Pre_GI: a dynamic catalogue and set of computational tools for the ontology and stratigraphy of horizontally transferred genomic islands in bacterial genomes

by

Rian Ewald Pierneef

Submitted in partial fulfillment of the requirements for the degree

Philosophiae Doctor

in the

Centre for Bioinformatics and Computational Biology

Department of Biochemistry

School of Biological Sciences

Faculty Natural and Agricultural Science

University of Pretoria

Pretoria

January 2016

Declaration of Originality

I, Rian Ewald Pierneef 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:

Acknowledgments

My academic journey at the University of Pretoria has been arduous yet worthwhile and enriching. I wish to express my deep gratitude towards the University for the opportunity to complete my degrees and further my career at this established and innovative institution. My path has been intertwined with numerous exceptional individuals and colleagues especially that motley group of individuals calling Bioinformatics home. It has been a true privilege and honor to work with these researchers and scientists whom have made great sacrifices to continue studying and working in the academic arena.

I wish to acknowledge and thank the National Research Foundation (NRF) and the University of Pretoria for the financial support granted to me. It has enabled me to continue on my chosen path and greatly alleviated the financial burden.

To my supervisor, Prof. Oleg Reva I am grateful for the trust and patience afforded me in my research and endeavors. You are a great scientist and researcher and thank you for all the help, guidance and advise. To the head of our Bioinformatics Centre, Prof. Fourie Joubert, I am ever grateful for allowing me into your lab. You are a true scholar and gentleman and serve as a role model to me.

My parents who have tirelessly supported and carried me I thank you. Your love and understanding has lightened the weight of this journey.

To my beautiful wife, I am truly blessed to have you in my life. Thank you for your patience and encouragement throughout.

Summary

The non-genealogical transfer of genetic information between prokaryotes is a frequent and omnipresent event. The acquisition of foreign genomic segments may aid organisms in adaptation to novel or extreme habitats with these rapid evolution events phenotypically beneficial to the recipient. These regions of atypical and foreign origin are vernacularly termed islands and are identified by their unique local genomic signature or composition which differs from the global host genomic signature. The SeqWord Genomic Island Sniffer program utilizes tetranucleotide frequency patterns and statistics to identify regions of probable horizontal transfer. Optimum parameter values were determined for this compositional-based island identifier to ensure acceptable levels of false negative and false positive occurrence. Post-identification island analysis is demonstrated with the aid of the LingvoCom package available from the SeqWord project. This island identifier was furthermore compared with other existing transfer detection packages to indicate relevance and reliability. The continued identification of islands in prokaryotic genomes requires a novel and functional repository with the ability to expand as newly sequenced archaeal/bacterial genomes are available. The amalgamation of a robust database, convenient interface and island analysis tools presents a novel avenue in prokaryotic island research. The Predicted Genomic Islands database currently houses 26,744 islands identified in 2,407 archaeal/bacterial genomes and is freely available from http://pregi.bi.up.ac.za/. The database serves as an island information hub and collection of analytical tools allowing users the ability to address a myriad of horizontal transfer and island ontology research questions. Inclusion of various novel island information criteria and analytical tools may distinguish this platform from extant island databases and tools. Novel island comparison against the current content enables a rapid yet reliable tool set in the age of brisk and economically efficient genome sequencing. The collection of all island information in a single set allows for various avenues of research and stratigraphy of islands by allowing for the deconstruction and inspection of layers in archaeal/bacterial island communities and exchange between island hosts. The capability of this island garage was previewed with novel island identification results and research directions in an attempt to convey the future potential. This collection of island information and tools may prove a reliable and innovative approach in variable fields of horizontal transfer and island research with numerous applications and associations.

Contents

List of Figures

List of Tables

List of Abbreviations

ABC - ATP-binding cassette

BLAST - Blast Local Alignment Search Tool

BLASTN - Nucleotide BLAST

BLASTP - Protein BLAST

BLASTX - Protein BLAST with translated nucleotide query

bp - Nucleotide base pair

BVTs - Bacteria to vertebrate transfers

CDS - Coding sequence

CU - Codon usage

D - Distance

DNA - Deoxyribonucleic acid

DUS - DNA uptake sequences

EST - Expressed sequence tags

FASTA - Text-based format for representing either nucleotide or peptide sequences, where nucleotides or amino acids are represented by single-letter codes

FNR - False negative rate

FPR - False positive rate

FTP - File transfer protocol

G+C - Guanine and cytosine nucleotide bases

GenBank - Rich format for storing sequences and associated annotations

GI(s) - Genomic island(s)

GRV - Global relative variance

GUI - Graphical user interface

HGT - Horizontal gene transfer

HMM - Hidden Markov Models

HT - Horizontal transfer

HUS - Hemolytic uremic syndrome

ICE - Integrative and conjugative elements

IHGSC - International Human Genome Sequencing Consortium

IVOMs - Interpolated Variable Order Motifs

kbp - Nucleotide kilobase pair

k-mer - Word size of length k

MCL - Markov clustering algorithm

MDS - Multidimensional scaling

MGE - Mobile genetic element

MpF - Mating-pair formation

NCBI - National Center for Biotechnology Information

NGS - Next-generation sequencing

OU - Oligonucleotide usage

OUP - Oligonucleotide usage pattern

OUV - Oligonucleotide usage variance

PAI - Pathogenicity island

PS - Pattern skew

RDMS - Relational database management system

REI - Resistance island

RNA - Ribonucleic acid

rrn - Ribosomal RNA operon

rRNA - Ribosomal ribonucleic acid

RV - Relative variance

SET - Serial Endosymbiotic Theory

SVG - Scalable vector graphics

SWGIS - SeqWord Genomic Island Sniffer

T-DNA - Transfer DNA

tRNA - Transfer RNA

tmRNA - Transfer messenger RNA

Ti - Tumor inducing

V - Variance

1 Chapter 1: Introduction

If you don't like bacteria, you're on the wrong planet. Stewart Brand

Horizontal (lateral) transfer (HT) is the non-genealogical transfer of genetic material from one organism to another which may cross specie and domain borders. HT is an evolutionary event that allows recipient organisms to adapt to fluctuating environments and ecological pressures [3, 39]. Prokaryotes reproduce asexually and as such lack genetic recombination and rapid alteration to sustain in demanding and wavering environments. The insertion of mobile genetic elements (MGE) vernacularly termed genomic islands (islands, GIs) are essential in the evolution of prokaryotes. Persistence and fixation of these islands in prokaryotic genomes suggest an advantage was conferred to the recipient host [137]. The majority of horizontally acquired genetic material may cause a detrimental effect on the chromosome of the recipient organism and therefore be discarded in a similar fashion than deleterious mutations with the neutral acquisitions survival dependent on chance events [149]. Recent pan-genomics studies infer the continuous sampling and/or shuffling of microbial genetic material in contrast to slow, progressive changes [69]. This is evident in the uprising of drug resistant and pathogenic bacterial strains [78]. The advent of a single HT event may alter the phenotype and virulence aspect of a microbe leading to a cascade of further accumulation of virulence elements. This reconciles with the hypothesis of genetic capitalism, which profess that a successful integration and selection of foreign genetic material increases the extent of possible HT events in the future [17]. The existence of HT events and islands is now known and accepted, still this was not always the case. HT was deemed an oddity and of no consequence to evolution with the only acceptance of HT being the flow of genetic material between an endosymbiotic organelle (mitochondria, chloroplast) and the nuclear genome [39]. Ironically the endosymbiotic theory would entail the genesis of eukaryotes and HT between organisms due to environmental dependencies.

1.1 The Original Horizontal Transfer Event

1.1.1 Symbiogenesis or Endosymbiotic Theory

The genesis of eukaryotic cells from prokaryotic cells is a principle step in the understanding of life and evolution. The formation of complex entities by means of addition of less complex entities is a cornerstone in evolution and advance of life. In its most primitive form the endosymbiotic theory proposes that mitochondria, plastids and other organelles are previously free-living bacteria that were engulfed by a cell as endosymbionts. This event is postulated as roughly 1,5 billion years ago with the SAR11 clade of Rickettsiales,

a member of the α-proteobacteria, the mitochondria and cyanobacteria as chloroplasts [86].

Currently the theory of symbiogenesis and origin of plastids is widely accepted and unchallenged due to the vast amount of comparative data, yet this was not always the case. Symbiogenesis was accepted for the first part of the century, but this changed during the First World War. The theory of *de novo* plastid evolution in non-plastid bearing cells or autogenous origin was favored above endosymbioses well into the 1970s [124].

The first formally defined version of the endosymbiotic theory was in 1905 by a Russian botanist Konstantin Sergejewiz Merezkovskij affiliated with a university in Kazan, the capital of the Kazan province in the Russia empire. His paper was the combination of three principles that were known at the time. These lines of evidence was the known principle of symbiosis, the prior findings of plastids that proliferate by division and the novel evidence obtained by the physiological comparison of plastids and cyanobacteria [124]. The article is the culmination of various known events in a single and tangible theorem. Merezkovskij clearly articulated the evidence in a simple yet astounding theory given the limited technology and data available in 1905. He coherently integrated various observations and deductions to support his views. At the time it was believed that chromatophores, pigment-containing cells, were organs or structures that had differentiated out of the colorless plasma of the cell body. Due to the greening of colorless parts of a plant when in contact with light it was concluded that chlorophyll originates *de novo* and because chlorophyll arises *de novo* the chromatophore, carrier of the chlorophyll, also arises *de novo* [127]. In 1885 it was demonstrated that although chlorophyll does arise *de novo* in certain cases, the carriers of the chlorophyll, plastids, are present as colorless leucoplasts and never arise *de novo* [142]. Merezkovskij then clearly explains why chromatophores are not organs and outlines why they are rather symbionts:

- *The continuity of chromatophores.* Cromatophores do not arise *de novo* but arise through the division of pre-existing plastids, which in turn arise from pre-existing plastids and so forth. Thus the first chromatophore migrated into a colorless organism. Thereby chromatophores are foreign bodies or symbionts.

- *Chromatophores are highly independent of the nucleus.* Chromatophores will continue to live even after a portion of a green plant cell is anucleated. They grow, multiply, assimilate CO_2 and produce synthetic starch grains in the absence of a nucleus and as such they are independent of the nucleus, behaving like independent organisms rather than organs [117].

- *The complete analogy of chromatophores and zoochlorellae. Zoochlorellae* is a genus of algae and is used to refer to green algae living symbiotically within an invertebrate

or protozoan. The only difference between chromatophores and *Zoochorellae* is that *Zoochorellae* can live and divide outside a host cell.

- *There are organisms that we can regard as free-living chromatophores.* Lower forms of Cyanophyceae may be viewed as such as there are only slight differences between a chromatophore and an *Aphanocapsa* or *Microcystis*. This similarity infers the likelihood that chromatophores are Cyanophyceae that invaded the plasma.

- *Cyanophytes actually live as symbionts in cell protoplasma.* Chromatophores are not organs, but foreign organisms that migrated into a cell. Chromatophores display similar behavior to *Zoochlorellae*, which are foreign organisms. There are extant organisms that are in all probability living antecedents of chromatophores. The invasion of the cytoplasm by cyanophytes and continued symbiotic existence is the only piece of the puzzle that needs to be revealed. The blue-green bodies contained by the flagellate *Cyanomonas americana* and various other well-known examples indicate the ease with which cyanophytes enter a symbiotic existence even in the presence of cell walls.

Thus the endosymbiont theory was formulated in 1905 by a Russian botanist with the focus on chromatophores, chloroplasts and cyanobacteria, equating plastids to "little green slaves" laboring for their hosts to produce nutrients from light [126]. The only literature to predate Merezkovskij was a casual footnote by Andreas Schimper [124]. Schimper had observed in 1883 that the division of chloroplasts resembled that of cyanobacteria and proposed in a footnote that green plants were the result of a symbiotic union of two organisms.

The endosymbiotic theory was extended to mitochondria in 1927 by Ivan E. Wallin in his book *Symbionticism and the Origin of Species*. Wallin concluded that the ubiquitous nature of bacteria-derived mitochondria indicated the necessity of bacterial symbionts as building blocks of evolutionary change and speciation [2] with mitochondria existing as bacterial organisms symbiotically combined with the tissues of higher organisms [153]. Wallin's theory was universally rejected and the leading American cell biologist, E. B. Wilson, remarked that the idea was "too fantastic for present mention in polite biological society" [57].

The endosymbiotic theory then entered a hiatus for roughly fifty years due to various factors including the discovery of chromosomal genes and the start of population genetics [2]. The theories of Merezkovskij and Wallin were also left stagnant due to the assumption that mitochondria and chloroplasts do not contain DNA. In the 1960s there was a re-suscitation of the theory due to electron microscopic comparisons between cyanobacteria and chloroplasts [138] and the confirmation of mitochondrial and plastid DNA [145].

In 1966 an American biologist, Lynn Margulis, wrote a theoretical paper "On the origin of mitosing cells" [141], that was rejected by 15 journals but eventually published in 1967. Today this paper is acknowledged as a bastion of the endosymbiotic theory. It presented a theory on the origin of discontinuity between eukaryotic and prokaryotic and the genesis of mitochondria and photosynthetic plastids from free-living cells with the eukaryotic cell the result of the evolution of ancient symbiosis [141]. She acknowledges that the paper did not present a novel theory. She formulated previous work to be consistent with data on biochemistry and cytology of the day and concluded that many aspects of the theory were verifiable by modern molecular biology techniques. The first step in the endosymbiotic theory was the oversupply of oxygen in the atmosphere. Survival for a heterotrophic anaerobe in this oxygen-containing atmosphere was possible by the ingestion of an aerobic prokaryotic microbe. Margulis and her son, Dorion Sagan, commented that "Life did not take over the globe by combat, but by networking". Margulis further proposed the Serial Endosymbiotic Theory (SET), represented in Figure 1, whereby eukaryotic cells evolved through a series of discrete symbiotic partnerships with various prokaryotic cells whereby mitochondria, chloroplasts and flagella evolved [57].

The contribution of Margulis, guided by previous researchers, is paramount in the field of symbiogenesis. Richard Dawkins stated "I greatly admire Lynn Margulis's sheer courage and stamina in sticking by the endosymbiosis theory, and carrying it through from being an unorthodoxy to an orthodoxy. I'm referring to the theory that the eukaryotic cell is a symbiotic union of primitive prokaryotic cells. This is one of the great achievements of twentieth-century evolutionary biology, and I greatly admire her for it" [4].

The endosymbiotic theory itself has evolved overtime with various other earlier theories for the origin of eukaryotic cells listed below [111]:

- Independent origin of archaebacteria, eubacteria and eukaryotes from a universal ancestor by C. R. Woese (1981, 1987).
- Evolution of the nucleus due to engulfment of an archaebacterial species, Lake *et. al* (1982, 1994).
- Eukaryotic cell evolution from an intermediate between archaebacteria and Gram-positive bacteria by T. Cavalier-Smith (1987).
- The eukaryotic nuclear genome a chimera which was formed by primary fusion of an archaebacterium and a bacteria by W. Zillig (1991).
- The eocyte group was indicated as the closest relatives to eukaryotes by M. C. Rivera and J. A. Lake (1992).

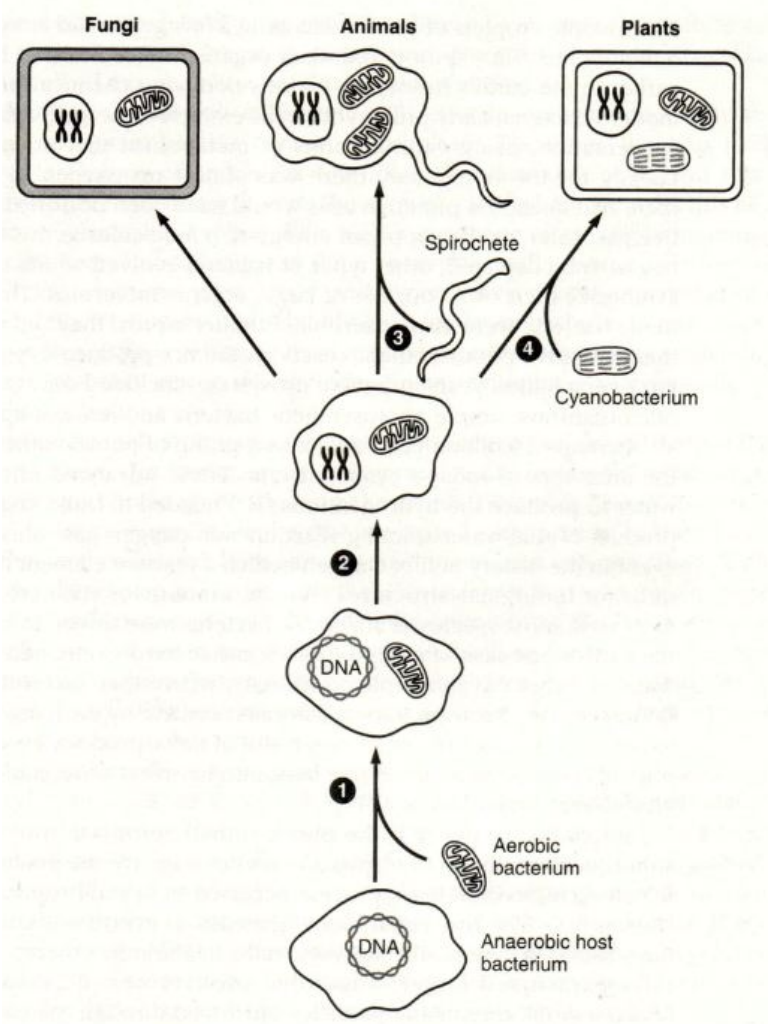


Figure 1: Evolution of eukaryotic cells by a series of discrete endosymbiotic events: 1. Mitochondria evolve; 2. Nucleus evolves from simpler DNA molecule; 3. Flagella evolve from spirochetes; 4. Chloroplasts from free-living cyanobacteria. Cell walls evolve independently [57].

The endosymbiotic theory is a form of saltational or leap evolution where there is a sudden change from one generation to the next. After the initial leap the endosymbionts transferred portions of their own DNA to that of the host by means of endosymbiotic gene transfer. As such the endosymbiotic theory unveils the very first instance of inter-domain horizontal gene transfer.

1.1.2 Alternate Endosymbiotic Theory: The Third Wheel

Why has the primary endosymbiosis event or acquisition of primary plastids not been repeated many times over? What was the initial selective pressure that directed the acquisition and retention of the cyanobacterium in the host? Only one case of a recent primary plastid endosymbiosis was identified, in the photosynthetic amoeba *Paulinella*

chromatophora [36]. The Archaeplastida, also called Plantae, are the founding lineage of photosynthetic eukaryotes and their nuclear genomes serve as a basin for primary endosymbiosis research. The large portion of cyanobacterium-derived components in these genomes arose from endosymbiotic gene transfer resulting in the movement of copious cyanobiont genes to the host chromosome with a second contributor of a large portion of foreign genes to the Archaeplastida identified as the obligate intracellular bacterium *Clamydia* [87]. In the tripartite endosymbiotic hypothesis the cyanobiont provided energy for a host cell whose energy resources where being drained by a intracellular chlamydial pathogen with the clamydial proteins themselves converting the cyanobacterial metabolites into host energy stocks and is referred to as the 'ménage à trois' hypothesis (Figure 2) [31].

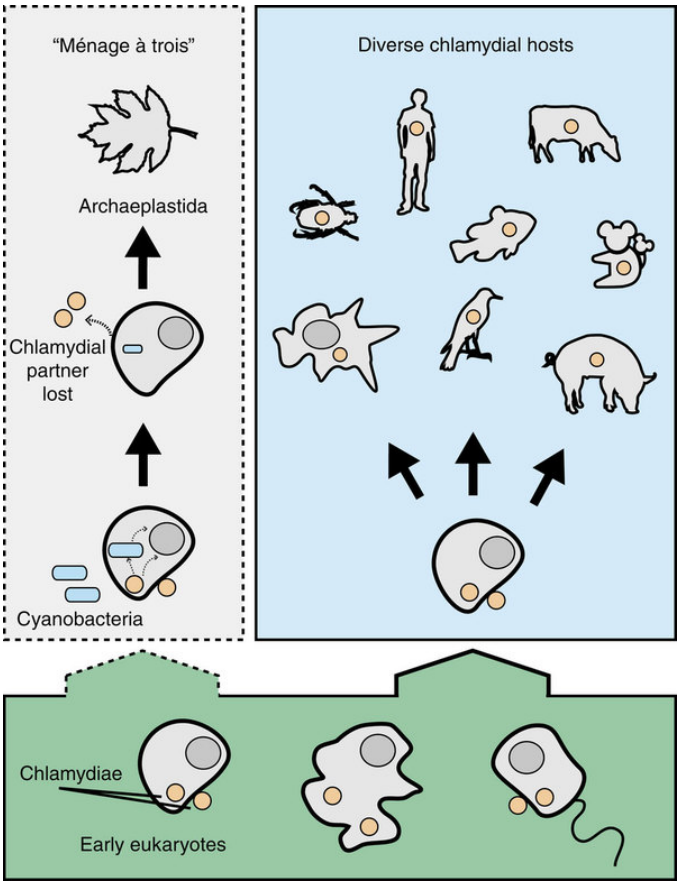


Figure 2: The 'ménage à trois' hypothesis and diverse chlamydial hosts [31].

According to this hypothesis the clamydial partner converted host glucose-1-phosphate to the bacterial metabolite ADP-glucose which was then polymerized to glycogen and processed for import. The engulfed cyanobacterium would offer the host immediate relief by means of ADP-glucose which is a by-product of cyanobacterium metabolism and thus prevent further host energy depletion. The host energy depletion would provided the initial selective pressure for the uptake and consolidation of the cyanobacterium in the

host and provide a lasting symbiosis by metabolically linking host and cyanobiont [31]. The long term amalgamation of host, cyanobiont and clamydial enzyme functions would lead to metabolic stabilization. Thereafter subsequent evolution and extensive HT would form the true photoautotrophic eukaryote that would be the direct ancestor of current archaeplastidial lineages and render the clamydial partner dispensable, leading to the loss of the clamydial symbiont [87].

The 'ménage à trois' hypothesis by Ball et al. was weighed by Domman et al. and found wanting. The elementary methods used by Ball et al. did not adequately model sequence evolution with better-fitting models not supporting the clamydial partner hypothesis [31]. Domman et al. concluded that in the absence of cytological evidence or support from more robust phylogenetic models the 'ménage à trois' hypothesis is at this point in time only a fantasy.

1.1.3 Secondary and Tertiary Endosymbiosis: Matryoshka dolls

The endosymbiotic theory or symbiogenesis is classified as primary endosymbiosis, but not all eukaryotes obtained the advantage of photosynthesis by engulfing a cyanobacterium. Secondary endosymbiosis is the process whereby an eukaryote acquired the ability of photosynthesis by engulfing an eukaryotic algae (product of primary endosymbiosis) and retaining the plastid (chloroplast) [133]. This process produces plastids that are surrounded by three or four membranes unlike the two membranes of primary plastids [146]. The number of membranes in secondary plastids are plausibly as a result of the mechanism by which they were engulfed. Phagocytosis will probably produce four-membraned structures and myzocytosis, which does not include the ingestion of the prey cell membrane, will likely lead to three-membrane structures [11]. These events formed multiple independent groups of photosynthetic organisms including two groups of algae, the cryptomonads and chlorarachniophytes that still contain a relict nucleus, a nucleomorph, from the originally engulfed algae [133]. Phylogenetic analysis and sequencing of the nucleomorph chromosomes indicated that the nucleomorph originated from a red algae in the cryptomonads and a green algae in the chlorarachniophytes [146]. The haptophytes and heterokonts do not contain a nucleomorph and display a more reduced secondary symbiont with plastids surrounded by four membranes originating from red algae [146]. The primary reasoning behind secondary endosymbiosis would have been the availability of the photosynthetic energy production center but after time the pathways and centers between host and symbiont would have become inter-twinned ensuring that the plastid became indispensable even when not capable of their original function. This is evident in colorless primary plastids and the non-photosynthetic apicoplasts found in Apicomplexa. The loss of photosynthetic ability by the plastid does not lead to elimination of the plastid in the host.

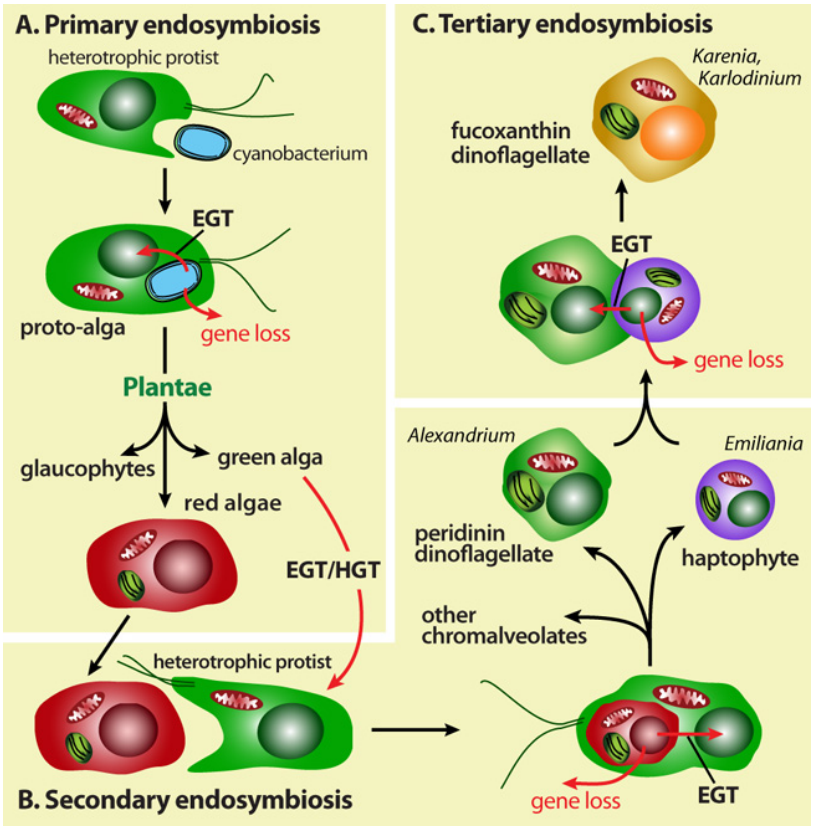


Figure 3: Primary, secondary and tertiary endosymbiosis [5].

Tertiary endosymbiosis is the next step in the hierarchical eukaryote dinner. The uptake of a photosynthetic symbiont that originated from secondary symbiosis by an eukarotic cell is defined as tertiary endosymbiosis as is particularly evident in some dinoflagellate species [146]. Within this group the existing red algae secondary plastid is replaced with another plastid of secondary origin leading to the exchange of peridinin as the main carotenoid for fluxoxanthin [30].

The nested Russian doll process of secondary and tertiary endosymbiosis provides eukaryotes with novel genes and thereby increases eukaryotic biodiversity.

1.2 Prokaryote Horizontal Transfer

Currently HT between related and distant species is accepted as a regular occurrence in prokaryote evolution and adaptation, yet this was not always the custom. Originally it was believed that micro-organisms evolved clonally with little or no genetic material exchanged [82]. The traditional dogmas regarding HT was challenged and revised by means of experimental and comparative analysis. In 1951 HT between avirulent and virulent *Corynebacterium diphtheriae* was described and concluded that the virulence was dependent on various factors including the degree of association with bacteriophages

[107]. Cross-species HT between a resistant *Escherichia coli* and a *Shigella* was induced by mixed culture and documented in 1960 [99]. In 1985, Michael Syvane proposed that the degree and influence of HT was much greater than what was generally accepted but ceded that such a theory was not proposed earlier due to the lack of known mechanisms for HT [148].

The importance of HT in bacterial evolution has been elevated to such a degree that various bacteriologists are questioning the existence of bacterial species [137]. The persistence and fixation of these horizontally transferred genes indicates that a selective advantage is conferred on recipient organisms [137]. Estimates for horizontally acquired DNA range from 0.5% to 25% [151], and 1.6% to 32.6% [3], which are vast sizes considering the size of a typical bacterial genome. The source and amount of HT has been linked to an organism's lifestyle with events following a different classification systems ranging from new genes, paralogs from existing genes and displacement from orthologs from another linage [39]. The role of HT in microbial evolution is thus playing a more vital role than previously thought [3]. Recent pan-genome studies strongly suggest that microbial genomes are continuously sampling and/or shuffling their genomic information, rather than undergoing slow, progressive changes [69], which is practically evident in the uprising of drug resistant strains. Furthermore, a single HT event may alter the phenotype and virulence characteristics, transforming a benign organism into a pathogen [17]. The hypothesis of genetic capitalism, which states that successful integration and selection of a foreign genetic element increases the number of possible genetic transfer events in the future, is in agreement to the uptake of multiple islands and modules of resistance determinants [17].

The composition of islands include more novel genes than the rest of the genome, indicating an adaptive and auxiliary function [72]. The number of hypothetical proteins in islands are often high and encoded functions are highly specific, thus enhancing the fitness of the species [128]. The complication of island investigations are amplified by the large spectrum of varieties in terms of genetic organization and functionality [72], and the variation in gene family gain and loss rate over gene families [8].

1.3 Mechanisms, Incorporation and Barriers of Prokaryotic Horizontal Transfer

Prokaryote HT is achieved by three principal mechanisms:

- Transformation. The procurement of naked DNA from the environment by a genetically competent organism. Transformation has been described as early as 1928 when Griffith reported the advent of a virulent organism by mixing heat-killed,

virulent pneumococci with live, non-virulent bacteria [109]. This was expanded in 1944 by Avery, MacLeod and McCarthy through the "transforming principle" which concluded that material from a deceased bacteria conveyed information to living, non-virulent bacteria and as such confirmed that the material transporting the information is DNA [79]. The conserved capacity of bacteria to obtain foreign DNA by transformation indicates a functionally important genetic trait [149]. Transformation is attributed to 3 non-mutually exclusive requirements that necessitate the uptake of foreign DNA [7]:

1. Diversity. Novel traits and functions may be conferred to the recipient.
2. Repair. DNA from closely related organisms might aid the repair of DNA damage.
3. Food. Carbon, nitrogen and phosphorous can be sourced from DNA.

The process of transformation requires the recipient or host to be physiologically "competent" or receptive to the uptake of extracellular DNA [52]. This time-limited state involves 20 to 50 proteins in response to changing environmental conditions, *e.g.* growth conditions, nutrient access, cell density or starvation [149]. The import of foreign, extracellular DNA is a complex task with various physical, time and physiological barriers. The first essential ingredient for transformation is the availability of stable, extracellular DNA in an environment. Extracellular DNA is provided on a continual basis to the environment by decomposing or disrupted cells, viral particles, excretion by living cells and active secretion by viable organisms [7, 149]. The degradation of extracellular DNA is affected by environmental conditions with stability of extracellular DNA influencing the bacterial exposure time and therefore the transformation frequency, *i.e.* the number of bacteria with HT DNA in relation to the total number of bacteria exposed, per unit time [149]. Gram-negative bacteria contain 3 physical barriers, *i.e.* outer membrane, cell wall and cytoplasmic membrane, and Gram-positive 2 physical barriers, *i.e.* cell wall and cytoplasmic membrane for extracellular DNA to cross. Only a single strand of DNA is adequately imported into the cytoplasm with the other strand degraded and released into the extracellular environment (Gram-positive) or the periplasmic space (Gram-negative) [7], illustrated in Figure 4.

Translocation across the membrane is not uniform across all species with *Neisseria* species, *Haemophilus influenza* and *Actinobacillus actinomycetemcomitans* selective in the DNA to be translocated, yet other species indifferent to the sequence ingested [149]. These sequence motifs are designated DNA uptake sequences (DUS) or uptake signal sequences (USS) and have been identified as 5'-GCCGTCTGAA-3' for *Neisseria* sp. and 5'-AAGTGCGGT-3' for both *Haemophilus influenza* and

Actinobacillus actinomycetemcomitans [7]. Sequence-specific binding receptors in the recipients are yet to be identified. Transformation in these sequence dependent recipients may be highly specific and species restricted [149]. In sequence-specific and non-specific organisms the single strand of DNA is linearly absorbed into the cytosolic space with a free end mandatory to start the transport process [52].

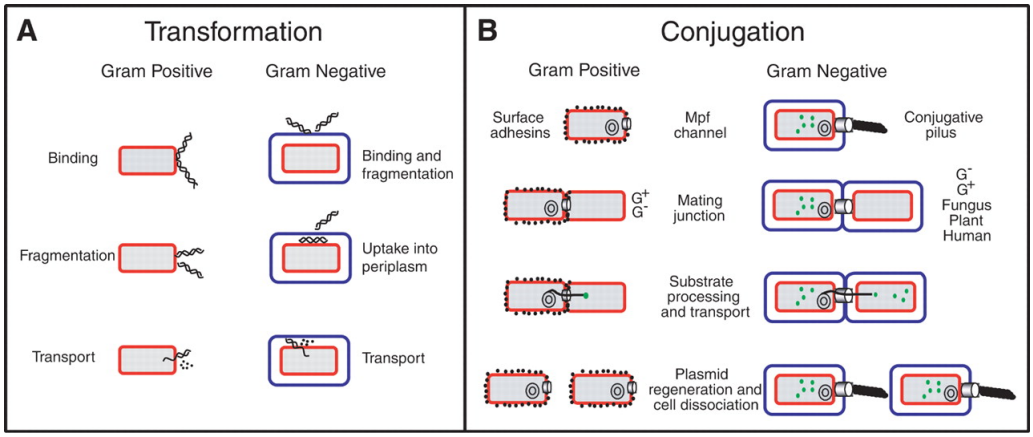


Figure 4: DNA conversion during Transformation and Conjunction [52].

The uptake of stable, extracellular DNA in a competent host occurs briskly when determined *in vitro*, ranging from 60 to 100 bp per second [149]. In Gram-negative bacteria the outer-membrane is probably traversed with the aid of secretins. Secretins are outer-membrane proteins involved in extrusion, secretion and transformation in Gram-negative-organisms by forming stable multimeric structures consisting of 12 or 14 subunits [7]. These donut-like multimers are components of the type IV pilus (T4P) and located in the outer-membrane with an aqueous central cavity which is 6.5 nm in diameter with ample space for DNA or a DNA-protein complex to pass through [52]. DNA presumably reaches the host periplasm by means of these secretin donut stacks, yet direct evidence is still lacking. Competent bacteria enroll systems akin to the T4P and type II secretion system (T2SS) to absorb DNA. There is a correlation between competence in bacteria with T4P and piliation but the narrow channel and hydrophobicity of pili eliminate them as a pipeline for transformation. Competence pseudopilus, a T4P like structure, are suggested to be participants in DNA transport during transformation by chaperoning foreign DNA to the transport machinery located at the cytoplasmic membrane [7]. In non-T4P organisms there exists machinery dedicated to pseudopilus construction. Pseodopilus formation in organisms with T4P would occur by means of the same components as the pilus and thus explain the correlation between pilus formation and competence [7]. Comparison on pseudopilin lengths to secretion-pilus and type IV pili indicate sizes ample to span the periplasm and cell wall (~55 nm) [52]. Pseudopilins may retract or pull back DNA from the bacterial surface to the cy-

tostolic membrane receptor to be transported across the cytostolic membrane [53], arbitrate transition of DNA across the secretin channel in Gram-negative bacteria or have a more passive position by forming a bridge across the cell wall enabling foreign DNA to reach the cytostolic membrane receptor [7]. Transport into the cytosol entails the employment of essentially two large polytopic membrane proteins (ComEC and ComEA) and a third membrane-bound ATPase (ComFA) in the case of Gram-positive organisms [52]. The Com prefix of these proteins relates to the required state of "competence" by organisms to transform extracellular DNA. The cytoplasmic DNA receptor ComEA functions in both binding and transport of DNA with ComEC essential for DNA transport by forming an aqueous channel across the cytoplasmic membrane [7]. The ATP-binding protein employed by Gram-positive bacteria may assist in the translocation of DNA through the ComEC channel [52] and aid in other translocation tasks such as the unwinding of double-stranded DNA or the gating of the ComEC channel [7]. The single-stranded DNA now present in the cytoplasm is to be integrated in the bacterial host genome. This is facilitated by a RecA-dependent process and is reliant on the interaction between the freshly translocated DNA and the cytoplasmic proteins [7]. There are various forms of integration between the single-stranded newly acquired DNA and the bacterial host genome. Homologous recombination requires lengthy segments (25 to 200 bp) of high sequence similarity to initiate pairing and strand exchange [149] and therefore is prevalent between closely related individuals as seen in the trading of genetic information between *Escherichia coli* strains [118]. Due to the similarity to the host genome these homologous recombination transfer events are difficult to distinguish from simple mutations and requires larger data sets for comparative analysis. Illegitimate (non-homologous) recombination is sequence independent and transfer may occur between distantly related individuals. The prevalence of illegitimate combinations is increased by the presence of a segment of homologous sequence in the donor DNA to the recipient which serves as a recombinational anchor for the RecA-dependent transfer [10]. Homologous and illegitimate recombination may be viewed as replacing transfers (Figure 5 A) from similar or distant lineages. Additive integration (Figure 5 B) entails non-replacing transfers that add substantial genetic material to the recipient genome. This includes recombination between two circular molecules or recombination between a circular molecule and the host chromosome due to single crossovers at small segments of high similarity [149]. The underlying structure of replacing and additive transfers are similar as seen in additive recombinations associated with replacing recombination at flanking regions [85]. Linear DNA integration into chromosomal DNA may furthermore lead to additive integration when exchange extends beyond segments of homology or sequence similarity resulting in substitution of sequences or addition of sequences and is defined as

homology-facilitated illegitimate recombination [149].

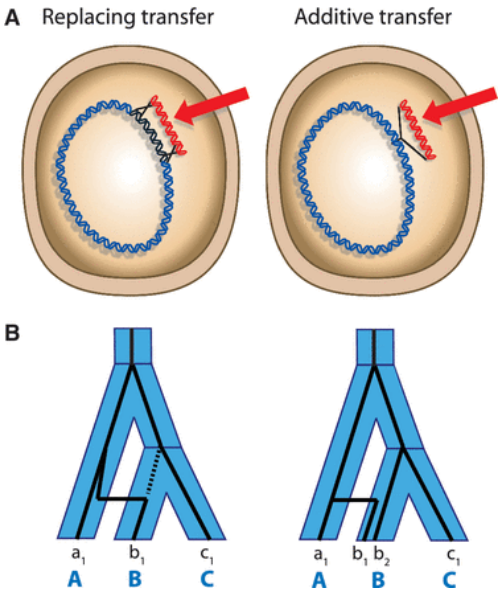


Figure 5: Replacing and Additive Transfer. A) Integration may follow replacing or additive transfer. B) Phylogenetic analysis reveals a clear distinction between replacing and additive transfers. Replacing transfers lead to loss events whereas in additive transfers no loss event is evident [85].

- Conjugation. The unidirectional and conservative transfer of DNA mediated by MGE, *e.g.* plasmids, transposons, after cellular contact. Conjugative transfer is regarded as having the highest potential capacity and evolutionary influence of the HT mechanisms [12]. In 1946 Lederberg and Tatum, unaware of the existence of plasmids, described conjugation for the first time when studying 2 *Escherichia coli* strains with different nutritional requirements [120]. Conjugation relies on independently replicating genetic elements (conjugative plasmids) or chromosomally integrative and conjugative elements (ICEs) as well as conjugative transposons [68]. In contrast to plasmids, it is currently believed ICEs are not able to survive in an extra-chromosomal state due to the inability to replicate autonomously [155]. These MGE serve as a "buffet" for bacterial populations to obtain traits, variation and recombination in varying environments with a greater probability of fixation than other transfer mechanisms due to the protracted retention time in the cell [91]. The relatively small size of these elements encourage transfer as larger elements, *e.g.* a whole chromosome, will require more than an hour to be transferred during which the interbacterial junction would have disintegrated and the process halted [149]. The donor cell advances transfer by synthesizing the multi-protein equipment connecting the donor and recipient cell and often processing the DNA before transfer to single-stranded with double-strand conversion accomplished by the replication machinery of the recipient cell [155]. Processing and preparation of DNA to be

transferred is widely conserved between bacterial species and is dependent on a relaxase and some auxiliary factors [52]. The three central steps in conjugation are mating-pair formation (MpF) followed by a signalling event or green light for transfer to proceed and finally the transfer of DNA [68]. Conjugation equipment consists of a cell-envelope crossing translocation channel with a pilus for Gram-negative bacteria and surface-localized protein adhesins for Gram-positive bacteria, as described in Figure 4 [52]. The Gram-negative conjugative pilus (mating-pair apparatus) is assembled by the T4SS wherein a coupling protein joins a trans-envelope protein complex to the plasmid's relaxasome which is located at the plasmid origin of transfer (*oriT*) [68]. The relaxasome is a protein-DNA complex at the *oriT* that nicks DNA when proteins are denatured chemically or cleaved proteolytically [149].

- Transduction. Transfer of DNA by means of a bacteriophage. The "one-night stand" of HT, transduction has been likened to sexual reproduction in bacteria [121]. Transduction was demonstrated in 1951 by Lederberg and Zinder in *Salmonella typhimurium* with a filter to prevent cell contact and conjugation [63]. This transfer mechanism is specific as bacteriophages have a narrow host range and requires the accidental packaging of host DNA into bacteriophage particles during replication [91]. Phages are defined as virulent or temperate. Virulent phages follow an exclusively lytic infectious cycle and temperate phages may follow a lytic infectious cycle but prefer to institute a lysogenic cycle as a prophage. Transduction is divided into generalized and specialized. Generalized transduction entails the transfer of genetic material by a phage to any segment of the chromosome whereas specialized transduction involves the transfer to restricted portions of the chromosome. Generalized transducing phages randomly acquire bacterial DNA after lysis of the donor cell by wrongfully absorbing the host bacterial DNA into the phage head [114]. The serendipitous genetic material may be infused in a recipient cell by means of recombination after the phage attaches to the recipient bacteria and inserts the phage contents. Specialized transducing phages amass genetic material due to the defective disengagement of a prophage from a bacterial chromosome leading to phage containing specific bacterial genetic material. Phage lambda (λ) is an excellent example of a specialized transducing phage which always inserts between the *gal* and *bio* region of the host chromosome and as such phage λ only transduces *gal* and *bio* genes [34].

The method of transfer is of no consequence if the transferred island is not securely inherited by the recipient organism. Figure 6 depicts the 5 steps that are essential in the stable transfer of islands by means of HT [143]:

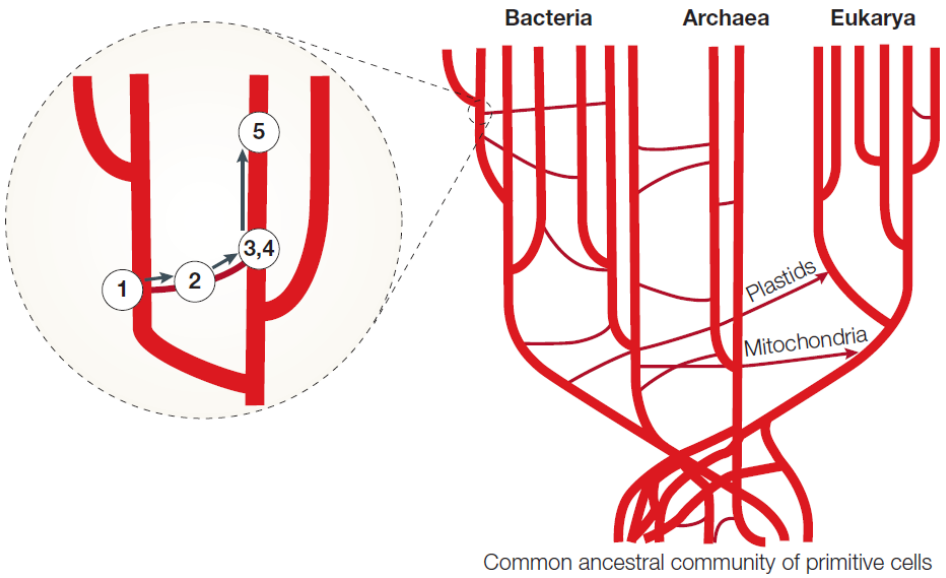


Figure 6: 5 Steps leading to the HT and durable incorporation of foreign genetic material [143].

1. Preparation of island for transfer. In order for an island to be successfully delivered it has to be correctly packaged in the current host. This is achieved by wrapping the island in phage particles, replication from an origin for transfer by conjunction, integron assembly or the static release following growth or cell death.

2. Transfer of island. This is achieved by the methods described above.

3. Entry of island in host.

4. Incorporation of island in host. After packaging, delivery and acknowledging receipt, the island is to be unpacked and stored. The island may be established in the recipient as a self-replicating element or recombination with/transposition into the recipient host chromosome.

5. Stable inheritance and continuation of island in new host.

Integration into the host chromosome is established by means of the following mechanisms [13, 65]:

- Homologous recombination. This is mechanism is favored between related taxa and is proposed to be the most important method of incorporation between closely related lineages. Homologous recombination will typically replace a similar sequence rather than introducing novel DNA in a recipient host.

- Persistence as an episome or integrating plasmid. These plasmids are integrated in the host chromosome, yet inevitably they function as an independent plasmid molecule in the host at some stage.

- Integration by MGE. This method may be important in HT between distantly related species and is personified by integrons, a unique class of MGE, that utilizes site specific integration.
- Illegitimate incorporation by means of an accidental double-strand break repair.

HT is not a free-flowing process and is subject to various barriers and roles played by the donor and recipient in addition to environmental and physical restraints. Transformation entails an active role by the recipient in contrast to conjugation where the donor assumes a positive role and the recipient a negative or limiting role in order to curb HT [149]:

- Surface exclusion. Prospective recipients restrict the frequency of conjugative transfer if in possession of a similar or related plasmid.
- Restriction. Restriction endonucleases divide foreign DNA into smaller fragments. Recipients with a restriction system show a reduction in the frequency of transconjugants from plasmids who are vulnerable. This barrier is breached to a certain extent as small plasmids and individual genes may not present all the restriction sites and as such evade cleavage. Furthermore, conjugative and transformative transfers involve single-stranded DNA which is more capable of averting the restriction system.
- Plasmid replication and establishment barriers. The ability of a plasmid to replicate and thus avert a replicon such as a chromosome eases HT to a recipient genome. Plasmids are varied in their host range. The size of the host range may be due to the replication proteins present and the absence of lagging-strand synthesis. The aggregation of single-stranded DNA-replication intermediates due to the lack of lagging-strand synthesis inhibits the amount of genetic material that plasmids may amass and thus transfer before it becomes recombinationally unstable. Narrow host range plasmids have been found wanting in terms of presence and ability of replication proteins. Mutations in the plasmid replication proteins alter the host range, suggesting that if a selective pressure was present the plasmid may increase its host range by simple mutations. This is supported by the difference in host ranges for plasmids within the same family, illustrating that the selective history of the plasmid influences the host range.

1.4 Interdomain Horizontal Transfer

1.4.1 Methods of Interdomain Transfer

HT to eukaryotes is achieved by similar methods as between prokaryotes, illustrated in Figure 7.

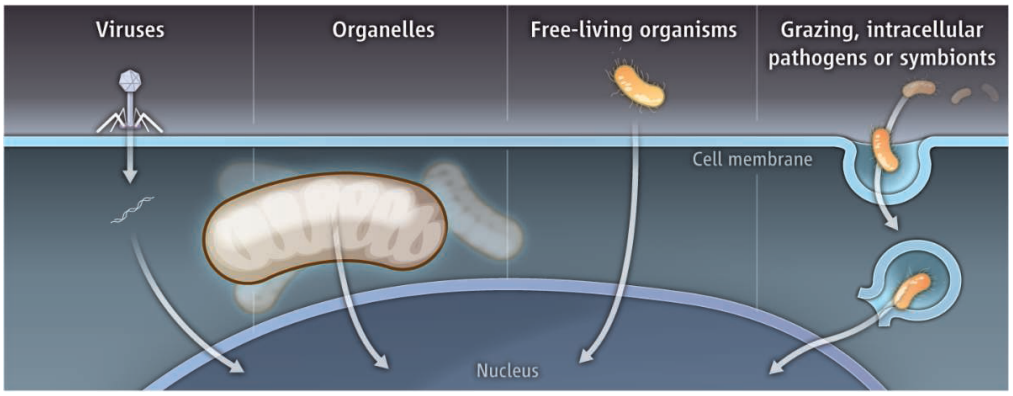


Figure 7: Pathways for HT from Prokaryotes to Eukaryotes [139].

The presence of multiple membranes complicates bacteria to vertebrate transfers (BVTs). Genes would have to cross cellular and nuclear membranes to reach a chromosome and only be heritable when present in a germ cell and facilitating a large fitness benefit to the host [108]. These methods of transfer are in essence similar to that found between prokaryotes. The largest proportion of interdomain transfer is evident between endosymbionts and their eukaryote hosts. The prolonged period of time in close proximity increases the probability of successful integration between host and symbiont [113]. This is particularly evident in the case of organelles such as mitochondria and chloroplasts.

1.4.2 Endosymbiont Bacteria to Eukaryote Horizontal Transfer

Transfers from an endosymbiont cell to the host cell will occur more frequently than that of free-living organisms due to the close and constant proximity of cells in both organisms. Transfers from endosymbionts in germ cells will have an even higher frequency as these transfers will be passed onto future generations. Mitochondria from an α-proteobacteria endosymbiont and chloroplasts from a cyanobacteria endosymbiont are found in reproductive cells and are transmitted to the progeny by germ cells with routine DNA transfer from these genomes to the nuclear genome [113]. Organelle-to-eukaryote transfer is clear in *Arabidopsis thaliana* chromosome 2 centromere. It contains 270 kb or 75% of the nuclear mitochondrial insert and the chromosomal sequence has 99% sequence identity with the mitochondrial genome, which indicates a recent transfer event [104].

Endosymbiont to eukaryote HT is highly frequent between *Wolbachia* endosymbionts and their hosts. *Wolbachia* cells are intracellular, maternally inherited and transferred through the egg cytoplasma ensuring a high probability of heritable HT to eukaryotic genomes [113]. *Wolbachia* HT has been observed in various eukaryotic orders and phyla:

- *Wolbachia* HT to the adzuki bean beetle, *Callosobruchus chinensis*, was the first *Wolbachia*-to-Arthropod transfer described [75]. Further research suggested that

roughly 30% of *Wolbachia* genes are present on the *Callosobruchus chinensis* nuclear genome, most likely from a single HT event, and that nearly half of the transferred genes were transcribed [76].

- The nuclear genomes of two distantly related, endosymbiont-free, filarial nematode species, *Acanthocheilonema viteae* and *Onchocerca flexuosa*, contain *Wolbachia* sequences [94]. Both these nematode species are *Wolbachia*-free, and as such there may have been rigorous HT between host and endosymbiont after which there was no need for the endosymbiont and discarded. Changes in nematode biology, *e.g.* host shift, making the endosymbiont obsolete may also explain the loss of endosymbiont [113].

- Roughly the entire *Wolbachia* genome was found to be transferred to the tropical fruit fly *Drosophila ananassae* (Figure 8) [60]. This insert is widely distributed and found in four lines of *Drosophila ananassae* from Asia and the Pacific [113].

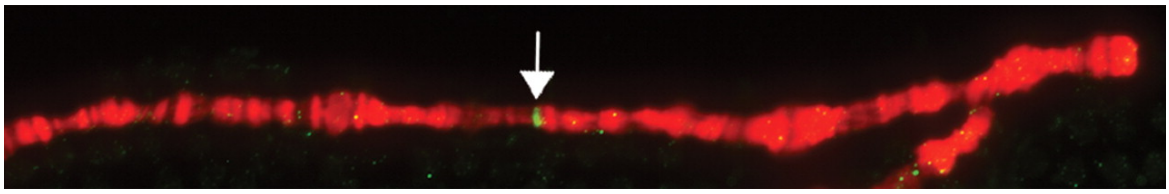


Figure 8: Fluorescence microscopy of chromosome 2L of *Drosophila ananassae*. DNA of *Drosophila ananassae* stained red with propidium iodide. Probe for a *Wolbachia* gene bound to a unique location, indicated in green [60].

- *Wolbachia* frequently infects the testes or ovaries of hosts and as such is able to enter the germline to enable transmission of transferred material to the host offspring [14].

The endosymbiont HT ratchet describes the accumulation of genes in the "host" nucleus with organelle origin with a similar ratchet used to explain the transfer of bacterial genes to the phagotrophic unicellular eukaryotes that feed on them [113]. Ratchets like these are applicable where two organisms have a close and lasting relationship (Figure 9).

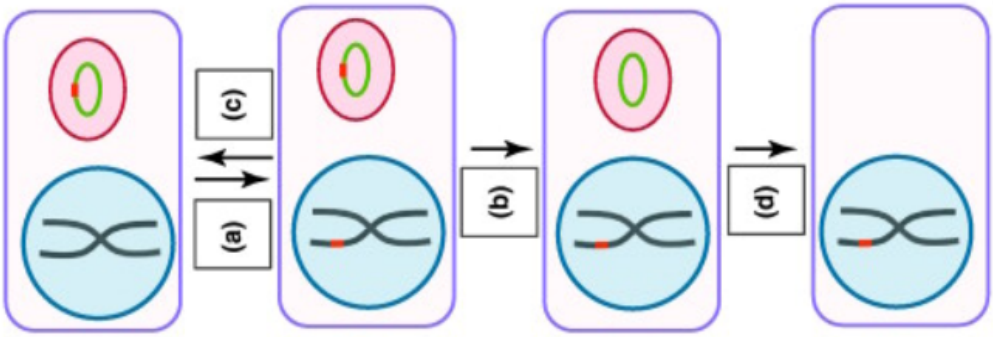


Figure 9: Endosymbiont HT ratchet. A gene (red) is transferred from the genome of the endosymbiont (green circular chromosome) to the nuclear genome (gray linear chromosome) in low frequency. (a) At even lower frequency transfers will occur that allows for the gene to be functional in the nuclear chromosome. As such either the nuclear or the endosymbiont version will be lost, with the loss of the system version nearly irreversible with the gene becoming fixed in the nuclear genome. (b) The loss of the nuclear version returns the endosymbiont to its original state and the process may be repeated. (c) Over time all genes that can be inserted in the nuclear genome will be inserted. (d) If maintenance of the endosymbiotic structure is no longer needed and HT has fulfilled all the needs of the host the endosymbiont may be lost [113].

1.4.3 Free-living Bacteria to Eukaryote Horizontal Transfer

Fungi

Fungi are likely to undergo a higher frequency of HT as many have a saprophytic or symbiotic lifestyle that involves close interactions with bacteria without some of the typical fences of HT such as the differentiation of germ line and soma [123]. HT may have been more important in fungal evolution than in other eukaryotes [140]. These events are more likely to be identified in fungi as they are one of the best sampled eukaryotic groups by means of fully sequenced genomes [123].

- In 2010 all sixty complete fungal genomes sequenced at the time were analyzed with fifty three (88%) of them displaying HT [123].
- Ten potentially horizontally transferred genes were identified in the genome of *Saccharomyces cerevisiae*. Two of these genes were further inspected with one gene found to be required for anaerobic synthesis of uracil, possibly transferred from Lactobacillales, and the other gene allowing for the utilization of sulfate from organic sources, possibly transferred from α-proteobacteria [26].
- An endoglucanase was reportedly transferred from the rumen bacteria *Fibrobacter succinogenes* to the rumen fungi *Orpinomyces joyonii* allowing for the colonization of a herbivorous rumen where cellulose and plant hemicellulose are the main raw nutritive substrates [88].

Algae

- *Bigelowiella natans* of the class chlorarachniophytes contained 2 genes obtained by means of HT from different bacteria [64]. It is suggested that HT occurs in *Bigelowiella natans* due to the mixotrophic, living both phagocytotically and photosynthetically, lifestyle [133].
- Diatoms are a major group of algae and is considered a large contributor to the primary productivity on Earth. Gene transfer from various bacterial sources was found to constitute 5% of the diatom gene stockpile, a level of transfer that is comparable to rates found in bacteria [23].
- The extremophilic unicellular red alga *Galdieria sulphuraria*, a member of the *Cyanidiophyceae* which inhabits volcanic hot sulfur springs, solfatara soils and anthropogenic hostile environments, displayed at least 75 independent gene acquisitions from archaea and bacteria [44]. The copious amount of gene transfer from prokaryotes enabled the adaptation of *Galdieria sulphuraria* to environments that are uninhabitable for eukaryotes with the large amount of transfers and diversity of archaeal/bacterial origins of these genes proposing that these transfers originated from free-living organisms [139].

Plants

- *Agrobacterium* is the usual suspect in bacteria-to-plant HT and the most commonly used in genetic engineering to introduce novel genetic material to a plant [125]. During infection with the bacterium a region of the tumor-inducing (Ti) plasmid, termed transfer DNA or T-DNA, is incorporated into the nuclear genome of the plant cell [1].
- T-DNA sequences were found to be present in all 291 tested accessions of cultivated sweet potato with detailed analysis indicating a probable *Agrobacterium* infection during the evolution and domestication of the popular crop [100].
- The insertion of T-DNA in a plant genome and the ensuing transfer by means of sexual reproduction is relevant in several species of the genera *Nicotiana* and *Linaria* [125].
- *Rhizobium*, *Sinorhizobium* and *Mesorhizobium* displayed the capacity to transform plant cells when furnished with suitable plasmids for DNA transfer [102]. Ti-like plasmids are not found contained in these bacteria naturally but HT of a Ti-plasmid may occur between them and *Agrobacterium* [1].

Asexually reproducing eukaryotes

Asexual reproduction denies organisms the opportunity of recombination through sexual reproduction. Muller's ratchet infers that asexual organisms will accumulate irreversible mutations. HT is thus deemed to be more prevalent in asexual organisms as it affords them a source of variation similar to recombination found in sexual organisms [113].

- *Hydra magnipapillata* is a freshwater cnidarian that may reproduce asexually through budding. Seventy-one candidates for HT has been identified with 70% of these candidates having expressed sequence tags (EST) support and the majority of these candidates only having bacterial homologs [56].
- Bdelloid rotifers are freshwater invertebrates that reproduce asexually. HT was found to be abundant in these small animals and one such candidate gene was over expressed in *Escherichia coli* yielding a functional enzyme [35].
- 148 genes in rumen Ciliates were found to be of bacterial or archaeal origin, especially Firmicutes, with the majority of these genes involved in metabolism, specifically in the metabolism of complex carbohydrates which is a rich food source in the rumen [43].

Sexually reproducing eukaryotes

- Six different gene families involved in cell-wall modification from four independent bacterial groups were identified in clade IV plant-parasitic nematodes [41]. Three of the bacterial groups are plant pathogenic soil bacteria associated with symbiotic relations with plant roots.
- The evolution of stinging cells in cnidarians was found to be as a consequence of the HT of a bacterial subunit of poly-y-glutamate [37]. Cnidarians use nematocytes to capture their prey.
- In the coffee berry borer beetle, *Hypothenemus hampei,* a gene encoding mannanase which hydrolyzes coffee berry galactomannan, the primary seed storage polysaccharide of coffee, was found to be of bacterial origin [80].

Humans

The human body contains more bacterial cells than host cells with numerous opportunities for HT due to proximity between cells and mechanisms available to prokaryotes for transfer of DNA [113]. HT between bacteria and humans, especially in the microbiome,

is of great interest to health and well-being with transfers having the potential to induce somatic cell mutations. Initial analysis on the draft human genome [38] indicated 223 proteins which may have been laterally transferred from a bacterial origin [92]. These proteins had no comparable similarity to proteins of any non-vertebrate eukaryote and the original human sequence was filtered to eliminate bacterial contamination [38]. The majority of these proposed genes contained introns and were confirmed as being from human origin by more than one observation thereby eliminating the possibility of contamination [131]. This is a considerable amount of BVTs and would suggest the permanent transfer of genes to the host due to infections. The human genome may therefore be a mere vesicle under continuous manipulation of foreign entities. Under this proposition it would be required that genes are transferred into the germ cell lineage with these transfers stably maintained in the host cell and spread throughout the population by means of a selective advantage to the host or a "selfish" nature entailing the ability to duplicate and transpose [92]. Detailed computational analysis of the draft human genome whittled the number down to 113 genes that were abundant in bacteria and only present in vertebrates in the eukaryote lineage [38]. The 110 genes discarded from the initial set of BVTs were found to be sparsely distributed in prokaryotes and as such not deemed characteristic bacterial proteins [108]. The condensed list of genes showed no preference for a bacterial donor and included genes involved in the metabolism of xenobiotics and stress response. The presence of two paralogues of monoamine oxidase, an enzyme central in the metabolism of neuromediators and target of psychiatric drugs, demonstrates the involvement of these laterally transferred genes in critical human physiological functions which would lead to fixation and maintenance during evolution due to the selective advantage provided to the host [38].

The International Human Genome Sequencing Consortium (IHGSC) findings with regard to the bacterial origin of numerous human genes did not go unnoticed and resulted in a "fresh skirmish in the genome wars" as it was dubbed in *The New York Times* [108]. The Celera version of the human genome [61] did not comment on or include any evidence of BVTs, HT or genes from a bacterial origin in their original paper. Salzberg et al. [92] threw the first punch with a detailed reanalysis of the Ensembl set [38] and included analysis on the Celera set [61]. Their work painted a different picture as to the frequency of BVTs in the human genome identifying 41 genes in the Ensembl set and 46 genes in the Celera set to be likely candidates for HT between bacteria and human. The existence of genes shared by prokaryotes and humans, yet unavailable in non-vertebrates is accounted for by a combination of evolutionary rate variation, the small sample of non-vertebrate genomes and most importantly the loss of these genes in non-vertebrate lineages [92]. Various further research and reviews show discrepancies and faults with the IHGSC results on the bacterial origin of human genes. In general it is proposed that there certainly are BVTs present in the human genome but that the frequency proposed

by the IHGSC paper is exaggerated and would likely lessen with the inclusion of more non-vertebrate sequence data and richer phylogenetic models. Figure 10 details the set of potential BVTs as proposed by different groups with only ten candidates accepted by all involved.

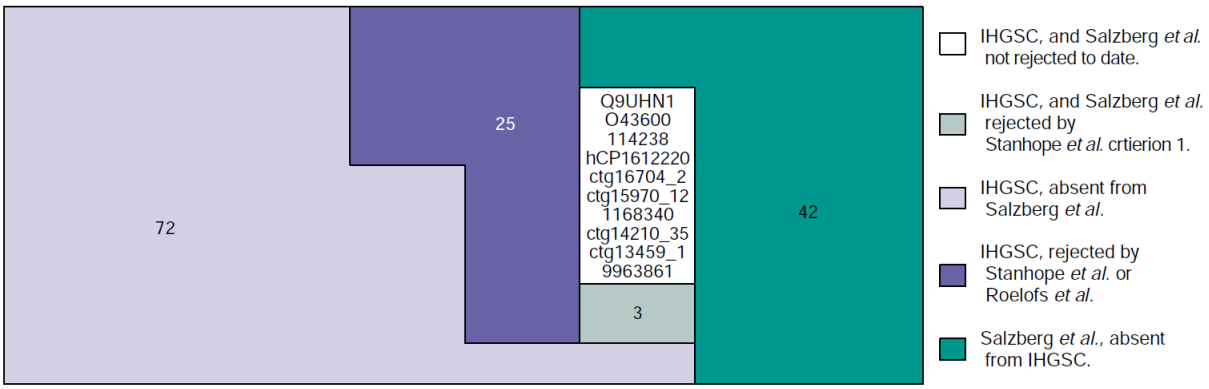


Figure 10: Set of potential genes transferred from bacteria to humans with overlaps and exclusions [108].

Challenges regarding multiple barriers, membranes and integration into the host germ line complicates inherited HT and BVTs. It is evident that inherited bacterial HT to a human host is not ubiquitous, but this does not dampen the opportunity for HT to somatic cells due to the large size and contact area of the human microbiome [113]. The majority of HT to somatic cells would go unnoticed in whole genome assemblies with these HT and somatic mutations having a critical effect on human health and disease. Proto-oncogene disruption by bacterial DNA has been proposed as an alternative to chronic inflammation as mechanism for cancer development as well as bacterial DNA transcription and expression [113]. Bacterial to human HT is not all doom and gloom with various advantages such as the production of vitamin K and the fermentation of non-digestible residues in the human gut [110].

1.5 Transfer, transfer everywhere

HT has been identified in multitudes of organisms from vastly differing taxonomic groups. The advent of the Next-generation sequencing (NGS) age will surely increase the identification of these events and the MGE that facilitate it. HT is an omnipresent and powerful tool in the evolution of life forms and the ability to adapt to various strenuous environments. The prevalence of HT events in early life forms to the current human genome reinforces the importance of MGE and HT in sustainability and biological progress. Identification and investigation of HT events and MGE will likely broaden current knowledge and hypotheses regarding evolution and adaptation in the biological realm.

1.6 Identification of Horizontal Transfer Events and Islands

The identification of islands in prokaryote genomes is an ongoing and developing field with various contrasting central dogmas. The two main approaches revolve around sequence composition and comparative genomics each with their own pros and cons. Sequence composition methods do not require extra information (sequences) as would the comparative methods and rely solely on the composition of the sequence under investigation to identify local areas that differ from the global pattern. Unfortunately sequence composition methods are prone to false positives and negatives. Highly expressed regions are known to be of variable composition and as such are often identified as probable regions of horizontal transfer when they are not. Dependence on sequence composition bias further increases the probability of a false-negative as horizontally acquired areas may reflect the host sequence composition. In essence all the sequence composition methods recruit k-mer (2-9) frequencies to indicate discrepancies in the sequence composition in a certain location [71].

- SIGI-HMM [97] predicts island by means of codon usage (CU) tables and Hidden Markov Models (HMM) for each gene in a sequence.
- IslandPath-DIMOB [103] is based on dinucleotide bias in eight or more consecutive open reading frames together with at least a single mobility gene for the prediction of islands.
- PAI-IDA [150] uses a combination of G+C content, dinucleotide frequency and CU as indicators of genome signature to identify islands.
- Centroid [55] splits a given genome or sequence into regions with a length of chosen window size and determines sequence composition through a k-mer (2-8) with a default of 5 nucleotides to obtain bias in a sequence signature.
- Alien_Hunter [152] employs Interpolated Variable Order Motifs (IVOMs) to identify compositional biases through variable order motif distributions, essentially long variable k-mers.
- PredictBias [96] incorporates G+C content, dinucleotide bias and regions with a minimum of six genes displaying codon biases to classify a segment of apparent lateral transfer.

Comparative methods are more robust against false positives and false negatives but do require multiple sequences of suitable origin for a detailed comparison. This method identifies clusters or areas in a genome that are not present in several other closely related genomes and as such are deemed to be laterally transferred. The success of this approach

relies on the set of genomes wherewith the comparison is done. Genomes too far separated will lead to difficulty in alignment and thus probable false positives with comparison of genomes too closely related overlooking possible transfer events in their communal history [71]. The choice of genomes to include in the comparative approach is the caveat in this approach. It should be noted that gene loss in genomes may influence results but that the increasing availability of sequenced genomes will improve results.

- IslandPick [70] automatically identifies a set for genome comparison and as such removes user bias after which MAUVE and then BLAST is used to predict unique and conserved regions.
- MobilomeFinder [51] focuses on bacterial tRNA and tmRNA genes, high insertion frequency regions, and MAUVE comparisons for the isolation of transfer events.

The use of 4-mer oligonucleotides in sequence comparison methods have been found to provide the best results in the determination of probable horizontal origin of clusters of genes [58]. The SeqWord Genomic Island Sniffer (SWGIS) program [77] enlists oligonucleotide and 4-mer frequencies in the determination of probable transfer events [77]. It has been found that the frequencies corresponding to oligonucleotides depend on physicochemical properties, and is influenced by CU [136]. Oligonucleotides are conserved signatures for bacterial genomes [135], as such genomic oligonucleotide usage (OU) composition is less variable within genomes than between, regardless of the region under investigation [46]. Alternating word lengths are analyzed to produce different oligonucleotide usage pattern (OUP) that are normalized by varying methods, with each OUP then characterized by unique statistical parameters [136]. These parameters include D - the distance between 2 patterns of the same type, PS - pattern skew defined as the distance between two patterns of the direct and reverse strands of the same DNA sequence, RV and GRV - oligonucleotide usage variances (OUV) normalized locally and globally respectively [46]. The OUV parameter is defined as variance of OU deviations normalized by the mononucleotide content of the sequence [135]. PS was determined to be universally conserved for complete bacterial genomes with OUV a more taxon-specific signature. Structural polymorphisms are effectively analyzed with the local OUP signatures [136]. These parameters are able to indicate regions altered by horizontal gene transfer as the parameters differ from the rest of the genome as seen in Figure 11. The genomes of every bacterial species is defined by an OUP or genomic signature, biases in frequencies of 2-7 bp oligonucleotides, that serves as a baseline for distinction between and within [77].

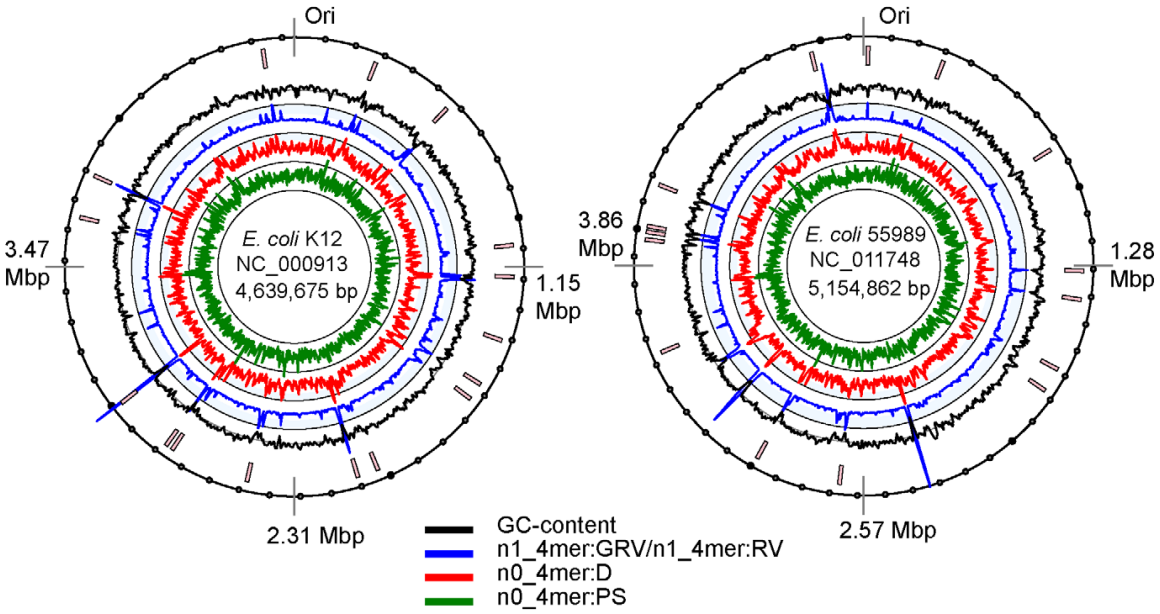


Figure 11: Distribution of islands identified by SWGIS in *Escherichia coli* strains.

1.7 Current Island Databases

1.7.1 HGT-DB (http://www.fut.es/~debb/HGT/)

The Horizontal Gene Transfer DataBase (HGT-DB) [89] could be regarded as the patriarch of prokaryote HT databases. Published in 2003 it is still available and functional today. The original version contained 88 bacterial and archaeal genomes including gene content and statistics, G+C content and deviations, amino-acid content and codon content and usage. HGT-DB identifies putatively foreign genes by means of G+C content, CU, length and amino-acid composition. A gene of interest is deemed to be horizontally acquired if the local G+C content or CU deviates by more than 1.5 standard deviations from the global mean, the gene length is greater than 300 bp and the local amino-acid composition does not deviate from the global amino-acid composition. Clusters of genes are furthermore regarded as laterally transferred if they deviate in G+C content regardless of their length or CU. These clusters would thus fall into the definition of an island. HGT-DB regards highly expressed genes such as ribosomal proteins to be false positive HT predictions and excludes them from the database. Highly expressed genes may adapt their CU to confer with the most available tRNAs and as such their CU is likely to differ from the mean.

1.7.2 ACLAME (http://aclame.ulb.ac.be)

The A CLAssification of Mobile genetic Elements database (ACLAME) [67] was originally published in 2004 and updated in 2009. The first version of ACLAME (version α)

employed TRIBE-MCL to produce clusters of true MGE proteins based on sequence similarity which were then manually curated. 5,069 proteins from 119 DNA bacteriophage genomes used in the first version resulted in 437 clusters with 3 or more members covering 2,501 proteins (50%). The top 300 clusters were then manually analyzed with 233 receiving functional annotation. The current release (version 0.4) includes MGE information on 457 bacteriophage genomes, 1,109 plasmids and 760 prophages. ACLAME includes reticulation or interweaving events between MGE based on a graph-based method originally used for phages. The communality of genetic modules between related groups of MGE is included in the database and defined as evolutionary cohesive modules.

1.7.3 PAIDB (http://www.paidb.re.kr)

The PAthogenicity Island DataBase (PAIDB) [90] first published in 2007 and updated in 2015 is a collection of reported pathogenicity islands (PAIs) and potential PAIs identified by a combination of feature-based and sequence-based analysis. The initial version included 112 PAIs which have been increased to 223 PAIs in PAIDB v2.0 which includes 88 resistance islands (REIs). PAIs reportedly promote disease development and REIs allow for a fitness advantage against antimicrobial agents. The updated version of PAIDB uses SIGI-HMM (measures codon adaption index) and IslandPath-DIMOB (measures dinucleotide bias in presence of mobility gene(s)) through the IslandViewer web server as the method of island prediction from which candidate PAIs and REIs are obtained.

1.7.4 IslandViewer (http://pathogenomics.sfu.ca/islandviewer)

IslandViewer [21] houses precomputed islands and allows for island prediction by 3 methods. The third release of IslandViewer is currently available, after being originally published in 2009, updated in 2013 and contains islands predicted in 2,794 genomes. This database incorporates the IslandPick, IslandPath-DIMOB and SIGI-HMM island prediction methods which are described by the authors as island predictors employing different yet complementary features. SIGI-HMM and IslandPath-DIMOB are sequence composition-based methods and IslandPick defined as a comparative genomics-based method.

1.8 Need for a Novel Database?

All the databases mentioned above target a specific area of island and prokaryotic research. They all appear to view islands as separate and mutually exclusive entities with little or no effect on other islands. The history and origin of islands are not explored and investigated in current databases, with the exception to some extent of ACLAME. This

methodology of island research set into motion the need for a comprehensive and encompassing tool to research islands in the host they are predicted in and the relationships they are involved in within the host and each other. A more detailed view on islands as a community and the interaction within the community and with the habitat they occur in was essential. Islands do not appear *de novo* and the footsteps or path taken by them needs to be retraced in order to understand and comprehend their existence. Current databases employ various methods of island prediction all with their own strengths and weaknesses. OUP and other OU statistics is an important island predictor, yet lacked a suitable database and collection of tools to analyze islands identified by alternative genomic signatures. The ability to compare newly predicted islands not only by means of prior prediction but by relational position to other islands was lacking and necessitated the development of a comparative tool incorporating all possible avenues.

1.9 Ontology and Stratigraphy

Island ontology is defined as the study 'of that which is' or existence of islands in prokaryotic genomes. This includes investigation into the existence of categories or groups of being. The identification of an island in combination with all relations and probable origin may give greater insight with regards to the history and reason of an island.

Stratigraphy is a term used to describe the geological study of layers or layering in rock and includes the investigation of fossils contained within rock strata. This term is used metaphorically to describe the flow of islands into a genome and as such adding information and layers to the existing content. These layers may be analyzed to determine probable time of insertion and origin in an attempt to unravel the flow of information. The identification of island or 'fossils' caught in the genome may provide insight with regards to the current genome content and existence of islands in a genome.

1.10 Aims and Objectives

The production of a viable and fruitful bioinformatic database and collection of tools for the adequate analysis of previously identified and novel islands is the cornerstone of this project. The ease and speed at which prokaryotic genomes are fully sequenced requires a convenient yet powerful starting point for the identification and analysis of predicted islands and comparison of novel and/or existing islands. The honing of composition-based approaches with which islands are predicted and collected is of vital importance in the warm-up to a reliable and progressive island analysis platform. Islands predicted require full inspection and deconstruction prior to catalogue and storage to enable the production of a unhindered pipeline to be used for further development. A "one stop

shop" is therefore envisaged to allow users functional and reliable research in the field of prokaryote mobilomics. This will be constructed on a relational database foundation with a graphical user interface (GUI) as merchant to guide and advise users. The ability to add or remove "inventory" is vital to the longevity of a scientific database in the current age of sequencing and needs to be addressed in the construction of a contemporary island resource. The relevance and accuracy of any scientific emporium should furthermore be tested and weighed after completion to indicate purpose and commitment. In summary the aims and objectives is as follows:

- Optimization of composition-based approaches employed in the identification and collection of islands in prokaryotes.
- Construction of a MySQL database housing information on islands identified by SWGIS hosted by sequenced archaeal/bacterial genomes.
- Establish the ability to scale and refurbish the database as the need arrives.
- Development of a web-based GUI in collaboration with the MySQL database to enable users island browsing, searching and retrieval capabilities.
- Database introspection and mining.
- Evaluation of the database as a contemporary and suitable island analysis package.

1.11 Discussion

The proposed methodology and implementation aims to produce a contemporary and dynamic suite of tools and collection of information to be extensively used in the field of HT and MGE research. This database aims to provide users with reliable and detailed information in conjunction with the ability to conduct further analysis on known and novel islands. The production of a scalable and expandable repository is of the essence in the age of affordable and rapid sequencing technologies. Development of a user-friendly interface with the ability to conduct novel island research is vital for database longevity and sustained use in an ever expanding biological arena. Availability of an alternative to current island databases and tool collections may improve present knowledge and understanding of HT and MGE. This set of tools and the collection of island information aims to allow users flexibility in the island research questions addressed and investigation approaches followed.

2 Chapter 2: Optimization of composition-based island prediction and collection

> We mostly don't get sick. Most often, bacteria are keeping us well. Bonnie Bassler

2.1 SWGIS overview

NGS allows for the complete identification of novel organism and strain genomic information in a relatively short period of time with a restrained cost. The ability to produce new genomic data is fast outperforming the capability of current sequence information mining in complete genomes. The identification of probable regions of HT based on compositional methods produce reliable results when used in combination with optimal parameters. This approach is deemed genome linguistics and employs text analysis algorithms in the detection of foreign regions of genetic material from varying origins [25, 129, 132].

The SeqWord Genomic Island Sniffer (SWGIS) program [77] is developed in the Python programming language and recruits various previously published routines in an effort to identify and isolate deviations in local OUP from the genomic global OUP in prokaryotes [46, 136]. Inconsistencies between local and global OUP may indicate genomic areas of probable HT.

SWGIS utilizes genome linguistics to detect discrepancies in local from global oligonucleotide frequencies and incorporates a set of combinatorial parametric measures with OU statistics to increase reliability of island prediction. It has been established that in archaeal/bacterial genomes frequencies of tetranucleotides provide the optimum results [46, 135, 136]. These OU parameters are designated n0_4mer for non-normalized tetranucleotide usage pattern and n1_4mer for normalized tetranucleotide usage pattern. It should be included that SWGIS allows for the use of word sizes from 2-mer to 7-mer with or without normalizations but the utilization of default parameters is strongly advised.

The SWGIS algorithm is grounded in the basic principle of superimposing values of several statistical parameters calculated for a sliding window to identify loci with a distinct OUP and furthermore distinguish between alternate categories of these atypical genomic fragments. In essence islands are identified by an alternative OUP (increased n0_4mer:D) with a lower internally normalized OUV (n1_4mer:RV) and an increase in globally normalized OUV (n1_4mer:GRV). PS (n0_4mer:PS) comparisons are used to identify and circumvent *rrn* operons as these regions display acute values of n0_4mer:PS. These parameters are measured by means of a sliding window approach. Genomic fragments of 8

kbp with a 2 kbp step are compared with the tetranucleotide usage values in the whole genome. The identification of a reliable increase in the local distance (n0_4mer:D) in conjunction with a significant decrease in n1_4mer:RV and increase of n1_4mer:GRV indicates an area of foreign acquisition and signals the window to move several positions back and repeat the analysis with steps of size 0.2 kbp in order to identify the coordinates of the foreign insert. Thresholds for parameter deviation may be specified to achieve acceptable false negative and false positive ratios.

SWGIS is able to identify multiple islands in numerous genomes in a single run. Complete archaeal/bacterial genomes in GenBank or FASTA format are required as input. GenBank format is advised as this enables greater power to SWGIS in the production of output files. Assorted output files are created for each genome which include a standard text file (.out) that contains a list of identified islands with coordinates and OU parameter values for each island. The usage of GenBank file format as input enables SWGIS to identify genes within the borders of each island and will be included in the .out file. It further allows for the production of a GenBank file for each island that would contain annotation data. SWGIS provides a FASTA file containing sequence information for all identified islands for both input formats and allows users the option to create graphical SVG files for each analyzed genome. These genomic atlas .svg files indicate the position of predicted islands in conjunction with the visualization of defection of OU parameter values for each island.

2.2 SWGIS parametric optimization

Islands identified and housed by PAIDB were used as a baseline and reference for a true positive island prediction. Empirical analysis indicated that D and V values below 1.5 increased false positive prediction of islands whereas values above 2 increased false negative rates. Factorial analysis was enrolled to determine the optimal combination of values in order to minimize false positive and false negative rates [144]. This resulted in regression equations for estimating expected values of false negative rates (FNR) and false positive rates (FPR) for D and V thresholds:

$$FNR = -0.628 + 0.118D + 0.392V$$

$$FPR = 0.752 - 0.121D - 0.173V$$

$$FNR + FPR = 0.124 - 0.003D - 0.219V$$

The determination of "true negative" category for islands is cumbersome and as such the sensitivity and specificity parameters were modified:

$$Sensitivity = \frac{1-FNR-FPR}{1-FPR}$$

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

Expected values of FNR, FPR and FNR+FPR under different parametric combinations are presented in Figure 12 A - C below.

A

FNR = -0.628 + 0.118D + 0.392V

D \ V	1.5	1.55	1.6	1.65	1.7	1.75	1.8	1.85	1.9	2
1.5	0.14	0.16	0.18	0.2	0.22	0.24	0.25	0.27	0.29	0.33
1.55	0.14	0.16	0.18	0.2	0.22	0.24	0.26	0.28	0.3	0.34
1.6	0.15	0.17	0.19	0.21	0.23	0.25	0.27	0.29	0.31	0.35
1.65	0.15	0.17	0.19	0.21	0.23	0.25	0.27	0.29	0.31	0.35
1.7	0.16	0.18	0.2	0.22	0.24	0.26	0.28	0.3	0.32	0.36
1.75	0.17	0.19	0.21	0.23	0.25	0.26	0.28	0.3	0.32	0.36
1.8	0.17	0.19	0.21	0.23	0.25	0.27	0.29	0.31	0.33	0.37
1.85	0.18	0.2	0.22	0.24	0.26	0.28	0.3	0.32	0.34	0.37
1.9	0.18	0.2	0.22	0.24	0.26	0.28	0.3	0.32	0.34	0.38
2	0.2	0.22	0.24	0.25	0.27	0.29	0.31	0.33	0.35	0.39

B

FPR = 0.752 - 0.121D - 0.173V

D \ V	1.5	1.55	1.6	1.65	1.7	1.75	1.8	1.85	1.9	2
1.5	0.31	0.3	0.29	0.28	0.28	0.27	0.26	0.25	0.24	0.22
1.55	0.3	0.3	0.29	0.28	0.27	0.26	0.25	0.24	0.24	0.22
1.6	0.3	0.29	0.28	0.27	0.26	0.26	0.25	0.24	0.23	0.21
1.65	0.29	0.28	0.28	0.27	0.26	0.25	0.24	0.23	0.22	0.21
1.7	0.29	0.28	0.27	0.26	0.25	0.24	0.23	0.23	0.22	0.2
1.75	0.28	0.27	0.26	0.25	0.25	0.24	0.23	0.22	0.21	0.19
1.8	0.27	0.27	0.26	0.25	0.24	0.23	0.22	0.21	0.21	0.19
1.85	0.27	0.26	0.25	0.24	0.23	0.23	0.22	0.21	0.2	0.18
1.9	0.26	0.25	0.25	0.24	0.23	0.22	0.21	0.2	0.19	0.18
2	0.25	0.24	0.23	0.22	0.22	0.21	0.2	0.19	0.18	0.16

C

FNR+FPR = 0.124 - 0.003D + 0.219V

D \ V	1.5	1.55	1.6	1.65	1.7	1.75	1.8	1.85	1.9	2
1.5	0.45	0.46	0.47	0.48	0.49	0.5	0.51	0.52	0.54	0.56
1.55	0.45	0.46	0.47	0.48	0.49	0.5	0.51	0.52	0.54	0.56
1.6	0.45	0.46	0.47	0.48	0.49	0.5	0.51	0.52	0.54	0.56
1.65	0.45	0.46	0.47	0.48	0.49	0.5	0.51	0.52	0.54	0.56
1.7	0.45	0.46	0.47	0.48	0.49	0.5	0.51	0.52	0.53	0.56
1.75	0.45	0.46	0.47	0.48	0.49	0.5	0.51	0.52	0.53	0.56
1.8	0.45	0.46	0.47	0.48	0.49	0.5	0.51	0.52	0.53	0.56
1.85	0.45	0.46	0.47	0.48	0.49	0.5	0.51	0.52	0.53	0.56
1.9	0.45	0.46	0.47	0.48	0.49	0.5	0.51	0.52	0.53	0.56
2	0.45	0.46	0.47	0.48	0.49	0.5	0.51	0.52	0.53	0.56

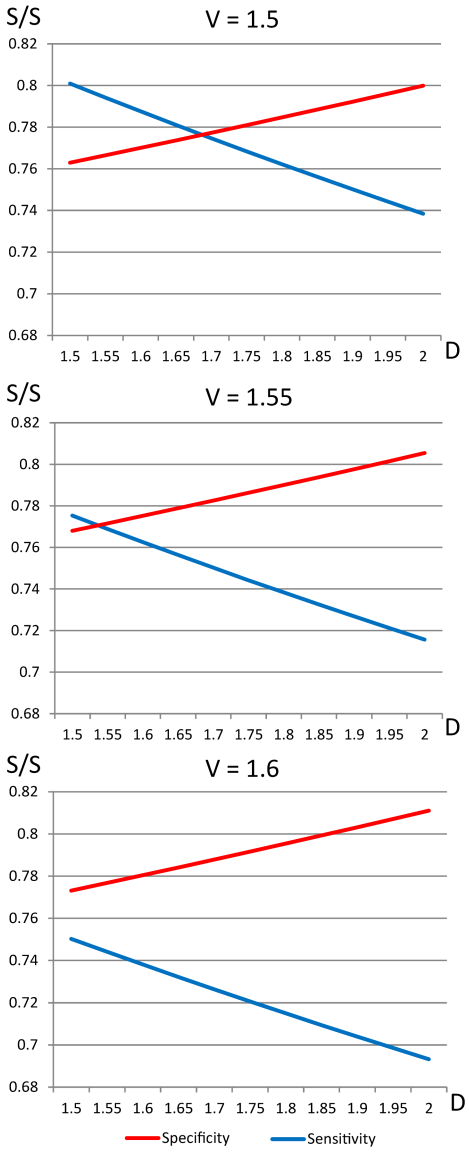


Figure 12: Parts A and B show FNR and FPR calculated for different combinations of D and V, respectively; and their sum in the part C. Parts D, E and F represent the expected specificity and sensitivity (S/S) for variable D thresholds depicted on the horizontal axis and fixed V thresholds. Vertical axes represent specificity and sensitivity values.

There is a gradual change in colors from the highest (red) to intermediate (yellow) to lowest or optimum (green) prediction rates in the figures presented above. FNR is at a

minimum for [D:1.5; V1.5] and increases when changed to [D:2.0, V:2.0]. This is indicative of a lower probability of neglecting an island by using parametric values [D:1.5, V:1.5] rather than [D:2.0, V:2.0]. The parametric values [D:1.5, V:1.5] result in a lower FNR and highest sensitivity yet it produces an increased FPR and low specificity. Reduction in FPR and increase in specificity is achieved with the parametric set [D:2.0; V:2.0] (Figure 12 B). Alterations in the cumulative FNR+FPR which is influenced by parameters D and V is displayed in Figure 12 C. Increase in parameter V leads to the progressive increase in FNR+FPR while a change in D has no effect as the increase in FNR is compensated by a correlated decrease in FPR. Optimization of specificity and sensitivity in island prediction is achieved by the adjustment of parameter D and the use of a constant and minimal V.

Specificity and sensitivity for altered D and fixed V parameter thresholds are shown in Figure 12 D - F. Optimal specificity/sensitivity ratio is achieved with parameter set [D:1.7; V:1.5]. This serves as the default parameter set for SWGIS. This default set of parameter values may be altered by users to reduce FPR or in case of adjustments required to specificity/sensitivity ratio for certain genomes.

2.3 SWGIS failures and problem resolving strategies

The performance of SWGIS may further be improved by analyzing the patterns of genomes where it performed poorly with respect to an independent island prediction program (IslandViewer). Genomes in Figure 13 below were identified as over or understating the amount of islands predicted by SWGIS. These genomes are graphically marked in the column FPR/FNR by red leftward and blue rightward bars depicting FNR and FPR over-ranges, respectively. FPR/FNR was calculated as follows:

$$FPR/FNR = (N_{SWGIS} - N_{IslandViewer})/N_{average}$$

N_{SWGIS} is the number of islands predicted by SWGIS at [D:1.5; V:1.5]; $N_{IslandViewer}$ is the maximum number of islands predicted by one of the IslandViewer programs and $N_{average}$ is the average number of islands predicted by all programs.

These predictions were investigated to determine probable causes of these failures. Genomes were searched for commonalities which may explain the excessive amount of islands predicted.

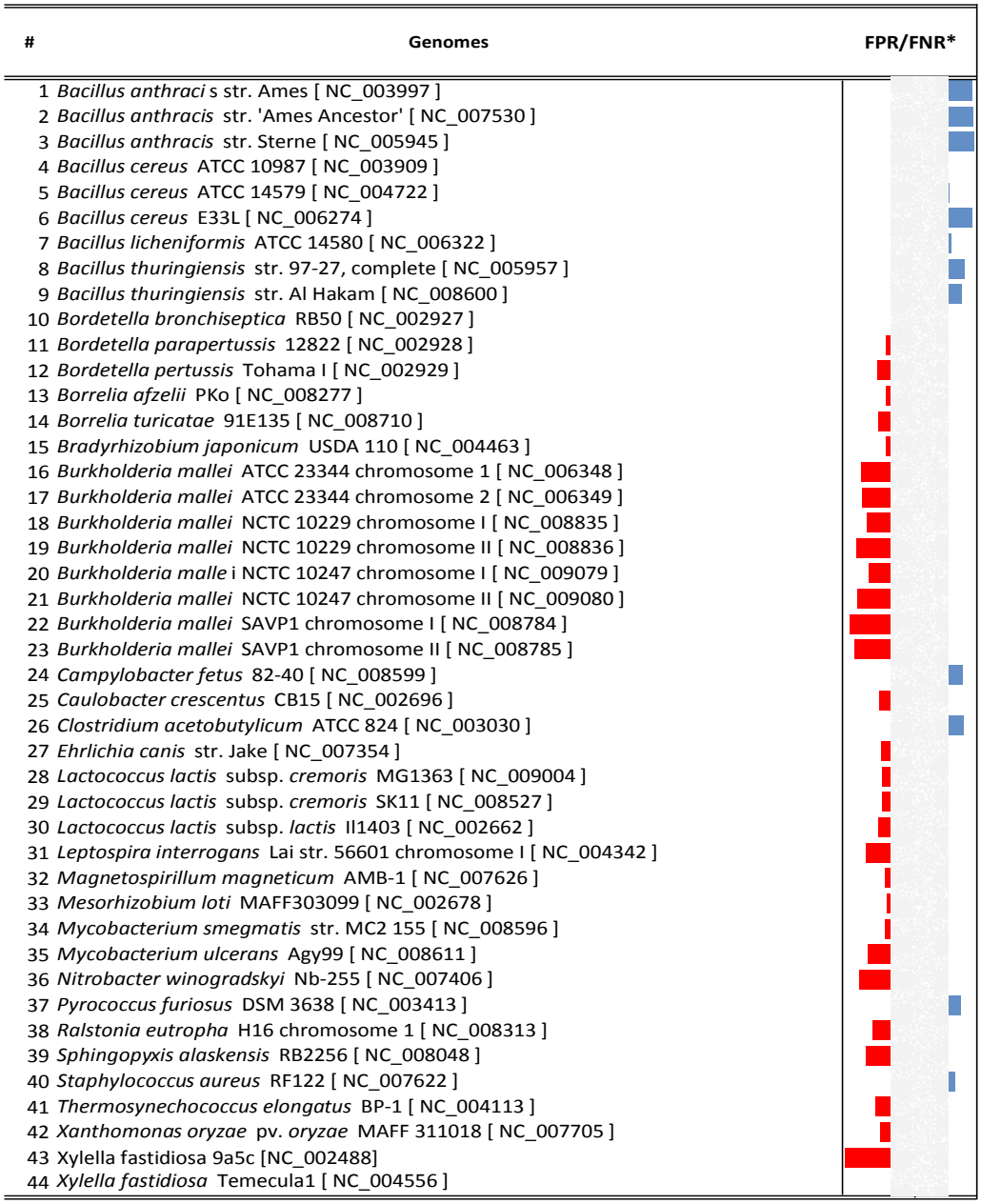


Figure 13: Genomes in which numbers of islands predicted by SWGIS were significantly over-ranged regarding to predictions by other programs that may indicate large FNR (red columns) or large FPR (blue columns).

2.3.1 False positives

In the genomes of *Bacillus cereus, Bacillus anthracis, Bacillus thurengiensis* and several others, mostly Firmicutes, SWGIS predicted considerably more islands than the Island-

Viewer programs. In Figure 14 multiple falsely selected *rrn* operons are included in the set of predicted islands hosted by *Bacillus cereus* ATCC 14579.

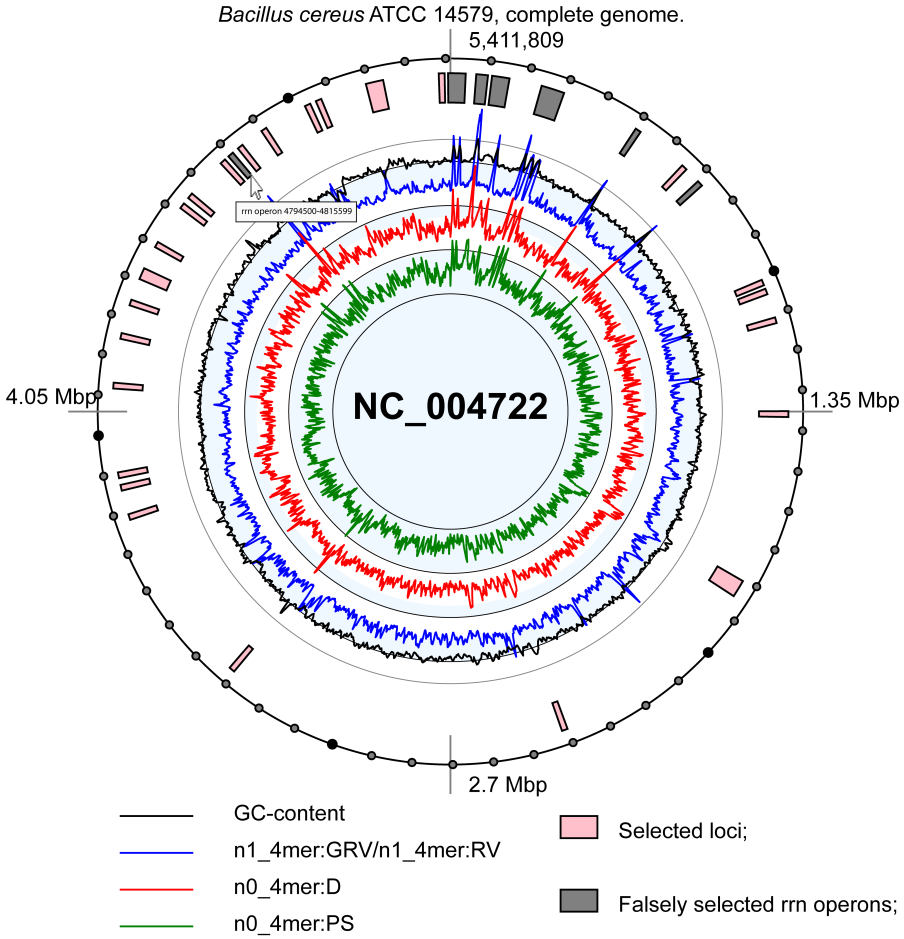


Figure 14: Multiple islands predicted by SWGIS in *Bacillus cereus* ATCC 14579 including falsely selected *rrn* operons.

Theoretically there are no genes or genomic fragments in archaea/bacteria which may not be subjected to HT and as such there is no standard on the rejection of falsely predicted islands. This complicates HT prediction and as such genomes displaying unusually high numbers of predicted islands were investigated for the presence of common properties to explain the excessive number of predicted islands. The genomes under question all displayed compositional polymorphism with large parts of their chromosomes characterized by alternative G+C and OU-bias. In particular, DNA molecules in the central area of *Bacillus cereus* chromosome are more AT rich and possess more pronounced intrinsic curvature; increased stacking energy; higher position preference; and a higher occurrence of palindromes [59]. It is proposed that these bacteria acquired one or even multiple giant islands which subsequently underwent fragmentation and spread across the chromosomes. The biological relevance of the commonality of this compositional polymorphism is of yet unknown, except for their common horizontal acquisition. It is possible that the majority

of the islands predicted in this set of bacteria might be false positive results. High FPR is avoided by more stringent SWGIS parameter settings and in this example an increase in the D threshold is suggested (Figure 12).

2.3.2 False negatives

Composition-based island prediction methods are designed to identify regions with atypical OUP in a given genome. This method may omit islands which have been acquired from a donor with a similar OUP or ancient acquisitions which have been altered by amelioration. SWGIS is not immune to this problem and fails to identify ancient insertions; fragments where OUP are indistinguishable from the core chromosomal sequence; and reliable OU calculations from short DNA inserts [46]. SWGIS detected relatively few islands in *Borrelia burgdorferi* B31; *Burkholderia mallei* ATCC 23344, *Burkholderia mallei* NCTC 10229 and *Burkholderia mallei* 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. These predictions were deemed inconsistent with those of the IslandViewer programs.

The organisms mentioned above do not harbor any taxonomical links between themselves. Even in other strains of the same species islands may be identified without any problems that are in contrast to the false positive prediction problem discussed in the previous section, which was characteristic for species and groups of related organisms. The reason for island prediction failure in *Xylella fastidiosa* 9a5c is that this organism has developed a mutator phenotype that eroded its chromosomal OUP specificity [135], it was thus impossible for SWGIS to make predictions. In the contrary, there were no problems with island identification in *Xylella fastidiosa* Temecula1 which shows a stable chromosomal OUP. False negative prediction in *Thioalkalimicrobium cyclicum* ALM1 is shown in Figure 15.

This overlooked region is a large 87,608 bp viral filamentous hemagglutinin gene with constituent repeats. This island is highlighted in Figure 15 on the genomic atlas and displays parameters associated with an island. This island was rejected by SWGIS as it considers giant genes with multiple repeats as a separate category of genomic elements with alternative OUP [134]. Visual inspection of genome maps provided by SWGIS enables the identification of these false negative predictions. SWGIS does not include giant genes in prediction results as these genes seldom undergo HT and would lead to numerous false positive predictions.

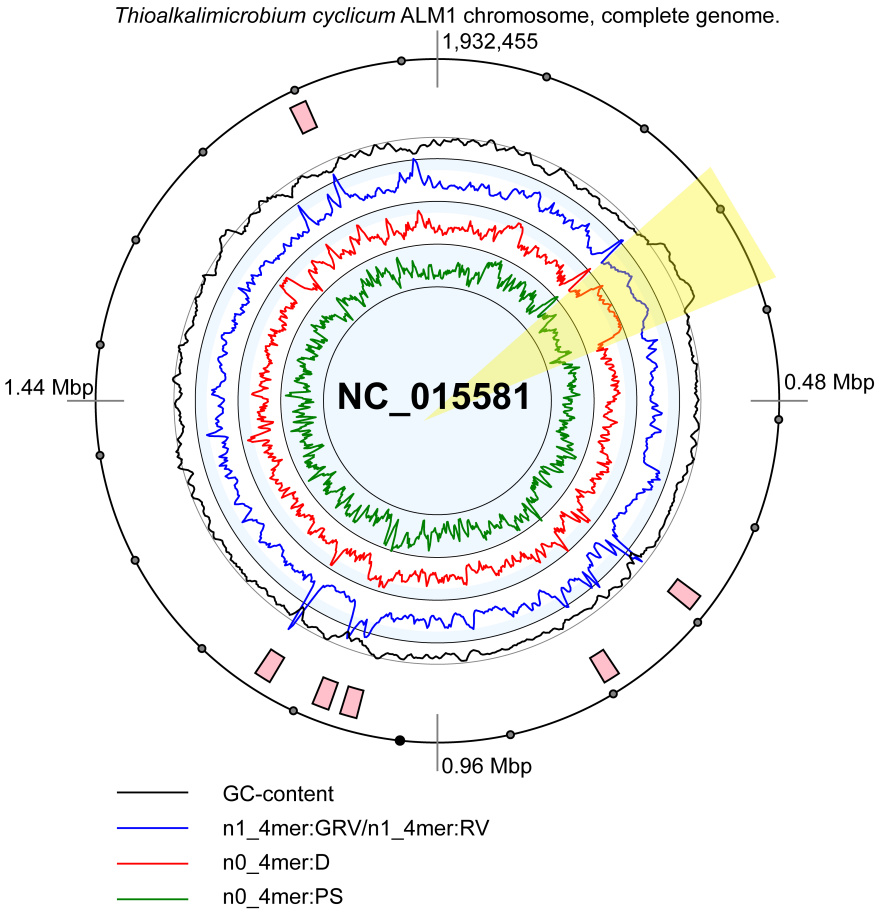


Figure 15: An insertion of a giant viral gene in the chromosome of *Thioalkalimicrobium cyclicum* ALM1 is highlighted on the atlas and was overlooked by SWGIS.

2.4 Continued analysis after prediction

The successful prediction of a region of HT is the first phase in the study of HT. Numerous computational methods exist to identify these regions of foreign acquisition, yet there is a lack of tools to further study the relationships between islands and their possible origins. SWGIS allows for various analysis of islands post prediction in an effort to expand knowledge on island origin, liaison and migration.

LingvoCom is a collection of such utilities, available from the SeqWord project, to analyze genome linguistics in genomic sequences ranging in size from small fragments to complete genomes in GenBank or FASTA format. This includes island files as produced by SWGIS. Alternatively it may extract DNA fragments from a whole genome by user defined coordinates and is freely available from http://www.bi.up.ac.za/SeqWord/lingvocom.

LingvoCom offers "3D-plot" and "d-matrix" functions to group islands or other DNA segments by compositional similarity or phylogenetic trees respectively. 3D-plotting is an implementation of the non-metric multidimensional scaling (MDS) algorithm [122]. SWGIS identified 12 islands in *Nitrosomonas europaea* ATCC 19718 and 11 islands in

Nitrosomonas eutropha C91 which were used as input for LingvoCom. The genomes of *Salmonella enterica* subsp. *enterica* Typhi Ty2, *Clostridium thermocellum* ATCC 27405 and *Acidovorax ebreus* TPSY were used as outgroups in this analysis. LingovoCom creates a graphical output file (Figure 16) that represents a 3D projection of islands and chromosomes which were compared by means of OUP similarity.

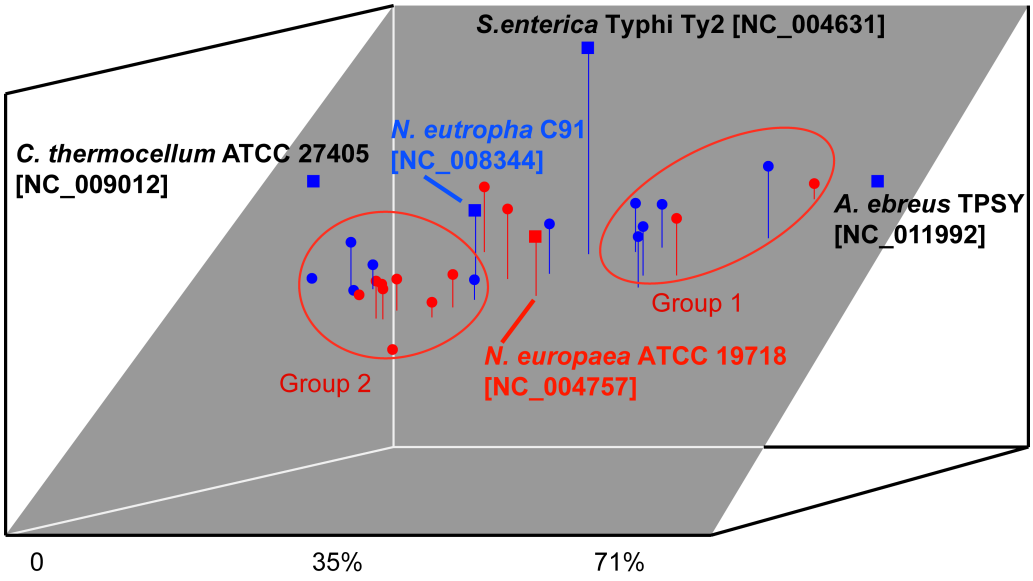


Figure 16: OUP comparison 3D projection of the two *Nitrosomonas* genomes, their islands and the three outgroup genomes of *Salmonella enterica*, *Clostridium thermocellum* and *Acidovorax ebreus*. Islands are depicted by red (hosted by *Nitrosomonas europaea* ATCC 19718) and blue (hosted by *Nitrosomonas eutropha* C91) circles; whereas the chromosomes are depicted by squares. Two groups of *Nitrosomonas* islands with similar patterns are outlined and encircled.

The islands contained in the two *Nitrosomonas* species were grouped by compositional similarity in two clusters with different probable origins. G+C-rich islands shared OUP similarity with *Acidovorax* whilst the AT-rich islands shared similarity with *Clostridium*. The OUP of the *Salmonella* specie is equally distant from the islands of the *Nitrosomonas* species and their chromosomes.

The "d-matrix" function is used to construct a matrix file (Phylip format) with distances calculated for OUP of DNA sequences. This functionality was used to produce a Neighbour Joining tree based on the matrix of distances calculated for OUP of *Nitrosomonas* islands and their chromosomes together with that of *Salmonella enterica* subsp. *enterica* Typhi Ty2 as the outgroup and is displayed in Figure 17. This approach indicates two probable origins for the islands identified in the *Nitrosomonas* species.

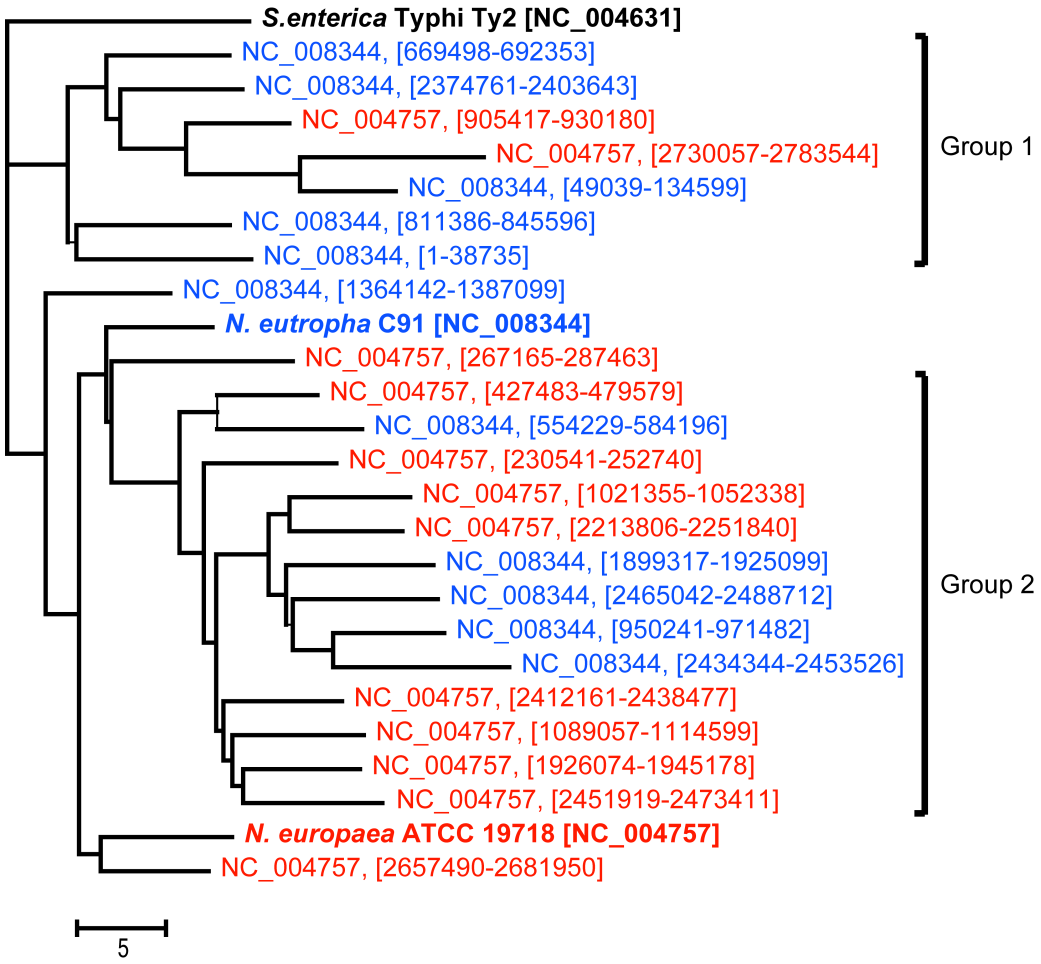


Figure 17: A dendrogram representation of two groups of *Nitrosomonas* islands with *Salmonella enterica* subsp. *enterica* Typhi Ty2 as an outgroup.

The process of genome amelioration alters the OUP of a foreign insert to resemble that of the genome in which it resides. It was illustrated in Figure 16 that these islands possibly originated from two different sources and with their OUP reverting to that of the current host chromosome. The islands in Figure 16 which are located closer to the chromosome indicate older inserts. Distantly placed islands reveal a recent acquisition as they retain the OU properties of the donor genome. LingovoCom enables the comparison of island OUP to that of their hosts in order to determine putative donors. In Figure 18 the OUP similarity conducted for an island [2,730,057-2,783,544] of *Nitrosomonas europeae* ATCC 19718 and its possible donor, *Acidovorax ebreus,* is shown. Two dark green spots on the plot represent OUP of the query (at the center point) and subject (on the horizontal axis) chromosomes. Light green circles depict $\frac{1}{2}$ of the distance between patterns calculated for the chromosomes. Islands of the query genome are shown as red small circles and those of the subject genome, as blue circles. The OUP of island [2,730,057-2,783,544] of group 1 (designated in Figure 16) is much more similar to *Acidovorax* than to its host chromosome, which indicates that this island may have originated from the *Acidovorax*

genus and underwent subsequent HT.

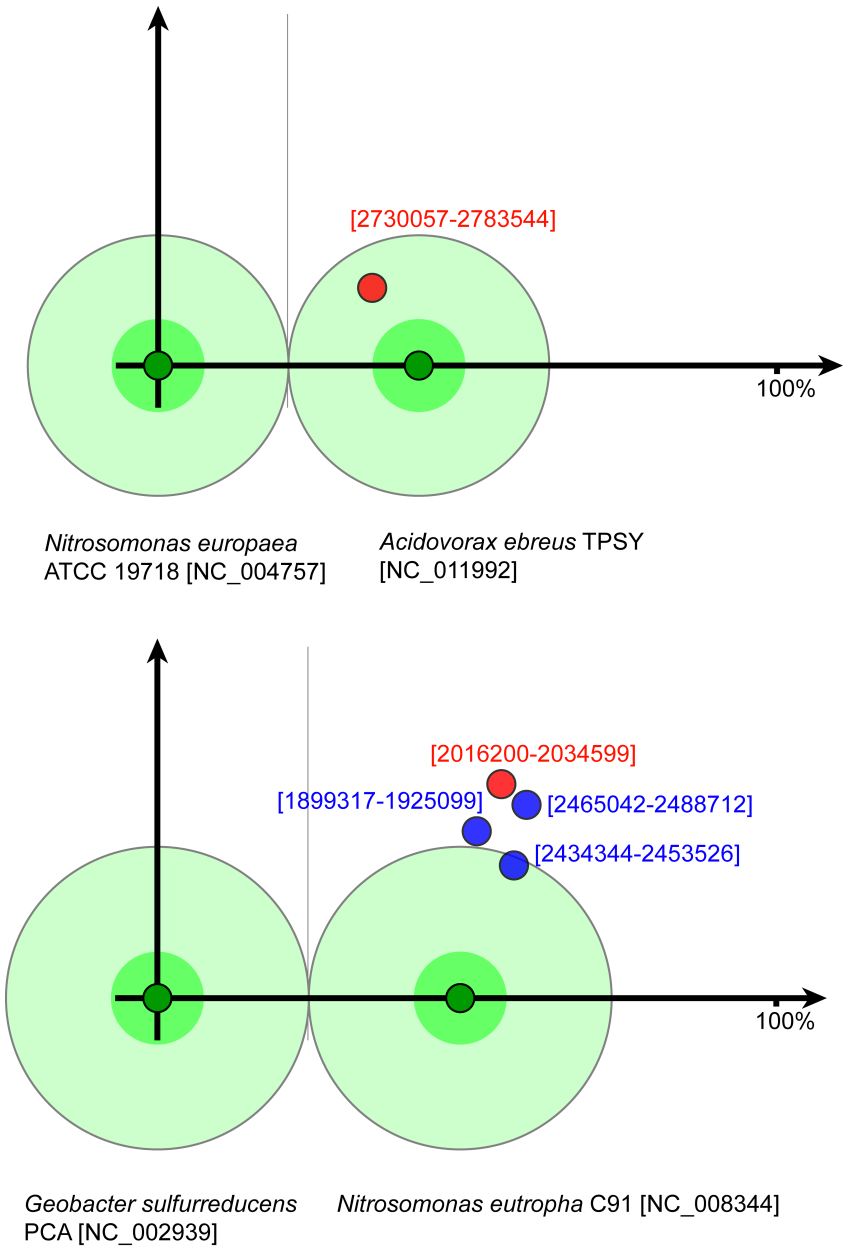


Figure 18: 2D projection of islands and their possible donor organisms as determined by calculating distances between their OUP. This method is used to determine donor-recipient relations between islands and groups of organisms which share a common OUP. Top) Depicts that *Acidovorax* is a possible donor of one island found in *Nitrosomonas eutropha* ATCC 19718; Bottom) Depicts *Nitrosomonas eutropha* C91's possibly ameliorated islands (blue circles), and an island (red circle) of *Geobacter sulfurreducens*, which is possibly of *Nitrosomonas eutropha* C91 origin.

An island which displayed compositional similarity with group 2 of *Nitrosomonas* (Figure 16), was found in the chromosome of *Geobacter sulfurreducens* [2,016,200-2,034,599]. 2D projection of OUP of these islands and their host chromosomes (Figure 18) indicate that *Nitrosomonas* may be a donor of HT genes for *Geobacter sulfurreducens*, or in both

organisms islands have originated from a common yet unknown source. Possible pathways of distribution of islands including PAI and the relative time of acquisition may be determined by this approach [78].

2.5 SWGIS comparison

SWGIS was compared to the predictors included in IslandViewer, i.e. IslandPick, SIGI-HMM and IslandPath [21], by means of re-identification of 51 PAIs found in 24 microorganisms to determine FNR. FNR is defined as the percentage of known island included in PAIDB [90], overlooked by any or all prediction programs. SWGIS was implemented with 4 different combinations of D and V parameters: [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] and outperformed other island prediction method even when implemented with the strictest of parameters ([D:2.0; V:2.0]). IslandViewer programs concurred on 69% of the 51 PAIs and SWGIS with 88%. Overlap between the different island identification programs is summarized in Figure 19.

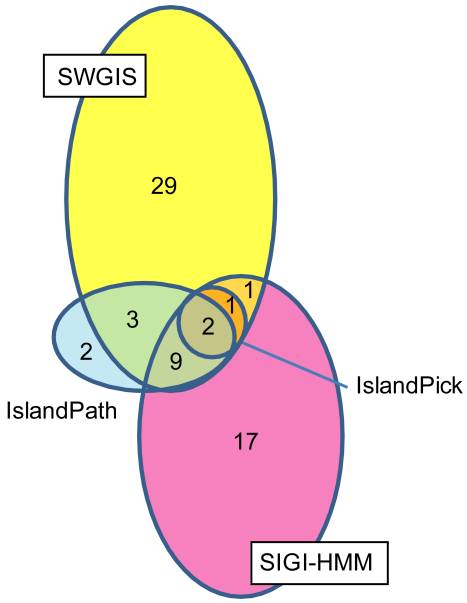


Figure 19: A Venn diagram of the overlapping predictions by SWGIS, SIGI-HMM, IslandPick and IslandPath.

All IslandViewer predicted islands were confirmed by SWGIS with only 2 IslandPath islands not included by SWGIS. In total only 4 PAIs were not predicted by either SWGIS or IslandViewer programs. The results of the comparison is displayed in Figure 20 with varying SWGIS parameters.

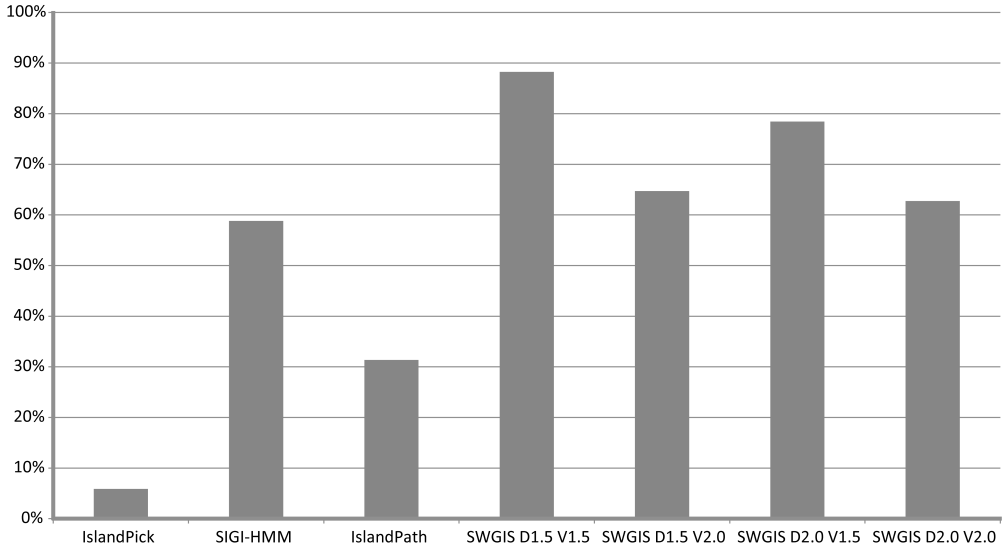


Figure 20: Re-identification of known PAIs by IslandViewer programs and SWGIS.

FPR is defined as the prediction of a genomic segment as being HT, whilst it was not. The identification of falsely predicted islands is more troublesome than FNR as there are no formal methods regarding FPR with reference to MGE. Regions of foreign origin which are of benefit to the host are altered by the process of amelioration. This is not observed for regions inclusive of *rrn* gene clusters; operons of ribosomal proteins; giant genes; and local tandem repeats which are always characterized by OUP which differ from the host [134, 136]. SWGIS employs overlay of different OU statistical parameters to identify regions of atypical genomic loci. Operons of ribosomal RNA genes display compositional properties akin to that of islands and as such frequently composition-based methods falsely concluded that these operons are islands. Optimization of OUP statistics as described above may decrease FPR, even so it is not always possible to remove all *rrn* operons from predictions due to the compositional specificity of numerous bacterial genomes. SWGIS is equipped with a database of 16s rRNA sequences obtained from different high level taxonomically grouped prokaryotes to allow for the filtering of the *rrn* genes during the prediction process. Transcription of horizontally acquired *rrn* operons may be destructive to recipients and therefore these elements are unlikely to undergo HT [112], still possibility of transfer may not be ruled out [116]. Furthermore, the periphery of *rrn* operons are favorable insertion hotspots for islands [20, 74]. Compositional similarity of *rrn* operons with islands located in their outskirts complicates the detection of a border which would set them apart and as such SWGIS returns the combined segment as a putative island. A set of 2,413 islands were discarded due to filtering with the database of 16S rRNA and re-examined for occurrence of MGE associated key words such as “integrase”, “transposase”, “phage” and “IS-element”. This identified 372 (15%) of the islands in the set to be falsely excluded and in fact true positives consisting of genes associated with MGE. SWGIS affords users the opportunity to inspect these rejected

islands by means of the genomic atlas output. Positions of rejected islands displaying sequence similarity to the 16S rRNA database are shown in grey (Figure 21) and exact location retrieved by mousing over the grey blocks. These locations may then be manually inspected for confirmation.

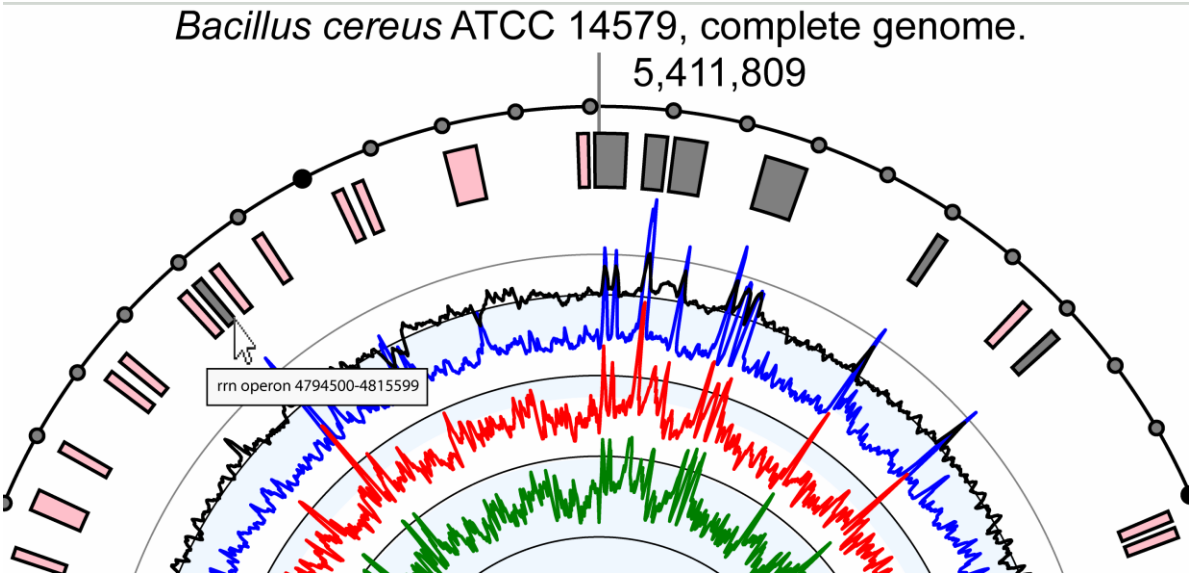


Figure 21: A graphical output of SWGIS in SVG format displaying positions of predicted islands in *Bacillus cereus* ATCC 14579 [NC_004722]. Pink blocks depict islands, whereas grey blocks depict genomic regions which comprise genes of 16S rRNA and segments that are falsely predicted.

Rates of false positives were determined by comparing islands predicted by SWGIS and IslandViewer tools. The FPR cannot be estimated directly, therefore frequencies of islands which were predicted by a single program and not by others were calculated and termed "unconfirmed predictions". Pre-calculated sets of islands from 169 bacterial chromosomes were obtained from the IslandViewer web resource. SWGIS searched for islands in the same 169 bacterial chromosomes with parameters [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]. Counts of predicted islands and frequencies of unconfirmed islands for each program is summarized in Figure 22 with SWGIS using the most relaxed parameters [D:1.5; V:1.5]. Confirmation of an island was subject to at least partial overlap of predicted genomic loci between programs.

None of the methods used for this comparison guarantees detection of all islands. This may explain the large amount of unconfirmed islands, additionally a large amount of these unconfirmed islands may be false positives. SWGIS identified the greatest amount of islands incorporating the less stringent parameter [D:1.5; V:1.5]. This included the highest rate of unconfirmed predictions.

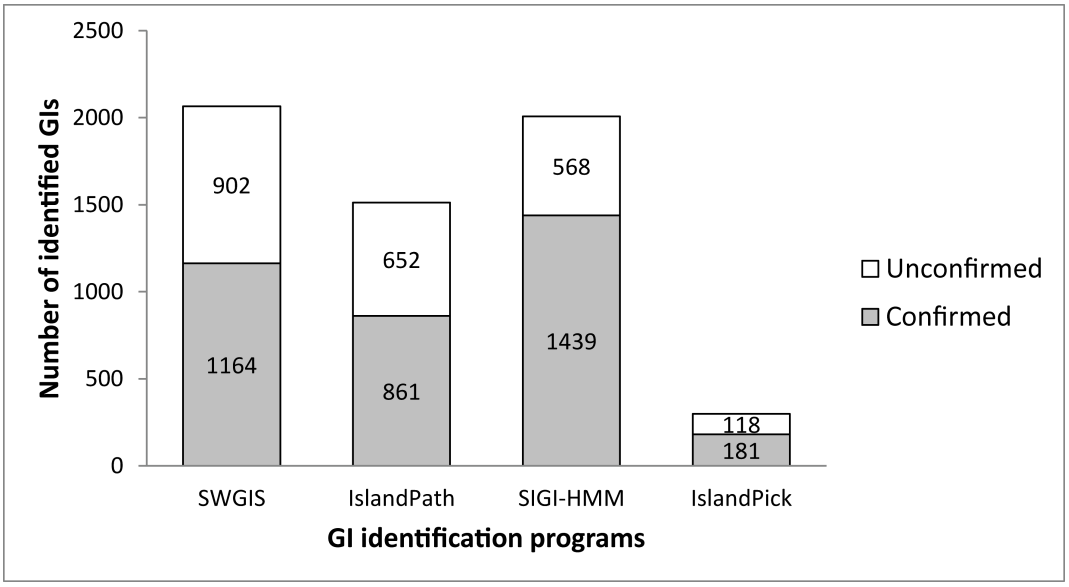


Figure 22: Counts of islands predicted only by one of four programs (unconfirmed) and confirmed by the others. Confirmation is obtained when two or more islands predicted by different programs at least partly overlapped.

Validation of these predictions was approached with an island gene content method. Predictions were inspected for inclusion of genes associated with MGE, *e. g.* "integrase", "phage", "IS-element" *etc.*, to justify island identification. The absence of MGE associated genes in a prediction should not be regarded as an indication of a false positive due to incorrect annotation, fragmentation of islands and others. It has been observed that in a collection of 1,252 true positive islands only 56% of them contained MGE identifier words and as such defined as "key positives" [77]. Therefore the amount of true positives amongst unconfirmed islands was estimated as:

Number of unconfirmed key positive islands $\times \frac{100}{56}$

FPR were estimated for the SWGIS set of islands predicted in the 169 genomes as mentioned above and presented in Table 1.

Estimated false positives were calculated as:

Unconfirmed $-$ Unconfirmed key word positive $\times \frac{100}{56}$

FPR rates were calculated as:

Estimated false positives $\div$ Number of islands predicted

FNR rates were calculated as the fraction of known islands that were not identified by SWGIS.

Table 1: Estimated FPR and FNR for islands predicted by SWGIS with different parameters.

SWGIS Parameter	[D:1.5; V:1.5]	[D:1.5; V:2.0]	[D:2.0; V:1.5]	[D:2.0; V:2.0]
Number of islands predicted	2,066	928	1,571	809
Unconfirmed	902	280	545	188
Unconfirmed key word positive	137	44	92	28
Estimated false positives	657	201	381	138
FPR	0.318	0.217	0.243	0.171
FNR	0.118	0.353	0.216	0.373

SWGIS is a robust and reliable island identifier yet it is acknowledged that it is not perfect. The process of amelioration burdens all island prediction methods and SWGIS is no exception as it fails to identify old or ancient inserts that have undergone amelioration to such an extent that their genomic signature reflects that of the host they reside in. Furthermore SWGIS is biased to identify only relatively large fragments that exceed 5 kbp.

2.6 Discussion

Genome linguistics, implemented by compositional comparison, of bacterial genomes is a valuable approach to handle large scale archaeal/bacterial comparative genome projects. Various computational tools based on compositional similarity have been proposed and applied in recent times [81, 25, 46, 98]. This method provides a reliable alternative in the identification of islands in complete prokaryotic genome sequences. SWGIS utilizes OU statistics which is comparable to other composition-based approaches for island identification such as SIGI-HMM [21] and GOTHAM [93]. Analysis of an archaeal/bacterial genome is completed within 5-10 minutes on average by SWGIS and as such it is appropriate for large scale identification of islands in multiple genomes.

Optimization of SWGIS parameters was enabled by the use of FNR/FPR statistics and case studies performed to identify failures and provide possible scenarios to resolve these failures. Factorial analysis of the proposed island identification algorithm provides users with reliable information on selecting customizable parameters to ensure acceptable FNR and FPR.

After island boundary prediction the LingvoCom package allows for the further analysis of islands. This package is suitable to infer possible phylogenetic relationships and may be used to determine probable origin and age of islands. The application of alignment-free composition-based genome comparison methods for phylogenetic inference and clustering has been ascertained [24, 73, 93]. The greatest obstacle affecting all genome linguistic

inferences including those conducted by LingvoCom is the absence of an adequate evolutionary model to clarify the changes in local and global OUP found in different time scales. The only model for the amelioration of bacterial DNA proposed by Lawrence and Ochman in 1997 [119] is rather basic and was shown not to be sufficiently accurate [154]. The absence of a reliable evolutionary model limits the use of composition-based approaches. This being said the comparison of OUP provides a possible approach in the reconstruction of donor-recipient relationships as encountered in the island community.

Comparison of different island identification approaches is rather precarious and disjoint. This disparity results from the flexibility of HT, which occurs through various mechanisms, and post-insertion pressures of fragmentation and amelioration. The ability of a method to correctly identify an island further depends on the length of an island, the genetic content and duration of stay in host genome. The ability to distinguish and remove false predictions is of utmost importance in the identification of islands. Not all genomic regions displaying alternate DNA composition to the rest of the genome are of foreign origin and some regions displaying compositional similarity to the host may be of lateral origin yet no calculations are available for FPR and FNR encountered with island prediction methods.

SWGIS exploits OU statistical parameters to identify islands and atypical genomic regions and employs sequence similarity searches against a database of 16S rRNA to discard *rrn* operons. These methods inevitable increase the percentage of unidentified islands and as such FNR. Nevertheless, SWGIS may significantly contribute to island identification and reveal regions of HT that have remained undetected by other island identification programs.

3 Chapter 3: Database construction, maintenance and expansion

> If you go far enough back, your genome connects you with bacteria, butterflies, and barracuda - the great chain of being linked together through DNA.
> Spencer Wells

The continued prediction and analysis of novel islands in newly sequenced archaeal/bacterial genomes by an optimized identifier is of little importance if this information is not kept in a reliable and accessible format. The need for a scalable and expandable database to house islands predicted by SWGIS and as such enable further research on predicted islands was evident. The construction of a database should be approached with caution and logic in order to develop a biologically valuable and lasting tool.

The NGS age brings forth the ability to speedily and affordably sequence genomes. This possible wealth of future islands requires the advent of novel techniques in the construction of a database. The computational intensity and resources required for certain island information generation desires the implementation of these new methods to ensure an expandable and contemporary database.

Development of a convenient GUI to access the database is critical in the production of a relevant island analysis repository. This may ensure the extensive use in possible future HT and island research and as such the biological application of the database.

3.1 Database Blueprint

MySQL is an open-source relational database management system (RDMS) and provides the backbone to the database. This system enables the storage of records in a memory efficient and structured manner. The power of this system is fully utilized by means of relationships between different tables or sets of records. The database blueprint is displayed in Figure 23. Tables are represented by each block and fields within tables are given. It should be noted that all tables are connected. This ensures that a "dead-end" is not reached in query and that all information included is inter-connected. This design ensures optimal accessibility and speed in the extraction of records with all information available if needed due to relational tables.

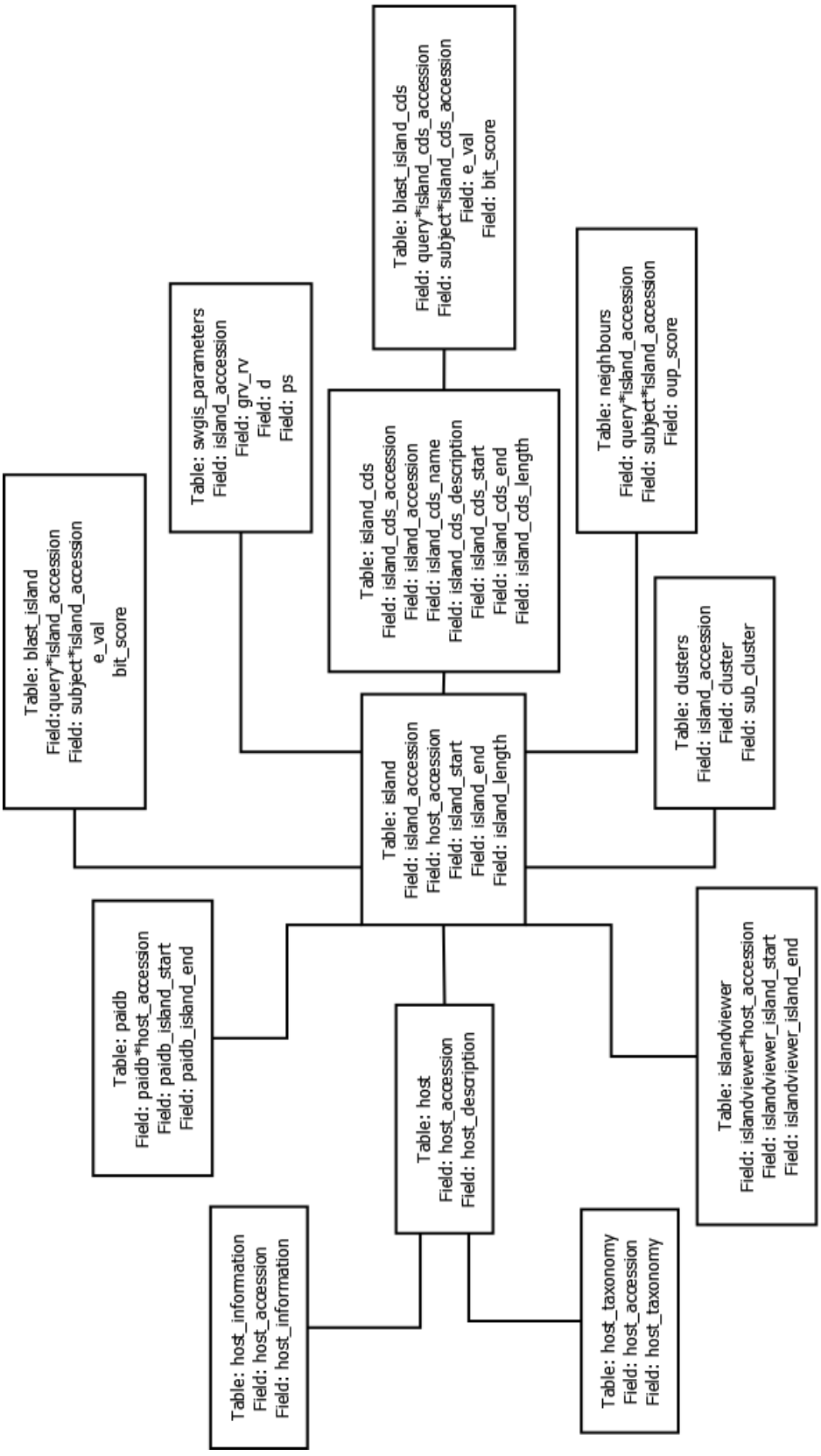


Figure 23: Schematic representation of MySQL database structure.

3.2 A Table in every Room

3.2.1 Table: host

The host table contains fields for host accession and host description. Archaeal/bacterial NCBI accession numbers are unique and as such is utilized as an identifier for each host within which island(s) were identified. The individuality of these accession numbers furthermore provide an ideal key in establishing relations with other tables. Host description refers to the host organism name as provided by the NCBI for specific host accession.

Relations:

- Table: host_taxonomy
- Table: host_information
- Table: island

Key:

- host_accession

3.2.2 Table: host_information

General information regarding host organism habitat, lifestyle, virulence and other traits is stored here and linked by means of the unique host accession.

Relations:

- Table: host

Key:

- host_accession

3.2.3 Table: host_taxonomy

Identified by unique host accession number this table provides detailed taxonomic information for each host.

Relations:

- Table: host

Key:

- host_accession

3.2.4 Table: island

This table records information on each predicted island. Each predicted island is given a unique identifier by means of combining the unique host accession with the island start position. SWGIS predicts islands with clear boundaries and as such it is highly unlikely that different islands will display the same exact lower boundary or start position. The combination of host accession and island start position enables the use of a unique island identifier. This table includes host accession as a means of relationship with the host table. Island global positioning is available by island_start, island_stop and the length through island_length.

Relations:

- Table: host
- Table: paid
- Table: islandviewer
- Table: blast_island
- Table: swgis_parameters
- Table: island_cds
- Table: neighbours
- Table: clusters

Key:

- host_accession
- island_accession

3.2.5 Table: paidb and islandviewer

The identification of overlaps for an SWGIS island to other databases is established with these tables. It is possible to determine overlaps of predicted islands by means of island locations and host accession, all of which are readily available.

Relations:

- Table: island

Key:

- host_accession

3.2.6 Table: blast_island

Sequence similarity results obtained by BLASTN for the entire length of a query island will be housed here. This information includes subject island and statistics on the alignment. The inclusion of unique subject island identifier allows for the rapid extraction of information regarding the subject hit if the need arises.

Relations:

- Table: island

Key:

- island_accession

3.2.7 Table: swgis_parameters

Parameters as calculated for each island is stored in this table.

Relations:

- Table: island

Key:

- island_accession

3.2.8 Table: swgis_parameters

Parameters as calculated for each island is stored in this table.

Relations:

- Table: island

Key:

- island_accession

3.2.9 Table: neighbours

Compositional similarity hits as calculated by OUP similarity between query and subject islands is available from here. This includes unique subject accession and hit score in percentage.

Relations:

- Table: island

Key:

- island_accession

3.2.10 Table: cluster

Cluster and subcluster information for each island is contained in this table.

Relations:

- Table: island

Key:

- island_accession

3.2.11 Table: island_cds

Contents of each predicted island is instilled in this table to aid in the retrieval of island information. This includes a unique island cds identifier. The cds identifier is based on the conjunction of an unique island identifier and unique cds start location. This table includes the island accession as a rapid means of retrieving island information. CDS locations and lengths are further included, in addition to cds name and description.

Relations:

- Table: island
- Table: blast_island_cds

Key:

- island_accession
- island_cds_accession

3.2.12 Table: blast_island_cds

Sequence similarity between genes in islands to be maintained here. This is inclusive of a unique gene identifier to retrace subject to the island it was found to reside in and BLASTP alignment statistics.

Relations:

- Table: island_cds

Key:

- island_cds_accession

3.3 Floating foundation

Database construction defined as above allows for the extension with newly identified islands as the design is robust yet simple. The incorporation of unique identifiers for host, island and island cds enables the exact pinpoint of a record to be established and circumvents the problem of duplicates. This design proposes to be an efficient retrieval construction with the maximum amount of information with the minimum amount of resources used. Database construction as detailed above allows for critical maintenance of current records and the addition of novel islands to ensure a contemporary and dynamic analytical tool in the field of prokaryote HT and MGE.

3.4 Maintenance and Expansion of Database

The age of affordable and agile sequencing requires a dynamic database. The ability to adapt to new data is of the essence and unwillingness or ineptness to update information is regarded as a critical shortfall for biological databases [156]. The great wealth of sequence information at the fingertips of researchers requires further analysis and deposit in an interest-bearing information bank. It is not just the prediction of islands but the furthered analysis of these MGE that will enlighten the biological reasoning and sustained application of HT in the prokaryotic realm.

Current sequencing capabilities delivers reliable genomic information in rapid time spans. These advances produce copious amounts of data to be analyzed and as such require novel approaches in the gathering and storage of relevant material. The calculation of all relationships and ontologies for segments of information further complicates advances as data is produced faster than what complete analysis can handle.

SWGIS is a robust and efficient island predictor capable of archaeal/bacterial genome analysis in 5-10 minutes. The continued analysis of these islands is computationally intensive and requires certain modifications to ensure the completion of thorough investigation with a reasonable period. This is of utmost importance in the update of the database to allow for the incorporation of newly acquired information before it is deemed redundant. The pipeline followed in database maintenance and expansion is described in Figure 24 below.

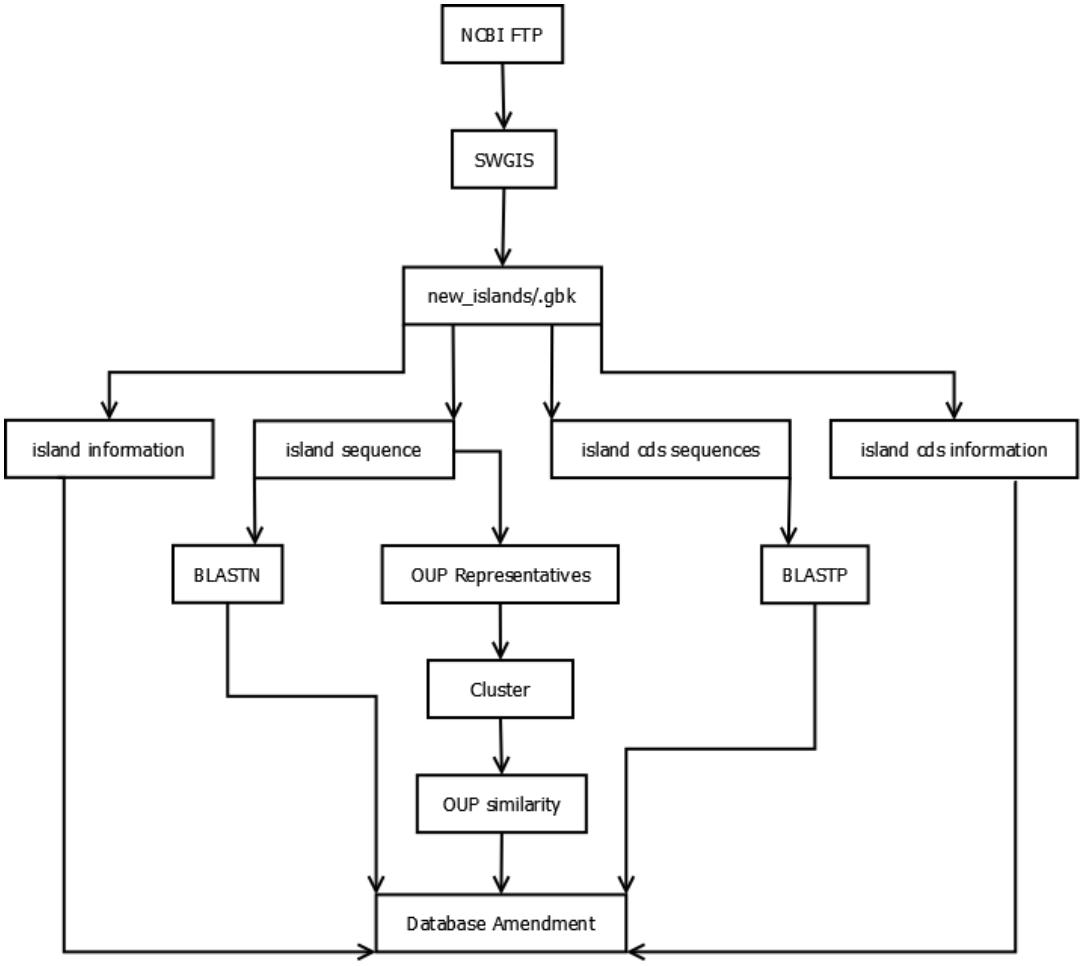


Figure 24: Schematic representation of database expansion with newly identified islands.

3.4.1 Novel island collection

Whole archaeal/bacterial genome sequences are readily available for download from the NCBI FTP site (ftp://ftp.ncbi.nlm.nih.gov/). This allows for the rapid retrieval of multiple reliable samples with which further analysis is possible. The collection of GenBank file format sequence data is preferred as this format enables SWGIS island prediction to include vital information in the identified island output files.

3.4.2 SWGIS island identification

SWGIS allows for the prediction of multiple islands in numerous genomes simultaneously. It is therefore not required to analyze each genomic sequence individually and thus multitudes of island identifications are possible in a relatively short period of time. Each island is represented by a GenBank file and this is all that is needed for expansion as all required information is available in a single file.

3.4.3 Island file dissection

Each island file is parsed to extract all relevant information which is to be added to the database. This includes unique identifiers for the island itself and all genes present within identified segment including locational information on all. Various comparisons regarding sequence and composition of island and genes within islands follow and therefore all sequence information is extracted and equipped with a unique identifier.

3.4.4 Island and cds information

All island information regarding host organism is determined. Island and cds locational information is extracted and therewith a unique identifier(s) ascertained. The use of GenBank file format enables SWGIS to include cds annotation and as such this information is available for extraction. This file format further incorporates SWGIS parameter values in the production file to be parsed for database amendment. These values are used in future donor-recipient predictions and is required for proposed island flow information.

3.4.5 Island and cds sequences

The addition of an island and all relevant information is of little importance if sequence comparison and analysis is not incorporated. All sequences for island and genes within an island is therefore extracted to be used in comparison with existing sequence data. This is accomplished in collaboration with unique identifiers for all to ensure reliability and precision in analysis.

3.4.6 Island and protein sequence comparisons

After sequence extraction it is possible to conduct sequence comparison on the island sequence and the protein sequences contained within an island. The unique identifiers allows for the retrieval of alignment statistics. BLASTN is used for the island sequence comparison to all island sequences currently housed in the database and BLASTP for cds

sequences against all proteins in the database. The inclusion of an e-value limit of 10^{-6} ensures the production of relevant comparison information. All results are automatically parsed to remove redundancy and ensure reliability.

3.4.7 Island compositional comparison

Island sequence comparison against the entirety of the database is accomplished with relative ease in appropriate time. Compositional comparison of newly identified islands against the totality of the database proved to be more arduous. Time expenditure was found to be excessive and as such a novel approach identified.

3.4.8 Island clustering and representatives

Clustering of copious amounts of elements into significant and related groups allows for numerous advantages and optimizations when working with big data. Islands were clustered with the Markov Clustering Algorithm (MCL) [15] and OUP similarity hits as a relational score. MCL is a suitable clustering algorithm for islands information as non-overlapping clusters are produced by this graph-based, deterministic, partitional algorithm that incorporates hard clustering. Random compositional similarity links between islands were removed by the implementation of a floor threshold of at least 75% OUP similarity. Biased over-representation of duplicate islands in closely related species was curbed by a ceiling threshold of 85% OUP similarity. The second threshold chiseled the original 69,176,627 OUP similarity hits above 75% to 62,670,254 OUP similarity hits between 75% and 85%. MCL constructed 34 clusters with no singletons. Islands with no OUP similarity and thus not included in the clustering where pooled into a cluster thus resulting in 35 clusters from 62,670,254 relational scores between 26,744 islands. Clusters with more than 50 islands were deemed large and subclustered with MCL to ensure individuality and distinctiveness of clusters and islands incorporated in a cluster. The first six clusters contained more than 50 elements and subclustering produced 134 subclusters and therefore a total of 163 individual clusters. Representatives or captains for each cluster/subcluster were deemed vital for probing compositional similarity searches through the database and to aid with amendments and additions to the extensive amount of island information. The node in each cluster/subcluster with the maximum amount of edges was designated as the cluster/subcluster representative, thus the island with the utmost OUP similarity hits between 75% and 85% to other island in a specific cluster/subcluster. Multiple representatives were required for large clusters as diverse members of a cluster/subcluster may not display any OUP similarity links. As such all clusters were inspected to determine islands not showing OUP similarity hits to the representative in said cluster. Additional cluster/subcluster representatives were deter-

mined for these isolated islands to ensure all members of a clustering is represented. This resulted in an omnipresent list of representatives for all clusters/subclusters to aid and direct searches and comparisons. General statistics regarding the taxonomic composition of clusters/subclusters are available for further research on the bias of certain islands in certain groups. This information may indicate the biologically relevant reason on why specific islands are grouped together and seem to be attracted to each other.

This method of OUP hit determination, cluster designation and database amendment is summarized in Figure 25.

A newly described island is first compared to a set of cluster/subcluster representatives to determine OUP similarity hits. This will indicate with which cluster(s)/subcluster(s) elements the novel island will be compared and as such alleviates the need to compare an island to all elements in the database. Cluster/subcluster representative hit is followed by compositional comparison of an island to all members of a cluster/subcluster by means of OUP similarity searches. This may include comparison to multiple clusters/subclusters. Compositional similarity hits are inspected for redundancy and novel island compositional comparison and cluster/subcluster information added to the database.

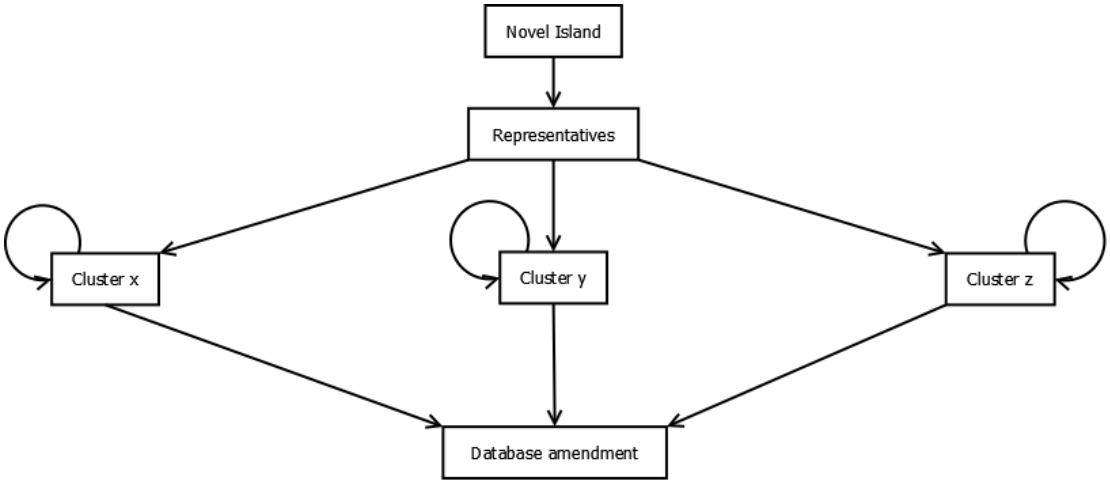


Figure 25: Novel island inclusion in the database aided by precomputed island cluster representatives.

3.5 GUI development - open house for viewing

Database construction and information integration requires a convenient and user-friendly interface. This was developed primarily in PHP with Python, HTML and command-line collaborations. The structure was based on the LAMP (Linux, Apache, MySQL and PHP) model which is suitable for the production of dynamic web sites. This model allows for the production of request pages "on the fly" and as such the amount of precomputed information is drastically diluted. Information is linked in such a fashion to allow users

an informative and pleasant research opportunity with accessible and detailed functions. Help pages and tool tips are incorporated to aid and direct users with page development. Great care was taken in the development of a web-resource free of stalemate pages.

3.6 Discussion

The construction of a biological database should be logical and expandable. The age of rapid and affordable sequencing bears the burden of vast amounts of data to be investigated and incorporated. The design of a database to house current content and absorb new knowledge should be simplistic yet robust.

The database blueprint required the construction of various interlinking tables and identifiers to extract information in a precise and speedy fashion. This included the assignment of unique identifiers and keys across the board. The maintenance and expansion of the database is subject to the initial design and requirements.

The insertion of island information requires the processing of records to extract that which is deemed necessary. Records undergo analysis and investigation prior to insertion in the database. Computational and time constraints required the implementation of novel techniques in the processing of island information. The sustained expansion of the database is necessary to ensure longevity and vitality in the biological arena.

A web-based interface was developed to allow users access to all available data. The interface serves as a user-friendly portal to island information and analytical tools.

4 Chapter 4: Pre_GI

> What you see is that the most outstanding feature of life's history is a constant domination by bacteria. Stephen Jay Gould

The Predicted Genomic Islands database (Pre_GI) is a detailed and extensive archive of prokaryotic islands freely accessible at http://pregi.bi.up.ac.za. The web-resource is aimed to enlighten users on island existence, similarity and history. The production of an alternative tool to investigate and deconstruct island movement is long overdue in the field of mobilomics and HT. This contemporary MGE database allows for the analysis of probable fluxes across taxonomic and environmental borders with the added ability to compare novel islands to the horde of information available in this package.

Pre_GI contains 26,744 probable islands from 2,407 archaeal and bacterial chromosomes and plasmids identified by SWGIS using default parameters. Database search and filter options are available for users to collect specific islands or browsing in general. These fields include NCBI accession number for host of island, host strain description, host taxonomic data and general information with regards to host habitat, isolation, peculiarity *etc.* Islands may be searched with regards to their gene content to produce lists of islands containing said genes with a specific gene annotation. All island host NCBI accession numbers are hyperlinked to retrieve all information regarding predicted islands in host of choice. Detected islands are graphically displayed in a rendered host page by means of a SVG graph. Island host lineage and general information is provided to assist the deconstruction of the HT event by defining the natural behavior, location and characteristics of the host. All visualized islands associated with the chosen host are listed below the graph together with locational information, SWGIS determined identifier parameters and inclusion of "key words" in gene annotation. Islands displaying genomic location overlaps to other island databases are indicated with links to predicted islands in the other database. Key word evidence and overlap to other databases using alternative island prediction methods increase probability of an island to be considered a true positive. Islands containing ribosomal proteins are marked as quite often areas containing ribosomal proteins are characterized by an alternative OUP and deemed to be of horizontal origin by SWGIS. These islands are incorporated as it has been shown that ribosomal HT cannot be excluded [116]. Compositional and sequence comparison hits of an island are available and in addition all sequence similarity hits of genes contained in the island. Island gene content allows for searches against the database for similarly annotated genes in other islands with an added link to the QuickGO browser for further information regarding gene ontology and annotation. All island sequences and gene sequences are readily available for download in FASTA format.

4.1 Current Content

4.1.1 Browse

Host

Pre_GI contains 2,407 host archaea/bacteria in which 26,744 islands are predicted. This is represented in host list from which the user is able to make a selection. The host list may be filtered by various fields in order to narrow the selection. These filters include host accession, host description, host lineage and host information and may be used in parallel. This enables the user to choose a specific organism of interest with specified characteristics. All host accessions are hyperlinked to retrieve identified islands residing in the organism of choice.

Host lineage and information

The incorporation of archaeal/bacterial host lineage and general information in collaboration with island sequence and compositional comparison results contributes a novel movement in HT and MGE research. Comprehensive island analysis is helped by host habitat and isolation information to provide a detailed picture of island ontology and origin. This added information may assist research in the logic of HT by including the natural behavior, location and characteristics of island hosts. Taxonomic tallies were calculated for the entire database and clusters/subclusters which may aid research on the propensity of islands in certain taxonomic groups.

Islands

A chosen host accession results in the display of all islands identified in said host. Host genome atlas graphically displays an island as a pink box. Hovering over a specified island on the SVG atlas displays the location of the island in the host sequence. The legend below the SVG atlas details information on the graph in conjunction with a definition box regarding the SWGIS parameters. Host taxonomy and general information together with a link to the NCBI with regards to the host accession is available to users below the atlas.

Predicted islands are listed and those containing ribosomal proteins are marked with an asterisk. Island location is displayed as start and stop with the start location hyperlinked to display all information regarding the content of chosen island. The file information for each island is available for display in the browser or as a download.

All SWGIS parameter results for each island is displayed with the added functionality of searching the database for islands with similar parameters. GRV is defined as the

globally normalized relative variance of the OU, the host, and RV as the relative variance of the OU, the island. An increase in the divergence between GRV and RV indicates probable HT. This divergence is presented by GRV_RV. Islands with a similar GRV_RV parameter may be searched by the hyperlink on this parameter. The parameter D details the distance between the host and island OUP. This value is indicative of the proposed age of an island as the process of amelioration alters foreign genomic fragments to resemble that of the host. Islands with a large D parameter are proposed to be recently acquired whereas those with a lower D parameter subjected to amelioration for a protracted period of time. It is possible to search the database for islands with a similar "age" by means of the hyperlinked D parameter value. The SWGIS parameter PS indicates the distance between the patterns of the direct and reverse strands of the same DNA sequence and allows for comparison of this parameter to other islands in the database through the PS parameter value hyperlink.

Compositional similarity of an island to the entirety of the database is accessible from the Neighbours link. These hits were obtained by means of an all-against-all compositional similarity search with floor threshold of 75%. Island of interest's taxonomy and general information is displayed and the subject hit list may be filtered to narrow or specify compositional similar subjects. Proposed donor-recipient movement is displayed and it is possible to retrieve all donors or all recipients from the subject hit list.

Cluster/subcluster details the elements of a grouping with which a chosen island aligned. Non-overlapping, distinct clusters were identified by means of the MCL and OUP similarity between islands as a relational score. Cluster/subcluster statistics on taxonomy of elements in the grouping is available for inspection. Donor-recipient movement for chosen island against the cluster/subcluster is identified with the aid of the SWGIS parameter D. It is possible to filter elements of a grouping by host characteristics or proposed donor-recipient flow to magnify specific relationships.

Sequence similarity in the form of BLASTN results with an e-value threshold of 10^{-6} is accessible through the hyperlink. Results include e-value and bit score in combination with the option to visualize the resulting alignment. Results may be filtered by subject characteristics or ordered by a column heading. The sequence similarity visualization provides users with either a BLASTN or BLASTP graph where it is possible to identify elements of high scoring sequence similarity by mousing over area of interest.

Key word confirmation is indicated when an island contents include MGE-associated gene(s). These gene annotations include transport, transposon, transposable element, transposase, integrase, is-element, phage and relaxase. Overlaps of an SWGIS-predicted island with that found in PAIDB and/or IslandViewer will be displayed and provides the option of accessing the database with which an overlap was identified. These identifiers increase the probability of a true positive island prediction.

Genes

Island content is available from the hyperlinked start location of the island. This includes gene content, gene location and annotation. CDS descriptions are hyperlinked to retrieve all similar descriptions from the database. This aids users in searching for other islands with a specific gene or cds annotation. All annotations are linked to the QuickGO site to retrieve gene ontology. Sequence similarity hits with e-value threshold of 10^{-6} for a protein of interest is accessible through the BLASTP link.

Island and gene comparison statistics

Compositional comparison for all-versus-all predicted islands was achieved by means of OUP calculations. The determination of pattern similarity is calculated by means of comparing 4-mer frequencies between 2 predicted islands. The inclusion of essential compositional similarity was administered by the implementation of a floor cut-off 75% OUP similarity. This ensured the probability of common ancestry or at least an involvement in common reticulation events [78]. All-against-all OUP similarity comparison resulted in a total of 69,176,627 compositional similarity hits with the inclusion of a floor threshold of 75%. Compositional similarity hits for each island is available together with percentage OUP similarity between query and subject.

Sequence comparison between all islands for the entirety of the island length was calculated with BLASTN with BLASTP employed to determine sequence similarity between all cds sequences included in all islands. Both BLASTN and BLASTP was performed with an e-value of 10^{-6} to enable true prediction of homology. All-against-all BLASTN comparison for the total length of the island sequence produced 3,692,401 hits and all-against-all BLASTP sequence similarity for genes predicted in all islands resulted in 138,590,509 hits all adhering to an e-value threshold of 10^{-6}. All island-to-island sequence comparison hits are available as graphical representation and as text output.

Combining compositional and sequence similarity links between islands with the added information on lineage and general lifestyle of the hosts they are detected in opens a novel avenue in HT and MGE research that incorporates biologically relevant information to deduce reason and logic for transfer.

Representatives

Representatives for all clusters/subclusters were deemed vital for database expansion and novel island comparison. These individuals were assigned as the element in a cluster/subcluster with the majority of compositional similarity hits to all other elements.

Clusters/subclusters were further inspected for elements not displaying a direct relationship to the representative and extra individuals identified to represent a cluster/subcluster to ensure an omnipresent list of islands detailing relationships to the majority of elements. Therefore certain clusters/subclusters may include more than one representative. The list of representatives is available for inspection and each individual hyperlinked to browse island content.

Gene annotation

The ability to search all islands for specific gene annotations or phrases provides the user with an novel tool. This enables users to identify islands based on specific genetic content and annotation rather than the host they reside in. Descriptions are linked to the QuickGO browser from EMBL-EBI (http://www.ebi.ac.uk/QuickGO/) for gene ontology retrieval. It is as such possible to identify all islands containing efflux proteins or beta-lactamases.

Island location

Pre_GI can be searched by means of user provided genomic locations. Current content of the database may be quizzed by means of location in a host to indicate presence of an island in a defined prokaryotic segment. Overlap of newly predicted islands with existing or presently available will be indicated to the user. This method is furthermore appropriate in the detection of genes of interest in segments pertaining to islands.

Island taxonomy statistics

General statistics regarding the taxonomy of hosts included in the database is provided. This enables introspection on the content of the database with regards to frequencies and distribution of taxonomic groupings. These statistics are further available for clusters/subclusters. Examples of these taxonomy statistics for domain and phylum are presented in Table 2 and Table 3 respectively.

Table 2: General statistics with regards to island host domain content of Pre_GI.

Domain	Number of hosts	Percentage of Pre_GI	Number of islands	Percentage of Pre_GI
Archaea	166	6.90 %	1,861	6.96 %
Bacteria	2,241	93.10 %	24,883	93.04 %

Table 3: General statistics with regards to island host phylum content of Pre_GI.

Phylum	Number of hosts	Percentage of Pre_GI	Number of islands	Percentage of Pre_GI
Acidobacteria	11	0.46 %	127	0.47 %
Actinobacteria	235	9.76 %	3,171	11.86 %
Aquificae	13	0.54 %	139	0.52 %
Bacteroidetes	90	3.74 %	1,245	4.66 %
Caldiserica	1	0.04 %	4	0.01 %
Chlamydiae	38	1.58 %	121	0.45 %
Chlorobi	11	0.46 %	95	0.36 %
Chloroflexi	21	0.87 %	155	0.58 %
Chrysiogenetes	1	0.04 %	18	0.07 %
Crenarchaeota	41	1.70 %	603	2.25 %
Cyanobacteria	71	2.95 %	613	2.29 %
Deferribacteres	5	0.21 %	58	0.22 %
Deinococcus-Thermus	24	1.00 %	150	0.56 %
Dictyoglomi	2	0.08 %	23	0.09 %
Elusimicrobia	3	0.12 %	22	0.08 %
Euryarchaeota	119	4.94 %	1,217	4.55 %
Fibrobacteres	1	0.04 %	24	0.09 %
Firmicutes	423	17.57 %	5,429	20.30 %
Fusobacteria	7	0.29 %	107	0.40 %
Gemmatimonadetes	1	0.04 %	8	0.03 %
Korarchaeota	1	0.04 %	3	0.01 %
Nanoarchaeota	1	0.04 %	4	0.01 %
Nitrospirae	3	0.12 %	43	0.16 %
Planctomycetes	7	0.29 %	70	0.26 %
Proteobacteria	1,139	47.32 %	12,241	45.77 %
Spirochaetes	61	2.53 %	565	2.11 %
Synergistetes	3	0.12 %	25	0.09 %
Tenericutes	47	1.95 %	233	0.87 %
Thaumarchaeota	2	0.08 %	17	0.06 %
Thermobaculum	2	0.08 %	15	0.06 %
Thermodesulfobacteria	2	0.08 %	12	0.04 %
Thermotogae	14	0.58 %	125	0.47 %
unclassified Archaea	2	0.08 %	17	0.06 %
unclassified Bacteria	1	0.04 %	4	0.01 %
Verrucomicrobia	4	0.17 %	41	0.15 %

4.1.2 Browsing example

An outbreak of *Escherichia coli* O157:H7 in Sakai City, Osaka, Japan was attributed to the consumption of white radish sprouts from one particular farm [47]. This enterohemorrhagic *Escherichia coli* was spread by lunch foods supplied to the elementary schools in Sakai and caused 121 cases of hemolytic uremic syndrome (HUS) in 12,680 symptomatic

patients with 3 fatalities [45]. *Escherichia coli* O157:H7 is a major food-borne infectious pathogen causing diarrhea, hemorrhagic colitis and HUS.

Pre_GI was tasked to identify and investigate islands predicted in the complete genome of *Escherichia coli* O157:H7 str. Sakai [NC_002695]. The list of 2,407 possible hosts were filtered with host description, host lineage and host information to retrieve the desired host for further investigation (Figure 26). Host description was chosen as "Escherichia coli", host lineage selected as "Proteobacteria" from the available list and the field host information was chosen as "Japan". This resulted in 2 options available to choose from, namely *Escherichia coli* O157:H7 str. Sakai plasmid pO157 [NC_002128] and *Escherichia coli* O157:H7 str. Sakai, complete genome [NC_002695]. *Escherichia coli* O157:H7 str. Sakai, complete genome [NC_002695] was selected for further analysis.

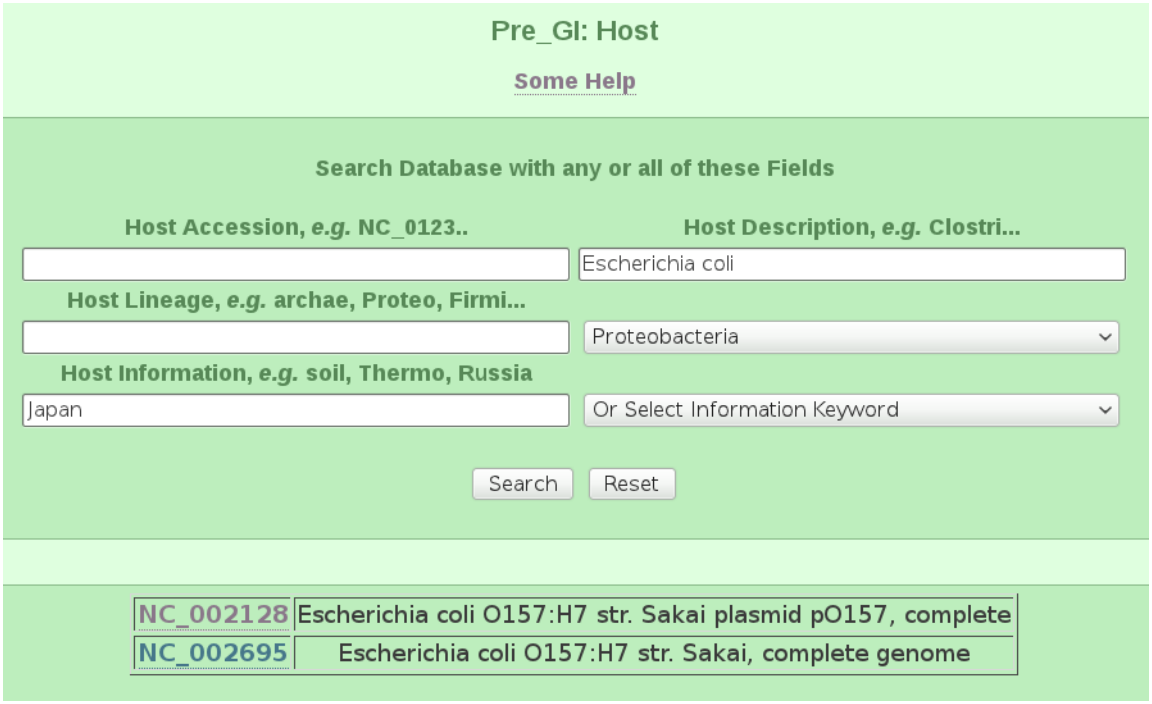


Figure 26: Pre_GI island host. Elements in the host list were filtered for a host description "Escherichia coli", host lineage "Proteobacteria" and host information "Japan" to obtain the desired host organism.

The genome atlas in Figure 27 of *Escherichia coli* O157:H7 str. Sakai, complete genome [NC_002695] indicates the location of identified islands in combination with SWGIS parameter deviations. No falsely selected *rrn* operons were present and a total of 29 islands identified. General information regarding host lineage and host information/characteristics are available to users with a hyperlink to the NCBI for further information regarding host sequence (Figure 28).

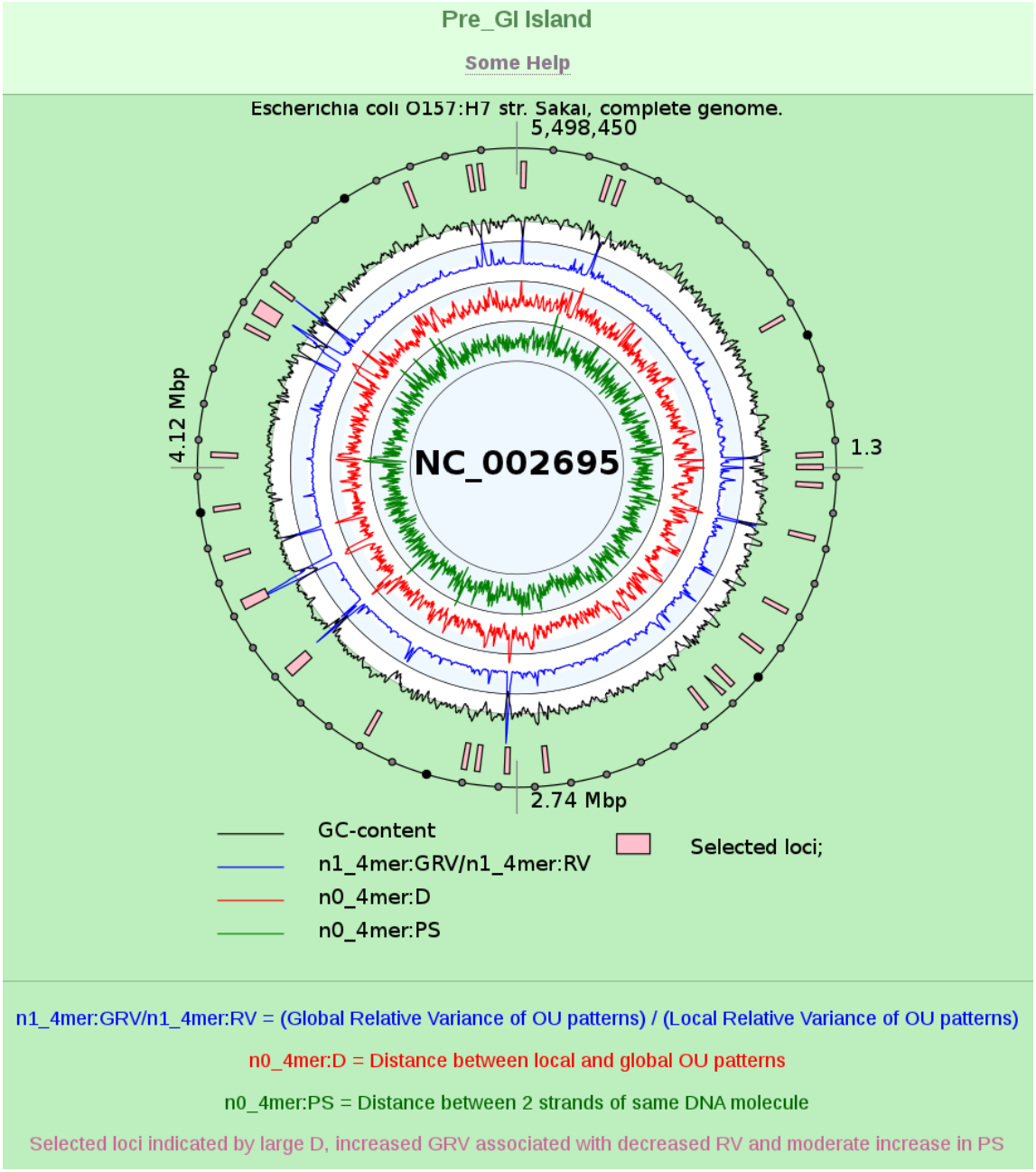


Figure 27: SVG genome representation with islands indicated by pink blocks in the periphery. The legend below the atlas describes the SWGIS parameter lines.

The genome atlas allows for the identification of island location by mousing over a selected island indicated in a pink block. SWGIS parameter descriptions below the graph may aid users interpretation. The example above contains no falsely selected *rrn* operons, false positives, as these would be indicated by grey blocks in SVG genome representation. This graphs indicates a fairly even distribution of islands across the *Escherichia coli* O157:H7 str. Sakai genome.

Figure 28: Host taxonomy and general information is freely available in combination with a hyperlink to the NCBI regarding host organism.

All 29 identified islands are displayed in a list below the host information. Islands containing ribosomal proteins or RNA related elements are marked with an asterisk to indicate to users that these predicted islands may be false positive selections. A section of this list is presented in Figure 29. Out of the 29 islands 10 were deemed to include ribosomal proteins or RNA related elements with only 2 of these islands not predicted by another independent prediction program. Key word confirmation was present for 18 of the 29 islands and other database (IslandViewer and PAIDB) confirmation was established for 25 of the 29 islands. The hyperlinks on database titles will reveal the island as available in the alternative island database(s).

Islands with an asterisk (*) contain ribosomal proteins or RNA related elements and may indicate a False Positive Prediction!

#	Start	End	Length	Island Text	GRV_RV	D	PS	Neighbours	Cluster	Sub Cluster	BLASTN	Key Word Confirmation	Other DB Confirmation	Download Island
1	12000*	35173	23174	Island text	2.90752	37.7161	29.0702	Neighbours	4	1	BLASTN		IslandViewer	12000.gbk
2	261235	283772	22538	Island text	1.5815	35.1562	22.0535	Neighbours	4	1	BLASTN		IslandViewer PAI DB	261235.gbk
3	301939	328730	26792	Island text	2.83691	39.811	23.6949	Neighbours	4	1	BLASTN	+	IslandViewer	301939.gbk
4	917509	940398	22890	Island text	1.76385	35.2195	38.3634	Neighbours	4	1	BLASTN		IslandViewer	917509.gbk
5	1329472	1353746	24275	Island text	3.03659	38.3438	24.5547	Neighbours	4	1	BLASTN		IslandViewer PAI DB	1329472.gbk
6	1364833	1389274	24442	Island text	2.22487	37.7023	26.6752	Neighbours	4	1	BLASTN	+	IslandViewer	1364833.gbk
7	1417912	1437301	19390	Island text	2.0738	31.3382	18.3153	Neighbours	4	1	BLASTN	+	IslandViewer PAI DB	1417912.gbk
8	1571389	1598158	26770	Island text	3.16705	41.3174	25.7377	Neighbours	4	1	BLASTN	+	IslandViewer	1571389.gbk
9	1794635	1816244	21610	Island text	1.86351	35.9924	33.8652	Neighbours	4	1	BLASTN	+	IslandViewer	1794635.gbk
10	1929215*	1950116	20902	Island text	2.47197	34.3172	16.9866	Neighbours	4	1	BLASTN		IslandViewer	1929215.gbk
11	2056545	2076939	20395	Island text	1.64893	32.5596	30.7818	Neighbours	4	1	BLASTN			2056545.gbk
12	2091890	2113599	21710	Island text	2.17314	32.6511	17.3416	Neighbours	4	1	BLASTN	+	IslandViewer PAI DB	2091890.gbk
13	2158314	2182071	23758	Island text	1.48961	36.4725	32.708	Neighbours	5	1	BLASTN		IslandViewer	2158314.gbk
14	2655445	2677869	22425	Island text	1.84131	34.1082	23.6703	Neighbours	4	1	BLASTN		IslandViewer	2655445.gbk
15	2769387	2799892	30506	Island text	4.46492	45.1761	19.758	Neighbours	4	1	BLASTN		IslandViewer	2769387.gbk

Figure 29: Pre_GI islands page displaying a section of the list containing 29 identified islands in *Escherichia coli* O157:H7 str. Sakai [NC_002695].

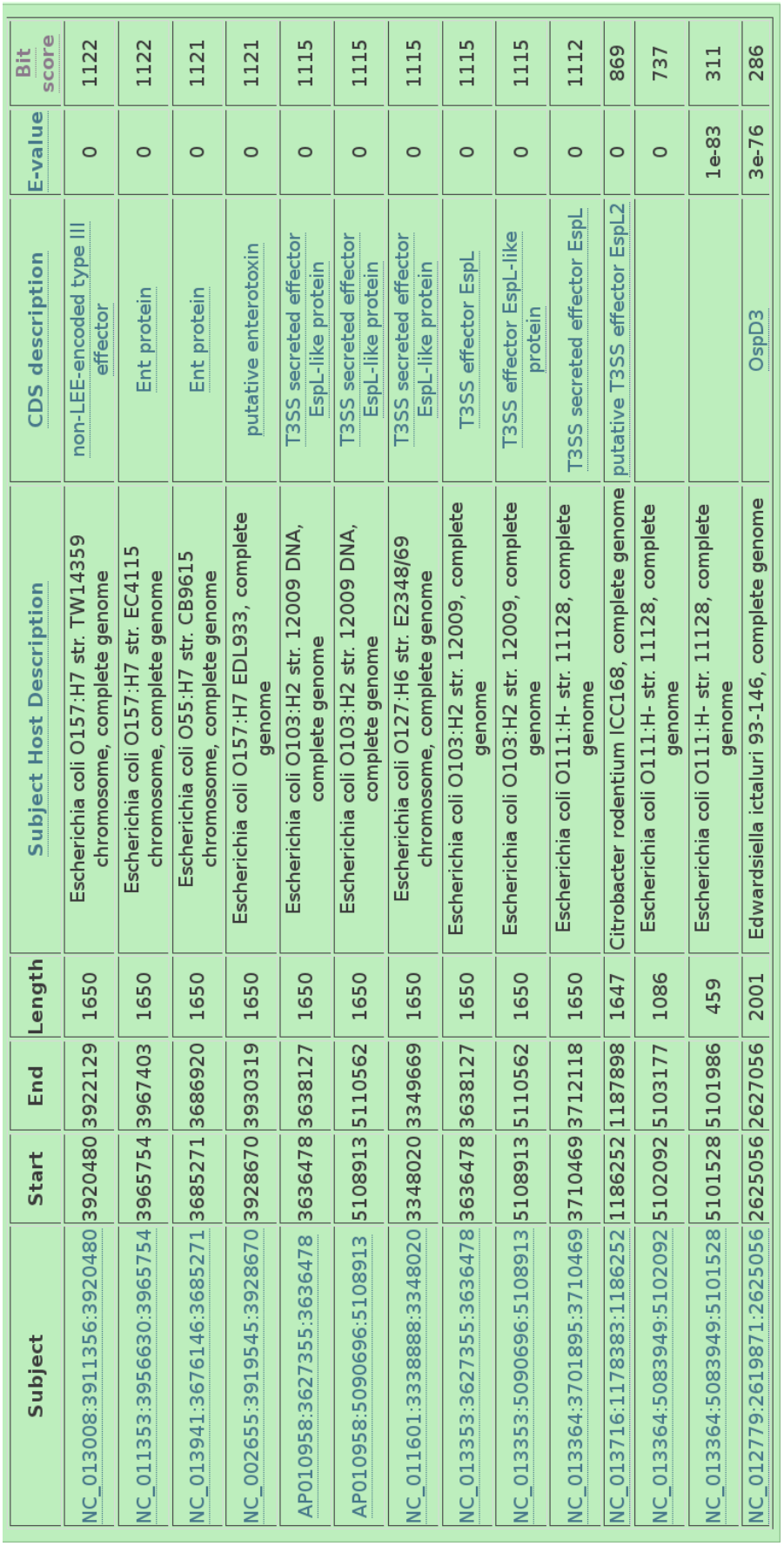
Island#21 identified in *Escherichia coli* O157:H7 str. Sakai displayed key word confirmation and indicated an overlap with IslandViewer. This island is located at position 3,852,233 - 3,878,876 and content is displayed in Figure 30 which includes an integrase and putative transposase in conjunction with a putative enterotoxin. Islands with a similar description to "putative enterotoxin", as retrieved by the cds annotation hyperlink, are presented in Table 4. BLASTP sequence comparison of the putative enterotoxin (Figure 31) reveals high scoring hits with various other *Escherichia coli* strains.

Start	End	Length	CDS description	QuickGO ontology	BLASTP
3852233	3853498	1266	integrase	QuickGO ontology	BLASTP
3854489	3855163	675	hypothetical protein		BLASTP
3855160	3855507	348	hypothetical protein		BLASTP
3855527	3856975	1449	hypothetical protein		BLASTP
3856776	3857720	945	hypothetical protein		BLASTP
3857829	3858377	549	putative virulence-related membrane protein	QuickGO ontology	BLASTP
3858950	3859150	201	hypothetical protein		BLASTP
3860922	3861149	228	hypothetical protein		BLASTP
3861082	3861276	195	hypothetical protein		
3861357	3863006	1650	putative enterotoxin	QuickGO ontology	BLASTP
3863614	3864603	990	hypothetical protein		BLASTP
3864652	3865326	675	hypothetical protein		BLASTP
3865761	3866549	789	hypothetical protein		BLASTP
3869375	3870265	891	putative transposase	QuickGO ontology	BLASTP
3870262	3870588	327	putative transposase	QuickGO ontology	BLASTP
3870594	3870710	117	hypothetical protein		
3870779	3871126	348	hypothetical protein		BLASTP
3871323	3872714	1392	hypothetical protein		BLASTP
3872782	3872940	159			BLASTP
3872937	3873287	351	hypothetical protein		BLASTP
3873500	3874903	1404	hypothetical protein		BLASTP
3875222	3875566	345	hypothetical protein		BLASTP
3875409	3876725	1317	putative low-affinity phosphate transport protein	QuickGO ontology	BLASTP
3877017	3878876	1860	glutathionylspermidine synthetaseamidase	QuickGO ontology	BLASTP

Figure 30: Island#21 in *Escherichia coli* O157:H7 str. Sakai [NC_002695] gene content.

Table 4: Islands identified as containing a putative enterotoxin through a gene annotation search.

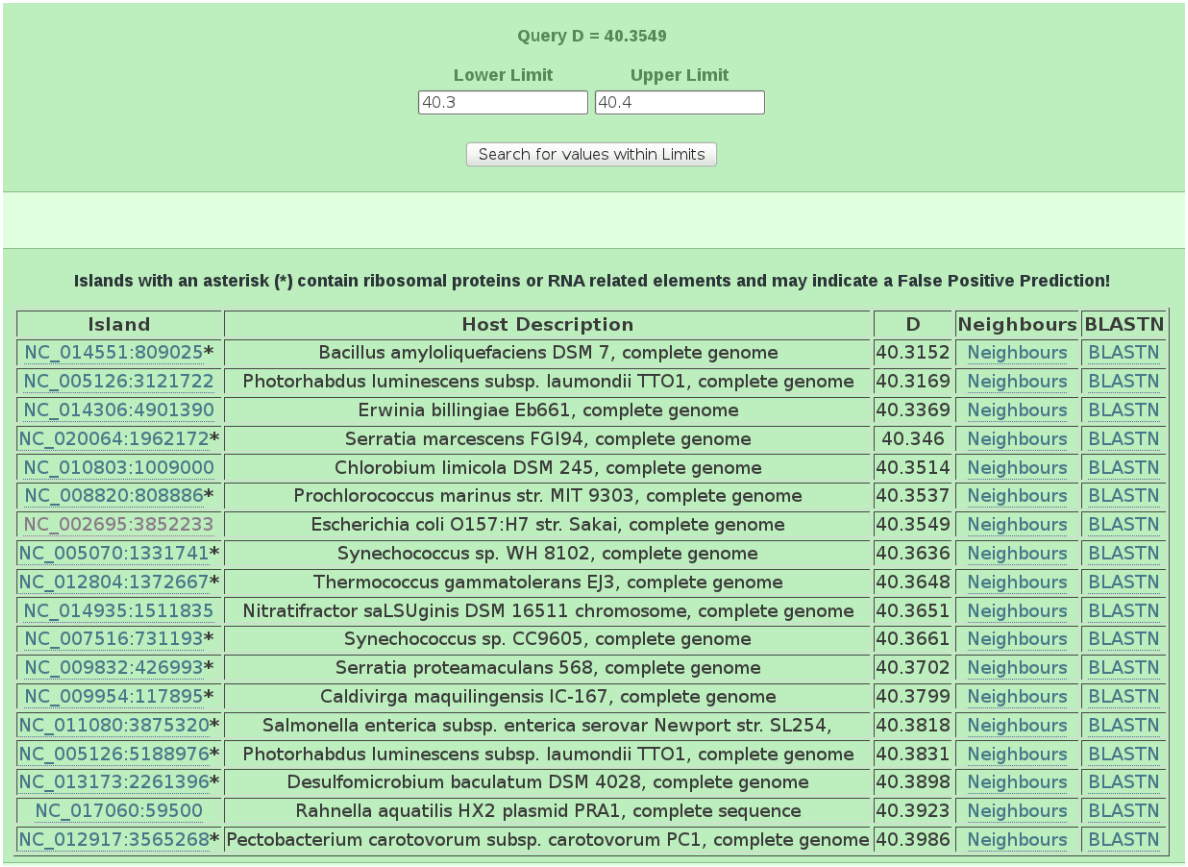
CDS description	Island	Host Description	Key word confirmation	Overlap with another database
putative enterotoxin	NC_014335:2718000	*Bacillus cereus* biovar *anthracis* str. CI	No	No
putative enterotoxin	NC_002655:3919545	*Escherichia coli* O157:H7 EDL933	Yes	Yes
putative enterotoxin protein SenB	NC_016822:2940000	*Shigella sonnei* 53G	Yes	No



Subject	Start	End	Length	Subject Host Description	CDS description	E-value	Bit score
NC_013008:3911356:3920480	3920480	3922129	1650	Escherichia coli O157:H7 str. TW14359 chromosome, complete genome	non-LEE-encoded type III effector	0	1122
NC_011353:3956630:3965754	3965754	3967403	1650	Escherichia coli O157:H7 str. EC4115 chromosome, complete genome	Ent protein	0	1122
NC_013941:3676146:3685271	3685271	3686920	1650	Escherichia coli O55:H7 str. CB9615 chromosome, complete genome	Ent protein	0	1121
NC_002655:3919545:3928670	3928670	3930319	1650	Escherichia coli O157:H7 EDL933, complete genome	putative enterotoxin	0	1121
AP010958:3627355:3636478	3636478	3638127	1650	Escherichia coli O103:H2 str. 12009 DNA, complete genome	T3SS secreted effector EspL-like protein	0	1115
AP010958:5090696:5108913	5108913	5110562	1650	Escherichia coli O103:H2 str. 12009 DNA, complete genome	T3SS secreted effector EspL-like protein	0	1115
NC_011601:3338888:3348020	3348020	3349669	1650	Escherichia coli O127:H6 str. E2348/69 chromosome, complete genome	T3SS secreted effector EspL-like protein	0	1115
NC_013353:3627355:3636478	3636478	3638127	1650	Escherichia coli O103:H2 str. 12009, complete genome	T3SS effector EspL	0	1115
NC_013353:5090696:5108913	5108913	5110562	1650	Escherichia coli O103:H2 str. 12009, complete genome	T3SS effector EspL-like protein	0	1115
NC_013364:3701895:3710469	3710469	3712118	1650	Escherichia coli O111:H- str. 11128, complete genome	T3SS secreted effector EspL	0	1112
NC_013716:1178383:1186252	1186252	1187898	1647	Citrobacter rodentium ICC168, complete genome	putative T3SS effector EspL2	0	869
NC_013364:5083949:5102092	5102092	5103177	1086	Escherichia coli O111:H- str. 11128, complete genome		0	737
NC_013364:5083949:5101528	5101528	5101986	459	Escherichia coli O111:H- str. 11128, complete genome		1e-83	311
NC_012779:2619871:2625056	2625056	2627056	2001	Edwardsiella ictaluri 93-146, complete genome	OspD3	3e-76	286

Figure 31: BLASTP results for a putative enterotoxin contained in island#21 residing in *Escherichia coli* O157:H7 str. Sakai [NC_002695].

Island#21 displayed a relatively large SWGIS D parameter which is indicative of a recently acquired foreign segment due to the sizable difference in OUP between the island and the host *Escherichia coli* O157:H7 str. Sakai. The database was searched for elements with a similar parameter by means of the hyperlinked parameter value. Results are displayed in Figure 32.



Query D = 40.3549

Lower Limit 40.3 Upper Limit 40.4

Search for values within Limits

Islands with an asterisk (*) contain ribosomal proteins or RNA related elements and may indicate a False Positive Prediction!

Island	Host Description	D	Neighbours	BLASTN
NC_014551:809025*	Bacillus amyloliquefaciens DSM 7, complete genome	40.3152	Neighbours	BLASTN
NC_005126:3121722	Photorhabdus luminescens subsp. laumondii TTO1, complete genome	40.3169	Neighbours	BLASTN
NC_014306:4901390	Erwinia billingiae Eb661, complete genome	40.3369	Neighbours	BLASTN
NC_020064:1962172*	Serratia marcescens FGI94, complete genome	40.346	Neighbours	BLASTN
NC_010803:1009000	Chlorobium limicola DSM 245, complete genome	40.3514	Neighbours	BLASTN
NC_008820:808886*	Prochlorococcus marinus str. MIT 9303, complete genome	40.3537	Neighbours	BLASTN
NC_002695:3852233	Escherichia coli O157:H7 str. Sakai, complete genome	40.3549	Neighbours	BLASTN
NC_005070:1331741*	Synechococcus sp. WH 8102, complete genome	40.3636	Neighbours	BLASTN
NC_012804:1372667*	Thermococcus gammatolerans EJ3, complete genome	40.3648	Neighbours	BLASTN
NC_014935:1511835	Nitratifractor saLSUginis DSM 16511 chromosome, complete genome	40.3651	Neighbours	BLASTN
NC_007516:731193*	Synechococcus sp. CC9605, complete genome	40.3661	Neighbours	BLASTN
NC_009832:426993*	Serratia proteamaculans 568, complete genome	40.3702	Neighbours	BLASTN
NC_009954:117895*	Caldivirga maquilingensis IC-167, complete genome	40.3799	Neighbours	BLASTN
NC_011080:3875320*	Salmonella enterica subsp. enterica serovar Newport str. SL254,	40.3818	Neighbours	BLASTN
NC_005126:5188976*	Photorhabdus luminescens subsp. laumondii TTO1, complete genome	40.3831	Neighbours	BLASTN
NC_013173:2261396*	Desulfomicrobium baculatum DSM 4028, complete genome	40.3898	Neighbours	BLASTN
NC_017060:59500	Rahnella aquatilis HX2 plasmid PRA1, complete sequence	40.3923	Neighbours	BLASTN
NC_012917:3565268*	Pectobacterium carotovorum subsp. carotovorum PC1, complete genome	40.3986	Neighbours	BLASTN

Figure 32: Similar valued D parameter islands with reference to island#21 in *Escherichia coli* O157:H7 str. Sakai.

The figure above relays the ease by which all parameters may be searched against. Resulting islands are hyperlinked to retrieve island information with compositional and sequence comparison results for an island available through the hyperlinks on "Neighbours" and "BLASTN" respectively.

Compositional similarity results for island#21 against the entirety of the database was retrieved by the "Neighbours" as displayed in Figure 29. These results were filtered by the host fields for subject hits from the Proteobacteria phylum with subject host information inclusive of the word "Enterohemorrhagic". It was further specified to retrieve only such results displaying a proposed movement from the subject to the query island#21. These subject host parameters and defined donor-recipient movement produced a single compositional similarity hit adhering to all fields and is presented in Figure 33.

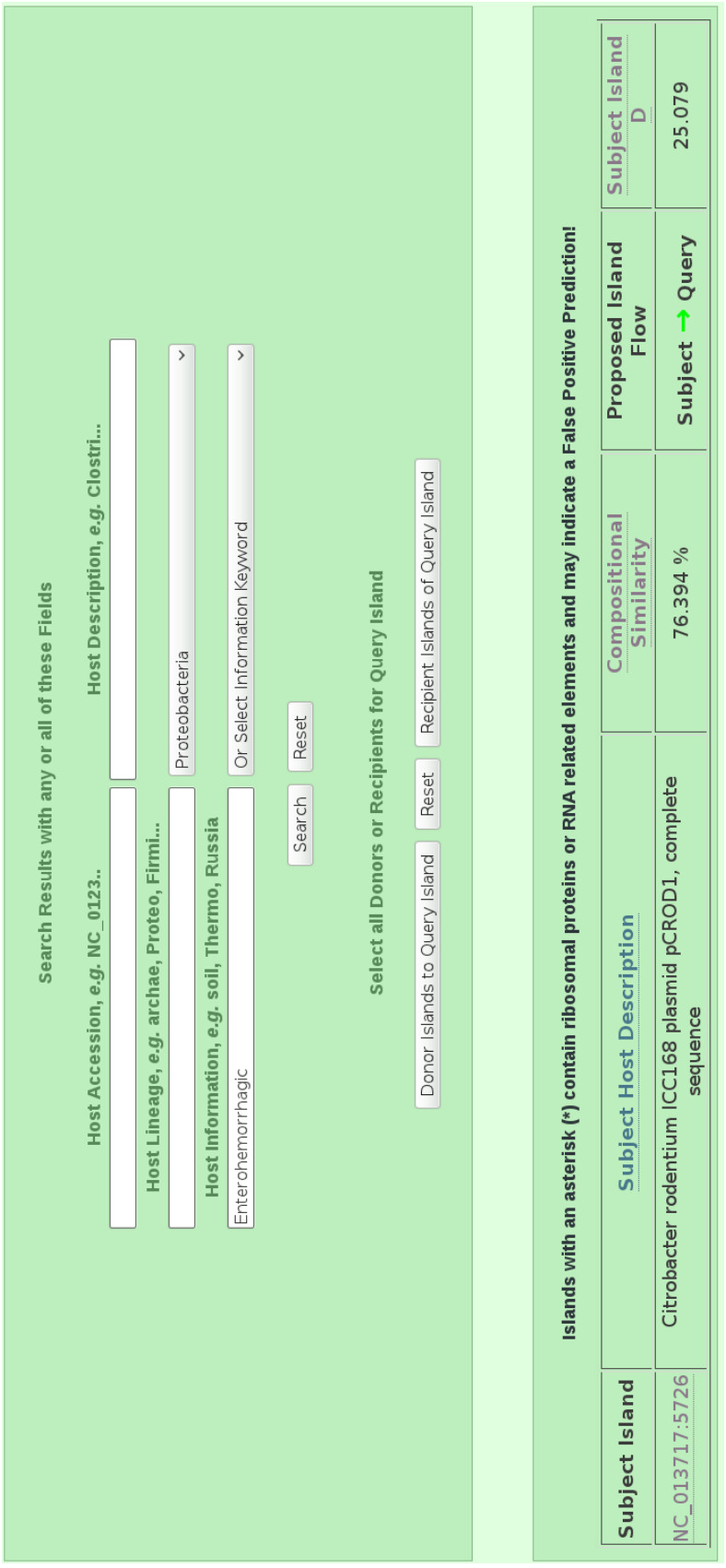


Figure 33: Compositional similarity hit to island#21 of *Escherichia coli* O157:H7 str. Sakai adhering to filters for host subject phylum, host subject information and proposed movement from a subject to the query.

The compositional similarity subject hit island was revealed to be island#1 in *Citrobacter rodentium* ICC168 plasmid pCROD1 located at position 5,726 - 30,536. The level of compositional similarity between these islands was calculated as being 76.394 %. *Citrobacter rodentium* host information revealed that it is the causative agent of transmissible murine colonic hyperplasia in mice. This disease is characterized by a hyperproliferation of the epithelial cells in the colon similar to that found in humans suffering from idiopathic inflammatory bowel disease. In addition this organism contains virulence factors similar to those found in enterohemorrhagic *Escherichia coli* and enteropathogenic *Escherichia coli*. *Citrobacter rodentium* is furthermore being used in models studying mucosal response to infection, colon tumor production, and virulence associated with pathogenic *Escherichia coli*.

Cluster results indicated that island#21 was included in cluster 4 subcluster 1. Subcluster statistics revealed that this subcluster contains 3,140 hosts with the vast majority of these from the phylum Proteobacteria. Representatives for this subcluster is available in Table 5.

Table 5: Representatives for cluster 4 subcluster 1 in which island#21 of *Escherichia coli* O157:H7 str. Sakai was placed.

Island	Host Description	Key word confirmation	Overlap with another database
NC_008150:1705152	*Yersinia pestis* Antiqua	Yes	No
NC_011149:4390315	*Salmonella enterica* subsp. *enterica* serovar Agona str. SL483	No	Yes
NC_011750:2928990	*Escherichia coli* IAI39	Yes	Yes
NC_015566:4049000	*Serratia* sp. AS12	No	No

Sequence compositional similarity hits for island#21 of *Escherichia coli* O157:H7 str. Sakai included multiple islands hosted by different strains of *Escherichia coli* as well as species of *Shigella*. High scoring sequence similarity was found against subject island#17 hosted by *Escherichia coli* O103:H2 str. 12009. This sequence similarity (BLASTP) is graphically displayed in Figure 34.

This example illustrates the ease and accessibility afforded by Pre_GI in gathering of island information. In the development of the database and GUI the inclusion of all information was enforced and user friendliness the endeavor. It is therefore possible to efficiently retrieve all information regarding a specific island in a logical fashion with no dead-end pages.

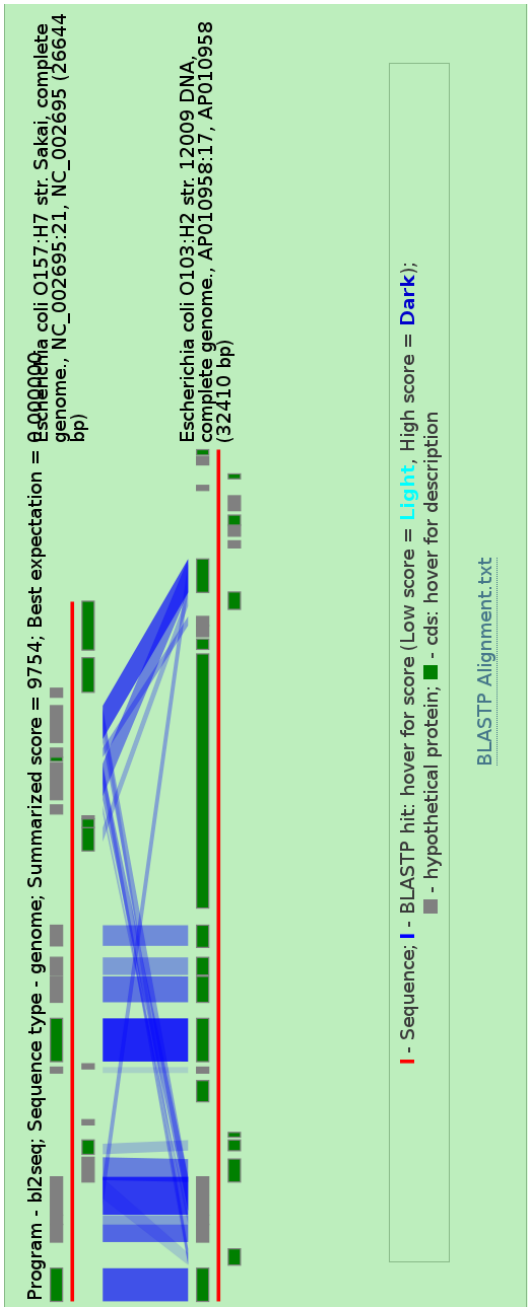


Figure 34: Sequence similarity (BLASTP) visualization as available in Pre_GI between query island#21 in *Escherichia coli* O157:H7 str. Sakai and subject island#17 in *Escherichia coli* O103:H2 str. 12009.

4.2 Cluster representative example

The list of cluster representatives was search for a specific island by implementing the provided filters. Host description was chosen as "Bacillus", host lineage selected from the provided list as "Firmicutes" and host information screened for the inclusion of the word "soil". This resulted in the retrieval of 2 island adhering to the specified parameters (Figure 35).

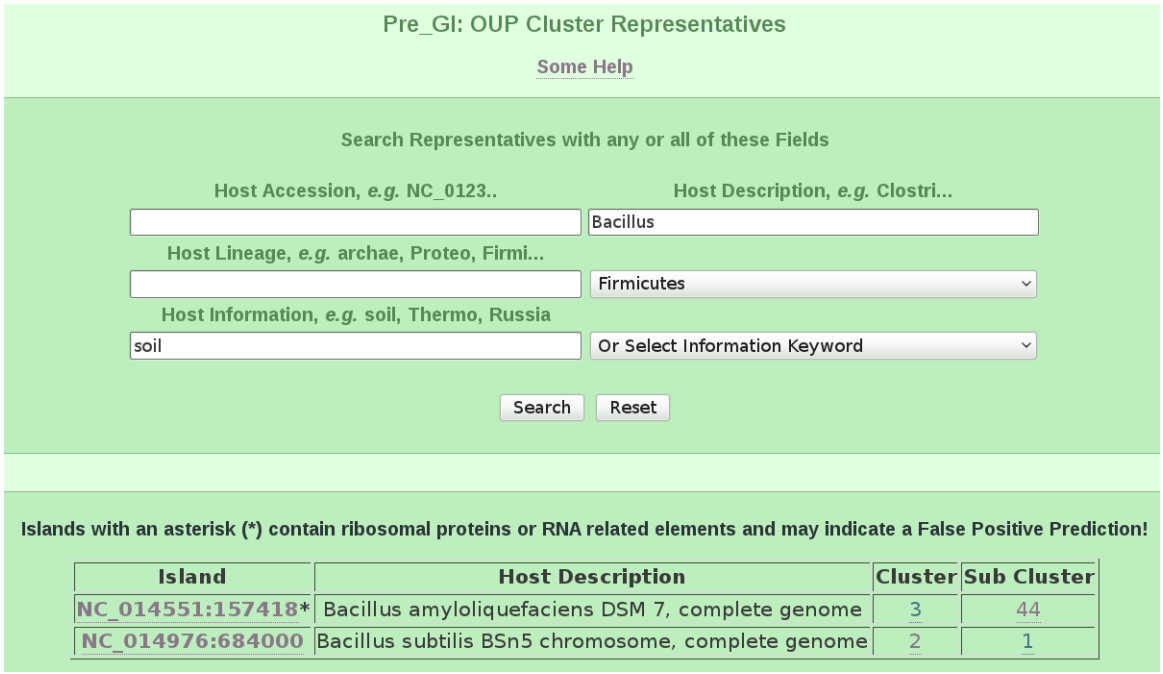


Figure 35: Cluster representative search result with various filters included.

The *Bacillus amyloliquefaciens* DSM 7, complete genome island (NC_014551:157418) contains a gene or genes with a ribosomal or RNA related annotation and is clearly marked with an asterisk to alert users to the probability of a false positive prediction. This island is a representative of cluster 3 subcluster 44. Clusters and subclusters may contain more than 1 representative to ensure that all elements of a grouping is connected. In the case of cluster 3 subcluster 44 there are a further 2 representatives.

Bacillus subtilis BSn5 chromosome, complete genome contains an island located at position 684,000 - 702,667 which functions as a representative for cluster 2 subcluster 1. This island is confirmed by the inclusion of a keyword and an overlap with the IslandViewer database. This island contains various efflux and antioxidant proteins and the host, *Bacillus subtilis* BSn5, is described as potentially inhibiting *Erwinia carotovora* subsp. *carotovora* strain SCG1. *Amorphophallus* soft rot disease is caused by *Erwinia carotovora* subsp. *carotovora* strain SCG1. Cluster 2 subcluster 1 is represented by a further 2 islands found in diverse host organisms. *Pyrococcus horikoshii* OT3, complete genome contains an representative of cluster 2 subcluster 1 located at position 372,000 - 432,099. This island is confirmed by the IslandViewer database and the host organism is described as an obligate anaerobic, hyperthermophilic archaeon which was isolated from a hydrothermal vent in the Pacific Ocean at a depth of 1,395 meters. *Shewanella* sp. MR-7, complete genome contains the third representative of cluster 2 subcluster 1. This island is located at position 3,585,601 - 3,609,298 and displays an overlap with the IslandViewer database. *Shewanella* sp. MR-7 is an environmental isolate from the Black Sea.

Cluster 2 subcluster 1 contains 8,181 islands consisting of 10% archaeal and 90% bacterial hosts. General subcluster composition statistics are displayed in Figure 36 below.

Phylum

Phylum	# of Hosts	# of Hosts as Percentage of Cluster
Actinobacteria	6	0.07 %
Aquificae	74	0.90 %
Bacteroidetes	545	6.66 %
Caldiserica	4	0.05 %
Chlamydiae	48	0.59 %
Chlorobi	18	0.22 %
Chloroflexi	20	0.24 %
Crenarchaeota	289	3.53 %
Cyanobacteria	224	2.74 %
Deferribacteres	49	0.60 %
Dictyoglomi	22	0.27 %
Elusimicrobia	13	0.16 %
Euryarchaeota	489	5.98 %
Firmicutes	3834	46.86 %
Fusobacteria	98	1.20 %
Nanoarchaeota	4	0.05 %
Nitrospirae	16	0.20 %
Proteobacteria	1780	21.76 %
Spirochaetes	340	4.16 %
Synergistetes	8	0.10 %
Tenericutes	199	2.43 %
Thaumarchaeota	5	0.06 %
Thermodesulfobacteria	8	0.10 %
Thermotogae	84	1.03 %
unclassified Bacteria	3	0.04 %
Verrucomicrobia	1	0.01 %

Figure 36: Cluster 2 subcluster 1 phylum composition statistics.

In this cluster 2 subcluster 1 the top 4 phylums with regards to the amount of islands in the gouping are Firmicutes, Proteobacteria, Bacteroidetes and Euryarchaeota. The 3 island representatives described above are from the phylums Firmicutes, Euryarchaeota and Proteobacteria respectively.

4.3 Gene annotation example

The ability to retrieve islands based on their gene content is a novel implementation in island databases. This allows users to specify a gene of interest rather than an host organism in order to retrieve islands. This tool eases the search of genes of interest

transported by HT and MGE. Users are furthermore able to track the movement of proteins between different organisms.

Vancomycin is an antibiotic used in the treatment of various bacterial infections and is listed as one of the most important medications required in a public health system. Resistance to vancomycin is intrinsically found in certain bacteria such as the *Leuconostoc* genus with acquired resistance found in various species of *Staphylococcus* and *Enterococcus*. Pre_GI gene annotation browser was tasked to retrieve all islands within the database with a gene content that included "resistant" and "Vancomycin" (Figure 37).

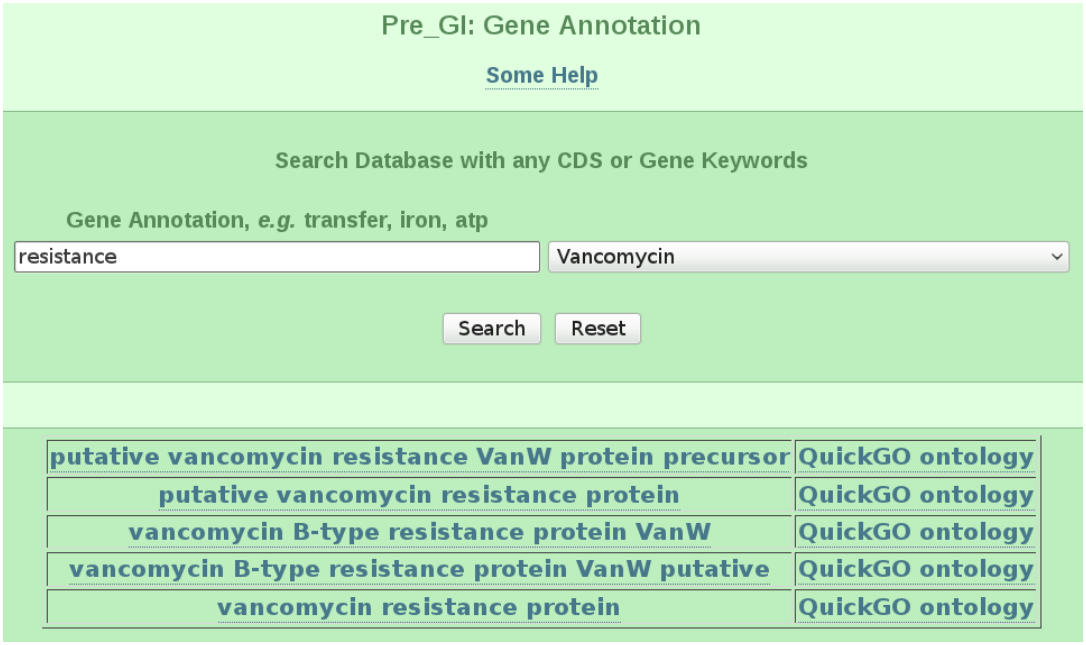


Figure 37: Gene annotation search for islands containing proteins with an annotation adhering to the filters "resistant " and "Vancomycin".

The results are tabulated with gene annotations hyperlinked to retrieve all islands containing a protein with such an annotation. The search displayed in Figure 37 resulted in the identification of 3 islands containing a similar annotation to "vancomycin B-type resistance protein VanW". One island was found hosted by *Leuconostoc gasicomitatum* LMG 18811, a species that intrinsically contains vancomycin resistant genes. *Enterococcus faecalis* V583 was found to host an island containing a vancomycin-resistant gene. This strain is one of the the first vancomycin-resistant strains isolated with a $\frac{1}{4}$ of the genome consisting of MGE. This island is located at position 2,198,027 - 2,270,099 and is also identified by the IslandViewer database. Vancomycin resistance in *Enterococcus* involves the alteration peptidoglycan synthesis pathway and variations in associated enzymes with clinical resistance traced to altered ligases producing D-alanine-D-lactate rather than D-alanine-D-alanine [16]. The close proximity of such a ligase to the vancomycin-resistant protein is displayed in Figure 38.

2212353	2212961	609	D-alanyl-D-alanine dipeptidase	QuickGO ontology	BLASTP
2212967	2213995	1029	D-alanine--D-lactate ligase	QuickGO ontology	BLASTP
2213988	2214959	972	D-specific alpha-keto acid dehydrogenase	QuickGO ontology	BLASTP
2214956	2215783	828	vancomycin B-type resistance protein VanW	QuickGO ontology	BLASTP
2215801	2216607	807	D-alanyl-D-alanine carboxypeptidase	QuickGO ontology	BLASTP
2216783	2218126	1344	sensor histidine kinase VanSB	QuickGO ontology	BLASTP
2218126	2218788	663	DNA-binding response regulator VanRB	QuickGO ontology	BLASTP
2218843	2219031	189	streptomycin resistance protein putative	QuickGO ontology	BLASTP

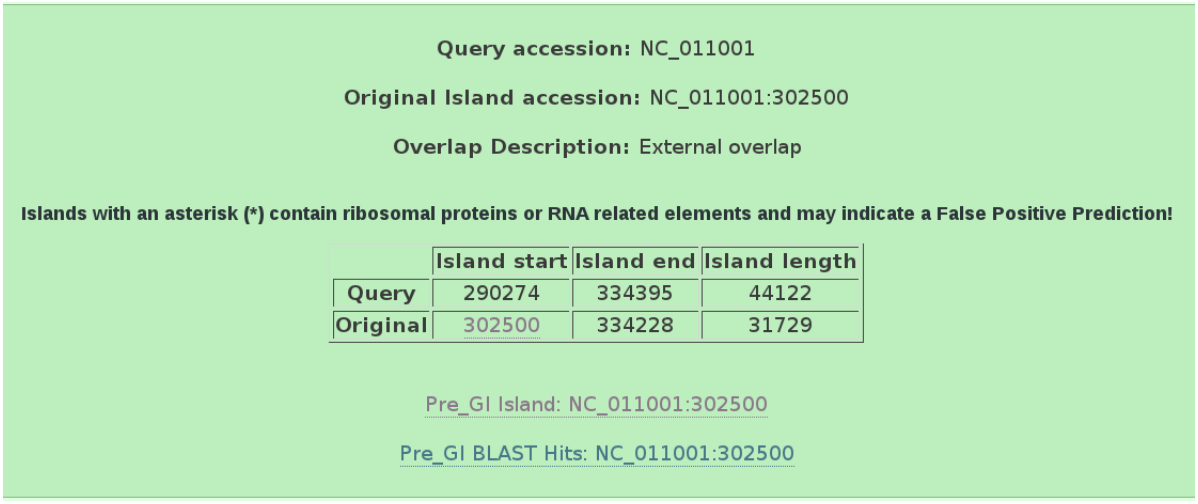
Figure 38: Gene content of an *Enterococcus faecalis* V583 island containing a vancomycin resistance gene.

The ability to search islands by their content and gene annotation is a valuable tool in antibiotic resistance research. This enables users to rapidly identify islands containing proteins of interest, circumventing the need to first identify archaeal/bacterial hosts and then manually inspecting all islands contents.

4.4 Locational search example

The ability to search Pre_GI for islands by means of island location is a further novel application in island databases. PAIDB includes REIs as a different class of island linked to pathogenesis by conferring resistance and promoting the emergence of multidrug resistance pathogens [90]. PAIDB indicated the presence of such a REI in *Burkholderia cenocepacia* J2315 chromosome 2 located at position 290,274 - 334,395. Pre_GI was tasked by means of a locational query to indicate the presence of an overlap in the database. The query result is displayed in Figure 39.

The query indicated that an overlap was present in Pre_GI. The overlap was deemed to be an external overlap as the PAIDB REI boundaries extended over the Pre_GI housed island boundaries. The island in Pre_GI with which an overlap was found is easily accessible through the hyperlinks. The island is identified as island#3 in Pre_GI host *Burkholderia cenocepacia* J2315 chromosome 2 with island content displayed in Figure 40. This Pre_GI island displays key word confirmation and an overlap with IslandViewer is further indicated. The "island text" hyperlink indicates that this island was identified by SWGIS in September of 2013. The current version of PAIDB was first published online in October of 2014 and as such this overlap between Pre_GI and PAIDB is not indicated in the current version of Pre_GI. The REI from PAIDB and the island in Pre_GI displayed below both contain a "putative porin" of the exact same length which is described by PAIDB as tested and deemed virulent.



Query accession: NC_011001

Original Island accession: NC_011001:302500

Overlap Description: External overlap

Islands with an asterisk (*) contain ribosomal proteins or RNA related elements and may indicate a False Positive Prediction!

	Island start	Island end	Island length
Query	290274	334395	44122
Original	302500	334228	31729

Pre_GI Island: NC_011001:302500

Pre_GI BLAST Hits: NC_011001:302500

Figure 39: Locational query of a PAIDB identified resistance island (REI) in *Burkholderia cenocepacia* J2315 chromosome 2 located at position 290,274 - 334,395 against the Pre_GI database.

The proteins tested and determined by PAIDB to be resistant and virulent residing in the REI found in host *Burkholderia cenocepacia* J2315 chromosome 2 were inspected and compared to Pre_GI. PAIDB described 4 resistance proteins located at the start of the REI. The locational query result for Pre_GI indicated an external overlap and after inspection it was concluded that the start region containing the resistance genes in the PAIDB REI was not included in the island identified by SWGIS. The virulence protein was located in the center of both islands and sequence similarity between the REI putative porin and the Pre_GI putative porin concluded 100% identity. Resulting BLASTP sequence similarity between the resistance proteins and virulence protein indicated by PAID in the REI of *Burkholderia cenocepacia* J2315 chromosome 2 and all proteins contained within the Pre_GI database is available in Table 6 below. The 4 PAIDB resistance proteins were excluded from the Pre_GI island yet highly similar proteins were available in the another island namely *Burkholderia cenocepacia* MC0-3 chromosome 2, island#2.

Table 6: Highest scoring hits of PAIDB resistance proteins and a virulent protein compared against Pre_GI by BLASTP.

PAIDB description	Pre_GI description	%id	e value
ArsR family regulatory protein	transcriptional regulator ArsR family	98.26	1e-53
putative arsenate reductase	protein tyrosine phosphatase	93.90	8e-87
putative sodium bile acid symporter family protein	arsenical-resistance protein	96.71	0.0
MarC family protein	multiple antibiotic resistance MarC-related protein	93.64	6e-91
putative porin	putative porin	100.00	0.0

The locational query application available in Pre_GI is a quick and reliable approach in the determination of identified island overlaps and may be used to inspect regions of a host genome for the presence of an island or the inclusion of specific proteins.

Start	End	Length	CDS description	QuickGO ontology	BLASTP
303100	303495	396	putative transposase	QuickGO ontology	BLASTP
303492	303839	348	putative transposase	QuickGO ontology	BLASTP
303891	305462	1572	putative transposase	QuickGO ontology	BLASTP
306046	307635	1590	putative AMP-binding protein	QuickGO ontology	BLASTP
307702	308637	936	putative 3-oxoacyl-acyl-carrier-protein synthase	QuickGO ontology	BLASTP
308634	309659	1026	putative 3-oxoacyl-acyl-carrier-protein synthase	QuickGO ontology	BLASTP
309643	309921	279	putative acyl carrier protein	QuickGO ontology	BLASTP
309952	311208	1257	putative acyl-CoA dehydrogenase	QuickGO ontology	BLASTP
311448	312281	834	epidemic strain marker regulator	QuickGO ontology	BLASTP
312675	312956	282	putative transcriptional regulator	QuickGO ontology	
312973	313272	300	repressor protein	QuickGO ontology	BLASTP
313367	313702	336	putative repressor protein	QuickGO ontology	BLASTP
313884	314672	789	putative amino acid transporter	QuickGO ontology	BLASTP
314763	315836	1074	putative branched-chain amino acid transporter	QuickGO ontology	BLASTP
315836	316570	735	ABC transporter ATP-binding protein	QuickGO ontology	BLASTP
316557	317312	756	ABC transporter ATP-binding protein	QuickGO ontology	BLASTP
317309	318487	1179	hypothetical protein		BLASTP
318540	319955	1416	putative amidase	QuickGO ontology	BLASTP
319952	320497	546	hypothetical protein		BLASTP
320550	321584	1035	putative porin	QuickGO ontology	BLASTP
321591	321815	225	hypothetical protein		BLASTP
322124	323473	1350	hypothetical protein		
323992	324261	270	hypothetical protein		BLASTP
324258	324746	489	putative acetyltransferase	QuickGO ontology	BLASTP
324761	325819	1059			BLASTP
326635	327486	852	hypothetical protein		BLASTP
328358	328813	456	hypothetical protein		BLASTP
329025	329495	471	putative universal stress protein	QuickGO ontology	BLASTP
329613	329888	276	hypothetical protein		BLASTP
330064	330498	435	putative heat shock protein	QuickGO ontology	BLASTP
330593	331258	666	hypothetical protein		BLASTP
331489	332139	651	putative phospholipid-binding protein	QuickGO ontology	BLASTP
332176	332379	204	hypothetical protein		
332471	334228	1758	putative sulfate transporter family protein	QuickGO ontology	BLASTP

Figure 40: Island#3 contents in host *Burkholderia cenocepacia* J2315 chromosome 2 available in Pre_GI that displayed a locational overlap with a REI identified in PAIDB.

4.5 Novel island(s) search and analysis

The production of a static repository is insufficient in the current age of prokaryotic genomic data generation. Pre_GI allows for the comparison of novel islands against the wealth of housed content by various means. These comparisons are available in numerous formats and functionality. Comparison of islands in their entirety enhances the applicability in island and MGE research. Novel islands to be compared to the database

may be predicted by any currently accepted method, although SWGIS is recommended as resulting predicted island file content includes parameters enabling the detection of possible donor-recipient movement.

Sequence similarity of a single novel island in nucleotide FASTA format is accomplished by means of BLASTN with default or user defined e-value cut-off. High scoring hits against the subject database is tabulated and all hits are accessible for further inspection of subject hit content and information. BLASTN visualization as available from the results is displayed in Figure 41. A novel island predicted by SWGIS at location 3,609,844 - 3,653,217 in host *Serratia marcescens* FGI94 [NC_020064] was searched against the database for sequence similarity. Various islands indicated sequence similarity with an example from *Escherichia coli* presented below.

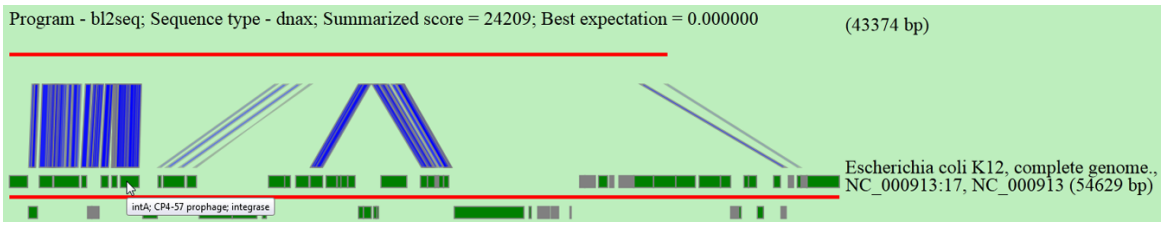


Figure 41: Novel island BLASTN high scoring hit against Pre_GI visualization. The query or novel island sequence is depicted by the upper red line and the subject hit island sequence by the lower red line. Green boxes indicate genes with blue lines representing high scoring sequence pairs.

Compositional similarity searches of a novel island against housed content is done by OUP similarity searches. The query nucleotide sequence is first compared against the 420 cluster/subcluster representatives to ascertain mutual compositional similarity above 75%. These representatives serve as a litmus test to indicate further compositional similarity searches to be done against all members of a marked cluster/subcluster. All compositional similarity hits are tabulated with hit subject island easily available to users for further analysis. *Hyphomicrobium nitrativorans* strain NL23 is a denitrifying bacterium isolated from the biofilm of a denitrification system treating marine water [28]. SWGIS identified an island in this novel bacterium located at 318,281 - 345,034. This island was compared against the database by means of compositional similarity. A high scoring hit of 88.65 % OUP similarity was found against an island housed by *Bradyrhizobium japonicum* USDA 110 located at position 8,401,060 - 8,453,099. This organism has been widely studied because of its nitrogen fixation capabilities. The compositional comparison hit list was filtered with regards to subject and the subject information. All hits pertaining to the word "biofilm" in the subject information category were extracted. This resulted in the identification of a high scoring hit to an island identified in *Pseudomonas aeruginosa* PAO1 with calculated OUP similarity of 81.92 %. *Pseudomonas aeruginosa* PAO1 are naturally found in biofilms and are renowned for their metabolic versatility.

Comparison of multiple prokaryotic GenBank island files is available to the user, with the use of SWGIS island GenBank files recommended. Sequence and compositional similarity results are available for up to 8 files in a single run. The use of SWGIS generated island GenBank files allows for the detection of probable donor-recipient movement and filtering of compositional comparison results by means of direction of movement. All information as described above for similarity searches is generated and viewable.

SWGIS predicted 8 novel islands in *Spiroplasma apis* B31T ATCC 33834. This bacterium has been linked to May disease in Honeybees and was isolated from a honeybee in France that was affected by May disease [27]. Tetracyclin was found to be effective in the control of *Spiroplasma apis* but not penicillin [29]. All 8 novel islands were uploaded and compared to Pre_GI (Figure 42). *Spiroplasma apis* island located at position 12,064 - 59,291 displayed high sequence similarity to a *Staphylococcus epidermidis* ATCC 12228 island detected at location 277,291 - 324,568. *Staphylococcus epidermidis* ATCC 12228 is resistant to various antibiotics including penicillin. Compositional comparison indicated a high affinity for probable movement of islands from *Spiroplasma apis* B31T ATCC 33834 to various species within the genus *Bacillus* including the species *Bacillus atrophaeus* and *Bacillus subtilis* at OUP similarity values of above 85 %.

Pre_GI offers users the ability to compare newly identified islands against the wealth of information contained within the database. This application is approachable with raw sequence data although a GenBank file format is suggested for added results. This database is not biased towards any avenue of research and all possible relations and ontology will be supplied to aid users in their niche of prokaryotic research.

4.6 Discussion

Pre_GI offers an alternative to currently used island databases. It aims not just to be a island repository but a island analysis toolbox with various applications and functions. This database includes a wealth of information regarding islands and the contents of these foreign segments to aid users in various fields of HT and MGE research. Island information and HT events will increasingly become more important in resistance and pathogenicity research, and in general evolution studies. The identification of islands in archaeal/bacterial genomes needs to be followed with an extensive investigation into all possible relations and avenues of movement of predicted islands. SWGIS in collaboration with Pre_GI offers researchers the ability to identify and analyze islands from newly sequenced archaeal/bacterial hosts.

Islands Uploaded

Island(s)	Island Host Description	Start	Stop	Length	Compositional Similarity	Sequence Similarity	GRV_RV	D	PS
NC_022998:1	Spiroplasma apis B31, complete genome	12064	59291	47228	OUP	BLASTN	2.434361	15.730367	24.200790
NC_022998:2	Spiroplasma apis B31, complete genome	174090	192599	18510	OUP	BLASTN	1.599384	15.649319	19.576075
NC_022998:3	Spiroplasma apis B31, complete genome	222333	241472	19140	OUP	BLASTN	1.538171	15.385863	17.718358
NC_022998:4	Spiroplasma apis B31, complete genome	266500	284099	17600	OUP	BLASTN	1.483150	15.649319	20.258253
NC_022998:5	Spiroplasma apis B31, complete genome	383515	403461	19947	OUP	BLASTN	1.984518	15.535830	25.124863
NC_022998:6	Spiroplasma apis B31, complete genome	876396	918599	42204	OUP	BLASTN	3.625581	16.790334	22.467495
NC_022998:7	Spiroplasma apis B31, complete genome	1059759	1078196	18438	OUP	BLASTN	2.822772	15.120379	27.957120
NC_022998:8	Spiroplasma apis B31, complete genome	1086838	1115063	28226	OUP	BLASTN	3.355257	17.718680	27.817707

Figure 42: 8 Novel *Spiroplasma apis* B31T (ATCC 33834) islands uploaded and compared to content of Pre_GI.

5 Chapter 5: Analysis of Current Pre_GI Content - Islandomics

> When antibiotics became industrially produced following World War II, our quality of life and our longevity improved enormously. No one thought bacteria were going to become resistant. Bonnie Bassler

The collection of all islands and island content affords the ability to do large scale analysis on islands and MGE as entities. The global analysis of islands enhances current knowledge and understanding of these elements as individuals and the communities they occupy. It may be possible to isolate islands from the host and regard them as living entities of their own residing in a host. Collections or groupings of islands may therefore be regarded as communities. The wealth of information available in Pre_GI offers researchers the opportunity to investigate these island communities.

The following sections aim to illustrate the application of Pre_GI with regards to future research made possible by the gathering of all island information in a single, accessible location.

5.1 Database introspection

The fate of an acquired island in a host may follow different paths. One such avenue would be fragmentation of the island as it has been proposed that after insertion certain islands may undergo cleavage. The availability of multitudes of islands and the compositional similarity between them as provided by the Pre_GI database enables investigation into the process of fragmentation after insertion. Empirical observations suggested that islands fragmented from a large insert would share compositional similarity of at least 80% and that the change in distance of OUP from the island to the host should not exceed 15%. Exclusion of islands with highly similar OUP to that of the host is forced by the mechanism of amelioration. All foreign DNA inserts are influenced by amelioration and after time all islands composition reflect that of the host and each other independent of the origin of the island. Within the set thresholds it was possible to isolate up to 10 groups of islands with distinct origins per host genome as shown in Figure 43. Host genome is represented by a data point with counts of predicted islands on the y-axis and number of distinct origins of the islands depicted on the x-axis.

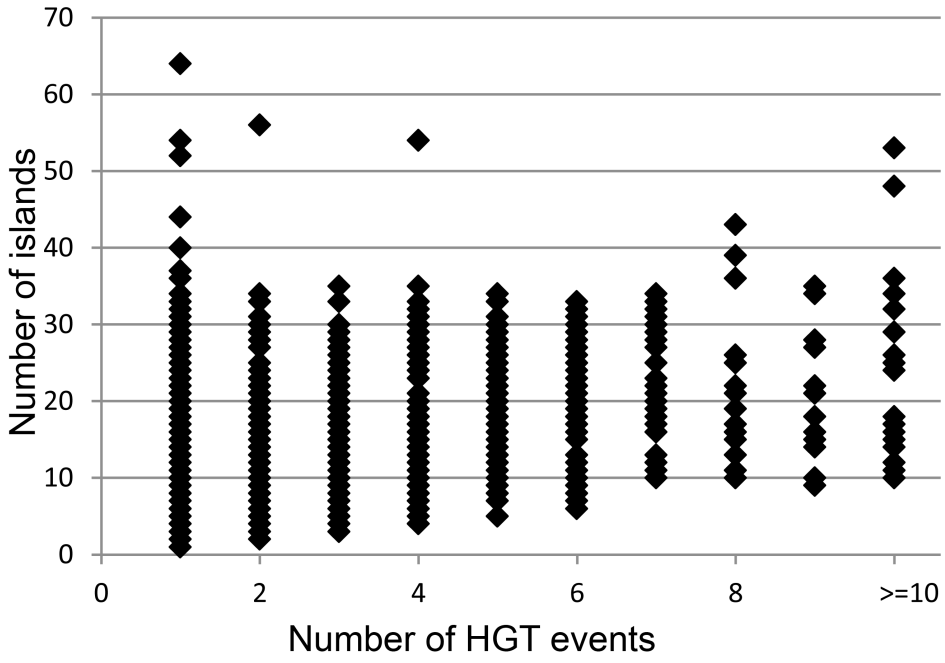


Figure 43: Dissemination of number of islands and causative transfer events. Data points indicate prokaryotic host genomes with amount of predicted islands per genome on the vertical axis. The number of non-overlapping acquisitions of the islands for a genome is depicted on the horizontal axis.

It was found that the amount of islands predicted per genome was independent of the number of origins of the islands in the genome. 64 Islands found in *Stigmatella aurantiaca* DW4_3_1 displayed uniform DNA composition and thus possibly resulted from the fragmentation of a large insertion. This is equivalent in the 54 islands of *Clostridium saccharoperbutylacetonicum* N1-4(HMT) and 52 islands of *Clostridium pasteurianum* BC1. Motley islands in genomes resulting from at least 10 different origins are presented in Table 7. Hosts acquiring genes from variable sources included multiple symbiotic/pathogenic organisms and extremophyles defined by unusual enzymatic activities which may result from the solution of genes from distinct origins.

Table 7: Genomes with islands of diverse origins.

Species and strain	Phylum	Short description*	No. Islands
Bacteroides fragilis 638R	Bacteroidetes	Gut microflora, symbiont, pathogen	35
Bacteroides salanitronis DSM 18170	Bacteroidetes	Gut microflora, symbiont	39
Bifidobacterium asteroides PRL2011	Actinobacteria	Gut microflora, symbiont	19
Cellvibrio japonicus Ueda107	Gammaproteobacteria	Soil bacterium	32
Corynebacterium diphtheriae 241	Actinobacteria	Pathogen	18
Corynebacterium diphtheriae VA01	Actinobacteria	Pathogen	16
Denitrovibrio acetiphilus DSM 12809	Deferribacteres	Marine bacterium, bio-degradation of pollutants	10
Desulfitobacterium dichloroeliminans LMG P-21439	Firmicutes	Soil bacterium, bio-degradation of pollutants	14
Desulfovibrio piezophilus C1TLV30	Deltaproteobacteria	Deep-sea sulfate reducer	20
Eubacterium limosum KIST612	Firmicutes	Carbon monoxide-utilizing acetogen	33
Fibrella aestuarina BUZ 2T	Bacteroidetes	Filamentous marine bacterium	28
Fibrobacter succinogenes subsp. *succinogenes* S85	Fibrobacteres	Cellulolytic organism	24
Geobacter uraniireducens Rf4	Deltaproteobacteria	Uranium bio-remediation organism	30
Granulicella mallensis MP5ACTX8	Acidobacteria	Tundra soil organism	26
Hahella chejuensis KCTC 2396	Gammaproteobacteria	Marine bacterium producing an algicidal agent	57
Nitrosococcus halophilus Nc4	Gammaproteobacteria	Marine bacterium	24
Nitrosococcus oceani ATCC 19707	Gammaproteobacteria	Marine bacterium	26
Octadecabacter antarcticus 307	Alphaproteobacteria	Polar marine bacterium	17
Octadecabacter arcticus 238	Alphaproteobacteria	Polar marine bacterium	12
Paenibacillus mucilaginosus KNP414	Firmicutes	Soil silicate degrading bacterium	36
Parabacteroides distasonis ATCC 8503	Bacteroidetes	Gut microflora, symbiont	45
Prevotella dentalis DSM 3688	Bacteroidetes	Oral microflora, symbiont	14
Pyrobaculum arsenaticum DSM 13514	Crenarchaeota	Arsenate-reducing hyperthermophile	20
Pyrobaculum oguniense TE7	Crenarchaeota	Hyperthermophile	19
Rothia dentocariosa ATCC 17931	Actinobacteria	Oral microflora, symbiont	13
Serratia symbiotica str. '*Cinara cedri*'	Gammaproteobacteria	Insect endosymbiont	17
Spirochaeta africana DSM 8902	Spirochaetales	Soda lake alkaliphilic anaerobe	12
Spirosoma linguale DSM 74	Bacteroidetes	Free-living non-pathogenic bacterium	53
Teredinibacter turnerae T7901	Gammaproteobacteria	Endosymbiont of marine shipworms	26

*Short description was taken from corresponding NCBI and GOLDCARD bioprojects.

Relational links total 69,176,627 for OUP similarity above 75% and 3,692,401 for BLASTN similarity with an e-value threshold of 10^{-6}. This relates to each island displaying compositional similarity to 2,586 and sequence similarity to 138 other islands with analogous or dissimilar hosts. Comparing the amount of compositional and sequence similarity links per islands reveals 2 distinct groupings of islands (Figure 44). These groups indicate a cluster of islands with a relatively small amount of BLASTN links (in average 150) and another cluster with a multiple BLASTN links (in average 1,500). The latter grouping indicates highly conserved sequences observed in islands, *e.g.* tRNA, ribosomal RNA, genes encoding gyrase subunits, etc.

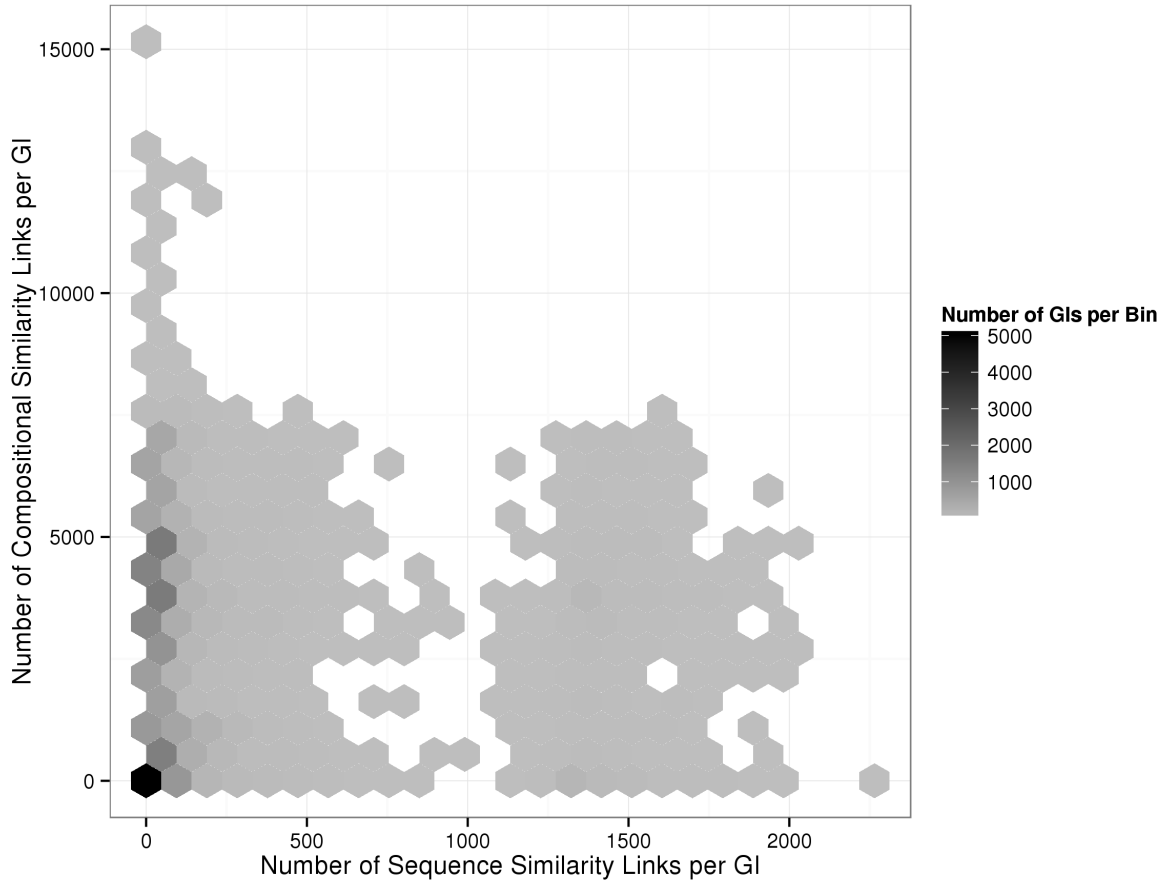


Figure 44: Distribution of compositional similarity (vertical axis) and sequence similarity (horizontal axis) per island in a hexagonal binning plot. Islands are binned in a hexagon according to the number of compositional similarity links and sequence similarity links. Distinct groupings are related to islands poor in sequence similarity links and islands rich in BLASTN links.

The extensive amount of OUP links between islands in relation to the amount of BLASTN links between islands is not unexpected. Fragmentation of former larger segments or entities (integrated plasmid, phage or transferred DNA region) would indicate shared OUP similarity due to the common origin even though no homologous genes are shared

between islands. In Figure 44 the grouping rich in BLASTN links (lower right) includes tRNA regions. Genes for tRNA are known as routine recombination attachment (*attP*) sites targeted by phage and conjugative plasmid integrases [62]. These areas or sites are accepted to be excluded from HT events yet are included in the SWGIS island sequence as the sliding window approach is unable to separate foreign genomic material from the insertion (*attP*) sites. The commonality of these attachment sites in different islands is a clear indication of homology of these MGE.

The large constituent of OUP similarity links (69,176,627) in islands enables research on similarity between islands in taxonomic groupings. All OUP records were assembled in such a fashion to distinguish between links found for islands within the same genome, amid different strains of the same species, between different species of the same genus and so forth up the taxonomic pyramid ending in domain. Results are displayed in Figure 45. The general trend observed is that of a diminishing average value of OUP similarity values amid distantly related groups.

The global trend displayed above adheres to what is expected of the sharing of compositional similarity between closely related organisms. Inconsistent with the expectation was the undulating local trend between certain taxonomic factions. Islands detected in different strains of the same species were less similar with regards to OUP than islands of distinct species within the same genus. This is repeated with a decrease on the level of genera but an increase on the level of family. This may be due to the method by which organisms are selected for sequencing in research projects. Generally different strains of the same species are chosen from varying habitats to truthfully represent the spectrum of genetic diversity. Studies relating to taxonomic diversity of a habitat often incorporate sequences of related species which co-exist in different conditions with the same limiting factor pressures specific for the chosen habitat. It seems plausible from Figure 45 that the sharing of habitats may have a greater influence than the taxonomic relatedness on the similarity of islands shared by different micro-organisms. This conforms to Karberg et al. [115] who proposed that taxonomic distant organisms pick from a common gene pool, described as a 'supraspecies pangenome', to explain the unexpectedly high synonymous codon usage similarity between horizontally acquired genes.

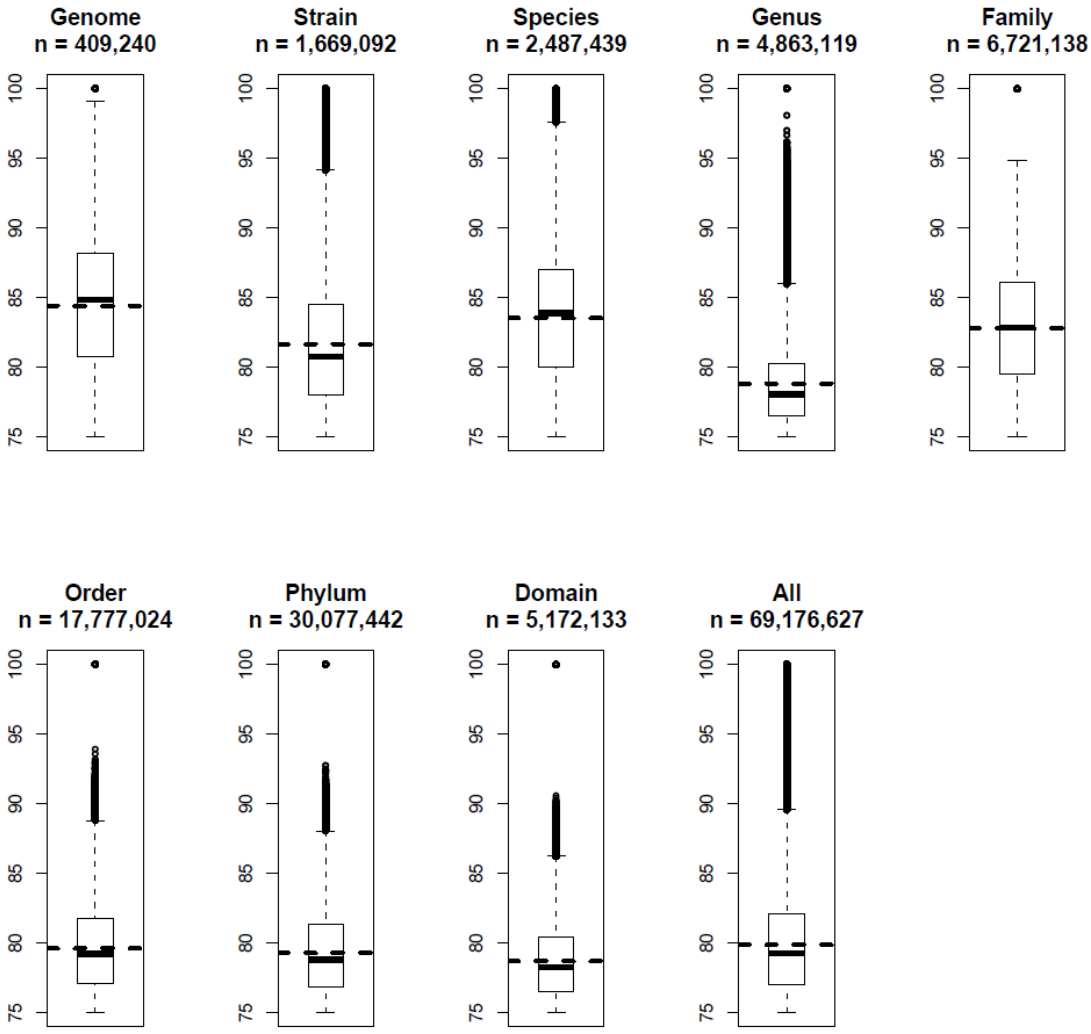


Figure 45: OUP similarity box plots from distinct taxonomic levels. OUP similarity divided into 8 categories: Genome - compositional similarity links between islands hosted by the same genome; Strain - links between islands of different strains of the same species; Species - different species of the same genus; Genus - distinct genera of the same family; Family - contrasting families of the same order; Order - different orders of the same phylum; Phylum - different phyla of the same domain; and Domain - OUP links for separate domains. Amount of OUP links for each grouping indicated above the box plot for the group with mean values of group displayed as dashed lines.

5.2 Island ebb and flow

Amelioration over time alters the nucleotide genomic signature of an island to reflect the nucleotide genomic signature of the host or carrier genome. This process enables the determination of time of acquisition or age of an island and possible donor genome of an island as for a protracted period of time the island signature will retain the global nucleotide signature of the donor genome sequence [77, 119]. Compositional homologous islands may indicate donor-recipient migration signaled by their individual compositional similarity with regards to the hosts they are located in. If island a predicted in genome

A displays compositional similarity by means of OUP to island b identified in genome B and island a has a genomic signature more reflective of host A than island b has of host B then it is probable that island b in host B was donated by island a located in host A. It should be noted that intermediate hosts or carriers of islands are not excluded in the recipient-donor process.

Proposed donor-recipient direction is available for island compositional similarity links as well as for islands in any cluster/subcluster. Direction of movement is predicted as defined above with a minimum difference in place to ensure a higher level of certainty on direction of flow. This methodology was implemented in LingvoCom, a SeqWord project island analysis toolkit used in search of possible island donor genomes, freely available from www.bi.up.ac.za/SeqWord/lingvocom/. An example is described in Figure 46 where a large island residing at 1,638,946 - 1,700,655 on the *Xyllella fastidiosa* 9a5c chromosome is considered.

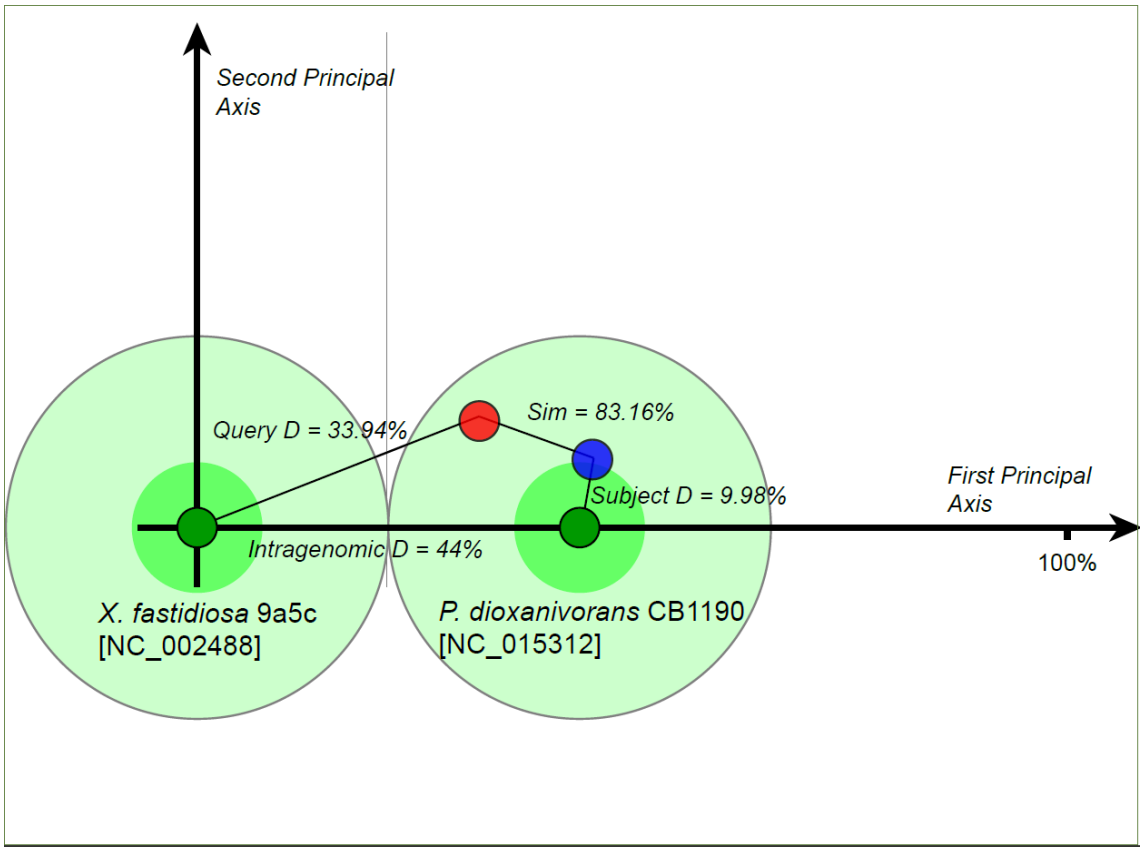


Figure 46: Proposed donor-recipient movement in collaboration with the LingvoCom 2D projection tool. The 2 dark green spots on the figure represent OUP of *Xylella fastidiosa* 9a5c (center) and *Pseudonocardia dioxanivorans* CB1190 (first principle axis) chromosomes. Light green circles announce $\frac{1}{2}$ of the distance between OUP calculated for the chromosomes. The island of *Xylella fastidiosa* is displayed as a small red circle and the island of *Pseudonocardia dioxanivorans* as a blue circle. Islands are plotted along the second principle axis in relation to the distance between OUP of the island and of the host chromosome.

Pre_GI indicated a possible *Pseudonocardia* origin for this island with the strain *Pseudonocardia dioxanivorans* CB1190 containing an island located at position 5,876,957 - 5,900,279 similar in composition. These islands from different hosts display a high OUP similarity to *Pseudonocardia dioxanivorans* CB1190 with a lower similarity to *Xyllella fastidiosa* 9a5c. It is therefor plausible to presume that the *Xyllella fastidiosa* 9a5c island was donated from *Pseudonocardia dioxanivorans* CB1190. Intermediate carriers should not be excluded from this donor-recipient movement. The detection of proposed flow between organisms is dependent on the current species sampling and sequencing and may change in the future.

5.3 Omnipresent island proteins

CDS descriptions and annotations found in the majority of islands may relay the overall gene structure of an island. Occurrence of protein descriptions in all currently housed islands are presented in Figure 47 as a word cloud to visualize the over-representation of certain words in cds annotations. The omnipresence of island and MGE key words is displayed and further visualized with a bar chart in Figure 48.

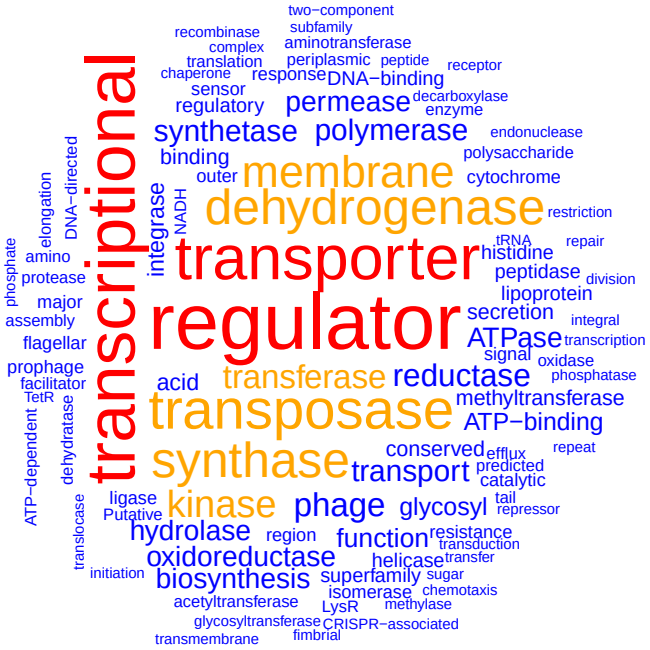


Figure 47: Word cloud of cds descriptions found in all islands. Color and size is related to number of occurrence. Red, large indicates a high frequency with yellow, medium representing an intermediate count and blue, small indicative of a low occurrence.

The availability of ample annotated islands in a single depository enables future research with regards to core and pan island protein content. Future research into this field could expand current knowledge regarding the "machinery" or "motors" included in an island which enables insertion and incorporation into a host genome.

The bar chart below (Figure 48) clearly displays the large incidence of certain gene annotation words. The top words were "regulator" and "transcriptional", seemingly opposite words in prokaryotes as regulators often encode repressor proteins and transcription is the first step in expression. Inclusion of certain proteins in an island may indicate the biological activity and viability of an island. The high frequency of transport related proteins and "phage" words are to be expected in elements and regions of HT.

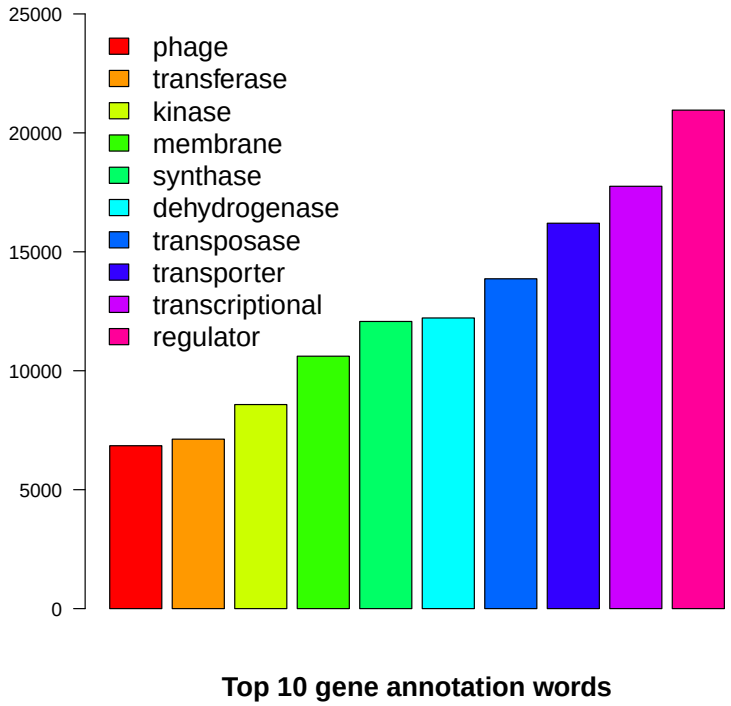


Figure 48: Bar chart of top 10 represented words in island gene annotations.

The collection of all BLASTP sequence similarity between genes contained in islands was inspected to identify proteins displaying gene sequence similarity links to the majority of island gene sequences and presented in Table 8. The number of sequence similarity hits excluded any pair located within the same island and was subjected to an e-value cut-off 10^{-6}. This list conveys the high incidence of certain protein sequence similarity to be found in islands.

Table 8: High frequency sequence similarity island genes.

Accession	Host Species	CDS Description	BLASTP Hits
NC_003366	*Clostridium perfringens*	probable amino acid ABC transporter	7767
NC_003551	*Methanopyrus kandleri*	ATPase subunit of a ABC-type transport system involved in lipoprotein release	7989
NC_004557	*Clostridium tetani*	glutamine transport ATP-binding protein glnQ	7880
NC_005213	*Nanoarchaeum equitans*	tRNA-His	9249
NC_007963	*Chromohalobacter salexigens*	ABC transporter related	7769
NC_008054	*Lactobacillus delbrueckii*	Spermidine/putrescine ABC transporter, ATP-binding protein	7796
NC_008261	*Clostridium perfringens*	amino acid ABC transporter, ATP-binding protein	7767
NC_012121	*Staphylococcus carnosus*	glutamate ABC transporter ATP-binding protein	7798
NC_014377	*Thermosediminibacter oceani*	amino acid ABC transporter ATP-binding protein, PAAT family	7854
NC_015428	*Lactobacillus buchneri*	phosphonate-transporting ATPase	7862

ABC transporters and related members display the most abundant protein similarity to be found in regions of HT. These proteins are part of one of the oldest and largest superfamilies represented in all biological divisions from prokaryotes to eukaryotes. These proteins are essential in prokaryotic viability and as such seem vital in island viability. These proteins are involved in resistance, transport of metals and stress responses.

Further research into island gene content and protein similarity may prove useful in the determination of novel island and HT key words. Biological activity and viability of an island in a host could be determined by the genetic content of an island. Core and pan island genome sets may be identified in an effort to enhance current understanding with regards to island structure and content.

5.4 Relatedness versus Habitat

The transfer of genetic material between unrelated species allow prokaryotes to adapt to altered habitats in a short period of time without the need to generate novel traits *de novo* [83]. HT is reliant on physical contact and therefore necessitates proximity and

overlap of habitat. In certain examples the sharing of habitat is of greater importance than taxonomic relatedness. Investigation into such possibilities was mentioned above and displayed in Figure 45. This assumption was further investigated by means of high scoring sequence similarity between distantly related organisms. Pre_GI was used to identify all BLASTN hits with an e-value of 0 and bit score larger than 1,000 between islands hosted by prokaryotes from different genera. These parameters were chosen to ensure the identification of highly similar islands which may have undergone a recent transfer event. Amelioration alters the genetic signature of an island and as such high scoring BLASTN values may be used to identify recent HT events. The inclusion of host habitat and taxonomic information in Pre_GI enables the overlap of host environmental and habitat information with sequence similarity.

Pre_GI was tasked to identify high scoring sequence similarity between different genera in the extreme and isolated Yellowstone National Park environment. *Geobacillus* sp. Y412MC52 contains an island located at position 1,909,275 - 1,954,266 which contains keyword confirmation and an overlap with IslandViewer which displays a high level of sequence similarity to an island found in *Alicyclobacillus acidocaldarius* subsp. *acidocaldarius* DSM 446 located at position 2,965,110 - 2,983,860. This is graphically displayed in Figure 49 below. These islands are hosted by organisms from different genera yet are located in a highly specific and volatile environment.

Figure 49: High scoring sequence similarity between an island hosted by *Geobacillus* sp. Y412MC52 and an *Alicyclobacillus acidocaldarius* subsp. *acidocaldarius* DSM 446 island. These organisms were isolated in the Yellowstone National Park, an extreme and explicit environment.

The example above was elaborated to the totality of the database in an attempt to identify habitat or environmental pools of HT. The combination of high scoring sequence similarity hits with host habitat information filtered by host genus highlights the importance of environment in HT events. The figures below indicate that there is a possible overlap of environments between islands hosted by different genera and as such may underline the importance of habitat and environment in HT events. The movement of genetic material between organisms seems more reliant on proximity than relatedness. Results are displayed below with Figure 50 and Figure 51 for grouping 1 based on high scoring sequence similarity.

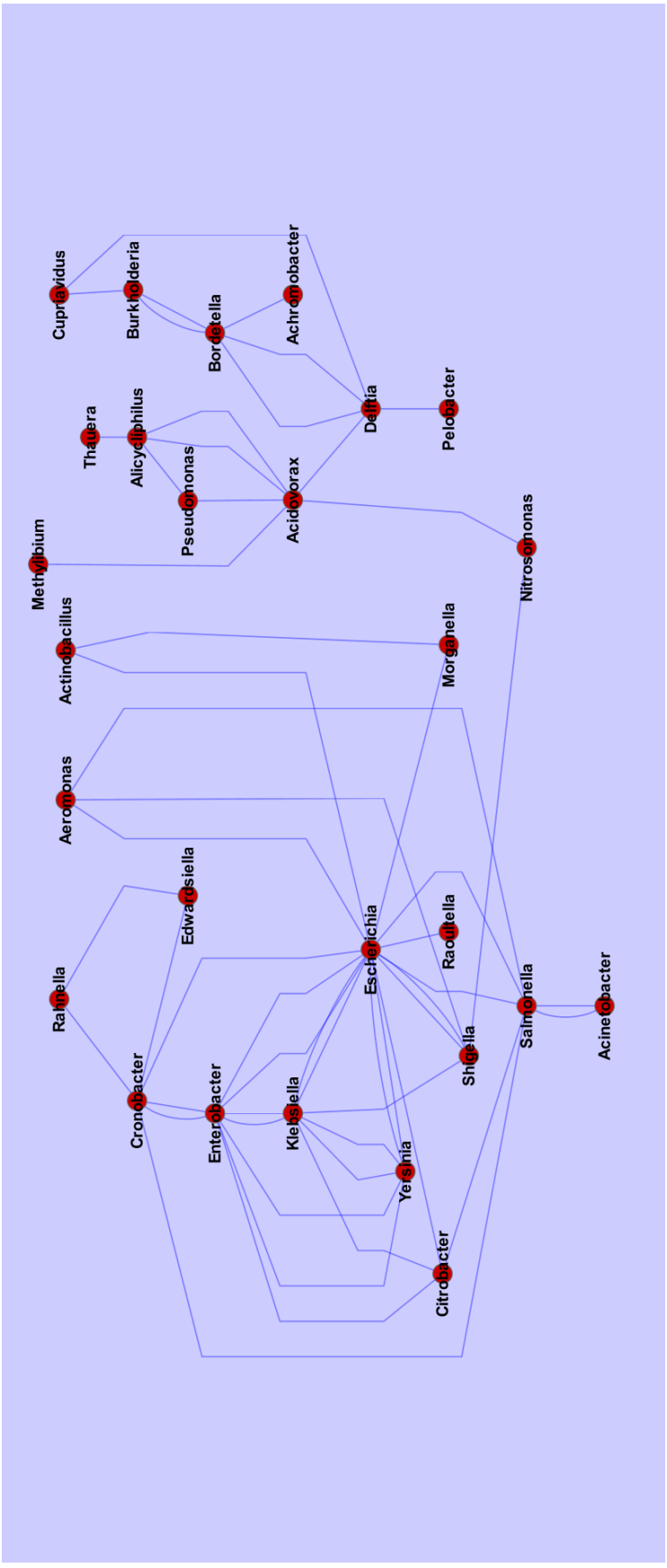


Figure 50: Sequence similarity links between islands from different genera.

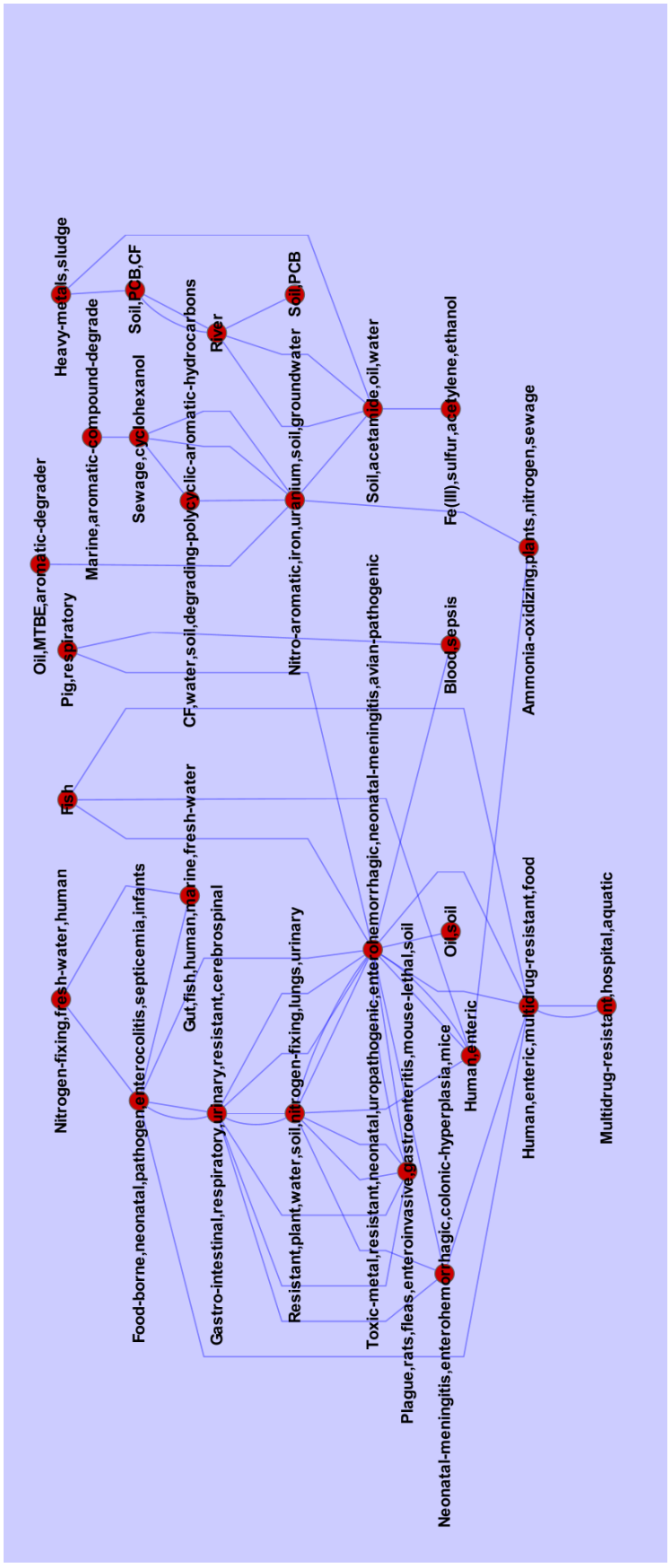


Figure 51: Environmental information for islands from different genera displaying high sequence similarity.

Grouping 2 genera and environmental information is presented in Figure 52 and Figure 53.

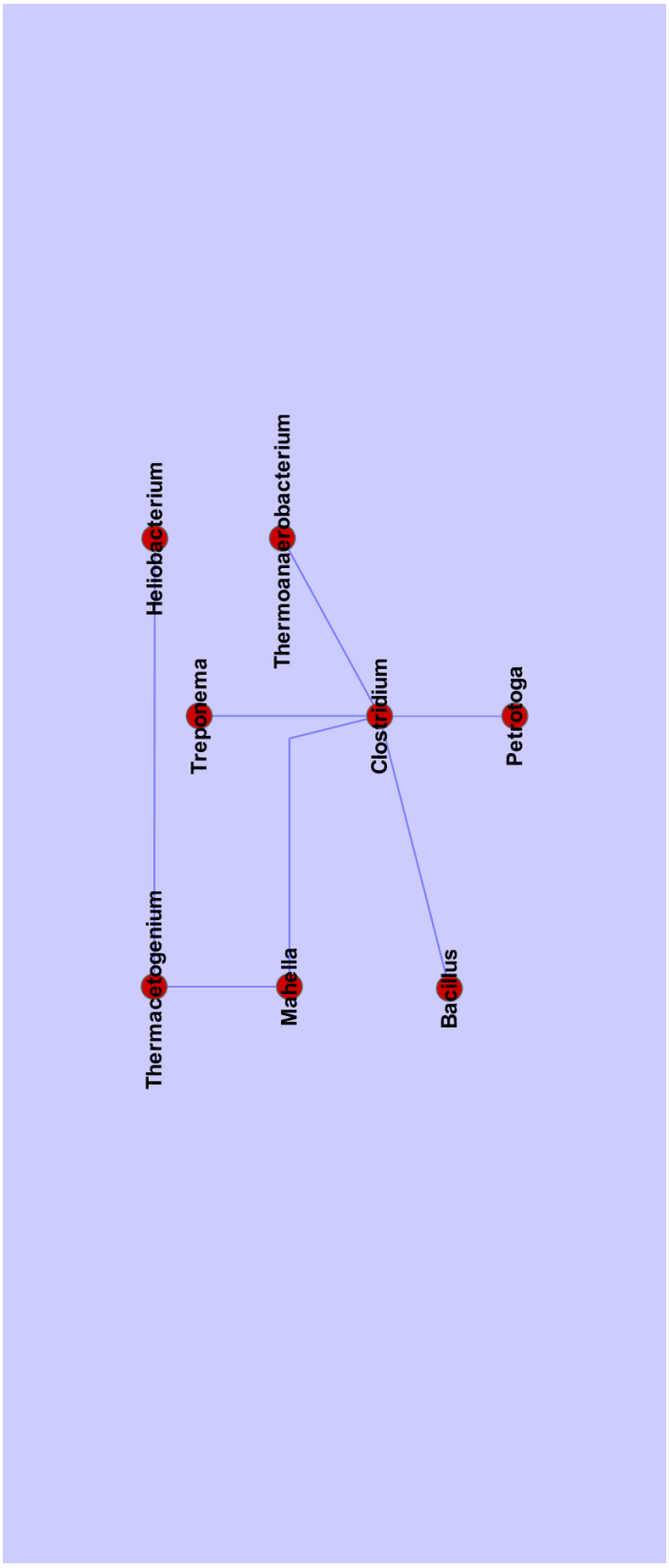


Figure 52: Grouping 2 taxonomic information for islands with high sequence similarity.

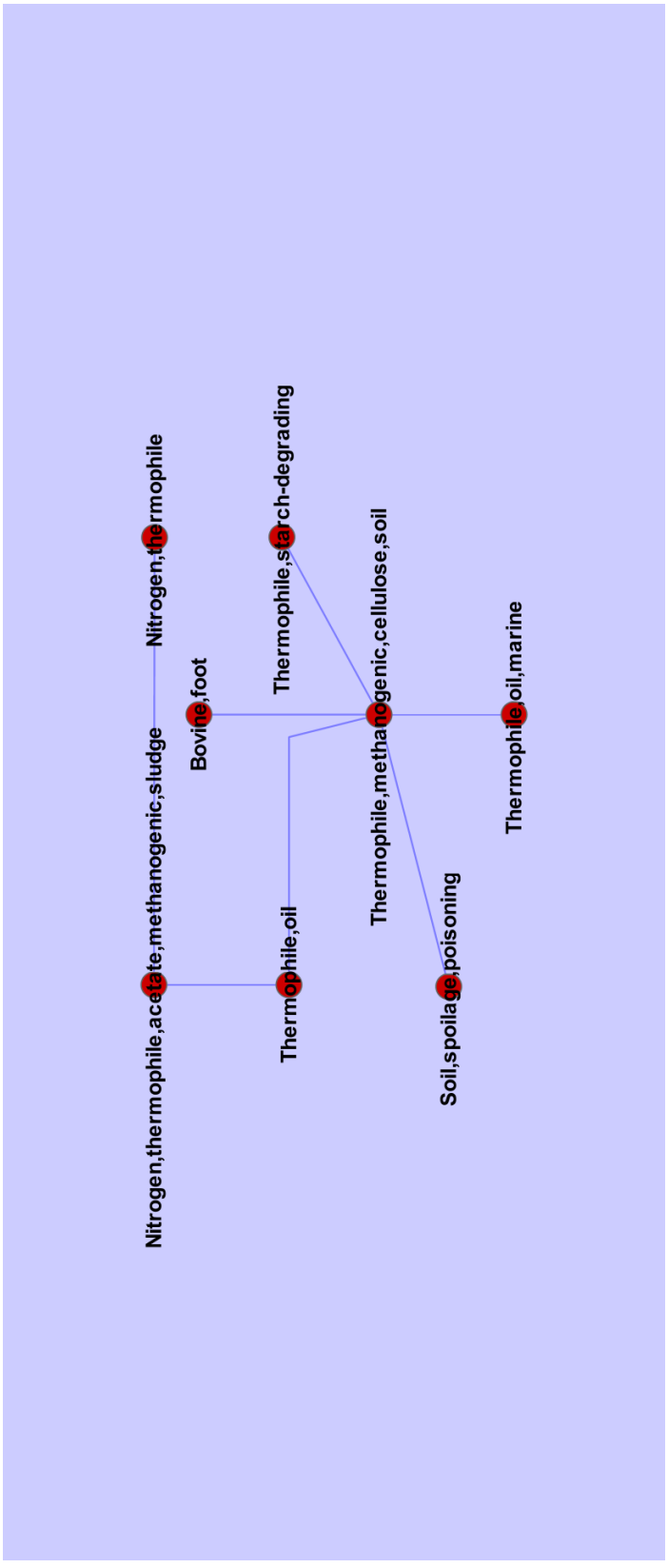


Figure 53: Environmental information for hosts of islands displaying high sequence similarity in grouping 2.

All other groupings are included below in Figure 54 and Figure 55.

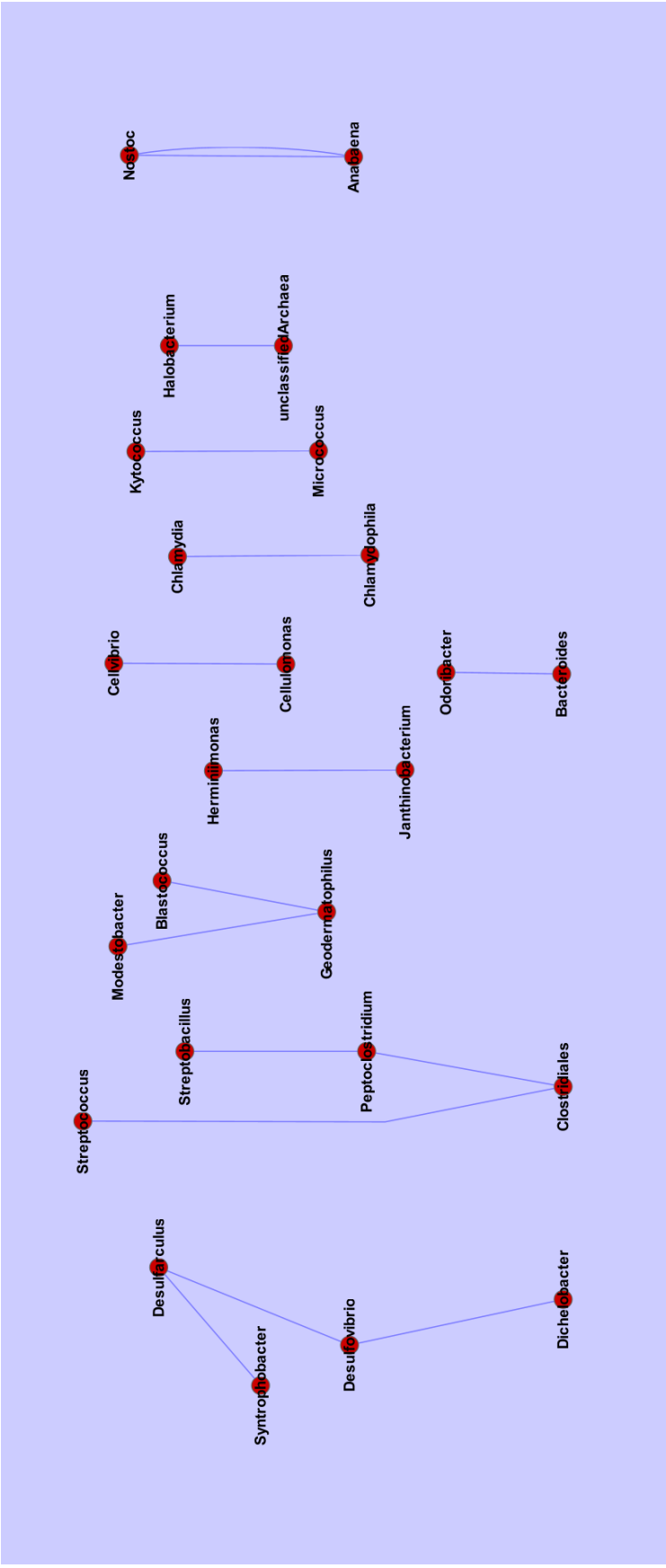


Figure 54: Taxonomic information for islands identified as sharing high sequence similarity.

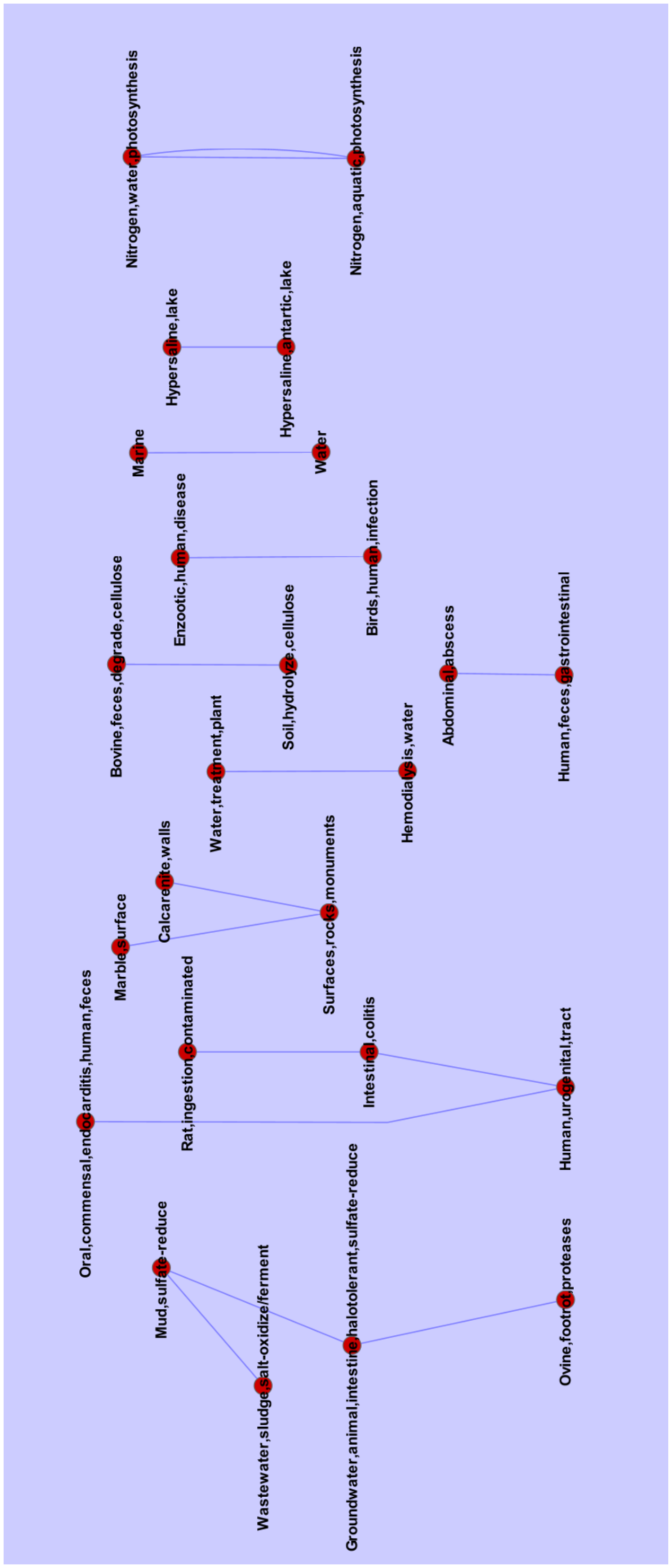


Figure 55: Host habitat information for islands displaying high sequence similarity.

The inclusion of newly sequenced genomes and their isolation information may alter these groups and environmental overlaps in the future.

5.5 Islands of Resistance

HT is a key driving force in the uprising of resistant bacteria. This enables these organisms to rapidly alter their repertoire and proliferate in an environment with multiple antibiotic substances. Pre_GI was inspected to reveal possible flow of resistance between diverse micro-organisms.

5.5.1 Flow detection

Pre_GI contains 2,712 proteins with a description related to "resistant". These genes were investigated to identify similar proteins in other islands and as such unveil possible movement of resistance between organisms. Movement of resistance genes between islands and organisms were determined by including sequence and compositional similarity hits between islands and genes. Islands with resistance genes were only included if the islands themselves displayed BLASTN hits and OUP hits to each other after sequence similarity was established between the proteins themselves. These islands were further filtered to include only those with a distinguishable direction of flow. This list was chiseled down to include only those hits between islands hosted by organisms within different phylums. This was included in order to ensure the exclusion of genes vertically inherited and thus increase the probability of HT. Detection of probable flow of resistant genes between phylums is displayed in Figure 56.

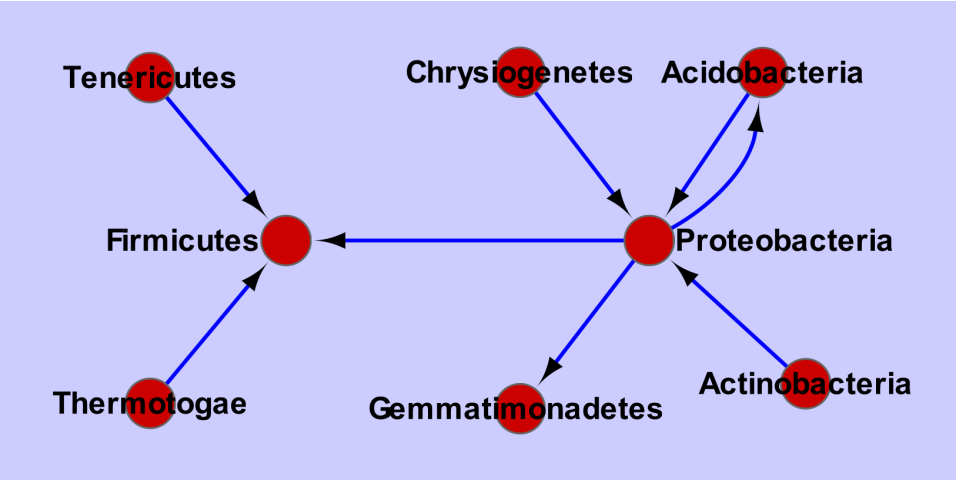


Figure 56: Movement of resistant genes between different phylums.

5.5.2 Island Groupings

The strict parameters described above concluded in the identification of 78 resistance proteins available in 44 islands that display HT between 8 different phyla. These HT could be divided into 8 groups based on the detection of flow between phyla. The cds descriptions in the figures to follow relate only to resistance proteins and their BLASTP hit cds description and do not entail all genes available in an island. Links displayed indicate that a sequence similarity between genes was detected as well as sequence and compositional similarity between the islands housing these proteins as a whole. Certain islands contain numerous resistant genes and as such these descriptions are separated by a semi-colon. Probable movement from donor to recipient is indicated by arrowheads.

Group 1 (Figure 57 and Figure 58) display a high affinity for multidrug resistance and related proteins shared between Proteobacteria, Firmicutes and Tenericutes. Figure 57 indicates a high incidence of ABC transport and ATP binding multidrug resistance proteins. These multidrug exporters extrude antibiotic and drug compounds by joining the hydrolysis of ATP to substrate transport across the cell membrane [6].

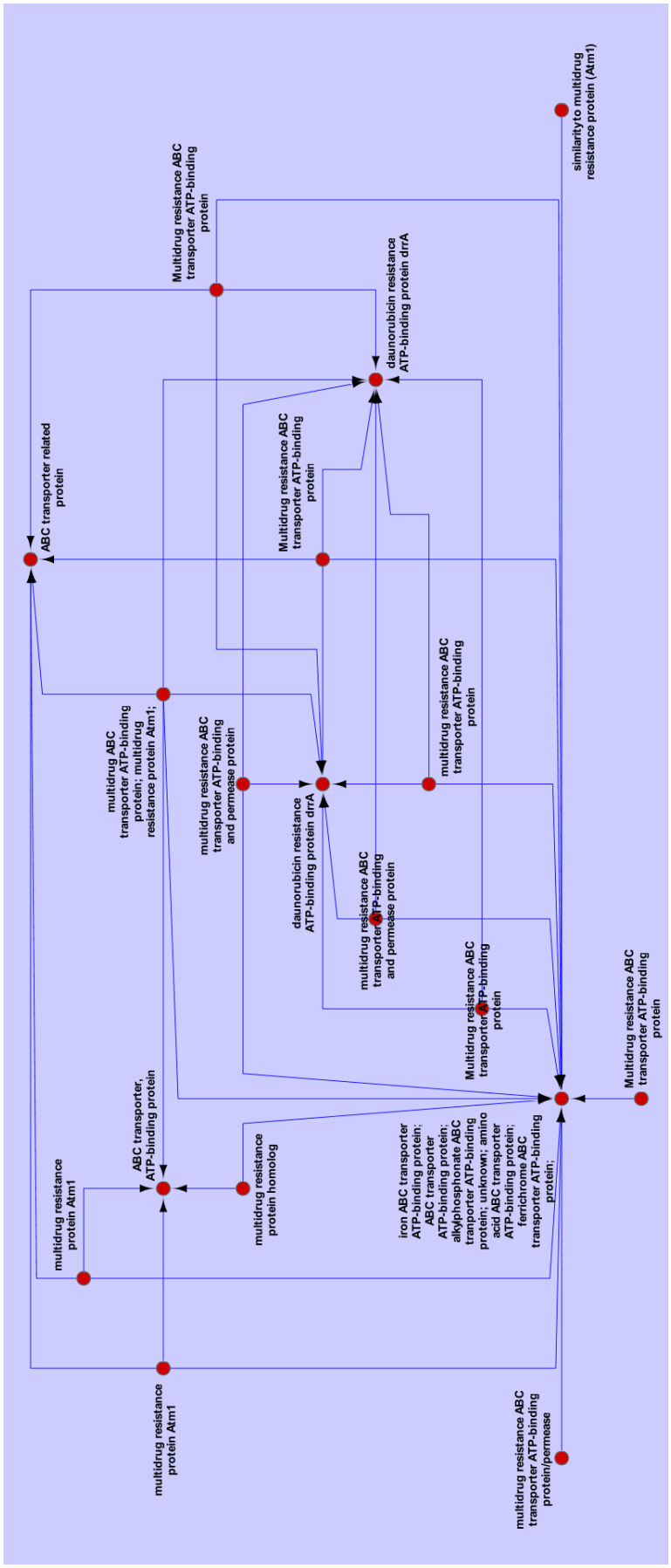


Figure 57: Group 1 movement and description of proteins.

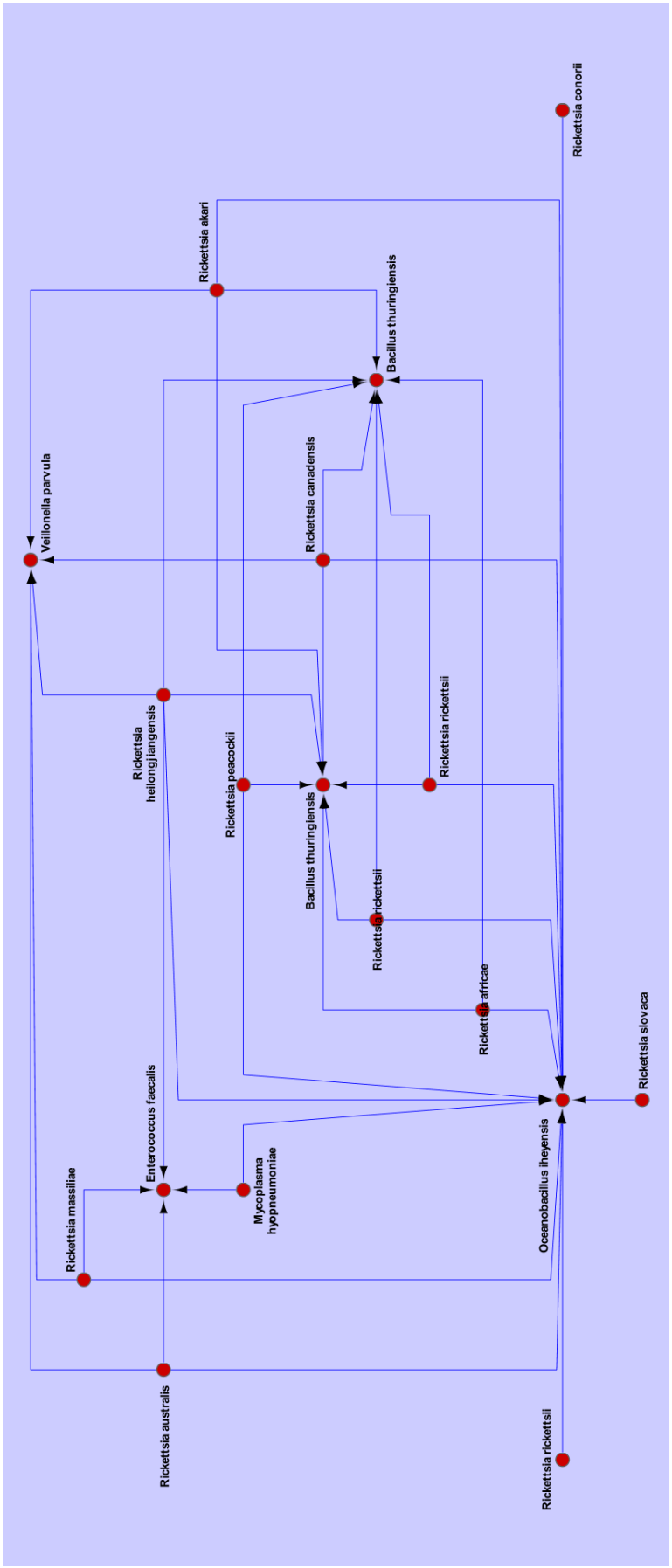


Figure 58: Group 1 movement of resistance genes between different species.

Grouping 2 contained movement of iron resistant proteins among 6 different genera and 4 phylums (Figure 59 and Figure 60).

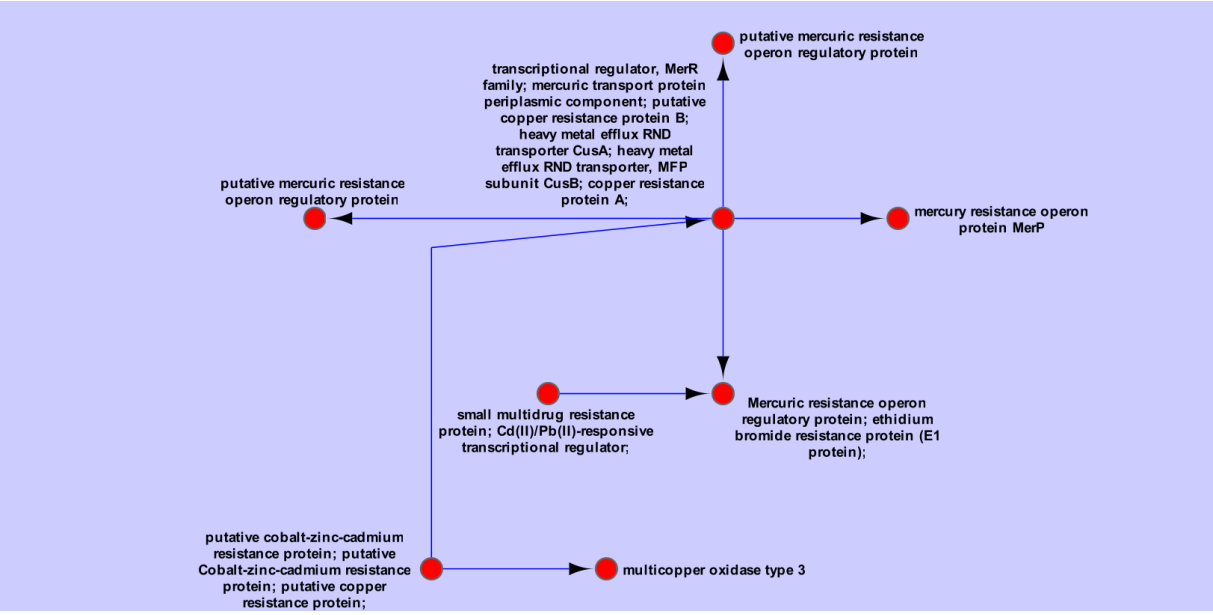


Figure 59: Grouping 2 movement and description of resistance proteins.

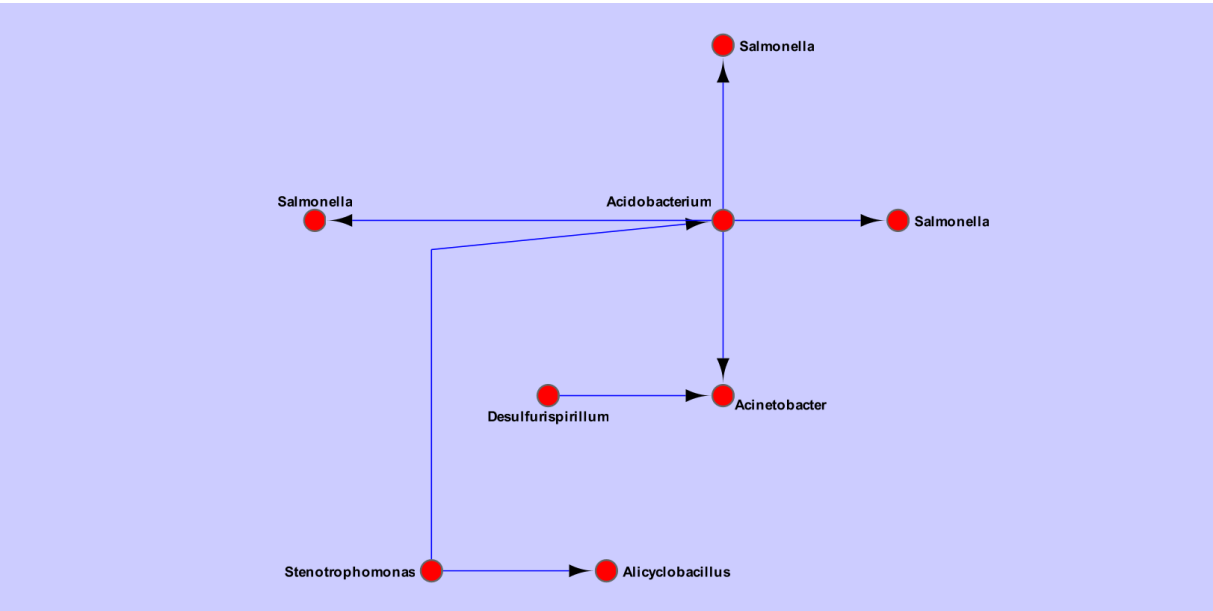


Figure 60: Group 2 movement of resistant genes between different genera.

Group 3 entailed the movement of arsenic resistance proteins between 3 genera and 3 disjoint phylums (Figure 61 and Figure 62).

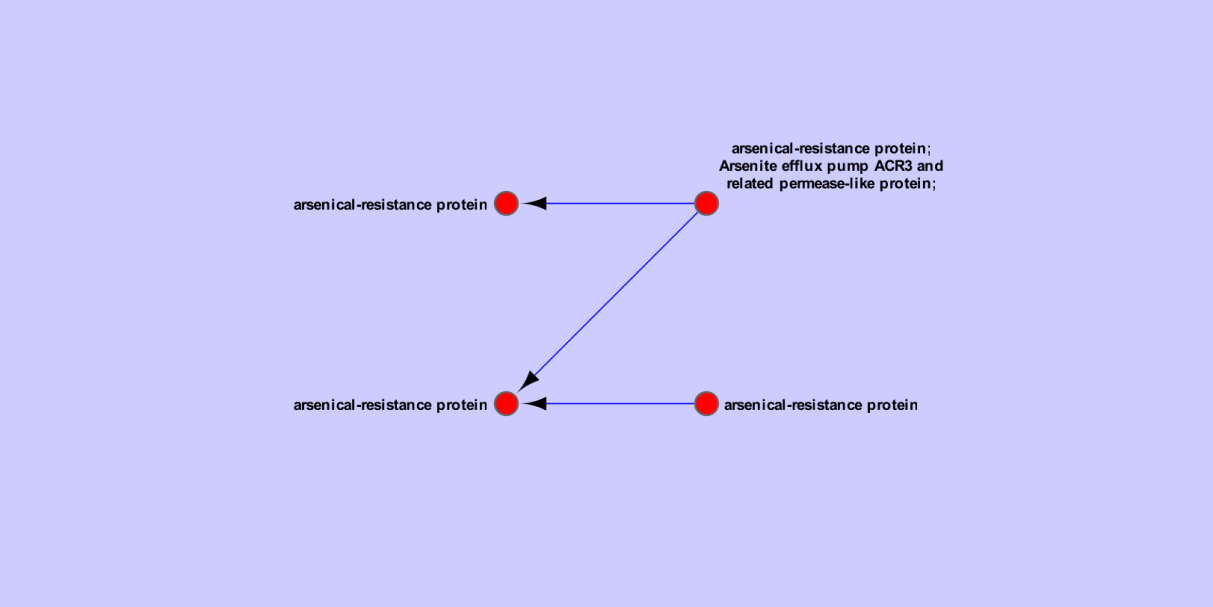


Figure 61: Movement of arsenic resistance proteins in group 3.

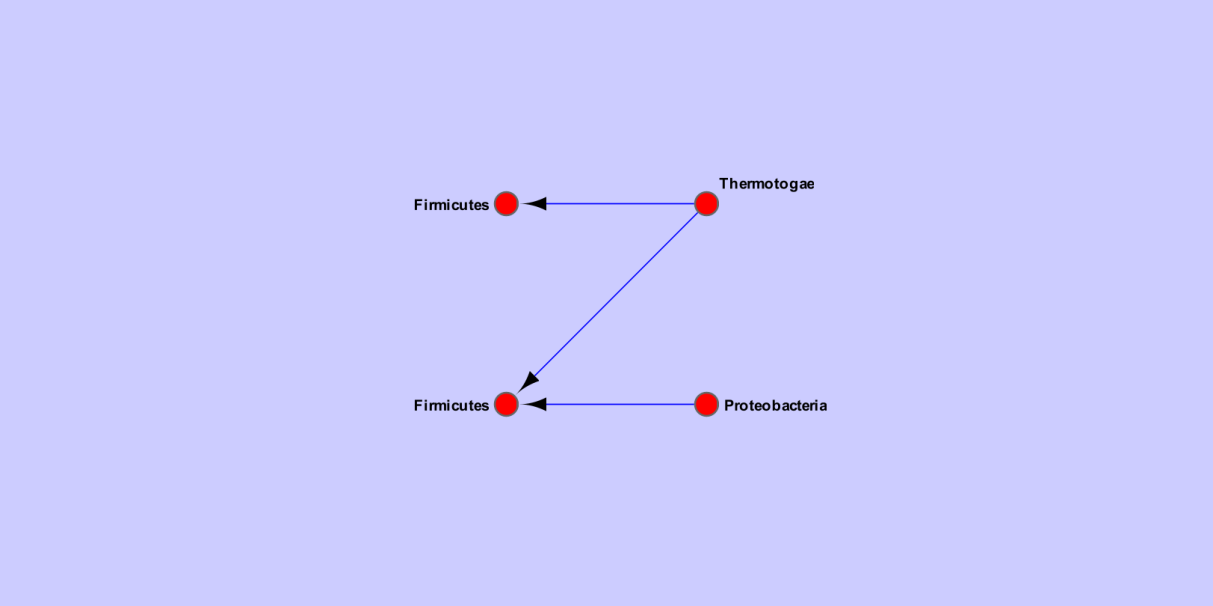


Figure 62: Group 3 movement of arsenic resistance related proteins between phyla.

Group 4 displays singular prediction of resistant gene flow by means of islands between different species (Figure 63 and Figure 64).

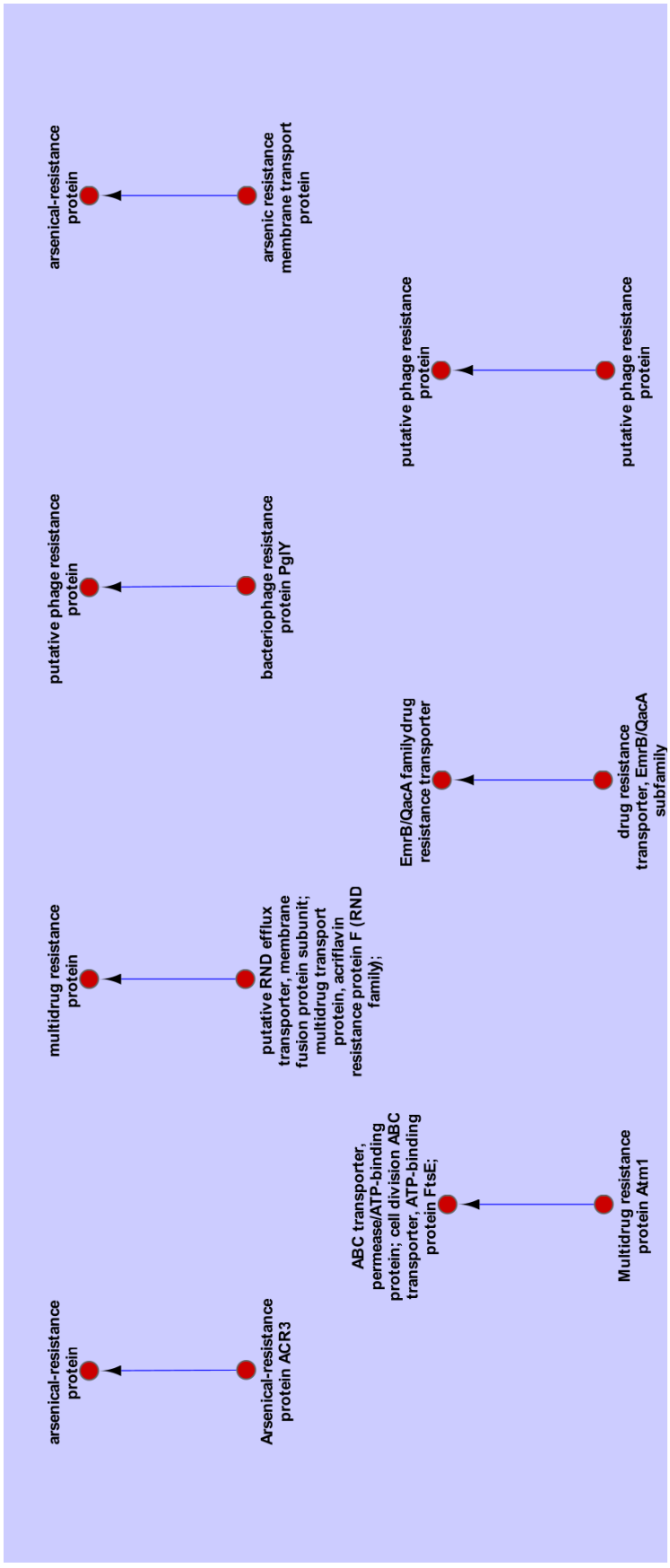


Figure 63: Group 4 displays singular flow of resistance related genes in non-overlapping genera.

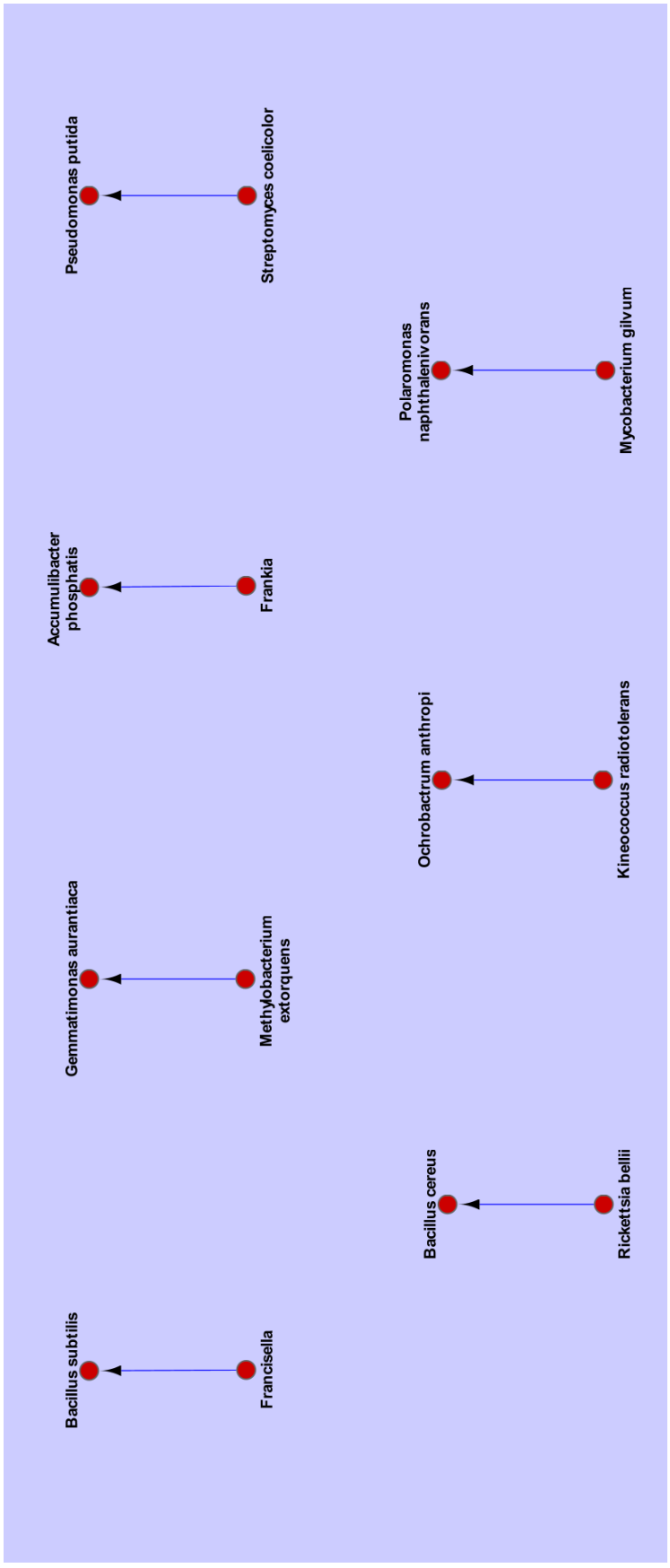


Figure 64: Group 4 movement of resistance related proteins between different species.

5.5.3 Protein Groupings

The islands identified above were further investigated to identify the specific proteins contained within them and the sequence similarity displayed to other islands from distinct phylums. It was ensured that the islands hosting these genes displayed sequence and compositional similarity as stipulated above. This resulted in the identification of hits between single proteins contained within island displaying proposed donor-recipient.

Five predominant island resistance protein groupings were isolated:

1. Multidrug ABC Transporters - This group of membrane proteins confer clinical resistance by transporting hydrophobic drugs and lipids across the cell membrane by coupling drug/lipid efflux with the energy obtained from the hydrolysis of ATP [6].

2. Mercury and other heavy metal resistance proteins - Mercury is the most toxic heavy metal to prokaryotes due to the high affinity for sulfur with microbial resistance hypothesized to have originated in geothermal environments [106].

3. Ethidium Bromide resistance - Ethidium bromide is a trypanocidal drug that influences nucleic acid synthesis by binding to DNA and RNA inhibiting DNA-polymerase and RNA-polymerase [32].

4. Arsenite resistance - Arsenic is an omnipresent toxic metalloid found in the environment due to geological, mining and agricultural human practices and natural events with resistance to arsenic a naturally occurring trait in prokaryotes in an effort to survive in arsenic rich environments [18].

5. Copper resistance - Various industrial, agricultural and mining activities elevate levels of the heavy metal and potentially toxic copper in the environment.

Group 1 (Table 9) consisted of multidrug resistant and related proteins with all elements connected through islands in distinct phylums. It should be emphasized that protein hits were only included if there was a difference in query and subject phyla. All proteins in this grouping were aligned with Clustal Omega [40] and displayed in Figure 65 and Figure 66 as a dendrogram.

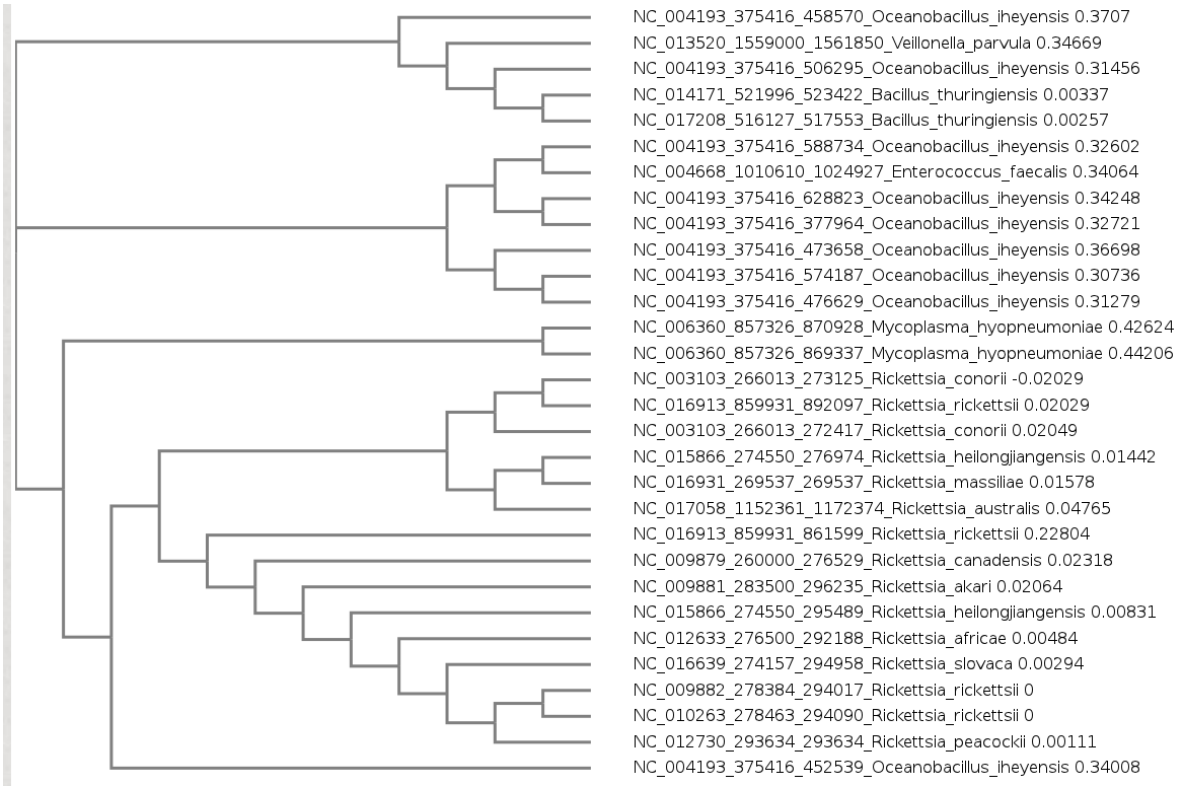


Figure 65: Dendrogram of proteins contained in grouping 1 with specie description.

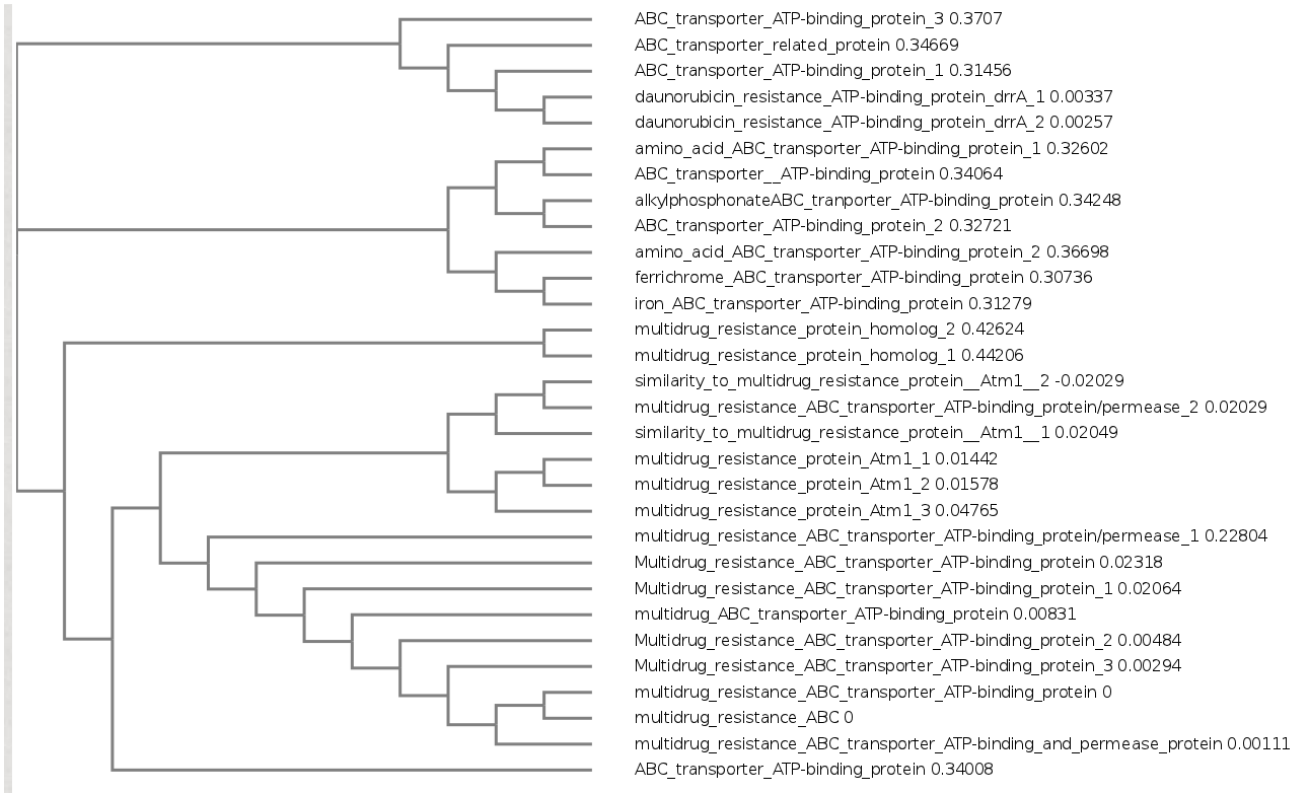


Figure 66: Dendrogram of proteins contained in grouping 1 with gene annotation.

Table 9: Resistance related protein grouping 1.

ID	Description	Species	Phylum
NC_016931:269537:269537	multidrug resistance protein Atm1	*Rickettsia massiliae*	Proteobacteria
NC_004193:375416:377964	ABC transporter ATP-binding protein	*Oceanobacillus iheyensis*	Firmicutes
NC_016639:274157:294958	Multidrug resistance ABC transporter ATP-binding protein	*Rickettsia slovaca*	Proteobacteria
NC_012633:276500:292188	Multidrug resistance ABC transporter ATP-binding protein	*Rickettsia africae*	Proteobacteria
NC_004668:1010610:1024927	ABC transporter, ATP-binding protein	*Enterococcus faecalis*	Firmicutes
NC_004193:375416:506295	ABC transporter ATP-binding protein	*Oceanobacillus iheyensis*	Firmicutes
NC_006360:857326:870928	multidrug resistance protein homolog	*Mycoplasma hyopneumoniae*	Tenericutes
NC_004193:375416:588734	amino acid ABC transporter ATP-binding protein	*Oceanobacillus iheyensis*	Firmicutes
NC_015866:274550:276974	multidrug resistance protein Atm1	*Rickettsia heilongjiangensis*	Proteobacteria
NC_010263:278463:294090	multidrug resistance ABC transporter ATP-binding and permease protein	*Rickettsia rickettsii*	Proteobacteria
NC_012730:293634:293634	multidrug resistance ABC transporter ATP-binding and permease protein	*Rickettsia peacockii*	Proteobacteria
NC_003103:266013:272417	similarity to multidrug resistance protein (Atm1)	*Rickettsia conorii*	Proteobacteria
NC_017208:516127:517553	daunorubicin resistance ATP-binding protein drrA	*Bacillus thuringiensis*	Firmicutes
NC_004193:375416:452539	ABC transporter ATP-binding protein	*Oceanobacillus iheyensis*	Firmicutes
NC_006360:857326:869337	multidrug resistance protein homolog	*Mycoplasma hyopneumoniae*	Tenericutes
NC_016913:859931:892097	multidrug resistance ABC transporter ATP-binding protein/permease	*Rickettsia rickettsii*	Proteobacteria
NC_004193:375416:628823	alkylphosphonate ABC tranporter ATP-binding protein	*Oceanobacillus iheyensis*	Firmicutes
NC_004193:375416:473658	amino acid ABC transporter ATP-binding protein	*Oceanobacillus iheyensis*	Firmicutes
NC_004193:375416:458570	ABC transporter ATP-binding protein	*Oceanobacillus iheyensis*	Firmicutes
NC_009882:278384:294017	multidrug resistance ABC transporter ATP-binding protein	*Rickettsia rickettsii*	Proteobacteria
NC_015866:274550:295489	multidrug ABC transporter ATP-binding protein	*Rickettsia heilongjiangensis*	Proteobacteria
NC_014171:521996:523422	daunorubicin resistance ATP-binding protein drrA	*Bacillus thuringiensis*	Firmicutes
NC_009881:283500:296235	Multidrug resistance ABC transporter ATP-binding protein	*Rickettsia akari*	Proteobacteria
NC_013520:1559000:1561850	ABC transporter related protein	*Veillonella parvula*	Firmicutes
NC_009879:260000:276529	Multidrug resistance ABC transporter ATP-binding protein	*Rickettsia canadensis*	Proteobacteria
NC_004193:375416:574187	ferrichrome ABC transporter ATP-binding protein	*Oceanobacillus iheyensis*	Firmicutes
NC_004193:375416:476629	iron ABC transporter ATP-binding protein	*Oceanobacillus iheyensis*	Firmicutes
NC_017058:1152361:1172374	multidrug resistance protein Atm1	*Rickettsia australis*	Proteobacteria
NC_003103:266013:273125	similarity to multidrug resistance protein (Atm1)	*Rickettsia conorii*	Proteobacteria
NC_016913:859931:861599	multidrug resistance ABC transporter ATP-binding protein/permease	*Rickettsia rickettsii*	Proteobacteria

Group 2 (Figure 67 and 68) housed members involved in mercury resistance and iron transport.

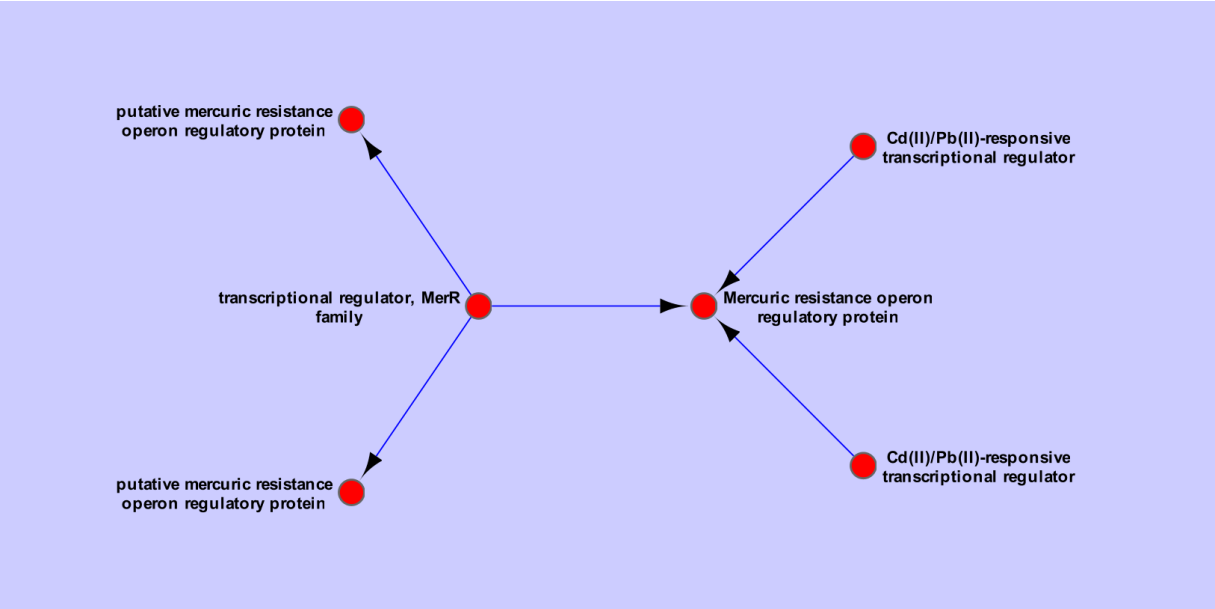


Figure 67: Group 2 protein descriptions and flow.

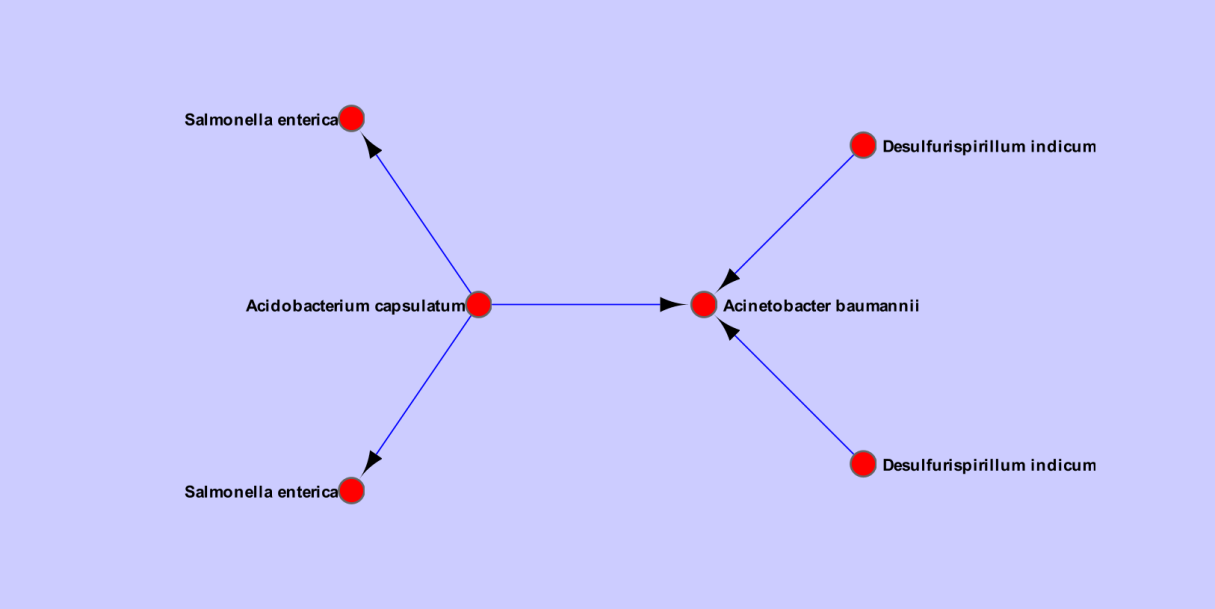


Figure 68: Group 2 proposed movement between different species.

Alignment of these protein sequences, displayed in Figure 69 and Figure 70, indicated the relationship between a transcriptional regulator MerR family protein found in island#9 hosted by *Acidobacterium capsulatum* ATCC 51196 and a putative mercuric resistance operon regulatory protein in island#2 hosted by *Salmonella enterica* subsp. *enterica* serovar Typhi str. CT18. *Acidobacterium capsulatum* ATCC 51196 was isolated from

acidic mine drainage in Japan and *Salmonella enterica* subsp. *enterica* serovar Typhi str. CT18 is a multidrug resistant strain which is human specific.

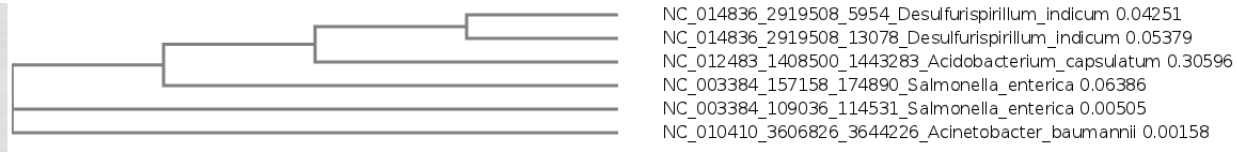


Figure 69: Dendrogram of group 2 proteins multiple sequence alignment with specie descriptions.

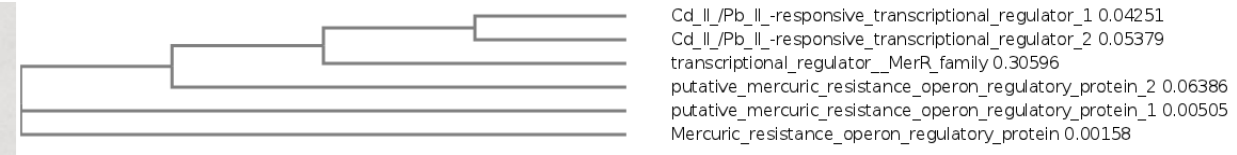


Figure 70: Dendrogram of group 2 proteins multiple sequence alignment with gene annotations.

Group 3 details the movement of a small multidrug resistance protein available in *Desulfurispirillum indicum S5* to ethidium bromide resistance proteins in island#5 hosted by *Acinetobacter baumannii* AYE (Figure 71 and Figure 72) and multiple sequence alignment of these 4 proteins in Figure 73 and Figure 74. *Acinetobacter baumannii* AYE is responsible for community-acquired infections and is highly resistant to antibiotics. This bacterium is commonly isolated from the hospital environment and hospitalized patients.

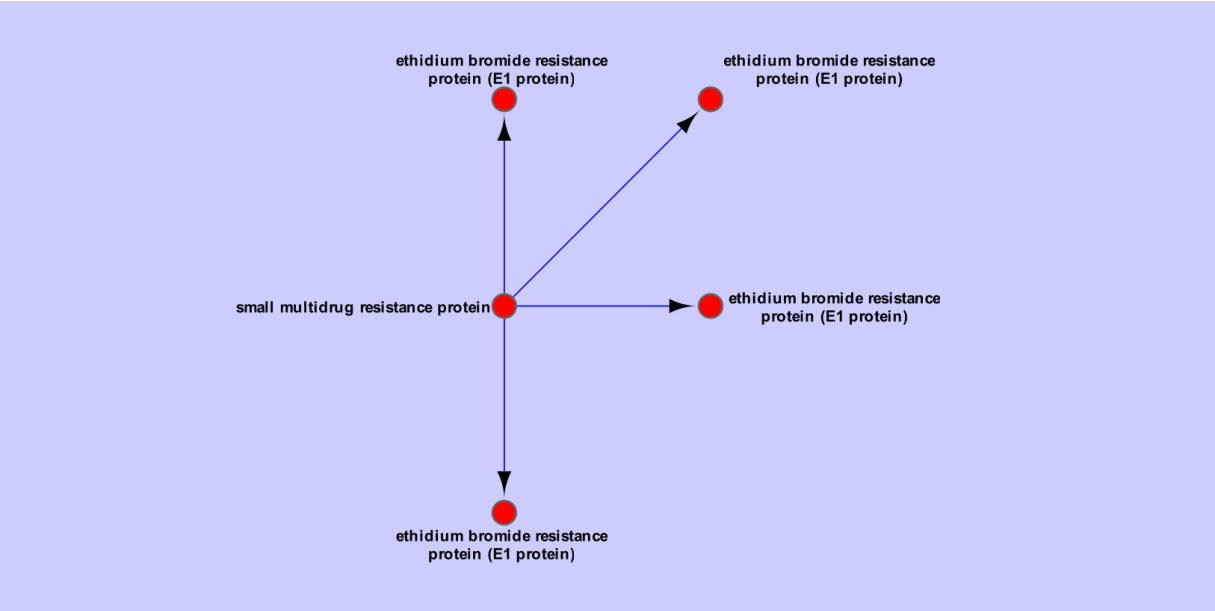


Figure 71: Proposed movement from a small multidrug resistant protein to multiple ethidium bromide resistance proteins.

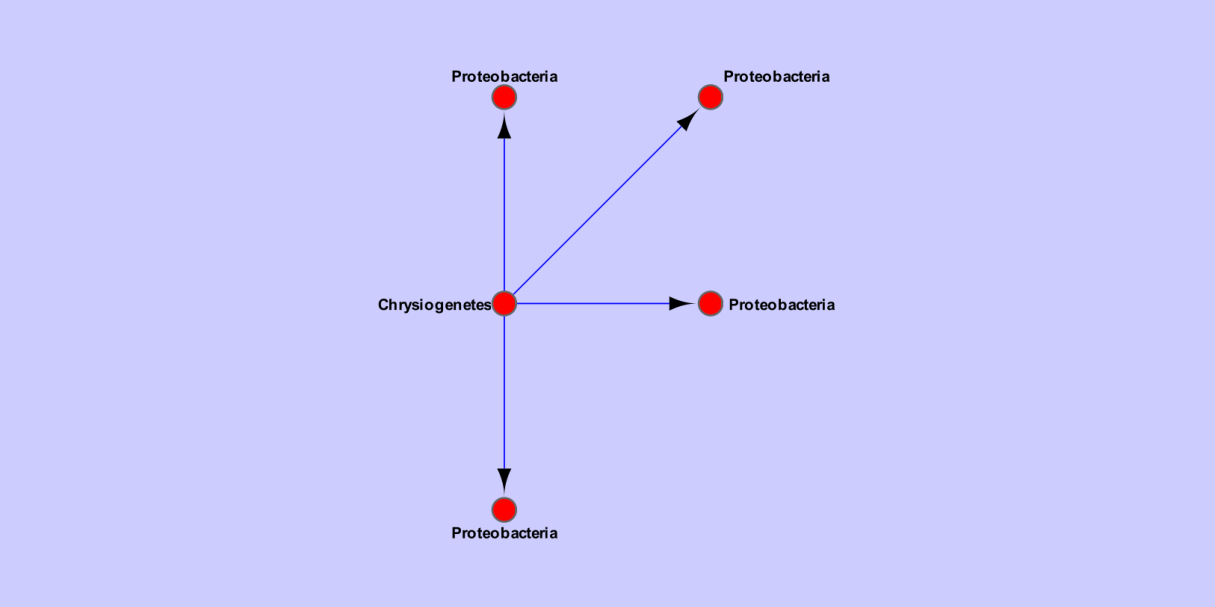


Figure 72: Group 3 probable movement between different phylums.

```
NC_010410:3606826:3620981_Acinetobacter_baumannii       MKGWLFLVIAIVGEVIATSALKSSEGFTKLAPSAVVIIGYGIAFYFLSLVLKSIPVGVAY
NC_010410:3606826:3651450_Acinetobacter_baumannii       MKGWLFLVIAIVGEVIATSALKSSEGFTKLAPSAVVIIGYGIAFYFLSLVLKSIPVGVAY
NC_010410:3606826:3655558_Acinetobacter_baumannii       MKGWLFLVIAIVGEVIATSALKSSEGFTKLAPSAVVIIGYGIAFYFLSLVLKSIPVGVAY
NC_014836:2919508:10162_Desulfurispirillum_indicum      MKGWLFLVIAIVGEVIATSALKSSEGFTKLAPSAVVIIGYGIAFYFLSLVLKSIPVGVAY
NC_010410:3606826:3675865_Acinetobacter_baumannii       MKGWLFLVIAIVGEVIATSALKSSEGFTKLAPSAVVIIGYGIAFYFLSLVLKSIPVGVAY
                                                        ************************************************************

NC_010410:3606826:3620981_Acinetobacter_baumannii       AVWSGLGVVIITAIAWLLHGQKLDAWGFVGMGLIIAAFLLARSPSWKSLRRPTPW
NC_010410:3606826:3651450_Acinetobacter_baumannii       AVWSGLGVVIITAIAWLLHGQKLDAWGFVGMGLIIAAFLLARSPSWKSLRRPTPW
NC_010410:3606826:3655558_Acinetobacter_baumannii       AVWSGLGVVIITAIAWLLHGQKLDAWGFVGMGLIIAAFLLARSPSWKSLRRPTPW
NC_014836:2919508:10162_Desulfurispirillum_indicum      AVWSGLGVVIITAIAWLLHGQKLDAWGFVGMGLIIAAFLLARSPSWKSLRRPTPW
NC_010410:3606826:3675865_Acinetobacter_baumannii       AVWSGLGVVIITAIAWLLHGQKLDAWGFVGMGLIIAAFLLARSPSWKSLRRPTPW
                                                        ********************************************************
```

Figure 73: Multiple alignment of a multidrug resistant protein found in *Desulfurispirillum indicum S5* 3 and ethidium bromide resistance proteins in *Acinetobacter baumannii* AYE.

```
ethidium_bromide_resistance_protein_(E1_protein)_1      MKGWLFLVIAIVGEVIATSALKSSEGFTKLAPSAVVIIGYGIAFYFLSLVLKSIPVGVAY
ethidium_bromide_resistance_protein_(E1_protein)_2      MKGWLFLVIAIVGEVIATSALKSSEGFTKLAPSAVVIIGYGIAFYFLSLVLKSIPVGVAY
ethidium_bromide_resistance_protein_(E1_protein)_3      MKGWLFLVIAIVGEVIATSALKSSEGFTKLAPSAVVIIGYGIAFYFLSLVLKSIPVGVAY
small_multidrug_resistance_protein                      MKGWLFLVIAIVGEVIATSALKSSEGFTKLAPSAVVIIGYGIAFYFLSLVLKSIPVGVAY
ethidium_bromide_resistance_protein_(E1_protein)_4      MKGWLFLVIAIVGEVIATSALKSSEGFTKLAPSAVVIIGYGIAFYFLSLVLKSIPVGVAY
                                                        ************************************************************

ethidium_bromide_resistance_protein_(E1_protein)_1      AVWSGLGVVIITAIAWLLHGQKLDAWGFVGMGLIIAAFLLARSPSWKSLRRPTPW
ethidium_bromide_resistance_protein_(E1_protein)_2      AVWSGLGVVIITAIAWLLHGQKLDAWGFVGMGLIIAAFLLARSPSWKSLRRPTPW
ethidium_bromide_resistance_protein_(E1_protein)_3      AVWSGLGVVIITAIAWLLHGQKLDAWGFVGMGLIIAAFLLARSPSWKSLRRPTPW
small_multidrug_resistance_protein                      AVWSGLGVVIITAIAWLLHGQKLDAWGFVGMGLIIAAFLLARSPSWKSLRRPTPW
ethidium_bromide_resistance_protein_(E1_protein)_4      AVWSGLGVVIITAIAWLLHGQKLDAWGFVGMGLIIAAFLLARSPSWKSLRRPTPW
                                                        ********************************************************
```

Figure 74: Multiple alignment of a multidrug resistant protein found in *Desulfurispirillum indicum S5* 3 and ethidium bromide resistance proteins in *Acinetobacter baumannii* AYE with gene annotations.

Group 4 was congested with arsenite and arsenical resistance proteins found in 4 islands from 4 different strains (Figure 75 and Figure 76).

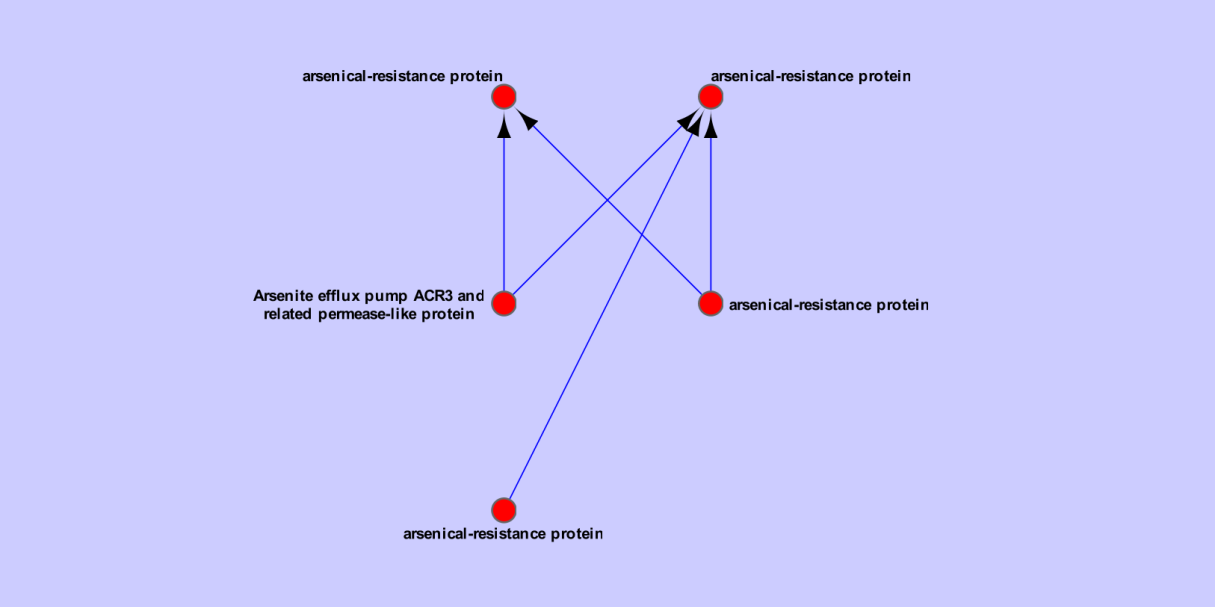


Figure 75: Group 4 movement of arsenic resistance.

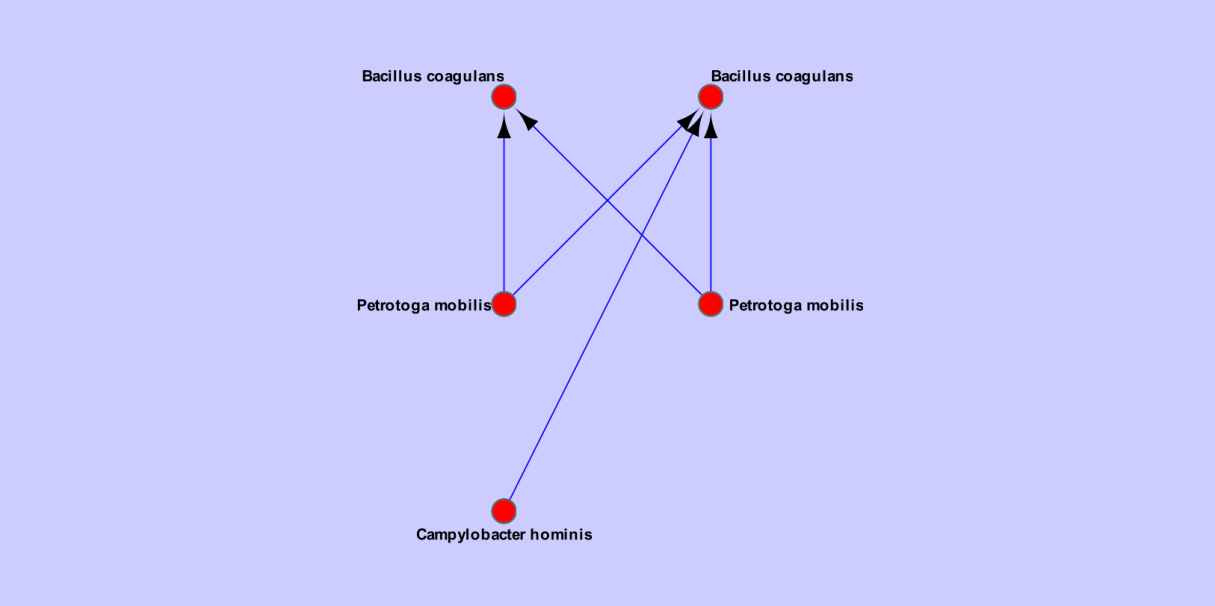


Figure 76: Group 4 movement between different species.

Campylobacter hominis ATCC BAA-381 was isolated from the faeces of a healthy human. Members of this genus are one of the most common causes of bacterial gastroenteritis (campylobacteriosis). *Bacillus coagulans* 36D1 and *Bacillus coagulans* 2-6 are lactic acid bacterium and have been identified as potential probiotics. *Petrotoga mobilis* SJ95 is a thermophile isolated from production water of a oil reservoir. Multiple sequence alignment of all 5 proteins are displayed in Figure 77 and Figure 78.

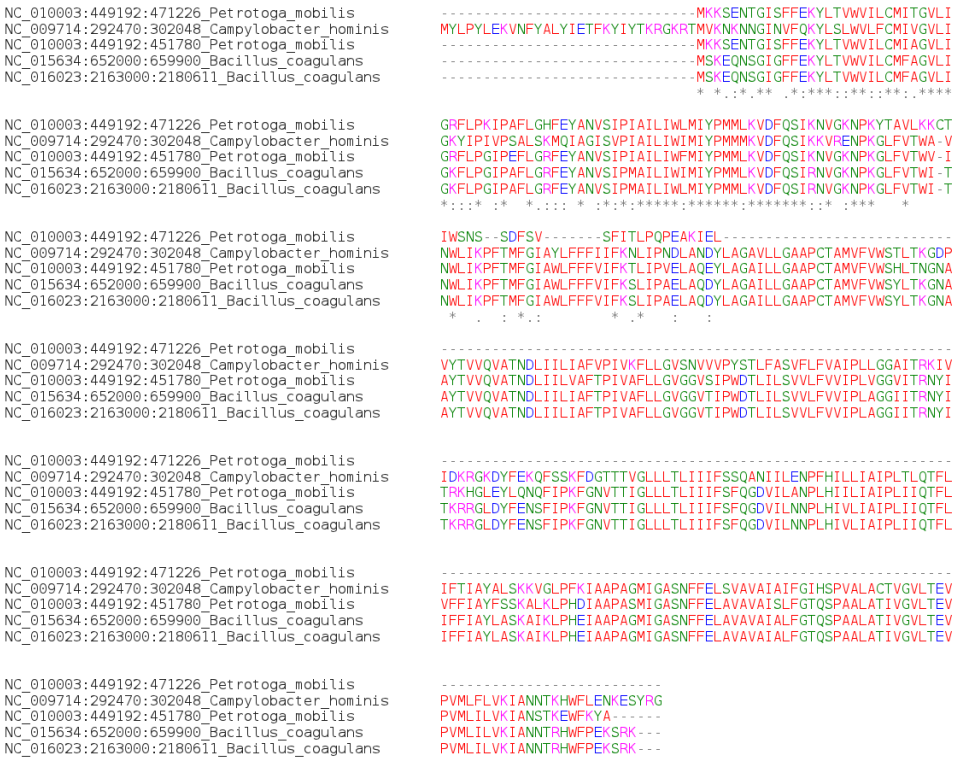


Figure 77: Multiple sequence alignment of arsenic related resistance proteins with specie description.

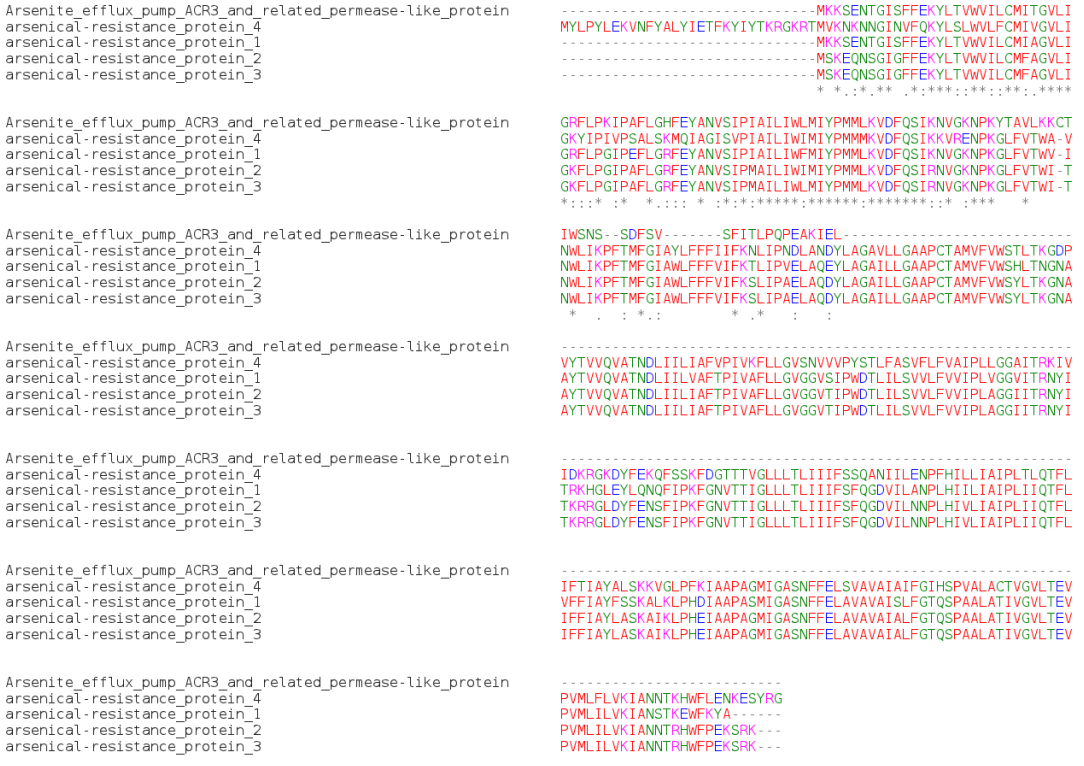


Figure 78: Multiple sequence alignment of arsenic related resistance proteins with cds descriptions.

Group 5 entails the movement of copper resistance proteins from a single source to 2 different species and genera of bacteria (Figure 79 and Figure 80).

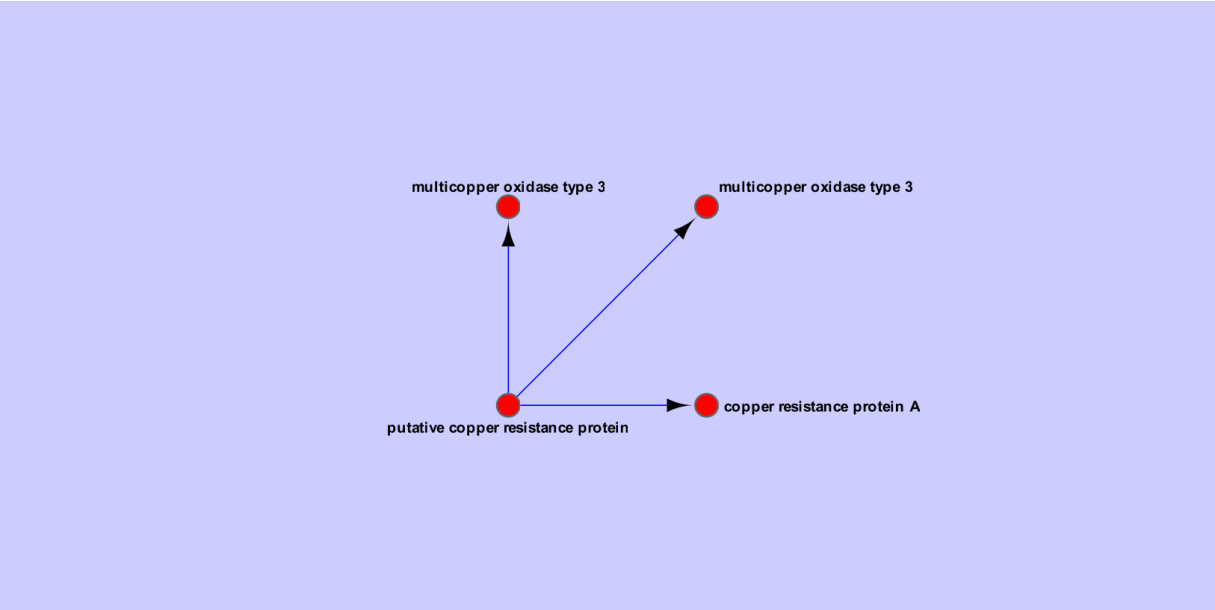


Figure 79: Movement of copper resistance proteins in group 5.

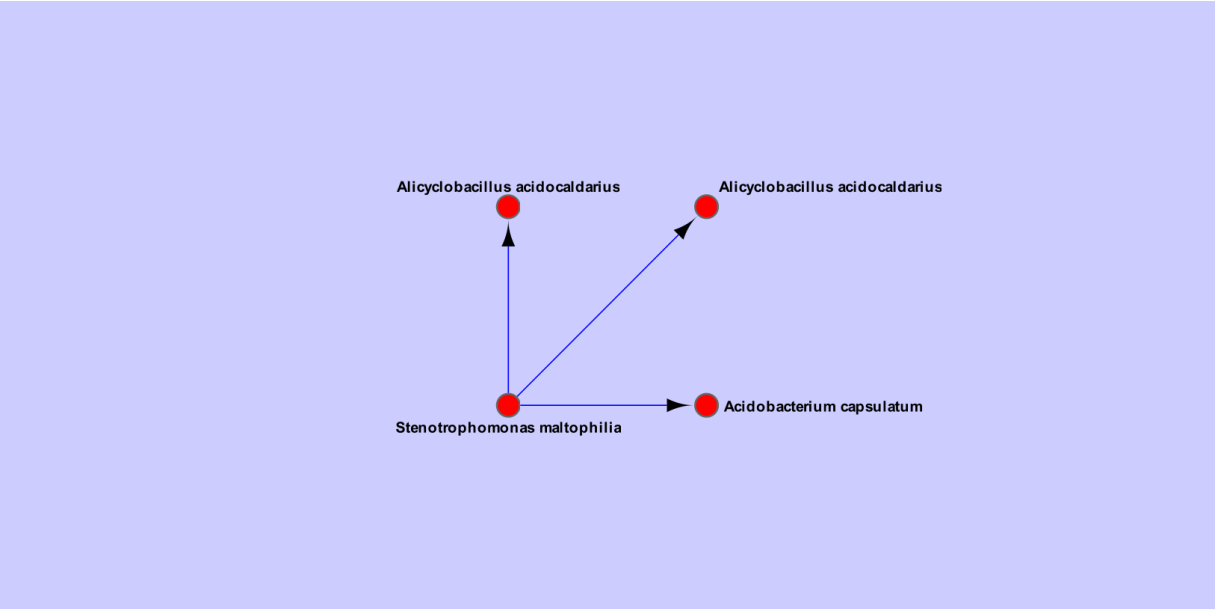


Figure 80: Movement of copper resistance between genera.

Stenotrophomonas maltophilia K279a was isolated from a blood infection and has resistance to numerous commonly used antibiotics. *Alicyclobacillus acidocaldarius* subsp. *acidocaldarius* DSM 446 is an acidophilic thermophile isolated from the Yellowstone National Park. *Acidobacterium capsulatum* ATCC 51196 was isolated from acidic mine drainage in Japan. The dendrogram of these copper resistance proteins (Figure 81 and

Figure 82) portrays the close relationship between *Acidobacterium capsulatum* ATCC 51196 and *Stenotrophomonas maltophilia* K279a.

Figure 81: Multiple sequence alignment dendrogram of group 5 copper resistance proteins with specie descriptions.

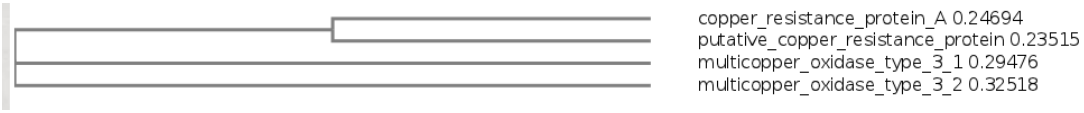


Figure 82: Multiple sequence alignment dendrogram of group 5 copper resistance proteins with gene annotations.

Multiple other detection of HT of resistance related proteins between distantly related organisms is displayed in Figure 83 and Figure 84.

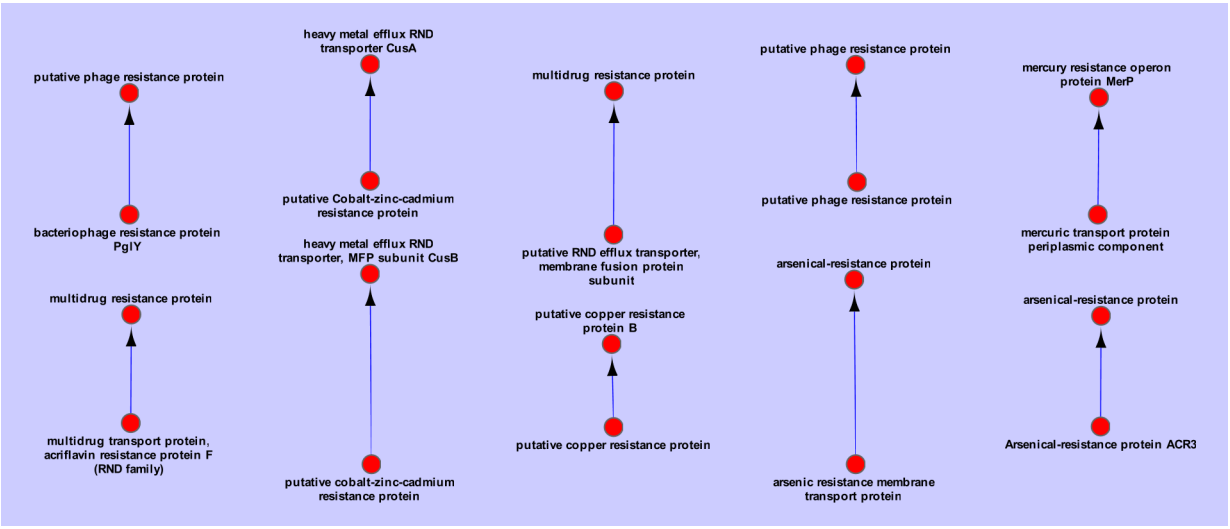


Figure 83: Numerous HT events related to resistance proteins.

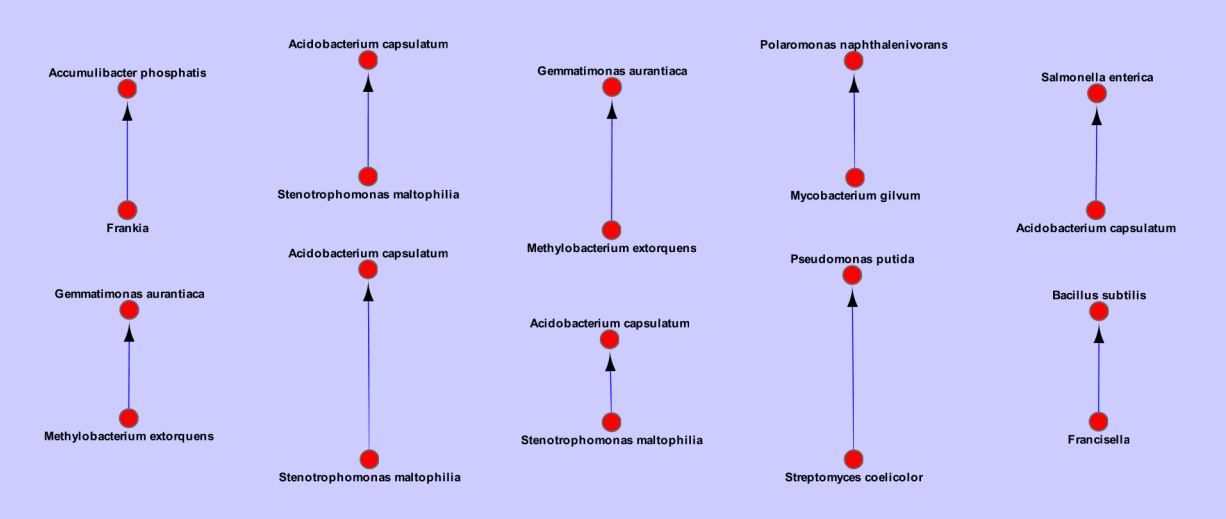


Figure 84: Movement of resistant genes between various species.

5.6 Discussion

The ability to remove islands from their hosts and regard them as individuals may greatly help in the understanding of these selfish segments of genetic material. The power of large data sets and inclusion of various known resources could improve the general picture of island existence. The individuality of islands enables characterization of the island community and may lead to a greater understanding of island ontology or being.

Host habitat and location is suggested to be of great influence in island existence and movement. The importance of environmental or habitat overlap between organisms for HT is illustrated above. Physical contact is a necessity for HT and as such habitat union is paramount. This should apply to all mechanisms of transfer inclusive of transformation as time of habitat occupation is not factored in.

The identification of resistance islands and the proposed movement between species will greatly aid in our understanding on the origin and capacity of these islands and the generation of resistant bacterial strains. The upsurge of antibiotic resistant species and strains is empowered by the availability of resistance islands to receptive bacteria in times of environmental pressures. The rapid alteration of a vulnerable species or strain may be resultant of HT. These movements leave proposed footprints which enables researchers to retrace probable pathways of travel and as such locate possible origins of resistance.

Current Pre_GI island information offers various analytical applications in various arenas of HT and MGE research. Current content is open to numerous novel avenues of research due to the inclusion of large amounts of information regarding island ontology, history and relations.

6 Chater 6: Case Studies and Applications

> For the first half of geological time our ancestors were bacteria. Most creatures still are bacteria, and each one of our trillions of cells is a colony of bacteria.
> Richard Dawkins

Pre_GI was developed to be a contemporary and analytical resource with various applications in the field of HT and MGE studies. The availability of extensive literature and newly sequenced genomes affords us the opportunity to develop novel pipelines of analysis with regards to genetic exchange in communities. The application of Pre_GI in these future areas of research is demonstrated below.

6.1 Testing Hypotheses in Literature

Various hypotheses of HT in prokaryotes is available in current literature and was compared to the information available in Pre_GI.

6.1.1 *Staphylococcus*

Staphylococcus aureus and *Staphylococcus epidermidis* are opportunistic pathogens associated with high levels of resistance, mortality and morbidity [42]. Méric et al. described high levels of transfer between these species emphasizing the movement of antibiotic and metal resistance. Pre_GI indicated 2 species of *Staphylococcus* that incorporate methicillin resistance proteins. *Staphylococcus lugdunensis* HKU09-01 hosted an island located at 1,723,939 - 1,746,135 that includes a protein involved in methicillin resistance that displayed highest sequence similarity to island#1 in *Staphylococcus epidermidis* RP62A plasmid pSERP. This plasmid harbors various proteins involved in antibiotic resistance, *e. g.* penicillinase repressor, streptothricin acetyltransferase and beta-lactamase, with both these islands including key word confirmation. The methicillin resistant strain *Staphylococcus aureus* subsp. *aureus* COL revealed high compositional similarity to *Staphylococcus epidermidis* RP62A plasmid pSERP island#1 and high sequence similarity to *Staphylococcus epidermidis* ATCC 12228 island#6 which houses iron transport proteins.

6.1.2 *Streptococcus*

MGE bring forth new phenotypic features, including antibiotic resistance, to recipient strains and HT has been well documented between beta-hemolytic *Streptococcus* strains with the emphasis on *Streptococcus dysgalactiae* subsp. *equisimilis*, *Streptococcus pyogenes*, *Streptococcus agalactiae* and *Streptococcus suis* [33]. Smyth et al. demonstrated

movement from *Streptococcus dysgalactiae* subsp. *equisimilis* to various *Streptococcus* strains. These results were tested by means of compositional similarity to indicate possible direction of flow between host organisms. Beta-hemolytic *Streptococcus dysgalactiae* subsp. *equisimilis* GGS_124 accommodates 5 regions of HT with all islands inclusive of transfer key words and displaying overlaps with IslandViewer. Various compositional similarity links between *Streptococcus dysgalactiae* islands and *Streptococcus suis* P1/7 are given. These links unfortunately do not contain enough certainty on the direction of movement to confidently ascertain donor and recipient. Notable *Streptococcus dysgalactiae* island#3 displays sequence similarity to island#2 in *Streptococcus suis* P1/7 which contains a metallo-beta-lactamase superfamily protein. These proteins are involved in the breakdown of antibiotics by resistant bacteria. Movement was predicted by Pre_GI from *Streptococcus dysgalactiae* island#1 (donor) to *Streptococcus agalactiae* A909 island#2 (recipient).

6.1.3 *Enterobacteriaceae*

An occurrence of interspecies transfer was reported in an Australian hospitalized patient with regards to a bla_{IMP-4} plasmid from *Enterobacter cloacae* to *Escherichia coli* [48] within a period of months. Pre_GI was inspected to reveal any similar transfers between *Enterobacter cloacae* and *Escherichia coli*. The database contains 4 strains of *Enterobacter cloacae* totaling 68 islands. The interspecies transfer described by Sidjabat et al. occurred in a relatively short period of time and as such amelioration was not integral. *Enterobacter cloacae* islands were examined for high scoring sequence similarity to a species of *Escherichia coli*. All islands predicted in *Enterobacter cloacae* displayed high homology and probable transfer to *Enterobacteriaceae* species. Island#30 in *Enterobacter cloacae* subsp. *cloacae* ATCC 13047 contained island confirmation key words and has an overlap with IslandViewer. The highest scoring BLASTN hit was found to *Escherichia coli* 55989 island#19 and is graphically presented in Figure 85.

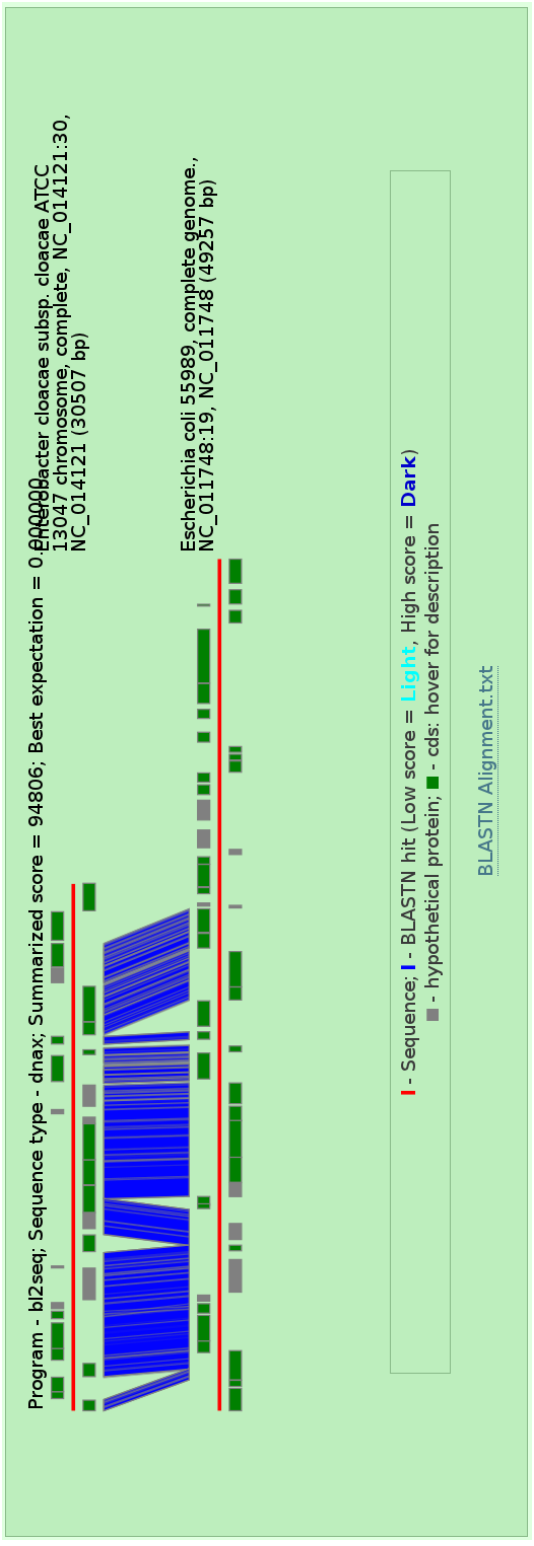


Figure 85: Highest scoring sequence similarity for *Enterobacter cloacae* subsp. *cloacae* ATCC 13047 island#30 against *Escherichia coli* 55989 island#19.

Escherichia coli 55989 island#19 includes key word confirmation and is also predicted by IslandViewer. This island contains various proteins related to resistance and efflux. *Enterobacter cloacae* SCF1 includes 2 islands (2 and 22) both containing key words and

predicted by IslandViewer that indicate a highest scoring sequence similarity to an *Escherichia coli* strain. The islands in *Enterobacter cloacae* strains mentioned above furthermore present high levels of compositional similarity to various *Escherichia coli* strains. The high level of similarity between islands of *Enterobacter cloacae* and *Escherichia coli* indicate probable transfer events with a large incidence of proteins related to resistance and pathogenicity.

6.1.4 *Streptococcaceae*

Streptococcus macedonicus is primarily isolated from foods akin to *Streptococcus thermophilus*, yet it is closely related to commensal opportunistic pathogens such as *Streptococcus bovis* and *Streptococcus equinus* [66]. Papadimitriou et al. propose that a plasmid isolated from the dairy strain *Streptococcus macedonicus* was amassed from *Lactococcus lactis* strains due to an ancestral transfer event that occurred in dairy products and as such conclude a dairy origin for *Streptococcus macedonicus* ACA-DC 198. Pre_GI was tasked to identify similar movement between species as stipulated by Papadimitriou et al. The strain of interest is included in the database with all comparisons contained. *Streptococcus macedonicus* ACA-DC 198 displayed 2 areas of HT with island#1 located at position 1,371,997 - 1,396,685 and island#2 at 1,914,481 - 1,946,907. PAIDB and IslandViewer contained a island which overlaps with island#1 and IslandViewer displayed an overlap for island#2. Pre_GI indicated that island#1 did not include a key word but overlap with 2 other databases increases the probability of a true positive prediction. This island displayed a compositional similarity of above 80% to an island found in *Lactococcus lactis* subsp. *lactis* IO-1 DNA with proposed movement from *Lactococcus lactis* to *Streptococcus macedonicus*. It should be noted that *Lactococcus lactis* subsp. *lactis* IO-1 DNA is a non-dairy lactic acid bacterium isolated from water in the drain pit of a kitchen sink. This is in contrast to Papadimitriou et al. whom concluded that the transfer event occurred in a milk environment. Pre_GI furthermore predicted transfer to *Streptococcus macedonicus* ACA-DC 198 from *Thermoanaerobacter mathranii* subsp. *mathranii* str. A3 and *Carnobacterium* sp. 17-4 both of which are found in diary products.

6.1.5 Proteobacteria

HT of *lux* genes, bacterial light production genes, have been identified to occur across the species barrier from members of *Aliivibrio* to *Shewanella* [50]. This was attempted in Pre_GI to identify possible transfer of genes from *Aliivibrio* to *Shewanella*. *Aliivibrio salmonicida* LFI1238 was identified as a potential donor with an island located at 871,445 - 892,418 displaying multiple compositional similarity links and direction of flow to *Shewanella* species as displayed in Figure 86.

Islands with an asterisk (*) contain ribosomal proteins or RNA related elements and may indicate a False Positive Prediction!

Subject Island	Subject Host Description	Compositional Similarity	Proposed Island Flow	Subject Island D
NC_010334:2131939*	Shewanella halifaxensis HAW-EB4, complete genome	81.2806 %	Subject ← Query	31.3267
NC_010506:4873487	Shewanella woodyi ATCC 51908, complete genome	77.0129 %	Subject ← Query	31.5108
NC_010506:1893000	Shewanella woodyi ATCC 51908, complete genome	77.6195 %	Subject ← Query	31.5236
NC_009052:1899954	Shewanella baltica OS155, complete genome	79.0472 %	Subject ← Query	31.5886
NC_010334:1701957	Shewanella halifaxensis HAW-EB4, complete genome	82.0006 %	Subject ← Query	31.739
NC_010506:1835910	Shewanella woodyi ATCC 51908, complete genome	82.0956 %	Subject ← Query	31.8124
NC_014012:1676983	Shewanella violacea DSS12, complete genome	86.9884 %	Subject ← Query	31.899
NC_009997:685726	Shewanella baltica OS195, complete genome	79.2249 %	Subject ← Query	31.9374
NC_008322:3585601*	Shewanella sp. MR-7, complete genome	75.0429 %	Subject ← Query	32.0176
NC_014012:3812754*	Shewanella violacea DSS12, complete genome	81.299 %	Subject ← Query	32.0464
NC_011663:180889	Shewanella baltica OS223 chromosome, complete genome	77.4479 %	Subject ← Query	32.168
NC_008577:1579950	Shewanella sp. ANA-3 chromosome 1, complete sequence	80.1164 %	Subject ← Query	32.2252
NC_010506:5195000	Shewanella woodyi ATCC 51908, complete genome	81.2163 %	Subject ← Query	32.4018

Figure 86: *Aliivibrio salmonicida* LFI1238 island displayed various possible donor relationships to *Shewanella* species.

6.1.6 Remarks

Pre_GI was able to concur on various reported examples of HT found in literature. This reveals the potential of Pre_GI as a verification tool in HT and MGE hypothesis research. Future studies may find validation of proposed transfer events between micro-organisms by comparison with this extensive database.

6.2 Identification and analysis of islands in newly sequenced genomes including simple eukaryotes

Newly sequenced archaeal/bacterial genomes provide researchers with mountains of information to be analyzed and the combination of SWGIS and Pre_GI aim to aid novel research in these genomes. Various islands in novel genomes, including a simple eukaryote, were identified and analyzed by means of the SWGIS/Pre_GI pipeline.

6.2.1 *Brucella canis*

Brucella canis strain SVA13 was found to be the causative agent for an outbreak of canine brucellosis in Sweden during August of 2013. Sweden is officially free of brucellosis with incidence and outbreaks acquired abroad. The 2013 outbreak was facilitated by a canine imported from Spain for breeding purposes. The whole genome of the causative agent was sequenced, assembled, analyzed and is available from the NCBI under accession numbers CP007629 for chromosome 1 and CP007630 for chromosome 2 [84]. SWGIS was used with default parameters to identify islands in *Brucella canis* SVA13 chromosome 1 and 2. This resulted in 6 candidate islands for CP007629 (Figure 87) and 1 candidate island for CP007630. These candidates were curated for falsely selected *rrn* operons concluding with 4 islands in CP007629 and none in CP007630.

Manual inspection of the SWGIS-produced GenBank files for each island indicated that all 4 islands of CP007629 contain island confirmation keywords. Islands of *Brucella canis* strain SVA13 identified by SWGIS was compared to the IslandViewer database for overlaps. All 4 SWGIS islands displayed overlaps to islands predicted by at least one method contained in the IslandViewer package. *Brucella canis* was not available in PAIDB for overlap detection.

All *Brucella canis* islands were compared with regards to compositional similarity to Pre_GI. This is enabled by the upload and compare function available for up to 8 files simultaneously. Special interest was placed on the probable movement of islands located in *Brucella canis*. All but island#4 indicated proposed donor and recipient movement to various diverse organisms. The majority of the movement was between *Brucella canis* and *Mesorhizobium*, *Bradyrhizobium*, and *Agrobacterium* with the latter organisms

identified as donors. *Brucella* species are mainly associated with animal and human diseases whilst the organisms from which transfer was predicted are all soil or plant root bacterium. This may be explained by the close interaction between animals and soil, plants. *Brucella* belongs to the order Rhizobiales which is the same as the proposed donors mentioned above. Recipient movement was furthermore detected from *Brucella canis* to *Agrobacterium tumefaciens*. This may indicate a circular movement of genetic material from animals to soil within the order Rhizobiales between organisms only related on the level of order. Movement from *Brucella* was also detected to *Pseudomonas* and *Salmonella*. These organisms are only related at the level of phylum with the multidrug resistant *Salmonella* recipient including mercury resistance genes in its repertoire. These comparisons provide a fresh outlook on the movement of genetic material between divergent species and hosts.

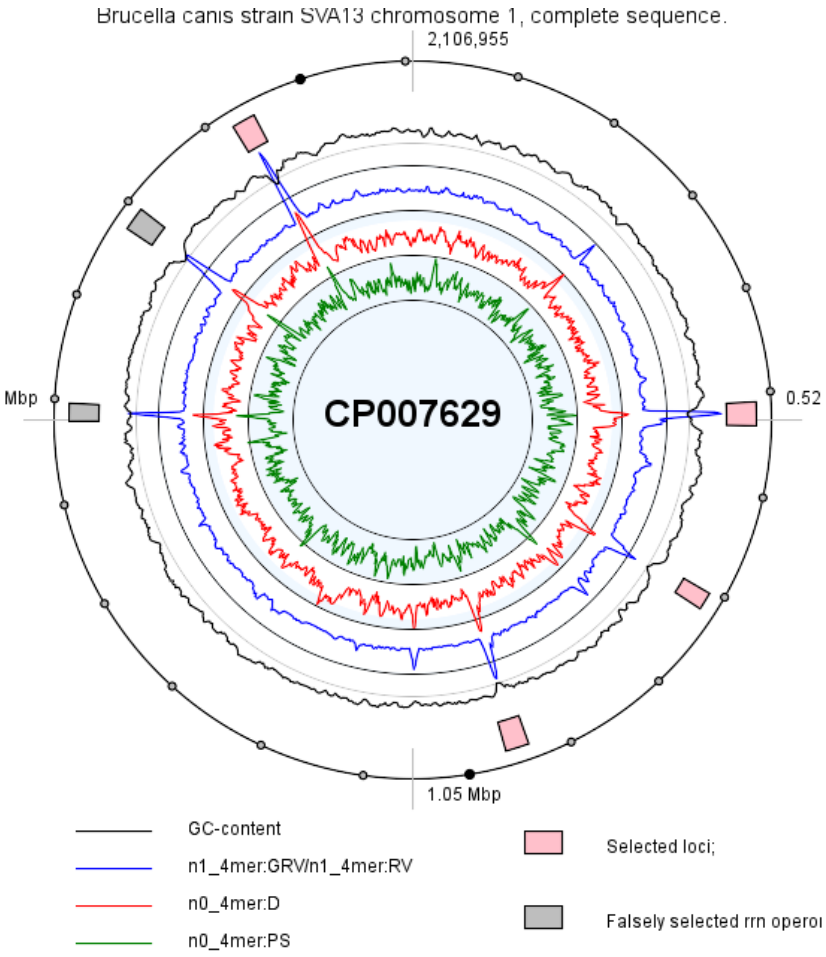


Figure 87: Graphical representation of 6 candidate islands predicted in *Brucella canis* strain SVA13, chromosome 1.

6.2.2 *Bacillus* sp. BH072

Bacterial strain BH072 was isolated from honey and identified as a new subspecies of *Bacillus* sp. displaying antifungal activity against mold [105]. The complete genome is available under accession CP009938 and available from the NCBI. SWGIS indicated 19 regions of horizontal transfer with 1 location overlapping with a region predicted by IslandViewer. This area was compared against Pre_GI through sequence similarity and the highest scoring hit found against an island hosted by *Bacillus amyloliquefaciens* subsp. *plantarum* UCMB5036 and visualized in Figure 88 below.

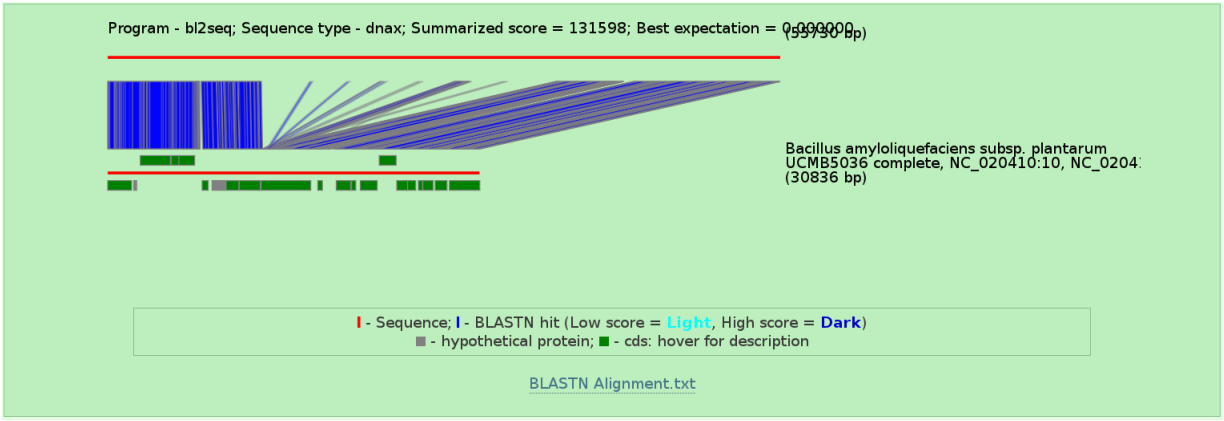


Figure 88: Highest scoring BLASTN hit for *Bacillus* sp. BH072 against Pre_GI. The red line at the top of the diagram indicates the query sequences of island 19 predicted in the genome of *Bacillus* sp. BH072. The red line at the bottom of the diagram is of the highest scoring hit which is located in *Bacillus amyloliquefaciens* subsp. *plantarum* UCMB5036.

Bacillus amyloliquefaciens produce antifungal and antibacterial substances with the island displayed as subject in the graph above containing protease and lyase proteins in addition to a predicted saccharopine dehydrogenase all of which are involved in antibiotic or antifungal properties.

6.2.3 *Staphylococcus aureus*

The type strain of *Staphylococcus aureus* subsp. *aureus* Rosenbach 1884 (DSM 20231^T) was first isolated in 1884 from human pleural fluid in Germany yet the genome of the type strain was only recently sequenced [19]. This invites various opportunities with regards to comparison with recently evolved strains and organisms. Shiroma et al. found MGE to comprise 15% to 20% of the strain chromosome. This strain is available in IslandViewer but does not contain any island information data. The complete genome sequences of *Staphylococcus aureus* subsp. *aureus* Rosenbach 1884 (DSM 20231^T) chromosome and

plasmid are available from GenBank under accession CP011526 and CP011527 respectively. SWGIS predicted no islands in the plasmid and 10 islands in the chromosome. The predicted islands are displayed in Figure 89 below.

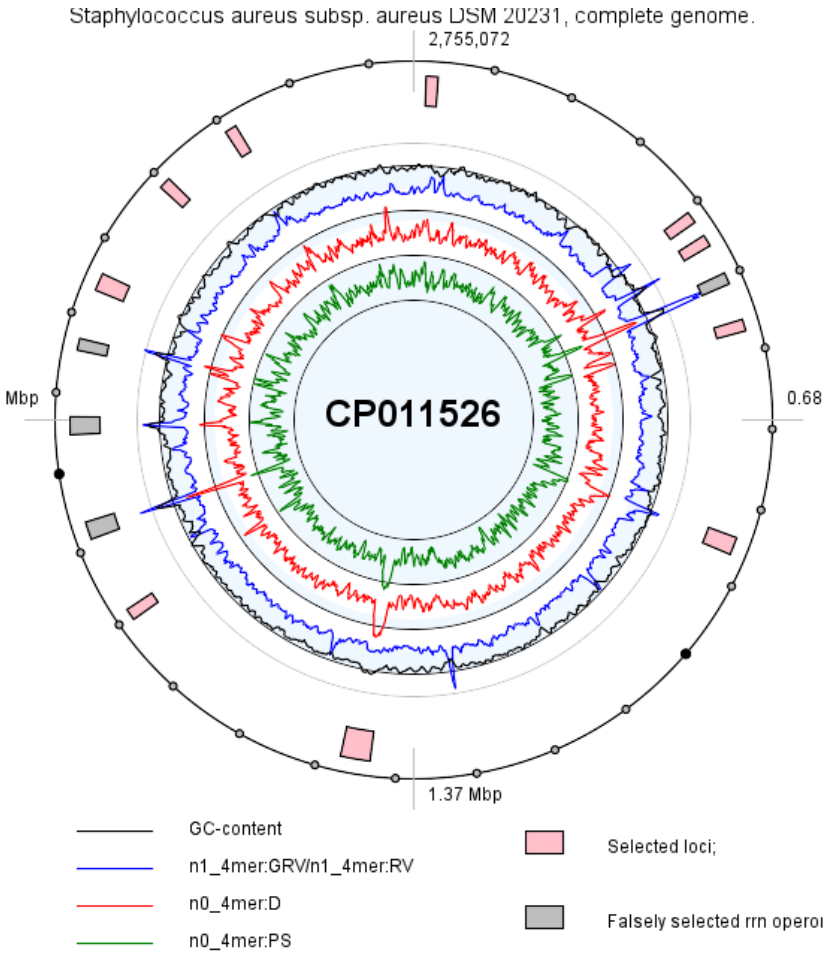


Figure 89: Islands found in *Staphylococcus aureus* subsp. *aureus* Rosenbach 1884 (DSM 20231^T) chromosome.

These islands were inspected for key word information and probable resistance proteins.

Query island#2 displayed multiple donor characteristics to various *Bacillus* islands. This is quite realistic as both these genera occupy the same habitat and an air of competition surrounds them and transfer is probable. The proposed movement from *Buchnera aphidicola* str. APS (*Acyrthosiphon pisum*) island#4 to query island#2 was of interest. This island in *Buchnera aphidicola* includes multidrug resistance proteins. *Buchnera aphidicola* is an aphid symbiont and found in the pea aphid.

Staphylococcus aureus subsp. *aureus* Rosenbach 1884 (DSM 20231^T) query island#3 displayed multiple high scoring sequence similarity hits to various organisms. A high scoring hit is presented in Figure 90 found against island#5 hosted by *Bacillus thuringiensis*

BMB171. *Bacillus thuringiensis,* famous for the BT toxin, is a pathogen of various insects.

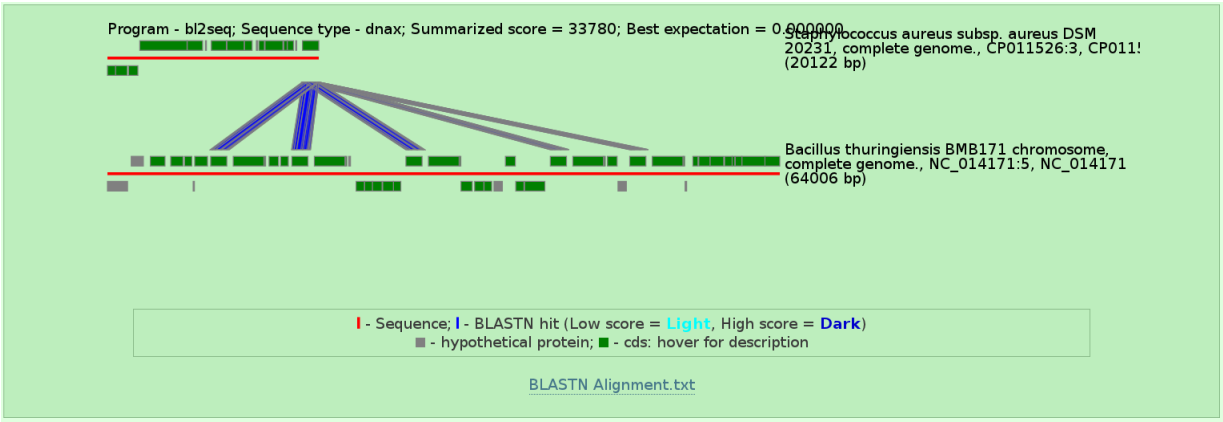


Figure 90: High scoring sequence similarity between islands in *Staphylococcus aureus* subsp. *aureus* Rosenbach 1884 (DSM 20231^T) and *Bacillus thuringiensis* BMB171.

Multiple fluxes were detected for this island with *Buchnera aphidicola* str. APS a persistent donor. Movement from *Staphylococcus aureus* island#3 to *Pasteurella multocida* 36950 island#1, one of the first pathogens to be studied in 1880s in France, was proposed with a compositional similarity of above 80% between query and subject.

Staphylococcus aureus island#4 similarly displayed high levels of compositional comparison and movement obtained from *Buchnera aphidicola* str. APS as well as *Rickettsia* and *Borrelia* both of which are transferred by ticks. Donor movement to *Pasteurella multocida* 36950 island#1 was again confirmed.

Staphylococcus aureus subsp. *aureus* MSSA476 is a hyper-virulent methicillin resistant strain isolated from the United Kingdom and displayed high sequence similarity to query island#6. Both these islands contain amino acid permease, thymidylate synthase, threonine dehydratase and alanine dehydrogenase. Query island#6 also accommodates quinolone resistance protein and a virulence factor.

Query island#7 contains a multidrug MFS transporter and displayed a high sequence similarity to *Staphylococcus haemolyticus* JCSC1435 island#6 and is displayed in Figure 91. *Staphylococcus haemolyticus* was isolated from a patient in Japan, 2000, and is highly resistant to various antibiotics. It has been proposed that *Staphylococcus haemolyticus* serves as a reservoir for resistance genes available to other virulent *Staphlococci.*

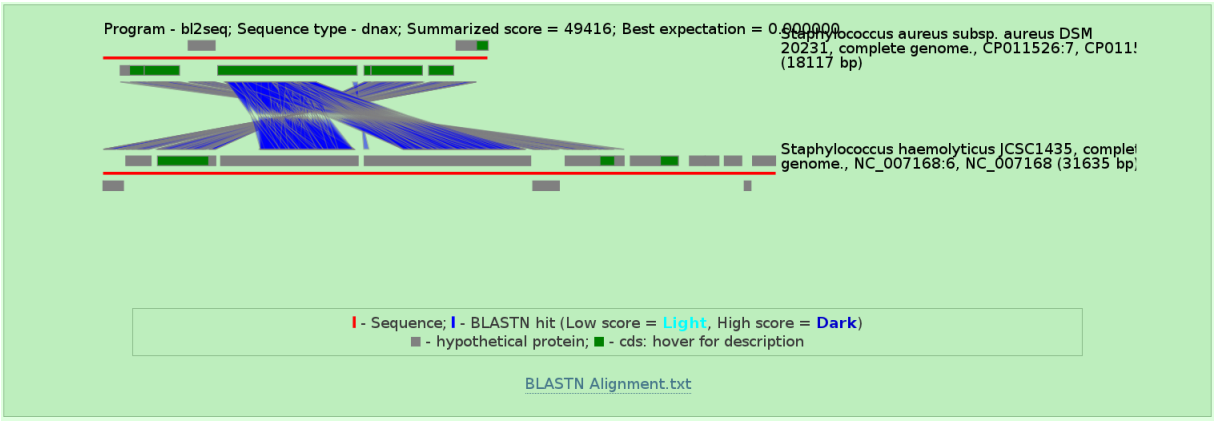


Figure 91: High sequence similarity between *Staphylococcus aureus* subsp. *aureus* Rosenbach 1884 (DSM 20231^T) query island#7 and *Staphylococcus haemolyticus* JCSC1435 island#6.

Proposed movement from query island#7 to *Vibrio cholerae* O1 biovar eltor str. N16961 island#2 is detected. This *Vibrio cholerae,* deemed an epidemic serogroup, was isolated in 1971 in Bangladesh and is distinguished from the classical biotype due to hemolysin production.

Query island#9 contains various proteins related to nitrates, zinc, host death and a methicillin resistance protein. A high scoring sequence similarity for this island was found against methicillin and vancomycin resistant *Staphylococcus aureus* subsp. *aureus* Mu50 island#14. This strain was isolated in 1997 from a Japanese patient whom did not respond to vancomycin. Island#14 contains various genes related to nitrates and zinc. Query island#9 furthermore displayed high sequence similarity to *Staphylococcus aureus* subsp. *aureus* USA300 island#13 which contains various nitrate and zinc proteins and a transcriptional regulator MerR family for mercuric resistance. Query island#9 displayed high compositional similarity to multiple *Staphylococcus aureaus* strains. OUP similarity of above 90% was established to *Staphylococcus aureus* subsp. *aureus* MRSA252 island#15 which contains a transcriptional regulator MerR family for mercuric resistance with movement proposed to be from the subject to the query.

Staphylococcus aureus island#10 contains multiple fibronectin-binding proteins which are associated with osteomyelitis. Osteomyelitis is a infection of the bone or bone marrow. High scoring sequence similarity between the query island and *Staphylococcus aureus subsp. aureus* ED98 was exposed. *Staphylococcus aureus subsp. aureus* ED98 was isolated in Northern Ireland and is associated with osteomyelitis particularly in chronically ill and immunocompromised individuals. Compositional similarity was also detected to *Staphylococcus lugdunensis* N920143 island#4 which is known to cause osteomyelitis.

6.2.4 *Galdieria sulphuraria*

The extremophilic red alga *Galdieria sulphuraria* was subjected to the SWGIS program to identify probable areas of vertical transfer. *Galdieria sulphuraria* evolution has been influenced by numerous horizontal transfer events from bacteria and archaea to produce a versatile, extremophilic red algae with a genome that includes various contributions from a pan-domain gene pool [44]. Schonknecht et al. found that the adaptation of *Galdieria sulphuraria* to heat may be expedited by the horizontal acquisition and development of archaeal ATPase gene families with high salinity tolerance afforded by HT from halophilic cyanobacteria. All acetate permeases seem to be derived from bacterial origin, certain amino acid-polyamine-organocation transporters from thermo-acidophilic archaea as well as acid phosphatases and b-galactosidases from bacteria [44].

The genome of *Galdieria sulphuraria* has been assembled to the level of scaffolds. This allows for the prediction of islands by means of SWGIS given that the scaffolds are of appropriate length. The joining of scaffolds into a single sequence for island identification circumvents the minimum length requirement and was applied. The usage of this format does hamper island comparison as no GenBank island file(s) are produced that would include added information, yet illustrates the opportunity to identify probable HT in newly sequenced simple eukaryotes with SWGIS. Island identification resulted in 23 candidate islands predicted of which all 23 were identified as islands and visualized in Figure 92.

All islands identified in *Galdieria sulphuraria* were subjected to BLASTX with e-value cut-off 10^{-6} against all genes contained within the Pre_GI database to identify genes of probable HT from archaea/bacteria or displaying homology to archaeal/bacterial proteins. Nineteen of the 23 islands contained genes that displayed high scoring sequence similarity to the database. Heat shock proteins were identified in numerous islands with sequence similarity to Pre_GI proteins from various genera such as *Acidithiobacillus*, *Acetobacter*, *Bacillus*, *Methanosarcina* to name a few. Numerous proteins affiliated with metals were identified mainly from diverse prokaryotic genera. In general 8,260 high scoring hits were found against proteins contained within archaeal/bacterial islands housed by Pre_GI.

The set of *Galdieria sulphuraria* islands were compared with all islands available in the database to detect sequence similarity. BLASTP with e-value cut-off 10^{-6} was utilized and identified 4 islands with high sequence similarity along the entire length of an island. Hits included genera *Geobacillus*, *Thermobacillus*, and *Methylacidiphilum*. *Galdieria sulphuraria* island#10 displayed high sequence similarity notable to the genus *Synechococcus* who are members of the *Cyanobacteria*. 2 *Synechococcus* strains were identified with one isolated from coastal water (*Synechococcus* sp. CC9902) and the other also isolated from

a marine environment highly tolerant of light intensity, up to intensities of 2 times the sun (*Synechococcus* sp. PCC 7002). Island#13 had a high sequence affinity with *Clostridium* species. *Clostridium* produce heat-resistant spores and is able to grow at high pH levels. Island#14 was shown to share sequence similarity with numerous genera including *Phytoplasma*, *Blattabacterium* and *Caldanaerobacter*. *Thermoanaerobacter tengcongensis* MB4 is a member of *Caldanaerobacter* and was isolated from a hot spring in China and showed high sequence similarity to island#14. *Clostridium* detailed the highest hits to island#16.

Compositional similarity between islands of *Galdieria sulphuraria* and islands contained in Pre_GI was calculated by means of OUP. The highest hits amongst all comparisons were commonly to *Methanosarcina*, *Bacillus*, *Borrelia*, *Methanocaldococcus* and *Sulfolobus*. OUP similarity of above 85% was calculated between *Sulfolobus tokodaii* str. 7 and *Galdieria sulphuraria* island#10. *Sulfolobus tokodaii* is described as an hyperthermophilic, acidophilic sulfur-metabolizing archaeon and was isolated from geothermal hot spring in Japan with an optimal growth at 80°C.

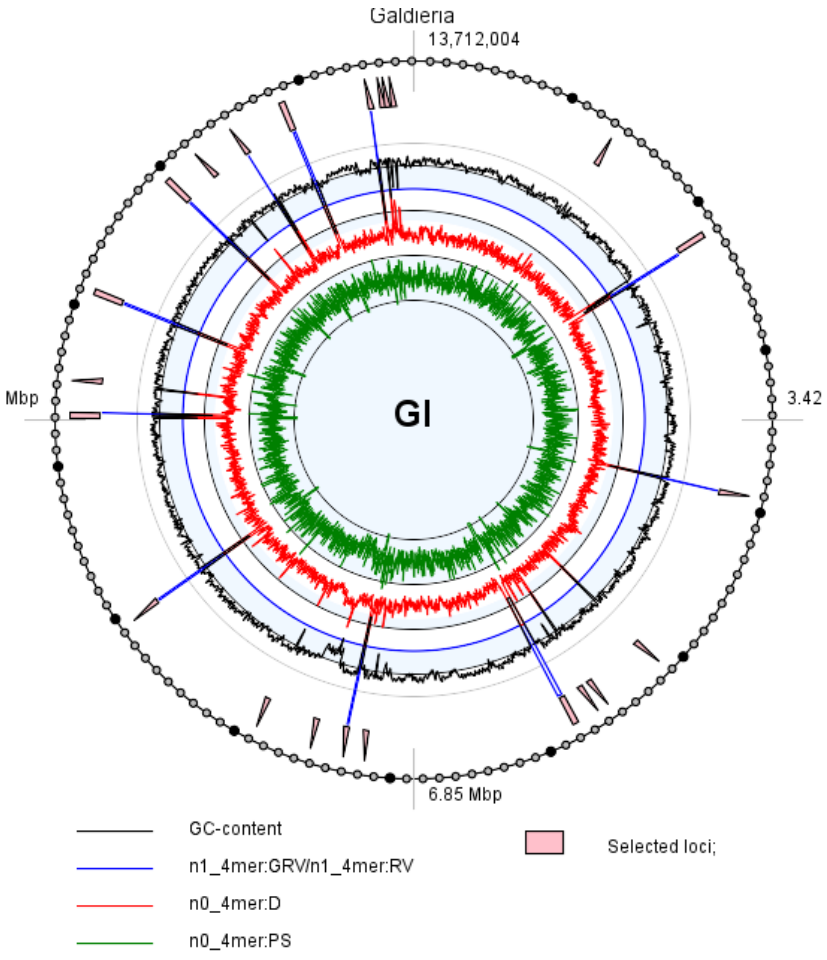


Figure 92: 23 Islands predicted in the eukaryotic extremophilic red alga *Galdieria sulphuraria* by SWGIS.

6.2.5 Remarks

The combination of SWGIS and Pre_GI provides users with the opportunity to identify and analyze islands in newly sequenced genomes. The speed and affordability of sequencing enables researchers to assemble diverse and economically important genomes. The detailed analysis and comparison of these newly sequenced genomes is vital. The identification of islands by SWGIS and the analysis of identified islands by Pre_GI is a novel tool in the NGS age.

6.3 Exchange in extreme communities - too hot to handle

The influence of HT on archaeal/bacterial communities and not just individuals may provide insight to their origin and current state. The vast amount of information from various taxa included in Pre_GI allows for analysis on communities as a whole. The inclusion of host lineage in the database may be vital in studies regarding these communities. The application of Pre_GI with regard to community HT studies is displayed below with the genus *Geobacillus* and *Bacillus* as examples.

Geobacillus is a motley genus within the family *Bacillaceae.* These organisms were historically placed within the large and diverse genus *Bacillus.* Small-subunit rRNA analysis revealed a highly diversified genus with ultimately 5 phylogenetically distinct groupings [22], with the majority of thermophilic aerobic spore-forming species binned in *Bacillus* groups 1 and 5 [101]. Nazina et al. proposed the novel genus *Geobacillus* to encompass organisms previously placed in group 5 of the genus *Bacillus* and be defined as earth or soil bacillus. Prior to the proclamation of the genus *Geobacillus* a large portion of these *Geobacillus* strains were narrated as belonging to *Bacillus stearothermophilus* even though there was disjunction with regards to physiology, temperature range and phenotypic characteristics of these strains [147]. Members of the genus *Geobacillus* have now been demonstrated to be omnipresent in a variety of habitats and extreme environments including cool [54] to cold areas [49]. Members of this kaleidoscopic genus have been isolated on all seven continents from 3,000 m above to 3,000 m below sea level [157]. This genus is of particular interest in biotechnology as a source of thermostable enzymes, natural products and the ability to ferment and digest a variety of substrates [147]. The relatively recent scientific construction of the genus *Geobacillus* and the ability to survive and thrive in various contrasting environments require further investigation into the content of *Geobacillus* HT and MGE as these events may explain ecological diversity and adaptation to these diverse locations. The identification and local/global comparison of foreign inserts in multiple *Geobacillus* strains (geo_islands) may elicit the origin and evolution of this species as influenced by environmental demand.

6.3.1 Strains

A pool of 29 geographically disjoint *Geobacillus* strains consisting of 12 complete and 17 draft genomes were used to obtain a clearer impression on the nature and influence of geo_islands on the genus *Geobacillus*. Sequence data is available from the NCBI and genomes used listed in Table 10.

Table 10: 29 Draft and complete genomes of *Geobacillus* used in geo_islands identification and comparison.

Strain	Accession	Status	Strain Information
G. thermoleovorans B23	BATY00000000	Draft	Oil-water. Japan. Thermophile
G. thermoleovorans CCB_US3_UF5	NC_016593.1	Complete	Hot spring. Malaysia. Thermophile
G. kaustophilus HTA426	BA000043.1	Complete	Deep sea sediment. Mariana Trench
Geobacillus sp. CAMR5420	JHUS01000000	Draft	Hemicellulolytic. Thermophile
G. kaustophilus Gblys	BASG00000000	Draft	Lysogenic. Phage infection
Geobacillus sp. MAS1	AYSF00000000	Draft	Hot spring. Pakistan
Geobacillus sp. A8	AUXP01000000	Draft	Deep mine water. South Africa. Thermophile
Geobacillus sp. CAMR12739	JHUR01000000	Draft	Iceland. Hemicellulolytic
Geobacillus sp. C56-T3	CP002050.1	Complete	Mesophile. Chemo-organotroph
Geobacillus sp. Y412MC61	NC_013411.1	Complete	Hot Spring. Yellowstone National Park
Geobacillus sp. Y412MC52	NC_014915.1	Complete	Hot Spring. Yellowstone National Park
Geobacillus sp. WSUCF1	ATCO00000000	Draft	Soil. Compost. USA. Thermophile
Geobacillus sp. GHH01	NC_020210.1	Complete	Soil. Germany. Thermophile
Geobacillus sp. JF8	NC_022080.4	Complete	Compost. Japan. PCBs degrader.
Geobacillus sp. G11MC16	ABVH00000000	Draft	Thermophile
G. thermodenitrificans NG80-2	NC_009328.1	Complete	Oil-water. China. Thermophile
G. caldoxylolyticus CIC9	AMRO00000000	Draft	Hot spring. Indonesia. Thermophile
G. thermoglucosidasius C56YS93	NC_015660.1	Complete	Hot Spring. Yellowstone National Park
G. thermoglucosidasius TNO-09.020	NZ_CM001483	Complete	Thermophile. Dairy
Geobacillus sp. Y4.1MC1	NC_014650.1	Complete	Hot Spring. Yellowstone National Park
Geobacillus sp. WCH70	NC_012793.1	Complete	Thermophile. Chemo-organotroph
G. caldoxylosilyticus NBRC 107762	BAWO01000000.1	Draft	Soil. Australia. Thermophile
Geobacillus sp. FW23	JGCJ01000000.1	Draft	Oil-water. India. Thermophile
G. stearothermophilus NUB3621	AOTZ01000000.1	Draft	Thermophile
G. thermoglucosidasius NBRC 107763	BAWP00000000	Draft	Japan
G. thermocatenulatus GS-1	JFHZ01000000.1	Draft	Oil. China. Thermophile
G. stearothermophilus 22	JQCS00000000	Draft	Hot spring. Russsia
G. stearothermophilus 53	JPYV00000000	Draft	Hot spring. Russia
Geobacillus sp. 12	Not Available	Draft	Not Available

6.3.2 Protein clustering and distribution

The set of 99,821 proteins present in the 29 genomes of *Geobacillus* were subjected to an OrthoMCL algorithm with an inflation parameter of 1.5, e-value threshold of 10^{-6} and

75% pair-wise alignment to produce orthologous protein clusters. These clusters were inspected to identify a core, softcore, shell and cloud set of proteins to exemplify the total *Geobacillus* genome. The core embodies the collection of conserved genes with a 100% attendance in all 29 studied *Geobacillus* genomes and contained 527 elements. The softcore is a collection of proteins available in 95% of the genomes under investigation and resulted in 1,862 proteins. The softcore allows for missing genes/proteins as many of the genomes are still in a draft phase. The combination of core and softcore epitomizes a collection of conserved genes in the strains under scrutiny and provides information on the evolutionary history of a lineage [158]. The shell would relate to proteins present in 10% to 90% of subjected genomes (3,515 genes) and the cloud to genes available in less than 10% of the 29 *Geobacillus* strains under investigation (8,218 proteins). The shell and cloud constitute the pliable genome and mirrors the evolutionary history of a sublineage in addition to the various adaptations enforced by an alternative lifestyle and environment [130]. The rate of loss and gain between the conserved (core and softcore) and flexible (shell and cloud) gene pools are estimated to differ. The conserved grouping would expect a slower rate of gene addition or subtraction and the flexible grouping a highly increased rate with HT more pronounced in the flexible clusters [9].

6.3.3 Clusters content and analysis

The core, softcore, shell and cloud were individually compared against the entire collection of proteins available in the Pre_GI database. This entails 656,806 proteins from diverse archaeal/bacterial sources and comparison was performed by means of BLASTP and e-value threshold of 10^{-6}. Results are available in Table 11 below. The bulk of proteins displaying no similarity to the *Geobacillus* elements, even at a less stringent threshold, were described as "hypothetical" and excluded from further analysis.

Table 11: Sequence similarity of cluster elements against the Pre_GI database and includes counts on zero similarity identified at less stringent BLASTP comparisons.

	Elements	Hits to Pre_GI proteins	Distinct Genera	No significant similarity
Core	527	492	46	8
Softcore	1,862	1,747	123	19
Shell	3,515	2,902	245	280
Cloud	8,218	4,753	303	1,855

Subject hits for each clustering were investigated with regards to the subject host genus and only the highest scoring hit for each protein in a cluster used to avoid over representation. Table 12 - 15 displays the top 10 genera with the highest number of individual hits to the *Geobacillus* clusters.

Table 12: Top 10 highest scoring genera against the *Geobacillus* protein core cluster.

Genus	Number of individual hits to core	Percentage of individual hits to core
Bacillus	261	53.05%
Geobacillus	100	20.33%
Anoxybacillus	26	5.29%
Clostridium	13	2.64%
Oceanobacillus	11	2.24%
Paenibacillus	9	1.83%
Staphylococcus	8	1.63%
Lactobacillus, Thermosediminibacter	5	1.02%
Streptococcus, Thermobacillus	4	0.81%
Enterococcus, Halobacteroides, Thermoanaerobacter	3	0.61%

Table 13: Top 10 highest scoring genera against the *Geobacillus* protein softcore cluster.

Genus	Number of individual hits to softcore	Percentage of individual hits to softcore
Bacillus	844	48.31%
Geobacillus	378	21.64%
Anoxybacillus	92	5.27%
Clostridium	46	2.63%
Oceanobacillus	45	2.58%
Paenibacillus	40	2.29%
Staphylococcus	28	1.60%
Thermobacillus	16	0.92%
Lactobacillus	14	0.80%
Desulfotomaculum, Thermoanaerobacter	13	0.74%

Table 14: Top 10 highest scoring genera against the *Geobacillus* protein shell cluster.

Genus	Number of individual hits to shell	Percentage of individual hits to shell
Geobacillus	1,339	46.14%
Bacillus	567	19.54%
Paenibacillus	107	3.69%
Clostridium	90	3.10%
Oceanobacillus	39	1.34%
Desulfotomaculum	33	1.14%
Staphylococcus	27	0.93%
Alicyclobacillus	24	0.83%
Thermaerobacter	21	0.72%
Anoxybacillus	20	0.69%

Table 15: Top 10 highest scoring genera against the *Geobacillus* protein cloud cluster.

Genus	Number of individual hits to cloud	Percentage of individual hits to cloud
Geobacillus	2,234	47.00%
Bacillus	945	19.88%
Paenibacillus	163	3.43%
Clostridium	144	3.03%
Anoxybacillus	66	1.39%
Desulfotomaculum	56	1.18%
Oceanobacillus	45	0.95%
Alicyclobacillus, Thermobacillus	29	0.61%
Brevibacillus	27	0.57%
Pseudomonas	26	0.55%

Figure 93 and 94 graphically represents the subject hit genus for each cluster as found against Pre_GI.

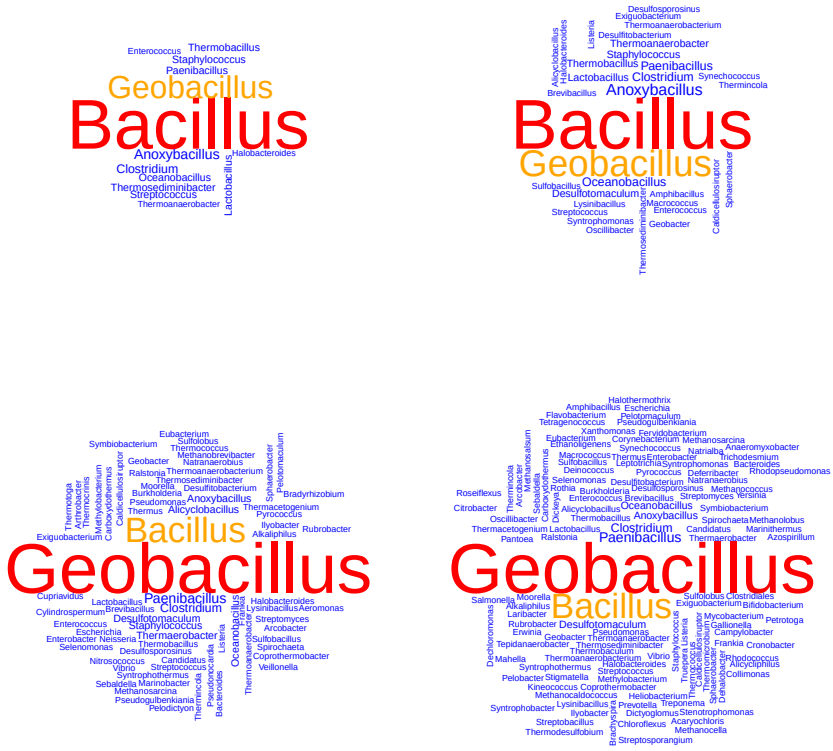


Figure 93: Word cloud representation of subject genera determined by BLASTP to each *Geobacillus* protein cluster. Core in the top left and softcore in the top right. Shell in the bottom left with cloud to the right. High frequency genera are displayed in red and a large font with intermediate frequency in yellow and an intermediate font. Lower frequency genera are represented by blue and a small font.

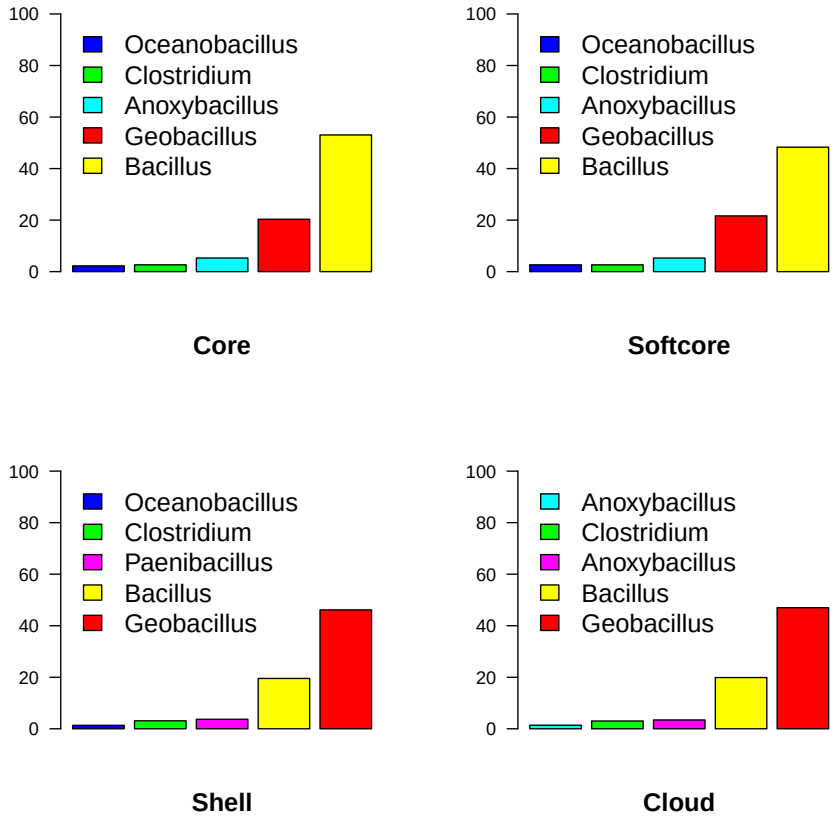


Figure 94: Bar charts of the top 5 genera in each cluster.

The tables presented above in combination with Figure 93 and Figure 94 indicate a probable movement from *Bacillus* in the core to *Geobacillus* in the cloud. The core and softcore is controlled by *Bacillus* whilst the shell and cloud is largely impacted by *Geobacillus*. The volume of distinct genera contributing to each cluster increases when moving from the conserved to the flexible genome.

The exposure to a variety of divergent archaea/bacteria may have altered these organisms to invade novel yet extreme environments and as such molded and branched the genus G*eobacillus* to produce a detached group of organisms. The results presented above may indicate a genesis of *Geobacillus* from *Bacillus* owing to various factors inclusive of changing environment and HT. The core or conserved segment of the *Geobacillus* genome is represented by *Bacillus* with the flexible contribution to the genome consisting of *Geobacillus*.

Pre_GI host information was incorporated to display a general trend regarding the clusters. Subject hit host information, as available from the database, may aid research with regards to the influence of environment on HT. The occurrence of environmental phrases and words are presented in Figure 95 and Figure 96.

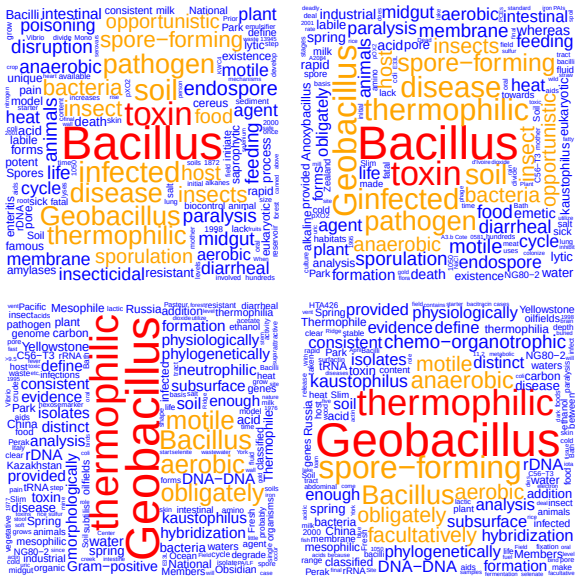


Figure 95: Word clouds on subject host information, habitat and general lifestyle for clusters. Core and softcore clusters are displayed in the top left and right respectively with shell and cloud in the bottom left and right. Red, large sized words indicate a high frequency with yellow, medium sized representing an intermediate frequency and blue, small sized relating to low frequency words.

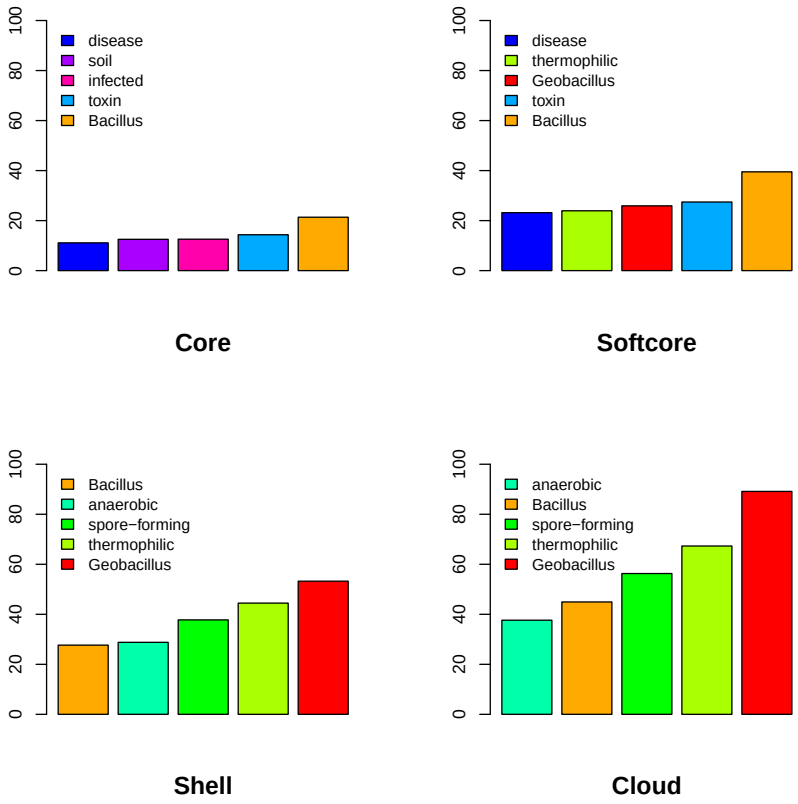


Figure 96: Bar charts of the top 5 words regarding host information in each cluster.

General extremophile characteristics seem more visible in the outer clusters (shell and cloud). There appears to be a gradual movement from a mesophilic lifestyle in the core to an extremophilic habitat in the cloud.

6.3.4 Islands in the soil/earth bacteria

All 29 draft and complete *Geobacillus* genomes were subjected to island identification by the SWGIS. Sequences were only assembled to the level of scaffolds and as such available exclusively in FASTA format. Regions of foreign insertion (geo_islands) were identified in the entire set of genomes and presented in Table 16.

Table 16: Number of geo_islands identified in the set of *Geobacillus* genomes.

Strain	Accession	Number of islands
G. thermoleovorans B23	BATY00000000	21
G. thermoleovorans CCB_US3_UF5	NC_016593.1	18
G. kaustophilus HTA426	BA000043.1	19
Geobaillus sp. CAMR5420	JHUS01000000	18
G. kaustophilus Gblys	BASG00000000	23
Geobacillus sp. MAS1	AYSF00000000	24
Geobacillus sp. A8	AUXP01000000	20
Geobacillus sp. CAMR12739	JHUR01000000	23
Geobacillus sp. C56-T3	CP002050.1	19
Geobacillus sp. Y412MC61	NC_013411.1	19
Geobacillus sp. Y412MC52	NC_014915.1	21
Geobacillus sp. WSUCF1	ATCO00000000	27
Geobacillus sp. GHH01	NC_020210.1	12
Geobacillus sp. JF8	NC_022080.4	13
Geobacillus sp. G11MC16	ABVH00000000	22
G. thermodenitrificans NG80-2	NC_009328.1	17
G. caldoxylolyticus CIC9	AMRO00000000	26
G. thermoglucosidasius C56YS93	NC_015660.1	13
G. thermoglucosidasius TNO-09.020	NZ_CM001483	19
Geobacillus sp. Y4.1MC1	NC_014650.1	14
Geobacillus sp. WCH70	NC_012793.1	7
G. caldoxylosilyticus NBRC 107762	BAWO01000000.1	22
Geobacillus sp. FW23	JGCJ01000000.1	23
G. stearothermophilus NUB3621	AOTZ01000000.1	18
G. thermoglucosidasius NBRC 107763	BAWP00000000	22
G. thermocatenulatus GS-1	JFHZ01000000.1	21
G. stearothermophilus 22	JQCS00000000	26
G. stearothermophilus 53	JPYV00000000	21
Geobacillus sp. 12	Not Available	19

A total of 567 geo_islands were identified with and average of 19 insertions per strain. It

should be noted that the lower number of geo_islands predicted in certain strains may be due to the status of the genome, *i.e.* complete or draft. *Geobacillus* protein cluster genes were aligned to all geo_islands to identify elements of the core, softcore, shell and cloud which are not present in any geo_island. These results are tabulated below (Table 17).

Table 17: Frequency of elements in a protein cluster present in geo_islands.

	In a geo_island	Percentage
Core	357	67.74%
Softcore	1,341	72.02%
Shell	2,915	82.93%
Cloud	5,835	71.00%

The table above and Figures 97 - 100 below indicate that the majority of elements across all 4 protein groupings (core, softcore, shell and cloud) that are found in a geo_island are of *Geobacillus* origin and those that are not present in a geo_island are from *Bacillus* origin. This may illustrate the high level of influence HT has had on the genus *Geobacillus.* The high frequency of *Bacillus* elements not found in geo_islands points to the probable evolutionary origin of *Geobacillus.* It therefore seems plausible that *Geobacillus* originated from *Bacillus* and was altered by HT and foreign inserts that are today associated with *Geobacillus.* It is thus these geo_islands that likely defined and fashioned *Bacillus* organisms into members of the genus *Geobacillus.*

All 567 geo_islands were compared to islands housed by Pre_GI for indications of sequence and compositional similarity. OUP similarity concluded that only 1 geo_island (GI:23|JGCJ01000000.1) did not present compositional similarity to an island in Pre_GI with only 1.24% of these hits not to a *Geobacillus* island. BLASTN indicated all but 1 geo_island (GI:22|AMRO0000000) shared sequence similarity to an island in the database. Only 2.12% of the highest scoring sequence similarity hits were not to an island hosted by *Geobacillus.*

The genera composition of proteins in an geo_island and not contained in a geo_island was obtained from the previous result and graphically presented in Figure 97 and Figure 98 for the core and softcore and Figure 99 and 100 for the shell and cloud.

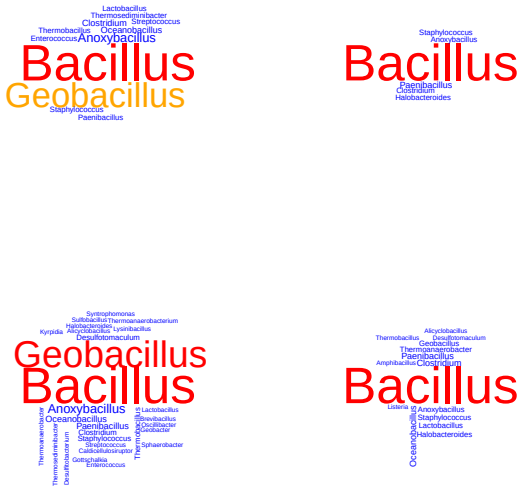


Figure 97: Genera composition of proteins from the core and softcore available in geo_islands and absent from geo_islands. The top row represents the core with elements in the set of geo_islands on the the left and elements absent to the right. The bottom row describes the inclusion in the softcore to the left and exclusion on the right. Large, red displayed genus indicate a high frequency with intermediate, yellow genus conveying a medium frequency and small, blue genus indicating a low frequency.

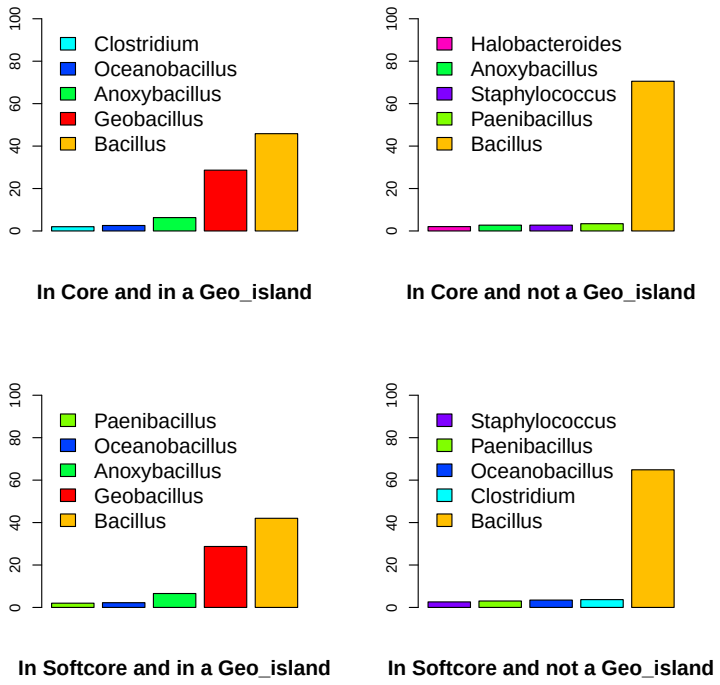


Figure 98: Bar chart of the top 5 genera in a geo_island and not in a geo_island for the core and softcore clusters.

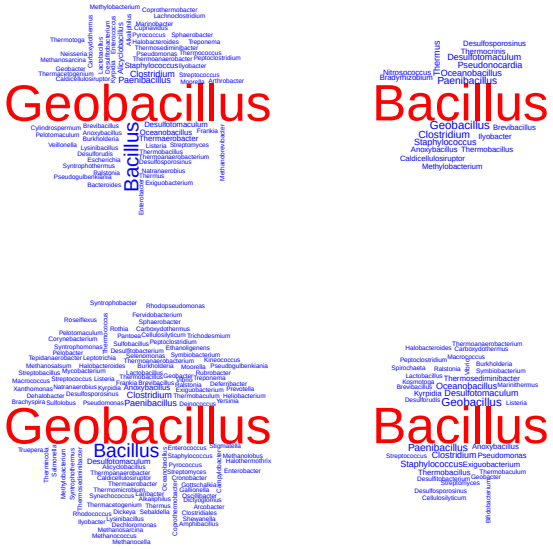


Figure 99: Genera composition of proteins from the shell and cloud available in geo_islands and absent from geo_islands. The top row represents the shell with elements in the set of geo_islands on the the left and elements absent to the right. The bottom row describes the inclusion in the cloud to the left and exclusion on the right. Red and large font genus indicates a high frequency with blue, smaller font genus indicating a lower frequency.

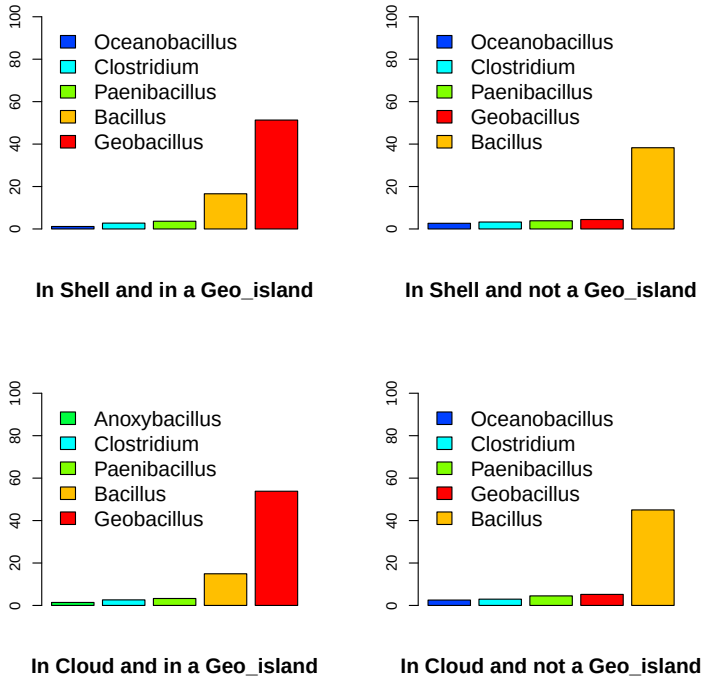


Figure 100: Bar charts of the top 5 genera in a geo_island and not in a geo_island for the shell and cloud clusters.

6.3.5 Remarks

Pre_GI may be used in the analysis of bacterial communities to identify the effect and frequency of HT within these groups. The inclusion of host lineage information in Pre_GI may ease and aid proposed future studies on the extent and influence of HT in different communities.

6.4 Reconstruction of large inserts through their fragments - divide and conquer

The insertion of a foreign region in a host genome may be followed by the fragmentation to produce multiple individual segments or islands. As presented earlier this method of divide and conquer has been proposed for numerous islands housed in the database. Certain islands may thus have been part of a much larger segment. If the original fragment can be reconstructed it is possible to determine the donor of this large HT event. This entails viewing a set of islands within a host as a possible genome which has been sliced an placed in various positions across a host genome. Islands currently housed in Pre_GI were tasked with this novel approach to determine viability a similar future studies.

6.4.1 Islands reassembly

Islands proposed as resulting from a single large insert were reassembled with SPAdes-3.6.0 [95]. All islands for an individual host were randomly fragmented into 10,000,000 250 bp reads in a mock shotgun approach. These reads were reassembled into scaffolds which were compared against the NCBI Genomes (chromosome) database with Megablast which is optimized for highly similar sequences.

6.4.2 Fragmented islands donors

NC_000909

Methanocaldococcus jannaschii DSM 2661, complete genome consists of 15 islands identified across the genome. These islands were reassembled in a 331,064 length sequence and compared against the NCBI with the resulting hits displayed in Figure 101. The highest hit was against *Methanocaldococcus* sp. FS406-22, complete genome [NC_013887] and obtained a max score of 17,165, total score of 3.359e+05 with 80% of the query covered, 89% identity and an e-value of 0.0.

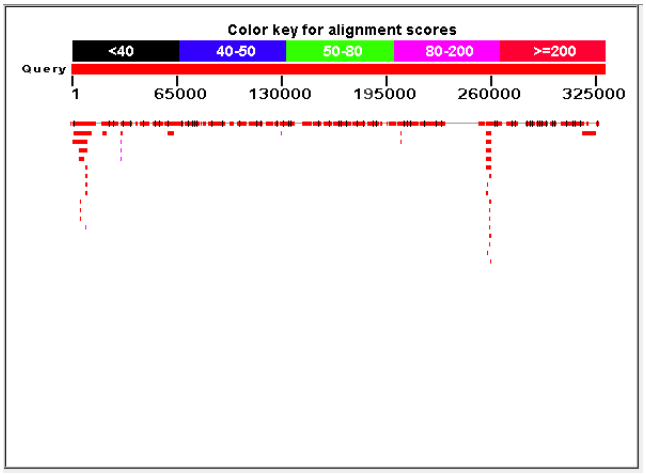


Figure 101: Megablast results for assembled *Methanocaldococcus jannaschii* DSM 2661, complete genome islands. The highest hit was found against *Methanocaldococcus* sp. FS406-22, complete genome [NC_013887].

NC_002662

Lactococcus lactis subsp. *lactis* Il1403, complete genome contained 4 islands with an reassembly length of 83,154. Megablast (Figure 102) indicated the highest hit to be to the query itself. This indicates that the reassembly methodology is functional. The next hit with regards to Max score was found to be against *Lactococcus* prophage bIL286, complete genome. All *Lactococcus lactis* subsp. *lactis* Il1403 islands contain keyword confirmation with a high prophage protein composition.

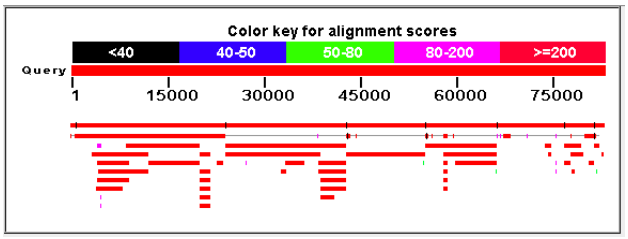


Figure 102: Reassembly of *Lactococcus lactis* subsp. *lactis* Il1403 islands compared to NCBI.

NC_002928

Bordetella parapertussis 12822, complete genome entails 5 islands which were assembled to a length of 161,568. The highest hit was found to be against itself yet the following 6 best hit were deemed to be against different strains of *Bordetella bronchiseptica,* the best of these being to *Bordetella bronchiseptica* 253, complete genome. In Figure 103 this hit is the second red line.

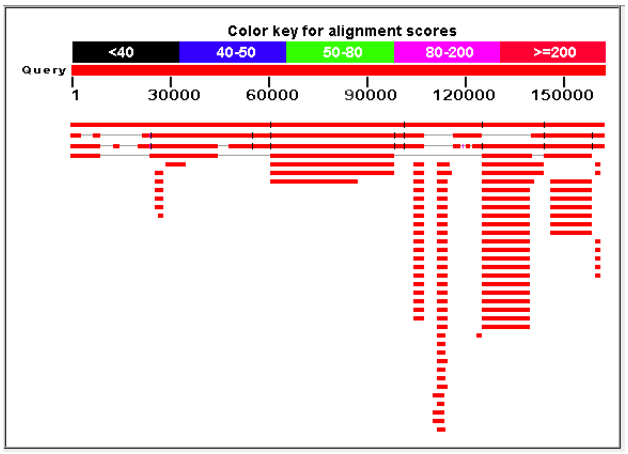


Figure 103: *Bordetella parapertussis* 12822, complete genome islands reassembly hits against NCBI.

NC_003063

Agrobacterium tumefaciens str. C58 chromosome linear, complete incorporates 6 islands and is a biovar 1 nopaline-producing strain isolated from a cherry tree tumor. The islands produced a sequence of length 154,781 which was compared to the NCBI and displayed in Figure 104.

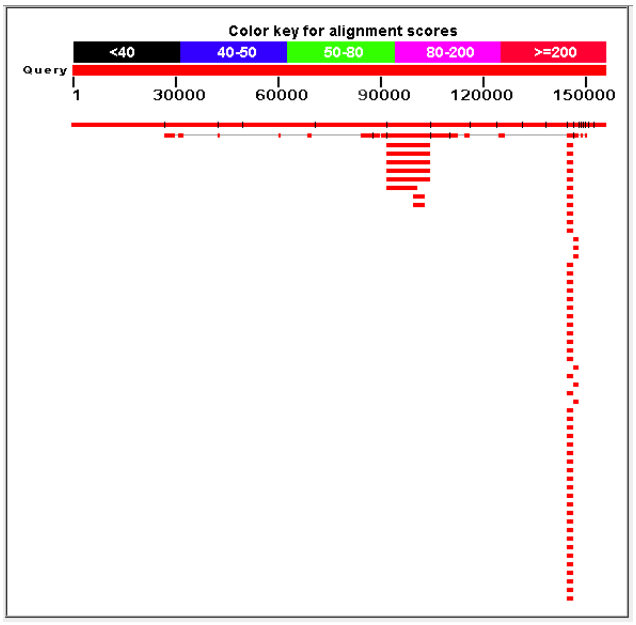


Figure 104: 6 *Agrobacterium tumefaciens* str. C58 chromosome linear, complete were assembled and compared to the NCBI by means of Megablast with sequence similarity hits displayed

Agrobacterium tumefaciens strain Ach5 chromosome linear, complete sequence displayed high sequence similarity to the query. This species is a phytopathogenic bacterium that

causes crown gall disease and was isolated from yarrow (*Achillea millefolium*). The highest scoring hit was against *Agrobacterium tumefaciens* str. C58 chromosome linear, complete and reinforces the reassembly. The second best hit represents *Agrobacterium tumefaciens* strain Ach5 chromosome linear, complete sequence.

NC_003078

Sinorhizobium meliloti 1021 plasmid pSymB, complete sequence contains 2 islands with a reassembled length of 77,877. This reassembly produced an alignment which covered 100% of the query with 99% to *Sinorhizobium meliloti* 2011 plasmid pSymB, complete sequence. The top hit was to the query sequence itself with the second hit to the 2011 plasmid differing only slightly with regards to the Max score. These alignments are displayed in Figure 105.

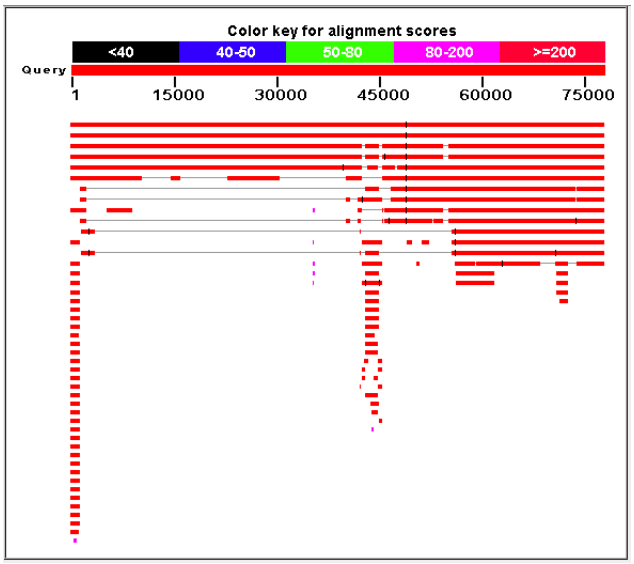


Figure 105: High scoring hits with islands of *Sinorhizobium meliloti* 1021 plasmid pSymB, complete sequence as the query.

NC_003212

Listeria innocua Clip11262, complete genome is a non-pathogenic soil organism containing 11 islands with a reassembled length of 274,012. The highest scoring reassembled hit was to itself. Further high scoring hits included different *Listeria innocua* and *Listeria monocytogenes* strains. 3 *Listeria innocua* Clip11262 islands contain a high frequency of bacteriophage proteins with the assembled query indicating high sequence similarity to *Listeria* phage B054, complete genome, a phage of *Listeria monocytogenes,* in Figure 106 (red line number 4 from the top).

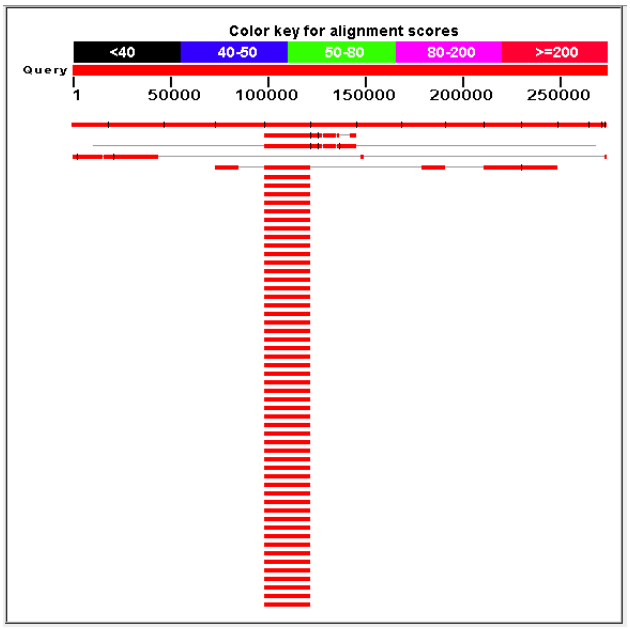


Figure 106: High scoring sequence similarity of assembled islands to various *Listeria innocua* and *Listeria monocytogenes* strains. These hits include similarity to *Listeria* phage B054, complete genome in row 4 of subject hits.

NC_003304

Agrobacterium tumefaciens str. C58 chromosome circular, complete contains 2 islands producing a reassembly of length 39,608. This sequence displayed high sequence similarity to various *Agrobacterium tumefaciens* and *Rhizobiaceae* strains. In particular a hit was found to *Rhizobium* sp. IRBG74 circular chromosome, complete genome which covered 97% of the query sequence with 86% identity.

6.4.3 Remarks

Pre_GI houses numerous islands resultant for a possible single large insertion followed by fragmentation. The availability of this information in a singular location allows for novel research and applications with regards to future island research.

6.5 Discussion

The examples shown aim to highlight the application of Pre_GI in various diverse research strategies that may be followed. The wealth of information residing in Pre_GI enables researchers the ability to follow various analytical pipelines and work flows. The wealth of information hoarded in Pre_GI may be approached from varying angles to address diversified research questions.

7 Chapter 7: Discussion

> The larger the island of knowledge, the longer the shoreline of wonder. Ralph W. Sockman

Horizontal transfer and mobile genetic elements is a fundamental process in evolution and adaptation influencing organisms as far back as the original formation plastids in the endosymbiotic theory. This theory has undergone amelioration and multiple changes in recent years yet the effect of the non-genealogical transfer between organisms in the symbiosis is generally accepted. Transfer between prokaryotes is achieved by various mechanisms with each relating to an environmental need and all requiring different forms of donor-recipient acceptance and willingness. The successful transfer of a foreign element does not imply the completion of the border crossing process as there are various hindrances to the complete integration of foreign genetic material in a new host organism. The ability to negate the physical and chemical complication regarding transfer and integration illustrates the profound need and importance of these events in archaeal/bacterial survival and adaptation. Foreign acquisitions have furthermore been identified to occur across taxonomic borders and involve organisms from diverse domains. The insertion of prokaryotic genetic information has been identified in various eukaryotes ranging from simple to complex eukaryotic organisms. The detection of transfer from bacteria to the human genome illuminates the omnipresence of these events.

The identification of horizontally acquired genetic information and events relating in a host organism are grounded in two approaches. Compositional and comparative methods incorporates different approaches and information in an attempt to isolate and identify these islands from the rest of the host genome continent. The SeqWord Genomic Island Sniffer (SWGIS) program is a composition-based island identifier that employs 4-mer oligonucleotides and various statistics regarding local and global patterns in a host to infer the presence of an island.

Current island collections have proved themselves to be valuable and informative congregations of island information. These databases provided users with a wealth of knowledge regarding island location, content and sequence with little to no information detailing the relationship between islands themselves. The need thus arose for the production of an encompassing island database to provide users the ability to investigate and analyze island information and the various relationships with which an island was proposed to be involved with. The incorporation of all available island information in conjunction with island relational information aims to be a vital cog in island and island event research.

The implementation of next-generation sequencing enables the rapid and economical determination of whole genome information. This drives the need for competent and brisk

sequence information analysis approaches and tools. Genome linguistics resorts to text analysis algorithms as a method of compositional sequence investigation and comparison. This approach is used by SWGIS to identify divergences from local to global patterns in genomic sequences as a means of island identification. Tetranucleotide frequencies and parameters in combination with a sliding window approach form the foundation of this comparative island predictor.

Island predictors and SWGIS are not immune to false positive or false negative predictions. Factorial analysis was used in the optimization of island prediction parameters to minimize false prediction rates and resulted in the formulation of regression equations for false negative and false positive rate estimation. This enables the determination of a single parameter set to ensure an optimal specificity/sensitivity ratio in SWGIS island prediction. These parameters may be varied in relation to user requirements as genome composition and content differs.

Post-identification analysis of islands is generally lacking and typically confined to gene content. SWGIS affords users the ability to perform various analysis and investigation on islands after prediction and annotation. The LingvoCom collection of utilities is a valuable island analysis and interpretation tool kit which forms part of the SeqWord project. These utilities include the grouping of islands by means of compositional similarity and the construction of phylogenetic trees for islands based on compositional comparison. The process of amelioration is used to quantify the period of time that a foreign region of DNA has spent in a host. This enables LingvoCom to determine donor-recipient relationships between organisms with regards to islands housed by them and the probable paths that were followed in the acquisition of novel sequence segments.

The ability to identify islands by SWGIS was compared to contemporary island identification tools to determine the reliability and potential of this island identifier. SWGIS was found to be a competent and inclusive island identifier outperforming other identification tools currently available. The prediction of islands is accompanied by the likelihood of false positive inclusions and false negative exclusions to the set of identified islands. False positive and negative rates with reference to island prediction is currently poorly stipulated and defined with no formal method available. Comparison of SWGIS identified islands to known islands enables the calculation of these false prediction and exclusion rates for SWGIS as an island identifier.

Current and continued identification and analysis of islands in archaeal/bacterial genomes requires a viable and expandable database. This repository was constructed with MySQL and the inclusion of various unique identifiers. The construction of the database was done in such a fashion to ensure speedy and reliable retrieval of information. The copious amount of information to be stored forced the construction of a inter-connected yet

transparent database structure. The allocation of novel island and island gene identifiers ensured redundancy and efficiency in record retrieval.

The incorporation of novel islands and accompanying information into the database ensures relevance and longevity. Newly sequenced genomes are subjected to the SWGIS for island identification. Resulting island files are fed to the expansion pipeline for incorporation. Computational and time constraints led to the development of a novel inclusion pipeline by means of island representatives. This approach enables the inclusion of novel islands and all relating information in an expandable and growing environment.

The ability to access and share information in a convenient and user-friendly fashion was established by the development of a graphical user interface (GUI).

Amalgamation of the MySQL database and the GUI culminated in the Predicted Genomic Island database (Pre_GI) which contains 26,744 islands identified by SWGIS in 2,407 archaeal/bacterial chromosomes and plasmids. This web-resource freely available from http://pregi.bi.up.ac.za and aims to be a novel alternative to currently used island databases.

Current island information and relational results are conveniently accessible for browsing. This includes the unique incorporation of island archaeal/bacterial host taxonomy and general information to aid in the formation of a holistic HT and environment image. All information regarding an island is presented to a user along with the annotated gene content for inspection. Island similarity results include sequence and compositional comparison against the totality of the current content with the added option of sequence similarity visualization. Various result filters are further available to users when browsing in order to obtain specific and precise information. General taxonomy statistics for Pre_GI content presents a novel inclusion of information with regards to current island databases.

The database allows for the retrieval of information by means of numerous approaches in order to afford users the opportunity to identify specific islands in diverse research directions. These approaches include browsing of the cluster representatives, searching the database by means of island location in a host and the identification of island with a specified gene content.

The ability to compare novel or excluded islands to the current content of Pre_GI is a principle component in HT and MGE research. This tool is available to users with raw island sequence data or preferably island GenBank file format. This format allows for valuable added information available in the calculated results. Sequence similarity for a user provided island is done across the totality of the database and results include an alignment visualization option. The computational intensity of compositional

comparisons required the implementation of a novel strategy. Predetermined island representatives are used for an initial indication of compositional similarity and serve as cluster/subcluster markers from where further comparisons are executed. This enables a speedy yet robust calculation of compositional similarity.

Future versions and updates will include numerous alternative island confirmation categories. The absence of similar islands in closely related strains may serve as an indication of a true positive prediction. The construction of the relational database allows for the rapid implementation of such an island confirmation criterion. Various other validation methods such as island mutation rate comparison to that of the original genome and the breaking of host synteny by insertion of an island may furthermore be incorporated.

It should be emphasized that at all times the direction of HT between organisms is a likely inference and not proven. The direction of movement is therefore deemed proposed or probable throughout. Furthermore the existence of an intermediary between organisms is not excluded. The determination of a true donor is complicated by the relatively small amount of sequenced prokaryote genomes available. Future versions and updates of Pre_GI will included newly sequenced genomes and as such current content and movement hypotheses are more than likely to change.

Pre_GI is a functional and reliable collection of island information and comparison tools. The database serves as an holistic island investigation platform with various applications and abilities in the field of HT and MGE research. Future expansion of the database will include novel island information and hypotheses as they become available.

The extensive amount of information embedded in Pre_GI may be analyzed in order to gain an enriched perspective on island communities and general behavior. It is therefore possible to analyze islands independent from their host in search of an improved understanding regarding island lifestyle and logic.

The process of amelioration alters the composition of an island to resemble that of the host it resides in. This hampers the identification of island, especially old inserts, and as such contributes to the false negative rates. On the upside it does provide researchers with an indication of the age or time of insertion of an island in a host. It is therefore possible to predict movement of islands between organisms and as such trace channels of donor-recipient movement between archaeal/bacterial hosts. The post-transfer destiny of an foreign insertion is varied. A possibility is the fragmentation of the insert and spread across the host genome. The compression of island information in a single resource enables the prediction of such events and the elucidation of HT frequency in host genomes. The abundant amount of compositional and sequence similarity information aids in the determination of island groupings. Inspection of all compositional similarity results between islands of different taxonomic backgrounds revealed a general trend

of decreasing similarity as was expected. Local variations to this global trend was not wonted. It is hypothesized that these deviations are due to the organism selection methods used for sequencing in research projects. These deviation reconciles with proposition that taxonomic distant organisms pick from a common gene pool.

The large collection of islands with annotated gene content available in Pre_GI enabled the detection of core genes encountered in the majority of all islands. It was found that genes relating to transfer or transport was in abundance. The high incidence of ABC transporter and related proteins may be associated with island viability and resistance.

Host information in combination with sequence similarity enables the detection of probable movement of islands between distantly related genera within a shared habitat. The physical properties of HT require a sharing of habitat in some point of time. High scoring sequence similarity links between islands of different genera was overlapped with the general habitat information of the host organisms in an attempt to highlight the importance of shared location to HT events. This may illustrate the importance of proximity rather than relatedness in the movement of islands between organisms in future research.

The ability to extract islands with a specific gene content by means of the island gene annotation tool in Pre_GI was used to identify probable islands of resistance. The inclusion of proposed donor-recipient movements between islands with compositional similarity may aid users in the detection of resistance pathways between organisms. Various analysis and comparisons are further possible due to the availability of island sequence information and content through the Pre_GI interface.

Pre_GI aims to be a reliable and contemporary source of information in a developing biological community. The comparison of independent HT hypothesis available in literature to the information contained within Pre_GI is intended to evaluate the database with regards to content and inference. Pre_GI may therefore possibly be used as a method of confirmation or negation in future island studies.

Novel sequencing projects are growing at an exponential rate and require speedy yet reliable analysis tools. The collaboration of SWGIS and Pre_GI provides users with a novel avenue in island prediction and analysis. The ability to identify and compare islands in various newly sequenced organisms, even simple eukaryotes, may aid research in the field of HT and MGE.

Group or community based island analysis could possibly broaden current knowledge and proposition with regards to prokaryotic evolution. The wealth of information included in Pre_GI may ease the burden of large scale island analysis by providing users with a single resource for all comparisons. The inclusion of host information and taxonomy is a novel advent in island databases and may provided users with the necessary information for relevant island inferences.

Pre_GI may be used in the advent of novel future research directions and investigations. The database aims to expand the current limits in island research by providing users with the opportunity to formulate and test various biological hypothesis. Pre_GI may further be used in collaboration with various other available resources in an attempt to elucidate novel HT information.

The culmination of aims and objectives as set out from the onset will hopefully increase current knowledge and guide future research in the field of horizontal transfer, mobile genetic elements and islands in general. The merger of an optimized island identifier and a contemporary expandable database enables access to a wealth of island information in conjunction with numerous analytical tools. This package may be used in various current or future island research directions and approaches. Therefore Pre_GI is presented as a comprehensive and reliable alternative to current island database with numerous novel capabilities in a bid to broaden island philosophy.

References

[1] R. Bock. The give-and-take of DNA: horizontal gene transfer in plants. *Trends in Plant Science*, 15(1):11 – 22, 2010.

[2] S. R. Bordenstein. Symbiosis and the origin of species. In *Insect symbiosis*. Bourtzis K, Miller T, editors. New York: CRC Press., 2003.

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

[4] J. Brockman. *Third Culture: Beyond the Scientific Revolution*. Simon and Schuster, 1995.

[5] C. X. Chan and D. Bhattacharya. The origin of plastids. *Nature Edu*, 3:84, 2010.

[6] G. Chang. Multidrug resistance ABC transporters. *FEBS letters*, 555(1):102–105, 2003.

[7] I. Chen and D. Dubnau. DNA uptake during bacterial transformation. *Nature Reviews Microbiology*, 2(3):241–249, 2004.

[8] O. Cohen and T. Pupko. Inference and characterization of horizontally transferred gene families using stochastic mapping. *Molecular biology and evolution*, 27(3):703–713, 2010.

[9] R. E. Collins and P. G. Higgs. Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome. *Molecular biology and evolution*, 29(11):3413–3425, 2012.

[10] J. de Vries and W. Wackernagel. Integration of foreign DNA during natural transformation of *Acinetobacter* sp. by homology-facilitated illegitimate recombination. *Proceedings of the National Academy of Sciences*, 99(4):2094–2099, 2002.

[11] C. F. Delwiche. Tracing the thread of plastid diversity through the tapestry of life. *the american naturalist*, 154(S4):S164–S177, 1999.

[12] K. M Derbyshire and T. A. Gray. Distributive Conjugal Transfer: New Insights into Horizontal Gene Transfer and Genetic Exchange in Mycobacteria. *Microbiology spectrum*, 2(1), 2014.

[13] C. Dutta and A. Pan. Horizontal gene transfer and bacterial diversity. *Journal of biosciences*, 27(1):27–33, 2002.

[14] A. Crisp et al. Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. *Genome Biology*, 16(1):50, 2015.

[15] A. J. Enright et al. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, 30(7):1575–1584, 2002.

[16] A. P. Kuzin et al. Enzymes of vancomycin resistance: the structure of d-alanine–d-lactate ligase of naturally resistant *Leuconostoc mesenteroides*. *Structure*, 8(5):463–470, 2000.

[17] A. Paauw et al. Evolution in quantum leaps: multiple combinatorial transfers of HPI and other genetic modules in *Enterobacteriaceae*. *PLoS One*, 5(1):e8662, 2010.

[18] A. R. Achour et al. Diversity of arsenite transporter genes from arsenic-resistant soil bacteria. *Research in microbiology*, 158(2):128–137, 2007.

[19] A. Shiroma et al. First complete genome sequences of *Staphylococcus aureus* subsp. *aureus* rosenbach 1884 (DSM 20231), determined by PacBio single-molecule real-time technology. *Genome announcements*, 3(4):e00800–15, 2015.

[20] B. Fernández-Gómez et al. Patterns and architecture of genomic islands in marine bacteria. *BMC genomics*, 13(1):347, 2012.

[21] B. K. Dhillon et al. IslandViewer 3: more flexible, interactive genomic island discovery, visualization and analysis. *Nucleic Acids Res.*, page gkv401, 2015.

[22] C. Ash et al. Phylogenetic heterogeneity of the genus *Bacillus* revealed by comparative analysis of small-subunit-ribosomal RNA sequences. *Letters in Applied Microbiology*, 13(4):202–206, 1991.

[23] C. Bowler et al. The phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature*, 456(7219):239–244, 2008.

[24] C. Chapus et al. Exploration of phylogenetic data using a global sequence analysis method. *BMC evolutionary biology*, 5(1):63, 2005.

[25] C. Dufraigne et al. Detection and characterization of horizontal transfers in prokaryotes using genomic signature. *Nucleic Acids Research*, 33(1):e6–e6, 2005.

[26] C. Hall et al. Contribution of Horizontal Gene Transfer to the Evolution of *Saccharomyces cerevisiae*. *Eukaryotic Cell*, 4(6):1102–1115, 2005.

[27] C. Ku et al. Complete genome sequence of *Spiroplasma apis* B31T (ATCC 33834), a bacterium associated with May disease of honeybees (*Apis mellifera*). *Genome announcements*, 2(1):e01151–13, 2014.

[28] C. Martineau et al. Complete genome sequence of *Hyphomicrobium nitrativorans* strain NL23, a denitrifying bacterium isolated from biofilm of a methanol-fed denitrification system treating seawater at the Montreal Biodome. *Genome announcements*, 2(1):e01165–13, 2014.

[29] C. Mouches et al. A spiroplasma of serogroup IV causes a May-disease-like disorder of honeybees in Southwestern France. *Microbial ecology*, 8(4):387–399, 1982.

[30] D. Bhattacharya et al. Photosynthetic eukaryotes unite: endosymbiosis connects the dots. *Bioessays*, 26(1):50–60, 2004.

[31] D. Domman et al. Plastid establishment did not require a chlamydial partner. *Nature communications*, 6, 2015.

[32] D. H. Bouanchaud et al. Elimination by ethidium bromide of antibiotic resistance in enterobacteria and staphylococci. *Journal of general microbiology*, 54(3):417–425, 1968.

[33] D. J. Smyth et al. Conjugative transfer of ICESde3396 between three beta-hemolytic streptococcal species. *BMC research notes*, 7(1):521, 2014.

[34] D. T. Suzuki et al. *An introduction to genetic analysis*. Number Ed. 3. WH Freeman and Company, 1986.

[35] E. A. Gladyshev et al. Massive Horizontal Gene Transfer in Bdelloid Rotifers. *Science*, 320(5880):1210–1213, 2008.

[36] E. C. M. Nowack et al. Chromatophore genome sequence of *Paulinella* sheds light on acquisition of photosynthesis by eukaryotes. *Current Biology*, 18(6):410–418, 2008.

[37] E. Denker et al. Horizontal gene transfer and the evolution of cnidarian stinging cells. *Current Biology*, 18(18):R858 – R859, 2008.

[38] E. S. Lander et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[39] E. V. Koonin et al. Horizontal Gene Transfer in Prokaryotes: Quantification and Classification. *Annual Review of Microbiology*, 55(1):709–742, 2001.

[40] F. Sievers et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology*, 7(1):539, 2011.

[41] G. J. E. Danchin et al. Multiple lateral gene transfers and duplications have promoted plant parasitism ability in nematodes. *Proceedings of the National Academy of Sciences*, 107(41):17651–17656, 2010.

[42] G. Méric et al. Ecological overlap and horizontal gene transfer in *Staphylococcus aureus* and *Staphylococcus epidermidis*. *Genome biology and evolution*, 7(5):1313–1328, 2015.

[43] G. Ricard et al. Horizontal gene transfer from Bacteria to rumen Ciliates indicates adaptation to their anaerobic, carbohydrates-rich environment. *BMC Genomics*, 7(1):22, 2006.

[44] G. Schönknecht et al. Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. *Science*, 339(6124):1207–1210, 2013.

[45] H. Fukushima et al. Clinical experiences in Sakai City Hospital during the massive outbreak of enterohemorrhagic *Escherichia coli* O157 infections in Sakai City, 1996. *Pediatrics International*, 41(2):213–217, 1999.

[46] H. Ganesan et al. 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, 2008.

[47] H. Michino et al. Massive outbreak of *Escherichia coli* O157: H7 infection in schoolchildren in Sakai City, Japan, associated with consumption of white radish sprouts. *American journal of epidemiology*, 150(8):787–796, 1999.

[48] H. Sidjabat et al. Interspecies transfer of *bla*IMP-4 in a patient with prolonged colonization by IMP-4-producing *Enterobacteriaceae*. *Journal of clinical microbiology*, 52(10):3816–3818, 2014.

[49] H. Takami et al. Genomic characterization of thermophilic *Geobacillus* species isolated from the deepest sea mud of the mariana trench. *Extremophiles*, 8(5):351–356, 2004.

[50] H. Urbanczyk et al. Phylogenetic analysis of the incidence of lux gene horizontal transfer in *Vibrionaceae*. *Journal of bacteriology*, 190(10):3494–3504, 2008.

[51] H. Y. Ou et al. MobilomeFINDER: web-based tools for in silico and experimental discovery of bacterial genomic islands. *Nucleic Acids Res.*, 35(suppl 2):W97–W104, 2007.

[52] I. Chen et al. The Ins and Outs of DNA Transfer in Bacteria. *Science*, 310(5753):1456–1460, 2005.

[53] I. Chen et al. A macromolecular complex formed by a pilin-like protein in competent *Bacillus subtilis*. *Journal of Biological Chemistry*, 281(31):21720–21727, 2006.

[54] I. M. Banat et al. *Geobacillus debilis* sp. nov., a novel obligately thermophilic bacterium isolated from a cool soil environment, and reassignment of *Bacillus pallidus* to *Geobacillus pallidus* comb. nov. *International journal of systematic and evolutionary microbiology*, 54(6):2197–2201, 2004.

[55] I. Rajan et al. Identification of compositionally distinct regions in genomes using the centroid method. *Bioinformatics*, 23(20):2672–2677, 2007.

[56] J. A. Chapman et al. The dynamic genome of Hydra. *Nature*, 464(7288):592–596, 2010.

[57] J. B. Hagen et al. Margulis and the question of how cells evolved. In *Doing Biology*. Benjamin Cummings, 1997.

[58] J. Becq et al. A benchmark of parametric methods for horizontal transfers detection. *PLoS One*, 5(4):e9989, 2010.

[59] J. Bohlin et al. Relative entropy differences in bacterial chromosomes, plasmids, phages and genomic islands. *BMC genomics*, 13(1):66, 2012.

[60] J. C. D. Hottop et al. Widespread Lateral Gene Transfer from Intracellular Bacteria to Multicellular Eukaryotes. *Science*, 317(5845):1753–1756, 2007.

[61] J. C. Venter et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001.

[62] J. Klockgether et al. Diversity of the abundant pklc102/pagi-2 family of genomic islands in *Pseudomonas aeruginosa*. *Journal of bacteriology*, 189(6):2443–2459, 2007.

[63] J. Lederberg et al. Recombination analysis of bacterial heredity. In *Cold Spring Harbor symposia on quantitative biology*, volume 16, pages 413–443. Cold Spring Harbor Laboratory Press, 1951.

[64] J. M. Archibald et al. Lateral gene transfer and the evolution of plastid-targeted proteins in the secondary plastid-containing alga *Bigelowiella natans*. *Proceedings of the National Academy of Sciences*, 100(13):7678–7683, 2003.

[65] J. R. Zaneveld et al. Are all horizontal gene transfers created equal? Prospects for mechanism-based studies of HGT patterns. *Microbiology*, 154(1):1–15, 2008.

[66] K. Papadimitriou et al. Acquisition through Horizontal Gene Transfer of Plasmid pSMA198 by *Streptococcus macedonicus* ACA-DC 198 Points towards the Dairy Origin of the Species. *PloS one*, 10(1), 2015.

[67] L. Raphaël et al. ACLAME: a CLAssification of Mobile genetic Elements, update 2010. *Nucleic Acids Res.*, page gkp938, 2009.

[68] L. S. Frost et al. Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology*, 3(9):722–732, 2005.

[69] L. X. Nouvel et al. Comparative genomic and proteomic analyses of two *Mycoplasma agalactiae* strains: clues to the macro-and micro-events that are shaping mycoplasma diversity. *BMC Genomics*, 11(1):86, 2010.

[70] M. G. I. Langille et al. Evaluation of genomic island predictors using a comparative genomics approach. *BMC Bioinformatics*, 9(1):329, 2008.

[71] M. G. I. Langille et al. Detecting genomic islands using bioinformatics approaches. *Nature Reviews Microbiology*, 8(5):373–382, 2010.

[72] M. Juhas et al. Genomic islands: tools of bacterial horizontal gene transfer and evolution. *FEMS microbiology reviews*, 33(2):376–393, 2009.

[73] M. K. Cheung et al. 2011 German *Escherichia coli* o104: H4 outbreak: whole-genome phylogeny without alignment. *BMC research notes*, 4(1):533, 2011.

[74] M. Strätz et al. System to study horizontal gene exchange among microorganisms without cultivation of recipients. *Molecular microbiology*, 22(2):207–215, 1996.

[75] N. Kondo et al. Genome fragment of *Wolbachia* endosymbiont transferred to X chromosome of host insect. *Proceedings of the National Academy of Sciences*, 99(22):14280–14285, 2002.

[76] N. Nikoh et al. *Wolbachia* genome integrated in an insect chromosome: Evolution and fate of laterally transferred endosymbiont genes. *Genome Research*, 18(2):272–280, 2008.

[77] O. Bezuidt et al. SeqWord Gene Island Sniffer: a program to study the lateral genetic exchange among bacteria. *World Academy of Science, Engineering and Technology*, 58:1169–11274, 2009.

[78] O. Bezuidt et al. 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, 2011.

[79] O. T. Avery et al. Studies on the chemical nature of the substance inducing transformation of pneumococcal types induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii. *The Journal of experimental medicine*, 79(2):137–158, 1944.

[80] R. Acuña et al. Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee. *Proceedings of the National Academy of Sciences*, 109(11):4197–4202, 2012.

[81] R. Chatterjeel et al. On detection and assessment of statistical significance of Genomic Islands. *BMC genomics*, 9(1):150, 2008.

[82] R. Jain et al. Horizontal gene transfer in microbial genome evolution. *Theoretical population biology*, 61(4):489–495, 2002.

[83] R. Jain et al. Horizontal gene transfer accelerates genome innovation and evolution. *Molecular biology and evolution*, 20(10):1598–1602, 2003.

[84] R. Kaden et al. Whole-genome sequence of *Brucella canis* strain sva13, isolated from an infected dog. *Genome announcements*, 2(4):e00700–14, 2014.

[85] S. C. Choi et al. Replacing and additive horizontal gene transfer in *Streptococcus*. *Molecular biology and evolution*, 29(11):3309–3320, 2012.

[86] S. D. Dyall et al. Ancient Invasions: From Endosymbionts to Organelles. *Science*, 304(5668):253–257, 2004.

[87] S. G. Ball et al. Metabolic Effectors Secreted by Bacterial Pathogens: Essential Facilitators of Plastid Endosymbiosis? *The Plant Cell Online*, 25(1):7–21, 2013.

[88] S. Garcia-Vallve et al. Horizontal Gene Transfer of Glycosyl Hydrolases of the Rumen Fungi. *Molecular Biology and Evolution*, 17(3):352–361, 2000.

[89] S. Garcia-Vallvé et al. HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Res.*, 31(1):187–189, 2003.

[90] S. H. Yoon et al. PAIDB v2. 0: exploration and analysis of pathogenicity and resistance islands. *Nucleic Acids Res.*, page gku985, 2014.

[91] S. J. Sørensen et al. Studying plasmid horizontal transfer in situ: a critical review. *Nature Reviews Microbiology*, 3(9):700–710, 2005.

[92] S. L. Salzberg et al. Microbial genes in the human genome: Lateral transfer or gene loss? *Science*, 292(5523):1903–1906, 2001.

[93] S. Ménigaud et al. GOHTAM: a website for 'genomic origin of horizontal transfers, alignment and metagenomics'. *Bioinformatics*, 28(9):1270–1271, 2012.

[94] S. N. McNulty et al. Endosymbiont DNA in Endobacteria-Free Filarial Nematodes Indicates Ancient Horizontal Genetic Transfer. *PLoS ONE*, 5(6), 2010.

[95] S. Nurk et al. Assembling genomes and mini-metagenomes from highly chimeric reads. *Research in computational molecular biology. Springer Verlag, Berlin, Germany*, pages 158–170, 2013.

[96] S. Pundhir et al. PredictBias: a server for the identification of genomic and pathogenicity islands in prokaryotes. *In silico biology*, 8(3-4):223–234, 2008.

[97] S. Waack et al. Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinformatics*, 7(1):142, 2006.

[98] T. Abe et al. Informatics for unveiling hidden genome signatures. *Genome research*, 13(4):693–702, 2003.

[99] T. Akiba et al. On the mechanism of the development of multiple-drug-resistant clones of Shigella. *Japanese journal of microbiology*, 4(2):219–227, 1960.

[100] T. Kyndt et al. The genome of cultivated sweet potato contains *Agrobacterium* T-DNAs with expressed genes: An example of a naturally transgenic food crop. *Proceedings of the National Academy of Sciences*, 112(18):5844–5849, 2015.

[101] T. N. Nazina et al. Taxonomic study of aerobic thermophilic bacilli: descriptions of *Geobacillus subterraneus* gen. nov., sp. nov. and *Geobacillus uzenensis* sp. nov. from petroleum reservoirs and transfer of *Bacillus stearothermophilus, Bacillus thermocatenulatus, Bacillus thermoleovorans, Bacillus kaustophilus, Bacillus thermodenitrificans* to *Geobacillus* as the new combinations *G. stearothermophilus, G. thermodenitrificans*. *International Journal of Systematic and Evolutionary Microbiology*, 51(2):433–446, 2001.

[102] W. Broothaerts et al. Gene transfer to plants by diverse species of bacteria. *Nature*, 433:775–787, 2005.

[103] W. W. Hsiao et al. Evidence of a large novel gene pool associated with prokaryotic genomic islands. *BMC Bioinformatics*, 1:e62, 2005.

[104] X. Lin et al. Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature*, 402:761–768.

[105] X. Zhao et al. Isolation and identification of antifungal peptides from *Bacillus* BH072, a novel bacterium isolated from honey. *Microbiological research*, 168(9):598–606, 2013.

[106] Z. Freedman et al. Mercury resistance and mercuric reductase activities and expression among chemotrophic thermophilic *Aquificae*. *Applied and environmental microbiology*, 78(18):6568–6575, 2012.

[107] V. J. Freeman. Studies on the virulence of bacteriophage-infected strains of *Corynebacterium diphtheriae*. *Journal of bacteriology*, 61(6):675, 1951.

[108] D. P. Genereux and J. M. Logsdon Jr. Much ado about bacteria-to-vertebrate lateral gene transfer. *Trends in Genetics*, 19(4):191 – 195, 2003.

[109] F. Griffith. The significance of pneumococcal types. *Journal of Hygiene*, 27(02):113–159, 1928.

[110] F. Guarner and J. Malagelada. Gut flora in health and disease. *The Lancet*, 361(9356):512–519, 2003.

[111] R. S. Gupta and G. B. Golding. The origin of the eukaryotic cell. *Trends in biochemical sciences*, 21(5):166–171, 1996.

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

[113] J. C. D. Hotopp. Horizontal gene transfer between bacteria and animals. *Trends in Genetics*, 27(4):157 – 163, 2011.

[114] H. Ikeda and J. Tomizawa. Transducing fragments in generalized transduction by phage P1: I. Molecular origin of the fragments. *Journal of molecular biology*, 14(1):85–109, 1965.

[115] K. A. Karberg. Similarity of genes horizontally acquired by *Escherichia coli* and *Salmonella enterica* is evidence of a supraspecies pangenome. *Proceedings of the National Academy of Sciences*, 108(50):20154–20159, 2011.

[116] K. Kitahara and K. Miyazaki. Revisiting bacterial phylogeny: natural and experimental evidence for horizontal gene transfer of 16s rrna. *Mobile genetic elements*, 3(1):e24210, 2013.

[117] G. Klebs. Untersuchungen aus dem bot. pages Institut zu Tubingen, Bd. II, 1887.

[118] J. G. Lawrence. Gene transfer in bacteria: speciation without species? *Theoretical population biology*, 61(4):449–460, 2002.

[119] J. G Lawrence and H. Ochman. Amelioration of bacterial genomes: rates of change and exchange. *Journal of molecular evolution*, 44(4):383–397, 1997.

[120] J. Lederberg and E. L. Tatum. Gene recombination in *Escherichia coli*. *Nature*, 158:558, 1946.

[121] J. Lederberg and E. L. Tatum. Sex in bacteria; genetic studies, 1945–1952. *Science*, 118(3059):169–175, 1953.

[122] P. Legendre and L. Legendre. Ordination in reduced space, 1998.

[123] M. Marcet-Houben and T. Gabaldon. Acquisition of prokaryotic genes by fungal genomes. *Trends in Genetics*, 26(1):5 – 8, 2010.

[124] W. Martin and K. V. Kowallik. Annotated English translation of Mereschkowsky's 1905 paper. *European Journal of Phycology*, 34:287–295, 1999.

[125] T. V. Matveeva and L. A. Lutova. Horizontal gene transfer from *Agrobacterium* to plants. *Frontiers in plant science*, 5, 2014.

[126] G. I. McFadden. Primary and secondary endosymbiosis and the origin of plastids. *Journal of Phycology*, 37(6):951–959, 2001.

[127] C. Mereschkowsky. Uber Natur und Ursprung der Chromatophoren im Pflanzenreiche. *Biol. Centralbl.*, 25:593–604, 1905.

[128] R. Merkl. A comparative categorization of protein function encoded in bacterial or archeal genomic islands. *Journal of molecular evolution*, 62(1):1–14, 2006.

[129] J. Mrazek and S. Karlin. Detecting Alien Genes in Bacterial Genomes. *Annals of the New York Academy of Sciences*, 870(1):314–329, 1999.

[130] W. C. Nelson and J. C. Stegen. The reduced genomes of Parcubacteria (OD1) contain signatures of a symbiotic lifestyle. *Frontiers in microbiology*, 6, 2015.

[131] C. P. Ponting. Plagiarized bacterial genes in the human book of life. *Trends in Genetics*, 17(5):235 – 237, 2001.

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

[133] J. Raymond and R. E. Blankenship. Horizontal gene transfer in eukaryotic algal evolution. *Proceedings of the National Academy of Sciences*, 100(13):7419–7420, 2003.

[134] O. Reva and B. Tümmler. Think big–giant genes in bacteria. *Environmental microbiology*, 10(3):768–777, 2008.

[135] O. N Reva and B. Tümmler. Global features of sequences of bacterial chromosomes, plasmids and phages revealed by analysis of oligonucleotide usage patterns. *BMC Bioinformatics*, 5(1):90, 2004.

[136] O. N Reva and B. Tümmler. Differentiation of regions with atypical oligonucleotide composition in bacterial genomes. *BMC Bioinformatics*, 6(1):251, 2005.

[137] M. A. Riley and M. Lizotte-Waniewski. Population genomics and the bacterial species concept. In *Horizontal Gene Transfer*, pages 367–377. Springer, 2009.

[138] H. Ris and R. N. Singh. Electron microscope studies on blue-green algae. *J Biophys Biochem Cytol*, 9(1):63–80, 1961.

[139] E. P. C. Rocha. With a Little Help from Prokaryotes. *Science*, 339(6124):1154–1155, 2013.

[140] U. L. Rosewich and H. C. Kistler. Role of Horizontal Gene Transfer in the Evolution of Fungi. *Annual Review of Phytopathology*, 38(1):325–363, 2000.

[141] L. Sagan. On the origin of mitosing cells. *Journal of Theoretical Biology*, 14(3):225–274, 1967.

[142] A. F. W. Schimper. Untersuchungen fiber die Chlorophyllkorper und die ihnen homologen Gebilde. *Jb wiss Botanik.*, 16:1–247, 1885.

[143] B. F Smets and T. Barkay. Horizontal gene transfer: perspectives at a crossroads of scientific disciplines. *Nature Reviews Microbiology*, 3(9):675–678, 2005.

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

[145] C. Stocking and E. Gifford. Incorporation of thymidine into chloroplasts of Spirogyra. *Biochem. Biophys. Res. Comm.*, 1(3):159–164, 1959.

[146] B. Stoebe and U. Maier. One, two, three: nature's tool box for building plastids. *Protoplasma*, 219(3-4):123–130, 2002.

[147] D. J. Studholme. Some (bacilli) like it hot: genomics of *Geobacillus* species. *Microbial biotechnology*, 8(1):40–48, 2015.

[148] M. Syvanen. Cross-species gene transfer; implications for a new theory of evolution. *Journal of theoretical Biology*, 112(2):333–343, 1985.

[149] C. M. Thomas and K. M. Nielsen. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature reviews microbiology*, 3(9):711–721, 2005.

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

[151] M. W. J. van Passel et al. An acquisition account of genomic islands based on genome signature comparisons. *BMC Genomics*, 6(1):163, 2005.

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

[153] I. E. Wallin. On the nature of mitochondria. VII. The independent growth of mitochondria in culture media. *American Journal of Anatomy*, 33(1):147–173, 1924.

[154] B. Wang. Limitations of compositional approach to identifying horizontally transferred genes. *Journal of molecular evolution*, 53(3):244–250, 2001.

[155] R. A. F Wozniak and M. K Waldor. Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. *Nature Reviews Microbiology*, 8(8):552–563, 2010.

[156] J. D. Wren and A. Bateman. Databases, data tombs and dust in the wind. *Bioinformatics*, 24(19):2127–2128, 2008.

[157] D. R. Zeigler. The *Geobacillus* paradox: why is a thermophilic bacterial genus so prevalent on a mesophilic planet? *Microbiology*, 160(Pt 1):1–11, 2014.

[158] Y. Zhang and S. M. Sievert. Pan-genome analyses identify lineage-and niche-specific markers of evolution and adaptation in *Epsilonproteobacteria*. *Frontiers in microbiology*, 5, 2014.