Exploring the boundaries of shallow phylogeny in the YESS group and the dynamics of gene cluster and operon formation in bacterial genomes

Fergal Martin, B. Sc.

Thesis submitted to the National University of Ireland Maynooth in fulfilment of the requirements for the Degree of Doctor of Philosophy

NUI MAYNOOTH
Ollscoil na hÉireann Má Nuad

Department of Biology
National University of Ireland
Maynooth
Co. Kildare

Supervisor: Dr. James McInerney
Head of Department: Prof. Kay Ohlendieck October 2009

Table of contents

I – Acknowledgements

To James, for putting up with me over the course of both my undergrad and my Ph.D. I learned a lot under your supervision. The advice and guidance was always there when I needed it.

To IRCSET, for providing me with funding so I could carry out my Ph.D.

To my lab mates, for making my time in the Bioinformatics lab so enjoyable. There was never a dull moment. Well actually there were lots of dull moments, but those aren’t the ones I remember.

To my family for supporting me from day one. From when I was a four year old wanting to become a digger driver, right up till when I first started talking about doing a Ph.D, you’ve always given me nothing but encouragement.

To Niki. Over the past few months you’ve cracked the metaphorical whip and, whether you realised it or not, helped me get through this. Thank you. Your constant optimism should be bottled and sold for real money.

To Lorraine, the two Noels and everyone else in the fencing club for all the support during my writing.

II – Declaration

This thesis has not been submitted in whole, or in part, to this, or any other University for any other degree and is, except where otherwise stated, the original work of the author.

Signed: ________________

Date: ________________

III – Index of figures

IV – Index of tables

V - Glossary

BLAST – Basic local alignment search tool

GFP – Green fluorescent protein

IG – Intervening gene.

HGT – Horizontal gene transfer.

LGT – Lateral gene transfer.

ME – Minimum evolution

ML – Maximum likelihood.

MLSA – Multilocus sequence analysis.

MP – Maximum parsimony.

MSA – Multiple sequence alignment.

MRCA – Most recent common ancestor.

NJ – Neighbor-joining.

PIM – Protein immobility model

SOM – Selfish operon model

SP – superpathway.

Amino acid abbreviations:

ala – alanine

arg – arginine

asn – asparagine

asp – aspartic acid

cys – cysteine

glu – glutamic acid

gln – glutamine

gly – glycine

his – histidine

ile – isoleucine

leu – leucine

lys – lysine

met – methionine

phe – phenylalanine

pro – proline

ser – serine

thr – threonine

try – tryptophan

tyr – tyrosine

val – valine

VI - Abstract

In this thesis I look at two different problems in bacterial genomic analysis. The first involves reconstructing the evolutionary history between a group of closely related bacteria. I addressed whether or not it is possible to separate such genomes into different genera, species and strains. Specifically, I addressed how different approaches such as the use of 16S rRNA phylogenetic trees, phylogenetic supertrees and concatenation of individual genes in order to construct phylogenetic trees compare with one another. What effect will problems associated with resolving shallow-phylogeny have on recovering a tree of life? Ultimately I show that for the group of genomes involved, different methods and data produce different results and that the true tree, if a tree-like structure does indeed exist for these genomes, is unrecoverable using such approaches.

In the second part of my thesis I examine the phenomenon of gene clustering in bacterial genomes. I present a software program, GenClust, for the identification, analysis and visualisation of gene clusters. I show how GenClust can be used to recover and analyse clusters of genes involved in amino acid biosynthesis across a large γ-proteobacterial dataset. Finally, I examine models of gene cluster and operon formation and test them with real data, using a combined approach of comparing clusters on both structural similarity and the underlying phylogenetic signals of the clustered genes. I provide a hypothesis for the selective forces driving cluster and operon formation in bacterial genomes.

Chapter 1 - Introduction

1.1 Phylogenetic methods:

1.1.1 Tree thinking:

The use of a tree like structure to describe the evolutionary relationships between organisms was first illustrated by Darwin in his 1959 book "The Origin of Species" (Darwin, 1859). The idea was further popularised by German biologist Ernst Haeckel (figure 1.1) (Haeckel, 1879). Haeckel depicted a more literal tree than the mathematical structures used today. Haeckel imposed a hierarchy based on what he believed the natural progression from simple to complex, with man resting at the top of the tree. Nevertheless, the phylogenetic trees we draw today are remarkably similar to the original idea pioneered by both Darwin and Haeckel.

In this section, I am going to discuss the features of phylogenetic trees along with some of the more commonly used methods for generating phylogenetic hypotheses. In particular I will examine some of the strengths and weaknesses of each method. Finally I will talk about methods of measuring both signal and conflict in phylogenetic hypotheses generated using these inference methods.

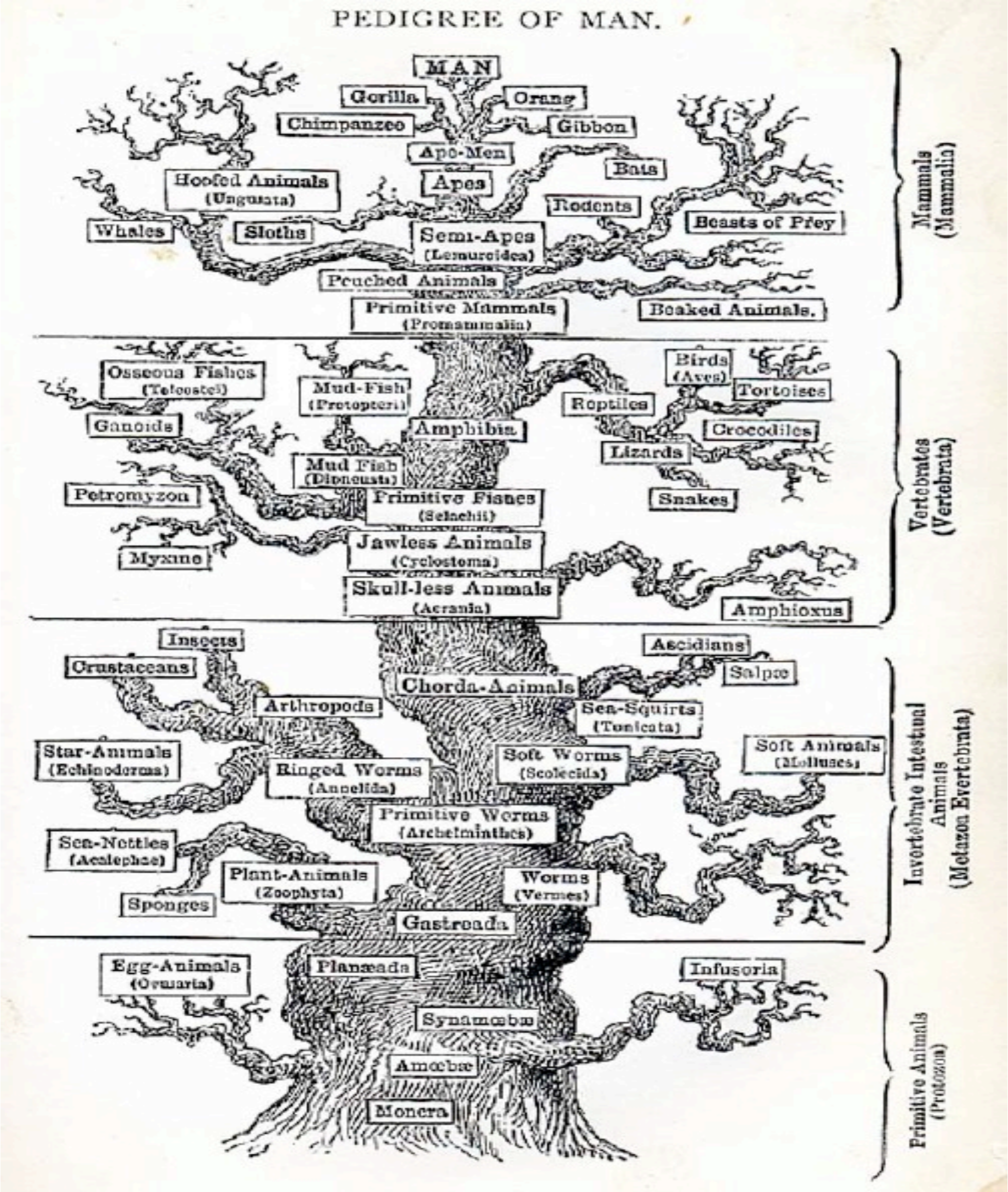


Figure 1.1: Haeckel's Tree of Life from the book "The Evolution of Man" (Haeckel, 1879).

1.1.2 From sequence data to phylogenetic trees:

While there are many types of data that can be used to determine the evolutionary relationships between a group of organisms, the most commonly used is molecular sequence data. The process of building a phylogeny describing the relationships between those sequences involves a number of steps. The sequences must be aligned using multiple sequence alignment (MSA) software such as Clustal, Muscle or Prank (Thompson et al., 2002; Edgar, 2004; Löytynoja and Goldman, 2008). Alignments are generated by inserting gap characters, generally denoted by a '-', into the sequences in order to bring positions that are considered to be conserved into alignment with one another. Once aligned, a 2d pairwise distance matrix can be generated to provide a measure of the distance of all the sequences in the alignment from one another. Due to differences in the algorithms of MSA software, the alignment generated is dependent to an extent on the software used.

After the alignment has been completed it can be input into phylogenetic inference software, for example Phyml or PAUP* (Guindon and Gascuel, 2003; Swofford, 2003). Different software use different algorithms for the inference of the evolutionary history of molecular sequence data. These range from simple algorithms such as neighbor-joining (Saitou and Nei, 1987) to more complex ones like maximum likelihood (Guindon and Gascuel, 2003). The important thing to consider about the resulting phylogeny is that, much like MSA, the result may be dependent on the method used. Different algorithms can produce different phylogenies for the same alignment. Some algorithms produce different phylogenies for a single alignment depending on the set of input parameters used (Keane, 2006). This is an important point to consider when generating a phylogeny. In addition, MSA software will produce a result regardless of the quality of the data that is used as input. If the sequences input to the software show little to no conservation then the resulting alignment will be poor and any conclusions drawn from it will be unreliable.

1.1.3 Structure of phylogenetic trees:

Modern day phylogenetic trees are mathematical structures that propose a model for the evolutionary relationships between a set of units, such as organisms or genes in a gene family (Page and Holmes, 1998). For the purpose of this introduction I will describe trees in terms of species trees. Species trees are a subclass of phylogenetic trees that describe the evolutionary history of a group of species. However, the description holds true for phylogenetic trees in general.

As a mathematical structure, formal definitions exist for each component of a phylogenetic tree (figure 1.2). Trees consist of branches, nodes and a topology. A branch defines a relationship between two nodes. Nodes can be subdivided into three classes: root nodes, internal nodes and external nodes. Root nodes represent the presumed most recent common ancestor (MRCA) for all the species represented on the tree. Unrooted trees do not have a root node and show only the relationships of the species relative to one another. Rooted trees have root nodes. Rooted nodes give trees direction. This

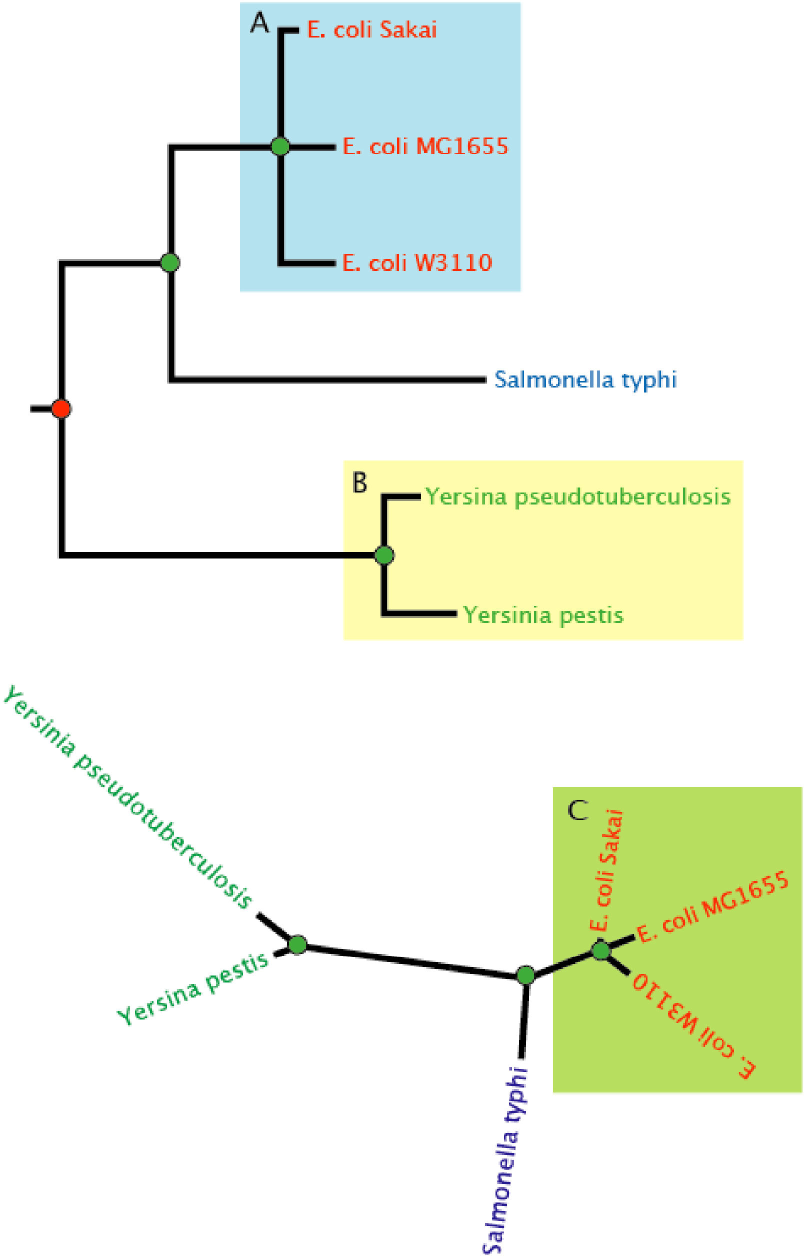


Figure 1.2: Features of phylogenetic trees. The top tree is rooted, as denoted by the root node, coloured red. Internal nodes are coloured green. Terminal nodes are represented by species name. The bottom tree is the same but unrooted. The three *E. coli* strains are an example of a monophyletic clade (a group of taxa to the exclusion of all others). Branch lengths vary, and show the relative rate of evolution of each node. Other features include a trifurcating, unresolved node (A), a bifurcating node (B) and a clan (C, the equivalent to a monophyletic clade for an unrooted tree).

direction is evolutionary time, since the root is defined as the node from which all other nodes descend. Often a root node is defined via the use of an outgroup. For a species tree, an outgroup can be defined as a species, or set of species, believed to be less closely related to the ingroup species than the ingroup species are to one another. An example would be using rodent sequences as an outgroup on a primate tree. Then the root node is defined as the MRCA for primates and rodents. If the tree is unrooted then this directional information is not present, however the relationships between the species are still represented in the topology. Therefore, there is a maximum of one root node per tree, located at the base. The root is the parent node for the entire tree. Internal nodes correspond to the set of nodes that are both parent and child nodes. Terminal nodes, more often called leaf nodes (but also tips, terminal taxa or operational taxonomic units) represent the extant data on a tree (Page and Holmes, 1998). Mathematically speaking, these nodes are the set of nodes that are child nodes but not parent nodes. The branching pattern of the tree is known as the topology.

Other common features of phylogenetic trees include: branch lengths (where the length of a branch corresponds to the rate at which it is evolving), clades (sub-groupings within the tree structure), resolution (whether the relationships at a node can be inferred or not) and balance (the level of bifurcation in the branching pattern). In the following sections I will describe the algorithms and models used in constructing a phylogeny.

1.1.4 Distance matrix methods:

Distance matrix methods were first introduced in 1967 (Cavalli-Sforza and Edwards, 1967; Fitch and Margoliash, 1967). Distance matrix methods convert an alignment into a matrix of pairwise distances, using some model for measuring the distance between two sequences. The resulting matrix is used to produce the branch ordering and branch lengths.

The most basic method for calculating the distance between characters in an alignment is to simply count the number of observed differences across each site. This is known as the p-distance. No account is taken for the possibility of multiple changes at a site. To improve on this idea, several models were developed to calculate distances between DNA and amino acid sequences.

The simplest model is the Jukes and Cantor (JC) model (Jukes and Cantor, 1969). The JC model assumes all four bases have equal frequencies and that all possible substitutions are equally likely. The distance between two DNA sequences is then calculated using the following formula:

$$d = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right)$$

where d is the distance, ln is the natural log and p is the number of nucleotide positions that differ between the two sequences.

An example of a more complex model is the general-time reversible (GTR) model (Lanave et al., 1984; Rodriguez et al., 1990). The GTR model has a total of 10 parameters: six substitution rate parameters and four base frequency parameters. The model is symmetric and thus time reversible:

$$Q = \begin{pmatrix} * & R_{AC}\pi_c & R_{AG}\pi_G & R_{AT}\pi_T \\ R_{AC}\pi_A & * & R_{CG}\pi_G & R_{CT}\pi_T \\ R_{AG}\pi_A & R_{CG}\pi_C & * & R_{GT}\pi_T \\ R_{AT}\pi_A & R_{CT}\pi_C & R_{GT}\pi_G & * \end{pmatrix}$$

where R_{ij} is the rate at which base i changes to base j and π_i is the frequency of base i. Since the GTR model allows for variable base frequencies and reversible models can come close to fitting real data, the model is more robust than the JC model (Rodriguez et al., 1990; Yang et al., 1994).

Many different models of varying complexity exist, such as Kimura's 2-parameter model (K2P) (Kimura, 1980), Hasegawa, Kishino and Yano (HKY85) (Hasegawa et al., 1985) and the LogDet model (Lockhart et al., 1994). Similarly many models exist for amino acid data, such as Dayhoff (Dayhoff et al., 1978) and BLOSUM (Henikoff and Henikoff 1992).

One of the main problems with models of sequence evolution in relation to phylogenetics is model mis-specification. If the wrong model is choosen when constructing a phylogenetic tree it can result in an incorrect phylogeny (Keane et al., 2006). A major problem occurs when no pre-existing model fits the data under analysis. The best-fit model is not nessecarily one that accurately describes the data (Keane et al., 2006), though tests to examine model mis-specification do exist (Goldman, 1993).

1.1.5 Neighbor-joining:

The neighbor-joining (NJ) algorithm for phylogenetic inference was first purposed by Saitou and Nei in 1987. The algorithm is compatible with any type of evolutionary distance data (Saitou and Nei, 1987).

The NJ algorithm works off the concepts of 'neighbors', with a pair of neighbors defined as a pair of taxa (external nodes) connected by a single internal node. The initial topology is star-like (i.e. no resolution of the relationships between the taxa). Taxa are clustered such that of all possible pairs, the pair with the smallest sum of branch lengths is chosen. The chosen pair is treated as a single unit. This process is repeated until all interior branches are found (Saitou and Nei, 1987).

The advantage of the NJ method is that it is computationally inexpensive and can often obtain the correct tree topology (Saitou and Nei, 1987). The method lacks the sophistication of more complex algorithms, as it does not consider anything other than distance when reconstructing a topology.

1.1.6 Minimum Evolution:

The principle of minimum evolution (ME) was first proposed by Cavalli-Sforza and Edwards in 1967, however the method was later refined to decrease computation time by Saitou and Imanishi (Cavalli-Sforza and Edwards, 1967; Saitou and Imanishi, 1987). For a given tree topology the length of each branch is computed. The branch lengths are then summed and the tree showing the smallest sum of branch lengths is considered the minimum evolution tree (Saitou and Imanishi, 1987). While ME bears a resemblance to maximum parsimony (section 1.1.7), it is actually much more similar to neighbor-joining, as both require distance matrices and NJ includes the principle of minimum evolution in its algorithm (Saitou and Imanishi, 1987; Saitou and Nei, 1987). ME and NJ were both found to produce similar results on test data sets, however ME has the advantage of searching more of tree space than NJ search and is thus more likely to find the best tree, though this makes it slower than NJ (Saitou and Imanishi, 1987).

1.1.7 Maximum parsimony:

Maximum parsimony is a character-based method of phylogenetic inference. Willi Hennig is attributed with the development of parsimony (Hennig, 1966). Hennig also proposed important concepts such as synapomorphic and symplesiomorphic characters. Synapomorphic characters are those that are shared by two or more groups, inherited from their last common ancestor (i.e. they are specific to that clade). Symplesiomorphic characters are characters shared by a number or groups, but that originated before the last common ancestor of those groups. Hennig believed that trees should only be constructed from synapomorphic characters (Hennig, 1966). Publications by Edwards and Cavalli-Sforza, who first used the technique to analyse gene frequency data, and Camin and Sokal, who used it for morphological characters, further popularised parsimony as an inference method (Edwards and Cavalli-Sforza, 1964; Camin and Sokal, 1965).

Maximum parsimony draws upon the principle of Occam's razor. According to Occam's razor, the explanation requiring the fewest assumptions is generally the correct one. The principle of parsimony is to reconstruct the evolution of a particular site using the fewest possible steps. Characters in an alignment are analysed on a site-by-site basis. Each candidate tree topology is scored based on the minimum possible number of changes in character states per site. The sum of these scores across all sites dictates how parsimonious a particular topology is. The tree requiring the least amount of changes in character states across all sites is considered the most parsimonious tree (Page and Holmes, 1998; Yang, 1996).

There are some obvious flaws inherent to maximum parsimony. Parsimony favours the minimum number of changes per site. However by trying to maximise similarity due to common ancestry, characters that do not fit a given topology are assumed to be homoplastic. Parsimony, by definition, does not take into account the possibility of varying rates of substitution. Because of this, parsimony is vulnerable to long-branch attraction (Felsenstein, 1978). Long branch attraction can occur between rapidly evolving branches on a tree (long branches). Such long branches can be placed together, sometimes incorrectly, on a tree simply because they are rapidly evolving. A related disadvantage of parsimony is that it does not allow a model of sequence evolution to be taken into account. Therefore if the sequences are evolving under some known process parismony cannot use this information to produce more accurate results.

1.1.8 Maximum likelihood:

Maximum likelihood is a robust method of phylogenetic inference (Whelan et al., 2001). Likelihood is defined as the probability of observing the data given a particular model (Page and Holmes, 1998). The data are fixed, the model is subject to change. In terms of molecular evolution, data refers to the alignment, while "the model" often refers to a particular tree topology combined with a model of sequence evolution.

Maximum likelihood works by calculating the lengths for the branches on a tree. This is achieved using a series of matrix multiplications, based upon the information contained in the model. Optimisation occurs by calculating the set of branch lengths, for a given topology, that maximises the likelihood of observing the data. For all branches one or possibly both of the nodes connected by the branch are unknown ancestral sequences. As such it is necessary to calculate every possible combination of ancestral states for the given topology. For a four-taxon tree there are 16 possible combinations of ancestral states. This optimisation is performed on all possible tree topologies. ML is a computationally expensive approach, particularly for large datasets (Whelan et al., 2001; Steel, 2005).

The maximum likelihood tree, therefore, is the tree that, in combination with the model, has the highest likelihood of explaining the observed data. It is important to note that the tree selected is simply the most likely tree. This is not necessarily equivalent to the correct tree. In particular, the choice of the model is extremely important in recovering the correct phylogeny. The model can consist of many parameters, such as transition/transversion ratio, base composition biases, correction for differing substitution rates, among site rate variation and the proportion of invariant sites (Page and Holmes, 1998). ML offers the advantage that the best values for each of the parameters can be estimated based on the data. It is possible to compare nested models (where one model is a special case of another model) to test whether one model is significantly better than the other (via a chi-squared test). An example of this is keeping a parameter fixed in one model and letting it vary in the other (Page and Holmes, 1998).

Models can be made parameter rich to more accurately model the evolution of the data. Increasing the complexity in the model increases the complexity of the increases the computation time of the analysis. Also, the more parameters that are present, the higher the chance of over-fitting the model to the data is. In order to speed up the likelihood calculations software such as ModelGenerator (Keane et al., 2006) and MODELTEST (Posada and Crandall, 1998) can be used to choose the model which best fits the data. This removes the need to test all topologies with all possible models of sequence evolution, under the assumption that the software chooses the model that will lead to the maximum likelihood. However, computation time still remains a problem as more taxa are added (Page and Holmes, 1998).

Maximum likelihood is a popular method because it produces consistent estimates of phylogeny, and if it is given a good model and enough data, maximum likelihood will find the correct tree (Whelan et al., 2001).

1.1.9 Bayesian inference of phylogeny:

Bayesian inference of phylogeny is a parameter-based method of calculating the probability of a data set (Huelsenbeck et al., 2001). Bayesian analysis differs from ML by incorporating a prior probability distribution into the calculation, i.e. it incorporates prior beliefs on the values of the parameters of the model that may be independent of the data. If all parameter values have the same prior probability then the prior probability distribution is flat. If the prior probability for a parameter is not flat then this implies the value of that parameter may have a significant impact on the analysis. The goal is to obtain a posterior probability distribution over all possible parameters. The posterior probability distribution is a combination of the prior probability distribution and the likelihood for each parameter value. The posterior probability distribution can be calculated using Bayes' theorem. Bayes' theorem states that, given a hypothesis H (in this case, a tree) and some data D, the posterior probability of the hypothesis given the data is:

$$\Pr ob(H \mid D) = \frac{\Pr ob(H) x \Pr ob(D \mid H)}{\Pr ob(D)}$$

If the prior distribution is flat then the posterior probability distribution will effectively mimic ML where the parameter values giving the maximum likelihood will also give the maximum posterior probability.

The posterior probability distribution can be computationally expensive to calculate, as it involves calculations of all possible branch length combinations and calculation of substitution model parameters (Huelsenbeck et al., 2001). In order to cut down the computational overhead, a Markov chain Monte Carlo (MCMC) approach can be used to calculate the posterior probability distribution. MCMC uses samples from a simulated distribution that is believed to be the posterior probability distribution instead of deriving the posterior distribution via integration (Shoemaker et al., 1999).

Some caution is needed when carrying out a Bayesian analysis (Huelsenbeck et al., 2002). It has been noted that support for nodes in trees derived through Bayesian analysis tend to have higher values than corresponding nodes in trees derived from the same data using ML and the precise cause of this trend is unclear (Huelsenbeck et al., 2002). As the prior probability distribution is a key part of Bayesian analysis, it is not surprising that the use of different priors has a large effect on the posterior probability distribution (Shoemaker et al., 1999).

1.1.10 The advent of whole genome DNA sequencing:

In 1995 the complete genome of *Haemophilus influenzae* was sequenced (Fleischmann et al., 1995). This was a major milestone in the field of molecular biology. The genomes of many other organisms followed soon after, with preference towards model organisms such as the nematode *Caenorhabiditis elegans* (The C. elegans Sequencing Consortium, 1998) and the fruit fly *Drosophila melanogaster* (Adams et al., 2000). The second major milestone in the sequencing of complete genomes came with the sequencing of the human genome (Lander et al., 2001; Venter et al., 2001).

As a consequence of the availability of whole genome sequences, the field of comparative genomics was born. To date there are 1,115 complete published genomes and 4,626 ongoing genome projects, spread across all three domains of life. With sequencing becoming faster and more affordable, these numbers are only a hint of what's to come in the next decade.

The emergence of comparative genomics changed the landscape of phylogenetics in general. Instead of being restricted to building phylogenies from one or a few genes, researchers were given the opportunity to use all, or at least a large fraction, of the genes in an organism when carrying out phylogenetic analyses. This brought many advantages and potential pitfalls.

Different methods exist for combining the information contained in different gene families into a single phylogeny. Popular methods include data concatenation, presence absence methods and supertree approaches.

1.1.11 Data concatenation and supermatrix approaches:

The principle of data concatenation is relatively simple (see figure 1.3). A set of genes, generally widely or universally distributed among the organisms under analysis, is selected. Genes are aligned individually using multiple sequence alignment software and then the alignments are concatenated together, creating a supermatrix. The ordering of the genes within the each supermatrix is conserved. A phylogenetic tree is then constructed based off the concatenated alignment. The resulting tree should display the combined signal of all the genes in the alignment and thus, at least in principle, should be more reliable than a tree constructed from an alignment of a single gene family.

Data concatenation and supermatrix approaches have gained widespread popularity (Baldauf et al., 1999; Bapteste et al., 2002; Ciccarelli et al., 2006). However, there are a number of things to consider when using concatenated data. Firstly, the topology of trees based on the individual genes in a concatenated alignment may not match the topology of the tree built from the concatenated alignment itself (Bapteste et al., 2008). Software has been recently been developed to test incongruence in concatenated genes, in order to assist the selection appropriate sets of genes for concatenation (Leigh et al., 2008). Secondly, concatenated data are biased towards producing strongly supported trees and

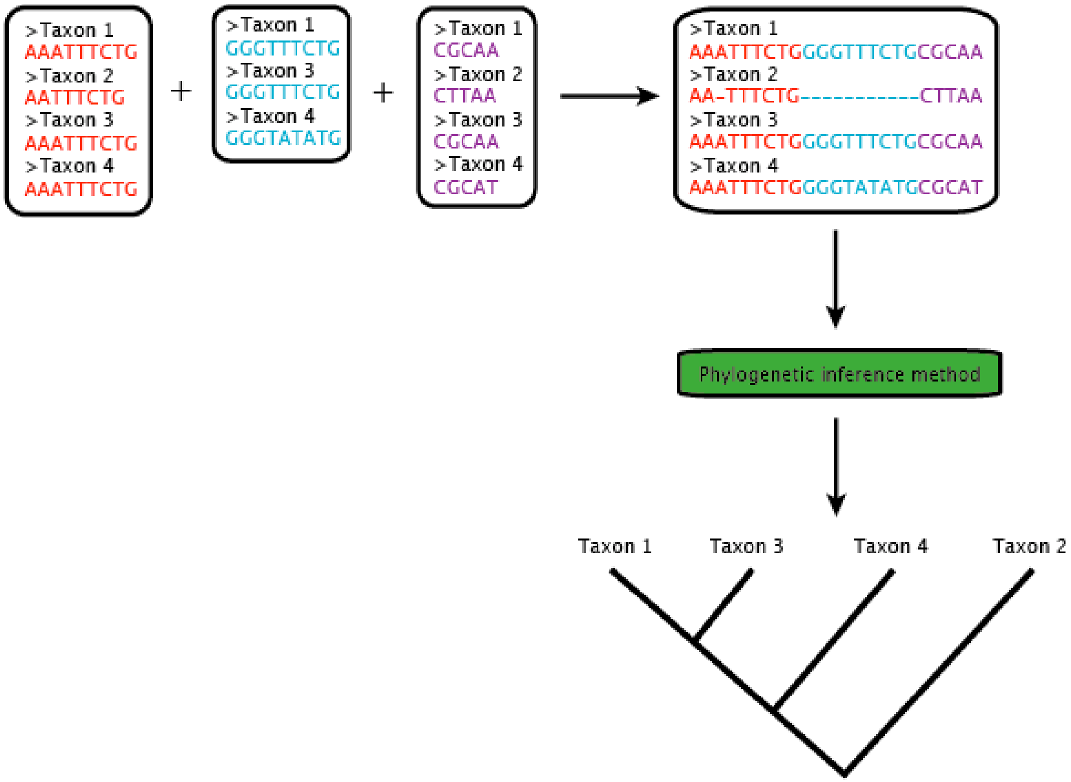


Figure 1.3: Data concatenation. The data contained in the three gene families are combined into a single, concatenated alignment.

the trees produced are dependant on the model of sequence evolution selected (Phillips et al., 2004; Keane et al., 2006).

1.1.12 Supertree construction:

In this thesis, I describe supertree analyses of genomic data, therefore, the following section contains some background on supertree methods, their strengths and weaknesses. A supertree is a tree that represents the phylogenetic relationships of a group of input trees (Wilkinson et al., 2004). Like data concatenation, supertree construction (see figure 1.4) is based on the principle of using the information contained in multiple data points to generate a phylogeny. The method differs from data concatenation in that a single alignment and phylogenetic tree are constructed for each gene family, and the information contained in the resulting trees is combined into a supertree. Overlap between taxon sets in the input trees allows relationships to be resolved in the final supertree. Sometimes bootstrapped data is used, with multiple alignments and trees per gene family, but the principle remains the same. A major advantage of supertrees is that there is no requirement that the set of taxa in each input tree are identical. This is an important feature, as the vast majority of gene families are not universally distributed.

The actual algorithm for constructing the final supertree from the input trees can vary. Many such algorithms exist and they can be broadly separated into strict and liberal supertrees methods (Wilkinson et al., 2004). Strict supertrees methods are those that resolve common or uncontested groupings among a set of input trees. Methods include

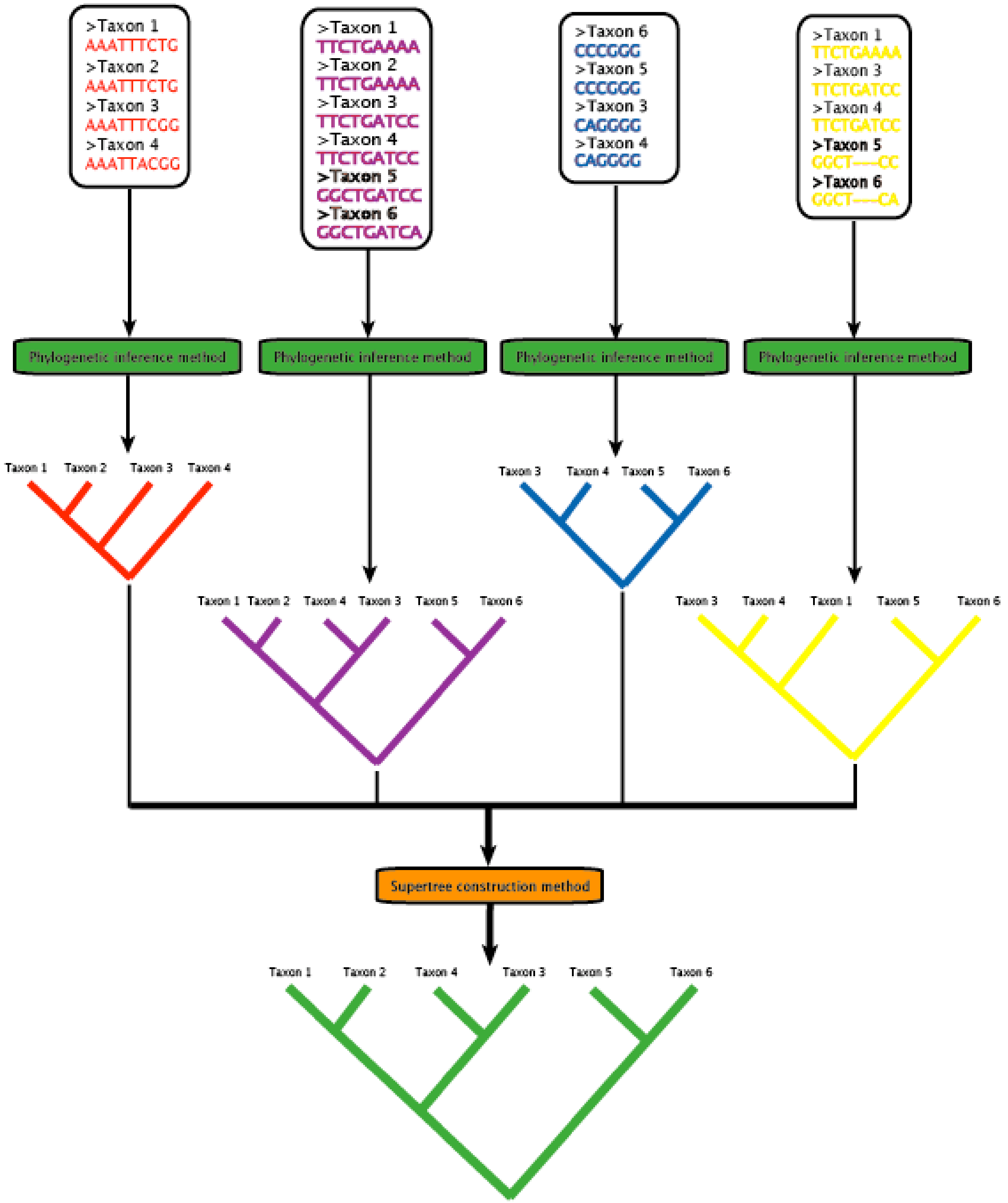


Figure 1.4: Supertree construction. Each alignment is used to construct a corresponding phylogenetic tree. The topological information contained in these trees is overlapped and used to construct a supertree.

strict, semi-strict and strict consensus merger. Strict supertree methods are decreasing in popularity in prokaryotic biology because the underlying processes of gene loss, gain and HGT mean the majority of gene phylogenies show some degree of conflict and this conflict is left unresolved in agreement supertrees.

Liberal supertree methods are those that have the maximum fit to the input trees under some objective function (Wilkinson et al., 2004). The objective function differs from method to method. Two popular methods are matrix representation using parsimony (MRP) (Baum, 1992; Ragan, 1992) and most similar supertree (MSSA) (Creevey et al., 2005). The purpose of these methods is to compare candidate supertrees to the input trees. The candidate supertree is pruned so that the leaf set matches that of the current input tree. The tree-to-tree distance is then measured between the pruned supertree and the current input tree. This process is repeated for all input trees. Optimisation selects for the supertree that agrees best with the relationships displayed in the input trees dictates the topology of the final supertree (or supertrees if multiple supertree topologies provided the same level of optimisation to the input trees). More recently ML supertrees have been described (Steel and Rodrigo, 2008). ML supertrees are a liberal supertree method using a ML approach. Steel and Rodrigo have demonstrated that taking an ML approach can produce stastically consistant results, unlike MRP, which can sometimes produce statistically inconsistant results (Steel and Rodrigo, 2008).

Liberal methods attempt to resolve the relationships in the supertree, even in the presence of conflict. It is important to note that a fully resolved tree can therefore be produced in the absence of any strong signal, though support values can be assigned to nodes to assess the strength of the underlying signal. Additionally, no optimisation algorithm is without weakness (Creevey et al., 2005).

Supertree methods have gained popularity in recent times and have been applied to a wide variety of phylogenies such as seabirds (Kennedy and Page, 2002), dinosaurs (Lloyd et al., 2008) and the origins of eukaryotes (Pisani et al. 2007).

1.1.13 Gene content methods:

Gene content methods compare the genetic repertoire of a set of genomes. This requires the identification of orthologous sets of genes. Orthology is a somewhat subjective matter, as it is only based on extant genes, and therefore inferred orthology is not always correct. However, a number of different schemes have been used to define orthologous genes, such as intergenomic best hits (Snel et al., 1999; Korbel et al., 2002) or sequence similarity to COG groups (Lin and Gerstein, 2000; Tatusov et al., 2001).

Once orthology has been assigned between genomes a presence/absence matrix can be constructed. Presence/absence matrices are often encoded as binary strings, with a ‘1’ denoting the presence of a gene in a particular genome, while ‘0’ denotes absence. Pairwise distances are then calculated between genomes and a phylogeny can be constructed using simple schemes such as neighbor-joining (McCann et al., 2008).

The problem with gene content methods is that they are not suitable for reconstructing prokaryotic phylogenies in general (Wolf et al., 2002). Variation in the rate of gene loss between different genomes has been shown to produce incorrect phylogenies (Wolf et al., 2001a). This is especially true for genomes that have undergone genome reduction. Parasitic and endosymbotic genomes often have drastically reduced genomes, and gene content methods are, by nature, not designed to take this into account. Some workarounds to this problem have been developed such as exclusion of genomes that have undergone reduction or normalisation of gene content based on the size of the reduced genome when calculating pairwise distances between genomes (Snel et al., 1999; Korbel et al., 2002).

While gene content methods are unsuitable for the construction of phylogenies, they are interesting in their own right and are a useful tool for studying similarities and differences between genomes (Wolf et al., 2002).

1.1.14 Measuring support and conflict in phylogenetic analyses:

Reconstructing a phylogeny for a set of sequences is a relatively straightforward process. It is important to be able to measure the quality of the signal both present in a phylogenetic tree and in the underlying alignment. In this section I will discuss some common methods for measuring the statistical significance of phylogenetic signal.

The permutation tail probability (PTP) test (Archie, 1989) is a commonly used test to evaluate the level of phylogenetic signal present in an alignment of characters. The PTP test is a randomisation procedure used to assess whether or not an alignment contains a hierarchical phylogenetic signal. The algorithm permutes the character assignments within each character, generating a new alignment (Archie, 1989). The randomisation of character assignments removes phylogenetic information from the newly generated alignment while keeping the character state distribution the same as the original alignment. This process is repeated multiple times and the observed number of steps on the minimum length tree generated from the original alignment is compared to the mean number of steps on the minimum length trees derived from the permuted alignments. This measures if the observed signal in the original alignment is significantly better than random. Alignments failing the PTP test are generally considered to be void of phylogenetic signal and are often removed from the analysis. However it should be noted that the PTP test is considered somewhat weak and it has been shown that alignments that have no signal can get highly significant scores.

Bootstrapping is a commonly used statistical measure (Efron, 1979). In a phylogenetic framework it measures the level of support for different nodes in a phylogenetic tree (Felsenstein, 1985). The algorithm works by randomly selecting sites in an alignment (with replacement) and generating a new alignment of the selected sites. The number of randomly selected sites for the new alignment is equal to the number of sites in the original alignment. Generally 100 or 1000 new alignments are built in this manner. A tree is built from each new alignment using some inference method. The information contained in these trees is then amalgamated using a consensus method into a single tree. Nodes in this tree are assigned a support value. That value is equal to the percentage of times the groupings supported by a given node are present in the set of trees generated during the bootstrapping process. High values imply that a node is strongly supported. Because each alignment generated during the bootstrap is based on a random sampling of the signal in the original data, the stronger the signal present in the original data the less conflict there will be between trees inferred from the generated alignments. This results in higher the support values for the nodes on the final bootstrapped tree. Likewise, weak or conflicting signals in the original data can less to poorly supported nodes on the final tree. Bootstrapping is an extremely valuable tool in assessing confidence in phylogenies, though it is important to note that conflicting phylogenies can sometimes attain high levels of support through bootstrapping (Phillips et al., 2004).

Paired-sites test are another method of measuring confidence in phylogenetic trees. These tests include the Kishino-Hasegawa (KH) test (Kishino and Hasegawa, 1989), the Shimodaira-Hasegawa (SH) test (Shimodaira and Hasegawa, 1999) and the approximately unbiased (AU) test (Shimodaira, 2002). The principle behind these tests is to decide whether or not one phylogenetic hypothesis (tree topology) is significantly better than other possible hypotheses at explaining the data. The KH test is a method of estimating the standard error and confidence intervals for the difference in log-likelihoods between two different phylogenetic trees representing the same data. Initially the test was developed to compare trees that were specified *a priori*, *i.e.* trees that were derived independently of the data. However the test was adapted to compare ML trees, for example the comparing the tree with the highest likelihood to the trees with the second or third highest likelihood (Goldman et al., 2000). It has been noted that a bias exists in the KH test that can lead to overconfidence being placed in an incorrect topology (Shimodaira and Hasegawa, 1999; Goldman et al., 2000). The SH test is similar in concept to the KH test but attempts to overcome the bias in the KH by using multiple comparisons (Shimodaira and Hasegawa, 1999). The SH test suffers from another kind of bias due to the fact that the number to trees included in the confidence set becomes large as the number of tree comparisons increases (Strimmer and Rambaut, 2002). Because of this the SH test is considered to be conservative (Shimodaira, 2002). To avoid these biases Shimodaira (2002) developed the AU test. Several sets of bootstrap replicates are generated with varying sequence length in each set. The AU test counts the number of times a hypothesis is supported by the bootstrap replicates in each set to obtain bootstrap probability values for different sequence lengths. It then calculates the approximately unbiased *P*-value based on the change in bootstrap probability values along the changing sequence lengths (Shimodaira, 2002). Like the SH test the AU test adjusts for the selection bias present in the KH test, but it is less conservative than the SH test and in general works better (Shimodaira, 2002).

It is imporant to bear in mind that any tree constructed using the methods described above is merely a point estimate and that trees have confidence interavals of varying sizes. Often many trees will describe the data equally well, even if their topology conflicts with one another it is possible that their confidence intervals will overlap.

1.2 Horizontal gene transfer and defining a bacterial species:

Much of the work reported in this thesis focuses on horizontal gene transfer (HGT). In the following sections I will discuss methods the importance of HGT in prokaryotic biology, the processes through which HGT occurs, preferences and barriers to HGT and the impact of HGT on phylogenetics.

1.2.1 Introduction to horizontal gene transfer:

In 2005 Andersson defined horizontal gene transfer (HGT) as "Lateral, or horizontal, gene transfer is the process of exchange of genetic material between distantly related species" (Andersson, 2005). This definition is perhaps too narrow in the sense that HGT can also occur between closely related species (Fraser et al., 2009), or strains of the same species (Majewski et al., 2000), or even within a single organism, as in the case of the transfer of genes from the chloroplast to the nucleus (Martin et al., 1998). In this thesis I define HGT as the transfer of genetic material from one bacterium to another via a means other than vertical (maternal) transfer.

In 1999 Doolittle said, with reference to HGT, "Thus, there is a problem with the very conceptual basis of phylogenetic classification" (Doolittle, 1999b). Now, a decade later, HGT is recognized as a prominent force in the evolution of bacterial genomes.

HGT is an incredibly versatile process. For example, HGT has occurred between *Wolbachia*, an endosymbiotic proteobacteria, and its multicellular, eukaryotic, insect hosts, with evidence of transfers ranging from short sequences (< 500bp) to the entire *Wolbachia* genome (> 1mb) (Kondo et al., 2002; Hotopp et al., 2007). HGT is important the evolution of bacterial metabolic networks, where horizontally transferred genes are integrated onto the periphery of the network and help the recipient adapt to changes in the environment (Pal et al., 2005). Genes located in functional clusters within the genome are subject to orthologous replacement via HGT, i.e. replacement of the original gene *in situ* with a horizontally transferred copy that carries out the same function (Omelchenko et al., 2003), with the implication that conserved synteny may not imply conserved evolutionary history. It is even estimated that 18 percent of the open reading frames (ORFs) in *Escherichia coli* have been introduced via HGT since its divergence with the Salmonella lineage 100 million years ago (Lawrence and Ochman, 1998).

HGT clearly a widespread process and for this reason it must be taken into consideration in any study involving bacterial phylogenetics.

1.2.2 Methods of HGT

There are three methods by which bacteria acquire genes horizontally: conjugation, transduction and transformation (reviewed in Syvanen and Kado, 1998; Ochman et al., 2000; Jain et al., 2002). The primary difference between these processes is the method of entry of the horizontally transferred DNA to the recipient.

Conjugation (figure 1.5 A), effectively bacterial sex, involves exchange of a plasmid from donor to recipient via a tubular structure known as a pilus. The pilus docks on the recipient cell and the plasmid is transferred through the pilus. Plasmids may contain entire cassettes of genes that imbue new properties on the host. For example, *Shigella* and enteroinvasive *E. coli* (EIEC) are *E. coli* strains that have acquired a virulence plasmid (VP) (Pupo et al., 2000). This plasmid is the source of their pathogenicity and there is currently much debate as to whether the VP was introduced ancestrally or whether there have been multiple independent acquisitions of virulence in *Shigella* and EIEC (Pupo et al., 2000; Escobar-Paramo et al., 2003; Yang et al., 2007).

Conjugation is not limited to closely related bacteria. *E. coli*, a proteobacterium, has been shown to conjugate with cyanobacteria (Wolk et al., 1984). *E. coli* can even conjugate with *S. cerevisiae*, a eukaryote, in an example of trans-kingdom conjugation (Heinemann and Sprague, 1989). It is clear that conjugation facilitates the transfer of genetic material over great phylogenetic distances. However, conjugation is naturally limited to organisms that are in close physical proximity to one another.

Transduction (figure 1.5 B) is the movement of genes from one bacterium to another via a bacteriophage. The premise is simple, the donor cell is infected with a phage, the chromosome of the donor cell fragments, fragments of the chromosome become packaged as new viral particles and, following cell lysis, go on to infect and recombine within a new, potentially distantly related bacterium. The actual amount of DNA transferred in a single transduction event is limited by the capsid size of the

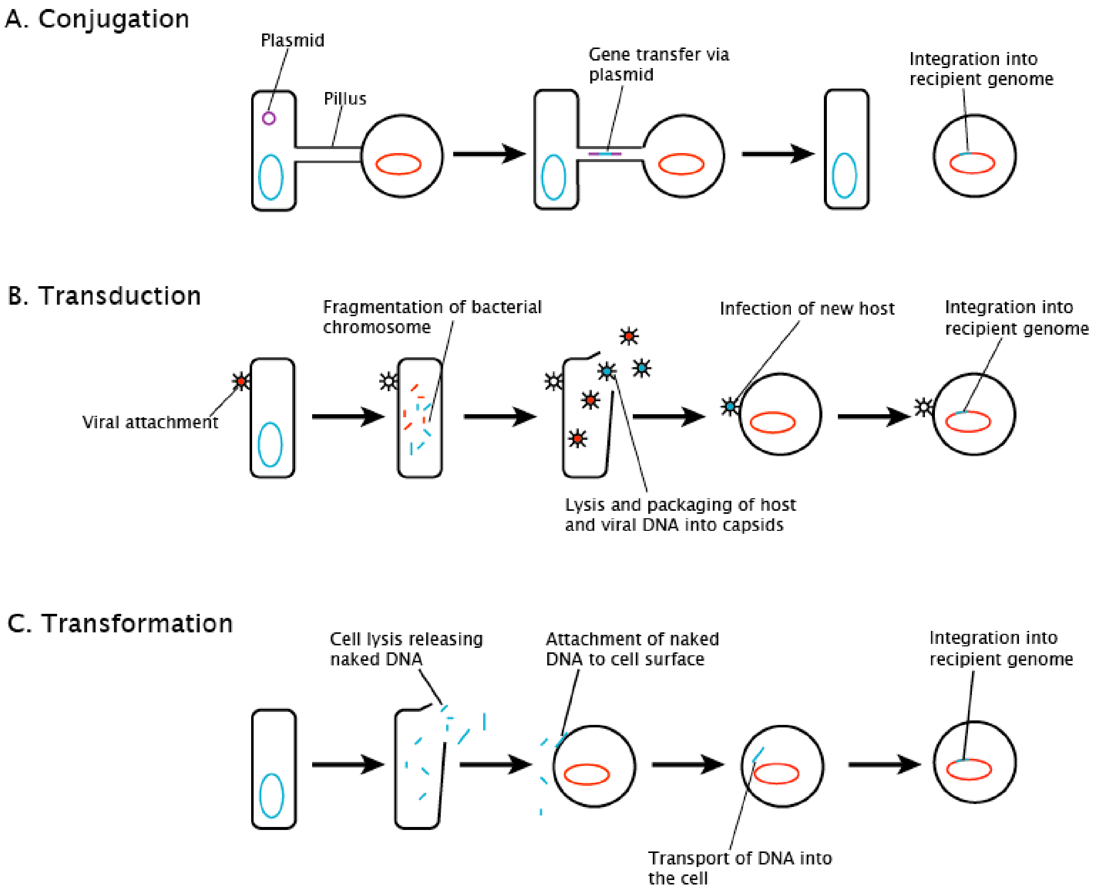

A. Conjugation
Plasmid
Pillus
Gene transfer via plasmid
Integration into recipient genome
B. Transduction
Fragmentation of bacterial chromosome
Infection of new host
Integration into recipient genome
Viral attachment
Lysis and packaging of host and viral DNA into capsids
C. Transformation
Cell lysis releasing naked DNA
Attachment of naked DNA to cell surface
Integration into recipient genome
Transport of DNA into the cell

Figure 1.5: The three types of horizontal gene transfer. Conjugation (A) involves the formation of a pillus and transfer of DNA via a plasmid. Transduction (B) is where DNA is transferred via a bacteriophage capsid. Transformation (C) is the uptake of naked DNA from the surrounding environment.

bacteriophage, but can range upwards of 100 kb (Ochman et al., 2000). It is possible to find evidence of transduction in bacterial genomes by looking at the sequence surrounding suspected horizontally transferred genes for prophage like inserts (Kunst et al., 1997). Unlike conjugation, transduction does not have strict physical or temporal constraints, in that the donor and recipient need never come into contact. Also, because the transfer occurs via the phage, phage encoded proteins mediate both the delivery and integration of the donor DNA to the recipient (Ochman et al., 2000). The limitation of this process lies in the fact that transduction can only occur with bacteria expressing receptors recognized by the carrier bacteriophage.

The third method of HGT in bacteria is transformation (figure 1.5 C). Transformation differs from the pervious two mechanisms in that it is solely controlled by the recipient. Transduction involves the uptake of naked DNA by the recipient. Some bacteria are perpetually competent at DNA uptake, while in others competence is regulated and occurs at certain physiological stages in their lifecycles (Ochman et al., 2000). Uptake involves the binding of naked DNA to the cell surface of the recipient and intake into the cell. Gram-positive and gram-negative bacteria have slightly different intake systems due to inherent differences in their membranes. The average size of the naked DNA bound to the cell surfaces of competent bacteria varies, though the upper limit appears lower than that of transduction (Dubnau, 1999). Another important factor to consider is that some bacteria require specific recognition sequences for effective transformation, while others show no preference for sequence composition but are capable of high levels of transformation (Ochman et al., 2000). Like transduction, this implies there is no requirement for physical or temporal proximity between donor and recipient.

1.2.3 Preferences and barriers to HGT:

Genes in bacterial genomes can be divided into two classes: informational and operational (Rivera et al., 1998). Informational genes are genes that are involved transcription, translation, replication and related processes. Operational genes are ones that are involved in house-keeping functions such as amino acid and nucleotide biosynthesis. In analyzing the likelihood of a gene to undergo HGT, it is important to consider which of these two classes the gene belongs to.

Informational genes are significantly less likely to undergo successful horizontal transfer than operational genes (Rivera et al., 1998). An explanation for this may lie in the complexity of the network in which a gene resides, dubbed the complexity hypothesis (Jain et al., 1999). The complexity hypothesis is based around the fact that the products of operational genes, on average, have far less interactions. For example, translation in *E. coli* involves interaction between at least 100 gene products, while many operational genes only involve a single enzyme-substrate interaction (Jain et al., 1999).

However, while the complexity hypothesis is attractive from a number of perspectives, it is important to remember that there is a difference between HGT of certain gene being unlikely as opposed to impossible. The 16S rRNA, the basis of countless phylogenies and long considered immune to HGT, can be horizontally transferred from *Proteus vulgaris* to *E. coli*, replacing the existing copy, with a growth rate reduction of 10-30 percent (Asai et al., 1999). Indeed, evidence of such transfers occurring outside the laboratory has also been documented, with the identification of a possible transfer of an rRNA operon into *Thermonospora chromogena* from *Thermobispora bispora* or a related organism (Yap et al., 1999). In a 2007 work by Sorek et al. demonstrated that for 246,045 genes, from 79 different prokaryotes, only 1,402 were impossible to transfer via transduction into *E. coli*. Informational genes accounted for a signification amount of the genes that resilient to transfers, in agreement with the complexity hypothesis. On the other hand, even for these 1,402 genes, in all cases it was possible to horizontally transfer orthologous copies of the genes from other species. While the study itself focused on the barriers to HGT, it is important to reflect on the fact that all of genes examined could be horizontally transferred. So while it is unlikely for informational genes to successfully undergo HGT, and even less likely for the transfer to be selected for, it is not impossible.

1.2.4 Horizontal gene transfer and phylogenetics:

Since the dawn of evolutionary biology one of the most fascinating goals is the recovery of the tree of life. While the true tree is unrecoverable, since our knowledge of species will never fully encompass those that did, currently and will exist, the desire to classify species into groups remains.

The first major break-through in building the tree of life came about with the advent of DNA sequencing. As previously discussed, it is believed that the 16S rRNA is unlikely to undergo HGT. This coupled with the properties of the gene being so widely distributed, with a universally conserved structure and both fast and slow evolving sites makes the 16S rRNA a seemingly ideal candidate on which to base the tree of life (Woese, 1987). As a result, bacterial species phylogenies have been created using a single gene, often the 16S rRNA or other genes considered to have properties similar to the 16S (Dauga, 2002; Purkhold et al., 2003; Paradis et al., 2005).

However, it has long been noted that individual gene trees are often incongruent with 16S rRNA phylogenies (Doolittle, 1999). A major factor in this is HGT. HGT does not conform to the path laid out by successive speciation events and thus creates problems in recovering the correct species phylogeny for any given bacterial group. Some have argued that HGT has been ascribed an 'inflated role' in evolution, and that its frequency has been overestimated, however it remains an important factor in any moderm day study of prokaryotic evolution (Kurland et al., 2003).

To overcome this inherent weakness of single gene phylogenies, different approaches have been adopted over the years. Two widely used methods are data concatenation and supertree construction. Both methods can be used to combine the information in many genes into a single phylogeny (see for example Ciccarelli et al., 2006; Pisani et al., 2007). Neither method is resistant to HGT and species phylogenies that are made via these methods are generally constructed under the assumption that the ‘true’ species signal is stronger than that of any HGT.

An ideal situation would be to identify all genes that show evidence of HGT and remove them for a dataset, when trying to establish a species phylogeny. This still has the inherent weakness that it is difficult to identify HGT with confidence when the phylogeny is not known in the first place (Suchard et al., 2003). While HGT is more readily identifiable when it occurs between distantly related species, it is much more difficult to identify when it is among closely related bacteria, such as different strains of the same species. Many attempts have been made to reliably identify HGT events (Lawrence and Ochman, 1998; Ragan, 2001a; Ragan, 2001b; Mirkin et al., 2003; Suchard et al., 2003) but ultimately the problem is a difficult one and each new method brings a set of strengths and weaknesses, with none offering a complete solution.

Regardless of whether estimates of rampant HGT in bacterial genomes are accurate or not (Lerat et al., 2005), HGT is accepted as an important force to consider when studying bacterial evolution.

1.2.5 What defines a bacterial species?

For higher organisms the concept of a species is clearly defined as it is underpinned by evolutionary and ecological processes (Gevers et al., 2005). In prokaryotes the process of defining a species is itself somewhat undefined. Classical prokaryotic species definitions arose from prokaryotic features of human interest. Pathogens, for example, were separated into species based on the diseases they caused. Other species were defined based on unique biochemical processes they possessed. These definitions were not theory based and therefore somewhat arbitrary (Gevers et al., 2005). Given the vast amounts of as yet unclassified data, there has been much controversy and interest surrounding how to define a bacterial species, especially in light of HGT (Fraser et al., 2009).

Currently prokaryotic species are defined using a consensus of genotypic and phenotypic properties (Vandamme et al., 1996; Stackebrandt et al., 2002). With sequence data available for an increasingly large volume of prokaryotes, genotypic characterisation is currently at the forefront of attempts for define prokaryotic species. Many methods exist for analysing sequences in this context.

DNA-DNA hybridisation (DDH), developed used in the 1970s, was the first method of genotypic characterisation. DDH measures the degree to which two genomes hybridise and as such provides a measure of both shared gene content and nucleotide sequence similarity (Gevers et al., 2005). Using a DDH approach genomes showing 70% DDH or greater are treated as the same species, though this level of hybridisation was calibrated based on previously recognised phenotypic-based species (Gevers et al., 2005). DDH approaches have the inherent shortfall of being unable to cope non-culturable strains, which comprise the majority of strains in the biosphere (Amann et al., 1995). In addition it is a time a difficult and time consuming process (Gevers et al., 2005).

Another approach is to look at the sequences of individual genes or sets of genes. The classical approach of using single gene phylogenies based on the 16S rRNA sequence or other genes with similar properties, as has been discussed earlier in the text, is undesirable for a number of reasons. These include potential HGT events and the fact that even small differences in sequence similarity of such genes can imply quite large differences in DDH values, with <97% sequence similarity of two 16S gene usually corresponding to <70% DDH (Fox et al., 1992). This makes it difficult to assign isolates to the same species based on high levels of 16S sequence similarity alone (Fox et al., 1992). To bypass these weaknesses, concatenated sequence data is generally used (in this framework called multilocus sequence analysis or MLSA), also discussed previously in the text, as it provides greater resolution for clustering isolates into groups.

The problem with these methods is that they are arbitrary in terms of species definition. They are based on some cut-off for sequence similarity with the assumption that such a cut-off exists and is universal to prokaryotes. This is unlikely to be the case. While clusters of prokaryotes are readily identifiable using these methods, at what depth does a cluster become a species? To advance the concept, models have been created to attempt to incorporate ecological, genomic and phenotypic data into the clusters resolved through MLSA in an attempt separate clusters into species (Gevers, 2005). These models provide theory based methods of species definition and while they are outside the context of this text to discuss in detail, they are likely a strong indication of the future direction of defining prokaryotic species (see Farser et al., 2009, for a review).

1.2.6 *Yersinia, Escherichia, Shigella* and *Salmonella*: The YESS group

The group consisting of *Yersinia*, *Escherichia*, *Salmonella* and *Shigella*, sometimes termed the YESS group (Canback et al., 2004; Comas et al., 2007), are facultatively-anaerobic, Gram-negative, rod shaped γ-proteobacteria that are catalase-positive and oxidase-negative (Brenner, 1984). The group contains many important human pathogens. This is reflected in the large number of fully sequenced YESS group genomes.

Yersinia pestis is the most noteworthy member of the *Yersinia* family as it is the causative agent of plague. *Y. pestis* infection can occur in three regions: the lymph nodes (bubonic plague), the blood (septicemic plague) and in the lungs (pneumonic plague). In the case of bubonic and septicemic plague symptoms include chills, fever, weakness, shock and internal bleeding. The formation of lumps, known as buboes, is specific to bubonic plague. Symptoms of pneumonic plague include fever, shortness of breath, chest pain, cough and bloody or watery sputum. An estimated 75 million people died in the 1300s due to the bubonic plague. Even in more recent times, outbreaks of plague caused by *Y. pestis* have occurred, such as the 1994 outbreak of plague in India (Shivaji et al., 2000). In addition to *Y. pestis*, two other types of highly pathogenic *Yersinia* exist: *Y. pseudotuberculosis* and *Y. enterocolitica* (Schubert et al., 1998). Pathogenicity is determined by a 70-kb virulence plasmid (VP) (Portnoy and Martinez, 1985). In *Y. pestis* full virulence requires two additional plasmids, a 100-kb plasmid and a 9.5-kb plasmid both encoding genes linked to pathogenicity (Pendrak and Perry, 1993; Perry et al., 1993).

Escherichia coli is perhaps the most well studied prokaryote; it is a model organism that has been critical in the advancement of the field of molecular biology. Under normal conditions *E. coli* is the dominant resident of the gastrointestinal tract of warm-blooded animals. It benefits the host by producing vitiman K_2 (Bentley et al., 1982) and out-competing pathogens bacteria for space (Hudault et al., 2001). There has been much interest in pathogenic strains of *E. coli*, with a lot of media attention surrounding outbreaks of *E. coli* infection (Kaper, 2005). There are many different pathotypes of pathogenic *E. coli*, including, but not limited to, enteropathogenic (EPEC), enterotoxigenic (ETEC), enteroinvasive (EIEC), enterohemorrhagic (EHEC) and enteroaggregative (EAEC) *E. coli* (Kaper, 2005). Many are associated with infections of the intestine, the main symptoms being fever, diarrhea and abdominal cramping (Kaper, 2005). Extra intestinal infection is also possible, uropathogenic *E. coli* (UPEC) are associated with infections of the urinary infections, while avian pathogenic *E. coli* (APEC) are associated with respiratory tract infection in poultry (APEC) (Kaper, 2005). While *E. coli* infection is not usually fatal, it is nonetheless important as demonstrated by the 2006 outbreak of *E. coli* infection from contaminated spinach in America and is it a major problem in the developing world.

Shigella is the etiological agent of bacillary dysentery or shigellosis. In the 1940s *Shigella* was defined as a genera containing four species: *Sh. boydii, Sh. dysenteriae, Sh.*

flexneri and *Sh. sonnei* (Ewing, 1949). Even then it was clear that *Shigella* closely resembled *E. coli*, with a few phenotypic characteristics such as the inability to ferment lactose and non-motility used for classification purposes. This classification system was insufficient to cover all pathogenic *E. coli*-like strains; some strains were found to have an incomplete set of phenotypic characteristics and as such were classified as pathogenic *E. coli* rather than *Shigella* (Pupo et al. 2000). Nowadays it is well known that *Shigella* are effectively *E. coli* strains that have acquired a VP (Pupo et al., 2000). *Shigella* strains cluster within the *E. coli* superfamily (Pupo et al., 2000). The retention of the genera *Shigella* is largely due to the medical importance of shigellosis of which there are an estimated 160 million cases worldwide a year, with approximately 1.1 million deaths, mainly in children under the age of five (Kotloff et al., 1999). The symptoms of shigellosis are similar to that of intestinal infection by pathogenic *E. coli* and include fever, diarrhea and abdominal cramping. From an evolutionary standpoint there is much debate over the origins of both *Shigella* and pathogenic *E. coli* and this will be discussed in more detail in the next section.

Salmonella is the causative agent of salmonellosis. Members of the genera *Salmonella* are intestinal parasites and intracellular pathogens in many different hosts including mammals, birds, reptiles, amphibians and plants (McQuiston et al., 2008). Nontyphoidal salmonellae are responsible for approximately 1.4 million cases of salmonellosis a year in the United States, with 400 of those fatal on average (Voetsch et al., 2004). Symptoms are similar to those described for *Shigella* and pathogenic *E. coli* infection. The history of the classification of *Salmonella* is a complicated one, with the genera currently divided into two species: *S. bongori* and *S. enterica* (McQuiston et al., 2008). *S. enterica* is further divided into six subspecies (Tindall et al., 2005). *S. bongori* was originally considered a subspecies (called *Salmonella subsp. bongori*) but was reclassified in the 1980s as a separate species (Reeves et al., 1989). Much work has been done to attempt to recover the phylogeny of *Salmonella*, using methods such as DNA-DNA hybridisation (Corsa et al., 1973), multilocus enzyme electrophoresis (Boyd et al., 1996), mircoarray data (Porwollik et al., 2002) and sequence-based methods (Boyd et al., 1996, McQuiston et al., 2008). The different methods produced conflicting results, however the recent analysis by McQuiston and colleagues using more robust sequence data and correcting for HGT seems to have produced a well supported phylogeny (McQuiston et al., 2008).

1.2.7 Single versus multiple origins of *Shigella*

The nature of the relationship between *Shigella* and *E. coli* is an ongoing debate (Pupo et al., 2000; Escobar-Paramo et al., 2003; Yang et al., 2007). It has long been known that been known that *Shigella* is closely related to *E. coli*, to the point where they can be placed in the same species (Brenner, 1984). The reason that *Shigella* continues to be treated as a separate genera is largely due to the serious nature of shigellosis (Pupo et al., 2000).

In 1997 Pupo et al. carried out a multi locus enzyme electrophoresis study at ten enzyme loci and the used sequence of the housekeeping gene *mdh* in an attempt to understand the relationships of pathogenic *E. coli* strains and *Shigella* strains (Pupo et al., 1997). They found that the *Shigella* strains formed a single cluster on their phylogenetic trees, grouping within *E. coli*. The strains showed a stronger clustering than many of the pathogenic *E. coli* strains. For this reason they suggested that it would be more appropriate to include the four *Shigella* 'species' as strains within *E. coli*. Furthermore they found that pathogenicity of *E. coli* and *Shigella* strains had likely arisen multiple times, with the acquisition of the virulence plasmid providing the potential for any strain of *E. coli* to become pathogenic.

Later, in 2000, Pupo et al. continued their investigation into the origin of pathogenicity in *E. coli* (Pupo et al., 2000). Using four chromosomal regions they built phylogenetic trees to assess with greater confidence whether or not pathogeneticy in *E. coli* could be traced back to a single evolutionary event (presumably an initial acquisition of an ancestral virulence plasmid) or whether it had arisen independently in multiple strains, as the 1997 data suggested. In the resulting phylogenetic trees they consistently found three separate clusters of *Shigella* strains. Additionally the clusters consisted of strains from more than one traditional *Shigella* species, implying the division of *Shigella* into four species was incorrect. Their conclusion was that the *Shigella* phenotype arose seven times. The explanation of the common characteristics of *Shigella* strains, in light of multiple independent origins of the phenotype, was convergent evolution. Pupo and colleagues suggested that upon acquisition of the virulence plasmid *Shigella* strains have a tendency to lose various catabolic pathways and motility. These losses are a product of the change in environment and as such it would make sense for convergence of phenotypic characteristics. They also speculated that enteroinvasive *E. coli* strains, which share many phenotypic characteristics with *Shigella* strains and also possess a virulence plasmid, may in fact be *E. coli* strains in an intermediary state of progressing to the *Shigella* phenotype. It was noted, however, that the enteroinvasive strains examined did not fall in any of the *Shigella* clusters and therefore the distinction was not arbitrary.

The multiple origins theory of the evolution of the *Shigella* phenotype was not to go unchallenged. In 2003 Escobar-Paramo and co-workers revisited the question of the origins of *Shigella* by looking at the evolution of four chromosomal genes and three virulence plasmid genes (Escobar-Paramo et al., 2003). Under the multiple origins theory it was expected that phylogenetic trees based on the genes from the plasmid and the genes from the chromosome would produce unrelated groups. This would correspond to the virulence plasmid being transferred multiple times horizontally, and therefore the plasmid genes would not conform to a maternal pattern of inheritance while the chromosomal genomes would. If on the other hand the virulence plasmid had been acquired only once, the single origin theory, then the evolutionary histories of the plasmid and chromosomal genes would agree with one another. The data presented by Escobar-Paramo et al. supported the single origin theory, as there was very little disagreement between the trees of genes from the plasmid and those from the chromosome, with any disagreement suggested to be due to partial plasmid gene transfer. They rejected the multiple origins hypothesis and concluded that a single, ancestral virulence plasmid arrived into an *E. coli* strain and this gave rise to a monophyletic group from which all *Shigella* and enteroinvasive *E. coli* strains descended.

The story was to take a further twist however in 2007 when Yang et al. investigated both hypotheses with more robust data (Yang et al., 2007). They constructed three trees: a large chromosomal tree using 23 housekeeping genes, a chromosomal tree based on 4 housekeeping genes but with a larger sampling of the *E. coli* superfamily and a virulence plasmid tree using 5 genes taken from outside the entry region of the plasmid. For both chromosomal trees they found groupings in agreement with the results of Pupo et al. (2000). For the tree based on the virulence plasmid genes they found that while the most of the strains grouped into the three main clusters defined in the chromosomal tree, the relationships between these clusters did not match. This conflicting topology derived from the virulence plasmid genes disagreed with the predictions of the single origin hypothesis. They concluded that the *Shigella* and enteroinvasive *E. coli* have multiple origins arising from multiple horizontal transfers of ancestoral virulence plasmids.

The current body of evidence lends more support the multiple origins theory. The increasing availability of complete genome sequences for *Shigella* and *E. coli* should lead to a definitive answer to the question of the origins of pathogenicity in *E. coli* in the near future.

1.3 Operons and gene clusters

Operons and gene clusters are examples of higher-level genomic organisation. In this section I will discuss the discovery of the *lac* operon, provide a formal definition of a gene cluster and examine examples of gene clusters and operons in prokaryotes and eukaryotes.

1.3.1 The discovery and mechanism of the *lac* operon:

In 1960 Jacob and Monod elucidated the system by which the genes involved in the breakdown of lactose are regulated (Jacob et al., 1960). Four years later they were awarded the Nobel Prize for medicine, shared with Andre Lwoff, for their discoveries "concerning genetic control of enzyme and virus synthesis". The regulatory system, known as the *lac* operon, has become a genetic paradigm.

The *lac* operon in *E. coli* is a complex yet elegant system for the regulation of genes involved in the conversion of lactose to glucose or galactose (Jacob and Monod, 1960). It consists of four genes: a repressor gene (*lacI*), a *β*-galactosidase (*lacZ*), a permease (*lacY*) and a transacetylase (*lacA*) (see figure 1.6). The distinguishing feature of the *lac* operon (and operons in general) is co-transcription. The *lacZ, Y* and *A* genes are co-transcribed into a single mRNA product. The system is induced in the presence of lactose. Under normal conditions the repressor protein will bind to the operator

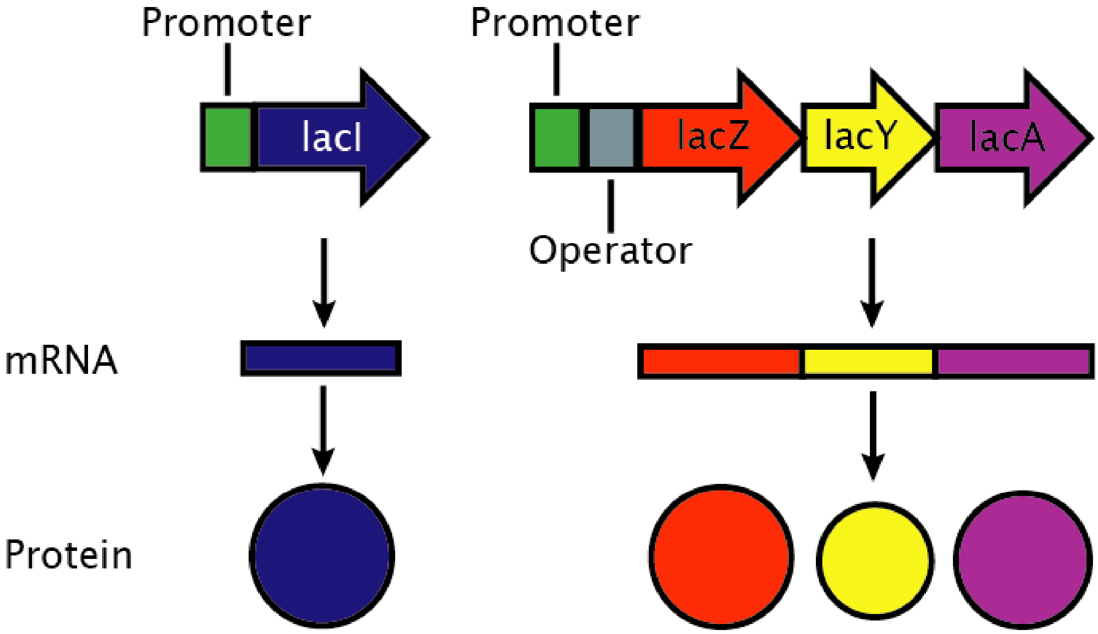


Figure 1.6: The structure of the *lac* operon. The *lacZ, Y* and *A* genes are under the control of the same promoter and are co-transcribed into a single mRNA. The repressor gene, *lacI*, is under the control of a different promoter and is transcribed separately.

region, blocking the binding of RNA polymerase and preventing the transcription of *lacZYA*. When lactose enters the cell small amounts of a lactose isomer, allolactose, are formed. Allolactose binds to the repressor, preventing it from binding the operator, allowing the transcription of *lacZYA*. As the products of the *lac* operon become active, they begin to break down lactose and allolactose, releasing the repressor and preventing further synthesis of the *lacZYA* genes.

The ultimate function of the *lac* operon is the formation of glucose, the preferred energy source of the cell. A further layer of regulation is present in the *lac* operon: the levels of glucose present in the cell also regulate the functioning of the system. The *lac* operon is at peak performance when glucose levels are low and lactose is present. This is achieved via a small molecule called cyclic adenosine monophosphate (cAMP) and its receptor protein, cyclic AMP receptor protein (CRP). cAMP levels are inversely proportional to glucose levels in *E. coli*. CRP and cAMP bind one another, forming a complex. This complex binds the promoter region of the *lac* operon and works as a transcriptional activator. Without the presence of this activator RNA polymerase binds weakly to the promoter and transcription is rarely initiated. In this way the CRP-cAMP complex positively regulates the *lac* operon and, in a broader sense, the relative levels of glucose and lactose regulate the system as a whole.

1.3.2 Operons in prokaryotic genomes:

The *lac* operon is no evolutionary singularity; operons are common across prokaryotes (Ermolaeva et al., 2001; Price et al., 2005a). Roughly half of all protein-coding genes in a typical prokaryotic genome are in operons (Price et al., 2006). Prokaryotes show relatively low levels of conservation in terms of gene order and their genomes are prone to rearrangements (Mushegian and Koonin, 1996; Watanabe et al., 1997; Dandekar et al., 1998). Thus operons are generally not conserved and only 5-25% of genes belonging to operons in a typical prokaryotic genome are in the same operon in two or more distantly related species (Wolf et al., 2001b).

The reason why operons form and persist is a hotly debated topic (Lawrence and Roth 1996; Pal and Hurst, 2004; Price et al., 2005b). Regardless, they are often distributed across species via vertical inheritance (Itoh et al., 1999; Overbeek et al., 1999; Wolf et al., 2001b) and can also be transferred to distantly related species via HGT (Lawrence and Roth, 1996; Omelchenko et al., 2003; Xie et al., 2003a). Omelchenko and colleagues, in a 2003 study of the relationship between HGT and operons, found many examples of new operon acquisition, paralogous operon acquisition and xenologous operon displacement via HGT. In addition they found many cases of what they termed mosaic operons. Mosaic operons are operons whose genes show different evolutionary histories.

Many metabolic processes, such as pathways for the biosynthesis of amino acids, are organised into operon structures (Goldschmidt and Cater, 1970; Xie et al., 2003b; Omokoko et al., 2008). Some of these, such as the tryptophan operon, have an ancient origin. It is believed that the operon is ancestral to bacteria and archea (Xie et al., 2003b). Two major evolutionary events have taken place since its formation. The first was the splitting of the operon in two, the second was when it was rejoined via a gene fusion event (Xie et al., 2003b). Other operons have formed more recently (Price et al., 2005a), demonstrating that operons are not a remnant of some ancient gene organisation strategy, but rather a dynamic process of birth and dissemination that is continually active in prokaryotic genomes (Price et al., 2006).

Perhaps the most interesting question regarding operons is how and why they form. Many models exist to explain the origins of operons and they will be discussed in detail later in the text.

1.3.3 Operons in eukaryotes:

Operons are defined as a cluster of genes under the control of a single promoter (Jacob et al., 1960). The wording of this definition is important, it does not state that a polycistronic mRNA is produced. Eukaryotes were originally thought not to possess operons. However, over the years, genetic structures that fall under the definition of operons have been discovered in a variety of eukaryotes (Muhich and Boothroyd, 1988; Lee, 1991; Spieth et al., 1993; Davis and Hodgson, 1997; Ganot et al., 2004). Much work has gone into examining operons in nematodes in particular, including anaylsis of operon conservation and the evolution of *trans*-splicing, the system by which the transcripts of such operons are resolved into mature mRNA (Guiliano and Blaxter, 2006).

Eukaryotic operons can be divided into two classes: those that produce polycistronic initial transcripts that are co-transcriptionally processed to form monocistronic mRNA, found in nematodes, flatworms and some primitive chordates (type 1 operons), and those that produce dicistronic transcripts that are translated in that form, found in flies, vertebrates and plants (type 2 operons) (Blumenthal, 2004).

Type 1 operons differ from prokaryotic operons in that the single initial polycistronic mRNA is split into monocistronic mRNA before translation. Potential operons are identified by finding genes in the same orientation with an unusually small amount of intervening DNA. This alone is not proof of an operon as it is difficult to eliminate the possibility of a promoter lying between the two genes (Blumenthal, 2004). Searching for the transcripts of potential operons is also a tricky task as there is often little accumulation the polycistronic precursor before it is *trans*-spliced (Blumenthal, 2004). In nematodes, such as *C. elegans,* the precursor is stable enough to be detected but this does not prove that the polycistronic mRNA leads to mature mRNAs. Nematode operons have however been successfully identified via the fact that the genes that are SL1 and SL2-like *trans*-spliced correlate strongly with true nematode operons (Blumenthal et al., 2002).

Type 2 operons, found in *Drosophila*, vertebrates and also plants, are dicistronic operons. Dicistronic operons are always composed of two genes that are transcribed into a dicistronic mRNA (Blumenthal, 2004). This transcript does not undergo *trans*-splicing, but is instead transported to the cytoplasm and translated. In this sense dicistronic operons are much more similar to their prokaryotic counterparts. The exact mechanism of translation of the second gene in a dicistronic mRNA is not fully understood. Matsuda and Dreher showed that the close spacing of AUG initiation codons can confer dicistronic character (Matsuda and Dreher, 2006). They found that for an overlapping dicistronic mRNA, when the AUG codons were 7 nucleotides apart, translation of both genes occurred. Raising the genetic distance between the codons increased the expression of the upstream gene while decreasing the expression of the downstream gene, eventually to the point where expression was converted from dicistronic to monocistronic.

Type 1 and type 2 operons both display patterns in terms of their gene content (Blumenthal, 2004). In an analysis of type 1 operons in *C. elegans*, Blumenthal and Gleason, 2003, showed that certain classes of genes were often found in operons, in particular genes encoding the machinery for expression, transcription, splicing and translation, whereas other classes, such as those related to perixisomes or cuticle formation, are never found in operons (Blumenthal and Gleason, 2003). Many of the genes of type 1 operons do have some functional relationship to one another (Page, 1997; Treinin et al., 1998; Furst et al., 2002). Similarly, for type 2 operons functional relationships can be observed between dicistronic gene pairs. The γ-glutamyl kinase and

γ-glutamyl phosphate reductase in tomato form a discistronic mRNA, believed to be of bacterial origin (Garcia-Rios et al., 1997). The stoned A and B proteins in *Drosophila*, which are co-localised in the nerve terminals, are encoded as a discistronic mRNA (Andrews et al., 1996).

Eukaryotic operons, particularly type 2 operons, share common features with prokaryotic operons. However, it is evident that while they fall under the technical definition of an operon, they have a markedly different rule set governing them compared to prokaryotic operons.

1.3.4 Gene clusters:

A more loosely defined and perhaps more mysterious level of organisation in both prokaryotes and eukaryotes is the concept of a gene cluster. For the purpose of this thesis I define a gene cluster as a group of functionally related genes in close physical proximity (figure 1.7). This definition is distinct from that of an operon in that it says nothing of regulation. A gene cluster could consist entirely of individually regulated genes, multiple operon structures or a mix of both. Under this definition all operons are gene clusters, but not all gene clusters are operons.

One of the most notable, and ancient, examples of a gene cluster is the Hox gene cluster (Ferrier and Holland, 2001). Hox genes dictate the identity of an embryo along the

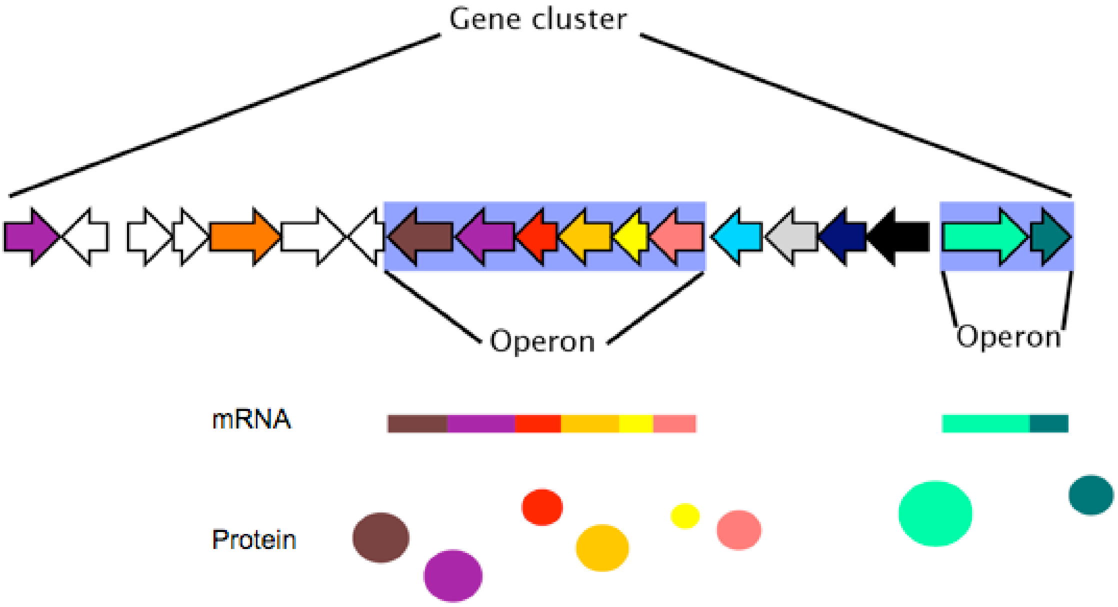


Figure 1.7: Gene clusters and operons. This example shows a hypothetical gene cluster. Genes are represented by arrows, with the arrow pointing in the direction of transcription. Genes that are coloured are functionally related. Genes that are the same colour are homologous. White genes denote those genes that have no functional relationship to the cluster. This cluster contains two operon structures, denoted by the coloured box behind each operon.

anterior-posterior axis. Hox genes contain a homeobox: a 180bp region that encodes a homeodomain. The homeodomain has the ability to bind DNA. Genes containing homeobox encode transcription factors involved in switching on large sets of genes. The clustering of Hox genes is widespread throughout the animal phyla (Ferrier and Holland, 2001). Perhaps the most interesting property of the Hox cluster is that the genes display co-linearity. The genes located towards the 3' end work on the anterior of the embryo, central genes are involved in the development of the mid section, while the 5' genes function in the development of the posterior region of the embryo (Lewis, 1978).

The *DAL* gene cluster is the largest metabolic gene cluster in yeast (Wong and Wolfe, 2005). It consists of six genes encoding proteins that allow *Saccharomyces cerevisiae* to use allantoin as a nitrogen source (Cooper, 1996). Unlike the Hox cluster, which is deemed ancient, the *DAL* cluster formed relatively recently, assembling through a series of near simultaneous genomic rearrangements in the ancestor of *S. cerevisiae* and *Saccharomyces castellii* (Wong and Wolfe, 2005). Wong and Wolfe traced the formation of the cluster to a reorganisation of the purine degradation pathway, which switched from utilising urate to allantion.

While gene clusters in eukaryotes are not unusual, clustering in prokaryotes is more pronounced. The phenylacetate degradation pathway provides a window onto just how dynamic the clustering of functionally related genes can be (Luengo et al., 2001). The genes of the phenylacetate degradation pathway (*paa* genes) encode proteins involved in the conversion phenylacetate into succinyl-CoA, connecting the pathway to the TCA cycle (Ismail et al., 2003). Diverse clusters of the 15 genes associated with the pathway can be found in many bacterial genomes (Luengo et al., 2003). Perhaps the most striking feature is the fact that no real structural identity exists for *paa* gene clusters. With the exception of *paaABCDE*, whose products form a complex with one another, the contents and order of genes in *paa* clusters differ from genome to genome. Even the paa clusters of *Escherichia coli* K12 and *Pseudomonas putida* U (Ferrandez et al., 1998; Olivera et al., 1998), which clearly display recent common ancestry, contain slightly different sets of genes and are under different regulation schemes (Ferrandez et al., 1998).

Like operons, there are several theories concerning how and why gene clusters form. They will be discussed in detail in the next section.

1.4 Models of operon and gene cluster formation

The selective forces driving the formation of gene clusters and operons are one of the most intriguing mysteries in modern genomics. Many models attempt to explain why genes are organised into clusters and operons. In this section I will discuss some of the more popular and recent models, in particular the evidence both for and against each of them.

1.4.1 The Natal model:

The Natal model is the simplest explanation of why functionally related genes are found in close physical proximity in a genome. It suggests that genes are clustered because they are born that way (Lawrence and Roth, 1996). The Natal model is a culmination of several observations on the nature of genes coding for biochemical pathways. In 1935 Graœneberg found that adjacent duplications were frequent in *Drosophila* (Graœneberg, 1935). This observation was given an evolutionary context by Lewis in 1951, who suggested that the divergence of duplicated genes could lead to functionally and physically linked gene clusters (Lewis, 1951). This idea of tandem gene duplication effectively growing a biochemical pathway is supported by the fact that the gene order of the *trp* and *his* operons in *S. typhimurium* reflects the order of the reactions involved in their biosynthesis (Lawrence and Roth, 1996). Pathways were considered to evolve based on the limiting effect of intermediate substrate, with gene duplication and divergence allowing the conversion of similar compounds into the limiting substrate.

Several problems exist when examining the Natal model in light of molecular sequence data. The major prediction of the Natal model is homology between clustered genes, given that genes are clustered via duplication. This is untrue for the majority of bacterial operons, whose sequences show no obvious homology (Lawrence and Roth, 1996). In addition to this, the Natal model does not account for the persistence of gene clusters (Lawrence and Roth, 1996). The fact that functionally related genes are found in close physical proximity implies that there is a selective advantage to keeping them together (Demerec and Hartman, 1956). Therefore, this advantage must still exist if two non-homologous, but functionally related, genes are brought together, by some mechanism such as recombination or horizontal gene transfer. This observation suggests an alternative route for genes to cluster.

Some examples of bacterial operons adhering to the Natal model are known to exist, for example the histidine operon (Fani et al., 1994). Also, gene clusters in eukaryotes, while considerably less frequent than those in prokaryotes, often fit the Natal model. The mammalian β-globin gene cluster evolved through duplication and divergence (Maniatis et al., 1980). Therefore, in a sense, the Natal model is not incorrect; rather, it is a model that accounts for a small percentage of gene cluster formation.

1.4.2 The Fisher model:

The Fisher Model postulates that clustering of genes offers the benefit that random recombination events will tend to disrupt co-adapted genes less often. The model is named after Ronald Fisher, who noted that co-adapted alleles had higher levels of linkage (Fisher, 1930). Later it was suggested that selection for co-adapted alleles could give rise to gene clusters (Bodmer and Parsons, 1962; Stahl and Murray, 1966). As the distance between the co-adapted alleles decreases, so too does the probability that recombination events could disrupt the co-adapted alleles.

The Fisher model fits well with observed gene clustering within bacteriophage genomes (Stahl and Murray, 1966). Genes in certain families of bacteriophages are arranged into functional clusters (Botstein, 1980; Campbell and Botstein, 1983; Casjens et al., 1992). Because the genes are arranged in discreet modules it is possible for recombination to occur on the limits of these functional clusters, generating new phage combinations, without disrupting the interacting genes contained within a cluster. Many of the genes within these clusters interact physically, in accordance with predictions from the Fisher model (Casjens, 1974; Casjens et al., 1992).

According to the Fisher model two conditions are required for genes to cluster (Lawrence and Roth, 1996). The first is the existence of multiple variants of co-adapted gene complexes. Second, recombination events must frequently disrupt these co-adapted gene complexes. If both these requirements are met, then selection can occur to cluster co-adapted alleles. However, while recombination occurs in eukaryotes during meiosis and sexual reproduction, it is less frequent in prokaryotes (Lawrence and Roth, 1996). Additionally, genes involved in metabolic pathways are frequently clustered and yet do not necessarily interact with one another, and therefore are not clustered due to coadaptation. For example the genes involved in histidine biosynthesis are clustered and show no evidence of physical interaction with one another (Martin et al., 1971). This rules out the Fisher model as a primary mechanism for the formation of gene clusters.

1.4.3 The Co-regulation model:

The co-regulation model draws on the fact that genes in an operon have the advantage of being co-transcribed. Co-regulation offers several possible benefits. All the genes in an operon are under the control of a single operator and are transcribed as a single mRNA transcript. They are therefore active and repressed at the same time. For the same reason they are present in equimolar amounts. Additionally there is a localised concentration of the gene products in prokaryotes, where transcription and translation are coupled (Svetic et al., 2004). In terms of genes coding for metabolic pathways, all of these factors, at least on the surface, suggest that having genes clustered into operons increases the efficiency of the associated biochemical reactions. Indeed, the discovery of operons led many to believe that co-regulation could be the driving force behind gene clustering (Pardee et al., 1959; Jacob et al., 1960; Jacob and Monod, 1962). In 2005 Price et al., advanced the Co-regulation model by suggesting that as the amount of regulatory information to control a set of functionally related genes increases, so too does the likelihood that the genes will form an operon (Price et al., 2005b). They supported this theory by observing that operons in *E. coli* and *B. subtilis* tend to have more conserved regulatory sequences than other genes. This is consistent with the fact that not all genes that reside in operons are functionally related (Rogozin et al., 2002).

However, the Co-regulation model has a number of substantial problems associated with it. Since the model provides no selective benefit for clustering until co-transcription, rare and precise chromosomal rearrangements would be required for every gene added to the operon (Lawrence and Roth, 1996), though this point has been questioned (Price et al., 2005b) given the high rates of rearrangements in some bacterial genomes (Papadopoulos et al., 1999) coupled with large population sizes. The Co-regulation model provides no real explanation for why genes are often clustered but not in a single operon. For example, metabolic genes in involved in phenylacetate degradation are located in a single cluster in both *Escherichia coli* K12 and *Pseudomonas putdia* U (Ferrandez et al., 1998; Olivera et al., 1998). In both genomes the single cluster is actually a collection of multiple operon structures. In cases like this, where physical proximity is selected for in the absence of co-regulation, the Co-regulation model breaks down. On top of this genes can be co-regulated without being clustered and the potential benefits of co-transcription are not necessarily benefits at all. Even when a single transcript is produced for a set of genes, the genes themselves can display different translation efficiencies (van de Guchte et al., 1991) and different mRNA half lives (Blundell et al., 1972) leading to different levels of protein produced from a single transcript (Whitfield et al., 1970). In addition to this, Zaslaver and his co-workers have found evidence of a complex temporal expression pattern for genes involved in amino acid biosynthesis in *E. coli* (Zaslaver et al., 2004). Their discovery of so-called 'Just-in-time' transcription, where the activation of promoters is precisely timed with the order of the steps in a metabolic pathway, further suggests that it is not necessarily beneficial to have a set of functionally related genes co-transcribed in a single mRNA. Taking all this into account, co-regulation is more likely to account for persistence of some gene clusters as opposed to driving their formation.

1.4.4 The Selfish Operon model:

In 1996 Lawrence and Roth proposed the Selfish Operon model (SOM). This was a major departure in thinking about why genes might cluster. They suggested that genes cluster in order to facilitate their own horizontal transfer as a group. In the case of genes for weakly selected functions the SOM provides an escape route from extinction. Such genes can be lost over time in their native host. However, if clustered, these genes may be passed to a new host via HGT, conferring or regaining function in the new host. In this sense the SOM is different to the Co-regulation and Fisher model. The clustering of genes is not related to the fitness of the host, but rather the fitness of the cluster itself (Lawrence and Roth, 1996). Any change in host fitness is associated with the function provided by the acquired cluster.

This alone is not enough to explain gene clusters. The selfish nature of clustering suggested a reason for observing clusters, but not how they form. Lawrence and Roth also provided a mechanism for cluster assembly. They drew their explanation on the fact that the spontaneous deletion of intervening DNA can bring functionally related genes into closer proximity (Demerec, 1960). Normally deletion of intervening DNA would cause deleterious effects, with the loss of any genes contained in the intervening DNA. However, if the DNA has been introduced to a new host horizontally, the intervening

DNA may be of no benefit to the host and quickly deleted, bringing the genes under selection closer together.

Further to this Lawrence and Roth suggested that co-transcription might also be a selfish property. As previously noted, some problems exist with the idea that co-transcription is the sole reason for genes existing in operons. In light of the SOM co-transcription is a logical property of genes in a cluster. Co-transcription means that genes do not need separate promoters. This is an advantage for the survival of the cluster, because if each gene had a separate promoter then the probability of the cluster being non-functional would increase and there would be no selection in the host for retention of the cluster. Along a similar line of thought the authors suggest that translational coupling and the high frequency of *trans*-acting regulatory proteins found adjacent to the operon they regulate are also selfish properties of gene clusters.

Several predictions arise from the SOM. Non-essential genes should cluster. Essential genes should not cluster. Recently introduced selfish operons should be detectable. Lawrence and Roth backed up their claims by providing an operon, the cobalamin biosynthesis operon, which exemplified the SOM, along with computer simulations that produced data in agreement with the model. Other evidence offers support for the SOM, such as the observation that many operons have been acquired via HGT (Omelchenko et al., 2003) and the fact that essential genes do not generally undergo HGT (Lerat et al., 2003).

In spite of this, a large body of evidence against the SOM has amassed. Two of the major predictions of the model, that non-essential genes should cluster while essential genes should not cluster, have been shown to be untrue. In 2004 Pal and Hurst carried out a study of essential and non-essential genes in *E. coli* (Pal and Hurst, 2004). Firstly, they found that essential genes have a slightly higher tendency to reside in operons when compared to non-essential genes. This result was in line with the predictions of the Co-regulation model and in direct conflict with the SOM. Secondly, they found that the clustering of essential pairs of functionally related genes was particularly pronounced. This conflicts with the prediction of the SOM that essential genes should not cluster. Price et al. in 2005 demonstrated that suspected HGT genes in *E. coli* are in general no more likely to be in an operon than to not be in an operon (Price et al., 2005b). Further to this they found that native genes formed new operons at the same rate as HGT genes and that there was no preference for HGT genes to form operons with other HGT genes rather than native genes. Lastly they found that essential genes formed many new operons. These results are not in line with the predictions of the SOM, which predicts that new operons should be acquired via HGT of a selfish cluster into the host genome.

1.4.5 The Protein Immobility model:

The Protein Immobility model (PIM) suggests that there is a thermodynamic advantage to clustering genes because, given that transcription and translation is coupled in prokaryotes, the close physical proximity of their products in the cytoplasm will allow biochemical reactions to proceed efficiently in a low nutrient environment (Svetic et al., 2004). The model assumes that because the cytoplasm of the cell is a dense population of macromolecules, the large size of soluble enzymes essentially fixes in the cytoplasm at the point of their expression. They provide a hypothetical biochemical reaction:

$$A \xrightarrow{E1} B \xrightarrow{E2} C$$

where E1 and E2 and enzymes, A and B are substrates and C is the product. The model assumes that the smaller molecules, A, B and C, are free to diffuse around the cell. E1 and E2, being much larger, are found in much higher concentrations around the point of their expression on the chromosome than elsewhere in the cytoplasm. Given that A is effectively constant through the cytoplasm, the limiting step in the reaction is the conversion of B to C. B will naturally be concentrated around E1, the point of its production, but will freely diffuse around the cytoplasm (the model assumes that the rate of intracellular diffusion of B is large). Under the PIM, the spatial proximity of E2 to the site of production of B will increase the efficiency at which B is converted to C. Since E1 and E2 are anchored relative to the genomic positions of their corresponding genes, shrinking the genetic distance between the genes for E1 and E2 should in theory facilitate more efficient conversion of B to C. The authors suggest that clustering genes in this way favours rapid growth in nutrient limited environments and that the model itself applies to an organism transitioning from stationary phase to an active growth phase.

One major advantage of the PIM is that it provides a straightforward selective advantage for the formation of gene clusters. Additionally, the selective advantage would be relatively weak; clusters would not be an absolute requirement, rather a slight advantage. This would help explain the diversity of homologous cluster structures observed in different genomes (Luengo et al., 2001). In many ways such an advantage parallels codon usage patterns, where a relatively weak selective advantage can have a major influence on the underlying genome (McInerney, 1998). In addition to the computer simulations carried out by the authors, some evidence exists to help support the PIM. The cytosol is crowded with macromolecules (Cayley et al., 1991; Zimmerman and Trach, 1991), with the experimental evidence that size effects protein mobility in the prokaryotic cytoplasm (Elowitz et al., 1999) along with possible binding and confinement effects (Konopka et al., 2006). The effects of macromolecular crowding should be particularly pronounced for proteins that form complexes (Arrio-Dupont et al., 2000).

The main problem with the PIM is that it has undergone very little testing with real data. While experimental and computational support for the model exists, there is little to nothing in the way of hard evidence that the model is correct. The model also makes many assumptions about the distribution of molecules in the cytosol that, while intuitive on some levels, are yet to be verified. A major prediction of the model - that genes that encode larger proteins should cluster more frequently - remains untested.

1.4.6 The Persistence Model:

The persistence model explains clustering in terms of the persistence of genes in bacterial genomes (Fang et al., 2008). Fang et al., suggest that there are two classes of frequently clustered genes: highly persistent genes and rare genes. Persistent are defined as genes present in the majority of organisms. This class of genes not only includes genes that are lethal when knocked out but also genes that drastically affect the fitness of an organism. Rare genes are defined as those that are not widely distributed.

Fang et al. suggest that the clustering of persistent genes is made possible through a constant flux of gene insertion and deletion events. They found that under computer simulation genes were less likely to be affected by a deletion event if they were clustered as opposed to uniformly distributed.

Fang et al.'s explanation for the high levels of clustering of rare genes is that they fitted the "Selfish gene hypothesis" (referring to the SOM), and found that sets of genes that had likely been introduced via a HGT event showed a high tendency to cluster.

The validity of the Persistence model is questionable. The observation that clustering leads to a decreased chance of deletion of a persistent gene fails to take into account that, following the assumption of their model that deletion is a random event, the probability of a deletion event removing a persistent gene is the same regardless of whether they are clustered or unclustered. Clustering genes makes the area of the genome lacking persistent genes larger, but the cluster itself increases in length with each gene added. As such there is no change in the probability that a random deletion event will remove a persistent gene.

So far no single model of either gene cluster or operon formation remains unchallenged. New models are being purposed regularly. The question of how and why genes are organised in a genome is still open.

In this thesis I present an examination of the boundaries of a bacterial species. This is achieved through an analysis of 27 completely sequenced YESS group genomes. Using a variety of methods, including a 16S rRNA phylogeny, data concatenation and supertree construction, I look at the kind of resolution achievable between closely related clusters, i.e. the four genera that comprise the YESS group, and within those clusters. In particular I examine whether the availability of whole genome data leads to more robust results and the level of congruence and conflict between the different approaches.

In addition to this I present software, GenClust, designed for finding clusters of genes in bacterial genomes. The user provides a set of genes and genomes of interest and GenClust identifies any potential homogolous clusters. I demonstrate the ease of use of GenClust with a simple examination of gene clusters associated with the superpathways of amino acid biosynthesis in 180 γ-proteobacterial genomes. The software is relatively easy to use, fast and results can be visualised.

Lastly I examine genes associated with the breakdown of phenylacetate in 108 different bacterial genomes. The degradation of phenylacetate is interesting from an evolutionary standpoint because the underlying genes are often found clustered in bacterial genomes, though their distribution is extremely patchy, implying possible high levels of HGT. I identify many new phenylactetate degradation gene clusters and examine the evolution history of both the clusters and the genes themselves. Using this information I compare the data to the current models of gene cluster formation and provide a perspective on the selective forces and mechanisms driving gene cluster formation.

Chapter 2 - Gene and genome trees conflict on many levels: an analysis of 27 YESS genomes

Note:

In relation to the paper published in Philosophical Transactions of the Royal society B series (Haggerty et al., 2009), I am joint first author. I produced all results presented in this thesis, with the following exceptions, which were jointly produced by L. Haggerty and myself: Identification of single gene families, alignment of single gene families and carrying out a PTP test of all alignments.

2.1 INTRODUCTION

Recently, it has been questioned whether or not there is a future for the Tree of Life metaphor (McInerney et al., 2008). Many have gone further and feel that the time has long since gone when this metaphor was useful (Doolittle and Bapteste, 2007). The central issue is that HGT has affected all or nearly all genes in every genome at one stage in their evolutionary history (Dagan and Martin, 2007; Dagan et al., 2008). The most recent estimate is that in each genome an average of 81±15% of the genes have experienced a HGT event at some stage (Dagan and Martin, 2007). In the next few years, we must precisely describe how the prokaryotic world, in particular, is structured and what exactly HGT has done.

There are two categories of HGT events: homology-dependent and homology-independent (though the most important factor is similarity level, not whether the sequences are homologous). Homologous recombination, according to Ochman *et al.*, occurs mainly within a bacterial species, but there is very little recombination (approximately 1%) between any given species and its close relatives (Ochman et al., 2005). However, the process of non-homologous recombination or the introduction of new genes that have no similarity to incumbent genes is mostly a process that involves organisms that we consider to be very far outside the species boundary. Non-homologous recombination also encompasses recombination events where regions with no significant similarity to anywhere in the recipient genome are carried into that genome by flanking regions that do have similarity to the recipient genome. Lawrence has put forward the theory that integration of foreign non-homologous DNA into a genome is a driver of speciation in prokaryotes (Lawrence, 2002).

On the question of what boundaries might exist that prevent a gene from being successfully incorporated into a recipient genome, Sorek *et al.*, have indicated that gene dosage and promoter structure might be barriers (Sorek et al., 2007). In contrast, McInerney and Pisani suggest that the barriers to HGT, if they exist, might be very low (McInerney and Pisani, 2007). However, these opinions relate to the artificial scenario where barriers to HGT have been measured *in vitro*.

While much of the focus on the issue of HGT has been on the long-term evolutionary history of prokaryotes, a number of studies have examined shallower relationships.

Ochman *et al.* analysed HGT at the shallower taxonomic levels and concluded that while there was relatively frequent HGT between homologous genes within species there was a much lower amount of HGT between homologs across the species boundary (Ochman et al., 2005). Given that new genomes are being sequenced on a daily basis, it is possible to examine what this structure means for microbiology. In particular, this might have an important consequence for our concept of a bacterial species.

2.1.1 What is a bacterial species?

What seems indisputable is that we can identify organisms that have synapomorphies, both genetic and phenotypic. Multi-Locus Sequence Analysis (MLSA) (Gevers et al., 2005) has shown that there is some structure among currently defined species (Kidgell et al., 2002; Achtman and Wagner, 2008; Buckee et al., 2008). However, this kind of analysis, which has been carried out extensively in thousands of isolates, has the limitation that it only examines the evolutionary history of a set of core genes. Not only does this limit the amount of information used in the analysis, core genes are not representative of the rest of the genes in a genome in terms of factors such as functional category and rate of mutation. For a modern system of classification to work, it must use complete genomes and be able to accommodate HGT.

The concept of prokaryotic species is difficult to address and there is considerable diversity of opinion on what constitutes a species among the prokaryotes. HGT might be considered to be a form of sex and therefore, all prokaryotes might be considered to be a single species. Alternatively, we might consider a species to be an ‘irreducible cluster’ of organisms (Staley, 2006) and this seems in many ways to be sensible. Staley has advanced the idea of a genomic-phylogenetic species concept (Staley, 2006). Doolittle has suggested that if a species concept is not needed, it should be let it go, whereas if it can be found it might be useful (Doolittle and Papke, 2006; Papke et al., 2007).

At the moment there is a polyphasic definition of a bacterial species. Depending on the data that are available, this polyphasic definition can involve the use of ribosomal RNA sequence identity, reciprocal DNA-DNA re-association values, biochemical traits and so forth. If there is a valid biological bacterial species concept, it may be possible to ask what drives speciation. Therefore from a number of perspectives it is interesting to explore evolution at the boundaries of recognized species and genera.

In this study I use as an example the YESS group of γ-proteobacteria and examine what kind of phylogenetic signals emerge when different parts of genomes, different genes and different analysis methods are used. At the time of writing 27 YESS genomes are fully sequenced, allowing an exploration of what happens if trees of genomes or subsets of genomes are inferred, given that the rate of homologous recombination and level of sequence similarity are expected to be high for many of the strains in this group.

2.1.2 A test dataset for exploring groups of the YESS group:

The YESS group of γ-proteobacteria, consisting of *Yersinia*, *Escherichia*, *Salmonella* and *Shigella,* are facultatively anaerobic Gram-negative rod-shaped bacteria that are catalase-positive and oxidase-negative (Brenner, 1984). The YESS group is of particular interest as many members are human pathogens. For instance, *Y. pestis* was the causative agent of the bubonic plague that killed an estimated 75 million worldwide during the 1300s. *Shigella* and enteroinvasive *E. coli* (EIEC) are the etiological agents of bacillary dysentery or shigellosis, of which there are an estimated 160 million cases worldwide a year, with approximately 1.1 million deaths, mainly in children under the age of five (Kotloff et al., 1999). *Salmonella* infection, known as salmonellosis, induces vomiting, diarrhea, fever and abdominal cramps and can last several days. Outbreaks of YESS group associated diseases are common (Tacket et al., 1985; Mahon et al., 1997; Lee et al., 2000; Varma et al., 2003) and consequently this group of prokaryotes is extensively sampled in genome sequencing projects.

The phylogenetic relationships of different *Shigella* strains have been the subject of intense debate in recent years. Joshua Lederberg famously said that Enterohemorrhagic *E. coli* (EHEC) were "*Shigella* in a little cloak of *E. coli* antigens". *Shigella* are essentially *E. coli* that have acquired a virulence plasmid (VP) (Sansonetti et al., 1981; Lan et al., 2001). There are two conflicting theories on the origin of *Shigella*. The multiple independent origin theory (Pupo et al., 2000) suggested that *Shigella* strains formed through multiple acquisitions of the VP, whereas the single origins theory

(Escobar-Paramo et al., 2003) claims that a single ancestral acquisition of the VP is responsible for the genus *Shigella*. There has been much debate on the issue, with the balance currently in favour of the multiple origins hypothesis (Yang et al., 2007)

The issue of defining the boundary between Shigella and E. coli typifies the kind of problem that will become more and more commonplace as sampling density increases. In the case of *E. coli* and *Shigella*, the boundary between the two genera is based almost solely on the medical importance of *Shigella* and it has been suggested that the four 'species' within the genus *Shigella* should simply be strains within the genus *Escherichia* (Pupo et al., 1997). The question is whether there are definable boundaries within the *E. coli/Shigella* group, or more precisely whether using the current methods and data are any sub-group boundaries identifiable? As the sequencing of bacterial genomes grows at a rapid pace, the classical species definitions will likely become outdated, so it is time to examine how current methods cope with denser taxon sampling.

2.1.3 Many methods and datatypes:

Phylogenies based on housekeeping genes such as *gyrB, tufA* and *atpD*, are often compared with those based on 16S rRNA phylogenies (Dauga, 2002; Purkhold et al., 2003; Paradis et al., 2005). The goal of comparing genes is to examine linkage disequilibrium or recombination or to overcome systematic biases (Cooper and Feil, 2004) that might be present in one molecule and not in another. Methodological problems that are encountered during phylogenetic analysis include artifacts related to both molecular and lineage-specific differences in evolutionary rates and mutational saturation (Doolittle, 1999a). These processes can sometimes be detected and if an appropriate model of sequence evolution is available, they can be overcome (Rodriguez-Ezpeleta et al., 2007). It should be noted however that HGT can occur in genes that have been cited as unlikely candidates, including ribosomal proteins (O'Neil et al., 1969). One study has even shown that it is possible to replace the 16S rRNA of *E. coli* with the corresponding sequence from *Proteus vulgaris*, though there is an associated drop in growth rate of between 10 to 30% (Asai et al., 1999).

The technique of data concatenation is often used in order to reconstruct phylogenetic relationships (Sanderson et al., 2003). This usually involves multiple gene sequences being concatenated and aligned as a single sequence. Using this greater number of genes is supposed to bring out the true phylogenetic relationships, the theory being that signal, even when it is weak, is cumulative, whereas homoplastic noise will be dispersive (Sanderson et al., 2003). However, in general data concatenation is usually based on small sets of genes. For example Ciccarelli *et al.* used only 31 genes, or less than 1% of the genes in the average genome (Dagan and Martin, 2006), in their data set, to determine the relationships for 191 species (Ciccarelli et al., 2006). Also, data concatenation can sometimes produce misleading results (Rokas et al., 2003; Phillips et al., 2004). This is not the fault of concatenation *per se*; however, concatenation generally leads to long sequences, so this is an importation factor to consider when using concatenated data.

Supertree methods of inferring phylogeny address the weakness of using a tree based on a single alignment by combining data from several input trees into a single representative phylogeny (Creevey and McInereny, 2005). Supertree methods offer the advantage that the leaf sets of the input trees need not match each other exactly, merely overlap. At the level of gene families, this means that it is not necessary for every organism under investigation to have a copy of every gene. Additionally, it is possible to carry out a *post hoc* analysis of agreement between input trees and supertrees in order to assess congruence (Creevey et al., 2004). These are key points in favour of phylogenetic supertrees.

Suitable gene families can be identified using accepted criteria for asserting homology; the phylogenetic relationships inferred from these homologs can be extracted and used to build the supertree. By using large numbers of gene families, the final supertree is based upon many more relationships between the genomes in a given data set than by simply using a small number of genes to build a phylogeny. On this basis, supertree based studies have become increasingly popular in recent times, see for example Pisani *et al.*, 2007, and Beiko *et al.*, 2005. However, there are limitations associated with supertree construction. Probably the biggest drawback is the inability of current software to handle gene families where paralogous sequences are present. This limits the number of gene families used to build the final supertree, given that paralogs are frequent, even in prokaryotic genomes. Some methods can be used to deal with this in part, such as deletion of lineage-specific duplication events, but ultimately the problem is still a serious one. Another problem with supertree methods is that the quality of the supertree is based on the quality of the input data, in this case the input trees. If the input trees themselves have low levels of support for the relationships they represent, or if they do not overlap sufficiently (Scornavacca et al., 2008), or if some organisms are not well represented, then the quality of the supertree will also suffer. However, unlike the issue of using single-gene families, these problems can be addressed to a certain extent by employing various methods to ensure the input trees are of sufficient quality for supertree construction, such as removing poorly aligned regions (Talavera and Castresana, 2007), removing alignments with little signal or removing very short alignments. These kinds of alignments and regions of alignments are expected to confound phylogenetic inference (Talavera and Castresana, 2007).

The purpose of this study is to demonstrate the difficulty associated with using genome data to construct a phylogeny of the YESS group using many of the methods listed in the previous paragraphs, namely single-gene phylogenies, data concatenation and supertree analysis. By doing this it is possible to test whether these organisms can be robustly classified, whether there is a meaningful phylogenetic tree that is agreed upon by a considerable amount of the data and whether there is general agreement across all methods and all data. The YESS group was specifically chosen to look at shallow-level relationships both inside and outside the species boundaries, as they are currently understood.

2.2 Materials and methods

2.2.1 Genome sequences:

The GOLD database (http://www.genomesonline.org/) was used to obtain the genome for 27 completed YESS group genomes. This included 8 *Yersinia*, 8 *Escherichia*, 5 *Salmonella* and 6 *Shigella* genomes. A full list of the individual genomes can be found in the Supplementary Table 1 (see S. I. 2.1).

2.2.2 Ribosomal RNA sequence analysis:

All 188 16S rRNA sequences from the 27 YESS group genomes were aligned using ClustalW v1.83. A total of 7 copies of the 16S gene were retrieved from each YESS genome, with the exception of *Yersinia pestis* Orientalis CO-92, which only had 6 copies of the gene. The alignment (S. I. 2.2) was inspected by eye and ambiguously aligned regions were removed. Using standard methods for finding the optimal model of nucleotide substitution (Keane et al., 2006), the HKY+I+G model was used for all subsequent phylogenetic analyses. A maximum likelihood tree was constructed using Multiphyl (Keane et al., 2007). Confidence in phylogenetic hypotheses was assessed using bootstrap resampling and results are presented following 100 bootstrap replicates.

2.2.3 Topological tests for randomly selected 16S rRNA sequences:

One copy of the 16S rRNA gene was selected at random from each of the 27 YESS group genomes. These 27 randomly selected sequences were aligned and a phylogenetic tree was constructed as in section 2.2.1. This process was repeated 100 times, producing a total 100 phylogenetic trees, each with one randomly selected copy of the 16S gene per genome.

A pairwise AU test was carried out between all 100 trees using CONSEL (Shimodaira and Hasegawa, 2001). For each pair the result of the AU test was examined from both sides. For example, in an test between tree A and tree B, the results of the test using the alignment for tree A and the results of the test using the alignment for tree B were examined in order to determine if tree A and tree B were significantly different. Both trees had to fall outside each other's confidence set for a significant difference in topology to exist. CONSEL performs eight different tests for significant difference between two topologies and if two trees had a score of 0.01 or greater for any of the tests then their topologies were not considered to be significantly different to one another. The results were displayed as a 100x100 symmetrical matrix, where '1' denoted no significant different in topology between a pair of trees existed, while '0' denoted a significant difference. Because no tree was significantly different to itself, the diagonal of the matrix consisted of '1's. See S. I. 2.3 for alignments, trees and CONSEL files.

2.2.4 Housekeeping gene analysis:

The three housekeeping genes *atpD, gyrB* and *trpB*, were retrieved from each genome using BLAST. The sequences were aligned using ClustalW v1.83 (Thompson et al., 2002). Upon inspection of the alignments no further changes were felt necessary because the sequences were strongly conserved and the alignments seemed sensible. Maximum likelihood phylogenetic trees were built using Multiphyl (Keane et al., 2007) with the model selection option turned on. A concatenated alignment of all three genes was also constructed and a maximum likelihood phylogenetic tree was constructed as for the individual genes. Bootstrap resampling was carried out on all trees to assess the level of support for nodes in the resulting trees. See S. I. 2.2 for alignment files.

2.2.5 Identification of single-gene families:

Gene families were identified using the RandomBLAST method as described in (Fitzpatrick et al., 2006). A total of 8,736 gene families were recovered. The set of gene families was then filtered to remove families with fewer than four sequences, which is the smallest number of sequences required to build a non-trivial phylogenetic tree. This left 4,693 gene families. Out of these families, 3,109 were found to be single-gene families, with at most one representative sequence from each of the 27 genomes.

2.2.6 Multiple sequence alignment of remaining single-gene families:

The corresponding amino acid sequences of the 3,109 single-gene families were used as input to ClustalW version 1.83 (Thompson et al., 2002) for multiple sequence alignment. A total of 3,109 alignments were produced. Each of the 3,109 alignments was input into Gblocks (Talavera and Castresana, 2007) to remove poorly aligned regions. A shell script was created to remove badly aligned regions in a more relaxed manner than the default Gblocks settings. The minimal length of a block was set to 8 amino acid positions, and the maximum number of allowed contiguous non-conserved amino acid position to 15. Gapped sites were not systematically removed; rather they were treated as any other site in the alignment. Perl scripts were written to remove alignments that had fewer than 150 residues following analysis by Gblocks. This left a total of 1,960 alignments.

The remaining alignments were converted to nexus format (Maddison et al., 1997) and a PAUP* block (Wilgenbusch and Swofford, 2003) for carrying out a PTP test was added to each nexus file. The nexus files were then executed in PAUP* and a PTP test was carried out on each alignment. The resulting p-values gave a measure of confidence in the strength of the signal within the alignment. Only alignments passing the PTP test, i.e. those with a p-score of <= 0.01 were retained. A total of 1,408 alignments were found to pass the PTP test. Nucleotide sequence alignments were then constructed based on these amino acid alignments (see S.I 2.2).

2.2.7 Construction of phylogenetic trees:

Maximum likelihood phylogenetic trees for the 1,408 alignments were constructed using MultiPhyl (Keane et al., 2007), with the model selection option turned on. This resulted in 100 bootstrapped trees for each alignment. Each set of 100 bootstrap replicates was then summarized as a majority-rule consensus tree using CONSEL (Shimodaira and Hasegawa, 2001). The default settings were changed so that only nodes receiving 70 percent support or greater shown to be resolved on the resultant output tree. This produced 1,408 consensus trees, one tree for each of the 1,408 alignments. These trees were used for the supertree analysis.

2.2.8 Supertree construction:

Clann (Creevey and McInerney, 2005) was used for supertree construction. A variety of different supertrees were constructed using the *dfit* optimization function. All other settings were left on their default values. Bootstrap resampling (100 replicates) of the input data was carried out and supertrees generated using these replicates were summarized using a majority-rule consensus method. See S. I. 2.8 for the 1,408 input tree file for Clann.

2.2.9 Input tree-to-supertree distances:

Clann (Creevey and McInerney, 2005) was used to measure the level of incongruence between the input trees and the dfit supertree. A score was generated for each of the 1,408 input trees in terms of dissimilarity to an appropriately pruned supertree. This score was based on the Robinson-Foulds distance metric (Robinson and Foulds, 1981).

In order to get a better understanding of the signal present in the input trees, a further 1,408 trees of 27 taxa were 'grown', under the Yule model of random tree generation (Yule, 1924), using BioPERL (http://www.bioperl.org). For each of the 1,408 original input trees, the set of taxa present in each tree was recorded and a tree with random branching order was grown using the set of taxa. Through this method each original input tree had a corresponding random tree with the same number of nodes. These 1,408 random trees were also scored against the dfit supertree using Clann. The scores for both the input trees derived from the single gene families and those that were randomly generated were divided into 10 bins and graphed. See S. I. 2.9 for the randomly generated trees and tree-to-supertree distance files.

2.2.10 Minimum-evolution tree:

The nucleotide data for the 1,408 single-gene families was aligned by translating the individual sequences into their corresponding amino acid sequences, aligning the proteins using ClustalW version 1.83 and putting the gap characters into the nucleotide sequences according to where they were found in the amino acid sequences. The sequences were then concatenated into a single alignment. The concatenated alignment then analysed using PAUP* (Wilgenbusch and Swofford, 2003) using the GTR distance matrix method with the optimality criterion set to minimum evolution (Rzhetsky and Nei, 1993). Minimum evolution was used as the dataset was too large for a ML analysis and ME offered the advantage of being able to fit a model to the data, unlike parsimony.

2.3 Results

2.3.1 16S rRNA gene tree:

All copies of the 16S rRNA gene were retrieved from the 27 genomes. This came to a total of 188 genes, which were then aligned. The broad topology of the 16S rRNA tree, as outlined in figure 2.1, is in line with expectations. *Yersina* and *Salmonella* both form monophyletic groups while *Shigella* groups within *Escherichia*. However, many of the other features of the tree are unusual. Firstly, in general the 16S rRNA genes within each genome do not form monophyletic groups with one another. *Shigella* is non-monophyletic, with multiple *Shigella* groupings within the *Escherichia* clade. The simplest interpretation of the data is that homogenisation of ribosomal RNA genes is not sufficiently rapid that each genome has its own unique kind of 16S gene. This means that a genome-of-origin cannot be assigned based on the sequence of the 16S rRNA gene. The alternative explanation is that 16S rRNA genes are being exchanged between strains by some recombination mechanism. The one distinct pattern is that for this collection of genomes, there are three kinds of 16S rRNA – a *Yersinia*-type of rRNA, a *Salmonella*-type of rRNA and an *Escherichia/Shigella*-type of rRNA.

2.3.2 Conflicting topologies of phylogenies of randomly selected 16S sequences:

Figure 2.2 shows the results of pairwise tests of conflicts between trees constructed from an alignment of one randomly selected 16S sequence per genome. If the topology of two trees constructed in this manner differ significantly in topology, i.e. there was a

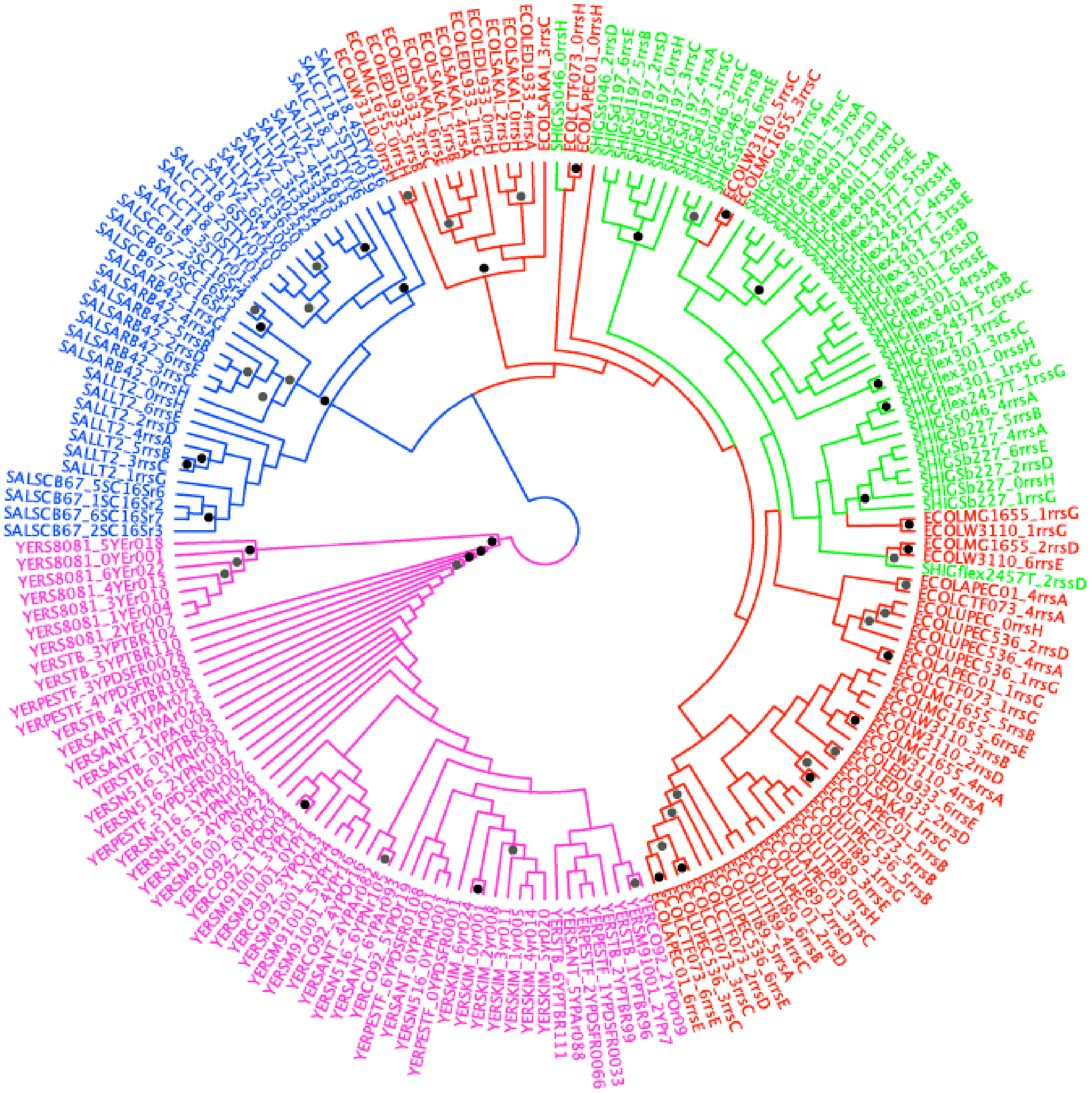


Figure 2.1: Phylogenetic tree of 188 16S rRNA sequences. Grey nodes denote > 50 percent bootstrap support, black nodes denote > 70 percent bootstrap support. The different colours for the different branches represent the different groups with *Yersina* in purple, *Escherichia* in red, *Shigella* in green and *Salmonella* in blue.

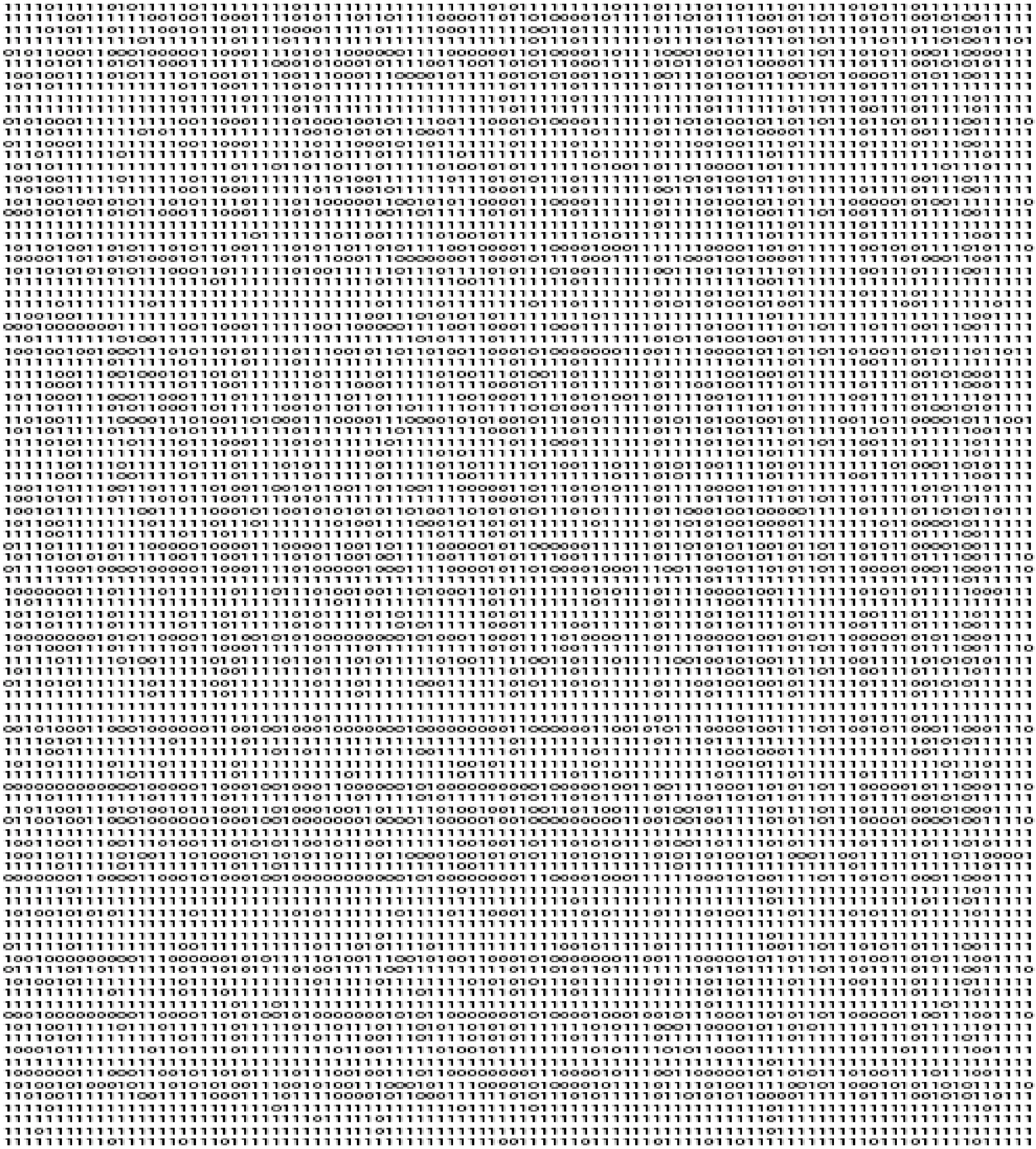


Figure 2.2: Pairwise tests of topological conflict between 100 16S rRNA trees. The matrix is symmetrical. ‘1’ denotes a pair of trees shows no significant difference in topology for at least one test. ‘0’ denotes a pair of trees that had significantly conflicting topologies under all tests.

significant difference in all eight of CONSEL's tests (technically 16 tests were carried out per entry, as the eight tests were carried out for both the alignments for each pair), then the corresponding entry was denoted with a '0'. Conversely if two trees were not deemed significantly different in topology by even a single test, the corresponding entry was denoted with a '1'. The matrix was symmetric with dimension 100x100. Therefore, after subtracting the diagonal (no tree shows conflict with itself) and dividing the results in half, the percentage of conflict was measured. For a total of 4950 unique pairs, 3687 showed no significant conflict in tree topology, while 1263 pairs of trees had significantly conflicting topologies. This equated to 25.5 percent conflict within the matrix. In tree terms, by selecting random copies of the 16S rRNA from each genome there was a one-in-four chance that a pair of trees derived from sets selected in this manner would have significant conflict in their topologies.

2.3.3 Concatenated atpD, gyrB and trpB tree:

Figure 2.3 (a-c) shows the trees for the three housekeeping genes *atpD*, *gyrB* and *trpB*. Once again, in all trees there is a monophyletic grouping of *Yersina*, a monophyletic grouping of *Salmonella,* and the *Shigella* sequences are mixed with the *E. coli* sequences. A closer analysis of these gene trees reveals some common features. Assuming a rooting on the split between *Yersinia* and the rest of the genomes, *Y. enterocolitica* is the deepest branch in each tree, followed by *Y. pseudotuberculosis*. *Y. pestis* Microtus and *Y. pestis* Mediaevalis group together in the *gyrB* tree and *tryB* tree. The relationships for the

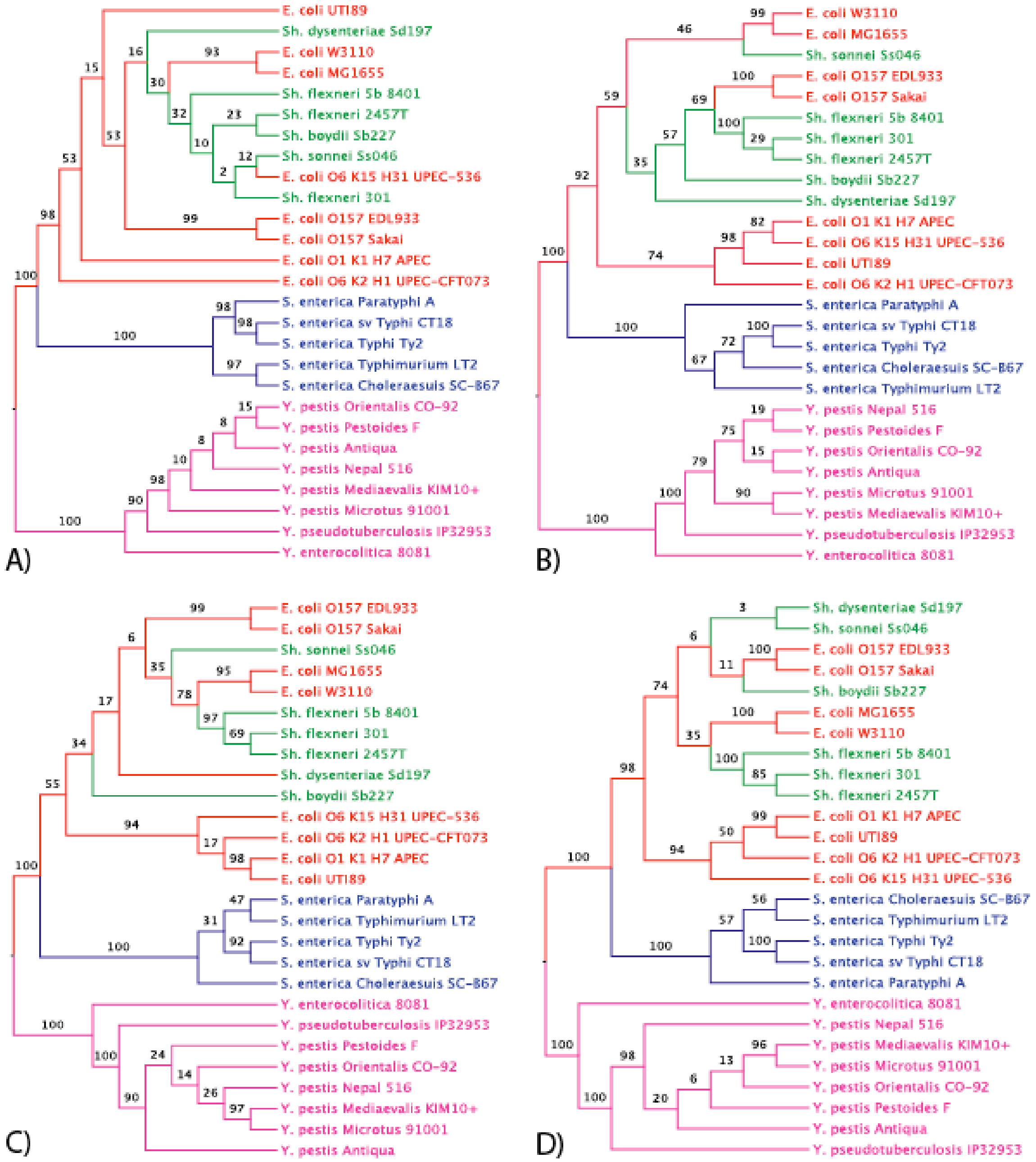


Figure 2.3: Phylogenetic trees for (A) *atpD*, (B) *gyrB*, (C) *trpB*, (D) concatenated alignment for *atpD, gyrB* and *trpB*.

Salmonella genomes show a similar level of conflict. *S. enterica* Typhi Ty2 and *S. enterica sv* Typhi CT18 group together on all the trees. The other three *Salmonella* strains are found located in different positions in each tree. In the *gyrB* and *trpB* trees, *E. coli* MG, *E. coli* W, *E. coli* 0157 and *E. coli* Sakai from a group outside the subclade formed by the remaining four *E. coli* and six *Shigella* strains. The *atpD* tree is different; with the 0157/Sakai group outside the *Shigella*, while *E. coli* 06 K15 moves from outside *Shigella* to a grouping with *Sh. sonnei*.

Using the CONSEL software (Shimodaira and Hasegawa, 2001) a number of analyses of the significance of the difference between the trees generated from the three housekeeping genes were carried out (see S. I. 2.4-2.6). For each of the three alignments, the topology of the maximum likelihood tree was tested to see whether its topology was within the confidence set of trees for the other two alignments. Eight different tests of significant difference were carried out for the maximum likelihood tree versus the other two trees. So, for example, for the *gyrB* alignment, the topology of the *gyrB* tree was compared to that of the *atpD* and *trpB* trees. This process was repeated for each of the three alignments giving a total of 24 tests. In total, for each alignment, the two trees that were not derived using that alignment were rejected by 23 out of 24 tests. The single exception among the 24 tests was where the Shimodaira-Hasegawa (SH) test did not consider the topology of the *gyrB* tree to be outside the confidence set of trees for the *trpB* alignment and therefore did not reject that topology (p=0.132). Notably, all other tests of the significance of difference for this alignment and tree combination rejected the topology of the *gyrB* tree. Interestingly, when the test is carried out for the concatenated alignment against the individual gene trees, only the *atpD* gene tree fell outside the confidence interval, doing so on all eight tests, while the *gyrB* and *trpB* genes only failed one of the eight tests each (see S. I. 2.7). This suggests that the *gyrB* and *trpB* genes are not significantly different in topology (under most tests) to the tree from the concatenated alignment. However, they do have a significant difference in topology with each other in all but one test. Little confidence can be placed in the tree for the concatenated data as a result.

2.3.4 Supertree of 1,408 single gene families:

Figure 2.4 shows a supertree constructed from 1,408 single-gene families derived from nucleotide alignments. The supertree recovered using these shows strong support across the majority of the tree. Some low support values exist in the *Yersinia* clade, but in general the tree has strongly supported relationships, probably indicative of the greater amount of signal in the single-gene families data.

2.3.5 Tree-to-supertree distances for 1,408 source trees:

One of the most interesting questions is whether or not the various phylogenetic trees used as input to generate the supertree are similar in topology to an appropriately pruned supertree. Tree-to-tree distances from the 1,408 ML input trees to the supertree were calculated using the Robinson-Foulds distances (Robinson and Foulds, 1981). The average input tree-to-supertree distance was 1.1733 (median 1.168, range 0.181-3.458)

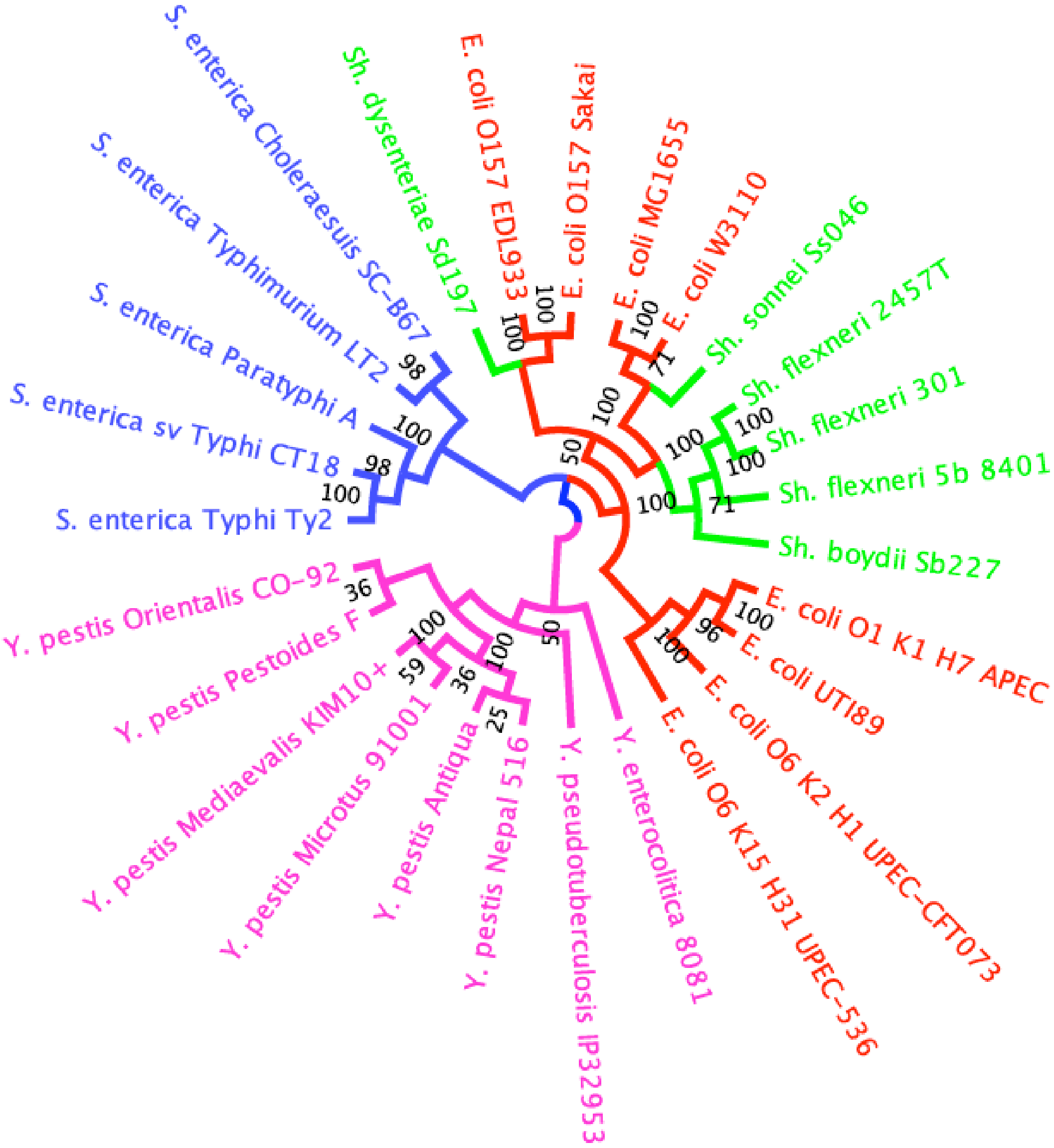


Figure 2.4: Supertree of 1,408 single-gene families using nucleotide data.

(see figure 2.5a). Input trees based upon families with larger numbers of sequences were in general responsible for much of the conflict observed, though it should be noted that since 587 of the 1,408 families were universally distributed across the 27 genomes, some of this is simply a reflection of the abundance of widely distributed genes in the data. In terms of phylogenetic conflict, of the families with the largest tree-to-tree distances compared to the supertree, many were found to be ribosomal or ribosome-associated proteins. Because of the high level of similarity in the sequences of these genes, there is little statistical support for the input trees. While the groups themselves may be well defined, the lack of resolution of the internal relationships within a group leads to conflict with the supertree.

When the scores of these trees were compared to scores for 1,408 randomly generated trees (figure 2.5b), there was a clear distinction in the scores. The average distance for the random trees to the pruned supertree topology was 2.53 (median 2.7, range 0.5-3.95). The peaks for the randomly generates trees versus the true input trees in figure 2.5 do not overlap, suggesting that while there is conflict in the true input trees they fit the supertree significantly better than the corresponding set of random tree. This was not surprising, but did suggest that, despite conflict persisting, there was underlying signal in the true input trees.

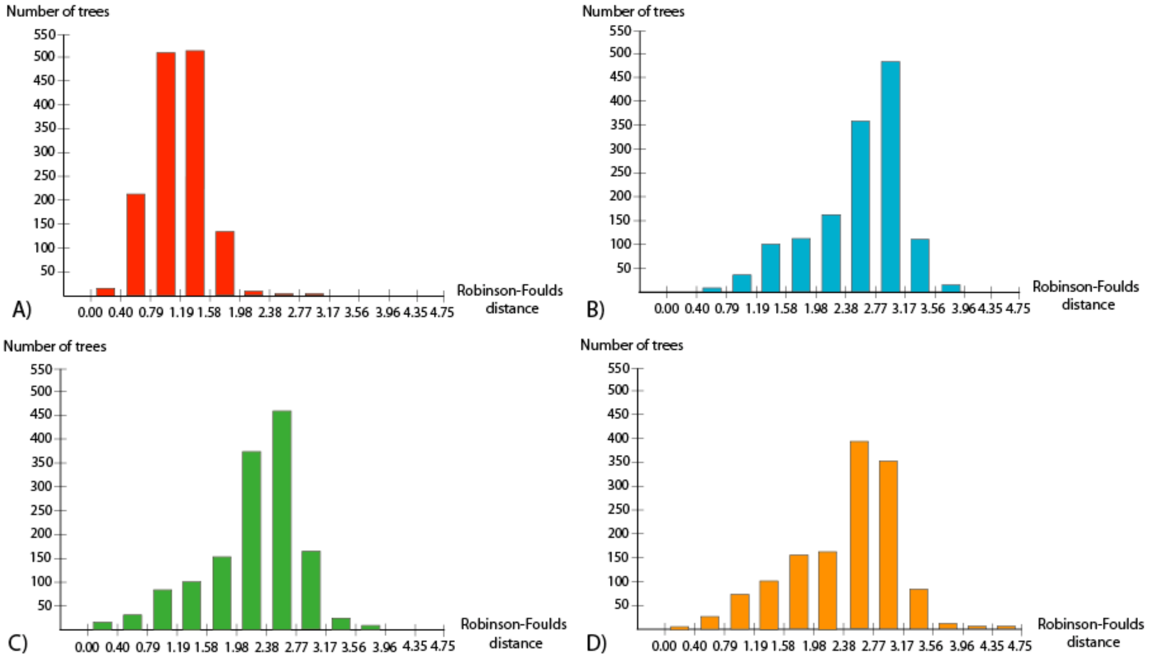


Figure 2.5: Tree-to-supertree distances. (A) Distances for the 1,408 input trees to the pruned supertree. (B) Distances for a corresponding set of 1,408 randomly grown trees, with the same size distribution as the original dataset to the supertree. (C) Distances for 1,408 randomly grown trees versus a supertree constructed from these random trees. (D) Distances from the original 1,408 input trees to the supertree constructed from the random trees in C.

2.3.6 Minimum evolution tree of concatenated data:

Figure 2.6 displays the tree recovered using the minimum evolution criterion for a concatenated alignment of the same 1,408 single-gene families used for the construction of the supertree. Minimum evolution was used instead of maximum likelihood due to the length of the alignment (1,537,155 bases). The concatenated data tree shows strong support for the majority of the nodes in the tree. Weak support is present only towards the base of the *Yersinia* clade and at the node separating the *Salmonella* clade with the *Escherichia*/*Shigella* clade.

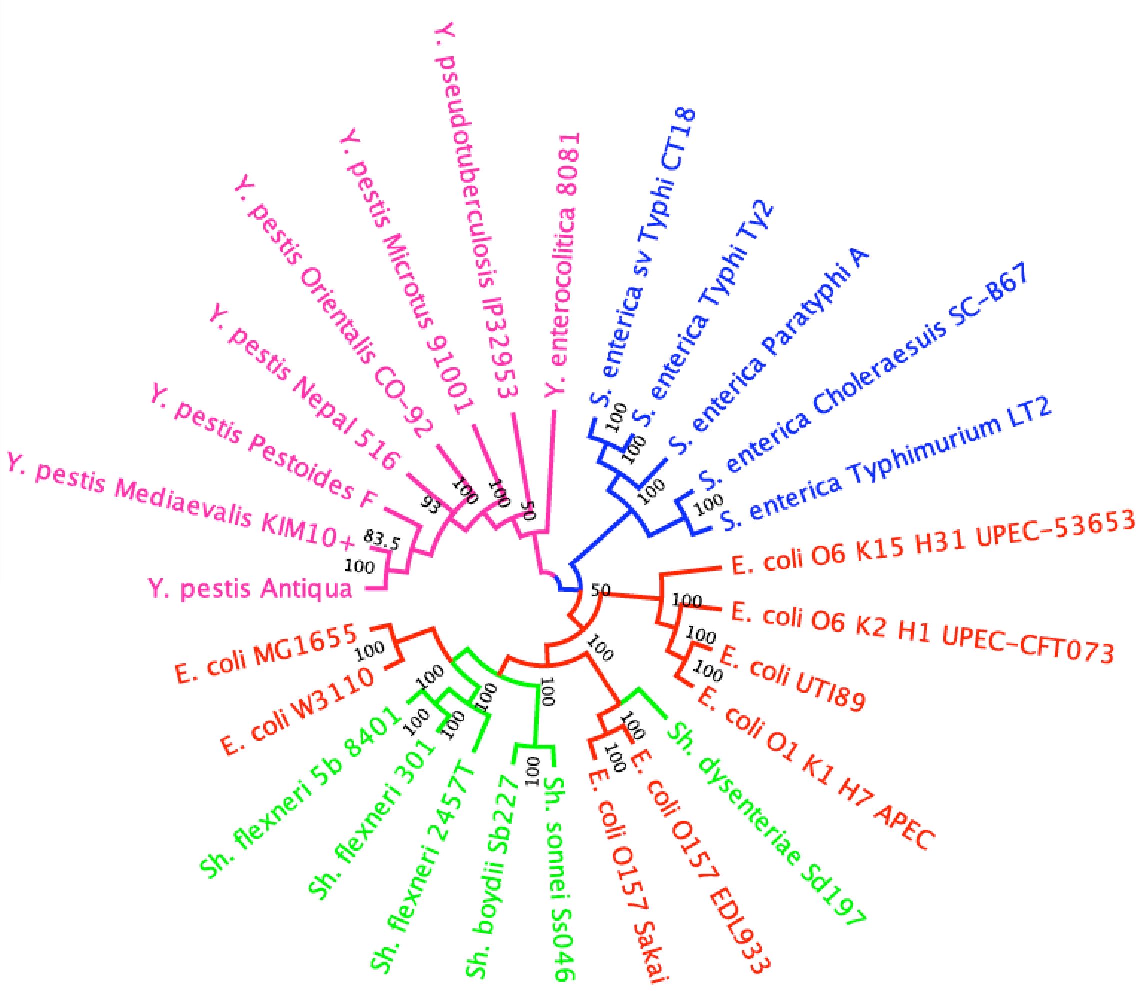


Figure 2.6: Minimum evolution tree built from an alignment of 1,408 single-gene families.

2.4 Discussion

This study examined a set of taxa that are known to be closely-related and where there is some confusion over whether there is a total of three or four valid genera within the group. This test dataset is emblematic of the issues that crop up in employing genome-scale data to answer questions concerning the evolutionary history of prokaryotes.

Three groups were consistently recovered from the analysis, irrespective of the method chosen to infer phylogenetic relationships. These were the *Yersinia* group, the *Salmonella* group and the *Escherichia/Shigella* group. As these groups were recovered consistently it implies that the effects of HGT between genera was not a major factor, or at least whatever HGT has occurred between the three groups has not been substantial enough to overcome the underlying signal. A single origin of *Shigella* (Escobar-Paramo et al., 2003) was not observed, rather, the data suggests multiple origins, in accordance with the findings of Pupo *et al.,* 2001. The three groups were not found to be a single homogenous entity; partitions existed. There were clear boundaries and none of the analysis methods broke these boundaries. These boundaries did not fully agree with traditional classification methods, in that *Escherichia/Shigella* were considered a single group. This result was in line with expectations (Pupo et al., 2000).

Looking within each of the groups, the story is clearly somewhat different. There were very few recurrent themes across different analyses and different datasets, with weak bootstrap value present for many of the internal relationships. Unlike Ochman *et al.,* 2005, the analysis was not limited to only those gene families that were found in all genomes; all gene-families were analysed, even those with a patchy distribution. One of the weak features of this kind of analysis is the sampling issue. Having relatively few fully sequenced genomes to work with means that the analysis is unlikely to have probed the boundaries of the groups as they are depicted in the figures in this study.

The 16S phylogeny produced a result that might be considered contrary to expectations. Only one taxon, *Salmonella enterica* paratyphi A, had a corresponding clade of all seven 16S genes. This clade was strongly supported, implying homogenisation of the 16S genes within this lineage. However, in general, homogenisation of all 16S sequences within a genome was not complete. In fact, multiple clades place copies of the 16S from different strains together with strong bootstrap support. As a result, depending on the copy of the 16S gene used when constructing a tree for all 27 genomes, different phylogenies, often with significantly conflicting topologies, can be produced. This finding is in line with previous studies of the homogenisation of 16S rRNA genes (Cilia et al., 1996). One possible reason for this is that homogenisation is not fast enough that all copies of the sequence are the same in each genome. Another speculation is that there has been recombination between strains and this is the reason for the absence of within-genome monophyly. Given the lack of resolution on the tree due to high sequence similarity and the fact that the predefined barriers within each group are somewhat arbitrary, it would not be counter-intuitive to think of recombination and homogenisation as occurring within each group as a whole, not as a process that is solely restricted to each individual strain. Indeed, similar results have been seen in Helicobacteria pylori, where recombination has been found to be frequent between unrelated strains (Falush et al., 2001). The three major groups are recovered on this 16S rRNA tree and this suggests that either homogenisation is rapid enough to avoid the intermingling of sequences across the three major groups, or that sequence divergence has been sufficient that homologous recombination is much less frequent across the genome-species divide.

When examining the results of concatenating the sequences of *atpD*, *gyrB* and *trpB* the same three major groups are recovered in each tree, but the internal relationships differ significantly (as judged by a number of tests using CONSEL) from tree to tree. In fact there is little to no agreement over the internal relationships of the groups. Concatenating the data and reconstructing a representative phylogeny produces a result that is a mixture of the information contained in three conflicting topologies, but it is not clear what this tree means and in fact, may suggest that it is meaningless. This kind of approach has been used previously to assess congruence with 16S rRNA phylogenies in a large number of *Streptomyces* and it has been reported that the results were "obviously superior to the 16S rRNA gene tree in both resolution power and topological stability" (Guo et al., 2008). The tree that is recovered from this concatenated alignment has low bootstrap support and this reflects the fact that the individual trees have conflicting histories. As an approach to understanding the evolutionary history of the YESS group, this method seems to be ambiguous.

Both the 1,408 gene nucleotide-based supertree and the minimum evolution tree of the concatenated nucleotide data fare much better with regard to support for the hypotheses that they display. The trees agree completely in terms of the relationships for the *Salmonella* clade and only minor differences exist in the *Escherichia/Shigella* clade with the position of *Sh. sonnei* and the relationships between the three *Sh. flexneri* strains changing between the two trees. It should be noted that even though the differences are minor, they receive strong support in both trees. The major area of difference between the trees is in the *Yersinia* clade. The supertree shows weak support for some of the internal relationships while the minimum evolution tree shows strong support for all the relationships bar the split between *Y. enterocolitica* and the rest of the clade. Do these trees have more meaning than the trees from the 16S rRNA gene or the housekeeping genes? This is a difficult question to answer and remains an open question.

What does this tell us about the YESS group? Surely this is not a uniquely difficult group to analyse, yet after a thorough examination of the data, apart from concluding that there are three, not four, major groups, there are as many questions as when the analysis started. It has been previously argued that a tree like phylogeny may exist only at the tips for prokaryotes and that the deeper branches may remain a mystery (Creevey el al., 2004). This study suggests that at the tips it may be impossible to derive a reliable phylogeny using current methods.

An important point to note is that this analysis is specific to the YESS group. It cannot be guaranteed that the same can be said about other bacterial groups. Despite the fact that there are now more than 1,800 sequenced prokaryotic genomes, sampling is still patchy, usually being driven by medical or economic factors and therefore, is perhaps not representative of most species. There may be phylogenetic bias in the collection of organisms being analysed. If in some cases we have sequenced closely related strains and in others we have sampled more distant strains, this can have an effect on our estimates of recombination rate and population structure.

Assessing deep-level phylogenetic relationships is fraught with difficulties related to HGT, hidden paralogy, model misspecification and erosion of phylogenetic signal, however, assessing shallow relationships is no less difficult.

In the next chapter I present a software tool for identifying gene clusters in bacterial genomes and use it to examine the chromosomal organisation of amino acid biosynthesis genes, covering not only the *YESS* group but also many other γ-proteobacteria.

Chapter 3 - GenClust: a software tool for identifying, analysing and visualising gene clusters

3.1 INTRODUCTION

The clustering of functionally related genes is a widespread feature of bacterial genomes (Lawrence and Roth, 1996; Olivera et al., 1998; Rison et al., 2002). As such there is considerable interest in identifying gene clusters that are conserved across multiple genomes, and various software tools exist that allow the user to achieve this (Rutherford et al., 2000; Schmidt and Stoye 2008; Jensen et al. 2009). However in there is a lack of flexibility in the existing software, with limitations on the number of genomes examined at one time (Rutherford et al., 2000), the ease of use of the software (Schmidt and Stoye 2008) and the ease of mining the relevant data for such clusters (Jensen et al. 2009). While the existing software all have their own merits, they lack a balance between usability and flexibility.

Here I introduce a software tool, GenClust, which attempts to combine usability and flexibility. GenClust is written in PERL (http://www.perl.org/) and designed for use with Unix based operating systems. The user needs only PERL and the BLAST package (Altschul et al., 1990) installed in order to run GenClust. GenClust takes a set of query sequences, i.e. a set of sequences that the user wishes to examine for evidence of clustering, and a set of bacterial genomes in Genbank full format (Benson et al., 2008). GenClust uses a BLAST based method to identify clusters of genes that are homologous to the query sequences provided by the user. GenClust provides a detailed analysis of cluster content including per-gene cluster frequency, pairwise cluster co-occurrence between genes, average cluster size with and without non-query sequences, a homologs to clusters ratio and cluster size versus frequency data. An Adobe Illustrator file (Adobe Systems, San José, California) showing a visual representation of all clusters in the form of arrow diagrams is also produced.

In order to provide flexibility in the software, the user has control over two important parameters. The first is the number of intervening genes (IGs) allowed. IGs are defined as genes that are not part of the users set of query genes (see figure 3.1). By varying this parameter the user can affect the tightness of clustering in the result files, by setting this number to 0 for example, the software would only retrieve clusters that did not contain any IGs. Raising this number decreases the tightness of the identified clusters, setting the parameter to 5 would allow anywhere between 0 and 5 IGs to be present between any two consecutive query genes. The second parameter the user can modify is the e-value used in the BLAST search. Making the value smaller increases the strictness of the search, helping ensure only highly significant hits are considered when identifying potential clusters. Conversely if the user values positional evidence of clustering over strict sequence similarity levels, then they may raise the e-value and thus potentially identify more distantly related clusters. By combining the values entered into these two parameters the user has a high level of control over what is identified as a gene cluster.

To demonstrate the robustness of GenClust, a sample data set of 180 γ-proteobacterial genomes was downloaded from Genbank. A set of query sequences was then downloaded

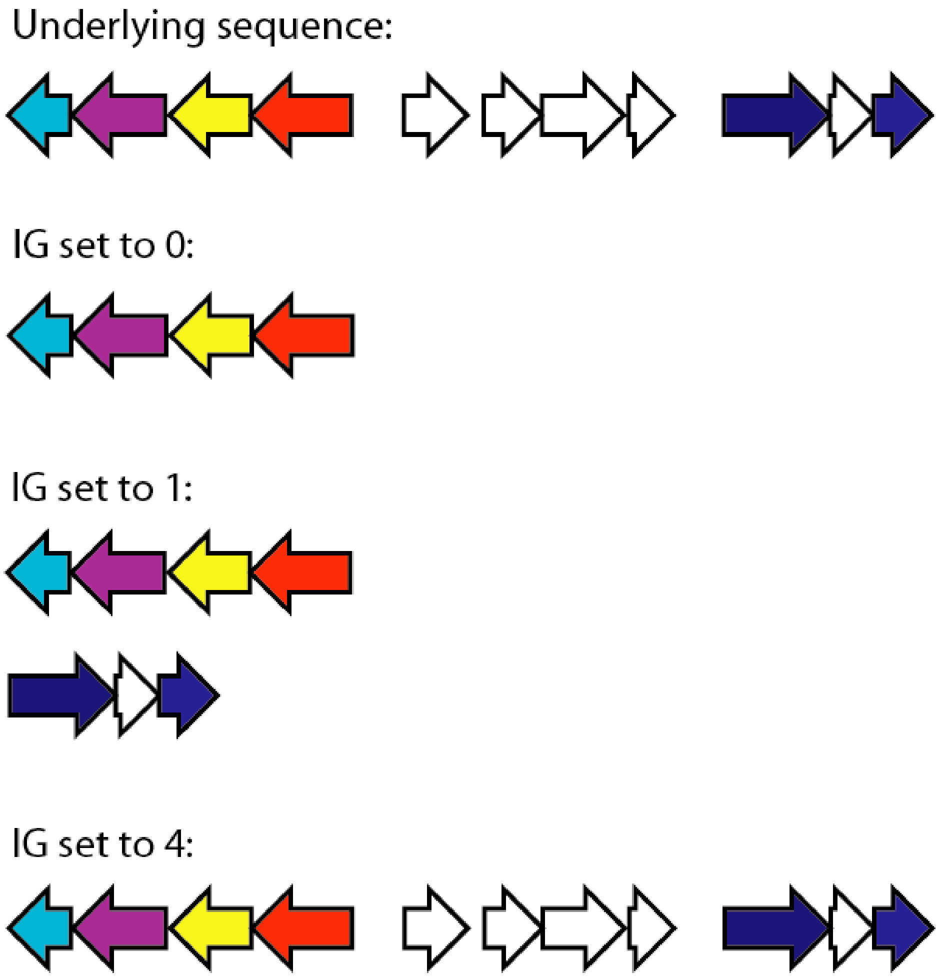


Figure 3.1: The IG parameter. The top cluster is a hypothetical set of clustered genes. Coloured genes are part of the query set provided by the user. White genes are IGs. Setting IG to 0 returns only returns the four tightly clustered genes on the left. Setting it to 1 returns two clusters, as a single IG separates the two coloured genes on the right hand side of the top cluster. Setting it to four recovers the full cluster as no more that four IGs separate any two consecutive coloured genes in the top cluster.

from Ecocyc (Keseler et al., 2005). These query sequences represented five different superpathways (SPs) of amino acid biosynthesis: aspartate and asparagine biosynthesis; leucine, valine, and isoleucine biosynthesis; lysine, threonine and methionine biosynthesis; phenylalanine, tyrosine, and tryptophan biosynthesis and serine and glycine biosynthesis. The number of genes involved in each SP ranged from four to twenty-one (see table 3.1 for more details).

These SPs of amino acid biosynthesis were selected for their diversity in terms of number of genes present and whether or not they were clustered in *Escherichia coli* K12. Because of the importance of the genes involved in these pathways, it was assumed they would be widely distributed and highly conserved across the γ-proteobacteria.

A large-scale analysis of each SP was carried out to determine how often the associated genes are found clustered across the γ-proteobacteria. In particular the results were examined to see if homologous clusters conformed to strict ancestral structures and whether any of the genes under observation showed no evidence of clustering or even avoidance of clustering. GenClust code, its manual and sample data can be found in S. I. 3.1.

Superpathway	**Total genes**	**Gene names**	**Clustered in *E. coli* K12**
asn, asp	7	*ansA, ansB, asnA, asnB, aspA, aspC, iaaA*	Not clustered
gly, ser	4	*glyA, serA, serB, serC*	Not clustered
ile, leu, val	16	*ilvA, ilvB, ilvC, livD, ilvE, ilvG1, ilvG2, ilvH, ilvI, ilvM, ilvN, leuA, leuB, leuC, leuD, tyrB*	Multiple clusters
lys, met, thr	21	*argD, asd, aspC, dapA, dapB, dapD, dapE, dapF, lysA, lysC, malY, metA, metB, metC, metE, metH, metK, metL, thrA, thrB, thrC*	Multiple clusters
phe, trp, tyr	20	*aspC, aroA, aroB, aroC, aroD, aroE, aroF, aroG, aroH, aroK, aroL, pheA, trpA, trpB, trpC, trpD, trpE, tyrA, tyrB, ydiB*	Multiple clusters

Table 3.1: Superpathways of amino acid biosynthesis in *E. coli* K12. The number of genes involved in each SP ranges from four to twenty-one. Two of the SPs have no corresponding gene clusters, while the remaining three are associated with multiple cluster structures.

3.2 Materials and methods

3.2.1 γ-proteobacterial data set:

180 γ-proteobacterial genomes were downloaded from the Genbank ftp server (ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria), see supplementary data 3.1 for a complete list. All files used in the analysis were in .gbk file format and corresponded to completely sequenced genomes.

3.2.2 Input file parsing:

GenClust was used to parse all relevant information from the .gbk files. This included retrieving all amino acid sequence data along with the corresponding nucleotide sequences. Further information corresponding to the start and end locations of the genes, along with their orientation was also stored for subsequent steps in the analysis.

3.2.3 Superpathways of amino acid biosynthesis:

The Ecocyc database was used to retrieve the amino acids sequences for the genes involved in each of the five SPs under consideration. For each SP, the corresponding amino acid sequences were downloaded in fastA format and stored individually in a query file.

3.2.4 Input files and parameters for GenClust analysis:

A ‘genomes.txt’ file is required to run GenClust, containing a list of all the .gbk file names and a corresponding alphanumeric abbreviation for each genome (used for ensuring unique names for each sequence during the BLAST phase). This is in addition to having all query files and genome files in the working directory (see S. I. 3.1 for a manual along with code and sample data). As there were five SPs in the analysis, each with its own set of query files, the analysis was run in five different directories. Each directory contained a set of query files corresponding to one of the SPs. The only difference in the five runs of GenClust was the set of query files in each directory. For simplicity’s sake the remainder of the discussion will refer to a single run of GenClust. The e-value parameter was set to e^{-5}. The IG parameter was set to 5. All subsequent steps in the analysis were carried out automatically by GenClust.

3.2.5 Database creation and BLAST search of query sequences:

All amino acid sequences coded by the 180 γ-proteobacterial genomes were concatenated into a single amino acid database. This database contained a total of 636,813 amino acid sequences. A BLAST analysis was carried out for each query file against the database, creating one BLAST output file per query sequence containing hits of all potential homologous sequences with an e-value less than or equal to e^{-5}.

3.2.6 Cluster identification and merging:

The genomic location of each sequence in the BLAST output files was compared to the location of all other sequences in all BLAST output files for a given SP. If two genes were found to fulfill the condition of the IG parameter, i.e. they were separated by no more than five genes that were non-homologous to any query sequences, then the genes were considered a linked pair. For each genome, all such linked pairs were identified and stored in a list. After this each pair was considered to be a single entity and the process was repeated to find sets of linked pairs that overlapped. Once all overlapping pairs were identified, they were merged and any redundant information (genes that appeared in two different pairs) was discarded. This process ensured the largest possible clusters, in accordance with the IG parameter, were always identified.

3.2.7 Parsing of genomic information and result files:

The genomic information for each cluster was retrieved from the files created during the input file parsing stage. This information is required for the visualisation of the clusters. Each gene in a cluster is drawn as an arrow diagram, with the direction of the arrow corresponding to the direction of transcription. Each gene in the cluster is scaled relative to its length in nucleotides. Non-coding sequence between genes are scaled in the same way. The colouring of each arrow corresponds to the query sequence to which it is homologous. IGs appear as white arrows. A legend is created at the bottom of the visualisation to show what colour represented each query sequence. For each SP, this information was output as an Adobe Illustrator format document. In cases where the number of clusters was too great to fit inside the maximum bounding box size of an Adobe Illustrator document (15000x15000 pixels), multiple output files were created for the pathway.

Multiple output files are created by GenClust for each run. These include a file called ‘analysisresults.txt’, which contains a variety of data on all the clusters identified, a file called ‘cooccurrences.txt’, containing pairwise cluster co-occurrence frequencies for all genes involved in the analysis, and ‘clusterfrequency.txt’, which contains the number of times each gene occurred in a cluster. In addition to these output files, the corresponding sequences are retrieved for each identified cluster and placed in fastA format files.

3.3 RESULTS

3.3.1 Runtime analysis of algorithm:

The total runtime in hours for GenClust can be seen in figure 3.2. The runtime analysis was carried out on three different datasets. The first dataset contained all 180 genomes, the second contained 100 genomes and the final dataset contained 50 genomes. The runtime for each dataset is plotted in figure 3.2, giving three points per SP, in order of increasing number of genomes. In each case the BLAST phase of the analysis was inconsequential in terms of overhead, with almost all computation time occurring at the cluster identification and assembly stage. For all SPs increase in runtime versus total BLAST hits followed a polynomial curve.

3.3.2 BLAST searches and cluster data:

BLAST searches carried out using the genes associated with each SP resulted in over a thousand hits per pathway (table 3.2). In general the number of genes involved in the SP correlated well with the total number of BLAST hits retrieved, with the exception of the *gly* and *ser* SP, which had a considerably higher than expected number of hits relative to the number of genes involved. This is attributed to multiple duplications of *glyA* homologs across the data.

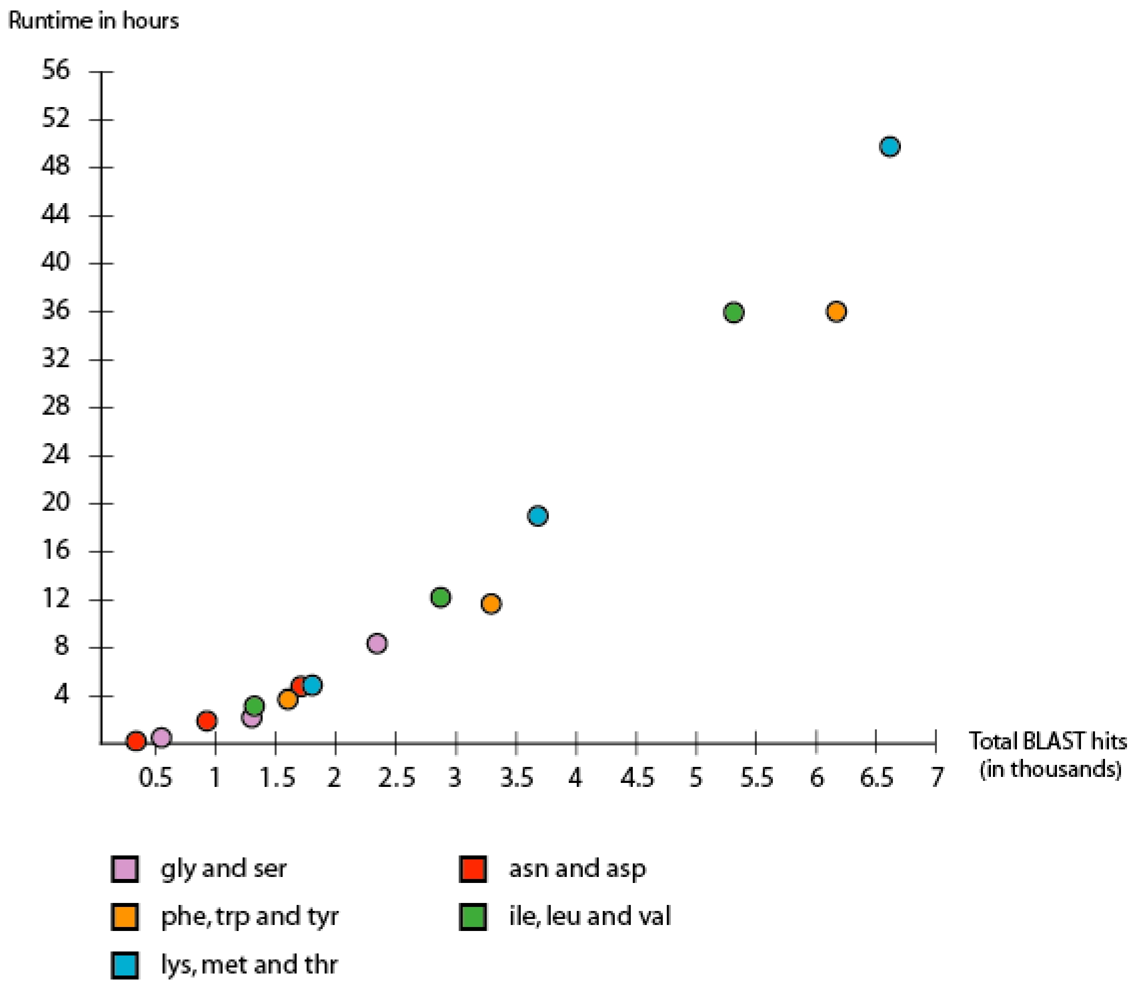


Figure 3.2: Runtime data for varying numbers of genomes. Each SP has three points, corresponding in order to the 50, 100 and 180 genome datasets. Tests run on Intel(R) Xeon(R) CPU E7340 2.40GHz processor with 4 gigabytes of RAM available.

SP	Total BLAST hits	Total clusters	Average genes without IGs	Average genes with IGs	Genomes without clusters
asn, asp	1,757	38	2	3.74	144
gly, ser	2,457	80	2.04	3.31	109
ile, leu, val	5,301	431	3.74	4.58	18
lys, met, thr	6,676	594	2.25	3.17	9
phe, trp, tyr	6,210	663	2.70	3.39	8

Table: 3.2: Numerical analysis of gene cluster data. This table shows BLAST and cluster data for the 180 γ-proteobacterial dataset. 'Total BLAST hits' refers to the total number of matches to all query sequences for a particular SP. 'Total clusters' is the number of clusters identified per SP. 'Average genes without IGs' is the average number of genes per cluster discounting IGs. 'Average genes with IGs' is the average number of genes per cluster with IGs included in the calculation. 'Genomes without clusters' is the number of genomes where no clusters were found for a particular SP

The total number of clusters per genome correlated strongly with the total BLAST hits. However in terms of the ratio of clusters to BLAST hits, two different patterns were observed. The two smaller SPs, the *asn* and *asp* SP and the *gly* and *ser* SP, had a ratio of 0.022 and 0.033 respectively. This contrasted with the three larger SPs, with ratios of 0.081, 0.089 and 0.11. This was not altogether unexpected, the more genes per SP, the more possibilities available for clustering. This is contradicted by the ratio for the *gly* and *ser* SP being higher than that of the *asn* and *asp* SP, but again the contradiction is explainable by high levels of duplication of *glyA* homologs, with tandem duplications providing instant clusters and accounting for a large proportion of the gene clusters of the *gly* and *ser* SP. Factoring out these duplications brings the numbers back in line with expectations based on gene number.

The average number of genes per cluster, excluding and including IGs, can be seen in figure 3.3 and figure 3.4 respectively. Note that when referring to excluding and including IGs, this only indicates whether or not the IGs were excluded or included in calculations, not that the clusters themselves excluded or included IGs. The clusters were identical for both sets of calculations, always allowing for the inclusion of IGs, with the maximum number of consecutive IGs set to five.

The *asn* and *asp* SP had the smallest average cluster size. No clusters of more that two of the genes involved in the SP were found, giving an average of exactly two genes per cluster, the minimum number of genes possible for a cluster to exist. If there is a barrier to larger clusters for these genes, the nature of that barrier is unclear. When IGs were

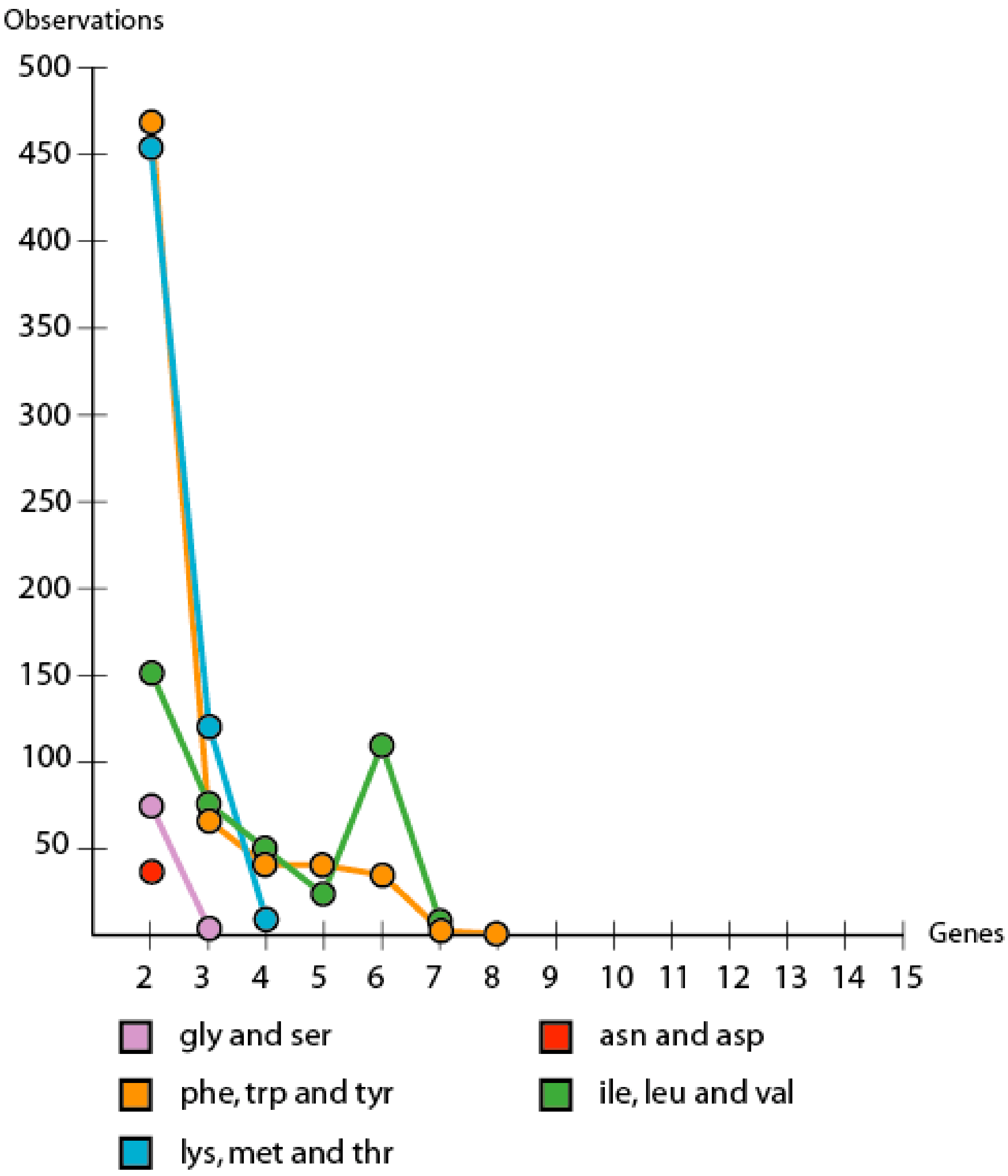


Figure 3.3: Cluster size versus frequency without IGs. The x-axis denotes the number of functionally related genes in a cluster. The y-axis represents the number of observations in the data.

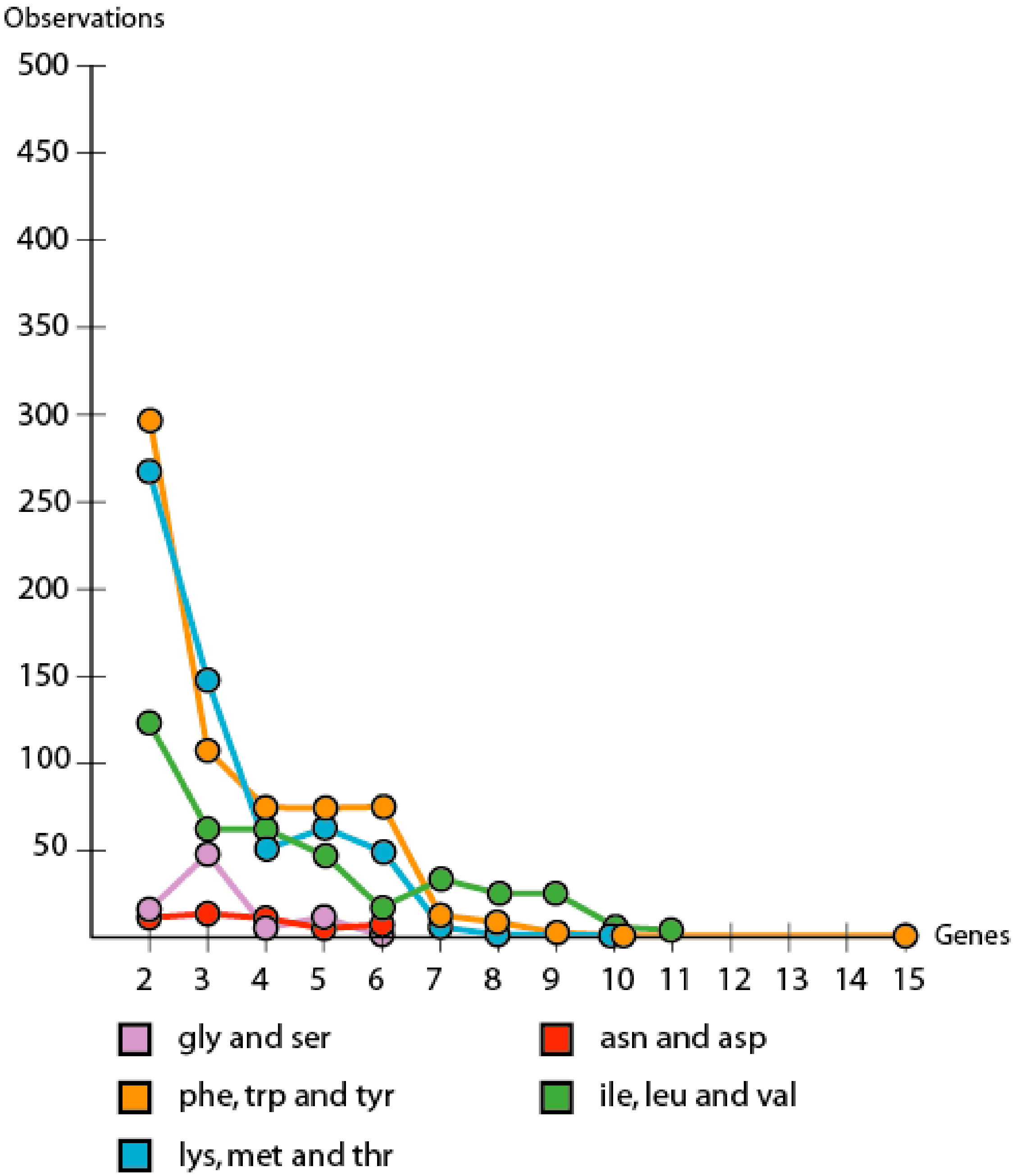


Figure 3.4: Cluster size versus frequency with IGs. The x-axis denotes the total genes, including IGs, in a cluster. The y-axis represents the number of observations in the data.

included in calculating average cluster size this figure changed to 3.74. The size of clusters including IGs ranged from two to six genes.

The *gly* and *ser* SP clusters averaged 2.04 genes per cluster. With only four genes involved in the SP and duplications of *glyA* homologs accounting for the majority of observed clusters, this number was in line with expectations. Three clusters of three genes were observed, and these appear to be a lineage specific further duplication of the *glyA* gene. Adding in IGs changed the average size to 3.31 and the maximum cluster size was six genes.

The *ile, leu* and *val* SP differed from all other SPs, with a significantly higher number of average genes per cluster at 3.74. Large clusters of up to six genes were common with 114 six-gene clusters and 8 seven gene clusters identified. Including IGs the average genes per cluster grew to 4.58 and the maximum cluster size was eleven genes.

The *lys, met* and *thr* SP had an average of 2.25, with the majority of clusters containing two genes. The maximum cluster size was four genes for this SP, observed eleven times. With IGs the average genes per cluster was 3.17 and the maximum cluster size was ten.

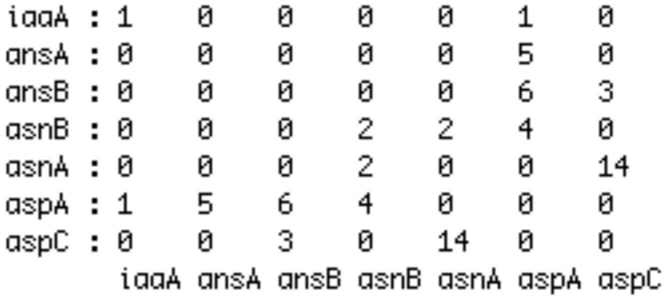
The *phe, trp* and *tyr* SP showed a slightly higher than expected level of clustering, with an average of 2.7 genes per cluster. This was due to the fact that many larger clusters were identified, with 39 clusters of six genes and 2 clusters of seven genes present in the dataset. Adding in the IGs the average cluster size was 3.39 and the maximum cluster size was 15. This was the largest cluster found in the data, in the genome of *Candidatus blochmannia floridanus*. In this cluster the *trpA, trpB, trpC, trpD* and *trpE* genes are flanked on either side by the *aspC* and *trpA* genes along with a number of IGs.

The absence of clusters in particular genomes correlated strongly with total clusters observed per SP. This implies an even spread of clusters across the dataset, as opposed to certain genomes showing high levels of clustering. The *asn* and *asp* SP, with the least number of identified clusters of all the SPs also had the highest number of genomes without an identifiable cluster. Conversely, the *phe, trp* and *tyr* SP, which had the highest number of identified clusters, had the lowest number of genomes without an associated cluster.

3.3.3 Cluster co-occurrence data:

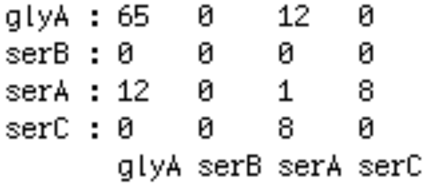
GenClust outputs a pairwise co-occurrence matrix for each SP. Each row and column represents a gene. The value in a cell represents the number of times a particular pair of genes was found in a cluster (the genes did not need to be side-by-side, merely both present). The diagonal of the matrix represents the number of times each gene was found in a cluster with a duplicate copy of itself. As such each matrix is symmetrical.

The output files for all five SPs can be seen in tables 3.3-3.7. For the asn and asp SP all genes were found in at least one cluster. The least commonly clustered gene was *iaaA*. The most frequently clustered pair of homologs was *asnA* and *aspC*, found together in a



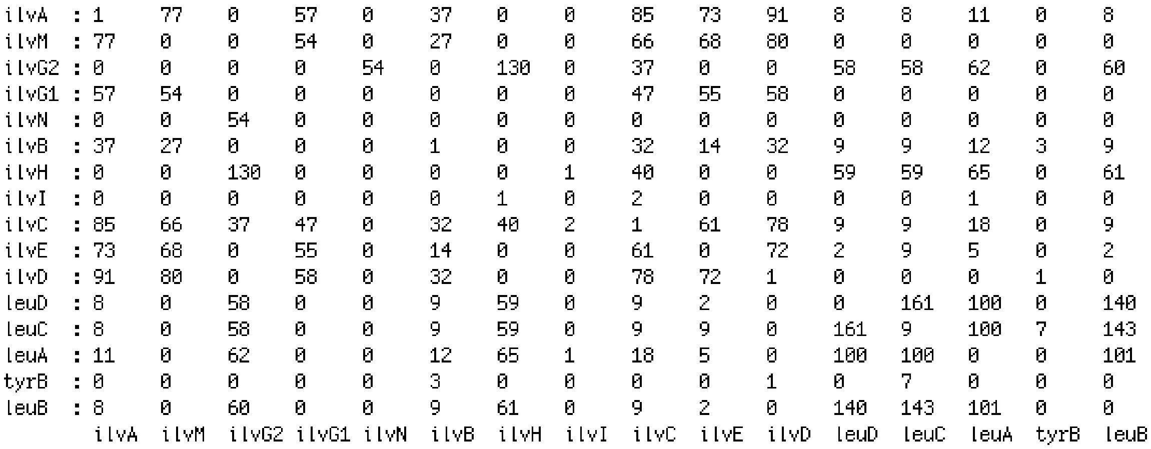
```
iaaA : 1    0    0    0    0    1    0
ansA : 0    0    0    0    0    5    0
ansB : 0    0    0    0    0    6    3
asnB : 0    0    0    2    2    4    0
asnA : 0    0    0    2    0    0    14
aspA : 1    5    6    4    0    0    0
aspC : 0    0    3    0    14   0    0
       iaaA ansA ansB asnB asnA aspA aspC
```

Table 3.3: Pairwise cluster co-occurrence data for the *asn* and *asp* SP.



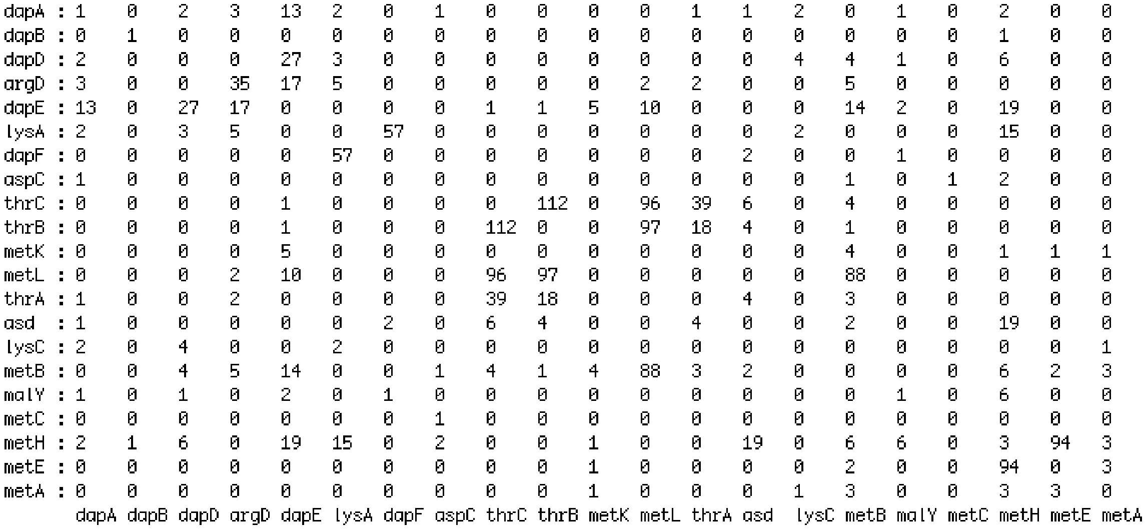
```
glyA : 65   0    12   0
serB : 0    0    0    0
serA : 12   0    1    8
serC : 0    0    8    0
       glyA serB serA serC
```

Table 3.4: Pairwise cluster co-occurrence data for the *gly* and *ser* SP.



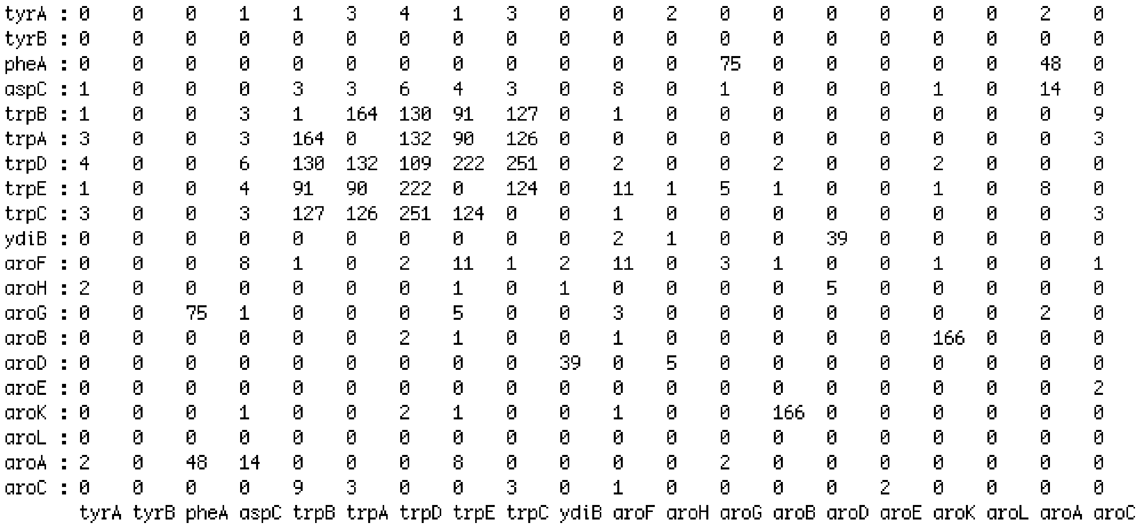
```
ilvA  : 1     77    0     57    0     37    0     0     85    73    91    8     8     11    0     8
ilvM  : 77    0     0     54    0     27    0     0     66    68    80    0     0     0     0     0
ilvG2 : 0     0     0     0     54    0     130   0     37    0     0     58    58    62    0     60
ilvG1 : 57    54    0     0     0     0     0     0     47    55    58    0     0     0     0     0
ilvN  : 0     0     54    0     0     0     0     0     0     0     0     0     0     0     0     0
ilvB  : 37    27    0     0     0     1     0     0     32    14    32    9     9     12    3     9
ilvH  : 0     0     130   0     0     0     0     1     40    0     0     59    59    65    0     61
ilvI  : 0     0     0     0     0     0     1     0     2     0     0     0     0     1     0     0
ilvC  : 85    66    37    47    0     32    40    2     1     61    78    9     9     18    0     9
ilvE  : 73    68    0     55    0     14    0     0     61    0     72    2     9     5     0     2
ilvD  : 91    80    0     58    0     32    0     0     78    72    1     0     0     0     1     0
leuD  : 8     0     58    0     0     9     59    0     9     2     0     0     161   100   0     140
leuC  : 8     0     58    0     0     9     59    0     9     9     0     161   9     100   7     143
leuA  : 11    0     62    0     0     12    65    1     18    5     0     100   100   0     0     101
tyrB  : 0     0     0     0     0     3     0     0     0     0     1     0     7     0     0     0
leuB  : 8     0     60    0     0     9     61    0     9     2     0     140   143   101   0     0
        ilvA  ilvM  ilvG2 ilvG1 ilvN  ilvB  ilvH  ilvI  ilvC  ilvE  ilvD  leuD  leuC  leuA  tyrB  leuB
```

Table 3.5: Pairwise cluster co-occurrence data for the *ile*, *leu* and *val* SP.



dapA :	1	0	2	3	13	2	0	1	0	0	0	0	1	1	2	0	1	0	2	0	0
dapB :	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
dapD :	2	0	0	0	27	3	0	0	0	0	0	0	0	0	4	4	1	0	6	0	0
argD :	3	0	0	35	17	5	0	0	0	0	0	2	2	0	0	5	0	0	0	0	0
dapE :	13	0	27	17	0	0	0	0	1	1	5	10	0	0	0	14	2	0	19	0	0
lysA :	2	0	3	5	0	0	57	0	0	0	0	0	0	0	2	0	0	0	15	0	0
dapF :	0	0	0	0	0	57	0	0	0	0	0	0	0	2	0	0	1	0	0	0	0
aspC :	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	2	0	0
thrC :	0	0	0	0	1	0	0	0	0	112	0	96	39	6	0	4	0	0	0	0	0
thrB :	0	0	0	0	1	0	0	0	112	0	0	97	18	4	0	1	0	0	0	0	0
metK :	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	4	0	0	1	1	1
metL :	0	0	0	2	10	0	0	0	96	97	0	0	0	0	0	88	0	0	0	0	0
thrA :	1	0	0	2	0	0	0	0	39	18	0	0	0	4	0	3	0	0	0	0	0
asd :	1	0	0	0	0	0	2	0	6	4	0	0	4	0	0	2	0	0	19	0	0
lysC :	2	0	4	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
metB :	0	0	4	5	14	0	0	1	4	1	4	88	3	2	0	0	0	0	6	2	3
malY :	1	0	1	0	2	0	1	0	0	0	0	0	0	0	0	0	1	0	6	0	0
metC :	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
metH :	2	1	6	0	19	15	0	2	0	0	1	0	0	19	0	6	6	0	3	94	3
metE :	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	2	0	0	94	0	3
metA :	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	3	0	0	3	3	0
	dapA	dapB	dapD	argD	dapE	lysA	dapF	aspC	thrC	thrB	metK	metL	thrA	asd	lysC	metB	malY	metC	metH	metE	metA

Table 3.6: Pairwise cluster co-occurrence data for the *lys, met* and *thr* SP.



```
tyrA : 0    0    0    1    1    3    4    1    3    0    0    2    0    0    0    0    0    0    2    0
tyrB : 0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
pheA : 0    0    0    0    0    0    0    0    0    0    0    0    75   0    0    0    0    0    48   0
aspC : 1    0    0    0    3    3    6    4    3    0    8    0    1    0    0    0    1    0    14   0
trpB : 1    0    0    3    1    164  130  91   127  0    1    0    0    0    0    0    0    0    0    9
trpA : 3    0    0    3    164  0    132  90   126  0    0    0    0    0    0    0    0    0    0    3
trpD : 4    0    0    6    130  132  109  222  251  0    2    0    0    2    0    0    2    0    0    0
trpE : 1    0    0    4    91   90   222  0    124  0    11   1    5    1    0    0    1    0    8    0
trpC : 3    0    0    3    127  126  251  124  0    0    1    0    0    0    0    0    0    0    0    3
ydiB : 0    0    0    0    0    0    0    0    0    0    2    1    0    0    39   0    0    0    0    0
aroF : 0    0    0    8    1    0    2    11   1    2    11   0    3    1    0    0    1    0    0    1
aroH : 2    0    0    0    0    0    0    1    0    1    0    0    0    0    5    0    0    0    0    0
aroG : 0    0    75   1    0    0    0    5    0    0    3    0    0    0    0    0    0    0    2    0
aroB : 0    0    0    0    0    0    2    1    0    0    1    0    0    0    0    0    166  0    0    0
aroD : 0    0    0    0    0    0    0    0    0    39   0    5    0    0    0    0    0    0    0    0
aroE : 0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    2
aroK : 0    0    0    1    0    0    2    1    0    0    1    0    0    166  0    0    0    0    0    0
aroL : 0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
aroA : 2    0    48   14   0    0    0    8    0    0    0    0    2    0    0    0    0    0    0    0
aroC : 0    0    0    0    9    3    0    0    3    0    1    0    0    0    0    2    0    0    0    0
       tyrA tyrB pheA aspC trpB trpA trpD trpE trpC ydiB aroF aroH aroG aroB aroD aroE aroK aroL aroA aroC
```

Table 3.7: Pairwise cluster co-occurrence data for the *phe*, *trp* and *tyr* SP.

cluster 14 times. In general clustering of pairs of genes in this SP was infrequent and out of 28 possible gene pairs (including a gene being paired with a duplicate copy of itself) only 9 were observed.

In the *gly* and *ser* SP the most frequent pair was a *glyA* homolog clustered with another copy of itself. This was observed 65 times. *serB* was never found in a cluster. The remaining two genes clustered infrequently. Out of 10 possible pairs, 4 were observed.

Many genes showed a strong tendency to co-occur together in the *ile, leu* and *val* SP. For example *leuA, leuB, leuC* and *leuD* all frequently co-occurred with one another. *ilvG2* and *ilvH* were also found together regularly, the pair were found in 130 different clusters together. 69 pairs out of a possible 128 combinations were observed. Homologs of all genes were found in one or more clusters.

Similarly, in the data for the *lys, met* and *thr* SP all genes were found in one or more clusters, with the *metC* gene clustering least frequently (only one co-occurrence observations with other genes in the SP in total). The most common combinations were *thrB* and *thrC*, observed 112 times, *thrB* and *metL*, observed 97 times and *thrC* and *metL*, observed 96 times. This implies that these three genes are likely to co-occur frequently. 67 out of 231 possible combinations were observed.

The *phe, trp* and *tyr* SP had two genes, *tyrB* and *aroL*, that were never found in a cluster despite having a total of 182 and 234 matching hits repsectively during the BLAST phase. Frequently co-occurring pairs included all combinations of the *trpA, trpB, trpC, trpD* and *trpE* genes with each other. For this set of genes the lowest number of observed cluster co-occurrences was for the *trpA* and *trpE*, which co-occurred 90 times. The highest co-occurrence was for *trpC* and *trpD*, which co-occurred 251 times. The combination of the *aroB* and *aroK* gene was also frequent, co-occurring 166 times. 53 of a possible 210 combinations were observed.

3.3.4 Cluster content and structure:

GenClust generated a visual representation of each gene cluster. This allowed for rapid analysis by eye of cluster content and structure. The *asn* and *asp* SP (figure 3.5 and supplementary data 3.3), as previously described, only has clusters of two of the seven genes associated with the SP. Several combinations of genes exist, though they represent a relatively small fraction of all potential pairs. In general there was only a single cluster in each of the in the 36 genomes containing a cluster for this SP, with the exception of *Hahella chejuensis* KCTC 2396 which contained three clusters, two of which appear to be gene duplications. A common cluster was homologs of *asnA* and *aspC* separated by several IGs; this structure was present in *Acinetobacter* sp ADP1, *H. chejuensis* KCTC 2396, some strains of the *Salmonella* lineage and *Serratia proteamaculans* 568. Though the pairing of *asnA* and *aspC* homologs is highly conserved in the *Salmonella* lineage, the actual layout of the IGs appears to differ from strain to strain.

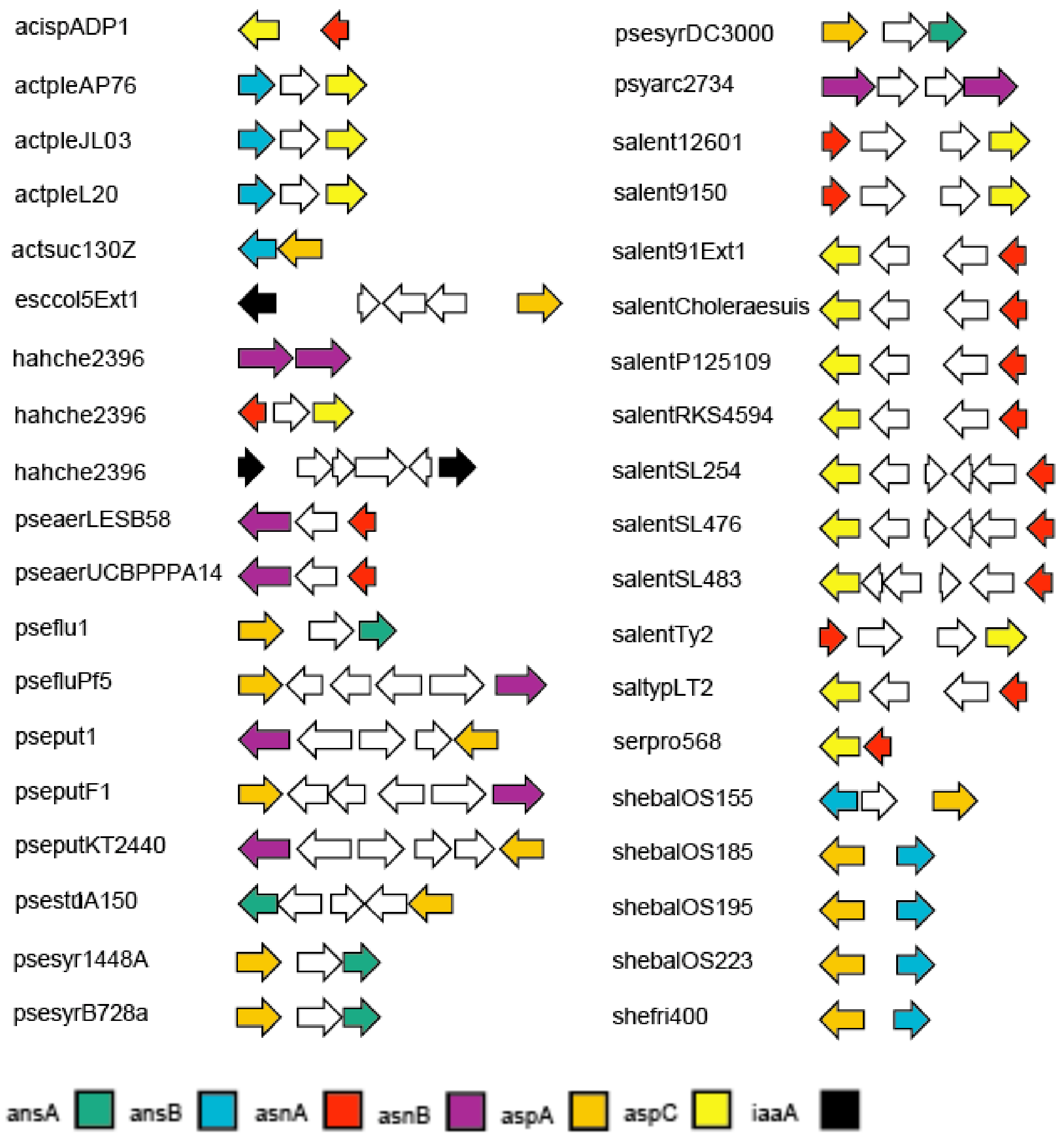


Figure 3.5: Clusters of the asn, asp SP. This figure shows all 38 clusters identified and visualised by GenClust. Taxon legend can be found in supplementary data 3.2.

The clusters associated with *gly* and *ser* SP (figure 3.6 and supplementary data 3.4) were dominated by a duplication of a *glyA* homolog. This duplication accounted for most of the clusters identified. Two other relatively common clusters were identifiable, *glyA* with *serA* and the *serA* with *serC*. The combination of *glyA* and *serC* was never observed, though given the relatively small number of clusters associated with the SP this was not altogether surprising. More interesting was the absence of *serB* from any cluster despite having 151 homologous sequences found during the BLAST searches. One potential explanation is that the *serB* gene is under strong selection to cluster with genes other than the ones associated with this SP.

The *ile*, *leu* and *val* SP displayed a large amount of variability in both cluster size and content (figure 3.7 and supplementary data 3.5). Despite this there is a modular nature to their construction. Common units include the *leuA, B, C* and *D* genes, the *ilvA, B, C, D, E, G1* and *M* genes and the *ilvG2* and *ilvH* genes. However many such units exist and the gene sets from the different units often overlap. For example while *ilvG2* and *ilvH* are often found together, another common combination has *ilvG2* paired with *ilvN*. This implies that while certain combinations are preferred, they are permutable. Many large clusters exist for this SP, consisting of discrete units. One large cluster present in *YESS* group consists of two units. The first unit contains *leuA, B, C* and *D* genes. The second unit contains the *ilvG2* and *ilvH* genes. These two units are separated by IGs. The number of IGs can vary depending on the genome. In the *Xanthomonas* lineage there is evidence of crossover between units. *ilvA*, *B* and *C* along with *leuA* form a unit along with a single IG that separates *ilvA* and *leuA* from *ilvB* and *ilvC*. All five genes are tightly packed and

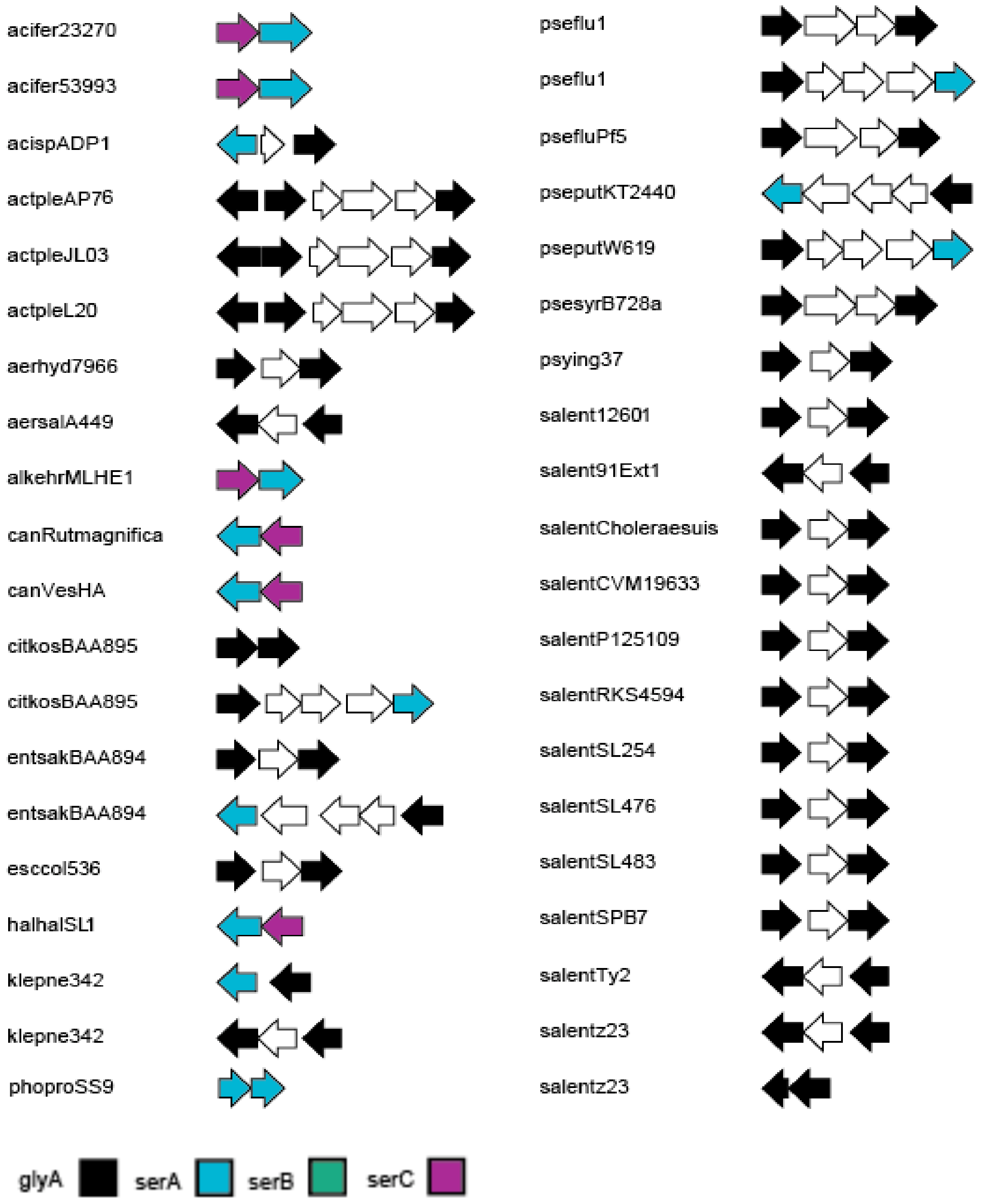


Figure 3.6: Clusters of the gly and ser SP. This figure shows a selection of clusters associated with the SP (the full set of clusters can be found in supplementary data 3.4). Taxon legend can be found in supplementary data 3.2.

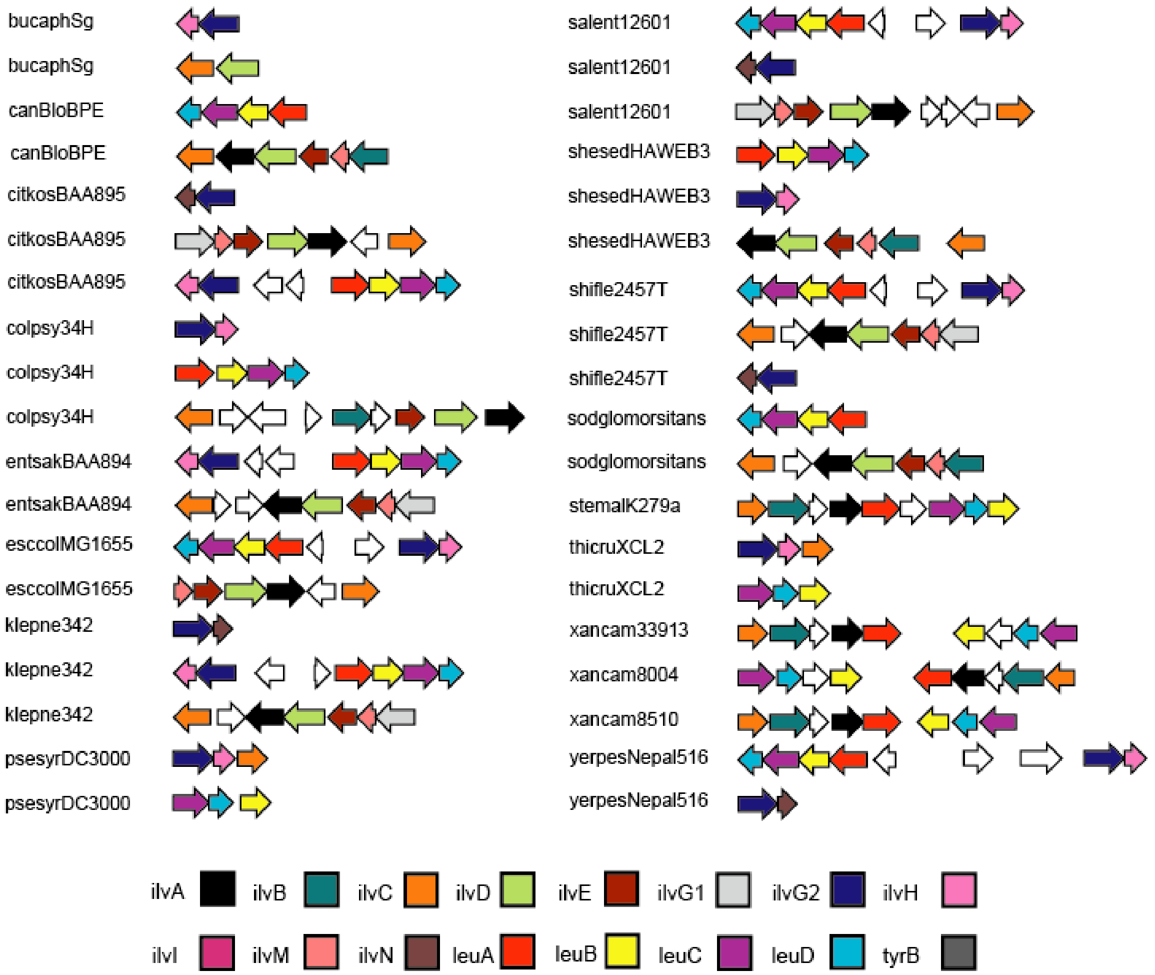


Figure 3.7: Clusters of the *ile*, *leu* and *val* SP. This figure shows a selection of clusters associated with the SP (the full set of clusters can be found in supplementary data 3.5). Taxon legend can be found in supplementary data 3.2.

transcribed in the same direction, indicating an operon structure. A second likely operon structure forms the remainder of the cluster, consisting of *leuB, C* and *D* and an IG. This second unit is separated from the first by a stretch of non-coding DNA and is transcribed in the opposite direction. While *leuA* is still part of the cluster as a whole it has been separated from the three other *leu* genes. This is curious, as the benefits of rearranging the standard configuration of the *leuA, B, C* and *D* operon are not obvious. One possibility is that it is a neutral or only slightly deleterious change.

The *lys, met* and *thr* SP showed diversity in terms of cluster content (figure 3.8 and supplementary data 3.6). While clusters were small, on average consisting of two functionally related genes, the combinations of genes within each cluster varied from genome to genome. Lineage specific patterns were observable, for example genomes from the *Pseudomonas* lineage contained clusters of *thrA* with *thrC*, *dapF* with *lysA*, *metH* with *asd* and *metE* with *metH*. However, even among the *Pseudomonas* genomes many other combinations of functionally related genes existed. Some clusters appeared frequently across the data, these included the pairing of the *metL* gene with the *metB* gene and clusters of *metL, thrB* and *metC*.

The clusters associated with the *phe, trp* and *tyr* SP, see figure 3.9 and supplementary data 3.7, showed patterns in line with the *lys, met* and *thr* SP. Many small clusters of different gene combinations were present. Similarly several combinations were frequently observed across the data such as the pairing of *aroB* with *aroK* and the pairing of *aroD* with *ydiB*. The notable exception was a large cluster containing *trpA, trpB, trpC,*

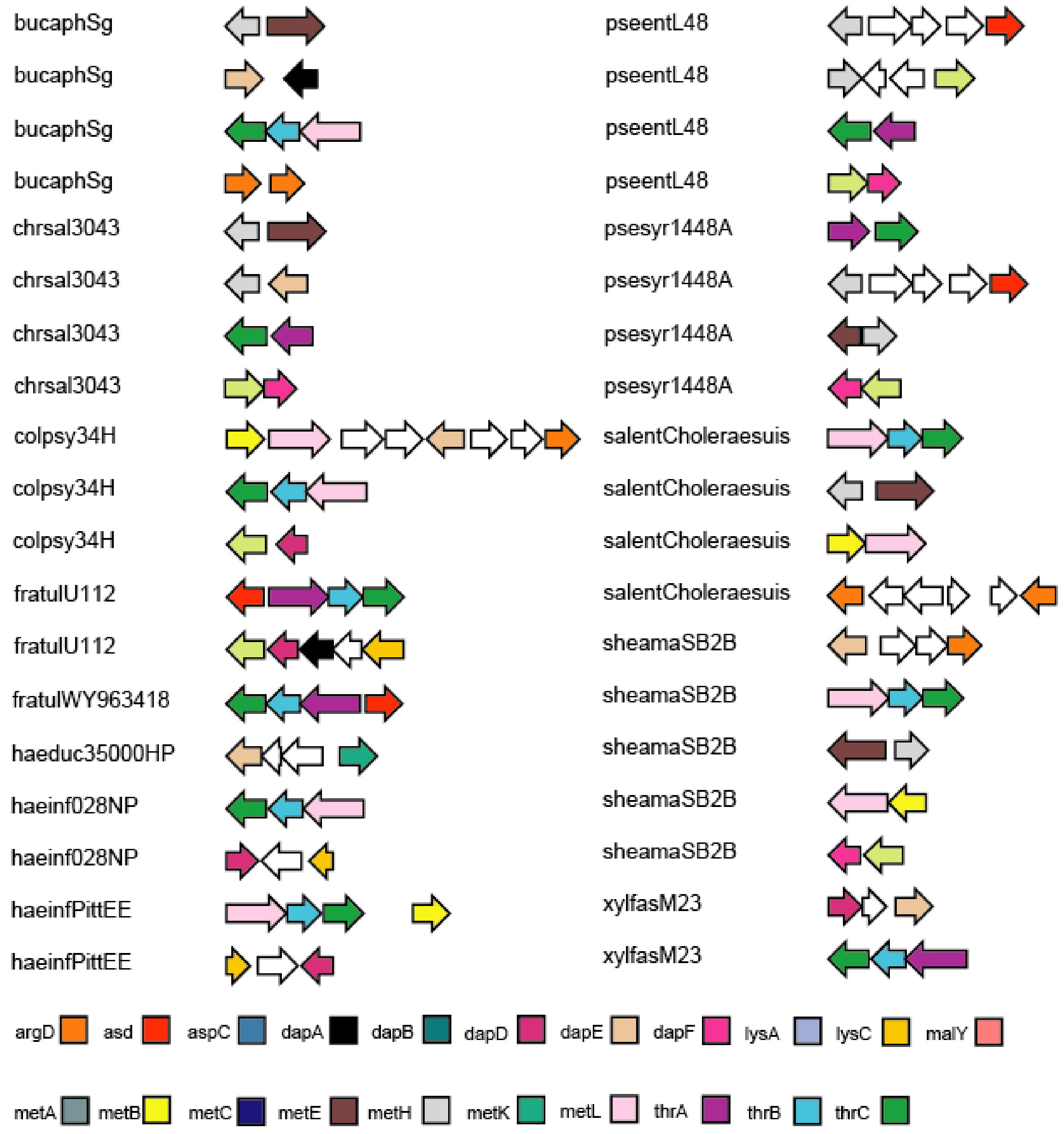


Figure 3.8: Clusters of the *lys, met* and *thr* SP. This figure shows a selection of clusters associated with the SP (the full set of clusters can be found in supplementary data 3.6). Taxon legend can be found in supplementary data 3.2.

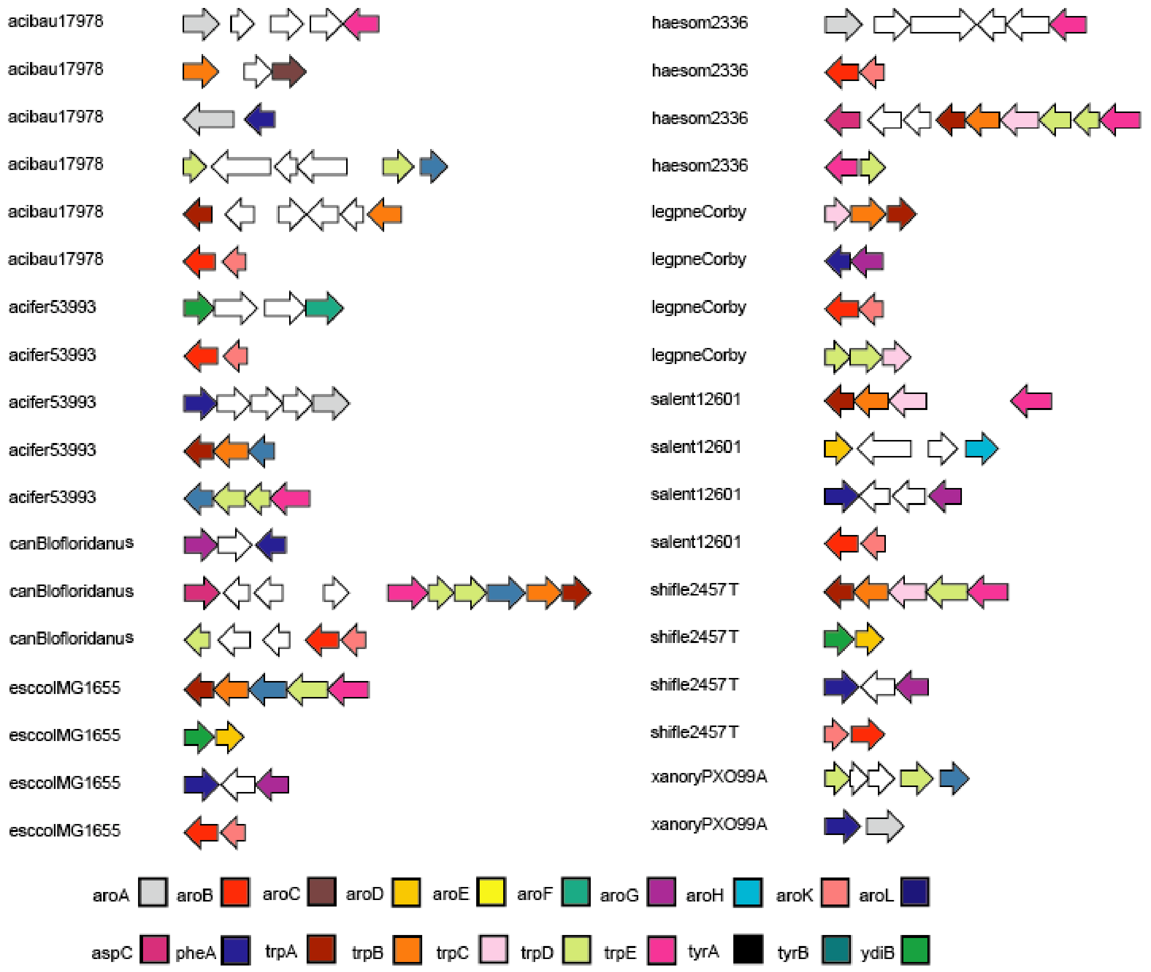


Figure 3.9: Clusters of the *phe*, *trp* and *tyr* SP. This figure shows a selection of clusters associated with the SP (the full set of clusters can be found in supplementary data 3.7). Taxon legend can be found in supplementary data 3.2.

trpD and *trpE*. This cluster was strongly conserved and present in most genomes, implying selection for conservation of gene content and order. The cluster was not immutable; some *Coxiella burnetii* strains have a deletion of the *trpD* gene from the cluster. Genomes from the *Pseudomonas* lineage did not contain the cluster at all, though sets of smaller clusters containing the genes were found to be present. Many genomes, for example *Pseudoalteromonas atlantica* and *Photobacterium profundum,* had a tandem duplication of the *trpD* gene again highlighting the dynamic nature of the structure and content of the *trp* gene cluster.

3.4 DISCUSSION

Runtime analysis of GenClust demonstrates a polynomial increase in computation time with increasing numbers of BLAST hits identified (figure 3.2). The algorithm for the identification of linked pairs is of order O(n^2) were x is the total number of BLAST hits to query sequences. During the identification process BLAST hits must undergo a pairwise comparison in order to find all possible linked pairs of genes (the diagonal is subtracted because a gene is not a linked pair with itself and the total comparisons are cut in half as all linked pairs are reciprocal). Additional overhead occurs after the initial BLAST linking phase, as linked pairs must be overlapped. This adds computation time of the same order, however the overhead is inexpensive in comparison to the BLAST phase in general as there are generally much fewer linked pairs to overlap. Iterative overlap between the linked pairs further increases the runtime, however the iterations are also inexpensive when compared to the initial BLAST linkage phase and the overhead is dependent on the sizes of the clusters identified in the end (large clusters requires more iterations to join all the linked pairs). Using all 180 genomes produced runtimes that ranged from 4 hours for the *asn* and *asp* SP to 50 hours for the *lys*, *met* and *thr* SP. Runtime correlated strongly with total BLAST hits identified, with the *asn* and *asp* SP and *lys*, *met* and *thr* SP having the minimal and maximal values of each respectively. This correlation held true for the two other test cases, where 100 and 50 genomes were randomly selected from the original set and the analysis was re-run to examine the effect of a reduced number of genomes on computation time. Reduction in genome number produced a minimal (the *asn* and *asp* SP) and maximal (the *lys*, *met* and *thr* SP) runtime of 2 and 18 hours respectively for the 100-genome dataset and a minimal (the *asn* and *asp*

SP) and maximal (*lys*, *met* and *thr* SP) runtime of 1 and 4.5 hours respectively for the 50-genome dataset.

Most of the clustering data produced by GenClust was in line with expectations based on the number of genes involved with each SP. The ratio of BLAST hits to the number of clusters identified correlated strongly with the total number of genes in the SP. The exception to this was the *gly* and *ser* SP, which had a higher number of clusters identified than the *asn* and *asp* SP despite having fewer genes involved. This is readily explainable by examining the content of the *gly* and *ser* clusters (figure 3.6). A duplication of the *glyA* gene homolog, not the true *glyA* gene itself, accounts for the majority of clusters identified was thus inflates the total number of clusters identified for this SP.

The average number of genes per cluster per SP was 2.55 when only functionally related genes were considered and 3.64 when IGs were included in the calculation. This implies that in general clusters of functionally related genes were small across the data, with most clusters consisting of only two functionally related genes. The increase in average cluster size when the IGs are included into the calculation implies that clusters are not tightly packed and that IGs are common in functionally related clusters, though it should be noted that just because an IG was not directly involved in an SP does not mean that it is not functionally related. Many IGs are likely to come from related pathways.

Examining the average cluster size on an SP-by-SP basis (see table 3.2) provides some interesting insights into variations in cluster sizes across the data. Considering only functionally related genes (figure 3.3) we see that for all SPs clusters of two functionally related genes are the most common, or exclusive in the case of the *asn* and *asp* SP. The total number of genes per SP does not correlate particularly well with cluster size in terms of individual SPs. The *lys, met* and *thr* SP has the most genes, at 21, however the largest clusters are of four functionally related genes. In contrast the *ile*, *leu* and *val* SP has 16 genes but the largest clusters contained 7 functionally related genes. This SP appears to have a tendency towards large clusters, with a spike in the graph for clusters containing 6 functionally related genes. This spike in the graph was unique, with the rest of the data for all the SPs tailing off quickly from high numbers of observations of two gene clusters to low numbers of observations of larger functionally related gene clusters.

Adding in IGs to the calculation (figure 3.4) reduced the sharpness of the drop in terms of observations versus cluster size. Most of the trends stayed the same, with larger clusters occurring less frequently. One exception to this was a spike for three gene clusters for the *gly* and *ser* SP. This is explainable by the high frequency of a three-gene cluster containing two copies of a *glyA* homolog separated by an IG. Another exception are the clusters for the *asn* and *asp* SP which display an almost level curve ranging from two to six genes. This implies that no real pattern exists in the clusters for this SP. Interestingly the spike for the *ile*, *leu* and *val* SP leveled out when IGs were included. If the spike was due to the presence of a strongly conserved ancestral cluster of six functionally related genes then it would be expected that this would still be reflected the data when IGs were included. However no such spike exists when IGs are included, with a gentle decrease in number of observations versus cluster size. One possible explanation comes from the idea of the larger clusters being composed of functional modules. The genes in these modules may be strongly conserved, but when modules are combined into larger clusters, they may be separated by varying numbers of IGs. Even the modules are not immutable, for example the *leuA, leuB, leuC* and *leuD* genes frequently cluster together in a potential operon but there are still examples where there is an IG in the middle of the module. The general trend across the data is that the presence of IGs in clusters is frequent and clustering of functionally related genes is not so strong as to exclude (potentially) non-functionally related genes.

Pairwise cluster co-occurrence data shows that not all genes co-occur with one another in the data (table 3.3-3.7). In fact the majority of possible co-occurrences are not seen in the data. The *phe, trp* and *tyr* SP showed the smallest number of unique co-occurring pairs, with only 25 percent of the total number of possible pairs observed. The *ile, leu* and *val* SP showed the highest level of cluster co-occurrence, with 54 percent of all possible pairings observed. This suggests there is a pattern to cluster content. It is likely that for SPs like the *phe, trp* and *tyr* SP, where several hundred clusters contain only a quarter of all possible co-occurrences of genes, that conservation of smaller ancestral clusters across the data limits the combination of genes within those clusters. Evidence of this can be seen in the visual analysis the cluster data (see figure 3.9 and supplementary data 3.7). Clusters such as the *trpA, B, C, D* and *E* genes are repeated many times, thus limiting potential variation. Contrastingly, The *ile, leu* and *val* SP displayed much greater variation. Clusters associated with this SP were generally larger (figure 3.7 and supplementary data 3.5). As previously discussed, some modules of genes were prevalent in the data for this SP, with larger clusters often consisting of combinations of these modules. Given the number of genes in the SP, sixteen, and the frequency of large clusters, it is not surprising that a large percentage of the genes are found to co-occur with one another. As such the co-occurrence data by itself does not imply that the *ile, leu* and *val* SP shows more variation in terms of cluster content than the other SPs, merely that large clusters relative to the number of genes examined increases the percentage of genes found to co-occur.

The overall picture produced by the data is one small clusters being the norm. Many of these two gene clusters are likely to be false positives, as in the case of the *glyA* homolog duplication, where clustered sequences represent homologs but not true orthologs of the query sequences. This problem reduces as cluster size increases, large clusters of homogous sequences, such as those found for the *ile, leu* and *val* SP, have a higher likelihood of representing true orthologous clusters.

Large clusters, as in the case of the *ile, leu* and *val* SP, show evidence of conserved sets of genes, themselves small clusters, coming together to form larger structures. While conservation was prevalent, considerable diversity still existed, and even strongly conserved ancestral clusters contained taxon specific IGs, duplications or deletions. This implies that selection for clustering is relatively strong but is not so strong that once a set of genes becomes clustered the structure of the cluster becomes unchangeable. GenClust allowed for rapid, large-scale identification and analysis of clusters of genes homologous to the initial set of amino acid biosynthesis superpathway genes present in *E. coli* K12.

In the next chapter I take a specific gene cluster, the *paa* cluster, and examine it in detail with respect to models of gene cluster and operon formation. I identify *paa* clusters across a wide range of bacterial genomes, using GenClust for initial identification of potential clusters and manual curation of the results to recover true *paa* gene clusters. I use a phylogenetic approach to analyse the evolutionary history of clustered *paa* genes for conflicting signals. I relate this analysis to how clusters and operons are prercieved to form.

Chapter 4 - Recurring cluster and operon assembly for Phenylacetate degradation genes

4.1 INTRODUCTION

The aerobic degradation of phenylacetic acid in *E. coli K12* occurs *via* a series of five reactions, involving eleven catabolic *paa* genes (Ismail et al., 2003), two of which are distant paralogs, with the rest showing no sequence homology to one another (figure 4.1). The first step of the pathway is catalysed by the product of the *paaK* gene, a CoA ligase that catalyses the conversion of phenylacetate into phenylacetyl-CoA. The second step involves a ring-oxygenase complex formed from the gene products of *paaABCDE*. This heteromer converts phenylacetyl-CoA into 2’-OH-phenylacetyl-CoA. The third step, where 2’-OH-phenylacetyl-CoA is converted to 3-hydroxyadipyl-CoA, is jointly catalysed by *paaJ*, *paaG* and *paaZ.* The fourth step sees the conversion of 3-hydroxyadipyl-CoA by *paaF* and *paaH* to -ketoadipyl-CoA. The final step is catalysed by *paaJ*, which converts -ketoadipyl-CoA to succinyl-CoA, thereby connecting phenylacetate degradation with the TCA cycle (Ismail et al., 2003). In addition to these 11 catabolic genes, *E. coli K12* has 3 other *paa* genes, two of which regulate the pathway (*paaX* and *paaY*), the other has an unknown function (*paaI*). Other *E. coli* strains such as *E. coli O157* and *E. coli O73* do not share homologs to all 11 catabolic genes, with no homologs found for *paaA*, *paaB*, *paaC*, *paaD, paaE, paaG* and *paaK* in either of these two genomes. However, previous studies have identified other bacteria as having homologs to *paa* genes, such as *Pseudomonas putida U* (Olivera et al., 1998). In

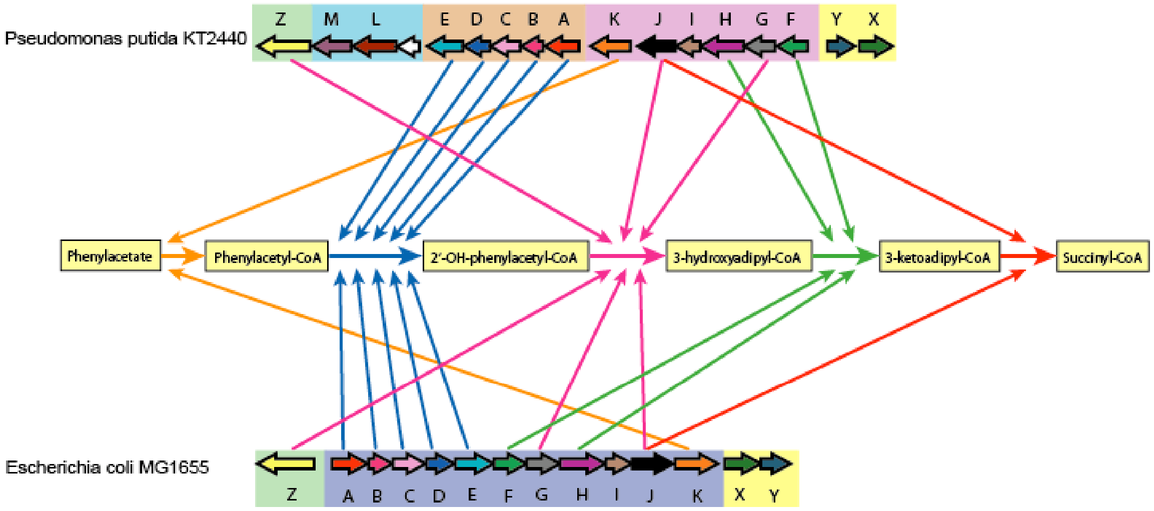


Figure 4.1: The phenylacetate degradation pathway and the *paa* gene clusters of *E. coli* K12 and *P. putida* KT2440. Steps in the pathways are colour coded by arrows, with genes encoding products involved in each step connected by a correspondingly coloured arrow.

addition to these 14 genes found in *E. coli K12*, a further three genes associated with the pathway were examined in this study. These were *paaL* and *paaM*, coding for a phenylacetic acid transporter protein and a phenylacetic acid specific porin respectively, and *tetR*, a transcription factor.

The genes involved in phenylacetate degradation in *E. coli K12* and *P. putida U* are located in clusters (Ferrandez et al., 1998; Olivera et al., 1998). In this study I define a gene cluster as a set of functionally related genes located in close physical proximity in a genome. The term operon refers to a set of genes under common regulatory control that are transcribed into a single mRNA and are all co-directional in orientation on the chromosome. An operon, therefore, is a more structured instance of a cluster. All operons by definition are also clusters, but not all clusters are operons. A gene cluster can consist entirely of independently transcribed genes or multiple operon structures or combinations of both. Clusters and operons are observed both in prokaryotes and eukaryotes, however, the system of operon processing in eukaryotes involves mRNA splicing, and is different to the system in prokaryotes (Spieth et al., 1993; Blumenthal et al., 1995).

Clustering of genes involved in the same metabolic pathway is a widespread phenomenon (Siefert et al., 1997; Dandekar et al., 1998; von Mering et al., 2002; Fani et al., 2005; Wong et al., 2005), and the polycistronic operon is a paradigm of prokaryotic genomic biology (Demerec and Hartman, 1959). However, the process of operon formation remains poorly understood and the precise link between clustering and operon formation has never been fully explained, though several models exist.

The simplest model is the natal model where clusters form via tandem gene duplications (Lawrence, 1997). However, many operons contain genes that are not homologous, but have some kind of functional link. As a general mechanism of operon formation, the natal model is inadequate.

The Fisher model postulates that clustering of genes into operons offers the benefit that random recombination events will tend to separate co-adapted genes less often if they are clustered together. This model has suffered criticism recently because of observations of orthologous replacement *in situ* of operon genes (Omelchenko et al., 2003; Price et al., 2006), which suggests that the primary reason for operon formation is unlikely to be the preservation of co-adapted alleles.

The co-regulation model (Jacob et al., 1960) states that operons are formed in order to facilitate the production of gene products in equal measures. This theory only accounts for operon maintenance. In order for an operon to spontaneously form, rare, highly specific recombination events must occur. However, it has recently been asserted that operon formation is driven by co-regulation (Price et al., 2005b). This assertion is largely due to the more complex regulatory regions associated with operons in some γ-proteobacteria compared with genes that are not in operons. However, this study only focused on operons and not on the broader issue of cluster formation.

The selfish operon model (SOM) suggests that operons in prokaryotes are in some respect like viruses or transposons and their formation facilitates their horizontal gene transfer (HGT) (Lawrence, 1997). The formation of an operon is therefore of no direct benefit to the organism but it means that the fitness of gene cluster itself is enhanced. An extension of the SOM posits that if HGT is indeed the main reason for operon formation, non-essential genes are more likely to be in operons/clusters than essential genes (Lawrence, 1997). However, Pál and Hurst have provided evidence that essential genes are more likely to be found in operons and clusters than non-essential genes, thereby presenting a significant problem to the SOM (Pál and Hurst, 2004).

A recent proposition has been made that gene clustering is due to the relative difficulty of protein movement through the cellular matrix (Svetic et al., 2004). This model, known as the protein immobility model (PIM), suggests that because transcription and translation are coupled in prokaryotes, the resulting physical proximity of enzymes minimizes the steady state level of reaction step intermediates thereby saving energy and reducing the amount of protein that needs to be produced. The PIM has not been tested using empirical data, but has been supported by computer simulation. An observation that indirectly supports the PIM is the study by Elowitz *et al*. that shows that protein diffusion is slower through the cytoplasm than through water, is adversely affected by the size of the protein, and is also reduced when expression levels are higher (Elowitz et al., 1999).

Lastly, Fang et al. have suggested that the clustering of genes is due to persistence (Fang et al., 2008). They observed that two types of genes show a high tendency to cluster within a genome, genes that are widely distributed, the ‘persistent’ genes, and genes that are very narrowly distributed, the ‘rare’ genes. The clustering of rare genes was explained by the SOM, as these rare genes were likely candidates for HGT. Fang et al. suggested that the clustering of persistent genes was due a constant flux of insertion and deletion events, with the probability such events disrupting a persistent gene, which would have a negative impact on an organisms fitness, decreasing when the genes are clustered. Fang et al. supported this assertion with a number of computer simulations.

Because *paa* genes show a patchy phylogenetic distribution and previously observed *paa* clusters have diverse structures that appear to be independent of the species phylogeny, it was clear that the phenylacetate degradation pathway was important to study from an evolutionary standpoint. Indeed, phenylacetate degradation has previously been identified as a potential model for understanding the evolution of metabolic pathways (Luengo et al., 2001). By examining the gene content of previously studied *paa* clusters a total of 17 genes are associated with the pathway including catabolic genes, regulatory genes, a transporter and an exporter. In this study I identify new *paa* gene clusters and examine the structure and distribution of *paa* gene clusters with respect to their evolution and implications for models of both cluster and operon formation.

4.2 MATERIALS AND METHODS:

4.2.1 Homolog identification

An iterative strategy for locating homologs to all 17 genes encoding proteins involved in the degradation of phenylacetate was implemented. Initially, the genomes for taxa containing known *paa* gene clusters, previously reported in the literature, were downloaded from GenBank (Benson et al., 2007). A BLAST-based (Altschul et al., 1997) similarity search strategy was used to extract all the known *paa* genes from these initial genomes and used them in order to find homologs in other completed bacterial genomes. These additional bacterial genomes were downloaded from GenBank, bringing the total number of genomes in the dataset to 102 (see supplementary information S. I. 4.1 for the full list).

Alignments were generated using Muscle v3.5 (Edgar, 2004) for genes where multiple homologs were found (see S. I. 4.2 for alignments). The exceptions were the *paaL* and *paaM* genes that were only found only in *P. putida KT2440*. This gave a total of 15 initial alignments. These alignments were then used as input for PSI-BLAST using the default parameters (Altschul et al., 1997), with the larger dataset of 102 bacterial genomes as the input database. This gave a comprehensive list of homologs, see table 4.1 for further information.

GenClust was used to analyse the PSI-BLAST results for homolgous gene clusters. GenClust iteratively identified sets of linked genes from the PSI-BLAST results. If two

Gene	From literature	After PSI-BLAST
paaA	16	25
paaB	16	25
paaC	16	25
paaD	16	25
paaE	13	23
paaF	8	177
paaG	11	23
paaH	9	145
paaI	12	19
paaJ	11	277
paaK	14	37
paaL	1	1
paaM	1	1
paaX	6	13
paaY	5	84
paaZ	9	399
tetR	8	14

Table 4.1: Homolog identification. The first column contains all 17 *paa* genes. The middle column represents the number of genes taken from previous studies on known *paa* genes. The third column contains the number of homologous genes identified after an iterative search using PSI-BLAST.

genes found in the result files generated by the PSI-BLAST searches came from the same genome and had no more than five intervening genes between them, then such genes were considered to be an initial linked pair. All initial linked pairs were identified and then merged by the software if they overlapped. In this way, clusters of various sizes were identified.

4.2.2 Construction of phylogenetic trees:

The 15 gene families were used to build phylogenetic trees. *paaL* and *paaM* were excluded from any further analysis as no homologs to these genes were identified. The amino acid sequences of all homologs were extracted from their genome files and each family was aligned using Muscle v3.5 (Edgar, 2004) with all settings at their default values. Model selection was performed on the alignments using ModelGenerator (Keane et al., 2006) and maximum likelihood phylogenetic trees were constructed based on the selected models using Phyml v3.0 (Guindon and Gascuel, 2003). Confidence in phylogenetic hypotheses was assessed using the bootstrap resampling approach (Felsenstein, 1988) (see S. I. 4.3 for phylogentic trees).

4.2.3 Visualisation of clusters on phylogenetic trees:

For each gene family, it was important to be able to visualise both the relationships among members of the family and their cluster context simultaneously. Visualisation of each gene cluster was achieved by extracting the necessary genomic location information for the cluster from the corresponding GenBank file. This was carried out automatically using GenClust (see chapter 3). Once this information was parsed from the GenBank file, the corresponding cluster was drawn using the postscript language (Adobe Systems, San José, California). Visual representations of the clusters were then merged with the phylogenetic trees. If, for instance, a cluster contained the genes *paaA* and *paaB*, then this cluster will appear on the *paaA* tree at the phylogenetic position of the *paaA* gene and on the *paaB* tree at the phylogenetic position of the *paaB* gene. Adobe Illustrator files (Adobe Systems, San José, California) can be found in S. I. 4.4 (note however that some trees, for example *paaJ*, were too large to visualise in this manner).

4.2.4 Identification of HGT events:

For the comparison of the evolutionary history of homologous genes a bootstrap resampling approach was used to detect potential HGT events. If two homologous genes from a pair of structurally similar clusters were found as close relatives on their corresponding phylogenetic tree then it was assumed that there was no evidence of HGT. If however the two genes did not group closely on the phylogenetic tree and instead grouped with homologous genes from clusters that showed no obvious structural similarity then a potential HGT event was inferred. The strength of confidence in both sister-group relationships and potential HGT events was determined by the bootstrap values for the nodes involved. A support value of 70 percent or higher for a particular grouping was considered strong support, while less than 70 percent support was considered weak support. The analysis of HGT events was carried out manually and further refined by considering the underlying species phylogeny of the taxa involved.

4.3 Results:

In order to test whether a cluster has been independently assembled more than once, the phylogenetic trees of both cluster and non-cluster homologs were examined. If a cluster has originated once and has never been subsequently perturbed, then for every gene in the cluster the corresponding phylogenetic tree will include a clade containing all the species in which the cluster is present. Given the prevalence of HGT (Kinsella et al., 2003) this clade does not have to correspond to any recognised phylogenetic group. The only relationships that are of importance are the relationships of the genes.

4.3.1 Variation in cluster and operon content and context:

Table 1 shows a summary of all 1,311 homologs identified via the PSI-BLAST searches, in terms of the frequency with which they were found in a *paa* gene cluster and if found in a cluster, how often they were in an operon. In the cases of *paaA*, *B*, *C* and *D* the genes were always found in an operon and obviously therefore, always in a cluster. For *paaE*, in 19 out of 23 instances it was found with other *paa* genes. *paaI* was always found in a cluster (19 occasions) and the majority of times (16 out of 19), in an operon. Similarly *paaX* and *tetR* were found relatively rarely (13 and 14 times respectively) and were usually found in clusters (11 out of 13 for *paaX*, 12 out of 14 for *tetR*) and 7 times each, they were in operons. *paaG* was found 23 times, 17 times in a cluster and 16 out of those 17 times it was found in an operon. *paaK* is found 37 times and in slightly more than 50% of the instances (21 of 37), it is in a cluster and the majority of times that it is in a cluster it is in an operon (19 of 21). The remaining five genes *paaF, paaH,*

Gene	In cluster	Not in cluster	In operon	Not in operon	Total genes	Alignment length (aa)
paaA	25	0	25	0	25	358
paaB	25	0	25	0	25	232
paaC	25	0	25	0	25	324
paaD	25	0	25	0	25	229
paaE	19	4	19	0	23	716
paaF	14	163	10	4	177	879
paaG	17	6	16	1	23	401
paaH	12	133	9	3	145	976
paaI	19	0	16	3	19	194
paaJ	11	266	11	0	277	1070
paaK	21	16	19	2	37	517
paaX	11	2	7	4	13	331
paaY	5	79	3	2	84	357
paaZ	22	377	8	14	399	1551
tetR	12	2	7	5	14	308
All	**263**	**1048**	**225**	**38**	**1311**	**-**

Table 4.2: Frequency of presence in a cluster and operon. This table shows the number of times potential *paa* genes were observed both inside and outside of clusters and operons.

paaJ, paaY and paaZ are more widely distributed and the majority of times these homologs are not found in clusters or operons. The gene that is least likely to be found in an operon is *paaY*, which is only found in an operon in 3 out of 84 instances. Interestingly, apart from *paaA, B, C, D* and *E,* where being in a cluster automatically means being in an operon, most other genes are found in an operon the majority of the times they are found in a cluster. The exception is *paaZ*, where for 14 out of 22 instances of the gene being in a cluster it is not in an operon.

Figure 4.2 shows the set of unique operons involving two or more *paa* genes found in all identified clusters. The most striking aspect of this analysis is the sheer diversity in terms of size, gene content and gene order among the operons. A total of 33 different operons were identified, ranging in size from 2 to 11 genes. Out of the 33 unique operons only two display identical gene content, one being *paaXY*, the other *paaYX*. This diversity is not surprising from a mathematical standpoint, given that 17 genes were examined in the study. Even for operons consisting of only 2 genes there are 289 possible permutations. Aside from the *paaABCDE* operon, which is clearly under strong selection (all 25 clusters form operons), no particular operon composition or configuration is dominant. This result seems to indicate that operon formation (apart from *paaABCDE*) is not dependent on the composition of the genes that are present. Operons seem to form, simply when members of the pathway are present and no single operon composition or order is obligatory.

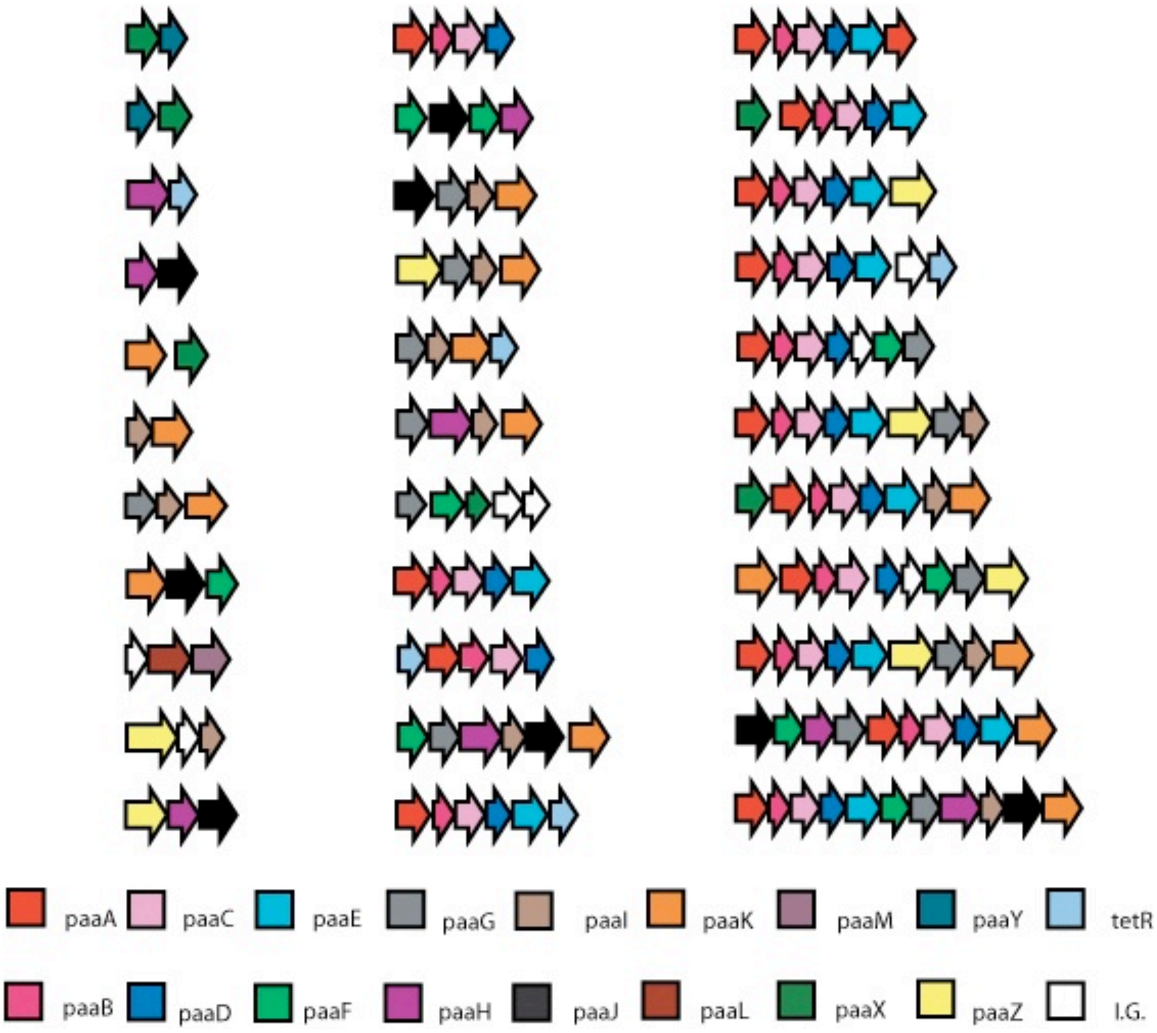


Figure 4.2: An exhaustive list of all observed operons in the dataset of 102 genomes examined. Each arrow represents a gene, with the name of the gene being given in the legend. I.G. refers to an intervening gene, which is a gene in the cluster that is not involved in the degradation of phenylacetate.

4.3.2 Co-occurrence and intra-cluster distance:

Figure 4.3 shows co-occurrence frequency and average cluster distance for *paa* genes. Some traits stand out. *tetR* is never found in a cluster with *paaX*, implying that *tetR* may serve a similar function to *paaX*. *tetR* is also a regulatory gene so this seems like a reasonable conjecture. *tetR* is also never found in a cluster with *paaL* and *paaM*. However, this is not unexpected as *paaL* and *paaM* are only in one cluster and this cluster contains a copy of *paaX*. Aside from these three instances, all other genes are found to co-occur in at least one cluster. In terms of average distances between genes in a cluster, some genes show a strong bias in terms of their location. *paaABCDE* is frequently found at the edge of a cluster or near an inserted non-paa gene. *paaABCDE* has no obvious affinity/bias to being close to any of the other *paa* genes. This high frequency of being close to the edge has the knock-on effect of funneling all the other genes to one side of *paaABCDE. paaF, paaG* and *paaH* are on average found in close proximity in a cluster. *paaI* and *paaK* are also found close to each other. *paaJ* is generally close to *paaF, paaG*, *paaH* and *paaI. paaL* and *paaM* are only found in one cluster and therefore no conclusions could be drawn about their location in the context of a cluster. *paaX, paaY* and *tetR*, who function as regulators of the pathway show a preference for being located towards the edge of the cluster, with *tetR* often being found close to an insertion, while *paaX* and *paaY* are on average located further away from insertions than any of the other genes. *paaZ* shows no strong relationship in terms of distance to any of the other genes but does have a tendency to be found close to the edge of a cluster or close to an

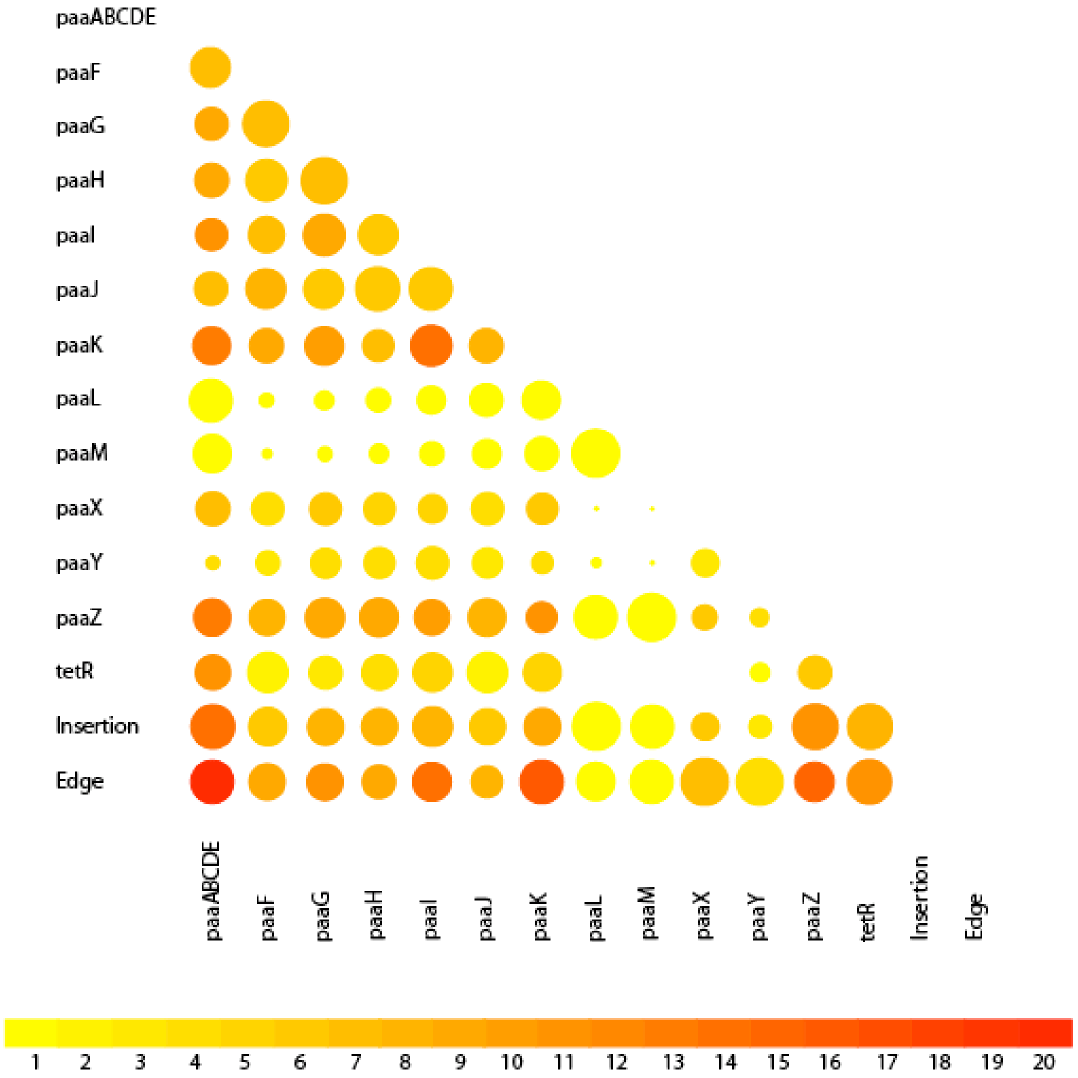


Figure 4.3: Co-occurrence and average distance of clustered genes. The colour chart denotes the number of times a pair of genes co-occurs in a cluster. The colouring of the edge row denotes total observations of each gene, as every cluster has an edge. *paaA, B, C, D* and *E* are treated as a single entity. The size of the circles is inversely proportional to the average distance between a pair of genes when clustered. Large circles imply a small average distance between a pair. Edge denotes the edge of a cluster. Insertion denotes a non-*paa* gene.

insertion. Excluding *paaL* and *paaM* from the analysis, the genes pairs that are on average furthest apart are *paaY* with *paaABCDE* and *paaY* with *paaZ*. The least frequently observed genes across the clusters were *paaL, paaM* and *paaY*. Despite some patterns mentioned above, no strong signal is prevalent across the data. If a single original cluster structure was responsible for all the clusters examined in this study, then it would have been expected that some signal of that structure would still exist in the extant clusters, but instead there is a diversity in cluster structure and content that can be most parsimoniously explained from the idea of multiple independent assemblies of the *paa* gene cluster.

4.3.3: Transcriptional orientation of *paa* genes:

Table 4.3 shows orientation data for all *paa* genes found to occur in a cluster. Three possible orientations were considered: head-to-head, where two *paa* genes were side by side and transcribed in opposite directions pointing towards one another, head-to-tail, where transcription of the pair occurred in the same direction, and tail-to-tail, where transcription of the pair occurred in opposite directions and pointing away from one another. All neighbouring pairs of *paa* genes were examined. The most common orientation was head-to-tail, observed a total of 193 times. Tail-to-tail pairings were much less frequent, occurring only 20 times. However by far the least observed orientation was the head-to-head which occurred only once between a neighbouring pair of *paa* genes.

Pairwise orientation	**Observed instances**
Head-to-tail	193
Tail-to-tail	20
Head-to-head	1

Table 4.3: The orientation of all observed *paa* gene pairs. This table refers to all instances in the data where two *paa* genes were found next to one another in the genome.

4.3.4 Analysis of gene clusters containing all 11 catabolic paa genes:

In order to establish how operons and clusters grow, the evolutionary history of the genes involved in the largest clusters was examined in detail. Of particular interest was whether for large clusters there was selection to keep co-adapted alleles together. Five clusters were identified in the dataset that were almost complete and were present in genomes that were not thought to be each others' closest relatives as judged using phylogenetic supertree methods based on completed genomes (Pisani et al., 2007). These included the clusters found in *E. coli*, *P. putida, Rhodoccoccus sp., Nocardia farcinica* and *Corynebacterium efficiens*. The evolutionary history of these clusters was examined in detail: phylogenetic trees and additional data are available as supplementary information (see S. I. 4.3 and 4.4).

Figure 4.1 shows the operon structures observed in *E. coli* K12 and *P. putida* KT2440. In *E. coli K12*, all fourteen genes involved in the pathway are clustered together and the cluster is broken into three operons (Ferrandez et al., 1998). *paaABCDEFGHIJK* are present in one operon, *paaXY* in another and the paaZ gene is transcribed by itself.

Superficially, the cluster in *P. putida* has high levels of similarity to the cluster in *E. coli K12* with simple rearrangements of the order of blocks of genes accounting for the majority of the observed differences, at first glance (see figure 4.1). In *P. putida* the gene cluster is arranged in five operons (Ferrandez et al., 1998) with *paaABCDE* being in one operon and *paaFGHIJK* being in a second, where both are merged in *E. coli*. *paaLM* and an unrelated gene are in another operon, *paaYX* is in an operon (the order is reversed in *E. coli*) and *paaZ* is transcribed by itself in the cluster. The gene content difference between the two clusters is the presence of *paaL*, a phenylacetic acid transporter, and *paaM*, a phenylacetic acid specific porin, along with an additional gene not known to be involved in phenylacetate degradation. *paaL* and *paaM* are only present in *P. putida* and in none of the other 102 genomes studied.

The phylogenetic analyses of the genes in these two clusters reveal a much greater degree of difference. The phylogenetic trees for all genes in these clusters were examined, with the expectation that the individual genes would be each other's closest relatives or at least reasonably closely related. For the *paaA, C, D, F, G, I, J, K and X* genes the *E. coli* and the *P. putida* copies grouped closely on a phylogenetic tree (see figure 4.4). Contrastingly, for *paaB, E, H, Y* and *Z* there was support for the separation of the two *E. coli* sequences from the *P. putida* sequences on their respective phylogenetic trees (figure 4.5). This result indicates that orthologous gene displacement has replaced a considerable number of genes in these clusters since the clusters separated from their common ancestor. Given the compositional similarity the most parsimonious explanation is that a complete cluster existed in the past and the two that exist in *E. coli* and *P. putida* today are descended with great modification, probably by rearrangement, insertion and orthologous displacement from the ancestral cluster. Of particular interest is the *paaABCDE* operon which is relatively invariable (see previous results), but from this analysis it is still subject to gene turnover and replacement. Notably these are the

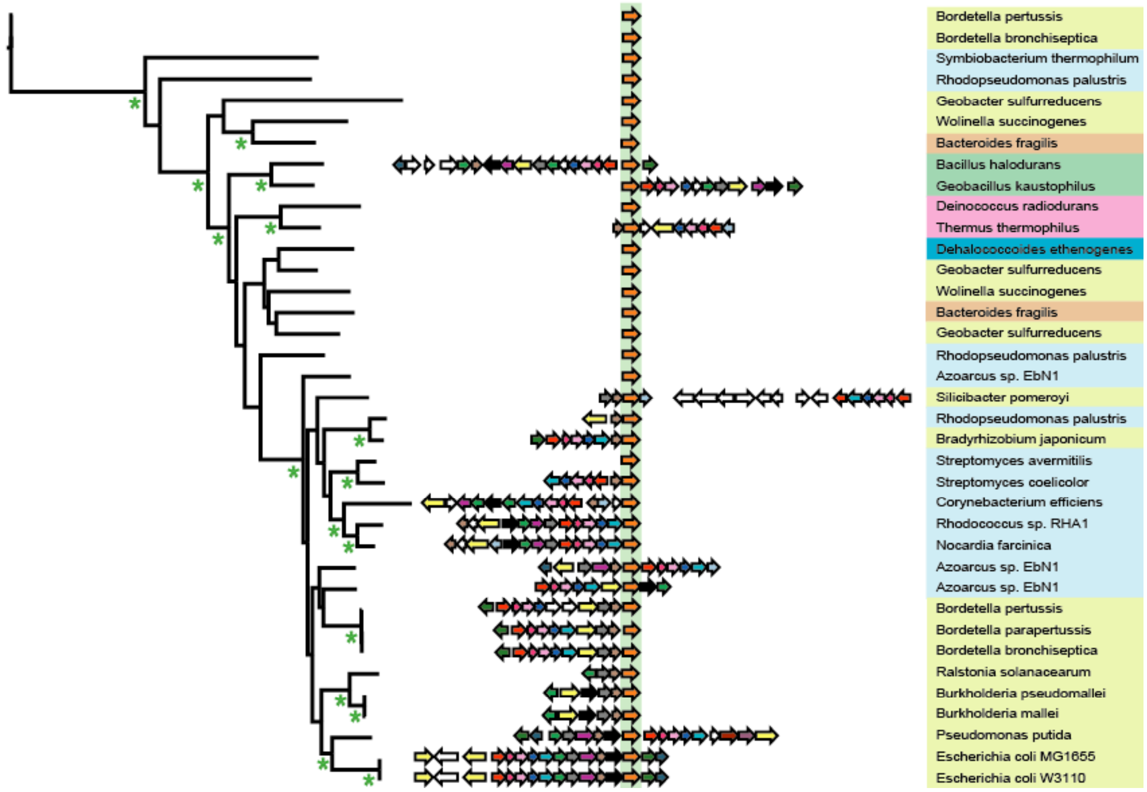


Figure 4.4: Phylogenetic tree for the *paaK* gene. On the left is the gene tree for *paaK*, in the middle are the clusters of genes in which the respective *paaK* genes are found, with the *paaK* genes aligned to one another and facing away from the tree. On the right are the taxon names (colouring representative of different bacterial groups). Strongly supported nodes (greater than 70 percent bootstrap support) are denoted with a '*'.

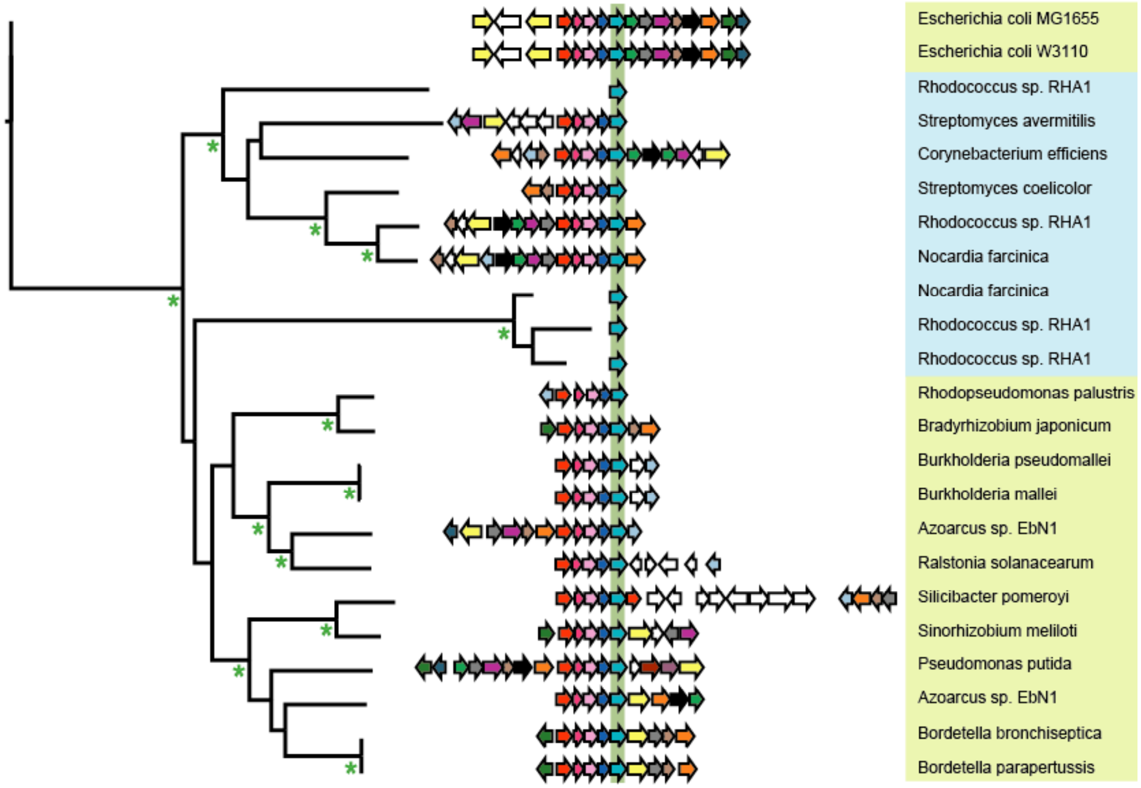


Figure 4.5: Phylogenetic tree for the *paaE* gene.

two most complete and similar clusters in the dataset. Extrapolating from this result and going further back through evolutionary history, assuming a similar rate of gene replacement, then it is likely that replacement of every single gene in this cluster – one at a time – can occur relatively rapidly.

4.3.5 The *Rhodococcus sp./Nocardia farcinica/Corynebacterium efficiens* clusters:

Rhodococcus sp. and *Nocardia farcinica* have two clusters that are very similar both in terms of gene content and orientation of genes within the cluster. In all phylogenetic analyses of the *paa* genes in the clusters, there is strong support for a sister group relationship between these two taxa (see figure 4.6). This suggests a recent common ancestor of both clusters. The *N. farcinica* cluster is split into four operons, the first is *paaI* by itself, the second contains a non-*paa* gene and *paaZ*, the third is *tetR* by itself and the fourth contains paa *J, F, H, G, A, B, C, D, E* and *K*. The *Rh. sp*. cluster is split into two operons, the difference being that *paaI* is in an operon with a non-paa gene and paaZ. This is followed by an operon consisting of *paaJ, F, H, G, A, B, C, D, E*, and *K*. These clusters are very different in terms of gene order when compared with either *E. coli* or *P. putida*.

The *Corynebacterium efficiens* cluster has some similarities to the *Rh. sp./N. farcinica* cluster. Firstly the gene content is almost identical, the only difference being that there are two copies of *paaF* in the *C. efficiens* cluster while *paaG* is absent. Secondly, all

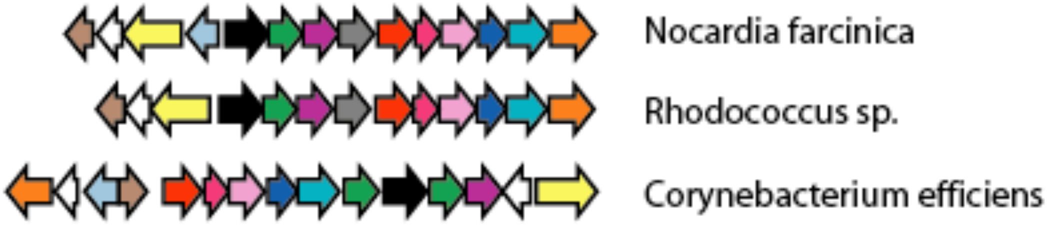


Figure 4.6: The *paa* gene clusters of *N. farcinica*, *Rh. sp.* and *C. efficiens*. The gene clusters of *N. farcinica* and *Rh. sp.* show share an almost completely conserved structure, with the presence of *tetR* in the *N. farcinica* cluster being the only difference. The *C. efficiens* cluster is more divergent but shares some subtle features in the layout of the genes and gene content. The non-paa genes located on the left-hand side of all three clusters are homologs of one another.

three clusters contain a gene of unknown function, and these three genes are homologs of one another. Thirdly, the *C. efficiens* cluster contains a copy of the *tetR* transcriptional regulator, as does the *N. farcinica* cluster. Interestingly, *Rh. sp.* also contains a copy of the *tetR* gene, but it does not lie in the *Rh. sp. paa* cluster. Examining the phylogenetic tree for the tetR gene shows that the non-clustered *Rh. sp. tetR* gene is sister to the clustered copy in *N. farcinica* (see the *tetR* gene tree in S. I. 4.3 for more detail). Lastly, there are subtle patterns of similarity in gene order with *paaZ, J, G, F* and *H* all in close proximity to one another in the three clusters, as were *paaA, B, C, D, E* and *K*.

When the phylogenetic relationships was reconstructed between the genes on the *C. efficiens* and the *Rh. sp./N. farcinica* clusters a sister group relationship was recovered for the *paaF, H, I, J, and K* genes with strong bootstrap support for this arrangement (see figure 4.4). However, for the *paaA, B, C, D, E and Z* genes there is strong support for grouping *Rh. sp./N. farcinica* with *Streptomyces coelicolor*, although in some cases the *C. efficiens* homolog is nearby on the tree (figure 4.5, 4.7 and 4.8). *S. coelicolor* has a *paa* cluster consisting of *paaK, I, A, B, C, D* and *E*. The results suggest that the *paaABCDE* operon in *Rh. sp./N. farcinica/S. coelicolor* are each others closest relatives for all the genes in the operon, while for the *paaK* and *paaI C. efficiens* groups while *Rh. sp*. and *N. farcinica*, to the exclusion of *S. coelicolor*.

An analysis of all five near-complete clusters does not support a single origin of these clusters and there are no genes that place *E. coli* or *P. putida* as sister-taxa to genes from

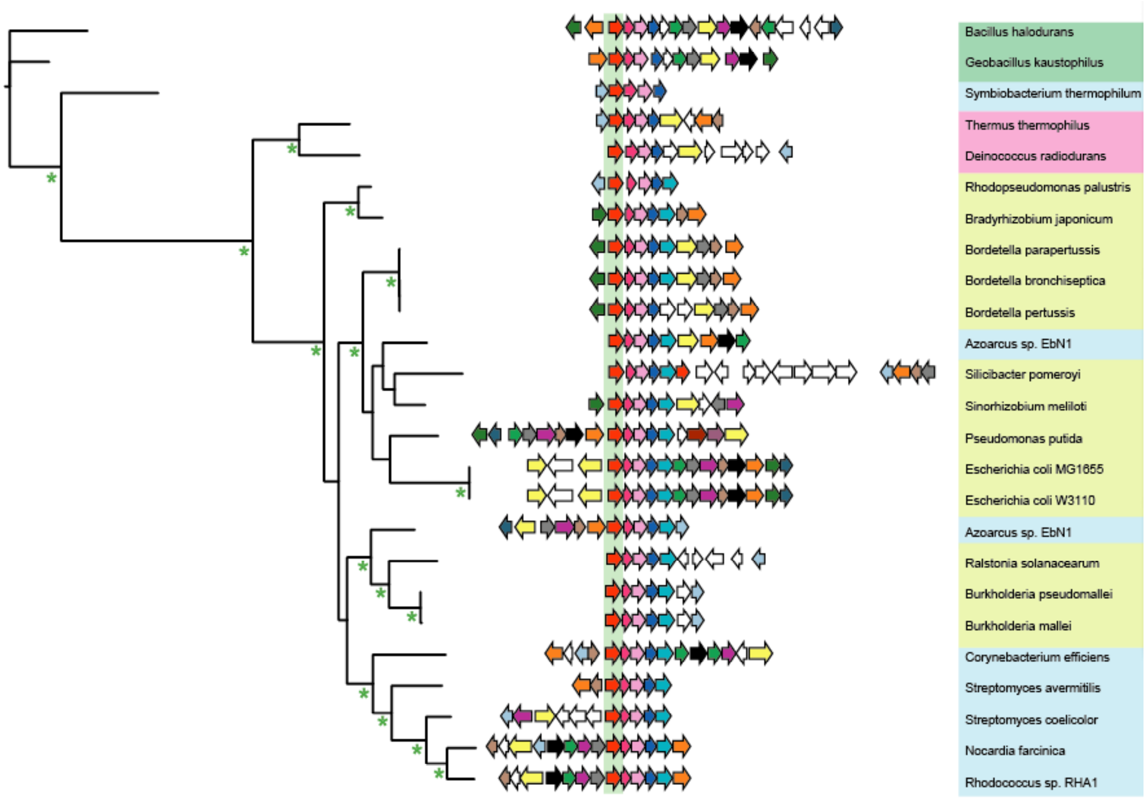


Figure 4.7: Phylogenetic tree for the *paaA* gene.

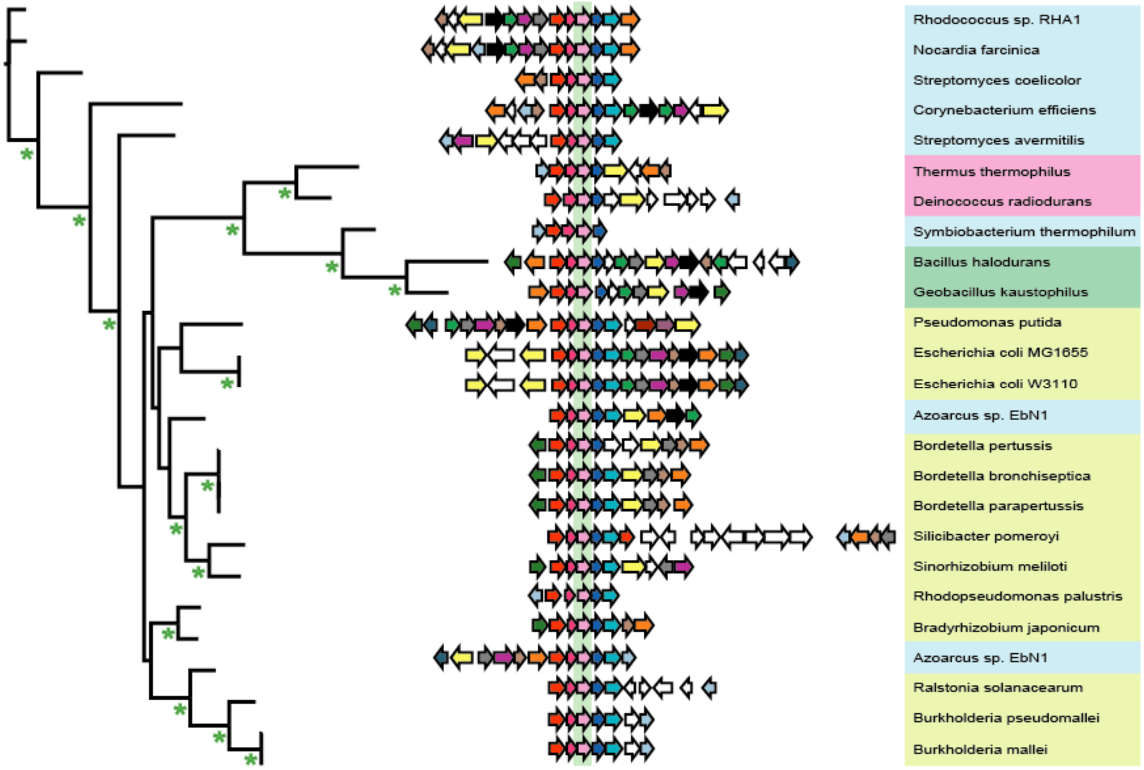


Figure 4.8: Phylogenetic tree for the *paaC* gene.

the *C. efficiens* or *Rh. sp./N. farcinica* genes. This indicates that formation of these near-complete clusters occurred independently on at least these two occasions, one assembly occurring in the proteobacteria and the other in the actinobacteria.

4.3.6 Comparative analysis of *paaK* and *paaC* gene trees:

A comparative analysis of the evolutionary histories of the *paaK* and the *paaC* genes can be seen in figures 4.5 and 4.6. The *paaC* gene is always found in an operon with *paaA*, *B* and *D*. Also, there is only one instance where this operon is not found in a cluster with other genes from the phenylacetate degradation pathway (i.e. in the case of *Symbiobacterium thermophilum*). The *paaK* gene is found in a cluster of more than two phenylactetate degradation genes approximately half of the times it is observed, the rest of the time, it is found as a single gene in the genome. There are four clans (Wilkinson et al., 2007) (the tree is only rooted for convenience, but is really unrooted) in which the *paaK* gene is at the edge of a cluster. Overall, it can be seen that the clusters for both genes dynamically grow, shrink and are rearranged (additional phylogenetic trees for every gene are supplied in S.I 4.3 and 4.4 and the reader should consult these trees).

To illustrate the variability in cluster context it is possible to take some examples from figure 4.8. In the *paaC* tree (figure 4.8), the two instances of this gene in *Azoarcus sp. EbN1* are located in completely different areas of the genome and both are part of a *paa* gene cluster. They are not particularly closely related genes, as evidenced by their phylogenetic positions. A reasonable speculation is that one or both of these genes was introduced into the genome via horizontal gene transfer. In contrast the two instances of *paaK* (figure 4.6) found in *A. sp*. are indeed each other's closest relatives, indicating a relatively recent gene duplication event. The *Thermus thermophilus* and *Deinococcus radiodurans* genes on both trees are nearest neighbors, suggesting a relatively recent common ancestor. This relative recentness of common ancestry might lead to the expectation that the cluster context of these two genes might be similar, however, the *D. radiodurans paaK* gene is not in a cluster, whereas the *D. radiodurans paaC* gene is in a cluster. Also, Deinococcus and Thermus are thought to form a bacterial clade (Garrity and Holt, 2001), so this orthology might be preserved since these two taxa shared a common ancestor.

On the *paaC* tree, there are three *Bordetella* clusters that are almost identical in terms of gene content and order. However, in one of the three genomes (that of *B. pertussis* Tohama I) there are two genes in the middle of the cluster that are not found in the other two strains. These two genes seem to have displaced the *paaE* gene in *B. pertussis* Tohama I, which lacks a copy of *paaE*. The other two *Bordetella* strains have copies of *paaE* in their clusters. The most parsimonious reconstruction, based on the *paaC* tree is that these two genes have been inserted into the cluster in *B. pertussis* Tohama I.

These observations demonstrate the enormous variability and rapid rate of assembly and disassembly of clusters as well as the semi-independent assembly of two near-complete clusters.

4.4 DISCUSSION

In this work, the evolutionary history of the genes involved in the phenylacetate degradation pathway has been analysed, with a view to understanding the origin and spread of functionally related gene clusters and operons.

The most surprising result from this study is the observed diversity in terms of both cluster and operon structure. Based on the different structures present in the data, the clustering of phenylacetate degradation genes has occurred repeatedly in several different lineages, the clusters themselves are mosaics and are generally composed of genes that have been acquired from other species, either recently or relatively recently. Often, strains of the same species have very different cluster structures and indeed in the case of *E. coli* and *P. putida*, even though the clusters look similar, many of the genes cannot trace their most recent common ancestor to the same point. In other words, orthologous gene displacement is quite common, as is illegitimate recombination. This has been reported previously (Omelchenko et al., 2003) and it indicates that the selective pressure to form clusters is not so strong that clusters, once formed, become immutable or that clusters continue to become larger.

In general, operon destruction as well as operon formation is seen to occur in the dataset and there are a total of 33 unique operon structures. This suggests that either the selective advantage that accrues as a result of operon formation is not very strong and recombination followed by random genetic drift can successfully break up operons (a neutralist explanation) or that if indeed a selection pressure exists that drives operon formation, there exists another opposing selection pressure to split operons. It is also possible that a selective advantage could exist to create an operon, but subsequently this advantage is no longer present as the environment changes. Irrespective of the explanation, it seems that for this particular pathway, the formation of large operons containing most or all of the genes is not necessarily hugely important, or perhaps it is not possible. The exception to the rule is seen in the *paaABCDE* operon, which is strongly conserved. The obvious explanation is that these proteins products physically interact and their existence in equimolar concentrations is necessary. Therefore, there is a gradient of selective pressure for co-regulation which is strongest for interacting proteins in our small dataset, less strong for proteins that do not physically interact and indeed co-regulation might be a selective disadvantage in some cases (in 14 out of 22 cases *paaZ* is in a cluster but not in an operon) and may lead to the successful destruction of an operon.

One strong bias present in the data is the general absence of head-to-head orientation of genes in these clusters (table 4.3). There is no obvious reason for the relative absence of head-to-head orientation of genes, since genes are frequently found on opposite strands of DNA and are often in a head-to-head orientation (personal observation). However, within these clusters, the number of times a head-to-head orientation of genes in this pathway is observed is 1 time out of a total of 214 observed *paa* gene pairs. Tail-to-tail orientations are more frequent, but also relatively rare, occurring only 20 times. Head to tail orientations dominate the data, with a total of 193 such pairs. By random chance we would expect to observe equal numbers of head-to-head, tail-to-tail and head-to-tail orientations of cistrons. Operon structures account for the majority of the bias in favour of a head-to-tail orientation as it is a requisite for membership in an operon. This does not explain the near absence of head-to-head arrangements observed and it is likely that there is a selective pressure that prevents particular arrangements of operons and single genes with respect to one another. Exactly what this selective pressure is remains unclear, but is possibly related to collisions of transcription apparatuses.

The study also sheds some light on the various models of cluster and operon formation. The expectation from the natal model of operon growth is that all genes in the operon are evolutionarily related. This theory is clearly insufficient to account for the observations in this analysis.

The selfish operon model (SOM) posits that operons exist so that they can be easily transferred via horizontal gene transfer. The analysis shows that there is evidence of gene replacement within a cluster and within an operon and this presents a difficulty with the hypothesis that operons exist in order to facilitate their transfer as a group. Additionally, the sheer diversity of operons present in the analysis is at odds with the SOM. There are 33 unique operon structures. Even the clusters of *E. coli K12/W* and *P. putida*, which are clearly homologous, differ in gene content, order, operon structure and show evidence of orthologous replacement via HGT. While it is not in doubt that there is an advantage to passing a set of genes horizontally, the results show little evidence of selfish operon style transfers. The only stable operon structure is that of *paaADCDE* and this is an example of an operon that cannot exist outside of a selfish operon framework, since the gene products form a complex with one another. In addition, Pál and Hurst have already shown that essential genes are more likely to be in an operon than non-essential genes and this is also incompatible with the SOM (Pál and Hurst, 2004).

The Fisher model states that cluster formation is a way of keeping co-adapted alleles together. It is clear from the analysis that the turnover rate of alleles is high and alleles do not seem to spend much time being inherited together and so this model is not compatible with the data.

The co-regulation model, while recently receiving some support from an analysis of operons only (Price et al., 2005b) is also insufficient to cover some of the observations of this analysis. Many genes present in a cluster, but not in an operon 38 times. Genes are located in operons 225 times, however, 119 of those times the operon is the *paaABCDE* operon, which contains genes that form a single heteromeric complex. The co-regulation model only governs operon maintenance and is strongly in operation for the maintenance of *paaABCDE* but is still insufficient to explain all the data.

The protein immobility model (PIM) fits with the idea that there is a small selective advantage for clustering genes together. The reason for this small selective advantage is the effect macromolecular crowding has on the movement of proteins in the cell. Macromolecular crowding tends to increase the speed of biochemical reactions (Ellis, 2001), whilst simultaneously limiting the ability of large proteins to move around the cell. While the cellular matrix is a dynamic environment, the movement of a protein through the cytoplasm of a prokaryote is slower than through water (Elowitz et al., 1999) and when several proteins are involved, this is likely to result in sufficient restriction of movement that a selective advantage accrues for an organism that synthesizes functionally related proteins in close proximity to one another. However, the PIM only covers the formation of clusters and does not cover operon formation and maintenance. Based on the data, operon formation and maintenance is not an inevitable consequence of cluster formation, perhaps simplifying transcription.

The persistence model is somewhat difficult to apply to the data as it does not consider clustering in terms of functional relatedness but rather in terms of how widely distributed genes are. Given that there is a high level of clustering of *paa* genes with one another this implies that there is selection for clustering based on membership of a common metabolic pathway, independent of how widely distributed the genes are. The different genes involved in the pathway show large variation in terms of their distribution, however, excluding *paaL* and *paaM,* which are specific to the *P. putida* cluster, all genes occur in multiple clusters. As such, the persistence model does not explain the observed data.

A number of studies have indicated that transcriptional control of independent transcription units (single genes and operons) is likely to have influenced genomic structure (Hershberg et al., 2005). This is reflected in the co-localisation of genes that are controlled by the same transcription factor. Additionally, the distribution and orientation of transcription units is not random (Warren and ten Wolde, 2004) and is associated with an optimisation process. In this study it is evident that while these genome optimisation processes are under way, the process of horizontal gene transfer and within-cluster gene content perturbation is continuous and at times fairly radical.

It is important to note that no one model of operon assembly completely covers the observations of this analysis. Perhaps a more robust model would be one that deals with cluster and operon formation as different levels of organisation. Operon formation occurs subsequent or at the same time as cluster formation, however, the data clearly show that operon formation is not absolutely necessary. The majority of genes in clusters also in operons, but this is likely to be a secondary advantage ensuring that they are transcribed at the same time. A more comprehensive model requires a component that provides a selective advantage for moving genes closer together in a genome and a separate component providing selective advantage for operon formation. In terms of the current models, the best fit would be a combination of the PIM and the co-regulation model.

Perhaps more important, however, is evolutionary history of the genes of the phenylacetate degradation pathway. The massive diversity of the clusters and operons observed, coupled with complete lack of correlation to phylogeny, provides an interesting insight into just how dynamic is the process rearranging the position of genes in a genome. While this is only a single pathway, the evidence still strongly implies the existence of a complicated underlying system in prokaryotes based upon a recombination selection balance. Even if phenylacetate degradation is unusual when compared to clusters associated with amino acid biosynthesis or other core pathways, it may provide a much deeper understanding of the principles of cluster and operon formation than static, widely distributed gene clusters ever could.

Chapter 5: Discussion

In their 2006 paper, Ciccarelli et al. stated that “reconstructing the phylogenetic relationships among all living organisms is one of the fundamental challenges in biology” (Ciccarelli et al., 2006). However, using an alignment of 31 concatenated genes they produced what they deemed a “highly resolved Tree of Life”. This tree came under scrutiny for the small number of genes used to derive the phylogeny, less than 1 percent of the genes in the average prokaryotic genome (Dagan and Martin, 2006).

The existence of a tree of life continues to be the center of intense debate. Some believe they have found it (Ciccarelli et al., 2006), many believe it cannot exist in the face of the processes that underpin evolution in the prokaryotic division (Dagan and Martin, 2006; Bapteste et al., 2008). Regardless of whether a tree of life exists, there is no denying that defining a tree like phylogeny for certain groups has proved troublesome.

The eukaryotic domain of the tree of life has its own difficulties, typified by debates such as the Ecdysozoa versus Coelomata hyopotheses (Philip et al., 2005; Phillipe et al., 2005) and the origins of the eukaryotic genetic apparatus itself (Cox et al., 2008). Even in the absence of high levels of HGT and genome re-organisation seen in the prokaryotic world, these problems do pose serious difficulties for defining a tree of life (Rokas and Carroll, 2006). It should be noted, however, that as methods advance and sampling density increases there is every chance that many of the ongoing debates concerning the eukaryotic groups will be settled one day.

For a tree of life to exist we must be able to define the relationships between all biological organisms while keeping within a tree-like framework. Nowhere is this more difficult than in the prokaryotic world, where HGT disrupts the underlying signal of vertical inheritance (Bapteste et al., 2008). In recent work, Bapteste et al. used careful methods for selecting core genes from which to build a prokaryotic phylogeny with (Bapteste et al., 2008). They found that only 0.7 percent of the average prokaryotic genome could be used to build a prokaryotic phylogeny and that even then it was safer to assume a "comb-like" structure rather than a tree like one.

In this thesis, I look at another problem in constructing a tree of life. What happens when we explore a small set of closely related genera, species and strains? Can a tree like structure be found for a group of closely related genomes, and if not, then at what point does a tree-like structure break down?

The answer is of course dependent on both the data and the methods, and for this reason different methods and different portions of the genome were examined for evidence of tree-like signal. Ultimately, the traditional classification of members of the YESS group was of little importance, the question merely distilled to finding divisions within a group of genomes.

Using different combinations of methods and data, different answers are produced, often with weak support. There is little reason to trust one method over another. Using all the single-gene families results in similar trees for the concatenated alignment and supertree, but are these methods more reliable than constructing phylogenies of 16S or housekeeping genes? Possibly, given that an issue with latter two approaches is that such genes are likely to be strongly conserved and thus provide little in the way of phylogenetic signal. However it is equally valid to argue that the single gene families are more likely to have undergone HGT and thus their phylogenies represent gene trees rather than species trees. Certainly by looking at tree-to-tree distances between the trees for the single-gene families and the supertree there is evidence of both signal and conflict. So in the end, the answer would appear to be that there is no answer, at least not using these data and methods.

This has implications for the tree of life problem. While the result is specific to this group of genomes, at this density of sampling, it highlights nonetheless an oncoming problem in the world of prokaryotic phylogenetics. At some point the resolution of shallow phylogeny will decrease as sampling density increases. When do we consider two genomes different? As sampling density increases how do we continue to separate organisms into strains, species and genera? Is it a better idea simply to let the current methods define the boundaries between genomes with as much resolution as possible, instead of trying to fit everything into the framework laid down by traditional classification methods? Perhaps, but change on that scale is unlikely to occur anytime soon. Instead, many will continue the search for a tree of life, regardless of the difficulties presented along the way.

Examining gene organisation and clustering within a genome presents a challenge that is no less difficult. In chapter 3 and chapter 4 I present studies of gene clustering from two different perspectives.

In chapter 3 the focus is identification and analysis of clusters across the γ-proteobacteria, using genes that are known to be widely distributed. From this perspective it is clear that clustering is commonplace in genes involved in amino acid biosynthesis. Of all the genes examined across the five superpathways of amino acid biosynthesis, only three showed no evidence of clustering. Looking at the co-occurrence data and the structure of the clusters themselves it is clear that some configurations are highly conserved/strongly selected for. For example, the *leuA, B, C* and *D* genes, which catalyse successive steps in their superpathway, are almost always found clustered in an operon with one another. A similar scenario can be seen with the *trpA, B, C, D* and *E* genes. This is likely because of the benefits of co-localisation and co-regulation, i.e. the products are present in the same place, at the same time.

The study presented in chapter 4 provides a different perspective on the process of gene clustering, specifically with relation to models of gene cluster and operon formation. Using phylogenetic analyses, a broader dataset and a set of genes showing a patchy phylogenetic distribution it was possible to test various hypotheses as to why genes cluster. Little evidence was found to support the Selfish Operon model or the Fisher model. Instead, given both the abundance of operons and gene clusters, it appears likely that a combination of co-regulation and co-localisation drive the clustering of of *paa* genes.

The movement of molecules through the cytoplasm is not yet well understood. Elowitz et al., have shown that the rate of diffusion is a multifactorial problem (Elowitz et al., 1998). They found that the rate of diffusion for green fluorescent protein (GFP) was linked to its concentration, which they suggested was possibly due to dimerisation of the GFP at higher concentrations. Interestingly they also found that a major reduction in the rate of diffusion was dramatically reduced by the addition of a six-histidine tag to the GFP, suggesting apparently small sequence changes can massively restrict motility. The effective viscosity of a bacterial cell is significantly higher than that of a eukaryotic cell, possibly due to the presence of a nucleoid (Mullineaux et al., 2006). Recently it has been shown that some protein clusters are localised in the cytoplasm, and this localisation requires regulation (Thompson et al., 2006).

Clearly, movement through the cytoplasm is not always as simple as random diffusion. One conjecture is that the default state for proteins in the cytoplasm is that they are essentially non-motile due to the effects of macromolecular crowding. Proteins that need to move to specific areas of the cell, such as those involved in cell division, often have their own transport system in place, again highlighting the fact that movement through the cytoplasm must be difficult (Collier and Shapiro, 2007). If this is the case, then perhaps the selection for clustering for the co-localisation of products is actually quite strong.

However, regardless of whether or not selection for clustering is strong, it is not so strong that clusters become immutable. A common observation in both chapter 3 and chapter 4 was that no cluster structure was constant. Even the *paaA, B, C, D* and *E* cluster, whose products form a complex with one another, had four cases where the *paaE* gene was present in the genome but not clustered and two cases where it was not present at all. Similarly, clusters such as the *leu* and *trp* gene clusters were not always conserved. This implies that even the clusters that are likely to be under the strongest selection can change.

Moving away from genes, such as the *leu* and *trp* genes, that showed a strong tendency to cluster and reside in operons, other genes demonstrated the fluidic nature of clustering. While only a subset of all potential clusters was observed across the superpathways examined in chapter 3, there was still huge diversity in the clusters observed, both in terms of cluster size and gene content. Differences in the distribution of non-functionally related genes within these clusters further highlighted this fact. When examining *paa* clusters, there is evidence of large clusters of genes being assembled multiple times independently. Co-occurrence data for *paa* genes shows that, with three exceptions (two of which relate to the fact that *paaL* and *M* are only present in one genome), almost all genes co-occur in at least one cluster.

Ultimately it appears clustering is commonplace in bacterial genome. Organising genes that are involved in the same biochemical pathway into close physical proximity in the genome is often beneficial. However, other processes, such as HGT can disrupt clusters.

This disruption could be selected for if the retention of the new genes introduced via HGT provides a greater increase in fitness than retention of the cluster structure. It is therefore unsurprising to see so many different cluster configurations in the data. The dynamic nature of the bacterial gene clusters mirrors the dynamic nature of bacterial genome.

The work presented in this thesis has added to the field of bacterial genomics in two areas. Firstly I have shown that shallow phylogeny is a serious problem in the YESS group, even with the availability of whole genome data. I have shown that different methods and data produce very different results for the same set of organisms. I have also demonstrated that using the 16S rRNA, a gene considered to be ideal for the recovery of species phylogenies, is unusable in this scenario due to likely horizontal gene conversion event between closely related strains within the YESS group. Previously, homogenisation of 16S rRNA genes was considered mainly to be a factor within individual genomes. This study suggests that it is likely to occur between closely related strains by horizontal processes.

Finally, I have provided a new model for gene clustering, taking observed data and previous models into account. I have shown that existing models do not fully explain the observed data, that different selection forces for cluster and operon formation exist and that this must be taken into account when creating a model to explain why functionally related genes co-localise. I demonstrate that the genes associated with the phenylacetate degradation pathway have a diverse range of associated clusters and operons. I have shown that large clusters of most or all *paa* genes have formed multiple times independently and this suggests that gene clustering is an extremely dynamic process, dictated by selective pressures in the environment and not a remnant of an ancestral genome organisation. However, a lot of questions still remain unanswered. I provide software that can be used to carry out future studies on gene clustering. Such studies will be required to refine our understanding of gene co-localisation in bacterial genomes.

Future Work:

A great deal of work remains to be done in the field of shallow genomics. In this thesis I studied the YESS group and found that there was no clear single phylogeny that described the group and concluded that this was a result of high levels of noise in the data due to HGT and gene conversion. However other questions still remain, such as how the problem applies to other closely related bacterial groups, what exactly the boundary is for resolving shallow phylogenies and what methods can we develop to increase the level of resolution of such phylogenies? It would be interesting to see whether the conflicting results were due to other problems not considered in the study, such as model mis-specification producing incorrect phylogenies. In particular using tests such as the Goldman test (Goldman, 1993) might shed further light on how much influence the methodology had on the conflict present in the derived phylogenies.

In this thesis I describe a new software tool, GenClust, for the identification of gene clusters in bacterial genomes and its use in the analysis of a metabolic pathway whose genes are sometimes found in large clusters. Many improvements could be made it the GenClust algorithm. At present it is designed to show the user all potential clusters. However it is likely that many clusters returned from the analysis are false positives. While they technically fit the criteria of the search, they are not the genes the user is interested in. As such it is then up to the user to manually curate the results to identify true clusters. In future versions of GenClust this situation will be improved. This can be achieved by taking factors such as query-to-hit coverage levels and percent identity of hit into account, to compliment the current e-value based assessment of homology. Allowing the user to set values for coverage and percent identity would allow them much greater control of the strictness of the search and could greatly reduce the number of false positive results and therefore the level of manual curation required. A statistical measure of cluster significance could be added to help reduce time spent analysing the results. Such a measure would take into consideration gene content, individual hit coverage, similarity and percent identity to come up with a significance value for all reported clusters. In addition this could be further complemented by the option for the user to specify a reference cluster to search for. The significance test could then take into consideration gene order in addition to allowing more emphasis to be placed on gene content. This would allow an even higher level of distillation of meaningful results.

Finally, many questions remain unanswered in terms of why gene clusters are so prevalent in bacterial genomes. I present in this thesis a model agrees with the observed data. Conversely I show that none of the previous models completely explain the observed data. This was achieved by looking at the genes associated with a single, though very interesting, metabolic pathway. Further insight would be gained by looking at other evolutionarily unconserved gene clusters. Much like excluding the constant sites in a multiple sequence alignment, excluding highly conserved gene clusters and focusing on the divergent clusters may further refine our understanding of why and how gene clusters form.

Chapter 6 – Bibliography

ACHTMAN, M. & WAGNER, M. (2008) Microbial diversity and the genetic nature of microbial species. *Nat Rev Microbiol,* 6**,** 431-40.

ADAMS, M. D., CELNIKER, S. E., HOLT, R. A., EVANS, C. A., GOCAYNE, J. D., AMANATIDES, P. G., SCHERER, S. E., LI, P. W., HOSKINS, R. A., GALLE, R. F., GEORGE, R. A., LEWIS, S. E., RICHARDS, S., ASHBURNER, M., HENDERSON, S. N., SUTTON, G. G., WORTMAN, J. R., YANDELL, M. D., ZHANG, Q., CHEN, L. X., BRANDON, R. C., ROGERS, Y. H., BLAZEJ, R. G., CHAMPE, M., PFEIFFER, B. D., WAN, K. H., DOYLE, C., BAXTER, E. G., HELT, G., NELSON, C. R., GABOR, G. L., ABRIL, J. F., AGBAYANI, A., AN, H. J., ANDREWS-PFANNKOCH, C., BALDWIN, D., BALLEW, R. M., BASU, A., BAXENDALE, J., BAYRAKTAROGLU, L., BEASLEY, E. M., BEESON, K. Y., BENOS, P. V., BERMAN, B. P., BHANDARI, D., BOLSHAKOV, S., BORKOVA, D., BOTCHAN, M. R., BOUCK, J., BROKSTEIN, P., BROTTIER, P., BURTIS, K. C., BUSAM, D. A., BUTLER, H., CADIEU, E., CENTER, A., CHANDRA, I., CHERRY, J. M., CAWLEY, S., DAHLKE, C., DAVENPORT, L. B., DAVIES, P., DE PABLOS, B., DELCHER, A., DENG, Z., MAYS, A. D., DEW, I., DIETZ, S. M., DODSON, K., DOUP, L. E., DOWNES, M., DUGAN-ROCHA, S., DUNKOV, B. C., DUNN, P., DURBIN, K. J., EVANGELISTA, C. C., FERRAZ, C., FERRIERA, S., FLEISCHMANN, W., FOSLER, C., GABRIELIAN, A. E., GARG, N. S., GELBART, W. M., GLASSER, K., GLODEK, A., GONG, F., GORRELL, J. H., GU, Z., GUAN, P., HARRIS, M., HARRIS, N. L., HARVEY, D., HEIMAN, T. J., HERNANDEZ, J. R., HOUCK, J., HOSTIN, D., HOUSTON, K. A., HOWLAND, T. J., WEI, M. H., IBEGWAM, C., et al. (2000) The genome sequence of Drosophila melanogaster. *Science,* 287**,** 2185-95.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. (1990) Basic local alignment search tool. *J Mol Biol,* 215**,** 403-10.

ALTSCHUL, S. F., MADDEN, T. L., SCHAFFER, A. A., ZHANG, J., ZHANG, Z., MILLER, W. & LIPMAN, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res,* 25**,** 3389-402.

AMANN, R. I., LUDWIG, W. & SCHLEIFER, K. H. (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev,* 59**,** 143-69.

ANDERSSON, J. O. (2005) Lateral gene transfer in eukaryotes. *Cell Mol Life Sci,* 62**,** 1182-97.

ANDREWS, J., SMITH, M., MERAKOVSKY, J., COULSON, M., HANNAN, F. & KELLY, L. E. (1996) The stoned locus of Drosophila melanogaster produces a dicistronic transcript and encodes two distinct polypeptides. *Genetics,* 143, 1699-711.

ARCHIE, J.W. (1989) A randomisation test for phylogenetic information in systematic data. Syst. Zoo., 38, 239–252.

ARRIO-DUPONT, M., FOUCAULT, G., VACHER, M., DEVAUX, P. F. & CRIBIER, S. (2000) Translational diffusion of globular proteins in the cytoplasm of cultured muscle cells. *Biophys J,* 78, 901-7.

ASAI, T., ZAPOROJETS, D., SQUIRES, C. & SQUIRES, C. L. (1999) An Escherichia coli strain with all chromosomal rRNA operons inactivated: complete exchange of rRNA genes between bacteria. *Proc Natl Acad Sci U S A,* 96, 1971-6.

BALDAUF, S. L. (1999) A Search for the Origins of Animals and Fungi: Comparing and Combining Molecular Data. *Am Nat,* 154, S178-S188.

BAPTESTE, E., BRINKMANN, H., LEE, J. A., MOORE, D. V., SENSEN, C. W., GORDON, P., DURUFLE, L., GAASTERLAND, T., LOPEZ, P., MULLER, M. & PHILIPPE, H. (2002) The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba. *Proc Natl Acad Sci U S A,* 99, 1414-9.

BAPTESTE, E., SUSKO, E., LEIGH, J., RUIZ-TRILLO, I., BUCKNAM, J. & DOOLITTLE, W. F. (2008) Alternative methods for concatenation of core genes indicate a lack of resolution in deep nodes of the prokaryotic phylogeny. *Mol Biol Evol,* 25, 83-91.

BAUM, B. R. (1992). Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon 41*, 3-10.

BEIKO, R. G., HARLOW, T. J. & RAGAN, M. A. (2005) Highways of gene sharing in prokaryotes. *Proc Natl Acad Sci U S A,* 102, 14332-7.

BENSON, D. A., KARSCH-MIZRACHI, I., LIPMAN, D. J., OSTELL, J. & WHEELER, D. L. (2008) GenBank. *Nucleic Acids Res,* 36, D25-30.

BENTLEY, R. & MEGANATHAN, R. (1982) Biosynthesis of vitamin K (menaquinone) in bacteria. *Microbiol Rev,* 46, 241-80.

BLUMENTHAL, T. (1995) Trans-splicing and polycistronic transcription in Caenorhabditis elegans. *Trends Genet,* 11, 132-6.

BLUMENTHAL, T. (2004) Operons in eukaryotes. *Brief Funct Genomic Proteomic,* 3**,** 199-211.

BLUMENTHAL, T., EVANS, D., LINK, C. D., GUFFANTI, A., LAWSON, D., THIERRY-MIEG, J., THIERRY-MIEG, D., CHIU, W. L., DUKE, K., KIRALY, M. & KIM, S. K. (2002) A global analysis of Caenorhabditis elegans operons. *Nature,* 417**,** 851-4.

BLUMENTHAL, T. & GLEASON, K. S. (2003) Caenorhabditis elegans operons: form and function. *Nat Rev Genet,* 4**,** 112-20.

BLUNDELL, M., CRAIG, E. & KENNELL, D. (1972) Decay rates of different mRNA in E. coli and models of decay. *Nat New Biol,* 238**,** 46-9.

BODMER, W. F. & P. A. PARSONS P. A. (1962) Linkage and recombination in evolution. Adv. Genet. 11: 1-100.

BOTSTEIN, D. (1980) A theory of modular evolution for bacteriophages. Ann. NYAcad. Sci. 354. 484-491.

BOYD, E. F., WANG, F. S., WHITTAM, T. S. & SELANDER, R. K. (1996) Molecular genetic relationships of the salmonellae. *Appl Environ Microbiol,* 62**,** 804-8.

BRENNER, D. J. (1984) Enterobacteriaceae. In *Bergey's Manual of Systematic Bacteriology*, eds. Krieg, N. R. & Holt, J. G. (Williams and Wilkins, Baltimore), Vol. 1, pp. 408–420.

BUCKEE, C. O., JOLLEY, K. A., RECKER, M., PENMAN, B., KRIZ, P., GUPTA, S. & MAIDEN, M. C. (2008) Role of selection in the emergence of lineages and the evolution of virulence in Neisseria meningitidis. *Proc Natl Acad Sci U S A,* 105**,** 15082-7.

C. ELEGANS SEQUENCING CONSORTIUM (1998) Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science,* 282**,** 2012-8.

CAMPBELL, A. & D. BOTSTEIN, D. (1983) Evolution of lambdoid phages, pp. 365-380 in *Lumbda II,* edited by R. W. HENDRIX, J. W. ROBERTS, F. W. STAHL, and R. A. WEISBERG. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, *NY.*

CAMIN, J. H. and SOKAL, R. R. (1965). A method for deducing branching sequences in phylogeny. *Evolution* 19, 311-326.

CANBACK, B., TAMAS, I. & ANDERSSON, S. G. (2004) A phylogenomic study of endosymbiotic bacteria. *Mol Biol Evol,* 21**,** 1110-22.

CASJENS, S. (1974) Bacteriophage lambda *FIZ* gene protein: role in head assembly. J. Mol. Biol. 90: 1-23.

CASJENS, S., HATFULL, G. & HENDRIX, R. (1992) Evolution of dsDNA tailed-bacteriophage genomes. Virology 3: 383-397.

CAVALLI-SFORZA, L. L. and EDWARDS, A. W. F. (1967). Analysis of human evolution. In *Genetics Today. Proceedings of the XI International Congress of Genetics. The Hague, The Netherlands, September 1963., pp 923-933. Edited by S. J. Geerts*. Oxford: Perganom Press.

CAYLEY, S., LEWIS, B. A., GUTTMAN, H. J. & RECORD, M. T., JR. (1991) Characterization of the cytoplasm of Escherichia coli K-12 as a function of external osmolarity. Implications for protein-DNA interactions in vivo. *J Mol Biol,* 222**,** 281-300.

CICCARELLI, F. D., DOERKS, T., VON MERING, C., CREEVEY, C. J., SNEL, B. & BORK, P. (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science,* 311**,** 1283-7.

CILIA, V., LAFAY, B. & CHRISTEN, R. (1996) Sequence heterogeneities among 16S ribosomal RNA sequences, and their effect on phylogenetic analyses at the species level. *Mol Biol Evol,* 13**,** 451-61.

COLLIER, J. & SHAPIRO, L. (2007) Spatial complexity and control of a bacterial cell cycle. *Curr Opin Biotechnol,* 18**,** 333-40.

COMAS, I., MOYA, A. & GONZALEZ-CANDELAS, F. (2007) From phylogenetics to phylogenomics: the evolutionary relationships of insect endosymbiotic gamma-Proteobacteria as a test case. *Syst Biol,* 56**,** 1-16.

COOPER, J. E. & FEIL, E. J. (2004) Multilocus sequence typing--what is resolved? *Trends Microbiol,* 12**,** 373-7.

COOPER, T. G. (1996) Regulation of allantoin catabolism in Saccharomyces cerevisiae. in The Mycota III: Biochemistry and Molecular Biology (ed. Marzluf, G.A.) 139–169. Springer, Berlin.

CROSA, J. H., D. J. BRENNER, D. J., EWING, W. H. & FALKOW, S. (1973). Molecular relationships among the salmonelleae. J. Bacteriol. 115:307-315.

COX, C. J., FOSTER, P. G., HIRT, R. P., HARRIS, S. R. & EMBLEY, T. M. (2008) The archaebacterial origin of eukaryotes. *Proc Natl Acad Sci U S A,* 105**,** 20356-61.

CREEVEY, C. J., FITZPATRICK, D. A., PHILIP, G. K., KINSELLA, R. J., O'CONNELL, M. J., PENTONY, M. M., TRAVERS, S. A., WILKINSON, M. & MCINERNEY, J. O. (2004) Does a tree-like phylogeny only exist at the tips in the prokaryotes? *Proc Biol Sci,* 271**,** 2551-8.

CREEVEY, C. J. & MCINERNEY, J. O. (2005) Clann: investigating phylogenetic information through supertree analyses. *Bioinformatics,* 21**,** 390-2.

DAGAN, T., ARTZY-RANDRUP, Y. & MARTIN, W. (2008) Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc Natl Acad Sci U S A,* 105**,** 10039-44.

DAGAN, T. & MARTIN, W. (2006) The tree of one percent. *Genome Biol,* 7**,** 118.

DAGAN, T. & MARTIN, W. (2007) Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc Natl Acad Sci U S A,* 104**,** 870-5.

DANDEKAR, T., SNEL, B., HUYNEN, M. & BORK, P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci,* 23**,** 324-8.

DAUGA, C. (2002) Evolution of the gyrB gene and the molecular phylogeny of Enterobacteriaceae: a model molecule for molecular systematic studies. *Int J Syst Evol Microbiol,* 52**,** 531-47.

DARWIN, C. (1859). On the origin of Species by means of natural selection or the preservation of favoured races in the struggle for life. Murray, London.

DAVIS, R. E. & HODGSON, S. (1997) Gene linkage and steady state RNAs suggest trans-splicing may be associated with a polycistronic transcript in Schistosoma mansoni. *Mol Biochem Parasitol,* 89**,** 25-39.

DAYHOFF, M. O., SCHWARTZ, R. M. & ORCUTT, B. (1978). A model of evolutionary change in proteins. Atlas of Protein Sequence and Structure 5, 345-352.

DEMEREC, M. (1960) Frequency of deletions among spontaneous and induced mutations in *Salmonella*. Proc. Natl. Acad. Sci. USA 46: 1075-1079.

DEMEREC, M. & HARTMAN, Z. E. (1956) Tryptophan mutants in *Salmonella typhimurium*. Carnegie Inst. Washington Publ. 612 5-33.

DEMEREC, M. & HARTMAN, P.E. (1959) Complex Loci in Microorganisms. Annual Review of Microbiology: 13:377-406.

DOOLITTLE, W. F. (1999a) Lateral genomics. *Trends Cell Biol,* 9**,** M5-8.

DOOLITTLE, W. F. (1999b) Phylogenetic classification and the universal tree. *Science,* 284**,** 2124-9.

DOOLITTLE, W. F. & BAPTESTE, E. (2007) Pattern pluralism and the Tree of Life hypothesis. *Proc Natl Acad Sci U S A,* 104**,** 2043-9.

DOOLITTLE, W. F. & PAPKE, R. T. (2006) Genomics and the bacterial species problem. *Genome Biol,* 7**,** 116.

DUBNAU, D. (1999) DNA uptake in bacteria. *Annu Rev Microbiol,* 53**,** 217-44.

EDGAR, R. C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics,* 5**,** 113.

EDWARDS, A. W. F. & CAVALLI-SFORZA, L. L. (1964). Reconstruction of evolutionary trees. In *Phenetic and Phylogenetic Classification*. Edited by V. H. Heywood & J. McNeill. London: Systematics Association Publ.

EFRON, B. (1979). Bootstrap methods: another look at the jackknife. Annals of Statistics 7:1–26.

ELLIS, R. J. (2001) Macromolecular crowding: an important but neglected aspect of the intracellular environment. *Curr Opin Struct Biol,* 11**,** 114-9.

ELOWITZ, M. B., SURETTE, M. G., WOLF, P. E., STOCK, J. B. & LEIBLER, S. (1999) Protein mobility in the cytoplasm of Escherichia coli. *J Bacteriol,* 181**,** 197-203.

ERMOLAEVA, M. D., WHITE, O. & SALZBERG, S. L. (2001) Prediction of operons in microbial genomes. *Nucleic Acids Res,* 29**,** 1216-21.

ESCOBAR-PARAMO, P., GIUDICELLI, C., PARSOT, C. & DENAMUR, E. (2003) The evolutionary history of Shigella and enteroinvasive Escherichia coli revised. *J Mol Evol,* 57**,** 140-8.

EWING, W. H. (1949) Shigella nomenclature. *J Bacteriol,* 57**,** 633-8.

FALUSH, D., KRAFT, C., TAYLOR, N. S., CORREA, P., FOX, J. G., ACHTMAN, M. & SUERBAUM, S. (2001) Recombination and mutation during long-term gastric colonization by Helicobacter pylori: estimates of clock rates, recombination size, and minimal age. *Proc Natl Acad Sci U S A,* 98**,** 15056-61.

FANI, R., BRILLI, M. & LIO, P. (2005) The origin and evolution of operons: the piecewise building of the proteobacterial histidine operon. *J Mol Evol,* 60**,** 378-90.

FANI, R., LIO, P., CHIARELLI, I. & BAZZICALUPO, M. (1994) The evolution of the histidine biosynthetic genes in prokaryotes: a common ancestor for the hisA and hisF genes. *J Mol Evol,* 38**,** 489-95.

FELSENSTEIN, J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* 27:401-410.

FELSENSTEIN, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39: 783–791.

FELSENSTEIN, J. (1988) Phylogenies from molecular sequences: inference and reliability. *Annu Rev Genet,* 22**,** 521-65.

FERRANDEZ, A., MINAMBRES, B., GARCIA, B., OLIVERA, E. R., LUENGO, J. M., GARCIA, J. L. & DIAZ, E. (1998) Catabolism of phenylacetic acid in Escherichia coli. Characterization of a new aerobic hybrid pathway. *J Biol Chem,* 273**,** 25974-86.

FERRIER, D. E. & HOLLAND, P. W. (2001) Ancient origin of the Hox gene cluster. *Nat Rev Genet,* 2**,** 33-8.

FISHER, R. A. (1930) *The Genetical Theory of Natural Selection*. Oxford University Press. Oxford.

FITCH, W. M. & MARGOLIASH, E. (1967) Construction of phylogenetic trees. *Science,* 155**,** 279-84.

FITZPATRICK, D. A., CREEVEY, C. J. & MCINERNEY, J. O. (2006) Genome phylogenies indicate a meaningful alpha-proteobacterial phylogeny and support a grouping of the mitochondria with the Rickettsiales. *Mol Biol Evol,* 23**,** 74-85.

FLEISCHMANN, R. D., ADAMS, M. D., WHITE, O., CLAYTON, R. A., KIRKNESS, E. F., KERLAVAGE, A. R., BULT, C. J., TOMB, J. F., DOUGHERTY, B. A., MERRICK, J. M. & ET AL. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science,* 269**,** 496-512.

FOX, G. E., WISOTZKEY, J. D. & JURTSHUK, P., JR. (1992) How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. *Int J Syst Bacteriol,* 42**,** 166-70.

FRASER, C., ALM, E. J., POLZ, M. F., SPRATT, B. G. & HANAGE, W. P. (2009) The bacterial species challenge: making sense of genetic and ecological diversity. *Science,* 323**,** 741-6.

FURST, J., RITTER, M., RUDZKI, J., DANZL, J., GSCHWENTNER, M., SCANDELLA, E., JAKAB, M., KONIG, M., OEHL, B., LANG, F., DEETJEN, P. & PAULMICHL, M. (2002) ICln ion channel splice variants in Caenorhabditis elegans: voltage dependence and interaction with an operon partner protein. *J Biol Chem,* 277**,** 4435-45.

GANOT, P., KALLESOE, T., REINHARDT, R., CHOURROUT, D. & THOMPSON, E. M. (2004) Spliced-leader RNA trans splicing in a chordate, Oikopleura dioica, with a compact genome. *Mol Cell Biol,* 24**,** 7795-805.

GARCIA-RIOS, M., FUJITA, T., LAROSA, P. C., LOCY, R. D., CLITHERO, J. M., BRESSAN, R. A. & CSONKA, L. N. (1997) Cloning of a polycistronic cDNA from tomato encoding gamma-glutamyl kinase and gamma-glutamyl phosphate reductase. *Proc Natl Acad Sci U S A,* 94**,** 8249-54.

GEVERS, D., COHAN, F. M., LAWRENCE, J. G., SPRATT, B. G., COENYE, T., FEIL, E. J., STACKEBRANDT, E., VAN DE PEER, Y., VANDAMME, P., THOMPSON, F. L. & SWINGS, J. (2005) Opinion: Re-evaluating prokaryotic species. *Nat Rev Microbiol,* 3**,** 733-9.

GOLDMAN N. 1993. Statistical tests of models of DNA substitution. *J. Mol. Evol.* 36:182.

GOLDMAN, N., ANDERSON, J. P. & RODRIGO, A. G. (2000) Likelihood-based tests of topologies in phylogenetics. *Syst Biol,* 49**,** 652-70.

GOLDSCHMIDT, E. P., CATER, M. S., MATNEY, T. S., BUTLER, M. A. & GREENE, A. (1970) Genetic analysis of the histidine operon in Escherichia coli K12. *Genetics,* 66**,** 219-29.

GRAŒNEBERG, H. (1935) Gene doublets as evidence for adjacent small duplications in *Drosophila*. Nature 140: 932

GUILIANO D.B. & BLAXTER, M.L. (2006) Operon Conservation and the Evolution of *trans*-Splicing in the Phylum Nematoda. PLoS Genet 2(11): e198. doi:10.1371/journal.pgen.0020198

GUINDON, S. & GASCUEL, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol,* 52**,** 696-704.

GUO, Y., ZHENG, W., RONG, X. & HUANG, Y. (2008) A multilocus phylogeny of the Streptomyces griseus 16S rRNA gene clade: use of multilocus sequence analysis for streptomycete systematics. *Int J Syst Evol Microbiol,* 58**,** 149-59.

HAECKEL, E. (1879). The Evolution of Man. London.

HALL, B. G. (2001). *Phylogenetic trees made easy: A how-to manual for molecular biologists.*, First edn: Sinauer Associates Inc.

HASEGAWA, M., KISHINO, H. & YANO, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol,* 22**,** 160-74.

HEINEMANN, J. A. & SPRAGUE, G. F., JR. (1989). Bacterial conjugative plasmids mobilize DNA transfer between bacteria and yeast. *Nature,* 340**,** 205-9.

HENIKOFF, S., & HENIKOFF, J.G. (1992). Amino Acid Substitution Matrices from Protein Blocks. PNAS 89: 10915–10919.

HENNIG, W. (1966). Phylogenetic Systematic. Urbana: University of Illinois Press.

HERSHBERG, R., YEGER-LOTEM, E. & MARGALIT, H. (2005) Chromosomal organization is shaped by the transcription regulatory network. *Trends Genet,* 21**,** 138-42.

HOTOPP, J. C., CLARK, M. E., OLIVEIRA, D. C., FOSTER, J. M., FISCHER, P., TORRES, M. C., GIEBEL, J. D., KUMAR, N., ISHMAEL, N., WANG, S., INGRAM, J., NENE, R. V., SHEPARD, J., TOMKINS, J., RICHARDS, S., SPIRO, D. J., GHEDIN, E., SLATKO, B. E., TETTELIN, H. & WERREN, J. H. (2007) Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science,* 317**,** 1753-6.

HUDAULT, S., GUIGNOT, J. & SERVIN, A. L. (2001) Escherichia coli strains colonising the gastrointestinal tract protect germfree mice against Salmonella typhimurium infection. *Gut,* 49**,** 47-55.

HUELSENBECK, J. P., LARGET, B., MILLER, R. E. & RONQUIST, F. (2002) Potential applications and pitfalls of Bayesian inference of phylogeny. *Syst Biol,* 51**,** 673-88.

HUELSENBECK, J. P. & RONQUIST, F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics,* 17**,** 754-5.

ISMAIL, W., EL-SAID MOHAMED, M., WANNER, B. L., DATSENKO, K. A., EISENREICH, W., ROHDICH, F., BACHER, A. & FUCHS, G. (2003) Functional genomics by NMR spectroscopy. Phenylacetate catabolism in Escherichia coli. *Eur J Biochem,* 270**,** 3047-54.

ITOH, T., TAKEMOTO, K., MORI, H. & GOJOBORI, T. (1999) Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Mol Biol Evol,* 16**,** 332-46.

JACOB, F. & MONOD, J. (1962) On the regulation of gene activity. Cold Spring Harbor Symp. Quant. Biol. 26: 193-211.

JACOB, F., PERRIN, D., SANCHEZ, C. & MONOD, J. (1960) [Operon: a group of genes with the expression coordinated by an operator.]. *C R Hebd Seances Acad Sci,* 250**,** 1727-9.

JAIN, R., RIVERA, M. C. & LAKE, J. A. (1999) Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A,* 96**,** 3801-6.

JAIN, R., RIVERA, M. C., MOORE, J. E. & LAKE, J. A. (2002) Horizontal gene transfer in microbial genome evolution. *Theor Popul Biol,* 61**,** 489-95.

JENSEN, L. J., KUHN, M., STARK, M., CHAFFRON, S., CREEVEY, C., MULLER, J., DOERKS, T., JULIEN, P., ROTH, A., SIMONOVIC, M., BORK, P. & VON MERING, C. (2009) STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res,* 37**,** D412-6.

JUKES, T. H. & CANTOR, C. R. (1969). Evolution of protein molecules. In *Mammalian Protein Metabolism*, pp. 21-132. Edited by M. N. Munro. New York: Academic Press.

KAPER, J. B. (2005) Pathogenic Escherichia coli. *Int J Med Microbiol,* 295**,** 355-6.

KEANE, T. M., CREEVEY, C. J., PENTONY, M. M., NAUGHTON, T. J. & MCLNERNEY, J. O. (2006) Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol Biol,* 6**,** 29.

KEANE, T. M., NAUGHTON, T. J. & MCINERNEY, J. O. (2007) MultiPhyl: a high-throughput phylogenomics webserver using distributed computing. *Nucleic Acids Res,* 35**,** W33-7.

KENNEDY, M. & PAGE, R. D. M. (2002) *Seabird supertrees: combining partial estimates of procellariiform phylogeny*. The Auk, 119 (1). pp. 88-108.

KESELER, I. M., COLLADO-VIDES, J., GAMA-CASTRO, S., INGRAHAM, J., PALEY, S., PAULSEN, I. T., PERALTA-GIL, M. & KARP, P. D. (2005) EcoCyc: a comprehensive database resource for Escherichia coli. *Nucleic Acids Res,* 33**,** D334-7.

KIDGELL, C., REICHARD, U., WAIN, J., LINZ, B., TORPDAHL, M., DOUGAN, G. & ACHTMAN, M. (2002) Salmonella typhi, the causative agent of typhoid fever, is approximately 50,000 years old. *Infect Genet Evol,* 2**,** 39-45.

KIMURA, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol,* 16**,** 111-20.

KINSELLA, R. J., FITZPATRICK, D. A., CREEVEY, C. J. & MCINERNEY, J. O. (2003) Fatty acid biosynthesis in Mycobacterium tuberculosis: lateral gene transfer, adaptive evolution, and gene duplication. *Proc Natl Acad Sci U S A,* 100**,** 10320-5.

KISHINO, H. & HASEGAWA, M. (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J Mol Evol,* 29**,** 170-9.

KONDO, N., NIKOH, N., IJICHI, N., SHIMADA, M. & FUKATSU, T. (2002) Genome fragment of Wolbachia endosymbiont transferred to X chromosome of host insect. *Proc Natl Acad Sci U S A,* 99**,** 14280-5.

KONOPKA, M. C., SHKEL, I. A., CAYLEY, S., RECORD, M. T. & WEISSHAAR, J. C. (2006) Crowding and confinement effects on protein diffusion in vivo. *J Bacteriol,* 188**,** 6115-23.

KORBEL, J. O., SNEL, B., HUYNEN, M. A. & BORK, P. (2002) SHOT: a web server for the construction of genome phylogenies. *Trends Genet,* 18**,** 158-62.

KOTLOFF, K. L., WINICKOFF, J. P., IVANOFF, B., CLEMENS, J. D., SWERDLOW, D. L., SANSONETTI, P. J., ADAK, G. K. & LEVINE, M. M. (1999) Global burden of Shigella infections: implications for vaccine development and implementation of control strategies. *Bull World Health Organ,* 77**,** 651-66.

KUNST, F., OGASAWARA, N., MOSZER, I., ALBERTINI, A. M., ALLONI, G., AZEVEDO, V., BERTERO, M. G., BESSIERES, P., BOLOTIN, A., BORCHERT, S., BORRISS, R., BOURSIER, L., BRANS, A., BRAUN, M., BRIGNELL, S. C., BRON, S., BROUILLET, S., BRUSCHI, C. V., CALDWELL, B., CAPUANO, V., CARTER, N. M., CHOI, S. K., CODANI, J. J., CONNERTON, I. F., DANCHIN, A. & ET AL. (1997) The complete genome sequence of the gram-positive bacterium Bacillus subtilis. *Nature,* 390**,** 249-56.

KURLAND, C. G., CANBACK, B. & BERG, O. G. (2003) Horizontal gene transfer: a critical view. *Proc Natl Acad Sci U S A,* 100**,** 9658-62.

LAN, R., LUMB, B., RYAN, D. & REEVES, P. R. (2001) Molecular evolution of large virulence plasmid in Shigella clones and enteroinvasive Escherichia coli. *Infect Immun,* 69**,** 6303-9.

LANAVE, C., PREPARATA, G., SACCONE, C. & SERIO, G. (1984) A new method for calculating evolutionary substitution rates. *J Mol Evol,* 20**,** 86-93.

LANDER, E. S., LINTON, L. M., BIRREN, B., NUSBAUM, C., ZODY, M. C., BALDWIN, J., DEVON, K., DEWAR, K., DOYLE, M., FITZHUGH, W., FUNKE, R., GAGE, D., HARRIS, K., HEAFORD, A., HOWLAND, J., KANN, L., LEHOCZKY, J., LEVINE, R., MCEWAN, P., MCKERNAN, K., MELDRIM, J., MESIROV, J. P., MIRANDA, C., MORRIS, W., NAYLOR, J., RAYMOND, C., ROSETTI, M., SANTOS, R., SHERIDAN, A., SOUGNEZ, C., STANGE-THOMANN, N., STOJANOVIC, N., SUBRAMANIAN, A., WYMAN, D., ROGERS, J., SULSTON, J., AINSCOUGH, R., BECK, S., BENTLEY, D., BURTON, J., CLEE, C., CARTER, N., COULSON, A., DEADMAN, R., DELOUKAS, P., DUNHAM, A., DUNHAM, I., DURBIN, R., FRENCH, L., GRAFHAM, D., GREGORY, S., HUBBARD, T., HUMPHRAY, S., HUNT, A., JONES, M., LLOYD, C., MCMURRAY, A., MATTHEWS, L., MERCER, S., MILNE, S., MULLIKIN, J. C., MUNGALL, A., PLUMB, R., ROSS, M., SHOWNKEEN, R., SIMS, S., WATERSTON, R. H., WILSON, R. K., HILLIER, L. W., MCPHERSON, J. D., MARRA, M. A., MARDIS, E. R., FULTON, L. A., CHINWALLA, A. T., PEPIN, K. H., GISH, W. R., CHISSOE, S. L., WENDL, M. C., DELEHAUNTY, K. D., MINER, T. L., DELEHAUNTY, A., KRAMER, J. B., COOK, L. L., FULTON, R. S., JOHNSON, D. L., MINX, P. J., CLIFTON, S. W., HAWKINS, T., BRANSCOMB, E., PREDKI, P., RICHARDSON, P., WENNING, S., SLEZAK, T., DOGGETT, N., CHENG, J. F., OLSEN, A., LUCAS, S., ELKIN, C., UBERBACHER, E., FRAZIER, M., et al. (2001) Initial sequencing and analysis of the human genome. *Nature,* 409**,** 860-921.

LAWRENCE, J. G. (1997) Selfish operons and speciation by gene transfer. *Trends Microbiol,* 5**,** 355-9.

LAWRENCE, J. G. (2002) Gene transfer in bacteria: speciation without species? *Theor Popul Biol,* 61**,** 449-60.

LAWRENCE, J. G. & OCHMAN, H. (1998) Molecular archaeology of the Escherichia coli genome. *Proc Natl Acad Sci U S A,* 95**,** 9413-7.

LAWRENCE, J. G. & ROTH, J. R. (1996) Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics,* 143**,** 1843-60.

LEE, S. J. (1991) Expression of growth/differentiation factor 1 in the nervous system: conservation of a bicistronic structure. *Proc Natl Acad Sci U S A,* 88**,** 4250-4.

LEE, T. M., CHANG, L. L., CHANG, C. Y., WANG, J. C., PAN, T. M., WANG, T. K. & CHANG, S. F. (2000) Molecular analysis of Shigella sonnei isolated from three well-documented outbreaks in school children. *J Med Microbiol,* 49**,** 355-60.

LEIGH, J. W., SUSKO, E., BAUMGARTNER, M. & ROGER, A. J. (2008) Testing congruence in phylogenomic analysis. *Syst Biol,* 57**,** 104-15.

LERAT, E., DAUBIN, V. & MORAN, N. A. (2003) From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria. *PLoS Biol,* 1**,** E19.

LERAT, E., DAUBIN, V., OCHMAN, H. & MORAN, N. A. (2005) Evolutionary origins of genomic repertoires in bacteria. *PLoS Biol,* 3**,** e130.

LEWIS, E. B. (1951) Pseudoallelism and gene evolution. Cold Spring Harbor Symp. Quant. Biol. 16: 159-174.

LEWIS, E. B. (1978) A gene complex controlling segmentation in Drosophila. *Nature,* 276**,** 565-70.

LIN, J. & GERSTEIN, M. (2000) Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. *Genome Res,* 10**,** 808-18.

LLOYD, G. T., DAVIS, K. E., PISANI, D., TARVER, J. E., RUTA, M., SAKAMOTO, M., HONE, D. W., JENNINGS, R. & BENTON, M. J. (2008) Dinosaurs and the Cretaceous Terrestrial Revolution. *Proc Biol Sci,* 275**,** 2483-90.

LOCKHART, P. J., STEEL, M. A., HENDY, M. D. & PENNY, D. (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol Biol Evol,* 11**,** 605-12.

LOYTYNOJA, A. & GOLDMAN, N. (2008) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science,* 320**,** 1632-5.

LUENGO, J. M., GARCIA, J. L. & OLIVERA, E. R. (2001) The phenylacetyl-CoA catabolon: a complex catabolic unit with broad biotechnological applications. *Mol Microbiol,* 39**,** 1434-42.

MADDISON, D. R., SWOFFORD, D. L. & MADDISON, W. P. (1997) NEXUS: an extensible file format for systematic information. *Syst Biol,* 46**,** 590-621.

MAHON, B. E., PONKA, A., HALL, W. N., KOMATSU, K., DIETRICH, S. E., SIITONEN, A., CAGE, G., HAYES, P. S., LAMBERT-FAIR, M. A., BEAN, N. H., GRIFFIN, P. M. & SLUTSKER, L. (1997) An international outbreak of Salmonella infections caused by alfalfa sprouts grown from contaminated seeds. *J Infect Dis,* 175**,** 876-82.

MAJEWSKI, J., ZAWADZKI, P., PICKERILL, P., COHAN, F. M. & DOWSON, C. G. (2000) Barriers to genetic exchange between bacterial species: Streptococcus pneumoniae transformation. *J Bacteriol,* 182**,** 1016-23.

MANIATIS, T., FRITSCH, E. F., LAUER, J. & LAWN, R. M. (1980) The molecular genetics of human hemoglobins. *Annu Rev Genet,* 14**,** 145-78.

MARTIN, R. G., BERBERICH, M. A., AMES, B. N., DAVIS, W. W., GOLDBERGER, R. F. & YOURNO, J. D. (1971) Enzymes and Intermediates of Histidine Biosynthesis in Salmonella typhimurium. In: Methods in Enzymology, Vol. 17B, H. Tabor and C. W. Tabor, eds. (Academic Press, New York, NY), pp. 3-44.

MARTIN, W., STOEBE, B., GOREMYKIN, V., HAPSMANN, S., HASEGAWA, M. & KOWALLIK, K. V. (1998) Gene transfer to the nucleus and the evolution of chloroplasts. *Nature,* 393**,** 162-5.

MATSUDA, D. & DREHER, T. W. (2006) Close spacing of AUG initiation codons confers dicistronic character on a eukaryotic mRNA. *Rna,* 12**,** 1338-49.

MCCANN, A., COTTON, J. A. & MCINERNEY, J. O. (2008) The tree of genomes: an empirical comparison of genome-phylogeny reconstruction methods. *BMC Evol Biol,* 8**,** 312.

MCINERNEY, J. O. (1998) Replicational and transcriptional selection on codon usage in Borrelia burgdorferi. *Proc Natl Acad Sci U S A,* 95**,** 10698-703.

MCINERNEY, J. O., COTTON, J. A. & PISANI, D. (2008) The prokaryotic tree of life: past, present. and future? *Trends Ecol Evol,* 23**,** 276-81.

MCQUISTON, J. R., HERRERA-LEON, S., WERTHEIM, B. C., DOYLE, J., FIELDS, P. I., TAUXE, R. V. & LOGSDON, J. M., JR. (2008) Molecular phylogeny of the salmonellae: relationships among Salmonella species and subspecies determined from four housekeeping genes and evidence of lateral gene transfer events. *J Bacteriol,* 190**,** 7060-7.

MIRKIN, B. G., FENNER, T. I., GALPERIN, M. Y. & KOONIN, E. V. (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol Biol,* 3**,** 2.

MUHICH, M. L. & BOOTHROYD, J. C. (1988) Polycistronic transcripts in trypanosomes and their accumulation during heat shock: evidence for a precursor role in mRNA synthesis. *Mol Cell Biol,* 8**,** 3837-46.

MULLINEAUX, C. W., NENNINGER, A., RAY, N. & ROBINSON, C. (2006) Diffusion of green fluorescent protein in three cell environments in Escherichia coli. *J Bacteriol,* 188**,** 3442-8.

MUSHEGIAN, A. R. & KOONIN, E. V. (1996) Gene order is not conserved in bacterial evolution. *Trends Genet,* 12**,** 289-90.

O'NEIL, D. M., BARON, L. S. & SYPHERD, P. S. (1969) Chromosomal location of ribosomal protein cistrons determined by intergeneric bacterial mating. *J Bacteriol,* 99**,** 242-7.

OCHMAN, H., LAWRENCE, J. G. & GROISMAN, E. A. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature,* 405**,** 299-304.

OCHMAN, H., LERAT, E. & DAUBIN, V. (2005) Examining bacterial species under the specter of gene transfer and exchange. *Proc Natl Acad Sci U S A,* 102 Suppl 1**,** 6595-9.

OLIVERA, E. R., MINAMBRES, B., GARCIA, B., MUNIZ, C., MORENO, M. A., FERRANDEZ, A., DIAZ, E., GARCIA, J. L. & LUENGO, J. M. (1998) Molecular characterization of the phenylacetic acid catabolic pathway in Pseudomonas putida U: the phenylacetyl-CoA catabolon. *Proc Natl Acad Sci U S A,* 95**,** 6419-24.

OMELCHENKO, M. V., MAKAROVA, K. S., WOLF, Y. I., ROGOZIN, I. B. & KOONIN, E. V. (2003) Evolution of mosaic operons by horizontal gene transfer and gene displacement in situ. *Genome Biol,* 4**,** R55.

OMOKOKO, B., JANTGES, U. K., ZIMMERMANN, M., REISS, M. & HARTMEIER, W. (2008) Isolation of the phe-operon from G. stearothermophilus comprising the phenol degradative meta-pathway genes and a novel transcriptional regulator. *BMC Microbiol,* 8**,** 197.

OVERBEEK, R., FONSTEIN, M., D'SOUZA, M., PUSCH, G. D. & MALTSEV, N. (1999) The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A,* 96**,** 2896-901.

PAGE, A. P. (1997) Cyclophilin and protein disulfide isomerase genes are co-transcribed in a functionally related manner in Caenorhabditis elegans. *DNA Cell Biol,* 16**,** 1335-43.

PAGE, R. D. M. & HOLMES, E. C (1998). *Molecular evolution: A phylogenetic approach*. Blackwell Science.

PAL, C. & HURST, L. D. (2004) Evidence against the selfish operon theory. *Trends Genet,* 20**,** 232-4.

PAL, C., PAPP, B. & LERCHER, M. J. (2005) Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet,* 37**,** 1372-5.

PAPADOPOULOS, D., SCHNEIDER, D., MEIER-EISS, J., ARBER, W., LENSKI, R. E. & BLOT, M. (1999) Genomic evolution during a 10,000-generation experiment with bacteria. *Proc Natl Acad Sci U S A,* 96**,** 3807-12.

PAPKE, R. T., ZHAXYBAYEVA, O., FEIL, E. J., SOMMERFELD, K., MUISE, D. & DOOLITTLE, W. F. (2007) Searching for species in haloarchaea. *Proc Natl Acad Sci U S A,* 104**,** 14092-7.

PARDEE, A. B., JACOB, F. & MONOD, J. (1959) The genetic control and cytoplasmic expression of “inducibility” in the synthesis of Pgalactosidase by *E. coli*. J. Mol. Biol. 1: 165-178.

PARADIS, S., BOISSINOT, M., PAQUETTE, N., BELANGER, S. D., MARTEL, E. A., BOUDREAU, D. K., PICARD, F. J., OUELLETTE, M., ROY, P. H. & BERGERON, M. G. (2005) Phylogeny of the Enterobacteriaceae based on genes encoding elongation factor Tu and F-ATPase beta-subunit. *Int J Syst Evol Microbiol,* 55**,** 2013-25.

PENDRAK, M. L. & PERRY, R. D. (1993) Proteins essential for expression of the Hms+ phenotype of Yersinia pestis. *Mol Microbiol,* 8**,** 857-64.

PERRY, R. D., LUCIER, T. S., SIKKEMA, D. J. & BRUBAKER, R. R. (1993) Storage reservoirs of hemin and inorganic iron in Yersinia pestis. *Infect Immun,* 61**,** 32-9.

PHILIP, G. K., CREEVEY, C. J. & MCINERNEY, J. O. (2005) The Opisthokonta and the Ecdysozoa may not be clades: stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the Coelomata than Ecdysozoa. *Mol Biol Evol,* 22**,** 1175-84.

PHILIPPE, H., LARTILLOT, N. & BRINKMANN, H. (2005) Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Mol Biol Evol,* 22**,** 1246-53.

PHILLIPS, M. J., DELSUC, F. & PENNY, D. (2004) Genome-scale phylogeny and the detection of systematic biases. *Mol Biol Evol,* 21**,** 1455-8.

PISANI, D., COTTON, J. A. & MCINERNEY, J. O. (2007) Supertrees disentangle the chimerical origin of eukaryotic genomes. *Mol Biol Evol,* 24**,** 1752-60.

PORTNOY, D. A. & MARTINEZ, R. J. (1985) Role of a plasmid in the pathogenicity of Yersinia species. *Curr Top Microbiol Immunol,* 118**,** 29-51.

PORWOLLIK, S., WONG, R. M. & MCCLELLAND, M. (2002) Evolutionary genomics of Salmonella: gene acquisitions revealed by microarray analysis. *Proc Natl Acad Sci U S A,* 99**,** 8956-61.

POSADA, D. & CRANDALL, K. A. (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics,* 14**,** 817-8.

PRICE, M. N., ARKIN, A. P. & ALM, E. J. (2006) The life-cycle of operons. *PLoS Genet,* 2**,** e96.

PRICE, M. N., HUANG, K. H., ALM, E. J. & ARKIN, A. P. (2005a) A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res,* 33**,** 880-92.

PRICE, M. N., HUANG, K. H., ARKIN, A. P. & ALM, E. J. (2005b) Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome Res,* 15**,** 809-19.

PUPO, G. M., KARAOLIS, D. K., LAN, R. & REEVES, P. R. (1997) Evolutionary relationships among pathogenic and nonpathogenic Escherichia coli strains inferred from multilocus enzyme electrophoresis and mdh sequence studies. *Infect Immun,* 65**,** 2685-92.

PUPO, G. M., LAN, R. & REEVES, P. R. (2000) Multiple independent origins of Shigella clones of Escherichia coli and convergent evolution of many of their characteristics. *Proc Natl Acad Sci U S A,* 97**,** 10567-72.

PURKHOLD, U., WAGNER, M., TIMMERMANN, G., POMMERENING-ROSER, A. & KOOPS, H. P. (2003) 16S rRNA and amoA-based phylogeny of 12 novel betaproteobacterial ammonia-oxidizing isolates: extension of the dataset and proposal of a new lineage within the nitrosomonads. *Int J Syst Evol Microbiol,* 53**,** 1485-94.

RAGAN, M. A. (1992) Phylogenetic inference based on matrix representation of trees. *Mol Phylogenet Evol,* 1**,** 53-8.

RAGAN, M. A. (2001a) Detection of lateral gene transfer among microbial genomes. *Curr Opin Genet Dev,* 11**,** 620-6.

RAGAN, M. A. (2001b) On surrogate methods for detecting lateral gene transfer. *FEMS Microbiol Lett,* 201**,** 187-91.

REEVES, M. W., EVINS, G. M., HEIBA, A. A., PLIKAYTIS, B. D. & FARMER, J. J., 3RD (1989) Clonal nature of Salmonella typhi and its genetic relatedness to other salmonellae as shown by multilocus enzyme electrophoresis, and proposal of Salmonella bongori comb. nov. *J Clin Microbiol,* 27**,** 313-20.

RISON, S. C., TEICHMANN, S. A. & THORNTON, J. M. (2002) Homology, pathway distance and chromosomal localization of the small molecule metabolism enzymes in Escherichia coli. *J Mol Biol,* 318**,** 911-32.

RIVERA, M. C., JAIN, R., MOORE, J. E. & LAKE, J. A. (1998) Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci U S A,* 95**,** 6239-44.

ROBINSON, D. & FOULDS, L. (1981) Comparison of phylogenetic trees. *Biosciences*, 53:131-147.

RODRIGUEZ, F., OLIVER, J. L., MARIN, A. & MEDINA, J. R. (1990) The general stochastic model of nucleotide substitution. *J Theor Biol,* 142**,** 485-501.

RODRIGUEZ-EZPELETA, N., BRINKMANN, H., ROURE, B., LARTILLOT, N., LANG, B. F. & PHILIPPE, H. (2007) Detecting and overcoming systematic errors in genome-scale phylogenies. *Syst Biol,* 56**,** 389-99.

ROGOZIN, I. B., MAKAROVA, K. S., MURVAI, J., CZABARKA, E., WOLF, Y. I., TATUSOV, R. L., SZEKELY, L. A. & KOONIN, E. V. (2002) Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Res,* 30**,** 2212-23.

ROKAS, A. & CARROLL, S. B. (2006) Bushes in the tree of life. *PLoS Biol,* 4**,** e352.

ROKAS, A., WILLIAMS, B. L., KING, N. & CARROLL, S. B. (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature,* 425**,** 798-804.

RUTHERFORD, K., PARKHILL, J., CROOK, J., HORSNELL, T., RICE, P., RAJANDREAM, M. A. & BARRELL, B. (2000) Artemis: sequence visualization and annotation. *Bioinformatics,* 16**,** 944-5.

RZHETSKY, A. & NEI, M. (1993) Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Mol Biol Evol,* 10**,** 1073-95.

SAITOU, N. & NEI, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol,* 4**,** 406-25.

SANDERSON, M. J., DRISKELL, A. C., REE, R. H., EULENSTEIN, O. & LANGLEY, S. (2003) Obtaining maximal concatenated phylogenetic data sets from large sequence databases. *Mol Biol Evol,* 20**,** 1036-42.

SANSONETTI, P. J., KOPECKO, D. J. & FORMAL, S. B. (1981) Shigella sonnei plasmids: evidence that a large plasmid is necessary for virulence. *Infect Immun,* 34**,** 75-83.

SCHUBERT, S., RAKIN, A., KARCH, H., CARNIEL, E. & HEESEMANN, J. (1998) Prevalence of the "high-pathogenicity island" of *Yersinia* species among *Escherichia coli* strains that are pathogenic to humans. Infect Immun;66:480–5.

SCORNAVACCA, C., BERRY, V., LEFORT, V., DOUZERY, E. J. & RANWEZ, V. (2008) PhySIC_IST: cleaning source trees to infer more informative supertrees. *BMC Bioinformatics,* 9**,** 413.

SCHMIDT, T. & STOYE, J. (2007) Gecko and GhostFam: rigorous and efficient gene cluster detection in prokaryotic genomes**.** Methods Mol. Biol . 396:165-82.

SHIMODAIRA, H. (2002) An approximately unbiased test of phylogenetic tree selection. *Syst Biol,* 51**,** 492-508.

SHIMODAIRA, H. & HASEGAWA, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16: 1114–1116.

SHIMODAIRA, H. & HASEGAWA, M. (2001) CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics,* 17**,** 1246-7.

SHIVAJI, S., BHANU, N. V. & AGGARWAL, R. K. (2000) Identification of Yersinia pestis as the causative organism of plague in India as determined by 16S rDNA sequencing and RAPD-based genomic fingerprinting. *FEMS Microbiol Lett,* 189**,** 247-52.

SHOEMAKER, J. S., PAINTER, I. S. & WEIR, B. S. (1999) Bayesian statistics in genetics – a guide for the uninitiated. *Trends Genet*. **15,** pp. 354–358.

SIEFERT, J. L., MARTIN, K. A., ABDI, F., WIDGER, W. R. & FOX, G. E. (1997) Conserved gene clusters in bacterial genomes provide further support for the primacy of RNA. *J Mol Evol,* 45**,** 467-72.

SNEL, B., BORK, P. & HUYNEN, M. A. (1999) Genome phylogeny based on gene content. *Nat Genet,* 21**,** 108-10.

SOREK, R., ZHU, Y., CREEVEY, C. J., FRANCINO, M. P., BORK, P. & RUBIN, E. M. (2007) Genome-wide experimental determination of barriers to horizontal gene transfer. *Science,* 318**,** 1449-52.

SPIETH, J., BROOKE, G., KUERSTEN, S., LEA, K. & BLUMENTHAL, T. (1993) Operons in C. elegans: polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions. *Cell,* 73**,** 521-32.

STACKEBRANDT, E., FREDERIKSEN, W., GARRITY, G. M., GRIMONT, P. A., KAMPFER, P., MAIDEN, M. C., NESME, X., ROSSELLO-MORA, R., SWINGS, J., TRUPER, H. G., VAUTERIN, L., WARD, A. C. & WHITMAN, W. B. (2002) Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *Int J Syst Evol Microbiol,* 52**,** 1043-7.

STAHL, F. W. & MURRAY, N. E. (1966) The evolution of gene clusters and genetic circularity in microorganisms. *Genetics,* 53**,** 569-76.

STALEY, J. T. (2006) The bacterial species dilemma and the genomic-phylogenetic species concept. *Philos Trans R Soc Lond B Biol Sci,* 361**,** 1899-909.

STEEL, M. (2005) Should phylogenetic models be trying to "fit an elephant"? *Trends Genet,* 21**,** 307-9.

STEEL, M., and Rodrigo, A. (2008). Maximum likelihood supertrees. *Syst. Biol.* 57:243–250.

STRIMMER, K. & RAMBAUT, A. (2002) Inferring confidence sets of possibly misspecified gene trees. *Proc Biol Sci,* 269**,** 137-42.

SUCHARD, M. A., KITCHEN, C. M., SINSHEIMER, J. S. & WEISS, R. E. (2003) Hierarchical phylogenetic models for analyzing multipartite sequence data. *Syst Biol,* 52**,** 649-64.

SVETIC, R. E., MACCLUER, C. R., BUCKLEY, C. O., SMYTHE, K. L. & JACKSON, J. H. (2004) A metabolic force for gene clustering. *Bull Math Biol,* 66**,** 559-81.

SWOFFORD, D. L. (2003) *PAUP *: Phylogenetic analysis using parsimony (* and other methods), version 4.0b 10.* Sinauer Associates Sunderland, Massachusetts.

SYVANEN, M. & KADO, C. I., eds. (1998) *Horizontal Gene Transfer.* Chapman & Hall, London.

TACKET, C. O., BALLARD, J., HARRIS, N., ALLARD, J., NOLAN, C., QUAN, T. & COHEN, M. L. (1985) An outbreak of Yersinia enterocolitica infections caused by contaminated tofu (soybean curd). *Am J Epidemiol,* 121**,** 705-11.

TALAVERA, G. & CASTRESANA, J. (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol,* 56**,** 564-77.

TATUSOV, R. L., NATALE, D. A., GARKAVTSEV, I. V., TATUSOVA, T. A., SHANKAVARAM, U. T., RAO, B. S., KIRYUTIN, B., GALPERIN, M. Y., FEDOROVA, N. D. & KOONIN, E. V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res,* 29**,** 22-8.

THOMPSON, J. D., GIBSON, T. J. & HIGGINS, D. G. (2002) Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics,* Chapter 2**,** Unit 2 3.

THOMPSON, S. R., WADHAMS, G. H. & ARMITAGE, J. P. (2006) The positioning of cytoplasmic protein clusters in bacteria. *Proc Natl Acad Sci U S A,* 103**,** 8209-14.

TINDALL, B. J., GRIMONT, P. A., GARRITY, G. M. & EUZEBY, J. P. (2005) Nomenclature and taxonomy of the genus Salmonella. *Int J Syst Evol Microbiol,* 55**,** 521-4.

TREININ, M., GILLO, B., LIEBMAN, L. & CHALFIE, M. (1998) Two functionally dependent acetylcholine subunits are encoded in a single Caenorhabditis elegans operon. *Proc Natl Acad Sci U S A,* 95**,** 15492-5.

VAN DE GUCHTE, M., KOK, J. & VENEMA, G. (1991) Distance-dependent translational coupling and interference in Lactococcus lactis. *Mol Gen Genet,* 227**,** 65-71.

VANDAMME, P., POT, B., GILLIS, M., DE VOS, P., KERSTERS, K. & SWINGS, J. (1996) Polyphasic taxonomy, a consensus approach to bacterial systematics. *Microbiol Rev,* 60**,** 407-38.

VARMA, J. K., GREENE, K. D., RELLER, M. E., DELONG, S. M., TROTTIER, J., NOWICKI, S. F., DIORIO, M., KOCH, E. M., BANNERMAN, T. L., YORK, S. T., LAMBERT-FAIR, M. A., WELLS, J. G. & MEAD, P. S. (2003) An outbreak of Escherichia coli O157 infection following exposure to a contaminated building. *Jama,* 290**,** 2709-12.

VENTER, J. C., ADAMS, M. D., MYERS, E. W., LI, P. W., MURAL, R. J., SUTTON, G. G., SMITH, H. O., YANDELL, M., EVANS, C. A., HOLT, R. A., GOCAYNE, J. D., AMANATIDES, P., BALLEW, R. M., HUSON, D. H., WORTMAN, J. R., ZHANG, Q., KODIRA, C. D., ZHENG, X. H., CHEN, L., SKUPSKI, M., SUBRAMANIAN, G., THOMAS, P. D., ZHANG, J., GABOR MIKLOS, G. L., NELSON, C., BRODER, S., CLARK, A. G., NADEAU, J., MCKUSICK, V. A., ZINDER, N., LEVINE, A. J., ROBERTS, R. J., SIMON, M., SLAYMAN, C., HUNKAPILLER, M., BOLANOS, R., DELCHER, A., DEW, I., FASULO, D., FLANIGAN, M., FLOREA, L., HALPERN, A., HANNENHALLI, S., KRAVITZ, S., LEVY, S., MOBARRY, C., REINERT, K., REMINGTON, K., ABU-THREIDEH, J., BEASLEY, E., BIDDICK, K., BONAZZI, V., BRANDON, R., CARGILL, M., CHANDRAMOULISWARAN, I., CHARLAB, R., CHATURVEDI, K., DENG, Z., DI FRANCESCO, V., DUNN, P., EILBECK, K., EVANGELISTA, C., GABRIELIAN, A. E., GAN, W., GE, W., GONG, F., GU, Z., GUAN, P., HEIMAN, T. J., HIGGINS, M. E., JI, R. R., KE, Z., KETCHUM, K. A., LAI, Z., LEI, Y., LI, Z., LI, J., LIANG, Y., LIN, X., LU, F., MERKULOV, G. V., MILSHINA, N., MOORE, H. M., NAIK, A. K., NARAYAN, V. A., NEELAM, B., NUSSKERN, D., RUSCH, D. B., SALZBERG, S., SHAO, W., SHUE, B., SUN, J., WANG, Z., WANG, A., WANG, X., WANG, J., WEI, M., WIDES, R., XIAO, C., YAN, C., et al. (2001) The sequence of the human genome. *Science,* 291**,** 1304-51.

VOETSCH, A. C., VAN GILDER, T. J., ANGULO, F. J., FARLEY, M. M., SHALLOW, S., MARCUS, R., CIESLAK, P. R., DENEEN, V. C. & TAUXE, R. V. (2004) FoodNet estimate of the burden of illness caused by nontyphoidal Salmonella infections in the United States. *Clin Infect Dis,* 38 Suppl 3**,** S127-34.

VON MERING, C. & BORK, P. (2002) Teamed up for transcription. *Nature,* 417**,** 797-8.

WARREN, P. B. & TEN WOLDE, P. R. (2004) Statistical analysis of the spatial distribution of operons in the transcriptional regulation network of Escherichia coli. *J Mol Biol,* 342**,** 1379-90.

WATANABE, H., MORI, H., ITOH, T. & GOJOBORI, T. (1997) Genome plasticity as a paradigm of eubacteria evolution. *J Mol Evol,* 44 Suppl 1**,** S57-64.

WHELAN, S., LIO, P. & GOLDMAN, N. (2001) Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet,* 17**,** 262-72.

WHITFIELD, H. J., JR., GUTNICK, D. L., MARGOLIES, M. N., MARTIN, R. G., RECHLER, M. M. & VOLL, M. J. (1970) Relative translation frequencies of the cistrons of the histidine operon. *J Mol Biol,* 49**,** 245-9.

WILGENBUSCH, J. C. & SWOFFORD, D. (2003) Inferring evolutionary trees with PAUP*. *Curr Protoc Bioinformatics,* Chapter 6**,** Unit 6 4.

WILKINSON M., THORLEY J. L., PISANI D., LAPOINTE F. J., MCINERNEY J. O. (2004) Some desiderata for liberal supertrees. In: Phylogenetic supertrees: Combining information to reveal the Tree of Life—Bininda-Emonds O. R. P., ed. Dordrecht, The Netherlands: Kluwer Academic. 227–246.

WILKINSON, M., MCINERNEY, J. O., HIRT, R. P., FOSTER, P. G. & EMBLEY, T. M. (2007) Of clades and clans: terms for phylogenetic relationships in unrooted trees. *Trends Ecol Evol,* 22**,** 114-5.

WOESE, C. R. (1987) Bacterial evolution. *Microbiol Rev,* 51**,** 221-71.

WOLF, Y. I., ROGOZIN, I. B., GRISHIN, N. V. & KOONIN, E. V. (2002) Genome trees and the tree of life. *Trends Genet,* 18**,** 472-9.

WOLF, Y. I., ROGOZIN, I. B., GRISHIN, N. V., TATUSOV, R. L. & KOONIN, E. V. (2001a) Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol Biol,* 1**,** 8.

WOLF, Y. I., ROGOZIN, I. B., KONDRASHOV, A. S. & KOONIN, E. V. (2001b) Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res,* 11**,** 356-72.

WOLK, C. P., VONSHAK, A., KEHOE, P. & ELHAI, J. (1984) Construction of shuttle vectors capable of conjugative transfer from Escherichia coli to nitrogen-fixing filamentous cyanobacteria. *Proc Natl Acad Sci U S A,* 81**,** 1561-5.

WONG, S. & WOLFE, K. H. (2005) Birth of a metabolic gene cluster in yeast by adaptive gene relocation. *Nat Genet,* 37**,** 777-82.

XIE, G., BONNER, C. A., BRETTIN, T., GOTTARDO, R., KEYHANI, N. O. & JENSEN, R. A. (2003a) Lateral gene transfer and ancient paralogy of operons containing redundant copies of tryptophan-pathway genes in Xylella species and in heterocystous cyanobacteria. *Genome Biol,* 4**,** R14.

XIE, G., KEYHANI, N. O., BONNER, C. A. & JENSEN, R. A. (2003b) Ancient origin of the tryptophan operon and the dynamics of evolutionary change. *Microbiol Mol Biol Rev,* 67**,** 303-42, table of contents.

YANG, J., NIE, H., CHEN, L., ZHANG, X., YANG, F., XU, X., ZHU, Y., YU, J. & JIN, Q. (2007) Revisiting the molecular evolutionary history of Shigella spp. *J Mol Evol,* 64**,** 71-9.

YANG, Z. H., GOLDMAN, N. & FRIDAY, A. (1994). Comparison of Models for Nucleotide Substitution Used in Maximum-Likelihood Phylogenetic Estimation. *Molecular Biology and Evolution* 11, 316-324.

YANG, Z. H. (1996). Phylogenetic analysis using parsimony and likelihood methods. *Journal of Molecular Evolution* 42, 294-307.

YAP, W. H., ZHANG, Z. & WANG, Y. (1999) Distinct types of rRNA operons exist in the genome of the actinomycete Thermomonospora chromogena and evidence for horizontal transfer of an entire rRNA operon. *J Bacteriol,* 181**,** 5201-9.

YUL E , G. U. (1924). A mathematical theory of evolution, based on the conclusions of Dr J. C. Willis. *Philos. Trans. Roy. Soc. London Ser. B* **213** 21–87.

ZASLAVER, A., MAYO, A. E., ROSENBERG, R., BASHKIN, P., SBERRO, H., TSALYUK, M., SURETTE, M. G. & ALON, U. (2004) Just-in-time transcription program in metabolic pathways. *Nat Genet,* 36**,** 486-91.

ZIMMERMAN, S. B. & TRACH, S. O. (1991) Estimation of macromolecule concentrations and excluded volume effects for the cytoplasm of Escherichia coli. *J Mol Biol,* 222**,** 599-620.

Chapter 7 – Appendices