Development of novel computational tools to infer the distribution patterns of bacterial accessory genomic elements and the implications of microevolution towards pathogenicity

by

Keoagile Ignatius Oliver Bezuidt

Submitted in partial fulfillment of the requirements for the degree

Philosophiae Doctor

in the

Bioinformatics and Computational Biology Unit

Department of Biochemistry

Faculty of Natural and Agricultural Sciences

University of Pretoria

Pretoria

September 2013

I, Keoagile Ignatius Oliver Bezuidt, declare that the thesis/dissertation, which I hereby submit for the degree PhD(Bioinformatics) in the Department of Biochemistry, at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

SIGNATURE:_____ DATE

Acknowledgements

This thesis is dedicated to the memories of my mother and cousin, J.D Mogase and B.M Mogase, respectively.

Special thanks are due to the following:

my supervisor, Prof Oleg Reva for his support and help to complete this work.

my co-supervisor Prof Burkhard Tuemmler for agreeing to supervise part of my work despite his many other professional commitments. His invaluable guidance and input together with the support for my visits to Germany are much appreciated.

Drs Jens Klockgether and Colin Davenport for their patience, help with the analysis, and being part of the weekly lengthy meetings with Burkhard.

the Clinical Research Group at Medizinische Hochschule Hannover, in particular Lutz Wiehlmann and Frauke Stanke.

fellow colleagues and students at the BCBU, in particular Fourie Joubert, Charles Hefer, Pieter Burger, Nanette Coetzer and Jeanre Smit.

my brothers, Clive and Juries for their support and best wishes.

my lovely Tiang for her support, encouragement, constant patience and putting up with me while I was completing this work.

the German Academic Exchange Service (DAAD), the National Research Foundation (NRF) and the Deutsche Forschungsgemeinschaft (SFB) for financial support.

Summary

Bacterial diversity has always been associated with micro-evolutionary events such as horizontal gene transfer and DNA mutations. Such events influence the rapid evolution of bacteria as a result of the environmental conditions which they encounter. They further establish beneficial phenotypic effects that allow bacteria to specialize in new habitats. Due to the increase in number of bacterial genomic sequences, studying microbial evolution has been made possible, and the impact of micro-evolution on bacterial diversity is becoming more apparent. To gain biological information from this ever increasing genomic data, a variety of computational tools are required. This thesis therefore, focuses on the development and application of computational approaches to identify genomic regions of divergence which have resulted from horizontal gene transfer or small mutational changes. The first and major part of the thesis describes the application of DNA patterns, termed oligonucleotide signatures to identify horizontally acquired genomic regions in prokaryotes. These DNA patterns are demonstrated to differentiate between signatures of the core genome and those which have been acquired through horizontal transfer events. DNA patterns are further demonstrated to: reveal the distribution patterns of horizontally acquired genomic elements, determine their acquisition periods, and predict their putative donor organisms. The second part of the thesis focuses on the evaluation of modern short read sequence data of geographically unrelated *Pseudomonas aeruginosa* to study their intraclonal genomic diversity. The work described in the thesis was purely *in silico* driven and performed at Hannover Medical School and the Bioinformatics and Computation Biology Unit at the University of Pretoria.

Contents

List of figures

List of Tables

List of Abbreviations

BLAST	Basic Local Alignment Search Tool
Bp	Nucleotide base pair
CAI	Codon Adaptation Index
CF	Cystic Fibrosis
CGR	Chaos Game Representation
Contigs	A multiple alignment of reads converted into a long contiguous genomic sequence
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CUTG	Codon Usage Tabulated from GenBank
D	Distance
DNA	Deoxyribonucleic Acid
DR	Direct Repeat
FASTA	text-based format for representing either nucleotide sequences or peptide sequences, in which nucleotides or amino acids are represented using single-letter codes
FNR	False Negative Rate
FNR	False negative rate
FPR	False Positive Rate
FPR	False positive rate
GC	Guanine and Cytocine nucleotide bases
Genbank	A rich format for storing sequences and associated annotations
GI	Genomic Island
GOLD	Genomes Online Database
GRV	Global Relative Variance
HGT	Horizontal Gene Transfer
HMM	Hidden Markov Models
Indel	Insertion or deletion of bases in the DNA of an organism
IS	Insertion Sequences
kbp	kilobase pair
k-mer	A word size of length *k*
LEE	Locus of Enterocyte Effacement
MGE	Mobile Genetic Elements
NCBI	National Center for Biotechnology Information

OU	Oligonucleotide Usage
PAI	Pathogenicity Island
PS	Pattern skew
Reads	DNA string of base pairs
RGP	Region of Genomic Plasticity
RNA	Ribonucleic Acid
rrn	Ribosomal RNA operon
rRNA	Ribosomal Ribonucleic Acid
RV	Relative Variance
SAGI	Staphylococcus aureus Genomic Island
SIGI-HMM	Score-based Identification of Genomic Islands using Hidden Markov Models
SNP	Single Nucleotide Polymorphism
SPI	Salmonella Pathogenicity Island
SVG	Scalable vector graphics
SWGIS	SeqWord Genomic Islands Sniffer
Tet	Tetracycline
tmRNA	Transfer-messenger RNA
Tn	Transposon
tRNA	Transfer Ribonucleic Acid
V	Variance
VCF	Variant Call Format

Chapter 1

1 Modern comparative and composition-based approaches – the state of the art

1.1 Introduction

Following the publication of the first two complete genome sequences of *Haemophilus influenza* (Fleischmann *et al.*, 1995) and *Mycoplasma genitalium* (Fraser *et al.*, 1995) in 1995, a total of 6342 bacteria have reportedly been completely sequenced and become publicly available in the Genomes OnLine Database (GOLD), accessed September the 5th 2013 (Kyrpides N. C., 1999). The increase in numbers of sequenced bacterial genomes from 2 in 1995 to 6342 in 2013 has been made possible by the advances in high-throughput sequencing technologies i.e. Roche-454 (Margulies *et al.*, 2005), Solexa – Illumina (Fedurco *et al.*,2006), ABI SOLiD (Shendure *et al.*, 2005) and Pacific Biosciences (Eid *et al.*, 2009). The high-throughput projects have provided fast and cost effective methodologies that allow sequencing of large DNA stretches which span entire genomes in a single run. These advancements indicate how the concept of genome sequencing has evolved ever since the times of Gilbert & Maxam (1973) and Sanger & Coulson (1975), when researchers were only interested in getting single runs that allowed reading of 200-300 nucleotides. During the conventional sequencing days, sequencing was costly and a single bacterial genome sequence was delivered in months or even years (Loman *et al.*, 2012).

The relatively inexpensive sequencing technologies have been producing large quantities of data by the day and opened up new perspectives in the field of genomics (Metzker, 2005). In addition to the reported number of completely sequenced genomes as mentioned above, 12189 more bacterial genomes are currently indicated as ongoing in GOLD, accessed September the 5th 2013 (Kyrpides N. C., 1999). The numbers mentioned earlier do not include the entire amount of available bacterial genomes and those of ongoing sequencing projects taking place in the other research facilities, as many do not make it onto GOLD. These great numbers of bacteria are sequenced in order to study processes by which they evolve and understand their range of functions. Several of these bacteria are used to study their virulence properties towards humans, plants and animals and also their importance in industrial use i.e. bioremediation. Metagenomics, the study that focuses on the genomic

analysis of genetic material recovered directly from communities in environmental samples also contributes towards the increase in number of sequenced bacteria and make it possible to study differences in genomes of organisms which are difficult to culture. Collectively these genomes are beneficial for comparative genomics studies as they make it feasible to study large groups of (related and unrelated) bacteria with the aim of providing a broader picture of genomic diversity. It is generally believed that the important information which underlies the differences between highly virulent and avirulent strains may be read in their genomes as these may be due to DNA replication and repair based events (Karlin *et al.*, 1997). Such analysis may be conducted by utilizing the available data to study and compare sets of strains and variants of the same species to get insight into their genomic evolutionary changes and identify factors which influence pathogenicity, persistence in host environments and, drug resistance. These also allow studying genomes in a comparative approach in order to understand the evolutionary events which are the most important on the level of micro-evolution. This review chapter will provide a brief overview of analytical approaches and studies conducted on the vast amounts of completely sequenced bacteria to gain insight into their genomic diversity and adaptation to different environments with the influence of microevolutionary events.

1.2 Bacterial microevolution and fitness traits

Bacteria are single celled microorganisms that inhabit a wide range of environmental niches. They can be found almost everywhere in the environment: in the air, soil, great depths of the sea, plants and intestinal tracts of animals and humans (Binnewies *et al.*, 2006). The factors and lifestyle complexities which allow bacteria to survive in such a variety of environments are captured in their relatively small genomes, of sizes typically between 0.6 to over 10Mbp (Binnewies *et al.*, 2006).The current high-throughput technologies have provided a platform to study the dynamics and mechanisms of microbial diversity in various habitats. Comparative analyses projects between different genomes have uncovered genetic variability - a factor known to be an important prerequisite for biological evolution in microbial lifestyles and diversification (Schloter *et al.*, 2000). This diversity reflects the adaptation of microbes to a wide range of environments with variable conditions i.e. poor nutrient source and higher or lower than usual temperature. The latter together with genetic variability result from genetic mechanisms such as point mutation, genetic rearrangements and horizontal gene transfer (HGT) (Morschhäuser *et al.*, 2000). These are microevolutionary mechanisms which contribute towards the evolution of microbes by generating new clone variants within a short

space of time (Morschhäuser *et al.*, 2000). Such factors allow bacteria to evolve progressively in response to the environments which they encounter in order to alter their adaptive and functional properties (Falush, 2009).

Microevolution is actively involved in the development of pathogenicity and drug resistance activities that are primarily caused by change of function mechanisms which depend on point mutations that confer selective advantage to bacteria in the new host niche i.e. host immune response and during antibiotic therapy (Sokurenko *et al.*, 1999; Casali *et al.*, 2012). The selective advantage mentioned in the latter may also be implicated by the colonization of the host by bacteria, as the dynamics of the (new) environment and host-bacteria interaction is essential for the onset of pathogenicity and severity of the disease caused (Wilson, 2012). Point mutations denote the substitution of one nucleotide by another frequently known as single nucleotide polymorphism – SNP, and also insertion or deletion – indel of a single nucleotide or sometimes two, three, etc nucleotides deleted or inserted. SNPs and indels which occur in protein encoding DNA sequences may result in missense mutations or frameshifts which may have an impact on the function of the protein, respectively. The contribution of the latter in the acquisition of antibiotic resistance by pathogens has been demonstrated in *Staphylococcus aureus* where high impact mutations in genes: *stp1* [Q12Stop] and *vraS* [G45R] selectively conferred resistance towards teicoplanin: a glycopeptide antibiotic (Renzoni *et al.*, 2011). Few other SNPs which confer fitness and drug resistance have similarly been reported in e.g genes: *rpoB*, *katG*, *pncA* and *embB* of *Mycobaterium Tuberculosis* (Casali *et al.*, 2012). These result in change of function mechanisms which arise from minor genetic alterations that allow bacteria to grow and spread in diverse host environments (Sokurenko *et al.*, 1999).

More SNPs with selective advantages towards pathogenicity have been detected in organisms such as *Pseudomonas aeruginosa* (Boucher *et al.*, 1997; Yu *et al.*, 1998); *Vibrio parahaemolyticus* (Okuda & Nishibuchi, 1998) and streptococci (Stockbauer *et al.*, 1999). *P. aeruginosa* are frequently associated with lung infections in cystic fibrosis and individuals have been shown to typically possess mutations in their MucA genes which result in the overproduction of exopolysaccharide alginate (Boucher *et al.*, 1997). The overproduction of alginate is advantageous for *P. aeruginosa* as it can persist in a favoured niche and promote chronic respiratory infections and inhibition of phagocytosis (Boucher *et al.*, 1997; Sokurenko *et al.*, 1999). The persistent *P. aeruginosa* isolates further gain additional adaptive

mutations in genes for O-antigen biosynthesis, type III secretion, twitching motility, exotoxin A regulation, multidrug efflux, osmotic balance, phenazine biosynthesis, quorum sensing, and iron acquisition relative to wildtype / early isolates (Smith *et al.*, 2006). These mutations mainly arise due to the disrupted DNA mismatch repair gene: *mutS* coupled with positive selection pressure in bacterial niches (Smith *et al.*, 2006; Hoboth *et al.*, 2009; Cramer *et al.*, 2011).

Another kind of a mutational phenotypic effect in microorganisms has been shown by studying gene *tdh1* of *V. parahaemolyticus* where a base change in the promoter region was demonstrated to increase the production of hemolysin (Okuda & Nishibuchi, 1998). In addition, another study conducted on the streptococcal pyrogenic exotoxins (spe-AC) isolated from different countries suggested that these genes represented functionally different variants which may be linked to different levels of pathogenesis as a result of non-synonymous amino acid exchanges (Norrby-Teglund *et al.*, 1994). Genetic rearrangements also contribute towards the phenotypic conversions of microorganisms. The effect of rearrangements has been experimentally demonstrated in the conversion of *P. aeruginosa* to the mucoid phenotype for a chronic pulmonary infection as observed in a region upstream of the exotoxin A gene (Sokol *et al.*, 1994). The latter suggests that the lungs of cystic fibrosis patients are initially colonized by nonmucoid *P. aeruginosa* which at the later stages of an infection are converted to mucoid phenotypes (Sokol *et al.*, 1994). Mutations together with genetic rearrangements lead to improved biological functions that are under selective pressure and also give rise to microbial diversification and phenotypic changes that allow bacteria to occupy variable niches (Schloter *et al.*, 2000).

Apart from the factors mentioned above, HGT defined as the transfer of genetic material between organisms in a manner other than by descent, has always been tightly linked with microbial adaptation and evolution ever since the era of incongruent and conflictive phylogenies (Boto, 2010). The process of HGT was illustrated to be an important factor for the dissemination of genes which confer virulence determinants and antibiotic resistance upon the conjugation experiment conducted in *Escherichia coli* by Lederberg and Tatum in 1946 (Lederberg & Tatum, 1946). The transfers and exchanges of genes are well documented mainly for bacteria. The HGT event can happen within a species as well as between bacteria from different species. However these exchanges have been shown to not only happen among bacteria but also between bacteria and different other domains such as archaea and eukarya in

all possible directions (Boto, 2010). HGT allow bacteria to acquire genetic material from distant or related organisms to colonize different environments or new hosts (Becq *et al.*, 2010). The intercellular transfers of genetic materials occur mainly because of the existence of mobile genetic elements (MGEs). These are elements which can move or be mobilized between bacterial cells. MGEs such as plasmids, transposons and bacteriophages are the key vectors for gene transfer between bacteria. The latter are known to contribute directly towards bacterial evolution and speciation (Elsas & Bailey, 2002). Additionally, these elements have been linked to the spread of adaptive and symbiotic traits which are involved in the survival/ fitness of host organisms and their neighbours (Elsas & Bailey, 2002; Rankin *et al.*, 2010).

1.2.1 Horizontal Gene transfer

Two strains of bacteria of the same species can differ by as much as 30% of the accessory parts of their genomes (Sueoka, 1962). These differences mostly result from mechanisms such as: insertions, deletions, transpositions, duplications, recombination and rearrangements of residues of mobile DNA sequences. Various comparative and statistically driven computational methods have been developed with the aim to decode the rearranged genomic structures as well as the characteristics of gene flow among different species. A list of various bioinformatics tools developed to search for sets of genes exchanged between bacterial genomes is presented in a review by Langille *et al*., (2010), however only a selected of few of such methods will discussed in this work below. Sequence data were found to display wide variations in their nucleotide compositions across bacterial species, as a result of an evolutionary factor that infers genome plasticity, known as horizontal gene transfer (Hacker *et al.*, 2003). It is increasingly becoming apparent that genetic materials within single and multi-celled organisms have been acquired by horizontal gene transfer since the early stages of life (Choi & Kim, 2007; Boto, 2010). The exchange of genetic material was found to have occurred in different domains of life as mentioned in the previous section (Choi & Kim, 2007; Boto, 2010). This mechanism effectively contributes to the evolution and diversity of bacterial species through the transmission of novel genomic segments, however not all of these undergo horizontal transfer as their preferential transfer is strongly correlated with gene function (Jain *et al.*, 1999). Informational genes, defined as the core and most conserved segments in a genome are present in almost all organisms. Such genes encode conserved proteins such as DNA and RNA polymerases and are therefore less likely transferred as genomes that lack their functional properties are rare (Ochman, 2001; Dutta & Pan, 2002). Apart from the rarity of their functional properties, these genes are mostly highly expressed

and genes with such expression levels have been shown to be less subject to HGT events (Park & Zhang, 2012). Operational genes, also defined as the accessory parts of the genome are highly likely to be transferred as they may provide their host bacterium with fitness traits (Rivera *et al.*, 1998; Jain *et al.*, 1999). The transfer of operational genes is a continual process that promotes prokaryotic diversity (Jain *et al.*, 1999; Ochman, 2001). During the horizontal gene transfer events the acquired DNA providing functions that are beneficial to the host are kept while the DNA providing less beneficial functions may be discarded (Lawrence, 1999). However, it does not necessarily mean that for each block of DNA to be taken up by bacteria through HGT the other has to be discarded.

Mobile genetic elements possess genes that contribute not only towards bacterial speciation but also carry with them factors that contribute to bacteria's fitness traits, secondary metabolism, antibiotic resistance and symbiotic interactions (Hsiao *et al.*, 2003; Dobrindt *et al.*, 2004; Mantri & Williams, 2004). Collectively these factors are known as the flexible gene pools. The flexible gene pools are named according to the types of functions they encode, those that encode virulence features are designated pathogenicity islands (PAI). PAI were first identified in human pathogenic strains of *E. coli*, the acquisition of genes of their sort have been shown to possess the ability to confer a virulence phenotype upon a normally avirulent strain (Ochman, 2001). PAIs are highly variable mobile DNA segments present only in one or more strains of a given species.

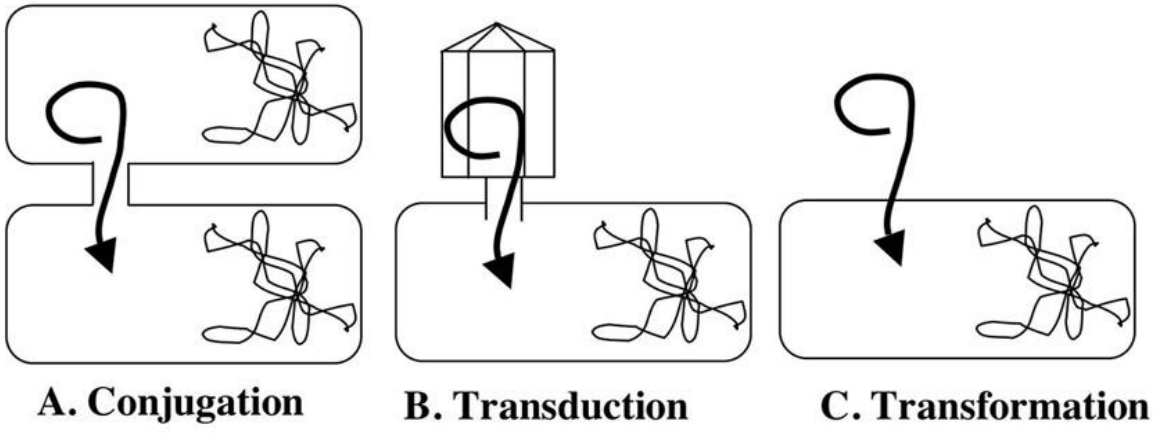


Figure 1.1: The mechanisms of horizontal gene transfer. The above figure denotes three mechanisms involved in the transfer of genetic material among bacterial species. These are: (a) Conjugation: a transfer that is mediated by plasmids (b) Transduction: a transfer that is mediated by bacteriophages. (c) Transformation: a transfer that is mediated by an uptake of free DNA.
Image adapted from http://www.bioscience.org/2009/v14/af/3515/fulltext.php?bframe=figures.htm.

These segments can transfer between environmental microorganisms, across species and even genus boundaries and influence virulence. Genomic elements similar to pathogenicity islands by general composition and organization are subsequently identified in non-pathogenic bacterial species, and termed genomic islands (Hacker & Carniel, 2001; Dobrindt *et al.*, 2004). Genomic islands (GIs) are multigene chromosomal subunits that confer bacterial multifunctional traits and are evident of horizontal transfer. The transfer of such subunits occurs through three mechanisms (Figure 1.1): (a) transformation, (b) conjugation and (c) transduction. These mechanisms mediate the movement and transfer of DNA segments intercellularly. Conjugation and transduction are the major players in HGT and require mobile elements such as plasmids and bacteriophages to transfer genetic elements, respectively (Hacker & Carniel, 2001). Upon transfer, these genetic elements integrate into their host's chromosomes through homologous or illegitimate recombination techniques or exist as conjugative plasmids (Dutta & Pan, 2002; Beiko *et al.*, 2005).

Conjugation is the process of DNA transfer mediated by plasmids. The process requires physical contact between the donor and recipient organisms which is initiated by independently replicating, self-transmissible, and mobilizable plasmids. The donor cell extends its pili structure to attach to the recipient and pull the two cells together to create a conjugation bridge where one strand of plasmid DNA is passed into the new host (Sia *et al.*, 1996). Apart from plasmids, this mechanism is also known to be mediated by conjugative transposons, elements which are known to encode proteins that facilitate their own transfer or the transfer of their host's DNA to recipient organisms.

Transduction is the process whereby DNA transfer is mediated by bacteriophages whereby physical contact between donor and recipient cells is not required (as in conjugation). Bacteriophages are known as independently replicating elements which facilitate the transfer of DNA fragments between organisms. These are known to package fragments of host DNA into their capsids by mistake through generalized transduction processes, which then get injected into newer hosts for recombination. However, host DNA packaging is limited to phages which are about 50-100kb in size. The packaged DNA is required to successfully recombine with the new host genome in order to survive and become functional. The latter recombination process is said to be only limited to bacterial members of the same species (Frost *et al.*, 2005).

Transformation, unlike the other mechanisms does not require any form of a vector to transport genomic elements between bacteria as it is mediated by the uptake of naked DNA in

the environment. The uptake usually takes place upon the release of DNA from decomposing and disrupted cells, viral particles, excretions from living cells (Thomas & Nielsen, 2005) or even through prey-derived HGT just as illustrated in *Bdellovibrio bacteriovorus* (Frost *et al.*, 2011). Natural transformable bacteria need to develop competence in order to uptake and integrate naked DNA into their genomes (Dröge *et al.*, 1999). During uptake, the naked DNA attaches to the specific cell receptors, thereafter one strand of the two gets degraded whereas the other gets transported into the cytoplasm and gets integrated with the chromosome of the new host (Lorenz & Sikorski, 2000).

DNA compositions comparisons between lineages have revealed that genes acquired by the above mechanisms display features that are distinct from those of their recipient genomes (Hacker & Carniel, 2001; van Passel *et al.*, 2005). Genes acquired by horizontal transfer can often display atypical sequence characteristics and a restricted phylogenetic distribution among related strains (Ochman *et al.*, 2000; Dutta & Pan, 2002). Although genes acquired by horizontal transfer appear atypical at the start, they overtime get affected by a process of amelioration and start to reflect the compositions of their host genomes. Bacterial species are variable in their overall GC content but the genes in genomes of particular species are fairly uniform with respect to their base composition patterns of codon usage and frequencies of oligonucleotides (Sueoka, 1962; Ochman *et al.*, 2000; Hsiao *et al.*, 2003). The phylogenetic aspect of similarity in base composition among closely related species arises from their common origin (Sueoka, 1962).

Similarity is also influenced by species-specific mutational pressures that act upon their genes to promote the maintenance of composition stability. The similarity of these compositions is conserved within and among lineages. Native/core genes in a bacterial organism exhibit homogeneous GC content and codon usage whereas foreign genes display atypical characteristic features that resemble horizontal transfer. GIs display characteristic features such as: high or low GC content, unusual codon usage, and atypical amino acid usage that differs from the rest of the chromosome (van Passel *et al.*, 2005; Dutta & Pan, 2002). These are known to be relatively large genomic fragments with sizes that range between 10-200kb, whereas fragments of sizes less than 10kb are termed genomic islets (Juhas *et al.*, 2009). GIs often prefer tRNA genes as integration hotspots and are also flanked by direct repeats (DR) of the same or variable lengths and insertion sequences (IS) (see Figure 1.2) (van Passel *et al.*, 2005; Dutta & Pan, 2002). These regions also contain transposase or integrase genes that

are required for chromosomal excision and integration (Auchtung *et al.*, 2005; Klockgether *et al.*, 2007), respectively.

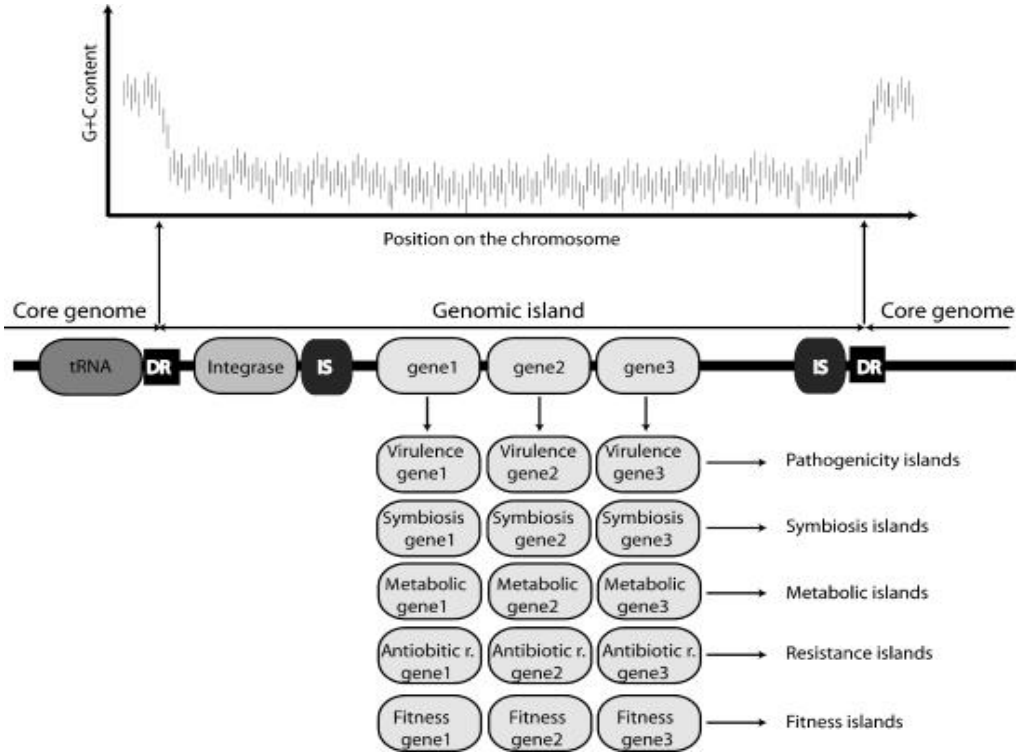


Figure 1.2: Schematic model of a genomic island of bacteria. This figure displays the characteristics of genomic islands and how they differ from the composition of their host. Genomic islands often display low or high GC content compared to that of the host genome. They are often inserted at the tRNA site and flanked by direct repeats (DR), insertion sequences (IS) and integrases. Genomic islands often harbour genes that posses different functions i.e. virulence and symbiosis genes. Image was adapted from Juhas *et al.* (2009).

1.3 Genomic fragments acquired through horizontal transfer events

The HGT associated genomic regions are classified according to their functional properties. Different classes of GIs are explained in detail in the sections below.

1.3.1 PAI-pathogenicity islands

Although amino acid mutations, gene loss via deletion, and nucleotide insertions contribute towards the differential patho-adaptive evolution of bacteria (Ochman *et al.*, 2000; Kisiela *et al.*, 2012), acquisition of PAIs through HGT has showed to play a major role in the emergence of virulence (McDaniel & Kaper, 1997). PAIs were first described in 1990 by Hacker et al (1990) as gene clusters which coded for hemolysin and fimbiriae only present in *E. coli* strains causing urinary tract infections, sepsis and meningitis. These PAIs were then

defined to be highly variable mobile DNA segments (10-200kb) present only in one or more pathogenic strains and absent in non-pathogenic strains of a given species (see Figure 1.2) (Hacker & Kaper, 2000). These are mainly associated with adherence factors, iron uptake systems, Toxins, Types I - VI secretion systems and antiphagocytotic determinants (Spanier & Cleary, 1980) which allow bacteria to undergo several host-cell infection cycles such as adherence to host cell surfaces, evasion from host immune response, and production of toxins.

PAIs are named according to their encoded functions or by the names of the species they are associated with. The pathogenicity island harboured by the enteropathogenic strains of *E. coli* is termed the locus of enterocyte effacement – LEE which is involved in the adherence to epithelial cells and formation of lesions (Blanc-Potard & Groisman, 1997). Individual bacterial strains may harbour multiple clusters of virulence genes that are acquired by horizontal gene transfer. *Salmonella enterica* serovars are a typical example of bacteria which possess several clusters of PAIs termed **S**almonella **P**athogenicity **I**slands SPIs 1-5 and each of these confers a different beneficial function (Rychlik *et al.*, 2009). SPI1 – invasion of host cells (Mills *et al.*, 1995), SPI2 – intracellular survival especially host immune cells (Cirillo *et al.*, 1998), SPI3 – intra-macrophage survival and virulence (Blanc-Potard & Groisman, 1997), SPI-4 – adhesion to epithelial cells and (Gerlach *et al.*, 2007), SPI5 – enteropathogenicity (Knodler *et al.*, 2002; Eswarappa *et al.*, 2008).

PAIs are often located within or adjacent to chromosomal tRNA regions, known to serve as integration hot spots for the mobilome (a group of mobile genetic elements in a genome). Plasmids and bacteriophages mainly encompass the mobility of virulent cassettes across species boundaries. The distributions of virulence associated gene clusters contribute towards microbial evolution and are important towards the development of novel pathogenic strains. McDaniel and Kaper illustrated that PAIs are able to convert bacteria from avirulent to fully virulent through an experiment which they conducted on *E. coli* K-12, indicating the importance of PAI towards the evolution of pathogenic microbes (McDaniel & Kaper, 1997). Plasmids have previously been mainly associated with the dissemination of antibiotic resistant genes across bacteria (O'Brien *et al.*, 1980). Their contribution towards bacterial virulence was previously shown for *Yersinia enterocolitica* (Portnoy *et al.*, 1981) and enteroinvasive shigellae (Sasakawa *et al.*, 1988). Bacteriophages also take part in the evolution of microbial pathogenicity as much as virulence plasmids. The experimental

measures which have previously been conducted on *Streptococcus pyogenes* CS112 indicated that the streptococcal pyrogenic exotoxins A and C (SPEs) which they carry are phage associated, as they were found to be located adjacent to phage insertion sites (Betley & Mekalanos, 1985; Johnson *et al.*, 1986; Goshorn & Schlievert, 1989). More analysis conducted on *E. coli* O157:H7 also revealed that the shiga toxins (stx1 and stx2) that they harbour are also phage related (O'Brien *et al.*, 1989). Moreover, the Vibrio cholerae genes which encode toxins responsible for watery diarrhea are known to be carried by a temperate CTX phage (Waldor *et al*., 1996).

1.3.2 Antibiotic resistance islands

A variety of bacteria develop resistance to antibiotics through mechanisms such as random DNA mutation, genetic rearrangements, or HGT mechanisms. The most common mechanism through which bacteria acquire drug resistance gene cassettes is conjugation. The acquisition of these resistance gene clusters enables bacteria to survive in lethal / hazardous niches and replicate in the presence of antibiotics. Antibiotic resistance genes are often carried by transposable elements which are frequently located in plasmids and highly associated with IS elements at their flanks (Bennett, 2008). Apart from transposons and plasmids, integrons have also been noted to take part in the spread of antibiotic resistance factors. The latter cannot self-transpose and therefore associate themselves with conjugative plasmids and transposons to aid in the transmission and regulation of resistance. Disease-causing bacteria in clinical environments are increasingly developing resistance to some of the drugs most commonly used for treatments i.e. Enterobacteriaceae and Pseudomonads. A clinical isolate of *P. aeruginosa* was reported to possess transposon: Tn*2401* known to be a transposable element which confers multiresistance towards aminoglycoside antibiotics such as: gentamicin, tobramycin, siosomicin, dibekacin, and amikacin (Schmidt *et al.*, 1983). The other transposons which are known to carry genes which confer resistance against: kanamycin, chloramphenicol, ampicillin, and erythromycin are known as: Tn*5*, Tn*9*, Tn*1*, Tn*917* respectively (Bennett, 2008).

Tetracyclines also form part of broad spectrum antibiotics which have always been effective towards the inhibition of protein synthesis in a wide range of microbes until superbugs emerged. Superbugs are pathogenic microbes which carry different kinds of resistance genes in their genome. Most tetracycline resistance genes are associated with resistance plasmids (Hartman *et al.*, 2003; Pezzella *et al.*, 2004) demonstrating their distribution among bacteria.

Genes which confer resistance to these antibiotics have been identified and are designated as tet of types A to G with close associations with class 1 integrons (Pezzella *et al.*, 2004; Agersø & Sandvang, 2005). Tetracycline resistance gene: tet(A) was found to be associated with transponson Tn1721 carried by plasmid pGFT1 of *Salmonella enterica* subsp. *enterica* serovar (Frech & Schwarz, 1998). Tet(A)- 1 an allele of tet(A) (Hartman *et al.*, 2003) was identified in both *Salmonella* spp and entero-invasive *E. coli* carried by plasmid pSSTA-1. Several other tet genes, designated tet(L), tet(H) and tet(O) were identified in *Actinobacillus* by Blanco *et al.* (2006) and were also found to be associated with plasmids p11745 and p9555.

1.3.3 Heavy metal resistance islands

A number of GIs comprising of possible virulence determinants were identified in the genome of *E. coli* TY2482, and at least one of these possessed a plasmid associated mercury resistance operon (Bezuidt *et al.*, 2011). Heavy metal resistance genes are often found in PAIs of different highly virulent microorganisms (Durante-Mangoni & Zarrilli, 2011; Levings, *et al.*, 2007). Such resistance genes are utilized by bacterial pathogens for protection from the exploitation of a transition metal such as copper by mammalian immune defenses (Osman & Cavet, 2011). Copper is known to be an essential toxicity component used by macrophages for killing pathogens within phagosomes once engulfed (Osman & Cavet, 2011; Percival, 1998; Gold *et al.*, 2008). However, the most frequent heavy metal resistance genes reported to be associated with PAIs and plasmids are the mercury resistance operons. Schottel et al. reported that in a collection of some 800 antibiotic-resistance plasmids isolated from clinical *E. Coli* strains, 25% of these possessed mercury resistance determinants (Schottel *et al.*, 1974). The latter serves as an indication that these determinants are frequent in plasmids and may also be associated with microbes which confer resistance to antibiotics. In fact, the *mer* operons are highly versatile (Silver & Phung, 1996) and it has recently been demonstrated experimentally that the sensitivity of mercury-sensing regulators may be re-directed by mutagenesis to sense other heavy metal pollutants (Hakkila *et al.*, 2011). The role of the *mer* operons in pathogenicity remains unclear, but their prevalence in pathogenic bacteria suggests that they may be important for alternative functions such as transport and detoxification of antibiotics and other detrimental compounds. Furthermore, the roles of mercury resistance genes in bacterial resistance toward clinical disinfectants have also been reported (Russell, 1999). In relation to the latter, an acquired antiseptic and disinfectant resistance of *Acinetobacter baumannii* was associated with the arsenic and mercury

resistance operons (Durante-Mangoni & Zarrilli, 2011) which could possibly be of plasmid origin. Not only are these genes important for clinical / pathogenic strains but several other genes with similar characteristics are frequently identified in bacteria that are found in polluted or mineral niches to aid in response to heavy metals (Diels *et al.*, 2009; Janssen *et al.*, 2010). These elements have also been shown to be encoded by plasmids which allow bacteria to grow in heavy metal contaminated habitats i.e Cupriavidus metallidurans (Diels *et al.*, 2009; Janssen *et al.*, 2010).

1.3.4 Symbiosis Islands

Bacteria have different modes of establishing mutual relationships with their host organisms, particularly multi-cellular organisms. These relationships mainly result from acquisitions of pathogenicity and symbiosis islands by bacterial species through horizontal transfer. Symbiosis and pathogenicity islands share similar structural properties and both use similar mechanisms for manipulating their multicellular hosts. Unlike pathogenicity islands, symbiosis islands do not cause infections nor cause tissue damage to their hosts instead they encode genes for mutualism (Uchiumi *et al.*, 2004). Cases for symbiosis are common for bacteria and plants, mainly between *Mesorhizobium*, *Rhizobium*, *Bradyrhizobium*, *Sinorhizobium*, *Azorhizobium* and legumes (MacLean *et al.*, 2007). *Rhizobia* are well known as symbiotic nitrogen-fixing soil bacteria that use leguminous plants as their hosts (Uchiumi *et al.*, 2004). Their mutual relationships are established by symbiotic genes that they possess which are usually clustered on large plasmids referred to as symbiotic plasmids (pSym) or within genomic islands referred to as symbiotic islands (SIs). Some SIs are associated with phage-related integrases indicating that they may be of bacteriophage origin especially the ones found in *M. loti* and *B. japonicum* (MacLean *et al.*, 2007). The symbiotic *M. loti* strain ICMP3153 together with the others have been illustrated to be capable of converting nonsymbiotic *Mesorhizobia* into symbiotic counterparts through acquisitions of SIs and pSym (Sullivan & Ronson, 1998; MacLean *et al.*, 2007).

Plants require nitrogen as a measure of alternating their metabolic pathways but are not capable of fixing it directly from the atmosphere, instead they depend on *Rhizobia* to initiate such processes. *Rhizobia* convert N2 gas to NH3 for their hosts (Uchiumi *et al.*, 2004) and in turn these bacteria get high energy plant-derived carbohydrates. *Rhizobia* fix nitrogen within the plants roots nodules which they form upon infection. These are made possible by the presence of nod (nodulation) and nif (nitrogen fixation) genes which are possessed within theirs pSym (Prakash & Atherly, 1984). The acquisition of GIs resembles evolutionary

mechanisms that shape up host-bacteria interactions, and the adaptation of bacteria to different host environments. *Rhizobia* seem to be leading a dual life-style since they possess factors that enable them to survive both in the soil and plants. Life in the soil allows them to exchange genes with other bacteria in the same niche to promote diversity (Sullivan *et al.*, 2002; Uchiumi *et al.*, 2004).

1.3.5 Auxillary metabolic islands

Genomic islands do not only ensure the survival and adaptation of host bacterial organisms to their ever changing environmental conditions. They also offer fitness traits to their vectors to ensure survival, fitness and a more successful infection. Such genes have only been identified in marine cyanophages known to infect cyanobacteria. The proteins which they encode are known as auxillary metabolic genes (AMGs) and are thought to increase phage fitness by modifying the metabolism of the host during infection (Thompson *et al.*, 2011). The cyanophages which are particularly known to infect bacteria such as Prochlorococcus and Synechococcus are the only phages to possess such genes (AMGs). These allow cynaphages to redirect host metabolism to increase the biosythensis of dNTP and to boost the host Pentose Phosphate Pathway and Photosynthetic reactions to produce NADPH for phage genomic DNA production.

1.4 Features and detection of Genomic Islands

Genomic islands are characterized by their sequence composition features that distinguish them from native genes in the genome (Daubin *et al.*, 2003). Most of the previously published methods for the detection of genomic islands search for genomic fragments that possess atypical compositional features (Philippe & Douady, 2003). The use of composition-based methods is most plausible as it does not involve phylogenetic approaches and DNA comparison between multiple species to detect tree incongruencies and abnormal sequence similarities. The other advantage of using this method is that analysis can be conducted directly from a single genomic sequence. At most, gene clusters that appear atypical in a bacterial genome are suspected of having been acquired from foreign sources (Lawrence & Ochman, 2002). The assumption is that directional mutation pressures within bacterial genomes may easily distinguish between native genes and genomic fragments which have been acquired from genomes with different mutational biases (Lawrence & Ochman, 2002). Recently acquired GIs display characteristic features of their donor genomes but over time

they start reflecting the base compositions of their new hosts as they start getting affected by the same directional mutation pressure as all the other neighbouring genes (Lawrence & Ochman, 1997; Dutta & Pan, 2002).

1.4.1 Phylogenetic inference of HGT

HGT has been said to lead to topological differences between phylogenetic trees constructed from sets of sequences of the same bacterial taxa (Beiko & Hamilton, 2006). The latter complicates the consruction of a HGT-free phylogenetic tree as many bacterial genomes are known to have been involved in gene exchange events. However, phylogeny based methods are among those that may accurately detect HGT events and identify directions of gene transfer between different nodes of bacterial taxa (Poptsova & Gogarten, 2007). These require a construction of a HGT-free tree which accurately describes the evolutionary relationships between the studied microorganisms. Apart from HGT, a reference tree should also be free from statistical biases and complications of paralogy in order to be efficiently used as a reference against trees of orthologous genes (Beiko & Hamilton, 2006). HGT detection phylogenetic-based methods are meant to find inconguencies between a studied bacteria species tree and trees of orthologous genes (Poptsova, 2009). These methods are may fail to detect HGT due to a wrong selection of families of orthologous genes and unreliable phylogenectic reconstructions.

1.4.2 Sequence composition- based approaches

The two main types of sequence composition approaches that search for GIs in prokaryotic genomes utilize nucleotide composition and genomic signature. The genomic signature method searches for atypical genomic regions by calculating genomic oligonucleotide usage frequencies whereas the nucleotide composition method determines GC content and codon usage bias. Genes which are associated with horizontal gene transfer are mainly A+T-rich (Daubin *et al.*, 2003). They are displaced in a similar codon usage direction (AT-rich) (Daubin *et al.*, 2003) and reflect the pattern of their donor genomes (Ermolaeva, 2001). Their differences can be visible in each codon position, particularly the third codon position as it is the most likely to change synonymously (Lawrence & Ochman, 1997; Daubin *et al.*, 2003). Ranjan *et al.*, (2007) have shown that codons with dinucleotide patterns such as: AA, AT, AG, TA and TC at their first two positions are more abundant in GC poor genomes. Those with bases such as: GG, GC, CT, CG and CC at the first two positions are more abundant in GC rich genomes (Rajan *et al.*, 2007). Moreover, Daubin *et al.*, (2003) examined the base

composition and codon usage in genes unique to genomes from several bacterial species and found that genes believed to be recently acquired have a relatively low GC content and atypical codon usage patterns when compared with surrounding genes, even in AT-rich genomes (Charkowski, 2004). Daubin *et al.*, (2003) further illustrated that the native genes in enterobacteria generally avoid the usage of codons: ATA, AGA and AGG which are abundant in their horizontally acquired genetic elements. Sharp and Li (1987) proposed a method to calculate the codon adaptation index (CAI) for each gene in a genome to quantify the degree of codon usage bias. The CAI compares the codon usage bias of a given set of genes relative to a reference pool of highly expressed genes (Sharp & Li, 1987) and also measures the dominating codon bias in genomes (Carbone *et al.*, 2003). Davids and Zhang (2008) effectively adopted the CAI method and implemented it in determining the differences in gene expression levels of horizontally transferred genomic islands in accordance to core (shared by all *E. coli* strains) and non-core (present in one strain and not all others) genes of different *E. coli* strains. Their analyses illustrated that core genes, although evolving slowly, have higher gene expression levels and an increased codon adaption index as compared to the highly evolving HGT genes. Lawrence and Ochman (1998) w11ere also able to identify genes in *E. coli* MG1655 which were horizontally acquired from organisms of similar composition but very different codon usage patterns using the CAI method. These genes showed a strong bias and low CAI due to the fact that they use codons not used and preferred by *E. coli* (Lawrence & Ochman, 1998). The latter observations explain why conserved genes use optimized codon usage patterns while putative or horizontally transferred genes do not. Based on the associations of mobile elements with atypical genomic features, the currently developed methods search genomes for genomic regions that possess (atypical) DNA features which are distinguishable from those of the native genes (Karlin *et al.*, 1997; Garcia-Vallvé *et al.*, 1999; Deschavanne *et al.*, 1999; Daubin *et al.*, 2003, Tsirigos & Rigoutsos, 2005).

Genomic signature methods are popular and currently the most preferred over other available methods for their practicality and sensitivity in predicting GIs (Hsiao *et al.*, 2003). The concept of searching for GIs by calculating oligonucleotide frequencies was first introduced and brought into practical use by Karlin and Burge (1995), following the work of Josse *et al.* (1961) and Russell *et al.* (1976). Karlin and Burge (1995) had illustrated that bacterial organisms from different species could be differentiated from one another by studying the distribution patterns of their oligonucleotides (see Figure 1.3) as small as dinucleotides.

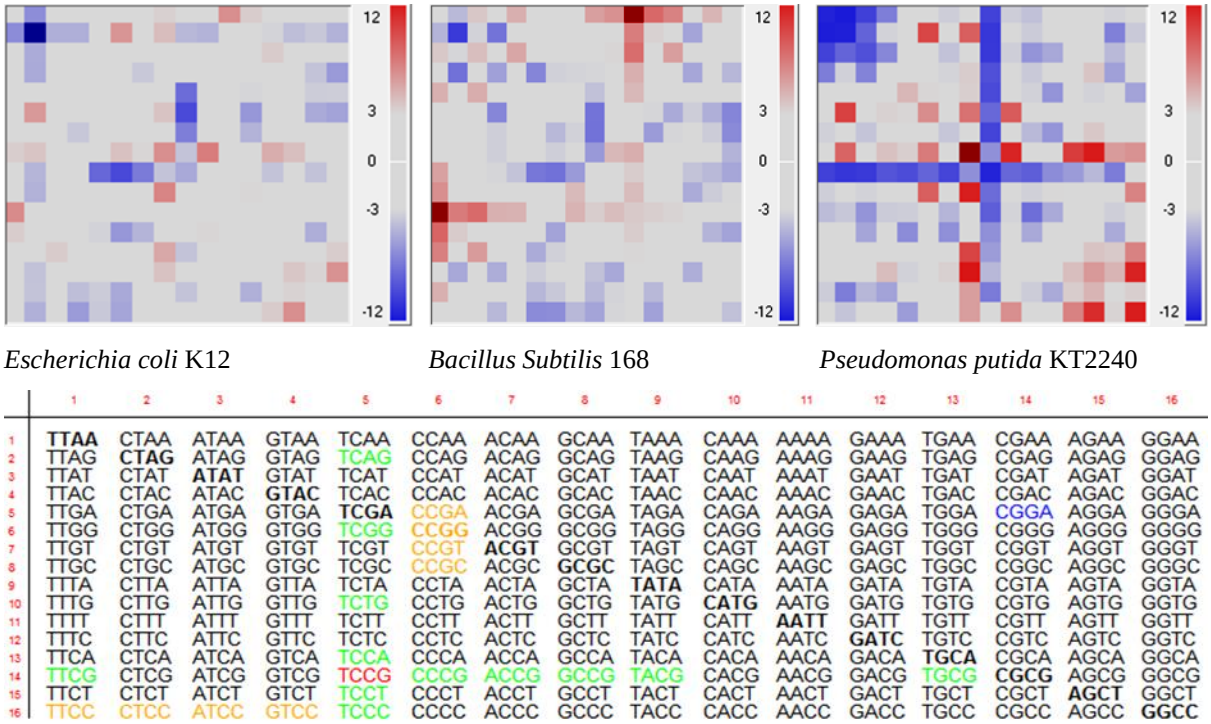


Figure 1.3: Species specific genomic signatures. The figure illustrates genomic signature determined for *Escherichia coli* K12, *Bacillus Subtilis* 168 and *Pseudomonas putida* KT2240 by calculating the frequencies of their tetranucleotides. The 256 combinations of their tetranucleotide patterns are displayed in a 16X16 matrix blocks. The different colours in blocks ranging from red-gray-blue depict overrepresentation - nonoccurrence – underrepresentation of the tetranucleotides in the genomes, respectively. Image adapted from Reva & Tümmler (2004).

Dinucleotide biases determined for successive 10 - 50kb segments of a bacterial genome were illustrated to be similar to each other and those of closely related organisms than to sequences from distant organisms (Karlin, 1998). This approach uses extensive statistical parameters to determine genomic segments that display significant differences in oligonucleotide usage patterns compared to the rest of the genome (Karlin & Burge, 1995). It detects horizontal transfer events using global sequence patterns rather than using individual genes, and does not require DNA similarity analysis or phylogenetic distributions (Karlin *et al.*, 1997). Oligonucleotides are simply defined as chains of overlapping short words of the same or different lengths. Patterns of frequencies of oligonucleotides in bacterial genomes are not random (Reva & Tümmler, 2004) and can be used to reveal different properties of DNA (Bohlin *et al.*, 2008).

Karlin and Burge (1995) determined distributions of dinucleotides by establishing the statistical formula: $\rho(xy) = f(xy) / f(x).f(y)$ of dinucleotide abundance values, where $f(x)$ and $f(y)$ denote the frequencies of mononucleotides x, y, and $f(xy)$ denotes the frequency of the dinucleotide xy. By using the formula it was observed that frequencies of dinucleotide

compositions were uniform across the entire genome, and were regarded as a stable property of DNA of a bacterial organism (Karlin *et al.*, 1994; Srividhya *et al.*, 2007). The occurrences of these patterns although not fully understood are suspected to be an influence of DNA structural properties such as base stacking energy, propeller twist angle, protein deformability, bendability, position preference or repair mechanisms (Baldi & Baisnée, 2000). On the other hand they could be a result of correlations of codon usage and environmental pressures exerted on the genome.

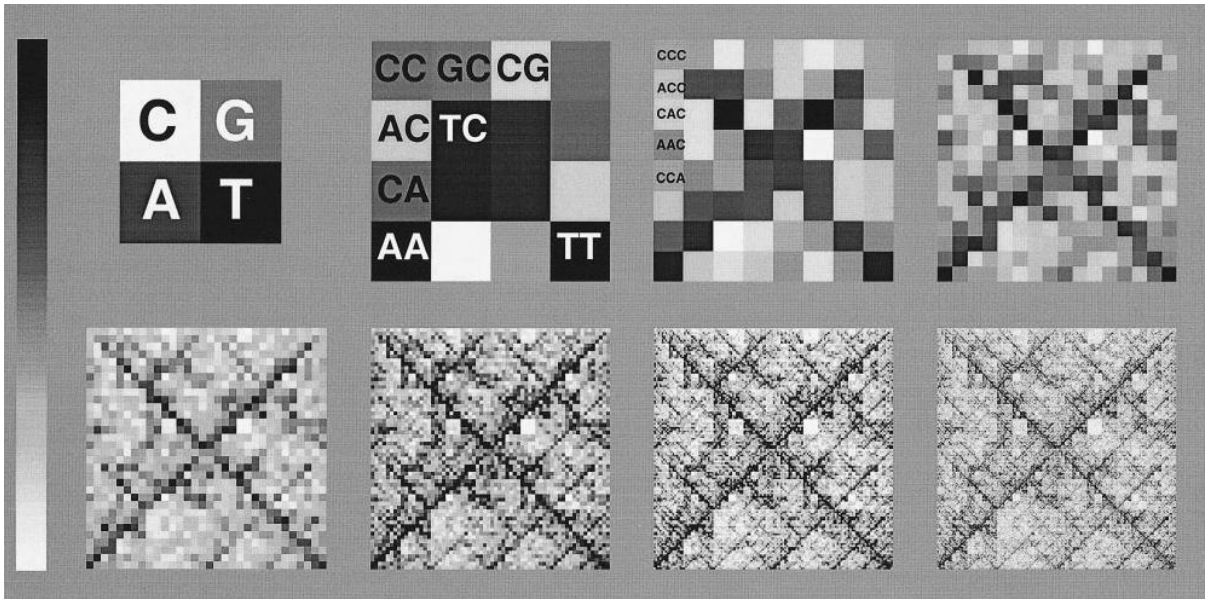


Figure 1.4: A chaos game representation of oligonucleotides. The figure presents up to eight bases long oligonucleotides determined for the genomes of *Archeoglobus fulgidus*. Image has been adapted from Deschavanne *et al.* (1999).

Methods have since been developed to visualize the signature patterns in bacteria upon the genomic signature findings explained above. These were developed with the aim to show that a genomic signature could be used as a tool to classify bacterial organisms which belong to the same species and also differentiate between those which do not. The differences in terms of oligonucleotide frequencies in genomes of different phyla lead to the concept of genomic signature, defined as the frequencies of the whole set of short oligonucleotides observed in a sequence (Karlin & Burge, 1995; Deschavanne *et al.*, 1999). These signatures also define the features for each genome. Genomes of related bacterial organisms constitute a similar signature, and can therefore be classified into groups. Deschavanne *et al.* (1999) were the first to use the Chaos Game Representation (CGR) concept initiated by Jeffrey (1990) to depict DNA signature patterns in bacterial genomes. CGR was originally developed to

visualize the underlying structures of DNA sequences from four letters 'a', 'c', 'g', 't'. Jeffery (1990) used CGR in the form of a square and labelled the vertices in each corner with bases 'a', 'c', 'g', 't' to visualize the first 6 bases: 'gaattc' of human beta globin region, chromosome 1. Deschavanne *et al.*, (1999) simplified the CGR algorithm so it could be used to represent DNA signature patterns in the form of fractal images (Figure 1.4), where every block in the image corresponds to the frequency of a specific word (oligonucleotide). The generated images are divided into four quadrants where each gets subsequently divided into four subquadrants, each containing a unique sequence pattern. Oligonucleotide frequencies are displayed by the intensity of each pixel where each pixel is associated with a specific word, the darker the pixel the higher the frequency of a pattern. The resolution of the generated CGR images correspond to the length of the word i.e. frequencies of a dinucleotide whose possible combinations is n=16 are illustrated on a 4 × 4 pixel image whereas eight-letter words are presented on a 256 × 256 pixel.

Following the concept of oligonucleotides, genomes were screened for local variations of dinucleotides with expectations to identify regions of interest where GIs might be located. The dinucleotide method was the most plausible until Deschavanne *et al.* (1999) and Pride *et al.* (2003) found that frequencies of words that are 2 and 3 bases long are poorly species-specific and do not allow a good discrimination between species since they are just an influence of codon usage preferences. They found longer words to be more species-specific even though their frequencies in genomes may appear to be more variable, and that their use may discriminate between different bacterial organisms. Tetranucleotides were indicated to carry a phylogenetic signal as they could be used to cluster bacterial organisms with similar oligonucleotide usage patterns (Pride *et al.*, 2003). These clusters were found to be congruent with the phylogenetic trees created using 16S rRNA genes as compared to clusters of shared dinucleotide and codon usage similarities (Pride *et al.*, 2003). Deschavanne *et al.*, (2000) also classified bacterial genomes using different lengths, and indicated that tetranucleotides are the best classifiers as compared to di- and trinucleotides and even much longer words. However, a recent study has indicated that longer *k*-mers, particulary heptanucleotide may perform better than tetranucleotides in taxonomic binning of metagenomic samples as they are conserved among closely similar organisms (Alsop & Raymond, 2013).

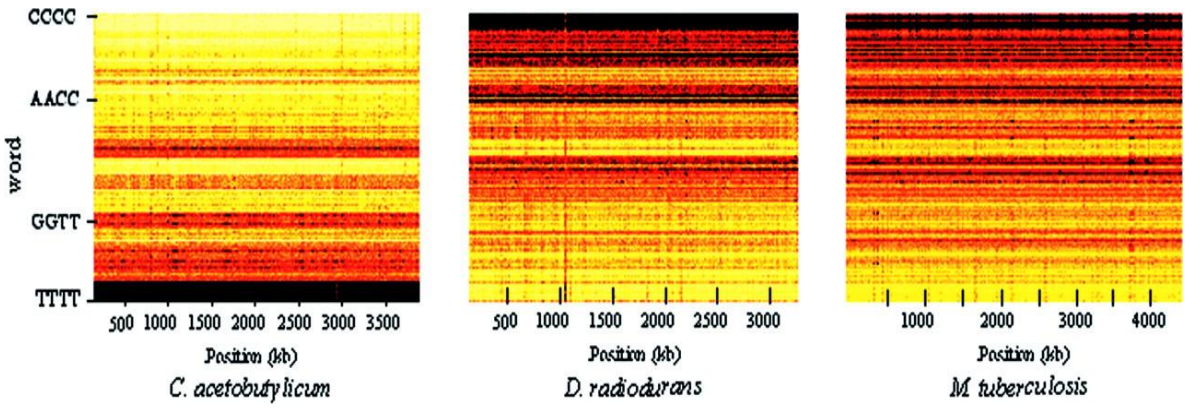


Figure 1.5: Genomic signatures with atypical tetranucleotide patterns. Tetranucleotides patterns determined for *C. acetobutylicum*, *D. radiodurans* and *M. tuberculosis*. Local signature patterns are illustrated as vertical lines (breaks), these are stacked in order to show the variations along each genome. Horizontal lines are illustrations of variations of patterns along the genome i.e. genomic regions with different signatures. Image adapted from Deschavanne *et al.* (1999).

The screening of local variations of usages of words along genomes is expected to detect the regions of interest where HGT might be located (Deschavanne *et al.*, 1999), see Figure 1.5. Horizontally transferred genomic elements are known to display the oligonucleotide characteristics of their source and their signatures can therefore be used to predict their genome of origin (Sandberg *et al.*, 2001). These GIs are first compared to each other based on their shared signatures and later get compared to those of publically available genome sequences. Although the latter sounds probable, it only applies when the (i) GIs are recent acquisitions and (ii) still highly retain the signatures of their donors (iii) and if their signatures differ from those of their hosts. Sequence composition-based approaches may be the most preferred method but may fail to identify ancient GI acquisitions that have already been ameliorated (see below) and GIs acquired from genomes with similar signatures as the host. They also mistake rRNA operons for GIs as they also display a composition which differs from those of the other genes in the host genomes. Amelioration is a process whereby GIs start to reflect the composition of their host organisms' over time due to the similar mutational pressures that affect the native genes in their environment (Lawrence & Ochman, 1997). The amelioration process occurs in order for acquired genomic elements to adapt to the host's replication, transcription or translation machineries which are optimized for the typical signatures of a species. Therefore, GIs that have been acquired for longer periods are as difficult to detect as atypical genomic fragments and reliable identification may require detailed phylogenetic analysis.

Several computational tools have been developed to aid the detection of GIs in bacteria through the use of sequence composition approaches as mentioned in the above section. Some approaches carry down genomic analysis and GI predictions using only genomic signatures whereas others combine nucleotide composition and genomic signature methods together with multiple-threshold framework to increase sensitivity of composition-based methods (Azad & Lawrence, 2011). Few of the currently available sequence composition tools are as follows:

IslandPath DIMOB - IslandPath DIMOB is a computational method that incorporates multiple DNA signals and mobile genetic elements annotation features to search for GIs (Hsiao *et al.*, 2003). The method first identified GIs by searching for bias in GC content and dinucleotide frequency distributions. These are subsequently searched for the presence of other features known to be associated with genomic islands, such as proximal tRNA and mobility genes (integrases and transposases). The search for genes which are associated with mobile genetic elements is conducted by the use of NCBI annotations and PFAM hidden markov models (HMMs). Genomic regions that have a dinucleotide bias and the presence of at least one mobility gene are tagged as being part HGT events.

SIGI-HMM - SIGI-HMM is a hidden markov model based method that searches for GIs and predicts their putative donors through codon usage analysis (Waack *et al.*, 2006). It implements a method which exploits taxon specific differences in the usage of codons when predicting GIs. Genes which exhibit unusual codon usage patterns as compared to the other genes in the genome are tagged as being part of a GI. Codon frequencies of these gens are then compared to the genes sequences of the Codon Usage Tabulated Genbank database (Nakamura *et al.*, 1999) which holds codon frequencies. A bacterial organism with a codon frequency that matches that of a GI is tagged as its putative donor. The database also contains word frequencies determined for highly expressed genes e.g rRNA operons, which normally are mistaken for GIs as they also possess atypical oligonucleotide usage (OU) patterns. These are used to filter out rRNA operons from a list of predicted GIs.

Alient Hunter - Alien Hunter is a tool which uses variable order compositional indices to search for GIs in bacterial genomes (Vernikos & Parkhill, 2006). However, it prefers longer *k*-mer's (i.e words of length 8mers) to shorter ones even though it is mentioned in the paper that although longer words are more informative, enough data is required in order to produce

reliable probability estimates (Vernikos & Parkhill, 2006). In the paper it is also mentioned that longer words are preferred over shorter as shorter ones are believed to not provide sufficient discrimination of regions with atypical patterns. In cases where longer words such as 8-mers are used and fail to provide reliable probability estimates the tool then recalculates for distribution patterns of shorter words with ranges from 2-7mers until reliable estimates are reached. HMM analyses are applied to refine the boundaries of the genomic regions that are suspected to from horizontal gene transfer events. Compositions of these regions are compared to those of their neighbours to identify particular regions in the genome where deviation starts to occur between GIs and native genomic regions.

PAI IDA - PAI IDA is a methods with incorporates GC bias, dinucleotide usage statistics and codon usage to identify GIs (Tu & Ding, 2003). The method employs a sliding window approach to search for atypical genomic regions which meet the latter three compositional criteria. The window W: local_pattern of length 20kb with steps of 5kb is used to calculate for the three criteria in search for genomic fragments that deviates from the whole genome G: global_pattern. The tool also employs a quartile method to determine cut-off values for GI predictions. These parametric values were derived from an initial training set of nine of the twenty-six known GIs from seven bacterial genomes. The nine GIs were chosen for they still possessed atypical compositions as compared to their host's global genomes and the other seventeen were discarded for their lack of compositional bias.

1.4.3 Sequence similarity-based approaches

Sequence similarity-based approaches were among the first to gain popularity as they were highly dependent on publicly available GIs and their associated gene features derived from genome comparative studies. Such studies compare the contents of genes in various bacterial organisms in search for those which share higher similarity to those in other species (Jain *et al.*, 1999; Kyrpides & Olsen, 1999). These revealed that genomic regions which were present in one or two genomes and not the other members of the same species frequently show the presence of tRNA and tmRNA in their proximal regions (Ou *et al.*, 2006). The latter together with the other mobile elements genes make up the common features that are associated with GIs (apart from GC and codon bias) as they are typically harboured in their flanks (Reiter *et al.*, 1989). The tRNA and tmRNA amongst other gene features known to be associated with GIs were initially used by sequence similarity based methods as BLAST (Altschul *et al.*, 1990) query sequences to search for homologues in other bacterial genomes to identify GIs

just as illustrated by Mantri and Williams (2004). Other feature-based GI detection methods search for phage and plasmid associated genes within bacterial genomes as they are mainly known to be vectors for mobilizing genetic elements within and between bacteria. All the known phage and plasmid annotations are first collected from public databases and further used to search for their associated features in all the publicly available genomes (Leplae *et al.*, 2004). Although the latter approaches are sound, they may suffer several drawbacks as they only determine foreign inserts on the basis of sufficient homology searches and sets of available well-annotated sequences (Rajan *et al.*, 2007). These therefore overlook GIs which lack features associated with mobile genetic elements. Some of the available resources that utilize similarity based approaches such as ACLAME (Leplae *et al.*, 2004) and PAIDB (Yoon *et al.*, 2005) will be mentioned in the online genomic islands resources section below.

Comparative genomics approaches have always been associated as the reliable methods to detect GIs in bacteria. These are more reliant on comparisons of two or more genomes of closely related bacterial organisms to search for genomic inserts which are present in one genome and not in the other related genomes. Although the comparative approach seems plausible, its only flaw is that closely related genomes are required for the method to be more probable, unfortunately, these are not always available. However, due to the increase of genomic data in public databases these methods will soon gain popularity as there will be enough genomes for comparison to search for GIs that are of environmental or medical importance. In the case of PAI detection, one needs to compare two or more genomes of related bacteria, pathogenic and non-pathogenic included in the set. The cluster of genes that only appear to be present in the pathogenic bacteria and not the others are likely to have been acquired by horizontal gene transfer events. The same approach applies for the search of other classes of GIs, these are tagged as alien genomic regions if they are only present in the query genome and not in the other multiple genomes of closely related bacteria. The current and popular method that successfully uses comparative genomics analysis (whole genome alignment approach) to identify GIs in prokaryotes is as follows:

IslandPick – IslandPick is reported to be the first completely automated method to identify GIs in prokaryotes using comparative genomics (Langille *et al.*, 2008). Input query genomes are only analysed if a sufficient number of related genomes are present in the database to compare. The latter is conducted by a flexible approach which evaluates genome relatedness

before analysis is undertaken. Upon the selection of suitable references, whole genome pair-wise alignments of the query are performed against each genome in the reference set using Mauve, a tool for multiple alignment of conserved genomic sequence with rearrangements (Darling *et al.*, 2004). The aligned pairs are subsequently searched for any rearrangements and duplications. Duplicated regions are removed if any are present then the regions which are unique to the query genome are tagged as possible GIs and are removed for further analysis. These possibly GI segments are aligned against the query genome and all reference genomes with BLAST. The unique regions with BLAST matches of less than 700 nucleotides are discarded, while the ones with less than sizes of 8kb are kept and labeled as putative GIs.

1.5 Online Genomic Islands resources

There are several online resources that have been developed to aid with the identification and comparisons of GIs identified using different methods due to the increased number of genome sequences. These provide the community with possibilities to view and manipulate pre-computed datasets generated by a wide range of computational methods. Not all of these contain only pre-computed datasets, some allow users to upload their own genomic datasets for analysis. These resources may not always be useful as some may not contain a genome which is of interest to the user, nor do they allow users to explore different parametric measures for analysis. Not all but some of the available resources for the predictions of GIs in prokaryotes will be discussed below. Only few resources were selected for discussion as many of these search for GIs using approaches such as BLAST based on searches against published GIs or virulence related factors i.e PAI-DB (Yoon *et al.*, 2007) and VFDB (Yang *et al*., 2008).

ACLAME - A Classification of mobile genetic elements ACLAME (http://aclame.ulb.ac.be/) is a comprehensive web resource which contains a collection and functional classes of mobile genetic elements (MGEs) comprising all known bacteriophages, plasmids and transposons (Leplae *et al.*, 2004). The protein sequences encoded within these MGEs were clustered into families according to their functional properties using TRIBE-MCL, a graph theory based Markov clustering algorithm. These clusters were then subsequently used to search against public databases for related sequences, and to also allow for the annotations of MGE encoded proteins with unknown functions. The similarity searches were performed with PSI-BLAST (Altschul *et al.*, 1997) against a Swissprot database (Boeckmann *et al.*, 2003), and the hits obtained from the search are linked to their corresponding clusters. Protein sequences in each

cluster were multiply aligned against one another to build Hidden Markov Models - HMM (Enright *et al.*, 2002). The models were then used for further searches against databases such as SCOP (Conte *et al.*, 2002) and NRDB-NCBI (Benson *et al.*, 2003) to identify remotely related proteins. The development of the database was led by difficulties in the systematic analysis of mobile genomic islands and the lack of MGEs in existing databases.

PAI-DB - The pathogenicity islands database PAIDB (http://www.gem.re.kr/paidb/) is a comprehensive web resource that contains information on all reported and candidate PAIs of prokaryotic genomes (Yoon *et al.*, 2007). The PAI-DB method for detecting PAIs primarily uses a homology-based method to detect pathogenicity islands in complete bacterial genomes. The detection is performed using sets of already reported PAI loci collected from the GenBank database and literature. The collected sets of known PAI loci were searched for nucleotide and protein sequence similarity against prokaryotic genomes using BLAT (Kent, 2002) and BLASTP (Altschul *et al.*, 1990) to search for genomic regions with PAI associated genes. Genomic regions sharing homology with either of the PAIs used in the search were considered pathogenic only if they were in possession of four or more virulence genes and also if they contained genes coding for transposases, IS element and integrases (Yoon *et al.*, 2005).

HGT-DB – Horizontal gene transfer database (HGT-DB) (http://genomes.urv.cat/HGT-DB/) is a nucleotide composition web resource that provides pre-calculated averages and standard deviations for: GC content, codon usage, relative synonymous codon usage and amino acid content of whole genome sequences of bacteria and archaea (Garcia-Vallve *et al.*, 2003). It also provides lists of gene clusters that possess compositional features which differ from those of their host genomes. It uses a set of statistical approaches to determine the genes that deviate from the mean GC and/or average codon usage of the host genome. Genes are marked as alien if they possess atypical GC compositions. The developers of HGT-DB do not provide the tool which they used during analysis for download, only the statistical method applied is explained on their paper (Garcia-Vallve *et al.*, 2003).

GIST – GIST (http://www5.esu.edu/cpsc/bioinfo/software/GIST) is a genomic island suite of tools for the detection of GIs in prokaryotes (Hasan *et al.*, 2012). The tool is made up of five commonly used composition-based methods and these are: Alien Hunter (Vernikos & Parkhill, 2006), IslandPath (Hsiao *et al.*, 2003), SIGI-HMM (Waack *et al.*, 2006),

INDeGenIUS (Shrivastava *et al.*, 2010) and Pai-Ida (Tu & Ding, 2003). Users are allowed to select their preferred methods for GI identification if they do not want use all five. GIST also provides an optimization tool called EGID: an ensemble algorithm for the improved genomic island detection, which it uses to evaluate the GIs predicted for a bacterial organism. EGID takes as input predicted GIs, evaluates them for shared gene features and overlaps and generates consensus predicted GIs.

IslandViewer – IslandViewer (http://www.pathogenomics.sfu.ca/islandviewer/) is the first web resource to incorporate sets of pre-computed prokaryotic GIs identified by three publicly available prediction methods (Langille & Brinkman, 2009, Dhillon *et al*., 2013): IslandPick (Langille *et al.*, 2008), IslandPath_DIMOB (Hsiao *et al.*, 2003) and SIGI-HMM (Waack *et al.*, 2006). The three methods utilized by the resource to search for GIs use different approaches, and these are explained in detail above. The use of all three methods to search for GIs allows the detection of regions missed by others, and also allows users to evaluate GI predictions *i.e.* a given GI is considered true positive if predicted by more than one tool. IslandViewer provides a user-friendly interface which allows viewing of the predicted GIs derived from the pre-calculated dataset or user defined sequences. User selected genomes are viewed as circular images with their predicted GIs highlighted in different colours which correspond to each prediction tool. IslandViewer chose to incorporate the three tools because they were freely available for download and that they were shown to have the highest precision (86-92%) and overall accuracy (86%) as compared to other tools such as: PAI_IDA (Tu & Ding, 2003), Centroid (Rajan *et al.*, 2007) and Alien Hunter (Vernikos & Parkhill, 2006) with as low as 38% precision values.

GOHTAM – GOHTAM (http://gohtam.rpbs.univ-parisdiderot.fr/) is a web resource for genomic origin of horizontal transfers, alignments and metagenomics (Ménigaud *et al.*, 2012). It provides services for the detection of GIs and their assignments to potential donor genomes. The tool consists of tetranucleotide signatures determined for the whole set of sequences of Genbank (release 188). User's genomes of choice (in either GenBank or FASTA format) are searched for GIs using a combination of genomic signature and codon usage methods these are subsequently assessed by comparing their tetranucleotide signatures to those in the species genomic signature database included in the tool in search of their donor organisms. GOHTAM also takes as input multi-fasta files of metagenomic sequences to compute for signature patterns. These are then compared to the signature database using a

neighbour-joining approach; sequences which are closest to one another are displayed in the form of a phylogenetic tree (based on signature patterns).

SWGB – SeqWord Genome Browser (http://www.bi.up.ac.za/SeqWord/mhhapplet.php) is a web resource developed to visualize the natural compositional differences of prokaryotic DNA sequences using OU statistics (Ganesan *et al.*, 2008). SWGB also allows the identification of divergent genomic regions through the analysis of biased tetranucleotide distributions in complete genome sequences. Several statistical parameters which were previously defined by Reva and Tümmler (2004, 2005) are implemented in the resource to determine OU and pattern skew differences between the patterns calculated for the local and global genome using a sliding window approach. These statistical parameters are also used to distinguish between mobile genomic islands and other elements characterized by an alternative OU (clusters of genes encoding ribosomal RNA and proteins, tandem multiple repeats and so on). Users may visually identify GIs by browsing through bacterial chromosomes and they may also group genomic fragments by their shared compositional properties.

1.6 Research objective

Horizontal gene transfer and DNA mutations have proved to be the contributing factors towards the evolution of bacteria. These have also been shown to play roles towards the emergence of pathogens, antibiotic resistance, and persistence of microbes within infected host cells. Different research groups are developing computational tools to study the mechanisms of horizontal gene transfer and its contributions towards disease outbreaks, as it is believed to be a major contributing factor towards bacterial evolution. Although tools are continuously being developed to search for GIs in bacteria, there still is a pressing need to create a system that monitors distributions of horizontally transferred GIs, to help understand and prepare for the emergence of new pathogens. The system in need should incorporate composition-based approaches to aid with the determination of GI distributions and origins, and also predict groups of bacteria which are likely to become pathogens. Environmental conditions are also known to contribute majorly towards bacterial evolution and genomic diversity. Current studies, particularly those of hospital-based infections only focus on studying microbial diversities between bacterial clones of related habitats i.e. *P. aeruginosa* isolates of cystic fibrosis patients in the same ward. There is however a need to conduct comparative genomics studies on the microevolution of bacterial clones of unrelated habitat

and geographic origin in order to understand the impact of niche differentiation on microbial pathogens which constitute a clone. The developed systems should be able to provide a common understanding about which of the evolutionary events are the most important on the level of micro-evolution

An important notion is that the origin of GIs cannot be determined solely by the traditional sequence similarity methods, i.e. by BLAST. Sequence similarity is based on functional conservations of the genes whereas the compositional similarity reflects the preferences of the replication/reparation system of an organism and its specific codon adaptation. Although composition-based methods are the most reliable in GI detection they suffer few drawbacks, i.e., OU pattern convergence between unrelated bacteria and amelioration of GIs over time. Composition-based methods however gain much more credibility when confirmed by sequence similarity comparison. A combination of the two approaches is beneficial in terms of discovering the donor-recipient interactions between microorganisms and timing their GI acquisitions. These are also practical for studying ontological links of GIs and studying their relations in terms of shared composition similarities. It is hypothesized that through the composition similarities shared by these, one might be able to create graph-based clusters and reconstruct GI distribution patterns in terms of relative acquisition periods and donor-recipient relations.

It is generally believed that genomes carry all the important information that may be utilized to determine the differences between highly virulent and avirulent strains and microbes from different habitats. This research project was designed to address the above-mentioned challenges and understand the micro-evolutionary events that shape up the genomes of such prokaryotes by comparing several groups of pathogens and environmental bacteria using several in-house programs and freely available tools. Horizontal gene transfer as one of the most important factors of micro-evolution can modify the clinical phenotype of a commensal bacterium to become pathogenic. The major focuses of this study is on the application and comparison of different tools for sequence composition-based predictions of GIs to determine their performances together with rates of specificity (true negative rate) and sensitivity (true positive rate). Upon the comparisons of different methods, SeqWord Genomic Islands Sniffer (SWGIS) an in-house tool that predicts GIs through the use of OU patterns was selected as the prediction method of choice due to its astounding performance. The other reason behind the use of SWGIS to carry out GI analysis throughout the entire study was due to the

following incorporated parametric measures that lack in existing methods: i) expected higher performance as it superimposes several statistical parameters to detect GIs and distinguishes them from other native loci with alternative OU patterns, ii) adjustability of the run parameters to optimize the sensitivity/specificity rate, iii) ability to analyse multiple complete and draft genomes in parallel. SNPs and other types of the genomic polymorphisms are other important factors which contribute towards the evolution of bacterial genomes therefore several in-house custom scripts were later used to analyse high-throughput sequence data to follow the microevolution of *P. aeruginosa* in CF patients from different geographic origins.

1.7 Aims

The first aim of this work was to implement a composition-based method that utilizes revised OU statistics and parametric measures to efficiently identify GIs in microbial communities. The pools of these identified GIs would subsequently be compared against each other and sets of freely available completely sequenced bacterial genomes based on their compositional properties to provide insight into their ontology and distribution patterns. Composition-based methods were selected for this work as they are believed to be practical for: the reconstruction of ontological links between GIs, studying donor-recipient relations between host organisms and their GIs, and determining the relative time of GI insertions.

The other aim of this work is to study the niche differentiation of the bacterial strains which constitute *P. aeruginosa* clones CHA and TB. The comparative analysis of the latter clones is expected to reveal the contributions of environmental conditions towards the genomic diversity of bacterial organisms on the clonal lineage level. This part of the study allowed the use of next generation technologies and an in depth assessment of the genomic islands and methods / patterns of gene exchange and islands present in the *P. aeruginosa* clones.

The specific aims were as follows:

Chapter 2 – Optimization and practical use of composition-based approaches towards identification of horizontally transferred genomic islands: The aim was to develop a user-friendly tool with revised and optimized composition-based parameters to effectively predict horizontally transferred genomic elements in prokaryotes for further use in Chapters: 3 and 4.

Chapter 3 – Analyses and visualization of genomic islands using composition-based

approaches: The chapter describes the further use of composition-based approaches to visualize and analyse genomic elements in terms of their shared ontological links, relative acquisition periods, and donor-recipient relations.

Chapter 4 – Mainstreams of horizontal gene exchange in enterobacteria: consideration of the outbreak of enterohemorrhagic E. coli O104:H4 in Germany in 2011: The chapter focuses on the use of composition and similarity based approaches to predict the origins of genomics islands harboured by *E. coli* O104:H4 known as the Germany 2011 outbreak strain. The results obtained from the work indicate the practical use of composition and similarity based methods towards understanding the nature of outbreak strains and distribution patterns of genomic islands.

Chapter 5 – Intraclonal genome diversity of *Pseudomonas aeruginosa* clones CHA and TB: The work was conducted to determine the effects of environment and selection towards genomic diversity of the highly virulent cystic fibrosis (CF) isolate CHA and two temporally and geographically unrelated clonal variants.

1.8 List of Manuscripts

The following are manuscripts which are relevant to this thesis:

1. **Bezuidt O**, Pierneef R, Mncube K, Lima-Mendez G, Reva ON. (2011). Mainstreams of horizontal gene exchange in enterobacteria: Consideration of the outbreak of enterohemorrhagic *E. coli* 0104:H4 in Germany in 2011. PLoS ONE 6(10): e25702.

2. Reva O, **Bezuidt O**. (2012). Distribution of horizontally transferred heavy metal resistance operons in recent outbreak bacteria. Mobile Genetic Elements, 2(2).

3. **Bezuidt OKI**, Klockgether J, Elsen S, Attree I, Davenport CF, Tümmler B. (2013). Intraclonal genome diversity of *Pseudomonas aeruginosa* clones CHA and TB.

A section on heavy metal resistance islands in Chapter 1 was extracted from the work described in Manuscript 2. Manuscripts 1 and 2 resulted from the work described in Chapters 3 and 4 and several other conference proceedings and book chapters which have not been added to the list. Manuscript 3 resulted entirely from the work described in Chapter 5.

Chapter 2

2 Optimization and practical use of composition-based approaches towards identification of horizontally transferred genomic islands

2.1 Introduction

The recurrent outbreaks of pathogens that possess new virulence factors and broad range antibiotic resistance gene cassettes reflect the importance of horizontal gene transfer (HGT) in the evolution of pathogenic bacteria (Smith *et al.*, 2000; Fernández-Gómez *et al.*, 2012). In many cases, the evolution of pathogens is mediated by mobile genetic elements (MGEs), which can easily be interchanged between bacterial taxa inhabiting the same or different environments (Kelly *et al.*, 2009). These MGEs possess / transfer DNA fragments known as genomic islands (GIs) that carry genes that can increase fitness of the bacterium. GIs are multigene chromosomal subunits that confer bacterial multifunctional traits and are evident of horizontal transfer. HGT does not only contribute towards virulence but also in the dissemination of factors that are involved in bacterial fitness traits, resistance, secondary metabolism and symbiotic interactions (Jain *et al.*, 2002; Dobrindt *et al.*, 2004; Mantri & Williams, 2004; Hsiao *et al.*, 2005). These factors are named according to the types of functions they encode. They possess a nucleotide composition which is atypical as compared to that of their hosts and can therefore be distinguished from the core genome through a composition-based analysis as described in Chapter 1, section 1.4.1.

This work illustrates the exhaustive use of compositional-based approaches to detect GIs in prokaryotes. The study also shows that composition-based approaches may produce reliable predictions when optimal parameters are introduced. The idea behind composition-based approaches is to use text analysis algorithms to distinguish between DNA fragments of different origin that may be used either for assignment of genomic fragments (Coenye & Vandamme, 2004) or for detection of horizontally acquired genomic islands (Mrazek & Karlin, 1999; Dufraigne *et al.*, 2005; Pride & Blaser, 2002). Methods used in this study for the identification of atypical DNA fragments in prokaryotes using composition-based statistics termed oligonucleotide usage (OU) statistics were implemented and explained in great depth in the previous work of Reva and colleagues (Reva & Tümmler, 2004; Reva & Tümmler, 2005; Ganesan *et al.*, 2008; Bezuidt *et al.*, 2011). These were recently revised and led to the development of SeqWord Genomic Island Sniffer (SWGIS), a computational tool

that examines and detects variances in frequencies of short chains of words (*k*-mers of length 2 to7) often referred to as 'oligonucleotides' in bacterial genomes (Bezuidt *et al.*, 2011).

The general idea of identifying GIs by GC-content and alterations in frequencies of higher order oligonucleotides is not new. A number of computational tools based on this approach have been proposed before (Mrazek & Karlin, 1999; Pride & Blaser, 2002; Ménigaud *et al.*, 2012; Abe *et al.*, 2003; Dufraigne *et al.*, 2005; Chatterjee *et al.*, 2008) in parallel with the tools that use sequence based similarity searches or combined approaches (Vernikos *et al.*, 2007; Lima-Mendez *et al.*, 2008; Langille & Brinkman, 2009; Wang *et al.*, 2010; Abby *et al.*, 2010; Wei & Guo, 2011). IslandViewer serves as the first web resource that integrates tools that use different approaches to search for GIs in prokaryotes (Langille & Brinkman, 2009). In our study SWGIS was compared to IslandViewer to help optimize its parametric values and for benchmark purposes. The optimization and benchmarking were performed by re-identifying sets of known (curated) GIs from PAIDB and predicting newer GIs in genomes which were acquired from NCBI. Sets of parametric measures determined from this study would be of practical use for other composition-based methods. These also allow the estimates of acceptable false positive and false negative rates.

2.2 Materials and Methods

2.2.1 Source of genome sequences

Sequences of bacterial genomes were downloaded from the NCBI FTP directory of bacterial genomes - ftp://ftp.ncbi.nih.gov/genomes/Bacteria/. Sets of known pathogenicity islands were obtained from the pathogenicity islands database PAIDB - www.gem.re.kr/paidb/. The pre-calculated sets of GIs of Islandviewer were obtained from the download site as follows: www.pathogenomics.sfu.ca/islandviewer/download.php.

2.2.2 OU pattern statistics for identification of atypical genomic regions

The oligonucleotide usage (OU) pattern was denoted as a matrix of deviations $\Delta_{[\xi 1 \ldots \xi N]}$ of observed from expected counts for all possible words of length *N*. In this work tetranucleotides were used, therefore N was 4. Oligonucleotides or words are distributed in sequences logarithmically and deviations of their frequencies from expectations may be found as follows:

$$\Delta_w = \Delta_{|\xi 1 \ldots \xi N|} = 6 \times \frac{\ln\left(\frac{C^2_{|\xi 1 \ldots \xi N||obs}\sqrt{C^2_{|\xi 1 \ldots \xi N||e} + C^2_{|\xi 1 \ldots \xi N||0}}}{C^2_{|\xi 1 \ldots \xi N||e}\sqrt{C^2_{|\xi 1 \ldots \xi N||obs} + C^2_{|\xi 1 \ldots \xi N||0}}}\right)}{\sqrt{\ln\left(\left[C^2_{|\xi 1 \ldots \xi N||0} \Big/ C^2_{|\xi 1 \ldots \xi N||e}\right] + 1\right)}} \quad (1)$$

where ξ_n is any nucleotide A, T, G or C in the 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 $C_{[\xi 1 \ldots \xi N]|e} = C_{[\xi 1 \ldots \xi N]|n}$ if OU is normalized by empirical frequencies of shorter constituent words of length n. The expected count of a word $C_{[\xi 1 \ldots \xi N]|e}$ of the length N in a L_{seq} long sequence normalized by frequencies of n-mers ($n < N$) is calculated as follows:

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

Where the $F_{[\xi 1 \ldots \xi n]}$ values are the observed frequencies of a particular word of length n in the sequence and ξ is any nucleotide A, T, G or C. For instance, the expected count of a word ATGC in a sequence of L_{seq} nucleotides normalized by frequencies of trinucleotides would be determined as follows:

$$C_{ATGC} = L_{seq} \times F_{ATG} \times \frac{F_{TGC}}{F_{TGA} + F_{TGT} + F_{TGG} + F_{TGC}} \quad (3)$$

Two approaches of normalization have been exploited where the *F* values are calculated for the complete genome (generalized normalization) or for a given sliding window (local normalization).

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

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

Application of ranks instead of relative oligonucleotide frequency statistics made the comparison of OU patterns less biased to the sequence length provided that the sequences are longer than the limits of 0.3, 1.2, 5, 18.5, 74 and 295 kbp for di-, tri-, tetra-, penta-, hexa- and heptanucleotides, respectively (Reva & Tümmler, 2005).

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

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

$$RV = \frac{\sum_{w}^{4^N} \Delta_w^2}{\left(4^N - 1\right) \times \sigma_0} \qquad (5)$$

where *N* is word length; Δ^2_w is the square of a word *w* count deviation (see equation 1); and σ_0 is the expected standard deviation of the word distribution in a randomly generated

sequence which depends on the sequence length (L_{seq}) and the word length (N):

$$\sigma_0 = \sqrt{0.02 + \frac{4^N}{L_{seq}}} \tag{6}$$

In this work RV (locally normalized) and GRV (globally normalized) variances were calculated for Δ_w normalized by the frequencies of nucleotides. The only difference between locally and globally normalized parameters is that for RV the frequencies of nucleotides were calculated in sliding windows, and for GRV the frequencies of nucleotides were calculated in the complete genome. The parameter V used for SWGIS optimization was calculated as follows:

$$V = GRV / RV \tag{7}$$

2.3 Results and Discussion

2.3.1 Design and Implementation

SWGIS is a computer program developed in the Python programming language to search for foreign GIs in prokaryotic genomes. It performs searches for such genomic fragments by determining intragenomic alterations between OU patterns computed for local DNA loci and the global genome. Several routines used in this program have been developed and published recently (Reva & Tümmler, 2005; Ganesan *et al.*, 2008) with several modifications described above in the materials and methods section. The parametric modifications which were applied to the tool made a major part of this current work. In SWGIS the different types of OU parameters were abbreviated in the form type_$l_w$mer as introduced previously (Reva & Tümmler, 2005). Types are n0 if they are not normalized by a mononucleotide frequency or n1 if they are normalized by a zero-order Markov method. For example, a non-normalized trinucleotide usage pattern is represented as type: n0_3mer, whereas as a normalized pentanucleotide usage pattern is represented as type: n1_5mer. In this work n0_4mer and n1_4mer parameters were used as it was recently found that frequencies of tetranucleotides are optimal for the identification of GIs in bacterial genomes (Reva & Tümmler, 2004; Reva

& Tümmler, 2005; Ganesan *et al.*, 2008). SWGIS only allows the use of word sizes that range from 2mer to 7mer and the change of normalizations from 0 to *N*-1 for an *N*mer to perform general analysis of compositional polymorphisms of prokaryotic genomes. Apart from the identification of GIs, SWGIS consists of several other optional scenarios which allow for the identification of ribosomal RNAs, ribosomal proteins (Reva & Tümmler, 2005) and giant genes (Reva & Tümmler, 2008). However, the parameters for identifications of all these were not fully tested and optimized as the focus of this study is only on GIs and not other topics. In the case of GI predictions users are encouraged to use SWGIS in its default settings as these are already set for identification of GIs. Parametric changes may only be applied according to the findings and discussions that will follow below.

The basic principle behind the SWGIS algorithm is to apply sets of several independent statistical parameters that allow the identification of genomic fragments with alternative OU patterns relative to that determined for the core genome. These deviant genomic fragments are searched through genomes in a sliding window fashion by the use of sets of OU statistical parameters (D, PS, RV and GRV) introduced in the above section (materials and methods). In this case, GIs are identified by an alternative oligonucleotide usage (increased n0_4mer:D) with lower internally normalized OU variance (n1_4mer:RV) and an increase in globally normalized OU variance (n1_4mer:GRV). Pattern skew (n0_4mer:PS) comparisons are used to filter out *rrn* operons characterized with extreme values of n0_4mer:PS. Parameters D and PS are measured as percentages of the theoretically possible maxima (see equation 4 in the materials and methods section); and V is a numeric value (equation 7). These parameters are measured in SWGIS by the use of a sliding window approach, whereby values of genomic fragments of 8 kbp with a 2 kbp step are compared with the tetranucleotide usage values of patterns of the same type in the whole genome. If the program recognizes a statistically reliable increase of local distance n0_4mer:D values accompanied by a significant decrease of n1_4mer:RV and an increase of n1_4mer:GRV, the window shifts several steps back and repeats the analysis, this time with steps of 0.2 kbp to precisely identify coordinates of foreign inserts. The tool allows users to set their own parametric values in order to achieve acceptable false negative and false positive ratios. Several recommendations of tested values that may be used in order to achieve the latter and reduce the false positive prediction rates will be discussed later in the section.

SWGIS was developed to identify GIs in multiple genomes in a single run. It takes as input

complete bacterial genomes in GenBank (preferable) or FASTA format. Several output files are created for each genome depending on the user's selected choice of output. In general, one of these is a standard text file with extension OUT, which contains a list of identified GIs with coordinates; OU statistical values calculated for each GI; and lists of gene names within the borders of the GIs. Genes are shown only if the annotation data was included in the source / input file. In addition to the above mentioned OUT file, the program outputs FASTA files with DNA sequences for GIs of every genome used in the analysis, and if the input was in GenBank, then output files in GenBank format are created for each GI in a given genome with annotations. Lastly, users may instruct the program to create graphical SVG files of genomic atlases with indicated positions of predicted GIs. Several examples of these atlases are shown below in the case study discussion. Users may view the SVG files by the use of vector graphic editors such as Adobe Illustrator or the Mozilla Firefox browser. Alternatively, users may convert SVG files to other graphical formats by using graphical format converters which are freely available online (www.fileformat.info/convert/image/svg2raster.htm). An HTML help file on how to use SWGIS is available on the SeqWord project Web-site which can be accessed from: www.bi.up.ac.za/SeqWord/sniffer/index.html; also from the same page users may download the latest version of the program.

2.4 SWGIS parametric optimization

An empirical analysis was performed to generate optimal parametric values to be used for D and V. From the analysis, the D and V values below 1.5 showed to predict increased false positive GIs while those above 2.0 overlooked many known GIs (preliminary data not shown). The next step in the analysis involved the use of factorial analysis (St-Pierre & Weiss, 2009) to determine the optimal combination of values for D and V to ensure minimal false positive and false negative rates.

2.4.1 False negative rate calculation

The parametric measures for SWGIS were optimized to attain better predictions through the re-identification of known PAIs from PAIDB (Yoon *et al.*, 2007), which were used as training data. The optimization and re-identification process was carried out on 51 of PAIDB PAIs which were obtained from 24 micro-organisms. The latter was conducted in parallel with three of the IslandViewer incorporated programs: IslandPick, SIGI-HMM, and Islandpath in their default settings (Langille & Brinkman, 2009). Results obtained from these tools were

later compared in order to workout values for false negative rates (FNR). FNR in this instance is defined as the percentage of the known PAIDB PAIs that were overlooked by either of the programs used in the study. SWGIS was run for 4 times with different combinations of D and V: [D:1.5; V:1.5]; [D:2.0; V:2.0]; [D:1.5; V:2.0] and [D:2.0; V:1.5] during the analysis in order to determine the most optimal values to be used to attain better predictions. Results are shown in Appendix A Supplementary Table S1. From the comparison of results attained from all programs, SWGIS showed to have outperformed individual IslandViewer methods even when the most stringent threshold values [D:2.0; V:2.0] were set. Jointly the IslandViewer programs identified 69% of the 51 PAIs while SWGIS identified 88% with [D:1.5; V:1.5], 78% with [D:2.0; V:1.5], 65% with [D:1.5; V:2.0] and 63% with [D:2.0; V:2.0].

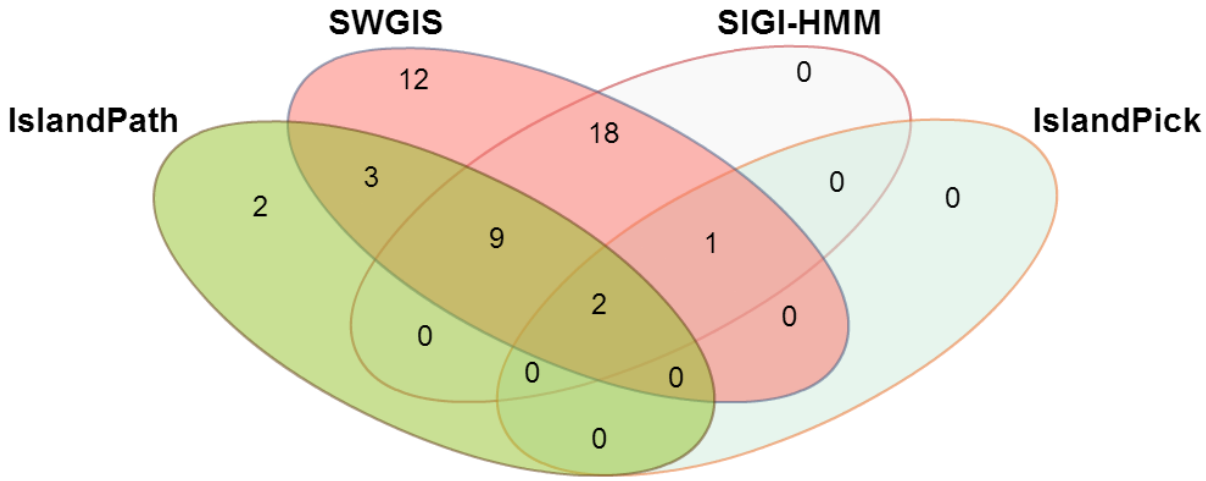


Figure 2.1: Comparison of SWGIS to the currently available genomic islands prediction methods. A Venn diagram of the 51 PAIDB PAIs that were re-identified by SWGIS, SIGI-HMM, IslandPick and IslandPath.

The total numbers of PAIs that were predicted by all the programs during analysis are summarized in the form of a venn diagram in Figure 2.1. The numbers displayed in the figure illustrate the overlapping and unique PAIs which were predicted by all four programs. SWGIS has the most overlapping and unique predictions as compared to the rest, with IslandPick having the least number of predictions. Two (SWGIS and SIGI-HMM) of the composition-based methods proved to be the most effective in predictions of the PAIDB PAIs as compared to IslandPick which utilizes a comparative analysis approach. All PAIs predicted by the IslandViewer programs except for 2 which were only predicted by IslandPath were also predicted by SWGIS [D:1.5; V:1.5]. Four PAIs were not detected by any of the 4

methods. These were overlooked by the individual sequence composition methods probably because they did not possess any compositional anomalies due to amelioration and were further missed by a comparative based method as they were found to be present in the query and reference genome sequences. The latter shows a major distinction between the different methods of GIs detection, i.e. identification by: BLAST (PAIDB identification), Nucleotide composition, and comparative genomics. FNR values determined for each program are shown in Table S1.

2.4.2 False positive rate calculation

Horizontally acquired genes which are maintained in the host are affected by the mutational pressures of their new host through a process known as amelioration (described in Chapter 1, section 1.4.2 in detail). Over time these start to resemble the codon usage properties and frequencies of constituent oligonucleotides of the native genes in their new host environment (Lawrence & Ochman, 1997; Marri & Golding, 2008). In the contrary, genomic loci comprising genes for ribosomal RNA; ribosomal protein gene clusters; giant genes; and local tandem repeats are always characterized by atypical OU patterns relative to the core sequence of the host genome (Reva & Tümmler, 2005; Reva & Tümmler, 2008). SWGIS uses a superimposition of different OU statistical parameters to distinguish between different types of such genomic loci. Operons of ribosomal RNA genes are also characterized by compositional properties of horizontally acquired GIs, the only factors that distinguishes them from the actual GIs are their increased n0_4mer:PS values. In many cases, composition-based methods mistakenly predict these operons as GIs due to their atypical features.

The setting for a more stringent value of n0_4mer:PS increases FNR, but even so it is not always possible to filter out all *rrn* operons because of compositional specificities of several bacterial genomes. However, SWGIS comes equipped with a small database of 16S rRNA sequences of bacteria from different high level taxonomic units to filter some of the *rrn* genes during the prediction process. The *rrn* operons are believed to resist horizontal transfer (Gürtler & Mayall, 2001), probably not because they cannot be mobilized by vectors, but because their transcription may be destructive to recipients. The possibility of *rrn* to undergo horizontal transfer is still discussed and cannot be ruled out as several cases of transfer possibilities have been reported (Yap *et al.*, 1999; Acinas *et al.*, 2004).

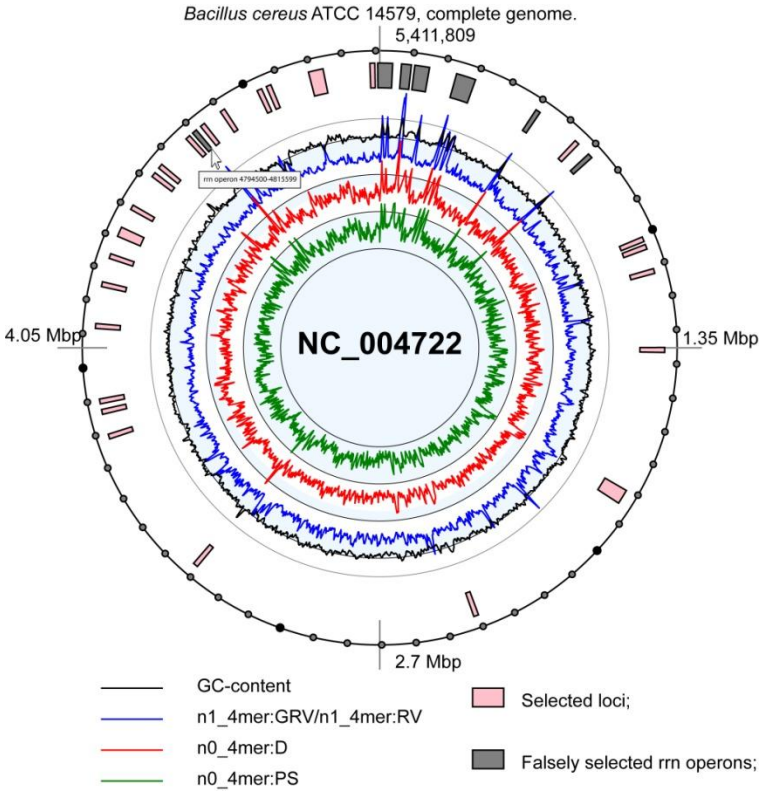


Figure 2.2. SVG of *Bacillus cereus* ATCC 14579 GIs and predicted by SWGIS. A graphical output of SWGIS in SVG format displaying positions of predicted GIs in *Bacillus cereus* ATCC 14579 [NC_004722]. Pink blocks depict GIs, whereas gray blocks depict genomic regions which comprise genes of 16S rRNA and segments that are falsely predicted. Histograms of the characteristic OU statistical parameters calculated for local windows along the chromosome are also shown.

Another issue of concern may be that the genomic regions which consist of *rrn* operons in many bacteria are favourable insertion hotspots for horizontally acquired GIs (Fernández-Gómez *et al.*, 2012; Strätz *et al.*, 1996). Their compositional similarity with those of GIs makes it very difficult to find a border which sets them apart, therefore SWGIS returns the whole block as a putative GI. These get labelled as "false positives" because of the possession of 16S rRNA genes and are subsequently searched through their annotation data for occurrences of mobile elements associated key words such as "integrase", "transposase", "phage" and "IS-element". Those which possess keywords which are associated with mobile elements are then labelled as true positives. GIs are very often not well annotated, and due to other factors such as fragmentation and deletions not all GIs comprise mobile elements associated genes; therefore the rate of falsely rejected true positives are not easy to predict. To resolve this dilemma, SWGIS returns a SVG atlas of predicted GIs in a given genome where positions of rejected GIs comprising 16S rRNA are shown in gray as in Figure 2.2. When opened in the Mozilla Firefox browser, the exact coordinates of selected regions may be retrieved by placing the mouse cursor above the coloured blocks. Users may then check

manually whether these putative GIs are true or false. The predicted GIs with *rrn* operons are currently undergoing a thorough inspection in order to develop a method that will successfully differentiate them from GIs.

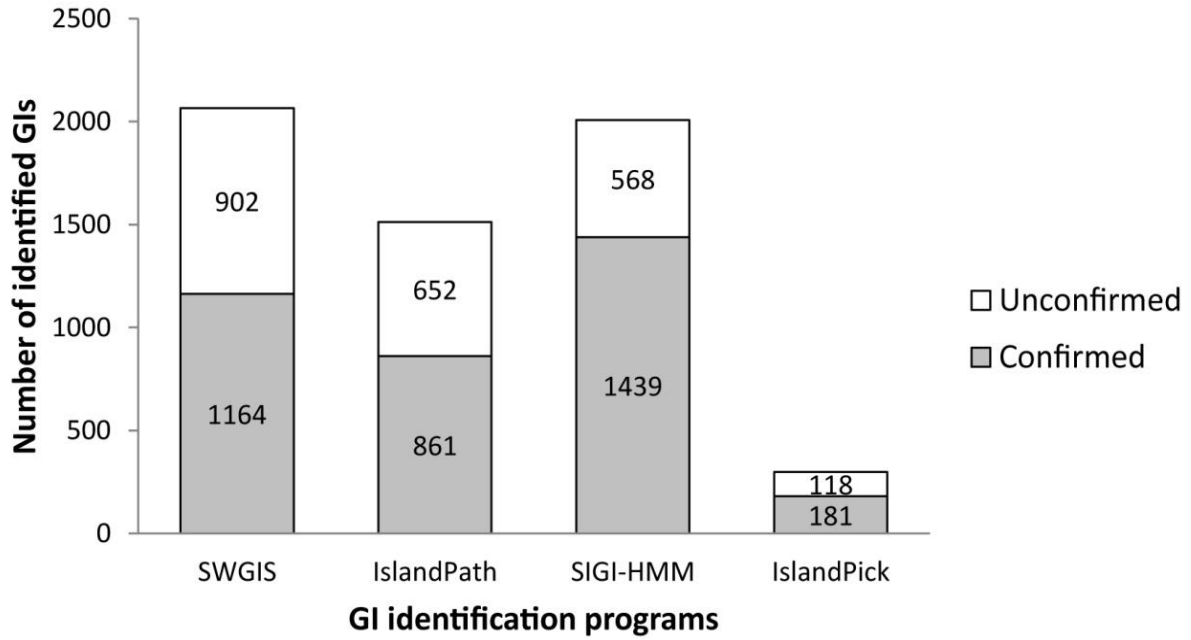


Figure 2.3. A histogram of GIs predicted by SWGIS and IslandViewer. Frequencies of GIs predicted by SWGIS together with the IslandViewer programs. The unconfirmed predictions are GIs which have been predicted only by one of the four programs, whereas confirmed are those which were predicted by two or more of the four programs.

Further GI predictions were carried out using both SWGIS and the IslandViewer tools for comparisons and to determine the rates of false positives. Contrary to FNR, rates for false positive predictions (FPR) are quite difficult to determine for GIs, as there is no any formal way to prove that a given genomic fragment has not been acquired horizontally. As FPR cannot be estimated straight away, we first calculated the statistics of unconfirmed predictions, i.e. the frequencies of GIs which were predicted only by one program and not the others. The analysis was performed using a new set of genomes obtained directly from NCBI and IslandViewer. Sets of pre-calculated GIs were downloaded from the IslandViewer web resource (www.pathogenomics.sfu.ca/islandviewer/download.php) and included in the analysis. SWGIS searched for GIs in the 169 bacterial chromosomes acquired from the NCBI bacterial genomes ftp site (see Appendix A Supplementary Table S2). These matched the genomes of the GIs which were obtained from the IslandViewer database and were analysed with sets of parameters as follows: [D:1.5; V:1.5]; [D:2.0; V:2.0]; [D:1.5; V:2.0] and [D:2.0; V:1.5]. Numbers of predicted GIs and frequencies of unconfirmed GIs for each program are

shown in Appendix A Supplementary Table S2 and summarized in Figure 2.3. Neither of the methods guarantees detection of all GIs harboured by a given genome, and that explains why great deals of unconfirmed GIs have been observed. On the other hand, many false positives are expected to be present among these unconfirmed GIs. SWGIS identified more GIs than the other methods with the following settings: [D:1.5; V:1.5], and also resulted in the highest rate of unconfirmed predictions relative to the other methods**.** For the assessment of SWGIS's performance and selecting the parametric criterion which ensures better GI predictions, we performed an estimate for FPR based on the rate of unconfirmed predictions.

Table 2.1. Prediction of GIs by SWGIS with different program run parameters and estimated FPR and FNR.

SWGIS parameters	**[*D*:1.5; *V*:1.5]**	**[*D*:1.5; *V*:2.0]**	**[*D*:2.0; *V*:1.5]**	**[*D*:2.0; *V*:2.0]**
Total GIs predicted	2066	928	1571	809
Unconfirmed GIs	902	280	545	188
Unconfirmed key positive GIs	137	44	92	28
False positive estimation*	657	201	381	138
FPR†	0.318	0.217	0.243	0.171
FNR	0.118	0.353	0.216	0.373

*Number of false positives was calculated as: "Unconfirmed GIs" – "Unconfirmed key positive GIs"×100/56;
†FPR was calculated as "False positive estimation"/ "Total GIs predicted", see Appendix A Supplementary Table S1.

To estimate the frequency of true positives in unconfirmed GIs, a search for the occurrences of mobile elements associated genes in a manner similar to the one mentioned above was performed in all gene annotation output files generated by SWGIS. Values for the estimated FNR and FPR are shown in Table 2.1. Note that in Table 2.1 FPR values are smaller than the corresponding rates of unconfirmed GIs in Table S2.

2.4.3 Optimization of parametric values by factorial experiment

Factorial experiment design [54] was applied to fit a model of two regression equations 8 and 9 estimating expected values of FNR and FPR for given D and V thresholds:

$$FNR = -0.628 + 0.118D + 0.392V \quad (8)$$

$$FPR = 0.752 - 0.121D - 0.173V \quad (9)$$

Sensitivity and specificity parameters were modified in this study taking into account that the "true negative" category of GIs cannot be calculated. Thus, the following two equations 10 and 11 were used:

$$Sensitivity = \frac{1 - FNP - FPR}{1 - FPR} \tag{10}$$

$$Specificity = \frac{1}{1 + FPR} \tag{11}$$

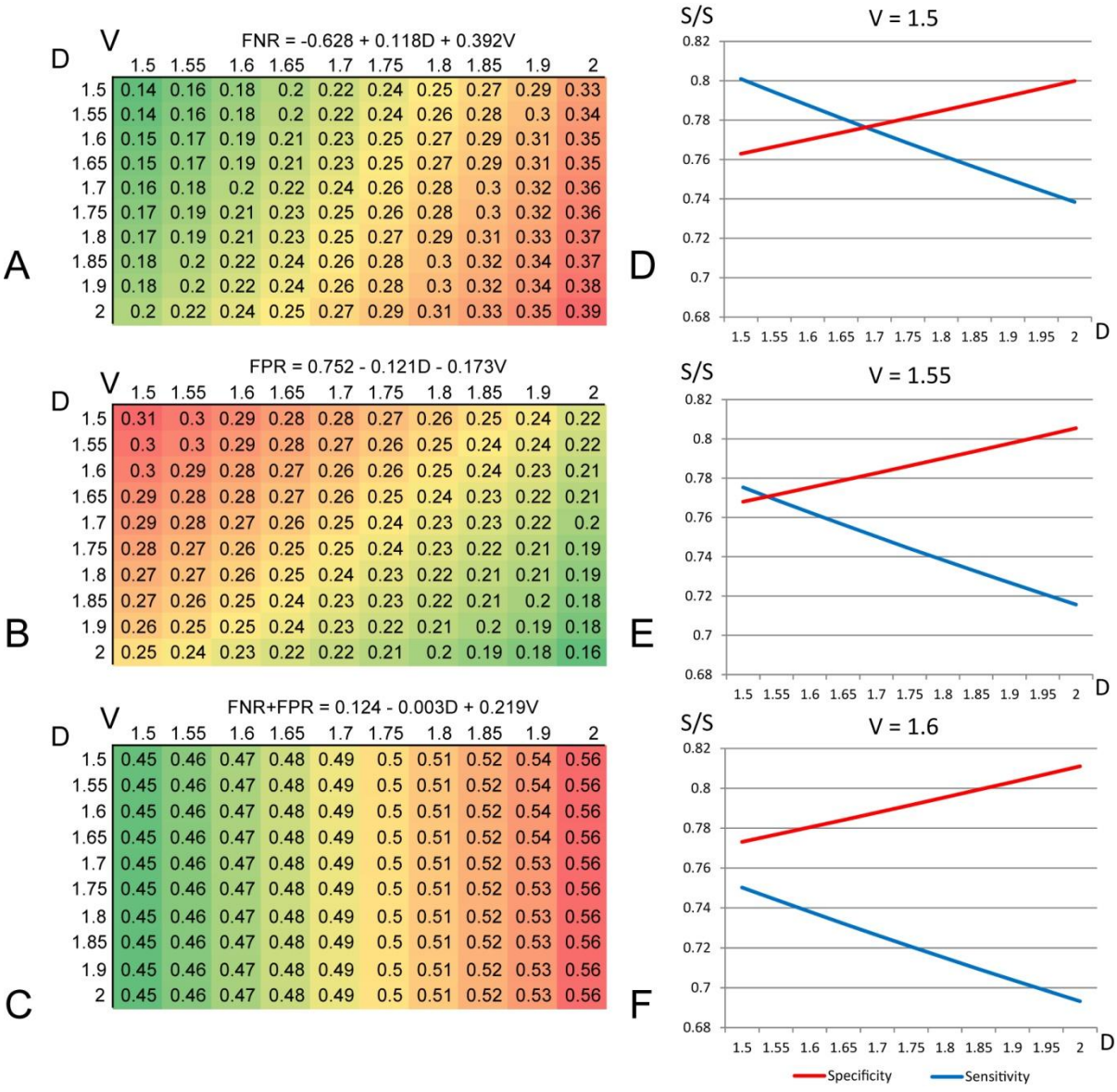


Figure 2.4: FNR and FPR calculated for different combinations of distance – D and variance – V values. Parts A and B show FNR and FPR calculated for different combinations of D and V, respectively; and their sum in the part C. Parts D, E and F represent ROC diagrams of expected specificity and sensitivity (S/S) calculated by equations 3 and 4 for variable D thresholds depicted on the horizontal axis and fixed V thresholds. Vertical axes represent specificity and sensitivity values calculated by equations 3 and 4, respectively.

Figure 2.4 A-C shows expected FNR, FPR and FNR+FPR values that are likely to be generated while different parametric combinations are in use. Colours change gradually as values move from the highest and worst (red) to intermediate (yellow) and smallest or optimal (green) prediction rates. FNR is at its lowest for [*D*:1.5; *V*:1.5] and gradually increases when it moves towards [*D*:2.0; *V*:2.0]. The latter observations indicate that there is a less chance of overlooking a GI when using the parametric values [*D*:1.5; *V*:1.5] and not [*D*:2.0; *V*:2.0] (Figure 2.4A). Although [*D*:1.5; *V*:1.5] results in smaller FNR and the highest sensitivity, it however generates an increased FPR and low specificity. The latter highlights that for one to attain a reduced FPR and increased specificity in GI prediction, a parametric set of [*D*:2.0; *V*:2.0] is the most favourable (Figure 2.4B). Changes in the cumulative FNR+FPR which depend on *D* and *V* are shown in Figure 2.4C. It was observed that an increase in *V* gradually increases FNR+FPR while a change in *D* has no effect as the increase in FNR is compensated by a similar decrease in FPR. Thus, optimization of the rates for specificity and sensitivity of GI identification by this approach may be achieved by an adjustment of *D* and keeping *V* constant and minimal.

Diagrams that illustrate the rates calculated for specificity and sensitivity with the variable and fixed values determined for *D* and *V*, respectively are shown in Figure 2.4D-F. The optimal specificity/sensitivity combination is achieved when the parameters are set for [*D*:1.7; *V*:1.5], these are currently set as default parameters for SWGIS. These parameters may be changed by users to reduce FPR or when the specificity/sensitivity of a studied genome requires a parametric adjustment as it will be discussed in the case study section.

2.5 Case study of SWGIS failures and problem solving strategies

The performance of SWGIS was further improved by re-analysing the patterns of genomes in which it performed poorly. Genomes in Table S2 are those in which SWGIS identified too many or too little GIs as compared to the IslandViewer tools. These genomes are graphically marked in the column FPR/FNR by red-leftward and blue-rightward bars depicting FNR and FPR values, respectively. FPR/FNR was calculated by equation 12:

$$FPR / FNR = \left(N_{SWGIS} - N_{IV}\right) / N_{av} \qquad (12)$$

where N_{SWGIS} is the number of GIs predicted by $SWGIS_{[D:1.5;\ V:1.5]}$; N_{IV} is the maximum number of GIs predicted by one of the IslandViewer programs and N_{av} is the average number of GIs predicted by all programs. To investigate the possible causes of failures, predictions of

GIs in several genomes were investigated.

2.5.1 False positives

In the genomes of *Bacillus cereus*; *B. anthracis*; *B. thurengiensis*; and those of the other organisms, mostly Firmicutes, SWGIS predicted much more GIs than the IslandViewer programs (Table S2). An example of the multiple GIs predicted in *B. cereus* ATCC 14579 including multiple falsely predicted *rrn* operons is shown in Figure 2.2. There is no standard way to rule out a falsely predicted GI, as there are no genes or genomic fragments which cannot undergo transfer between bacteria, at least in theory. Nevertheless, we studied these genomes and searched for the presence of common genomic properties which may explain their excessive number of GIs. As a result, these genomes showed to exhibit common compositional polymorphism. Large parts of their chromosomes are characterized by alternative GC-content and frequencies of longer nucleotides. In particular, DNA molecules in the central area of *B. cereus* chromosome are more AT rich and possess more pronounced intrinsic curvature; increased stacking energy; higher position preference; and a higher occurrence of palindromes (Bohlin *et al.*, 2012). At some point, these organisms acquired one or several giant GIs which at later stages got fragmented and spread across the chromosomes. However, the sharing of this compositional polymorphism by all sequenced *B. cereus* and their relatives of *B. thuringiensis* and *B. anthracis* may be of yet unknown biological meaning, except for their common horizontally acquired GIs. In this particular case, many of the GIs found in these microorganisms might be false positives. To avoid a high FPR, more stringent parameter settings should be set, preferably an increased D value (see Figure 2.4).

2.5.2 False negatives

Composition-based methods are customized to identify GIs as regions with atypical OU patterns in a given genome. This approach overlooks GIs which have either been acquired from sources with the OU pattern similar to that of the host organism or ancient acquisitions which have already been affected by DNA amelioration. SWGIS also suffers from such drawbacks, it fails to identify: ancient insertions; fragments whose OU patterns are indistinguishable from the core chromosomal sequences; and genomic islets (Ganesan *et al.*, 2008).

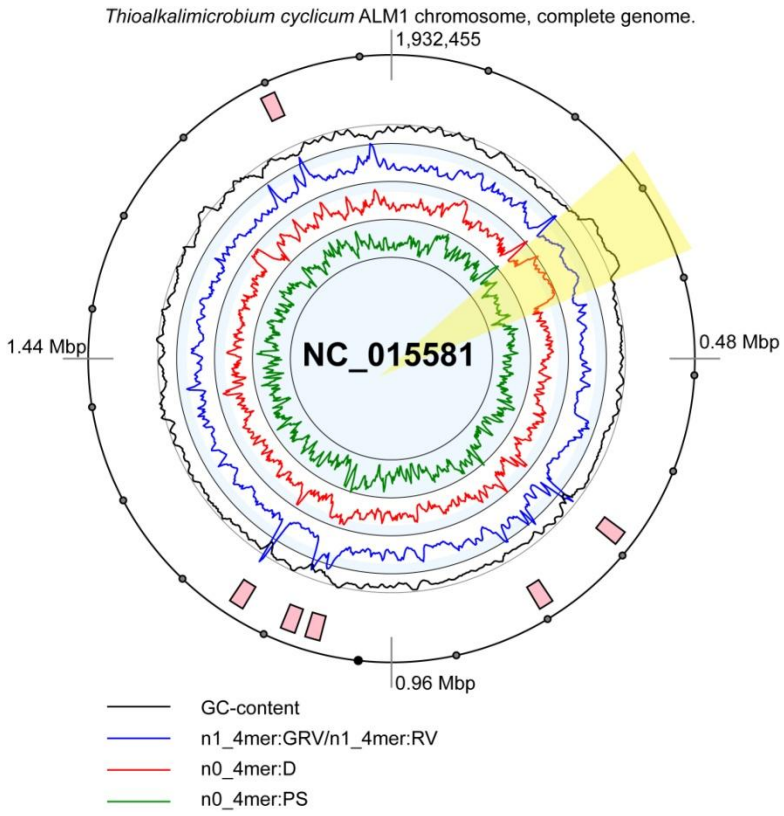


Figure 2.5: SVG representation of a giant viral gene inserted in the genome of *Thioalkalimicrobium cyclium* ALM1. An insertion of a giant viral gene that is highlighted on the atlas was overlooked by SWGIS. Similar genes are rarely subjected to horizontal transfer.

SWGIS detected only a few GIs in *Borrelia burgdorferi* B31; *Burkholderia mallei* ATCC 23344, NCTC 10229 and NCTC 10247; *Halobacterium* sp. NRC-1; *Mycobacterium ulcerans* Agy99; *Nitrobacter hamburgensis* X14; *Sphingopyxis alaskensis* RB2256; chromosome 2 of *Vibrio cholerae* O1 biovar eltor str. N16961; and *Xylella fastidiosa* 9a5c (Table S2). Predictions in these were found to be inconsistent with those of IslandViewer. The reason for failure to predict GIs in *X. fastidiosa* 9a5c is that this organism has developed a mutator phenotype which has eroded its chromosomal OU pattern specificity (Reva & Tümmler, 2004). In the contrary, there were no problems with predictions of GIs in the genome of *X. fastidiosa* Temecula1 as it possesses a stable chromosomal OU pattern.

Another example is of an overlooked GI in *Thioalkalimicrobium cyclicum* ALM1 as shown in Figure 2.5. It is a large 87,608 bp long viral filamentous hemagglutinin gene with multiple constituent repeats. This GI can clearly be seen on the genomic atlas (Figure 2.5). The reason for overlooking this region was that SWGIS considers giant genes with multiple repeats as a separate category of genomic elements with alternative OU patterns (Reva & Tümmler,

2008). The majority of these genes are not horizontally transferred. Including these giant genes by default to the SWGIS prediction output would result in too many false positives as these genes are rarely subjected to horizontal transfer.

2.6 Conclusion

Compositional comparison of bacterial genomes is a prospective approach to cope with large scale genome comparison projects. Many computational tools based on composition similarity analysis have been proposed over the past decade and proved to be useful (Abe *et al.*, 2003; Dufraigne *et al.*, 2005; Ganesan *et al.*, 2008; Chatterjee *et al.*, 2008; Lima-Mendez *et al.*, 2008; Langille & Brinkman, 2009; Wang *et al.*, 2010; Abby *et al.*, 2010; Wei & Guo, 2011; Hasan *et al.*, 2012; Ménigaud *et al.*, 2012). These showed to be reliable in the detection of GIs in complete genome sequences and binning of short metagenomic DNA reads. SWGIS employs the revised OU statistics which were previously introduced in Reva *et al.* papers (Reva & Tümmler, 2004; Reva & Tümmler, 2005; Ganesan *et al.*, 2008). It is comparable with the other composition-based methods for GI identification; particularly SIGI-HMM (Waack *et al.*, 2006), which employs Hidden Markov Models and GOHTAM (Ménigaud *et al.*, 2012) which uses both the chaos game model and codon bias statistics. SWGIS has been scaled to analyse multiple genomes in a single run and was customized to meet the user's needs. After the optimization procedures were carried out, a total of 1,674 GenBank files of bacterial chromosomes and plasmids downloaded from NCBI database (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/) were used as input to search for GIs. A total of 18,235 GIs were identified together with the 2,413 regions containing *rrn* operons. The output files from the run are available for viewing by SeqWord Sniffer GI Browser (www.bi.up.ac.za/SeqWord/mhhapplet.php). On average it took SWGIS approximately 5-10 min to analyse one bacterial chromosome. The whole task for analysing 1,674 bacterial replicons was completed in 2 weeks on a regular intel 8 core i7-2600 3.40HGZ desktop computer with 8 GiB RAM. FNR/FPR statistics were also implemented to aid with the selection of optimal parameters for GI identification. A case study was performed to investigate the failures of composition-based GI detection methods and to consider possible ways to overcome these failures. The comparison of different approaches for horizontal transfer identification is rather problematic. The predictions retrieved by different programs overlap only partly (see Appendix A Supplementary Tables S1 and S2). This discrepancy results from the extreme versatility of HGT, which occurs through three different

mechanisms: conjugation, transduction and transformation (Jain *et al.*, 2003; Hacker & Carniel, 2001). Then the integrated elements fall under the pressure of fragmentation and amelioration. Efficiency of different methods strongly depends on the lengths of GIs, their genetic content and introgression period. The best result may be achieved when the outputs of several programs are combined as it was implemented in the IslandViewer web-portal (Langille & Brinkman, 2009) and later in GIST (Hasan *et al.*, 2012). In this work it was shown that SWGIS may significantly contribute towards the identification of GIs, which in most cases remained undetected by the IslandViewer programs (Table S1 and S2).

An important issue of identification of GIs is the ability to distinguish and filter out false predictions. It has been reported that not all genomic loci showing alternative DNA compositions were horizontally transferred (Reva & Tümmler, 2004; Bezuidt *et al.*, 2011; Reva & Tümmler, 2005; Koski *et al.*, 2001), but no calculations up to now have been done to estimate the rates of false negative and false positive predictions attributed to different methods. There is no consensus at the moment which designates GIs as false positives. Prophinder (Lima-Mendez *et al.*, 2008), Islander (Mantri & Williams, 2004), SIGI-HMM (Waack *et al.*, 2006) and IslandPath/DIMOB (Hsiao *et al.*, 2003) search for genes associated with horizontally transferred GIs (transposases, integrases, viral capsid proteins, etc) to confirm the lateral origin of corresponding genomic fragments. In contrast, GOHTAM (Ménigaud *et al.*, 2012) simply returns a whole list of atypical regions found in a genome together with their annotation data, and then allow users to decide themselves which of these are horizontally transferred. SWGIS employs superposition of OU statistical parameters to distinguish between GIs and other categories of atypical genomic loci (Reva & Tümmler, 2005; Bezuidt *et al.*, 2011). It additionally performs a BLAST similarity search of the predicted DNA fragments against a database of 16S rRNA sequences to discard *rrn* operons. A drawback of all these discriminating approaches is that they unavoidably increase the percentage of overlooked GIs. The factorial analysis of the proposed GI identification algorithm was performed in this work to allow users to make an informative choice in selecting customizable parameters to ensure acceptable FNR and FPR. The latter proved to be of importance for parametric customizations of composition-based methods, and is therefore highly recommended.

Chapter 3

3 Analyses and visualization of genomic islands using composition-based approaches

3.1 Introduction

Upon the optimization of the SWGIS parameters as described in the previous chapter, we felt the need to develop several custom composition-based methods to visualize and analyse the predicted GIs to further understand their phylogenetic relations. These methods are practical for creating clusters of GIs on the basis of their shared pattern similarities and study their distributions in bacteria of different taxonomic backgrounds. GIs which share similar compositions cluster together and may possibly be used to detect their putative "donor" organisms or route of acquisition / transmission. Frequent exchanges of similar groups of GIs between different groups of bacteria however make it impossible to detect their original donor organisms as these slowly lose their native signatures due to the new host's mutational biases which act upon them. The clusters of GIs which share similar compositional features may shed light on their distribution patterns and highlight groups of organisms which prefer the acquisition or share certain types of GIs. Compositional analysis between lineages have uncovered that recently acquired GIs possess atypical DNA characteristics which are distinct from those of their host genomes (Hacker & Carniel, 2001; van Passel *et al.*, 2005; van Passel, 2011). These appear to be distinct because recently acquired GIs are known to still retain the compositions of their former hosts. Therefore GIs which arise from similar donor genomes may possess similar compositional properties, and may possibly be traced back to their probable former host organisms. The Lingvocom utility (described below in subsection 3.3.4) was developed to compare pattern distances of GIs and their host genomes in search for possible putative donors.

The aspect of base composition similarity among GIs arises from their common origin (Sueoka, 1962), i.e. from the same lineage of plasmids or phages, or from the same former host organism. Similarity is influenced by the species specific mutational pressure that acts upon the whole genome to maintain composition stability. Recently acquired GIs resemble features of the donor genome, but over time they get affected by the driving forces of the host's replication/transcription machineries (driving force) and slowly start to resemble compositional features of its native genes. Old GI inserts that have almost completely been

affected by these amelioration driving forces remain undetected by composition-based approaches as they cannot be distinguished from the native genes anymore. We however showed in the previous chapter that through the use of non-stringent parametric measures composition-based approaches may still detect some of the GIs undergoing amelioration (although completely ameliorated GIs remain undetected). In this current work we illustrate the implementation of a composition-based stratigraphic method which determines the relative acquisition periods of GIs to differentiate between recent and ancient acquisitions.

3.2 Materials and Methods

3.2.1 Source of genome sequences

Sequences of bacterial and plasmid genomes were downloaded from the NCBI FTP directories of plasmid and bacterial genomes (ftp://ftp.ncbi.nih.gov/genomes).

3.2.2 Identification and clustering of genomic islands

Analyses of horizontally transferred genomic elements were carried out using the SeqWord Genomic Island Sniffer (SWGIS) algorithm, introduced in the previous chapter. In brief, SWGIS identifies horizontally transferred genomic elements by the use of OU statistical parameters. It searches for GIs by calculating and superimposing the four OU pattern parameters: *D* – distance between local and global OU patterns; *RV* and *GRV* variances; and *PS* (pattern skew). Horizontally transferred GIs are characterized by a significant pattern deviation (large D), significant increase in GRV associated with decreased RV and a moderately increased PS (Ganesan *et al.*, 2008). The SWGIS parameters used in this study to search for the genomic islands were set as follows: [*D*:1.5; *V*:1.5]. GIs identified in the analysis were compared against one another and clustered by their OU pattern similarity. The clusters were created based on the assumption that GIs which originated from the same source share compositional similarity. Compositional similarity was measured as 100% – *D*. GIs with a pattern similarity above 75% were often found to share homologous blocks of DNA sequences. Hence a pattern similarity index of 75% was chosen as a threshold for the clustering of GIs.

3.2.3 Graphical representation of GI clusters

Graphs generated in this study were created by the *circo* algorithms of Graphviz distributed under the terms of the Eclipse Public Licence [http://www.graphviz.org/].

3.2.4 Sequence similarity comparison

Local BLAST was performed using NCBI standalone BLAST utilities to search for

homologous blocks of DNA sequences between compositional similar GIs within the clusters (Altschul *et al.*, 1990).

3.2.5 Inferring donor-recipient relations

The predictions of donor-recipient relations between bacterial genomes were determined by projecting their GIs onto a two-dimensional plot using the following equation:

$$Y = \frac{d_1^2 - d_2^2 + D^2}{2D} \qquad (1)$$

$$X = \sqrt{D^2 - Y^2}$$

$d_{1(1)}$ is the distance between OU patterns calculated for a given GI and its host genome **A**, $d_{1(2)}$ – determines the pattern distance between the same GI used in $d_{1(1)}$ against that of the counterpart genome **B**, D – determines the distance between host **A** and counterpart **B** genomic OU patterns (see illustrations in Fig. 1). $d_{2(1)}$ determines pattern distances between the GI of genome **B** against that of genome **A**, $d_{2(2)}$ is for calculating distances between the same GI in $d_{2(1)}$ against its host genome **B** (see illustration in Figure 3.1). Note that plotting of two GIs from two different genomes by equation 1 makes sense only if the distance between GIs is small (< 75% in this study), or at least if two GIs may be linked by intermediate nodes. An example of how the donor-relations may be determined using the above equation is as explained below.

Figure 3.1 illustrates how the donor recipient relationships may be determined for GIs of genomes A and B. X shows the distance between host chromosomes depicted by dark green circles. The diameters of the circles outline intergenomic variability of local OU patterns between the two studied genomes. The radius or shaded light green areas around genome circles (A and B genomes) and an intermediate gray vertical line depict the half-distance (D = 45%) between chromosomal OU patterns. GIs of the organisms A and B are shown as red and blue circles, respectively. Y values for the GIs are based on the distance values calculated between OU patterns of these GIs and their host chromosomes by the use of equation 1 above. D11is the OU pattern distance determined for the red GI and its host genome (A) whereas D12 is the distance for the red GI and its counterpart genome (B).

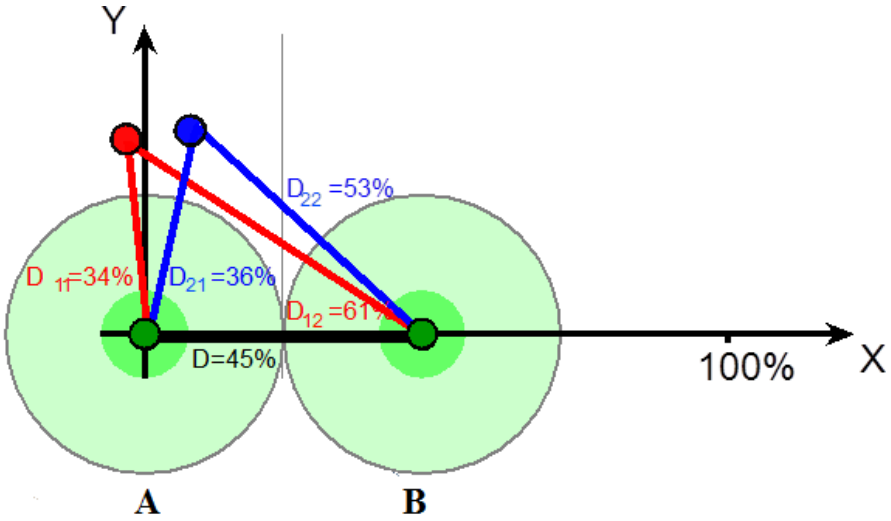


Figure 3.1: A 2D representation of donor-recipient relations between genomes and GIs. An example of a 2D based method that was applied in our work for the determination of donor-recipient relations between genomes and their constituent GIs. The method applies mainly when the genomes in comparison possess GIs whose OU patterns are shared.

The 34% distance indicates that D11 retains the composition of its host genome (A), they therefore share a similar pattern, whereas D12 indicates that the same GI deviates by 61% from the pattern of its counterpart genome (B). The blue GI (D22) has a pattern which differs from that of host genome (B) by 53%, but has a pattern which is similar to that of its counterpart genome (A) by 36%. The latter indicates that the blue GI may have been donated to genome B by A as it still retains its composition.

3.3 Results and Discussion

3.3.1 Detection of bacterial genomic islands

To facilitate a large scale analysis of bacterial genomes, SWGIS was utilized to search for GIs in a set of prokaryotic genomes representing different taxonomic classes. A total of 1,237 sequenced bacterial chromosomes and plasmids were searched for GIS using the following parameters: D >= 1.5 sigma deviation per genome; GRV/RV >= 1.5 and PS <= 55%. The PS parameter was set at 55% to set GIs apart from clusters of ribosomal RNA genes as some tend to exhibit extremely high PS values relative to the actual GIs. Predicted sets of GIs were subsequently searched for sequence similarity against a local database of 16S rRNA using nucleotide BLAST to further filter out the *rrn* operons which did not possesses increased PS values. In total, 11 870 putative GIs resulted from the filtration step and were subsequently compared against one another for shared pattern similarity in order to create graph-based clusters.

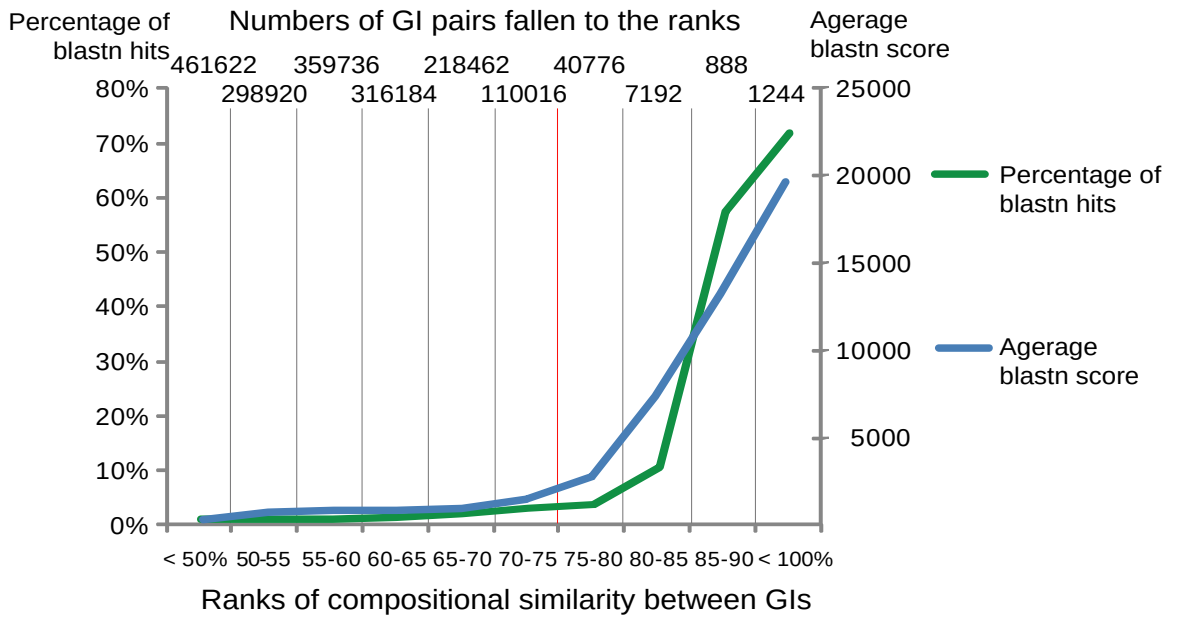


Figure 3.2: Diagram with BLASTn ranks of compositional similarity values between GIs. Percentage of BLASTn hits and average scores calculated for pairs of GIs ranked by compositional similarity values. The threshold value used for clustering of GIs is depicted by a vertical red line. Image adapted from Bezuidt *et al.*, (2011).

3.3.2 Clustering of genomic islands

Clusters of GIs were created based on their shared OU patterns where compositional similarity was measured as 100% – D. This was performed assuming that GIs that share compositional similarity may indicate a common provenance. The level of sequence similarity for homologous GIs is expected to be higher than that of a randomly selected pair of GIs. To validate our hypothesis, pairs of GIs from the predicted set were searched for similarity in composition and sequence using OU statistics and bl2seq, respectively. These pairs were further ranked by their compositional similarity values as shown in Figure 3.2. For each rank the numbers of GIs producing significant BLASTn hits (scored above 100) were counted, and also the average BLASTn scores for each rank were subsequently calculated. Similarity of randomly generated DNA sequences is expected to be around 50%. A thorough analysis indicated that the pairs of GIs which shared similarity between 50% and 70% were almost equally random. In the contrary, a further increase in shared compositional similarity between GIs is associated with a steep increase in percentage of pairs of GIs sharing sequence similarity. Compositional similarity above 90% is largely attributed to GI duplicates. Hence a pattern similarity index of 75% was chosen as an optimal threshold for creating clusters of related GIs, which may or may not share sequence similarity as determined by BLAST searches. In total 1,305 clusters were created; however, 1,158 of the

total clusters were singletons. The biggest cluster consisted of 1,119 GIs, see Figure 3.3.

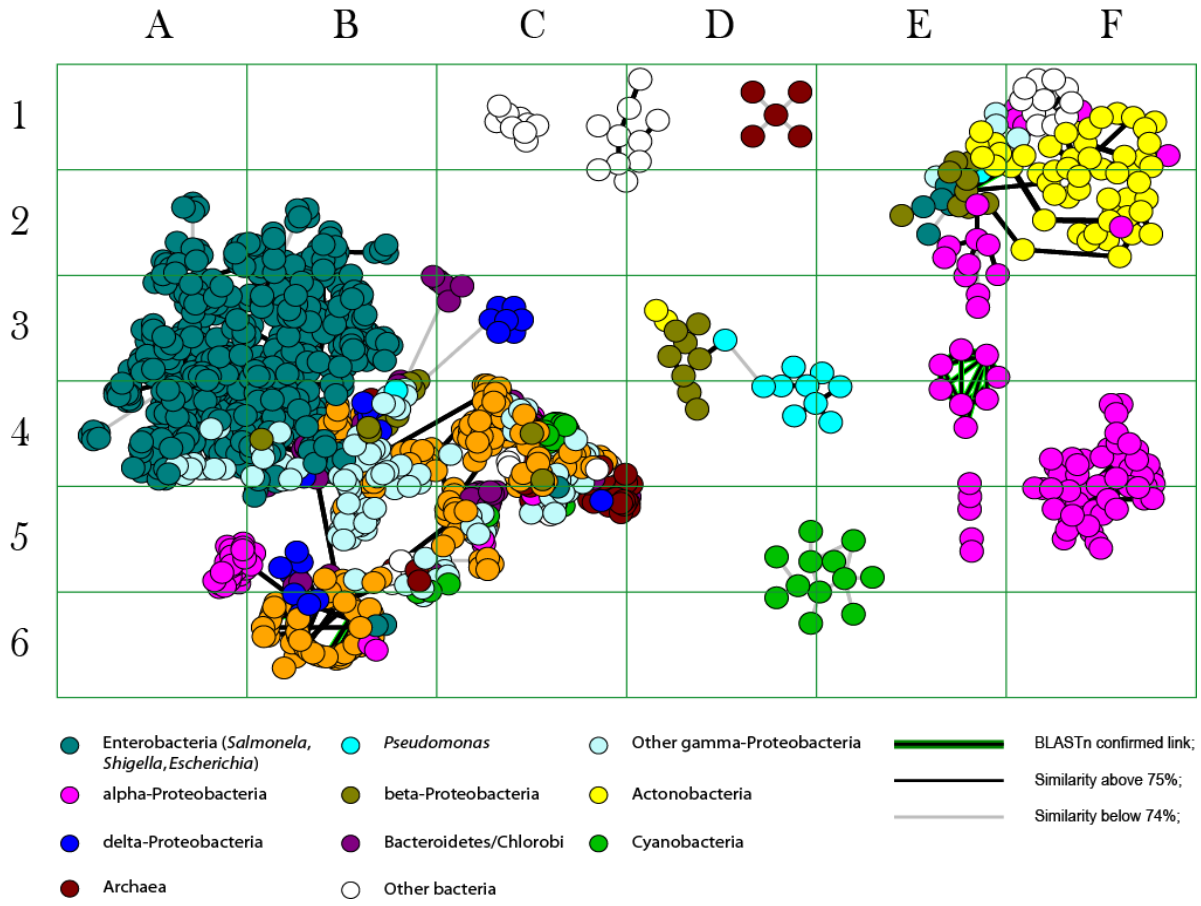


Figure 3.3: Clusters of GIs from different bacterial classes. Each node represents one GI. For more details visit the interactive map at http://www.bi.up.ac.za/SeqWord/maps/map.html (tested on Mozilla Firefox 5.0).

Although the biggest cluster was consisting of 1,119 GIs, it does not necessarily mean that all these share a similar composition, as the lesser similar groups of GIs form sub-clusters. The elements of subclusters may not share similarity in either composition or sequence with members of the other cluster as their associations are established by one or few of their intermediate elements. GIs (designated as nodes) which shared 75% compositional similarity or more were connected by edges. A custom Python script coupled with Graphviz and graph pruning criterion implementations was developed and used to create these clusters. The pruning method was implemented as follows: if three nodes in a graph are interlinked, the edge representing the smallest similarity percentage gets pruned. The script consequently determines sequence similarity between linked GIs by bl2seq (a perl script that comes with the old blastall package obtainable from NCBI) and generates a graphical output as shown in Figure 3.4. GIs with composition similarity above 75% are connected by black coloured edges whereas those with lesser similarities are connected by gray coloured edges. The black

coloured of GIs with shared sequence similarities as determined by BLAST2seq are highlighted by green colours at the edges.

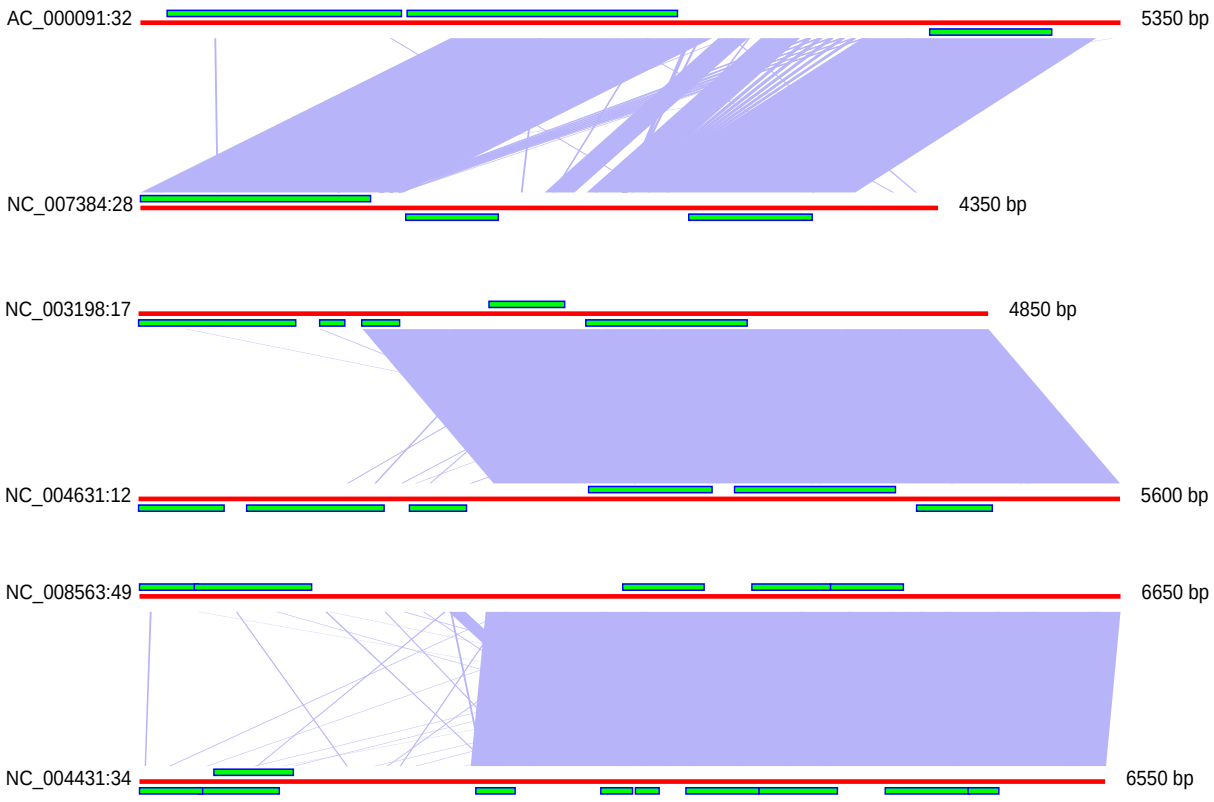


Figure 3.4: BLAST2Seq representation of three pairs of homologous GIs found in different enterobacterial genomes. Comparison of DNA sequences of three pairs of homologous GIs found in different enterobacterial genomes. Despite DNA similarity depicted by bl2seq high-scoring segment pairs, the prediction of coding genes shown as green bars is not consistent. Image adapted from Bezuidt *et al.*, (2011).

3.3.3 Stratigraphic analysis of genomic islands

Inserts of foreign DNA that have been incorporated into a genome for a longer time start to reflect the compositions of their hosts due to amelioration (Lawrence, 1997). Horizontally acquired GIs which are used by host organisms evolve much faster to optimize the transcription and translation of their constituent genes (Pál *et al.*, 2005). Selfish genes and non-coding sequences do not get affected by amelioration as much as the functional genes; however their compositions also change over time (see section 3.1 above). We hypothesize that the comparison of GIs of the same origin which are distributed in organisms which belong to the same species would result in different *D*-values as a result of their different acquisition periods. Through the application of the stratigraphic method, one might be able to tell the differences in acquisition periods of these individual GIs, and also tell which of the organisms were first to acquire certain groups of GIs. In Figure 3.5 the differences in *D*-

values are depicted by a gray colour gradient. The darker colours (towards 50) depict GIs which have been acquired recently, while the lighter colours (towards 15) depict ancient acquisitions – see the Figure 3.5 gradient for colours and values.

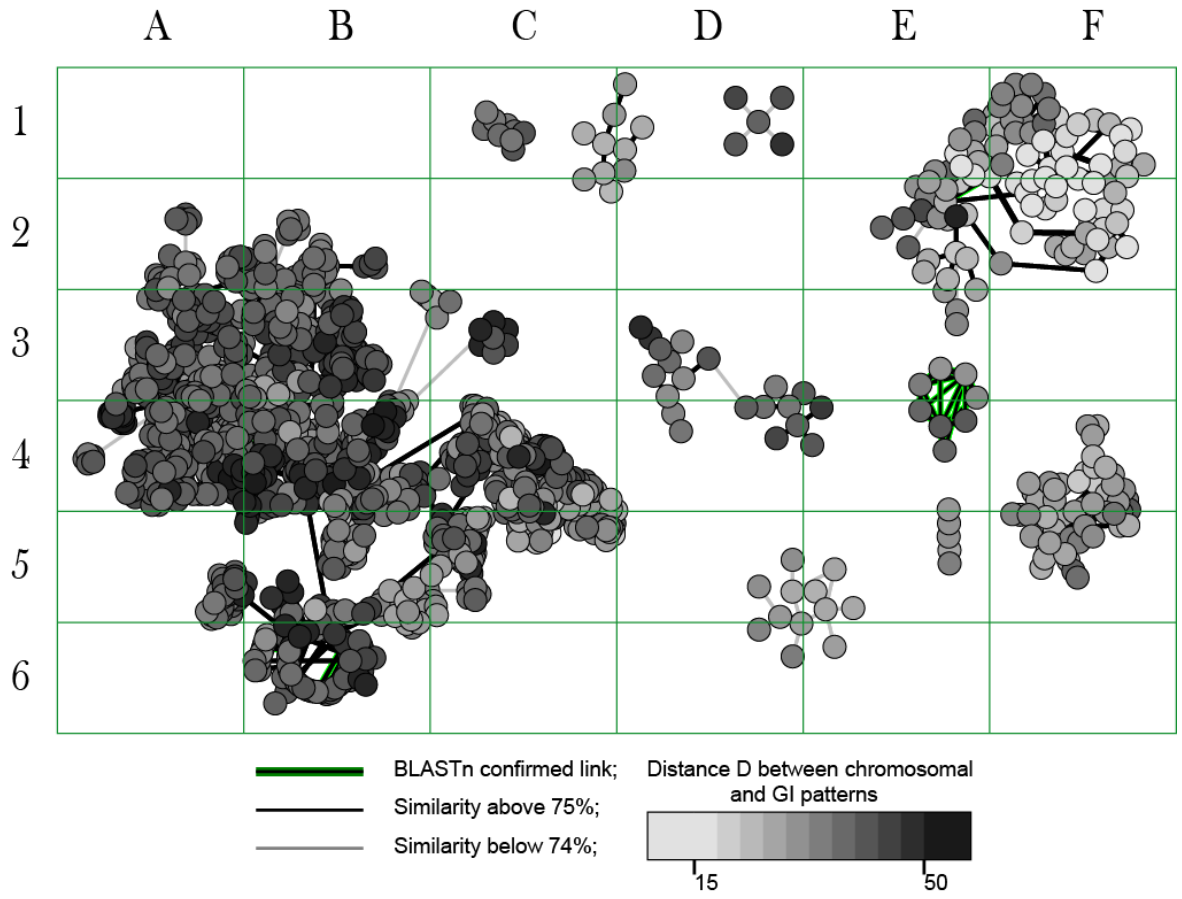


Figure 3.5: Stratigraphic analysis of GI inserts. Each node represents one GI. For more details visit the interactive map at http://www.bi.up.ac.za/SeqWord/maps/map.html (tested on Mozilla Firefox 5.0).

The GIs as mentioned earlier in the previous section are depicted as nodes, whereas their associations are illustrated by edges. The dark gray nodes denote GIs that still strongly exhibit the signatures of their former host organisms as their OU pattern distances highly deviate from those of their new hosts. The intermediate and lighter gray nodes are GIs that have been acquired for longer periods and are therefore slowly losing their native signatures to resemble those of their current hosts due to amelioration. Although these have lighter gray colours, it does not necessarily mean that they have completely lost their native compositions, they may still be matched with their former hosts. The problem regarding determination of donor-recipient relations only arises when GIs have completely ameliorated to an extent where their compositions cannot be differentiated from those of the neighbouring native genes in host genomes. The method also allows the possibility to create associations between

GIs of different acquisition periods. These offer possibilities of determining their routes of transmission and to also study the relationships of their hosts.

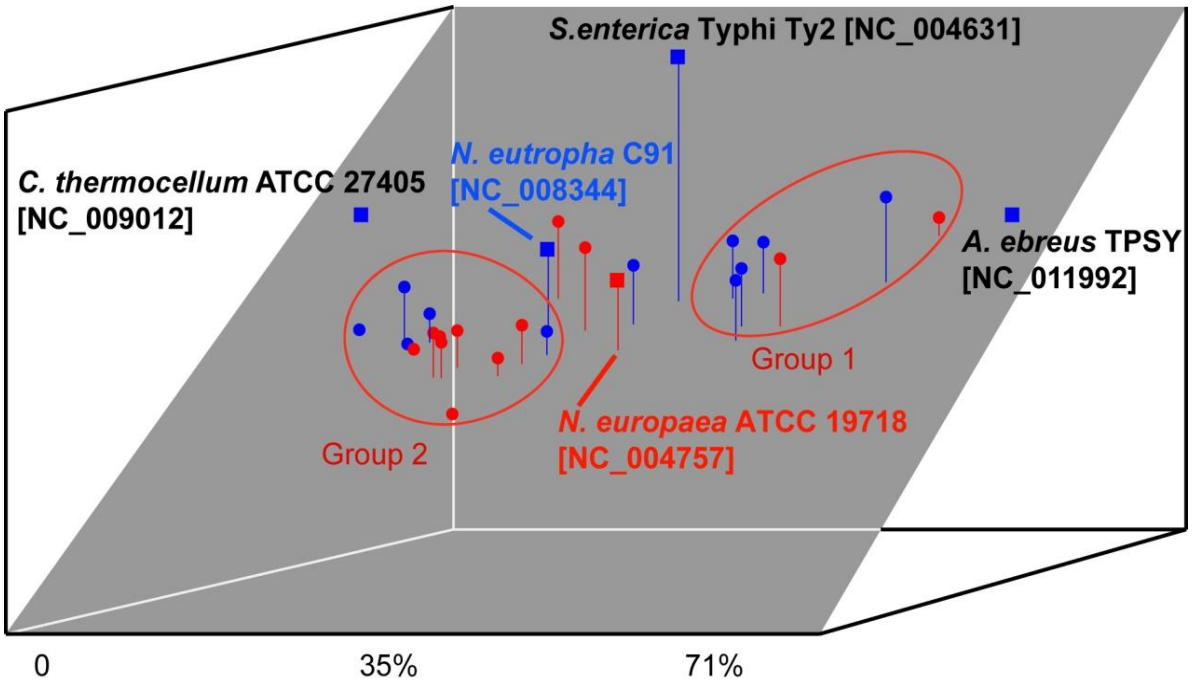


Figure 3.6: A 3D projection of the OU patterns determined for the two *Nitrosomonas* genomes, their GIs and the three outgroup genomes of *S. enterica*, *C. thermocellum* and *A. ebreus*. GIs are depicted by red (from *N. europaea* ATCC 19718) and blue (*N. eutropha* C91) circles; whereas the chromosomes are depicted by squares. Two groups of Nitrosomonas' GIs with similar patterns are outlined and encircled.

3.3.4 Further analysis of identified GIs by LingvoCom

The prediction of GI coordinates in prokaryotic genomes is only the first phase of studying horizontal gene transfer events. While a number of computational methods for predictions of GIs already exist, there is lack of those to further study ontological relationships between compositionally similar GIs, reconstruct their distribution patterns, and identify their possible origins. The SWGIS program is accompanied by several in-house composition-based tools to allow further analysis of its predicted GIs. One such tool is LingvoCom, a collection of utilities for the analysis of genome linguistics in DNA sequences with ranges from small fragments to complete bacterial genome sequences. This tool was developed by members of the SeqWord research group using the Python programming language. It takes as input raw GenBank or FASTA files or may be utilized to analyse the predicted GIs in GenBank or FASTA formats as generated by SWGIS. Alternatively, it may be ordered to extract DNA fragments from a given genome by the use of user defined genomic coordinates. The tool may also be used by other programs as a Python module or can be given arguments through a

command line. This utility is freely available, its download link and user-specifications may be accessed from the LingvoCom web-site (www.bi.up.ac.za/SeqWord/lingvocom).

3.3.5 Grouping of GIs by OU pattern similarity

LingvoCom offers functions such as the "3D-plot" and "d-matrix" for the grouping of GIs or other DNA sequences by compositional similarity and build phylogenetic trees, respectively. The 3D-plotting function is an implementation of the nonmetric multidimensional scaling (MDS) algorithm (Legendre & Legendre, 1983; Legendre & Legendre, 2012). As an example of how the function may be utilized, files for *Nitrosomonas europaea* ATCC 19718 [NC_004757] and *Nitrosomonas eutropha* C91 [NC_008344] in Genbank format and the coordinates of their 12 and 11 respective GIs as predicted by SWGIS were used as input for LingvoCom to create plots. In addition to these files, three genomes of *Salmonella enterica* subsp. *enterica* Typhi Ty2 [NC_004631], *Clostridium thermocellum* ATCC 27405 [NC_009012] and *Acidovorax ebreus* TPSY [NC_011992] were included in the analysis for composition comparisons between their genomes and *Nitrosomonas* GIs. The *Nitrosomonas* GIs were compared against each other; their hosts; and three other genomes included in the analysis for composition similarity to create association plots. GIs of the two *Nitrosomonas* strains were grouped by their shared compositional similarity into two clusters, which possibly illustrate different routes of origination. There is also a chance that these organisms may have possibly donated GIs to each other as two of the NC_004757 GIs share a high pattern similarity with NC_008344. Among these, all GC-rich GIs shared OU pattern similarity with *Acidovorax* while AT-rich GIs shared similarity with *Clostridium*. OU pattern calculated for *Salmonella* is equally distant from GIs of *Nitrosomonas* and their chromosomes. An SVG output file that was created by LingvoCom which depicts a 3D projection of GIs and chromosomes used in the analysis is shown in Figure 3.6. The parameters and commands used may be checked on the LingvoCom Web-site (www.bi.up.ac.za/SeqWord/lingvocom).

The "d-matrix" function constructs a matrix file of distances calculated for OU patterns of DNA sequences. The distance matrix file is saved in Phylip format and may be used immediately to calculate a phylogenetic tree by using Phylip's *neighbour.exe*, *fitch.exe* or *kitsch.exe* (http://evolution.gs.washington.edu/phylip.html). The distance matrix is also suitable for other types of analysis including principal component (PCA) and principal coordinate (PCoA) analysis (Legendre & Legendre, 1983; Legendre & Legendre, 2012);

however, reformatting of the Phylip matrix file may be needed depending on the third-party program in use. A Neighbour Joining tree which was created based on the matrix of distances calculated for OU patterns of *Nitrosomonas* GIs; their chromosomes together with that of *Salmonella enterica* subsp. *enterica* Typhi Ty2 as the outgroup is shown in Figure 3.7. The LingvoCom d-matrix and 3D-plot functions have both demonstrated that the GIs of *Nitrosomonas* may have been acquired from two different sources as they have both resulted in the formation of two groups of compositionally similar GIs. Further analyses by the LingvoCom utilities may provide the possibilities to identify possible donors of such groups of GIs.

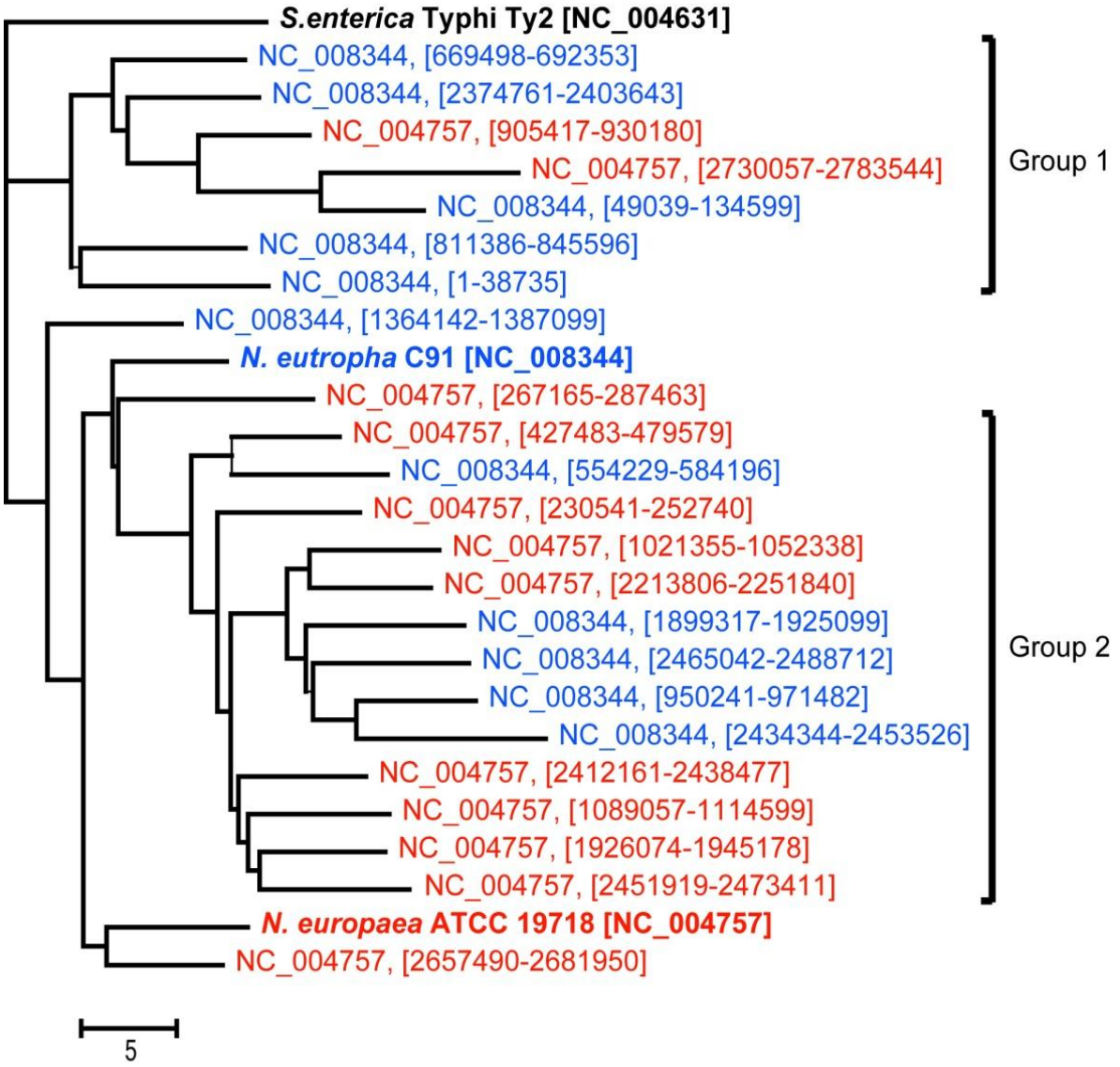


Figure 3.7: A dendrogram representation of two groups of *Nitrosomonas*' GIs based on the distance matrix of D-values (see equation 1). The N. eutropha C91 genome and it GIs are colored in blue whereas N. europaea and its GIs are colored in red.

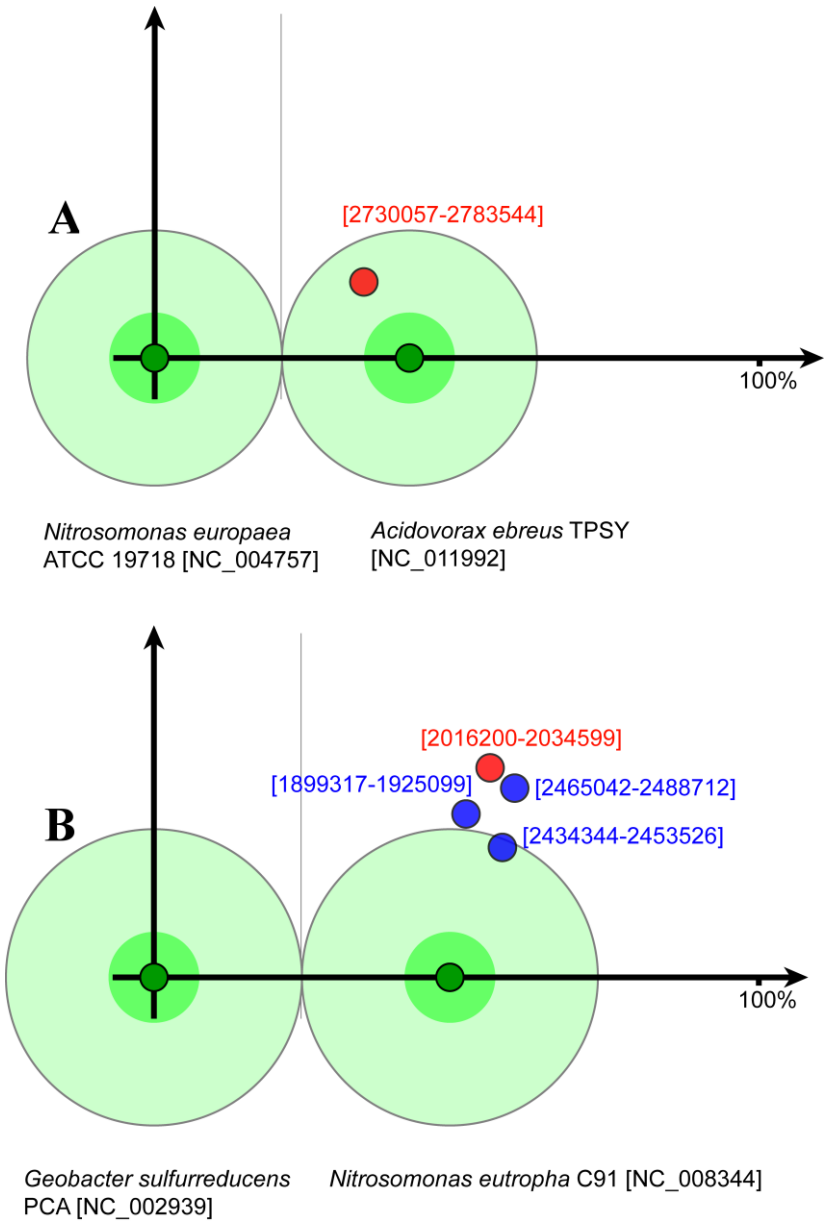


Figure 3.8: A 2D projection of the donor-recipient relations of GIs and genomes of *N. europaea* ATCC 19718, *A. ebreus* TPSY, *G. Sulfurreducens* PCA and *N. eutropha* C91. This method is used to determine donor-recipient relations between GIs and groups of organisms which share a common OU pattern. A) Depicts that *Acidovorax* is a possible donor of one GI found in *N. eutropha* ATCC 19718; B) Depicts *N. eutropha* C91's possibly ameliorated GIs (blue circles), and a GI (red circle) of *G. sulfurreducens*, which is possibly of *N eutropha* C91 origin.

3.3.6 Donor-recipient relationships

As a result of genome amelioration, GIs resemble OU patterns of both their donor organisms and new hosts (Marri & Golding, 2008). A while after the immobilization of a GI, properties of the donor organism slowly diminish in order to start reflecting those of the recipient organism. Some of the *Nitrosomonas* GIs (red and blue circles) in Figure 3.6 are drawing closer towards the OU patterns of their host chromosomes (red and blue squares). Those GIs,

which on the diagram are located closer to the *Nitrosomonas* chromosomes, designate older inserts, and the ones which are distant are recent acquisitions for they still retain OU properties of their donor genomes. The 3D plotting function is much more instrumental in terms of creating associations between groups of compositionally similar GIs to compare against several groups of organisms in search for their putative donors. If either of the GIs show a close association with any given genome in terms of shared composition similarity, these get further looked at to determine donor recipient relationships. Figure 3.8A shows the OU pattern similarity conducted for the GI [2,730,057-2,783,544] of *N. europeae* ATCC 19718 and its possible donor – *A. ebreus* as their close associations were first illustrated in the 3D-plot depicted in Figure 3.6 earlier. Two dark green spots on the Figure 3.8A 2D plot represent OU patterns of the query (at the centre point) and subject (on the horizontal axis) genomes. Light green circles depict the ½'d distance between OU patterns calculated for the two analysed genomes. GIs of the query genome are shown as red small circles and those of the subject genome, as blue circles. The OU pattern of GI [2730057-2783544] of group 1 designated in Figure 3.6 as the only red circle in the far right corner is much more similar to *Acidovorax* than to its host chromosome. The latter may be an indication that this GI may be originating from the *Acidovorax* lineage (Figure 3.8A). A *Geobacter sulfurreducens* GI [2016200-2034599] shares compositional similarity with the *Nitrosomonas* group 2 GIs. A 2D projection of OU patterns calculated for these GIs and their host chromosomes (Figure 3.8B) indicates that *Nitrosomonas* is possibly the donor of horizontally transferred genes for *G. sulfurreducens*, or that GIs in both organisms originate from a common source. The possible pathways of distribution of GIs including PAIs in bacteria and their relative times of insertion were analysed by this approach in our recently published work (Thomas & Nielsen, 2005).

3.4 Conclusion

This work illustrates the importance and practical use of composition-based approaches to study the distribution patterns of horizontally transferred genomic elements in prokaryotes. These are practical for creating clusters of GIs generated from different sources; estimate the relative time of GI insertions; and reconstruct donor-recipient relations between GIs and their host genomes just implemented in the LingvoCom utilities. Donor-recipient relations may be determined if two bacterial genomes with diverse OU patterns contain GIs sharing similar OU patterns. The utilities of the LingvoCom package allow further analysis of predicted GIs to infer their phylogenetic interrelations. The interrelations studies are practical for predicting

putative origins and estimating the relative acquisition periods of GIs. Output files generated by SWGIS may be used as an input for LingvoCom to create an analytical pipeline. The LingvoCom tool is easy to use, platform independent and scalable for the analysis of multiple input files.

The idea of applying alignment free composition-based genome comparison methods towards a phylogenetic inferring and clustering is not new (Ménigaud *et al.*, 2012; Chapus *et al.*, 2005; Volkovich *et al.*, 2010; Cheung *et al.*, 2011). The biggest problem affecting all composition-based phylogenetic inferences including those conducted by LingvoCom and all the other currently available computational tools is the absence of an adequate evolutionary model which can explain the effects and changes of genomic OU patterns occurring in different evolutionary time scales. The only model for amelioration of bacterial DNA proposed by Lawrence and Ochman in 1997 (Lawrence & Ochman, 1997) is rather basic, and according to an independent study this was shown not to be sufficiently accurate (Wang, 2001). Apart from the absence of such evolutionary models, the use of composition-based approaches have primarily proved to be useful towards the reconstruction of donor-recipient relationships between micro-organisms exchanging DNA fragments. The only other program apart from LingvoCom that allows the search for possible donor organisms is GOHTAM (Ménigaud *et al.*, 2012). It compares the frequencies of tetranucleotides calculated for the query sequences against a large database of patterns pre-calculated for all publicly available genomes. Prediction of possible origins of horizontally acquired genomic elements may be improved by the combination of SWGIS, LingvoCom and GOHTAM, whereby the predicted GIs (FASTA files) of SWGIS may be submitted directly to GOHTAM for analysis to further compare with LingvoCom.

Chapter 4

4 Mainstreams of horizontal gene exchange in enterobacteria: consideration of the outbreak of enterohemorrhagic *E. coli* O104:H4 in Germany in 2011

Oliver Bezuidt[1], Rian Pierneef[1], Kingdom Mncube[1], Gipsi Lima-Mendez[2,3] and Oleg N. Reva[1*]

[1]University of Pretoria, Dep. Biochemistry, Bioinformatics and Computational Biology Unit, Pretoria, South Africa.

[2]Laboratoire de Bioinformatique des Ge´nomes et des Re´seaux (BiGRe), Universite´ Libre de Bruxelles, CP263, Bvd du Triomphe, 1050 Bruxelles, Belgium

[3] Current address: Department of Applied Biological Sciences, Faculty of Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium/ VIB, Belgium

Adapted from PLoS One (2011) , 6 (10), e25702.

4.1 Abstract

4.1.1 Background

The enterohemorrhagic *Escherichia coli* O104:H4 caused a severe outbreak in Europe in 2011. The strain TY-2482 sequenced from this outbreak allowed the discovery of its closest relatives but failed to resolve ways in which it originated and evolved. On account of the previous statement, may we expect similar upcoming outbreaks to occur recurrently or spontaneously in the future? The inability to answer these questions shows limitations of the current comparative and evolutionary genomics methods.

4.1.2 Principal Findings

The study revealed oscillations of gene exchange in enterobacteria, which originated from marine γ-Proteobacteria. These mobile genetic elements have become recombination hotspots and effective ‘vehicles’ ensuring a wide distribution of successful combinations of fitness and virulence genes among enterobacteria. Two remarkable peculiarities of the strain TY-2482 and its relatives were observed, these are: i) retaining the genetic primitiveness by these strains as they somehow avoided the main fluxes of horizontal gene transfer which effectively penetrated other enterobacteria, ii) and acquisition of antibiotic resistance genes in

a plasmid genomic island of β-Proteobacteria origin which ontologically is unrelated to the predominant genomic islands of enterobacteria.

4.1.3 Conclusions

Oscillations of horizontal gene exchange events were reported to have resulted from a counterbalance between the acquired resistance of bacteria towards existing mobile vectors and the generation of new vectors in the environmental microflora. We hypothesized that TY-2482 may originate from a genetically primitive lineage of *E. coli* that has evolved in confined geographical areas and was brought by human migration or cattle trade onto an intersection of several independent streams of horizontal gene exchange. Development of a system for monitoring the new and most active gene exchange events was proposed.

4.2 Introduction

The evolution of pathogenic microorganisms is generally linked to horizontal gene exchange (Hacker & Kaper, 2000). Identification of the laterally acquired genomic islands (GIs) and the reconstruction of gene exchange events in a historical perspective is a very difficult task. However, there is a necessity to resolve this problem in order to help understand the nature in which disease causing genes arise and get distributed in bacteria. The recurrent outbreaks of pathogens that possess new virulence factors and broad range antibiotic resistance gene cassettes reflect the importance of horizontal gene transfer (HGT) in the evolution of pathogenic bacteria. In many cases, the evolution of pathogens is mediated by mobile genetic elements (MGEs). These can easily be interchanged between bacterial taxa inhabiting the same or different environments. Outbreaks of gastrointestinal and nosocomial infections seem to be the ones which take a heavy death toll.

In 2011 over 80 people of whom 27 died have reportedly been infected by a multidrug-resistant strain of *Klebsiella pneumoniae* at the Maasstad Hospital in Rotterdam, Netherlands (Potron *et al.*, 2011). Three biologists in the same hospital were held accountable for failing to control the spread of the disease as they had underestimated the seriousness of the outbreak (Sheldon, 2012). Earlier in the same year, an *Escherichia coli* O104:H4 outbreak which was initially reported in North of Germany resulted in 1,730 infections and 18 deaths (Brzuszkiewicz *et al.*, 2011). The recurrences of such instances illustrate our limited knowledge on the basic principles of the evolutionary trends of bacterial pathogens. Strains of

the microbes which were reported to be involved in these outbreaks were identified in only a few days through the use of next generation sequencing techniques. The recent advances in high throughput sequencing techniques and bioinformatics algorithms offer a platform to conduct a large scale genome analysis (Langille & Brinkman, 2009). Although these techniques make it possible to attain sequences in a short time span and conduct large scale analysis, the development of tools with effective data mining in complete genomes still lag behind.

The North Germany outbreak occurred in May 2 (2011) where a rare enterohemorrhagic *E. coli* O104:H4 caused haemolytic-uremic syndrome. The infection spread fast through many other European countries and sickened thousands of people. This strain showed an increased level of lethality associated with the production of Shiga toxin and resistance against many antibiotics. Several other isolates (TY-2482, LB226692, 01-09591, GOS1 and GOS2) from the same outbreak were later identified, sequenced and annotated (Brzuszkiewicz *et al.*, 2011; Manrique *et al.*, 2011; Mellmann *et al.*, 2011). Based on the unique combination of genomic features these strains were suggested to represent a new pathotype: Entero-Aggregative-Haemorrhagic *E. coli* (EAHEC) (Brzuszkiewicz *et al.*, 2011). In our study, strain TY-2482 was analysed in greater detail as it was among the first emerged strains of the outbreak to be sequenced and made available to public. The presence of several new virulence determinants in this strain which are absent in its closely related counterpart: *E. coli* strain 55989, – a causative agent of the 2002 outbreak in central Africa, – suggested the involvement of HGT in its evolution. This could illustrate that strain TY-2482 harbours sets of virulence determinants similar to those of strain 55989 and ones which have been acquired from different sources. The newer sets of PAIs were investigated in order to identify their putative sources; associations with PAIs from various bacterial organisms; and further understand their contributions in virulence.

The ability of microbes to cause diseases in their hosts has always been linked to acquisitions of virulence genes. Factors such as toxins, adhesins, polysaccharide capsule synthesis proteins and iron uptake system are mainly located in accessory genomic regions known as pathogenicity islands (PAIs). These genomic regions were first identified in the human pathogenic *E. coli* species and have since been known to carry virulence-associated genes (Oelschlaeger *et al.*, 2002). The effects of such genes are triggered mainly by altered environments and nutritional signals (Somerville & Proctor, 2009). These may also be

dependent on the genetic background of the host organism, which includes the organization of the core genome and presence/absence of other virulence determinants (Escobar-Páramo *et al.*, 2004). The acquisitions and distributions of PAIs are not restricted to a certain group of organisms, as these may be scattered throughout different groups of bacteria, related or not. The same virulence determinants from *E. coli* may be found in *Shigella* and *Salmonella* just as many other homologous GIs have been reported to be present in even more distant organisms such as *Yersinia* and *Bordetella*. In order to fully understand the complexities of GI distribution patterns in enterobacteria and determine mechanisms which underlie the virulence of *E. coli* TY-2482, innovative composition-based approaches were applied. These are as follows: identification of GIs; clustering of GIs by OU pattern similarity; stratigraphic analysis of GIs to determine their relative acquisition time; determination of donor-recipient relations between bacteria and their constituent GIs. The latter innovative composition-based measures were used in this study for they have showed to be of practical use as illustrated in the previous chapters (2 and 3).

4.3 Materials and Methods

4.3.1 Source of genome sequences

Sequences were downloaded from the NCBI FTP directories of plasmid and bacterial genomes (ftp://ftp.ncbi.nih.gov/genomes). The draft sequence of TY-2482 was obtained from the FTP site: ftp://ftp.genomics.org.cn/pub/Ecoli_TY-2482.

4.3.2 Identification and analysis of genomic islands

SeqWord Genomic Island Sniffer (SWGIS) was used for the detection of genomic islands (GIs). The following are sets of parameters which were used during the analysis: D >= 1.5 sigma deviation per genome; GRV/RV [V] >= 1.5 and PS <= 55%. The GIs identified in the study were further analysed for donor-recipient relations using LingvoCom, a collection of utilities which allow for the further analysis of GIs with shared compositional similarities and the search for their potential donor organisms.

4.3.3 Graphical representation of GI clusters

The GI composition and stratigraphy clusters generated in this study were calculated by the *circo* algorithms of Graphviz distributed under the terms of the Eclipse Public Licence [http://www.graphviz.org/].

4.3.4 Sequence similarity comparison

Local BLAST was performed by using the NCBI standalone BLAST utilities.

4.3.5 Markov Clustering Algorithm

Protein coding sequences located in GIs were clustered using Markov Clustering Algorithm (MCL) introduced by van Dongen as a graph-based, deterministic, partitional algorithm which incorporates hard clustering where flow or movement in a graph is simulated by algebraic methods (Van Dongen, 2008). MCL was used with an inflation parameter of 1.8. Similarity scores for creating clusters were obtained with an initial all-against-all BLASTp alignment and an *E*-value cut-off of 0.0001. The MCL requires a degree of relatedness for clustering, thus bit-scores were deemed a favourable indication of similarity. MCL preformed sufficiently and it consequently clustered the coding sequences into clearly defined classes. Co-occurrence of genes from different clusters which constituted the same GIs was checked by a Chi-square test (Fleiss *et al.*, 2013). Functional similarity between predicted gene modules was calculated as described previously (Lima-Mendez *et al.*, 2008).

4.4 Results

4.4.1 Genomic islands identification

To facilitate a large scale analysis of bacterial genomes, SeqWord Genomic Island Sniffer (SWGIS) was utilized. Amongst the 1,237 sequenced bacterial chromosomes which were searched for GIs by SWGIS as mentioned in chapter 3, was a draft genome of *E. coli* TY-2482, one of the first reported causative agents of the 2011 Germany *E. coli* O104 outbreak. The 11,870 putative GIs which resulted from the rRNA filtration step were searched for shared composition and sequence similarity with the GIs of enterobacteria to create associations. The draft genome of *E. coli* TY-2482 was predicted to harbour seventeen GIs and sixteen of these shared composition similarity with those in *E. coli* 55989. Strain 55989, a causative agent of the 2002 central Africa outbreak is currently considered to be an ancestor of the newly isolated *E. coli* TY-2482 as they both share over 90% nucleotide identity (Rohde *et al.*, 2011). The strain was isolated from a stool sample of an HIV infected adult who suffered from a persistent watery diarrhea and may lack some of the determinants which are present in TY-2482 (Mossoro *et al.*, 2002).

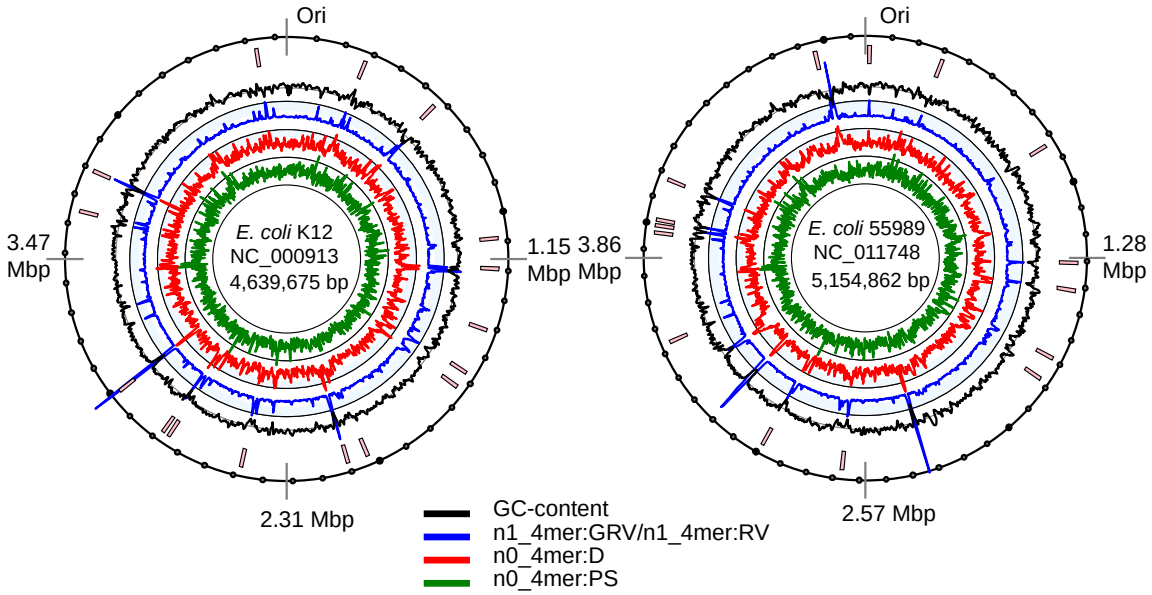


Figure 4.1: Graphical representation of GIs in commensal and enterohemorrhagic E. coli. Distribution of GIs in the genomes of commensal *E. coli* K12 and enterohemorrhagic *E. coli* 55989 as predicted by SeqWord Sniffer with parameters set as D >= 1.5 sigmas and GRV/RV >= 1.5.

Figure 4.1 depicts positions of GIs predicted in commensal *E. coli* K12 and enterohemorrhagic *E. coli* 55989. The chromosomes of these two *E. coli* strains possess approximately the same number of GIs positioned in more or less similar genomic regions (Figure 4.1). This is in agreement with an observation previously reported that gene uptake and loss in *E. coli* take place at precisely the same locations known as integration hotspots (Touchon *et al.*, 2009). Thus, a general overview of the presence of similar GIs possessed by both pathogenic and non-pathogenic bacteria may not aid much in understanding why some strains are virulent and others are not.

4.4.2 Clustering of genomic islands

Groups of GIs from different bacteria were compared with those of enterobacteria in search of shared compositional similarities. These were subsequently clustered by the use of an in-house python Graphviz tool with a measure of distance: 100% – D. The cluster in Figure 4.2 mainly represents GIs of enterobacteria together with few from distantly related bacteria. These can be viewed from the SeqWord project interactive web-page (www.bi.up.ac.za/SeqWord/). The figure illustrates that the *Escherichia* GIs share the highest similarity with those of *Shigella, Salmonella* and several other more distant genera. These GIs are grouped into several sub-clusters, many of which are polyphyletic, meaning that they comprise GIs found in organisms belonging to different genera.

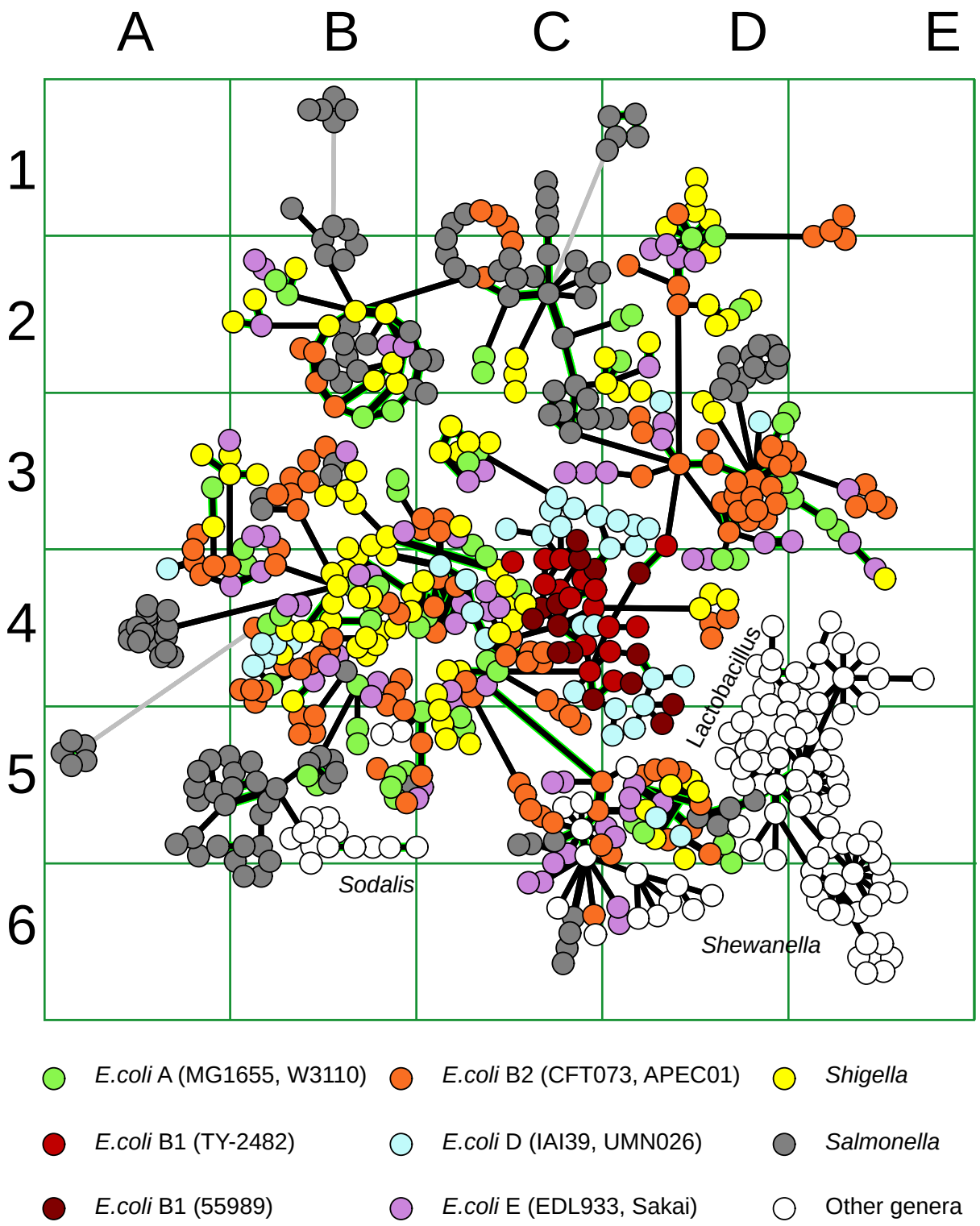


Figure 4.2: Clusters created from the GIs that share OU pattern similarity. Each node corresponds to one GI. Two nodes of GIs that share 75% or more OU pattern similarity are linked by a black edge. Gray edges represent similarity below 75%. Links between nodes of GIs which were confirmed by BLASTn to share similarity in sequence are highlighted in green.

GIs of *E. coli* of the phenotypic groups A, B2 and E share the same sub-clusters with GIs of *Shigella* and very often with those of *Salmonella*. On the contrary, GIs of *E. coli* strains TY-

2482 and 55989 (group B1) are clustered all together in cells C4-D4 in Figure 4.2. The GIs of *E. coli* in group D are moderately variable and cluster with different bacterial groups.

4.4.3 Stratigraphic analysis of genomic islands

Figure 4.3 depicts a similar cluster as in Figure 4.2 but with nodes illustrated in variable gray colours that correspond with the ones on the distance gradient displayed at the bottom of the figure. The different gray colours are measures of composition distances between GIs and their host genomes that were obtained using distance - D. The darker the colour of the GI node the more distant it is from its host in terms of composition. GIs which are distant from their hosts are recent acquisitions and still possess the signatures of their previous sources. The lighter coloured GI nodes resemble compositions which are closely similar to those of their hosts; they therefore result in smaller D-values as they have been acquired for longer periods.

The stratigraphic analysis indicates that the GIs of enterobacteria were acquired not simultaneously but sequentially in different spans of time, and many of these seem to have been acquired for longer periods. The central part of the graph (cell C4) contains the most ancient inserts. By comparing the graphs in Figures 4.3 and 4.4 one may conclude that the genomes of *E. coli* B1 contain predominantly old GIs which were inherited from an ancestor that is also common for *Shigella*. The only relatively recent acquisition is GI #12 of *E. coli* 55989 (in Figure 4.1 counted clockwise from the chromosomal origin of replication) and its corresponding island in *E. coli* TY-2482 which encodes a secreted autotranspoter toxin. This toxin is an important virulence factor for uropathogenic *E. coli* (Guyer *et al.*, 2002) and may explain to some extend the pathogenicity of strains TY-2482 and 55989. It however cannot be brought to conclusion that this toxin had a major influence on the outbreak caused by TY-2482. Figure 4.3 illustrates that this GI is not a recent acquisition as it has been present in more ancient *E. coli* strains. The most recently acquired GIs of enterobacteria are grouped in cells C5-D6, and these comprise LEE and SPI-10 PAIs. Other recent acquisitions are in cells D3-E3 (SHI-1, SHI-7, SRL, PAI I and PAI II) and A4 (SPI-4, SPI-8). The stratigraphic analysis reveals a distribution pattern / or movement of GIs from organisms with old inserts to those with newer ones. In order to get a much clearer picture of how these may be distributed, a further analysis of GIs that exhibit different acquisition periods and interlinked by their shared composition similarities may be required.

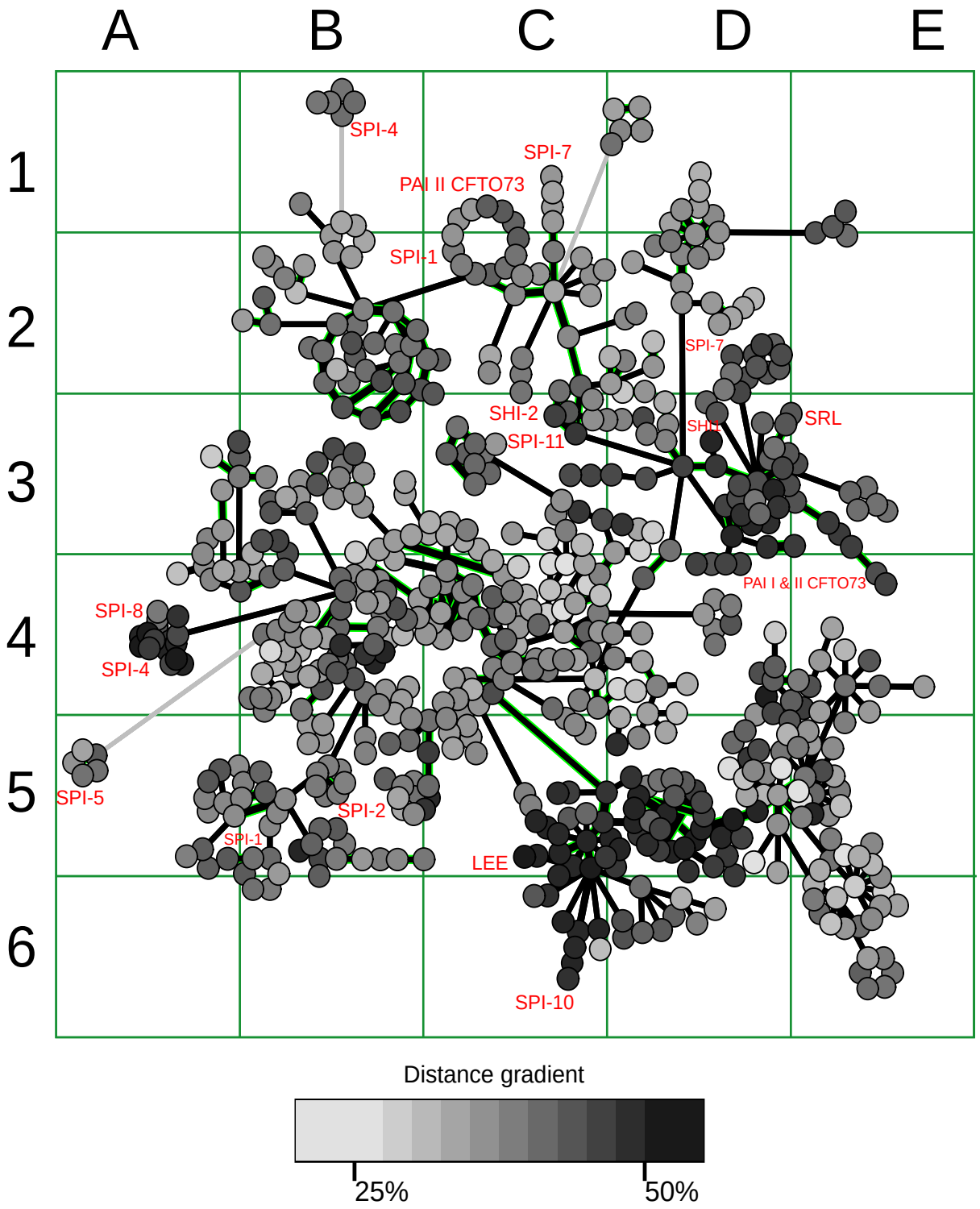


Figure 4.3: Stratigraphic analysis determined for enterobacterial GIs. Layout of GIs is the same as in Fig. 2. Gradient colours depict divergences of GI patterns from the complete genome patterns of their host organisms. The older inserts are depicted by lighter colours, as their OU patterns are much closer to their hosts. Known pathogenicity islands from PAI DB are mapped on the plot.

4.4.4 Donor-recipient relations

Donor-recipient relationships between microorganisms determine the sharing / transfer of GIs which are predominant in one organism and not the other. Such groups of GIs get dispersed across phylogenetically diverse organisms that dwell in the same / different habitats through the help of vectors. Phages and plasmids have been reported to have a major contribution towards horizontal gene exchange. Vectors such as conjugative plasmids contain hundreds of functional genes which allow their transfer to new hosts in single evolutionary events (Thomas & Nielsen, 2005). In many instances plasmids are seen to comprise phage integrases. These allow the possibilities to actively integrate genetic cassettes into chromosomes of unrelated organisms (Klockgether *et al.*, 2007) and mobilize fragments they excise from random organisms which they sequentially infect. Plasmids undergo the same amelioration process as other horizontally transferred GIs. They acquire traits of their most predominant host organisms and consequently start to reflect their composition. Composition comparisons among GIs and their hosts offer the possibility to establish donor-recipient relations and determine a pattern of HGT. As mentioned in the stratigraphic analysis section, nodes of GIs from different sources which are interlinked based on their shared compositions provide possibilities to study donor-recipient relations between their hosts.

The following GIs: *Shewanella denitrificans* OS217 [NC_007954:14] and *Salmonella enterica* ATCC 9150 [NC_006511:5] were chosen as subjects for the donor-recipient relationships analysis as they share similarities in both sequence and composition. In Figure 4.3 these are represented by two linked nodes in cell D5 with different acquisition periods. The *Shewanella denitrificans* GI is a much more ancient acquisition as compared to the recent one for *Salmonella enterica.* The pattern distances calculated for both host chromosomes and their GIs are shown in Figure 4.4A. Both GIs exhibit pattern distances which are closer to the *Shewanella* chromosome. This organism was therefore designated to be the donor of GI NC_006511:5 harboured by *Salmonella*. The determination of donor-recipient relations is reasonably possible for host organisms whose patterns are substantially dissimilar just as illustrated with *Shewanella* and *Salmonella*. The donor-recipient relations for a pair of homologous GIs from *Salmonella* and *Sodalis* (Figure 4.4B; these GIs in Figure 4.3 are in cell B5) were hard to determine, as their pattern distances deviate from both hosts. It was therefore assumed that the GI movement may have possibly been from *Salmonella* to *Sodalis* rather than the other way round. The donor-recipient relations determined for

Shewanella and *Sodalis* confirmed *Shewanella* to be the donor (Figure 4.4C).

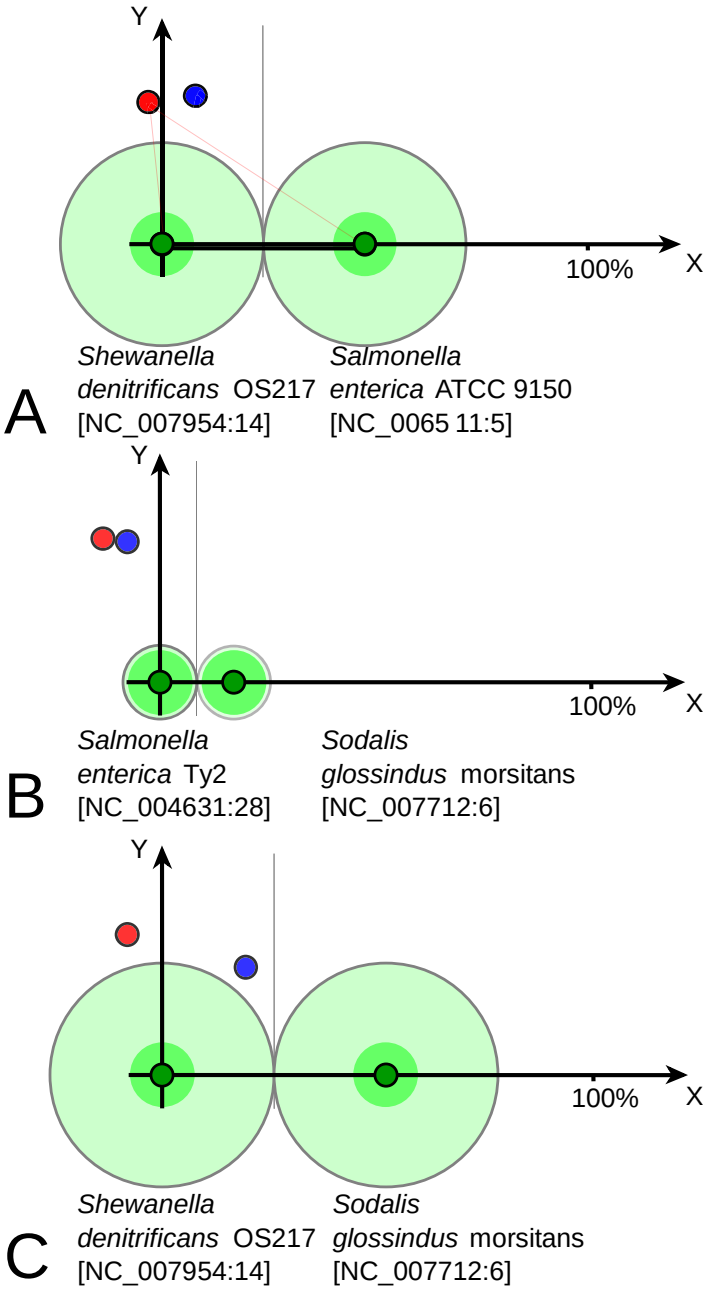


Figure 4.4: Donor recipient relations determined for *S. enterica* ATCC 9150, *S. enterica* Ty2, *S. denitrificans* OS217 and *S. glossindus* morsitans. A) Comparison of GIs NC_007954:14 and NC_006511:5 from *S. denitrificans* OS217 (3177400..3199999) and *S. enterica* ATCC 9150 (858550..886299). B) Comparison of GIs NC_004631:28 and NC_007712:6 from *S. enterica* Ty2 (2847050.. 2866799) and *S. glossindus morsitans* (1186850..1242349). C) Comparison of GIs NC_007954:14 and NC_007712:6 from *S. denitrificans* OS217 and *S. glossindus* morsitans.

4.4.5 Categories of genes distributed by horizontal transfer

Protein coding sequences of enterobacterial GIs in Figure 4.2 were clustered using MCL - Markov Clustering Algorithm. The MCL preformed sufficiently and it consequently clustered the coding sequences into favourable classes of functional categories. The top 22 clusters each comprising more than 50 genes from different GIs were annotated and categorized into functional groups (Table 4.1). Their functional importance and contributions towards the virulence and fitness of enterobacteria will be discussed in detail below.

Table 4.1. Annotation of genes of the top 22 MCL clusters.

Cluster #	Gene annotation	Functional group	Number of genes
1	ABC-transporters, ATP-binding proteins	Transport	252
2	Transposases	Selfish phage and plasmid related	213
3	Histidine kinase sensor response regulators	Translation and transcription regulation	162
4	Glycosyl transferases	Polysaccharide and O-antigen biosynthesis	153
5	IS1, iso-IS1	Selfish phage and plasmid related	153
6	Glucose epimerases and dehydratases	Polysaccharide and O-antigen biosynthesis	129
7	IS600 transposases	Selfish phage and plasmid related	93
8	Fimbrial proteins	Membrane and adhesive proteins	88
9	IS2 transposases and integrases	Selfish phage and plasmid related	85
10	Fimbrial periplasmic chaperons	Membrane and adhesive proteins	78
11	GGDEF diguanylate cyclases	Translation and transcription regulation	77

12	Outer membrane usher proteins	Membrane and adhesive proteins	75
13	Integrases	Selfish phage and plasmid related	73
14	Aminotransferases	Polysaccharide and O-antigen biosynthesis	71
15	Dehydrogenase reductases	Enzymes of unknown specificity	67
16	Acetyle transferases	Polysaccharide and O-antigen biosynthesis	65
17	RHS family proteins	RHS family proteins	64
18	Resolvases and recombinases	Selfish phage and plasmid related	63
19	Thymidylyl and uridylyl transferases	Polysaccharide and O-antigen biosynthesis	63
20	IstB transposases	Selfish phage and plasmid related	60
21	Transcription regulators	Translation and transcription regulation	53
22	IS1 ORF1	Selfish phage and plasmid related	52

GIs often comprise one to several genes of the same cluster. Chi-square test showed that the frequency of co-occurrences of genes from different clusters is on the level of a random combination except for the clusters 4, 6, 7 and 20 (all are of the polysaccharide and O-antigen biosynthesis functional group) showing a tendency for co-occurrence in GIs and clusters 11 and 13 which almost always form a pair of fimbrial periplasmic chaperon and outer membrane usher protein in GIs.

Distribution of genes of different clusters and functional groups in enterobacterial GIs is shown in Figure 4.5. Groups of ABC-transporters and RHS family proteins are represented by clusters 1 and 18 (Table 4.2), respectively. The group ‘Outer membrane and fimbrial

proteins' is a combination of clusters 9, 11 and 13; and the group 'Polysaccharide biosynthesis proteins' is a combination of clusters 4, 6, 7, 15, 17 and 20. Phage and plasmid associated transposases, integrases, resolvases, helicases and IS-elements are most widely distributed in all GIs, illustrating their roles in transmissions of GIs in bacteria. Classes of selfish mobility genes are not depicted in Figure 4.5 as they are present in almost every GI. Transcriptional and translational regulators are very important in virulence development. Acquisition or loss of GIs often has complex effects on the host bacterium, including regulatory effects on many core genes (Ritter *et al.*, 1995). GGDEF diguanylate cyclases may have a profound effect on the bacterial behavioural response towards environmental stimuli (Sun *et al.*, 2011). Diguanylate cyclases additionally form the messenger molecule cyclo-di-GMP, and this is directly involved in the microbial lifestyle and regulation of biofilm formation.The three other groups of genes which were found to appear in the enterobacterial PAIs in addition to the ones defined in Table 4.1 are as follows: efflux pump proteins, type III secretion proteins and invasive pili.

ABC-transporters are the second most abundant protein-coding genes in all sequenced genomes following transposases (Aziz *et al.*, 2010). Transporters could be required by bacteria to take up substances from the environment or to pump them out when unwanted and can be very useful for bacteria to adapt and survive in a new niche. These genes are however not distributed randomly among GIs as they create functional modules that are required by organisms in relation to the substrates with which they occur. A separate group of transporters in cell C4 (Figure 4.5) comprises iron uptake and transport proteins. A well developed iron uptake system is a prerequisite for virulent and commensal bacteria (Ejrnæs, 2011; Garénaux *et al.*, 2011). All these GIs are old inserts acquired by the *E. coli* ancestor. Outer membrane and fimbrial proteins are also important virulence factors (Oelschlaeger *et al.*, 2002; Rowley *et al.*, 2011). Genes encoding outer membrane proteins have been reported to be prevalent in bacterial chromosomes and plasmids by Nogueira *et al.*, (2009). These authors associated mobile outer membrane proteins with cooperative trait determinants important for shaping the microbial social behaviour of both pathogenic and commensal bacteria. Fimbrial biosynthesis is performed through a chaperone/usher dependant pathway whereby an operon encodes at a minimum three different proteins: a fimbrial subunit, a chaperone and an usher. These fimbrial operons represent hypervariable DNA regions and are frequently involved in HGT as observed in enterobacteria. Outer membrane and fimbrial proteins are a prerequisite for initial adhesion to the host epithelium.

Polysaccharide biosynthesis contributes to biofilm formation and O-antigen variability of the cell surface. Genes involved in O-antigen and polysaccharide biosynthesis occur as clusters on chromosomes with evidence of interspecies transfer among *Shigella*, *Salmonella* and *Escherichia* (cells D5-E5, B2, C2 and B4 in Figure 4.5). The contributions of O-antigens towards the virulence and pathogenicity of enterobacteria is broadly discussed in literature (Wang & Reeves, 1998; Plainvert *et al.*, 2007; Liu *et al.*, 2008; Lukáčová, 2008). Recombination or retrotransposon hotspot (RHS) protein family has been shown to be involved more with ligand-binding and chromosomal cell-surfaces rather than with rearrangements (Hill *et al.*, 1994; Jackson *et al.*, 2009). These proteins constitute several relatively recent HGT in *Escherichia* (cells E3-E4) and *Salmonella* (cell A4 in Figure 4.5).

Type III secretion proteins, invasive pili and efflux pumps are the most notorious virulence factors of enterobacteria (Bugarel *et al.*, 2011; Karasova *et al.*, 2010; Ogura *et al.*, 2009; Rendón *et al.*, 2007). Despite an obvious sequence similarity of type III secretion proteins present in different GIs, DNA compositional analysis showed that these virulence factors evolved independently for several periods by gene recombination. Type III secretion system in the SPI-2 PAI of *Salmonella* (cell B5) is well characterized (Karasova *et al.*, 2010; Bhowmick *et al.*, 2011). This is a relatively old gene insert in contrast to the type III secretion proteins of the PAI LEE in *Escherichia* which are of recent horizontal gene exchange events (C5-D5 in Figure 4.4). Much older inserts of type III secretion proteins were found in *E. coli* GIs shown in cell C4 of Figure 4.5. Genes encoding invasive pili are present in 3 separate clusters in cells B5, B4 and C1. These have most likely evolved independently by gene rearrangements. Efflux pump proteins are involved in drug resistance (Turner *et al.*, 2003). There are different families of genes that encode efflux pump proteins. Many of these are distributed among GIs of enterobacteria but not all of them are depicted in Figure 4.5. Cells B3-C4 depict a group of conserved GIs which comprise a pair of efflux proteins denoted as multidrug resistant proteins K and Y that are associated with a conserved two component sensor transcriptional regulator. All these GIs are old inserts harboured by *E. coli* of different phenotypic groups. A significant sequence conservation of these genes shows that these are important for survival of *E. coli* and also specialized with regard to the function they contribute.

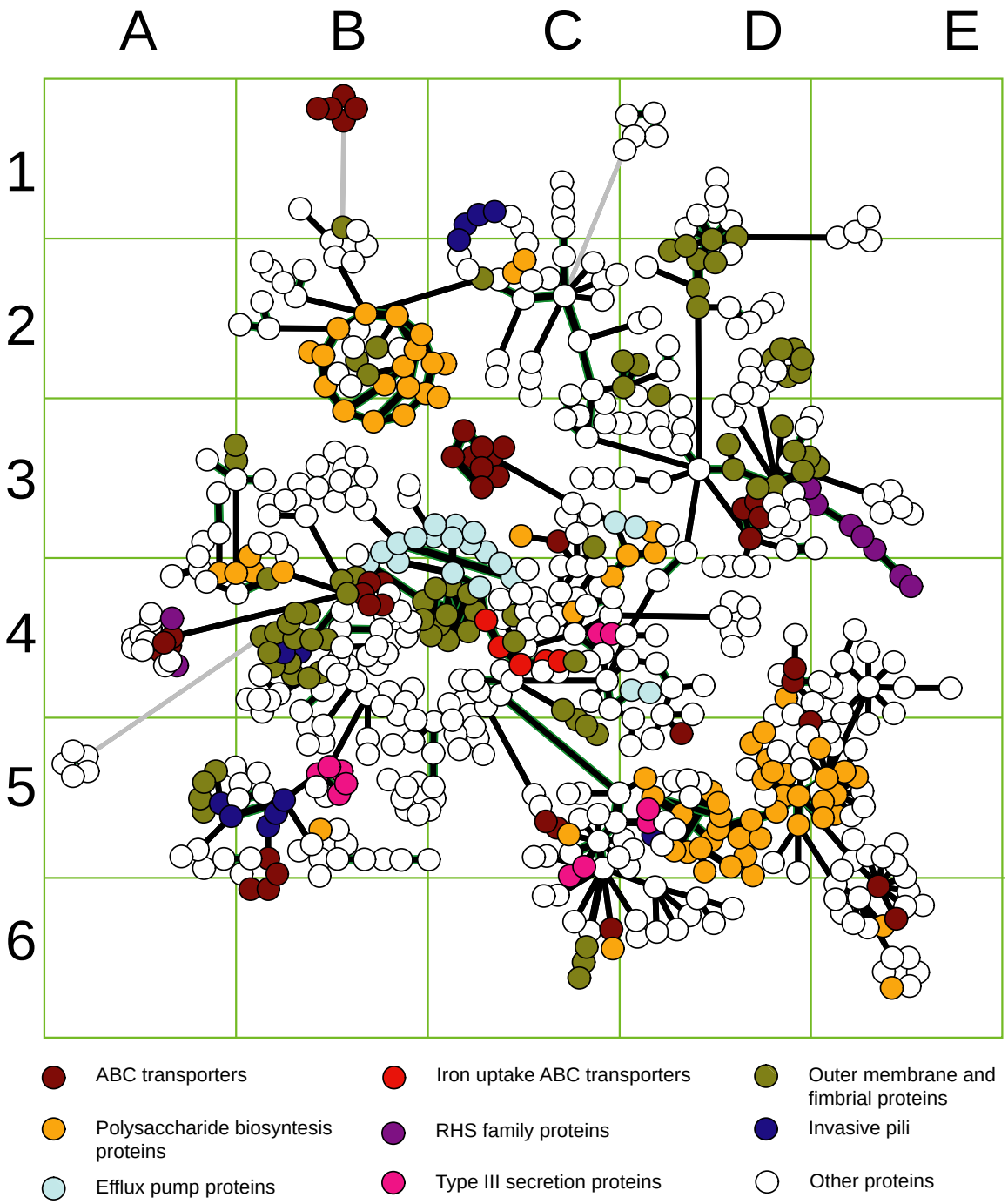


Figure 4.5: An illustration of functional groups of genes distributed among the GIs of enterobacteria. The illustration of distribution patterns of genes of different functional groups of the GIs which were identified in enterobacteria. Layout of GIs is the same as in Figures 4.2 and 4.3.

Many GIs represent a combination of modules which provide host organisms with specific functionality. The functional relatedness between GIs was statistically proven by calculating SIG values, as previously described (Lima-Mendez *et al.*, 2008). The analysis of functional modules of genes has been proposed recently as a tool for reconstructing the evolution and phylogenetic links between prophages. OU pattern comparison (Figure 4.5) demonstrated that the GIs of the same functional module may re-appear many times independently, probably at

each oscillation of a new vector. The evolution of phages should be separated from that of functional gene modules even if they are 'cargo elements' of the same phages.

4.5 Genomic islands of the pathogenicity plasmids of the EAHEC strains

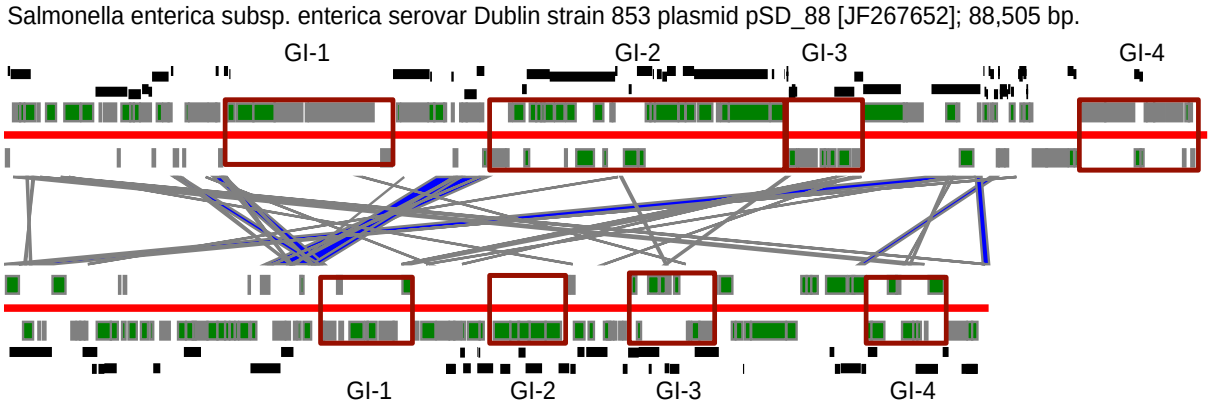


Figure 4.6: A bl2seq representation of comparative analysis conducted between TY-2482 contigs and pathogenicity plasmids pSD_88 and 55989p. Protein coding genes are shown by green bars and hypothetical genes by gray bars. The red horizontal lines separate plasmid genes by their direction of transcription. BLAST hits between two plasmids are depicted by blue connecting stripes. Predicted GIs in plasmid sequences are framed and named respectively. Corresponding positions of the contigs of TY-2482 which were mapped against plasmid sequences by BLASTn are depicted by black bars.

The genome of enterohemorrhagic *E. coli* 55989 strain comprises a large plasmid 55989p of 74,482 bp which most likely contributes towards its virulence (Brzuszkiewicz *et al.*, 2011). Many contigs of isolate TY-2482 share sequence similarity with both this plasmid and that of the multidrug resistance plasmid pSD_88 of *S. enterica* ssp. *enterica* Dublin (Figure 4.6). The shared similarities suggest that TY-2482 most likely possesses a hybrid pathogenicity plasmid established from both 55989p and pSD_88. These two plasmids are thought to have ascended from a common ancestor as they both possess the same fragments of housekeeping genes (Figure 4.5). Both plasmids are made up by combinations of at least four horizontally transferred GIs whose OU patterns differ from each other and their host plasmid's core sequences. The only similarity found in plasmids 55989p and pSD_88 was between their core sequences. The contigs of TY-2482 partially overlap with both plasmids' core sequences and GIs: 2, 3 and 4 of pSD_88 and GIs: 2, 3 and 4 of 55989p. The transfer of mobile elements by these plasmids is suspected to have been promoted by the presence of more than 20 transposases and IS-elements in their sequences.

Table 4.2. BLAST results of GIs of other bacterial genomes which showed a significant sequence similarity to GI-2 of *S. enterica ssp. enetrica* Dublin plasmid pSD_88.

Genomic island	Annotation*	Score	S%†	D%‡
Escherichia coli TY-2482 [contigs 00033_1, 00499_1 and 00546_1]	mer, βl, tet	9501	76	32
Salmonella enterica ssp. *enterica* Typhi CT18 plasmid, NC_003384:1 [157158..177822]	mer, βl, tet	4301	86	42
Comamonas testosteroni CNB-2, NC_013446:1 [2861000..2881851]		3811	80	36
Nitrosomonas europaea ATCC 19718, NC_004757 [905417..925311]	mer, tet	3313	39	36
Acinetobacter baumannii ACICU, NC_010611 [243356..264198]		2961	57	61
Acidovorax ebreus TPSY, NC_011992 [2277714..2295906]	mer, tet	2584	84	23
Shewanella frigidimarina NCIMB 400, NC_008345 [4109193..4146599]	mer	2469	17	53
Nitrosomonas eutropha C91, NC_008344:2 [1089057..1108455]	mer	2311	27	32
Salmonella enterica subsp. *enterica* Typhi CT18 plasmid, NC_003384:2 [107500..125897]	mer	2250	83	39
Delftia acidovorans SPH-1, NC_010002 [2932123..2951187]		2106	82	30
*Marinobacter aquaeole*i VT8, NC_008740 [1414926..1435099]		2063	53	44
Shigella dysenteriae Sd197, NC_007606 [3752925..3771840]		1804	41	33
Cupriavidus metallidurans CH34, NC_007973 [3238831..3258471]	tet	1790	83	27
Comamonas testosteroni CNB-2, NC_013446:2 [2776491..2801174]		1658	70	37
Gamma proteobacterium HdN1, NC_014366 [586254..605651]		1538	82	29
Corynebacterium urealyticum DSM 7109,		1485	79	29

NC_010545 [1781393..1801974]				
Salmonella enterica ssp. *enterica* Heidelberg SL476, NC_011083 [1525960..1544455]	βl	1383	48	37
Alicycliphilus denitrificans K601, NC_015422 [3514585..3535258]	mer	1306	82	22
Nitrosomonas eutropha C91, NC_008344:1 [230541..250543]	mer	1294	23	34

Remarks:
* mer – mercury resistance operon; βl – beta-lactamase; tet – tetracycline efflux protein.
† S% – compositional similarity of GIs to OU pattern of GI-2 of *S. enterica* Dublin plasmid pSD_88.
‡D% – distance between GIs' OU patterns and OU patterns of the host chromosomes.

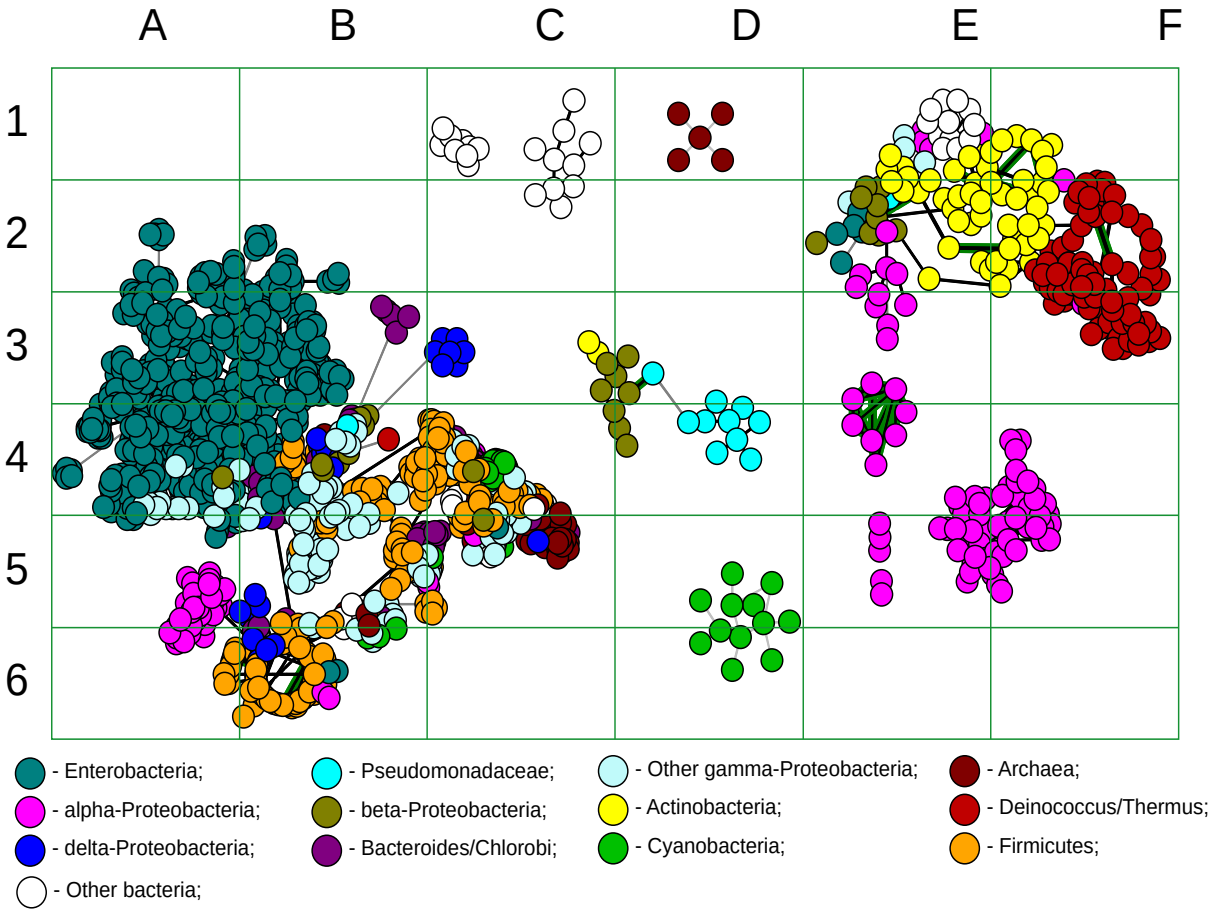


Figure 4.7: A graphical representation of similar GIs in different bacterial groups. An overview of clusters of GIs with shared compositional similarities identified from different bacterial classes.

The virulence of strain TY-2482 is very likely facilitated by the presence of GI-2 of pSD_88 which comprises genes encoding drug resistance proteins (broad range beta-lactamase and

tetracycline efflux protein TetA) and mercury detoxification proteins. In contrast to the other GIs of plasmids pSD_88 and 55989p, GI-2 of pSD_88 shows no compositional similarity to GIs of enterobacteria as depicted in Figure 4.2. Figure 4.7 is an overview of clusters of GIs from different bacterial classes; the major cluster of enterobacterial GIs shown in Figure 4.2 is placed in cells A2-B4. Newly found drug-resistance GIs fell into another quite distant cluster shown in cells E1-F3. The central and oldest elements of this cluster are mycobacterial GIs which spread to *Deinococcus* / *Thermus* and *Acidobacteria*. The latter homologous GIs were found to be predominant in beta-Proteobacteria and these share greatest compositional and gene content similarity with plasmid associated GIs of *Salmonella* and *Escherichia* (cell E2 in Figure 4.6). Many of these GIs contain a mercury resistance operon which may have been inherited from a mycobacterial plasmid isolated from *Mycobacterium abscessus* (Ripoll *et al.*, 2009).

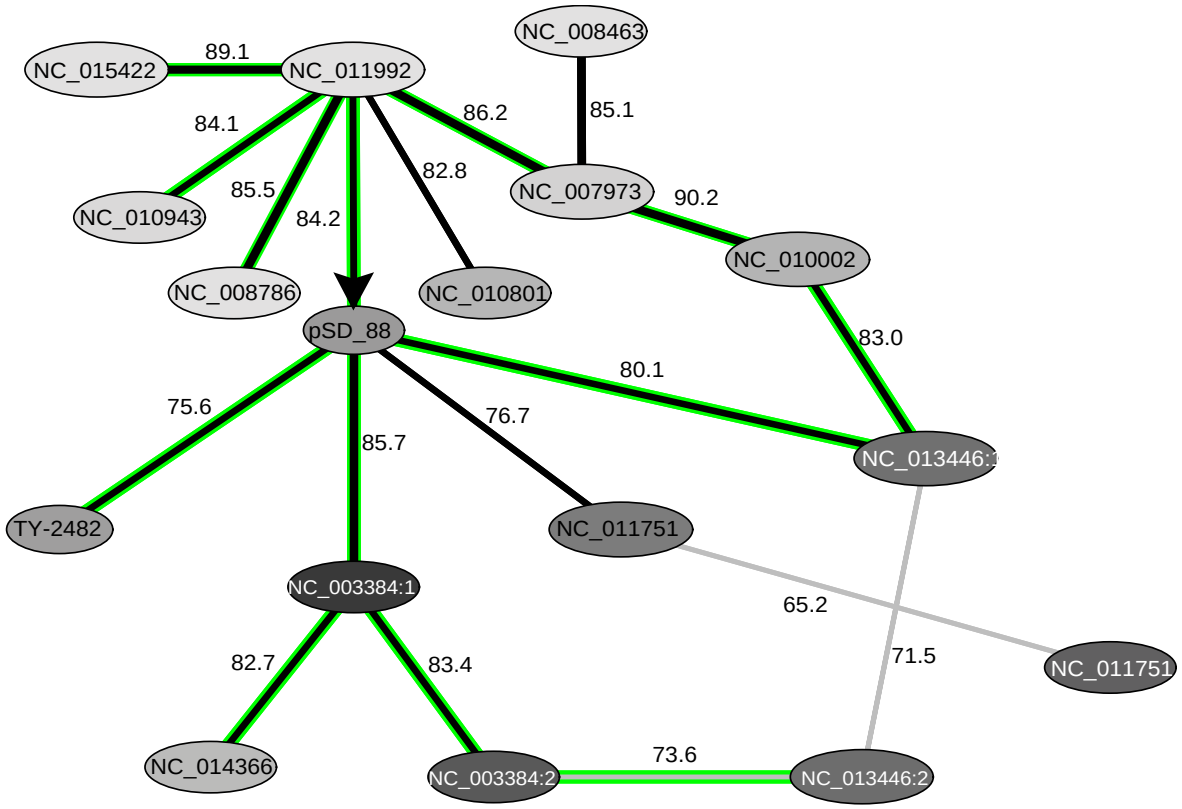


Figure 4.8: Stratigraphic analysis and OU pattern similarity relations between GIs showing similarity to the mercury resistance GI of plasmid pSD_88. Each node corresponds to one GI named as in Table 4.1. Two nodes of GIs that share 75% or higher OU pattern similarity are linked by black edges. Gray edges represent similarity below 75%. Links between GIs which were determined by BLASTn to share sequence similarity are highlighted in green. Values for OU pattern similarity shared between nodes are shown above the edges. Gradient colours depict divergences of GIs patterns from the complete genome patterns of their host organisms. The older inserts are depicted by lighter colours, as their OU patterns are much closer to their hosts. Putative direction of transfer of genetic material from NC_011992 to pSD_88 is depicted by an arrow (see also Figure 4.9).

GI-2 of pSD_88 was shown to share sequence similarity with DNA fragments of organisms of different classes after a nucleotide BLAST search was performed against a database of predicted GIs (Table 4.2). GIs of enterobacteria which were identified by BLAST to be similar to GI-2 of *S. enterica* ssp. *enterica* Dublin plasmid pSD_88 were found to be in possession of large mercury resistance operons accompanied by genes which encode broad range beta-lactamase and tetracycline efflux proteins (Fluit, 2005). A nucleotide BLAST search also retrieved similar heavy metal resistance operons in GIs of several γ- and β-Proteobacteria. OU pattern comparison revealed that many of these do not share composition similarities and are probably of different origins. GIs of γ-Proteobacteria are recent acquisitions as their compositions were revealed to be remarkably distant from those of their hosts. Homologous GIs in β-Proteobacteria are older inserts. Possible donor-recipient interactions between these GIs and their hosts are visualized in Figure 4.8.

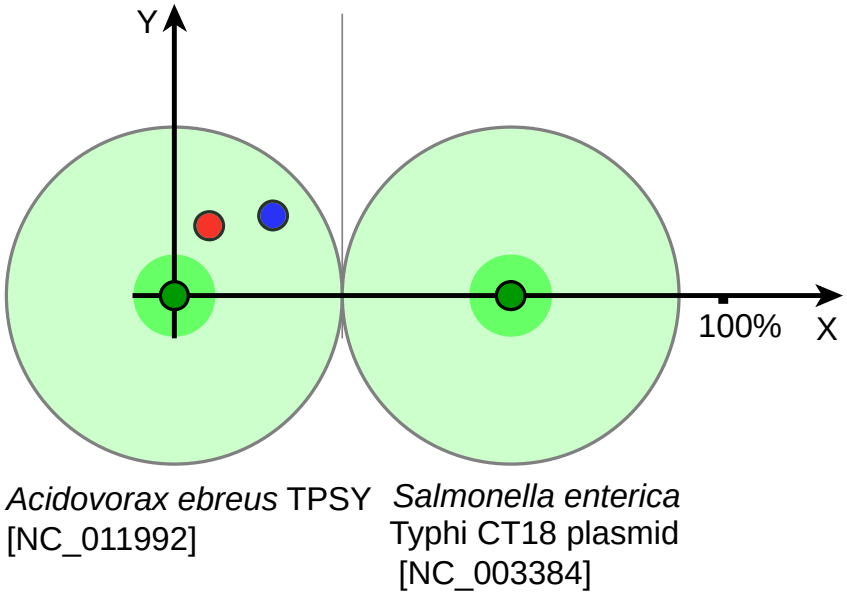


Figure 4.9: Donor-recipient relationship determined between GIs from *A. ebreus TPSY* and *S. enterica Typhi CT18*. Meanings of the graph elements were explained in Figure 4.4.

The stratigraphic analysis shown in Figure 4.8 illustrates the acquisition time frames of these heavy metal resistance genes by enterobacteria. According to the analysis, enterobacteria have likely acquired these genes from β-Proteobacterium, *Acidovorax ebreus* (Figure 4.9). *A. ebreus* TPSY and many other β-Proteobacterium that possess mercury resistance operons inhabit environmental areas which are contaminated with heavy metals (Rosewarne *et al.*,

2010). Pathogenic bacteria utilize mercury resistance genes for resistance against drugs and disinfectants (Levings *et al.*, 2007; Durante-Mangoni & Zarrilli, 2011). A large part of plasmid pSD_88 with neighbouring beta-lactamase encoding genes was also identified in the genome of uncultured γ-Proteobacterium HdN1 [NC_014366], demonstrating how easily this GI can be integrated into a chromosome.

4.6 Discussion

Horizontal gene exchange plays an important role in the dispersion of virulence factors and emergence of new pathogens. Genes encoding toxins and drug resistance proteins are often located in plasmids and highly variable genomic fragments which are presumably of lateral origin (Hacker & Carniel, 2001; Alonso *et al.*, 2001; Oelschlaeger *et al.*, 2002; Canchaya *et al.*, 2003; Turner *et al.*, 2003; Yoon *et al.*, 2005; Fluit, 2005; Pallen & Wren, 2007; Becq *et al.*, 2007; van Passel *et al.*, 2008). HGT is sometimes considered to be a random process of casual gene exchange events which may happen at any time between any organisms. The composition-based approaches introduced in this work have uncovered two important features of HGT: (i) donor recipient relations of GIs and their hosts; (ii) and the relative acquisition periods of GIs. The stratigraphic analysis conducted in the study has indicated that enterobacterial GIs are results of different acquisition time periods. Many of the GIs which they possess were determined to be ancient inserts, see Figure 4.3.

Prophages have evidently been illustrated to be substantially accountable for exchanges of virulence determinants in enterobacteria (Canchaya *et al.*, 2003). Pallen and Wren (Pallen & Wren, 2007) have considered enterobacterial PAIs to be a separate class of mobile elements as they are clearly distinguishable from prophages and plasmid inserts. However, their statement does not make it clear whether PAIs transfer between genomes independently on their own or whether they rely on plasmids and phages as vectors. In bacterial genomes prophages adjoin a variety of mobile genetic inserts in similar recombination hotspots (see Figure 4.1). It is very difficult to separate two adjacent GIs even by manual inspection of the genetic content. Reports also mention significant technical difficulties in distinguishing between prophages and other mobile elements due to the lack of reliable genetic markers (Herzer *et al.*, 1990; Clermont *et al.*, 2000; Gordon *et al.*, 2008). Many GIs are in fact mosaic chimeras of several independently evolved mobile genetic elements which are either composed of a single replicon as reported in lambdoid phages possessing “cargo elements” of foreign genes (Pallen & Wren, 2007); or if they simply co-occur in the same host

chromosome. Chimerism may be a reason why 50% of predicted GIs formed singletons when clustered by compositional similarity. In this work the focus was on the analysis of GIs with shared compositional similarity and not to distinguish between different types of mobile genetic elements.

Phylogenetic relationships between *E. coli* and *Shigella* species have always been a challenge to establish. A complete genome comparison study conducted by Touchon *et al.*, (Touchon *et al.*, 2009) confirmed the existence of four well-separated phylogenetic groups of *E. coli* referred to as: A, B1, B2, E and D (paraphyletic). These have previously been defined based on the comparison of restriction digest patterns and various DNA markers (Herzer *et al.*, 1990; Clermont *et al.*, 2000; Gordon *et al.*, 2008). The authors also illustrated that groups: A, B1 and E of *Shigella* and *E. coli* which occupy the same cluster (Figure 4.2) share the most recent common ancestor in contrast to *E. coli* of groups B2 and D. Our analysis revealed that OU compositions of GIs of *Shigella* and *E. coli* of groups A, E and B2 are much more variable as compared to those in the *E. coli* strains of group B1which are the causative agents of the recent enterohemorrhagic outbreaks. Strains 55989, IAI1 and the new isolate TY-2482 appear to be genetically 'primitive' organisms, as they all are in possession of the oldest GIs which probably have been inherited from an ancestor common for both *Escherichia* and *Shigella*. Centric phylogenetic positions of O104:H4 strains in relation to other groups of *E. coli* were also demonstrated by other authors in a minimum-spanning tree on the basis of allelic profiles of their core genes (Mellmann *et al.*, 2011). General assumption is that enterohemorrhagic strains are natural inhabitants of cattle intestines and may potentially cause infections to humans upon exposure (Ferens & Hovde, 2011; Mainil & Daube, 2005; Sasaki *et al.*, 2011). Inhabitation of cattle intestines does not explain how genomes of these microorganisms were confined from the mainstreams of horizontal gene exchange. Neither does it explain why the widespread bovine *E. coli* do not cause outbreaks among humans on a daily basis upon contact. Even if the outbreak strains and O104:H4 cattle *E. coli* are closely related, their relation is however not straightforward. A plausible explanation for the outbreak may be that these strains were confined in areas with very limited or difficult access to mobile genetic elements until recently. The *E. coli* strains isolated from faeces of healthy humans in industrialized countries were distinctive from those isolated in remote tropical islands (Escobar-Páramo *et al.*, 2004). Genetic naivety of these organisms makes them more vulnerable to the effects of newly acquired virulence factors. In strains 55989 and TY-2482 these factors are associated predominantly with plasmids, except for one chromosomal insert

containing the genes for secreted autotranspoter toxin and probably Shiga toxin subunits. Contig 00415_1 of *E. coli* TY-2482 is 98% similar to the fragment of *stx2* Shiga toxin gene of *E. coli* O157:H7 Thai-12. It correlates with the reported synthesis of the Shiga toxin by the outbreak strains (Mellmann *et al.*, 2011). Only the short and probably non-functional 564 bp fragment of the 4,426 bp long sequence of stx2 AB was found to be present in strain 55989. This DNA fragment is located 10 kbp downstream of GI #8 (Figure 4.1). Contig 00415_1 is not associated with either of the predicted GIs of TY-2482. However, it has to be noted that the prediction of GIs from randomly concatenated contigs cannot ensure the detection of all mobile elements.

There were several sequential fluxes of global gene exchange as depicted in Figures 4.2 and 4.3 in clusters A3-B3, B4, B2, D3 and C5-D5. These events affected genomes of *Shigella* and *E. coli* of groups A and B2, but somehow did not get to *E. coli* of B1. The former organisms are thought to have utilized their acquired genes well for pathogenicity and/or commensalism; and probably gained immunity against further mobilome insertions. The GIs of these organisms, be it pathogens or commensals, constitute same clusters which they also share with GIs of *Salmonella*. However it does not necessarily mean that these organisms have arbitrarily exchanged genes with one another. Alternative scenarios may be that these organisms were all infected from the same source, or that *Escherichia* acted as a donor of GIs to *Salmonella*, or vice versa. The OU patterns of these organisms are highly similar to each other to an extent that makes it almost impossible to give preference to either scenario. All the GIs of *E. coli*, *Shigella* and *Salmonella* share similarity in composition with the genome of *Shewanella*. Homology between GIs of these organisms was confirmed by nucleotide BLAST. A sequence similarity search conducted by BLAST revealed a homologous cluster of polysaccharide biosynthesis genes shared by GIs of *Salmonella enterica* ATCC 9150 (NC_006511:5) and *Shewanella denitrificans* OS217 (NC_007954:14). These islands were further analysed for similarity in OU composition and *Salmonella enterica* ATCC 9150 was revealed to have acquired its GI from the *Shewanella* lineage (Figure 4.4A). Common GIs of *E. coli* and *Salmonella* shown in Figure 4.5: cells D5-E6 including LEE and SPI-10 PAIs are recent acquisitions and share a high degree of similarity in composition and sequence with environmental *Shewanella*. However, the virulence genes harboured by these PAIs are not of *Shewanella* origin but seem to be virulence determinants which are only associated with enterobacteria. A global search for shared compositional similarity through the database of putative GIs illustrated that the root of transmission for the entire set of GIs most likely is

Vibrio, particularly *V. fischeri* plasmids (data not shown). GIs harboured by these organisms proved to be the oldest inserts, which thus supports the fact that most mobilome originate from marine γ-Proteobacteria. *Shewanella* may either be the best transmitter of these mobilome to other organisms, or simply the closest recipient. Further studies have to be conducted in order to elucidate the intrinsic donor-recipient relations between these micro-organisms.

The role played by environmental micro-flora towards the development of pathogenic microorganisms should be re-considered. Our current understanding of this process is limited to the fact that pathogenic or commensal bacteria may easily acquire new virulence factors from the environment. It is therefore generally accepted that antibiotic resistance genes which are widely spread in clinical isolates result from native genes of environmental bacteria. Particularly from bacteria which encode factors that aid to withstand highly toxic and polluting reagents (Alonso *et al.*, 2001; Riesenfeld *et al.*, 2004). This study indicated that environmental micro-organisms may also provide pathogens with newly generated vectors (plasmids, phages and integrons) which accelerate genomic rearrangements and the exchange of virulence factors. The effectiveness of a vector decreases in time as bacteria acquire immunity towards it. The mechanisms which bacteria acquire this immunity are through the development of a sophisticated anti-phage CRISPR system or through the presence of remnants of non-functional vectors which block specific insertion sites in chromosomes (Kiewitz *et al.*, 2000; Touchon *et al.*, 2011). Although a CRISPR system provides immunity against vectors, newly generated environmental vectors are able to evade the system. Such newly generated vectors accelerate HGT and contribute towards disease outbreaks that occur periodically. The spread of vectors is illustrated in Figure 4.3 by gradient colour changes depicted in clusters of enterobacterial GIs which vary from dark gray in cells D5-E6 and A4 to light gray in C4-D4 and intermediate shades in D3, B5 and B2. Old vectors continue their spread towards the new genetically naïve strains of microorganisms. The latter is most probably an explanation of what happened with the *E. coli* strains of group D (Figure 4.2). Strains IAI39 and UMNO26 mainly compose of ancient GI inserts and few of their newly acquired GIs share similarity with various clusters of enterobacterial GIs. Horizontal gene exchange has driven these two phylogenetically unrelated strains (Touchon *et al.*, 2009) to share a similar phenotype and genetic markers, as a result these joined into group D. It is quite possible that even the most primitive strains 55989 and TY-2482 of group B1 will

undergo the same evolutionary process.

In this study it was illustrated that the exchange of genetic materials between bacteria is not a random event but a strictly directed process with chains of donor and recipient organisms. Of all the GIs analysed in our study there was no single case of transmission which appeared to have taken place from enterobacteria to *Shewanella*. However there were several cases where enterobacteria showed to have transmitted GIs enriched with virulence determinants to various other organisms. *Sodalis glossinidius*, a maternally transmitted intracellular endosymbiont of the tsetse flies (Dale *et al.*, 2001) serves as one of the organisms possessing GIs which have been acquired from *Salmonella*. It was also found to comprise a type III secretion system identical to that of *S. enterica* SPI-1 PAI (Dale *et al.*, 2001). A total of 30 GIs were identified in *S. glossinidius morsitans* [NC_007712] and it is rather unusual for an intracellular symbiont to possess such a number of elements. Twelve of the 30 GIs share similarity in composition and sequence with each other and the GIs of *Salmonella*. Figure 4.4B illustrates that some of the (cell B5 in Figure 4.3) GIs of *Sodalis* are of *Salmonella* origin. These GIs are not duplicates of the same inserts but form a clear time series of independent acquisitions. The latter is another example of genetic vector oscillation and also implies that this intracellular organism regularly interacts with environmental micro-flora. Acquisition of pathogenicity determinants by this organism may have influenced the spread of virulence factors and invasion of new host cells and tissues. Probably the *E. coli* strains of group B2 evolved following the same scheme upon the acquisition of virulence determinants which allowed them to access extra-intestinal tissues to cause systemic infections in humans and animals (Oelschlaeger *et al.*, 2002). *E. coli* strains are potential donors of PAIs to other environmental microorganisms. Few of their related inserts were found to be present in *Erwinia*, a plant pathogen which causes septicaemia and nosocomial infections in immunologically compromised humans (Marklein *et al.*, 1981; Weiner & Werthamer, 1973). Additional studies have to be conducted to outline bacterial species which potentially have inherited pathogenic determinants from enterobacteria.

E. coli TY-2482, one of the causative agents of a deadly 2011 outbreak in Europe, possesses a plasmid-associated PAI. The PAI consists of broad range beta-lactamase; tetracycline efflux pump and a cluster of mercury resistance genes. Ontologically these factors are not related to either of the prevalent GIs of enterobacteria shown in Figure 4.2. However, this PAI is not completely novel, as there are a number of papers reporting similar inserts in rapidly evolving

plasmids commonly found in the drug resistant strains of *Salmonella* (Ripoll *et al.*, 2009; Levings *et al.*, 2007; Durante-Mangoni & Zarrilli, 2011). OU pattern analysis indicated that there are at least two GIs in TY-2482 sharing the same source. These have also been reported to be present in *E. coli* UMNO26 with the exception of antibiotic resistance genes which are possessed by TY-2482 (Figure 4.7). In contrast to the gradual oscillation of genomic inserts in genomes of enterobacteria observed in Figure 4.3, new GIs seem to be rapidly acquired from distant organisms (see Fig. 4.7, 4.8 and 4.9). The evolution and acquisition of these GIs is in consistence with a drastic rise in mercury resistant bacteria of the coastal environment in India (Ramaiah & De, 2003). It has been reported that in a period of 5 years (between 1997 and 2003) the number of microorganisms isolated from sea water which tolerated 10 ppm of $HgCl_2$ increased from none to 75 - 95% of CFU in the polluted zones of Mumbai (Ramaiah & De, 2003). It was concluded that the sharp rise in mercury tolerance could be linked to general ocean pollution by human industrial activity. It however is plausible that old dormant vectors which have been confined in isolated eco-niches of naturally polluted areas for long periods managed to break through their borders to finally exchange genes to tolerate chemical pollutions with enterobacteria. These are the genes that may have possibly driven bacteria to also tolerate a wide range of antibiotics. Vectors which carry such genes have established a new channel of horizontal gene exchange with all human commensals that are naïve and highly vulnerable.

4.7 Conclusion

This work highlights the importance of composition-based approaches in terms of studying distribution patterns of GIs among bacteria. The latter provides a possibility of creating a network of similar GIs which are harboured by different classes of bacteria. The network approach has showed to be practical for determining donor-recipient interactions between associated GIs and their host microorganisms. Stratigraphic analysis of the GIs with compositional associations allows for the determination of their distribution patterns and acquisition periods. The compositional and stratigraphical analysis of the data in Table 4.2 showed that the heavy metal resistance genetic cassettes which are widely distributed among environmental bacteria have been acquired recently by many new micro-organisms, probably resulting from general ocean pollution. The increased pollution has triggered several parallel currents of HGT. For instance, *A. baumannii* ACICU [NC_010611] has acquired its GI recently from an unknown α-Proteobacteria. The GI in *N. europaea* C91 [NC_008344:1] shows a strong compositional similarity to Firmicutes GIs harboured by *Clostridium*

thermocellum, *Desulfitobacterium hafniense* and *Carboxydothermus hydrogenoformans*; however, neither of them was found to be a donor of these GIs. The GIs in *N. europaea* C91 [NC_008344:2] and *N. europaea* ATCC 19718 [NC_004757], as well as the recent inserts of metal resistance gene cassettes in *S. frigimarina* NCIMB400 [NC_008345] showed no significant compositional similarity to each other, nor to the available sequenced bacterial genomes. Each channel of HGT appears to have its own range of recipient species, which therefore may result in different consequences in terms of evolvement of new human pathogens. The effect may be even more profound if several independent fluxes of gene exchange vectors interfere in the same organism. Our further work will aim at summarizing the ontological data for newly identified GIs and highlighting the most active mainstreams of gene exchange in a scale of the whole bacterial world.

Chapter 5

5 Intraclonal genome diversity of *Pseudomonas aeruginosa* clones CHA and TB

Oliver K. I. Bezuidt,[1,2] Jens Klockgether,[1*] Sylvie Elsen, [3,4,5,6] Ina Attree,[3,4,5,6] Colin F. Davenport,[1] Burkhard Tümmler[1,7]

[1]Klinische Forschergruppe, Klinik für Pädiatrische Pneumologie, Allergologie und Neonatologie, Medizinische Hochschule Hannover, D-30625 Hannover, Germany

[2]Bioinformatics and Computational Biology Unit; Department of Biochemistry; University of Pretoria; Pretoria, South Africa

[3]INSERM, UMR-S 1036, Biology of Cancer and Infection, Grenoble, France

[4]CNRS, ERL 5261, Bacterial Pathogenesis and Cellular Responses, Grenoble, France

[5]UJF-Grenoble 1, F-38041 Grenoble, France

[6]CEA, DSV/iRTSV, F-38054 Grenoble, France

[7]Biomedical Research in Endstage and Obstructive Lung Disease Hannover (BREATH), Member of the German Center for Lung Research

Adapted from BMC Genomics (2013), 14:416 doi:10.1186/1471-2164-14-416

5.1 Abstract

5.1.1 Background

Adaptation of *Pseudomonas aeruginosa* to different living conditions is accompanied by microevolution resulting in genomic diversity between strains of the same clonal lineage. In order to detect the impact of colonized habitats on *P. aeruginosa* microevolution we determined the genomic diversity between the highly virulent cystic fibrosis (CF) isolate CHA and two temporally and geographically unrelated clonal variants. The outcome was compared with the intraclonal genome diversity between three more closely related isolates of another clonal complex.

5.1.2 Results

The three clone CHA isolates differed in their core genome in several dozen strain specific nucleotide exchanges and small deletions from each other. Loss of function mutations and

non-conservative amino acid replacements affected several habitat- and lifestyle-associated traits, for example, the key regulator GacS of the switch between acute and chronic disease phenotypes was disrupted in strain CHA. Intraclonal genome diversity manifested in an individual composition of the respective accessory genome whereby the highest number of accessory DNA elements was observed for isolate PT22 from a polluted aquatic habitat. Little intraclonal diversity was observed between three spatiotemporally related outbreak isolates of clone TB. Although phenotypically different, only a few individual SNPs and deletions were detected in the clone TB isolates. Their accessory genome mainly differed in prophage-like DNA elements taken up by one of the strains.

5.1.3 Conclusions

The higher geographical and temporal distance of the clone CHA isolates was associated with an increased intraclonal genome diversity compared to the more closely related clone TB isolates derived from a common source demonstrating the impact of habitat adaptation on the microevolution of *P. aeruginosa*. However, even short-term habitat differentiation can cause major phenotypic diversification driven by single genomic variation events and uptake of phage DNA.

5.1.4 Key words

Pseudomonas aeruginosa, microevolution, habitat adaptation, genome diversity

5.2 Background

Pseudomonas aeruginosa is a metabolically versatile gamma-proteobacterium that preferentially thrives in aquatic habitats and the rhizosphere (Selezska *et al.*, 2012). This opportunistic pathogen is the most dominant bacterium causing chronic airway infections in cystic fibrosis (CF) (George *et al.*, 2009) and has become one of the most important causative agents of nosocomial infections, particularly in intensive care units (de Bentzmann & Plésiat, 2011).

The 5.2 – 7 Mbp *P. aeruginosa* genome is a mosaic of a conserved core and variable regions of genome plasticity (RGPs) (Mathee *et al.*, 2008). The core genome is characterized by a conserved synteny of genes (Tümmler, 2006). Clonal complexes differ from each other in clone-typical segments of core and accessory genome (Wiehlmann *et al.*, 2007) and a nucleotide divergence in the core genome of 0.5 – 0.7 % (Spencer *et al.*, 2003).

Intraclonal whole-genome variation in *P. aeruginosa* has mainly been studied in isolates from CF lungs that had been collected from the same patient longitudinally or at one time point (Kresse *et al.*, 2003; Smith *et al.*, 2006; Cramer *et al.*, 2011; Yang *et al.*, 2011; Chung *et al.*, 2012). The paired isolates from one patient typically differed due to a few dozens of single nucleotide substitutions (SNPs) and small insertions/deletions (indels) in the core genome, a few RGPs in the accessory genome and occasionally one large deletion or inversion. Close to 1,000 de novo SNPs and indels, however, were gained in hypermutable strains defective in DNA repair (Cramer *et al.*, 2011; Chung *et al.*, 2012).

Whereas genome microevolution of *P. aeruginosa* in the atypical habitat of the CF lungs has been investigated for several clones, only a single clone has so far been assessed in its genome diversity between strains of unrelated habitat and geographic origin (Klockgether *et al.*, 2011). The two examined clone PA14 strains from California and Germany were found to be of the same genome size and differ from each other in 0.0035 % of their nucleotide sequence. Since these data alone do not allow any general conclusions, we wanted to explore the impact of habitat, history and geographic origin on intraclonal genome diversity of *P. aeruginosa* in more depth. For that purpose two complementary scenarios of habitat differentiation were chosen. The three selected clone CHA strains were isolated from freshwater and CF patients at geographically distant sites within a 15-year period and represent the distant clone strain set. Conversely, the three selected clone TB strains were isolated during a local outbreak and represent the closely related clone strain set. Clones CHA and TB were chosen because we wanted to include the highly pathogenic strains CHA (Toussaint *et al.*, 1993) and TBCF10839 (Tümmler *et al.*, 1991) in the comparative genome analysis. CHA and TBCF10839 are the only known *P. aeruginosa* strains which can escape killing by leucocytes. TBCF10839 can persist and grow in leucocytes (Klockgether *et al.*, 2013), whereas CHA kills leucocytes by type III secretion-dependent oncosis (Dacheux *et al.*, 1999; Dacheux *et al.*, 2000; Dacheux *et al.*, 2001). Genome sequencing was expected to provide an explanation why CHA and TBCF10839, but not the other two clone CHA and two TB strains could undermine the major antipseudomonal defence mechanism in humans.

Genome sequencing revealed higher nucleotide divergence and a more variable composition of the accessory genome amongst the less closely related clone CHA strains than amongst the more highly related clone TB strains. Strain-specific SNPs were preferentially detected in

habitat-associated fitness loci. Conservation of small non-coding RNA loci followed clone-specific patterns with about 7 % (clone TB) or 11 % (clone CHA) not conserved. Clone-specific traits were also found for the accessory genomes of the analysed strains, but especially for clone CHA strains which were equipped with several strain-specific DNA elements, the majority of which appeared to be of phage origin. Phage-like DNA also differentiated the accessory genome of the clone TB wound isolate TB63741 from its relatives of CF-origin, indicating that uptake and integration of phage elements is a major driving force of intraclonal diversification of *P. aeruginosa* during adaptation to different habitats.

5.3 Methods

5.3.1 Bacterial strains

P. aeruginosa strains 491, TBCF10839, TBCF121838 and TB63741 were isolated from patients seen at the Medizinische Hochschule Hannover. Strain PT22 was retrieved from the river Ruhr close to Mülheim. Strain CHA was isolated from a patient seen at the CF clinic in Grenoble. First subcultures were maintained in LB supplemented with 15% (w/v) glycerol at -80°C until use.

5.3.2 Strain genotyping

P. aeruginosa strains were genotyped by a custom-made microarray following the protocol published previously (Wiehlmann *et al.*, 2007).

5.3.3 DNA preparation

P. aeruginosa genomic DNA was prepared from cells grown in LB medium following a protocol optimized for Gram-negative bacteria (Ausubel *et al.*, 1998).

5.3.4 Illumina genome analyser sequencing

After preparing genomic DNA libraries according to the manufacturer's instructions, sequencing-by-synthesis was performed at GATC-Biotech (Constance, Germany) for each library with an Illumina Genome Analyser II generating 36 bp sequence reads. Illumina Genome Analyser Pipeline Version 0.2 software was applied to qualify reads passing default signal quality filters. Obviously incorrect reads with homooligomers > 13 bases in length (not present in the *P. aeruginosa* genome) or an ‘N’-base call in at least three positions were excluded from the analysis (Cramer *et al.*, 2011). All sequence data from this study have been

submitted to the Sequence Read Archive (SRA) of the EBI (strain TB63741: study accession no. ERP001300; clone CHA strains CHA, PT22 and 491: study accession no. ERP001750).

5.3.5 Sequence and read alignment

36 bp reads data of the strains were individually mapped to the PAO1 reference genome (NC_002516.2) using the accurate alignment software Novoalign V2.07.00 (Novocraft Technologies, 2010). The command: novoalign –d Indexed_reference_genome –f Reads.fastq –o SAM > out.sam, was used during the mapping to create "sam" formatted alignment files. Two pools of data consisting of the PAO1 mapped and unmapped reads were then extracted directly from the three alignment files using a custom script. Unmapped reads representing non-PAO1 DNA and the mapped reads representing the PAO1 DNA were assigned to not-in-reference and in-reference read pools, respectively.

5.3.6 Sequence variation sites analysis

Clone CHA strains with genomic positions indicating single nucleotide variants relative to the PAO1 reference were extracted from the novoalign "in-reference" alignment files using SAMtools (Li *et al.*, 2009). The variant call format (vcf) output files generated by SAMtools were further filtered for low quality variants. Variants with minimum coverage of six reads with minimum base calling quality (Q) of 30 at the respective position, a minimum SNP-call quality (QUAL) of 160 (QUAL = -10 $\log_{10}$ (probability of wrong call) (Danecek *et al.*, 2011)) and with more than 67% of all quality reads calling the SNP were retained. These variants were then compared against each other to identify sets of strain specific SNPs through the use of an in-house SNP filter pipeline.

The SAMtools derived sequence variants output files were further searched for predictions of small indels. The top candidates (QUAL ≥ 160) were verified by manual inspection of the alignment. Predicted indels were removed that did not pass the following criteria: minimum coverage of more than five high quality reads (Q ≥30 at the candidate position) and more than 95% of reads flag the indel. Predicted indels and SNPs were subsequently annotated using SNPeff version 1.9.5 (Cingolani *et al.*, 2012) to identify their effect on coding DNA sequences.

5.3.7 De novo assembly

The not-in-reference pools of sequence reads characterized as Clone CHA accessory genome were assembled to larger contigs with the *de novo* assembler Velvet version 1.0.12 (Zerbino

& Birney, 2008). Commands used during the assembly process are as follows: velveth 63741_cov5_23 23 63741_reads.fas; velvetg 63741_cov5_23 -cov_cutoff 5.0 -max_coverage 300. The assembler parameters were set for a minimum read coverage of 5 and *k-mer* size of 23 to construct reliable contigs. These criteria were set for the analysis as they were demonstrated to maximise the tradeoff between base pairs incorporated and average and maximum contig size after thorough empirical testing. Assembled contigs of strain triplets were aligned against one another by BLASTn (1e-5 E-value threshold) to search for similarity between the sequences. Contigs that lacked similarity with others were designated as strain-specific DNA. These candidates were further validated using alignments of the short read data sets from both other strains using Novoalign. Contigs covered by reads were not considered to be strain-specific.

Validated strain-specific contigs were aligned using BLASTx against the UniProt database (Apweiler *et al.*, 2004) to identify sets of known (present in other *P. aeruginosa*) and novel (not present in other *P. aeruginosa*) genes in their accessory genomes.

5.3.8 Detection of horizontally transferred genomic elements in clone CHA

Assembled contigs of the three clone CHA strains were aligned against all known *P. aeruginosa* genomic islands and insertions in regions of genome plasticity using BLASTn (1e-10 E-value threshold). Alignment results for all the searches were then visualized by GenomeGraphs (Durinck *et al.*, 2009), an integrated genomic data visualization package for R (www.r-project.org) to help determine which of the known horizontally transferred genomic elements are completely/partially present in the three clone CHA strains.

5.3.9 Check for conservation of predicted sRNAs

Uncovered regions of the reference were extracted from the alignment results for the individual strains and checked for intersection with the 557 sRNA loci described for the PAO1 reference (Gómez-Lozano *et al.*, 2012). Complete or partial absence (> 10 % not conserved) was confirmed by visual inspection of alignment/coverage for these loci using the Integrative Genomics Viewer (Thorvaldsdóttir *et al.*, 2013).

5.4 Results

5.4.1 Origins of the *P. aeruginosa* clone CHA and clone TB strains

The clone CHA strains CHA, 491 and PT22 were isolated from sites in Grenoble, Hannover

and Mülheim in 1990, 2005 and 1992, respectively. Strain PT22 was isolated from a river, whereas strains CHA and 491 are CF airway isolates. Strain CHA was recovered from a critically ill CF patient with advanced lung disease and chronic *P. aeruginosa* infection (Toussaint *et al.*, 1993). Strain 491 was the first clone CHA isolate from respiratory secretions of a female CF patient with normal lung function (Wiehlmann *et al.*, 2012). The strain was successfully eradicated from the patient's airways by antipseudomonal chemotherapy and no further clone CHA strain has since been identified in the patient's respiratory secretions. The three clone TB strains were isolated from a burn wound (strain TB63741) and two unrelated CF patients (strains TBCF10839 and TBCF121838 (Klockgether *et al.*, 2013)) during a local outbreak at Hannover Medical School in summer 1983.

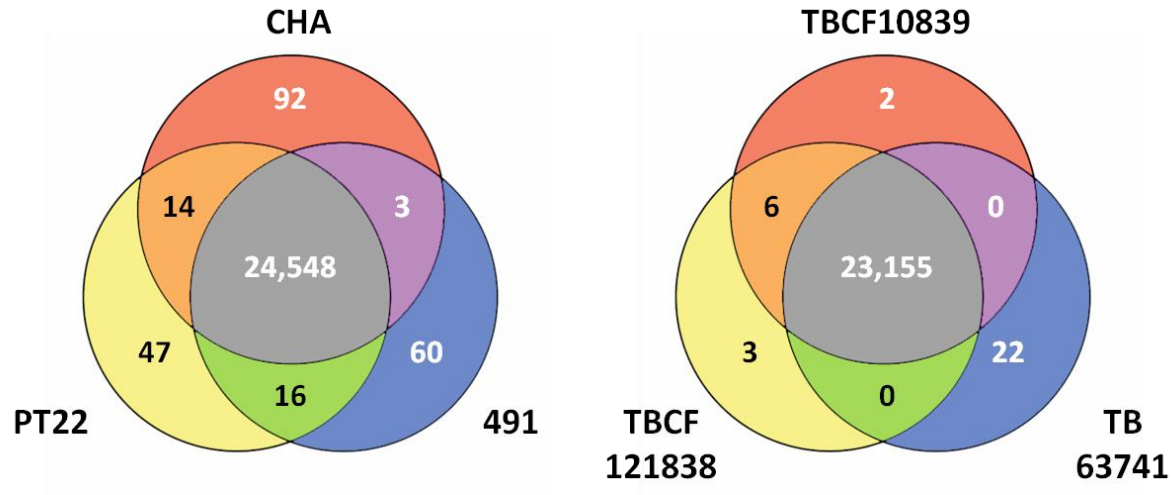


Figure 5.1: Venn diagrams of SNPs in clones CHA (left) and TB (right) of *P. aeruginosa*. SNP numbers are based on the alignment to the *P. aeruginosa* PAO1 reference sequence.

5.4.2 Shotgun genome sequencing

Fragment libraries of CHA, 491, PT22 and TB63741 were sequenced with the Illumina Genome Analyser II generating 36 bp reads as previously reported for strains TBCF10839 and TBCF121838 (Klockgether *et al.*, 2013). Reads passing quality criteria (Cramer *et al.*, 2011) were mapped to the PAO1 genome sequence ((Stover *et al.*, 2000); NCBI sequence NC_002516.2) in order to detect SNPs, indels and PAO1 loci absent in clones CHA and TB. Contigs representing the non-PAO1 loci of the accessory genome were de novo assembled from reads that could not be mapped to the PAO1 reference.

5.4.3 Comparison of the clone CHA genomes with the PAO1 genome

5.4.3.1 Replacement islands

The *P. aeruginosa* core genome harbours a few loci that are subject to diversifying selection. These make up one categorical group of the accessory genome known as replacements islands. Such are gene clusters that compose of bacterial fitness associated elements such as Lipopolysaccharide (LPS) O antigen, pyoverdine, pili, and flagella which are highly divergent between different strains. Clone CHA is equipped with LPS serotype 06, pyoverdine type IIa, a type-a2 flagellin and a novel type I pilin variant.

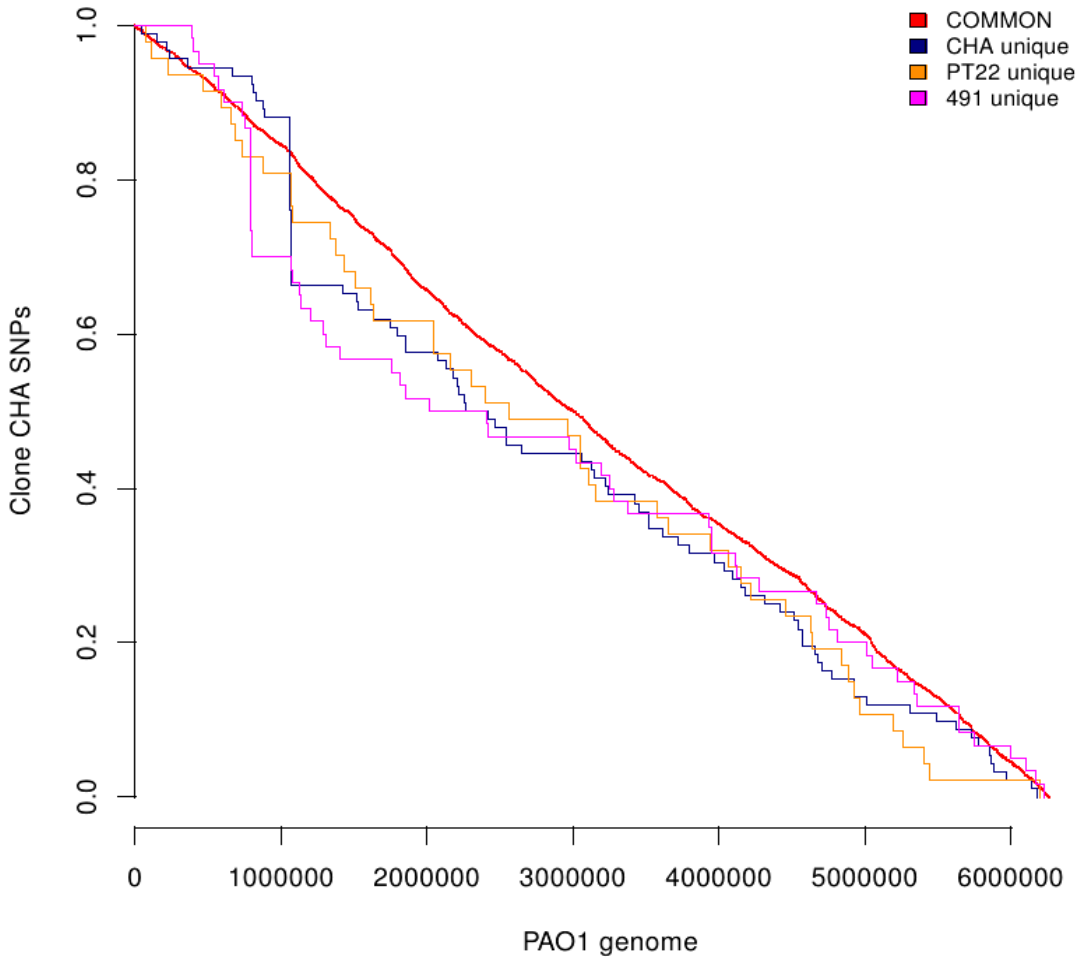


Figure 5.2: Kaplan-Meier curves of the proportions of *P. aeruginosa* clone CHA SNPs. Common synonymous and non-synonymous SNPs found in a) all three clone CHA strains and b) each of the three strains were plotted against genome position in *P. aeruginosa* PAO1. A flat horizontal line indicates that no SNPs were found in that region, while vertical lines illustrate a hotspot of SNPs at this genomic location. The red line shows that SNPs common to all three are evenly distributed throughout the genomes.

5.4.3.2 Common SNPs

The three clone CHA genomes shared 24,548 nucleotide exchanges (Figure 5.1 and Additional file 1 on the BMC Genomics website) compared to the PAO1 reference sequence,

which were evenly distributed in the genome (Figure 5.2). 503 of these lead to a non-conservative replacement of an amino acid as defined by a Dayhoff similarity index (Dayhoff, 1978) of less than 5 (Additional file 2 on the BMC Genomics website).

Table 5.1 lists these amino acid changes in the 25 proteins whose function have been experimentally demonstrated in *P. aeruginosa* (annotation class I, (Winsor *et al.*, 2011)). Besides a few proteins involved in DNA replication or secondary metabolism, the remaining proteins are transcriptional regulators, members of two-component systems, virulence effectors or are directly or indirectly involved in secretion or biofilm formation. Non-conservative amino acid replacements were neither observed in any enzyme of the core or intermediary metabolism nor in any component of the basic transcriptional or translational apparatus. This comparison of the PAO1 and clone CHA genomes suggests that diversifying selection with impact on protein function has preferentially affected *P. aeruginosa* genes that encode elements for communication with the environment.

5.4.3.3 Indels

Nineteen small indels (< 4 bp) were identified in the coding region of the clone CHA genomes (Table 5.2), 14 of which were already known from other completely sequenced *P. aeruginosa* strains. The three frameshifts in the last codons of PA3124 and PA4161 or the stop codon of PA5282 are neutral sequence variations and the three in-frame indels in PA2091, PA2302, and PA3462 should modulate the function of the encoded gene products to only minor extent, but the majority of the other 13 out-of-frame indels are probably loss-of-function mutations.

Five of the 19 indels are as yet undescribed in the Pseudomonas Genome Database (August 2012). Two of these have no functional consequences as mentioned above (PA3124, PA5282) and one destroys the reading frame of a chemotaxis transducer (PA4915). The remaining two mutations are located in the first ORFs of RGP2 and RGP7, both of which are known to carry clone-specific accessory elements and to be hotspots of genome mobility (Klockgether *et al.*, 2011). The frameshifts inactivate transposase/integrase genes and thus should fix these tRNA-associated genomic islands in the clone CHA genomes.

5.4.3.4 Gain and loss of start and stop codons

The loss of three start and stop codons each and the gain of eight premature stop codons were

noted in all three analysed clone CHA genomes (Table 5.3). Interestingly another premature stop codon was introduced into ORF PA0977 in all three strains at the same position but by divergent nucleotide exchanges, a transversion in two strains and a transition in the third strain, respectively. Two further nonsense mutations were exclusively identified in strain CHA (Table 5.3). The mutations affected transcriptional regulators, hypotheticals, glycolate oxidase and Glu-tRNA(Gln) amidotransferase operons. Thus basic bacterial functions of metabolism and translation are impaired or lost in *P. aeruginosa* clone CHA; i.e. glycolate utilization and the transamidation of misacylated Glu-tRNA^{Gln} to correctly charged Gln-tRNA^{Gln}.

Table 5.1. Non-conservative amino acid exchanges (Dayhoff matrix index < 5) in selected proteins* of clone CHA strains.

Locus_tag	Name	encoded product	aa exchange
PA0247	PobA	p-hydroxybenzoate hydroxylase	T98M
PA0595	OstA	organic solvent tolerance protein precursor involved in outer membrane biogenesis	M907T
PA0831	OruR	transcriptional regulator of ornithine utilization	W197C
PA1148	ToxA	exotoxin A precursor	F22S
PA1712	ExsB	exoenzyme S synthesis protein B	R52G
PA1717	PscD	type III secretion export protein	V346E
PA1718	PscE	type III secretion export protein	C40G
PA2236	PslF	glycosyl transferase, Psl exopolysaccharide biosynthesis	Y247D
PA3061	PelD	membrane-bound c-di-GMP-specific receptor regulating Pel exopolysaccharide production	Y208H
PA3063	PelB	Pel exopolysaccharide biosynthesis	W791L
PA3344	RecQ	ATP dependent DNA helicase	R571C
PA3805	PilF	pilus biogenesis, outer membrane pilotin for localization and multimerization of secretin PilQ	L243P
PA3810	HscA	molecular chaperone	R285G
PA3910	EddA	extracellular DNA degradation protein	P368L
PA3946	RocS1	sensor of two-component system controlling *cupC* fimbrial and efflux pump gene expression	I399S
PA4085	CupB2	periplasmic chaperone	H242L
PA4086	CupB1	major pilus subunit	Q102T; V154E
PA4776	PmrA	two component regulatory system modulating resistance to cationic antimicrobial peptides	L71R
PA4777	PmrB	two component regulatory system modulating resistance to cationic antimicrobial peptides	Y345H
PA5483	AlgB	two component response regulator controlling alginate biosynthesis	L382R
PA5484	KinB	two component sensor kinase (negative regulation of alginate production, positive regulation of virulence-associated phenotypes)	Y50H

PA5493	PolA	DNA polymerase I	C882R

Exchanges are given in comparison to protein sequences from the PAO1 reference.
*The function of the encoded gene product has been experimentally demonstrated (annotation class I).

Table 5.2. Small indels in the clone CHA genome compared to the PAO1 genome.

Indel-pos.[a]	Change	Locus_tag	Annotation	Indel known
288750	-AT	PA0257	put. integrase/transposase, first ORF of RGP2	no
740420	+C	PA0683	HxcY, type II secretion system protein	yes[1]
995238	+T	PA0912	Hypothetical protein	yes[1]
1060785	+T	PA0977	Hypothetical, phage-like, first ORF in RGP7	no
1116214	+C	PA1029	Hypothetical protein, homology to antitoxin	yes[1]
1697856	+G	PA1559	Hypothetical, part of PmrA regulated operon	yes[2]
1835046	+C	PA1685	MasA, enolase-phosphatase E-1, part of methionine salvage pathway	yes[2]
2301796	-GGC	PA2091	Hypothetical protein	yes[3]
2355772	+G	PA2139	Hypothetical protein	yes[2]
2356683	-C	PA2141	Hypothetical protein	yes[2]
2533912	+GTC	PA2302	AmbE, non-ribosomal peptide synthetase	yes[3]
2753523	+C	PA2452	Similar to enterobactin esterase	yes[2]
3083197	+G	PA2727	Similar DNA helicase	yes[2]
3506327	-C	PA3124	Transcriptional regulator; deletion in last codon	no
3873151	-CCC	PA3462	Sensor kinase of two-component system	yes[1]
4657418	-A	PA4161	FepG, ferric enterobactin transport protein; last codon, no change of coding sequence	yes[4]
4888195	+G	PA4360	Hypothetical, chromosome segregation protein, SMC-like; disruption of start codon	yes[5]
5515497	-A	PA4915	Chemotaxis transducer	no
5945963	+C	PA5282	Major facilitator transporter	no

[a]: position according to PAO1 reference sequence NC_002516.
Indel is listed in the Pseudomonas Genome Database for completely sequenced *P. aeruginosa* genomes: [1]for strains PA14, 2192, C3719, PACS2, 39016; [2]for strains PA14, 2192, C3719, PACS2, 39016, PA7; [3]for strains 2192, C3719, PACS2; [4]for strains PA14, PACS2, 39016; [5]for strains PA14, 2192, C3719, PA7, 39016.

Table 5.3. SNPs causing gain or loss of start and stop codons in *P. aeruginosa* clone CHA genomes.

Locus_tag	SNP-pos.[a]	SNP	Pos. in aa-seq.	Length of aa-seq.	Annotation
Clone CHA [CHA, PT22 and 491] common SNPs – stop codons gained:					
PA0089	325546	C-T	308	321	Transcriptional activator GpuR
PA1261	1369435	G-A	220	225	Probable transcriptional regulator
PA1427	1553550	G-T	147	188	Hypothetical protein
PA2691	3045894	G-A	87	402	Conserved hypothetical protein
PA4482	5013957	C-A	96	97	Glu-tRNA(Gln) amidotransferase subunit C
PA4982	5598104	G-A	58	999	Probable two-component sensor
PA5342	6010696	C-T	121	267	Probable transcriptional regulator
PA5353	6020049	G-A	356	409	Glycolate oxidase subunit GlcF
Clone CHA [CHA, PT22 and 491] common SNPs - stop codons lost:					
PA2456	2756650	A-G	114	114	Hypothetical protein
PA2566	2900372	T-G	396	396	Conserved hypothetical protein
PA6439	5206722	A-G	96	96	Hypothetical protein
Clone CHA [CHA, PT22 and 491] common SNPs - start codons lost:					
PA0819	895825	T-C	1	98	Hypothetical protein
PA2778	3136962	A-G	1	292	Hypothetical protein
PA5525	6218101	T-C	1	247	Probable transcriptional regulator
Divergent nucleotide exchange – stop codon gained:					
PA0977	1060555	A-C/T A-C in strains PT22 and 491, A-T in strain CHA	93	108	Hypothetical protein
Strain-specific SNPs in strain CHA only - stop codons gained:					
PA0734	802084	C-T	52	91	Hypothetical protein
PA5487	6178179	T-A	625	672	Hypothetical protein

[a]: position according to PAO1 reference sequence NC_002516.

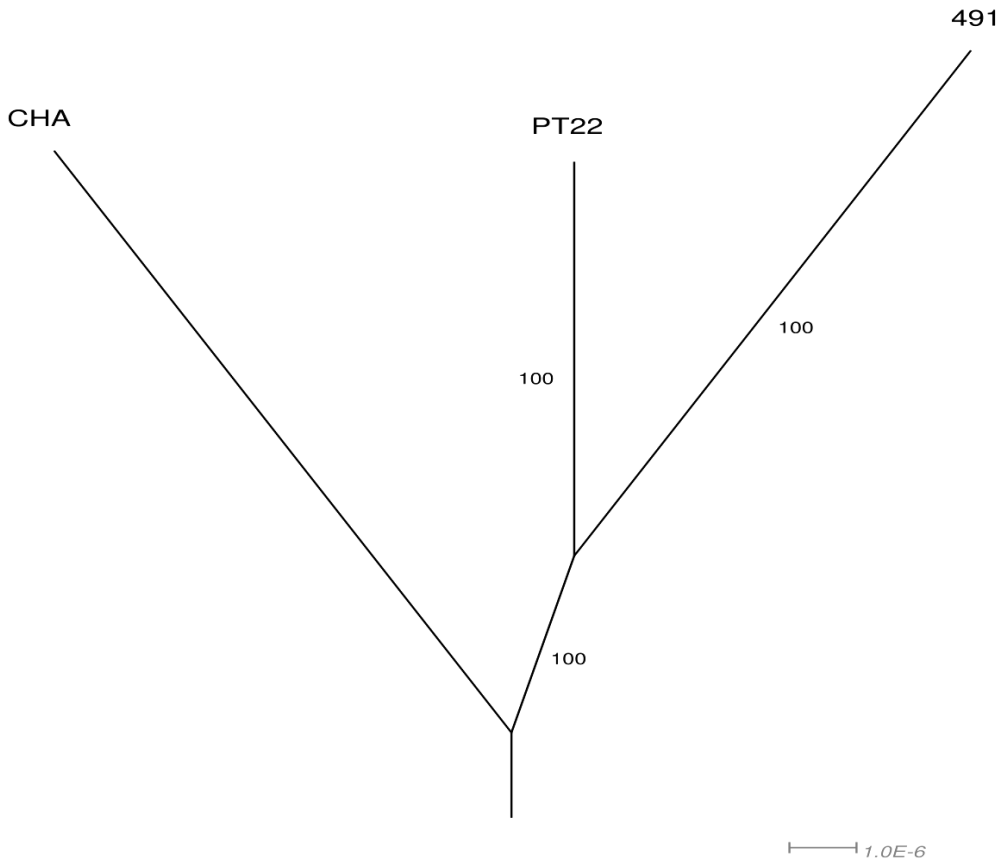


Figure 5.3. Phylogenetic network for *P. aeruginosa* clone CHA isolates based on identified SNPs. All SNPs mentioned in Figure 1 were incorporated into three pseudosequences derived from the PAO1 reference sequence by the script SequenceReplacer (available on request from the authors). The network was produced using the uncorrected P distance measure with normalisation followed by the NeighbourNet algorithm in the program Splitstree [62]. The scale indicates the number of substitutions per site. Numbers on the branches are 100 bootstrap resampling values which give a measure of the confidence of the displayed tree topology. A network for clone TB is not shown as the isolates display up to two orders of magnitude less divergence than clone CHA strains, which cannot be visualised appropriately.

5.4.3.5 SNPs shared by two clone CHA strains

Thirty one of 33 SNPs that were found in two, but not in the third CHA strain, are located in two regions of genomic mobility that are prone to horizontal gene transfer (Klockgether *et al.*, 2011) suggesting that these SNPs differentiate variants of phage-related sequences. The only two SNPs sensu stricto were identified in intergenic sequences (Additional file 3 on the BMC Genomics website).

5.4.3.6 Strain specific SNPs

The frequency of SNPs shared by two of the three strains was extremely low, but several dozen unique SNPs were found in each of the individual strains indicating some distinct microevolution in the clonally distant strain set (Figure 5.3). For instance, 47 strain-specific

SNPs were identified in the environmental isolate PT22 (Additional file 4 on the BMC Genomic website). The 34 SNPs in coding regions target genes encoding enzymes, transporters, transcriptional regulators and hypotheticals.

The genome of the CF isolate 491 carries 60 strain specific SNPs (Additional file 4 on the BMC Genomics website). The clade of strain 491 acquired non-synonymous SNPs in 31 ORFs including genes that should play a role during the colonization of CF airways. Serine-to-asparagine substitutions were present in the two-component response regulator AlgB which activates the transcription of the *algD* alginate biosynthesis operon (Leech *et al.*, 2008) and the cytoskeleton ATPase MreB which is essential for the maintenance of cell shape, chromosome segregation and polar localization of proteins (Cowles & Gitai, 2010). The most drastic change was the substitution of arginine by tryptophan R771W in the usher protein CupC3 that is essential for the assembly of CupC1 fimbriae (Ruer *et al.*, 2007). With 8 of the 60 strain-specific nucleotide exchanges in ORF PA0728, this gene encoding a phage-like integrase, was identified as a SNP hotspot in strain 491, and the unique SNPs were not evenly distributed over the whole genome (Figure 5.2).

Strain CHA carries most unique SNPs among the three sequenced isolates, i.e. 13 intergenic SNPs, 31 synonymous SNPs, 46 non-synonymous SNPs and two SNPs generating a stop codon (Additional file 4 on the BMC Genomic website). The predicted amino acid sequence was changed in 37 proteins including seven enzymes, six transporters and 15 ones of unknown function. Moreover, the clinically highly virulent strain CHA had acquired missense mutations in seven genes that are key for pathogenicity and adaptation to a habitat such as the CF lungs, i.e. A5G MucA, A651P PelB, R101H ExsA, R156H Tse2, L116F WspA, D514Y PA4036, E721K CbrA. The latter three missense mutations affect the chemotaxis operon WspABCDEF and two sensor kinases of two-component systems. CbrA has been demonstrated to be a global regulator of metabolism; motility; virulence; and antibiotic resistance (Nishijyo *et al.*, 2001; Li & Lu, 2007; Yeung *et al.*, 2011). Hence the E721K mutation in CbrA should be a pleiotropic modifier of the bacterial phenotype.

Table 5.4. Strain-specific losses of PAO1 DNA.

Locus_tag	Description
PA0977-0987 (RGP7)	region only partially conserved in all strains; ORFs PA0980-0981 absent in strain CHA only, ORFs PA0986-0987 absent in 491 only
PA0927-0928 (*ldhA*, *gacS*)	start of *ldhA* (278 nt) and end of *gacS* (146 nt) missing in strain CHA
PA1907	partial deletion (183 nt) in strain 491
PA2136	partial deletion (first 30 nt) in strain 491
PA2177	partial deletion (356 nt) in strain PT22

Complementation experiments demonstrated that the change of an alanine by glycine in the N-terminus of anti-sigma factor MucA (A5G) leads to the mucoid phenotype, whereas the (by definition) non-conservative exchanges L382R in AlgB and Y50H in KinB of the alginate regulon (Damron & Goldberg, 2012) were not causative for mucoidy in strain CHA (data not shown). The non-conservative missense mutation in the *pel* operon may influence the biosynthesis of the Pel exopolysaccharide and thereby modify the antibiotic tolerance and stability of the bacterial biofilm (Mann & Wozniak, 2012). The unique ability of strain CHA among functionally characterized *P. aeruginosa* to induce oncosis of neutrophils and macrophages is critically dependent on its active type III secretion system (Dacheux *et al.*, 2001). Whether the undescribed arginine-to-histidine substitution R101H in ExsA, the regulator of the type III secretion regulon, has an effect on the regulon's activity, is unclear. A non-conservative R156H mutation was observed in Tse2, a recently discovered substrate of the type VI secretion system of *P. aeruginosa* (Hood *et al.*, 2010). Toxin Tse2 inhibits the growth of competing bacterial cells, but the impact of the change on the competitive fitness of strain CHA is uncertain since Tse2 is not expressed in the CHA background (data not shown). Comparable to strain 491, hotspots of strain-specific nucleotide exchanges could also be found in strain CHA, as ORFs PA0982 and PA0977, both located in a region known for genomic instability (Klockgether *et al.*, 2004), had acquired nine and three SNPs, respectively.

Twelve, six and five strain-specific SNPs were identified in intergenic regions of strains CHA, PT22 and 491; one of which in each strain affected different sRNAs. Seven strain CHA - specific SNPs were found in the intergenic regions of PA0977-PA0978 (four SNPs) and PA0983-PA0984 (three SNPs) and thus located in the same region prone to genomic instability as 12 of the strains' unique intragenic SNPs (in PA0977 and PA0982).

5.4.3.7 PAO1-DNA absent in clone CHA strains

The clone CHA genome lacks 117 PAO1 ORFs (2.1% of all ORFs) the majority of which encode pyocins, phage elements or functionally yet uncharacterized gene products (see Additional file 5 on the BMC Genomics website). Twelve PAO1 ORFs only partially aligned with clone CHA sequence reads indicating that sequence variation is unusually high in these ORFs. All three clone CHA genomes also lack the small non-coding RNA gene (sRNA) *phrD*, that is part of a phage-like insertion in PAO1, and 39 of the 513 intergenic sRNA loci identified recently (Gómez-Lozano *et al.*, 2012) [34]. Another 21 of these loci were only partially covered by sequence reads of the clone CHA strains (Additional file 6 on the BMC Genomic website). Intraclonal differences were observed for two sRNA loci. The sRNA pant78 was absent in strain 491 only while pant106 was present in strains PT22 and 491 but absent in strain CHA. Both these pant-sRNAs are located in RGP-insertions in PAO1 (RGP5 or RGP7, respectively) and thus likely contributed to mobile DNA elements.

Strain-specific intragenic deletions of PAO1 coding sequence were observed for two ORFs in strain 491 and one ORF in strain PT22 (Table 4 and Appendix B Supplementary Figure S1). Strain CHA showed a 426 bp deletion and, due to that, lacks the last 146 nucleotides of the global regulator *gacS* (PA0928) and the first 278 nucleotides of the adjacent lactate dehydrogenase *ldhA* (PA0927). This two-gene spanning deletion generated a double mutant of key genes of lifestyle and metabolism of *P. aeruginosa* (Goodman *et al.*, 2009; Petrova *et al.*, 2012).

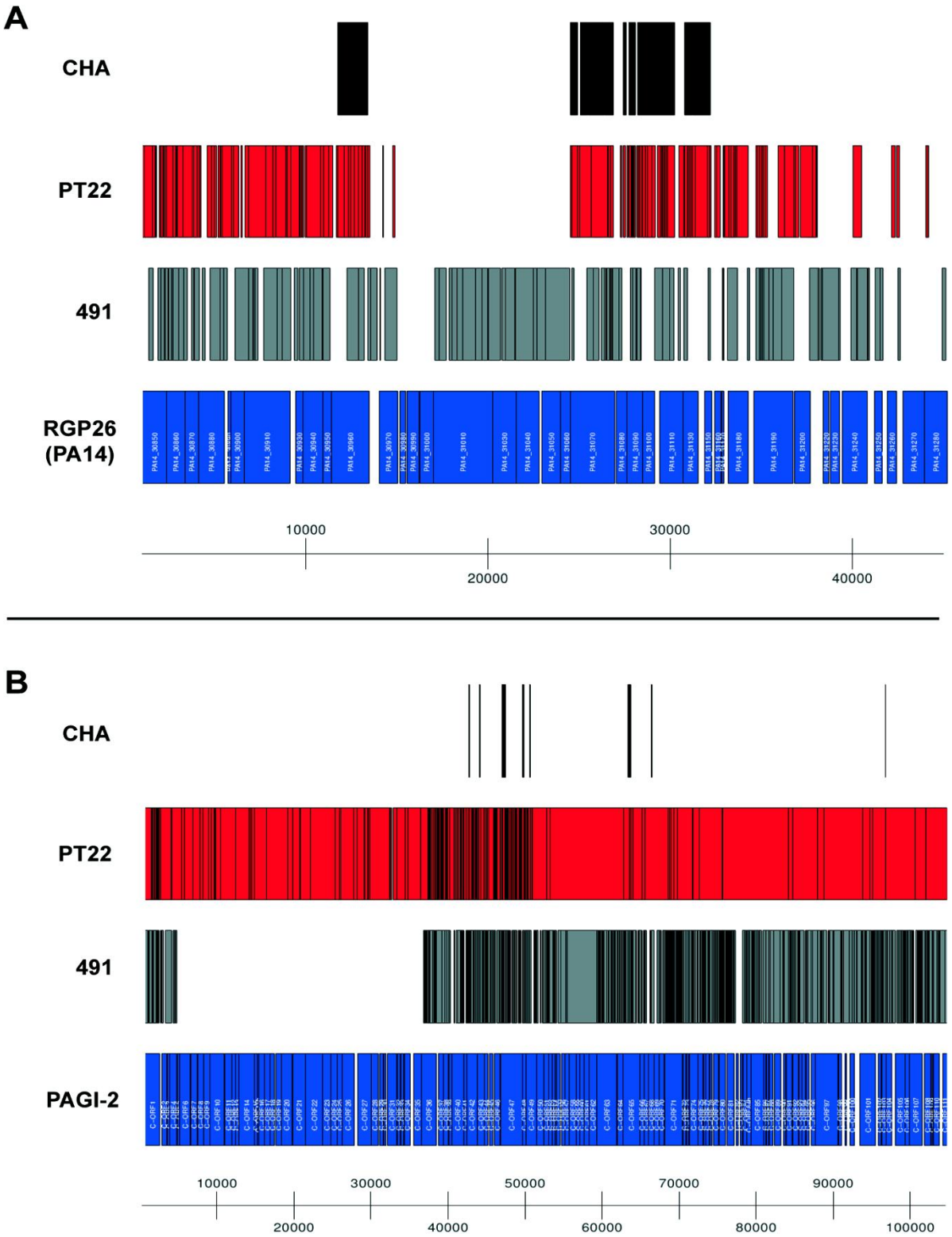


Figure 5.4: Diversity of the *P. aeruginosa* clone CHA accessory genome. As examples, BLAST alignments of de novo assembled not-in-reference accessory genome contigs from all three clone CHA isolates to the PA14 Region of Genome Plasticity (RGP) 26 (panel A) and the PAGI-2 genomic island (panel B) are shown. Contigs from CHA are coloured black, those from PT22 are red and contigs from 491 are gray, while the dark blue boxes represent the annotated ORFs from the *P. aeruginosa* PA14 genome and the PAGI-2 genomic island, respectively. For details on the annotated ORFs, please refer to the respective original publications ((Mathee *et al.*, 2008) for RGP26 from PA14, (Larbig *et al.*, 2002) for PAGI-2). PT22 harbours a complete copy of the PAGI-2 island, while 491 has a partial copy and it is absent in CHA. Figures were produced using the R package Genomegraphs (Durinck *et al.*, 2009).

5.4.4 The clone CHA accessory genome

5.4.4.1 Accessory DNA elements known from other *P. aeruginosa* clones

The clone CHA strains share several genomic islands with the transmissible Liverpool epidemic strain LESB58 (Winstanley *et al.*, 2009) (Table 5.5). CHA, PT22 and 491 harbour copies of LES-prophage 1, LESGI-2 and LESGI-4 of the LES strain and a copy of an RGP29-insertion in the completely sequenced strain PACS2. The three strains moreover share a few ORFs known from insertions in RGPs 6, 9, 27, 36 and 62 in other *P. aeruginosa* genomes (Klockgether *et al.*, 2011) (Table 5.5), although none of these insertions is completely conserved in the clone CHA genomes. Otherwise interstrain diversity is pronounced among the three sequenced clone CHA strains. Each strain carries its specific set of accessory elements. Individual variants were identified for the partially covered RGP26 (Figure 5.4A) and RGP77 insertions in strain PA14 or PA7, respectively, and for the mobile PAGI-2/pKLC102-type genomic islands. The clone CHA strains also harbour different sets of phage phiCTX-like genes. Variants of this phage either containing or lacking the cytotoxin gene *ctx* have been described for *P. aeruginosa* (Nakayama *et al.*, 1999), and apparently such different variants have been acquired by the clone CHA lineage, as the *ctx* gene is conserved in PT22 and 491, but not in strain CHA.

The environmental isolate PT22 is endowed with the largest accessory genome. It carries several ORFs of RGP42 and RGP63 and nearly identical copies of the genomic islands LESGI-3 of strain LESB58 (Winstanley *et al.*, 2009) and PAGI-2 of strain C (Larbig *et al.*, 2002) (Figure 5.4B, Table 5.5). Strain 491 harbours variants of PAGI-2 and LESGI-3 and phage sequences that are homologous to ORFs in LES-prophages 3 and 6, the latter of which is also found in strain CHA.

Table 5.5. Accessory DNA elements from other P. aeruginosa genomes detected in strains CHA, PT22, and 491.

From defined genomic islands

Name	No. of ORFs	Present parts
PAGI-2 (RGP29)	111	strain PT22: complete 105 kb island (> 99.9 %); strain 491: ORFs C1-4; C36–111 (80 – 100 %)
PAGI-5 (RGP7)[a]	121	strain 491: complete 99.4 kb island (> 99.2 %)
PAGI-6 (RGP87)	47	**phage CTX-like ORFs 6PG1–28 and 6PG32–38 (86.8 – 100 %)**
PAGI-8 (RGP62)	12	**ORFs 8PG1; 8PG7-8 (85.6 – 95.2 %)**
pKLC102 (RGP7)[a]	105	strain CHA: ORFs CP1–3; CP12–14; CP27; CP30-31; CP34–44; CP50–54; CP57–83; CP87–89; CP102-103 (94.8 – 99.3 %); strain PT22: ORFs CP1–3; CP9–14; CP18–21; CP26–27; CP29–30; CP33–45; CP47–56; CP58–83; CP87–93; CP102-103 (84.3 – 99.5 %)
LESGI-2 (RGP85)	18	**complete 31.7 kb island (98.8 – 100 %)**
LESGI-3 (RGP27)	115	strain PT22: complete 110.6 kb island (90.4 – 100 %); strain 491: PLES_26051–26061; PLES_26211–26221; PLES_26421–27102 (81.7 – 100 %)
LESGI-4 (RGP23)[b]	31	**complete 39.4 kb island (97.4 – 100 %)**
LES-prophage 1 (RGP 3)	19	**complete 14.8 kb island (81.6 – 100 %)**
LES-prophage 3 (RGP82)	51	strain 491: homologs to 18 ORFs (88.3 – 98.4 %)
LES-prophage 6 (RGP10)	12	strains CHA and 491: PLES_41181 – 41241 (90.7 – 100 %); PLES_41191 only partially covered

From other RGP insertions

RGP (host strains)	No. of ORFs	Present parts
RGP6 (2192)	41	**PA2G_05961-05962 (> 99.7 %)**
RGP9 (2192)	14	**PA2G_00059-00065; PA2G_00072 (95.1 – 100 %)**

RGP26 (PA14)	39	strain CHA: PA14_30960; PA14_31070 – 31150 (84.8 – 95.2 %); strain PT22: PA14_30850–30960; PA14_31070–31200 (81.1 – 98.6 %); strain 491: PA14_30850–30970; PA14_31110–31250 (79.7 – 97.2 %)
RGP27 (PACS2)	74	strain 491: PAERPA_01003080–3085; PAERPA_01003110; PAERPA_01003119–3120; PAERPA_01003136–3154 (84.5 – 100 %)
RGP29 (PACS2)	10	**complete RGP-insertion (98.5 – 100 %)**
RGP35 (2192)	43	strain 491: PA2G_02937–2942; PA2G_02953; PA2G_02956–02957; PA2G_02961–02963; PA2G_02965; PA2G_02969; PA2G_02972-02973 (92.2 – 100 %)
RGP36 (PA14)	31	**PA14_15620-15630; PA14_15650-15660 (96.4 – 99.7 %)**
RGP42 (2192)	11	strain CHA: PA2G_05286-05290 (97.1 – 99.5 %); strain 491: PA2G_05286–05292 (95.4 – 100 %)
RGP42 (PA7)	54	strain PT22: PSPA7_5339-5340 (85.1 – 89.1 %)
RGP63 (PA7)	72	strain PT22: PSPA7_0075 (86.3 – 90.5 %); PSPA7_0108-0114 (> 99.9 %)
RGP77 (PA7)	53	strain CHA: PSPA7_3708; PSPA7_3723; PSPA7_3726-3734 (83.4 – 93.3 %); strain PT22: PSPA7_3696-3708; PSPA7_3723; PSPA7_3726–3735; PSPA7_3738-3747 (79.5 – 100 %); strain 491: PSPA7_3696-3708; PSPA7_3723; PSPA7_3726–3729; PSPA7_3731-3733; PSPA7_3738-3740; PSPA7_3747 (79.3 – 100 %)

Present parts printed in bold are conserved in all three clone CHA strains. Pairwise % nucleotide identity of the corresponding sequence contigs is given in brackets.
[a]: majority of assigned contigs mapped on both PAGI-5 and pKLC102 references which share a large set of highly homologous genes.
[b]: contigs also mapped on island PAGI-1, a variant of LESGI-4.

5.4.4.2 Novel strain-specific genes

ORFs were designated as 'novel genes' if they had yet not been described in completely sequenced *P. aeruginosa* genomes deposited in databases by June 1st, 2012. The number of novel genes correlated with the genome size of the strain, i.e. least genes were identified in strain CHA and most genes were detected in strain PT22 (see Additional files 7 – 9 on the BMC Genomics website).

The strain CHA genome incorporated a truncated variant of the *Pseudomonas* phage B3 (Braid *et al.*, 2004) and an *aacC1* gene that confers resistance to aminoglycoside antibiotics. The *aacC1* sequence contig probably originated from an enterobacterial integron that has the highest homology to the enterobacterial type I integron harboured by plasmid p1658/97 (Zienkiewicz *et al.*, 2007).

Annotation uncovered 114 strain-specific ORFs in the CF isolate 491 (see Additional file 9 on the BMC Genomic website). Most ORFs to which a function could be ascribed encode enzymes of DNA metabolism or mobility or elements of conjugation and type IV secretion. The closest ortholog or homolog was identified for all ORFs in beta- or gamma-proteobacteria that have been classified in the pre-16S rDNA taxonomic era as 'honorary pseudomonads' because they share lifestyle, habitat and metabolic versatility with the 'class I' pseudomonads *P. aeruginosa*, *P. putida*, *P. fluorescens* and *P. syringae* (Palleroni, 2003). Twenty-five ORFs are shared with the metal-resistant Burkholderiales *Herminiimonas arsenicoxydans* (Muller *et al.*, 2006). These genes are part of PAGI-2 like islands harboured by strain 491 (Figure 5.4B) and the beta-proteobacterium, but none of them as annotated as a metal-resistance contributor.

167 strain-specific ORFs were identified in the aquatic isolate PT22 (see Additional file 8 on the BMC Genomic website). Like in strain 491, closest orthologs and homologs were detected exclusively among beta- and gamma-proteobacteria, but other genera, namely Acidovorax, Azoarcus, Cupriavidus, Ralstonia (26% of ORFs) and the true pseudomonads (47% of ORFs) were frequent among the closest relatives of PT22 ORFs. The function could be predicted for a larger proportion of ORFs than in the CF isolates, and a greater variety of functions could be addressed which is reflected by a much more diverse spectrum of functional categories/gene ontologies for the PT22-specific ORFs than for those specific for strains CHA or 491 (see Appendix B Supplementary Figure S2). The strain-specific

accessory genome of strain PT22 encodes enzymes of lipid and sulphur metabolism, the two-component system *armRS*, a heme lyase and a cytochrome C oxidase and multiple transporters including an efflux pump and a P-type ATPase for heavy metal ions (Additional file 8 on the BMC Genomic website). Moreover a paralog of the *P. aeruginosa* gene *mvaT* was identified. MvaT belongs to the H-NS family of small DNA-binding proteins that are global regulators of gene expression. Five homologues have been identified in P. putida and two homologues *mvaT* and *mvaU* have been identified in the P. aeruginosa core genome (Li *et al.*, 2009). *P. aeruginosa* PT22 is thus the first known *P. aeruginosa* strain with three *mvaT* homologues.

5.4.4.3 Comparison of the clone TB genomes with the PAO1 genome

In contrast to the analysed clone CHA strains, little intraclonal genomic diversity was observed for the three clone TB strains that were sampled during a local outbreak at Hannover Medical School. As reported earlier, only five individual nucleotide exchanges and one deletion each in a pilus assembly gene could be detected in the two CF airways isolates TBCF10839 and TBCF121838 (Klockgether *et al.*, 2013). Though many phenotypic differences were observed, also the accessory genome differed by only one 81 kb *Ralstonia pickettii* PAGI-2 like genomic island absent in the first but present in the latter isolate (Klockgether *et al.*, 2013).

Sequencing of a third clone TB isolate, the wound isolate TB63741, revealed some more intraclonal diversity, but still less than observed for the three clone CHA strains. TB63741 lacked six nucleotide exchanges that were detected for both TB CF-isolates, but carried 22 individual SNPs not seen in any of the two CF isolates (Figure 5.1 and Additional file 10 on the BMC Genomic website). TB63741 did not harbour any deletion in a *pil* gene, but it had acquired a 9-bp in-frame deletion in a two component sensor gene and two frame-shift mutations in a phage gene and in *oprD* (see Additional file 10 on the BMC Genomic website). The porin OprD transports basic amino acids and peptides but it also takes up the antipseudomonal agent imipenem. Loss-of-function mutations in *oprD* as seen in the clinical isolate TB63741 are a common mechanism of imipenem resistance (Li *et al.*, 2012).

Similar to the clone CHA lineage, the conservation of described non-coding sRNA loci does not differ within the clone TB lineage apart from one exception. The sRNA *oprD* and 30 pant-sRNAs are absent in the three genomes, of another 10 pant-sRNA loci significant parts were lacking (see Additional file 6 on the BMC Genomic website). The phage DNA-

associated sRNA pant78, present in both CF-isolates but absent in TB63741 made up the only intraclonal difference regarding sRNAs in clone TB.

Comparison of the sRNA conservation in clonal lineages CHA and TB revealed clone-specific patterns. While *oprD* and 20 pant-sRNA loci from PAO1 were completely absent (and four more partially) in both lineages, clone CHA lacked 17 pant-sRNAs which were present in clone TB. Six pant-sRNAs, however, were absent in clone TB but fully conserved in clone CHA. For another 23 pant-sRNA loci conservation patterns were partially divergent in the two clonal lineages (see Additional file 6 on the BMC Genomic website). According to that, varying spectra of small non-coding RNA genes in *P. aeruginosa* might contribute significantly to interclonal diversity but only to a small degree to diversity between clonal variants, if sRNA genes are parts of strain-specific acquisition of mobile DNA elements.

Clone TB is endowed with a large accessory genome including the genomic islands PAGI-1, PAGI-2, PAGI-5 and PAGI-6 (Klockgether *et al.*, 2013). The wound isolate TB63741 lacks the 81 kb TBCF121838-specific *R. pickettii* genomic island and numerous phage-like ORFs of phage Pf1 and of genomic island LESGI-1 which were present in both CF isolates. Conversely, TB63741 has incorporated more than 300 kbp that are absent in the two CF strains. Virtually all this DNA is of phage origin including LES-prophage 2 and 3 sequence (Winstanley *et al.*, 2009), of which 67.3 or 76.2 %, respectively, of the DNA were found in TBCF63741 with nucleotide identities ranging from 80 to 100 %. The closest homologues of accessory genome ORFs were found in other *P. aeruginosa* clones, other *Pseudomonas* taxa or in 'honorary' pseudomonads (see Additional file 11 on the BMC Genomic website). The shuffling of phage DNA apparently was the major driving force of microevolution of clone TB during the outbreak.

5.5 Discussion

5.5.1 Comparison of the sequenced clone CHA and clone TB genomes

This study compared the intraclonal genome diversity of *P. aeruginosa* isolates derived from common and divergent sources. Consistent with our expectation higher genomic variation was found among the clonal isolates with a more diverse spatiotemporal origin.

Sequence variation was low among the three clone TB strains that had been sampled in summer 1983 during a local outbreak. The two CF isolates belong to a small epidemic that tripled the prevalence of *P. aeruginosa* – positive patients at the CF clinic (Tümmler *et al.*,

1991). Despite individual profiles of phenotype, strains TBCF10839 and TBCF121838 show only minute differences in their genome sequence (Klockgether *et al.*, 2013). Strain TB63741 was isolated from a patient with severe burns who had been treated at the intensive care unit for burns from which clone TB had initially spread to surgical wards and later to the CF clinic. The ancestors of the TB63741 strain had incorporated numerous phages into the clone TB genome that were absent in the isolates from the CF lungs indicating that highly colonised burn wounds themselves and/or the associated hospital environment had tolerated or favoured the uptake of phages.

The three clone TB isolates had descended from a common source and the individual clades had diverged from each other by at most two years. In contrast, the three sequenced clone CHA isolates were sampled from spatially and temporarily distinct habitats. Correspondingly, the sequence of the core genome and the composition of the accessory genome were significantly more diverse among the three clone CHA than among the three clone TB strains. In particular, the numerous strain specific SNPs in absence of pairwise shared SNPs demonstrate the distinct microevolution of the clone CHA strains (Figure 5.3). Conversely, shared *de novo* mutations and comparably very few individual *de novo* mutations highlight the close relatedness of the two clone TB CF isolates.

The environmental isolate PT22 was endowed with the largest accessory genome of the investigated strains. PT22 was collected from the river Ruhr at a site with substantial anthropogenic pollution and contamination with industrial sewage (Ruhrgütebericht 1992). Consistent with its source, the genomic islands of PT22 encoded genes for the detoxification of xenobiotics and the efflux of heavy metal ions. PT22 carried a copy of PAGI-2 which also exists in CF isolates and *Cupriavidus metallidurans* CH34 that had been sampled from an industrial site polluted with heavy metal ions (Klockgether *et al.*, 2007; Diels *et al.*, 2009).

The CF airways isolates 491 and CHA were retrieved from patients with the extremes of the general state of health that are feasible with CF as the underlying predisposing condition. The clinically highly pathogenic strain CHA was isolated from a CF patient with end-stage lung disease, whereas strain 491 was recovered from an individual with normal anthropometry and excellent lung function. Strain 491 was eradicated by antipseudomonal chemotherapy and no clone CHA strain has yet been re-isolated from the patient's respiratory secretions in the last seven years. 491 had gained numerous elements of genomic mobility that may confer some global fitness to the strain, but only a few amino acid substitutions in traits that may facilitate

the colonization of CF airways. In other words, the microevolution of the 491 clade does not point to any pronounced selection of the 491 ancestry to accommodate itself to the CF lung habitat.

Conversely, the ancestors of the strain CHA isolate had selected numerous non-conservative amino acid substitutions in elements of chemotaxis, exopolysaccharide biosynthesis, motility and virulence. In addition, the genes *gacS and ldhA* were destroyed by a deletion. The lactate dehydrogenase LdhA has recently been demonstrated in strains PA14 and PAO1 to be indispensible for microcolony formation in biofilms (Petrova *et al.*, 2012). Hence, deletion of the 3' end of *ldhA* could alter biofilm formation although strain CHA displayed mucoid growth on agar plates (data not shown). The GacS/GacA two-component system controls the reciprocal expression of acute and chronic virulence determinants (Goodman *et al.*, 2009; Gómez-Lozano *et al.*, 2012). The deletion of *gacS* should abrogate this control. Consistent with this interpretation, strain CHA strongly expresses the pathways for alginate biosynthesis, a hallmark of a chronic infection, and the virulence effectors and structural elements of type III secretion, a hallmark of an acute infection (mRNA microarray data from bacteria grown to stationary phase, data not shown). Deletions and point mutations in key determinants of virulence and the control thereof thus established a genetic repertoire in the strain CHA isolate that is distinct from 491 and PT22 and should translate into the observed high pathogenic potential in the predisposed human host. This microevolution towards virulence seems to be quite specific for the inhabited CF lungs because strain CHA was inconspicuous in standard *P. aeruginosa* worm and fly infection models (Fauvarque *et al.*, 2002). Strain CHA apparently acquired signatures of a host-specific pathogen, whereas the 491 and PT22 clades retained the balance between environmental organism and opportunistic pathogen.

The clone CHA and TB genomes share numerous prophages and genomic islands with the virulent and transmissible LES clone, which has caused substantial morbidity in the CF patient population in the UK (Winstanley *et al.*, 2009). The relatedness of their genomes may explain why these clones are prone to nosocomial spread among predisposed human hosts and why virulent clades with uncommon pathogenicity traits have evolved in these clonal complexes. Subsequent evolvement of pathogenicity arising from such genomic predisposition proceeded differently than in the highly virulent examples TBCF10839 and CHA.

In the case of TBCF10839 only few sequence variations clearly differentiated its genome from that of the other two less virulent TB strains, mainly a loss-of-function mutation in TBCF10839 (Chang *et al.*, 2007). While lacking of type IV pili on the surface and being impaired in twitching motility, TBCF10839 was metabolically more active (Klockgether *et al.*, 2013), produced more outer membrane transporters and secreted more virulence effectors (Arevalo-Ferro *et al.*, 2004) than its clonal variants. Apparently the loss of PilQ induced a global response in the TB background that is far beyond pilus biogenesis. Any further mutations that are necessary to generate the unique ability of TBCF10839 to grow in neutrophils must have already existed in the clone TB lineage. Strain CHA, however, exhibits numerous strain-specific gain- or loss-of-function mutations in global regulators or key pathogenicity factors that should be involved in the specific virulence features of strain CHA like its capability to cause oncosis of neutrophils (Dacheux *et al.*, 1999; Dacheux *et al.*, 2000; Dacheux *et al.*, 2001). Evolvement of the specific pathogenicity traits likely occurred by a series of microevolution events in this case.

5.6 Conclusions

Intraclonal genome diversity in the two investigated strain triplets presented in a low number of strain-specific de novo mutations in the core genome and a variable composition of the accessory genome. Shared SNPs were mainly observed between the two most closely related clone TB isolates from the outbreak. The number of strain-differentiating single nucleotide substitutions ranged from 7 to 154 SNPs for the most and the least related strain pair of clone TB and CHA, respectively. Correspondingly the intraclonal sequence variation of the *P. aeruginosa* core genome was 200- to 3000-fold lower than the interclonal sequence variation of 0.3 – 0.5%. In contrast to the highly conserved core genome a strain-specific signature was noted for the repertoire of phage-related sequences and genomic islands in the distantly related clone CHA strain trio. Strains shared islands and prophages that have first been reported in the transmissible LES strain, but they were distinct in their PAGI-2/pKLC102-type islands that recruit their cargo from the extensive gene pool of the honorary pseudomonads. According to the annotation this cargo as well as the strain specific SNPs confer individual traits on the respective strains to cope with the demands of their habitat from which they were isolated.

Chapter 6

6 Concluding Discussion

The cost effective ultra high-throughput sequencing techniques are increasingly producing immense amounts of bacterial genome sequences of different backgrounds. The integration of these data sets provides possibilities and biological insights to understand the evolutionary patterns that influence bacterial diversity and adaptation to a variety of environments. Comparative analysis studies have highlighted the differences between related and unrelated bacteria and the evolutionary processes that influence speciation. These studies have also advanced our understanding of the evolutionary mechanisms and the types of genes that might be required by bacteria for behavioural alterations. Comparisons of bacteria with the basis of their gene content have revealed that their genomes constitute species and strain specific sets of genes. The strain specific sets of genes are those which are known to have been acquired from other unrelated bacteria through horizontal gene transfer events. These are genes that provide bacteria with fitness traits which enhance their adaptational capacity.

Apart from the roles which are played by horizontally transferred genomic elements, several other factors such as accumulations of DNA mutations (SNPs and indels) were also determined to have an influence in bacterial evolution. These have capabilities to change the phenotypic characteristics of an organism from avirulent to virulent through single nucleotide exchanges and insertions or deletions of one or more nucleotides. Many of these nucleotide exchanges are influenced by the interaction of bacteria with their environments mainly to improve their survival chances. Such changes coupled with gene acquisitions may however have increased beneficial effects on bacterial lifestyles. Both these entities result in many variations between closely related organisms especially those from environments with different conditions. We therefore formulated a study that addresses micro-evolutionary changes between organisms which belong to the same clonal lineage but isolated from different geographic regions. The idea was influenced by the fact that many other studies focus on variations of clonal complexes which were isolated from the same environments. The study was conducted on the three *Pseudomonas aeruginosa* clone CHA isolates of unrelated geographical origins. Two (CHA and 491) of the three strains were isolated from CF patients whereas the third (PT22) was an isolate from a polluted river. These were compared to spatiotemporally related strains of clone TB which were reported to have caused

a nosocomial outbreak at the Hannover Medical School. The three TB strains were isolated from different patients. Comparative analyses of the three clone CHA isolates revealed increased strain specific differences in their core genomic DNA mutations and horizontally acquired genomic elements whereas clone TB revealed very few differences in SNPs and indels. The PT22 strain had an increased number of accessory genes, mainly those that encode enzymes as compared to 491 and CHA. Strain CHA, the most virulent isolate of clone CHA had the smallest genome; smallest number of accessory genes (phage associated); and a deletion in a global regulator as compared to the other two strains. The differences which were observed in the study illustrate that habitat specific factors are the major driving forces behind the genetic makeup of most bacteria in support of their adaptation and enhanced lifestyles.

The detection of horizontally transferred genomic elements has increasingly become of importance in the fields of genomics and biomedical research as these possess important factors such as metabolic, virulence and antibiotic resistance genes. The acquisitions of such factors by bacteria have been reported to contribute towards many of the recurrent outbreaks and multidrug resistance; these thus pose a threat towards mankind. Bacteria do not necessarily acquire foreign elements to cause diseases or outbreaks, such are mainly important for survival; adaptation to a variety of environments; and interactions with hosts. These elements may allow bacteria to either cause damages to their hosts due to fluctuating environmental conditions or provide some benefits in the form of mutualism. Horizontally transferred genomic elements are highly regarded as important factors which may result in dramatic behavioural changes in bacteria. Several studies have indicated genomic features which are associated with horizontally acquired genomic elements. Computational tools have therefore been developed to search for genomic regions which resemble such features. As the detection of such factors is only a stepping stone in studying horizontal gene transfer events in bacteria, we felt the need to expand on the subject by offering compositional methods that aid in determining the distribution pathways of GIs in bacteria; their relative acquisition periods; and the relationships between host genomes and their similar constituent GIs. The idea to develop such methods was also influenced by the fact that many of the current GI prediction tools do not provide additional criteria to further analyse GIs and illustrate how these are interconnected. They only provide genomic coordinates of such entities and their annotations, and not any other additional information. SeqWord Genomic Island Sniffer (SWGIS); the LingvoCom utilities; together with several in-house customized scripts were

therefore developed in order to allow further analysis of prokaryotic GIs.

The SWGIS tool offers a platform to search for GIs in prokaryotic genomes using a wide range of parametric values which were optimized based on comparisons with the other publicly available prediction tools. The tool also provides a variety of output options, with SVG images created for individual genomes indicating their OU histograms and highlighted genomic island regions. The GIs predicted for a set of prokaryotic genomes are subsequently compared to one another by the use of an in-house custom GI-comparison script coupled with GraphViz functionalities in search for their shared compositions to create a graph based cluster. The idea of creating graph based clusters was to gain an insight about the ontological relationships of GIs which are harboured by different classes of bacteria. Ontological relationships coupled with the stratigraphical analysis allows for the determination of directional distributions of GIs. The stratigraphical analysis has also illustrated that sets of GIs in individual organisms were not necessarily acquired through single horizontal events, as these are depicted by variable gray colour gradients. The links created between ancient and recent inserts illustrate the movements of ancient GIs from older to newer naïve host organisms for beneficial purposes. Further analysis of such linked GIs indicated the possibilities to reveal the donor-recipient relationships between genomes and their constituent GIs.

The donor-recipient relationships of the GIs with shared compositional similarities were illustrated through the use of the LingvoCom utilities. LingvoCom compares GIs and a variety of host organisms for shared compositional similarities. Groups of GIs which are similar to one another form clusters around genomes with shared compositions and these are therefore predicted to be their source organisms. The LingvoCom utilities also provide functionalities for creating phylogenetic inferences between GIs in order to represent those which may have been acquired from common ancestors. The method also shows that GIs which are possessed by an individual organism form separate clusters and indicate that they may have been acquired from different sources, and not from a single donor.

As a test subject, the developed methods were successfully applied on *E. coli* TY-2482 to determine the evolutionary relationship of its GIs together with those of the other bacteria and their contributions towards the 2011 Germany outbreak. Through compositional analysis this strain indicated to share a high composition similarity with the EAEC 55989 *E. coli* strain,

which was isolated in Central Africa during an outbreak. These additionally share a high proportion of GIs. The intriguing thing is that these GIs were indicated by the stratigraphy method as ancient acquisitions, which indicates that the *E. coli* TY-2482 strain did not cause an outbreak as a result of recently acquired GIs. The analyses have also indicated that one of the TY-2482 GIs shares composition and sequence similarities with plasmids 55989p and pSD_88 of *E. coli* 55989 and *S. enterica* ssp. *enterica* Dublin, respectively. The GI was found to possess a range of heavy metal resistance genes which were indicated to have been initially acquired from marine γ-Proteobacteria through a series of donor-recipient relations analyses. The results obtained from the latter study have illustrated the practical value of composition-based approaches in the field of horizontal gene transfer. These have been demonstrated to offer more than just the prediction of coordinates which are associated with atypical genomic fragments. They allow for the determination of phylogenetic relatedness of GIs and the search for their source organisms. Such approaches have even gained popularity in metagenomic studies as they are used to assign short genomic fragments to their probable host genomes on the basis of their compositions. Comparative genomics studies have been argued to be the most favourable in terms of detecting GIs but do not have the necessary capabilities to find appropriate donors of such entities. The latter reflects the importance of composition methods; their vast applications; and an increased potential in genomic studies.

Bibliography

Abby, S. S., Tannier, E., Gouy, M., & Daubin, V. (2010). Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests. *BMC Bioinformatics , 11*, 324.

Abe, T., Kanaya, S., Kinouchi, M., Ichiba, Y., Kozuki, T., & Ikemura, T. (2003). Informatics for unveiling hidden genome signatures. *Genome Res , 13* (4), 693-702.

Acinas, S. G., Marcelino, L. A., Klepac-Ceraj, V., & Polz, M. F. (2004). Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. *Journal of Bacteriology , 186* (9), 2629-2635.

Agerso, Y., & Sandvang, D. (2005). Class 1 integrons and tetracycline resistance genes in *Alcaligenes*, *Arthrobacter*, and *Pseudomonas* spp. isolated from pigsties and manured soil. *Applied and environmental microbiology , 71* (12), 7941-7947.

Alonso, A., Sanchez, P., & Martinez, J. L. (2001). Environmental selection of antibiotic resistance genes. *Environmental Microbiology , 3* (1), 1-9.

Alsop, E. B., & Raymond, J. (2013). Resolving prokaryotic taxonomy without rRNA: Longer Oligonucleotide Word Lengths Improve Genome and Metagenome Taxonomic Classification. *Plos One*, *8*(7), e67337.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol , 215* (3), 403-410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res , 25* (17), 3389-3402.

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S. *et al.* (2004). UniProt: the universal protein knowledgebase. *Nucleic acids research , 32* (suppl 1), D115--D119.

Arevalo-Ferro, C., Buschmann, J., Reil, G., Gorg, A., Wiehlmann, L., Tümmler, B. *et al.* (2004). Proteome analysis of intraclonal diversity of two *Pseudomonas aeruginosa* TB clone isolates. *Proteomics , 4* (5), 1241-1246.

Auchtung, J. M., Lee, C. A., Monson, R. E., Lehman, A. P., & Grossman, A. D. (2005). Regulation of a *Bacillus subtilis* mobile genetic element by intercellular signaling and the global DNA damage response. *Proc Natl Acad Sci U S A , 102* (35), 12554-12559.

Ausubel, F., Brent, R., Kingston, R., Moore, D., Seidman, J., & Smith, J. i Struhl K (Eds). 1998. *Current protocols in molecular biology , 1*.

Azad, R. K. & Lawrence, J. G. (2011). Towards more robust methods of alien gene detection.. *Nucleic Acids Res, 39*(9), e56.

Aziz, R. K., Breitbart, M., & Edwards, R. A. (2010). Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Res , 38* (13), 4207-4217.

Baldi, P., & Baisnée, P. F. (2000). Sequence analysis by additive scales: DNA structure for sequences and repeats of all lengths. *Bioinformatics , 16* (10), 865-889.

Becq, J., Churlaud, C., & Deschavanne, P. (2010). A benchmark of parametric methods for horizontal transfers detection. *PLoS One , 5* (4), e9989.

Becq, J., Gutierrez, M. C., Rosas-Magallanes, V., Rauzier, J., Gicquel, B., Neyrolles, O. *et al.* (2007). Contribution of horizontally acquired genomic islands to the evolution of the tubercle bacilli. *Molecular biology and evolution , 24* (8), 1861-1871.

Beiko, R., Harlow, T., & Ragan, M. (2005). Highways of gene sharing in prokaryotes. *Proceedings of the National Academy of Sciences of the United States of America , 102* (40), 14332-14337.

Beiko, R. G. & Hamilton, N. (2006). Phylogenetic identification of lateral genetic transfer events.. *BMC Evol Biol,* Volume 6, 15.

Bennett, P. (2008). Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria. *British journal of pharmacology , 153* (S1), S347--S357.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Wheeler, D. L. (2003). GenBank. *Nucleic Acids Res , 31* (1), 23-27.

Betley, M. J., & Mekalanos, J. J. (1985). Staphylococcal enterotoxin A is encoded by phage.

Science , 229 (4709), 185-187.

Bezuidt, O., Pierneef, R., Lima-Mendez, G., Reva, O. N., & others. (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), e25702.

Bhowmick, P. P., Devegowda, D., Ruwandeepika, H. D., Karunasagar, I., & Karunasagar, I. (2011). Presence of *Salmonella* pathogenicity island 2 genes in seafood-associated *Salmonella* serovars and the role of the sseC gene in survival of *Salmonella enterica* serovar Weltevreden in epithelial cells. *Microbiology , 157* (1), 160-168.

Binnewies, T. T., Motro, Y., Hallin, P. F., Lund, O., Dunn, D., La, T. *et al.* (2006). Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. *Functional & integrative genomics , 6* (3), 165-185.

Blanco, M., Gutierrez-Martin, C. B., Rodriguez-Ferri, E. F., Roberts, M. C., & Navas, J. (2006). Distribution of tetracycline resistance genes in *Actinobacillus pleuropneumoniae* isolates from Spain. *Antimicrob Agents Chemother , 50* (2), 702-708.

Blanc-Potard, A.-B., & Groisman, E. A. (1997). The *Salmonella* selC locus contains a pathogenicity island mediating intramacrophage survival. *The EMBO journal , 16* (17), 5376-5385.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E. *et al.* (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res , 31* (1), 365-370.

Bohlin, J., Skjerve, E., & Ussery, D. W. (2008). Investigations of oligonucleotide usage variance within and between prokaryotes. *PLoS Comput Biol , 4* (4), e1000057.

Bohlin, J., van Passel, M. W., Snipen, L., Kristoffersen, A. B., Ussery, D., & Hardy, S. P. (2012). Relative entropy differences in bacterial chromosomes, plasmids, phages and genomic islands. *BMC genomics , 13* (1), 66.

Boto, L. (2010). Horizontal gene transfer in evolution: facts and challenges. *Proceedings of the Royal Society B: Biological Sciences , 277* (1683), 819-827.

Boucher, J., Yu, H., Mudd, M., & Deretic, V. (1997). Mucoid *Pseudomonas aeruginosa* in

cystic fibrosis: characterization of muc mutations in clinical isolates and analysis of clearance in a mouse model of respiratory infection. *Infection and immunity , 65* (9), 3838-3846.

Braid, M. D., Silhavy, J. L., Kitts, C. L., Cano, R. J., & Howe, M. M. (2004). Complete genomic sequence of bacteriophage B3, a Mu-like phage of *Pseudomonas aeruginosa*. *J Bacteriol , 186* (19), 6560-6574.

Brzuszkiewicz, E., Thurmer, A., Schuldes, J., Leimbach, A., Liesegang, H., Meyer, F.-D. *et al.* (2011). Genome sequence analyses of two isolates from the recent *Escherichia coli* outbreak in Germany reveal the emergence of a new pathotype: Entero-Aggregative-Haemorrhagic *Escherichia coli* (EAHEC). *Archives of microbiology , 193* (12), 883-891.

Bugarel, M., Martin, A., Fach, P., & Beutin, L. (2011). Virulence gene profiling of enterohemorrhagic (EHEC) and enteropathogenic (EPEC) *Escherichia coli* strains: a basis for molecular risk assessment of typical and atypical EPEC strains. *BMC microbiology , 11* (1), 142.

Canchaya, C., Proux, C., Fournous, G., Bruttin, A., & Brussow, H. (2003). Prophage genomics. *Microbiology and Molecular Biology Reviews , 67* (2), 238-276.

Carbone, A., Zinovyev, A., & Kepes, F. (2003). Codon adaptation index as a measure of dominating codon bias. *Bioinformatics , 19* (16), 2005-2015.

Casali, N., Nikolayevskyy, V., Balabanova, Y., Ignatyeva, O., Kontsevaya, I., Harris, S. R. *et al.* (2012). Microevolution of extensively drug-resistant tuberculosis in Russia. *Genome research , 22* (4), 735-745.

Castang, S., & Dove, S. L. (2010). High-order oligomerization is required for the function of the H-NS family member MvaT in *Pseudomonas aeruginosa*. *Molecular microbiology , 78* (4), 916-931.

Chang, Y.-S. T., Klockgether, J., & Tümmler, B. (2007). An intragenic deletion in pilQ leads to nonpiliation of a *Pseudomonas aeruginosa* strain isolated from cystic fibrosis lung. *FEMS microbiology letters , 270* (2), 201-206.

Chapus, C., Dufraigne, C., Edwards, S., Giron, A., Fertil, B., & Deschavanne, P. (2005). Exploration of phylogenetic data using a global sequence analysis method. *BMC evolutionary*

biology , 5 (1), 63.

Charkowski, A. O. (2004). Making sense of an alphabet soup: the use of a new bioinformatics tool for identification of novel gene islands. Focus on "identification of genomic islands in the genome of *Bacillus cereus* by comparative analysis with *Bacillus anthracis*". *Physiol Genomics , 16* (2), 180-181.

Chatterjee, R., Chaudhuri, K., & Chaudhuri, P. (2008). On detection and assessment of statistical significance of Genomic Islands. *BMC Genomics , 9*, 150.

Cheung, M. K., Li, L., Nong, W., & Kwan, H. S. (2011). 2011 German *Escherichia coli* O104:H4 outbreak: whole-genome phylogeny without alignment. *BMC Res Notes , 4*, 533.

Choi, I.-G., & Kim, S.-H. (2007). Global extent of horizontal gene transfer. *Proc Natl Acad Sci U S A , 104* (11), 4489-4494.

Chung, J. C., Becq, J., Fraser, L., Schulz-Trieglaff, O., Bond, N. J., Foweraker, J. *et al.* (2012). Genomic Variation among Contemporary *Pseudomonas aeruginosa* Isolates from Chronically Infected Cystic Fibrosis Patients. *Journal of bacteriology , 194* (18), 4857-4866.

Cingolani, P., Platts, A., Coon, M., Nguyen, T., Wang, L., Land, S. J. *et al.* (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly , 6* (2), 80-92.

Cirillo, D. M., Valdivia, R. H., Monack, D. M., & Falkow, S. (1998). Macrophage-dependent induction of the *Salmonella* pathogenicity island 2 type III secretion system and its role in intracellular survival. *Molecular microbiology , 30* (1), 175-188.

Clermont, O., Bonacorsi, S., & Bingen, E. (2000). Rapid and Simple Determination of the *Escherichia coli* Phylogenetic Group. *Applied and Environmental Microbiology , 66* (10), 4555-4558.

Coenye, T., & Vandamme, P. (2004). Use of the genomic signature in bacterial classification and identification. *Systematic and applied microbiology , 27* (2), 175-185.

Conte, L. L., Brenner, S. E., Hubbard, T. J., Chothia, C., & Murzin, A. G. (2002). SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res , 30* (1), 264-267.

Cowles, K. N., & Gitai, Z. (2010). Surface association and the MreB cytoskeleton regulate pilus production, localization and function in *Pseudomonas aeruginosa*. *Molecular microbiology , 76* (6), 1411-1426.

Cramer, N., Klockgether, J., Wrasman, K., Schmidt, M., Davenport, C. F., & Tümmler, B. (2011). Microevolution of the major common *Pseudomonas aeruginosa* clones C and PA14 in cystic fibrosis lungs. *Environmental Microbiology , 13* (7), 1690-1704.

Dacheux, D., Attree, I., Schneider, C., & Toussaint, B. (1999). Cell death of human polymorphonuclear neutrophils induced by a *Pseudomonas aeruginosa* cystic fibrosis isolate requires a functional type III secretion system. *Infection and immunity , 67* (11), 6164-6167.

Dacheux, D., Goure, J., Chabert, J., Usson, Y., & Attree, I. (2001). Pore-forming activity of type III system-secreted proteins leads to oncosis of *Pseudomonas aeruginosa*-infected macrophages. *Molecular microbiology , 40* (1), 76-85.

Dacheux, D., Toussaint, B., Richard, M., Brochier, G., Croize, J., & Attree, I. (2000). *Pseudomonas aeruginosa* cystic fibrosis isolates induce rapid, type III secretion-dependent, but ExoU-independent, oncosis of macrophages and polymorphonuclear neutrophils. *Infection and immunity , 68* (5), 2916-2924.

Dale, C., Young, S. A., Haydon, D. T., & Welburn, S. C. (2001). The insect endosymbiont *Sodalis glossinidius* utilizes a type III secretion system for cell invasion. *Proceedings of the National Academy of Sciences , 98* (4), 1883-1888.

Damron, F. H., & Goldberg, J. B. (2012). Proteolytic regulation of alginate overproduction in *Pseudomonas aeruginosa*. *Molecular microbiology , 84* (4), 595-607.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A. *et al.* (2011). The variant call format and VCFtools. *Bioinformatics , 27* (15), 2156-2158.

Darling, A. C., Mau, B., Blattner, F. R., & Perna, N. T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research , 14* (7), 1394-1403.

Daubin, V., Lerat, E., & Perrière, G. (2003). The source of laterally transferred genes in bacterial genomes. *Genome Biol , 4* (9), R57.

Davids, W., & Zhang, Z. (2008). The impact of horizontal gene transfer in shaping operons

and protein interaction networks--direct evidence of preferential attachment. *BMC Evol Biol , 8*, 23.

Dayhoff, M. (1978). Atlas of Protein Sequence and Structure. National Biomedical Research Foundation. *Washington, DC* .

Dayhoff, M. O. (1973). *Atlas of protein sequence and structure* (Vol. 5). National Biomedical Research Foundation.

de Bentzmann, S., & Plesiat, P. (2011). The *Pseudomonas aeruginosa* opportunistic pathogen and human infections. *Environmental microbiology , 13* (7), 1655-1665.

Deschavanne, P., Giron, A., Fagot, J. V., & Fertil, B. (2000). Genomic Signature is Preserved in Short DNA Fragments. *IEEE International Symposium on Bio-Informatics and Biomedical Engineering (BIBE'00)* (6), 161.

Deschavanne, P., Giron, A., Vilain, J., Fagot, G., & Fertil, B. (1999). Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Molecular Biology and Evolution , 16* (10), 1391-1399.

Diels, L., Van Roy, S., Taghavi, S., & Van Houdt, R. (2009). From industrial sites to environmental applications with *Cupriavidus metallidurans*. *Antonie Van Leeuwenhoek , 96* (2), 247-258.

Dobrindt, U., Hochhut, B., Hentschel, U., & Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms. *Nat Rev Microbiol , 2* (5), 414-424.

Dobrindt, U., Hochhut, B., Hentschel, U., & Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms. *Nature Reviews Microbiology , 2* (5), 414-424.

Droge, M., Puhler, A., & Selbitschka, W. (1999). Horizontal gene transfer among bacteria in terrestrial and aquatic habitats as assessed by microcosm and field studies. *Biology and Fertility of Soils , 29* (3), 221-245.

Dufraigne, C., Fertil, B., Lespinats, S., Giron, A., & Deschavanne, P. (2005). Detection and characterization of horizontal transfers in prokaryotes using genomic signature. *Nucleic Acids Research , 33* (1), e6--e6.

Durante-Mangoni, E., & Zarrilli, R. (2011). Global spread of drug-resistant *Acinetobacter baumannii*: molecular epidemiology and management of antimicrobial resistance. *Future microbiology , 6* (4), 407-422.

Durinck, S., Bullard, J., Spellman, P. T., & Dudoit, S. (2009). GenomeGraphs: integrated genomic data visualization with R. *BMC bioinformatics , 10* (1), 2.

Dutta, C., & Pan, A. (2002, Feb). Horizontal gene transfer and bacterial diversity. *Horizontal gene transfer and bacterial diversity. , 27 (1 Suppl 1)* , 27-33.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G. *et al.* (2009). Real-time DNA sequencing from single polymerase molecules. *Science , 323* (5910), 133-138.

Ejrnæs, K. (2011). Bacterial characteristics of importance for recurrent urinary tract infections caused by *Escherichia coli*. *Dan Med Bull , 58* (4), B4187.

Elsas, J. D., & Bailey, M. J. (2002). The ecology of transfer of mobile genetic elements. *FEMS microbiology ecology , 42* (2), 187-197.

Enright, A. J., Dongen, S. V., & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res , 30* (7), 1575-1584.

Ermolaeva, M. D. (2001). Synonymous codon usage in bacteria. *Curr Issues Mol Biol , 3* (4), 91-97.

Escobar-Paramo, P., Clermont, O., Blanc-Potard, A.-B., Bui, H., Le Bouguenec, C., & Denamur, E. (2004). A specific genetic background is required for acquisition and expression of virulence factors in *Escherichia coli*. *Molecular biology and evolution , 21* (6), 1085-1094.

Eswarappa, S. M., Janice, J., Nagarajan, A. G., Balasundaram, S. V., Karnam, G., Dixit, N. M. *et al.* (2008). Differentially evolved genes of *Salmonella* pathogenicity islands: insights into the mechanism of host specificity in *Salmonella*. *PLOS one , 3* (12), e3829.

Falush, D. (2009). Toward the use of genomics to study microevolutionary change in bacteria. *PLoS genetics , 5* (10), e1000627.

Fauvarque, M.-O., Bergeret, E., Chabert, J., Dacheux, D., Satre, M., & Attree, I. (2002). Role and activation of type III secretion system genes in *Pseudomonas aeruginosa*-induced

Drosophila killing. *Microbial pathogenesis , 32* (6), 287-295.

Fedurco, M., Romieu, A., Williams, S., Lawrence, I., & Turcatti, G. (2006). BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res , 34* (3), e22.

Ferens, W. A., & Hovde, C. J. (2011). *Escherichia coli* O157: H7: animal reservoir and sources of human infection. *Foodborne pathogens and disease , 8* (4), 465-487.

Fernández-Gómez, B., Fernández-Guerra, A., Casamayor, E. O., González, J. M., Pedrós-Alió, C., & Acinas, S. G. (2012). Patterns and architecture of genomic islands in marine bacteria. *BMC genomics , 13* (1), 347.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R. *et al.* (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science , 269* (5223), 496-512.

Fleiss, J. L., Levin, B., & Paik, M. C. (2013). *Statistical methods for rates and proportions.* John Wiley \& Sons.

Fluit, A. C. (2005). Towards more virulent and antibiotic-resistant *Salmonella*? *FEMS Immunology & Medical Microbiology , 43* (1), 1-11.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D. *et al.* (1995). The minimal gene complement of Mycoplasma genitalium. *Science , 270* (5235), 397-403.

Frech, G., & Schwarz, S. (1998). Tetracycline resistance in *Salmonella enterica* subsp. *enterica* serovar Dublin. *Antimicrob Agents Chemother , 42* (5), 1288-1289.

Frost, L. S., Leplae, R., Summers, A. O., & Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology , 3* (9), 722-732.

Gomez-Lozano, M., Marvig, R. L., Molin, S., & Long, K. S. (2012). Genome-wide identification of novel small RNAs *in Pseudomonas aeruginosa*. *Environmental Microbiology , 14* (8), 2006-2016.

Ganesan, H., Rakitianskaia, A., Davenport, C., Tümmler, B., & Reva, O. (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* (1), 333.

Garcia-Vallve, S., Palau, J., & Romeu, A. (1999). Horizontal gene transfer in glycosyl hydrolases inferred from codon usage in *Escherichia coli* and *Bacillus subtilis*. *Molecular biology and evolution , 16* (9), 1125-1134.

Garcia-Vallve, S., Guzman, E., Montero, M. A., & Romeu, A. (2003). HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Res , 31* (1), 187-189.

Garénaux, A., Caza, M., & Dozois, C. M. (2011). The Ins and Outs of siderophore mediated iron uptake by extra-intestinal pathogenic *Escherichia coli*. *Vet Microbiol , 153* (1-2), 89-98.

George, A. M., Jones, P. M., & Middleton, P. G. (2009). Cystic fibrosis infections: treatment strategies and prospects. *FEMS microbiology letters , 300* (2), 153-164.

Gerlach, R. G., Jackel, D., Stecher, B., Wagner, C., Lupas, A., Hardt, W.-D. *et al.* (2007). *Salmonella* Pathogenicity Island 4 encodes a giant non-fimbrial adhesin and the cognate type 1 secretion system. *Cellular microbiology , 9* (7), 1834-1850.

Gold, B. et al., (2008). Identification of a copper-binding metallothionein in pathogenic mycobacteria.. *Nat Chem Biol, 4*(10), 609-616.

Goodman, A. L., Kulasekara, B., Rietsch, A., Boyd, D., Smith, R. S., & Lory, S. (2004). A Signaling Network Reciprocally Regulates Genes Associated with Acute Infection and Chronic Persistence in *Pseudomonas aeruginosa*. *Developmental cell , 7* (5), 745-754.

Goodman, A. L., Merighi, M., Hyodo, M., Ventre, I., Filloux, A., & Lory, S. (2009). Direct interaction between sensor kinase proteins mediates acute and chronic disease phenotypes in a bacterial pathogen. *Genes & development , 23* (2), 249-259.

Gordon, D. M., Clermont, O., Tolley, H., & Denamur, E. (2008). Assigning *Escherichia coli* strains to phylogenetic groups: multi-locus sequence typing versus the PCR triplex method. *Environmental microbiology , 10* (10), 2484-2496.

Goshorn, S. C., & Schlievert, P. M. (1989). Bacteriophage association of streptococcal pyrogenic exotoxin type C. *J Bacteriol , 171* (6), 3068-3073.

Gürtler, V., & Mayall, B. C. (2001). Genomic approaches to typing, taxonomy and evolution of bacterial isolates. *International journal of systematic and evolutionary microbiology , 51* (1), 3-16.

Guyer, D. M., Radulovic, S., Jones, F.-E., & Mobley, H. L. (2002). Sat, the secreted autotransporter toxin of uropathogenic *Escherichia coli*, is a vacuolating cytotoxin for bladder and kidney epithelial cells. *Infection and immunity , 70* (8), 4539-4546.

Hacker, J., & Carniel, E. (2001). Ecological fitness, genomic islands and bacterial pathogenicity. *EMBO reports , 2* (5), 376-381.

Hacker, J., & Kaper, J. B. (2000). Pathogenicity islands and the evolution of microbes. *Annu Rev Microbiol , 54*, 641-679.

Hacker, J., Bender, L., Ott, M., Wingender, J., Lund, B., Marre, R. *et al.* (1990). Deletions of chromosomal regions coding for fimbriae and hemolysins occur in vitro and in vivo in various extra intestinal *Escherichia coli* isolates. *Microbial pathogenesis , 8* (3), 213-225.

Hacker, J., Hentschel, U., & Dobrindt, U. (2003). Prokaryotic chromosomes and disease. *Science , 301* (5634), 790-793.

Hakkila, K. M. et al., (2011). Cd-specific mutants of mercury-sensing regulatory protein MerR, generated by directed evolution.. *Appl Environ Microbiol, 77*(17), 6215-6224.

Hartman, A. B., Essiet, I. I., Isenbarger, D. W., & Lindler, L. E. (2003). Epidemiology of tetracycline resistance determinants in *Shigella* spp. and enteroinvasive *Escherichia coli*: characterization and dissemination of tet(A)-1. *J Clin Microbiol , 41* (3), 1023-1032.

Hasan, M. S., Liu, Q., Wang, H., Fazekas, J., Chen, B., & Che, D. (2012). GIST: Genomic island suite of tools for predicting genomic islands in genomic sequences. *Bioinformation , 8* (4), 203.

Herzer, P. J., Inouye, S., Inouye, M., & Whittam, T. S. (1990). Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of *Escherichia coli*. *Journal of bacteriology , 172* (11), 6175-6181.

Hill, C. W., Sandt, C. H., & Vlazny, D. A. (1994). Rhs elements of *Escherichia coli*: a family of genetic composites each encoding a large mosaic protein. *Molecular microbiology , 12* (6),

865-871.

Hoboth, C., Hoffmann, R., Eichner, A., Henke, C., Schmoldt, S., Imhof, A. *et al.* (2009). Dynamics of adaptive microevolution of hypermutable *Pseudomonas aeruginosa* during chronic pulmonary infection in patients with cystic fibrosis. *Journal of Infectious Diseases , 200* (1), 118-130.

Hood, R. D., Singh, P., Hsu, F., Guvener, T., Carl, M. A., Trinidad, R. R. *et al.* (2010). A Type VI Secretion System of *Pseudomonas aeruginosa* Targets a Toxin to Bacteria. *Cell host & microbe , 7* (1), 25-37.

Hsiao, W. W., Ung, K., Aeschliman, D., Bryan, J., Finlay, B. B., & Brinkman, F. S. (2005). Evidence of a large novel gene pool associated with prokaryotic genomic islands. *PLoS genetics , 1* (5), e62.

Hsiao, W., Wan, I., Jones, S. J., & Brinkman, F. S. (2003). IslandPath: aiding detection of genomic islands in prokaryotes. *Bioinformatics , 19* (3), 418-420.

Huson, D. H., & Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular biology and evolution , 23* (2), 254-267.

Jackson, A., Thomas, G., Parkhill, J., & Thomson, N. (2009). Evolutionary diversification of an ancient gene family (rhs) through C-terminal displacement. *BMC genomics , 10* (1), 584.

Jain, R., Rivera, M. C., & Lake, J. A. (1999). Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A , 96* (7), 3801-3806.

Jain, R., Rivera, M. C., Moore, J. E., & Lake, J. A. (2003). Horizontal gene transfer accelerates genome innovation and evolution. *Molecular Biology and Evolution , 20* (10), 1598-1602.

Jain, R., Rivera, M. C., Moore, J. E., & Lake, J. A. (2002). Horizontal gene transfer in microbial genome evolution. *Theoretical population biology , 61* (4), 489-495.

Janssen, P. J. et al., (2010). The complete genome sequence of Cupriavidus metallidurans strain CH34, a master survivalist in harsh and anthropogenic environments.. *PLoS One, 5*(5), e10433.

Jeffrey, H. J. (1990). Chaos game representation of gene structure. *Nucleic Acids Res , 18* (8), 2163-2170.

Johnson, L. P., Tomai, M. A., & Schlievert, P. M. (1986). Bacteriophage involvement in group A streptococcal pyrogenic exotoxin A production. *J Bacteriol , 166* (2), 623-627.

Josse, J., Kaiser, A., & Kornberg, A. (1961). Enzymatic synthesis of deoxyribonucleic acid. *J Biol Chem , 236*, 864-875.

Juhas, M., Der Meer, V., Roelof, J., Gaillard, M., Harding, R. M., Hood, D. W. *et al.* (2009). Genomic islands: tools of bacterial horizontal gene transfer and evolution. *FEMS microbiology reviews , 33* (2), 376-393.

Karasova, D., Sebkova, A., Havlickova, H., Sisak, F., Volf, J., Faldyna, M. *et al.* (2010). Influence of 5 major *Salmonella* pathogenicity islands on NK cell depletion in mice infected with *Salmonella enterica* serovar Enteritidis. *BMC microbiology , 10* (1), 75.

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

Karlin, S. (1998). Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr Opin Microbiol , 1* (5), 598-610.

Karlin, S. S., & Burge, C. (1995). Dinucleotide relative abundance extremes: a genomic signature. *Trends in genetics , 11* (7), 283-290.

Karlin, S., & Burge, C. (1995). Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet , 11* (7), 283-290.

Karlin, S., Ladunga, I., & Blaisdell, B. E. (1994). Heterogeneity of genomes: measures and values. *Proc Natl Acad Sci U S A , 91* (26), 12837-12841.

Karlin, S., Mrazek, J., & Campbell, A. (1997). Compositional biases of bacterial genomes and evolutionary implications. *Journal of Bacteriology , 179* (12), 3899-3913.

Kelly, B., Vespermann, A., & Bolton, D. (2009). The role of horizontal gene transfer in the evolution of selected foodborne bacterial pathogens. *Food and Chemical Toxicology , 47* (5), 951-968.

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. *Genome Res , 12* (4), 656-664.

Kiewitz, C., Larbig, K., Klockgether, J., Weinel, C., & Tümmler, B. (2000). Monitoring genome evolution ex vivo: reversible chromosomal integration of a 106 kb plasmid at two tRNALys gene loci in sequential *Pseudomonas aeruginosa* airway isolates. *Microbiology , 146* (10), 2365-2373.

Kisiela, D. I., Chattopadhyay, S., Libby, S. J., Karlinsey, J. E., Fang, F. C., Tchesnokova, V. *et al.* (2012). Evolution of *Salmonella enterica* virulence via point mutations in the fimbrial adhesin. *PLoS pathogens , 8* (6), e1002733.

Klockgether, J., Cramer, N., Wiehlmann, L., Davenport, C. F., & Tümmler, B. (2011). *Pseudomonas aeruginosa* genomic structure and diversity. *Frontiers in microbiology , 2*.

Klockgether, J., Miethke, N., Kubesch, P., Bohn, Y.-S., Brockhausen, I., Cramer, N. *et al.* (2013). Intraclonal diversity of the *Pseudomonas aeruginosa* cystic fibrosis airway isolates TBCF10839 and TBCF121838: distinct signatures of transcriptome, proteome, metabolome, adherence and pathogenicity despite an almost identical genome sequence. *Environmental microbiology , 15* (1), 191-210.

Klockgether, J., Reva, O., Larbig, K., & Tümmler, B. (2004). Sequence analysis of the mobile genome island pKLC102 of *Pseudomonas aeruginosa* C. *Journal of bacteriology , 186* (2), 518-534.

Klockgether, J., Wuerdemann, D., Reva, O., Wiehlmann, L., & Tümmler, B. (2007). Diversity of the abundant pKLC102/PAGI-2 family of genomic islands in *Pseudomonas aeruginosa*. *Journal of bacteriology , 189* (6), 2443-2459.

Klockgether, J., Würdemann, D., Reva, O., Wiehlmann, L., & Tümmler, B. (2007). Diversity of the abundant pKLC102/PAGI-2 family of genomic islands in *Pseudomonas aeruginosa*. *J Bacteriol , 189* (6), 2443-2459.

Knodler, L. A., Celli, J., Hardt, W.-D., Vallance, B. A., Yip, C., & Finlay, B. B. (2002). *Salmonella* effectors within a single pathogenicity island are differentially expressed and translocated by separate type III secretion systems. *Molecular microbiology , 43* (5), 1089-1103.

Koski, L. B., Morton, R. A., & Golding, G. B. (2001). Codon bias and base composition are poor indicators of horizontally transferred genes. *Molecular Biology and Evolution , 18* (3), 404-412.

Kresse, A. U., Dinesh, S. D., Larbig, K., & Romling, U. (2003). Impact of large chromosomal inversions on the adaptation and evolution of *Pseudomonas aeruginosa* chronically colonizing cystic fibrosis lungs. *Molecular microbiology , 47* (1), 145-158.

Kyrpides, N. C. (1999). Genomes OnLine Database (GOLD 1.0): a monitor of complete and ongoing genome projects world-wide. *Bioinformatics , 15* (9), 773-774.

Kyrpides, N. C., & Olsen, G. J. (1999). Archaeal and bacterial hyperthermophiles: horizontal gene exchange or common ancestry? *Trends in Genetics , 15* (8), 298-299.

Langille, M. G., Hsiao, W. W., & Brinkman, F. S. (2008). Evaluation of genomic island predictors using a comparative genomics approach. *BMC bioinformatics , 9* (1), 329.

Langille, M., & Brinkman, F. (2009). IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics , 25* (5), 664-665.

Larbig, K. D., Christmann, A., Johann, A., Klockgether, J., Hartsch, T., Merkl, R. *et al.* (2002). Gene islands integrated into tRNA(Gly) genes confer genome diversity on a *Pseudomonas aeruginosa* clone. *J Bacteriol , 184* (23), 6665-6680.

Lawrence, J. G. (1999). Gene transfer, speciation, and the evolution of bacterial genomes. *Curr Opin Microbiol , 2* (5), 519-523.

Lawrence, J. G., & Ochman, H. (1997). Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol , 44* (4), 383-397.

Lawrence, J. G., & Ochman, H. (2002). Reconciling the many faces of lateral gene transfer. *Trends Microbiol , 10* (1), 1-4.

Lawrence, J. G., & Ochman, H. (1998). Molecular archaeology of the *Escherichia coli* genome. *Proceedings of the National Academy of Sciences , 95* (16), 9413-9417.

Lederberg, J., & Tatum, E. L. (1946). Gene recombination in *Escherichia coli*. *Nature , 158*, 558.

Leech, A. J., Sprinkle, A., Wood, L., Wozniak, D. J., & Ohman, D. E. (2008). The NtrC family regulator AlgB, which controls alginate biosynthesis in mucoid *Pseudomonas aeruginosa*, binds directly to the algD promoter. *Journal of bacteriology , 190* (2), 581-589.

Legendre, P., & Legendre, L. (2012). *Numerical ecology* (Vol. 20). Elsevier.

Legendre, P., & Legendre, L. (1983). Ordination in reduced space. *Developments in Environmental Modelling* , 387-480.

Leplae, R., Hebrant, A., Wodak, S. J., & Toussaint, A. (2004). ACLAME: a CLAssification of Mobile genetic Elements. *Nucleic Acids Res , 32* (Database issue), D45--D49.

Levings, R. S., Partridge, S. R., Djordjevic, S. P., & Hall, R. M. (2007). SGI1-K, a variant of the SGI1 genomic island carrying a mercury resistance region, in *Salmonella enterica* serovar Kentucky. *Antimicrobial agents and chemotherapy , 51* (1), 317-323.

Li, C., Wally, H., Miller, S. J., & Lu, C.-D. (2009). The multifaceted proteins MvaT and MvaU, members of the H-NS family, control arginine metabolism, pyocyanin synthesis, and prophage activation in *Pseudomonas aeruginosa* PAO1. *Journal of bacteriology , 191* (20), 6211-6218.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N. *et al.* (2009). The sequence alignment/map format and SAMtools. *Bioinformatics , 25* (16), 2078-2079.

Li, H., Luo, Y.-F., Williams, B. J., Blackwell, T. S., & Xie, C.-M. (2012). Structure and function of OprD protein in *Pseudomonas aeruginosa*: From antibiotic resistance to novel therapies. *International Journal of Medical Microbiology , 302* (2), 63-68.

Li, W., & Lu, C.-D. (2007). Regulation of carbon and nitrogen utilization by CbrAB and NtrBC two-component systems in *Pseudomonas aeruginosa*. *Journal of bacteriology , 189* (15), 5413-5420.

Lima-Mendez, G., Van Helden, J., Toussaint, A., & Leplae, R. (2008). Reticulate representation of evolutionary and functional relationships between phage genomes. *Molecular biology and evolution , 25* (4), 762-777.

Liu, B., Knirel, Y. A., Feng, L., Perepelov, A. V., Senchenkova, S. N., Wang, Q. *et al.* (2008). Structure and genetics of *Shigella* O antigens. *FEMS microbiology reviews , 32* (4), 627-653.

Loman, N. J., Constantinidou, C., Chan, J. Z., Halachev, M., Sergeant, M., Penn, C. W. *et al.* (2012). High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nature Reviews Microbiology* .

Lorenz, M. G., & Sikorski, J. (2000). The potential for intraspecific horizontal gene exchange by natural genetic transformation: sexual isolation among genomovars of *Pseudomonas stutzeri*. *Microbiology , 146* (12), 3081-3090.

Lukáová, M. B. (2008). Role of structural variations of polysaccharide antigens in the pathogenicity of Gram-negative bacteria. *Clinical microbiology and infection , 14* (3), 200-206.

MacLean, A. M., Finan, T. M., & Sadowsky, M. J. (2007). Genomes of the symbiotic nitrogen-fixing bacteria of legumes. *Plant physiology , 144* (2), 615-622.

Mainil, J., & Daube, G. (2005). Verotoxigenic *Escherichia coli* from animals, humans and foods: who's who? *Journal of Applied Microbiology , 98* (6), 1332-1344.

Mann, E. E., & Wozniak, D. J. (2012). *Pseudomonas* biofilm matrix composition and niche biology. *FEMS microbiology reviews , 36* (4), 893-916.

Manrique, M., Pareja-Tobes, P., Pareja-Tobes, E., Pareja, E., & Tobes, R. (2011). *Escherichia coli* EHEC Germany outbreak preliminary functional annotation using BG7 system.

Mantri, Y., & Williams, K. P. (2004). Islander: a database of integrative islands in prokaryotic genomes, the associated integrases and their DNA site specificities. *Nucleic Acids Res , 32* (Database issue), D55--D58.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A. *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature , 437* (7057), 376-380.

Marklein, G., Waschkowski, G., Reichertz, C., & others. (1981). [Septicaemia caused by *Erwinia herbicola* in an 8-year-old boy (author's transl)]. *Klinische Padiatrie , 193* (5), 394.

Marri, P. R., & Golding, G. B. (2008). Gene amelioration demonstrated: the journey of nascent genes in bacteria. *Genome , 51* (2), 164-168.

Mathee, K., Narasimhan, G., Valdes, C., Qiu, X., Matewish, J. M., Koehrsen, M. *et al.* (2008). Dynamics of *Pseudomonas aeruginosa* genome evolution. *Proceedings of the National Academy of Sciences , 105* (8), 3100-3105.

McDaniel, T. K., & Kaper, J. B. (1997). A cloned pathogenicity island from enteropathogenic *Escherichia coli* confers the attaching and effacing phenotype on *E. coli* K-12. *Molecular microbiology , 23* (2), 399-407.

Mellmann, A., Harmsen, D., Cummings, C. A., Zentz, E. B., Leopold, S. R., Rico, A. *et al.* (2011). Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104: H4 outbreak by rapid next generation sequencing technology. *PloS one , 6* (7), e22751.

Ménigaud, S., Mallet, L., Picord, G., Churlaud, C., Borrel, A., & Deschavanne, P. (2012). GOHTAM: a website for Genomic Origin of Horizontal Transfers, Alignment and Metagenomics. *Bioinformatics , 28* (9), 1270-1271.

Metzker, M. L. (2005). Emerging technologies in DNA sequencing. *Genome research , 15* (12), 1767-1776.

Mills, D. M., Bajaj, V., & Lee, C. A. (1995). A 40 kb chromosomal fragment encoding Salmonella typhimurium invasion genes is absent from the corresponding region of the *Escherichia coli* K-12 chromosome. *Molecular microbiology , 15* (4), 749-759.

Morschhauser, J., Kohler, G., Ziebuhr, W., Blum-Oehler, G., Dobrindt, U., & Hacker, J. (2000). Evolution of microbial pathogens. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences , 355* (1397), 695-704.

Mossoro, C., Glaziou, P., Yassibanda, S., Lan, N. T., Bekondi, C., Minssart, P. *et al.* (2002). Chronic diarrhea, hemorrhagic colitis, and hemolytic-uremic syndrome associated with HEp-2 adherent *Escherichia coli* in adults infected with human immunodeficiency virus in Bangui, Central African Republic. *Journal of clinical microbiology , 40* (8), 3086-3088.

Mrazek, J., & Karlin, S. (1999). Detecting Alien Genes in Bacterial Genomesa. *Annals of the New York Academy of Sciences , 870* (1), 314-329.

Muller, D., Simeonova, D. D., Riegel, P., Mangenot, S., Koechler, S., Lievremont, D. *et al.*

(2006). *Herminiimonas arsenicoxydans* sp. nov., a metalloresistant bacterium. *International journal of systematic and evolutionary microbiology , 56* (8), 1765-1769.

Nakamura, Y., Gojobori, T., & Ikemura, T. (1997). Codon usage tabulated from the international DNA sequence databases. *Nucleic acids research , 25* (1), 244-245.

Nakamura, Y., Gojobori, T., & Ikemura, T. (1999). Codon usage tabulated from the international DNA sequence databases; its status 1999. *Nucleic acids research , 27* (1), 292-292.

Nakamura, Y., Itoh, T., Matsuda, H., & Gojobori, T. (2004). Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nature genetics , 36* (7), 760-766.

Nakayama, K., Kanaya, S., Ohnishi, M., Terawaki, Y., & Hayashi, T. (1999). The complete nucleotide sequence of phi CTX, a cytotoxin-converting phage of *Pseudomonas aeruginosa*: implications for phage evolution and horizontal gene transfer via bacteriophages. *Mol Microbiol , 31* (2), 399-419.

Nishijyo, T., Haas, D., & Itoh, Y. (2001). The CbrA--CbrB two-component regulatory system controls the utilization of multiple carbon and nitrogen sources in *Pseudomonas aeruginosa*. *Molecular microbiology , 40* (4), 917-931.

Nogueira, T., Rankin, D. J., Touchon, M., Taddei, F., Brown, S. P., & Rocha, E. P. (2009). Horizontal gene transfer of the secretome drives the evolution of bacterial cooperation and virulence. *Current Biology , 19* (20), 1683-1691.

Norrby-Teglund, A., Holm, S. E., & Norgren, M. (1994). Detection and nucleotide sequence analysis of the speC gene in Swedish clinical group A streptococcal isolates. *Journal of clinical microbiology , 32* (3), 705-709.

O'Brien, A. D., Marques, L. R., Kerry, C. F., Newland, J. W., & Holmes, R. K. (1989). Shiga-like toxin converting phage of enterohemorrhagic *Escherichia coli* strain 933. *Microb Pathog , 6* (5), 381-390.

O'Brien, T., Ross, D., Guzman, M., Medeiros, A., Hedges, R., & Botstein, D. (1980). Dissemination of an antibiotic resistance plasmid in hospital patient flora. *Antimicrobial agents and chemotherapy , 17* (4), 537-543.

Ochman, H. (2001). Lateral and oblique gene transfer. *Curr Opin Genet Dev , 11* (6), 616-619.

Ochman, H., Lawrence, J., Groisman, E., & others. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature , 405* (6784), 299-304.

Oelschlaeger, T., Dobrindt, U., & Hacker, J. (2002). Pathogenicity islands of uropathogenic E. coli and the evolution of virulence. *International journal of antimicrobial agents , 19* (6), 517-521.

Ogura, Y., Ooka, T., Iguchi, A., Toh, H., Asadulghani, M., Oshima, K. *et al.* (2009). Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic *Escherichia coli*. *Proceedings of the National Academy of Sciences , 106* (42), 17939-17944.

Okuda, J., & Nishibuchi, M. (1998). Manifestation of the Kanagawa phenomenon, the virulence-associated phenotype, of *Vibrio parahaemolyticus* depends on a particular single base change in the promoter of the thermostable direct haemolysin gene. *Molecular microbiology , 30* (3), 499-511.

Osman, D. & Cavet, J. S., (2011). Metal sensing in Salmonella: implications for pathogenesis.. *Adv Microb Physiol,* Volume 58, 175-232.

Ou, H.-Y., Chen, L.-L., Lonnen, J., Chaudhuri, R. R., Thani, A. B., Smith, R. *et al.* (2006). A novel strategy for the identification of genomic islands by comparative analysis of the contents and contexts of tRNA sites in closely related bacteria. *Nucleic acids research , 34* (1), e3--e3.

Pál, C., Papp, B., & Lercher, M. J. (2005). Horizontal gene transfer depends on gene content of the host. *Bioinformatics , 21* (suppl 2), ii222--ii223.

Pallen, M. J., & Wren, B. W. (2007). Bacterial pathogenomics. *Nature , 449* (7164), 835-842.

Palleroni, N. J. (2003). Prokaryote taxonomy of the 20th century and the impact of studies on the genus *Pseudomonas*: a personal view. *Microbiology , 149* (1), 1-7.

Pan, A., Chanda, I. & Chakrabarti, J. (2011). Analysis of the genome and proteome composition of Bdellovibrio bacteriovorus: indication for recent prey-derived horizontal gene

transfer.. *Genomics, 98*(3), 213-222.

Park, C. & Zhang, J. (2012). High expression hampers horizontal gene transfer.. *Genome Biol Evol, 4*(4), 523-532.

Percival, S. S. (1998). Copper and immunity.. *Am J Clin Nutr, 67*(5), 1064S--1068S.

Petrova, O. E., Schurr, J. R., Schurr, M. J., & Sauer, K. (2012). Microcolony formation by the opportunistic pathogen *Pseudomonas aeruginosa* requires pyruvate and pyruvate fermentation. *Molecular microbiology , 86* (4), 819-835.

Pezzella, C., Ricci, A., DiGiannatale, E., Luzzi, I., & Carattoli, A. (2004). Tetracycline and streptomycin resistance genes, transposons, and plasmids in *Salmonella enterica* isolates from animals in Italy. *Antimicrob Agents Chemother , 48* (3), 903-908.

Philippe, H., & Douady, C. J. (2003). Horizontal gene transfer and phylogenetics. *Curr Opin Microbiol , 6* (5), 498-505.

Plainvert, C., Bidet, P., Peigne, C., Barbe, V., Medigue, C., Denamur, E. *et al.* (2007). A new O-antigen gene cluster has a key role in the virulence of the *Escherichia coli* meningitis clone O45: K1: H7. *Journal of bacteriology , 189* (23), 8528-8536.

Poptsova, M. (2009). Testing phylogenetic methods to identify horizontal gene transfer.. *Methods Mol Biol,* Volume 532, 227-240.

Poptsova, M. S. & Gogarten, J. P. (2007). The power of phylogenetic approaches to detect horizontally transferred genes.. *BMC Evol Biol,* Volume 7, 45.

Portnoy, D., Moseley, S., & Falkow, S. (1981). Characterization of plasmids and plasmid-associated determinants of *Yersinia enterocolitica* pathogenesis. *Infection and immunity , 31* (2), 775-782.

Potron, A., Kalpoe, J., Poirel, L., & Nordmann, P. (2011). European dissemination of a single OXA-48-producing *Klebsiella pneumoniae* clone. *Clinical Microbiology and Infection , 17* (12), E24--E26.

Prakash, R., & Atherly, A. (1984). Reiteration of genes involved in symbiotic nitrogen fixation by fast-growing *Rhizobium japonicum*. *Journal of bacteriology , 160* (2), 785-787.

Pride, D. T., & Blaser, M. J. (2002). Identification of horizontally acquired genetic elements in *Helicobacter pylori* and other prokaryotes using oligonucleotide difference analysis. *Genome Letters , 1* (1), 2-15.

Pride, D., Meinersmann, R., Wassenaar, T., & Blaser, M. (2003). Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome research , 13* (2), 145-158.

Rajan, I., Aravamuthan, S., & Mande, S. S. (2007). Identification of compositionally distinct regions in genomes using the centroid method. *Bioinformatics , 23* (20), 2672-2677.

Ramaiah, N., & De, J. (2003). Unusual rise in mercury-resistant bacteria in coastal environs. *Microbial ecology , 45* (4), 444-454.

Rankin, D., Rocha, E., & Brown, S. (2010). What traits are carried on mobile genetic elements, and why & quest. *Heredity , 106* (1), 1-10.

Reiter, W.-D., Palm, P., & Yeats, S. (1989). Transfer RNA genes frequently serve as integration sites for prokaryotic genetic elements. *Nucleic acids research , 17* (5), 1907-1914.

Rendon, M. a., Saldana, Z., Erdem, A. L., Monteiro-Neto, V., Vazquez, A., Kaper, J. B. *et al.* (2007). Commensal and pathogenic *Escherichia coli* use a common pilus adherence factor for epithelial cell colonization. *Proceedings of the National Academy of Sciences , 104* (25), 10637-10642.

Renzoni, A., Andrey, D. O., Jousselin, A., Barras, C., Monod, A., Vaudaux, P. *et al.* (2011). Whole genome sequencing and complete genetic analysis reveals novel pathways to glycopeptide resistance in *Staphylococcus aureus*. *PLoS One , 6* (6), e21577.

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

Reva, O. N., & Tümmler, B. (2004). Global features of sequences of bacterial chromosomes, plasmids and phages revealed by analysis of oligonucleotide usage patterns. *BMC Bioinformatics , 5*, 90.

Reva, O., & Tümmler, B. (2008). Think big--giant genes in bacteria. *Environ Microbiol , 10* (3), 768-777.

Riesenfeld, C. S., Goodman, R. M., & Handelsman, J. (2004). Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. *Environmental microbiology , 6* (9), 981-989.

Ripoll, F., Pasek, S., Schenowitz, C., Dossat, C., Barbe, V., Rottman, M. *et al.* (2009). Non mycobacterial virulence genes in the genome of the emerging pathogen *Mycobacterium abscessus*. *PloS one , 4* (6), e5660.

Ritter, A., Blum, G., Emody, L., Kerenyi, M., Bock, A., Neuhieri, B. *et al.* (1995). tRNA genes and pathogenicity islands: influence on virulence and metabolic properties of uropathogenic *Escherichia coli*. *Molecular microbiology , 17* (1), 109-121.

Rivera, M. C., Jain, R., Moore, J. E., & Lake, J. A. (1998). Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci U S A , 95* (11), 6239-6244.

Rohde, H., Qin, J., Cui, Y., Li, D., Loman, N. J., Hentschke, M. *et al.* (2011). Open-source genomic analysis of Shiga-toxin--producing *E. coli* O104: H4. *New England Journal of Medicine , 365* (8), 718-724.

Rosewarne, C. P., Pettigrove, V., Stokes, H. W., & Parsons, Y. M. (2010). Class 1 integrons in benthic bacterial communities: abundance, association with Tn402-like transposition modules and evidence for coselection with heavy-metal resistance. *FEMS microbiology ecology , 72* (1), 35-46.

Rowley, G., Skovierova, H., Stevenson, A., Rezuchova, B., Homerova, D., Lewis, C. *et al.* (2011). The periplasmic chaperone Skp is required for successful *Salmonella* Typhimurium infection in a murine typhoid model. *Microbiology , 157* (3), 848-858.

Ruer, S., Stender, S., Filloux, A., & de Bentzmann, S. (2007). Assembly of fimbrial structures in *Pseudomonas aeruginosa*: functionality and specificity of chaperone-usher machineries. *Journal of bacteriology , 189* (9), 3547-3555.

Russell, A. D. (1999). Bacterial resistance to disinfectants: present knowledge and future problems.. *J Hosp Infect*, Dec, Volume 43 Suppl, S57--S68.

Russell, G., Walker, P., Elton, R., & Subak-Sharpe, J. (1976). Doublet frequency analysis of fractionated vertebrate nuclear DNA. *Journal of molecular biology , 108* (1), 1-20.

Rychlik, I., Karasova, D., Sebkova, A., Volf, J., Sisak, F., Havlickova, H. *et al.* (2009).

Virulence potential of five major pathogenicity islands (SPI-1 to SPI-5) of *Salmonella enterica* serovar Enteritidis for chickens. *BMC microbiology , 9* (1), 268.

Sandberg, R., Winberg, G., Branden, C. I., Kaske, A., Ernberg, I., & Coster, J. (2001). Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier. *Genome Res , 11* (8), 1404-1409.

Sanger, F., & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol , 94* (3), 441-448.

Sasakawa, C., Kamata, K., Sakai, T., Makino, S., Yamada, M., Okada, N. *et al.* (1988). Virulence-associated genetic regions comprising 31 kilobases of the 230-kilobase plasmid in *Shigella flexneri* 2a. *Journal of bacteriology , 170* (6), 2480-2484.

Sasaki, Y., Tsujiyama, Y., Kusukawa, M., Murakami, M., Katayama, S., & Yamada, Y. (2011). Prevalence and characterization of Shiga toxin-producing *Escherichia coli* O157 and O26 in beef farms. *Vet Microbiol , 150* (1-2), 140-145.

Schloter, M., Lebuhn, M., Heulin, T., & Hartmann, A. (2000). Ecology and evolution of bacterial microdiversity. *FEMS microbiology reviews , 24* (5), 647-660.

Schottel, J. et al., (1974). Volatilisation of mercury and organomercurials determined by inducible R-factor systems in enteric bacteria.. *Nature, 251*(5473), 335-337.

SCHMIDT, F., KRATZ, J., & WIEDEMANN, B. (1983). Identification of Tn2401, a transposon encoding multiresistance to aminoglycosides. *Journal of General Microbiology , 129* (5), 1527-1536.

Selezska, K., Kazmierczak, M., Musken, M., Garbe, J., Schobert, M., Haussler, S. *et al.* (2012). *Pseudomonas aeruginosa* population structure revisited under environmental focus: impact of water quality and phage pressure. *Environmental Microbiology , 14* (8), 1952-1967.

Sharp, P. M., & Li, W. H. (1987). The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res , 15* (3), 1281-1295.

Sheldon, T. (2012). Dutch microbiologists are disciplined for delays during pneumonia outbreak. *BMJ , 344*.

Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M. *et al.* (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science , 309* (5741), 1728-1732.

Shrivastava, S., Reddy, C. V., & Mande, S. S. (2010). INDeGenIUS, a new method for high-throughput identification of specialized functional islands in completely sequenced organisms. *Journal of biosciences , 35* (3), 351-364.

Sia, E. A., Kuehner, D. M., & Figurski, D. H. (1996). Mechanism of retrotransfer in conjugation: prior transfer of the conjugative plasmid is required. *Journal of bacteriology , 178* (5), 1457-1464.

Silver, S. & Phung, L. T. (199. Bacterial heavy metal resistance: new surprises.. *Annu Rev Microbiol*, Volume 50, 753-789.

Smith, E. E., Buckley, D. G., Wu, Z., Saenphimmachak, C., Hoffman, L. R., DArgenio, D. A. *et al.* (2006). Genetic adaptation by *Pseudomonas aeruginosa* to the airways of cystic fibrosis patients. *Proceedings of the National Academy of Sciences , 103* (22), 8487-8492.

Smith, J. M., Feil, E. J., Smith, N. H., & others. (2000). Population structure and evolutionary dynamics of pathogenic bacteria. *Bioessays , 22* (12), 1115-1122.

Sokol, P. A., Luan, M.-Z., Storey, D. G., & Thirukkumaran, P. (1994). Genetic rearrangement associated with in vivo mucoid conversion of *Pseudomonas aeruginosa* PAO is due to insertion elements. *Journal of bacteriology , 176* (3), 553-562.

Sokurenko, E. V., Hasty, D. L., & Dykhuizen, D. E. (1999). Pathoadaptive mutations: gene loss and variation in bacterial pathogens. *Trends in microbiology , 7* (5), 191-195.

Somerville, G. A., & Proctor, R. A. (2009). At the crossroads of bacterial metabolism and virulence factor synthesis in Staphylococci. *Microbiology and Molecular Biology Reviews , 73* (2), 233-248.

Spanier, J. G., & Cleary, P. P. (1980). Bacteriophage control of antiphagocytic determinants in group A streptococci. *J Exp Med , 152* (5), 1393-1406.

Spencer, D. H., Kas, A., Smith, E. E., Raymond, C. K., Sims, E. H., Hastings, M. *et al.* (2003). Whole-genome sequence variation among multiple isolates of *Pseudomonas*

aeruginosa. *Journal of bacteriology , 185* (4), 1316-1325.

Srividhya, K. V., Alaguraj, V., Poornima, G., Kumar, D., Singh, G. P., Raghavenderan, L. *et al.* (2007). Identification of prophages in bacterial genomes by dinucleotide relative abundance difference. *PLoS ONE , 2* (11), e1193.

Stockbauer, K. E., Magoun, L., Liu, M., Burns, E. H., Gubba, S., Renish, S. *et al.* (1999). A natural variant of the cysteine protease virulence factor of group A Streptococcus with an arginine-glycine-aspartic acid (RGD) motif preferentially binds human integrins alphavbeta3 and alphaIIbbeta3. *Proceedings of the National Academy of Sciences , 96* (1), 242-247.

Stover, C., Pham, X., Erwin, A., Mizoguchi, S., Warrener, P., Hickey, M. *et al.* (2000). Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. *Nature , 406* (6799), 959-964.

St-Pierre, N. R., & Weiss, W. P. (2009). Technical note: Designing and analyzing quantitative factorial experiments. *J Dairy Sci , 92* (9), 4581-4588.

St-Pierre, N., & Weiss, W. (2009). Technical note: Designing and analyzing quantitative factorial experiments. *Journal of dairy science , 92* (9), 4581-4588.

Strätz, M., Mau, M., & Timmis, K. N. (1996). System to study horizontal gene exchange among microorganisms without cultivation of recipients. *Molecular microbiology , 22* (2), 207-215.

Sueoka, N. (1962). On the genetic basis of variation and heterogeneity of DNA base composition. *Proceedings of the National Academy of Sciences of the United States of America , 48* (4), 582.

Sullivan, J. T., & Ronson, C. W. (1998). Evolution of rhizobia by acquisition of a 500-kb symbiosis island that integrates into a phe-tRNA gene. *Proceedings of the National Academy of Sciences , 95* (9), 5145-5149.

Sullivan, J. T., Trzebiatowski, J. R., Cruickshank, R. W., Gouzy, J., Brown, S. D., Elliot, R. M. *et al.* (2002). Comparative sequence analysis of the symbiosis island of *Mesorhizobium loti* strain R7A. *J Bacteriol , 184* (11), 3086-3095.

Sun, Y.-C., Koumoutsi, A., Jarrett, C., Lawrence, K., Gherardini, F. C., Darby, C. *et al.*

(2011). Differential control of *Yersinia pestis* biofilm formation in vitro and in the flea vector by two c-di-GMP diguanylate cyclases. *PLoS One , 6* (4), e19267.

Tümmler, B. (2006). Clonal variations in *Pseudomonas aeruginosa*. In *Pseudomonas* (pp. 35-68). Springer.

Thomas, C., & Nielsen, K. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature reviews microbiology , 3* (9), 711-721.

Thompson, L. R. et al., (2011). Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism.. *Proc Natl Acad Sci U S A, 108*(39), E757--E764.

Thorvaldsdottir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics , 14* (2), 178-192.

Touchon, M., Charpentier, S., Clermont, O., Rocha, E. P., Denamur, E., & Branger, C. (2011). CRISPR distribution within the *Escherichia coli* species is not suggestive of immunity-associated diversifying selection. *Journal of bacteriology , 193* (10), 2460-2467.

Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P. *et al.* (2009). Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS genetics , 5* (1), e1000344.

Toussaint, B., Delicattree, I., & Vignais, P. M. (1993). *Pseudomonas aeruginosa* Contains an IHF-like Protein That Binds to the algD Promoter. *Biochemical and biophysical research communications , 196* (1), 416-421.

Tu, Q., & Ding, D. (2003). Detecting pathogenicity islands and anomalous gene clusters by iterative discriminant analysis. *FEMS microbiology letters , 221* (2), 269-275.

Tümmler, B., Koopmann, U., Grothues, D., Weissbrodt, H., Steinkamp, G., & Von Der Hardt, H. (1991). Nosocomial acquisition of *Pseudomonas aeruginosa* by cystic fibrosis patients. *Journal of clinical microbiology , 29* (6), 1265-1267.

Turner, S. A., Luck, S. N., Sakellaris, H., Rajakumar, K., & Adler, B. (2003). Molecular epidemiology of the SRL pathogenicity island. *Antimicrobial agents and chemotherapy , 47* (2), 727-734.

Uchiumi, T., Ohwada, T., Itakura, M., Mitsui, H., Nukui, N., Dawadi, P. *et al.* (2004). Expression islands clustered on the symbiosis island of the *Mesorhizobium loti* genome. *Journal of bacteriology , 186* (8), 2439-2448.

Van Dongen, S. (2008). Graph clustering via a discrete uncoupling process. *SIAM Journal on Matrix Analysis and Applications , 30* (1), 121-141.

van Passel, M. W. (2011). Tracing common origins of Genomic Islands in prokaryotes based on genome signature analyses. *Mobile Genetic Elements , 1* (3), 247-249.

van Passel, M. W., Bart, A., Thygesen, H. H., Luyf, A. C., van Kampen, A. H., & van der Ende, A. (2005). An acquisition account of genomic islands based on genome signature comparisons. *BMC genomics , 6* (1), 163.

van Passel, M. W., Marri, P. R., & Ochman, H. (2008). The emergence and fate of horizontally acquired genes in *Escherichia coli*. *PLoS computational biology , 4* (4), e1000059.

Vernikos, G. S., & Parkhill, J. (2006). Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. *Bioinformatics , 22* (18), 2196-2203.

Vernikos, G. S., Thomson, N. R., & Parkhill, J. (2007). Genetic flux over time in the *Salmonella* lineage. *Genome Biol , 8* (6), R100.

Volkovich, Z., Kirzhner, V., & Barzily, Z. (2010). *Genome Clustering: From Linguistic Models to Classification of Genetic Texts* (Vol. 286). Springer.

Waack, S., Keller, O., Asper, R., Brodag, T., Damm, C., Fricke, W. F. *et al.* (2006). Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC bioinformatics , 7* (1), 142.

Wang, B. (2001). Limitations of compositional approach to identifying horizontally transferred genes. *Journal of Molecular Evolution , 53* (3), 244-250.

Wang, G., Zhou, F., Olman, V., Li, F., & Xu, Y. (2010). Prediction of pathogenicity islands in Enterohemorrhagic *Escherichia coli* O157: H7 using genomic barcodes. *FEBS letters , 584* (1), 194-198.

Wang, G., Zhou, F., Olman, V., Li, F., & Xu, Y. (2010). Prediction of pathogenicity islands in enterohemorrhagic *Escherichia coli* O157:H7 using genomic barcodes. *FEBS Lett , 584* (1), 194-198.

Wang, L., & Reeves, P. R. (1998). Organization of *Escherichia coli* O157 O antigen gene cluster and identification of its specific genes. *Infection and immunity , 66* (8), 3545-3551.

Wei, W., & Guo, F.-B. (2011). Prediction of genomic islands in seven human pathogens using the Z-Island method. *Genet Mol Res , 10* (4), 2307-2315.

Weiner, M., & Werthamer, S. (1973). Opportunistic infections with *Erwinia*-like species. *New York state journal of medicine , 73* (18), 2256.

Wiehlmann, L., Cramer, N., Ulrich, J., Hedtfeld, S., Weissbrodt, H., & Tümmler, B. (2012). Effective prevention of *Pseudomonas aeruginosa* cross-infection at a cystic fibrosis centre--Results of a 10-year prospective study. *International Journal of Medical Microbiology , 302* (2), 69-77.

Wiehlmann, L., Wagner, G., Cramer, N., Siebert, B., Gudowius, P., Morales, G. *et al.* (2007). Population structure of *Pseudomonas aeruginosa*. *Proceedings of the National Academy of Sciences , 104* (19), 8101-8106.

Wilson, D. J. (2012). Insights from genomics into bacterial pathogen populations. *PLoS pathogens , 8* (9), e1002874.

Winsor, G. L., Lam, D. K., Fleming, L., Lo, R., Whiteside, M. D., Nancy, Y. Y. *et al.* (2011). *Pseudomonas* Genome Database: improved comparative analysis and population genomics capability for *Pseudomonas* genomes. *Nucleic acids research , 39* (suppl 1), D596--D600.

Winstanley, C., Langille, M. G., Fothergill, J. L., Kukavica-Ibrulj, I., Paradis-Bleau, C., Sanschagrin, F. *et al.* (2009). Newly introduced genomic prophage islands are critical determinants of in vivo competitiveness in the Liverpool Epidemic Strain of *Pseudomonas aeruginosa*. *Genome research , 19* (1), 12-23.

Yang, J. et al., (2008). VFDB 2008 release: an enhanced web-based resource for comparative pathogenomics.. *Nucleic Acids Res*, 36, D539--D542.

Yang, L., Jelsbak, L., Marvig, R. L., Damkiaer, S., Workman, C. T., Rau, M. H. *et al.* (2011).

Evolutionary dynamics of bacteria in a human host environment. *Proceedings of the National Academy of Sciences , 108* (18), 7481-7486.

Yap, W. H., Zhang, Z., & Wang, Y. (1999). Distinct types of rRNA operons exist in the genome of the actinomycete *Thermomonospora chromogena* and evidence for horizontal transfer of an entire rRNA operon. *Journal of bacteriology , 181* (17), 5201-5209.

Yeung, A. T., Bains, M., & Hancock, R. E. (2011). The sensor kinase CbrA is a global regulator that modulates metabolism, virulence, and antibiotic resistance in *Pseudomonas aeruginosa*. *Journal of bacteriology , 193* (4), 918-931.

Yoon, S. H., Park, Y.-K., Lee, S., Choi, D., Oh, T. K., Hur, C.-G. *et al.* (2007). Towards pathogenomics: a web-based resource for pathogenicity islands. *Nucleic Acids Res , 35* (Database issue), D395--D400.

Yoon, S., Hur, C., Kang, H., Kim, Y., Oh, T., & Kim, J. (2005). A computational approach for identifying pathogenicity islands in prokaryotic genomes. *Bmc Bioinformatics , 6* (1), 184.

Yu, H., Hanes, M., Chrisp, C., Boucher, J., & Deretic, V. (1998). Microbial pathogenesis in cystic fibrosis: pulmonary clearance of mucoid *Pseudomonas aeruginosa* and inflammation in a mouse model of repeated respiratory challenge. *Infection and immunity , 66* (1), 280-288.

Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research , 18* (5), 821-829.

Zienkiewicz, M., Kern-Zdanowicz, I., Goebiewski, M., Zyliñska, J., Mieczkowski, P., Gniadkowski, M. *et al.* (2007). Mosaic structure of p1658/97, a 125-kilobase plasmid harbouring an active amplicon with the extended-spectrum beta-lactamase gene blaSHV-5. *Antimicrob Agents Chemother , 51* (4), 1164-1171.

7 Appendices

7.1 Appendix A

7.1.1.1 Supplementary tables

Table S1. Re-identification of known PAIs available from PAIDB by different programs

Genomes	#	Name	Start	Stop	IslandViewer			SWGIS			
					IslandPick	SIGI-HMM	IslandPath	D 1.5 V 1.5	D 1.5 V 2.0	D 2.0 V 1.5	D 2.0 V 2.0
NC_004668, Enterococcus faecalis V583	1	Not named	445282	582786	0	1	1	1	1	1	1
NC_004431, *Escherichia coli* CFT073	1	PAI I CFT073	3406225	3450866	0	1	1	1	1	1	1
	2	PAI II CFT073	4913367	4971660	0	0	1	1	0	1	0
NC_002655, *Escherichia coli* O157:H7 EDL933	1	LEE	4649752	4692545	0	1	1	1	1	1	1
NC_002695, *Escherichia coli* O157:H7 Sakai	1	LEE	4580769	4623562	0	1	1	1	1	1	1
NC_006570, Francisella tularensis subsp. tularensis Schu 4	1	FPI	1374701	1408279	0	0	0	1	1	1	1
	2	FPI	1768045	1801623	0	0	0	1	1	1	1
NC_004917, Helicobacter hepaticus ATCC 51449	1	HHGI1	223218	294244	0	0	0	1	0	0	0
NC_000915, Helicobacter pylori 26695	1	cag PAI	547328	585350	0	0	0	1	0	1	0
NC_000921, Helicobacter pylori J99	1	cag PAI	510500	548694	0	0	0	1	0	1	0
NC_003112, Neisseria meningitidis MC58	1	IHT-A	75694	109777	0	0	0	1	1	1	1
	2	IHT-C	1827738	1860290	0	1	1	1	1	1	1
NC_004578, Pseudomonas syringae pv. tomato DC3000	1	Hrp PAI	1502395	1551861	0	0	0	0	0	0	0

NC_006905, Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67	1	SPI-5	1155936	1161624	0	0	0	0	0	0	0
	2	SPI-11	1350481	1366166	0	1	0	1	1	1	1
	3	SPI-2	1497670	1539498	0	1	0	1	1	1	1
	4	SPI-12	2354604	2365678	0	1	0	1	0	1	0
	5	SPI-1	2960260	3003747	0	1	0	1	1	1	1
	6	SPI-3	3890879	3903697	0	0	1	0	0	0	0
	7	SPI-4	4411902	4438599	0	1	0	1	1	1	1
NC_003198, Salmonella enterica subsp. enterica serovar Typhi CT18	1	SPI-6	302092	360757	0	1	0	1	0	1	0
	2	SPI-5	1085068	1092563	0	0	0	1	1	0	0
	3	SPI-2	1624920	1666524	0	1	0	1	1	1	1
	4	SPI-9	2743495	2759190	0	0	0	0	0	0	0
	5	SPI-1	2858736	2900586	0	1	0	1	1	1	1
	6	SPI-8	3132530	3139414	0	1	0	1	1	1	1
	7	SPI-3	3883613	3900553	0	0	0	1	0	0	0
	8	SPI-4	4322993	4346383	0	1	0	1	1	1	1
	9	SPI-7	4409511	4543148	1	1	1	1	1	1	1
	10	SPI-10	4683605	4716538	1	1	0	1	1	1	1
NC_004631, Salmonella enterica subsp. enterica serovar Typhi Ty2	1	SPI-2	1314607	1356216	0	1	0	1	1	1	1
	2	SPI-1	2844593	2886443	0	1	0	1	1	1	1
	3	SPI-7	4394302	4526050	1	1	1	1	1	1	1
NC_003197, Salmonella typhimurium LT2	1	SPI-5	1175321	1184389	0	0	0	1	0	0	0
	2	SPI-2	1461740	1501800	0	1	0	1	1	1	1
	3	SPI-1	3005842	3050120	0	1	0	1	1	1	1
	4	SPI-3	3948576	3965191	0	0	0	0	0	0	0

	5	SPI-4	4477849	4501259	0	1	0	1	1	1	1
NC_002505, Vibrio cholerae O1 biovar eltor str. N16961	1	VSP-I	175343	189380	0	1	1	1	1	1	1
	2	VSP-II	523156	530602	0	1	1	1	1	1	1
	3	VPI	873242	914124	0	1	1	1	1	1	1
	4	VPI-2	1895692	1952861	0	1	1	1	1	1	1
NC_003919, Xanthomonas axonopodis pv. citri str. 306	1	Hrp PAI	462225	488334	0	0	0	1	0	0	0
NC_007086, Xanthomonas campestris pv. campestris str. 8004	1	Hrp PAI	3596348	3619441	0	1	0	1	0	1	0
NC_003902, Xanthomonas campestris pv. campestris str. ATCC 33913	1	Hrp PAI	1424335	1447429	0	1	0	1	0	1	0
NC_007508, Xanthomonas campestris pv. vesicatoria str. 85-10	1	Hrp PAI	459479	494176	0	1	0	1	0	1	0
NC_006834, Xanthomonas oryzae pv. oryzae KACC10331	1	Hrp PAI	58273	90451	0	0	1	0	0	0	0
NC_003143, Yersinia pestis CO92	1	HPI	2134947	2171362	0	0	1	1	1	1	1
NC_004088, Yersinia pestis KIM	1	HPI	2645436	2681339	0	0	1	1	1	1	1
NC_005810, Yersinia pestis biovar Medievalis str. 91001	1	HPI	1799736	1828106	0	0	0	1	1	1	1
NC_006155, Yersinia pseudotuberculosis IP32953	1	HPI	1913940	1949837	0	0	0	1	1	1	1
False negative rates					94.12%	41.18%	68.63%	11.76%	35.29%	21.57%	37.25%

Table S2. Numbers of genomic islands identified by different programs

#	Genomes	Total number of predicted Gis							Unconfirmed predictions						
		IslandViewer			SWGIS				IslandViewer			SWGIS			
		IslandPick	SIGI-HMM	IslandPath	D 1.5 V 1.5	D 1.5 V 2.0	D 2.0 V 1.5	D 2.0 V 2.0	IslandPick	SIGI-HMM	IslandPath	D 1.5 V 1.5	D 1.5 V 2.0	D 2.0 V 1.5	D 2.0 V 2.0
1	Acidovorax avenae subsp. citrulli AAC00-1 [NC_008752]	0	35	18	32	24	25	20	0	10	5	9	5	5	5
2	Acidovorax sp. JS42 [NC_008782]	25	34	11	27	19	17	17	7	7	1	9	6	7	7
3	Acinetobacter baumannii ATCC 17978 [NC_009085]	4	4	4	8	7	5	4	1	3	2	6	5	4	3
4	Anaeromyxobacter dehalogenans 2CP-C [NC_007760]	7	18	2	21	16	12	10	3	4	0	11	7	3	3
5	Arthrobacter sp. FB24 chr. 1 [NC_008541]	0	12	3	16	14	10	9	0	2	1	8	7	5	5
6	Azoarcus sp. EbN1 [NC_006513]	0	22	24	15	14	10	10	0	2	11	4	4	3	3
7	Bacillus anthracis str. 'Ames Ancestor' [NC_007530]	3	0	5	38	19	9	7	0	0	3	34	18	8	7
8	Bacillus anthracis str. Ames [NC_003997]	3	0	5	35	20	9	7	0	0	3	31	19	8	7
9	Bacillus anthracis str. Sterne [NC_005945]	3	0	5	41	21	9	7	0	0	3	37	20	8	7
10	Bacillus cereus ATCC 10987 [NC_003909]	3	4	10	30	17	11	8	1	0	5	26	15	10	7
11	Bacillus cereus ATCC 14579 [NC_004722]	5	2	9	32	21	12	11	0	2	6	29	21	12	11
12	Bacillus cereus E33L [NC_006274]	0	1	5	33	23	10	10	0	0	4	33	23	10	10
13	Bacillus licheniformis ATCC 14580 [NC_006322]	0	3	8	27	27	12	12	0	0	0	16	17	6	6
14	Bacillus thuringiensis serovar konkukian str. 97-27 [NC_005957]	0	1	7	34	21	8	8	0	1	7	34	21	8	8
15	Bacillus thuringiensis str. Al Hakam [NC_008600]	2	2	5	29	20	13	10	0	1	2	27	19	13	10
16	Bordetella bronchiseptica RB50 [NC_002927]	0	25	6	16	15	12	12	0	8	0	5	4	3	3
17	Bordetella parapertussis 12822 [NC_002928]	0	11	2	8	7	6	6	0	4	1	4	4	4	4
18	Bordetella pertussis Tohama I [NC_002929]	0	11	18	7	7	6	6	0	4	12	3	3	3	3
19	Borrelia afzelii PKo [NC_008277]	0	8	0	4	3	3	2	0	8	0	4	3	3	2
20	Borrelia burgdorferi B31 [NC_001318]	0	3	0	0	0	0	0	0	3	0	0	0	0	0
21	Borrelia garinii PBi chr. linear [NC_006156]	0	6	0	6	3	2	2	0	6	0	6	3	2	2
22	Borrelia turicatae 91E135 [NC_008710]	0	24	0	10	5	6	2	0	21	0	8	5	5	2
23	Bradyrhizobium japonicum USDA 110 [NC_004463]	0	71	36	27	27	8	8	0	12	8	1	1	1	1
24	Burkholderia cenocepacia AU 1054 chr. 1 [NC_008060]	10	10	10	13	12	11	11	7	1	1	3	3	3	3
25	Burkholderia cenocepacia AU 1054 chr. 2 [NC_008061]	1	14	9	11	11	7	7	0	1	1	1	1	1	1
26	Burkholderia cenocepacia AU 1054 chr. 3 [NC_008062]	0	1	3	3	3	1	1	0	0	3	2	2	1	1

27	Burkholderia cenocepacia HI2424 chr. 1 [NC_008542]	2	12	9	10	10	10	10	0	1	2	3	3	3	3
28	Burkholderia cenocepacia HI2424 chr. 2 [NC_008543]	1	16	8	13	12	8	8	1	2	1	2	1	1	1
29	Burkholderia cenocepacia HI2424 chr. 3 [NC_008544]	1	2	3	5	5	2	2	0	0	2	3	3	1	1
30	Burkholderia cepacia AMMD chr. 1 [NC_008390]	1	15	5	14	14	11	11	0	0	0	3	3	3	3
31	Burkholderia cepacia AMMD chr. 2 [NC_008391]	1	13	2	12	11	5	5	0	3	1	2	2	1	1
32	Burkholderia cepacia AMMD chr. 3 [NC_008392]	7	8	2	4	4	2	2	6	3	2	2	2	1	1
33	Burkholderia mallei ATCC 23344 chr. 1 [NC_006348]	0	2	16	5	5	3	3	0	0	13	1	1	2	2
34	Burkholderia mallei ATCC 23344 chr. 2 [NC_006349]	0	3	17	5	5	2	2	0	1	14	2	2	0	0
35	Burkholderia mallei NCTC 10229 chr. I [NC_008835]	0	3	14	5	5	2	2	0	0	9	1	1	0	0
36	Burkholderia mallei NCTC 10229 chr. II [NC_008836]	0	4	18	4	4	3	3	0	0	15	1	1	1	1
37	Burkholderia mallei NCTC 10247 chr. I [NC_009079]	0	2	15	6	6	2	2	0	0	10	2	2	1	1
38	Burkholderia mallei NCTC 10247 chr. II [NC_009080]	1	4	19	4	4	3	3	0	0	15	2	2	2	2
39	Burkholderia mallei SAVP1 chr. I [NC_008784]	0	1	10	3	3	0	0	0	0	9	2	2	0	0
40	Burkholderia mallei SAVP1 chr. II [NC_008785]	0	4	20	5	5	3	3	0	0	17	2	2	1	1
41	Burkholderia pseudomallei 1106a chr. I [NC_009076]	2	17	10	16	15	10	10	0	2	1	3	3	2	2
42	Burkholderia pseudomallei 1106a chr. II [NC_009078]	4	11	8	10	10	6	6	2	1	3	3	3	1	1
43	Burkholderia pseudomallei 1710b chr. I [NC_007434]	7	25	9	18	18	11	11	0	4	0	3	3	2	2
44	Burkholderia pseudomallei 1710b chr. II [NC_007435]	5	13	6	11	11	6	6	1	5	1	4	4	2	2
45	Burkholderia pseudomallei 668 chr. I [NC_009074]	3	14	12	15	15	13	13	1	1	2	3	3	2	2
46	Burkholderia pseudomallei 668 chr. II [NC_009075]	1	10	7	9	9	5	5	1	2	2	3	3	1	1
47	Burkholderia pseudomallei K96243 chr. 1 [NC_006350]	2	19	10	15	15	11	11	1	3	0	3	3	3	3
48	Burkholderia pseudomallei K96243 chr. 2 [NC_006351]	3	6	4	11	10	6	6	1	1	0	6	5	2	2
49	Burkholderia sp. 383 chr. 1 [NC_007510]	5	12	2	14	14	8	8	1	1	0	4	4	3	3
50	Burkholderia sp. 383 chr. 2 [NC_007511]	0	11	2	12	12	8	8	0	2	0	3	3	1	1
51	Burkholderia sp. 383 chr. 3 [NC_007509]	0	13	0	9	10	3	3	0	5	0	3	4	1	1
52	Burkholderia thailandensis E264 chr. I [NC_007651]	5	12	14	13	13	11	11	0	0	1	0	0	0	0
53	Burkholderia thailandensis E264 chr. II [NC_007650]	8	0	9	8	8	5	5	6	0	4	3	3	2	2
54	Burkholderia xenovorans LB400 chr. 1 [NC_007951]	0	38	18	22	19	7	7	0	6	2	4	4	3	3
55	Burkholderia xenovorans LB400 chr. 2 [NC_007952]	0	15	11	14	13	7	7	0	2	4	3	3	2	2
56	Burkholderia xenovorans LB400 chr. 3 [NC_007953]	0	4	3	9	8	2	2	0	0	2	5	4	0	0
57	Campylobacter fetus subsp. fetus 82-40 [NC_008599]	0	1	2	8	7	0	0	0	0	2	7	6	0	0
58	Campylobacter jejuni RM1221 [NC_003912]	5	0	3	8	6	4	4	0	0	0	6	5	4	4
59	Candidatus Protochlamydia amoebophila UWE25 [NC_005861]	0	1	6	5	5	0	0	0	0	5	3	3	0	0
60	Caulobacter crescentus CB15 [NC_002696]	0	14	7	4	4	2	2	0	10	5	1	1	1	1
61	Clostridium acetobutylicum ATCC 824 [NC_003030]	0	2	6	23	12	21	11	0	1	5	23	12	21	11
62	Clostridium difficile 630 [NC_009089]	0	14	7	22	14	22	14	0	6	0	13	6	13	6
63	Ehrlichia canis str. Jake [NC_007354]	0	20	0	10	7	9	7	0	17	0	7	5	6	5
64	Ehrlichia chaffeensis str. Arkansas [NC_007799]	0	13	0	12	7	11	5	0	12	0	11	6	9	3

65	Ehrlichia ruminantium str. Gardel [NC_006831]	0	11	0	16	10	15	10	0	10	0	15	9	14	9
66	Ehrlichia ruminantium str. Welgevonden [NC_005295]	0	12	0	0	0	0	9	0	12	0	0	0	0	9
67	Ehrlichia ruminantium str. Welgevonden [NC_006832]	0	10	0	15	10	15	0	0	10	0	15	10	15	0
68	Erythrobacter litoralis HTCC2594 [NC_007722]	0	11	8	6	6	3	3	0	3	2	1	1	1	1
69	Francisella tularensis subsp. holarctica [NC_007880]	0	0	9	15	9	4	5	0	0	7	13	8	4	5
70	Frankia alni ACN14a [NC_008278]	0	13	11	38	28	21	19	0	6	1	23	15	11	9
71	Frankia sp. CcI3 [NC_007777]	0	14	25	25	21	12	12	0	3	12	6	6	4	4
72	Geobacillus kaustophilus HTA426 [NC_006510]	4	31	16	28	21	16	16	0	1	0	6	6	1	1
73	Gramella forsetii KT0803 [NC_008571]	0	3	8	15	10	6	4	0	1	6	14	8	5	3
74	Haemophilus ducreyi 35000HP [NC_002940]	0	3	4	10	9	8	8	0	2	4	9	8	8	8
75	Haemophilus influenzae 86-028NP [NC_007146]	3	4	3	7	7	6	6	2	0	1	6	6	6	6
76	Haemophilus somnus 129PT [NC_008309]	2	0	5	16	15	13	13	1	0	3	15	15	13	13
77	Halobacterium sp. NRC-1 [NC_002607]	0	8	6	0	0	0	0	0	2	2	0	0	0	0
78	Halorhodospira halophila SL1 [NC_008789]	0	11	6	12	7	9	7	0	0	1	5	2	2	2
79	Helicobacter acinonychis str. Sheeba [NC_008229]	6	0	4	3	3	0	0	3	0	3	1	1	0	0
80	Hyphomonas neptunium ATCC 15444 [NC_008358]	0	11	3	9	8	3	3	0	2	1	3	2	1	1
81	Lactobacillus johnsonii NCC 533 [NC_005362]	8	0	4	7	5	3	3	4	0	3	5	4	3	3
82	Lactococcus lactis subsp. cremoris MG1363 [NC_009004]	3	6	16	10	9	5	5	0	1	9	6	6	4	4
83	Lactococcus lactis subsp. cremoris SK11 [NC_008527]	3	1	13	11	9	7	7	1	0	11	7	6	5	5
84	Lactococcus lactis subsp. lactis Il1403 [NC_002662]	1	1	11	9	7	5	5	0	0	9	7	6	5	5
85	Legionella pneumophila str. Lens [NC_006369]	1	1	5	7	4	5	4	1	0	5	6	3	4	3
86	Legionella pneumophila subsp. pneumophila str. Philadelphia 1 [NC_002942]	4	1	3	7	7	4	4	3	0	2	6	6	4	4
87	Leifsonia xyli subsp. xyli str. CTCB07 [NC_006087]	0	19	11	18	13	15	13	0	3	1	5	2	4	2
88	Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130 chr. I [NC_005823]	0	0	10	8	6	3	3	0	0	7	5	3	1	1
89	Leptospira interrogans serovar Lai str. 56601 chr. I [NC_004342]	0	0	15	9	7	2	2	0	0	23	4	2	0	0
90	Magnetospirillum magneticum AMB-1 [NC_007626]	0	48	15	20	19	18	19	0	13	3	0	0	0	0
91	Maricaulis maris MCS10 [NC_008347]	0	8	6	7	7	4	4	0	1	0	2	2	1	1
92	Mesorhizobium loti MAFF303099 [NC_002678]	0	40	17	17	16	5	5	0	16	5	1	1	1	1
93	Methylobacillus flagellatus KT [NC_007947]	0	9	3	13	9	2	2	0	5	2	11	7	1	1
94	Mycobacterium avium 104 [NC_008595]	6	26	13	21	20	2	2	1	6	6	6	6	1	1
95	Mycobacterium avium subsp. paratuberculosis K-10 [NC_002944]	3	16	9	12	10	6	6	0	6	6	3	3	0	0
96	Mycobacterium bovis AF2122/97 [NC_002945]	0	4	7	12	10	1	1	0	1	4	6	5	1	1
97	Mycobacterium bovis BCG str. Pasteur 1173P2 [NC_008769]	0	4	6	11	10	1	1	0	1	3	5	5	1	1
98	Mycobacterium smegmatis str. MC2 155 [NC_008596]	0	33	17	14	14	3	3	0	13	7	3	3	2	2
99	Mycobacterium sp. JLS [NC_009077]	0	29	16	16	16	12	12	0	6	4	2	2	2	2
100	Mycobacterium sp. KMS [NC_008705]	0	27	19	17	16	10	10	0	7	5	3	3	2	2

101	Mycobacterium sp. MCS [NC_008146]	0	23	20	14	16	10	10	0	7	8	3	3	2	2
102	Mycobacterium ulcerans Agy99 [NC_008611]	1	10	23	4	4	1	1	0	6	18	2	2	1	1
103	Mycobacterium vanbaalenii PYR-1 [NC_008726]	0	29	20	18	16	7	7	0	7	8	4	4	2	2
104	Myxococcus xanthus DK 1622 [NC_008095]	0	0	19	48	43	27	26	0	0	7	34	30	18	17
105	Neisseria meningitidis Z2491 [NC_003116]	1	6	2	15	13	7	7	1	1	0	9	7	4	4
106	Nitrobacter hamburgensis X14 [NC_007964]	0	18	15	7	7	0	0	0	10	9	3	3	0	0
107	Nitrobacter winogradskyi Nb-255 [NC_007406]	0	4	25	4	4	2	2	0	0	20	1	1	1	1
108	Nocardia farcinica IFM 10152 [NC_006361]	0	0	12	23	19	10	9	0	0	5	17	13	7	6
109	Nocardioides sp. JS614 [NC_008699]	0	24	14	20	19	16	15	0	4	0	7	6	3	3
110	Nostoc sp. PCC 7120 [NC_003272]	18	1	9	14	11	6	4	14	1	5	9	7	5	4
111	Novosphingobium aromaticivorans DSM 12444 [NC_007794]	0	9	5	7	7	2	2	0	3	2	2	2	1	1
112	Paracoccus denitrificans PD1222 chr. 1 [NC_008686]	0	22	5	11	10	8	8	0	6	0	0	0	0	0
113	Paracoccus denitrificans PD1222 chr. 2 [NC_008687]	0	8	5	7	6	3	3	0	3	1	3	2	0	0
114	Photobacterium profundum SS9 chr. 2 [NC_006371]	0	17	9	5	5	0	0	0	14	6	2	2	0	0
115	Polaromonas sp. JS666 [NC_007948]	0	23	10	13	14	8	8	0	11	4	4	4	2	2
116	Pyrobaculum islandicum DSM 4184 [NC_008701]	0	0	8	15	8	10	6	0	0	3	11	5	7	4
117	Pyrococcus furiosus DSM 3638 [NC_003413]	0	4	4	18	9	7	5	0	0	0	14	6	4	4
118	Ralstonia eutropha H16 chr. 1 [NC_008313]	0	35	4	14	13	9	9	0	21	0	2	2	2	2
119	Ralstonia eutropha H16 chr. 2 [NC_008314]	0	24	3	13	9	5	5	0	11	0	3	3	2	2
120	Ralstonia eutropha JMP134 chr. 1 [NC_007347]	0	19	4	12	12	9	9	0	8	1	3	3	3	3
121	Ralstonia eutropha JMP134 chr. 2 [NC_007348]	0	9	1	8	8	4	4	0	3	0	3	3	2	2
122	Ralstonia metallidurans CH34 chr. 2 [NC_007974]	0	11	3	12	9	5	5	0	0	0	7	6	2	2
123	Ralstonia solanacearum GMI1000 [NC_003295]	0	29	8	19	16	10	10	0	6	0	3	2	2	2
124	Rhodobacter sphaeroides 2.4.1 chr. 1 [NC_007493]	0	5	7	6	4	5	4	0	2	2	1	0	1	0
125	Rhodobacter sphaeroides 2.4.1 chr. 2 [NC_007494]	0	4	3	6	4	4	4	0	0	0	1	1	1	1
126	Rhodobacter sphaeroides ATCC 17029 chr. 1 [NC_009049]	0	5	4	5	4	3	3	0	0	0	0	0	0	0
127	Rhodobacter sphaeroides ATCC 17029 chr. 2 [NC_009050]	0	6	4	5	6	5	6	0	0	1	2	2	2	2
128	Rhodococcus sp. RHA1 [NC_008268]	0	8	25	20	19	6	6	0	4	11	11	10	5	5
129	Rhodopseudomonas palustris BisA53 [NC_008435]	0	19	9	9	9	9	9	0	2	3	3	3	2	2
130	Rhodopseudomonas palustris BisB18 [NC_007925]	0	13	9	10	10	6	6	0	5	4	2	2	1	1
131	Rhodopseudomonas palustris BisB5 [NC_007958]	12	10	7	7	7	3	3	3	1	1	1	1	1	1
132	Rhodopseudomonas palustris CGA009 [NC_005296]	0	9	6	5	5	3	3	0	5	5	3	3	2	2
133	Rhodopseudomonas palustris HaA2 [NC_007778]	20	16	7	7	7	5	5	12	4	1	1	1	1	1
134	Rhodospirillum rubrum ATCC 11170 [NC_007643]	0	9	4	10	10	5	5	0	4	2	4	4	2	2
135	Rubrobacter xylanophilus DSM 9941 [NC_008148]	0	11	5	13	7	11	7	0	0	1	5	0	3	0
136	Saccharopolyspora erythraea NRRL 2338 [NC_009142]	0	18	18	33	26	16	15	0	1	3	17	13	6	6
137	Silicibacter pomeroyi DSS-3 [NC_003911]	0	13	7	14	13	7	7	0	0	0	4	4	4	4
138	Sinorhizobium meliloti 1021 [NC_003047]	7	6	5	7	7	5	5	2	0	2	2	2	2	2

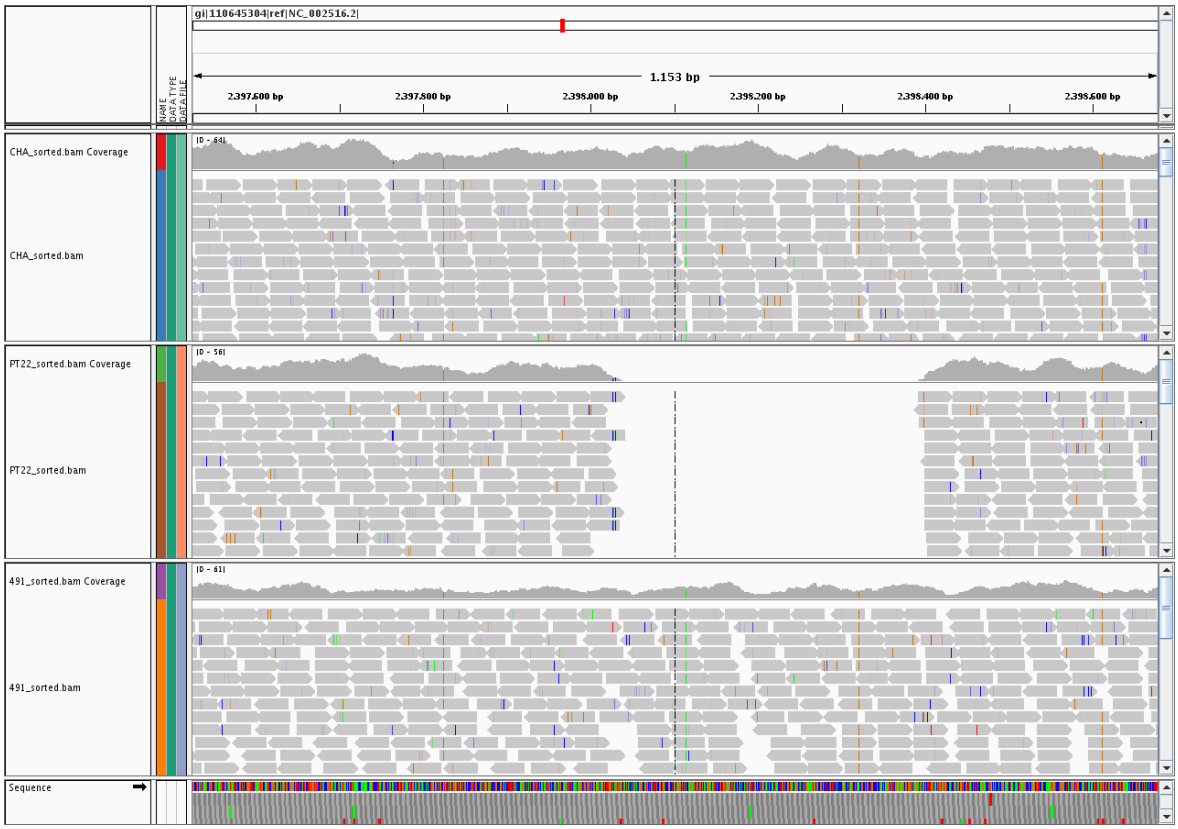
139	Sphingopyxis alaskensis RB2256 [NC_008048]	0	17	8	2	2	1	1	0	8	2	1	1	1	1
140	Staphylococcus aureus RF122 [NC_007622]	0	1	5	19	12	7	6	0	0	3	17	10	5	4
141	Staphylococcus aureus subsp. aureus MRSA252 [NC_002952]	0	3	10	19	14	8	7	0	2	6	15	10	7	5
142	Staphylococcus aureus subsp. aureus Mu50 [NC_002758]	1	4	7	19	13	10	8	0	1	2	16	11	7	6
143	Staphylococcus aureus subsp. aureus N315 [NC_002745]	0	4	8	18	12	10	7	0	2	4	16	10	8	5
144	Staphylococcus epidermidis ATCC 12228 [NC_004461]	0	3	7	19	13	11	9	0	1	5	16	10	8	6
145	Staphylococcus haemolyticus JCSC1435 [NC_007168]	0	0	10	19	12	8	7	0	0	8	17	10	8	7
146	Streptococcus agalactiae 2603V/R [NC_004116]	1	10	4	10	10	8	7	0	5	1	7	7	7	7
147	Streptomyces avermitilis MA-4680 [NC_003155]	0	11	20	37	27	18	18	0	2	7	24	16	11	11
148	Streptomyces coelicolor A3(2) [NC_003888]	0	29	17	44	42	27	28	0	4	3	21	19	8	9
149	Sulfolobus solfataricus P2 [NC_002754]	12	6	24	27	18	19	9	7	0	10	19	12	15	5
150	Sulfolobus tokodaii str. 7 [NC_003106]	0	14	7	25	18	17	13	0	6	2	18	13	14	10
151	Symbiobacterium thermophilum IAM 14863 [NC_006177]	0	10	13	17	15	15	15	0	0	1	9	8	6	6
152	Synechococcus sp. CC9311 [NC_008319]	0	15	1	26	21	22	20	0	0	0	14	9	11	8
153	Synechococcus sp. CC9605 [NC_007516]	0	29	6	26	18	25	18	0	2	1	4	2	4	2
154	Synechococcus sp. CC9902 [NC_007513]	0	11	1	16	14	11	11	0	1	0	7	5	3	3
155	Synechococcus sp. WH 8102 [NC_005070]	0	23	4	19	14	19	14	0	3	0	4	1	4	1
156	Thermosynechococcus elongatus BP-1 [NC_004113]	0	2	12	5	5	1	1	0	1	10	2	2	1	1
157	Thiobacillus denitrificans ATCC 25259 [NC_007404]	0	20	10	13	13	7	7	0	3	1	2	2	1	1
158	Thiomicrospira crunogena XCL-2 [NC_007520]	0	3	7	7	3	0	0	0	0	4	5	3	0	0
159	Verminephrobacter eiseniae EF01-2 [NC_008786]	0	48	15	36	31	16	16	0	13	6	7	6	4	4
160	Vibrio cholerae O1 biovar eltor str. N16961 chr. II [NC_002506]	0	7	5	0	0	0	0	0	1	0	0	0	0	0
161	Vibrio parahaemolyticus RIMD 2210633 chr. I [NC_004603]	8	16	7	16	11	4	2	3	6	1	6	5	0	0
162	Xanthomonas axonopodis pv. citri str. 306 [NC_003919]	6	24	15	22	20	14	14	2	2	2	7	6	3	3
163	Xanthomonas campestris pv. campestris str. 8004 [NC_007086]	6	34	9	23	22	16	16	1	6	0	3	3	2	2
164	Xanthomonas campestris pv. campestris str. ATCC 33913 [NC_003902]	4	36	10	27	24	15	15	1	5	0	5	5	3	3
165	Xanthomonas campestris pv. vesicatoria str. 85-10 [NC_007508]	6	31	24	29	27	12	12	1	5	2	7	6	2	2
166	Xanthomonas oryzae pv. oryzae KACC10331 [NC_006834]	11	8	23	9	9	5	5	6	1	15	3	3	2	2
167	Xanthomonas oryzae pv. oryzae MAFF 311018 [NC_007705]	11	10	27	8	8	4	4	6	3	21	3	3	2	2
168	Xylella fastidiosa 9a5c [NC_002488]	3	21	2	0	0	0	0	0	16	0	0	0	0	0
169	Xylella fastidiosa Temecula1 [NC_004556]	0	15	0	13	8	11	8	0	12	0	10	5	8	5

Total number of predicted and unconfirmed Gis	**302**	**2053**	**1509**	**2425**	**1981**	**1322**	**1192**	**115**	**583**	**636**	**1216**	**896**	**619**	**519**
Percentage of unconfirmed Gis								**38.08%**	**28.40%**	**42.15%**	**50.14%**	**45.23%**	**46.82%**	**43.54%**

7.2 Appendix B

7.2.1.1 Supplemental figures and legends

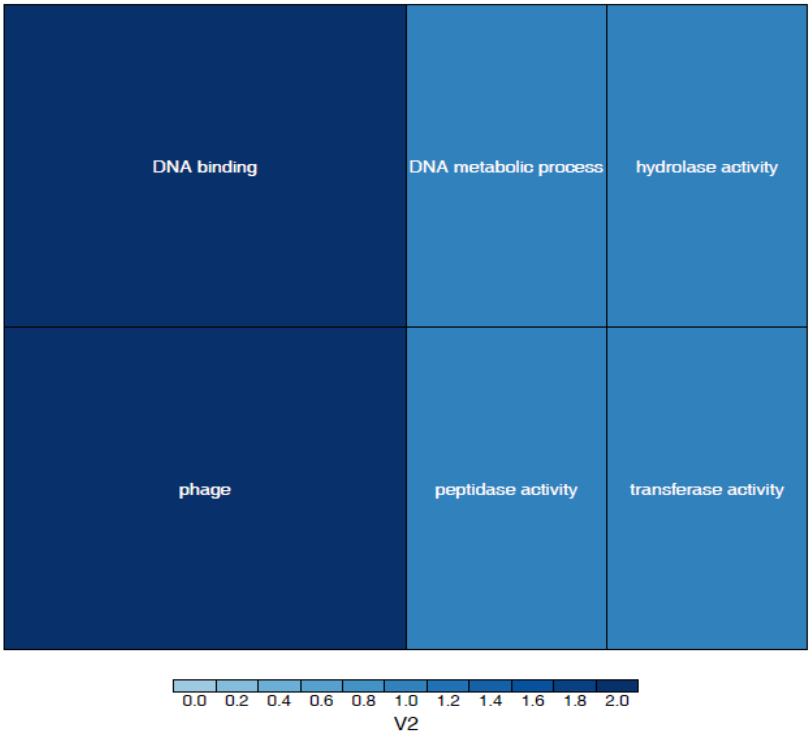
A strain-specific deletion in the genome of isolate PT22 was visualized by displaying read alignment results in the Integrative Genome Viewer: Thorvaldsdottir H, Robinson JT, Mesirov JP: Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 2012 Apr 19 doi: 10.1093/bib/bbs017.



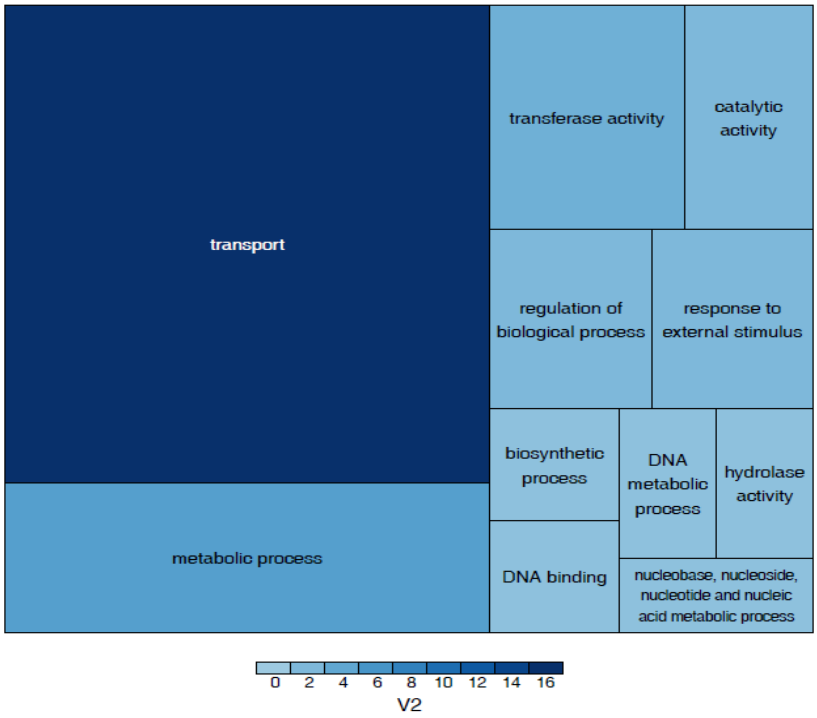
Supplementary Figure S1: Visualization of a strain-specific deletion in PT22. Alignments of strain CHA (top), PT22 (middle) and 491 (bottom) reads to the PAO1 reference (NC_002516.2) are displayed by the Integrative Genomics Viewer (Thorwaldsdottir et al., 2012). Positions 2,397,600 – 2.398,600 of the reference are shown, which are fully covered in CHA and 491 while in PT22 a block of 356 bp is uncovered.

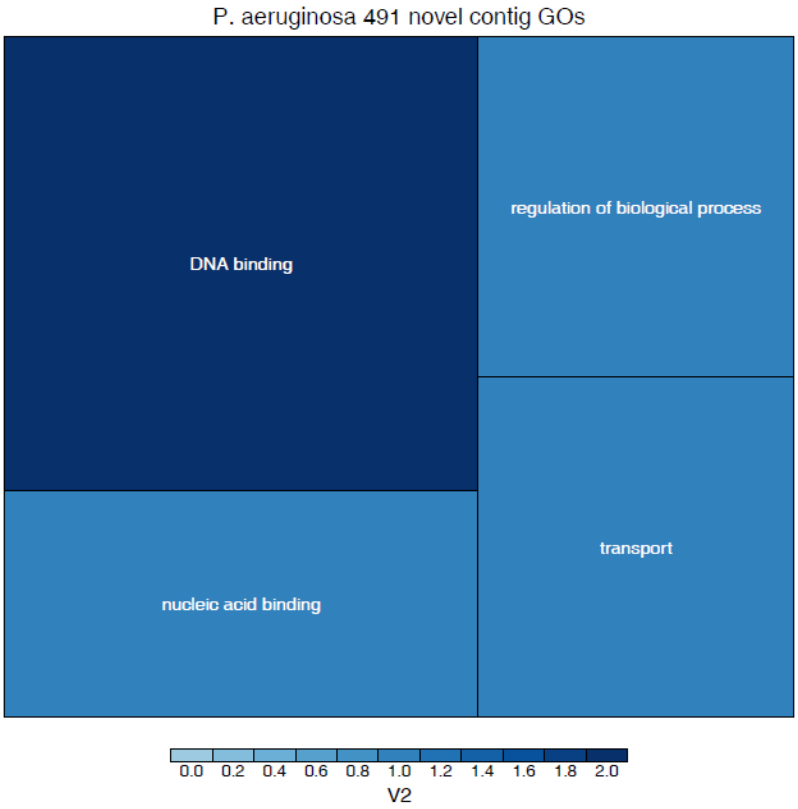
The GO gene ontology based figures for Clone CHA were created using Treemaps: Tennekes, M., de Jonge, E. (2011). Top-down data analysis with treemaps. Proceedings of the International Conference on Information Visualization Theory and Applications, IVAPP 2011, Algarve, Portugal.

P. aeruginosa CHA novel contig GOs



P. aeruginosa PT22 novel contig GOs





Supplementary Figure S2: Treemaps (Tennekes et al. 2011) of functional categories as defined by the GO gene ontology for all strain-specific contigs from the three accessory genomes. A larger and darker square indicates a higher proportion of genes of that functional category. PT22 has the most diverse unique accessory genome, while those of CHA and 491 are very restricted. Only those contigs containing genes with GO terms were included.