counterpart to the Pasteur strain that carries pXO2 but not pXO1. This strain is often used for vaccine development. Under starvation conditions this group of bacteria initiate a pathway that leads to endospore formation, a process that is thoroughly studied and is a model system for prokaryotic development and differentiation. Spores are highly resistant to heat, cold, dessication, radiation, and disinfectants, and enable the organism to persist in otherwise inhospitable environments. Under more inviting conditions the spores germinate to produce vegetative cells. This organism was the first to be shown to cause disease by Dr. Louis Pasteur (the organism, isolated from sick animals, was grown in the laboratory and then used to infect healthy animals and make them sick). This organism was also the first for which an attenuated strain was developed as a vaccine. Herbivorous animals become infected with the organism when they ingest spores from the soil whereas humans become infected when they come into contact with a contaminated animal. PA/LF and PA/EF complexes are internalized by host cells where the LF (metalloprotease) and EF (calmodulin-dependent adenylate cyclase) components act. At high levels LF induces cell death and release of the bacterium while EF increases host susceptibility to infection and promotes fluid accumulation in the cells.

Figure 5. 2 Example BLASTP visualizations of two GIs loci pairs (A & B) between Bacillus anthracis str. Sterne (NC_005945) and Bacillus anthracis str. A0248 (NC_012659).

We performed additional analysis on these strains to identify potential HGT partners outside of the *Bacillus* genus. Okinaka et al. (1999), suggests the sources of pXO1 may have a rare and perhaps unique origin of replication. The authors evaluate various gene clusters within the pXO1 plasmid and provide evidence of possible HGT from diverse organisms belonging to *Staphylococcus aureus*, *Streptococcus pyogenes*, *Bordetella pertussis* and *Clostridium* species, among others. We included these reference genera to obtain some estimation of organisms which seem to have strong influences related to the spread of genes among these *B. anthracis,* or which might play important roles as intermediate gene flow agents, specifically linked to *B. thuringiensis*, which has been reported as the carrier for the pXO2 plasmid in the case for virulent strain *Bacillus anthracis str. A0248*. ***Figure 5.3 to 5.6*** shows the results of the Flux Visualizer execution steps.

The available plasmids from *B. thuringiensis* were used together with both *B. anthracis* strains to obtain a schematic of the OUP similarities among these organisms. Interestingly, we found all of the plasmids shared sequence similarities with *Bacillus anthracis str. A0248* and only one plasmid (*Bacillus thuringiensis serovar konkukian str. 97-27 plasmid,* NC_006578) shared sequence similarity to *Bacillus anthracis str. Sterne*. As per ***Figure 5.3***, this indicated possible restricted plasmid interaction to *Bacillus anthracis str. Sterne* (NC_005945) which might have played a role in the obstruction of the pXO2 capsule genes. However, this plasmid (NC_006578) was also found to uniquely contain the only BLAST match available in the database from the *Bacillus thuringiensis* species *to Bacillus anthracis str. 'Ames Ancestor' plasmid, pXO2* (NC_007323) as per ***Figure 5.4****.* The *'Ames Ancestor'* strain is considered to be the "gold standard" for virulent *B. anthracis* strains. It is therefore more likely that *Bacillus anthracis str. Sterne* (NC_005945) obtained the pXO2 genes at some stage, and subsequently experienced gene loss of the entire integron.


Flux Visualizer

Genomic island distribution and flow analysis

Legend

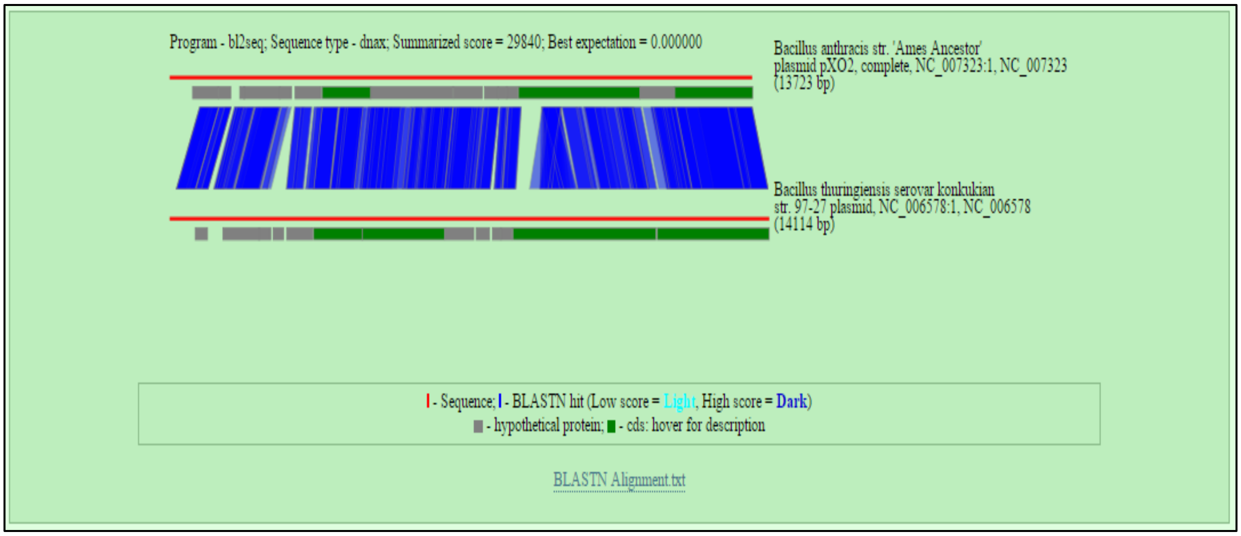
Bacillus anthracis

Bacillus thuringiensis

Group average pattern similarity: 83.18%

Sequence pattern similarity: 87.07%

Figure 5. 3 Flux Visualizer image of B. thuringiensis (yellow) plasmid gene fluxes to both Bacillus anthracis strains (red).



Program - bl2seq; Sequence type - dnax; Summarized score = 29840; Best expectation = 0.000000

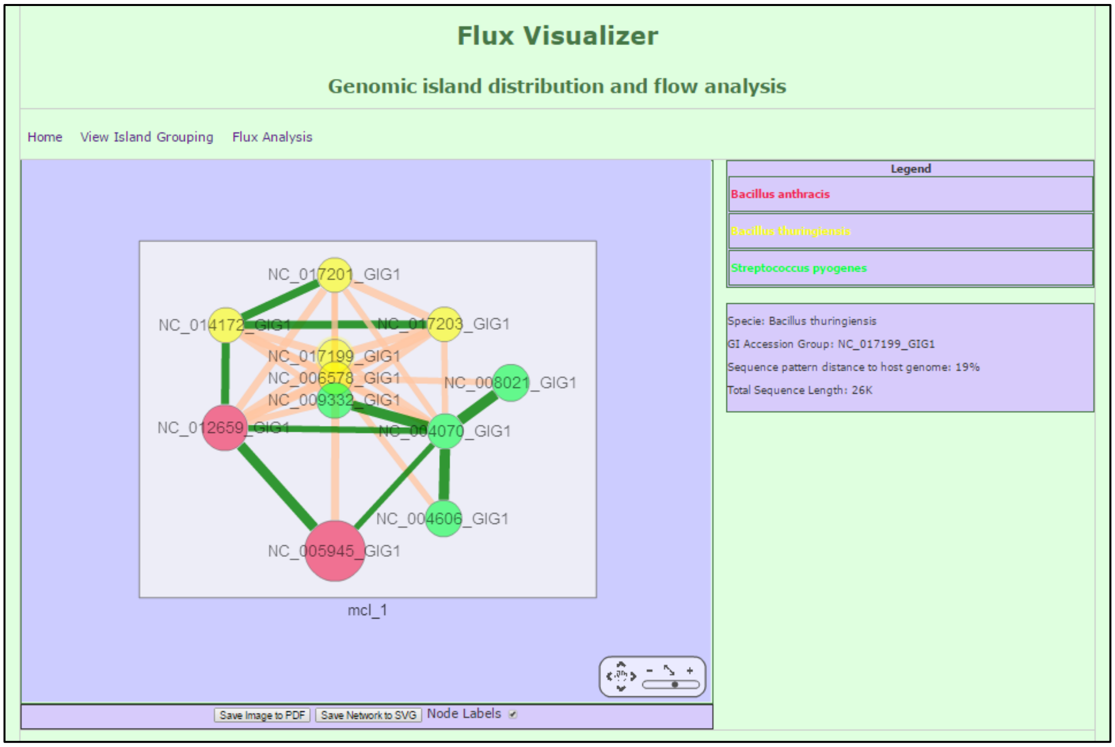
Bacillus anthracis str. 'Ames Ancestor'
plasmid pXO2, complete, NC_007323:1, NC_007323
(13723 bp)
Bacillus thuringiensis serovar konkukian
str. 97-27 plasmid, NC_006578:1, NC_006578
(14114 bp)

🟥 - Sequence; 🟦 - BLASTN hit (Low score = Light, High score = Dark)
⬛ - hypothetical protein; 🟩 - cds: hover for description

BLASTN Alignment.txt

Figure 5. 4 BLASTN visualization between Bacillus anthracis str. 'Ames Ancestor' plasmid, pXO2 (NC_007323) and Bacillus thuringiensis serovar konkukian str. 97-27 plasmid NC_006578)

The distribution of HGT between all available *Streptococcus pyogenes* showed sequence composition similarities to the two *B. anthracis* strains and the related *B. thuringiensis* plasmids as per **Figure 5.5**. In a similar manner we observed the distribution of *Staphylococcus aureus* and *Clostridium* HGT also centralized among the considered *Bacillus* organisms as per ***Figure 5.6 and 5.7**.* The gene sharing relationships for *Bacillus anthracis str. Sterne* in all these considered fluxes were seemingly disconnected to plasmid interactions and due to this limited connectivity, we may consider this strain as less compatible for HGT to the *B. thuringiensis* plasmids, compared to its virulent counterpart. However, since we did not observe any significant separation of the *Bacillus* organisms into second or third clusters, it remains unclear where the toxin gene cluster in pXO1 originated from. In contrast, we observed a number of BLAST confirmations (as denoted by the green highlighted lines) between *Bacillus anthracis* and afore mentioned families. Therefore HGT between these organisms did indeed occur at some stage and the specific gene clusters of the pXO1 plasmid probably consist of several gene sets from various donor families.



Legend

Bacillus anthracis

Bacillus thuringiensis

Streptococcus pyogenes

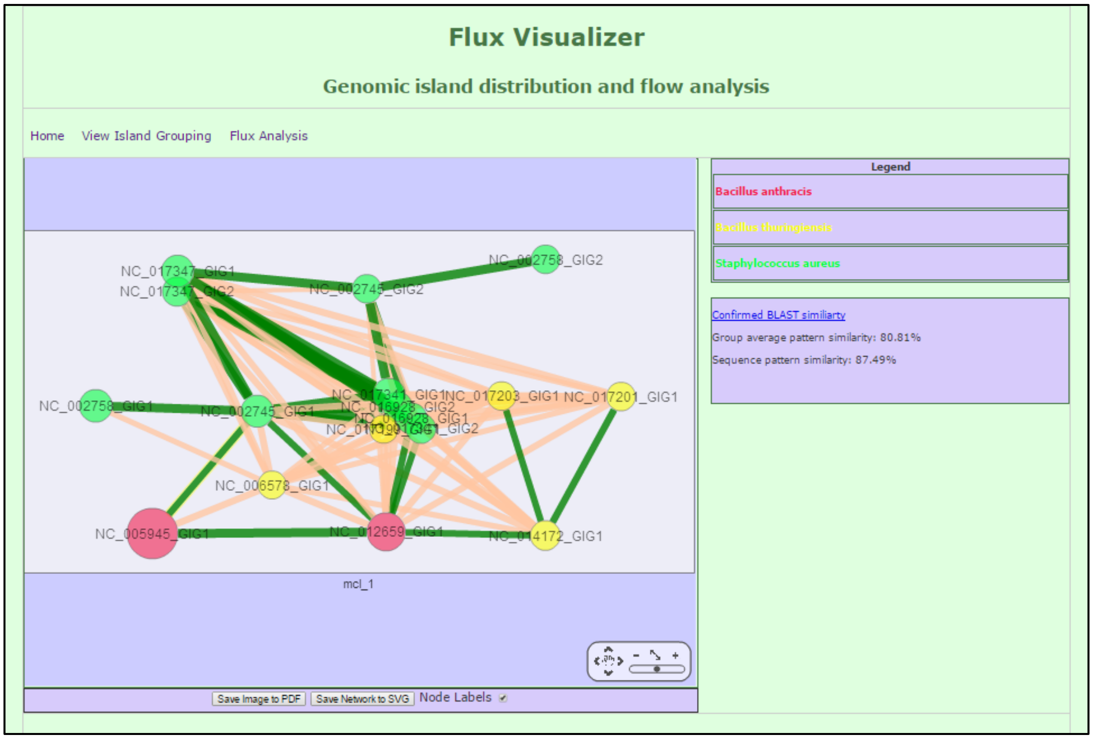
Specie: Bacillus thuringiensis

GI Accession Group: NC_017199_GIG1

Sequence pattern distance to host genome: 19%

Total Sequence Length: 26K

Figure 5. 5 Flux Visualizer image of B. thuringiensis (yellow) plasmid gene fluxes to Streptococcus pyogenes (green) and Bacillus anthracis strains (red).



Legend

Bacillus anthracis

Bacillus thuringiensis

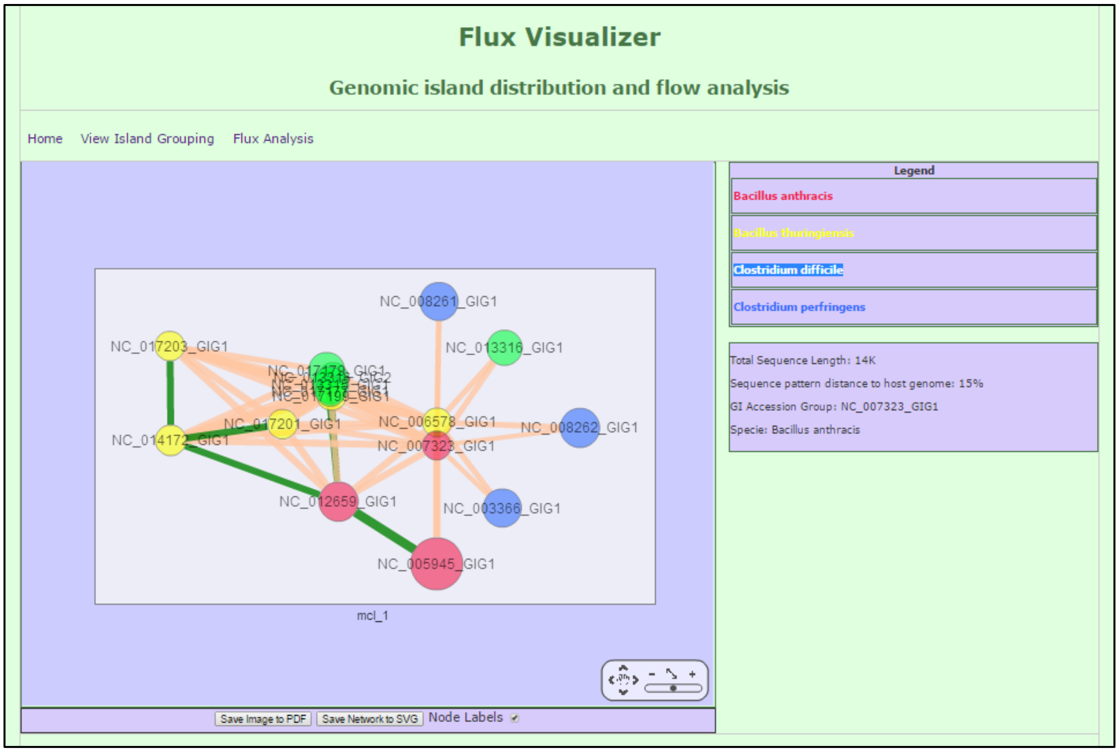
Staphylococcus aureus

Confirmed BLAST similiarty

Group average pattern similarity: 80.81%

Sequence pattern similarity: 87.49%

Figure 5. 6 Flux Visualizer image of B. thuringiensis (yellow) plasmid gene fluxes to Staphylococcus aureus (green) and Bacillus anthracis strains (red).



Flux Visualizer

Genomic island distribution and flow analysis

Legend

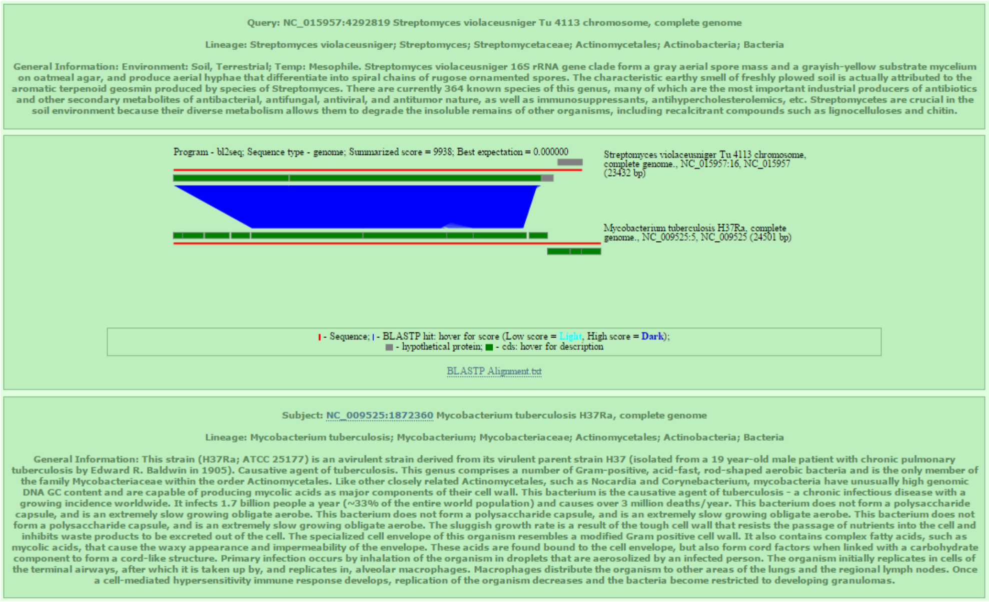
- Bacillus anthracis
- Bacillus thuringiensis
- Clostridium difficile
- Clostridium perfringens

Total Sequence Length: 14K
Sequence pattern distance to host genome: 15%
GI Accession Group: NC_007323_GIG1
Specie: Bacillus anthracis

Figure 5. 7 Flux Visualizer image of B. thuringiensis (yellow) plasmid gene fluxes to Clostridium difficile & perfringens (green & blue) and Bacillus anthracis strains (red).

5.2 Case study for Streptomyces & Mycobacterium

There are several known species of *Streptomyces*, many of which are important producers of industrial antibiotics. *Streptomyces* is also known as a prolific source of novel secondary metabolites which has impacts on wide range of biological activities, including antibiotic resistances (Emerson, et al., 2012). Secondary metabolisms offer the ability to stimulate production of secondary metabolites which are not critically required for the survival of the organism. *Streptomyces* are considered to have crucial influences on soil environments due to their capacity to degrade the insoluble remains of other organisms. In contrast, *Mycobacterium* includes pathogens known to cause serious and often fatal diseases in mammals, including tuberculosis. Mycobacterial infections are particularly difficult to treat as the bacteria are naturally resistant to a number of antibiotics and they can survive long periods of exposure to various harsh environments (Feltcher, et al., 2010). Since these two families have shown to share a significant number of BLAST matches between their GI groups, we postulate that *Streptomyces* may serve as a potential reservoir of genes, related to the secondary metabolisms of these *Mycobacteria.* ***Figure 5.8*** shows an example of BLASTP visualization between *Streptomyces violaceusniger Tu 4113* (NC_015957) and *Mycobacterium tuberculosis H37Ra* (NC_009525), the attenuated mutant of the virulent strain *H37Rv.*



Query: NC_015957:4292819 Streptomyces violaceusniger Tu 4113 chromosome, complete genome

Lineage: Streptomyces violaceusniger; Streptomyces; Streptomycetaceae; Actinomycetales; Actinobacteria; Bacteria

General Information: Environment: Soil, Terrestrial; Temp: Mesophile. Streptomyces violaceusniger 16S rRNA gene clade form a gray aerial spore mass and a grayish-yellow substrate mycelium on oatmeal agar, and produce aerial hyphae that differentiate into spiral chains of rugose ornamented spores. The characteristic earthy smell of freshly plowed soil is actually attributed to the aromatic terpenoid geosmin produced by species of Streptomyces. There are currently 364 known species of this genus, many of which are the most important industrial producers of antibiotics and other secondary metabolites of antibacterial, antifungal, antiviral, and antitumor nature, as well as immunosuppressants, antihypercholesterolemics, etc. Streptomycetes are crucial in the soil environment because their diverse metabolism allows them to degrade the insoluble remains of other organisms, including recalcitrant compounds such as lignocelluloses and chitin.

Program - bl2seq; Sequence type - genome; Summarized score = 9938; Best expectation = 0.000000

Streptomyces violaceusniger Tu 4113 chromosome, complete genome., NC_015957:16, NC_015957 (23432 bp)

Mycobacterium tuberculosis H37Ra, complete genome., NC_009525:5, NC_009525 (24501 bp)

🟥 - Sequence; 🟦 - BLASTP hit: hover for score (Low score = Light, High score = Dark);
■ - hypothetical protein; 🟩 - cds: hover for description

BLASTP Alignment.txt

Subject: NC_009525:1872360 Mycobacterium tuberculosis H37Ra, complete genome

Lineage: Mycobacterium tuberculosis; Mycobacterium; Mycobacteriaceae; Actinomycetales; Actinobacteria; Bacteria

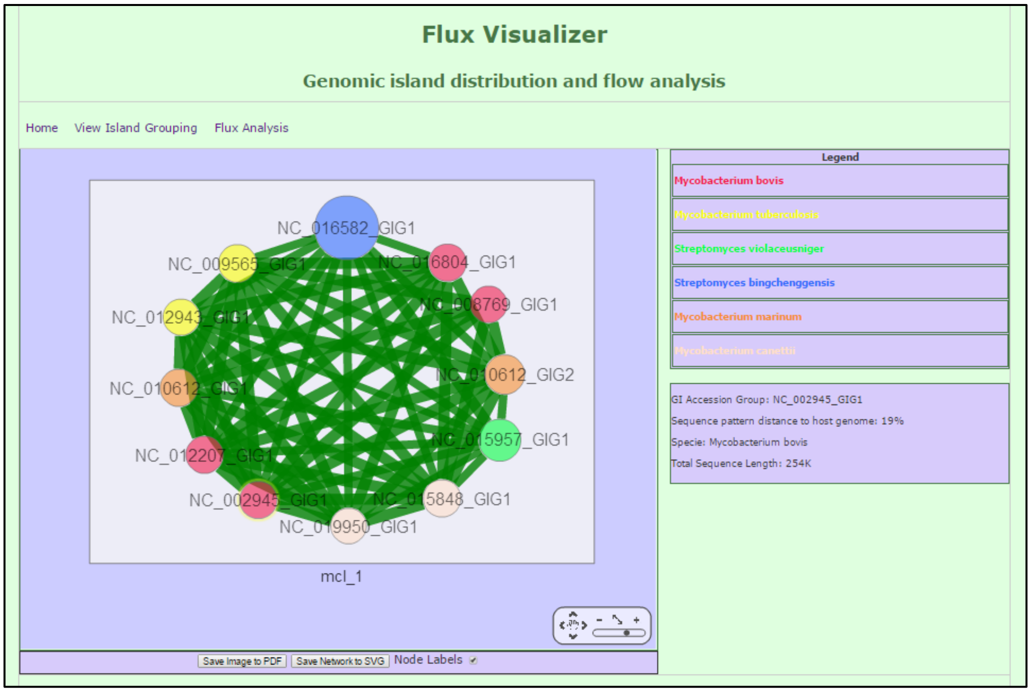
General Information: This strain (H37Ra; ATCC 25177) is an avirulent strain derived from its virulent parent strain H37 (isolated from a 19 year-old male patient with chronic pulmonary tuberculosis by Edward R. Baldwin in 1905). Causative agent of tuberculosis. This genus comprises a number of Gram-positive, acid-fast, rod-shaped aerobic bacteria and is the only member of the family Mycobacteriaceae within the order Actinomycetales. Like other closely related Actinomycetales, such as Nocardia and Corynebacterium, mycobacteria have unusually high genomic DNA GC content and are capable of producing mycolic acids as major components of their cell wall. This bacterium is the causative agent of tuberculosis - a chronic infectious disease with a growing incidence worldwide. It infects 1.7 billion people a year (~33% of the entire world population) and causes over 3 million deaths/year. This bacterium does not form a polysaccharide capsule, and is an extremely slow growing obligate aerobe. This bacterium does not form a polysaccharide capsule, and is an extremely slow growing obligate aerobe. This bacterium does not form a polysaccharide capsule, and is an extremely slow growing obligate aerobe. The sluggish growth rate is a result of the tough cell wall that resists the passage of nutrients into the cell and inhibits waste products to be excreted out of the cell. The specialized cell envelope of this organism resembles a modified Gram positive cell wall. It also contains complex fatty acids, such as mycolic acids, that cause the waxy appearance and impermeability of the envelope. These acids are found bound to the cell envelope, but also form cord factors when linked with a carbohydrate component to form a cord-like structure. Primary infection occurs by inhalation of the organism in droplets that are aerosolized by an infected person. The organism initially replicates in cells of the terminal airways, after which it is taken up by, and replicates in, alveolar macrophages. Macrophages distribute the organism to other areas of the lungs and the regional lymph nodes. Once a cell-mediated hypersensitivity immune response develops, replication of the organism decreases and the bacteria become restricted to developing granulomas.

Figure 5. 8 The BLASTP visualizations between Streptomyces violaceusniger Tu 4113 (NC_015957) and Mycobacterium tuberculosis H37Ra (NC_009525).

We evaluated the highest ranking *Mycobacterium* and *Streptomyces* GI group combinations as per **Table 5.1**. Only two organisms from *Streptomyces* were found to share a significant amount of genes to the rest of the *Mycobacterium* GI groups, namely *Streptomyces violaceusniger Tu 4113* (NC_015957) and *Streptomyces bingchenggensis BCW-1* (NC_016582). The references from these organisms were used as input to the Flux Visualizer system. As per **Figure 5.9**, we observed extreme cases of HGT with confirmed BLAST similarities between all of the included organisms. Upon closer inspection we observed that the BLAST hits between these organisms related exclusively to genes appointed to secondary metabolite biosynthesis as per **Table 5.2**. The gene descriptions were mainly assigned to 'Polyketide synthase modules and related proteins'. This suggested that these two *Streptomyces* strains served as active HGT hotspots to the *Mycobacterium* species, specifically for genes related to secondary metabolite biosynthesis.

No BLAST Similarities	GI Group Combinations
49	Streptomyces violaceusniger - Mycobacterium marinum
42	Streptomyces violaceusniger - Mycobacterium bovis
40	Streptomyces violaceusniger - Mycobacterium marinum
38	Streptomyces bingchenggensis - Mycobacterium marinum
35	Streptomyces violaceusniger - Mycobacterium tuberculosis
34	Streptomyces violaceusniger - Mycobacterium canettii
33	Streptomyces bingchenggensis - Mycobacterium bovis
32	Streptomyces violaceusniger - Mycobacterium canettii
32	Streptomyces bingchenggensis - Mycobacterium marinum
31	Streptomyces violaceusniger - Mycobacterium bovis
31	Streptomyces violaceusniger - Mycobacterium bovis
31	Streptomyces violaceusniger - Mycobacterium bovis
31	Streptomyces violaceusniger - Mycobacterium tuberculosis
31	Streptomyces violaceusniger - Mycobacterium tuberculosis
25	Streptomyces bingchenggensis - Mycobacterium tuberculosis

Table 5. 1 No of BLAST similarities among GI group combinations of Mycobacteria and Streptomyces.



GI Accession Group: NC_002945_GIG1
Sequence pattern distance to host genome: 19%
Specie: Mycobacterium bovis
Total Sequence Length: 254K

Figure 5. 9 Extreme HGT confirmed by BLAST similarities (green lines) from the Flux Visualization of secondary metabolite genes shared between Mycobacterium and Streptomyces.

Streptomyces Accession	Mycobacterium Accession	Functional Group	No Blast Matches
NC_015957	NC_012207	Secondary metabolites	37
NC_015957	NC_008769	Secondary metabolites	37
NC_015957	NC_002945	Secondary metabolites	37
NC_015957	NC_019950	Secondary metabolites	38
NC_015957	NC_015848	Secondary metabolites	39
NC_015957	NC_009565	Secondary metabolites	42
NC_015957	NC_016804	Secondary metabolites	51
NC_015957	NC_010612	Secondary metabolites	108
NC_016582	NC_010612	Secondary metabolites	83
NC_016582	NC_012943	Secondary metabolites	30
NC_016582	NC_012207	Secondary metabolites	29
NC_016582	NC_008769	Secondary metabolites	29
NC_016582	NC_002945	Secondary metabolites	29
NC_016582	NC_019950	Secondary metabolites	28
NC_016582	NC_015848	Secondary metabolites	27

Table 5. 2 Blast matches between Streptomyces and Mycobacterium reveal exclusive functional preferences to Secondary metabolites

5.3 Discussion

The evaluation of BLAST match frequencies between all available GI group combinations revealed a large proportion of organisms sharing only one gene between their groups. These genes were distributed mainly among functional classifications for HGT transference and integration, with over-represented COG classes among “Transcription”, “Translation” and “Nucleotide transport and metabolism”. In contrast, we found examples of extreme gene sharing indices among GI group combinations, specifically for two *B. anthracis* strains, due to their similar GI recombination loci. As mentioned, this may be expected because these organisms share very similar evolutionary backgrounds. Nevertheless, using the Flux Visualizer system we were able to conclude that the HGT relationships between them were shaped by parallel donors and at some time the avirulent *Sterne* strain, underwent a major gene loss event (probably to ensure other genetic material could be integrated, if you take into account the size of the *Bacillus anthracis str. Sterne* GI group). Furthermore, we showed that, from two *Streptomyces* strains, genes related to secondary metabolisms were primarily shared by many *Mycobacteria*, where the aforementioned were used as a reservoir of genes to provide novel survival mechanisms for the various harsh environments of the *Mycobacteria*. Here the Flux Visualizer has proved to be valuable in obtaining an understanding of the possible gene spread outcomes, and facilitated meaningful results necessary to understand the gene sharing relationships in both sets of organisms in this case study. Both of these candidate sets were evaluated and subjected to analyses with the Flux Visualizer system to reveal the roles of HGT among them. This was set out as one of the major goals to achieve for implementation of this system (as per Chapter 1). Hence, we successfully showed that through the evaluation and grouping of fragments of GIs, we were able to simulate the HGT relationship structures through the use of graph theory and clustering techniques, among a multitude of bacterial organisms.

Concluding remarks

The Flux Visualizer website is a tool aimed at visualizing bacterial genetic vector distributions through analyzing GI compositional similarities, GI fragment BLAST matches and shared gene functional classifications. The following achievements and discoveries resulted from this project:

1. We have investigated the level of fragmentation of GIs in the Pre_GI database. The working hypothesis was that the number of events of horizontal gene transfer were overestimated in the Pre_GI database due to fragmentation of GIs. This study demonstrated that many GIs in bacterial genomes shared significant levels of compositional similarity and were either acquired independently at the same time from a single source, or, that which is more likely, were transferred all together within a genetic vector (conjugative plasmid of phage) and fragmented in the bacterial chromosome. An attempt was made to group available GIs from the Pre_GI database to estimate the real number of HGT events and to enable more accurate evaluations of the impact of GI acquisitions on bacterial evolution by the analysis of the functions of the acquired genes. Number of GI groups per genome varied from 1 to 10 following Poisson distribution (see ***Figure. 2.4***)
2. A system for visualization of HGT events based on oligonucleotide usage similarities was successfully implemented as set out by the second goal of this project. The web address for the Flux Visualizer system is http://flux.bi.up.ac.za. Here users may use two links from the landing page of the website to navigate to the separate interfaces. The first interface allows the users to explore the grouping of GIs per organism, and the second interface allows the users to visualize donor-recipient relationships based on the user's selections of interest as shown in Chapter 3.
3. Annotations of all the identified GIs from the Pre_GI database were checked and the COG category descriptions were assigned to their genes. Then, using Markov Clustering we allocated groups of GIs in such a fashion as to represent accurate gene sharing structures between recipient organisms. We showed that GI groups could be separated based on genetic utility from the functional categories of the COG database as set out in the third goal of this project. This revealed that the COG categories combinations of NQ, NU, NV, INQ and MNO may be viewed as trademarks for HGT characterized by virulence gene exchanges. Additionally, we analyzed gene networks based on successive gene BLAST matches per COG functional category to formulate suggestions to include in the users analyses. Users of the Flux Visualizer website are therefore

able to include selections based on their own interests or suggested inclusions per specified functional category.

4. The system developed for this project aids in finding patterns of gene exchanges among a large number of diverse bacterial species. This may prove valuable in identifying pathogenic gene spreads and assist in predicting when and where virulence gene acquisitions occur. As per this projects' last goal, we demonstrated the successful use of the system and further extrapolated that it might be used for specific organisms of interest as in the case of the *B. Anthracis.* We also showed that the visualization of HGT relationships in this way, assists in painting a fuller picture, as was evident in the case for the wealth of secondary metabolism genes which were provided by the *Streptomyces* strains to various *Mycobacterium* species*.*

Results of this project were summarized and prepared as a manuscript for submission to PloS One journal.

Bibliography

Alsmark, C. et al., 2013. Patterns of prokaryotic lateral gene transfers affecting parasitic microbial eukaryotes. *Genome Biology,* 14(19).

An Diep, B. et al., 2006. Roles of 34 Virulence Genes in the Evolution of Hospital- and Community-Associated Strains of Methicillin-Resistant Staphylococcus aureus. *Virulence Genes and MRSA,* 193, pp. 1495 - 1503.

Ashburner, M. et al., 2000. Gene Ontology: tool for the unification of biology. *Nature Genetics,* 25, pp. 25-29.

Baldan, R. et al., 2012. Factors Contributing to Epidemic MRSA Clones Replacement in a Hospital Setting. *PLOS ONE,* 7(8).

Becq, J., Churlaud, C. & Deschavanne, P., 2010. A Benchmark of Parametric Methods for Horizontal Transfers Detection. *PLoS ONE,* 5(4).

Bezuidt, O. et al., 2011. Mainstreams of Horizontal Gene Exchange in Enterobacteria: Consideration of the Outbreak of Enterohemorrhagic E. coli O104:H4 in Germany in 2011. *PLoS ONE,* 6(10).

Bondy, J.A., & Murty, U. S. R, 2008. Graph Theory with applications. Springer.

Boyd, F. E., Almagro-Moreno, S. & Parent, M. A., 2009. Genomic islands are dynamic, ancient integrative elements in bacterial evolution. *Trends in Microbiology,* 17(2), pp. 47-53.

Brandes, U., Gaertler, M. & Wagner, D., 2008. Engineering Graph Clustering: Models and Experimental Evaluation. *ACM Journal of Experimental Algorithmics,* 12(1.1).

Chaffron, S., Rehrauer, H., Pernthaler, J. & von Mering, C., 2010. A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Research,* 20, pp. 947–959.

Deschavanne, P. J. et al., 1999. Genomic Signature: Characterization and Classification of Species Assessed by Chaos Game Representation of Sequences. *Mol. Biol. Evol,* 16(10), pp. 1391-1399.

Devault, A. M., Golding, B., Waglechner, N. & Enk, J. M., 2014. Second-Pandemic Strain of Vibrio cholerae from the Philadelphia Cholera Outbreak of 1849. *The New England Journal of Medicine,* 370, pp. 334-340.

Djordjevic, S. P., Stokes, H. W. & Chowdhury, P. R., 2013. Mobile elements, zoonotic pathogens and commensal bacteria: conduits for the delivery of resistance genes into humans, production animals and soil microbiota. *Frontiers in Microbiology,* 4(86).

Emerson, R., de Lima, P. et al., 2012. Antibiotics produced by Streptomyces. *The Brazilian Journal of INFECTIOUS DISEASES,* 16(5), pp. 466–471.

Encinas, D. et al., 2014. Plasmid Conjugation from Proteobacteria as Evidence for the Origin of Xenologous Genes in Cyanobacteria. *Journal of Bacteriology,* 196(8).

Feltcher, M. E., Sullivan, J. T. & Braunstein, M., 2010. Protein export systems of Mycobacterium tuberculosis: novel targets for drug development?. *Future Microbiology,* 5, pp. 1581–1597.

Fernández-Gómez, B. et al., 2012. Patterns and architecture of genomic islands in marine bacteria. *BMC Genomics,* 13(347).

Foggia, P., Percannella, G., Sansone, C. & Vento, M., 2008. Benchmarking graph-based clustering algorithms. *Image and Vision Computing,* 27, pp. 979–988.

Gatica , J. & Cytryn, E., 2013. Impact of treated wastewater irrigation on antibiotic resistance in the soil microbiome. *Environ Sci Pollut Res,* 20, pp. 3529–3538.

Goldsmith, C.E., Yap, J.M., Moore, J.E., 2013. Integrity of bacterial genomic DNA after autoclaving: possible implications for horizontal gene transfer and clinical waste management. *Journal of Hospital Infection*, Volume 83, pp. 247 - 249.

Guzzi, P. H. & Cannataro, M., 2012. *CytoMCL:* a Cytoscape plugin for fast Clustering of Protein Interaction Networks*, Proc. 25th International Symposuim on Computer-Based Medical Systems (CBMS)*.

Hamilton , G. et al., 2008. The SeqWord Genome Browser: an online tool for the identification and visualization of atypical regions of bacterial genomes through oligonucleotide usage. *BMC Bioinformatics,* 9(333).

Hao, W. & Golding, B., 2006. The fate of laterally transferred genes: Life in the fast lane to adaptation or death. *Genome Research,* 16, pp. 636–643.

Hao, W. & Golding, B., 2009. Does Gene Translocation Accelerate the Evolution of Laterally Transferred Genes?. *Genetics Society of America,* 182, pp. 1365–1375.

Ho Sui, S. J. et al., 2009. The Association of Virulence Factors with Genomic Islands. *PLoS ONE,* 4(12).

Jane, W. & Frederick, C. M., 2011. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS Microbiology Reviews,* 35, pp. 957–976.

Karlin, S., 2001. Detecting anomalous gene clusters and pathogenicity islands in diverse in bacterial genomes. *TRENDS in Microbiology,* 9(7), pp. 335 - 343.

Kloesges, T., Popa, O., Martin, W. & Dagan, T., 2011. Networks of Gene Sharing among 329 ProteobacterialGenomes Reveal Differences in Lateral Gene Transfer Frequency at Different Phylogenetic Depths. *Mol. Biol. Evol.,* 2(28), pp. 1057–1074.

Koehler, T. M., 2002. Bacillus anthracis genetics and virulence gene regulation.. *Current Topics in Microbiology and Immunology.,* 271, pp. 143-164.

Langille, G. M. & Brinkman, F. S., 2009. IslandViewer: An Integrated Interface for Computational Identification and Visualization of Genomic Islands. *Bioinformatics,* 5(25), pp. 664–665.

Langille, M. G. I., Hsiao, W. W. L. & Brinkman, F. S. L., 2010. Detecting genomic islands using bioinformatics approaches. *Nature Reviews Microbiology,* 8, pp. 373 - 382.

Lefeuvre, P. et al., 2013. Constraints on Genome Dynamics Revealed from gene distribution among Ralstonia solanacearum species. *PLOS ONE,* 8(5).

Makarova, K. S., Wolf, Y. I. & Koonin, E. V., 2013. Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Research,,* 41(8), pp. 4360–4377.

Martin, W., 1999. Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. *BioEssays,* 2(21), pp. 99-104.

Merkl, R., 2006. A Comparative Categorization of Protein Function Encoded in Bacterial or Archeal Genomic Islands. *Journal of Molecular Evolution,* 62, pp. 1-14.

Morris, J. H. et al., 2011. clusterMaker: a multi-algorithm clustering plugin for Cytoscape. *BMC Bioinformatics,* 436(12).

Newman, M., 2010. *Networks An Introduction,* New York: Oxford University Press.

Okinaka, R. et al., 1999. Sequence and Organization of pXO1, the Large Bacillus anthracis Plasmid Harboring the Anthrax Toxin Genes. *Journal of Bacteriology,* 181(20), pp. 6509–6515.

Park, C. & Zhang, J., 2012. High Expression Hampers Horizontal Gene Transfer. *Genome Biol. Evol,* 4(4), pp. 523–532.

Pavlovic, G. et al., 2004. Evolution of genomic islands by deletion and tandem accretion by site-specific recombination: ICESt1-related elements from Streptococcus thermophilus. *Microbiology,* 150, pp. 759–774.

Podell, S., Gaasterland, T. & Allen, E. E., 2008. A database of phylogenetically atypical genes in archaeal and bacterial genomes, identified using the DarkHorse algorithm. *BMC Bioinformatics, 419(9).*

Popa, O. & Dagan, T., 2011. Trends and barriers to lateral gene transfer in prokaryotes. *Current Opinion in Microbiology,* 14, pp. 615–623.

Popa, O. et al., 2011. Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Research,* 21, pp. 599–609.

Pride, D. T., Meinersmann, R. J., Wassenaar, T. M. & Blaser, M. J., 2003. Evolutionary Implications of Microbial Genome Tetranucleotide Frequency Biases. *Genome Research,* Volume 13, pp. 145-157.

Reva, O. N. & Tümmler, B., 2005. Differentiation of regions with atypical oligonucleotide composition in bacterial genomes. *BMC Bioinformatics,* 6(251).

Roos, T. E. & van Passel, MWJ., 2011. A quantitative account of genomic island acquisitions in prokaryotes. *BMC Genomics,* 12(427).

Sabino, R. et al., 2011. Isolates from hospital environments are the most virulent of the Candida parapsilosis complex. *BMC Microbiology,* 11(180).

Seiffert, S. N., Hiltya, M., Perretenb, V. & Endimiania, A., 2013. Extended-spectrum cephalosporin-resistant gram-negative organisms in livestock: An emerging problem for human health?. *Drug Resistance Updates,* 16, pp. 22– 45.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. Ideker T., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research, 13(11), pp. 2498 - 2504.*

Shen, S. et al., 2004. Evidence for a Hybrid Genomic Island in Verocytotoxin-Producing Escherichia coli CL3 (Serotype O113:H21) Containing Segments of EDL933 (Serotype O157:H7) O Islands 122 and 48. *Infection and immunity,* 3(72), pp. 1496–1503.

Shuman, S. & Glickman, M. S., 2007. Bacterial DNA repair by non-homologous end joining. *Nature Reviews Microbiology,* 5(November), pp. 852 - 861.

Smets, B. F. & Barkay, T., 2005. Horizontal Gene Transfer: Perspectives at a crossroad of scientific disciplines. *Nature Reviews Microbiology,* 3(September), pp. 675 - 678.

Tamminen, M., Virta, M., Fani, R. & Fondi, M., 2012. Large-Scale Analysis of Plasmid Relationships through Gene-Sharing Networks. *Mol. Biol. Evol.,* 29(4), pp. 1225–1240.

Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V., 2000. The COG Database: a tool for genome scale analysis of protien fuctions and evolution. *Nucleic Acids Research,* 28(1), pp. 33-36.

Thi Le, P., Pontarotti, P. & Raoult, D., 2014. Alphaproteobacteria species as a source and target of lateral sequence transfers. *Trends in Microbiology,* 22(3), pp. 147.

Thomas, C. M. & Nielsen, K. M., 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature Reviews MICROBIOLOGY,* 3(September), pp. 711 - 721.

Ulrich, D., Jorg, H., Ute, H. & Bianca, H., 2004. Genomic islands in pathogenic and environmental microorganisms. *Nature Reviews Microbiology.,* 2.5(May), pp. 414 - 424.

Van Dogen, S. M., 2000. *Graph clustering by flow simulation, PhD Thesis, Utrecht University.*

Wiedenbeck, J., Frederick C.M., 2011. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *Federation of European Microbiological Societies*, Volume 35, pp.957 - 976.

Wiezera, A. & Merklb, R., 2005. A comparative categorization of gene flux in diverse microbial species. *Genomics,* 86 , pp. 462 – 475.

Williams, K. P., 2002. Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nucleic Acids Research,* 30(4), pp. 866 - 875.

Y. J., Goldsmith , C. & Moore, . J., 2013. Integrity of bacterial genomic DNA after autoclaving: possible implications for horizontal gene transfer and clinical waste management. *Journal of Hospital Infection,* 83, pp. 247-249.

Zhang, W., Wang, X., Zhao, D. & Tang, X., 2012. Graph Degree Linkage: Agglomerative Clustering on a Directed Graph. In: Computer Vision – ECCV 2012, pp. 428-441.

Zhao, E.Y., 2013. Genomic comparison of Salmonella typhimurium DT104 with non-DT104 strains. *Molecular Genetics Genomics,* 288, pp. 549–557.

Zhou, C. E. et al., 2007. MvirDB—a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. *Nucleic Acids Research,* Volume 35, pp. 391 - 394.

Zongfu, W., Weixue, W. & Min, T., 2013. Comparative genomic analysis shows that Streptococcus suis meningitis isolate SC070731 contains a unique 105 K genomic island. *Gene,* 535, pp. 156–164.