

LATERAL GENE TRANSFER IN OPERONS AND ITS EFFECTS ON
NEIGHBOURING GENES

LATERAL GENE TRANSFER IN OPERONS AND ITS EFFECTS ON
NEIGHBOURING GENES

By
ASHER PASHA, B. Sc. (Hons.)

A Thesis
Submitted to the School of Graduate Studies
in Partial Fulfilment of the Requirements
for the Degree
Master of Science

McMaster University

© Copyright by Asher Pasha, September 2012

MASTER OF SCIENCE (2012)
(Biology)

McMaster University
Hamilton, Ontario

TITLE: Lateral Gene Transfer in Operons and Its Effects
on Neighbouring Genes

AUTHOR: Asher Pasha, B. Sc. Hons. (McMaster University)

SUPERVISOR: Dr. G. Brian Golding

NUMBER OF PAGES: ix, 127

Abstract

Prokaryotes evolve, in part, by lateral gene transfer (LGT). This transfer of genetic material is likely important in the evolution of operons, a group of genes that are transcribed as a single mRNA. Genes that are transferred may then be integrated into genomes by homologous recombination. In this thesis, it was proposed that homologous recombination is the mechanism of integration of laterally transferred genes into operons. To investigate this proposal, a phylogenetic tree of *Bacillus* was inferred using DNA sequence alignments. LGT was inferred using a parsimony algorithm, and operons were inferred using OperonDB. Homologous recombination breakpoints were identified by permutation tests, GENECONV and maximum chi square algorithm. The results indicate that there is evidence for integration of functionally annotated genes into operons by homologous recombination. There are several laterally transferred genes that have recombination breakpoints before the start codon or after the stop codon of the genes. It was also proposed in this thesis that LGT causes an increase in the rate of evolution of genes that are neighbours of laterally transferred genes. To investigate this proposal, genes that are neighbours of laterally transferred genes in *Bacillus* were identified. These genes were classified as upstream or downstream genes to the LGT event. Genes that are not neighbours of laterally transferred genes were also identified as a control. Selection and the rate of evolution was studied using maximum likelihood models implemented in CodeML of PAML. Genes under positive selection were inferred using likelihood ratio tests. The results indicate that only a few neighbouring genes were under positive selection, and the rate of evolution of the neighbouring genes was slightly higher than that of the non-neighbouring genes. The high rates of evolution of the neighbouring genes are likely due to relaxed selection on the neighbouring genes.

Acknowledgements

This work was made possible by an NSERC grant to Dr. Golding.

I am thankful to my supervisory committee: Dr. G. B. Golding and Dr. T. M. Finan. I am also thankful to the committee chair: Dr. B. Evans.

I am also thankful to the members of Dr. Golding's lab: Wilfred Haerty, Terri Porter, Yifei Huang, Melanie Lou, Wilson Sung, Vasile Catana, Carolyn Lenz, Tara Shadoway and Hamid Mohamadi. I am thankful to Dr. Evan's lab: Dr. B. Evans, Adam Bewick, Ana-Hermina Ghenu and Jane Shen, I am also thankful to Eric Collins and Judith (Jake) Szamosi.

I am thankful to my family: Edwin Pasha (father), Cynthia Pasha (mother) and Shirley Pasha (sister).

I am thankful to my friends and mentors: Sarah Chiang (Ph. D.), Rev. Peter Paul, Sartaj Singh and Hobbe Smit.

Last, but not the least, I am also thankful to you, the reader of this thesis.

Contents

I	INTRODUCTION	1
1	Introduction	2
1.1	Mechanisms of LGT	2
1.1.1	LGT	2
1.1.2	Transformation	2
1.1.3	Conjugation	3
1.1.4	Transduction	3
1.1.5	Nanotubes	4
1.1.6	Recombination	4
1.2	Detection of LGT	5
1.2.1	Laboratory Methods	5
1.2.2	Computational Methods	5
1.2.3	Accuracy of Detection	7
1.3	Rates of LGT	8
1.4	Barriers to LGT in Bacteria	10
1.4.1	Exclusion of DNA	10
1.4.2	CRISPR elements	11
1.4.3	Sequence Similarity	11
1.4.4	Complexity Hypothesis	12
1.5	Importance of LGT	12
1.5.1	Role in Evolution	12
1.5.2	Role in Medicine	12
1.5.3	Importance in Phylogenetics	13
1.5.4	Profile of Laterally Transferred DNA	14
1.5.5	Speciation	15
1.6	LGT and Eukaryotes	15
II	LATERAL GENE TRANSFER	17
2	Insertions of Genes in Operons	18
2.1	Abstract	18
2.2	Introduction	19

2.2.1	Evolution of Operons	19
2.2.2	Explanations for Operon Formation	21
2.2.3	Prediction of Operons	22
2.2.4	Homologous Recombination in Bacteria	25
2.2.5	Detection of Homologous Recombination in Bacteria	26
2.3	Method	29
2.3.1	Phylogenetic Tree	29
2.3.2	Operon Database	30
2.3.3	Detection of LGT	31
2.3.4	Detection of Homologous Recombination	32
2.4	Results	36
2.4.1	Phylogenetic tree	36
2.4.2	Operon Prediction	36
2.4.3	Prediction of LGT	39
2.4.4	Prediction of HR	43
2.5	Discussion	45
2.5.1	Choice of Organism	45
2.5.2	Phylogenetic Tree	46
2.5.3	Operons	47
2.5.4	Detection of LGT	48
2.5.5	Detection of Homologous recombination	48
3	The Effect of LGT on Neighbouring Genes	53
3.1	Abstract	53
3.2	Introduction	54
3.2.1	Evolution of Prokaryotes	54
3.2.2	Factors Effecting Selection	56
3.2.3	Detecting selection	57
3.2.4	Detection of sites under positive selection	61
3.2.5	Processes that affect neighbouring genes	61
3.3	Method	62
3.3.1	Phylogenetic Tree	62
3.3.2	Detection of LGT	63
3.3.3	Identification of Non-neighbouring Genes	63
3.3.4	Identification of Neighbouring Genes	63
3.3.5	Detection of Rates of Evolution and Positive Selection	64
3.3.6	Likelihood Ratio Test	64
3.3.7	Rates of Evolution and Magnitude of Selection	65
3.3.8	Wilcoxon Rank Sum Test and Permutation Test	65
3.3.9	Summary	65
3.4	Results	68
3.4.1	Phylogenetic tree	68
3.4.2	Detection of LGT	68

3.4.3	Non-neighbouring, upstream and downstream genes	68
3.4.4	Likelihood Ratio Tests	68
3.4.5	Rate of evolution	71
3.4.6	Type of selection	73
3.5	Discussion	77
3.5.1	Phylogenetic Tree	77
3.5.2	Inference of LGT	78
3.5.3	Neighbouring and Non-Neighbouring genes	79
3.5.4	Rate of Evolution and Selection	81
 III CONCLUSION		 84
4	Summary	85
4.1	LGT into Operons by Homologous Recombination	85
4.2	Effect of LGT on Neighbouring genes	87
 IV BIBLIOGRAPHY		 89
 V APPENDICES		 107
A	Insertion of Genes into Operons	108
B	Effects of LGT on Neighbouring Genes	118

List of Tables

2.1	Table shows the following results of OperonDB and LGT detection.	39
2.2	Pathogenic genes	42
3.1	Likelihood ratio tests	71
3.2	Permutation tests on tree lengths	73
3.3	Selection on upstream genes	74
3.4	Selection on non-neighbouring genes	75
3.5	Selection on downstream genes	75
3.6	Wilcoxon rank sum tests on selection data	76
3.7	Permutation tests on dN/dS ratios	76
A.1	Bacteria and their accession numbers	108
A.2	Results of OperonDB	109
A.3	Laterally transferred genes	114
B.1	Tree lengths of upstream genes	121
B.2	Tree lengths of non-neighbouring genes	122
B.3	Tree lengths of downstream genes	122

List of Figures

2.1	Summary of methods	34
2.2	Summary of methods (continued)	35
2.3	The phylogenetic tree showing posterior probabilities	37
2.4	The phylogenetic tree showing branch lengths	38
2.5	Location of genes in operons	41
2.6	An example of lateral gene transfer into operons	44
2.7	Another example of lateral gene transfer into operons	45
3.1	Summary of methods	66
3.2	Summary of methods (continued)	67
3.3	The phylogenetic tree with the probabilities	69
3.4	The phylogenetic tree with branch lengths	70
3.5	Tree lengths of upstream genes	72
3.6	Tree lengths of non-neighbouring genes	73
3.7	Tree lengths of downstream genes	74
3.8	Families of non-neighbouring genes	80

Part I
INTRODUCTION

Chapter 1

Introduction

1.1 Mechanisms of LGT

1.1.1 LGT

Variations in prokaryotic genomes can be generated by spontaneous mutations (Ochman *et al.*, 2000), but the rapid rate of evolution of microbial genomes cannot be explained by mutations alone. Lateral gene transfer (LGT) is a process by which DNA is transferred into an organism by mechanisms other than vertical inheritance. However, LGT can explain the ability of prokaryotes to evolve rapidly and exploit different environments. For example, the rapid spread of antibiotic resistance in hospitals and throughout the community are largely due to LGT of antibiotic resistance genes (Davies, 1996). If the function of laterally transferred gene is beneficial to bacteria, then it can be retained in the population. Genes that are not beneficial are lost from the population (Lawrence, 1999a). Moreover, LGT and recombination can remove or replace deleterious genes having deleterious mutations (Muller, 1932).

1.1.2 Transformation

The mechanisms of LGT are transformation, transduction, conjugation and via nanotubes (Dubey and Ben-Yehuda, 2011). Transformation is the uptake of extracellular DNA (Griffith, 1928). DNA from a distantly related organism can be acquired by transformation (Ochman *et al.*, 2000). For natural transformation, the bacteria must be in a state of competence, which involves the functioning of 20 to 50 proteins (Cohan, Roberts and King, 1991). Some organisms such as *Neisseria gonorrhoeae* and *Haemophilus influenzae* require special DNA uptake sequences for transformation. The uptake sequences are present in the genomes of these organisms at higher than expected frequencies (Ochman *et al.*, 2000). Moreover, *N. gonorrhoeae* and *H. influenzae*, are always in the state of competence, whereas other species, such as *Bacillus subtilis* and *Streptococcus pneumoniae*, become competent only at certain stages in their life cycle (Ochman *et al.*, 2000). *B. subtilis* is competent only at high population density whereas, most pathogenic bacteria are always

competent (Jain *et al.*, 2002). In *S. pneumoniae*, a seventeen amino acid long competence simulating peptide (CSP) is involved in competence (Barraclough, Balbi and Ellis, 2012). Quorum sensing of CSP in the medium induces competence in *S. pneumoniae* and results in uptake of DNA from the medium. It can also lyse the cells and release their DNA if the cells are not able to become competent (Barraclough, Balbi and Ellis, 2012).

The competence in bacteria is regulated by environmental factors such as altered growth conditions, nutrient access, cell density and starvation. Persistence of stable DNA in the environment is a prerequisite for natural transformation. DNA is released into the environment after the death of bacteria or lysis of bacteria by bacteriolytic viruses (Thomas and Nielsen, 2005). The stability of DNA depends on the physical conditions of the environment. For example, approximately $1\mu\text{g}$ of extracellular DNA can be recovered for each gram of soil (Ogram, Sayler and Barkay, 1987). When competent bacteria are exposed to the extra cellular DNA, the DNA is converted to single stranded DNA and taken up by the bacteria. Restriction enzymes cannot degrade foreign DNA in bacteria because the foreign DNA is taken up as single stranded DNA (Berndt, Meier and Wackernagel, 2003). The foreign DNA can stay in bacteria as plasmid or integrate into the chromosome of the bacteria.

1.1.3 Conjugation

Conjugation is the process by which DNA is transferred from one cell to another by cell-to-cell contact (Tatum and Lederberg, 1947). Self-transmissible or mobile plasmids are usually transferred via conjugation (Heinemann, Sprague Jr *et al.*, 1989). Plasmids can also be integrated into genomes. However, parts of chromosomes can be excised and transferred to other cells via conjugation (Thomas and Nielsen, 2005). Complete chromosomes are unlikely to be transferred by conjugation because they usually require more than one hour to transfer, and the cell-to-cell contacts break down before the transfer is complete (Thomas and Nielsen, 2005). DNA is usually transferred as a single stranded polymer through the conjugation pore (Berndt, Meier and Wackernagel, 2003). Plasmids can be self transmissible, such as those that are used by type IV secretion systems. Plasmids that are not self transmissible can be mobilized by the help of a self transmissible plasmid (Thomas and Nielsen, 2005). Conjugations require the donor cell to be present near the recipient cell, unlike transformations which do not require the presence of the donor cell in the environment.

1.1.4 Transduction

Transduction is the process by which DNA is introduced into the host bacterium by a bacteriophage (Zinder and Lederberg, 1952). The bacteriophage replicates in the host and transfers the DNA from one host to another. In the case of generalized transduction, the bacteriophage packages random DNA; whereas in specialized transduction, it packages DNA adjacent to its attachment site (Morse, Lederberg and Lederberg, 1956). The amount of DNA that can be successfully packaged depends on the size of the capsid of the bac-

terioophage. Usually, the amount of packaged DNA is approximately 100 kbp (Ochman *et al.*, 2000). Successful transformation requires that the new host cell has the receptors for the bacteriophage and that the foreign DNA does not get degraded in the new host cell (Ochman *et al.*, 2000). Similar to transformation, transduction does not require the donor and the recipient to be present at the same location. Recently, it was discovered that gene transfer agents (GTAs) can transfer DNA in oceans (McDaniel *et al.*, 2010). GTAs are viral-like particles produced by α -Proteobacteria. Marine bacteria acquire DNA by LGT facilitated by GTAs (McDaniel *et al.*, 2010). Gene transfer agents, such as viruses, have small geographical ranges and can only affect an organism within their range (Lawrence and Hendrickson, 2003).

1.1.5 Nanotubes

DNA and proteins can also be transferred via nanotubes connecting cells (Dubey and Ben-Yehuda, 2011). These nanotubes are between 30 to 130 nm in diameter. Macromolecules such as proteins (including green fluorescent protein and anti-biotic resistance protein) and DNA can be transferred without the requirement of special proteins. In contrast, conjugational transfers of genes minimally require the *tra* gene products (Dubey and Ben-Yehuda, 2011). The nanotubes are composed of membrane constituents. They can be an array of tubes connecting cells or one thick tube connecting two cells. They can be found in both gram positive and gram negative bacteria (Dubey and Ben-Yehuda, 2011). However, they are only observed when the bacteria is grown on solid medium, not liquid medium (Dubey and Ben-Yehuda, 2011).

1.1.6 Recombination

Laterally transferred genes that are not on plasmids have to integrate into chromosomes by homologous recombination (Lederberg and Tatum, 1946) or by additive integration (Moore and Haber, 1996). Homologous recombination depends on the presence of highly similar regions of DNA sequence. The foreign DNA must have a length of 25 to 200 base pairs and be a sequence with high similarity to the host genome (Lovett *et al.*, 2002). Since homologous recombination depends on the similarity of the sequence, it is not surprising to find out that as the divergence between the bacteria increases, the rate of homologous recombination decreases (Lovett *et al.*, 2002). There are two types of homologous recombination (Chan, Beiko and Ragan, 2006). In reciprocal recombination, the recombinant region is exchanged between the two similar DNA sequences. In nonreciprocal recombination, the recombinant region is replaced by the transferred region (Chan, Beiko and Ragan, 2006). Both processes can happen after DNA mismatch repair.

Foreign DNA can also be integrated by homology facilitated illegitimate recombination (De Vries and Wackernagel, 2002). Homology facilitated illegitimate recombination requires short segments of DNA, similar to the recipient, to anchor the foreign DNA in the chromosome of the recipient (De Vries and Wackernagel, 2002). DNA can also be joined by a non-homologous end joining (NHEJ) mechanism which does not require DNA sim-

ilarity (Popa *et al.*, 2011). NHEJ requires DNA **Ku** protein and the ATP dependent DNA ligase (LigD). This mechanism can bypass the DNA similarity requirement for integration of DNA. Hence, bacteria can integrate DNA from distantly related sources using the NHEJ mechanism (Popa *et al.*, 2011).

The NHEJ DNA repair mechanism does not exist in *Escherichia coli* (Chayot *et al.*, 2010). For many years, *Escherichia coli* had been used as a control for NHEJ experiments. However, Chayot *et al.* (2010) discovered that *E. coli* can repair double strand breaks in DNA by an alternative end-joining (A-EJ) mechanism that does not require Ku and LigaseD. A-EJ requires RecBCD and LigaseA to join double strand breaks. Hence, it can also facilitate the integration of laterally acquired DNA. Chayot *et al.* (2010) also suggested that A-EJ is an ancestral mechanism to alternative non-homologous end joining mechanism in some eukaryotes.

1.2 Detection of LGT

1.2.1 Laboratory Methods

LGT is extensively documented in the literature. However, it is important to detect LGT to understand its significance. In the laboratory, it is possible to observe the change in deficient strains in the presence of a donor strain or foreign DNA. Prior to gene transfer, the recipient strain lacks the features present in the donor strain. After LGT, the recipient strain will have the features provided by the donor strain. Thus, the change can be explained by LGT because of the similarity of the donor strain and the recipient strain (Ochman *et al.*, 2000). Genetic analysis, carried out by sequencing the region and aligning it with the host genome, can provide evidence for LGT (Ochman *et al.*, 2000).

1.2.2 Computational Methods

Computational methods can also be used to detect laterally transferred genes. It can be assumed that inconsistencies in gene trees can be due to LGT (Lawrence and Hartl, 1992). The conflict in phylogenetic trees can be due to variation in evolutionary rates, high levels of homoplasy (Lawrence and Hartl, 1992). To test whether the conflict in topology is due to LGT, and not any other phenomenon, data can be bootstrapped. Identity coefficients can be calculated and ranked. The correlation between the ranks can be calculated by Spearman statistics. Spearman statistic of +1 indicates a perfect positive correlation, while a statistic of -1 indicates a negative correlation (Lawrence and Hartl, 1992). A distribution of similarity coefficients can be calculated by repeating the bootstrap process for the bootstrapped datasets. In case of LGT, the distribution of similarity coefficients should be smaller than that of bootstrapped dataset. If LGT is excluded, the difference should be reduced significantly (Lawrence and Hartl, 1992).

LGT can also be detected by studying differences in codon usage and base pair composition of a gene in relation to other genes (Ragan, 2001). DNA composition may be different due to differences in the purine-to-pyrimidine ratio, oligonucleotide frequencies

or codon usage. For example, there may be GC rich islands in a region with low GC content, indicating a possible LGT event from a genome that has higher GC content. Previous studies used GC content at first and third codon positions to identify laterally transferred genes in *E. coli* MG1655 genome (Lawrence and Ochman, 1997). They suggested that 17.6% of open reading frames had originated by LGT since the divergence of *Escherichia* and *Salmonella* about 100 Mya. In those studies, atypical GC content needed to be two or more standard errors different than the mean for all ORFs in the respective genome (Koski, Morton and Golding, 2001). Codon bias can be measured by the χ^2 statistic of the codon usage (Lawrence and Ochman, 1997). However, χ^2 statistic can only tell whether there is a codon bias and not the direction of the bias (Sharp and Li, 1987). If the genes have atypical composition, then it has high codon bias and high CAI (Lawrence and Ochman, 1997). In a recent study by Koski, Morton and Golding (2001), it was shown that GC content at the first and the third position of codons is a better indicator of LGT than the GC content of codons at the third position alone. It was also found out that using DNA composition to identify LGT misses approximately 300 laterally transferred genes between *E. coli* and *S. typhi* (Koski, Morton and Golding, 2001). It can be concluded by analysis of conflicting phylogenetic topologies of gene trees with the species tree and chromosomal positions of orthologs, that DNA composition and the codon usage are not the best indicators of LGT. Either of them should not be used as a sole criterion for detecting LGT (Koski, Morton and Golding, 2001).

Once the foreign gene is integrated into the host genome, it is affected by the same mutational and selection bias as the native genes of the host. This leads to amelioration of foreign genes in the host genomes (Lawrence and Ochman, 1997). Over long time periods, the base pair composition of the transferred genes can reflect the base pair composition of the host genome. Thus, it is possible to correctly identify recently transferred genes from genomes that have different base pair compositions, but not the ancient gene transfers even if the base pair composition of the genomes differs in the ancestral population (Lawrence and Ochman, 1997). Moreover, some laterally transferred genes may not have codon and nucleotide bias (Gogarten and Townsend, 2005).

Another method of detecting LGT is conflicting topologies of gene tree and species tree (Gogarten, Doolittle and Lawrence, 2002). It is estimated that between 1.6 and 32.6 per cent of genes in each microbial genome originate due to lateral gene transfer (Koonin, Makarova and Aravind, 2001). However, this method is not very effective in differentiating between laterally transferred genes and massive parallel deletions (Boto, 2010). Moreover, methods of phylogeny inference can bias the detection of laterally transferred genes (Daubin, Moran and Ochman, 2003). Based on phylogenies, it is difficult to distinguish between LGT and duplication when distantly related taxa are compared. Moreover, apparent phylogenetic conflict could be due to noise in phylogenetic signal (Gogarten and Townsend, 2005). If the laterally transferred genes are perfectly conserved, phylogenetic tree cannot be resolved, and the conflict between phylogenetic trees cannot be determined (Gogarten and Townsend, 2005). Therefore, phylogenetic conflict does not always imply LGT.

Phylogenetic methods and composition methods may be insufficient to detect laterally

transferred genes (Azad and Lawrence, 2007). Phylogenetic methods often suffer from biased mutation rates, gene deletion and long branch attraction problems. Whereas base pair composition methods have no accurate threshold to separate laterally transferred genes from native genes (Azad and Lawrence, 2007). It was observed that many genes are transferred as genomic islands (Azad and Lawrence, 2007). Therefore, positional information of the genes is important for detection of laterally transferred genes. Besides phylogenetic methods and composition methods, another method of detection of LGT is to cluster atypical genes and then classify the genes as laterally transferred based on their position in the genome (Azad and Lawrence, 2007). Genes can be clustered using the Jensen-Shannon divergence measure that calculates the divergence between two probability distributions of discrete random variable (Azad and Lawrence, 2007).

Laterally transferred genomic islands can be identified by identifying individual genes in the island as laterally transferred. However, new research presents another method of detection of laterally transferred islands (Arvey *et al.*, 2009). The chromosome is divided recursively into smaller and smaller fragments. The fragments are characterized as laterally transferred based on base pair composition of the fragments (Arvey *et al.*, 2009).

In most studies, it is assumed that genes are transferred as a whole. However, this is not always the case, and there are ways to detect partial gene transfers (Boc and Makarenkov, 2011). One of them includes detection of LGT by phylogenetic conflict. A rooted phylogenetic tree can be constructed. Then a sliding window is used to get parts of genes or genomes. A phylogenetic tree is constructed for each segment and compared against the species tree (Boc and Makarenkov, 2011). Besides phylogenetic conflict, other methods of detection of LGT can also be applied using data from the sliding windows (Boc and Makarenkov, 2011).

If a gene is large (greater than 250 bp), and is very atypical in composition, then it can be easily identified as laterally transferred gene. If the gene is small, and its base pair composition does not deviate significantly from the mean base pair composition, then classifying the gene as laterally transferred gene is questionable (Azad and Lawrence, 2011). Two thresholds for base pair composition of laterally transferred genes can be used. Genes between the two thresholds can be classified as ambiguous genes. Short genes, not necessarily classified as ambiguous, can be reclassified as laterally transferred or native based on information about the genes flanking them (Azad and Lawrence, 2011). If a gene is preliminarily classified as native and exists between two laterally transferred genes, then it can be reclassified as laterally transferred. Similarly, if the gene is preliminarily classified as laterally transferred and exists between two neighbouring genes then it can be reclassified as laterally transferred gene (Azad and Lawrence, 2011). These assumptions are based on the observation that most genes are transferred as islands rather than single units (Azad and Lawrence, 2011).

1.2.3 Accuracy of Detection

Due to the large number of methods to detect LGT, it is difficult to determine which methods are the most accurate in detecting LGT (Azad and Lawrence, 2005). Moreover, dif-

ferent methods of detection of LGT may give different results because different methods test different null hypotheses (Azad and Lawrence, 2005). The accuracy and error rate from many methods cannot be determined. There is also inadequate number of genomes sequenced for some species, making it difficult to determine the accuracy of the laterally transferred genes (Azad and Lawrence, 2005).

To test the accuracy of methods of detection of LGT artificial genomes can be simulated using generalized hidden Markov models. Lateral gene transfer can be simulated in these genomes (Azad and Lawrence, 2005). Then, methods to detect LGT can be applied to artificial genomes, and the performance of these methods can be evaluated (Azad and Lawrence, 2005). Type I error occurs when the null hypothesis is true, but it is rejected. Type II error occurs when the null hypothesis is false, but it is accepted. The results indicate that nucleotide composition based tests show low type II error rates. In this case, only a small number of genes identified as laterally transferred are, in fact, native. However, these methods have high type I error rates, which shows that many genes are incorrectly identified as laterally transferred (Azad and Lawrence, 2005). Open reading frames that are smaller than 250 bp are often incorrectly classified as laterally transferred (Azad and Lawrence, 2005).

1.3 Rates of LGT

When the sequencing and assembly of genomes started, it was observed that GC content among bacterial genomes varies significantly (Lawrence and Ochman, 1997). The variance of the GC content was due to directional mutation pressure which is the difference in mutation rates among the four bases. Phylogenetically similar bacteria can be organized into groups of related clades indicating that the base pair composition is approximately similar in closely related species (Lawrence and Ochman, 1997). Moreover, the base pair composition is approximately similar over entire bacterial chromosome (Lawrence and Ochman, 1997). When a foreign gene is introduced via LGT, it can be detected by differences in GC content from the recipient genome (Lawrence and Ochman, 1997). Over time, the foreign gene ameliorates and the base pair composition of the foreign gene becomes similar to that of the recipient genome. The rate of amelioration can be used to estimate the relative time and the rate of gene gain and gene loss in bacteria (Lawrence and Ochman, 1997). The rate of gene gain and gene loss is approximately as high as the rate of mutation (Lawrence and Ochman, 1997).

Recent estimates of the rate of LGT are also available. Lawrence and Ochman (1998) estimated that a minimum of 17.6% of open reading frames in *E. coli* had originated via LGT. The terminus region of the chromosome has a relatively higher number of laterally transferred genes. If deletions were kept in consideration, the rate of LGT is estimated to be 16 kb per million years per lineage based on studies on divergence of *E. coli* from *Salmonella* lineages (Lawrence and Ochman, 1998). These two organisms diverged approximately 100 million years ago. The laterally transferred genes were identified using biases in base pair composition and codon usage. In addition, chromosomal position was also used to support the conclusion that a gene is laterally transferred (Lawrence and

Ochman, 1998).

The high rates of LGT in prokaryotes suggest that evolution of prokaryotes is mainly due to LGT. It has been estimated that the rate of lateral gene transfer is as high as the rate of gene deletion (Hao and Golding, 2004). The stable sizes of genomes suggest that most newly acquired genes are lost before speciation. Moreover, there is a poor correlation between LGT and branch lengths if the branch lengths are estimated from 16S RNA data (Hao and Golding, 2004). The actual rate of insertions is higher than that estimated using bacterial genomes because orthologous replacement of genes cannot be detected easily. Also, the rate of gene insertions is higher than the rate of nucleotide substitution (Hao and Golding, 2006). LGT rates are lower at ancient branches, but higher at the tips of phylogeny (Hao and Golding, 2006). The rate of insertions is approximately five times greater than the nucleotide substitution in the *B. cereus* group even if ORFans, genes that are not found in any other organism, are excluded. The rate of LGT is very high in strains of the same species (Hao and Golding, 2006). The rate of LGT over different branches of a phylogenetic tree can also be studied. The rate of lateral gene transfer is estimated to be faster at the tips of phylogeny than at the ancient branches (Marri, Hao and Golding, 2007). Moreover, recently acquired genes have been shown to be under relaxed selection and evolve at a very rapid rate (Marri, Hao and Golding, 2007). Since the recently acquired genes also have a higher proportion of synonymous and non-synonymous substitutions as compared to other non-laterally transferred genes. Most of the genes that are laterally transferred have low GC content (Marri, Hao and Golding, 2007).

Among the genes that are transferred, informational genes are the least likely to transfer, while metabolic genes are the most likely to transfer (Hao and Golding, 2008b) but LGT of informational genes still can occur because there is evidence for lateral transfer of ribosomal RNA genes (Hao and Golding, 2008b). If informational genes are removed from the analysis, the rate variation increases slightly. However, if most conserved genes are removed, the estimates of rate variation increases significantly. The effect of removal of informational genes in estimates of rate variation is similar to the effect of removal of random genes. Therefore, the complexity hypothesis, which states that metabolic genes are readily transferred, cannot fully explain the rate variation of LGT (Hao and Golding, 2008b).

The high rates of LGT are not due to false diagnosis of gene truncation because a study involving different E-values of BLAST and match lengths showed that the rate of LGT is still high even at low match lengths (Hao and Golding, 2008a). However, unique genes, duplications, and orthologous replacements by LGT were ignored in the study (Hao and Golding, 2008a). The *B. cereus* group has a higher number of truncated genes than other *Bacillus* organisms (Hao and Golding, 2008a).

Translocated genes undergo lateral transfer at a higher rate (Hao and Golding, 2009). Moreover, the rate of translocation is higher in recently acquired genes. (Hao and Golding, 2009). A higher proportion of laterally transferred genes are found on the leading DNA strand than on the lagging strand, suggesting a strand bias in the transfer or retention of laterally transferred genes (Hao and Golding, 2009). Essential genes are also more likely to be found on leading DNA strand (Hao and Golding, 2009). However, genes that are

laterally transferred on the leading strand may not always be essential. The strand bias in lateral gene transfer may be explained by considering that prophages integrate genes by a mechanism that places the gene on the leading strand (Hao and Golding, 2009). The high rates of LGT suggest that bacteria are ‘sampling’ genes, and not acquiring the genes for the rest of their life (Ochman *et al.*, 2000).

Not all genes and genomes undergo LGT at the same rate. Popa *et al.* (2011) used directed graphs to show that metabolic genes are transferred at a higher frequency. Many edges of the directed graphs connect pathogens, donors and recipients suggesting high rates of lateral transfers. Moreover, directed graphs of some taxa are more connected than others suggesting unequal rates of LGT across taxa. For example, β and γ proteobacteria have highly connected directed graphs, suggesting a high frequency of LGT (Popa *et al.*, 2011).

The frequency and mechanism of LGT also depends on the environment of the bacteria (John, Aharon and Christopher, 2011). Research suggests that there are differences in the rate of LGT in bacteria that reside in a thermal environment to those that live in a saline environment. DNA is easily degraded in a thermal environment, but it is preserved in a saline environment (John, Aharon and Christopher, 2011). It was shown that extremophiles have higher rates of LGT suggesting that LGT is beneficial for the survival of bacteria in extreme environments (John, Aharon and Christopher, 2011). Natural selection has favoured the GC rich genome of halophilic bacteria perhaps because of a possible protection against thymine dimerization in saline environments (John, Aharon and Christopher, 2011). Donors of laterally transferred genes cannot always be identified in studies involving saline environments such as the dead sea possibly due to bias in genome sequencing (John, Aharon and Christopher, 2011).

The rate of LGT varies among lineages. For example, the rate of LGT in intracellular pathogens is low (Lawrence and Hendrickson, 2003). Genes that are not beneficial are not retained. Therefore, the genomes of intracellular pathogens are smaller than free living bacteria (Lawrence and Hendrickson, 2003). Some sequences are frequently found on one strand but not the other (Lawrence and Hendrickson, 2003). There is a selection against such sequences. For example, the octamer GGGCAGGG is counter selected from appearing on the wrong strand, suggesting a strand bias in LGT (Lawrence and Hendrickson, 2003).

1.4 Barriers to LGT in Bacteria

1.4.1 Exclusion of DNA

There are mechanisms that prevent LGT. It was observed with *E. coli* that if plasmid F is present, then *E. coli* is not able to receive genes via conjugation (Achtman *et al.*, 1980). This phenomenon is termed as surface exclusion. In case of the F plasmid, surface exclusion works by reducing the receptiveness to the F pilus and by preventing DNA entry into the cell (Achtman *et al.*, 1980). Another phenomenon that prevents LGT is restriction enzymes that degrade double stranded DNA in cells (Berndt, Meier and Wackernagel, 2003). Although, DNA enters a bacterium as single stranded DNA in most cases, the re-

striction enzymes system can still degrade the foreign DNA that has been converted to double stranded form. Apart from restriction enzymes, some plasmids do not have a broad host range (Thomas and Nielsen, 2005). The plasmids that are not able to replicate in the host cell are lost from the cell after a few cell divisions (Thomas and Nielsen, 2005).

1.4.2 CRISPR elements

Acquired immunity from viral DNA in prokaryotes is mediated by clustered regularly interspaced short palindromic repeats (CRISPRs) (Jansen *et al.*, 2002). These are conserved repeats of 24 bp to 47 bp separated by derived DNA sequences from foreign phages or plasmids (Mojica *et al.*, 2000). The foreign DNA is integrated into host genomes as spacers between CRISPR elements. After integration of the spacers, the host bacterium is resistant to phages having similar sequences (Heidelberg *et al.*, 2009). If these spacers are deleted, then the resistance towards these phages is lost. Moreover, mutations in these spacers can cause a loss in resistance (Vale and Little, 2010). The resistance to phage DNA is mediated by CRISPR-associated (Cas) proteins (Haft *et al.*, 2005). Cas proteins are DNA binding proteins that are found upstream of CRISPR elements. Their function is similar to RNAi proteins. However, Cas proteins only bind to DNA (Mojica *et al.*, 2009).

Coevolution of bacterial host and bacteriophage can be studied using CRISPRs. When phage infects a bacterium, the result is a CRISPR element in the genome of the bacterium (Barrangou *et al.*, 2007). If the mutated sequence of the bacterium is presented, then a new spacer is added. If the new spacer is maintained in the bacterial genome, then the order of sequences of CRISPR elements presents a history of phage infections in the bacterium (Barrangou *et al.*, 2007). However, if the spacers were continuously added to genomes, the genomes would have grown infinitely large suggesting that there is a cost associated with the acquisition of CRISPR mediated immunity (Lennon *et al.*, 2007). A large number of CRISPR elements in a genome increase the size of the genome and decreases the replication rate of DNA. Therefore, there is a selection against a large number of CRISPR elements in the genome. Moreover, this causes one to doubt the reliability of spacer elements' ability to record the past host-pathogen interactions (Lennon *et al.*, 2007).

1.4.3 Sequence Similarity

The frequency of LGT is correlated with the similarity of genome sequences of the donor and the recipient cells (Majewski and Cohan, 1998). If the genome sequences of the donor and the recipient are similar, the probability of integration of acquired DNA in the recipient cell is higher because of high probability of successful homologous recombination. The high rate of gene transfer in closely related bacteria can be explained by integration of genes within the recipient genome by homologous recombination (Marri, Hao and Golding, 2006). Distantly related bacteria usually do not have enough sequence similarity for homologous recombination. Thus, dissimilarity of DNA sequence serves as a barrier to LGT (Popa *et al.*, 2011). In *Bacillus*, homologous recombination requires two regions of similarity (Majewski and Cohan, 1999b). Whereas in *E. coli*, only one region of similarity

is enough (Lovett *et al.*, 2002). The requirement of sequence similarity can be bypassed by NHEJ (Aravind and Koonin, 2001).

1.4.4 Complexity Hypothesis

Complexity can also serve as a barrier to LGT (Wellner, Lurie and Gophna, 2007). Complexity hypothesis states that informational genes, which are involved in DNA replication, transcription, translation and multiple molecular pathways, are less susceptible to LGT (Jain, Rivera and Lake, 1999). Genes that are involved in functions such as DNA binding, pathogenicity and cell surface functions are transferred more often than other genes (Jain, Rivera and Lake, 1999). Informational genes can also get transferred but not at a very high rate (Jain, Rivera and Lake, 1999). Moreover, there is no evidence for the transfer of ribosomal genes between distant species (Andam, Williams and Gogarten, 2010). Genes that are highly connected in the biological pathways are more likely to be essential. In general, genes with low duplicability have smaller probability of LGT (Jain, Rivera and Lake, 1999). Besides barriers to LGT, there could also be barriers to retention of genes (Wellner, Lurie and Gophna, 2007). For example, if the native gene is a part of a protein complex, the laterally transferred homolog of the gene may not be retained in the bacterial population and can eventually be lost due to negative selection (Jain, Rivera and Lake, 1999). Similarly, if the gene has lesser number of interactions, it is transferred or retained in the recipient genome more often. Also, recently transferred genes show relatively fewer interactions within the protein network (Cohen, Gophna and Pupko, 2011).

1.5 Importance of LGT

1.5.1 Role in Evolution

In eukaryotes, many genes have been transferred from mitochondria and plastids during endosymbiotic events (Martin *et al.*, 2002). If plant hybridization is considered as a LGT event, then the role of LGT in evolution of plants is very significant (Seehausen, 2004). LGT is also important in evolution of archaea and bacteria (Rest and Mindell, 2003). Studies have shown that there is high frequency of interdomain gene transfer between bacteria and archaea (Rest and Mindell, 2003). However, other studies have shown that LGT is more frequent in closely related organisms than in distantly related organisms (Marri, Hao and Golding, 2006). Moreover, the detection of LGT is biased towards detection of recent transfers. Thus, the impact of ancient gene transfers event cannot be easily studied (Boto, 2010).

1.5.2 Role in Medicine

LGT is important in medical research because the antibiotic resistance genes spread from antibiotic resistant strains to antibiotic susceptible strains via LGT (Davies, 1996). It was recognized in 1960 that LGT was the cause of antibiotic resistant strains. The evolution of

antibiotics resistance is not *de novo* (Akiba *et al.*, 1960). Antibiotic resistance genes allow the bacteria to exploit environments that have a high concentration of antibiotics (Davies, 1996).

LGT is also responsible for evolution of virulence in bacteria that were not previously virulent (Sasakawa *et al.*, 1988). For example, *Shigella* and *Yersinia* has virulence plasmids that may have originated by LGT (Portnoy, Moseley and Falkow, 1981). Moreover, *E. coli* can also acquire virulence due to LGT of virulence plasmids (McDaniel and Kaper, 1997). The virulence genes flanked by short, direct repeats, similar to those that are related to integration of mobile genetic elements, suggests that they may have been transferred by LGT (Cheetham and Katz, 1995).

LGT also provides the recipient with genes required for metabolic functions (Lawrence and Roth, 1996). The transfer of genes that code for metabolic functions allows the bacteria to exploit unique niches. Moreover, gene clusters and operons in some metabolic pathways can also be transferred by LGT (Lawrence and Roth, 1996). The complexity hypothesis predicts that if an informational gene and a non- informational gene get transferred into bacteria (Jain, Rivera and Lake, 1999). The product of non-informational genes, such as those that carry a metabolic function, are more likely to be retained and utilized by the bacteria (Jain *et al.*, 2002).

1.5.3 Importance in Phylogenetics

Based on high frequency of LGT, it is not possible to infer a single universal phylogeny. Moreover, accuracy of the phylogeny depends on the method of phylogenetic inference (Daubin, Moran and Ochman, 2003). It is difficult to distinguish between LGT and gene duplication when distant lineages are studied. (Daubin, Moran and Ochman, 2003). Accurate inference of phylogenetic tree requires conserved genes that are vertically inherited (Wolf *et al.*, 2002). Despite high rates of LGT, data suggests that there is a conserved set of genes that is inherited vertically and retains the true phylogenetic signal. These genes can be used to construct phylogenetic trees (Wolf *et al.*, 2002). Vertical inheritance with periodic selection is still important in evolution of microbial genomes (Lawrence and Hendrickson, 2003).

There are many barriers to the lateral transfer of core genes between organisms, making the impact of LGT not significant when inferring a phylogenetic tree (Kurland, Canback and Berg, 2003). The pan-genome concept can be applied to study genes in bacterial genomes (Medini *et al.*, 2005). It was discovered that 8 per cent of the genes in a typical bacterial genome can be found in 99 per cent of the sequenced genomes and can be considered core genes. Therefore, these genes can be used to infer a universal phylogenetic tree. However, the core genes in the bacteria can still be replaced by laterally transferred genes (Lawrence and Hendrickson, 2005). It is estimated that all bacteria share approximately 100 to 150 core genes, whereas all prokaryotes share only 30 to 50 genes (Charlebois and Doolittle, 2004). The number of core genes increases if the number of organisms is reduced. Organisms that have low frequencies of LGT have a smaller pan-genome size but a larger number of core genes in common (Charlebois and Doolittle, 2004).

Gene transfer between closely related strains of bacteria can reinforce phylogenetic signal (Andam and Gogarten, 2011). LGT between closely related bacteria does not cause conflicting tree topologies (Andam and Gogarten, 2011). Therefore, phylogenetic tree may still be a reliable way to understand the evolution of prokaryotes. In some cases in which a tree cannot be fully resolved due to reasons such as very low sequence diversity, bipartitions can be studied (Zhaxybayeva, Lapierre and Gogarten, 2004). Moreover, the number of bipartitions may be smaller than the number of topologies analyzed. Therefore, it speeds up the analysis (Zhaxybayeva, Lapierre and Gogarten, 2004).

1.5.4 Profile of Laterally Transferred DNA

Most studies about LGT assume that a gene is the basic unit of lateral transfer. However, recent evidence suggests that this may not always be the case. Chan *et al.* (2009) showed that protein domains can also be independently transferred and get integrated into genomes by homologous recombination. The recombination breakpoints within the protein coding regions were discovered (Chan *et al.*, 2009), and it was concluded that the recombination breakpoints were not uniformly distributed. They lie outside small protein domains suggesting that protein domains were transferred and integrated into the recipient genome (Chan *et al.*, 2009).

Some genes get transferred more often than the others. One explanation of the bias in transfer of genes is called the complexity hypothesis, which is discussed in section 1.4.4 (Jain, Rivera and Lake, 1999). Another explanation is the selfish operon theory (Lawrence and Roth, 1996). Selfish operon theory suggests that laterally transferred operons are beneficial to bacteria because they can contribute to metabolic pathways that code for traits such as antibiotic resistance (Lawrence and Roth, 1996). Most genes that are recently transferred are ORFans (Daubin and Ochman, 2004). They have high AT content and codon usage that is similar to phages and plasmids (Daubin *et al.*, 2003). Most of the laterally transferred genes have a high turnover rate. Beneficial genes can be fixed in populations by selective sweeps (Majewski and Cohan, 1999a).

Most of the gene transfer between bacteria is between closely related bacteria as discussed in section 1.4.3 (Marri, Hao and Golding, 2006). In some cases, LGT can serve as a genetic life raft and genes can be saved from extinction by LGT (Gogarten, Fournier and Zhaxybayeva, 2008). For example, if a bacterial lineage is dying and some of its genes can be saved by LGT via mechanisms such as transposition. The genes can get transferred to different lineages of bacteria and integrate via nonhomologous end joining. Thus, they are prevented from extinction (Andam and Gogarten, 2011).

Some biological processes are carried out by the interaction of several proteins (Gophna and Ofran, 2011). If a single gene is transferred, it may not retain its normal function. Therefore, genes that have comparatively fewer interactions are more likely to be transferred and retained in new species (Cohen, Gophna and Pupko, 2011). Proteins that are involved in many protein-protein interactions have a higher number of interaction sites (Cohen, Gophna and Pupko, 2011). If the protein has multiple interaction sites, then it is more likely to be retained in the recipient genome and form new interactions (Gophna and

Ofran, 2011). Therefore, connectivity of a gene in metabolic pathways is a very important factor in LGT.

1.5.5 Speciation

It is very difficult to neatly classify bacteria into species. However, bacteria can be classified into species based on ecological divergence or barriers to recombination (Barracough, Balbi and Ellis, 2012). A suggested DNA similarity threshold for speciation in *Bacillus* is 1% similarity in 16S rRNA sequences (Schloss and Handelsman, 2005), and 6% similarity for protein coding genes (Venter *et al.*, 2004). The problem with this rRNA approach is that it relies on a single gene. Based on recombination analysis, species arise when recombination is limited. Moreover, speciation can also occur if there is geographic or ecological isolation between bacteria (Cohan, 2006). However, geographic or ecological isolation is not common in bacteria due to their high dispersal rate (Whitaker and Banfield, 2006). LGT can expand the niche of bacterial and counter the effects of geographic and ecological isolation (Barracough, Balbi and Ellis, 2012). Speciation in bacteria is still an open question.

1.6 LGT and Eukaryotes

There has been a lot of focus on LGT in prokaryotes. However, it does not necessarily mean that there are no examples of LGT in eukaryotes (Gogarten, 2003). There are examples of bacterial genes that are found in diplomonads (Andersson *et al.*, 2003). There are also some examples of genes from *Wolbachia*, an intracellular pathogen, in their host insects (Hotopp *et al.*, 2007). There are some cases of lateral transfer of homing endonuclease genes in fungi (Koufopanou, Goddard and Burt, 2002). The importance of LGT in the evolution of eukaryotes cannot be ignored.

There is evidence for LGT in fungi, but the mechanisms of LGT are not fully understood (Garcia-Vallve, Romeu and Palau, 2000). The transfers could be due to anastomosis, a process in which germline cells of fungi fuse (Roca *et al.*, 2004). The presence of introns in eukaryotes, including fungi, decreases the error of detection of LGT in fungi because if a gene is acquired from bacteria then it does not have introns and, therefore, should easily be detected (Kondrashov *et al.*, 2006). Moreover, intron splicing serves as a barrier to transfer of genes from other eukaryotes to fungi because other eukaryotes have different mechanisms of gene splicing, which are not necessarily present in fungi (Fitzpatrick, 2012). The rates of gene transfer in fungi are lower than that in prokaryotes and should not affect the inference of fungal tree of life (Fitzpatrick, 2012).

It was believed that Bdelloid rotifers were asexual and evolved by mutations only, which could not explain their diversity. Recently, it was discovered that they have bacterial genes in their genomes, suggesting that there might have been high rates of LGT in the evolutionary history of Bdelloid rotifers (Gladyshev, Meselson and Arkhipova, 2008). The laterally transferred genes were identified using alien index (AI). Most of these genes are non functional and are located in telomeric regions (Gladyshev, Meselson and Arkhipova, 2008).

The observed pattern of gene distribution cannot be explained by vertical inheritance and multiple gene losses. Therefore, LGT is the most supported phenomenon to explain the gene distribution in *B. rotifers* (Gladyshev, Meselson and Arkhipova, 2008).

If LGT is assumed to be the transfer of DNA from one cell to another, then LGT also plays a role in transmission of cancer and chemotherapy resistance (Holmgren, 2010). When tumor cells die, they leave behind fragmented DNA. The DNA can be acquired by other cells and hence the resistance to chemotherapy can be transferred (Holmgren, 2010).

In the past few decades, LGT has sparked a lot of interest among researchers. LGT and speciation has drawn parallels with mythology of ancient Greeks. An interesting analogy is that of the ship of Theseus paradox (Zhaxybayeva, Lapierre and Gogarten, 2004). Similar to the ship of Theseus, in which the old lumber of the ship was replaced by the new lumber, and it was argued whether the ship was the same or not, it can also be argued that if most genes of a species are replaced by genes from other species, then is it the same species or some other species (Zhaxybayeva, Lapierre and Gogarten, 2004)? It also questions the phylogenetics placement of the species if it is argued that the species doesn't stay the same after massive lateral transfers (Zhaxybayeva, Lapierre and Gogarten, 2004).

Part II

LATERAL GENE TRANSFER

Chapter 2

Insertions of Genes in Operons

2.1 Abstract

Multiple genes that are transcribed as a single mRNA are known as operons. Operons can be detected by studying distances between genes, conservation of genes order across different taxa, functional relations of genes, and experimental evidence. Genes can be integrated into genomes by homologous recombination and nonhomologous end joining. The rate of homologous recombination is higher than that of nonhomologous end joining. Homologous recombination can be detected by permutation tests, conflicts in phylogenetic topology, differences in phylogenetic compatibility, and Bayesian methods.

In this research, we propose that homologous recombination is the mechanism of integration of laterally transferred genes into operons. To test this, orthologs of conserved genes in *Bacillus* were searched from the genomes. The genes were aligned in MUSCLE. Phylogenetic trees were inferred in MrBayes. Operon structure was inferred using OperonDB, which inferred operons based on the degree of conservation of gene order across different taxa. Genes were clustered using an MCL algorithm. Duplicated genes were removed from the analysis. Genes in clusters were represented by binary digits, and ancestral states were inferred using parsimony. LGT was inferred using the ancestral states. Homologous recombination was detected using permutation algorithms. The algorithms used were GENECONV and maximum chi square. GENECONV was used with default values. Stepwise implementation of maximum chi square was used with three different lengths for half-windows sizes. The results indicated that there were approximately ten percent of genes laterally transferred into *Bacillus*. Out of those, nearly ten percent of genes were laterally transferred into operons. Results of GENECONV indicated that there were several transfers of genes originating from within *Bacillus* species and transferring into other *Bacillus*. There were a few genes where the source of recombination could not be identified. Results of the maximum chi square algorithm indicated that there was at least one laterally transferred gene in which recombination breakpoints were detected before the start codon and after the stop codon of the gene. The results indicated that genes were transferred from a non-*Bacillus* organism to a *Bacillus* organism. A similar case was detected using two steps of stepwise program that implements maximum chi square algorithm. The

annotations indicated that most genes that were transferred into operons had biological function. Overall, the results indicated that biologically functional genes can be integrated into operons by the process of homologous recombination. The evidence may be difficult to find due to mutations and amelioration.

2.2 Introduction

2.2.1 Evolution of Operons

Previously, it was shown that there are laterally transferred genes in prokaryotic operons (Price, Arkin and Alm, 2006). However, the mechanism of transfer of genes into operons is not known (Price, Arkin and Alm, 2006). In this research, it is proposed that homologous recombination is the mechanism of integration of laterally transferred genes into operons. This study focuses on transfer of biologically functional genes from a distant organism into operons by LGT followed by homologous recombination.

Jacob *et al.* (1960) coined the term operon to describe a group of genes that were regulated by a single operator based on research conducted on metabolism of lactose in *Escherichia coli* and the synthesis of β -galactosidase and galactoside permease. Nearly 50 years later, operons still fascinate researchers and are redefined as a group of genes that are transcribed into a single mRNA (Price, Arkin and Alm, 2006). They usually encode proteins in the same biochemical or functional pathway, but that is not always the case (Itoh *et al.*, 1999). Approximately half of all protein coding genes in prokaryotes are found in operons (Price, Arkin and Alm, 2006). Genes in operons are separated by less than 50 base pairs, with the peaks of the distribution at -4, -1 and 10 base pairs of intergenic distance. The negative distance indicates an overlap of genes in operons (Salgado *et al.*, 2000). Operons may have originated in thermophilic organisms because the close spacing of genes facilitates the interaction and coexpression of genes in the same metabolic pathway and protects the proteins from thermal degradation (Glansdorff, 1999).

New operons form mainly by rearrangements, deletions and less importantly by LGT (Price, Arkin and Alm, 2006). Rearrangements can move genes close to each other and can result in the formation of operons (Itoh *et al.*, 1999). If two genes are physically close to each other, then a part of the intergenic DNA can be deleted, and an operon can form (Lawrence, 1999b). However, strong evidence for operon formation by deletion is lacking (Lawrence, 1997). Lateral transfers can place a gene close to another gene. The process does not require deletion of a large part of intergenic region for operon formation (Lawrence, 1997). If a foreign promoterless gene is integrated adjacent to a native gene, a two-gene operon forms. The relative age of foreign genes has been calculated and compared to the age of operons (Price, Arkin and Alm, 2006). It was found that the relative age of the operon was nearly equal to the age of the foreign gene. The results indicate that operons can form as a result of LGT. In *E. coli*, it was discovered that the foreign gene is often downstream of the native gene. Therefore, the foreign gene can be transcribed from the promoter of the native gene. In *Bacillus subtilis*, there was no significant preference for the foreign gene to be downstream (Price, Arkin and Alm, 2006).

Operons can also be formed by deletion of genes (Lawrence, 1999b). If the intergenic region and the promoter regions of a gene are deleted, then two genes will be transcribed into one mRNA. This leads to the formation of a new operon given that both genes are on the same strand. Deletion of genes can be inferred by parsimony (Fitch, 1971). Another mechanism of formation of operons is rearrangement (Itoh *et al.*, 1999). Gene rearrangements can place a gene adjacent to another gene so that both genes can be transcribed into a single mRNA (Itoh *et al.*, 1999).

Often the start codon of a gene overlaps with the stop codon of the upstream gene (Eyre-Walker, 1995). The genes may be closely placed to keep the size of genome small. There may also be translational coupling in which ribosome can move from the stop codon of the upstream gene to the start codon of the downstream gene (Mira, Ochman and Moran, 2001). The overlaps of one and four bases are so common that they are called “canonical” overlaps. These spacings are not due to errors in annotation. If they were due to errors in annotation, then the spacing would be a multiple of three, and that is not true in most cases (Price, Arkin and Alm, 2006). The codon adaptation index (CAI) can be used to investigate the relationship between expression level of genes and intergenic distances (Sharp and Li, 1987). CAI value is correlated to the gene expression level. Based on CAI values, it was discovered that highly expressed genes have large intergenic distances between them (Eyre-Walker, 1995). One explanation of highly expressed genes not terminating closely to the genes downstream is that the arrangement will avoid the interference of ribosome at the end of the upstream gene with the ribosome at the start of downstream gene (Eyre-Walker, 1995).

There are other factors involved in the determination of intergenic distance between genes because the correlation between the CAI value of the downstream gene and intergenic distance is weak. Eyre-Walker (1995) suggested the cotranscription of genes will eliminate the need for transcription control elements. Translational control would regulate the distance between genes. Genes that have an overlap of stop codon of the downstream gene with the start codon of the upstream gene may have an evolutionary advantage due to selection for small genome size. For example if there were a mutation in the stop codon of the upstream gene, then the stop codon of the downstream gene would still be functional given that it is inframe (Eyre-Walker, 1995). Microarray data was used to study gene spacing and the rate of gene expression in *E. coli* and *B. subtilis*. It was found that highly expressed operons have distances more than 20 base pairs between genes (Price, Arkin and Alm, 2006). Microarray data are more accurate than CAI values because microarray data measures the amount of mRNA.

Itoh *et al.* (1999) suggested that the structure of operons should be conserved in the course of evolution. There are approximately 100 operons common between *B. subtilis* and *E. coli*. Genes in an operon can shuffle randomly in the genomes. Operons are lost by deletion of one to many genes or inversions or insertions between the genes (Price, Arkin and Alm, 2006). For example, the *yebI-yebL* operon of *E. coli* could possibly have been lost by inversion (Price, Arkin and Alm, 2006). Operons are also lost by the replacement of a gene in an operon by LGT. For example, in the case of *ribE2* gene, a copy of *ribE2* gene was transferred into the *E. coli* genome and the native *E. coli* gene was deleted (Vitreschak *et al.*,

2002). The functional pathway coded by operons can influence the survival of operons. Operons consisting of genes involved in energy production have higher rates of survival than those that are involved in amino acid transport and metabolism (Price, Arkin and Alm, 2006).

2.2.2 Explanations for Operon Formation

When the first genetic maps were constructed, it was observed that genes with similar functions were clustered together (Lawrence, 1999b). Here, a cluster of genes means that the physical distance between the genes, measured in number of base pairs, is small as opposed to a protein cluster in which genes are related by function or sequence similarity scores. The nonrandom clustering of genes can be explained by any one of five models of evolution of gene clusters (Lawrence, 1999b). These include the natal model, the Fisher model, the molarity model, the coregulation model, and the selfish operon model (Lawrence, 1999b).

The simplest of the models to explain operons is the natal model. According to the natal model, genes may be found in clusters because clusters originate by duplication and divergence (Lawrence, 1999b). This model can explain some clusters in eukaryotes. However, it cannot explain the existence of operons in prokaryotic genomes because clustered genes in prokaryotes often code for distant families of proteins, suggesting that operons were formed from the genes that were not closely related (Orengo *et al.*, 1993). Another problem with the natal model is that there is a lack of pairing of duplicated genes in prokaryotic genomes (Dandekar *et al.*, 1998). Most duplicated genes do not cluster as pairs in bacteria; hence gene clusters could not have been originated by ancient duplications (Dandekar *et al.*, 1998). However in eukaryotes, genes may be physically clustered due to duplications. Some examples include the *hox* cluster and the globin cluster (Lawrence, 2002b).

According to the Fisher model, coadapted alleles can be placed in proximity. The rate of deleterious recombination events would decrease if the genes were in clusters (Fisher, 1930). This model is applicable to populations that are very variable and have high rates of recombination (Lawrence, 1999b). However, gene clusters in most bacterial lineages do not support this model because there is no evidence for coadapted alleles other than those that physically interact. There are suggestive recombination events that moved some genes into proximity with each other and dispersed other genes (Lawrence, 1999b).

There is growing evidence suggesting spatial segregation of proteins in cytoplasm (Losick and Shapiro, 1999). If genes are in close physical proximity, high levels of local concentrations of proteins can be produced in bacteria as suggested by the molarity model (Lawrence and Ochman, 1998). Imbalance in the concentrations of proteins can also be prevented (Pál and Hurst, 2004). However, the proximity of genes does not imply that the genes are necessarily in operons. Therefore, the molarity model can explain neither the formation nor the distribution of genes in operons (Lawrence and Ochman, 1998). There is also evidence for functionally unrelated genes to exist in the same operon (Itoh *et al.*, 1999).

The coregulation model suggests that genes are in clusters because there is selection for coregulation from a single promoter (Lawrence, 1999b). The model requires close physical

proximity and cotranscription of previously linked genes. Existence of unlinked genes in operons cause difficulty in operon formation by coregulation (Itoh *et al.*, 1999). This model explains the maintenance of operons in which genes are close to each other and are transcribed into a single mRNA. Formation of operons can be explained by considering that formation and organization of operons reduces regulatory sequences (Price *et al.*, 2005). Therefore, the evolution of operons is more probable than the evolution of independent regulatory sequences when the regulation of genes is complex (Price *et al.*, 2005).

The selfish operon model suggests that gene clusters are beneficial to themselves because they can spread by both vertical inheritance and LGT, while single genes may not function independently. Transfer of a cluster of genes that confer a selectable function is maintained (Lawrence, 1999b). The selfish operon model implies that the close arrangement of genes is selfish as it facilitates the transfer of function when genes acting in the same pathway get transferred as a unit (Lawrence, 1999b). It is implied that metabolic and nutrient transport genes should be in operons but not informational genes (Lawrence, 1999b). The transfer of an operon or a gene cluster implies that phenotypes could be transferred. Hence, genes that code for proteins in the same pathway should be in the same operon or gene cluster (Lawrence, 1999b).

The organization of genes in clusters can be explained by alternate hypotheses. For example, the origin of ordered clusters is related to the timing of expression of genes (Ussery *et al.*, 2001). The clusters may allow for transcription and translation coupling of products in a pathway. The gene clusters can prevent the imbalance of protein concentration (Pál and Hurst, 2004). Genes that form parts of protein complexes can also be found in clusters and operons. If that is the case, then more genes should be clustered together and found in operons (Pál and Hurst, 2004).

2.2.3 Prediction of Operons

Large scale sequencing of genomes has made possible the computational prediction of operons. There are at least five different types of computational methods to predict operon structure (Brouwer, Kuipers and van Hijum, 2008). These include intergenic distance methods, conserved gene clusters methods, relation of function methods, sequence elements and experimental evidence based methods (Brouwer, Kuipers and van Hijum, 2008).

The intergenic distance methods can be used to predict whether the genes are in operons (Salgado *et al.*, 2000). If a set of genes is a part of the same operon, then the intergenic distance between the genes is smaller than the intergenic distance between the genes that are not in operons (Brouwer, Kuipers and van Hijum, 2008). In *E. coli*, if the intergenic distances between the genes that are transcribed in the same direction are studied, then a significant number of genes show a short distance of approximately 10 base pairs (Salgado *et al.*, 2000). Operon prediction can be refined based on functional classification of the genes in operons assuming that genes in the same operons have related functions. Based on intergenic distance and functional classes (Riley, 1993), approximately 630 to 700 operons in *E. coli* are found (Salgado *et al.*, 2000). The intergenic distance between highly expressed genes is large even if they are found in operons (Eyre-Walker, 1995). Therefore,

this method is not the most accurate for prediction of operons.

In 1997, the sequencing of prokaryotic genomes revealed that some genes in prokaryotes are conserved in gene order on chromosome (Siefert *et al.*, 1997). At least 16 clusters of genes were conserved in gene order in four organisms that were sequenced (Siefert *et al.*, 1997). It can be argued that genes conserved in synteny are in operons (Ermolaeva, White and Salzberg, 2001). The order of genes in ribosomal protein operons is well conserved. If the order of a gene cluster is conserved in distantly related organisms, then it can be concluded, with high probability, that the conserved gene cluster is indeed an operon (Ermolaeva, White and Salzberg, 2001). In some cases, the gene order is not conserved. For example, Itoh *et al.* (1999) has shown that the order of some genes within the operons of *E. coli* and *B. subtilis*, compared to 11 other complete genome sequences, is not conserved. LGT and rearrangements within the operons can also be taken into account when predicting operons (Perteza *et al.*, 2009).

Often, genes that are found in operons have related functions (Brouwer, Kuipers and van Hijum, 2008). Their products may be a part of the same protein complex or the same biochemical pathway. Riley (1993) has extensively documented the functions of genes in *E. coli*. The functional pathways can be represented by a graph that can be transversed to produce sub graphs of closely interacting proteins (Zheng *et al.*, 2002). These sub graphs can be used to predict operons in microbial genomes (Zheng *et al.*, 2002). It is also interesting to note that the change of GC content is higher at the boundary of operons than within the operons (Zheng *et al.*, 2002). These results suggest lateral transfer of operons (Lawrence and Roth, 1996). Cluster of orthologous groups (COG) database is a classification of proteins in families based on similarity in sequence and function (Tatusov, Koonin and Lipman, 1997). Gene ontologies (GO) database show well defined relations between genes using common vocabulary (Ashburner *et al.*, 2000). It can be used to improve operon prediction (Brouwer, Kuipers and van Hijum, 2008). Operon prediction methods can use functional relations based on Riley's functional annotation, metabolic pathways, COG and GO to predict operon structure (Brouwer, Kuipers and van Hijum, 2008).

Genomic features such as promoter sequences, transcription factor binding sites, certain DNA motifs and transcription terminators can be used to predict the start and the end of operons (Brouwer, Kuipers and van Hijum, 2008). Transcription factor binding sites (TFBS), commonly detected using position specific weight matrices, can be used to predict operons (Emma, Khushwant and Simon, 2008). However, it assumes that there are no TFBS within operons (Emma, Khushwant and Simon, 2008). The upstream region of a gene that is not in an operon is different from the upstream regions of a gene that is in an operon (Janga *et al.*, 2006). In *E. coli*, the region for binding of sigma 70 can be studied along with characteristics of the promoter region such as low melting temperature and high AT content. A weight matrix for the region -35 to -10 can be used to study density of promoter signal in the region -200 to -10 base pairs upstream the genes (Janga *et al.*, 2006). The resulting operon prediction is more accurate for *E. coli* than for *B. subtilis* because *B. subtilis* has more sigma factors than *E. coli*. Hence, *B. subtilis* requires more sophisticated analysis (Janga *et al.*, 2006). Genome based features can also be used to improve the accuracy of the operon prediction. For example, operon prediction can be made using gene

conservation and intergenic distance between genes and similarity of CAI can be used to improve the accuracy of prediction (Price *et al.*, 2005). Genomic features in the operons such as TTTT motifs in intergenic regions of operons can also be used to improve the operon prediction when it is used with decision tree based classifiers and linear classifiers (Dam *et al.*, 2007).

Gene expression data can be used to predict which genes are in operons (Brouwer, Kuipers and van Hijum, 2008). Based on data from DNA microarrays, genes that are coexpressed could be in operons (Sabatti *et al.*, 2002). However, DNA microarray data are not available for all organisms. Moreover, operons cannot necessarily be predicted with high sensitivity using microarray data (Sabatti *et al.*, 2002). There could be secondary promoters within operons. Transcriptional attenuation and mRNA degradation can also bias the results. These would result in inaccuracy due to measurement of signals from the microarray data (Sabatti *et al.*, 2002). Moreover, multiple microarray experiments (such as change of media or heat shock) are required to study the expression of genes (Sabatti *et al.*, 2002). However, microarray technology can be used to improved computational prediction of operons by using intergenic distance and coexpression data to predict operons (Roback *et al.*, 2007).

Methods based on statistics and machine learning theory can also be used to predict operons. Support vector machines (SVMs) are data classification algorithms that use statistical learning (Zhang *et al.*, 2006). The prediction is based on the observation that genes within operons share common features. Prediction works by clustering genes that are close to each other and are functionally related (Zhang *et al.*, 2006). The prediction uses intergenic distances, the number of common pathways, the number of conserved pairs, phylogenetic profiles and the number of domains interactions in a SVM. The result is an accurate prediction of gene pairs within operons and transcription boundaries (Zhang *et al.*, 2006).

The performance and accuracy of computational methods of operon prediction can be measured (Brouwer, Kuipers and van Hijum, 2008). Experimentally verified collections of operons for *E. coli* and *B. subtilis* are available. The accuracy of operon prediction algorithms can be compared by verifying the predictions against experimentally verified databases (Brouwer, Kuipers and van Hijum, 2008). However, there are some potential problems with this method of calculating the accuracy of operon prediction. For example, in case of *B. subtilis*, three different experimentally verified collections of operon databases are available (Brouwer, Kuipers and van Hijum, 2008). These are Itoh collection, operon database (ODB) and the database of transcriptional regulation in *Bacillus subtilis* (DBTBS). The results in these databases may differ leading to inaccurate estimate of performance of computational methods (Brouwer, Kuipers and van Hijum, 2008). Moreover, the performance and accuracy of operon prediction can be estimated by different methods. These include predictions of gene pairs within operons and boundaries of transcriptional units. It is interesting to note that operons boundaries are less accurately predicted than gene pairs within the operons (Brouwer, Kuipers and van Hijum, 2008).

2.2.4 Homologous Recombination in Bacteria

Homologous recombination is common in bacteria (Didelot and Maiden, 2010). Although genetic material is frequently replaced and exchanged between closely related species of bacteria, the results of this process cannot be directly observed. If the bacteria that exchanged DNA are closely related, then the recombinant region can be detected. An important correlate of homologous recombination is DNA repair, in which damaged DNA is replaced by new recombinant DNA (Denamur *et al.*, 2000).

The rates of homologous recombination cannot be directly calculated without the knowledge of evolutionary history of the organism being studied (Feil *et al.*, 2001). Relative rates of homologous recombination can be inferred based on the rate of mutations. Statistics can be calculated from the genome sequences and can be used to infer the frequency of homologous recombination. The relative rates of homologous recombination can be inferred using linkage disequilibrium as a summary statistic (Perez-Losada *et al.*, 2006). For example in *Neisseria gonorrhoeae*, homologous recombination is 29.3 times faster than the mutations (Perez-Losada *et al.*, 2006). Often, different studies on relative rates of homologous recombination present conflicting data because of the method being used to estimate the rate of recombination (Didelot and Maiden, 2010). The accuracy of estimation of recombination rate depends on genetic diversity (Posada, Crandall and Holmes, 2002). Moreover, the rate of recombination can be estimated by simulating recombination on a coalescent tree (Stumpf and McVean, 2003). Rate of recombination cannot always be estimated because the effective population size is rarely fixed. In case of bacteria, the effective population size is very difficult to determine accurately (Stumpf and McVean, 2003).

Rate of recombination is not constant across different species of bacteria (Vos, 2009). The bacteria that are always competent show a high rates of recombination. Some bacteria, such as *E. coli*, *Neisseria meningitidis*, *Staphylococcus aureus*, *Streptococcus pneumoniae* and *Streptococcus pyogenes*, show higher rates of recombination than other bacteria (Posada, Crandall and Holmes, 2002). It is interesting to note that the large central region on the linear chromosome of *Streptomyces* is conserved while the terminal regions have relatively high rates of recombination (Doroghazi and Buckley, 2010). The rate of recombination is high within species in contrast to recombination from out groups into *Streptomyces*. This may be due to the high level of divergence between *Streptomyces* and other taxa (Doroghazi and Buckley, 2010). Genes from distant organisms are lost unless they are advantageous to the host (Doroghazi and Buckley, 2010).

DNA damage and double stranded breaks can be introduced in bacteria by ionizing radiation (Shinohara and Ogawa, 1995). The mechanisms for repair of broken DNA are slightly different in *E. coli* and *B. subtilis*. However, both involve homologous recombination (Shinohara and Ogawa, 1995). In both organisms, the RecA protein searches for homology regions whereas RecBCD functions as a helicase and an exonuclease. In *E. coli*, double-stranded breaks induces SOS response in the majority of cells in the population. In *B. subtilis*, SOS response is initiated only in a subpopulation (Simmons *et al.*, 2009). In both bacteria, SOS response is controlled by LexA repressor. *RecN* is highly expressed in SOS response of *E. coli*, whereas *RecN* expression in *B. subtilis* seems independent of SOS

response (Simmons *et al.*, 2009). Although, homologous recombination is a very efficient pathway to repair broken DNA, yet *B. subtilis* is also capable of repairing DNA by NHEJ, but the NHEJ pathway only repairs double stranded breaks during outgrowth of spores or in stationary phase cells (Simmons *et al.*, 2009). Therefore, NHEJ is not a common pathway to DNA repair in *B. subtilis*.

Unless there is positive selection, large scale DNA transfer is not frequently observed (Didelot and Maiden, 2010). The number of base pairs that are integrated by homologous recombination ranges from a few hundred base pairs to a few thousand base pairs. However, mosaic genes, formed due to recombination of short fragments of genes, have also been observed in bacterial genomes (Didelot and Maiden, 2010). Recombination is high in regions of genomes that are under positive selection (Didelot and Maiden, 2010). However, it cannot be concluded that recombination only takes place in regions under positive selection because recombination in regions under negative selection cannot be detected due to the loss of the recombinant region (Didelot and Maiden, 2010).

Recombination also depends on the physical distance between the bacteria. If the bacteria are physically close to each other, then DNA can transfer between the two bacteria and integrates into the genome of the recipient (Vos, 2009). Secondly, homologous recombination requires sequence similarity between the DNA of the host and the recipient (Majewski and Cohan, 1999b). There can also be negative selection against DNA that is integrated into the genome by recombination. Due to the negative selection, recombinant DNA can be removed from the population of bacteria (Didelot and Maiden, 2010).

2.2.5 Detection of Homologous Recombination in Bacteria

There are many different methods to detect homologous recombination in genomes (Posada, Crandall and Holmes, 2002). These include distance methods, phylogenetic methods, compatibility methods and substitution methods. The performance of all these methods depends on a given data set (Posada, Crandall and Holmes, 2002). It is difficult to detect recombination in highly similar or highly divergent sequences. Variation in the rates of evolution of different sites can also bias the results (Posada, Crandall and Holmes, 2002). Recent recombination events are easier to detect than ancient recombination because of mutations obscuring the recombinant region (Posada, Crandall and Holmes, 2002).

Recombination can be detected by DNA distance methods (Weiller, 1998). Recombination breakpoints can be detected by studying the change in the correlation between the DNA distance before the recombination breakpoint and after the recombination breakpoint. A C++ program PhylPro implements this algorithm (Weiller, 1998). A sliding window calculates the DNA distance upstream of a potential recombination breakpoint, while another window calculates the distance downstream of the breakpoint. If there is low correlation between the distances, then the results suggest that the potential site is a recombination breakpoint (Weiller, 1998). Another method that uses DNA distances and phylogenetic trees is implemented in TOPAL 2.0 (McGuire and Wright, 2000). This algorithm calculates pairwise distances in each half of the sliding window and builds a matrix of distances. The matrix is normalized and trees are inferred using least sum of squares. The difference

in least sum of squares between the two halves of a window is the test statistic (McGuire and Wright, 2000). If the difference in least sum of squares is relatively high, then recombination breakpoints are detected. The statistical significance can be verified by parametric bootstrapping (McGuire and Wright, 2000).

In the inference of a phylogenetic tree, it is assumed that the region used for inference is not a recombinant region. If the recombinant region is used for inference of a tree, then the inference would not be accurate (Posada, Crandall and Holmes, 2002). However, this property can be used to detect recombination. The recombinant region produces a phylogenetic tree that conflicts with the species tree (Posada, Crandall and Holmes, 2002). The trees generated using the recombinant region conflict in topology with the tree generated using the entire alignment. In contrast, the trees generated using the nonrecombinant regions, may not conflict with the tree generated using the entire alignment (Grassly and Holmes, 1997). These methods cannot accurately distinguish between recombination and convergent evolution (Grassly and Holmes, 1997). Low rates of evolution and a small window size for phylogenetic trees can also cause inaccuracies in detection of recombination. The C program PLATO implements this algorithm (Grassly and Holmes, 1997). A method implemented in the recombination detection program (RDP) makes a UPGMA tree to select reference sequences, and then uses it to infer recombination by selecting three sequences. Two sequences, A and B, are closely related to each other (Martin and Rybicki, 2000). If the probability that A or B is closely related to the third sequence C, based on identity scores from sliding windows, is low then there may be a recombination breakpoint (Martin and Rybicki, 2000). Inaccuracy occurs if the reference sequence cannot be accurately and unambiguously identified (Martin and Rybicki, 2000). Another program, RECPARS, can detect recombination breakpoints using phylogenetic methods (Hein, 1993). RECPARS infers phylogenies of different segments, and infer recombination by calculating a change in topology between different segments. Most parsimonious state sequence is searched, and a change in topology of tree suggests a recombination breakpoint (Hein, 1993).

Compatibility methods can also be used to detect recombination breakpoints (Jakobsen and Easteal, 1996). Sequences are aligned, and uninformative sites are removed. A compatibility matrix is constructed where each element represents phylogenetic compatibility or a lack of phylogenetic compatibility (Jakobsen and Easteal, 1996). The recombinant region has sites that are compatible with other sites within the region and not compatible with sites outside the region (Jakobsen and Easteal, 1996). The regions can be visually inspected on a compatibility matrix. Monte Carlo randomization can be used to show that the resulting observation is not due to chance (Jakobsen and Easteal, 1996). However, these compatibility methods require at least four sequences in the alignment (Jakobsen and Easteal, 1996). The method of recombination detection by compatibility is implemented in the C program Partimatrix (Jakobsen, Wilson and Easteal, 1997).

Permutation tests can be used to detect recombinant regions. In these methods, polymorphic sites are identified in an alignment (Sawyer, 1989). Sequences are considered pairwise and the bases are permuted. If the distribution of the bases after permutations matches those before permutations, then the region can be classified as a recombinant region (Sawyer, 1989). A program called GENECONV can detect recombination using this

method (Sawyer, 1989). GENECONV can detect recombination even if the source of recombinant fragment is not found in the alignment (Sawyer, 1989). Another algorithm called maximum chi square can use permutation test to detect recombination (Smith, 1992). This algorithm cannot detect recombinant regions that are not in the alignment. It is less sensitive and more accurate than GENECONV (Smith, 1992). Both tests should be used with detecting recombination by permutation tests.

Recombination breakpoints can also be detected by hidden Markov models (HMM) (Husmeier and McGuire, 2003). In HMM, the hidden state represents tree topology at a given site. A state transition from one topology to another suggests a recombination breakpoint (Husmeier and McGuire, 2003). Therefore, an HMM state sequence needs to be inferred to detect recombination breakpoints. Parameters of the HMM, such as branch lengths, can be estimated using heuristics (HMM-heuristics), maximum likelihood (HMM-ML) or Bayesian statistics (HMM-Bayes) (Husmeier and McGuire, 2003). HMM-heuristics is not very accurate because the branch length is estimated from a small part of the alignment. HMM-ML is often not accurate due to over-fitting by ML estimator. HMM-Bayes can be used to estimate parameters by MCMC simulation (Husmeier and McGuire, 2003). The last algorithm is implemented in C++ program BARCE (Husmeier and McGuire, 2003). BARCE analysis is limited to only four sequences in the alignment because the number of topologies for hidden states increase as the number of sequences increases. Moreover, BARCE cannot detect recombination where the branch length changes without any change in the topology. These can include recombination between closely related sequences (Husmeier and McGuire, 2003). BARCE cannot take into account the rate of evolution of sequences. However, it can be implemented using factorial HMM (FHMM) having two hidden states: one for change in the topology and another for change in the rate of evolution (Husmeier and McGuire, 2003).

Previously, recombination detection methods use DNA sequences to detect recombination. However, it is possible to use both DNA and protein sequences to detect recombination breakpoints using mixture models (MMs) and phylogenetic HMMs (phylo-HMM) (Boussau, Guéguen and Gouy, 2009). In phylo-HMMs, the hidden state is the topology. In MMs, multiple models are used to calculate the likelihood of trees (Boussau, Guéguen and Gouy, 2009). Alignments with more than one topology consistent over long stretches of DNA indicate a probable recombination event (Boussau, Guéguen and Gouy, 2009). Based on simulation results, the phylo-HMM method performed better than MM. A problem with this approach includes inability to detect recombination that does not change the topology of the tree. Recombination breakpoints that are close (less than 200) to the start or the end of a gene cannot be accurately detected by MM, but the accuracy improves by using phylo-HMM (Boussau, Guéguen and Gouy, 2009). Overall, phylo-HMM is better than MM in recovering recombinant fragments.

An evaluation of the methods can be made with simulations and real data. Recombination can be simulated using coalescent with recombination models, in which the waiting time for recombination is given by the exponential distribution (Posada and Crandall, 2001). Alignments were obtained, and different methods of detection of LGT were tested. The results indicated that maximum chi square and GENECONV are the most accurate

of recombination detection programs given that sequence divergence is moderate (Posada and Crandall, 2001). The maximum chi square method performed better than phylogenetic methods due to a number of factors that affect phylogenetic methods such as different rates of evolution at different sites, noise in phylogenetic signal, and lack of resolution of trees (Posada and Crandall, 2001).

The methods of detection of recombination can also be evaluated by empirical data (Posada, 2002). HIV-1 shows high rates of recombination and can be used to test methods of recombination. The results indicate that MaxChi and recombination detection program (RDP) are very accurate in detection of recombination (Posada, 2002). If data sets are relatively divergent, GENECONV can result in false positives. Ancient cases of recombination can be detected more accurately by phylogenetic methods than by permutation tests. Moreover, the frequency of recombination is higher in HIV-1 than previously estimated (Posada, 2002).

2.3 Method

2.3.1 Phylogenetic Tree

To study the evolutionary relationship between bacteria and infer LGT in the bacteria, a phylogenetic tree was inferred. Genome sequences of 47 genomes of *Bacillus*, *Anoxybacillus*, *Geobacillus*, and *Oceanobacillus* were downloaded from NCBI: `ftp://ftp.ncbi.nih.gov/genomes/`. Genomes of *Bacillus* were chosen because of their diversity and the availability of high quality genome sequence data. The genomes and their accession numbers are listed in the appendix section A. The DNA sequences of conserved genes: *gtlX*, *nusA*, *pheS*, and *rpoA* were used to construct the phylogenetic tree (Hao and Golding, 2008b). DNA sequences were used because they might have sufficient information to resolve the branches of *B. cereus* group. DNA sequences could reveal synonymous substitutions, whereas protein sequences only reveal nonsynonymous substitutions. The orthologs of these genes were searched in the genomes by reciprocal best hit of BLAST (Altschul *et al.*, 1997). The reciprocal best hit of BLAST assumed that if a query was searched in a database, and the first hit obtained was in turn used as a query, then the first hit of this reciprocal search should result in the original query. If that were the case, then the genes were assumed to be orthologs (Altschul *et al.*, 1997).

When the orthologs were found, the sequences were extracted and aligned in MUSCLE (Edgar, 2004). MUSCLE is a multiple sequence alignment tool that uses an iterative method of alignment. It aligns the sequences pairwise and obtains similarity scores. A guide tree is inferred using UPGMA method based on these scores. Multiple sequence alignments were inferred based on the tree. Iteratively, the trees and the alignments were refined, and the best multiple sequence alignment is obtained (Edgar, 2004). The alignments obtained for each gene were concatenated. Concatenation of alignments resulted in a combined dataset that would increase the accuracy of phylogenetic inference. The tree was inferred using MrBayes (Huelsenbeck and Ronquist, 2001). MrBayes is a program that infers phylogenies using a Bayesian statistical framework (Huelsenbeck and Ronquist, 2001). Posterior

probabilities are calculated using metropolis coupled Markov chain Monte Carlo simulation (Huelsenbeck and Ronquist, 2001). MCMC analysis was run on multiple CPU cores (Altekar *et al.*, 2004). *MrBayes* allows the user to choose different models of substitution (Huelsenbeck and Ronquist, 2001). In this study, the GTR + Γ + I model was used. The number of characters in the alignment was 4,778. *Oceanobacillus iheyensis* HTE831 was chosen as an outgroup. In the MCMC analysis, the number of generations was 10,000,000. The sampling frequency was 100. The burn-in number was set to 25,000.

2.3.2 Operon Database

Due to a lack of experimental data on operon in prokaryotic genomes, it was necessary to use computational prediction to infer structure in microbial genomes. At the time of writing, a number of computational algorithms and databases existed for operon prediction. OperonDB provided an algorithm for computational prediction of operons (Ermolaeva, White and Salzberg, 2001). In OperonDB, a gene pair was defined as two genes on the same strand that were separated by less than 200 base pairs (Ermolaeva, White and Salzberg, 2001). The algorithms found conserved gene pairs and calculated the probability that the conserved gene pair were part of an operon. The probability was high if the gene pairs or clusters were conserved in order and orientation across different genomes, especially the ones that were distantly related (Ermolaeva, White and Salzberg, 2001). Bacterial genomes have high rates of rearrangements, insertions and deletion. If the gene order were conserved, then it indicated with high probability that the gene pair or cluster would be a part of an operon (Ermolaeva, White and Salzberg, 2001).

In an update to OperonDB, the prediction algorithm was improved resulting in increased accuracy (Perteau *et al.*, 2009). “Gene pair” was redefined to allow the rearrangement of genes and insertion of genes within operons (Perteau *et al.*, 2009). A threshold could determine the maximum number of genes that could separate the two genes in the “gene pair” (Perteau *et al.*, 2009). The degree of conservation of genes would be identified using the Homology teams program. The modified version of the Homology teams program allowed other genes to exist between two genes conserved in synteny across taxa (Perteau *et al.*, 2009). The new algorithm of OperonDB increases the sensitivity of the prediction without causing a decrease in the specificity (Perteau *et al.*, 2009).

OperonDB was used to predict operons in the *Bacillus* genomes. All-against-all BLAST (Altschul *et al.*, 1997), required for OperonDB, was carried out on all protein sequences of all 47 genomes, with the following parameters (-p blastp -e 1e-5 -v 0 -F f) OperonDB (Perteau *et al.*, 2009). OperonDB used the BLAST outputs to infer the degree of conservation of genes across different genomes (Perteau *et al.*, 2009). A Perl script was used to format the input files for the OperonDB program. The output of the OperonDB was a list of gene pairs, confidence values, the number of other genomes where the homologous gene pair is found, and the list of homologous GI values for those gene pairs. The confidence value determined the probability of a gene pair to exist in the same operon (Perteau *et al.*, 2009). The cut-off for confidence value in this study was 80 as suggested in the documentation for OperonDB (Perteau *et al.*, 2009). The data was sorted into operons in which the operon

boundaries were defined as the change of strandedness or as a gene that was not in an operon. Operon boundaries are required because they are used to infer laterally transferred operons. Operons that were completely laterally transferred were excluded from the study. The focus of this study was only on genes that were transferred into operons but not lateral transfer of complete operons into genomes.

2.3.3 Detection of LGT

In order to detect LGT by the presence and the absence of genes, protein families were inferred by a clustering algorithm. The Markov clustering (MCL) algorithm was used to cluster genes into families (van Dongen, 2000). MCL is a graph clustering algorithm that clusters proteins by representing them as vertices of a graph. The similarity scores are represented by weighted edges (van Dongen, 2000). BLAST e-value or bit-score could be used as a statistic for the weight of the edges. A transition matrix showing the pairwise similarity between different genomes could be generated from the graph and transformed into a Markov matrix. The sum of columns of a Markov matrix is always equal to one (van Dongen, 2000). Random walks, known as flows, were simulated in the graph by multiple iterations. Randoms walks were stopped when the algorithm converged (van Dongen, 2000). An inflation parameter controls the size of clusters. If the inflation parameter is high, the clusters are small and vice versa (van Dongen, 2000). After each iteration, the matrix was rescaled to a Markov matrix (van Dongen, 2000). Therefore, MCL reinforced edges that were heavily weighted and removes edges that did not have heavy weight until a naturally clustered graph emerges.

All-against-all BLAST was carried out on all genomes (Altschul *et al.*, 1997). The BLAST parameters were E-value of 1×10^{-10} . The proteins were clustered into families by the MCL algorithm (van Dongen, 2000). The inflation parameter was set to 2.5. The bit score cut-off was 50. Families with duplications were removed to avoid misdetection of duplicate genes as laterally transferred genes. We used protein family information to make binary presence or absence matrices. If a protein existed in a genome, then it was represented by the binary digit of 1. If the protein did not exist in the genome, then it was represented by the binary digit of 0. Parsimony was used to infer the ancestral state at each node in the phylogenetic tree inferred previously (Fitch, 1971). Parsimony assumes that the probability of change of state from absent to present and from present to absent were equal. Ancestral states were inferred to minimize the number of evolutionary changes (Fitch, 1971). LGT was inferred using ancestral state information that was obtained by parsimony using the PHYLIP program pars (Felsenstein, 1989). A change of state from 0 to 1 was inferred as LGT, whereas a change of state from 1 to 0 was inferred as deletion (Felsenstein, 1989). The genes that were not in operons were removed from consideration. If all genes in an operon were laterally acquired, then the operon was classified as laterally transferred operon. It was not used in further analysis.

2.3.4 Detection of Homologous Recombination

Homologous recombination can be detected by several methods including permutation tests. There were two programs that implemented permutation tests. These included GENECONV (Sawyer, 1989) and MaxChi (Smith, 1992). GENECONV excludes monomorphic sites from the alignment. Then, it searches for fragments that are pairwise identical or have high score for base pair matches for a particular pair of sequences. The significance of the recombination breakpoints at both ends of fragments is computed using permutation tests (Sawyer, 1989). The P-value of the GENECONV is the proportion of permuted alignments that have scores greater than or equal to the score of the original alignment (Sawyer, 1989). GENECONV can detect recombination from a source that is within an alignment and from a source that is not in the alignment (Sawyer, 1989). MaxChi creates a 2 by 2 matrix that consists of the number of differences to the right and to the left of the current position in the pairwise alignment. The values of chi square for this matrix is calculated. A suggestive recombination breakpoint is the position where this chi square value is maximum in the alignment. The P-value is calculated by calculated the maximum chi squared value of randomized alignments. This P-value is the proportion of trials with maximum chi squared values greater than that obtained using the original data (Smith, 1992). Both GENECONV and the maximum chi square algorithm can be used independently to detect homologous recombination breakpoints in the same data.

A database of all prokaryotic genome sequences was downloaded from NCBI: `ftp://ftp.ncbi.nih.gov/genomes/`. All genes that were laterally transferred into operons were used for the study of recombination. The protein sequences of the genes were obtained. TBLASTN (Altschul *et al.*, 1997) was used to search the protein sequences against a six-frame translated database of prokaryotic sequences. TBLASTN was used to avoid the bias caused due to differences in the annotations across different taxa (Altschul *et al.*, 1997). In this way, genes that were similar to *Bacillus* genes were obtained, and the source of recombinant DNA in *Bacillus* was identified. Only highly similar sequences were required. Therefore, only the top hits of TBLASTN search were studied. At most, the five top hits from *Bacillus* and at most the five top hits from sequences other than *Bacillus* were studied, including 1,500 base pairs of flanking sequences at both ends of each hit. It was assumed that the recombination breakpoints lied in the flanking regions of the gene. The 1,500 base pair limit was arbitrarily chosen under the assumption that they would provide sufficient data to detect any homologous recombination signal that were present. If the sequences were found on the complementary strand, the sequences were reversed, and the compliments were taken. Multiple sequence alignments were carried out on the sequences using MUSCLE (Edgar, 2004).

MUSCLE alignments were used to detect recombination breakpoints. Recombination breakpoints were inferred using two different algorithms: GENECONV (Graham, McNeney and Seillier-Moiseiwitsch, 2005) and maximum chi square (Smith, 1992). GENECONV could detect homologous recombination from a source that is not included in the alignment. It also ran global and pairwise permutations (Sawyer, 1989). GENECONV was used with default parameters on the DNA sequence alignments. The number of random permutations was set to 10,000. The output from GENECONV was parsed based on the location

of breakpoints and the length of the recombinant region. Unique alignments with a P-value of less than 0.05 were counted.

In addition to *GENECONV*, another algorithm known as maximum chi method was used to detect recombination breakpoints. The maximum chi algorithm was implemented in a program called *Stepwise* that allowed the user to run maximum chi method multiple times and study multiple hypotheses predicting different locations of recombination breakpoints in a stepwise manner (Graham, McNeney and Seillier-Moiseiwitsch, 2005). Maximum chi algorithm was run with three different settings for half-window sizes. The settings were 20 base pairs, 30 base pairs and 50 base pairs width of half-window. The default half-window width was 30 base pairs. The number of permutations was kept at a default value of 1,000 permutations. For each setting, two steps of the algorithm were run on the alignments. If there were no recombination breakpoints found in the first step, then the second step could not be run, and the analysis was stopped. The function of the genes was studied using NCBI annotations.

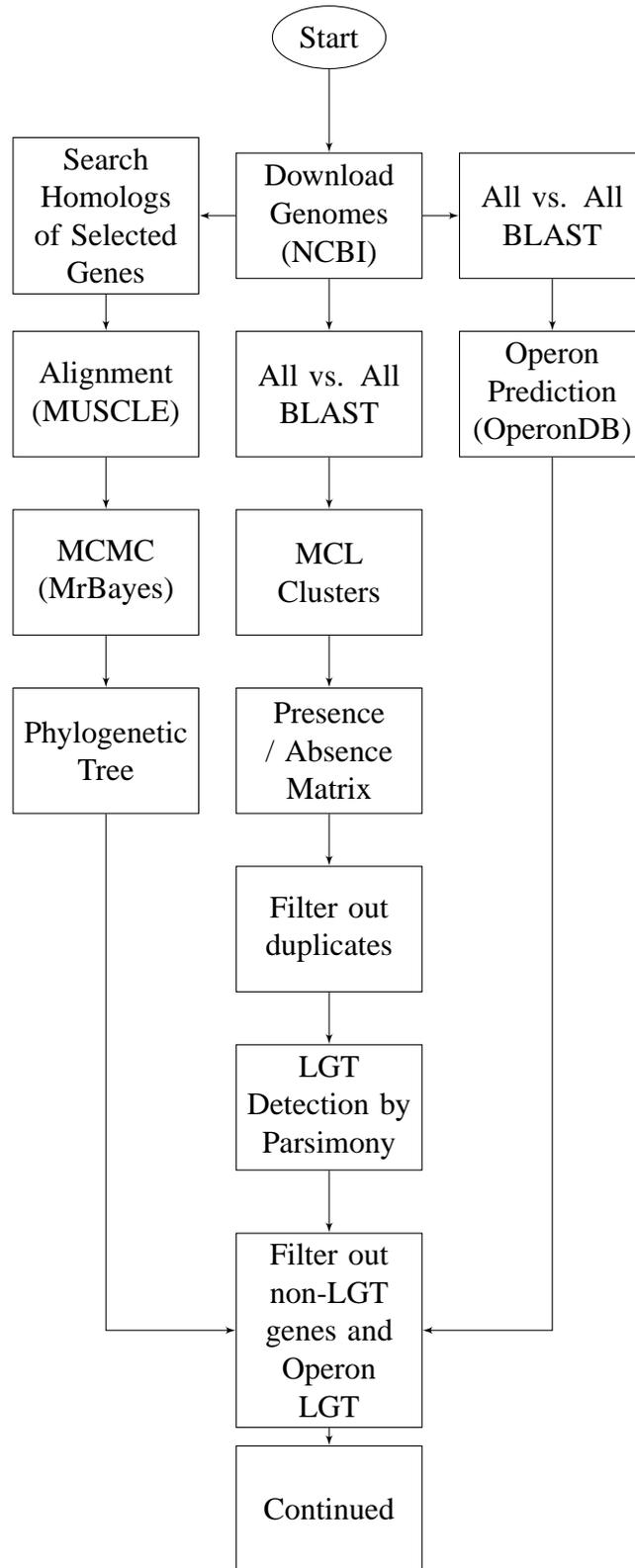


Figure 2.1: Summary of methods

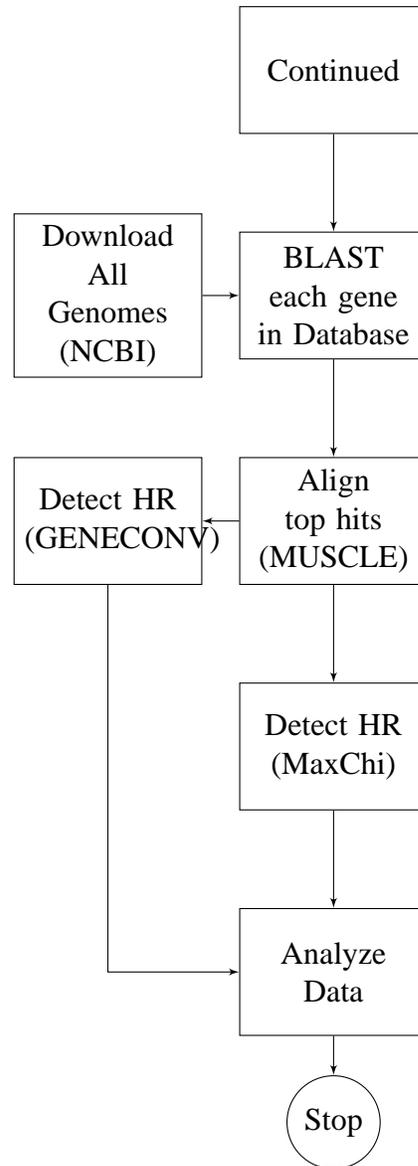


Figure 2.2: Summary of methods continued from the previous page

2.4 Results

2.4.1 Phylogenetic tree

Phylogenetic tree was inferred using MrBayes (Huelsenbeck and Ronquist, 2001). The log likelihood values at the end of the analysis were approximately constant over generations. The values indicated that the MCMC chain had converged. In this research, two runs of MCMC were carried out, and both had converged at the end of the analysis. The tree indicating posterior probabilities is shown in figure 2.3. The tree indicating branch lengths is shown in figure 2.4.

As shown in the figures, the branch lengths of species in *Bacillus cereus* group are small. DNA based Bayesian tree was able to resolve most of the strains of *B. cereus*. Subtree of the strains of *Bacillus anthracis* could not be resolved. As shown in figure 2.3, most clades have a posterior probability of greater than 0.95 with the exception of a few clades. These included clades in the *B. cereus* group particularly the *B. anthracis* clade. Figure 2.3 indicates that branch lengths of strains in the *B. cereus* group were extremely small. Figure 2.3 indicates that Alkali bacillus was grouped together in one clade with a posterior probability of 0.95. The strains of *B. subtilis* were also in a single clade supported with a posterior probability of 1.0. The strains of *Bacillus licheniformis* were very closely related.

Figure 2.4 indicates branch lengths of the consensus tree obtained from MrBayes. *O. iheyensis* HTE 831 had the largest branch length of 0.69. *O. iheyensis* is a marine bacterium (Takami, Takaki and Uchiyama, 2002). The branch lengths in cereus group are very small indicating that the species in the cereus group are closely related and recently diverged (Helgason *et al.*, 2000).

2.4.2 Operon Prediction

As indicated in the methods, gene pairs that were part of operons were predicted using OperonDB (Pertea *et al.*, 2009). The operon structure data is shown in the appendix, section A. The data indicate the number of operons in genomes, the total number of genes in all operons, the number of protein coding genes in operons, and the size of genome (in base pair of DNA). It was observed that there are approximately 300 to 500 operons in each genome. The total number of genes is approximately 1200 to 1900, resulting in approximately 3 to 4 genes in each operon on average. Smaller genomes had lesser number of operons with the exceptions of Geobacillus. In Geobacillus, the size of the genome is small, but the number of operons is not smaller than the average number of operons per genome. Geobacillus had lesser number of genes than other Bacillus. However, the percentage of genes in operons was higher in Geobacillus than that in other Bacillus species.

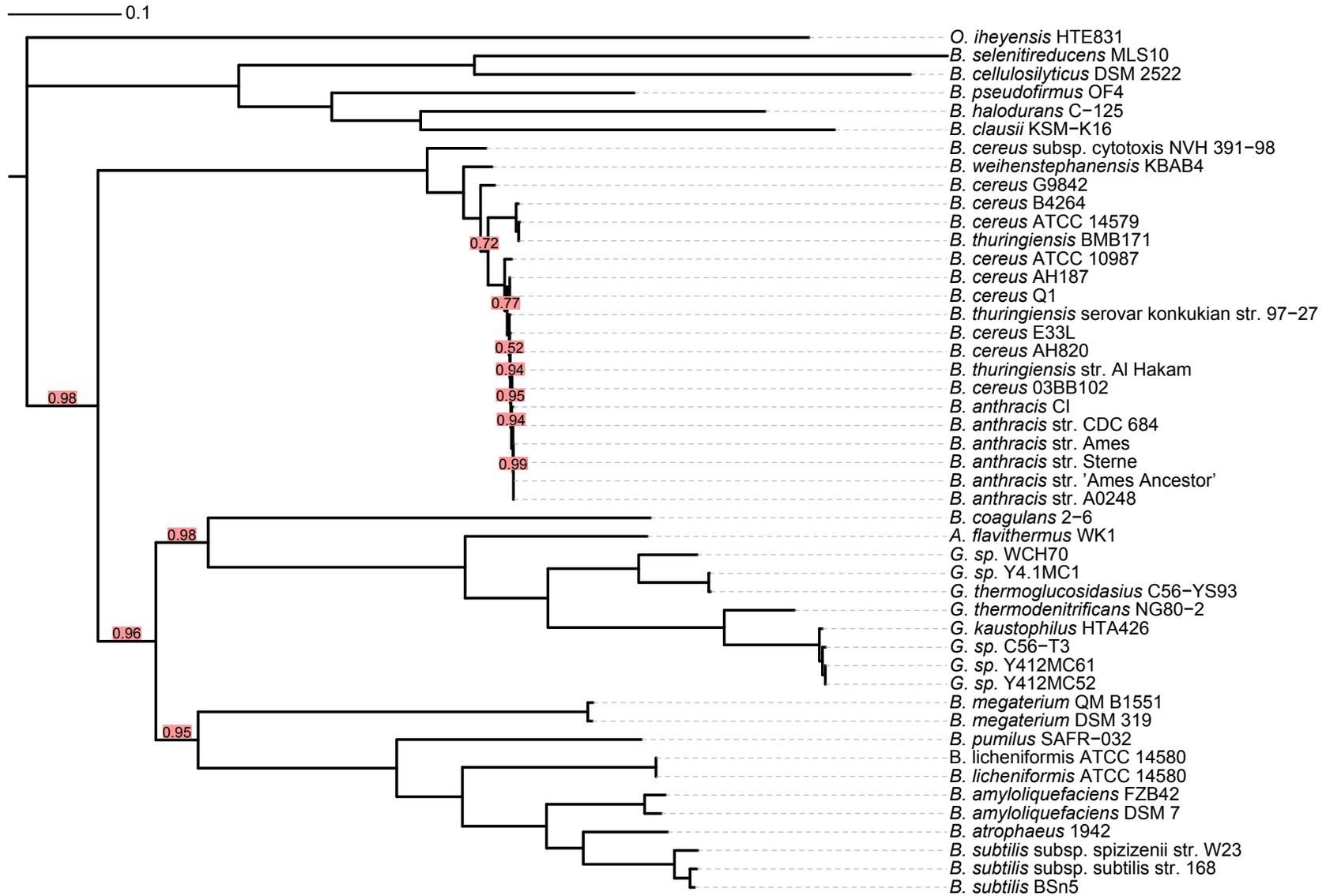


Figure 2.3: The phylogenetic tree shows posterior probabilities that are less than 1.0. The tree was plotted using iTOL (Letunic and Bork, 2006). The scale bar indicates branch lengths.

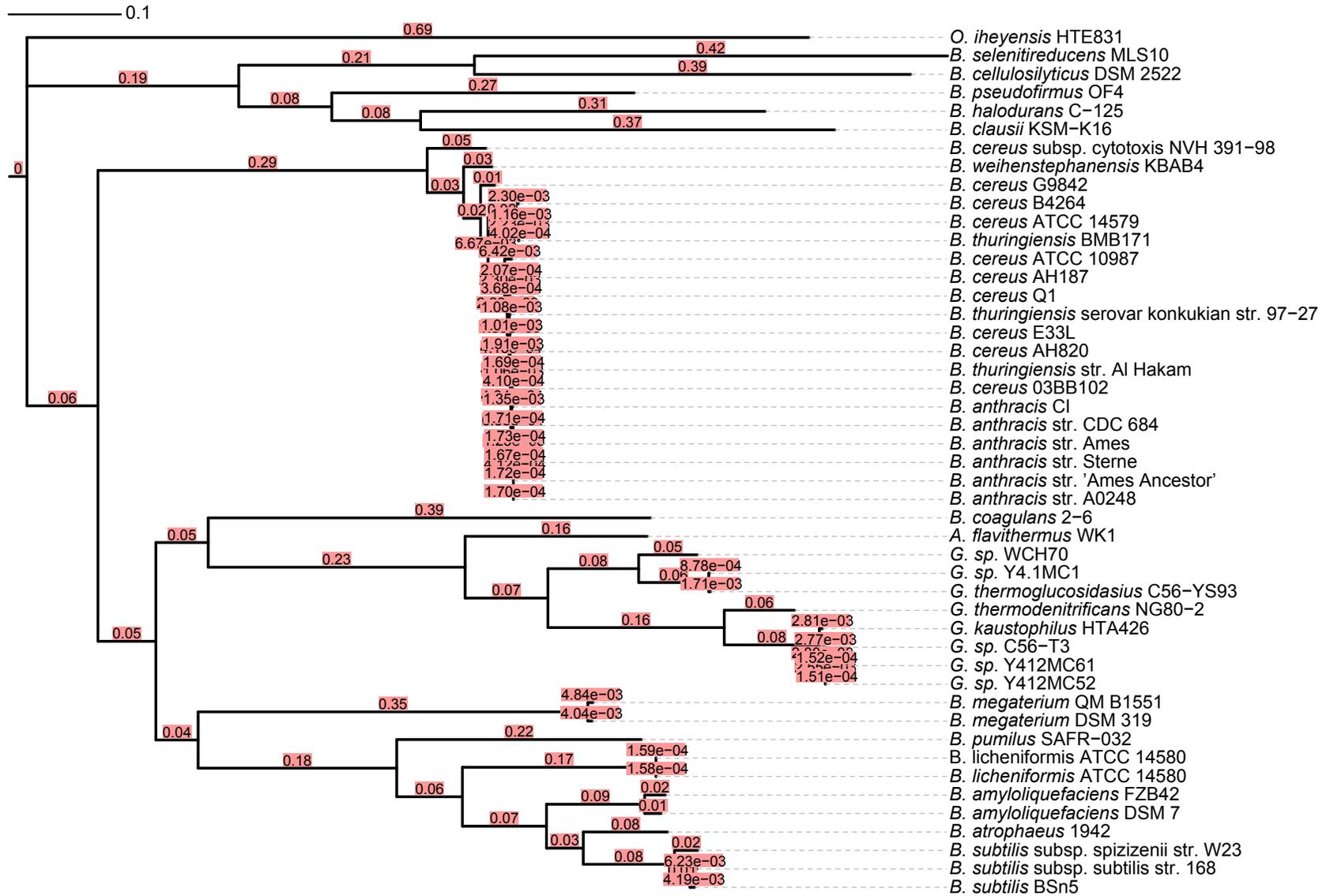


Figure 2.4: The phylogenetic tree shows branch lengths. The tree was plotted using iTOL (Letunic and Bork, 2006). The scale bar indicates branch lengths.

2.4.3 Prediction of LGT

The number of protein families obtained by MCL clustering algorithm was 10,277. Out of these 10,277 families, there were 2,610 families with multicopy genes and 7,667 families with single copy genes. We excluded the data that didn't have at least one nonambiguous change of state from 0 to 1. This set of laterally transferred genes was further refined by removing laterally transferred operons. The number of laterally transferred genes in operons was 2,072. The data are shown in the appendix section.

As shown in these data, approximately 10 to 20 percent of all genes in *Bacillus* were laterally transferred, even after ignoring ambiguous genes and ORFans. This indicates high rates of LGT in *Bacillus*. The data indicate that most genes are not transferred into operons. Less than 10 percent of laterally transferred genes are integrated into operons. However, these results do not imply that there is a low rate of transfer of complete operons into genomes. The lateral transfer of complete operons was not investigated in this research. The results are shown in the following table.

Table 2.1: Table shows the following results of OperonDB and LGT detection.

Genome	No. of operons	LGT	LGT in Operons
<i>A. flavithermus</i> WK1	366	381	61
<i>B. amyloliquefaciens</i> DSM 7	430	594	49
<i>B. amyloliquefaciens</i> FZB42	351	551	15
<i>B. anthracis</i> CI	452	1203	40
<i>B. anthracis</i> str. 'Ames Ancestor'	415	1454	39
<i>B. anthracis</i> str. A0248	410	1370	39
<i>B. anthracis</i> str. Ames	420	1466	39
<i>B. anthracis</i> str. CDC 684	451	1425	42
<i>B. anthracis</i> str. Sterne	437	1273	43
<i>B. atrophaeus</i> 1942	359	649	15
<i>B. cellulosilyticus</i> DSM 2522	466	521	62
<i>B. cereus</i> 03BB102	445	1318	39
<i>B. cereus</i> AH187	451	1338	41
<i>B. cereus</i> AH820	444	1357	40
<i>B. cereus</i> ATCC 10987	388	1315	28
<i>B. cereus</i> ATCC 14579	431	1205	37
<i>B. cereus</i> B4264	439	1227	36

Genome	No. of operons	LGT	LGT in Operons
<i>B. cereus</i> E33L	431	1183	41
<i>B. cereus</i> G9842	434	1262	9
<i>B. cereus</i> Q1	442	1241	41
<i>B. cereus</i> subsp. cytotoxis NVH 391-98	347	666	24
<i>B. clausii</i> KSM-K16	524	465	55
<i>B. coagulans</i> 2-6	370	229	26
<i>B. halodurans</i> C-125	439	449	26
<i>B. licheniformis</i> ATCC 14580	484	850	72
<i>B. licheniformis</i> ATCC 14580	486	839	72
<i>B. megaterium</i> DSM 319	367	1220	17
<i>B. megaterium</i> QM B1551	369	1243	19
<i>B. pseudofirmus</i> OF4	416	320	17
<i>B. pumilus</i> SAFR-032	420	427	51
<i>B. selenitireducens</i> MLS10	335	331	33
<i>B. subtilis</i> BSn5	378	660	20
<i>B. subtilis</i> subsp. spizizenii str. W23	365	619	25
<i>B. subtilis</i> subsp. subtilis str. 168	366	713	21
<i>B. thuringiensis</i> BMB171	423	1076	37
<i>B. thuringiensis</i> serovar konkukian str. 97-27	436	1169	45
<i>B. thuringiensis</i> str. Al Hakam	357	984	26
<i>B. weihenstephanensis</i> KBAB4	439	1252	41
<i>G. kaustophilus</i> HTA426	456	547	86
<i>G. sp.</i> C56-T3	423	601	61
<i>G. sp.</i> WCH70	393	458	78
<i>G. sp.</i> Y4.1MC1	437	699	90
<i>G. sp.</i> Y412MC52	441	672	80
<i>G. sp.</i> Y412MC61	435	669	74
<i>G. thermodenitrificans</i> NG80-2	427	488	95
<i>G. thermoglucosidasius</i> C56-YS93	458	744	74
<i>O. iheyensis</i> HTE831	422	278	51

The figure 2.5 shows the distribution of laterally transferred genes in operons. Moreover, genes that encode for toxins can also be identified in operons using NCBI annotations. Table 2.2 indicates the genes that encode for toxin in operons of *Bacillus*.

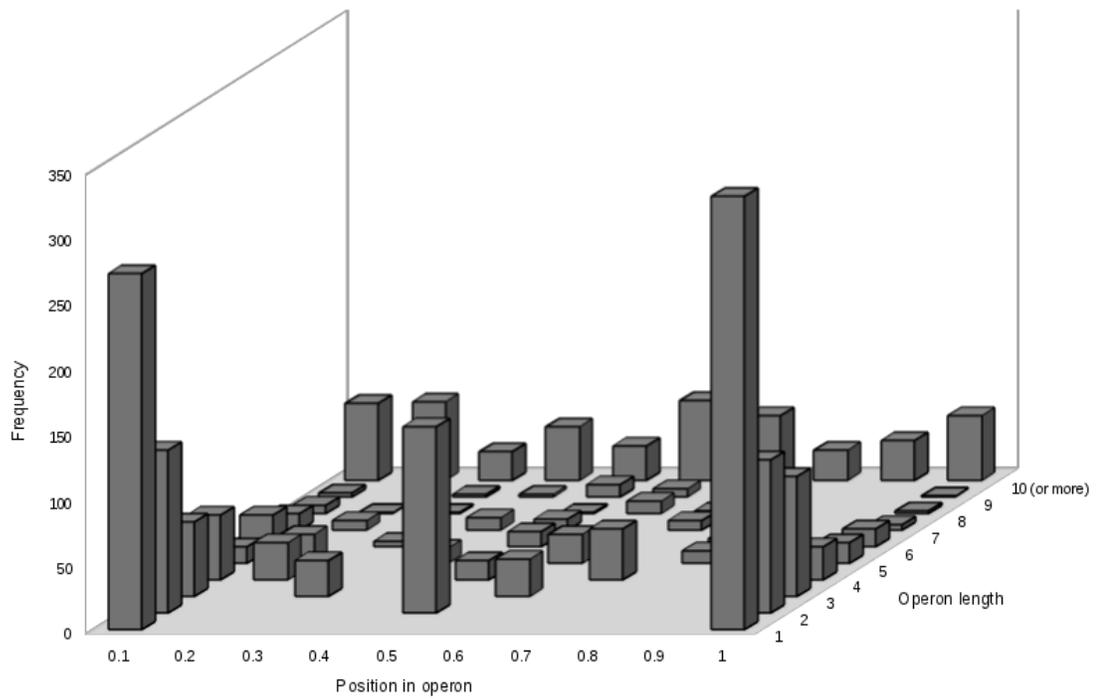


Figure 2.5: The figure shows the locations of genes in operons. The x-axis indicates locations of laterally transferred genes. “0” indicates a gene was transferred in the beginning of an operon, and “1” indicates that a gene was transferred at the end of an operon. The z-axis indicates the number of genes in operons before lateral gene transfer.

Table 2.2: GIs and descriptions of genes annotated as toxin genes

Genome	GI	Annotation
<i>A. flavithermus</i> WK1	212638048	Toxin-antitoxin addiction module toxin component MazF
<i>B. amyloliquefaciens</i> DSM 7	308172333	Antitoxin EndoAI EndoA inhibitor
<i>B. atrophaeus</i> 1942	311071137	endoribonuclease toxin
<i>B. cellulositicus</i> DSM 2522	317127076	transcriptional modulator of MazE/toxin, MazF
<i>B. coagulans</i> 2 6	336115424	transcriptional modulator of MazE/toxin, MazF
<i>B. cytotoxicus</i> NVH 391 98	152974084	transcriptional modulator of MazE/toxin, MazF
<i>B. megaterium</i> DSM319	295702427	antitoxin endoAI
<i>B. megaterium</i> QM B1551	294497062	antitoxin EndoAI (EndoA inhibitor)
<i>B. pseudofirmus</i> OF4	288554846	MazE of toxin/antitoxin system-like protein
<i>B. pseudofirmus</i> OF4	288554847	MazF of toxin/antitoxin system-like protein
<i>B. pumilus</i> SAFR 032	157693387	toxin regulator
<i>B. selenitireducens</i> MLS10	297582854	transcriptional modulator of MazE/toxin, MazF
<i>B. selenitireducens</i> MLS10	297583120	toxin secretion/phage lysis holin
<i>B. selenitireducens</i> MLS10	297585259	toxin secretion/phage lysis holin
<i>B. subtilis</i> 168	16077533	endoribonuclease toxin
<i>B. subtilis</i> BSn5	321314140	endoribonuclease toxin
<i>B. subtilis</i> spizizenii W23	305673180	endoribonuclease toxin
<i>B. weihenstephanensis</i> KBAB4	163938244	transcriptional modulator of MazE/toxin, MazF
<i>G. thermoglucosidasius</i> C56 YS93	336236990	transcriptional modulator of MazE/toxin, MazF
<i>G.</i> Y412MC61	261418555	transcriptional modulator of MazE/toxin, MazF
<i>G.</i> Y412MC61	261419886	transcriptional modulator of MazE/toxin, MazF

2.4.4 Prediction of HR

GENECONV was used to detect recombination breakpoints flanking all genes that were inferred to be laterally transferred into operons. The number of alignments that had at least one significant ($P < 0.05$) recombination breakpoint before the start site and one significant ($P < 0.05$) recombination breakpoint after the end site was 124. In these alignments, the source and the destination of the recombinant fragment was from a *Bacillus* species to another *Bacillus* species. Further analysis of these genes revealed that they were, originally, from 45 different families.

The alignments were used to detect if some recombination sequences originated from organisms other than *Bacillus*. There were 58 alignments indicating that the sources of recombinant fragments were non-*Bacillus* organisms, and the recipients were also non-*Bacillus* organisms.

GENECONV could also detect recombination in which the source sequence was not in the alignment. Results indicated that there were two unique genes in which the source of the recombinant region was unknown. One gene was from *Bacillus halodurans* C-125, and it was described as ‘hypothetical protein BH0716’ in the annotation. The other was from *B. licheniformis* ATCC 14580, and the protein was described as ‘hypothetical protein BLi04182’ in the annotation. The function of these proteins is unknown.

Stepwise implementation of the maximum chi square algorithm was also used to detect recombination (Graham, McNeney and Seillier-Moiseiwitsch, 2005). Three different settings for half-window size of maximum chi method were used. They were half-window sizes of 20 base pair, 30 base pair (default) and 50 base pair. When the half-window size was 20 base pair, 1405 alignments did not have any significant ($P < 0.10$) recombination breakpoints. However, 660 alignments had at least one significant recombination breakpoint. In those alignments, the recombination breakpoints could be found anywhere in alignment. There were 25 alignments that had at least one recombination breakpoint at one end of the laterally transferred genes. There were 21 alignments that had recombination breakpoints after the other end of laterally transferred genes. In the second step of stepwise program, the known recombination breakpoints were used to search for more recombination breakpoints (Graham, McNeney and Seillier-Moiseiwitsch, 2005). After the second step, there was one breakpoint found before the start of a laterally transferred gene.

When the half-window size was 30 base pair, 1041 alignments did not have any significant recombination breakpoints, as opposed to 1024 alignments that had at least a single recombination breakpoint. There was one alignment where the recombination breakpoints was before the start codon and after the stop codon. However, the result did not indicate the origin of the gene. The breakpoint, before one end of the gene, suggested that the donor of the gene was *Myxococcus xanthus* DK 1622; whereas, the breakpoint after the end of stop codon suggested that the donor was *Stigmatella aurantiaca* DW4/3-1. The laterally transferred gene had recombination breakpoints at both ends indicating that the gene might have been integrated by homologous recombination. There were 50 alignments that had recombination breakpoints before one end of the laterally transferred gene but not after the end of the gene. There were 30 alignments with recombination breakpoints after the gene but not before. After the second step, there was another recombination breakpoint found

before the start of one laterally transferred gene. A closer examination indicated that the laterally transferred gene had recombination breakpoints detected after the first step of the algorithm. The combined results indicated that one gene had recombination breakpoints before the stop codon and after the start codon. The results justified the used of stepwise approach of recombination detection (Graham, McNeney and Seillier-Moiseiwitsch, 2005). The origin of the gene was *Lysinibacillus sphaericus* C3-41. The recipient organism was *Geobacillus thermodenitrificans* NG80-2.

When the half-window size was 50 base pair, 748 alignments did not have any recombination breakpoints, whereas 1317 alignments had at least one recombination breakpoint. The number of alignments that had breakpoints before laterally transferred genes was 57, whereas the number of alignments that had recombination breakpoints after laterally transferred genes was 47. After step two, there were five additional alignments that had at least one recombination breakpoint before the laterally transferred gene. There were three additional alignments that had recombination breakpoints after the laterally transferred genes.

There were two genes in which evidence for double recombination is found. The results were also confirmed with another database of operons. The first gene was a peptide deformylase gene (GI: 138894692) that was present in *G. thermodenitrificans* NG80-2. The gene was in an operon based on the Database of prOkaryotic OpeRons (DOOR) database (ID: 308856; Dam *et al.*, 2007) The figure 2.6 shows the gene in other organisms.

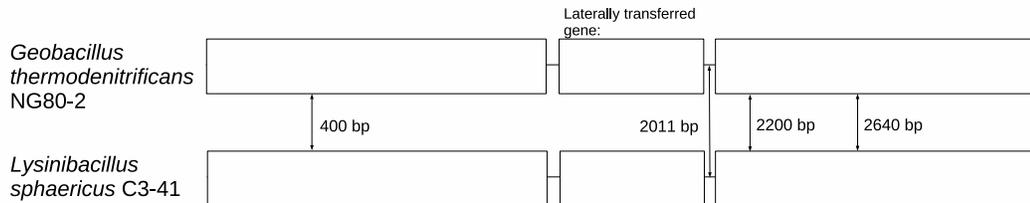


Figure 2.6: The figure shows the laterally transferred gene and its source organism. The gene in the centre is the laterally transferred gene, and the genes flanking the laterally transferred genes, are gene in similar pathways. The gene on the left side of the laterally transferred gene is a primosomal protein. The gene on the right side of the laterally transferred gene is methionyl-tRNA formyltransferase.

The second gene was the ABC transporter permease (GI: 138895062). It was transferred into *G. thermodenitrificans* NG80-2. The gene was in operons of DOOR database (ID: 308934; Dan *et al.*, 2007). The figure 2.7 shows the gene in other organisms.

The results of *GENECONV* indicated that there were at least 124 genes that had recombination breakpoints. Out of those 124 genes, 40 genes were annotated as hypothetical proteins and 84 genes were functionally annotations. In order to study the transfer of function from a non-Bacillus organism to a Bacillus species, results of maximum chi square method were studied. Results from two steps of the stepwise program were combined. When the half-window size was 20, there were 45 genes that had at least one recombination breakpoint at either end. Out of those 45 genes, 12 genes were annotated as hypothetical proteins. There were 33 genes that indicated a transfer of functionally annotated genes into operons

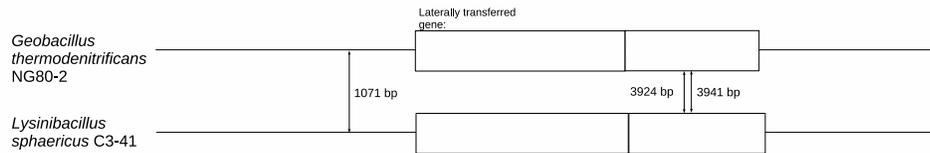


Figure 2.7: The figure shows the laterally transferred gene and its source organism. The gene on the left is the laterally transferred gene. It is the ABC transporter permease gene. The gene on the right is the ABC transporter ATP-binding protein.

of *Bacillus*. They were from non-*Bacillus* sources. When the half-window size was 30, there were 79 genes that had at least one recombination breakpoint at either end. Out of those 79 genes, 18 genes were annotated as hypothetical proteins. There were 61 genes that indicated a transfer of functionally annotated genes into operons of *Bacillus*. They were from non-*Bacillus* sources. When the half-window size was 50, there were 108 genes that had at least one recombination breakpoint at either end. Out of those 108 genes, 29 genes were annotated as hypothetical proteins. There were 79 genes that indicated a transfer of functional genes into operons of *Bacillus*. They were from non-*Bacillus* sources. In conclusion, the results suggest that it is possible for functionally annotated genes to integrate into functionally appropriate operons by homologous recombination.

2.5 Discussion

2.5.1 Choice of Organism

Bacillus was chosen because there were many high quality, complete and annotated genome sequences available in NCBI database. It is gram positive, rod-shaped, aerobic or facultative anaerobic bacteria that form highly resistant spores in response to stress (Earl, Losick and Kolter, 2008). Pathogenic species include some strains of *B. anthracis*, *B. cereus* and *B. thuringiensis* (Ravel and Fraser, 2005). Virulence in *B. anthracis* is encoded by plasmids pX01 and pX02 (Okinaka, Pearson and Keim, 2006). The plasmid pX01 has genes that encode for toxins, whereas pX02 has cap (capsule) genes. Both plasmids are required for virulence (Okinaka, Pearson and Keim, 2006). However, *B. cereus* strains CA and CI also have these two plasmids and are capable of causing anthrax disease (Okinaka, Pearson and Keim, 2006). Centre for Disease Control (CDC) defines *B. anthracis* as a bacterium that is a) capsule producing b) nonmotile c) susceptible to γ -phage d) non-haemolytic e) susceptible to penicillin (Okinaka, Pearson and Keim, 2006). *B. cereus* CI and CA are not susceptible to γ -phage. Therefore, they cannot be classified as *B. anthracis* (Okinaka, Pearson and Keim, 2006). Some strains of *B. anthracis* do not have pX01 and pX02. The ancestor of *B. anthracis* might have had pX01 and pX02, and some of the recent strains had lost them (Okinaka, Pearson and Keim, 2006). *B. cereus* ATCC 10987 was responsible for food poisoning caused by cheese in Canada (Rasko *et al.*, 2004). *B. thuringiensis* serovar konkukian str causes tissue neurosis in *Homo sapiens* (Hernandez *et al.*, 1998).

Different species of *Bacillus* have distinct habitats, such as fresh water, soil and marine (Luis *et al.*, 2011). Fresh water species include *B. sp m3-13* and *B. holodurans*. Soil species include *B. anthracis* and *B. cereus*. Marine and deep-sea species include *O. ihenensis* and *G. kausophilis* (Takami, Takaki and Uchiyama, 2002). These habitats lead to a diverse set of genes in the genomes of *Bacillus*. The genes that are involved in environmental adaptations show a high degree of variation in different species of *Bacillus* (Takami *et al.*, 2004). Marine species are resistant to osmotic pressure and salinity (Luis *et al.*, 2011). Due to this diversity and sequence availability of *Bacillus*, we decided to use *Bacillus* in our research.

2.5.2 Phylogenetic Tree

In the nineteenth century, Charles Darwin suggested that tree of life was similar to a coral (Darwin, 1891). Only the surface of a coral is alive. Similar to a coral, the organisms at the tips of the lineages are alive, whereas those that are not at the tips are extinct (Darwin, 1891). Evolution of eukaryotes can be modelled by a phylogenetic tree. Similarly, most genes in bacteria evolve by vertical descent. Therefore, a vertical like of descent exists and can be modelled by a phylogenetic tree (Wolf *et al.*, 2002).

Alignment of the sequences, required for the tree, was done using MUSCLE (Edgar, 2004). The advantage of using MUSCLE is that it is an iterative method and avoids the problem of sub-optimal alignments caused by progressive methods such as CLUSTALW (Edgar, 2004). The tree inferred in this research is a Bayesian tree. Bayesian methods calculate the posterior probability of observing data given the model (Li, Pearl and Doss, 2000). Bayesian methods of inferring phylogenies were proposed as alternate to maximum likelihood methods (Lartillot and Philippe, 2004). The tree inferred in this study showed that it was difficult to resolve species of the cereus group. However, the difference in the phenotype of *B. cereus* and *B. anthracis* can be explained by 200 unique genes between the two organisms (Hao and Golding, 2004). Moreover, based on Centre for Disease Control (CDC) definition of *B. anthracis*, the two species (*B. anthracis* and *B. cereus*) should not be considered as same species because most strains of *B. cereus* are motile and not susceptible to γ -phage, whereas most strains of *B. anthracis* are nonmotile and susceptible to γ -phage (Okinaka, Pearson and Keim, 2006).

In this study, *O. iheyensis* was chosen as an out-group because it had a unique niche compared to other species of *Bacillus* (Takami, Takaki and Uchiyama, 2002). As shown in the results, it had the longest branch length. However, a long branch length does not always imply that the species is an out-group because the long branch could be due to the high rate of evolution and might not necessarily be a true out-group (Gogarten, Murphey and Olendzenski, 1999).

Despite being statistically robust, the method of inference of phylogenetic tree used in this study is not perfect. It can be argued that single genes do not have enough information to resolve all branches of the tree (Wolf *et al.*, 2002) as in the case of strains of *B. anthracis*. In *B. anthracis*, it can be observed that the tree presented in the results section could not be resolved even if DNA sequences were used. All the genes that were concatenated may not

follow the same model (Wolf *et al.*, 2002).

2.5.3 Operons

Operons can be predicted by studying the intergenic distance between genes. If the intergenic distance between two genes is small, then the probability of having a promoter and a ribosome binding site between the genes decreases (Price, Arkin and Alm, 2006). Hence, it can be concluded that genes are in operons. However, research suggests that intergenic distance between *E. coli* genes that are on the same strand is positively correlated to the expression levels of the downstream gene but not the upstream gene (Eyre-Walker, 1995). The large intergenic distance could possibly prevent interference between ribosomes (Eyre-Walker, 1995). He suggests that the use of intergenic distances to predict operon structure may not necessarily result in a good prediction in these circumstances.

OperonDB was used because of a high specificity of prediction. The sensitivity of prediction was not very high in the first version of the database (Brouwer, Kuipers and van Hijum, 2008) because it assumed that synteny of conserved gene pairs is identical across different genomes (Ermolaeva, White and Salzberg, 2001). Moreover, it did not account for LGT within gene pairs (Ermolaeva, White and Salzberg, 2001). However, the second version of the database allows two laterally transferred or rearranged genes between gene pairs (Perteau *et al.*, 2009). The specificity of the prediction is still close to 98%, while the sensitivity of the prediction has been increased by at least 20% as compared to the first version of the database (Perteau *et al.*, 2009). Moreover, the genomes of *Bacillus* are more conserved in synteny than the genome of *E. coli* probably due to a low number of insertion sequences (IS) and low rate of rearrangements in *Bacillus* (Itoh *et al.*, 1999). Moreover, IS elements should not affect the LGT in operon analysis because LGT of IS is less frequent than transposition of IS (Lawrence, Ochman and Hartl, 1992). Therefore, we used the second version of the database.

The prediction algorithm of OperonDB is based on conservation of gene order in operons across species (Perteau *et al.*, 2009). Although the rate of rearrangements in prokaryotic genomes is high, the order of some genes may still be conserved in the genomes (Dandekar *et al.*, 1998). If the order of genes is conserved, then there is a high probability that the protein products physically interact. However, it should not be a problem for the prediction of operons using OperonDB because the second version of the database allows the gene pair to be separated by up to two genes (Perteau *et al.*, 2009). Hence, it will take into account the shuffling of genes within an operon and should not significantly affect the sensitivity of the operon prediction. Moreover, gene order is more conserved in gram-positive bacteria than in gram-negative bacteria (Itoh *et al.*, 1999). Our results of approximately 3-5 genes per operon are, therefore, reasonable and expected.

Besides operons, there are other regulatory and transcriptional structures discovered in bacterial genomes (Che *et al.*, 2006). These include regulons, modulons and stimulons. A group of operons that are regulated by the same transcription machinery is called a regulon (Kremling *et al.*, 2000). A group of regulons that are controlled by a global regulator and responds to the physiological states of the cell is known as modulon (Kremling *et al.*,

2000). A stimulon is a group of operons, regulons, or modulons that respond to common environmental stimulus (Kremling *et al.*, 2000).

2.5.4 Detection of LGT

Genes need to be clustered into families in order to distinguish divergent family paralogs from LGT. There are a number of methods to detect LGT. However, not all methods give the same results (Lawrence and Ochman, 2002). The lack of consensus among different methods is because different methods are testing different features in the data (Lawrence and Ochman, 2002). For example, nucleotide composition based methods can detect recent lateral transfers between species having different nucleotide composition. It cannot detect ancient genes transfer that can be detected by phylogenetic approaches (Lawrence and Ochman, 2002). The success of transfer of laterally transferred genes by any method depends on three factors: *a*) Transfer of the gene into the cell *b*) Integration of the gene into the chromosome *c*) Retainability of the gene by positive selection (Popa *et al.*, 2011). LGT was detected by studying patterns of presence and absence of genes in bacterial genomes. The method cannot be used to detect the replacement of orthologous genes (Hao and Golding, 2004). Moreover, the method underestimates the number of insertions and deletions (Hao and Golding, 2004). However, other methods, such as codon usage bias and nucleotide composition bias, are not very accurate methods to detect LGT (Koski, Morton and Golding, 2001).

The patterns of presence and absence of genes does not always lead to the correct conclusion about LGT. Multiple parallel deletions of genes are rare and unlikely, but they are possible. LGT is used as an explanation if the patterns of presence and absence of genes can be explained more parsimoniously by LGT than by multiple parallel deletions. For example, it was assumed that *E. coli* gained *lac* operon via LGT based on the pattern of distribution of genes (Stoebel, 2005). However, it was discovered later that the *lac* operon was lost from Salmonella concluding that the gain of the operon was unlikely in *E. coli* (Stoebel, 2005). Moreover, parsimony methods are biased because they underestimate the number of events (Hao and Golding, 2004). In addition, ORFans are excluded from this study (Hao and Golding, 2004).

2.5.5 Detection of Homologous recombination

In the absence of homologous recombination, it is expected that there would be: *a*) high linkage disequilibrium *b*) treelike phylogeny *c*) same phylogeny at all sites (Didelot and Maiden, 2010). Recombination was detected by the maximum chi square algorithm (Smith, 1992). The accuracy of the method was previously evaluated using both simulated and empirical data. Results of analysis of computer simulations (Posada and Crandall, 2001) and empirical data (Posada, 2002) showed that these methods are very accurate. It would be difficult to detect homologous recombination if the divergence of sequences were too high or too low (Posada, 2002). In this research, the divergence of sequences in the alignment was high (between 5% to 20% differences). In such cases, the maximum chi square algorithm

and GENECONV can be used (Posada, Crandall and Holmes, 2002).

This research showed that there is evidence for LGT within operons even if the number of laterally transferred genes is underestimated by parsimony (Hao and Golding, 2004). The results indicated that recombination breakpoints were detected in the operons before the start codon and after the stop codon of laterally transferred genes. The results of GENECONV indicated that, in most cases, the origin of recombinant regions is from *Bacillus* species, and it is transferred to another *Bacillus* species. Most of the genes were not annotated as ‘hypothetical proteins’ and were functionally annotated with functions appropriate for the operons. This does not necessarily mean that the functions of the genes were experimentally verified in the operons. However, based on sequences of genes and their functions in closely related bacteria, it could be inferred that the genes had functions in the recipient genome, as well. Genes that are not functional in bacterial genomes are lost from the population. Therefore, the laterally transferred genes might be expressed in the operons of bacteria. Since the genes were transferred from another *Bacillus* species, there is a higher probability that the genes were functional in the recipient genome.

In other cases, the origin of recombinant region was from non-*Bacillus* species, and it was transferred to a *Bacillus* species. Those cases were detected by stepwise implementation of the maximum chi square algorithm (Smith, 1992). The results of the maximum chi square algorithm indicated that there were recombination breakpoints before the start and after the end of the laterally transferred gene. Those recombination breakpoints were between a *Bacillus* sequence and a non-*Bacillus* sequence. Therefore, there was evidence for homologous recombination of laterally transferred genes from a distantly related source. Usually, the rate of homologous recombination between closely related sequences is high (Lawrence, 2002a). However, DNA from distantly related sources can be introduced by the nonhomologous end joining mechanism (Lawrence, 2002a). However, the observed breakpoints are most likely due to homologous recombination and not by the nonhomologous end joining mechanism. The rate of homologous recombination in bacteria is as high as the mutation rate, but the probability of nonhomologous end joining is very low (Lawrence, 2002a). Similar to the results of GENECONV, analysis of annotations indicated that most of the genes were not annotated as ‘hypothetical proteins’. However, in the case of the maximum chi square algorithm, the genes were integrated into genomes from a distant source. When searched for using BLASTP, the laterally transferred genes were flanked by functionally related genes in the distantly related organism suggesting that transfer was not by the nonhomologous end joining mechanism, but by homologous recombination.

It may be that the recombination signals in cases where the recombinant region is from non-*Bacillus* species to a *Bacillus* species are lost due to mutations. It is also expected that the recombination signal might be deleted after the genes are integrated into the operons, thus maintaining a small intergenic distance between the genes. The loss of a recombination signal could also be due to amelioration of laterally transferred genes. Laterally transferred genes ameliorate to reflect the codon bias and base pair composition of host genomes (Lawrence and Ochman, 1997). As a result, it might be difficult to detect recombination breakpoints. However, recombination breakpoints of newly transferred genes should be detected easily.

As shown in the results, there are some cases where recombination breakpoints are detected in one end of the gene but not the other suggesting that there may have been recombination signals at both ends of the genes, but the signal at one end may have been lost due to mutations, deletions or amelioration. It is known that the rate of integration of DNA increases by a factor of 10^5 if there is at least one end that is homologous between the donor and the recipient (De Vries and Wackernagel, 2002). In another study that did not look at operons specifically, the rate of recombination within a group was higher than outside of the group (Doroghazi and Buckley, 2010). This results suggested that the rate of recombination was higher when the divergence was low (Doroghazi and Buckley, 2010). However, recombination between different groups is still possible in operons because operons are relatively more conserved than other genes (Price, Arkin and Alm, 2006). Homologous recombination is more likely if the divergence is less than 25% (Thomas and Nielsen, 2005).

As noted in the results, recombination breakpoints could be upstream from the start codon and downstream from the stop codon. It could be suggested that the breakpoints are actually in the genes upstream or downstream laterally transferred genes. However, the presence of recombination breakpoints within genes need not disrupt the structure and function of the genes. In a recent study by Chan *et al.* (2009), it was shown that protein domains are units of LGT. Recombination breakpoints were detected within the coding sequences of the genes. These results suggested that the presence of recombination breakpoints, as long as the breakpoints are outside protein domains, should not affect the structure and function of the gene (Chan *et al.*, 2009). The breakpoints could not be due to nonhomologous recombination because the rate of homologous recombination is 10^9 times greater than that of nonhomologous recombination (De Vries and Wackernagel, 2002).

Despite high rates of LGT in bacteria, LGT and recombination was not detected at a high frequency in these results. One of the reasons might include the nature of the bacteria that was being studied. For example, analysis of 39 strains of *B. anthracis* showed that variation between strains of *B. anthracis* is low, and there is a lack of evidence for recombination (Zwick *et al.*, 2011). The lack of recombination also leads to high linkage disequilibrium as detected by the LDhat program. Moreover, Zwick *et al.* (2011) studied recombination throughout the genome, and the results cannot be applied to recombination in operons. The degree of sequence conservation in operons is high. Therefore, the probability of homologous recombination is also higher.

In our research, lack of detection of recombination breakpoint may be due to the nature of the organisms studied. The *B. cereus* group forms three major clades: clade 1, consisting of mostly *B. cereus*; clade 2, consisting of *B. thuringiensis*; clade 3, consisting of *Bacillus mycoides* and *Bacillus weihenstephanensis* (Didelot *et al.*, 2009). We did not include species that did not have final, annotated genome sequence. The results also indicated that the rate of recombination in clade 3 is three times higher than in clade 2 (Didelot *et al.*, 2009). Moreover, in some species of cereus group, the rate of recombination is four to seven times less than point mutations, but the effect of recombination on their evolution is higher due to a larger number of nucleotide changes per recombination event as compared to mutations (Didelot *et al.*, 2009). Clade 3 has the highest rate of recombinations and a high number of recombinant fragments from sources other than the cereus group (Dide-

lot *et al.*, 2009). *B. anthracis* is the most clonal among the cereus group. This may have resulted in the lack of evidence for recombination within operons. However, there is evidence for LGT within operons of the cereus group. Clonal isolation and a non-zero rate of homologous recombination suggest that the few transfers of genes into operons may be due to homologous recombination. A lack of evidence of homologous recombination may be due to a lack of sequence divergence.

The process of homologous recombination in *Bacillus* requires at least five steps (Majewski and Cohan, 1998). The first step is the acquisition of the donor's DNA by the recipient. This involves methods such as transformation and conjugation. The uptake of DNA requires uptake sequences in some bacteria, but in *Bacillus* and *Streptococcus*, there is no requirement for DNA uptake sequences (Majewski and Cohan, 1998). The second step is the escape of acquired DNA from the recipient's restriction system. The 3' end of DNA is slightly degraded during transformation. However, there is no evidence that the DNA restriction system is the rate-limiting step of recombination in *Bacillus* (Majewski and Cohan, 1998). The other steps include formation of donor-recipient DNA heteroduplex; escape of heteroduplex from the mismatch repair system of the recipient; and finally, the production of functional gene product from the donor DNA (Majewski and Cohan, 1998). The fourth step involves the correction of mismatches in DNA by the mismatch repair system. It is known to prevent to integration of foreign DNA into the genome (Majewski and Cohan, 1998). For example, if the **MutS** protein of the mismatch repair system is dysfunctional or absent, the rate of recombination is high, at least in *E. coli*. (Majewski and Cohan, 1998).

In *Bacillus*, it is the third step in recombination that is the rate limiting step in the process of recombination (Majewski and Cohan, 1998). Recombinational isolation in *Bacillus* increases exponentially with an increase in sequence divergence. DNA sequence divergence leads to difficulties in the formation of heteroduplex. Hence, recombination from a distantly related source is prevented (Majewski and Cohan, 1998). In our research, the lack of evidence of recombinant fragments from donors other than *Bacillus* may be due to the inability of divergent DNA to form a heteroduplex in *Bacillus*. However, the sequences in operons are usually more conserved than the sequences that are not part of operons. It is possible that there may have been higher rates of homologous recombination within operons and evidence of homologous recombination could not be detected by current methods or may have been lost due to mutations and amelioration.

It can be argued that the recombination breakpoints identified in this study are artifacts of the repair of double stranded DNA breaks using homologous recombination. However, this is not true. The recombination breakpoints were identified near the laterally transferred genes. The probability that the recombination breakpoint were due to homologous recombinational transfer of genes is greater than the probability of homologous recombinational repair of double stranded breaks. Moreover, the NHEJ pathway can also be used to repair DNA (Weaver, 1995), and *Bacillus* can repair double stranded breaks by the NHEJ pathway (Simmons *et al.*, 2009). The breakpoints detected were not due to repair of double stranded breaks by the NHEJ pathway because the pathway is limited to sporulation and the stationary phase of the life cycle of *Bacillus*. NHEJ is not a very common way to repair DNA in *Bacillus* (Simmons *et al.*, 2009).

This research indicates that there is some evidence that homologous recombination is a mechanism of integration of laterally transferred genes into operons. The genes in operons are usually conserved in sequence. Hence, it is highly possible that sequence similarity of genes could have led to homologous recombination. It is possible that the breakpoints discovered are due to differences in substitution rate. However, the large number of breakpoints on at least one end of the laterally transferred genes suggested that there might have been homologous recombination within operons. The results suggest that the rate of recombination could have been higher than the rate of nonhomologous end joining.

Future projects may include the study recombination and its effects on the structure of operon. It could be studied whether homologous recombination is the mechanism by which genes are rearranged within the operons. Moreover, mathematical models of rates of recombination in operons can be inferred. Other projects may include detection of laterally transferred gene fragments and their mechanism of transfer into operons. Laterally transferred gene fragments can be detected by methods that use a sliding window to detect laterally transferred genes.

Chapter 3

The Effect of LGT on Neighbouring Genes

3.1 Abstract

The neutral theory of evolution proposes that most substitutions in DNA sequence of an organism are neutral with respect to the fitness of the organism. The theory predicts that the rate of synonymous substitutions would be higher than the rate of nonsynonymous substitutions because nonsynonymous substitutions may cause a deleterious change in the product of the genes. Some genes have high rates of evolution and are under relaxed selection. These include some laterally transferred genes, duplicated genes and genes involved in pathogenicity and immune system. In contrast, other genes have low rates of evolution and are under strong purifying selection. These genes include genes for essential functions such as cell division, replication, transcription and translation. There are biological processes that can affect the rate of evolution and selection on the genes neighbouring or linked to a gene under strong positive or negative selection. These include hitchhiking and background selection.

In this research, we investigate the effects of laterally transferred genes on neighbouring genes. We propose that laterally transferred genes affect the neighbouring genes by increasing the rate of evolution of neighbouring genes. The change in the rate of evolution of neighbouring genes may be due to relaxed selection on these neighbouring genes. A phylogenetic tree was constructed using high quality sequences of *Bacillus* genomes. Genes were clustered into families. The duplicated genes were ignored. The distribution of presence and absence of genes were used to infer ancestral states of the genes by parsimony principle. Hence, LGT was inferred by a change of state from absent to present. The genes that were neighbours of laterally transferred genes were obtained, based on the sequence of the genomes. The neighbouring genes that were laterally transferred or duplicated were removed. Genes that were not neighbours of laterally transferred genes were also obtained to act as controls. Again, duplicate genes and laterally transferred genes were removed. The rate of evolution and type of selection on both sets of genes were inferred by maximum likelihood methods as implemented in the CodeML program of the PAML package.

Different models of codon evolution were tested. Gene families that were under positive selection were inferred using likelihood ratio tests. Tree lengths, obtained by calculating the sum of all branch lengths, were used to indicate the rate of evolution. The dN/dS values were compared by calculating a weighted average across different site categories.

The results indicated that the rate of evolution of genes that were neighbours of laterally transferred genes, were slightly higher than the rate of evolution of non-neighbouring genes. The result could be explained by relaxed selection on the neighbouring genes due to LGT. The difference in dN/dS of neighbouring genes and non-neighbouring genes is statistically significant. It could be due to laterally transferred genes by an unknown mechanism or there could be a tendency of genes to be transferred into regions with high rates of evolution and relaxed selection.

3.2 Introduction

3.2.1 Evolution of Prokaryotes

The type and the degree of selection on a gene can be detected by calculating the ratio, dN/dS or ω , of the number of nonsynonymous substitutions per nonsynonymous sites (dN) to the number of synonymous substitutions per synonymous sites (dS) (Yang and Bielawski, 2000). If selection is neutral, then the probability of fixation of a synonymous substitution is equal to the probability of fixation of nonsynonymous substitution. Therefore, dN/dS is equal to one when selection is neutral (Hurst et al., 2002). If dN/dS is less than one, then negative selection is occurring; in which case some changes in amino acid sequence are deleterious (Hurst et al., 2002). If dN/dS is greater than 1, then positive selection is occurring, and some amino acids changes are permitted by selection (Hurst et al., 2002).

In 1964, it was realized that the number of possible alleles of a gene based on permutations of mutations is astronomical (Kimura and Crow, 1964). However, some mutations would result in a dramatic change, in protein structure, while other would have only a minor change (Kimura and Crow, 1964). Amino acids generally evolve at a very slow rate (Kimura et al., 1968). However, the rate of evolution of the entire genome was very high. When nucleotide sequences were studied, it was discovered that most substitutions were synonymous (Kimura et al., 1968). The high rate of nucleotide substitutions was explained by considering that most substitutions had no effect on the fitness of organisms and were neutral (Kimura et al., 1968). These observations later gave rise to the neutral theory of molecular evolution. Moreover, it was discovered that the rate of amino acid substitution was nearly constant in homologous proteins (Kimura, 1969). The rate of amino acid substitution was estimated based on comparison of alpha haemoglobin of mammals to that of carp. Most substitutions in proteins were due to fixation of random mutations (Kimura, 1969). Random genetic drift played a more important role than previously thought. Substitution rate depended on time measured in years and was independent of generation time, living conditions and genetic background (Kimura, 1969). It was estimated that only one amino acid substitution per amino acid site occurs every 800 million years (Kimura, 1969).

Most amino acids substitutions were not due to natural selection, but were slightly deleterious or neutral (Kimura, 1969). It was suggested that not all evolutionary change was because of Darwinian natural selection, and genetic changes were fixed in a population mainly by genetic drift (King and Jukes, 1969). Out of 549 different single base pair changes that could occur in codons, 134 were synonymous, 392 were nonsynonymous, and 23 resulted in stop codons (King and Jukes, 1969). Small errors in DNA replication could result in mutations in DNA sequence, which were fixed in the population by genetic drift or removed from the population by natural selection (King and Jukes, 1969).

In the 1970s, the neutral theory of evolution was extended to include fixation of very slightly deleterious mutations in a population (Ohta, 1973). Such mutations had a higher probability of removal than retention in the population. However, they could still exist in a population because another beneficial mutation could compensate for the deleterious effects of the first mutation. Hence, they got fixed in the population by genetic drift (Ohta, 1973). Functionally important proteins and structurally constrained proteins evolved at a much slower rate than less important proteins, suggesting that the rate of evolution was not constant throughout the genome (Ohta, 1974). It was suggested that slightly deleterious mutations were more common than previously thought (Ohta, 1974). Homozygosity was explained by considering hitchhiking of neutral substitutions on the region of chromosome experiencing directional selection (Smith and Haigh, 1974). Also, by studying mRNA sequences for Beta-hemoglobin, it was observed that synonymous substitutions occurred three times more often than nonsynonymous substitutions consistent with the neutral theory of evolution (Jukes, 1978).

In the 1980s, pseudogenes were defined as the genes that had lost their normal expression, and it was shown that pseudogenes evolved rapidly (Li *et al.*, 1981). Beneficial substitutions in genes made up a minor fraction of the total number of substitutions (Ohta, Gillespie *et al.*, 1996). Most substitutions were nearly neutral, and their evolution was due to genetic drift (Kimura, 1981).

In the 1990s, it was observed that the rate of nonsynonymous substitution increases in the genes that had acquired new functions (Ohta, 1994). Fixations in such genes might be caused by natural selection rather than genetic drift (Ohta, Gillespie *et al.*, 1996). The amino acid substitutions were episodic in nature. It was proposed that populations adapt to an environment at the local optima, but when the environment changed, the population had to adapt to the new local optima by causing a change in substitution rate. A shift in the environment could move the population off the local peak of fitness in a given mutational landscape (Gillespie, 1994).

Positive selection could inflate the rate of evolution of proteins above the level expected from polymorphisms alone (Nei, Suzuki and Nozawa, 2010). In most cases, a laterally transferred gene that did not interact with the protein complexes of the host organism was neutral with respect to fitness of the organism (Omer *et al.*, 2010). If RNA polymerase β subunit (*rpoB*) of *Bacillus subtilis* were expressed in *Escherichia coli*, then the growth rate of *E. coli* did not decrease suggesting that acquisition of *rpoB* was neutral (Omer *et al.*, 2010).

3.2.2 Factors Effecting Selection

There are several factors that can affect the efficiency of selection. I will discuss several in turn. The rate of expression of genes can be measured computationally by the codon adaptation index (CAI) (Sharp and Li, 1987). CAI is based on the observation that there is a bias in the synonymous codon usage, in highly expressed genes. There is a positive correlation between codon bias and level of expression of genes (Sharp and Li, 1987). Synonymous codons in highly expressed genes are under strong selection and have high values of CAI. Some recently transferred genes have low CAI values indicating that they are not adapted to the host genome (Sharp and Li, 1987).

The rate of expression of genes can, therefore, affect the selection on genes (Higgs, Hao and Golding, 2007). Highly expressed genes should use codons that utilize abundant tRNAs in cells. They should use amino acids with low energetic cost of synthesis. They are under strong selection (Higgs, Hao and Golding, 2007). The rate of evolution of highly expressed genes is slower than that of genes that are not highly expressed. For genes with extreme CAI values, synonymous substitutions per synonymous site decreases (Retchless and Lawrence, 2007).

Previously, it was reported that there is a decrease in dN/dS over time in bacterial genomes, especially in laterally transferred genes (Castillo-Ramirez *et al.*, 2011). The genomes of recently (after 1950) diverged strains of *Staphylococcus aureus* and *Clostridium difficile* can be divided into core regions and noncore regions (Castillo-Ramirez *et al.*, 2011). The core regions are conserved regions with a low rate of recombination, while the noncore regions have a high rate of recombination and LGT. Analyses of SNPs in recently studied genomes indicate that SNPs in the core region are highly clustered and have a lower proportion of synonymous changes than those in the noncore regions (Castillo-Ramirez *et al.*, 2011). The core regions have higher dN/dS than the noncore regions. The reason for the low number of synonymous site changes in the core region. If recombination is considered, then a larger proportion of nonsynonymous site can be expected. The majority of the SNPs that were transferred into genomes, have a larger proportion of nonsynonymous changes as compared to the core regions of the recently emerging genomes.

Previously, the rates of evolution and selection on laterally transferred genes were studied (Hao and Golding, 2006). Genes that are recently transferred have higher dN/dS ratios in *Bacillus* (Hao and Golding, 2006). High dN/dS ratios could be due to transfer of genes that are evolving at a higher rate or relaxed selection constrains or amelioration of recently acquired genes (Hao and Golding, 2008b). The rates of insertions and deletion of genes are high in prokaryotes resulting in high gene turnover rate (Gogarten and Townsend, 2005). Most laterally transferred genes are nearly neutral while others increase the fitness of bacteria (Gogarten and Townsend, 2005). The rate of evolution and dN/dS ratio of recently transferred genes are also higher in *Streptococcus* (Marri, Hao and Golding, 2006).

The rate of evolution of duplicated genes is higher than that of other genes in an organism (Ohta, 1994). A high rate of amino acid substitution is reflected in a high dN/dS ratio, indicating that the duplicated genes are diverging rapidly (Gu *et al.*, 2002). The high rates of evolution in recently duplicated genes suggest that either the duplicated genes are under positive selection, or the functional constraint has been reduced (Van de Peer *et al.*, 2001).

Evidence of positive selection cannot be easily found in ancient duplicated genes (Van de Peer *et al.*, 2001); and hence, the reduction of functional constraint is usually invoked.

The size of the population can also affect the dN/dS value (Jordan *et al.*, 2002). Lower average dN/dS in *E. coli* genome, as compared to pathogenic bacteria may be an indication that purifying selection in a large population (*E. coli*) is more effective than in a small population (pathogenic strains) (Jordan *et al.*, 2002). Essential genes have lower dN/dS values than non-essential genes indicating a higher level of purifying selection on essential genes (Jordan *et al.*, 2002). The dN/dS ratio also depends on the functional classes. Some functional classes, based on COG annotations, have higher dN/dS values than those in other functional classes (Jordan *et al.*, 2002).

The dN/dS value is higher between closely related species than in distantly related species (Rocha *et al.*, 2006). This effect can be explained by considering the time lag between acquisition of slightly deleterious mutations and removal of them from the genomes (Rocha *et al.*, 2006). The dN/dS ratio decreases over time and then reaches a plateau. The high dN/dS ratio could be due to a change in niche or hitchhiking (Rocha *et al.*, 2006). This all suggests that dN/dS ratios should be interpreted with care.

Hitchhiking is a process by which a locus increases in frequency because it is linked to another locus under positive selection. By hitchhiking nonsynonymous mutations can increase in frequency and reach fixation (Rocha *et al.*, 2006). Computer simulations indicate that hitchhiking increases nonsynonymous substitutions and results in an increase of the dN/dS ratio (Rocha *et al.*, 2006).

The genes that cause pathogenicity have high rates of evolution and are under positive selection (Williamson, 2003). In HIV-1, the *env* gene codes for envelope proteins and shows high rate of nonsynonymous mutation and is under positive selection. The ancestral sequences of *env* can be reconstructed accurately due to high availability of sequence data of *env* (Williamson, 2003).

3.2.3 Detecting selection

There are two different classes of methods for detecting type and degree of selection on genes. The first type of method requires counts of synonymous and nonsynonymous substitutions in two sequences Yang and Bielawski (2000). The data is corrected for multiple substitutions at the same site. An equal rate of transitions and transversions and uniform codon usage are also assumed by some methods (Yang and Bielawski, 2000). However, these assumptions are not always accurate. The rate of transitions is generally greater than the rate of transversions. Moreover, codon usage can have unpredictable effects on estimation of ω (Yang and Bielawski, 2000).

The second class of methods includes likelihood methods (Yang and Bielawski, 2000). These methods allow the estimation of parameters, such as the ratio of transition vs. transversion κ , using maximum likelihood methods. If the transition vs. transversion ratio is ignored, then sites with synonymous substitutions are underestimated, resulting in over estimation of dS and under estimation of ω value (Yang and Bielawski, 2000). If the codon usage bias is ignored, then sites with synonymous substitutions are over estimated, leading

to under estimation of dS and overestimation of ω value Yang and Bielawski (2000). These methods also reconstruct ancestral sequences on a phylogeny to estimate ω value (Yang and Bielawski, 2000). A brief overview of some methods of estimation of dN/dS ratio developed over that past few years is given below.

The synonymous and nonsynonymous rates can be calculated by counting the number of synonymous substitutions per synonymous site and nonsynonymous substitutions per nonsynonymous site (Li, Wu and Luo, 1985). Nucleotide sites can be classified into three types: *a*) non degenerate *b*) two-fold degenerate and *c*) fourfold degenerate based on the results of nucleotide substitution (Li, Wu and Luo, 1985). A site is nondegenerate if all substitutions at the site are nonsynonymous. It is two-fold degenerate if one of the three substitutions at the site are synonymous. It is fourfold degenerate if all substitutions at the site are nondegenerate (Li, Wu and Luo, 1985). The likelihood of change of the codon is calculated by reconstructing ancestral states by parsimony (Fitch, 1971). The expected frequencies of codon change are also calculated. Observed and expected changes are studied, and dN/dS is calculated (Li, Wu and Luo, 1985).

The methods of calculation of dN/dS tend to underestimate the number of nonsynonymous substitutions, when the number is large (Nei and Gojobori, 1986). Counting the number of synonymous and nonsynonymous substitutions in sequences is not the most accurate way of determining dN/dS because there may be multiple substitutions at some sites (Nei and Gojobori, 1986). However, the synonymous and nonsynonymous changes can still be counted, and multiple substitutions can be corrected by assuming that each pathway of changes occurs at an equal probability and can be weighed equally. The proportion of synonymous and nonsynonymous substitutions can be calculated and corrected using Jukes Cantor formula (Jukes and Cantor, 1969). Finally, the dN/dS ratio can be calculated (Nei and Gojobori, 1986).

Transitions often occur at a higher rate than transversions. Hence, the dS value could be overestimated because most transitions at two-fold degenerate sites are synonymous (Li, 1993). The Li, Wu and Luo (1985) method can be modified to include the transition vs. transversion rate bias. Sites are classified into nondegenerate, two-fold degenerate and fourfold degenerate (Li, 1993). Codons are classified based on the type of sites and transition and transversion rates (Li, 1993). Transitions and transversion can be estimated from Kimura's two-parameter model (Kimura, 1980). Finally, dN and dS can be calculated by counting 1/3 of the two-fold degenerate sites as synonymous and 2/3 as nonsynonymous (Li, 1993).

Instead of counting synonymous and nonsynonymous substitutions, the dN/dS ratio can be estimated using a Markov process model of codon substitution (Goldman and Yang, 1994). Codons can be modelled as states of the Markov process. A 61 by 61 rate matrix can be constructed by considering a change of nucleotides in codons, a transition/transversion ratio (κ), a scaling factor, the distance between amino acids and the variability of sequences (Goldman and Yang, 1994). The parameters are inferred by a maximum likelihood method. The Markov process can be used to estimate the dN/dS ratio (Goldman and Yang, 1994). The method includes a transition versus transversion rate bias and a codon usage bias. However, it does not allow for multiple substitutions per codon because the probability of

such an event is very small based on this model (Goldman and Yang, 1994).

A method for detection of positive selection on amino acids is given by Suzuki and Gojobori (1999). In this method, a phylogenetic tree is constructed by neighbour joining method. For each codon, the ancestral codon is inferred by parsimony or maximum likelihood. Average number of synonymous and nonsynonymous substitutions is estimated over time. Finally, the dN/dS ratio is calculated (Suzuki and Gojobori, 1999).

Yang and Nielsen (2000) calculates dN/dS ratio by taking into account transition vs. transversion rate bias and nucleotide or codon usage bias. This is an approximation method suitable for large datasets. The HKY85 model is used to correct for transition vs. transversion rate bias and nucleotide and codon usage bias (Hasegawa, Kishino and Yano, 1985). Different pathways of evolution of codons are estimated based on their relative probabilities because equal weights of pathways of codon evolution tend to overestimate ω if it is greater than one and underestimate ω if it is less than one (Yang and Nielsen, 2000). The method works by calculating probabilities of different pathways and correcting for multiple substitutions. These steps are repeated until the algorithm converges (Yang and Nielsen, 2000). This method is faster than the maximum likelihood method (Goldman and Yang, 1994) and more accurate than the Nei and Gojobori (1986). The latter is less accurate for short sequences and high ω ratios (Nei and Gojobori, 1986).

The neutral theory of molecular evolution suggests that most mutations should have no effect on the fitness of the organisms and are fixed in the population due to random drift (Yang, 2002). It is implied that the dN/dS ratio will be equal to one in those cases. However, the dN/dS ratio is less than one for most genes due to averaging of dN/dS over the full length of the alignment (Yang, 2002). This effect can be minimized by focusing on a single branch and using different rate categories for different sites (Yang, 2002). Two different branch-site models can be used. These include 3-ratio models that assume two different omega values for two different branches and all the other branches have the same omega value (Yang, 2002). In the free ratio model, omega values are independent of each other for each branch in the phylogeny (Yang, 2002). Hence, the problem of averaging omega values over entire phylogeny and gene lengths can be avoided (Yang, 2002). There is evidence for episodic positive selection acting along some ancient branches of phylogenetic trees (Chang and Donoghue, 2000). The ancestral sequences can be artificially engineered, and selection can be studied in a laboratory.

Noncoding regions, such as those that control the transcription of genes, may be under selection (Wong and Nielsen, 2004). It is important to determine the type of selection on noncoding regions. It can be assumed that the rate of substitutions in noncoding regions correlates with the rate of synonymous substitution in the coding regions (Wong and Nielsen, 2004). The appropriate unit to model evolution is assumed to be the nucleotide instead of the codon. Selection can be inferred using an HKY85 matrix (Hasegawa, Kishino and Yano, 1985). The Greek letter ζ is used to represent the type of selection. The selection is negative if ζ is less than one; neutral if ζ is equal to one; and positive if ζ is greater than one (Wong and Nielsen, 2004). Based on this model, there is little evidence for positive selection in noncoding regions even if the noncoding regions are between genes that are under positive selection. Hence, the method lack power to detect selection. The

method assumes that there is no selection on the rate of synonymous substitutions, which is questionable when there is codon usage bias (Wong and Nielsen, 2004).

Likelihood methods have some limitations. These methods can only detect positive selection by considering the dN/dS ratio. Selection that is due to multiple alleles, such as heterozygote advantage, cannot be detected by the standard maximum likelihood methods (Yang and Bielawski, 2000). Maximum likelihood methods also assume consistency of selection pressure over sites (Yang and Bielawski, 2000).

In addition, if there is recombination within genes, likelihood methods do not give correct results (Yang and Bielawski, 2000). Standard maximum likelihood methods assume that the tree topology is accurate. However, the tree topology may not be accurate because the species tree, after recombinations because of recombination events (Anisimova, Nielsen and Yang, 2003). High rates of recombination can increase type I error rate up to 90% (Anisimova, Nielsen and Yang, 2003). Trees inferred from recombinant regions having a starlike topology can lead to false positives in recombination. In summary, recombination can lead to false positive detection of positive selection.

In the standard likelihood ratio methods, models with multiple rate categories are used, and different ω ratios can be distributed among different sites. These models include the following:

Models	Model Type	Description
M0	One ratio	One ω ratio for all sites
M1a	Neutral	Two ratios: $\omega < 1$ and $\omega = 1$
M2a	Positive selection	Three ratios: $\omega < 1$ and $\omega = 1$ and $\omega > 1$
M3	Discrete	Three ratios: $\omega > 0$ for all sites
M7	Beta distributed	Multiple ratios (Default = 10): $\omega < 1$ for all
M8	Beta and ω	Similar to M7 with one class of $\omega > 1$

These models are nested. There is an alternate model for each null model (Nielsen and Yang, 1998). Likelihood ratio tests can be used to study which model best fits the data. The Likelihood ratio test can be done on the following models:

Test	Null Model	Alternate Model
M0 vs. M3	M0	M3
M1a vs. M2a	M1a	M2a
M7 vs. M8	M7	M8

At low sequence divergence, the likelihood ratio test is conservative and cannot always detect positive selection (Nielsen and Yang, 1998). If the sample size is small, a maximum likelihood estimator has a positive bias (Nielsen and Yang, 1998). The distribution of ω becomes narrower if the number of nonsynonymous sites increases (Anisimova, Bielawski and Yang, 2001). The degree of sequence divergence has a significant effect on the power of the likelihood ratio test. If the sequence divergence is low, maximum likelihood parameters under M3 are overestimated (Anisimova, Bielawski and Yang, 2001). The power of maximum likelihood ratio test increases as the number of synonymous sites increases. It peaks

at an intermediate number of synonymous sites, and decreases as sequences become highly divergent (Anisimova, Bielawski and Yang, 2001). The length of sequences also affects the power of the test. The test has higher power, if the sequences are longer (Anisimova, Bielawski and Yang, 2001). Not all models detect positive selection with equal efficiency. M8 can detect more sites under positive selection than M3. Therefore, multiple models should be implemented (Anisimova, Bielawski and Yang, 2001).

3.2.4 Detection of sites under positive selection

A number of different rate categories can be used to calculate ω using the maximum likelihood method (Anisimova, Bielawski and Yang, 2002). Posterior probabilities can be calculated using Bayes theorem. Simulations suggest that the accuracy of Bayes predictor can be increased by using a large number of lineages because the prediction is not accurate for a small number of closely related lineages. Tree lengths can indicate the sequence divergence. Genes are considered to be under positive selection if the likelihood ratio test suggests that positive selection model fits on the genes, and at least one of the estimates of ω is greater than one (Anisimova, Bielawski and Yang, 2002).

3.2.5 Processes that affect neighbouring genes

In bacterial genomes, there are processes on certain genes that affect neighbouring genes. The processes include genome rearrangements by insertion sequences and deletion of pseudogenes (Kuo and Ochman, 2010). Insertion sequences are common in prokaryotic genomes (Siguier, Filee and Chandler, 2006). It is observed that less than three percent of the bacterial chromosome consists of insertion sequences (Siguier, Filee and Chandler, 2006). These insertion sequences are responsible for genome rearrangements and translocation of genes that are neighbouring the insertion sequences (Siguier, Filee and Chandler, 2006).

Genetic hitchhiking is the process by which a locus increases in frequency, in a population, by being in linkage disequilibrium with a gene under positive selection (Smith and Haigh, 1974). Although hitchhiked locus may not necessarily be physically close to the gene it is linked to, the effect is generally more pronounced on neighbouring genes. Regions of the genome that have low rates of recombination can have low variation in the population because genetic hitchhiking may have caused an increase in the frequency of the locus in the population (Fay and Wu, 2000). If recombination is absent, hitchhiking can remove a majority of variations in the population (Fay and Wu, 2000). Therefore, a gene under positive selection can affect the neighbouring loci if they are genetically linked to the gene under positive selection.

Besides hitchhiking, background selection can also act on neutral variation in DNA sequence (Charlesworth, Morgan and Charlesworth, 1993). When neutral variation is removed from the population by being linked to deleterious mutations, the effect is called background selection. If the rate of recombination is high, the background selection is low (Charlesworth, Morgan and Charlesworth, 1993). For background selection to work effec-

tively, deleterious alleles must be under strong, negative selection. Otherwise, the allele acts like a nearly neutral allele and does not affect the linked neutral variation (Charlesworth, Morgan and Charlesworth, 1993). Moreover, if the population size is small, the removal of linked variation is rapid. If the linkage is tight, background selection is even more rapid (Charlesworth, Morgan and Charlesworth, 1993). Therefore, background selection can affect the evolution of neighbouring alleles if they are linked to the gene under negative selection.

Genetic draft is defined as a stochastic force acting on closely linked locus due to selection on other loci (Gillespie, 2001). Population size can affect the rate of genetic draft. If the advantageous mutation is weakly selected, then the rate of its substitution decreases with an increase in the population size. However, if the mutation is deleterious, its substitution rate decreases as the population size increases (Gillespie, 2001). Genetic draft is a term to encompass both hitchhiking and background selection.

Another process that can affect laterally transferred genes is specialized transduction (Morse, Lederberg and Lederberg, 1956). Specialized transduction is the process that allows a bacteriophage to transfer neighbouring genes from one bacterium to another (Morse, Lederberg and Lederberg, 1956). The process does not directly affect the rate of evolution of neighbouring genes. However, if the bacteriophage genes are laterally transferred then the flanking genes often get transferred to another host. There, they have a higher rate of evolution, and some of them are under relaxed selection (Marri, Hao and Golding, 2007). Hence, bacteriophage can also affect neighbouring genes.

Pseudogenes are described as those genes that have lost their normal biological function. It was believed that evolution of pseudogenes is random because of loss of expression (Kuo and Ochman, 2010). However, a recent study suggests that the product of pseudogenes may have deleterious effects on the bacterium and require the use of transcriptional and translational energy (Kuo and Ochman, 2010). Most pseudogenes, defined as short open reading frames between conserved genes, have only a few mutations (Kuo and Ochman, 2010). After a few mutations, pseudogenes are deleted from the genomes. There is a high deletion rate of pseudogenes in order to maintain small genome size and remove genes that reduce fitness (Lawrence, Hendrix and Casjens, 2001). Deletion of pseudogenes may be due to their deleterious effects and not due to a fitness advantage of small genome size because large (8 - 11 Mbp) chromosomes also exist in nature (Lawrence, Hendrix and Casjens, 2001). Deletion mutation is the most common mutation leading to pseudogenization. Moreover, large deletions seem to be the mechanism of deletion of pseudogenes and it may delete neighbouring genes, as well (Kuo and Ochman, 2010).

3.3 Method

3.3.1 Phylogenetic Tree

Genome sequences of 47 genomes of *Bacillus*, *Anoxybacillus*, *Geobacillus*, and *Oceanobacillus* were downloaded from NCBI: <ftp://ftp.ncbi.nih.gov/genomes/>. Orthologs of conserved genes *gltX*, *nusA*, *pheS*, and *rpoA* were searched in

all the genomes using the reciprocal best hit of BLAST (Altschul *et al.*, 1997). These genes were selected because they were highly conserved and have not been laterally transferred (Hao and Golding, 2008b). The DNA sequences of these genes were aligned using *MUSCLE* (Edgar, 2004). The DNA alignments were concatenated, and the resulting data set was used to construct a phylogenetic tree in MrBayes (Huelsenbeck and Ronquist, 2001). MCMC analysis was run on multiple CPU cores (Altekar *et al.*, 2004). The total number of characters in the alignment was 4,778. The model used in MrBayes was GTR + Γ + I. The outgroup was *Oceanobacillus iheyensis* HTE831. The number of MCMC generations was 10,000,000, the sampling frequency was 100, and the burnin number was set to 25,000.

3.3.2 Detection of LGT

Genes were clustered into families by MCL (van Dongen, 2000). An all-against-all *BLAST* was carried out on all 47 genomes (Altschul *et al.*, 1997). MCL was used with a bit score cut-off of 50 bits and an inflation parameter of 2.5 (van Dongen, 2000). The results of MCL clustering were converted into a presence-absence matrix where 1 represented the presence of a gene and 0 represented the absence of a gene. Genes with duplicates and multicopy genes were removed because duplicated genes might be under divergent selection based on likelihood ratio tests (Bielawski and Yang, 2004). The presence-absence matrix was used to infer LGT using parsimony (Felsenstein, 1989). Standard parsimony assumes that the rate of gain of a trait or gene was equal to the rate of loss of trait (Fitch, 1971). A change of state from 0 to 1 was considered LGT, while a change of state from 1 to 0 was considered deletion. If the ancestral states of a family were ambiguous, then those genes were not included in the study. As a result, lists of laterally transferred genes in each of 47 genomes were obtained.

3.3.3 Identification of Non-neighbouring Genes

In this study, non-neighbouring genes were defined as genes that were not laterally transferred, based on the parsimony method. They were not multi-copy based on MCL clustering. They were not upstream or downstream of a laterally transferred gene, based on NCBI annotation and LGT detection by parsimony (Felsenstein, 1989). The non-neighbouring genes served as a control set in the study. Non-neighbouring genes might not necessarily have conserved sequences, conserved chromosomal positions or copies in all 47 genomes.

3.3.4 Identification of Neighbouring Genes

To study the effects of LGT on neighbouring genes, the genes that were upstream and downstream of laterally transferred genes had to be identified. NCBI annotations and the list of laterally transferred genes were used to identify neighbouring genes. An upstream neighbouring gene was defined as a gene that was upstream of a laterally transferred gene. It was not a duplicated gene. It was not laterally transferred itself. It was not immediately

downstream of another laterally transferred gene. A downstream neighbouring gene was defined as a gene that was downstream of a laterally transferred gene. It was not a duplicated gene. It was not itself laterally transferred. It was not immediately upstream of another laterally transferred gene. These neighbouring genes as obtained by using NCBI annotations, lists of laterally transferred genes and multicopy genes as inferred using MCL (van Dongen, 2000).

3.3.5 Detection of Rates of Evolution and Positive Selection

Rates of evolution and positive selection were detected using CodeML (Yang, 2007). CodeML is part of the PAML software package for studying molecular evolution (Yang, 2007). PAML version 4.5 was used with improved M1 and M2 models of codon evolution (Yang, 1997), referred to as M1a and M2a models in the new versions (Yang, 2007). In M1 and M2, ω_0 was fixed at 0. In M1a and M2a, ω_0 could vary between 0 and 1. CodeML was used to calculate tree lengths, ω values and log likelihoods of all families under models M0, M1a, M2a, M3, M7 and M8 (Yang, 2007). Protein alignments of each gene family for all non-neighbouring genes, upstream genes and downstream genes were done in MUSCLE (Edgar, 2004). The alignments were converted into codon alignments using DNA sequences from NCBI. PAL2NAL was used to convert protein alignments into codon alignments (Suyama, Torrents and Bork, 2006). The advantage of this program is that it works even if there are mismatches and inconsistencies between protein sequences and DNA sequences. It could also detect poly-A tails (Suyama, Torrents and Bork, 2006). The alignments were used to make neighbour joining trees using the PHYLIP software package (Felsenstein, 1989). The codon alignments and the neighbour joining trees were used to study rates of evolution and selection in CodeML (Yang, 2007). CodeML was run with models M0, M1a, M2a, M3, M7 and M8 (Yang, 2007).

3.3.6 Likelihood Ratio Test

Likelihood ratio tests were used to determine which of the models: M0, M1a, M2a, M3, M7 and M8 fitted the data (Anisimova, Nielsen and Yang, 2003). The tests were M0 vs. M3, M1a vs. M2a, M7 vs. M8. In these tests, M0, M1a and M7 were the null models, while M3, M2a and M8 were the alternate models (Anisimova, Nielsen and Yang, 2003). The degrees of freedom were calculated by taking the difference of the number of parameters between the alternate model and the null model. The test statistic, D, was calculated as follows: $D = 2 * (\ln L_A - \ln L_0)$, where $\ln L_A$ is the log likelihood of the alternate model and $\ln L_0$ is the log likelihood of the null model. A χ^2 test with significance of 0.05 was conducted to determine which model best explained the data. The test M0 vs. M3 was used to determine whether the data could be explained by a model with multiple site categories (M3) or a model with a single site category (M0) (Anisimova, Nielsen and Yang, 2003). The tests, M1a vs. M2a and M7 vs. M8 were used to detect positive selection. If M2a and M8 were supported over M1a and M7 respectively, then the genes in question were likely to be under positive selection (Anisimova, Nielsen and Yang, 2003). In summary, likelihood

ratio tests were done for conserved genes, upstream genes and downstream genes.

3.3.7 Rates of Evolution and Magnitude of Selection

The rate of evolution was also determined by *CodeML* (Yang, 2007). However, the trees of non-neighbouring genes could not be compared directly with the trees of upstream or downstream genes because the number of species included might be different. The problem was solved by dropping the leaves of the trees that were not common in both downstream or upstream trees. The leaves were dropped using the ‘ape’ package (Paradis, Claude and Strimmer, 2004). The R statistics package was used to drop leaves and reorganize the phylogenetic trees (Team, 2011). Extreme outliers were removed from the data.

The dN/dS values were obtained from *CodeML*, and weighted averages of the dN/dS values were obtained for all models and all genes. The averages were weighted according to the proportion of sites in each site category. The weighted averages were used to avoid the bias caused by high dN/dS values shared by only a few sites in the genes. Extreme outliers were removed from the data.

3.3.8 Wilcoxon Rank Sum Test and Permutation Test

In order to test whether the distributions of the rate of evolution and dN/dS values were significantly different between upstream genes vs. non-neighbouring genes, downstream vs. non-neighbouring genes and upstream vs. downstream genes, two-sided permutation tests were used. In permutation tests, the order of samples was randomized, and the difference between the two distributions were studied. Permutation tests do not assume that the data is distributed according to the normal distribution. However, it is assumed that the data can be randomized. The permutation tests were carried out in the R statistics program (Team, 2011) using the ‘coin’ package’ (Hothorn *et al.*, 2006). The tests were conducted using 1000 Monte Carlo resamples. The P-values and the confidence intervals are shown in the results.

In addition to permutation tests, the two-sided, unpaired Wilcoxon rank sum tests were used to test whether the distributions of dN/dS were significantly different between upstream genes and non-neighbouring genes, downstream genes and non-neighbouring genes, and upstream genes and downstream genes (Wilcoxon, 1945). Since independent samples were used, unpaired Wilcoxon rank sum tests were carried out. The P-value of Wilcoxon rank sum test was calculated in R statistics package (Team, 2011).

A flowchart summarizing the methods follows:

3.3.9 Summary

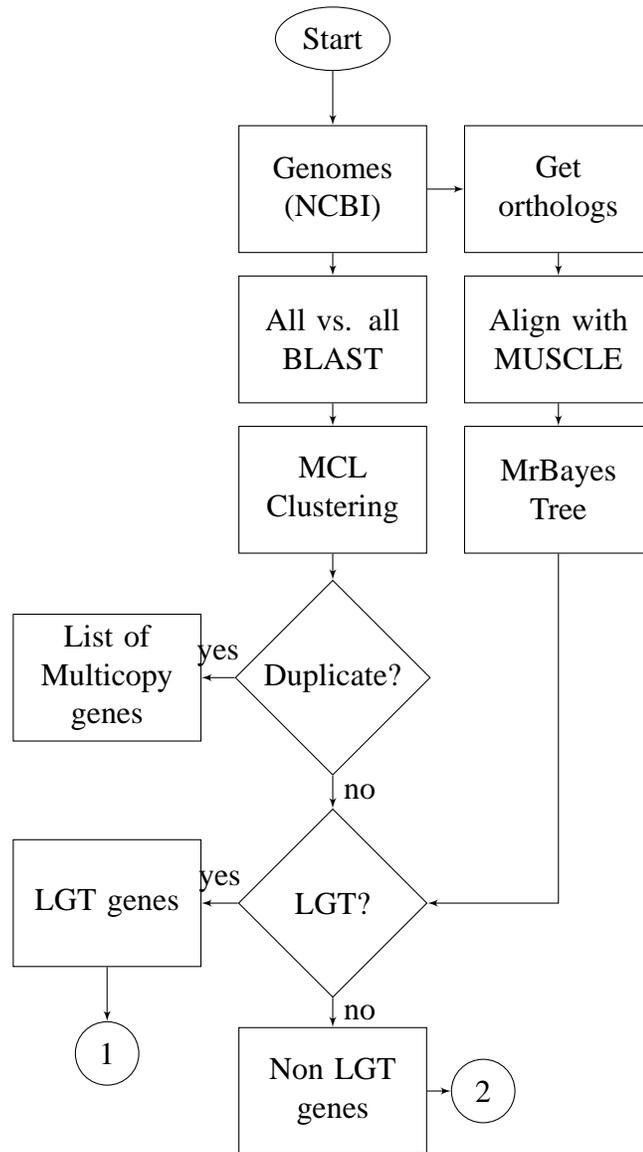


Figure 3.1: Summary of methods

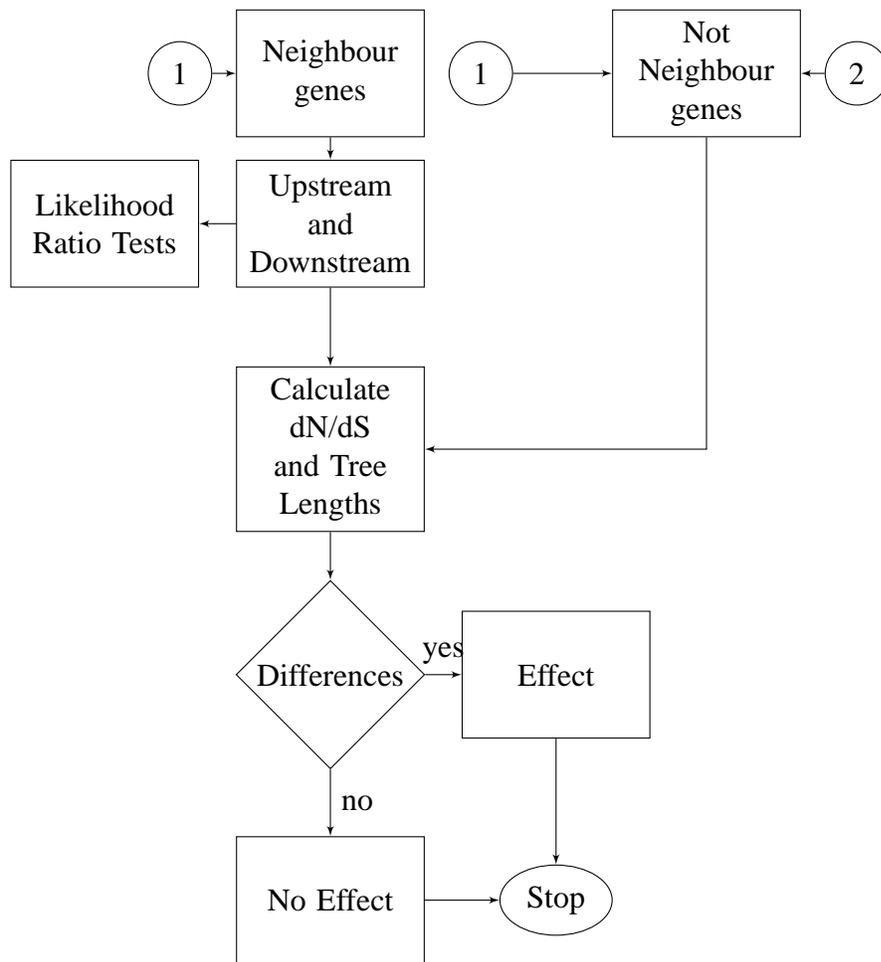


Figure 3.2: Summary of methods (continued)

3.4 Results

3.4.1 Phylogenetic tree

A phylogenetic tree was inferred using *MrBayes* (Huelsenbeck and Ronquist, 2001). The graphs of the log likelihood indicated that the log likelihoods were constant over samples showing that the topology had converged. Two runs of MCMC analysis were carried out, and both had converged at the end. The tree indicating posterior probabilities is shown in figure 3.3. The tree indicating branch lengths is shown in figure 3.4.

3.4.2 Detection of LGT

MCL clustering algorithm was used to cluster proteins into families (Enright, Van Dongen and Ouzounis, 2002). The number of clusters obtained was 10,277. Multicopy gene families were separated from these clusters. There were 2,610 families with multicopy genes. The number of gene families with single-copy genes was 7,667. These included families with laterally transferred genes, deletions and other complicated scenarios such as LGT followed by deletions or deletion followed by LGT. The number of laterally transferred genes in each genome was inferred by parsimony (Fitch, 1971), and the results are shown in the appendix.

3.4.3 Non-neighbouring, upstream and downstream genes

The genes defined as non-neighbouring were estimated from the 1,608 families that did not have any laterally transferred genes. The number of families of non-neighbouring genes was 1,255. This number does not include unique genes in all 47 genomes. Most of these families had representatives in less than 30 genomes. A phylogenetic algorithm requires at least three different sequences. Therefore, only the families that had at least three sequences were selected for further analysis. The number of gene families selected for further analysis was 890.

A number of gene families for upstream genes and downstream genes were also obtained. There were 241 gene families with at least three genes and upstream of laterally transferred genes. The number of gene families that were downstream of laterally transferred genes were 218 also consisted of at least three genes.

3.4.4 Likelihood Ratio Tests

Likelihood ratio tests were used to study which of the selection models best explained the data (Nielsen and Yang, 1998). The results of the likelihood tests are shown in the table 3.1.

The results of the likelihood ratio tests of models M0 vs. M3 indicated that for all three data sets (upstream, non-neighbouring and downstream), a larger number of genes fitted the model with multiple sites (M3) than the model with one site (M0). The proportion of gene families that fitted the multiple site model (M3) as opposed to the single site model (M0) was even higher in non-neighbouring genes than in upstream genes and downstream

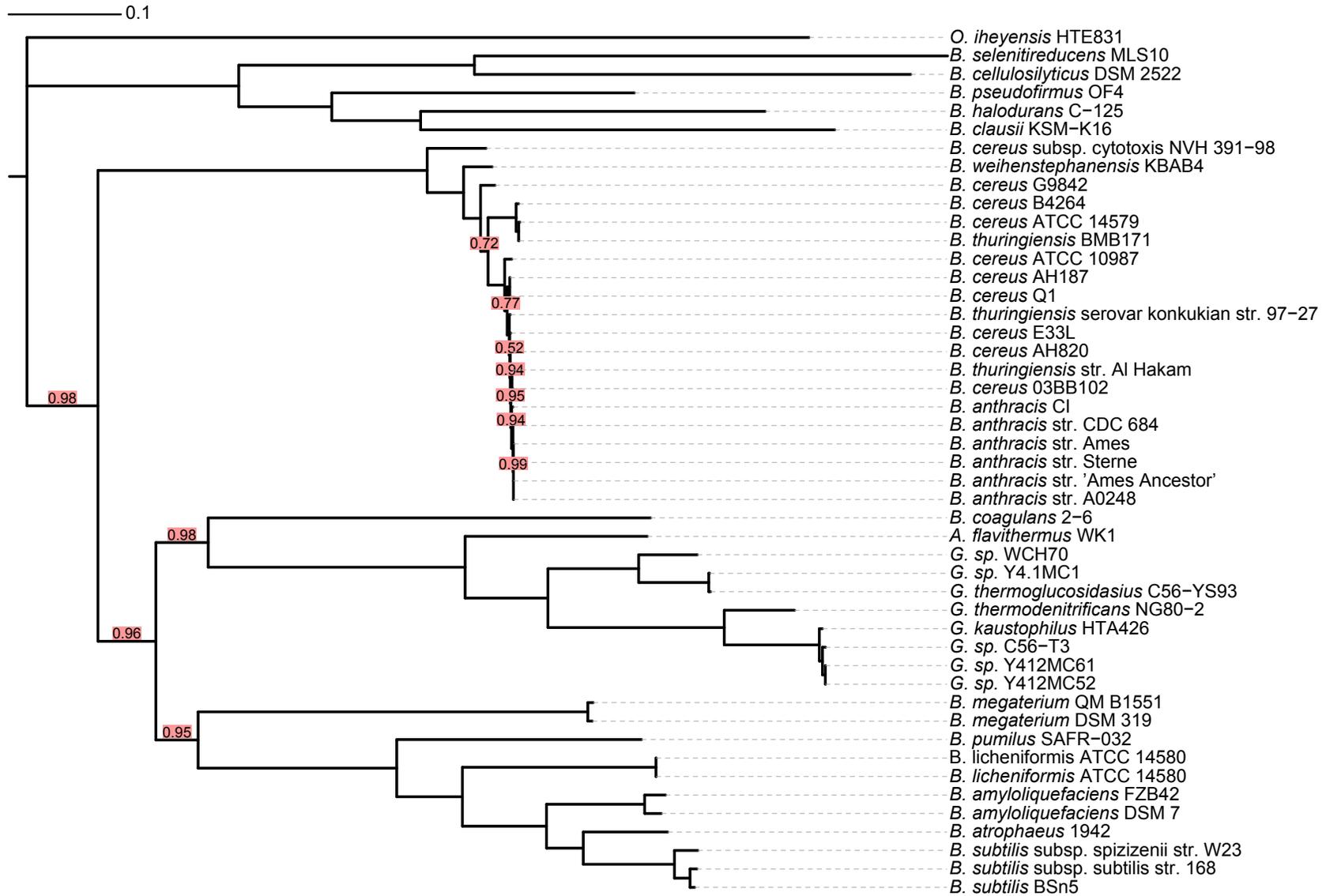


Figure 3.3: Phylogenetic tree showing posterior probabilities that were less than 1.0 plotted using iTOL (Letunic and Bork, 2006). The scale bar indicates branch lengths.

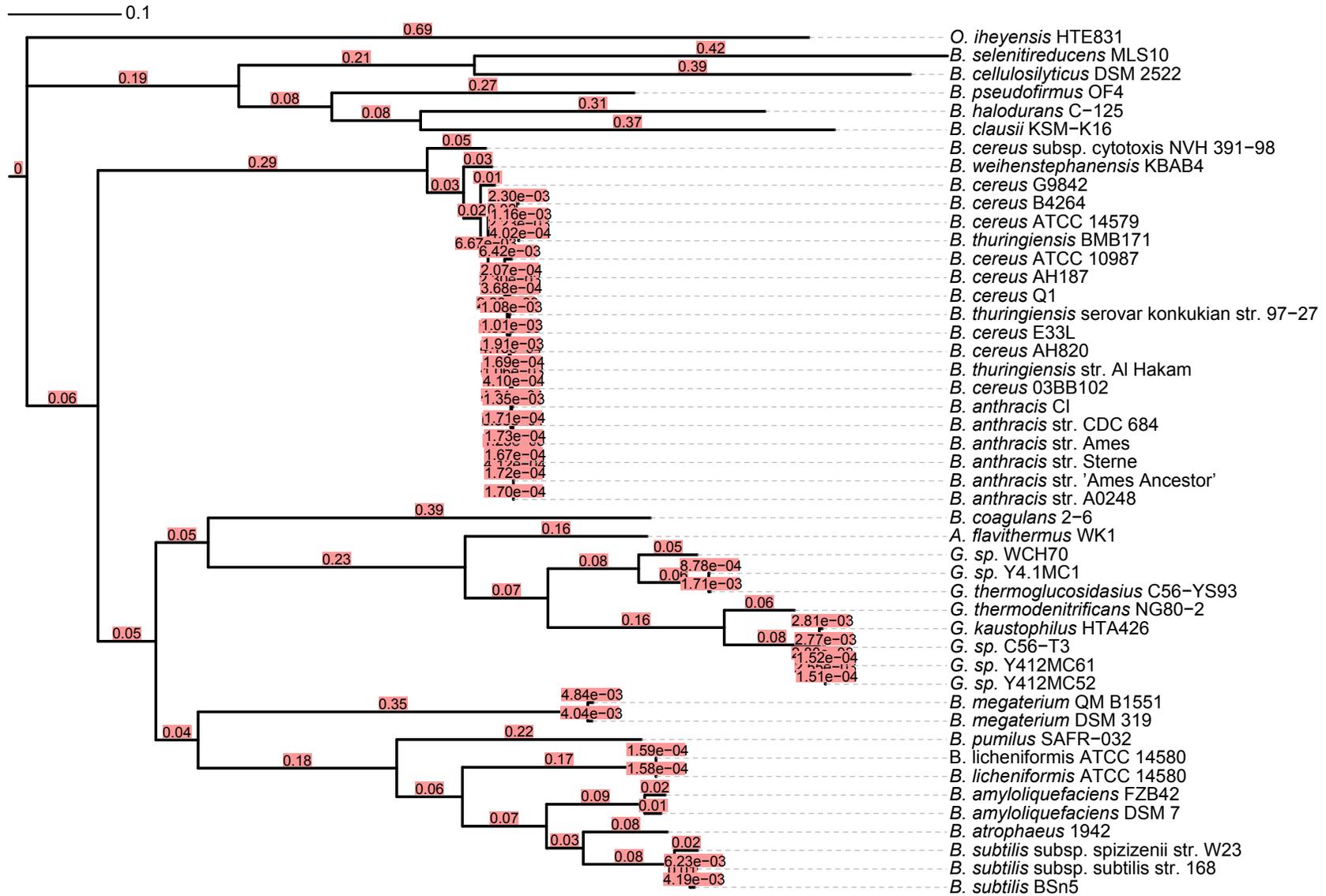


Figure 3.4: Phylogenetic tree showing branch lengths plotted using iTOL (Letunic and Bork, 2006). The scale bar indicates branch lengths.

Table 3.1: Table showing results of likelihood ratio tests

Model	Upstream	Non-neighbouring	Downstream
M0	76	37	82
M1a	233	881	214
M2a	5	7	4
M3	162	851	136
M7	211	793	199
M8	27	95	19

genes. Genes under positive selection were detected by testing M1a vs. M2a and M7 vs. M8. The more complicated model (M8) detected more genes potentially under positive selection than the lesser complicated and more conservative model (M2a). The results indicate that there is a small proportion of genes in which positive selection could potentially be detected. In M1a vs. M2a test, a larger proportion of upstream and downstream genes were potentially under positive selection than that in non-neighbouring genes.

3.4.5 Rate of evolution

The rate of evolution was studied using the tree lengths as calculated by CodeML (Yang, 2007). The tree length was defined as the sum of all branch lengths of the tree. Box plots showing branch lengths for upstream, non-neighbouring and downstream genes are given below (Figures 3.5 3.6 3.7).

The box plot of the tree lengths of the genes upstream of laterally transferred genes had a mean and median tree length was less than one. There were a few outliers in the data indicating that, in some cases, the rate of evolution was higher than average. The mean and median were relatively equal across different models of codon evolution (M0, M1a, M2a, M3, M7 and M8).

The box plot of tree lengths of the genes that were not neighbours of laterally transferred genes had the mean and median of tree lengths were between 0.5 and 1.0. As compared to upstream genes, the non-neighbouring genes lacked outliers that were evolving at a very high rate. However, similar to the upstream genes, the mean and median of tree length distributions were relatively equal across different models of codon evolution (M0, M1a, M2a, M3, M7 and M8).

The box plot of tree lengths for the genes downstream of laterally transferred genes had mean and median of the tree lengths were close to 1.0. Similar to upstream genes, but in contrast to non-neighbouring genes, there were some outliers in the data. The mean and median of tree lengths were consistent across different models (M0, M1a, M2a, M3, M7 and M8).

To determine if whether the difference is significant, different parametric and nonparametric tests could be used. Some tests such as a t-test assume that the data is normally

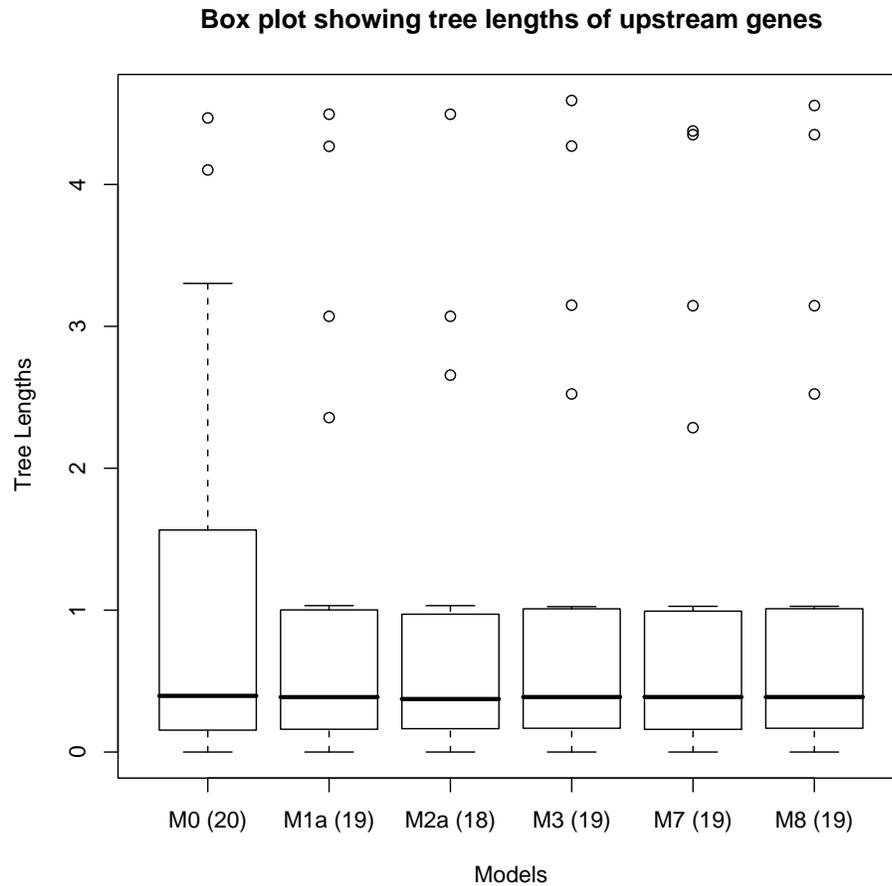


Figure 3.5: A box plot showing tree lengths of upstream genes. The numbers in the brackets indicate sample sizes

distributed. To avoid this assumption, permutation tests were used. Permutations tests were done to study whether that the difference between the distribution of tree lengths of non-neighbouring genes and upstream or downstream genes was significant. The permutations test randomized the data by making an assumption that the data could be randomized. The following table showed the results of permutation tests showing the P-value.

The results of the permutation test indicate the there is no significant difference in the distribution the tree lengths of upstream, downstream and non-neighbouring genes. All P-values are less than 1 but greater than 0.10. Some of then were close to 0.10 such as the P-value of M7 model of test between downstream genes vs. non-neighbouring genes that is 0.251. Others were close to one such as the P-values of M3 model of upstream genes and downstream genes. In that case, it was 0.967.

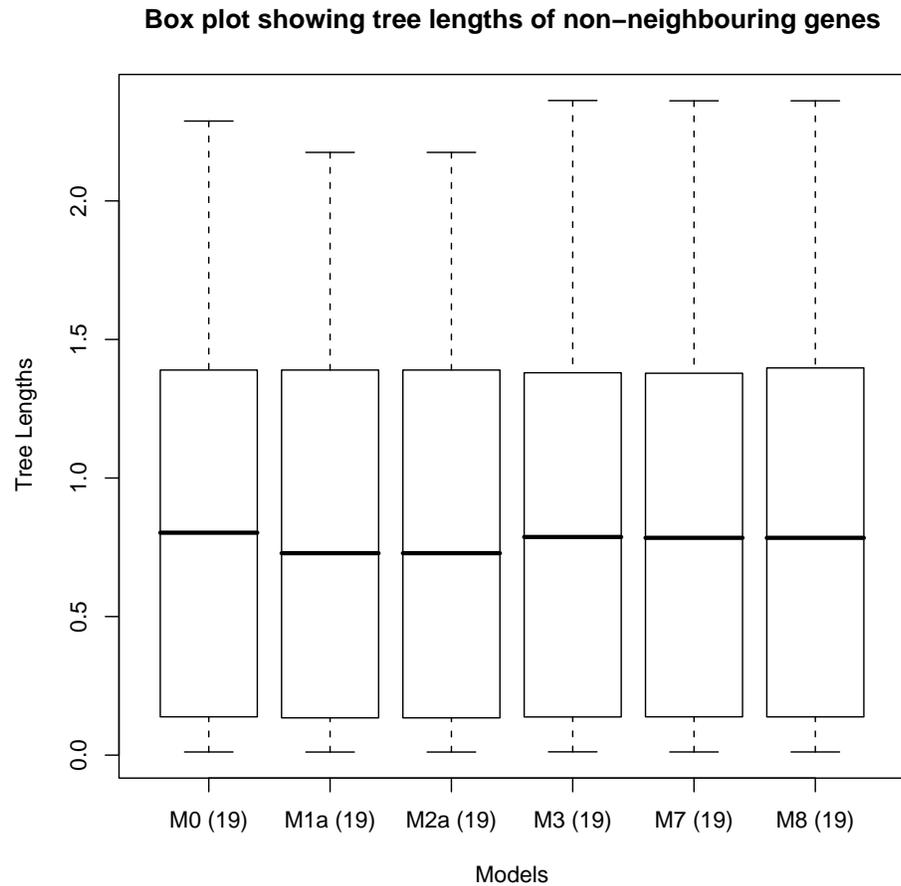


Figure 3.6: A box plot showing tree lengths of non-neighbouring genes. The numbers in the brackets indicate sample sizes

Table 3.2: Table showing results of permutations test (P-values) calculated from tree lengths (Non-N means non-neighbouring).

Model	Up vs. Non-N	Down vs. Non-N	Up vs. Down
M0	0.541	0.375	0.845
M1a	0.638	0.387	0.744
M2a	0.938	0.681	0.739
M3	0.634	0.631	0.967
M7	0.668	0.251	0.5
M8	0.639	0.617	0.959

3.4.6 Type of selection

The dN/dS values were studied for each model. The type of selection was determined for the genes. The following tables show the summary statistics of dN/dS values for different

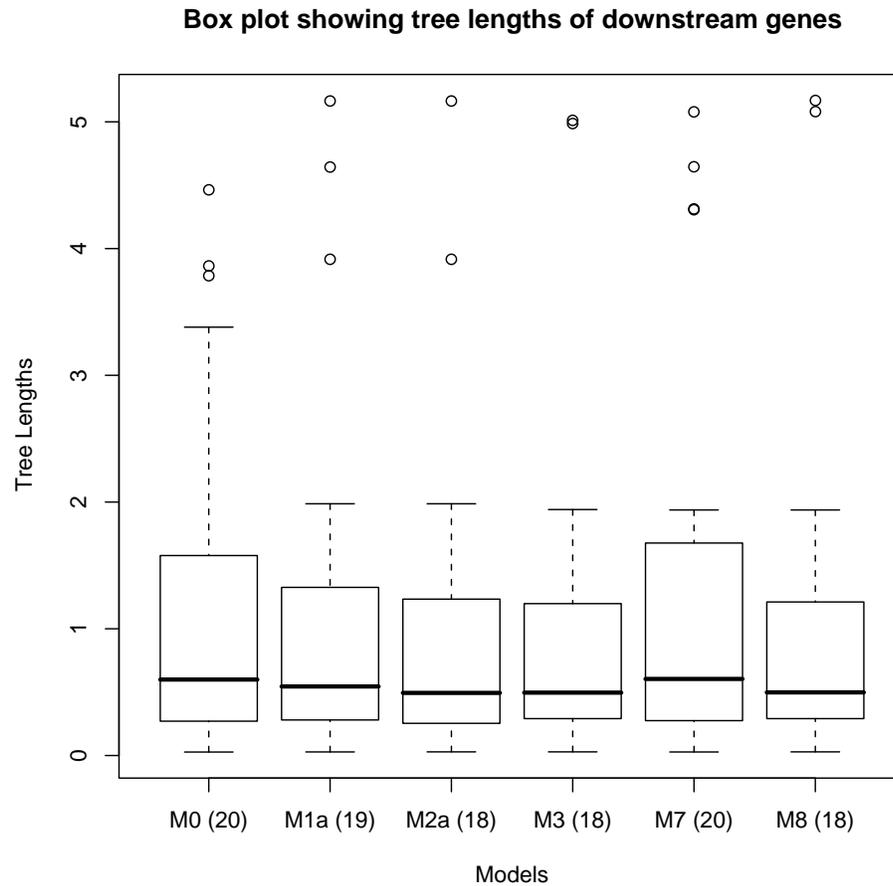


Figure 3.7: A box plot showing tree lengths of downstream genes. The numbers in the brackets indicate sample sizes.

models.

Table 3.3: Table showing summary statistics of dN/dS values of upstream genes

Model	N	Minimum	Median	Mean	Maximum	Variance
M0	168	0.00199	0.0632	0.0824	0.323	0.00426
M1a	161	0.00773	0.1445	0.172	0.656	0.0158
M2a	204	0.00772	0.136	0.166	0.746	0.0173
M3	99	0.0118	0.0799	0.120	0.580	0.0114
M7	100	0.0136	0.0855	0.110	0.379	0.00604
M8	112	0.0136	0.104	0.147	0.600	0.0168

The table 3.3 showed that average dN/dS values were higher in complicated models.

For example, median and mean of M8 model were higher than the median and mean of the corresponding null model M7. The results indicate that model M8 infer higher dN/dS ratios than M7. The maximum values suggest that positive selection models (M2a and M8) did find sites that were under positive selection, but the mean and median of the distribution of average were less than one in all models.

Table 3.4: Table showing summary statistics of dN/dS values of non-neighbouring genes

Model	N	Minimum	Median	Mean	Maximum	Variance
M0	835	0.00149	0.0404	0.0471	0.170	0.00109
M1a	840	0.00947	0.145	0.168	0.590	0.0111
M2a	862	0.00208	0.142	0.166	0.590	0.0111
M3	616	0.00317	0.0564	0.0686	0.264	0.00215
M7	682	0.00402	0.0568	0.0663	0.228	0.00193
M8	629	0.00404	0.0645	0.0818	0.348	0.00369

Similar to the upstream genes, the results in table 3.4 indicate that the mean and median of average omega value distribution of complicated positive selection models (such as M8) was higher than the null model (such as M7) indicating that sophisticated models infer more positive selection than the null models. Compared to the upstream genes, the non-neighbouring genes had lower dN/dS ratios on average. Even the column, showing the maximum of average dN/dS values, indicates that genes upstream genes had higher dN/dS values as compared to dN/dS of non-neighbouring genes.

Table 3.5: Table showing summary statistics of dN/dS values of downstream genes

Model	N	Minimum	Median	Mean	Maximum	Variance
M0	139	0.00301	0.0554	0.0731	0.283	0.00323
M1a	133	0.00303	0.124	0.153	0.478	0.0123
M2a	192	0.00302	0.109	0.151	0.701	0.0181
M3	84	0.00797	0.0705	0.106	0.537	0.0101
M7	88	0.00801	0.0728	0.107	0.511	0.00787
M8	90	0.00801	0.0761	0.132	0.729	0.0195

The results in table 3.5 indicate that alternate models for detection of positive selection infer higher dN/dS than the null models similar to upstream and non-neighbouring genes. Compared to the non-neighbouring genes, the downstream genes had higher dN/dS ratios in most cases. If average omega values of upstream genes and downstream gene were compared, then the results indicated that the average of dN/dS was higher in upstream genes in most cases. It was verified that the difference between the dN/dS ratios was

significant based on Wilcoxon rank sum tests. The results of the Wilcoxon rank tests are given below in table 3.6.

Table 3.6: Table showing results of Wilcoxon rank sum tests (P-values) calculated from average omega (Non-N indicates non-neighbouring).

Model	Up vs. Non-N	Down vs. Non-N	Up vs. Down
M0	7.73e-13	3.791e-07	0.2046
M1a	0.7153	0.03641	0.2278
M2a	0.1891	0.0002600	0.09694
M3	6.056e-07	0.001936	0.2334
M7	1.708e-09	4.087e-05	0.3302
M8	8.697e-09	0.003259	0.08755

The results of the Wilcoxon rank sum tests indicate that most comparisons resulted in a statistically significant difference indicated by P-value less than 0.05. The significance of the difference was very high when upstream genes and downstream genes were compared with the non-neighbouring genes. The difference between upstream genes vs. downstream genes was not as statistically significant as the difference between upstream genes vs. non-neighbouring genes and downstream genes vs. non-neighbouring genes. The significance of the difference was not very high for upstream genes vs. downstream genes. In summary, there is some evidence for higher dN/dS ratios in the genes neighbouring laterally transferred genes, with upstream genes having slightly higher dN/dS ratios than those of downstream genes.

There was a difference between average dN/dS ratios of non-neighbouring genes and the average dN/dS ratios of upstream/downstream genes. To determine whether the difference in the distributions was significant, permutations tests were also carried out between non-neighbouring, upstream and downstream genes (table 3.7).

Table 3.7: Table showing the results of permutation tests (P-values) calculated from average dN/dS ratios (Non-N indicates non-neighbouring).

Model	Up vs. Non-N	Down vs. Non-N	Up vs. Down
M0	2.2e-16	2.2e-16	0.183
M1a	0.683	0.133	0.158
M2a	0.977	0.1	0.256
M3	2.2e-16	2.2e-16	0.379
M7	2.2e-16	2.2e-16	0.856
M8	2.2e-16	2.2e-16	0.453

The results of the permutation tests indicate that the difference between the distributions

of average omega values of upstream genes vs. non-neighbouring genes and downstream genes vs. non-neighbouring genes is significant. However, the difference between the distributions of average omega values of upstream genes vs. downstream genes was not statistically significant indicating that the degree of selection between upstream genes and downstream genes were affected to the same extent by laterally transferred genes. The degree of selection on the neighbouring genes was statistically different from that of non-neighbouring genes. The P-value of permutation tests in most cases was significantly less than 0.05. Compared to the unpaired Wilcoxon rank sum test, the P-values of the permutation tests were lower.

3.5 Discussion

3.5.1 Phylogenetic Tree

To construct a reliable phylogenetic tree, genes conserved over long evolutionary periods were used (Hao and Golding, 2008b). The genes must not be laterally transferred or duplicated because these genes could not result in a tree that accurately represented the speciation of the organisms. The genes that were chosen in this study were *gtlX*, *nusA*, *pheS*, and *rpoA* (Hao and Golding, 2008b). The genes had been used in the phylogenetics of *Bacillus* previously and had shown reliable results (Hao and Golding, 2008b). The orthologs of these genes were found in other organisms using reciprocal best hit of BLAST (Altschul *et al.*, 1997).

The model that was used for phylogenetic tree estimation was GTR + Γ + I. The GTR model is a general time reversible model (Tavaré, 1986). It assumes that evolution of DNA sequence is a time reversible process in which the substitution process, from present to future, is probabilistically same as from the future to present (Tavaré, 1986). GTR models use several different rate parameters. The likelihood method assumed that the rate of evolution of each site is independent and all sites evolve at the same rate (Yang, 1993). One way to overcome this problem is by assuming that some sites are invariable, and other sites are evolving at a single rate. Another approach is to assume that the rates of substitution over sites is Γ distributed, where the shape of the distribution can be determined by a single parameter known as α (Yang, 1993). If the variation in rate is high, then the value of α would be low and vice versa (Yang, 1993).

This research included soil dwelling organisms in the *B. cereus* group. Phylogenetically, the group can be divided into three clades: clade 1, consisting of mostly *B. cereus*; clade 2, consisting of *B. thuringiensis*; clade 3, consisting of *B. mycoides* and *B. weihenstephanensis* (Didelot *et al.*, 2009). The three clades are not monophyletic because of LGT from one clade to another. Clade 3 is the most diverse and shows higher rates of LGT, while *B. anthracis* is nearly clonal (Didelot *et al.*, 2009). In this research, only a small number of high quality genomes of the cereus group were used. As noted earlier in Didelot *et al.* (2009), this study could separate the species neatly into three monophyletic clades because of high similarity of the species.

3.5.2 Inference of LGT

To determine the neighbouring genes and non-neighbouring genes of laterally transferred genes, the laterally transferred genes must be identified accurately. An approach based on the distribution of genes in organisms was used. It was assumed that if a gene were found in an organism, but not in its nearest neighbours, then the gene would have been laterally transferred gene or alternatively there might have been deletions in the neighbouring organisms. Based on a parsimony principle, a single lateral gene transfer event is more likely than multiple deletions in closely related genes (Fitch, 1971). Hence, parsimony was used to infer LGT (Fitch, 1971). To study the distribution of genes, we classified the genes into families in which closely related genes were in the same family and distantly related or unrelated genes were in different families. An algorithm was used that could find natural clusters in graphs in which nodes represented proteins and edges represented similarity between proteins (Enright, Van Dongen and Ouzounis, 2002). The ancestral states, inferred by parsimony, were used to infer changes of states. Hence, laterally transfer genes were identified similar to Hao and Golding (2004).

Due to the nature of the clustering algorithm and gene families, there were no single gene clusters. A family was defined as a group of closely related genes with at least two genes in the family. The clustering was based on the sequence similarity, which might not necessarily mean similarity in structure or function. Another bias in parsimony was that it tends to underestimate the number of events (Hao and Golding, 2004). For example, there might be two independent transfers of a gene in two sister taxa, but parsimony will infer it as a single LGT event at the node of the taxa (Fitch, 1971). There is also a bias towards changes at the tips of phylogeny (Hao and Golding, 2004).

Despite limitations, parsimony is more reliable than parametric methods for detection of laterally transferred genes. Parametric methods can be classified into gene based methods and window based methods (Becq, Churlaud and Deschavanne, 2010). Parametric methods compare GC contents, codon usage bias, tetranucleotide and oligonucleotide frequencies between different genomes and determine the host and the recipient genome (Becq, Churlaud and Deschavanne, 2010). Windows based methods have high sensitivity and can detect genes isolated laterally transferred gene fragments. Whereas gene based methods lack sensitivity. They have lower false positive rates than that of windows based methods (Becq, Churlaud and Deschavanne, 2010). Different methods give different results. Results of comparative analysis of different methods indicated that GC content based methods are more sensitive to the origin of laterally transferred genes. Whereas tetranucleotide methods are more accurate regardless of the origin of the genes. Tetra-nucleotide methods are accurate whether they were used in a window based method or a gene based method (Becq, Churlaud and Deschavanne, 2010). In general, parametric methods are able to detect LGT in genomes that are not sequenced or assembled completely. Whereas gene distribution methods such as parsimony require complete and assembled genome for the detection of laterally transferred genes and deleted genes. In our research, genomes that were complete sequenced, assembled and annotated were used. The completeness and accuracy of genomes influence the results of gene distribution methods.

The accuracy of phylogenetic methods, Bayesian methods, and first order Markov

methods can also be evaluated (Cortez, Lazcano and Becerra, 2005). To test the accuracy of methods is important because different methods can give different sets of laterally transferred genes (Cortez, Lazcano and Becerra, 2005). The accuracy of detection methods were tested by creating simulated LGT into genomes and using different methods to identify the laterally transferred genes (Cortez, Lazcano and Becerra, 2005). Non-parametric methods, including phylogenetic methods, Bayesian methods and methods based on the distribution of genes, were better than parametric methods (Cortez, Lazcano and Becerra, 2005).

3.5.3 Neighbouring and Non-Neighbouring genes

In this research, genes that were neighbours of laterally transferred genes and genes that were not neighbours of laterally transferred genes were studied. There were similarities and differences between both sets of genes. Both sets of genes were highly conserved genes because they did not have duplications and were not laterally transferred. Duplications were excluded because the rate of evolution and dN/dS ratios might be higher in duplicated genes (Ohta, 1994). It is well known that the rate of evolution and dN/dS ratios of laterally transferred genes are higher than that of conserved genes (Marri, Hao and Golding, 2007). If duplicated genes and laterally transferred genes were included in the analysis, they could have caused confounding effects in the results. Therefore, it was decided not to include them.

The number of non-neighbouring gene families in 47 genomes was 890 families. If there exists a core genome that is vertically inherited, then there are some genes that are vertically inherited (Wolf *et al.*, 2002). Moreover, there is a strand bias in LGT and more laterally transferred genes are found on the leading strand (Hao and Golding, 2009). Conserved structures such as operons do not have high rates of LGT within them (Itoh *et al.*, 1999). In our research, the core set of genes included those that might have been inherited vertically. They might be on lagging strand or form a part of an operon. These genes might not be essential because most essential genes are more frequently found on the leading strand (Hao and Golding, 2008b). Moreover, the average size of each family was less than 25 genes because most of the genes in the families were removed due to the presence of laterally transferred genes in the flanking regions. Most families had a small or a large number of genes. A small number indicated the lack of conservation of genes and high rate of rearrangements, while a larger number of genes indicated a large number of conserved genes and low rate or rearrangements. The histogram below (figure 3.8 shows the number of genes in each family.

The number of gene families that were upstream laterally transferred genes were 241, while the number of gene families downstream laterally transferred genes were 218. These low numbers were due the high rate of LGT in some regions of the genomes. Approximately half of genes in bacterial genomes are in operons (Price, Arkin and Alm, 2006). Most genes that were neighbours of laterally transferred genes were, therefore, in the regions of low operons density and on the leading strand. It was difficult to find neighbouring genes of laterally transferred genes due to the high frequency of LGT. The selfish operons

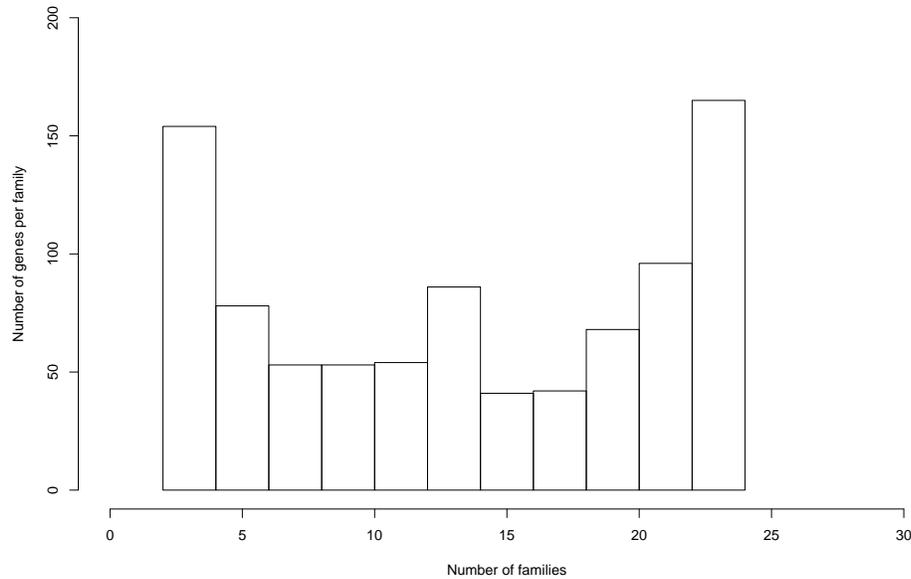


Figure 3.8: A histogram showing the distribution of number of families vs. number of genes per family in the non-neighbouring genes

model proposes that genes are transferred in clusters (Lawrence, 1999b). It implies that the neighbouring genes of a laterally transferred gene are themselves other laterally transferred genes because the gene may be part of a laterally transferred operon. Moreover, if there were a native gene between two laterally transferred genes, then the native gene would not be included in this study due to the ambiguous nature of the native gene. It is both upstream and downstream of a laterally transferred gene. For these reasons, the number of gene families of upstream and downstream genes was small.

The types of selection on genes were studied using likelihood ratio tests of model M0 vs. M3, M1a vs. M2a, and M7 vs. M8. The results indicated that the number of gene families under positive selection were very low. The low dN/dS value and the lack of positive selection can be explained by the degree of conservation of the gene families being studied. Highly conserved genes are under strong negative selection (Ohta, 1974). Moreover, these genes are generally not laterally transferred. Laterally transferred genes have higher rates of evolution (Hao and Golding, 2006). Therefore, vertical inheritance of these genes suggested that there would be a lack of evidence for positive selection. Moreover, duplicated genes were not included in this study. Duplicated genes may be under positive selection and have a high rate of evolution (Gu *et al.*, 2002). There may be effects on selection and dN/dS due to the time lag between acquisition of deleterious mutations and removal of them because these genes are not laterally transferred and are conserved (Rocha *et al.*, 2006). Therefore, the results of this research indicate that both sets of genes, neighbours and non-neighbours of laterally transferred genes, are most likely under negative selection.

3.5.4 Rate of Evolution and Selection

The rate of evolution was inferred based on tree lengths. Previously, tree lengths had been used to infer high rates of evolution in laterally transferred genes (Yang, 2007). In our research, tree lengths of neighbours of laterally transferred genes and non-neighbouring genes could not be directly compared. The number and the names of species had to be the same in the trees from the neighbouring genes and non-neighbouring genes. An R statistics package was used to drop leaves from non-neighbouring genes in order to compare the rates of evolution (Team, 2011). The results indicate that the rates of evolution of the upstream genes, downstream genes and non-neighbouring genes were approximated similar. The difference between the distributions of the tree lengths of different sets of genes was not statistically significant in most cases based on permutation test and Wilcoxon rank sum test (Team, 2011).

There is a difference in the distribution of the rates of evolution of different datasets. Closer examination revealed that the genes that were upstream of laterally transferred genes and the genes that are downstream of laterally transferred genes had more outliers than the non-neighbouring genes. These results indicate that, in some cases, laterally transferred genes might increase the rate of evolution of neighbouring genes. In other cases, the neighbouring genes might be essential and conserved. Therefore, the increase in rate of evolution might be negatively selected. Permutation tests and Wilcoxon rank sum test were done after extreme outliers were removed from the results. Extreme outliers might result due to inaccuracy of CodeML in the calculation of the rate of evolution, or the rate of evolution was indeed very high. The former could occur because the number of sequences in the genomes were small and were highly conserved. A very high rate of evolution of genes was unlikely because the genes were conserved.

The results indicate that non-neighbouring genes had lower dN/dS values as compared to the genes that were neighbours of laterally transferred genes. These results are consistent with previous studies in which genes that are essential and highly conserved have a lower rate of evolution than genes that are not essential (Ohta, 1974). Moreover, essential genes and genes that are important in biological pathways are under strong negative selection (Ohta, 1974). For example, if there is a single amino acid change (non-synonymous substitution) in the protein, evolution would act against the change and remove the bacterium from the population. Hence, the rate of nonsynonymous substitution per nonsynonymous site would be lower than expected by chance alone. Similarly, if there would be a synonymous change in the gene sequence, the amino acid sequence would not change. There would be no change in the structure and function of the protein. Hence, high rates of synonymous substitution per synonymous site would be tolerated in essential and necessary genes. The low rate of nonsynonymous substitutions per nonsynonymous site and high rate of synonymous substitutions per synonymous site lead to low dN/dS value (Yang, 2007). The results of this study indicate that the dN/dS values were less than one for most of the gene families regardless of whether they were in non-neighbouring set or neighbouring set. The results indicate that negative selection is relaxed in genes that are neighbours of laterally transferred genes. The average dN/dS value of neighbouring genes is higher than that of non-neighbouring genes. The difference between the distribution of average dN/dS

values was statistically significant as indicated by Wilcoxon rank sum test and permutation tests.

The results indicate that there is a significant difference in the degree of selection on the genes that were neighbours of laterally transferred genes and the genes that were not neighbouring laterally transferred genes. The differences in selection could be due to a number of reasons. These could include hitchhiking and background selection. Hitchhiking and background selection are known to affect genes that are linked to a gene that is under strong selection (Charlesworth, Morgan and Charlesworth, 1993).

An explanation for relaxed selection or positive selection in genes that are neighbours of laterally transferred genes is genetic hitchhiking. Genetic hitchhiking may have caused divergence of the neighbouring genes. In some cases, it may cause relaxed selection and an increase in dN/dS in the hitchhiked gene (Rocha *et al.*, 2006). If it is assumed that the laterally transferred gene and its neighbouring genes are in linkage disequilibrium, then relaxed selection may be due to genetic hitchhiking. Laterally transferred genes may be in linkage disequilibrium to their respective neighbouring genes. Hence, the apparent relaxed selection may be due to genetic hitchhiking.

Besides hitchhiking, background selection can also affect the rate of substitution of a locus that is linked to a negatively selected locus (Charlesworth, Morgan and Charlesworth, 1993). In this research, background selection may be the cause of the difference in the dN/dS ratio between neighbouring genes of laterally transferred genes and genes that were not neighbours of laterally transferred genes. Therefore, background selection could act on the genes neighbouring laterally transferred genes.

Based on this research, gene hitchhiking and background selection may be responsible for relaxed selection on genes that were neighbours of laterally transferred genes. Lateral transfer of genes could lead to relaxed selection on neighbouring genes. The data presented in this research showed that the distribution of dN/dS values are significantly higher in genes that are neighbours of laterally transferred genes than that in non-neighbouring genes. The rates of evolution of neighbouring genes are slightly higher than that of non-neighbouring genes. The high rates of evolution could be explained by relaxed selection on neighbouring genes due to lateral transfer of genes in region nearby. The molecular mechanism by which laterally transferred genes affect neighbouring genes is still unknown.

The results indicate that the genes neighbouring laterally transferred genes have higher dN/dS ratios. Moreover, previous research indicated that laterally transferred genes have a higher rate of evolution and higher dN/dS values (Marri, Hao and Golding, 2007). It could be possible that the genes were transferred to the regions where the rate of evolution and dN/dS values were higher than expected by chance alone. The genes neighbouring the laterally transferred genes might have higher rates of evolution and higher dN/dS ratios even before the genes were transferred. There could be no change after the transfer of genes. The relaxed negative selection detected in this study might be present even before the genes were laterally transferred. This problem was not further studied in our research. It is also possible that laterally transferred genes were selectively retained in the regions of high rates of evolution. The high rates of evolution in those regions decreased the probability of gene conversion because of sequence divergence.

The methods of detection of selection can be improved to detect selection on different branches using multiple different models and multiple parameters. Such methods can detect episodic selection resulting in an increase in accuracy of prediction of genes and sites under positive selection. A future study could be designed to investigate the reasons and mechanism of transfer in these regions and what makes these regions different from other regions in genomes. Finally, this research can be extended to include species other than *Bacillus* to confirm the results of this study.

Part III
CONCLUSION

Chapter 4

Summary

4.1 LGT into Operons by Homologous Recombination

Prokaryotes can evolve rapidly by the process of LGT, which is the non-vertical transmission of genes from one cell to another (Ochman *et al.*, 2000). It occurs by mechanisms such as transformation, conjugation, and transduction and via nanotubes (Dubey and Ben-Yehuda, 2011). DNA can be integrated into a chromosome by homologous recombination and nonhomologous end joining (Thomas and Nielsen, 2005). LGT can be detected by studying deviations in codon usage, nucleotide composition, conflict in phylogenetic trees and distribution of genes (Ragan, 2001). The rate of LGT can be higher than the rate of spontaneous mutation (Hao and Golding, 2004). LGT is common in bacteria living in diverse niches including oceans (McDaniel *et al.*, 2010). Metabolic genes are transferred and retained more readily than informational genes (Hao and Golding, 2008b). The rate of LGT is higher in closely related organisms than in distantly related organisms (Hao and Golding, 2006). LGT has an important role in the evolution of novel traits in bacteria.

An operon is a group of genes that are controlled by a single control region (Jacob *et al.*, 1960). Operons are very common in prokaryotes and may have originated in thermophilic organisms (Glansdorff, 1999). Genes can be organized into new operons by deletion of a part of intergenic sequence, rearrangements and LGT (Itoh *et al.*, 1999). Operons can be disorganized by rearrangements and LGT (Price, Arkin and Alm, 2006). The order and number of genes in any particular operon is nearly conserved across taxa. However, the intergenic spacing of genes within the operons may not always be less than 50 base pairs, because highly expressed genes tend to be widely spaced (Eyre-Walker, 1995). Some models that can explain the formation of operons include the natal model, the Fisher model, the molarity model, the coregulation model and the selfish operon model. The natal model states that genes are in clusters because of duplication and divergence (Lawrence, 1999b). The Fisher model suggests that the rate of deleterious recombination events decreases if genes are in proximity (Fisher, 1930). The molarity model states that genes in clusters result in high localized concentration of gene product (Losick and Shapiro, 1999). The coregulation model suggests that genes are in operons because they are regulated by a single operator (Jacob *et al.*, 1960). The selfish operon model suggests that genes are in

clusters because they can transfer a pathway by LGT (Lawrence, 1999b).

Homologous recombination is integration or replacement of DNA with the help of sequence homology (Didelot and Maiden, 2010). The rate of recombination in bacteria is as high as the rate of mutations and in some taxa. It is even higher than the rate of mutations (Perez-Losada *et al.*, 2006). The rate of homologous recombination varies across taxa. Recombination is more frequent between the species that are closely related (Didelot and Maiden, 2010). Recombination can be detected by computational methods such as distance methods, phylogenetic methods, compatibility methods, permutation tests and Bayesian methods. Distance methods test the difference of the genetic distance between recombinant regions and nonrecombinant regions (McGuire and Wright, 2000). Phylogenetic methods test the difference between phylogenies of recombinant regions and nonrecombinant regions (Grassly and Holmes, 1997). Compatibility methods test the phylogenetic compatibility between different regions (Jakobsen and Easteal, 1996). Permutation tests check that the sequence follows a certain distribution throughout the alignment (Sawyer, 1989). Bayesian methods detect recombination based on machine learning methods (Husmeier and McGuire, 2003).

This research focused on lateral transfer of genes into operons. The purpose of this research was to investigate the evidence of homologous recombination as a mechanism for integration of laterally transferred genes into operon. A phylogenetic tree of *Bacillus* was constructed using conserved DNA sequences that were aligned in MUSCLE (Edgar, 2004). A tree was inferred using MrBayes with GTR + Γ + I model and ten million generations of MCMC analysis (Huelsenbeck and Ronquist, 2001). Operon structure was inferred using OperonDB (Pertea *et al.*, 2009). The prediction is based on the degree of conservation of genes across different taxa and was shown to be very accurate (Pertea *et al.*, 2009). The phylogenetic tree was used to detect LGT. Genes were classified into families using Markov clustering algorithm (Enright, Van Dongen and Ouzounis, 2002). The presence and absence of genes were represented by binary characters 1 and 0 respectively. Parsimony was used to infer ancestral states (Fitch, 1971). Using the ancestral states, LGT was inferred. Laterally transferred genes in operons were selected for detection of homologous recombination. Every gene was searched in the database of all prokaryotes, and the sequences that were closely related to the gene were obtained from *Bacillus* and non-*Bacillus* species. Up to 1500 bps region upstream and 1500 bps region downstream of the genes were included in the sequences. The sequences were aligned in MUSCLE (Edgar, 2004). Homologous recombination was detected using the maximum chi square (Smith, 1992) and GENECONV algorithms (Sawyer, 1989).

The tree indicated that most clades were supported with a posterior probability greater than 0.8. Strains of *B. anthracis* and *B. cereus* were very closely related and had very short branch lengths. *O. iheyensis* was chosen as an outgroup because of its unique niche (Takami, Takaki and Uchiyama, 2002). Based on inferences using OperonDB, there were approximately 300 to 500 operons in each genome and approximately three to five genes in each operon. A phylogenetic tree was constructed to infer genes that were laterally transferred. Approximately ten percent of genes were inferred to be transferred laterally into each genome. Moreover, approximately ten percent of laterally transferred genes were

transferred into pre-existing operons. Based on results from GENECONV, the transfer of genes into operons was from sources that were closely related to the recipient cells. There was evidence of homologous recombination within the operons. The evidence indicated that the recombination breakpoints were upstream and downstream the laterally transferred genes. However, there was a lack of evidence of homologous recombination from an outgroup to *Bacillus* species. The maximum chi square algorithm was used to analyze the results further. With a half window size of 50 base pairs, there was a gene that was transferred from outgroup to *Bacillus* species and the recombination breakpoints were detected upstream and downstream with a P-value of less than 0.10. With both GENECONV and Max chi, there was a significant proportion of genes where the recombination breakpoint was detected either upstream or downstream of the gene but not in both locations, suggesting that there might have been double recombination event. The evidence for a recombination breakpoint from one end might have been lost due to mutations.

4.2 Effect of LGT on Neighbouring genes

The neutral theory of evolution states that most mutations are neutral with respect to the fitness of the organism (Kimura *et al.*, 1968). Most synonymous substitutions have no effect on the structure of the protein. The rate of synonymous substitutions is much higher than the rate of nonsynonymous substitutions. Slightly deleterious mutations can be fixed in a population after a beneficial mutation that compensates for the negative effects of the deleterious mutation (Ohta, 1974). Hitchhiking can lead to fixation of mutations that are linked to sites under positive selection. It was also observed that amino acid substitutions are episodic and most likely due to a change in the environment (Ohta, Gillespie *et al.*, 1996).

The ratio of the number of nonsynonymous substitutions per nonsynonymous site (dN) to the number of synonymous substitutions per synonymous site (dS) can be used the study the type and the degree of selection (Hurst *et al.*, 2002). The dN/dS ratio is affected by many factors such as the rate of expression of genes, which affects codon usage (Sharp and Li, 1987). Due to energetic requirements, some codons are under strongly deleterious selection (Higgs, Hao and Golding, 2007). Recombination can increase the observed value of dN/dS (Castillo-Ramirez *et al.*, 2011). Genes that are laterally transferred (Hao and Golding, 2006) or duplicated (Gu *et al.*, 2002) have higher dN/dS ratios than that of vertically inherited genes. Purifying selection is higher in larger populations (Jordan *et al.*, 2002). Moreover, the time lag between acquisition of slightly deleterious mutations and removal of them can cause a bias in dN/dS ratios (Rocha *et al.*, 2006). Genes involved in pathogenicity and immune systems are under positive selection (Williamson, 2003).

There are a number of methods to estimate dN/dS ratio. A maximum likelihood method can be used to estimate dN/dS ratios (Goldman and Yang, 1994). The evolution of codons can be considered as continuous time Markov chain and a 61 by 61 rate matrix can be constructed (Goldman and Yang, 1994). Codon bias and transition/transversion ratio bias can also be included in the model. The likelihood method uses a phylogenetic tree to estimate dN/dS ratio. Different models, such one ratio (M0), Neutral (M1a), positive selection

(M2a), discrete (M3), beta distributed (M7) and beta with ω (M8) can be used to study dN/dS. Usually a null model that does not allow for positive selection is compared with an alternate model that allows for positive selection (M0 vs. M3, M1a vs. M2a, and M7 vs. M8). If the null model is rejected and the omega value is greater than one, then there is evidence for positive selection (Anisimova, Bielawski and Yang, 2001). Sites that are under positive selection can also be inferred using Bayesian statistics (Anisimova, Bielawski and Yang, 2002).

The processes and elements that could affect neighbouring genes include insertion sequences (Siguier, Filee and Chandler, 2006), genetic hitchhiking (Fay and Wu, 2000), background selection (Charlesworth, Morgan and Charlesworth, 1993), genetic draft (Gillespie, 2001), specialized transduction (Ochman *et al.*, 2000) and pseudogenes (Kuo and Ochman, 2010). In this research, it was proposed that laterally transferred genes also affect the rate of evolution and selection on the neighbouring genes. A phylogenetic tree of *Bacillus* was inferred using MUSCLE alignment of conserved genes (Edgar, 2004). The tree was inferred using MrBayes (Huelsenbeck and Ronquist, 2001). Genes were organized into families using MCL (Enright, Van Dongen and Ouzounis, 2002). LGT was detected using parsimony algorithm (Fitch, 1971). The duplicated genes were removed from the analysis. Gene that were upstream or downstream laterally transferred genes were selected for further analysis. From these genes, duplicated genes, laterally transferred genes and genes that were neighbours of laterally transferred genes were removed. These were the upstream and downstream sets. The non-neighbouring genes consisted of genes that were not in upstream and downstream set, nor duplicated and nor laterally transferred. Rates of evolution and selection on all three sets were inferred using PAML4.5 (Yang, 2007). The significance of difference in the rate of evolution and selection on genes was studied using unpaired Wilcoxon rank sum tests and permutation tests.

The phylogenetic tree indicated that most species of *Bacillus* can be resolved into bifurcating nodes. MCL clustered *Bacillus* genes into 10,277 families. There were 2,610 multicopy genes and 7,667 single copy genes excluding ORFans in 47 genomes. The number of families that were upstream from laterally transferred genes were 241, and they consisted of at least three genes. The number of families that were downstream of laterally transferred genes were 218, and they consisted of at least three genes. The number of non-neighbouring gene families was 890. Likelihood ratio tests indicated that only a very small percentage (less than 20) of genes was under positive selection. It was observed that the rate of evolution of neighbouring genes was slightly higher than that of non-neighbouring genes. The difference was not significant based on unpaired Wilcoxon rank sum tests and permutation tests. There were some genes families in neighbouring genes in which the rate of evolution was much higher than average. However, it was observed that dN/dS values in neighbouring genes were higher than non-neighbouring genes suggesting relaxed selection on neighbouring genes. Based on unpaired Wilcoxon rank sum tests and permutation tests, the difference between dN/dS of neighbouring genes and non-neighbouring genes was significant. In conclusion, the results suggested that genes that were neighbours of laterally transferred genes were under less strong selection than those that were non-neighbours.

Part IV
BIBLIOGRAPHY

Bibliography

- Achtman, M., P. Manning, B. Kusecek, S. Schwuchow, and N. Willetts (1980). Genetic analysis of F sex factor cistrons needed for surface exclusion in *Escherichia coli*. *J Mol Biol.* 138(4), 779–795.
- Akiba, T., K. Koyama, Y. Ishiki, S. Kimura, T. Fukushima, et al. (1960). On the mechanism of the development of multiple-drug-resistant clones of Shigella. *Jpn J Microbiol.* 4(2), 219–227.
- Altekar, G., S. Dwarkadas, J. Huelsenbeck, and F. Ronquist (2004). Parallel metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics* 20(3), 407–415.
- Altschul, S., T. Madden, A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* 25(17), 3389–3402.
- Andam, C. and J. Gogarten (2011). Biased gene transfer in microbial evolution. *Nature Reviews Microbiology* 9(7), 543–555.
- Andam, C., D. Williams, and J. Gogarten (2010). Biased gene transfer mimics patterns created through shared ancestry. *Proc Natl Acad Sci U S A* 107(23), 10679–10684.
- Andersson, J., A. Sjogren, L. Davis, T. Embley, and A. Roger (2003). Phylogenetic analyses of diplomonad genes reveal frequent lateral gene transfers affecting eukaryotes. *Curr Biol.* 13(2), 94–104.
- Anisimova, M., J. Bielawski, and Z. Yang (2001). Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol Biol Evol.* 18(8), 1585–1592.
- Anisimova, M., J. Bielawski, and Z. Yang (2002). Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Mol Biol Evol.* 19(6), 950–958.
- Anisimova, M., R. Nielsen, and Z. Yang (2003). Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* 164(3), 1229–1236.

- Aravind, L. and E. Koonin (2001). Prokaryotic homologs of the eukaryotic DNA-end-binding protein Ku, novel domains in the Ku protein and prediction of a prokaryotic double-strand break repair system. *Genome Res.* 11(8), 1365–1374.
- Arvey, A., R. Azad, A. Raval, and J. Lawrence (2009). Detection of genomic islands via segmental genome heterogeneity. *Nucl. Acids Res.* 37(16), 5255–5266.
- Ashburner, M., C. Ball, J. Blake, D. Botstein, H. Butler, J. Cherry, A. Davis, K. Dolinski, S. Dwight, J. Eppig, et al. (2000). Gene Ontology: tool for the unification of biology. *Nat Genet.* 25(1), 25–29.
- Azad, R. and J. Lawrence (2005). Use of artificial genomes in assessing methods for atypical gene detection. *PLoS Comput Biol.* 1(6), e56.
- Azad, R. and J. Lawrence (2007). Detecting laterally transferred genes: use of entropic clustering methods and genome position. *Nucl. Acids Res.* 35(14), 4629–4639.
- Azad, R. and J. Lawrence (2011). Towards more robust methods of alien gene detection. *Nucleic Acids Res.* 39(9), e56.
- Barracough, T., K. Balbi, and R. Ellis (2012). Evolving concepts of bacterial species. *Evolutionary Biology* 39(2), 148–157.
- Barrangou, R., C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D. Romero, and P. Horvath (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315(5819), 1709–1712.
- Becq, J., C. Churlaud, and P. Deschavanne (2010). A benchmark of parametric methods for horizontal transfers detection. *PloS one* 5(4), e9989.
- Berndt, C., P. Meier, and W. Wackernagel (2003). DNA restriction is a barrier to natural transformation in *Pseudomonas stutzeri* JM300. *Microbiology* 149(4), 895–901.
- Bielawski, J. and Z. Yang (2004). A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. *J Mol Evol.* 59(1), 121–132.
- Boc, A. and V. Makarenkov (2011). Towards an accurate identification of mosaic genes and partial horizontal gene transfers. *Nucl. Acids Res.* 39(21), e144.
- Boto, L. (2010). Horizontal gene transfer in evolution: facts and challenges. *Proc. R. Soc. B.* 277(1683), 819–827.
- Boussau, B., L. Guéguen, and M. Gouy (2009). A mixture model and a hidden markov model to simultaneously detect recombination breakpoints and reconstruct phylogenies. *Evol Bioinform Online.* 5, 67–79.

- Brouwer, R., O. Kuipers, and S. van Hijum (2008). The relative value of operon predictions. *Brief Bioinform.* 9(5), 367–375.
- Castillo-Ramirez, S., S. Harris, M. Holden, M. He, J. Parkhill, S. Bentley, and E. Feil (2011). The impact of recombination on dN/dS within recently emerged bacterial clones. *PLoS Pathog.* 7(7), e1002129.
- Chan, C., R. Beiko, and M. Ragan (2006). Detecting recombination in evolving nucleotide sequences. *BMC bioinformatics* 7(412).
- Chan, C., A. Darling, R. Beiko, and M. Ragan (2009). Are protein domains modules of lateral genetic transfer? *PLoS One* 4(2), e4524.
- Chang, B. and M. Donoghue (2000). Recreating ancestral proteins. *Trends Ecol Evol.* 15(3), 109–114.
- Charlebois, R. and W. Doolittle (2004). Computing prokaryotic gene ubiquity: rescuing the core from extinction. *Genome Res.* 14(12), 2469–2477.
- Charlesworth, B., M. Morgan, and D. Charlesworth (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics* 134(4), 1289–1303.
- Chayot, R., B. Montagne, D. Mazel, and M. Ricchetti (2010). An end-joining repair mechanism in *Escherichia coli*. *Proc Natl Acad Sci U S A* 107(5), 2141–2146.
- Che, D., G. Li, F. Mao, H. Wu, and Y. Xu (2006). Detecting uber-operons in prokaryotic genomes. *Nucleic Acids Res.* 34(8), 2418–2427.
- Cheetham, B. and M. Katz (1995). A role for bacteriophages in the evolution and transfer of bacterial virulence determinants. *Mol Microbiol.* 18(2), 201–208.
- Cohan, F. (2006). Towards a conceptual and operational union of bacterial systematics, ecology, and evolution. *Philos Trans R Soc Lond B Biol Sci.* 361(1475), 1985–1996.
- Cohan, F., M. Roberts, and E. King (1991). The potential for genetic exchange by transformation within a natural population of *Bacillus subtilis*. *Evolution*, 1393–1421.
- Cohen, O., U. Gophna, and T. Pupko (2011). The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. *Mol Biol Evol.* 28(4), 1481–1489.
- Cortez, D. Q., A. Lazcano, and A. Becerra (2005). Comparative analysis of methodologies for the detection of horizontally transferred genes: a reassessment of first-order markov models. *In Silico Biol.* 5(5-6), 581–592.
- Dam, P., V. Olman, K. Harris, Z. Su, and Y. Xu (2007). Operon prediction using both genome-specific and general genomic information. *Nucleic Acids Res.* 35(1), 288–298.

- Dandekar, T., B. Snel, M. Huynen, and P. Bork (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci.* 23(9), 324–328.
- Darwin, C. (1891). *The origin of species by means of natural selection: or the preservation of favoured races in the struggle for life*. John Murray, Albemarle Street.
- Daubin, V., E. Lerat, G. Perrière, et al. (2003). The source of laterally transferred genes in bacterial genomes. *Genome Biol.* 4(9), R57.
- Daubin, V., N. Moran, and H. Ochman (2003). Phylogenetics and the cohesion of bacterial genomes. *Science* 301(5634), 829–832.
- Daubin, V. and H. Ochman (2004). Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*. *Genome Res.* 14(6), 1036–1042.
- Davies, J. (1996). Origins and evolution of antibiotic resistance. *Microbiologia* 12(1), 9–16.
- De Vries, J. and W. Wackernagel (2002). Integration of foreign DNA during natural transformation of *Acinetobacter* sp. by homology-facilitated illegitimate recombination. *Proc Natl Acad Sci U S A.* 99(4), 2094–2099.
- Denamur, E., G. Lecointre, P. Darlu, O. Tenaillon, C. Acquaviva, C. Sayada, I. Sunjevaric, R. Rothstein, J. Elion, F. Taddei, M. Radman, and I. Matic (2000). Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. *Cell* 103(5), 711–721.
- Didelot, X., M. Barker, D. Falush, and F. Priest (2009). Evolution of pathogenicity in the *Bacillus cereus* group. *Syst Appl Microbiol.* 32(2), 81–90.
- Didelot, X. and M. Maiden (2010). Impact of recombination on bacterial evolution. *Trends Microbiol.* 18(7), 315–322.
- Doroghazi, J. and D. Buckley (2010). Widespread homologous recombination within and between *Streptomyces* species. *The ISME J.* 4(9), 1136–1143.
- Dubey, G. and S. Ben-Yehuda (2011). Intercellular nanotubes mediate bacterial communication. *Cell* 144(4), 590–600.
- Earl, A., R. Losick, and R. Kolter (2008). Ecology and genomics of *Bacillus subtilis*. *Trends Microbiol.* 16(6), 269–275.
- Edgar, R. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.* 32(5), 1792–1797.
- Emma, L., S. Khushwant, and H. Simon (2008). Predicted transcription factor binding sites as predictors of operons in *Escherichia coli* and *Streptomyces coelicolor*. *BMC Genomics* 9(79).

- Enright, A., S. Van Dongen, and C. Ouzounis (2002). An efficient algorithm for large-scale detection of protein families. *Nucl. Acids Res.* 30(7), 1575–1584.
- Ermolaeva, M., O. White, and S. Salzberg (2001). Prediction of operons in microbial genomes. *Nucleic Acids Res.* 29(5), 1216–1221.
- Eyre-Walker, A. (1995). The distance between *Escherichia coli* genes is related to gene expression levels. *J Bacteriol.* 177(18), 5368–5369.
- Fay, J. and C. Wu (2000). Hitchhiking under positive darwinian selection. *Genetics* 155(3), 1405–1413.
- Feil, E., E. Holmes, D. Bessen, M. Chan, N. Day, M. Enright, R. Goldstein, D. Hood, A. Kalia, C. Moore, J. Zhou, and B. Spratt (2001). Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proc Natl Acad Sci U S A.* 98(1), 182–187.
- Felsenstein, J. (1989). PHYLIP-phylogeny inference package (version 3.2). *Cladistics* 5, 164–166.
- Fisher, R. (1930). The genetical theory of natural selection.
- Fitch, W. (1971). Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Biology* 20(4), 406–416.
- Fitzpatrick, D. (2012). Horizontal gene transfer in fungi. *FEMS Microbiol Lett* 329(1), 1–8.
- Garcia-Vallve, S., A. Romeu, and J. Palau (2000). Horizontal gene transfer of glycosyl hydrolases of the rumen fungi. *Mol Biol Evol.* 17(3), 352–361.
- Gillespie, J. (1994). Substitution processes in molecular evolution. iii. deleterious alleles. *Genetics* 138(3), 943–952.
- Gillespie, J. (2001). Is the population size of a species relevant to its evolution? *Evolution* 55(11), 2161–2169.
- Gladyshev, E., M. Meselson, and I. Arkhipova (2008). Massive horizontal gene transfer in bdelloid rotifers. *Science* 320(5880), 1210–1213.
- Glansdorff, N. (1999). On the origin of operons and their possible role in evolution toward thermophily. *J Mol Evol.* 49(4), 432–438.
- Gogarten, J. (2003). Gene transfer: gene swapping craze reaches eukaryotes. *Curr Biol.* 13(2), R53–R54.
- Gogarten, J., W. Doolittle, and J. Lawrence (2002). Prokaryotic evolution in light of gene transfer. *Mol Biol Evol.* 19(12), 2226–2238.

- Gogarten, J., G. Fournier, and O. Zhaxybayeva (2008). Gene transfer and the reconstruction of life's early history from genomic data. *Space Science Reviews* 135(1), 115–131.
- Gogarten, J., R. Murphey, and L. Olendzenski (1999). Horizontal gene transfer: pitfalls and promises. *Biol Bull.* 196(3), 359–362.
- Gogarten, J. and J. Townsend (2005). Horizontal gene transfer, genome innovation and evolution. *Nature Reviews Microbiology* 3(9), 679–687.
- Goldman, N. and Z. Yang (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* 11(5), 725–736.
- Gophna, U. and Y. Ofra (2011). Lateral acquisition of genes is affected by the friendliness of their products. *Proc Natl Acad Sci U S A* 108(1), 343–348.
- Graham, J., B. McNeney, and F. Seillier-Moisewitsch (2005). Stepwise detection of recombination breakpoints in sequence alignments. *Bioinformatics* 21(5), 589–595.
- Grassly, N. and E. Holmes (1997). A likelihood method for the detection of selection and recombination using nucleotide sequences. *Mol Biol Evol.* 14(3), 239–247.
- Griffith, F. (1928). The significance of pneumococcal types. *J Hyg* 27(02), 113–159.
- Gu, Z., D. Nicolae, H. Lu, and W. Li (2002). Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet.* 18(12), 609–613.
- Haft, D., J. Selengut, E. Mongodin, and K. Nelson (2005). A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput Biol.* 1(6), e60.
- Hao, W. and G. Golding (2004). Patterns of bacterial gene movement. *Mol Bio Evol.* 21(7), 1294–1307.
- Hao, W. and G. Golding (2006). The fate of laterally transferred genes: life in the fast lane to adaptation or death. *Genome Res.* 16(5), 636–643.
- Hao, W. and G. Golding (2008a). High rates of lateral gene transfer are not due to false diagnosis of gene absence. *Gene* 421(1-2), 27–31.
- Hao, W. and G. Golding (2008b). Uncovering rate variation of lateral gene transfer during bacterial genome evolution. *BMC Genomics* 9(235).
- Hao, W. and G. Golding (2009). Does gene translocation accelerate the evolution of laterally transferred genes? *Genetics* 182(4), 1365–1375.
- Hasegawa, M., H. Kishino, and T. Yano (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* 22(2), 160–174.

- Heidelberg, J., W. Nelson, T. Schoenfeld, and D. Bhaya (2009). Germ warfare in a microbial mat community: CRISPRs provide insights into the co-evolution of host and viral genomes. *PLoS One* 4(1), e4169.
- Hein, J. (1993). A heuristic method to reconstruct the history of sequences subject to recombination. *J Mol Evol.* 36(4), 396–405.
- Heinemann, J., G. Sprague Jr, et al. (1989). Bacterial conjugative plasmids mobilize DNA transfer between bacteria and yeast. *Nature* 340(6230), 205–209.
- Helgason, E., O. Økstad, D. Caugant, H. Johansen, A. Fouet, M. Mock, I. Hegna, and A. Kolstø (2000). *Bacillus anthracis*, *Bacillus cereus*, and *Bacillus thuringiensis*—one species on the basis of genetic evidence. *Appl Environ Microbiol.* 66(6), 2627–2630.
- Hernandez, E., F. Ramisse, J. Ducoureau, T. Cruel, and J. Cavallo (1998). *Bacillus thuringiensis* subsp. *konkukian* (serotype H34) superinfection: case report and experimental evidence of pathogenicity in immunosuppressed mice. *J Clin Microbiol.* 36(7), 2138–2139.
- Higgs, P., W. Hao, and G. Golding (2007). Identification of conflicting selective effects on highly expressed genes. *Evolutionary Bioinformatics Online* 3, 1.
- Holmgren, L. (2010). Horizontal gene transfer: you are what you eat. *Biochem Biophys Res Commun.* 396(1), 147–151.
- Hothorn, T., K. Hornik, M. Van De Wiel, and A. Zeileis (2006). A lego system for conditional inference. *The American Statistician* 60(3), 257–263.
- Hotopp, J., M. Clark, D. Oliveira, J. Foster, P. Fischer, M. Torres, J. Giebel, N. Kumar, N. Ishmael, S. Wang, J. Ingram, R. Nene, J. Shepard, J. Tomkins, S. Richards, D. Spiro, E. Ghedin, B. Slatko, H. Tettelin, and W. J.H. (2007). Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science* 317(5845), 1753–1756.
- Huelsenbeck, J. and F. Ronquist (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17(8), 754–755.
- Hurst, L. et al. (2002). The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet.* 18(9), 486.
- Husmeier, D. and G. McGuire (2003). Detecting recombination in 4-taxa DNA sequence alignments with Bayesian hidden Markov models and Markov chain Monte Carlo. *Mol Biol Evol.* 20(3), 315–337.
- Itoh, T., K. Takemoto, H. Mori, and T. Gojobori (1999). Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Mol Biol Evol.* 16(3), 332–346.

- Jacob, F., D. Perrin, C. Sanchez, J. Monod, and S. Edelman (1960). The operon: a group of genes with expression coordinated by an operator. *C. R. Acad. Sci.* 250(6), 1727–1729.
- Jain, R., M. Rivera, and J. Lake (1999). Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl. Acad. Sci. USA.* 96(7), 3801–3806.
- Jain, R., M. Rivera, J. Moore, and J. Lake (2002). Horizontal gene transfer in microbial genome evolution. *Theor Popul Biol.* 61(4), 489–495.
- Jakobsen, I. and S. Easteal (1996). A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. *Comput Appl Biosci.* 12(4), 291–295.
- Jakobsen, I., S. Wilson, and S. Easteal (1997). The partition matrix: exploring variable phylogenetic signals along nucleotide sequence alignments. *Mol Biol Evol.* 14(5), 474–484.
- Janga, S., W. Lamboy, A. Huerta, and G. Moreno-Hagelsieb (2006). The distinctive signatures of promoter regions and operon junctions across prokaryotes. *Nucleic Acids Res.* 34(14), 3980–3987.
- Jansen, R., J. Embden, W. Gaastra, and L. Schouls (2002). Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol.* 43(6), 1565–1575.
- John, S., O. Aharon, and H. Christopher (2011). Differences in lateral gene transfer in hypersaline versus thermal environments. *BMC Evolutionary Biology* 11, 199.
- Jordan, I., I. Rogozin, Y. Wolf, and E. Koonin (2002). Microevolutionary genomics of bacteria. *Theor Popul Biol.* 61(4), 435–447.
- Jukes, T. (1978). Neutral changes during divergent evolution of hemoglobins. *J Mol Evol.* 11(3), 267–269.
- Jukes, T. and C. Cantor (1969). Evolution of protein molecules. pp. 21–123.
- Kimura, M. (1969). The rate of molecular evolution considered from the standpoint of population genetics. *Proc Natl Acad Sci U S A.* 63(4), 1181–1188.
- Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol.* 16(2), 111–120.
- Kimura, M. (1981). Possibility of extensive neutral evolution under stabilizing selection with special reference to nonrandom usage of synonymous codons. *Proc Natl Acad Sci U S A.* 78(9), 5773.
- Kimura, M. et al. (1968). Evolutionary rate at the molecular level. *Nature* 217(5129), 624–626.

- Kimura, M. and J. Crow (1964). The number of alleles that can be maintained in a finite population. *Genetics* 49(4), 725.
- King, J. and T. Jukes (1969). Non-darwinian evolution. *Science* 164, 788–798.
- Kondrashov, F., E. Koonin, I. Morgunov, T. Finogenova, M. Kondrashova, et al. (2006). Evolution of glyoxylate cycle enzymes in metazoa: evidence of multiple horizontal transfer events and pseudogene formation. *Biol Direct.* 1, 31.
- Koonin, E., K. Makarova, and L. Aravind (2001). Horizontal Gene Transfer in Prokaryotes: Quantification and Classification 1. *Annu. Rev. Microbiol.* 55(1), 709–742.
- Koski, L., R. Morton, and G. Golding (2001). Codon bias and base composition are poor indicators of horizontally transferred genes. *Mol. Biol. Evol.* 18(3), 404–412.
- Koufopanou, V., M. Goddard, and A. Burt (2002). Adaptation for horizontal transfer in a homing endonuclease. *Mol Biology Evol.* 19(3), 239–246.
- Kremling, A., K. Jahreis, J. Lengeler, and E. Gilles (2000). The organization of metabolic reaction networks: a signal-oriented approach to cellular models. *Metabolic Engineering* 2(3), 190–200.
- Kuo, C. and H. Ochman (2010). The extinction dynamics of bacterial pseudogenes. *PLoS Genet.* 6(8), e1001050.
- Kurland, C., B. Canback, and O. Berg (2003). Horizontal gene transfer: a critical view. *Proc. Natl. Acad. Sci. USA.* 100(17), 9658–9662.
- Lartillot, N. and H. Philippe (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol.* 21(6), 1095–1109.
- Lawrence, J. (1997). Selfish operons and speciation by gene transfer. *Trends Microbiol.* 5(9), 355–359.
- Lawrence, J. (1999a). Gene transfer, speciation, and the evolution of bacterial genomes. *Curr Opin Microbiol.* 2(5), 519–523.
- Lawrence, J. (1999b). Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. *Curr Opin Genet Dev.* 9(6), 642–648.
- Lawrence, J. (2002a). Gene transfer in bacteria: speciation without species? *Theor Popul Biol.* 61(4), 449–460.
- Lawrence, J. (2002b). Shared strategies in gene organization among prokaryotes and eukaryotes. *Cell* 110(4), 407–413.
- Lawrence, J. and D. Hartl (1992). Inference of horizontal genetic transfer from molecular data: an approach using the bootstrap. *Genetics* 131(3), 753–760.

- Lawrence, J. and H. Hendrickson (2003). Lateral gene transfer: when will adolescence end? *Mol Microbiol.* 50(3), 739–749.
- Lawrence, J. and H. Hendrickson (2005). Genome evolution in bacteria: order beneath chaos. *Curr Opin Microbiol.* 8(5), 572–578.
- Lawrence, J., R. Hendrix, and S. Casjens (2001). Where are the pseudogenes in bacterial genomes? *Trends Microbiol.* 9(11), 535–540.
- Lawrence, J. and H. Ochman (1997). Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol* 44(4), 383–397.
- Lawrence, J. and H. Ochman (1998). Molecular archaeology of the Escherichia coli genome. *Proc Natl Acad Sci U S A.* 95(16), 9413–9417.
- Lawrence, J. and H. Ochman (2002). Reconciling the many faces of lateral gene transfer. *Trends Microbiol.* 10(1), 1–4.
- Lawrence, J., H. Ochman, and D. Hartl (1992). The evolution of insertion sequences within enteric bacteria. *Genetics* 131(1), 9–20.
- Lawrence, J. and J. Roth (1996). Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* 143(4), 1843–1846.
- Lederberg, J. and E. Tatum (1946). Gene recombination in Escherichia coli. *Nature* 158, 558–558.
- Lennon, J., S. Khatana, M. Marston, and J. Martiny (2007). Is there a cost of virus resistance in marine cyanobacteria? *The ISME journal* 1(4), 300–312.
- Letunic, I. and P. Bork (2006). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* 23(1), 127–128.
- Li, S., D. Pearl, and H. Doss (2000). Phylogenetic tree construction using Markov chain Monte Carlo. *Journal of the American Statistical Association*, 493–508.
- Li, W. (1993). Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J Mol Evol.* 36(1), 96–99.
- Li, W., T. Gojobori, M. Nei, et al. (1981). Pseudogenes as a paradigm of neutral evolution. *Nature* 292(5820), 237–239.
- Li, W., C. Wu, and C. Luo (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol.* 2(2), 150–174.
- Losick, R. and L. Shapiro (1999). Changing views on the nature of the bacterial cell: from biochemistry to cytology. *J Bacteriol.* 181(14), 4143–4145.

- Lovett, S., R. Hurley, V. Suter Jr, R. Aubuchon, and M. Lebedeva (2002). Crossing over between regions of limited homology in *Escherichia coli*: RecA-dependent and RecA-independent pathways. *Genetics* 160(3), 851–859.
- Luis, A., M. Gabriel, E. Luis, S. Valeria, H. Luis, and O. Gabriela (2011). Understanding the evolutionary relationships and major traits of bacillus through comparative genomics. *BMC Genomics* 11(332).
- Majewski, J. and F. Cohan (1998). The effect of mismatch repair and heteroduplex formation on sexual isolation in bacillus. *Genetics* 148(1), 13–18.
- Majewski, J. and F. Cohan (1999a). Adapt globally, act locally: the effect of selective sweeps on bacterial sequence diversity. *Genetics* 152(4), 1459–1474.
- Majewski, J. and F. Cohan (1999b). DNA sequence similarity requirements for interspecific recombination in *Bacillus*. *Genetics* 153(4), 1525–1533.
- Marri, P., W. Hao, and G. Golding (2006). Gene gain and gene loss in streptococcus: is it driven by habitat? *Mol Biol Evol.* 23(12), 2379–2391.
- Marri, P., W. Hao, and G. Golding (2007). The role of laterally transferred genes in adaptive evolution. *BMC evolutionary biology* 7(Suppl 1), S8.
- Martin, D. and E. Rybicki (2000). RDP: detection of recombination amongst aligned sequences. *Bioinformatics* 16(6), 562–563.
- Martin, W., T. Rujan, E. Richly, A. Hansen, S. Cornelsen, T. Lins, D. Leister, B. Stoebe, M. Hasegawa, and D. Penny (2002). Evolutionary analysis of arabisopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc. Natl. Acad. Sci. USA.* 99(19), 12246–12251.
- McDaniel, L., E. Young, J. Delaney, F. Ruhnau, K. Ritchie, and J. Paul (2010). High frequency of horizontal gene transfer in the oceans. *Science* 330(6000), 50.
- McDaniel, T. and J. Kaper (1997). A cloned pathogenicity island from enteropathogenic *Escherichia coli* confers the attaching and effacing phenotype on *E. coli* K-12. *Mol Microbiol.* 23(2), 399–407.
- McGuire, G. and F. Wright (2000). TOPAL 2.0: improved detection of mosaic sequences within multiple alignments. *Bioinformatics* 16(2), 130–134.
- Medini, D., C. Donati, H. Tettelin, V. Massignani, and R. Rappuoli (2005). The microbial pan-genome. *Curr Opin Genet Dev.* 15(6), 589–594.
- Mira, A., H. Ochman, and N. Moran (2001). Deletional bias and the evolution of bacterial genomes. *Trends Genet.* 17(10), 589–596.

- Mojica, F., C. Diez-Villasenor, J. Garcia-Martinez, and C. Almendros (2009). Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* 155(3), 733–740.
- Mojica, F., C. Diez-Villasenor, E. Soria, and G. Juez (2000). Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Mol Microbiol* 36(1), 244–246.
- Moore, J. and J. Haber (1996). Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in *Saccharomyces cerevisiae*. *Mol Cell Biol*. 16(5), 2164–2173.
- Morse, M., E. Lederberg, and J. Lederberg (1956). Transduction in *Escherichia coli* K-12. *Genetics* 41(1), 142–146.
- Muller, H. (1932). Some genetic aspects of sex. *Am Nat* 66(703), 118–138.
- Nei, M. and T. Gojobori (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol*. 3(5), 418–426.
- Nei, M., Y. Suzuki, and M. Nozawa (2010). The neutral theory of molecular evolution in the genomic era. *Annu Rev Genomics Hum Genet*. 11, 265–289.
- Nielsen, R. and Z. Yang (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148(3), 929–936.
- Ochman, H., J. Lawrence, E. Groisman, et al. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* 405(6784), 299–304.
- Ogram, A., G. Sayler, and T. Barkay (1987). The extraction and purification of microbial DNA from sediments. *J Microbiol Methods* 7(2), 57–66.
- Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature* 246(5428), 96–98.
- Ohta, T. (1974). Mutational pressure as the main cause of molecular evolution and polymorphism. *Nature* 252, 351–354.
- Ohta, T. (1994). Further examples of evolution by gene duplication revealed through DNA sequence comparisons. *Genetics* 138(4), 1331–1337.
- Ohta, T., J. Gillespie, et al. (1996). Development of neutral and nearly neutral theories. *Theor Popul Biol*. 49(2), 128–142.
- Okinaka, R., T. Pearson, and P. Keim (2006). Anthrax, but not *Bacillus anthracis*? *PLoS Pathog* 2(11), e122.

- Omer, S., A. Kovacs, Y. Mazor, and U. Gophna (2010). Integration of a foreign gene into a native complex does not impair fitness in an experimental model of lateral gene transfer. *Mol Biol Evol.* 27(11), 2441–2445.
- Orengo, C., T. Flores, W. Taylor, and J. Thornton (1993). Identification and classification of protein fold families. *Protein Eng.* 6(5), 485–500.
- Pál, C. and L. Hurst (2004). Evidence against the selfish operon theory. *Trends Genet.* 20(6), 232–234.
- Paradis, E., J. Claude, and K. Strimmer (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20(2), 289–290.
- Perez-Losada, M., E. Browne, A. Madsen, T. Wirth, R. Viscidi, and K. Crandall (2006). Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. *Infect Genet Evol.* 6(2), 97–112.
- Pertea, M., K. Ayanbule, M. Smedinghoff, and S. Salzberg (2009). OperonDB: a comprehensive database of predicted operons in microbial genomes. *Nucl. Acids Res.* 37(suppl 1), D479–D482.
- Popa, O., E. Hazkani-Covo, G. Landan, W. Martin, and T. Dagan (2011). Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Res.* 21(4), 599–609.
- Portnoy, D., S. Moseley, and S. Falkow (1981). Characterization of plasmids and plasmid-associated determinants of *Yersinia enterocolitica* pathogenesis. *Infection and immunity* 31(2), 775–782.
- Posada, D. (2002). Evaluation of methods for detecting recombination from DNA sequences: empirical data. *Mol Biol Evol.* 19(5), 708–717.
- Posada, D. and K. Crandall (2001). Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *PNAS* 98(24), 13757–1363.
- Posada, D., K. Crandall, and E. Holmes (2002). Recombination in evolutionary genomics. *Annu Rev Genet.* 36(1), 75–97.
- Price, M., A. Arkin, and E. Alm (2006). The life-cycle of operons. *PLoS Genet.* 2(6), e96.
- Price, M., K. Huang, E. Alm, and A. Arkin (2005). A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucl. Acids Res.* 33(3), 880–892.
- Price, M., K. Huang, A. Arkin, and E. Alm (2005). Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome Res.* 15(6), 809–819.
- Ragan, M. (2001). Detection of lateral gene transfer among microbial genomes. *Curr Opin Genet Dev.* 11(6), 620–626.

- Rasko, D., J. Ravel, O. Økstad, E. Helgason, R. Cer, L. Jiang, K. Shores, D. Fouts, N. Tourasse, S. Angiuoli, et al. (2004). The genome sequence of *Bacillus cereus* ATCC 10987 reveals metabolic adaptations and a large plasmid related to *Bacillus anthracis* pXO1. *Nucleic Acids Res.* 32(3), 977–988.
- Ravel, J. and C. Fraser (2005). Genomics at the genus scale. *Trends Microbiol.* 13(3), 95–97.
- Rest, J. and D. Mindell (2003). Retroids in archaea: phylogeny and lateral origins. *Mol Biol Evol.* 20(7), 1134–1132.
- Retchless, A. and J. Lawrence (2007). Temporal fragmentation of speciation in bacteria. *Science* 317(5841), 1093–1096.
- Riley, M. (1993). Functions of the gene products of escherichia coli. *Microbiol Rev.* 57(4), 862–952.
- Roback, P., J. Beard, D. Baumann, C. Gille, K. Henry, S. Krohn, H. Wiste, M. Voskuil, C. Rainville, and R. Rutherford (2007). A predicted operon map for *Mycobacterium tuberculosis*. *Nucleic Acids Res.* 35(15), 5085–5095.
- Roca, M., L. Davide, L. Davide, M. Mendes-Costa, R. Schwan, and A. Wheals (2004). Conidial anastomosis fusion between *Colletotrichum* species. *Mycological research* 108(11), 1320–1326.
- Rocha, E., J. Smith, L. Hurst, M. Holden, J. Cooper, N. Smith, and E. Feil (2006). Comparisons of dNdS are time dependent for closely related bacterial genomes. *J Theor Biol.* 239(2), 226–235.
- Sabatti, C., L. Rohlin, M. Oh, and J. Liao (2002). Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acids Res.* 30(13), 2886–2893.
- Salgado, H., G. Moreno-Hagelsieb, T. Smith, and J. Collado-Vides (2000). Operons in *Escherichia coli*: genomic analyses and predictions. *Proc Natl Acad Sci U S A* 97(12), 6652–6657.
- Sasakawa, C., K. Kamata, T. Sakai, S. Makino, M. Yamada, N. Okada, and M. Yoshikawa (1988). Virulence-associated genetic regions comprising 31 kilobases of the 230-kilobase plasmid in *Shigella flexneri* 2a. *J Bacteriol.* 170(6), 2480–2484.
- Sawyer, S. (1989). Statistical tests for detecting gene conversion. *Mol Biol Evol.* 6(5), 526–538.
- Schloss, P. and J. Handelsman (2005). Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl Environ Microbiol.* 71(3), 1501–1506.

- Seehausen, O. (2004). Hybridization and adaptive radiation. *Trends Ecol Evol.* 19(4), 198–207.
- Sharp, P. and W. Li (1987). The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 15(3), 1281–1295.
- Shinohara, A. and T. Ogawa (1995). Homologous recombination and the roles of double-strand breaks. *Trends Biochem Sci.* 20(10), 387–391.
- Siefert, J., K. Martin, F. Abdi, W. Widger, and G. Fox (1997). Conserved gene clusters in bacterial genomes provide further support for the primacy of RNA. *J Mol Evol.* 45(5), 467–472.
- Siguier, P., J. Filee, and M. Chandler (2006). Insertion sequences in prokaryotic genomes. *Curr Opin Microbiol.* 9(5), 526–531.
- Simmons, L., A. Goranov, H. Kobayashi, B. Davies, D. Yuan, A. Grossman, and G. Walker (2009). Comparison of responses to double-strand breaks between *Escherichia coli* and *Bacillus subtilis* reveals different requirements for SOS induction. *J Bacteriol.* 191(4), 1152–1161.
- Smith, J. (1992). Analyzing the mosaic structure of genes. *J Mol Evol.* 34(2), 126–129.
- Smith, J. and J. Haigh (1974). The hitch-hiking effect of a favourable gene. *Genet Res.* 23(1), 23–35.
- Stoebel, D. (2005). Lack of evidence for horizontal transfer of the lac operon into *Escherichia coli*. *Mol Biol Evol.* 22(3), 683–690.
- Stumpf, M. and G. McVean (2003). Estimating recombination rates from population-genetic data. *Nat Rev Genet.* 4(12), 959–968.
- Suyama, M., D. Torrents, and P. Bork (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 34(suppl 2), W609–W612.
- Suzuki, Y. and T. Gojobori (1999). A method for detecting positive selection at single amino acid sites. *Mol Biol Evol.* 16(10), 1315–1328.
- Takami, H., Y. Takaki, G. Chee, S. Nishi, S. Shimamura, H. Suzuki, S. Matsui, and I. Uchiyama (2004). Thermoadaptation trait revealed by the genome sequence of thermophilic geobacillus kaustophilus. *Nucleic Acids Res.* 32(21), 6292–6303.
- Takami, H., Y. Takaki, and I. Uchiyama (2002). Genome sequence of *Oceanobacillus iheyensis* isolated from the Iheya Ridge and its unexpected adaptive capabilities to extreme environments. *Nucleic Acids Res.* 30(18), 3927–3935.

- Tatum, E. and J. Lederberg (1947). Gene recombination in the bacterium *Escherichia coli*. *J Bacteriol.* 53(6), 673–684.
- Tatusov, R., E. Koonin, and D. Lipman (1997). A genomic perspective on protein families. *Science* 278(5338), 631–637.
- Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect. Math. Life Sci* 17, 57–86.
- Team, R. D. C. (2011). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.
- Thomas, C. and K. Nielsen (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol.* 3(9), 711–721.
- Ussery, D., T. Schou Larsen, K. Trevor Wilkes, C. Friis, P. Worning, A. Krogh, and S. Brunak (2001). Genome organisation and chromatin structure in *Escherichia coli*. *Biochimie* 83(2), 201–212.
- Vale, P. and T. Little (2010). CRISPR-mediated phage resistance and the ghost of coevolution past. *Proc. R. Soc. B* 277(1691), 2097–2103.
- Van de Peer, Y., J. Taylor, I. Braasch, and A. Meyer (2001). The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. *J Mol Evol.* 53(4), 436–446.
- van Dongen, S. (2000). *Graph clustering by flow simulation*. Ph. D. thesis, University of Utrecht.
- Venter, J., K. Remington, J. Heidelberg, A. Halpern, D. Rusch, J. Eisen, D. Wu, I. Paulsen, K. Nelson, W. Nelson, et al. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *science* 304(5667), 66–74.
- Vitreschak, A., D. Rodionov, A. Mironov, and M. Gelfand (2002). Regulation of riboflavin biosynthesis and transport genes in bacteria by transcriptional and translational attenuation. *Nucleic Acids Res.* 30(14), 3141–3151.
- Vos, M. (2009). Why do bacteria engage in homologous recombination? *Trends Microbiol.* 17(6), 226–232.
- Weaver, D. (1995). What to do at an end: DNA double-strand-break repair. *Trends Genet.* 11(10), 388–392.
- Weiller, G. (1998). Phylogenetic profiles: a graphical method for detecting genetic recombinations in homologous sequences. *Mol Biol Evol.* 15(3), 326–335.
- Wellner, A., M. Lurie, and U. Gophna (2007). Complexity, connectivity, and duplicability as barriers to lateral gene transfer. *Genome Biol* 8(8), R156.

- Whitaker, R. and J. Banfield (2006). Population genomics in natural microbial communities. *Trends Ecol Evol.* 21(9), 508–516.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin* 1(6), 80–83.
- Williamson, S. (2003). Adaptation in the env gene of hiv-1 and evolutionary theories of disease progression. *Mol Biol Evol.* 20(8), 1318–1325.
- Wolf, Y., I. Rogozin, N. Grishin, and E. Koonin (2002). Genome trees and the tree of life. *TRENDS in Genetics* 18(9), 472–479.
- Wong, W. and R. Nielsen (2004). Detecting selection in noncoding regions of nucleotide sequences. *Genetics* 167(2), 949–958.
- Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol Biol Evol.* 10(6), 1396–1401.
- Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer applications in the biosciences: CABIOS* 13(5), 555–556.
- Yang, Z. (2002). Inference of selection from multiple species alignments. *Curr Opin Genet Dev.* 12(6), 688–694.
- Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24(8), 1586–1591.
- Yang, Z. and J. Bielawski (2000). Statistical methods for detecting molecular adaptation. *Trends Ecol Evol.* 15(12), 496–503.
- Yang, Z. and R. Nielsen (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* 17(1), 32–43.
- Zhang, G., Z. Cao, Q. Luo, Y. Cai, and Y. Li (2006). Operon prediction based on SVM. *Comput Biol Chem.* 30(3), 233–240.
- Zhaxybayeva, O., P. Lapierre, and J. Gogarten (2004). Genome mosaicism and organismal lineages. *TRENDS in Genetics* 20(5), 254–260.
- Zheng, Y., J. Szustakowski, L. Fortnow, R. Roberts, and S. Kasif (2002). Computational identification of operons in microbial genomes. *Genome Res.* 12(8), 1221–1230.
- Zinder, N. and J. Lederberg (1952). Genetic exchange in salmonella. *J Bacteriol.* 64(5), 679–684.
- Zwick, M., M. Thomason, P. Chen, H. Johnson, S. Sozhamannan, A. Mateczun, and T. Read (2011). Genetic variation and linkage disequilibrium in bacillus anthracis. *Sci Rep.* 1(169).

Part V
APPENDICES

Appendix A

Insertion of Genes into Operons

Table A.1: A table showing the genomes used in the study and their NCBI accession number

Genome	Accession
Anoxybacillus flavithermus WK1	NC_011567
Bacillus amyloliquefaciens DSM 7	NC_014551
Bacillus amyloliquefaciens FZB42	NC_009725
Bacillus anthracis CI	NC_014335
Bacillus anthracis str. 'Ames Ancestor'	NC_007530
Bacillus anthracis str. A0248	NC_012659
Bacillus anthracis str. Ames	NC_003997
Bacillus anthracis str. CDC 684	NC_012581
Bacillus anthracis str. Sterne	NC_005945
Bacillus atrophaeus 1942	NC_014639
Bacillus cellulosilyticus DSM 2522	NC_014829
Bacillus cereus 03BB102	NC_012472
Bacillus cereus AH187	NC_011658
Bacillus cereus AH820	NC_011773
Bacillus cereus ATCC 10987	NC_003909
Bacillus cereus ATCC 14579	NC_004722
Bacillus cereus B4264	NC_011725
Bacillus cereus E33L	NC_006274
Bacillus cereus G9842	NC_011772
Bacillus cereus Q1	NC_011969
Bacillus cereus subsp. cytotoxis NVH 391-98	NC_009674
Bacillus clausii KSM-K16	NC_006582
Bacillus coagulans 2-6	NC_015634

Genome	Accession
Bacillus halodurans C-125	NC_002570
Bacillus licheniformis ATCC 14580	NC_006270
Bacillus licheniformis ATCC 14580	NC_006322
Bacillus megaterium DSM 319	NC_014103
Bacillus megaterium QM B1551	NC_014019
Bacillus pseudofirmus OF4	NC_013791
Bacillus pumilus SAFR-032	NC_009848
Bacillus selenitireducens MLS10	NC_014219
Bacillus subtilis BSn5	NC_014976
Bacillus subtilis subsp. spizizenii str. W23	NC_014479
Bacillus subtilis subsp. subtilis str. 168	NC_000964
Bacillus thuringiensis BMB171	NC_014171
Bacillus thuringiensis serovar konkukian str. 97-27	NC_005957
Bacillus thuringiensis str. Al Hakam	NC_008600
Bacillus weihenstephanensis KBAB4	NC_010184
Geobacillus kaustophilus HTA426	NC_006510
Geobacillus sp. C56-T3	NC_014206
Geobacillus sp. WCH70	NC_012793
Geobacillus sp. Y4.1MC1	NC_014650
Geobacillus sp. Y412MC52	NC_014915
Geobacillus sp. Y412MC61	NC_013411
Geobacillus thermodenitrificans NG80-2	NC_009328
Geobacillus thermoglucosidasius C56-YS93	NC_015660
Oceanobacillus iheyensis HTE831	NC_004193

Table A.2: A table showing columns: A: Total number of operons in genome, B: total number of genes in operons, C: number of protein coding genes in organism, D: size of genome in bp

Genome	A	B	C	D
A. flavithermus WK1	366	1496	2831	2846746
B. amyloliquefaciens DSM 7	430	1608	3893	3980199
B. amyloliquefaciens FZB42	351	1272	3693	3918589
B. anthracis CI	452	1554	5221	5196054
B. anthracis str. 'Ames Ancestor'	415	1440	5208	5227419
B. anthracis str. A0248	410	1444	5040	5227419
B. anthracis str. Ames	420	1442	5328	5227293
B. anthracis str. CDC 684	451	1523	5579	5230115

Genome	A	B	C	D
<i>B. anthracis</i> str. Sterne	437	1532	5289	5228663
<i>B. atrophaeus</i> 1942	359	1288	4186	4168266
<i>B. cellulosityticus</i> DSM 2522	466	1693	4266	4681672
<i>B. cereus</i> 03BB102	445	1515	5398	5269628
<i>B. cereus</i> AH187	451	1526	5424	5269030
<i>B. cereus</i> AH820	444	1538	5474	5302683
<i>B. cereus</i> ATCC 10987	388	1319	5602	5224283
<i>B. cereus</i> ATCC 14579	431	1488	5234	5411809
<i>B. cereus</i> B4264	439	1521	5398	5419036
<i>B. cereus</i> E33L	431	1541	5134	5300915
<i>B. cereus</i> G9842	434	1522	5488	5387334
<i>B. cereus</i> Q1	442	1535	5192	5214195
<i>B. cereus</i> subsp. <i>cytotoxis</i> NVH 391-98	347	1347	3833	4087024
<i>B. clausii</i> KSM-K16	524	1879	4096	4303871
<i>B. coagulans</i> 2-6	370	1306	2971	3073079
<i>B. halodurans</i> C-125	439	1563	4065	4202352
<i>B. licheniformis</i> ATCC 14580	484	1793	4173	4222597
<i>B. licheniformis</i> ATCC 14580	486	1814	4192	4222645
<i>B. megaterium</i> DSM 319	367	1379	5100	5097447
<i>B. megaterium</i> QM B1551	369	1393	5116	5097129
<i>B. pseudofirmus</i> OF4	416	1437	3921	3858997
<i>B. pumilus</i> SAFR-032	420	1577	3678	3704465
<i>B. selenitireducens</i> MLS10	335	1325	3255	3592487
<i>B. subtilis</i> BSn5	378	1318	4145	4093599
<i>B. subtilis</i> subsp. <i>spizizenii</i> str. W23	365	1307	4062	4027676
<i>B. subtilis</i> subsp. <i>subtilis</i> str. 168	366	1328	4176	4215606
<i>B. thuringiensis</i> BMB171	423	1468	5079	5330088
<i>B. thuringiensis</i> serovar <i>konkukian</i> str. 97-27	436	1557	5117	5237682
<i>B. thuringiensis</i> str. Al Hakam	357	1301	4736	5257091
<i>B. weihenstephanensis</i> KBAB4	439	1550	5155	5262775
<i>G. kaustophilus</i> HTA426	456	1758	3497	3544776
<i>G. sp.</i> C56-T3	423	1656	3315	3650813
<i>G. sp.</i> WCH70	393	1581	3128	3464618
<i>G. sp.</i> Y4.1MC1	437	1716	3605	3840330
<i>G. sp.</i> Y412MC52	441	1743	3420	3628883
<i>G. sp.</i> Y412MC61	435	1668	3407	3622844
<i>G. thermodenitrificans</i> NG80-2	427	1710	3392	3550319
<i>G. thermoglucosidasius</i> C56-YS93	458	1823	3676	3893306
<i>O. iheyensis</i> HTE831	422	1609	3500	3630528

An example of OperonDB output:

This is a gene pair from *B. subtilis* (NC_000964). The first two numbers (16077069 16077070) are GI's of the genes, the third number (90) is the confidence value that indicates the probability of the gene pair be in an operon. The fourth number (43) is the number of genomes (other than itself) in which the gene pair is conserved. All the numbers that follow are the GI's of the genes that are conserved in other genomes.

```
16077069 16077070 90 43 308171892 308171893 154684519
154684520 229600569 229601421 47525255 47525256 30260196
30260197 227812679 227812680 301051742 301051743 49183040
49183041 317126742 317126743 225862058 225862059 217957582
217957583 218901207 218901208 42779082 42779083 30018279
30018280 218235327 218234602 52145208 52145207 218895142
218895143 222093775 222093776 56961783 56961784 336112680
336112681 152973855 152973856 15612564 15612565 52078492
52078493 52783856 52783857 295702243 295702244 294496876
294496877 288554606 288554607 157690799 157690800 297582339
297582340 305672699 305672700 118475779 118475780 296500839
296500840 49479264 49476685 163938014 163938015 297528377
297528378 56418536 56418537 138893680 138893681 336233547
336233548 239825585 239825586 319765158 319765159 261417501
261417502 312109152 312109153 23097456 23097457
```

An example of an operon:

An example of an operon is shown. This is an operon from *B. subtilis* (NC_000964). It is the first operon in the genome based on OperonDB prediction. The first line is the header: representing genome and operon number. The next two lines are the genes in the operons. In the first line, the first number (410) represents the start base pairs of the gene, the second number (1750) is the end base pairs of the gene, the next column indicates strand (+), followed by the length of gene (446 amino acids), and the GI number (16077069) and finally the name of the gene (dnaA). Gene function and COG category are not shown due to lack of space. These lines are similar to lines in protein table (*.ptt) of GenBank.

```
>NC_000964_1_operon
410..1750 + 446 16077069 dnaA
1939..3075 + 378 16077070 dnaN
```

MCL output

The output from MCL program is given below (van Dongen, 2000). Note: Data is forced to fit the page by adding ‘ n’ characters in appropriate places

```
Output of command: mcxdeblast --score=b --sort=a --bcut=50
--m9 Blast
```

```
Index [Blast.tab] is sorted by alphabetic order
[/home/asher/GradSchool/Software/MCL/bin/mcxdeblast] 561
secondary elements not seen as primary element
[/home/asher/GradSchool/Software/MCL/bin/mcxdeblast] I
added all of them
[/home/asher/GradSchool/Software/MCL/bin/mcxdeblast] There
were 197727 elements in all
```

```
Output of command: mcxassemble -b Blast -q -r max --map
```

```
[mclIO] reading <Blast.map>
.....
[mclIO] read native interchange 197727x197727 matrix with
197727 entries
[mclIO] reading <Blast.raw>
.....
[mclIO] read raw interchange 197727x197727 matrix from
stream <Blast.raw>
[mclIO] writing <Blast.sym>
.....
[mclIO] wrote native interchange 197727x197727 matrix with
30648382 entries to stream <Blast.sym>
```

```
Output of command: mcl Blast.sym -scheme 2 -I 2.5 --append
-log=yes -o out.Blast.I25
```

```
[mclIO] reading <Blast.sym>
.....
[mclIO] read native interchange 197727x197727 matrix with
30648382 entries
[mcl] pid 27811
ite ----- chaos time hom(avg,lo,hi) expa expb expc fmv
1 ..... 77.35 265.34 0.97/0.05/6.04 2.16 0.76 0.76 18
2 ..... 64.48 66.53 0.87/0.04/25.65 2.04 0.87 0.66 25
3 ..... 65.21 46.40 0.79/0.06/24.19 1.57 0.79 0.52 19
4 ..... 62.90 31.05 0.76/0.10/14.41 1.19 0.72 0.37 14
5 ..... 40.34 17.71 0.75/0.10/15.86 1.08 0.59 0.22 7
6 ..... 25.94 6.50 0.74/0.13/12.84 1.02 0.62 0.14 3
```

7	9.23	1.84	0.78/0.18/2.90	1.01	0.66	0.09	0
8	9.68	1.06	0.82/0.20/2.33	1.00	0.66	0.06	0
9	5.78	0.75	0.87/0.24/1.07	1.00	0.71	0.04	0
10	4.27	0.53	0.93/0.21/1.25	1.00	0.72	0.03	0
11	4.35	0.38	0.96/0.20/1.30	1.00	0.76	0.02	0
12	3.16	0.29	0.98/0.27/1.00	1.00	0.84	0.02	0
13	2.44	0.26	0.99/0.33/1.00	1.00	0.92	0.02	0
14	2.15	0.25	0.99/0.33/1.00	1.00	0.95	0.02	0
15	2.86	0.23	0.99/0.38/1.00	1.00	0.97	0.02	0
16	1.93	0.22	0.99/0.34/1.00	1.00	0.98	0.02	0
17	1.72	0.21	1.00/0.44/1.00	1.00	0.99	0.02	0
18	1.99	0.21	1.00/0.41/1.00	1.00	0.99	0.02	0
19	0.99	0.19	1.00/0.48/1.00	1.00	0.99	0.02	0
20	0.31	0.20	1.00/0.76/1.00	1.00	1.00	0.02	0
21	0.42	0.21	1.00/0.71/1.00	1.00	1.00	0.02	0
22	0.43	0.21	1.00/0.69/1.00	1.00	1.00	0.02	0
23	0.25	0.20	1.00/0.76/1.00	1.00	1.00	0.02	0
24	0.08	0.21	1.00/0.92/1.00	1.00	1.00	0.02	0
25	0.00	0.20	1.00/1.00/1.00	1.00	1.00	0.02	0
26	0.00	0.21	1.00/1.00/1.00	1.00	1.00	0.02	0

[mcl] jury pruning marks: <19,82,91>, out of 100

[mcl] jury pruning synopsis: <43.8 or poor> (cf -scheme, -do log)

[mclIO] writing <out.Blast.I25>

.....

[mclIO] wrote native interchange 197727x10277 matrix with 197727 entries to stream <out.Blast.I25>

[mcl] 10277 clusters found

[mcl] output is in out.Blast.I25

Last command with no output: clmformat --dump -tab Blast.tab -lump-count 500 -imx Blast.sym -dump dump.out.Blast.I25 -icl out.Blast.I25 -dir fmt.Blast.I25

LGT in Operons

Table A.3: A table showing A: Number of genes that are laterally transferred into genomes
 B: Number of genes laterally transferred in operons excluding lateral transfer of operons

Genome	A	B
A. flavithermus WK1	381	61
B. amyloliquefaciens DSM 7	594	49
B. amyloliquefaciens FZB42	551	15
B. anthracis CI	1203	40
B. anthracis str. 'Ames Ancestor'	1454	39
B. anthracis str. A0248	1370	39
B. anthracis str. Ames	1466	39
B. anthracis str. CDC 684	1425	42
B. anthracis str. Sterne	1273	43
B. atrophaeus 1942	649	15
B. cellulolyticus DSM 2522	521	62
B. cereus 03BB102	1318	39
B. cereus AH187	1338	41
B. cereus AH820	1357	40
B. cereus ATCC 10987	1315	28
B. cereus ATCC 14579	1205	37
B. cereus B4264	1227	36
B. cereus E33L	1183	41
B. cereus G9842	1262	9
B. cereus Q1	1241	41
B. cereus subsp. cytotoxis NVH 391-98	666	24
B. clausii KSM-K16	465	55
B. coagulans 2-6	229	26
B. halodurans C-125	449	26
B. licheniformis ATCC 14580	850	72
B. licheniformis ATCC 14580	839	72
B. megaterium DSM 319	1220	17
B. megaterium QM B1551	1243	19
B. pseudofirmus OF4	320	17
B. pumilus SAFR-032	427	51
B. selenitireducens MLS10	331	33
B. subtilis BSn5	660	20

Genome	A	B
<i>B. subtilis</i> subsp. <i>spizizenii</i> str. W23	619	25
<i>B. subtilis</i> subsp. <i>subtilis</i> str. 168	713	21
<i>B. thuringiensis</i> BMB171	1076	37
<i>B. thuringiensis</i> serovar <i>konkukian</i> str. 97-27	1169	45
<i>B. thuringiensis</i> str. Al Hakam	984	26
<i>B. weihenstephanensis</i> KBAB4	1252	41
<i>G. kaustophilus</i> HTA426	547	86
<i>G. sp.</i> C56-T3	601	61
<i>G. sp.</i> WCH70	458	78
<i>G. sp.</i> Y4.1MC1	699	90
<i>G. sp.</i> Y412MC52	672	80
<i>G. sp.</i> Y412MC61	669	74
<i>G. thermodenitrificans</i> NG80-2	488	95
<i>G. thermoglucosidasius</i> C56-YS93	744	74
<i>O. iheyensis</i> HTE831	278	51

Recombination breakpoints

The following is an output from stepwise implementation of maximum chi square algorithm. The alignment is from a protein annotated as isochorismate synthase (GI: 308174886) from the organism *Bacillus amyloliquefaciens* DSM 7. The organism *Bacillus amyloliquefaciens* DSM 7 is represented as "bc0" in the output. The origin of the gene appeared to be *Myxococcus xanthus* DK 1622 chromosome indicated as NC_008095. The second genome is *Stigmatella aurantiaca* DW4/3-1 chromosome represented as NC_014623. The recombination breakpoints were detected at locations 933, 935, 937, 3135 and 3,136 with statistical significance at 10 percent level. The breakpoints are most likely due to homologous recombination.

There were 5 site-specific MaxChi statistics significant at the 10 percent level (90th percentile = 15.017, 95th percentile = 15.429):

Number	Location	MaxChi	pairs
1	933	15.022	(bc0_____ :NC_008095_)
2	935	15.022	(bc0_____ :NC_008095_)
3	937	15.022	(bc0_____ :NC_008095_)
4	3135	15.022	(bc0_____ :NC_014623_)
5	3136	15.022	(bc0_____ :NC_014623_)

Notes - "Location" is the polymorphic site just before the proposed breakpoint.

- MaxChi statistics significant at the 5 percent level indicated by a *

There is another example of an output from stepwise implementation of maximum chi square algorithm. The alignment is from inner-membrane translocator (GI: 297528882) from *Geobacillus* sp. C56-T3 chromosome. The protein is a membrane protein as indicated by the annotation. The recombination breakpoints were found before the start codon of gene in *Geobacillus* sp. C56-T3 chromosome and after the stop codon of the gene. They were near characters 1168 and 3140 respectively. The gene might have been originating from *Lysinibacillus sphaericus* C3-41 chromosome represented as NC_010382 or from *Thermaerobacter marianensis* DSM 12885 chromosome represented as NC_014831. The half-widow size was 50 bps.

There were 3 site-specific MaxChi statistics significant at the 10 percent level (90th percentile = 14.663, 95th percentile = 16.071):

Number	Location	MaxChi	pairs
1	1168	15.487	(NC_010382_:bc84_____)
2	3134	15.429	(bc84_____:NC_014831_)
3	3140	15.429	(bc84_____:NC_014831_)

Notes - "Location" is the polymorphic site just before the proposed breakpoint.

- MaxChi statistics significant at the 5 percent level indicated by a *

The following is an example where the stepwise approach was more advantageous than a single step approach. The origin of the gene was from *Lysinibacillus sphaericus* C3-41 chromosome represented as NC_010382. The recipient organism was *Geobacillus thermodenitrificans* NG80-2 chromosome represented as "bc87". The gene, that was transferred was a homolog of ABC transporter permease. The first part shows the step 1 of the algorithm. The second part is the step 2 of the algorithm. In step 2, breakpoints near the start were detected.

The step 1:

Step 1:

There were 3 site-specific MaxChi statistics significant at the 10 percent level (90th percentile = 19.817, 95th percentile = 21.696):

Number	Location	MaxChi	pairs
1	3924	20.376	(bc87_____:NC_010382_)

2	3941	25.452*	(bc87_____:NC_010382_)
3	3942	20.376	(bc87_____:NC_010382_)

Notes - "Location" is the polymorphic site just before the proposed breakpoint.

- MaxChi statistics significant at the 5 percent level indicated by a *

The step 2:

There were 1 site-specific MaxChi statistics significant at the 10 percent level (90th percentile = 19.461, 95th percentile = 20.376):

Number	Location	MaxChi	pairs
1	1071	19.817	(bc87_____:NC_010382_)

Notes - "Location" is the polymorphic site just before the proposed breakpoint.

- MaxChi statistics significant at the 5 percent level indicated by a *

Appendix B

Effects of LGT on Neighbouring Genes

Species of tree lengths

The data lists the species that were dropped from the trees. ‘Non-N’ means non-neighbouring genes. ‘NA’ means no species were dropped. After the species were dropped, the names of the species were same in each point in upstream, downstream and non-neighbouring dataset.

Drop upstreams: NC_014171 NC_012581 NC_007530 NC_011772

Drop downstreams: NA

Drop Non-N: NC_006582 NC_002570 NC_014829 NC_014219 NC_014019 NC_013411
NC_015660 NC_009674 NC_014171 NC_010184 NC_011772 NC_012581 NC_007530
NC_004193

Drop upstreams: NC_014171 NC_009674

Drop downstreams: NC_011969

Drop Non-N: NC_004193 NC_000964 NC_014551 NC_015634 NC_013411 NC_006510
NC_014915 NC_014650 NC_009674 NC_010184 NC_014171 NC_004722 NC_008600
NC_011969 NC_003997 NC_003909 NC_002570

Drop upstreams: NA

Drop downstreams: NC_011725 NC_011773 NC_004722

Drop Non-N: NC_013791 NC_011567 NC_012793 NC_014915 NC_014206 NC_006510
NC_000964 NC_014551 NC_009674 NC_011725 NC_004722 NC_006274 NC_011773
NC_002570

Drop upstreams: NC_003909

Drop downstreams: NC_003997 NC_011969

Drop Non-N: NC_000964 NC_014551 NC_015634 NC_014019 NC_002570 NC_014650
NC_013411 NC_014915 NC_006510 NC_009674 NC_010184 NC_011772 NC_014171
NC_004722 NC_003909 NC_011969 NC_003997 NC_004193

Drop upstreams: NA

Drop downstreams: NC_011658 NC_003909

Drop Non-N: NC_014206 NC_014976 NC_004193 NC_013791 NC_002570 NC_014829
NC_010184 NC_011772 NC_011969 NC_003909 NC_011658 NC_009674

Drop upstreams: NA

Drop downstreams: NC_015660

Drop Non-N: NC_014551 NC_014650 NC_015660 NC_011567 NC_014103

Drop upstreams: NC_015660

Drop downstreams: NC_006270 NC_014103 NC_006510 NC_009848

Drop Non-N: NC_015660 NC_014650 NC_006510 NC_009848 NC_006270 NC_009725
NC_014551 NC_014103 NC_005957 NC_010184 NC_003909 NC_004722 NC_011658
NC_011772 NC_012472 NC_013791

Drop upstreams: NA

Drop downstreams: NA

Drop Non-N: NC_009725 NC_014551 NC_009848 NC_014019 NC_014206 NC_014915
NC_006510 NC_011772 NC_004722 NC_010184 NC_011658 NC_011969 NC_012581
NC_003997 NC_007530 NC_005957 NC_006274 NC_004193

Drop upstreams: NA

Drop downstreams: NC_009848 NC_009674 NC_006274 NC_010184 NC_014171
NC_005957 NC_014019

Drop Non-N: NC_014829 NC_009848 NC_014639 NC_014976 NC_014019 NC_013411
NC_006510 NC_015660 NC_009674 NC_010184 NC_014171 NC_006274 NC_008600
NC_005957 NC_004193

Drop upstreams: NA

Drop downstreams: NC_012472 NC_011773 NC_012659 NC_005945

Drop Non-N: NC_009328 NC_012793 NC_014019 NC_009848 NC_004722 NC_011773
NC_012472 NC_012659 NC_005945 NC_006274 NC_015634

Drop upstreams: NC_003997

Drop downstreams: NC_011567 NC_014650 NC_015660 NC_011725 NC_014335
NC_009674

Drop Non-N: NC_002570 NC_000964 NC_014551 NC_011567 NC_014915 NC_006510
NC_014650 NC_015660 NC_009674 NC_014335 NC_011725 NC_014171 NC_003997
NC_006274 NC_014219

Drop upstreams: NA

Drop downstreams: NA

Drop Non-N: NC_009848 NC_009725 NC_014639 NC_014479 NC_015634 NC_013791
NC_012581 NC_014103

Drop upstreams: NA

Drop downstreams: NA

Drop Non-N: NC_014219 NC_006322 NC_014915 NC_012793 NC_014551 NC_014019
NC_010184 NC_011658 NC_011772 NC_011969 NC_014171 NC_003997 NC_007530
NC_006582

Drop upstreams: NC_008600

Drop downstreams: NC_011725 NC_004722

Drop Non-N: NC_015634 NC_011567 NC_014019 NC_011725 NC_014171 NC_004722
NC_006274 NC_008600

Drop upstreams: NC_012793

Drop downstreams: NC_008600 NC_011658 NC_011772 NC_011969 NC_006274
NC_014103

Drop Non-N: NC_014915 NC_012793 NC_014650 NC_014103 NC_009674 NC_010184
NC_011772 NC_011658 NC_011969 NC_006274 NC_008600 NC_014206

Drop upstreams: NA

Drop downstreams: NC_012793

Drop Non-N: NC_006582 NC_013791 NC_012793 NC_014915 NC_014103

Drop upstreams: NC_005957 NC_012472 NC_008600

Drop downstreams: NC_004722

Drop Non-N: NC_011567 NC_014206 NC_006510 NC_004722 NC_008600 NC_012472
NC_005957 NC_009674

Drop upstreams: NC_009725 NC_014479 NC_014639

Drop downstreams: NC_008600 NC_014335

Drop Non-N: NC_006322 NC_009725 NC_014479 NC_014639 NC_009848 NC_014019
NC_014171 NC_008600 NC_014335 NC_006270

Drop upstreams: NA

Drop downstreams: NC_014171 NC_003909

Drop Non-N: NC_002570 NC_014829 NC_014171 NC_003909 NC_004193

Drop upstreams: NC_011773 NC_012472 NC_012659

Drop downstreams: NC_014171

Drop Non-N: NC_014171 NC_012472 NC_011773 NC_012659

Tree lengths

The table below indicates the number of gene families, minimum, median, mean, maximum and variance of the tree lengths of the genes that are upstream of laterally transferred genes. Tree lengths that were equal to zero were not included because they did not represent a topology. Moreover, the extreme outliers were removed from the dataset before any calculations. The minimum tree lengths were small because the genes might be highly conserved. The mean and median tree lengths were approximately constant across different models. The variance was high because upstream genes might either be from regions that had a variety of rates of evolution, or they were rapidly evolving (such as genes for pathogenicity) or highly conserved (such as regulatory genes that are commonly found upstream of operons and higher structures).

Table B.1: Table showing summary statistics of tree-length distribution of the upstream genes

Model	N	Minimum	Median	Mean	Maximum	Variance
M0	20	0.00002	0.39580	1.13300	4.46900	2.079086
M1a	19	0.000022	0.386800	1.050000	4.494000	2.005926
M2a	18	0.00002	0.37350	0.88930	4.49400	1.536020
M3	19	0.00002	0.38740	1.07000	4.59200	2.084294
M7	19	0.000021	0.387500	1.047000	4.376000	1.999425
M8	19	0.00002	0.38750	1.07500	4.55700	2.096653

The table below show the number of genes, minimum, median, mean, maximum and variance of the genes that were non-neighbouring. Extreme outliers and numbers that were equal to zero were removed. The minimum tree length was higher than that of upstream genes. However, the median, mean and maximum were lower than that of upstream genes. The results indicated that non-neighbouring genes evolved at a slower rate as compared to upstream genes when the name and numbers of species were corrected. Moreover, the variations in non-neighbouring genes were lesser than that of the upstream genes indicating that the non-neighbouring genes consisted of homogeneous set of genes. These might include genes that were involved in essential cell survival function and might be found in operons. Moreover, the results were consistent across different models, indicating that the data was not affected by the choice of model.

The table below shows the number of genes families, minimum, median, mean, maximum and variance. The tree lengths that were equal to zero and extreme outliers were removed from the data set. The minimum, median, mean and maximum tree lengths were higher than that in non-neighbouring set indicating that the rate of evolution was higher in downstream genes than that in non- neighbouring genes. Similar to upstream genes, the variance was higher than the non-neighbouring genes. Whereas the median and mean of the distribution were approximately similar, suggesting that laterally transferred genes affected

Table B.2: Table showing summary statistics of tree-length distribution of the non-neighbouring genes

Model	N	Minimum	Median	Mean	Maximum	Variance
M0	19	0.01137	0.80260	0.89000	2.28800	0.6354727
M1a	19	0.01113	0.72870	0.86310	2.17500	0.5928212
M2a	19	0.01113	0.72870	0.86080	2.17500	0.5890382
M3	19	0.01183	0.78710	0.89370	2.36200	0.6459752
M7	19	0.01138	0.78410	0.89130	2.36100	0.643324
M8	19	0.01138	0.78410	0.89600	2.36100	0.6496485

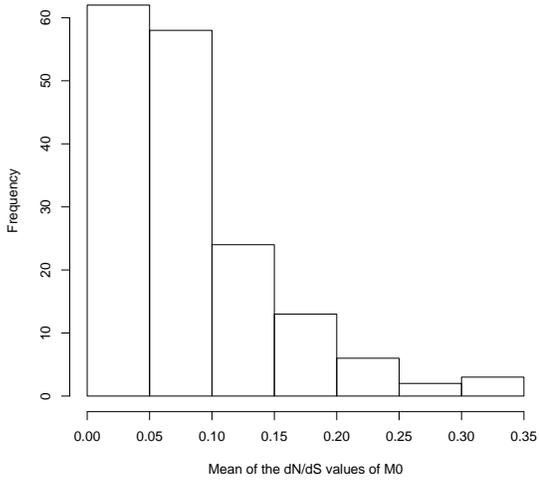
the upstream and downstream genes equally. Similar to upstream and non-neighbouring genes, the choice of codon substitution model did not affect the results.

Table B.3: Table showing summary statistics of tree-length distribution of the downstream genes

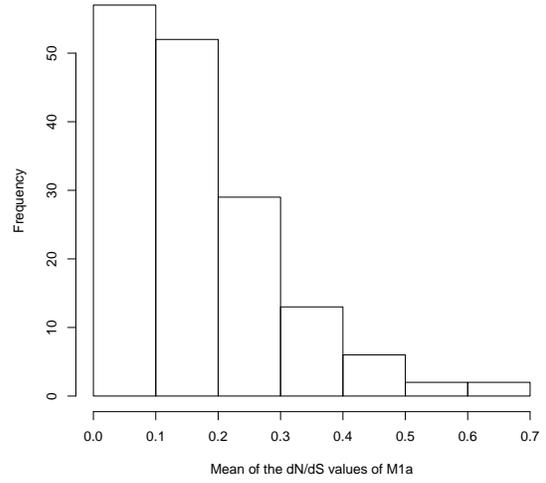
Model	N	Minimum	Median	Mean	Maximum	Variance
M0	20	0.02745	0.59930	1.22400	4.46400	2.079693
M1a	19	0.02824	0.54480	1.22300	5.16400	2.519825
M2a	18	0.02884	0.49400	1.03600	5.16400	1.940562
M3	18	0.02884	0.49620	1.08700	5.01100	2.285694
M7	20	0.02791	0.60410	1.38900	5.07800	2.941847
M8	18	0.02884	0.49800	1.10100	5.16800	2.402136

Distributions of weighted averages of dN/dS ratios

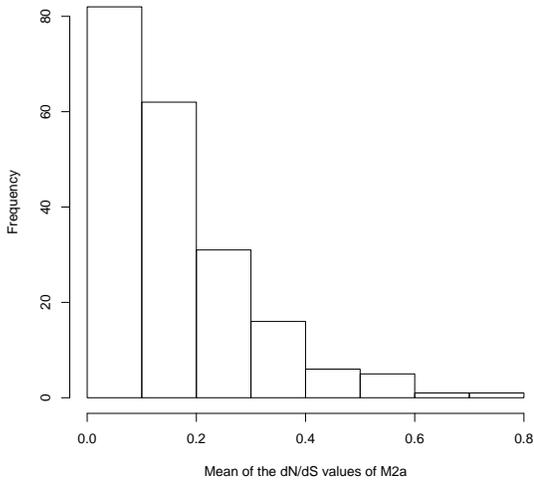
Histogram of the dN/dS values of the upstream genes



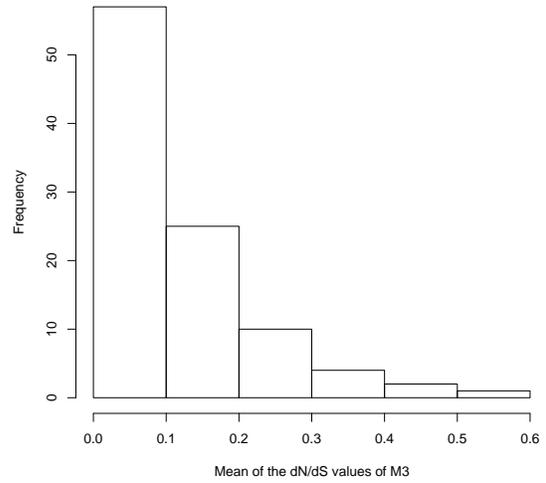
Histogram of the dN/dS values of the upstream genes

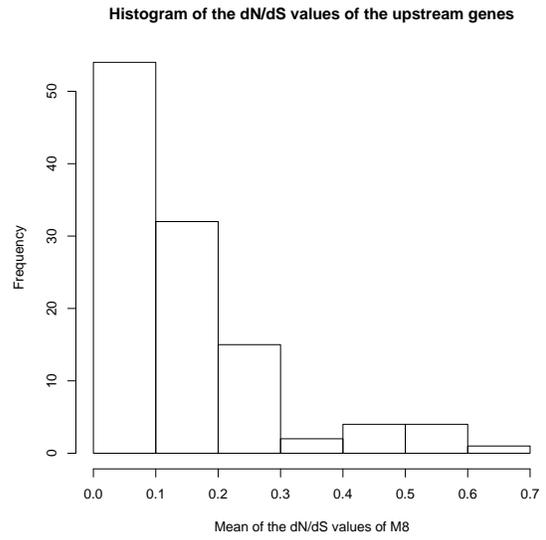
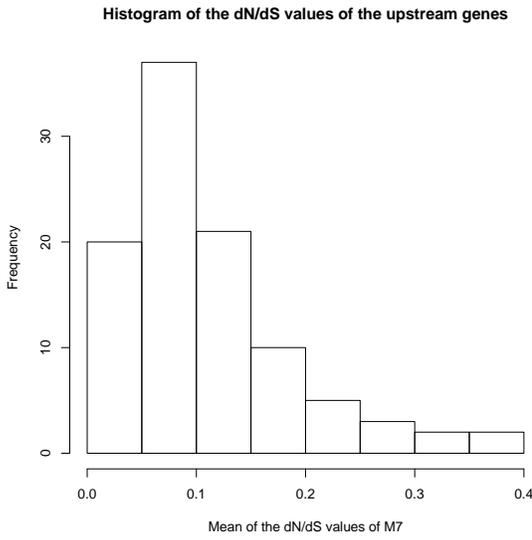


Histogram of the dN/dS values of the upstream genes

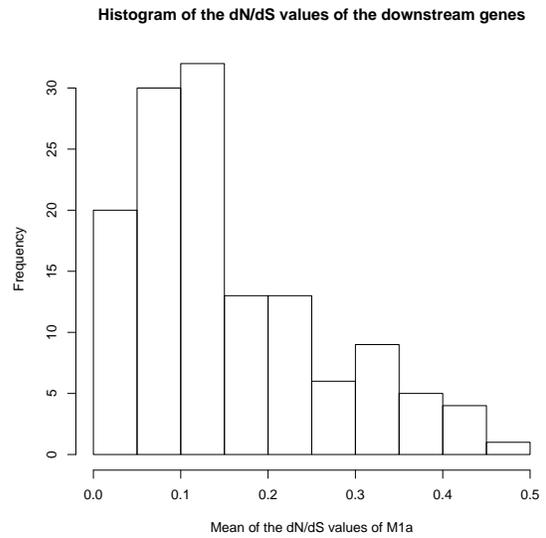
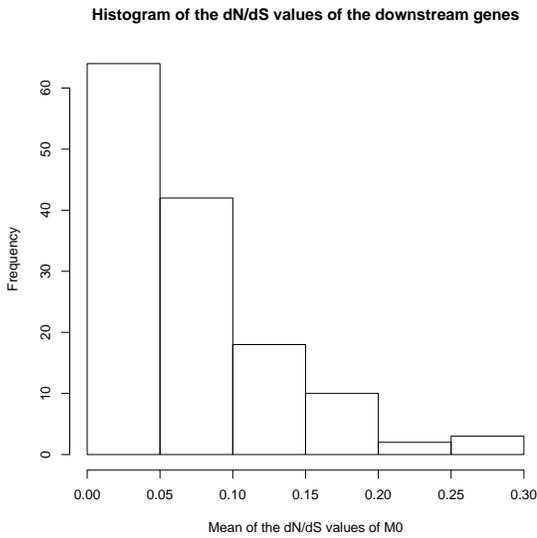


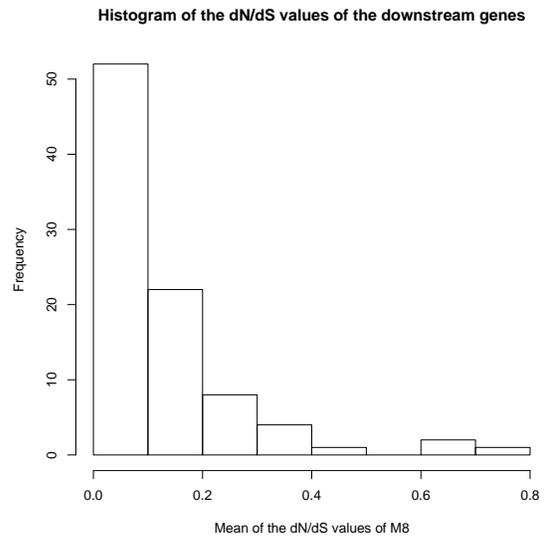
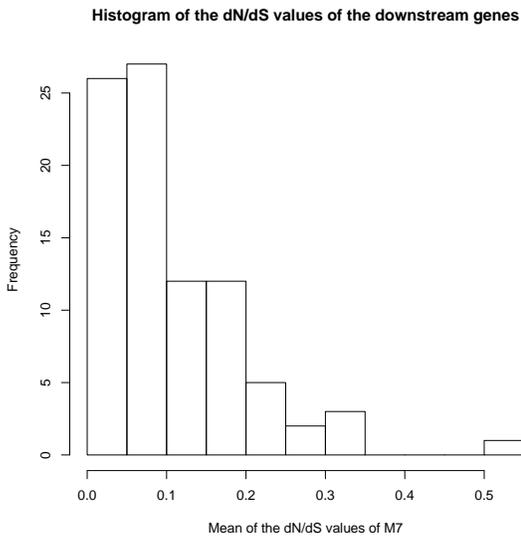
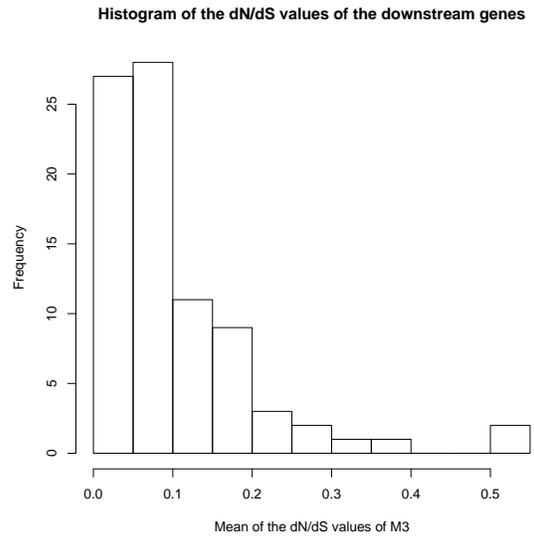
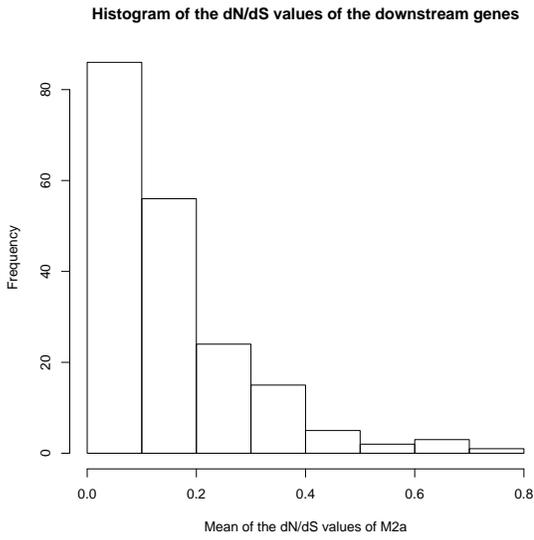
Histogram of the dN/dS values of the upstream genes





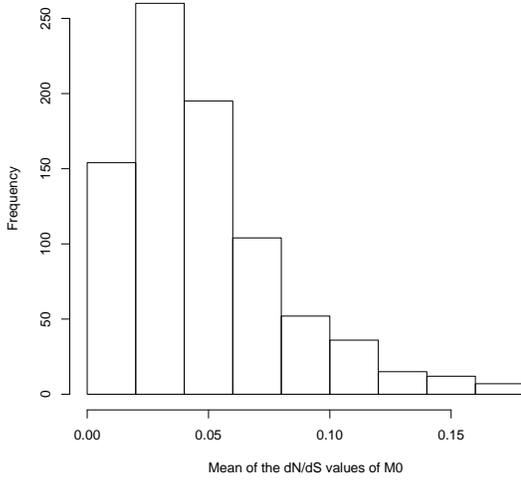
The above figures show the distribution of weighted averages of the dN/dS ratios. These are the figures for upstream genes.



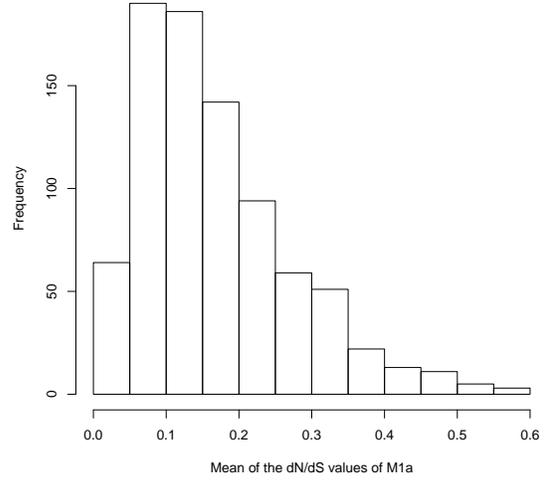


The above figures show the distribution of weighted averages of the dN/dS ratios. These are the figures for downstream genes.

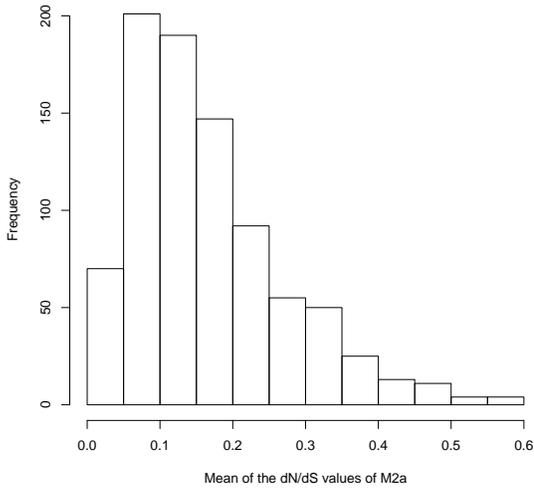
Histogram of the dN/dS values of the non-neighbouring genes



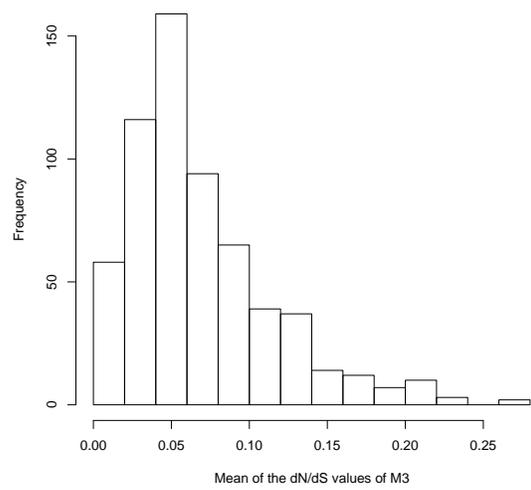
Histogram of the dN/dS values of the non-neighbouring genes

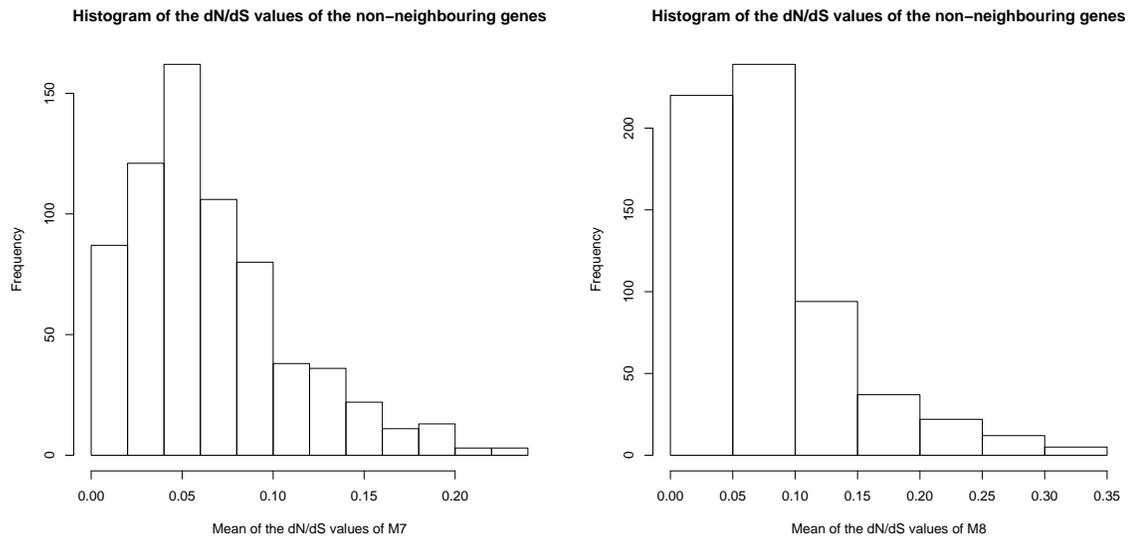


Histogram of the dN/dS values of the non-neighbouring genes



Histogram of the dN/dS values of the non-neighbouring genes





The above figures show the distribution of weighted averages of the dN/dS ratios. These are the figures for non-neighbouring genes.