Computational Ortholog Prediction: Evaluating Use Cases and Improving High-Throughput Performance

by

Matthew Daratha Whiteside

B.Sc., (Hons., Bioinformatics), University of Waterloo, 2006

Thesis Submitted In Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Molecular Biology and Biochemistry
Faculty of Science

SIMON FRASER UNIVERSITY

Spring 2013

Approval

Name: **Matthew Daratha Whiteside**

Degree: **Doctor of Philosophy**

Title of Thesis: ***Computational Ortholog Prediction: Evaluating Use Cases and Improving High-Throughput Performance***

Examining Committee: **Chair:** Dr. Ralph Pantophlet
Assistant Professor, Faculty of Health Sciences

Dr. Fiona S.L. Brinkman
Senior Supervisor
Professor, Department of Molecular Biology and Biochemistry

Dr. Jack Chen
Supervisor
Associate Professor, Department of Molecular Biology and Biochemistry

Dr. Margo M. Moore
Supervisor
Professor, Department of Biological Sciences

Dr. Ryan D. Morin
Internal Examiner
Assistant Professor, Department of Molecular Biology and Biochemistry

Dr. Rosemary J. Redfield
External Examiner
Professor, Department of Zoology
University of British Columbia

Date Defended/Approved: March 8th, 2013

Partial Copyright Licence

Abstract

Orthologs are genes that diverged from an ancestral gene when the species diverged. High-throughput computational methods for ortholog prediction are a key component of many computational biology analyses. A fundamental premise in these analyses is that orthologs (when predicted correctly) are functionally equivalent and can be used to transfer gene annotations across species. Currently, many existing ortholog prediction methods generate a sizeable number of incorrect ortholog predictions, especially in cases of complex gene evolution. My thesis examines the functional equivalence hypothesis further and presents one solution that increases the precision of ortholog prediction.

To examine the use of orthologs in computational analysis, I conducted and evaluated three projects that employ ortholog prediction in distinct ways. In these projects, orthologs were used to (1) identify conserved, unique genes in metazoan species, (2) validate predicted gene regulatory modules in *Pseudomonas aeruginosa*, and (3) construct a transcriptional regulatory network in *Aspergillus fumigatus*. I identified factors affecting ortholog prediction in these specific use cases, demonstrating how successive gene duplications, incomplete genomes and rapid evolution of gene regulation can impact the results for such analyses.

To improve ortholog prediction, I evaluated and augmented an existing method called Ortholuge. Ortholuge is a computational method that increases the precision of ortholog prediction in a high-throughput setting. I evaluated the performance of Ortholuge, showing that its approach of classifying orthologs based on their relative phylogenetic divergence does identify orthologs that are more functionally equivalent. I compared Ortholuge to contemporary methods QuartetS and OMA, and showed that Ortholuge consistently identifies functionally-equivalent orthologs across a range of taxonomic distances. I also further developed Ortholuge's functionality by reducing run-time, increasing accuracy and improving usability through a number of modifications. Lastly, to make Ortholuge results available to the research community, I developed a database of Ortholuge ortholog predictions for bacteria and archaea species. This online

database provides high-level visualization of orthologs and the ability to easily run complex queries to retrieve genes that are shared or unique between specified taxa.

Overall, this work contributes an enhanced method for precise high-throughput ortholog identification and increases our understanding of the functional equivalences between orthologs.

Keywords: Orthology; Comparative Genomics; Bioinformatics; Phylogenomics; Evolution

Dedication

Aan mijn liefde Joske.

Dank u voor jou eeuwige steun. Je geeft me de moed om mijn dromen te volgen.

Acknowledgements

I would like to thank the many people who have helped me in the completion of my PhD thesis. The first is my senior supervisor, Dr. Fiona Brinkman. You have my sincerest gratitude. Thank you for this opportunity, which has taught me so much. My PhD has been more than an education; it has been a life changing experience. I would like to thank past and present members of the Brinkman lab. You have made this experience truly unforgettable. Thank you so much for your support and shared laughs during these past six years. I am very thankful for the invaluable guidance and contribution of my supervisory committee; Dr. Margo Moore and Dr. Jack Chen. I would like to acknowledge all project collaborators. Your help and the opportunities you have provided me have made this work possible. I would like to extend thanks to:

- Drs. Jinko Graham and Brad McNeney, Jeong Eun Min –
  OL.locfdr Project
- Drs. Melissa Frederic and Michel Leroux –
  Metazoan Project
- Dr. Margo Moore, Linda Pinto and Jason Catterson –
  Aspergillus fumigatus Iron-Limitation Project
- Geoff Winsor and Matthew Laird –
  OrthologeDB Project

Thank you to the funding agencies: the Michael Smith Foundation for Health Research and Simon Fraser University for their financial support during my PhD. Finally, I would like to thank my family. Without your support and encouragement, none of this would be possible.

Table of Contents

List of Tables

List of Figures

List of Acronyms

AC	Adenylyl cyclase
AES	Australian epidemic strain (*Pseudomonas aeruginosa*)
BLAST	Basic local alignment search tool
bp	base pairs
CF	Cystic fibrosis
ECD	Empirical cumulative distribution
FN	False negative
FP	False positive
fRMA	Frozen Robust Multi-Array Analysis
GEO	Gene Expression Omnibus
GPCR	G-protein coupled receptor
HGT	Horizontal gene transfer
ISA	Iterative search algorithm
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	KEGG orthology
LCA	Last common ancestor
LES	Liverpool epidemic strain (*Pseudomonas aeruginosa*)
POG	*Pseudomonas* ortholog group
RBB	Reciprocal best BLAST
RBBH	Reciprocal best BLAST hit
RP	Ribosomal proteins
SCL	Subcellular localization
SSD	Supporting species divergence
TF	Transcription factor
TFBS	Transcription factor binding site
TM	Transcriptional module
TN	True negative
TP	True positive
TRN	Transcriptional regulatory network

Glossary

Comparative Genomics Analysis	A type of bioinformatics analysis that uses an evolutionary framework to study function or genome structure across multiple genomes.
Graph-based Ortholog Prediction	Computational ortholog prediction methods that use pair-wise sequence similarities computed between all genes in two genomes to predict orthologs. BLAST is typically used to measure sequence similarities.
Metazoan-associated genes	Orthologous genes widely conserved across and exclusive to metazoan species.
Negative Selection	Evolutionary selection pressure that removes deleterious mutations from a population.
Non-SSD Orthologs	Predicted orthologs with phylogenetic divergence that is not equivalent to the species-level of divergence.
Orthologs	Homologous genes that diverged from a common ancestral gene in the last common ancestor when the species diverged.
Outgroup	A species or genome that diverged prior to the species under investigation. Outgroup genes are used as reference in the phylogenetic analysis of a set homologous genes.
Paralogs (In-paralog / Out-paralog)	Homologous genes that diverged due to a gene duplication event. For in-paralogs, the duplication occurred in the individual species lineages after the species diverged and for out-paralogs, the gene duplication occurred in the ancestral genome prior to the divergence of the species.
Phylogenetic Distance	A measure of the degree of separation between genomes. Distances can be expressed in units of time, mutation rates or generations.
Phylogenetic tree-based Ortholog Prediction	Computational ortholog prediction methods that build a phylogenetic gene tree for a set of homologous genes and then resolve the gene tree against a provided species tree to infer orthologs and paralogs.
Positive Selection	Evolutionary selection pressure that retains advantageous mutations in an organism.
Pseudomonas aeruginosa epidemic strain	Highly prevalent *Pseudomonas aeruginosa* strains infecting Cystic Fibrosis patients.
Reciprocal best BLAST	A graph-based ortholog prediction method that identifies as orthologs, the reciprocal top BLAST hits in two genomes.
Regulon	A group of genes transcriptionally regulated by the same set of

	regulators.
SSD Orthologs	Supporting-species-divergence orthologs (SSD orthologs) refer to predicted orthologs with Ortholuge phylogenetic distance ratios that is equivalent to the species level of divergence.
Synteny	The conservation of the order of genes on a chromosome passed down from a common ancestor (The genetic definition of co-localization on the same chromosome is not used in this thesis).
Transcriptional Module	A group of genes that are co-regulated under the same conditions through a common transcription factor binding site motif profile.

1. Introduction to Homology and Comparative Genomics

Comparative genomics is the study of the genome content and structure across different species. The comparison of multiple genomes helps reveal the evolutionary origins of genes and the selective forces acting on them, allowing researchers to make inferences about the function of the genes. Typically, a comparative genomics analysis will first compare the similarities of protein or RNA sequences of genes in two or more species and then uses this information to identify corresponding genes between species. The corresponding genes can be used for downstream analysis to, for example, transfer of functional annotations, investigate selection, or highlight functional elements in the genome.

Comparative genomics is a relatively young field that emerged with the development of high-throughput genome sequencing and ensuing publication of thousands of genome sequences. While high-throughput genome sequencing has enabled comparative genomic analysis, the rapid explosion of genome sequences has also driven the advancement of the comparative genomics field, as the amount of genetic information has outpaced traditional methods of gene function characterization, and new methods based on function transfer between corresponding genes are needed to annotate newly sequenced genomes.

Development in the comparative genomics field has proceeded down two tracks. On a practical level, new computer algorithms were needed to identify the differences and similarities in the protein and DNA sequences of genes, as well as in the arrangement of genes within the genome. In addition to the practical aspects, conceptual models of gene correspondence are as important to developing robust comparative genomics applications. Definitions and methods for classifying gene relationships are needed to robustly interpret the gene similarities and differences. Broadly, homology is the study of similarity due to common ancestry. In the context of

genomics, a homolog is a gene that is related to another gene through descent from a common ancestor. Gene relationships are defined by the intervening gene evolution events that gave rise to the genes. This thesis examines and develops both aspects of comparative genomics.

1.1. Gene Evolution and Its Impact on Gene Function

Identifying gene evolution events and understanding how they affect a gene's function, are critical to making reliable functional inferences in comparative genomics. In the following subsections, I describe the major types of gene evolutionary processes, how genes are classified based on these evolution events and finally the current understanding of their impacts on gene function.

1.1.1. Overview of the Mechanisms of Gene Evolution

Studying genes in different species, as in comparative genomic analysis, requires an understanding of the major processes in gene evolution. In their approximate order of contribution to gene evolution, outlined below are the fundamental processes that give rise to novel genes or alter existing genes:

i. *Vertical descent with modification*: during speciation, each of the daughter species inherits a copy of the gene from the ancestral species. Depending on selection constraints on the genes in each species, these copies can incur significant mutations and evolve distinct functions. Typically though, genes that arise through speciation have lower mutation rates than genes arising from other forms of gene evolution (Koonin 2005).

ii. *Gene duplication with modification*: duplication of the DNA region containing a gene can occur through a number of different events, including chromosomal duplication, errors in homologous recombination, or retrotransposition. The change in gene copy number is frequently associated with a change in selective pressure. Often one or both genes accumulate mutations faster than single copy genes (Koonin 2005; Kaessmann 2010).

iii. *Gene loss*: gene loss can occur when the DNA region containing a gene is deleted, such as through homologous recombination or through the accumulation of deleterious mutations in a gene's DNA producing a non-functional pseudogene (Mira, Ochman, and Moran 2001).

iv. *Horizontal gene transfer*: horizontal gene transfer (HGT) occurs mainly in the bacteria and archaea and accounts for a large proportion of gene novelty in these species domains. Three main mechanisms of HGT have been identified: (i) natural transformation (uptake of free DNA by competent bacteria), (ii) transduction (introduction of foreign DNA through the infection and integration of a bacteriophage genome into the host's genome) and (iii) conjugation (transfer of mobile DNA elements through pili appendages between physically-connected bacteria) (Thomas and Nielsen 2005).

v. *Fusion and fission*: A gene fusion event is defined as two separate genes merging into a single transcriptional unit, while fission is the splitting of one gene into two transcriptional units. Many of the same mechanisms as gene duplication produce gene fusions and fissions (including homologous recombination, segmental translocation and retrotransposition), but instead of generating two separate gene copies, the genes' DNA is either juxtaposed or separated (Kaessmann 2010).

1.1.2. Definition of Orthology and Paralogy

In order to robustly use comparative genomics-based analysis to draw conclusions about gene functions, we need to understand the correspondence between genes and to do this we need a consistent framework to describe genes' evolutionary relationships. In this section, I define the key terms used in comparative genomics to classify genes.

At the broadest level, the term homolog refers to any genes that share a common origin. The main processes that contribute to gene evolution are vertical descent and gene duplication, and these are the principle processes that are used to classify genes. Orthologs are defined as homologous genes that diverged due to speciation in last common ancestor (and not an arbitrary ancestor). They must have diverged from same ancestral gene in the last common ancestor (Koonin 2005). Alternatively, genes that have diverged due to a gene duplication event are classified as paralogs (Koonin 2005). Paralogs are further sub-classified based on the relative timing of the gene duplication. Duplications that occurred before the divergence of the species are classified as out-paralogs. In-paralogs are defined as genes that have duplicated after the species diverge (Sonnhammer and Koonin 2002). Note that the definitions of in-paralog and ortholog are not mutually exclusive. A pair of in-paralogous genes that have duplicated subsequent to species divergence can also satisfy the definition for ortholog: specifically,

that the genes diverged from a common gene in the last common ancestor of the species. The term co-ortholog is sometimes used to describe these many-to-many relationships where there are more than one valid orthologs due to a recent gene duplication event (Sonnhammer and Koonin 2002).

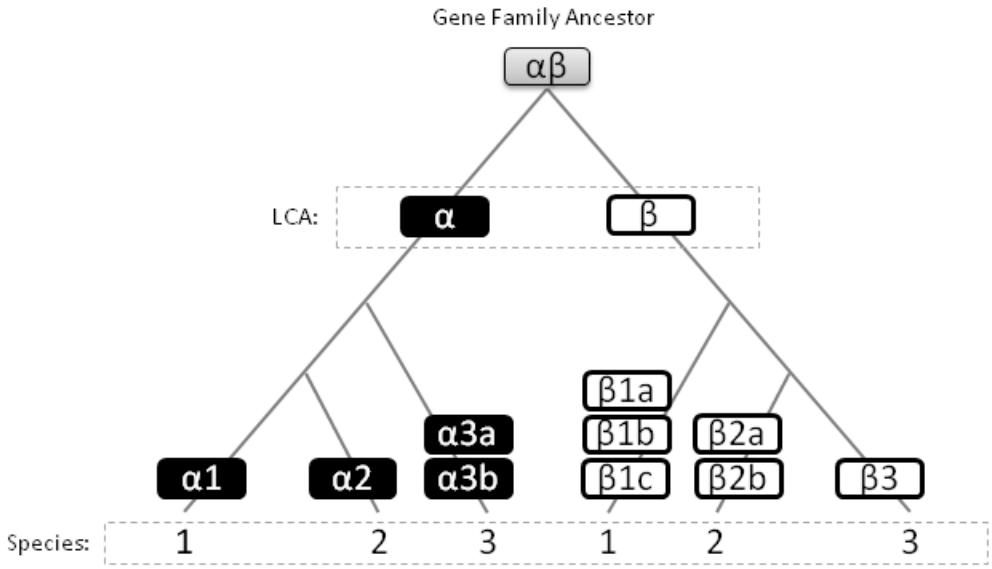


Figure 1.1 Defining Orthologous and Paralogous Genes

This hypothetical phylogenetic gene tree highlights all types of homologous relationships. Genes are represented by the boxes and in the last common ancestor (LCA) of species 1, species 2 and species 3, there were two ancestral genes α and β. Genes α1 and α2 in species 1 and 2 have both diverged from a common ancestral gene α in the LCA. These genes diverged when the species diverged and are examples of orthologs. Genes α1 and β3 are examples of paralogs because they arose from a gene duplication of the αβ gene ancestor in the LCA. They are further classified as out-paralogs because the gene duplication occurred prior to the divergence of the species 1 and 3. When the gene duplication occurs subsequent to species divergence, for example genes α3a and α3b, they are classified as in-paralogs. In-paralogs can also satisfy the criteria for orthologs, namely they diverged from a common ancestral gene in the LCA. The term co-ortholog is used to describe this type of many-to-many ortholog relationship (genes β1a, β1b, β1c and β2a, β2b are examples of co-orthologs).

Horizontal gene transfer is a major contributor of genetic novelty in bacteria and archaea. The term xenolog is used to describe genes that are homologous because they are derived from the horizontal gene transfer of the same genetic element in two or more species (Koonin 2005).

The basis for these classifications is solely the intervening evolutionary events between genes and although these terms have no explicitly defined functional correspondence, there are implications related to the mode of divergence and their impacts on gene functions (Koonin 2005). The effects on gene function of vertical descent and gene duplication are discussed in the next section.

1.1.3. Correspondence between Mode of Gene Evolution and Functional Divergence

Orthologs are derived from a common ancestral gene through vertical descent, while paralogs are derived from a common ancestral gene through gene duplication. These distinct evolutionary paths are thought to have significant impact on the expected similarity of the genes function. The widely accepted conjecture is that duplicated genes rapidly diverge in function or are lost (functional divergence is necessary for their preservation in genome), and therefore orthologs are functionally more similar than genes that have diverged by duplication, or paralogs (Koonin 2005).

In actuality, the difference in functional conservation between orthologs and paralogs is not clear cut. Multiple examples of functional divergence exist for both orthologs and paralogs (Studer and Robinson-Rechavi 2009). Many of the evolutionary processes that drive the functional divergence in duplicated genes also act on orthologs. One of the forces that drive functional change in genes is asymmetrical rates of evolution (Conant and Wolfe 2008; Studer and Robinson-Rechavi 2009). The implicit assumption of the ortholog function conjecture is that paralogs diverge more than orthologs in the same amount of time. Rates in evolution can change due to relaxation of negative selection or an increase in positive selection. The additional copy of a gene is thought to change the selective pressures acting upon the genes. In the neofunctionalization model of gene duplication, the duplicated copy experiences relaxed negative selection, allowing mutations to accumulate. These mutations can alter the genes original function and confer a selective advantage that is propagated in the population through positive selection (Conant and Wolfe 2008). In support of this, highly asymmetrical rates of evolution have been observed for many paralogs. However, asymmetrical rates of evolution are not exclusive to paralogs. Evolution in orthologs is

also frequently asymmetrical although it appears not to the same degree (Studer and Robinson-Rechavi 2009).

The validity of ortholog function conjecture has recently been examined on a larger scale (Nehrt *et al.* 2011; Altenhoff *et al.* 2012; Dessimoz *et al.* 2012; Thomas *et al.* 2012). In one study, manually-annotated, experimentally-derived gene ontology annotations were compared between orthologs and paralogs in 13 species. These studies showed that, after controlling for the level of divergence, orthologs tend to more often have functions that are conserved than paralogs. The difference in function conservation, however was relatively weak (Altenhoff *et al.* 2012).

The ortholog function conjecture is not a definitive proposition. Orthologs and paralogs both can diverge in function. Current evidence supports that paralogs more often have dissimilar functions. The degree to which is yet to be precisely determined, but it appears the degree of functional divergence in orthologs and paralogs is highly variable between gene families and species (Studer and Robinson-Rechavi 2009).

Fates of Genes after Duplication

Gene duplication has contributed significantly to the evolution of phenotypic novelty. The relative contribution of gene duplication, however, varies between taxa and gene families (Koonin 2005). In eukaryotes, it is more pervasive than in prokaryotic species, but gene expansions comprise a sizeable proportion of the genes in most species (for example, 5-33% in a study of 21 prokaryotic of the genes were derived from lineage-specific gene duplications (Jordan 2001)). Multiple theoretical models have been developed to explain how gene duplication gives rise to novel functions.

The most common fate for gene duplications in an organism is that they are lost. In the short-term, preservation of gene duplications occurs under two circumstances (Conant and Wolfe 2008). First, a gene duplication can provide an immediate benefit to the organism through gene dosage effects where the higher mRNA content confers a selective advantage (Francino 2005). Alternatively, gene duplication can also be preserved in cases where it is evolutionarily neutral. Even if preserved in the short-term, over time the most common fate of duplicated genes is the accumulation of loss-of-function mutations and pseudogenization. However, if those mutations result in

functional diversification, the duplicated gene can be positively selected for and preserved long-term in the population. The theoretical models explain how new functions arise from gene duplication. The models are divided into two main categories: subfunctionalization and neofunctionalization (Conant and Wolfe 2008; Kaessmann 2010).

Neofunctionalization

In the neofunctionalization model type, one of the duplicated genes maintains the ancestral gene function, while the other sister gene evolves a novel function (Conant and Wolfe 2008). In the MDN (mutation during non-functionality) model, one of duplicate genes is superfluous and acquires a random mutation that confers a distinct function (Conant and Wolfe 2008). With the IAD (innovation, amplification, divergence) model, the ancestral gene has side-activity that is not optimized (for example, a promiscuous enzyme that is optimized for one reaction, but is able to catalyze other reactions sub-optimally). After duplication, the sub-optimal activity is optimized through mutation in one of the sister genes, while the main function is preserved by the other gene copy (Conant and Wolfe 2008). The IAD modelled is favored over the MDN model because of the relative improbability of a gene to acquire a novel function through random mutation. A pre-existing side-function that is optimized after release from the constraint of main gene function is a more likely scenario (Conant and Wolfe 2008).

Subfunctionalization

In subfunctionalization, the ancestral gene has multiple functions, which after duplication are divided amoung the sister genes. In the DDC (duplication degeneration complementation) model of subfunctionalization, after a period of neutral evolutionary drift, the duplicated sister genes acquire complementary mutations that render mutually exclusive sets of the ancestral functions inactive in each of the sister genes. Individually, a single sister gene is no longer able to perform all of the ancestral gene functions. Like the IAD neofunctionalization model, the EAC (escape from adaptive conflict) presupposes that the ancestral gene has multiple functions that cannot be simultaneously optimized (Hittinger and S. B. Carroll, 2007; Conant and Wolfe, 2008). The gene duplication releases the genes from this “adaptive conflict”, allowing each of the sister genes to selectively optimize one of the functions of the ancestral gene

(corresponding with the loss of the other function). The key difference between EAC and DDC subfunctionalization models is that EAC functional divergence occurs through positive selection of separate functions of the ancestral gene, while in the DDC model, ancestral gene functions in the duplicated genes are lost through random mutation (Hittinger and Carroll 2007).

1.2. Detection of Orthologs

Ortholog prediction methods have been classified into two categories: tree-based, which use phylogenetic trees to resolve orthologs, and graph-based, which use pair-wise sequence similarities computed across the entire genome. Hybrid approaches have also been developed.

1.2.1. Phylogenetic Tree-based Detection of Orthologs

Tree-based methods predict orthologs and paralogs by building phylogenetic gene trees and then reconciling the phylogenetic gene trees with a supplied species tree. To achieve high-throughput prediction, the analysis steps must be automated. The steps in a typical tree-based ortholog prediction pipeline are outlined below (Dufayard *et al.* 2005; Kuzniar *et al.* 2008; Kristensen *et al.* 2011):

1. Gene sequences for the phylogenetic tree must be collected. Most programs use a sequence similarity search such as BLAST to select which genes from the genomes to use in the tree.
2. After obtaining the gene sequences, a multiple sequence alignment must be constructed. Ideally, low quality regions in the alignment are filtered out.
3. A phylogenetic tree is built from the multiple sequence alignment. Tree building methods fall into two types: distance based (e.g. UPGMA, neighbour-joining) and character-based (e.g. maximum parsimony). Distance based trees are computationally less intensive to build, but they are also less accurate in many situations (Kristensen *et al.* 2011). Approximate maximum-likelihood methods, however, reduce some of the computational burden of character-based methods (Price, Dehal, and Arkin 2010). Some ortholog prediction tools use methods such as bootstrapping (Zmasek and Eddy 2002) or build a consensus tree from multiple distinct tree building methods to improve confidence in the final tree (Vilella *et al.* 2009). Other

methods examine features such as branch length to identify potential problems (Dufayard *et al.* 2005).

4. The gene tree is reconciled with a supplied species tree. By inferring duplication and gene loss events along the branches, the gene tree is made to match the species tree.
5. Once a reconciled gene tree is produced, ortholog and paralog are inferred by examining the presence or absence of duplication events between the leaves of the tree. A gene tree that is congruent with the species tree is inferred to contain orthologs.

Generally, a phylogenetic approach is preferred for identifying orthologs (Kuzniar *et al.* 2008; Kristensen *et al.* 2011). The methods can be less prone to error in presence of gene duplication and loss. However, tree-based methods have several limitations. There is usually a significant computational cost associated with phylogenetic tree building. Most tree-based methods do not scale well as the number of genomes and sequences increases (Kuzniar *et al.* 2008). The resource MetaPhOrs is one of the largest repositories for orthologs predicted using phylogenetic tree-based methods. It contains ortholog predictions across 829 genomes. This scale of ortholog prediction is achieved by mining pre-computed phylogenetic gene trees from other sources (Pryszcz, Huerta-Cepas, and Gabaldón 2011). Another limitation of the tree-base methods is the sensitivity of automated tree building to biases in the data or errors in the multiple sequence alignment. For example, a well-documented issue with tree building is long-branch attraction. Long branches are grouped together in a maximum parsimony approach even though they may not be sister taxa (O'Connor *et al.* 2010). Finally, the requirement for a species tree is a significant limitation of tree-based methods. In many methods, the species tree must be a perfectly resolved, rooted, bifurcating tree. A few tools have developed methods for computing the root or working with multi-furcating trees (Zmasek and Eddy 2001; Dufayard *et al.* 2005).

1.2.2. Graph-based Detection of Orthologs

Graph-based methods use pair-wise sequence similarities to predict orthologs (paralogs are not explicitly detected in this approach). The formative premise behind the graph-based methods is that because the orthologs diverged from a common ancestral gene, they should appear as the reciprocally most similar genes from their respective genomes (Kuzniar *et al.* 2008). Because of its speed, BLAST is the most commonly

used method for determining the best match. In the typical approach, BLAST will be run twice, using each genome as the query and subject. The top BLAST hits will be recorded for each gene in the genomes and orthologs are declared as the pair of genes that are the reciprocal best BLAST hits of each other (this approach is referred to as RBB). Differences in implementations can include which BLAST metric is used to determine top hit: E-value, percent identity or bit score, and how ties or multiple top hits are dealt with (Kuzniar *et al.* 2008).

In practice, the RBB procedure works well and can outperform more complex ortholog predictions methods in some cases (Altenhoff and Dessimoz 2009). Most of the graph-based ortholog tools use RBB but then incorporate additional steps to improve accuracy, detect other types of homologous genes, or perform multi-species ortholog prediction. Graph-based approaches are computationally less intensive and more straightforward to automate, making them comparatively proficient at computing orthologs for large datasets (Kuzniar *et al.* 2008). Their major drawback is their limited phylogenetic view (they only consider the top matches). RBB or other similar methods can fail in situations where there is differential gene loss or recombination of protein domains that can change the top hit to a non-orthologous gene (Fulton *et al.* 2006).

1.2.3. Hybrid Approaches

Ortholog prediction approaches have been developed that use elements of both the phylogenetic tree-based methods and the graph-based methods. In most tools that can be considered a hybrid approach, this involves using a RBB procedure to identity candidates for phylogenetic gene tree building step (Li *et al.* 2006; Vilella *et al.* 2009). An example of a hybrid approach is the Phylogenetic Orthologous Groups (PHOGs) database, which uses a species tree to guide the creation of ortholog groups (Merkeev, Novichkov, and Mironov 2006). Groups are constructed using RBB in a hierarchical fashion, starting with the most closely related species and moving up the species tree to include species with a larger phylogenetic range. Orthologs from the earlier groups are used as seeds in the later groups.

1.2.4. Resolving Complex Ortholog Relationships

Graph-based methods search for genes that are reciprocal best matches to identify orthologs. While this simplicity provides several advantages, it also limits their utility. Graph-based methods only consider the top hits and do not detect recent gene duplications or in-paralogs that occur in the orthologs gene lineage. Also, graph-based methods work in a pair-wise fashion. To extend graph-based methods to multiple species, ortholog predictions from the pair-wise analyses need to be grouped.

Detection of In-paralogs

In order to identify orthologs, phylogenetic tree-based methods must also identify paralogs in the process. This is not true of graph-based methods. A procedure, however, has been developed for graph-based methods to identify in-paralogs (recent gene duplications). The graph-based method Inparanoid has an additional step in its ortholog prediction pipeline to explicitly detect in-paralogs (O'Brien, Remm, and Sonnhammer 2005). Pair-wise similarity scores are computed for all genes in two genomes, including between genes from the same genome. Reciprocal best matches from the two genomes are labelled as orthologs and form initial ortholog groups. Additional genes are added to an ortholog group if the similarity score between the gene and one of the orthologs is greater than the score between the orthologs. These added genes are labelled in-paralogs. The rationale is that since in-paralogs are duplicated genes that arose after speciation and the divergence of the orthologous genes, in-paralogs should be more similar in sequence than the two orthologs (O'Brien, Remm, and Sonnhammer 2005). Implementation details of the Inparanoid in-paralog detection method are provided in section 4.3.

Grouping Orthologs from Multiple Species

Phylogenetic tree-based ortholog prediction methods inherently produce multi-species ortholog groupings because they examine the phylogeny of all homologous genes simultaneously. Graph-based methods, however, are limited to pair-wise ortholog prediction. In order to extend the pair-wise graph-based methods to multiple species, the orthologs from the pair-wise analysis must be grouped. Several methods for creating ortholog groups have been developed. Ortholog grouping methods can be

characterized based on whether they are hierarchical or not, exact or approximate and the requirements for inclusion of a gene in a group.

Orthology is relative to species included and their last common ancestor, so the phylogenetic range of the species used can alter the orthologous genes belonging to a group (Jensen *et al.* 2008; Powell *et al.* 2012a). For example, an out-paralog can become an in-paralog, if the phylogenetic range of the species increases, shifting the last common ancestor of all species deeper in the phylogenetic gene tree. An additional factor related to the species distance, is that ortholog prediction is more accurate between closely-related species (Jensen *et al.* 2008). A hierarchical approach to ortholog grouping computes multiple levels of groups with each level increasing in the phylogenetic range of the species genomes that are used to create the groups. In some approaches, lower level groups are used as seeds for the upper levels. Hierarchical ortholog groups allow the user to select the group level and scope that fits their application. While it is more flexible, hierarchical ortholog groups are also more computationally intensive, since the ortholog grouping procedure must be run for each level.

Ortholog relationships grow combinatorially as the number of species' genomes used in the analysis increases. To reduce the computational burden, heuristic clustering approaches have been applied to the problem of grouping orthologs. A popular heuristic clustering algorithm for forming ortholog groups is Markov Clustering (used in the method OrthoMCL) (Li, Stoeckert, and Roos 2003). It uses random simulated Markovian walks that trace the ortholog relationships in the ortholog connection network to determine the group structure (a long random walk will spend most of the time in densely connected regions or clusters). MCL includes an inflation factor that helps determine the degree of connectivity of the groups. The MCL approach is very efficient in comparison to exact ortholog grouping methods; however, it does not always recapitulate the biologically-correct ortholog groups. MCL works best when the clusters are of the same approximate size which tends to constrain the ortholog group size (Altenhoff *et al.* 2011). This limitation is in conflict with the natural variation in gene family size.

When determining which predicted orthologs should be added to a group, a transitive approach is the most straight-forward. Genes linked by a predicted ortholog relationship to any of the genes in a group, are added to that group. However, all ortholog prediction methods incur some number of false positives. Ortholog grouping methods must take this into account in order to mitigate distinct ortholog groups being merged due to false ortholog predictions. The ortholog relationships in an ortholog groups are assumed to be transitive, so the impact of a single false positive can be significant because it can create a number of implicit false ortholog connections if two non-orthologous groups are combined. Conversely, many valid ortholog connections will be missed in ortholog prediction, so the expectation that ortholog groups will be perfectly connected is violated frequently in practice (i.e. all genes in a group have a predicted ortholog relationship with all other genes in the group). Because of the false negatives and false positives in ortholog prediction, most ortholog grouping methods have developed strategies to improve the accuracy of ortholog group predictions.

Several ortholog grouping methods use a rules-based approach that for example, restricts orthologs in a group based on whether they align over a significant proportion of the gene with a certain percent identity in a sequence alignment with other members of the group (O'Brien, Remm, and Sonnhammer 2005; Schneider, Dessimoz, and Gonnet 2007; Ostlund *et al.* 2010; Altenhoff *et al.* 2011). Another example of a rule includes the use of gene synteny to identify the best ortholog when there are multiple top matches (Winsor *et al.* 2009).

Connectivity is another feature that is used to evaluate the correctness of ortholog groups. The ortholog groups in OMA ortholog database are restricted to groups that are maximal cliques (genes must be orthologously connected to all other group members) (Altenhoff *et al.* 2011). In practice, the maximal clique property is frequently violated, so more sophisticated connectivity criteria have also been incorporated into ortholog group prediction. The minimum-cut of a graph is the number of edges that need to be removed in order to separate a connected graph into two distinct unconnected subgraphs. Invalid predicted groups often appear as two densely-connected sets of orthologs bridged by a small number of false ortholog predictions in the graph representation of the ortholog group. The invalid ortholog predictions in these situations will typically have a low minimum-cut, and so a minimum-cut threshold can be

used to split predicted ortholog groups that appear to be incorrectly merged (Altenhoff *et al.* 2011).

The initial orthologs used to build the ortholog groups can also impact the final accuracy. The Clusters of Ortholog Groups (COG) approach first identifies ortholog triangles (i.e. three genes forming a maximally-connected ortholog group). Additional genes are then added to the seed triangles if they are orthologous to any of the genes. The rationale behind the COG approach is that a triangle is more reliable than a pair-wise ortholog prediction, but still not too restrictive in cases where there are missing ortholog relationships (Tatusov *et al.* 2001; Tatusov *et al.* 2003).

The large number of distinct approaches for computing ortholog groups is a testament to the challenges in grouping predicted orthologs into biologically-correct groups. The simplistic or straight-forward approach to creating ortholog groups by linking all genes connected by an orthologous relationship can produce incorrect ortholog groups when false ortholog predictions are present in the data. Additional quality control measures or more sophisticated ortholog grouping procedures are needed to limit the fusion of distinct ortholog groups due to false predictions.

1.3. Factors Affecting Ortholog Prediction Accuracy

Gene Loss and Incomplete Genomes

Graph-based methods are often favored for large-scale applications because they are typically less computationally intensive and do not require species trees. However these methods have a significant methodological limitation. In cases where there are reciprocal gene losses of the orthologs in their respective genomes, two paralogs can become reciprocal best matches (Figure 1.2). Standard graph-based methods will report these cases as orthologs (Fulton *et al.* 2006). The true orthologs in a genome can be missing due to either gene loss or incomplete genomes. Missing genes account for a significant proportion of false positive ortholog predictions in graph-based methods (Kuzniar *et al.* 2008).

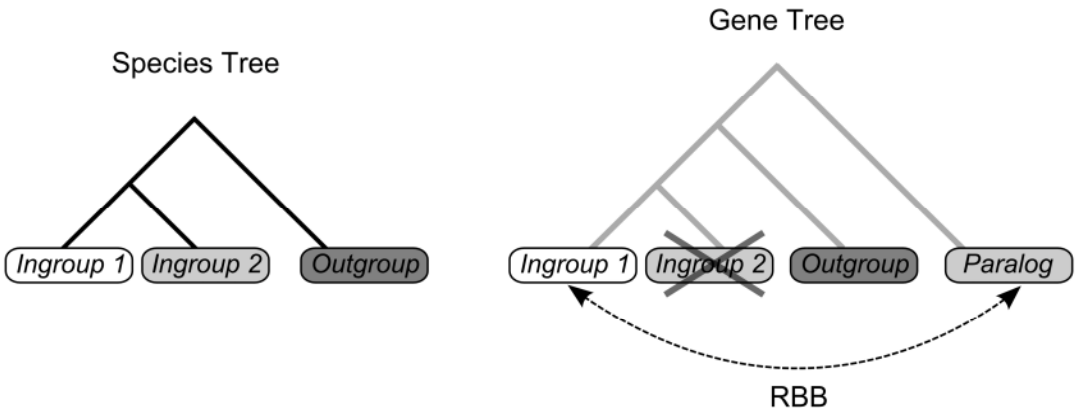


Figure 1.2 False Positive Produced by the RBB Ortholog Prediction Method

A depiction of how the RBB method can generate a false ortholog prediction. A species tree and corresponding phylogenetic gene tree are shown for three hypothetical orthologous genes labelled Ingroup 1, Ingroup 2 and Outgroup. RBB can mistakenly identify a paralog as an ortholog when the true ortholog is missing and there exists a paralog that can form a RBB relationship with the remaining ortholog in the other species.

Gene Fusion and Fission

Protein domain recombination is a common evolutionary process (Kuzniar *et al.* 2008). Gene fusions and fissions create novel chimeric proteins consisting of multiple non-homologous regions. Chimeric protein sequences are a potential source of error, especially when predicting ortholog groups. For example, a hybrid gene produced by a gene fusion can cause non-homologous genes to be mistakenly assigned to a common ortholog group, when both genes have similarity to a separate region of the hybrid gene (Kuzniar *et al.* 2008; Sjölander *et al.* 2011). Similarly, errors in the annotation of the gene sequence or alternative splicing, promiscuous protein domains, and low complexity regions can also complicate ortholog group prediction by causing non-homologous genes to display significant sequence similarity (Kuzniar *et al.* 2008).

Issues Identifying the Nearest Neighbour with the BLAST Algorithm

Reciprocal best BLAST (RBB) is a common procedure for identifying putative orthologs. One of the strengths of the method is the BLAST algorithm. BLAST can rapidly identify the reciprocally most similar genes from the two comparison genomes. BLAST uses a heuristic algorithm to detect similarity in a large sequence search space. The heuristic approach is much faster than computing optimal alignment. However, the

heuristic can return as the top hit a gene that is not the nearest phylogenetic neighbour, especially in cases where there is a close paralog (Wall, Fraser, and Hirsh 2003). Hits are ranked by the BLAST score; an artificial measure of sequence similarity that does not necessarily rank genes or proteins according the degree of phylogenetic relatedness (Altschul *et al.* 1997). Also, BLAST constructs a local alignment, identifying regions of the protein or genes that have significant sequence similarity (Altschul *et al.* 1997) . A global alignment over the entire length of the sequence would be more appropriate for identifying the nearest phylogenetic neighbour. These limitations of can result in the identification of a close paralog as the best BLAST hit instead of the nearest phylogenetic neighbour. This can disrupt the detection of valid ortholog pairs. If the forward BLAST hit misses the true ortholog and identifies a paralog, while the reverse BLAST hit identifies the ortholog, a RBB relationship will not be formed and a valid ortholog pair will be excluded (Wall, Fraser, and Hirsh 2003).

Horizontal Gene Transfer

Xenologs are homologous because they are derived from a common horizontally transferred genetic element. Since the insertion of HGT elements occurs after species divergence, xenologs will often appear as phylogenetically close homologs in an ortholog prediction analysis (Kuzniar *et al.* 2008). Careful examination of phylogenetic incongruence by comparing the patterns of presence and absence of genes in relation to the species tree is necessary to distinguish between orthologs and xenologs. Currently, no ortholog prediction method has the ability to detect xenologs (Kuzniar *et al.* 2008).

1.4. Improving Ortholog Prediction

1.4.1. Review of Existing Strategies for Improved Ortholog Prediction

Phylogenetic-based Ortholog Prediction without a Species Tree

Two methods have developed strategies for predicting orthologs from a phylogenetic tree without a species tree; a major limitation of phylogenetic-based ortholog prediction. Correlation Coefficient-based Clustering (COCO-CL) builds a correlation matrix of the phylogenetic distances between all sets of genes (Jothi *et al.*

2006). The method then clusters each gene in the correlation matrix. In each stage of the clustering when two clusters are combined, the number of overlapping species represented in the clusters is examined, and if there are duplicate species, a gene duplication event is inferred. The program LOFT also uses clustering and distributions of overlapping species to distinguish between orthologs and paralogs (Van der Heijden *et al.* 2007). In LOFT, the genes in a phylogenetic gene tree are hierarchically clustered and if the clustered branches contain mutually exclusive sets of species, the genes are considered orthologs (this approach is referred to as the "species overlap rule").

Overcoming Limitations of the BLAST algorithm in Identifying the Nearest Neighbour

In certain situations, the BLAST algorithm can identify a close paralog as the top BLAST hit instead of an ortholog (refer to section1.3). In the context of the ortholog identification, incorrectly identifying paralog as the top hit in place of an ortholog, can prevent the detection of valid orthologs by the reciprocal best BLAST procedure (Wall, Fraser, and Hirsh 2003). The reciprocal smallest distance (RSD) algorithm was developed to address this issue. It obtains a number of top hit candidates using BLAST, and then performs a global alignment followed by maximal likelihood estimation of the evolutionary distance between all top hit candidates. The nearest phylogenetic neighbours are identified as the pair of genes with the reciprocal smallest evolutionary distance (Wall, Fraser, and Hirsh 2003). This method is more robust to the issue of excluding valid orthologs when there is close paralog. In an analysis of *Saccharomyces cerevisiae* and *Candida albicans* genomes, the RSD algorithm produced 2777 putative ortholog pairs, while the RBBH algorithm produced 1824 pairs (Wall, Fraser, and Hirsh 2003). This algorithm is available in the ortholog tool: RoundUp (Deluca *et al.* 2006; DeLuca *et al.* 2012).

A Domain-Centric Approach

Gene fusions and fissions can complicate ortholog prediction. The merger of non-homologous regions into a single gene can act as a bridge between two or more non-homologous genes during ortholog prediction. Similarly, the fission of homologous regions in a single gene can conceal ortholog relationships. To overcome the complications that arise from the frequent domain rearrangements in proteins, some

ortholog methods have adopted a domain-centric approach for predicting orthology. These methods use conservation of domain architecture to inform or set boundaries on ortholog prediction (Zmasek and Eddy 2002; Uchiyama 2006; Kuzniar *et al.* 2008; Chen *et al.* 2010; Sjölander *et al.* 2011). For example, DODO (DOmain based Detection of Orthologs) first groups proteins based on their domain composition and then sub-classifies the domain group genes into orthologous groups using the RBBH method (Chen *et al.* 2010). The Resampled Inference of Orthologs (RIO) method uses a domain-based sequence alignment to align the input gene sequences and produce phylogenetic trees (which are then compared to a species tree to predict orthologs) (Zmasek and Eddy 2002). Another method takes a drastically different approach. The Hierarchical grouping of Orthologous and Paralogous Sequences (HOPS) database treats protein domains as the unit of orthology. It constructs phylogenetic trees with the sequences of protein domains from Pfam families (which are then used to predict orthologs) (Storm and Sonnhammer 2003).

Synteny

Synteny is the conservation of orthologous genes local ordering or neighbourhood (Dewey 2011). It has been used as a metric for assessing ortholog prediction because genome segments with conserved order in different species suggests that the contained genes diverged from the same set of ancestral genes (Hulsen *et al.* 2006; Lemoine, Lespinet, and Labedan 2007; Dewey 2011). Conserved gene neighbourhoods can be also be an indicator of related gene functions, as selective pressure can act to keep functionally related genes clustered on the chromosome (this pattern is more evident in bacteria) (Hulsen *et al.* 2006). This non-phylogenetic measure is used to complement standard ortholog prediction in several methods. In OrthoParaMap, genes in gene families identified through a phylogenetic analysis are classified as orthologs or paralogs based on whether they are located in syntenic or a duplicated region of the genome (Cannon and Young 2003). In MSOAR the ortholog prediction and gene order calculation are integrated (Fu and Jiang 2008; Shi, Zhang, and Jiang 2010). Using a combinatorial algorithm, the most parsimonious evolutionary scenario that minimizes the number of gene duplications and genome rearrangements is determined to arrive at the final ortholog assignments.

While synteny can be useful in distinguishing orthologs from paralogs, it has limitations. Gene order is highly fluid and synteny is lost relatively quickly (in bacteria and archaea the syntenic loss is even more rapid) (Lemoine, Lespinet, and Labedan 2007; Dewey 2011). This limits the phylogenetic distance between the species that can be analyzed using synteny. Tandem or segmental duplications can also create problems. Methods must also be able to distinguish between valid orthologs in syntenic regions and paralogs that appear to have conserved gene order because they are part of larger segmental duplication (Kristensen *et al.* 2011).

Examining the Phylogenetic Context of Predicted Orthologs

Tree-based methods predict orthologs by considering the divergence of all homologs in relation to the species tree. Because of this global phylogenetic approach, they are often more robust to erroneous ortholog prediction due to gene loss or asymmetrical rates of evolution. While graph-based methods have several advantages over tree-based methods, these methods typically do not consider the phylogenetic context and can generate false positive ortholog predictions when the ortholog is missing (Fulton *et al.* 2006; Kuzniar *et al.* 2008). Two heuristic-based methods have developed approaches to mitigate the misprediction due to missing ortholog genes. OMA and QuartetS are graph-based approaches that use additional outgroup genes to examine the broader phylogeny of a predicted ortholog pair in order to assess if the predicted orthologs are more likely paralogs (Altenhoff *et al.* 2011; Yu *et al.* 2011). QuartetS builds a gene tree consisting of four genes; the two predicted orthologs and two genes from a reference genome. If the branch lengths in the tree are exceptionally large, a paralog is inferred. OMA does not build a tree. It instead searches all genomes for outgroup orthologs that indicate the predicted orthologs are derived from distinct ancestral genes (so called “witnesses of non-orthology”).

1.4.2. The Ortholuge Method

Ortholuge, originally developed in the Brinkman Laboratory at Simon Fraser University, is a high-throughput method that improves the specificity of ortholog prediction (Fulton *et al.* 2006). It provides the benefits of graph-based methods including scalability, but limits false positives generated by missing orthologs, because it considers

the phylogenetic context of predicted orthologs. Ortholuge first predicts orthologs using the graph-based approach commonly called Reciprocal Best BLAST (RBB), but adds a second step where phylogenetic trees are built for each proposed orthologous gene/protein pair rooted with a suitable outgroup. This phylogenetic analysis is completed for all predicted orthologs and, coupled with a statistical analysis (Min *et al.* 2011), is used to flag orthologs that have diverged unusually versus what would be expected for the species. Many of the unusually diverging predicted orthologs are paralogs mispredicted as orthologs, or orthologs that have diverged more rapidly in one of the species (Fulton *et al.* 2006). The remaining orthologs are more likely to have retained similar functions and may be better suited for many comparative genomic analyses. The phylogenetic approach used in Ortholuge does not require a separate species tree, making it especially suited for microbial genomes where species tree construction can be complicated by horizontal gene transfer and widely differing degrees of divergence between the species being compared. Ortholuge is described in detail in section 3.

1.5. Applications of Orthologs

The increase in genomic sequencing throughput has generated a rapid increase in the number of available genome sequences (Langille *et al.* 2012; Pruitt *et al.* 2012). To make effective use of this growing resource, computational tools and databases for comparative genomics analysis must keep pace, ideally without sacrificing accuracy or performance. Computationally predicted orthologs are integral to many comparative genomics analyses. Orthologs, related genes between species that have diverged as a result of speciation, are thought to more likely have similar functions than paralogs, which are homologous genes that have arisen through gene duplication (Koonin 2005). This ortholog functional conservation hypothesis or conjecture is the basis for many comparative genomics methods using computationally predicted orthologs to infer gene functions across species. Outlined in Table 1.1 are the major types of comparative genomic analyses that require orthologs. This table indicates the types of comparative genomic analysis that explicitly depend on the predicted orthologs having similar or equivalent functions. Limitations of the ortholog functional conjecture should be

considered when interpreting the results from these types of comparative genomic analyses.

Table 1.1 Comparative Genomic Analyses that use Orthologs

Categories of comparative genomic analysis that use orthologs. The ortholog function conjecture (OFC) is an underlying assumption of many comparative genomic analyses that use orthologs. Analyses that rely on this assumption are indicated in the table.

Type	Description	Rely on OFC[a]?
Direct Inference of Gene Function	To annotate newly sequenced genomes or infer functions of uncharacterized genes, functional annotations are transferred from characterized genes in other genomes between genes that are orthologs.	Yes
Associative Inference of Gene Function	Indirect approaches examine the co-occurrence of orthologs or rearrangements of operons involving orthologs in multiple genomes to infer functions of uncharacterized genes. In co-occurrence methods, the underlying principle is that genes essential for a particular function should always appear in genomes with other genes required for that same function. The general function of a gene may be inferred through the annotations of associated co-occurring genes. Similarly, functionally-related genes are often clustered together in operons. By examining the orthologs that have recombined in operons in other genomes, the general function of a gene of interest may be inferred.	Yes
Detection of Transcriptional Regulatory Elements	Non-coding regions of genomes that are conserved across multiple species may suggest that those regions are under selection and may be functional (i.e. such as transcription factor binding site). The upstream regions of orthologs are often used in methods such as phylogenetic foot printing to detect conserved DNA signatures that may be regulatory elements.	No[b]
System-level Modeling	System-level models such as transcriptional regulatory networks or metabolic networks require large-scale gene function data. This level of data is only available in a few model organisms. In less well-studied organisms, system-level models are reconstructed from the orthologous gene interactions in these model organisms.	Yes
Measuring Sequence Level Evolutionary Change	Although there is not a strict requirement for orthologs (analyses involving paralogs can be informative if user is aware of the homology), orthologs are frequently used to measure the rate of evolution in gene families. Analysis will determine, for example, positive or negative selective forces acting on genes in specific species lineages.	No[c]
Protein Structure Modeling	The three dimensional structures of proteins can be highly conserved amoung homologs (often the structure is more conserved than the sequence). Comparative protein structure modeling infers the structures of uncharacterized proteins by aligning the protein sequence to homologous structural templates.	No[d]

Type	Description	Rely on OFC[a]?
Detecting Protein Domain Sequence Signatures	A domain is a discrete independently folding unit of a protein. Domains can have characteristic sequences that can be revealed through a multiple sequence alignment of homologous protein domains.	No[d]
Detecting Gene Level Differences	To determine the genetic elements that are conserved or distinct between taxonomic clades, phyletic profiles of orthologs are frequently used. For example, phyletic profiling is often used to identify genes unique to pathogens but absent from related non-pathogenic species.	Yes

[a] Is the ortholog function conjecture (OFC) critical to the application of orthologs in the method?
[b] Although there is no requirement for genes to be functionally equivalent, genes need to have the same regulatory mechanisms for approach to work.
[c] Genes with distinct function can be used, especially when trying to determine the sequence changes associated with a change in function.
[d] There is no requirement for genes to be functionally equivalent, however genes must have homologous domains or identical structures for approach to work.

1.6. Goal of Present Research

The expansion of available genome sequences has drastically increased the scale and scope of analyses that use ortholog prediction. Ortholog prediction is routinely conducted on a genome-wide scale across thousands of genomes and is critical to the annotation of genes in newly sequenced genomes. Ortholog prediction is also a key component in a variety of comparative genomics analysis.

To address the increasing role and demands of ortholog prediction, my research focused on two aspects related to orthology: improving the accuracy and power of computational ortholog prediction and the application of predicted orthologs in comparative genomics analysis.

Chapter 2 describes three distinct use cases of orthologs, in which I predicted and then employed orthologs to carry out three types of comparative genomics analyses. By referencing these use cases, my research goal became the identification of problem areas in ortholog prediction and the examination of the effectiveness of orthologs in these studies. Through the evaluation of these projects, I wanted to identify limitations of orthologs in comparative genomics analysis.

Ortholuge is an ortholog prediction approach that was previously developed to improve the accuracy of computation ortholog prediction in a high-throughput setting. When ortholog prediction accuracy is important, a tool like Ortholuge is extremely useful. My second research goal was to augment computational ortholog prediction by evaluating and enhancing the Ortholuge method. I evaluated the performance of Ortholuge, compared it to contemporary methods, and then further developed its functionality. This work is in Chapters 3 and 4.

To make Ortholuge results widely available to the research community, my last research goal was to develop a database of Ortholuge ortholog predictions for all fully sequenced bacteria and archaea species. The database will be routinely updated to keep up with the growing genomic data. The database development is described in Chapter 5.

Overall, the aim of this work is to advance comparative-genomic based research by building upon and disseminating the results of a promising ortholog prediction method; Ortholuge and by evaluating some of the methods in which predicted orthologs are used.

The work and analyses reported in this thesis has been carried out by me. In two cases, I received significant help in the completion of a thesis objective. In those sections, I outline in detail my contributions (see section 4.4 and chapter 5).

2. Use Cases of Orthologs in Comparative Genomics Analysis

2.1. Preamble

A component of my thesis involved the application of comparative genomics methods to address biological questions. In three separate projects, orthologs were used to transfer gene annotation, identify conserved genes and build prototype biological networks. In addition to reporting the biological discoveries, challenges in employing orthologs were also identified in these use cases. In the following sections, the biological results and methodological challenges are described individually for each project.

2.2. Case I: Identifying Metazoan-Associated Genes using Ortholog-based Phyletic Patterns

Portions of this chapter have been submitted for publication in the article "Identification of 526 Conserved Metazoan Genetic Innovations Exposes a New Role for Cofactor E-like in Neuronal Microtubule Homeostasis", co-authored by M.Y. Frederic, V.F. Lundin, M.D. Whiteside, D.K. Tu, S.Y.C. Kang, D.L. Baille, J.M. Bellanger, H. Hutter, F.S.L. Brinkman and M.R. Leroux in PLoS Genetics © 2013 Frederic et al.

2.2.1. Introduction

Multicellular organisms arose independently in the metazoan animal lineage. The molecular basis of multicellularity is of interest from a scientific viewpoint and also a medical viewpoint as many human diseases are the result of dysfunction in cell-cell communication or cellular proliferation. To uncover some of the molecular mechanisms

that support multicellularity, we used phyletic profiling to identify metazoan-associated genes that appear exclusively in and are widely conserved across metazoan species. This study also examines the effectiveness and challenges encountered in a large-scale phyletic profile-based comparative genomic analysis.

2.2.2. Methods

Identification of Metazoan-associated Orthologs

Ortholog groups computed using the OrthoMCL methodology were obtained from OrthoMCL-DB (Chen *et al.* 2006). This ortholog analysis consisted of 138 species; 26 metazoans and 112 non-metazoans. The 26 metazoan species represented a wide cross-section of the metazoan species tree. They included 12 Chordate species, 8 Arthopod species, 3 Nematode species, 1 Platyhelminth species and 2 additional metazoan species. To obtain a list of metazoan-associated ortholog groups from the OrthoMCL groups, groups were selected using two main criteria. First, groups were required to have sufficient coverage across metazoan species. Groups had to have at minimum, orthologs in:

i. 9 of 11 Chordate species
ii. 6 of 8 Arthopod species
iii. 2 of 3 Nematode species
iv. 1 of 2 additional metazoans (*Nematostella vectensis* & *Ciona intestinalis*)
v. A combined total of orthologs in at least 20 of 24 metazoan species

This criterion ensured a high degree of conservation throughout the metazoan species tree while permitting a limited number of false negatives (i.e. orthologs missing due to errors by OrthoMCL or incomplete genome information (Li, Stoeckert, and Roos 2003)). The platyhelminth species *Schtistosoma* was excluded from our analysis because of its low degree of genome completeness and the lack of a closely-related species in OrthoMCL-DB that could compensate for the low platyhelminth genome coverage. The second criterion ensures that the group's orthologs are found exclusively in metazoan species, while accommodating a limited number of falsely predicted non-metazoan orthologs. Groups could have at maximum, orthologs in 2 of 112 non-

metazoan species provided that those orthologs had 3 or less reciprocal best BLAST hit (RBBH) connections to the metazoan orthologs in the group. These weakly connected, singular non-metazoan orthologs likely represent false predictions by OrthoMCL (Li, Stoeckert, and Roos 2003). The resulting list of ortholog groups was divided into two sets based on the presence or absence of a *Trichoplax adherens* ortholog in group.

Determining Coverage in Genomes

Percent coverage of conserved eukaryotic genes (CEGs) was employed as an indicator of overall genome coverage in the metazoan species used in the study. CEGs were isolated following the methodology of Parra *et al.* (Parra *et al.* 2009). Briefly, genes conserved in the well-studied model organisms; *Homo sapiens*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Saccharyomyces cerevisiae*, *Arabidopsis thaliana* and *Schizosaccharomyces pombe* with a copy number of one were considered CEGs. Orthologs of the CEGs were identified in the species of interest and a final value indicating CEG coverage was computed.

Functional Analysis

To identify pathways and functional categories that were over- or under-represented with metazoan-associated human genes, a hypergeometric test was used. Multiple hypothesis correction was performed using a Benjamini-Hochburg procedure and results were considered significant if the corrected *p*-value was less than 0.05. Pathways were obtained from InnateDB and functional categories from KEGG BRITE database (Kanehisa and Goto 2000; Kanehisa *et al.* 2008). The background gene sets used in the tests comprised all human genes with pathway or category annotations in these databases. Differences in the proportions of human metazoan-associated genes with or without a *Trichoplax adherens* ortholog in each category were tested using the Fisher's exact test.

2.2.3. Results and Discussion

Identification of Metazoan-associated Orthologs

The goal of this study was to identify orthologs that were conserved across metazoans and absent in non-metazoan species. This comparative analysis should

identify genes that emerged early in metazoan lineage during the evolution of multicellularity and have been maintained in metazoan species. We used the ortholog prediction tool; OrthoMCL (Li, Stoeckert, and Roos 2003), to identify metazoan-associated orthologs. OrthoMCL is a highly accurate tool capable of performing automated genome-wide prediction of orthologs across multiple species (Li, Stoeckert, and Roos 2003; Chen *et al.* 2007). It initially predicts orthologous gene pairs by reciprocal best BLAST (RBB) analysis and then clusters the RBB pairs into highly connected multi-species ortholog groups. Extending pair-wise ortholog prediction to multiple species is computationally intensive. At the time of this analysis, OrthoMCL-DB contained the most phylogenetic diverse set of multi-species ortholog groups (pair-wise ortholog prediction tools such as Ortholuge would require grouping ortholog predictions on a massively large scale). Ortholog predictions for 138 species, including 26 metazoan species were obtained from OrthoMCL-DB (Chen *et al.* 2006). The metazoan species are widely dispersed and represent several distinct phylogenetic clades.

One of the challenges in identifying conserved orthologs across species is accommodating the varying degrees of genome completeness. We wanted to avoid missing valid metazoan-associated orthologs in cases where a potentially conserved gene was omitted due to incomplete genome sequence. We performed a cursory assessment of genome completeness in the metazoan species using a list of widely conserved, low-copy number core eukaryotic genes (Table 2.1). The presence or absence of these core eukaryotic genes (CEG) provides an approximation of the gene coverage (Parra *et al.* 2009). Coverage of the CEGs was on average 94% in the metazoan genomes used in this study, but was observed as low as 79% in the case of *Ciona intestinalis*. To accommodate this source of error, we adopted flexible criteria for selecting metazoan-associated orthologs. We divided the species into their taxonomic clades, and for an ortholog to be classified as metazoan-associated, we required that it be found in the majority of species in every metazoan clade, but could be missing from a few species in separate clades (see Methods section for further details). Because this modified criteria ensures that the gene is strongly represented in all metazoan clades, we feel that we are more often selecting for groups where an ortholog is absent due to incomplete genome sequence rather than gene loss.

Table 2.1 Coverage of Core Eukaryotic Genes in Metazoan Species

Gene space coverage was assessed by determining percent coverage of core eukaryotic genes (CEGs) in the genomes of the metazoan species used in the study. Genomes whose CEG coverage falls in the lower quartile are highlighted in grey.

Metazoan Genome	Common Name	Gene Space Coverage (%)
Aedes aegypti	Yellow Fever Mosquito	96.1
Anopheles gambiae str. PEST	African Malaria Mosquito	96.6
Apis mellifera	Western Honey Bee	91.2
Acyrthosiphon pisum	Pea Aphid	92.4
Brugia malayi	Filarial Nematode Worm	96.1
Bombyx mori	Silk Moth	91.7
Caenorhabditis briggsae AF16	-	98.5
Caenorhabditis elegans [a]	-	100.0
Ciona intestinalis	Transparent Sea Squirt	79.0
Canis lupus familiaris	Dog	98.8
Culex pipiens	Northern House Mosquito	94.9
Drosophila melanogaster [a]	Fruit Fly	100.0
Danio rerio	Zebrafish	96.6
Gallus gallus	Chicken	88.3
Homo sapiens [a]	Human	100.0
Monodelphis domestica	Gray Short-Tailed Opossum	97.6
Mus musculus [a]	Mouse	100.0
Nematostella vectensis	Starlet Sea Anemone	97.3
Ornithorhynchus anatinus	Duckbill Platypus	84.1
Pediculus humanus	Human Louse	97.8
Pan troglodytes	Chimpanzee	98.3
Rattus norvegicus	Rat	96.8
Schistosoma mansoni	Blood Fluke	91.0
Trichoplax adhaerens	-	95.9
Tetraodon nigroviridis	Spotted Green Pufferfish	95.9
Takifugu rubripes	Japanese Pufferfish	96.1

[a] Genome was used to isolate the CEGs and therefore will have 100% coverage.

OrthoMCL is an effective tool for indentifying metazoan-associated orthologs. It does, however, generate small numbers of falsely predicted orthologs (false positives). In the OrthoMCL data, singular non-metazoan genes would occasionally be clustered with a group of metazoan-conserved orthologs. These genes would often only be predicted to be orthologs to one or two other metazoan species genes in the group based on existence of an RBB relationship. From a phylogenetic perspective, it is more likely that these weakly-related, singleton genes are false positives than true orthologs. To prevent these groups from being excluded in our list of metazoan-associated orthologs, a second provision was added; an ortholog group could contain a predicted ortholog from 1-2 non-metazoan species (out of a possible 112 non-metazoan species), provided that the non-metazoan predicted orthologs were weakly connected to the metazoan species in the group (i.e. had limited number of RBB connections to the metazoan orthologs). This allowance maintains a high level of stringency, but includes ortholog groups containing apparent falsely-predicted non-metazoan orthologs (see Methods section for further details).

Applying these criteria, 526 groups of orthologs were identified as being conserved in metazoans and absent in non-metazoans (Appendix A.). Some species, especially vertebrates, have undergone extensive gene duplication, resulting in multiple proteins per ortholog group. In the 526 groups, there were 898 human proteins (on average 1.7 proteins per group) and 577 *Caenorhabditis elegans* proteins (average 1.1 proteins per group). *Trichoplax adherens* is one of the earliest diverging multicellular animals and likely represents a primitive metazoan mode of life. Because of *T. adherens*' unique position in the metazoan tree, we distinguished between ortholog groups that either contain or lack a *T. adherens* ortholog. Of the 526 metazoan-associated ortholog groups, 326 groups have a *T. adherens* ortholog and 180 lack a *T. adherens* ortholog.

Functions of Metazoan-associated Genes

In the 526 metazoan-associated ortholog groups identified, 64 are essentially uncharacterized based on an examination of the human and *C. elegans* gene function annotations. Metazoan-associated genes represent conserved and potentially critical mechanisms for multicellular systems.

Many of the annotated metazoan-associated genes clearly have a role in the development and maintenance of multicellularity. When we examined the functional categorization of the human metazoan-associated genes, we found that these genes were disproportionately associated with categories such as cell-cell communication, development, cell motility, and signal transduction (Figure 2.1). Several organ systems including the nervous, endocrine and circulatory systems were also overly represented.

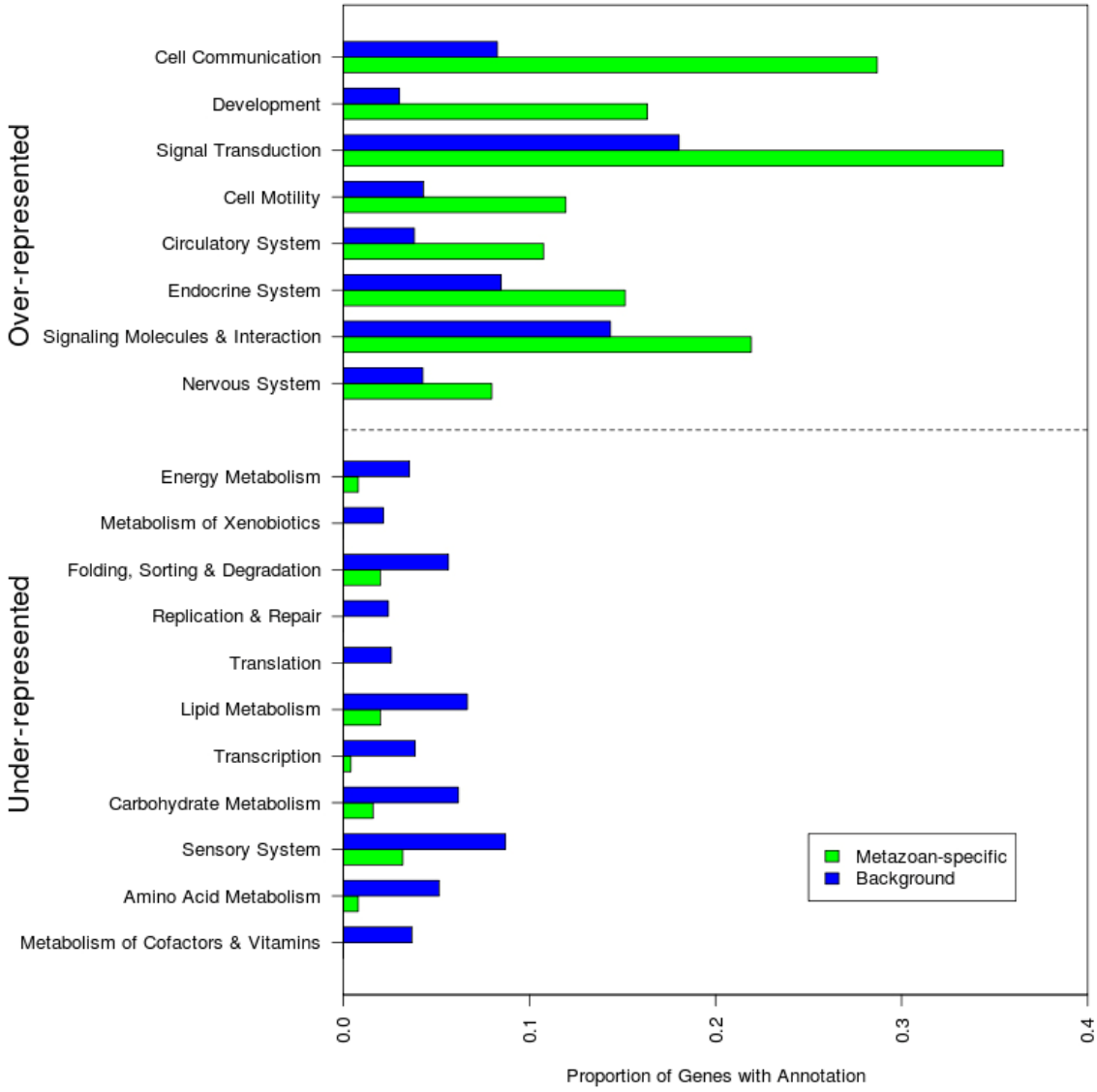


Figure 2.1 KEGG BRITE Functional Categories Containing Disproportionate Numbers of Human Metazoan-Associated Orthologs

KEGG BRITE functional categories were tested for over- or under-representation of human metazoan-associated orthologs. The bar graph shows the proportion of the total human metazoan-associated genes found in a functional category. Only functional categories that were statistically over- or under-represented are shown. The background is the proportion of all human genes in the functional category.

A number of functional categories were under-represented; containing disproportionately few metazoan-associated genes. Core processes such as transcription, replication and repair, amino acid & carbohydrate metabolism were under-

represented because the majority of genes belonging to these categories were found in many non-metazoans and likely evolved prior to divergence of metazoan lineage. In other under-represented categories, such as the sensory system, the majority of genes were only found in a subset of metazoan species, suggesting that the genes in this category evolved later in the metazoan lineage.

T. adherens is the earliest branching metazoan species used in our comparative analysis. It is a morphologically simple organism with only four described cell types and lacks specialized neuronal, muscle and sensory cells (Srivastava *et al.* 2008). To gain insight into the evolution of the nervous system, we also identified the gene complement that was found in all metazoans except for *T. adherens*. We found 180 metazoan-associated ortholog groups that lack a *T. adherens* ortholog. Interestingly, the only functional category that had a significant difference in proportion of genes with a *T. adherens* ortholog compared to genes without a *T. adherens* ortholog was Glycan Biosynthesis & Metabolism. Glycan biosynthesis and metabolism contained a higher proportion of genes that were not found in *T. adherens*. All other categories contained a significant proportion of genes that have a *T. adherens* ortholog as well as many genes that lack a *T. adherens* ortholog. This includes the genes belonging to the nervous system category. The implication of this result is that a significant number of the orthologs belonging to the functional categories associated multicellularity already existed in the earliest diverging metazoan; *T. adherens*. It also suggests that these functional categories associated multicellularity underwent significant gene expansion subsequent to the divergence of *T. adherens*.

To refine which functional components are conserved in and exclusive to metazoans, we also examined metazoan-associated gene functions at the pathway level. We identified known human pathways that contain a significant proportion of metazoan-associated orthologs (Appendix B.). Components of several pathways were found to be unique to metazoans. These pathway components are involved in multiple biological processes, including cell adhesion, cell-cell communication, development, cell motility, cell death, and the endocrine, immune and nervous systems.

Metazoan-associated Genes in Signalling Pathways

Several mitogen-activated protein kinase (MAPK) ortholog groups are associated only with metazoans. These orthologs appear to form a reduced but potentially functional MAPK pathway. The orthologs comprise most of the JNK MAPK pathway (Johnson and Lapadat 2002) and represent most major MAPK kinases (JNK, MKK, MKKK and MAPKAPK). All metazoan species, including *T. adherens*, have a single ortholog for these kinases, except for the chordate species which have multiple copies.

Several gene families involved in cAMP signaling are also conserved and exclusive to metazoans. cAMP is a ubiquitous second messenger that integrates numerous extracellular signals. In the protypical cAMP signaling pathway, G-protein coupled receptors (GPCR) detect exogenous signals such as hormones, metabolites, neural ligands or growth factors. Upon ligand binding, the G-protein subunit of the GPCR activates or represses an adenylyl cyclase (AC) enzyme, which produce cAMP and in turn activate protein kinase A (Patel 2001; Willoughby and Cooper 2007; Dessauer 2009). Numerous signalling pathways converge in the cAMP signaling cascade. The specificity of the cAMP response comes from the numerous distinct AC-GPCR interactions. In humans, there are 10 ACs, and over 900 GPCR genes. Our comparative analysis shows that all metazoan species have orthologous AC genes that fall into two ortholog groups; one group that contains the membrane bound adenylyl cyclases which has multiple copies per species ranging from two in *T. adherens* to eight in humans (ADCY1-8), and another group that contains a soluble adenylyl cyclase with only one copy found in each species (ADCY9 in humans). GPCRs are a common signal receptor/transducer in neuroendocrine processes. *T. adherens* lacks a recognizable nervous system, but a number of GPCRs involved in human neuroendocrine pathways have orthologs in *T. adherens* and are furthermore conserved in and exclusive to metazoans (Appendix C.). In higher metazoans, these metazoan-associated GPCRs detect glycoproteins (the human orthologs bind follicular stimulating hormone, thyroid stimulating hormone and luteinizing hormone/choriogonadotropin), neurotransmitters (in humans, the orthologs are the metabolotropic receptors GRM1-7, GABBR2) and a neuropeptide (orthologous to the human galanin GPCR) (Schiöth and Fredriksson 2005; Kamesh, Aradhyam, and Manoj 2008). The *C. elegans* neuropeptide GPCR ortholog is the allatostatin/galanin-like gene *npr-9*. In *C. elegans* Npr-9 is expressed in neurons

where it helps regulate foraging behavior (Bendena *et al.* 2008). Although *T. adherens* does not have a nervous system and contains no neuronal cells, it has been shown to exhibit behavioral responses to stimuli (Srivastava *et al.* 2008). These conserved neuroendocrine pathway components; the adenylyl cyclases and GPCRs may be part of a primitive stimulis response signaling system that existed in the last common ancestor of *T. adherens* and higher metazoans. Numerous other GPCRs have no ortholog in *T. adherens*, but are conserved across cnidarians (e.g. *Nematostella vectensis*) and bilaterians.

Gene Duplication in Metazoan-Associated Genes

The pathway analysis identified components of human pathways that are conserved across metazoans and are also unique to metazoans based on the presence or absence of an ortholog. As a whole, many of the metazoan-associated pathway components appear to be downstream functional modules that process or detect extracellular signals (e.g. map kinases, adenylyl cyclases, GPCRs). These conserved functional modules have been incorporated into a range of biological processes, such as cell-cell communication, development, adhesion, immunity, and the nervous system, and are also active in several cell types. Also, the metazoan-associated orthologs often are part of large gene families. These observations; the wide-spread integration and high level of duplication of the metazoan-associated orthologs, suggest that gene subfunctionalization followed by functional divergence has been a significant evolutionary process that has contributed to the development of unique and increasingly complex metazoan species (Arendt 2008).

The phyletic pattern-based search used in this study identified a number of potential metazoan-associated genes. While the phyletic search ensures that orthologs are conserved in all metazoan species, it does not take into account gene copy number. An important consideration when drawing conclusions about the larger biological functions conserved in metazoans, is how to interpret the functional equivalency of genes that have undergone duplication (especially when successive duplications have occurred). The conservation of genes belonging to the JNK-MAPK and GPCR-AC pathways would suggest that these signal transducing mechanisms evolved from ancestral pathways in the last common ancestor of metazoans and that they were

potentially critical components in the development of multicellularity. The duplication of GPCR and MAPK genes, however, has created multiple copies and branches of the pathways that respond to diverse stimuli and generate distinct outcomes across the metazoan species.

2.2.4. Case I Conclusions

Using phyletic profiling, 526 genes were found conserved in and unique to metazoan species. Based on *H. sapiens* gene function annotations, many of the metazoan-associated genes make up molecular processes involved in cell-cell communication, multicellular development and signal transduction as well as biological systems such as the nervous, endocrine and immune systems. The functional equivalence of the metazoan-associated genes, however, is uncertain. While gene families are conserved, frequent gene duplications appears to have lead to specialization of gene functions within a gene family in the individual species.

One of the challenges encountered with using a phyletic profiling approach to identifying metazoan-associated genes was dealing with missing genome data. Most metazoan genome sequences were incomplete to some degree. As well, the OrthoMCL tool used to identify the ortholog groups generates a limited number of false ortholog predictions. To avoid excluding valid metazoan-associated genes, a flexible approach was adopted to account for these sources of error. If the distribution of the gene was sufficient in all metazoan clades, it was not required to be strictly found in every species genome. Similarly, non-metazoan orthologs were ignored if they were weakly connected to the metazoan genes in the ortholog group (based on RBB relationships). Use of a flexible approach greatly increased the number of metazoan-associated genes that were identified.

2.3. Case II: Gene Expression Differences in Epidemic *Pseudomonas aeruginosa* Strains

2.3.1. Introduction

Pseudomonas aeruginosa, an opportunistic pathogen, is a critical threat to Cystic Fibrosis (CF) patients. The viscous mucus produced in the lungs of CF patients creates a unique microaerobic environment that *P. aeruginosa* can colonize and establish chronic infections. Inflammation resulting from the infections causes progressive damage to the patients' lungs and is the major contributor to the mortality and morbidity in CF patients. Recent genotyping of CF patient isolates has shown that a large proportion of infections in multiple locations appear to be caused by the same clones. These highly prevalent clones have not been found in the wider environment suggesting that they have acquired mechanisms of enhanced transmissibility and are being transferred between patients. Several highly prevalent clones or epidemic strains have been reported. The best studied epidemic strains include the Liverpool epidemic strain (LES) (Jones *et al.* 2001; Scott and Pitt 2004) and the Australian epidemic strains 1 & 2 (AES-1, AES-2) (Armstrong *et al.* 2003; O'Carroll *et al.* 2004). In addition to increased infectivity, these strains are characterized by distinctive phenotypes; superinfectivity, increased induction of quorum sensing, increased and prolonged protease activity and biofilm hyperproduction (Salunkhe *et al.* 2005; Tingpej *et al.* 2007; Kukavica-Ibrulj *et al.* 2008; Manos *et al.* 2008; Winstanley *et al.* 2009). LES, AES-1 and AES-2 are also associated with poor patient outcomes (Salunkhe *et al.* 2005; Naughton *et al.* 2011).

Studies into the evolution of *P. aeruginosa* have identified two major sources of variability across strains; novel genes acquired through horizontal gene transfer and adaptive mutation in the regulation of the core genome. The core genome is the set of genes found in the majority of strains thought to be indispensible to the species. Both sources of variation have been shown to contribute to the virulence of epidemic strains (Salunkhe *et al.* 2005; Manos *et al.* 2008; Manos *et al.* 2009; Winstanley *et al.* 2009). To date, three independent transcriptome analyses have been conducted on epidemic strains: LES, AES-1 and AES-2 (Salunkhe *et al.* 2005; Manos *et al.* 2008; Manos *et al.* 2009). Limited by the microarray content, these studies focused on identifying adaptations in the regulation of the core genome (the microarray is based on laboratory

strain PAO1 isolated from a burn wound, however a large proportion of the *P. aeruginosa* genome, including many virulence factors are conserved across strains). The genome of one epidemic strain isolate has been sequenced: LESB58. The sequencing identified 11 genomics islands (Winstanley *et al.* 2009). The genetic and gene expression studies did not find one particular adaptation that could solely account for the increase in transmissibility. Instead the virulence of epidemic *P. aeruginosa* was determined to be multi-factorial. These studies also demonstrated that there is considerable variation across epidemic strains and even among isolates from a single epidemic strain. This variation makes it difficult to determine the relative importance of the changes in epidemic strains.

With the spectre of *P. aeruginosa* strains with enhanced transmissibility, identifying the determinants behind the increase in transmissibility is critical to treatment of patients and development of future drugs. In this study, we conduct a meta-analysis of the available gene expression data for LES, AES-1 and AES-2. The gene expression meta-analysis is combined with a systems-level functional analysis and comparative genomics-based prediction of regulons. Previous transciptomics experiments have shown that there are gene expression changes associated independently with each epidemic strain, and also that there is considerable variability across epidemic strains. By focusing on gene regulatory adaptations that have been selected for in multiple epidemic strains, we hope to identify the changes that are driving the increased infectivity in epidemic strains. Using a systems biology approach, specific biological functions are assigned to the gene expression differences found in epidemic strains. Computational comparative genomics methods are also employed to predict the regulons involved in the gene expression changes.

2.3.2. Methods

Microarray Data & Processing

Gene expression data was collected from three separate studies (see Appendix D.: data was obtained from the Gene Expression Omnibus (GEO) or requested from the authors). All studies were conducted on the same Affymetrix platform (GEO accession: GPL84) which is based on PAO1 transcriptome. The data includes isolates from the

Liverpool Epidemic strain, Australian Epidemic strain 1 and Australian Epidemic strain 2 as well non-clonal and PAO1 samples used as a comparison (Salunkhe *et al.* 2005; Manos *et al.* 2008; Manos *et al.* 2009). Only samples grown in Luria-Bertani (LB) liquid media were used in the meta-analysis (LB media was the only common test condition across the three experiments. LB is commonly used as a non-specific media for planktonic *P. aeruginosa* growth). Samples from experiments GSE10304 and GSE6122 were grown to mid-log phase at 35˚C, while samples from the Salunkhe *et al.* study were grown to late log phase at 37˚C. As described in the source studies, an isolate is determined to belong to a particular strain when the PFGE analysis of the *SpeI* digestion shows less than a three band difference from other strain isolates. A non-clonal strain is defined as infecting less than three patients. The quality of the microarrays was assessed and samples with serious spatial artifacts or intensity distributions that deviated significantly were discarded.

Frozen Robust Multi-Array Analysis (fRMA), a specialized preprocessing method designed for multi-batch sample processing, was used to for microarray background correction, normalization and summarization of probe set intensities (McCall, Bolstad, and Irizarry 2010). fRMA is a variant of the Robust Multi-Array Average (RMA) microarray preprocessing procedure. It provides improvements over standard RMA for meta-analysis by estimating probe-wise and batch-wise variation separately; and also by normalizing against a large reference database of microarray experiments. These additions allow for better resolution of the biological variation from technical sources of variation. Available Affymetrix PAO1 microarray experiments were collected from GEO (GEO platform accession: GPL84) and used to construct a reference fRMA database according to the authors specifications. Standard RMA background correction was performed, followed by fRMA's quantile normalization and batch summarization.

Differential Gene Expression Testing

An initial test was performed to identify individual genes that are differentially expressed across the epidemic strains compared to the non-frequent clones and PAO1. The LIMMA package from Bioconductor was used to perform this test; it uses a linear model and an empirical Bayes procedure to estimate gene expression levels. The LIMMA *p*-values were corrected for multiple testing using Storey's *q*-value method

(Storey and Tibshirani 2003; Gentleman *et al.* 2004; Smyth 2004). Genes with a false discovery rate of 0.1 were selected as significantly differentially expressed.

Group-wise differential expression between epidemic and non-frequent clones and PAO1 was tested using the global test statistic for gene sets representing KEGG, PseudoCyc and PseudoCAP pathway annotations and operons obtained from pseudomonas.com, regulons from PRODORIC and predicted transcriptional modules identified in this study (Kanehisa and Goto 2000; Romero and Karp 2003; Winsor *et al.* 2005; Grote *et al.* 2009; Winsor *et al.* 2010). The global test (available through Bioconductor) tests whether the gene expression differences in a group are significantly correlated with the epidemic status of the isolates (Goeman *et al.* 2004). Defaults were used for the global test arguments, with the following exceptions; the microarray's data source was added as a nuisance variable to the test and conserved directionality was required for the operon tests. The Benjamini-Hochburg procedure was used to adjust for multiple hypothesis testing and gene sets with a corrected p-value < 0.05 were declared as significant in this step.

A significant result from the global test does not necessarily indicate that all epidemic strains are contributing to the result. An additional filtering procedure was used to isolate significantly differentially-expressed sets that are supported by multiple strains. The influence on the test result of each isolate can be extrapolated within the global test statistical framework and expressed as a p-value. Gene sets that were declared significant in the first step were then examined further to determine the number of strains contributing to the significant result. Specifically, a test was declared significant, if the influence values of more than half of the isolates in a strain group have a Benjamini-Hochburg-based false discovery rate of 0.05 and this was found for at least three of the five strain groups. The strain groups are LES, AES-1, AES-2, PAO1, and non-frequent clones. Similarly, the contribution of the genes within the gene sets to the global test statistic is also of interest. The global test provides a framework for quantifying gene influences, as well as a procedure to control family-wise error rate for a natural hierarchical sub-grouping of the genes within the gene sets (Meinshausen 2008). Genes with a multiple hypothesis corrected p-value of <= 0.05 were declared as effect-causing genes within a gene set.

Predicting Transcriptional Modules

We used *in silico* methods to predict additional *P. aeruginosa* transcriptional modules (TMs). In particular, we wanted to identify TMs that were associated with the differentially expressed gene sets discovered in the previous steps of the analysis. Transcriptional modules are sets of genes that are transcriptionally co-regulated under the same conditions. The co-regulation is due to the genes having the same transcription factor binding site motif profiles. The PAO1 laboratory strain has the largest number of available gene expression experiments and was used as a model for predicting *P. aeruginosa* TMs by clustering genes with similar expression profiles. PAO1 experiments representing 20 microarray experiments were collected from Gene Expression Omnibus (GEO) (Barrett *et al.* 2011) and pre-processed as described previously. The gene expression data matrix showed a characteristic checker board pattern indicating that overlapping sets of genes were co-expressed under different sets of conditions (data not shown). A biclustering method, Iterative Search Algorithm (ISA), was selected over global gene expression clustering methods, because the ISA algorithm clusters conditions and genes simultaneously without trying to optimize all groups at once and is able to more effectively capture transient co-expression patterns (Bergmann, Ihmels, and Barkai 2003; Ihmels, Bergmann, and Barkai 2004). ISA can also accept user-specified genes as seeds to search for associated biclusters; a feature that allowed us to direct the search toward TMs of interest. We used the ISA implementation available through EISA Bioconductor package (Csárdi, Kutalik, and Bergmann 2010). Differentially-expressed genes from the individual gene tests, and effect-causing genes from the gene set tests were used as seeds for the ISA algorithm. We used a range of gene and condition search thresholds and merged modules if there was a high degree of overlap (corresponding to a module correlation of 0.8).

TFBS Motif Identification in the Predicted Transcriptional Modules

DNA motifs that are statistically over-represented in the upstream regions of module genes may represent potential regulatory elements that are controlling the co-expression of the transcriptional module. These potential transcription factor binding sites (TFBS) provide an independent source of *in silico*-based evidence that supports the grouping the TM genes. We obtained the intergenic DNA regions upstream of all transcriptional units in each module (up to a maximum length of 500 bp) and then

searched for conserved motifs using the program MEME (the acronym MEME stands for Multiple EM for Motif Elicitation) (Bailey and Elkan 1994; Bailey *et al.* 2010). The following MEME options were used: one motif instance per transcriptional unit was searched for on either strand, with a size between 5-35 bp. A total of three separate motifs were identified per module. A 1st-order Markov model of PAO1 inter-genic regions was used to define background nucleotide frequency. Position specific priors were also computed for each module and used in the search. Each of the three discovered motifs was tested for over-representation among the module genes. Motifs instances in the PAO1 genome were identified using the program FIMO (FIMO stands for Find Individual Motif Occurrences. A discovered motif instance required a *p*-value < 1e-3) (Grant, Bailey, and Noble 2011) and the hypergeometric test was used to determine if a motif was more often associated with module genes versus all other non-module genes in the genome. We also extended the motif search to orthologous transcriptional units in other *Pseudomonas* species (an operon was deemed orthologous if the majority of its genes were orthologous). Motifs that are conserved across species suggests that the DNA elements are functional (although lack of cross-species conservation does not indicate the absence of a valid functional DNA element; many *P. aeruginosa* regulatory elements will not be found in other *Pseudomonas* species). The FIMO program and hypergeometric test was used to identify motifs and test over-representation as described above (Grant, Bailey, and Noble 2011).

Functional Characterization for the Predicted Transcriptional Modules

To elucidate the biological functions associated with the predicted TMs, enrichment of pathway and gene annotations was computed. Annotations for KEGG pathways, PseudoCyc pathways, PseudoCAP pathways, PseudoCAP functional categories, and the Gene Ontology (GO) were obtained from pseudomonas.com for the TM genes and the hypergeometric statistic was used to test for enrichment of these functional groups in each TM (Ashburner *et al.* 2000; Romero and Karp 2003; Harris *et al.* 2004; Winsor *et al.* 2005; Csárdi, Kutalik, and Bergmann 2010; Winsor *et al.* 2010). The tests were declared significant if the Benjamini-Hochburg corrected *p*-value was < 0.05.

2.3.3. Results and Discussion

In this meta-analysis, we compare the microarray data from three separate epidemic strains LES, AES-1 and AES-2 against a panel of PAO1 and non-clonal *P. aeruginosa* microarrays, which were pooled to provide a diverse comparison set (Salunkhe *et al.* 2005; Manos *et al.* 2008; Manos *et al.* 2009). Differential expression was examined at multiple levels using a series of individual gene and gene group tests. The group-based tests expand upon gene-based testing by examining specific biological systems for changes in expression. Table 2.2 provides an overview of the types of analyses conducted and number of significant results obtained for each analysis. At the gene level, 24 genes were found to be differentially expressed in epidemic strains, 4 up- and 20 down-regulated in epidemic *P. aeruginosa* (Table 2.3). Using the group-based differential expression test, we found 73 differentially expressed pathways and 20 differentially expressed operons, 4 up- and 16 down-regulated (Appendix E., Appendix F., respectively). Pathway expression changes were not strictly up- or down-regulated. The group-based test requires gene sets to be defined *a priori* and in this analysis, biological systems annotated by automated means; KEGG and PseudoCyc pathways and operons, as well as expert curated systems; PseudoCAP pathways, were used. Gene expression changes were also investigated at a regulon level (groups of genes co-regulated by the same set of transcriptional regulators). We applied the group-based statistic to test experimentally-obtained and *in silico*-predicted regulons for differential expression. Six regulons tested positive for differential expression (Table 6).

Table 2.2 Summary of Differential Expression Testing Results

Differential expression testing was performed for individual genes as well as at a systems level. Summarized below are the units tested and number of significant differentially expressed results obtained in epidemic strains of *P. aeruginosa* compared to non-clonal strains under the same conditions.

Test Unit	Differentially Expressed	Comments
Genes	24	4 up-regulated / 20 down-regulated genes in epidemic strains
Pathways	73	Pathways from multiple sources. After redundant pathways were merged, 73 formed 40 distinct pathway groups
Operons	20	4 up-regulated / 16 down-regulated in epidemic strains

Test Unit	Differentially Expressed	Comments
Regulons	6	2 experimentally identified regulons, plus 4 computationally predicted regulons

Table 2.3 Differentially Expressed Genes in Epidemic P. aeruginosa

Annotated *P. aeruginosa* genes that are differentially expressed in epidemic strains compared to non-clonal and PAO1 strains (hypothetical differentially expressed genes are not listed). The false discovery rate *q*-value is computed from the *p*-value given by LIMMA's moderated *t* statistic.

Locus	Gene	Description	Log2 Fold Change	False Discovery Rate (q-value)
PA1429		probable cation-transporting P-type ATPase	1.514	0.006
PA4042	*xseB*	exodeoxyribonuclease VII small subunit	-0.921	0.013
PA1557	*ccoN2*	Cytochrome c oxidase, cbb3-type, CcoN subunit	-1.956	0.080
PA3172		probable hydrolase	-0.961	0.080
PA0366		probable aldehyde dehydrogenase	0.875	0.080
PA0102		probable carbonic anhydrase	1.224	0.080
PA1556	*ccoO2*	Cytochrome c oxidase, cbb3-type, CcoO subunit	-1.587	0.080
PA5263	*argH*	argininosuccinate lyase	-0.877	0.080
PA1581	*sdhC*	succinate dehydrogenase (C subunit)	-1.006	0.080
PA4329	*pykA*	pyruvate kinase II	-0.795	0.080
PA2446	*gcvH2*	glycine cleavage system protein H2	-1.559	0.081
PA3527	*pyrC*	dihydroorotase	-1.492	0.097
PA2639	*nuoD*	NADH dehydrogenase I chain C,D	-1.136	0.097
PA1183	*dctA*	C4-dicarboxylate transport protein	-2.217	0.098

In order to be comprehensive, we tested biological systems from multiple sources, including PseudoCAP, PseudoCyc, and KEGG. Between the databases, there is not a consistent definition of a pathway in terms of both the scope and content, thus producing a complex redundancy in our group-based results. Also, pathways within a

single database can overlap significantly. This overlap and redundancy in our test groups resulted in multiple pathways testing positive for differential expression due to the effects of the same set of genes. We found that identifying the pathway genes that are significantly contributing to the test statistic was critical to the interpretation of the system-based result. Within the global test, it is possible to extrapolate the influence of the individual genes and samples. We clustered pathways based on their influential genes and organized the data based on this clustering to more easily identify this redundant pathway effect. We use the term 'effect-causing' genes for genes that have influence *p*-values < 0.05 (after multiple hypothesis correction). We use the sample influence to select differentially expressed gene sets that are supported by multiple epidemic strains (see Methods).

The results from the pathway and operon group-wise tests and individual gene tests, in many cases, support each other by implicating similar biological systems. The results highlight the importance of metabolic adaptation. Multiple metabolic pathways are down-regulated, suggesting that in epidemic strains, there is a redirection of resources to pathways that are optimal for the CF lung environment. Quorum sensing is one of the few systems that was found in this analysis to be up-regulated in epidemic strains. Quorum sensing allows a bacterial population to determine its cellular density through the release and detection of signalling molecules in the environment. Many behaviors are altered when the cellular density changes such as motility, biofilm production, and metabolic pathways. *P. aeruginosa* has three distinct quorum sensing systems (Bredenbruch *et al.* 2006), but only the *Pseudomonas* quinolone signal (PQS) system was found consistently up-regulated. Other elements were found differentially expressed including LPS & biofilm structural components, motility, bacteriophage gene clusters and DNA repair (LPS or Lipopolysaccharides are major structural components of the outer cell wall in gram-negative bacteria. Biofilms are an aggregate of bacteria that have adhered to a surface. The bacteria in a biofilm are often embedded in an extracellular matrix).

Gene Expression Changes Reflect Pathoadaptation of an Opportunistic Pathogen

Many of the changes in epidemic strains are a reduction in expression and are in metabolic and energy respiration systems. This trend is consistent with the process of

pathoadaptation of opportunistic pathogens. Opportunistic pathogens such as *P. aeruginosa* possess virulence factors that enable it to survive in host environments, but its fitness in this environment is suboptimal (Sokurenko, Hasty, and Dykhuizen 1999). The CF lung presents a unique and variable environment. It has limited oxygen availability with high concentrations of amino acids, nitrogen (in form of nitrate), nucleotides, and selective pressure applied through antibiotic treatment. The genome of versatile *P. aeruginosa* contains a large repertoire of respiration, metabolic, and transport genes. *P. aeruginosa* goes through a process of adaptation by mutation of its resident genes. These change-of-function mutations confer a selective advantage by driving metabolic flux through optimal pathways and saving resources and energy by down-regulation or loss-of-function mutations in unneeded pathways (Sokurenko, Hasty, and Dykhuizen 1999). We believe that the significant level of down-regulated systems found in this analysis is a reflection of this process of adaption in epidemic strains.

Testing Transcriptional Modules for Differential Expression

To explore if gene expression changes occur at a regulon level, we used the same group-based differential expression statistic to test experimentally obtained and *in silico*-predicted regulons for differential expression. A limited number of experimentally validated regulons for *P. aeruginosa* is available from PRODORIC database. We also carried out a targeted search for predicted regulons that were associated with epidemic strain gene expression changes. To find relevant transcriptional modules in epidemic *P. aeruginosa*, we collected available gene expression data and then used the genes from the differentially expressed pathways, operons and individual gene tests as seeds around which to build co-regulated gene clusters. The transcriptional module predictions were further validated by searching for conserved DNA motifs upstream of the genes and then determining if those motifs were also conserved in other *Pseudomonas* species orthologs upstream regions (Appendix G.). Non-coding DNA signatures that are conserved across diverse species are highly likely to be functional and may represent common regulatory sites. Enrichment of DNA motifs in genes in the TMs was computed and compared to the rest of the genome for all species (Figure 2.2). Significant enrichment was found in the PAO1 genome for all predicted modules. DNA motifs of several of the transcriptional modules were conserved in other *Pseudomonas* species orthologs.

In total 24 initial transcriptional modules were found. When tested for differential expression using the *global* test statistic, 6 transcriptional modules tested positive (Table 2.4). For the remaining TMs, the lack of statistical support for differential expression suggests that the gene expression changes in epidemic strains are not occurring at a regulon level.

Among the 6 differentially expressed transcriptional modules, we recovered two that correspond to the ANR and DNR regulons (referred to as modules 1 and 2, respectively). The ANR transcription factor is activated under microaerobic conditions and turns on genes involved in aerotaxis (*aer*), fermentation (*adhA*, *arcD*, *arcA*), cytochrome c oxidase (*ccoN2*), heme biosynthesis (*hemN*) (Rompf *et al.* 1998; Trunk *et al.* 2010). The ANR transcription factor also activates DNR (Trunk *et al.* 2010). DNR senses NO and activates genes involved in the denitrifying pathway (*nar/nir/nor/nos* operons) (Trunk *et al.* 2010; Giardina *et al.* 2011). Module 4 contains multiple motility and cytochrome genes. Module 6 contains a set of functionally diverse genes involved in motility (*filCD*, *pilAGH*, *cheYZ*, *fliA*, *and flgM*), ribosomal protein genes, oxidative phosphorylation (*sdhAB*, *nuoB*, *azu*, *cycB*, *and atpDEFH*). Similar genes are activated in nutrient-induced dispersion from biofilms (Sauer *et al.* 2004). Module 7 is best characterized by quorum sensing genes (*rsaL*, *lasI* and *rhlI*). Finally, module 9 consists of only 8 genes, most of which are hypothetical proteins and probable transporters. Two of the genes encode urease subunits. Urease catalyzes the breakdown of urea into ammonia.

Table 2.4 Differentially Expressed Transcriptional Modules

Predicted transcriptional modules significantly differentially expressed in epidemic versus non-clonal and PAO1 strains, as determined by the group-wise global test (FDR <= 0.05). The contribution of the genes can be extrapolated from the test result. Also shown are the effect causing genes for each module (contribution FDR <= 0.05).

Module	Description	False Discovery Rate	Effect Causing Genes in Module	
			Up-regulated in Epidemic Strains	Down-regulated in Epidemic Strain
1	ANR regulon	5.97E-005	PA1429	PA1557 (ccoN2)
2	DNR regulon	1.80E-004	PA0315	PA1123, PA4863, PA4867 (ureB), PA4864 (ureD),

				PA0295
4	Motility regulon	4.39E-003		PA1123, PA4396, PA0731
6	Diverse gene functions: motility, oxidative phosphorylation, ribosome	4.94E-003		PA1092 (fliC), PA1095, PA1094 (fliD), PA3351 (flgM), PA4430
7	Quorum sensing regulon	2.66E-002	PA1431 (rsaL)	
9	Urease regulon	2.66E-002	PA2018	PA4867 (ureB)

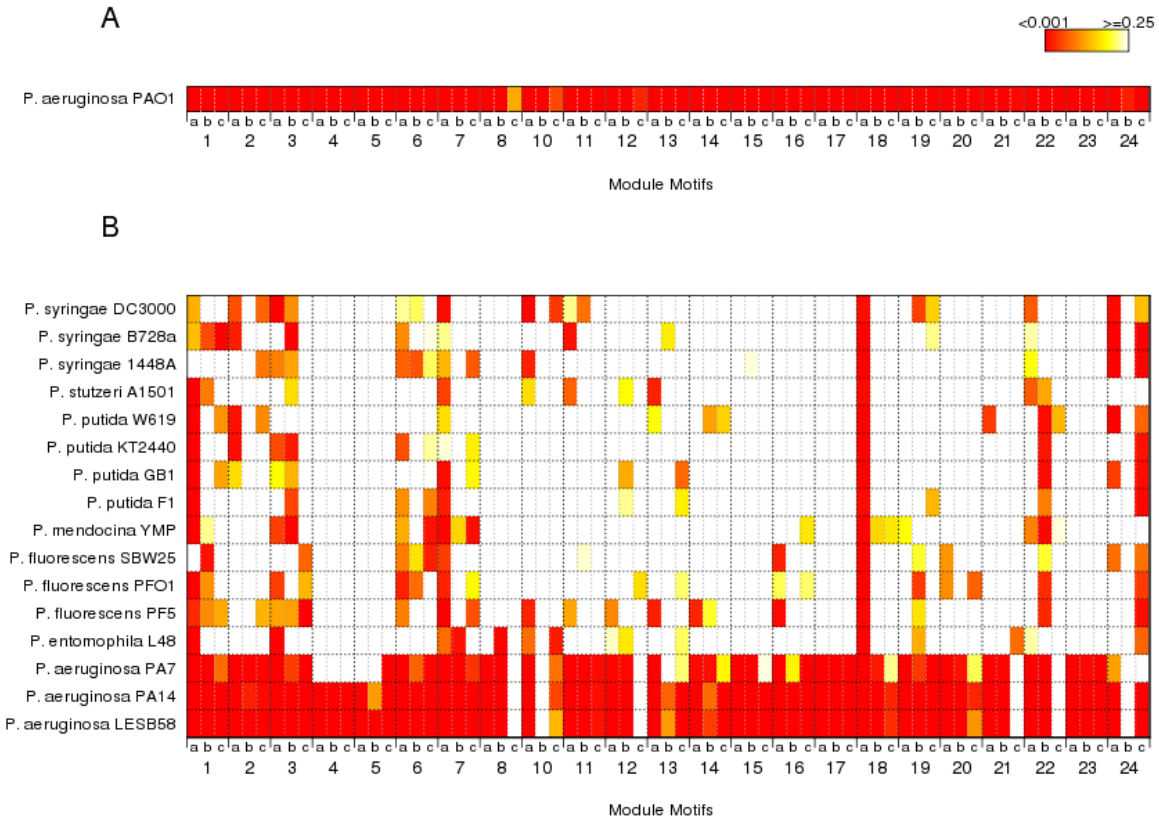


Figure 2.2 Motif Enrichment for Predicted Transcriptional Modules

To validate the transcriptional modules predictions, conserved motifs were searched for in the upstream DNA regions of module genes. The top three motif candidates (labeled a, b & c) were tested for enrichment in *P. aeruginosa* (A) and other *Pseudomonas* species (B). Intensity of the red in the cell indicates the strength of the enrichment based on a hypergeometric distribution *p*-value. The statistic tests the hypothesis that the motif is more often associated with module genes than genes outside module.

Conservation of Gene Regulation in the *Pseudomonas* Genus

Pseudomonas species are ubiquitous, metabolically versatile organisms. Of the seven species used in this study, *P. aeruginosa* is the only human pathogen (other species, such as *P. syringae,* are common plant pathogens). Pseudomonads are genetically diverse. On average 94.20% *P. aeruginosa* PAO1 genes are conserved in other *aeruginosa* strains. The ortholog conservation is significantly lower between distinct species of Pseudomonads (an average 58.78% of *P. aeruginosa* genes have orthologs in other *Pseudomonas* species, Table 2.5). Between *P. aeruginosa* strains, the core genome sequence is highly conserved. Comparing the DNA sequences of orthologs in the *P. aeruginosa* PAO1 and LESB58, the only available epidemic genome sequence, the average percent identity is 98.83% (95% confidence interval of the mean is [98.67, 98.98]). Aligning the upstream non-coding regions of the orthologs, the average percent identity is 72.85% (95% confidence interval is [72.17, 73.54]) for these strains. At the genus level, when comparing non-coding upstream elements of orthologs, there appears to be little conservation of non-coding DNA across different *Pseudomonas* species (reliable automated DNA sequence alignments were not able to be produced).

Table 2.5 Proportion of Orthologous Genes in Pseudomonas Species

The proportion of protein-coding genes in *P. aeruginosa* PAO1 with an ortholog in select other *Pseudomonas* species. Genes forming in-paralogous relationships were not counted in the ortholog total.

Comparison Genome	Percent of *P. aeruginosa* PAO1 Genes that have Orthologs in Comparison Genome
P. aeruginosa LESB58	95.93
P. fluorescens Pf-5	68.87
P. entomophila L48	60.85
P. putida KT2440	59.79
P. mendocina YMP	58.98
P. syringae pv. tomato str. DC3000	54.62
P. stutzeri A1501	49.54

Using *Pseudomonas* genomic data to validate clinically relevant transcriptional modules in *P. aeruginosa* has potential pitfalls. Firstly, how well the regulatory mechanisms are conserved in the *Pseudomonas* genus will determine the transcriptional modules that can be validated using a comparative genomics approach. Secondly, this approach is limited to validating transcriptional modules that are part of the core *Pseudomonas* genus gene set. Pathogenic genetic elements are frequently located on genomic islands and are not conserved across species. In the study, 24 putative transcriptional modules were isolated using the effect-causing genes from the meta-analysis as seeds to search for other co-regulated genes. Using the effect-causing genes as the starting seeds will help retrieve modules that have a role in the organism's pathogenicity (though, only a subset of the 24 modules were differentially expressed on a module-wide level in the epidemic strains). Of the 24 modules obtained, 23 had at minimum 5 orthologs in other *Pseudomonas* species. Additionally, 7 of the modules had orthologs in multiple other *Pseudomonas* species that were over-represented in the same upstream DNA motifs that were found for the *P. aeruginosa* modules genes, suggesting that the regulatory *cis* elements are conserved. And of these modules exhibiting cross-species conservation of the *cis* regulatory elements, three had significant differential expression in the epidemic strains compared to the non-epidemic strains of *P. aeruginosa*. In this study, a comparative genomics approach involving bacterial species from diverse evolutionary niches was used to validate putative transcriptional modules associated with infection in a human pathogen. While many of the putative modules either had no orthologs in the non-pathogen species or had divergent gene regulatory mechanisms, a small of number of modules were validated using this approach. In this case, the lack of suitable comparison species limited the utility of a comparative genomics approach for investigating gene regulation.

2.3.4. Case II Conclusions

In this study, a meta-analysis and system-based approach is used to identify gene expression differences in three epidemic *P. aeruginosa* strains. By focusing on expression changes that are common to multiple epidemic *P. aeruginosa* strains and absent in non-clonal *P. aeruginosa*, we have uncovered adaptations that emerged exclusively and repeatedly in multiple epidemic strains. There is considerable gene

expression variation between isolates and strains. This selective approach should help identify some of the critical adaptations that are increasing the infectivity of epidemic strains.

Through a systems-based approach and a meta-analysis, we generate novel insights into the common gene expression changes among epidemic strains. These insights represent high quality hypotheses as to the regulatory adaptations supporting the epidemic strains increased infectivity. Further work could also build upon these results. A system-based meta-analysis involving a larger panel of epidemic strains, microarray test conditions representing CF lung conditions and isolates from early infections that have a greater number of transmissibility determinants intact, could provide additional insights into the gene expression changes occurring in epidemic *P. aeruginosa*.

Using systems-based tests, several differentially expressed biological functions were identified in epidemic strains. The results highlight the importance of metabolic adaptation. Multiple metabolic pathways are down-regulated, suggesting that in epidemic strains, there is a redirection of resources to pathways that are optimal for the CF lung. Quorum sensing is one of the few systems that were found in this analysis to be up-regulated in epidemic. We investigated the gene regulatory mechanisms of these altered pathways and genes, by finding associated transcriptional modules. The ANR and DNR regulons appear to play a role in the adaptation of epidemic *P. aeruginosa*. The numerous systems found to be differentially expressed suggest that the increased infectivity in epidemic *P. aeruginosa* is multi-factorial. Follow-up studies are needed to confirm the importance of these changes for infection.

2.4. Case III: Comparative Genomics-based Regulatory Analysis in *Aspergillus fumigatus*

2.4.1. Introduction

Aspergillus fumigatus is an opportunistic fungal pathogen and the causative agent of aspergillosis; the most common invasive fungal infection in immuno-compromised patients (Schrettl *et al.* 2004; Hissen *et al.* 2005). Iron acquisition is

essential to the virulence of *A. fumigatus*. Iron is extremely limited in the host environment and mechanisms that sequester and import iron are necessary for the growth of *A. fumigatus* in the host (Schrettl *et al.* 2004). Two iron acquisition systems have been described in *A. fumigatus*: (i) a reductive iron mechanism that reduces ferric to ferrous iron and then uptakes the ferrous iron and (ii) a siderophore-assisted ferric iron uptake pathway (siderophores are iron-specific chelators). Siderophore uptake is essential to the virulence of *A. fumigatus*, while the contribution of reductive iron mechanism to infectivity is negligible (Schrettl *et al.* 2004; Hissen *et al.* 2005; Schrettl *et al.* 2007). These systems are both up-regulated during iron starvation. They are, however, only part of a larger iron-starvation response. Over 1100 genes in *A. fumigatus* have been found to be differentially-regulated in response to iron availability. The transcription factor, *sreA*, was found to regulate a limited number of these genes, including the siderophore biosynthesis pathway (Schrettl *et al.* 2008). The broader iron response in *A. fumigatus* has not been characterized.

Because of the importance of iron sequestration to pathogenicity and the large diverse response to iron starvation, we wanted to investigate the gene expression changes that occur in *A. fumigatus* due to iron availability. Jason Catterson, a member of Dr. Margo Moore's Laboratory at Simon Fraser University, carried out a microarray study of the fungal pathogen *A. fumigatus'* transcriptional response to low iron conditions (Catterson 2008). To characterize the iron response, a systems-level analysis of the microarray data was performed to look for biological systems regulated by iron concentration. There is only limited gene regulatory data for *A. fumigatus*, so a comparative genomics approach was used in place of directly interrogating experimentally obtained regulatory interactions.

2.4.2. Methods

Microarray Data

Microarray log-fold values and significance *p*-values representing the change in gene expression between iron-limited and iron-replete conditions were obtained from the following reference: Catterson, 2008. The microarrays compare the RNA transcript levels at 2, 4 and 6 hours after the addition *A. fumigatus* conidia to iron-limited and iron-

replete media (minimal essential media + 10% human serum with or without 50μM $FeCl_3$ added). Genes with a log-fold value greater than 1.4 up- or down-regulated and a *p*-value less than 0.05 were considered significantly differentially expressed at a particular time point. Applying this threshold filter produced datasets consisting of 180, 789 and 499 genes at the 2, 4 and 6 hour time-points. These datasets were used in the downstream analyses.

Transcriptional Regulatory Network Analysis

Transcriptional regulatory network (TRN) models from the fungal species: *Saccharomyces cerevisiae* and *Candida albicans* were used in the analysis of the transcriptional response of *A. fumigatus*. The *S. cerevisiae* TRN was constructed from the results of a comprehensive ChIP-on-chip analysis by Lee *et al.* involving 106 transcription factors in *S. cerevisiae* (Lee 2002). The network comprises 3658 nodes, representing the regulators and target genes, and 8262 edges, representing interactions between the regulators and the genes they regulate. The *C. albicans* TRN was also constructed from a ChIP-on-chip study (Lavoie *et al.* 2010). The *C. albicans* study was, however, less comprehensive, identifying gene targets in the *C. albicans* genome for six transcription factors. The *C. albicans* TRN contains 1534 nodes and 1998 edges.

Orthologs between *A. fumigatus* and *S. cerevisiae* and *A. fumigatus and C. albicans* were computed using the reciprocal best BLAST procedure. Transcriptional regulatory interactions from the model species were transferred to *A. fumigatus* through the ortholog gene mappings.

Subnetworks that are correlated with the gene expression changes in the three time points were identified in the larger TRNs. Cytoscape and the plugin jActiveModules were used to visualize and search for significant correlated subnetworks, respectively (Ideker *et al.* 2002; Shannon *et al.* 2003).

Transcription Factor Over-Representation Analysis

Predicted transcription factor-gene interactions in *A. fumigatus* were obtained from the *S. cerevisiae* ChIP-on-chip study by transferring the interactions from the *S. cerevisiae* orthologs to the *A. fumigatus* genes. Each of the transcription factors were tested to determine if their target genes were statistically over-represented in the list of

genes differentially expressed under iron-limited conditions. The hypergeometric test with Bonferroni-Hochburg multiple hypothesis correction was used to determine the significance. Transcription factor groups with a *p*-value less than 0.05 were declared significant.

DNA Pattern-based Search for Transcription Factor Binding Sites

The DNA sequences upstream of the *A. fumigatus* genes were extracted up to −1500 bp or to the next ORF. Motifs representing the binding sites of transcriptional regulators of RP genes in *S. cerevisiae* and *C. albicans* were obtained from Hogues et al. 2008. The Regulatory Sequence Analysis Tools (RSAT) suite of tools was used to detect the *cis*-elements in *A. fumigatus* upstream regions of RP genes (Thomas-Chollier *et al.* 2011). Instances of the motifs in the RP genes were compared to the frequency in the rest of the genome and the hypergeometric test was used to test for over-representation. After Bonferroni-Hochburg adjustment, the *p*-value cut-off for declaring a motif enriched amoung the *A. fumigatus* RP genes was 0.05.

De Novo DNA Motif Discovery

Two complementary *de novo* methods were used to discover conserved DNA motifs upstream of the RP genes in *A. fumigatus.* Using the RSAT Oligo-Analysis tool, over-represented oligomers between 5-6 base pairs in length were searched for in the RP genes upstream regions (Thomas-Chollier *et al.* 2011). Over-representation was determined by comparing the frequency of the oligomers in the upstream gene regions of the RP genes to frequency in the remainder of the genome. The hypergeometric distribution was used to test for enrichment. Overlapping oligomers were aligned and merged using the RSAT tool Pattern Assembly and these merged oligomers were reported. The second approach, with the tool MEME, identified statistically significant DNA motifs represented as position-specific scoring matrices in the *A. fumigatus* RP gene *cis* regions (Bailey *et al.* 2006; Bailey *et al.* 2010). For both approaches, 0.05 was used as the threshold for detection. Program defaults were used in the MEME tool.

2.4.3. Results and Discussion

Human cells, tissues and fluids have limited free iron (Weinberg 2009). To persist in the host, most pathogens including *A. fumigatus*, need a continuous source of iron (Catterson 2008; Weinberg 2009). Defining the response of *A. fumigatus* to iron-limited conditions could elucidate the transcriptional mechanisms that support *A. fumigatus* infection. To look for the biological systems regulated by iron concentration, a systems-level analysis of the microarray data was undertaken. Systems resources have not been developed for *A. fumigatus* (such as transcriptional regulatory networks), so a comparative genomics approach was used to build putative networks for *A. fumigatus*. Resources from other fungal species were mapped to *A. fumigatus* by predicting orthologous genes between the species.

Transcriptional Regulatory Subnetworks Correlated with Changes in Iron Availability

A common representation of the transcriptional regulatory interactions between transcription factors (TFs) and genes they regulate is a network graph. Gene regulation is often highly complex. It can involve regulation cascades as well as negative and positive feedback. A network graph encapsulates much of the structure of the regulatory interactions. In this study, prototype *A. fumigatus* gene regulatory networks were constructed from regulatory information produced by two ChIP-on-chip experiments involving an analysis of 106 TF in *S. cerevisiae* and 6 TF in *C. albicans* (Lee 2002; Lavoie *et al.* 2010). Orthology was used to map interactions predicted in the ChIP-on-chip experiments to genes in *A. fumigatus*.

Standard microarray analysis is typically insufficient to identify the transcription factors driving the changes in gene expression. An activated TF can show little to no differential expression and can have different expression patterns than the genes they regulate. Additional information, such as the gene regulatory structure as provided by a transcriptional network, is needed to implicate the specific TFs. To identify components of the regulatory network involved in the iron response, genes in the transcriptional regulatory networks were overlaid with the differential expression data from *A. fumigatus* microarray experiment and subnetworks of co-expressed genes were extracted from the

larger network with the program jActiveModules (part of the Cytoscape suite of add-on tools (Ideker *et al.* 2002; Shannon *et al.* 2003)).

There was only limited clustering of the differentially expressed genes in the gene regulatory network and only a few subnetworks were found that had correlated expression across the time points examined. A similar significant subnetwork was discovered in both the *S. cerevisiae* and *C. albicans*-derived networks (Figure 2.3 and Figure 2.4, respectively). Testing the differentially expressed genes for enrichment in a particular transcription factor, also implicated many of the same TFs as identified in the network analysis (Table 2.8).

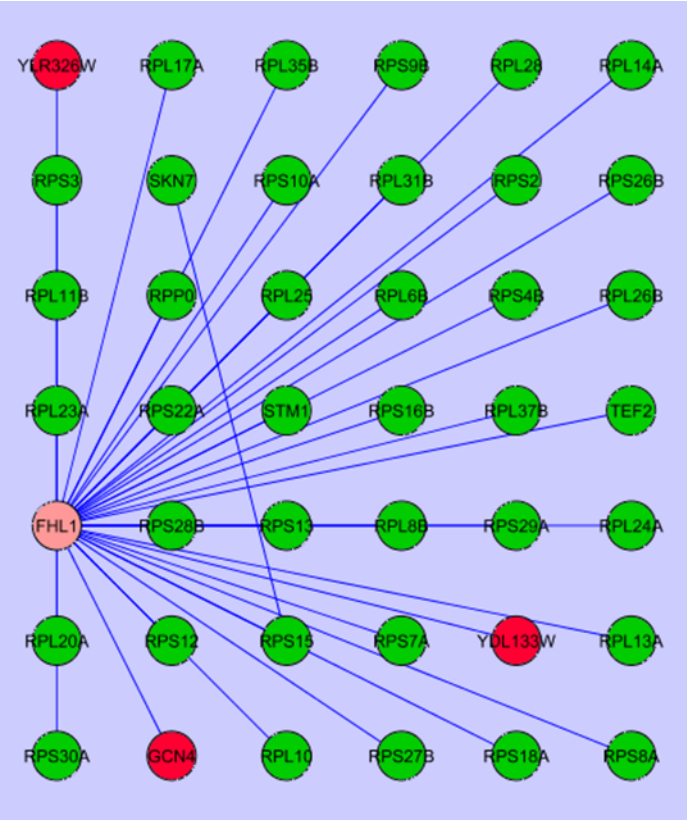


Figure 2.3 A Down-Regulated Subnetwork in the S. cerevisiae-Derived Transcriptional Regulatory Network

A subnetwork in the transcriptional regulatory network constructed from the *S. cerevisiae* gene regulatory interactions showing significant change in expression between iron-limited and iron-replete conditions. Gene expression data for the *A. fumigatus* genes was mapped to the orthologous genes in the *S. cerevisiae* network.

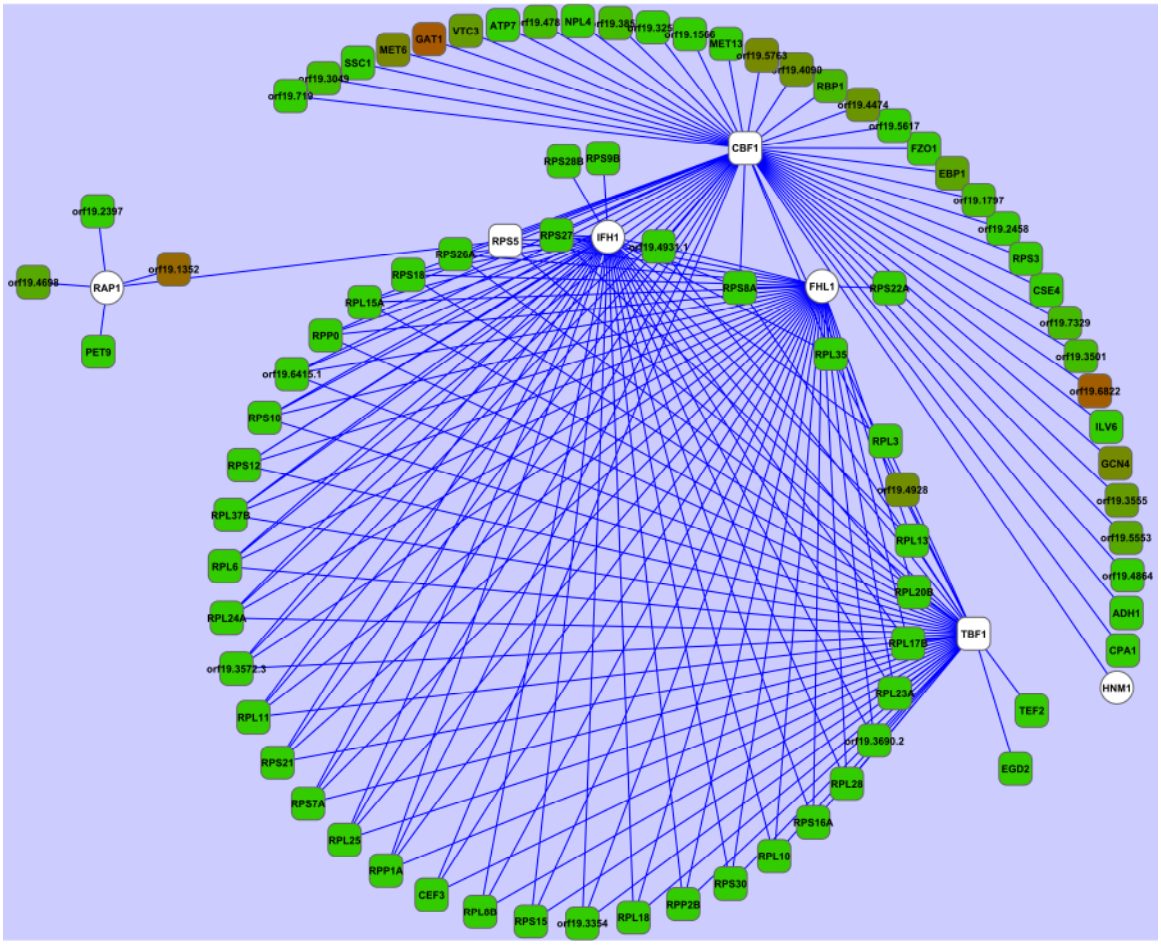


Figure 2.4 A Down-Regulated Subnetwork in the C. albicans-Derived Transcriptional Regulatory Network

A subnetwork in the transcriptional regulatory network constructed from *C. albicans* gene regulatory interactions showing significant change in expression between iron-limited and iron-replete conditions. Gene expression data for the *A. fumigatus* genes was mapped to the orthologous genes in the *C. albicans* network.

Table 2.6 Over-Represented Transcription Factors Associated with Differentially Expressed Genes

Using *S. cerevisiae* transcription factor-gene interactions obtained from the ChIP-on-chip study by Lee *et al.* (2002), *A. fumigatus* gene expression data was mapped to the gene interactions via ortholog mappings and over-represented transcription factors associated with the differentially expressed genes were identified (over-representation tests were completed for each time point in the time course). The threshold for declaring a transcription factor significantly over-represented was 0.05 after Benjamini-Hochburg correction.

S. cerevisiae Transcription Factor	*A. fumigatus* Ortholog	% Genes Regulated by TF that are Differentially Expressed	Adjusted *p*-value	Time Point (hr)
Fhl1	no ortholog	42.1	6.04E-12	4
Rap1	no ortholog	30.4	3.87E-7	4
Ndd1	no ortholog	28.6	0.0438	4
Gat3	no ortholog	31.1	0.0476	4
Hap4	no ortholog	30.4	0.0481	4

Fhl1 is a central regulator in both networks, which in *S. cerevisiae* is involved in the regulation of ribosomal protein (RP) genes. Based on the *A. fumigatus* microarray data, RP genes are coordinately repressed under iron limited conditions. In *S. cerevisiae* RP genes are tightly regulated and are repressed under environmental stress or nutrient limitation (Martin, Soulard, and Hall 2004; Hogues *et al.* 2008; Wapinski *et al.* 2010). One of the mechanisms of RP gene repression in *S. cerevisiae* is carried out by the transcription factor Gcn4 and in the subnetwork, the *A. fumigatus* Gcn4 ortholog CpcA is inversely up-regulated with the RP genes down-regulation. The *S. cerevisiae* transcription factor, Gcn4, is global regulator that coordinates amino-acid and purine biosynthesis. The regulation of Gcn4 itself, involves a complex translational level control mechanism that is activated in the presence of uncharged tRNAs. Gcn4 also represses RP gene expression (Joo *et al.* 2011). An activated Gcn4 binds the RP gene activator Rap1 preventing the histone acetylation at the RP gene promoters required for full expression (Joo *et al.* 2011). This interaction helps redirect resources toward the stress response during amino-acid starvation. The similarities, however, between *S. cerevisiae* and *A. fumigatus* RP gene regulation do not extend beyond the inverse correlation of Gcn4/CpcA and the RP gene expression. There appears to a major reorganization of

the RP gene transcriptional regulation in these two species. Many of the central RP gene regulators in *S. cerevisiae* have no corresponding ortholog in *A. fumigatus* (Table 2.7).

Table 2.7 Orthology of the Ribosomal Protein Transcription Factors in S. cerevisiae and A. fumigatus

S. cerevisiae Transcription Factor	Description [a]	*A. fumigatus* Ortholog
Fhl1	Regulator of RP transcription. Binds DNA directly at highly active RP genes and indirectly through Rap1 motifs at others.	no ortholog
Rap1	Regulates diverse processes. Required for high level of transcriptional activation of RP genes.	no ortholog
Ifh1	Coactivator that regulates transcription of ribosomal protein (RP) genes (with Fhl1).	no ortholog
Crf1	Corepressor involved in repression of RP gene transcription via the TOR signaling pathway (with Fhl1).	no ortholog
Hmo1	Non-essential regulator of RP transcription. Coordinates RP gene expression with Pol I-dependent rRNA expression.	Afu1g04550
Esa1	Subunit of histone acetyltransferase complex involved in activation of RP gene transcription.	Afu2g05530
Rpd3	Histone deacetylase associated with repression of RP gene transcription.	no ortholog

[a] Descriptions obtained from *Saccharomyces* Genome Database.

Substitution of the Transcription Factors Involved RP Gene Regulation

A major reorganization of RP transcription factors occurred in fungi (Hogues *et al.* 2008; Lavoie *et al.* 2010). In the *Saccharomyces* species lineage, a number of RP transcription factors have acquired new roles, regulating different sets of genes. In *C. albicans*, which is part of the same taxonomic family: Saccharomycetaceae as *S. cerevisiae*, Tbf1 is an essential regulator of RP transcription, but in *S. cerevisiae* Tbf1 binding is mainly concentrated to telomere regions. Telomere maintenance in *C. albicans* is performed by Rap1 (Hogues *et al.* 2008; Lavoie *et al.* 2010). The Rap1 ortholog in *S. cerevisiae* binds a wide range of target genes including most RP genes. Cbf1, a transcription factor involved in the regulation of sulfur starvation in both *C.*

albicans and *S. cerevisiae*, also binds many of the RP genes in *C. albicans*. Similarly, Hmo1 binding of RP genes is unique to *S. cerevisiae*. Other RP transcription factors such as Ifh1 and Fhl1 have largely conserved targets in the two species (Lavoie *et al.* 2010).

The transcription factor substitutions that have occurred in *C. albicans* and *S. cerevisiae* also appear to extend to *A. fumigatus* and may be more widespread in *A. fumigatus*. While Rap1, Fhl1, Ifh1, Hmo1 have orthologs in *C. albicans*, *A. fumigatus* lacks orthologs to these *S. cerevisiae* RP regulators. Tbf1 is an essential regulator of RP gene expression in *C. albicans*. Our transcription factor binding site analysis suggests that Tbf1 is also involved in the regulation of RP genes in *A. fumigatus*. Instances of the consensus sequence for Tbf1 were found to be enriched in upstream regions of RP genes (Table 2.8). When a *de novo* search for conserved *cis* elements in the RP genes was conducted, a Tbf1-like binding site motif was also recovered (Table 2.9 and Table 2.10). There are some differences in Tbf1 binding site motif recovered for *A. fumigatus* and sites in *S. cerevisiae* and *C. albicans*. The *A. fumigatus* Tbf1 binding site `yTCGCTTAGCG` appears to consist of a single `TTAGGG`-like motif with additional upstream conserved nucleotides. In *C. albicans* the Tbf1 binding site consists of a highly conserved 18 base pair palindrome and in *S. cerevisiae*, the binding site is made up of multiple `TTAGGG` motifs with no consistent spacing (Lavoie *et al.* 2010). The differences could imply that there has been a change in the binding specificity of the Tbf1 in *A. fumigatus*. The *de novo* search also recovered multiple other conserved motifs for the RP genes. These motifs have no significant similarity to RP gene regulator binding sites in *C. albicans* or *S. cerevisiae*, suggesting that they may be additional transcription factors regulating the RP genes in *A. fumigatus*.

Table 2.8 Enrichment of S. cerevisiae Transcription Factor Binding Site Motifs in A. fumigatus RP Genes

Using the reported DNA consensus patterns for *S. cerevisiae* transcription factor binding sites, instances of the patterns in the upstream regions of *A. fumigatus* genes were counted and then tested for over-representation amoung the *A. fumigatus* RP genes. Only transcription factors involved in RP regulation in *S. cerevisiae* were considered. The adjusted *p*-value threshold for the over-representation analysis was 0.05.

Transcription Factor	Motif Consensus [a]	% of RP Genes with Motif Upstream	Adjusted *p*-value
Rap1	CCCnnACA	37.8	0.0166
Ifh1	CyrGGCnG	26.8	0.422
Tbf1	TAGGGy	76.8	2.31E-18
Cbf1	CACGTG	24.4	0.0153

[a] IUPAC characters in consensus sequence: n = any nucleotide, y=T or C, r=A or G.

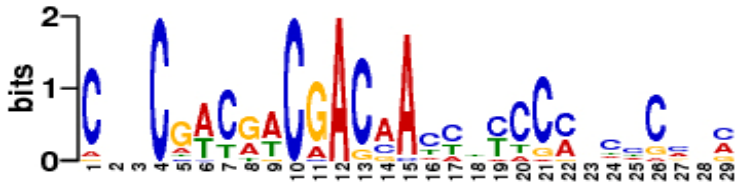
Table 2.9 Over-represented Oligomers Upstream of A. fumigatus RP Genes

Significantly over-represented oligomers in the upstream regions of *A. fumigatus* RP genes were obtained and then overlapping oligomers were assembled into 8 distinct patterns. The significance index is computed as negative logarithm of the E-value. Higher values indicate more exceptional patterns (and values above 0 are considered significant).

Pattern	Oligomer	Reverse Complement	Significance Index
1	TTCGACAA	TTGTCGAA	13.98
2	GCTCGCTAAGCGAAAAT	ATTTTCGCTTAGCGAGC	13.98
3	GACGACAA	TTGTCGTC	13.98
4	AGCCCTAA	TTAGGGCT	12.07
5	GCTCGCTTAGCGAGC	GCTGCGTAAGCGAGC	11.40
6	CTCGCTAGCC	GGCTAGCGAG	8.93
7	GCACACA	TGTGTGC	4.33
8	GAAATTCG	CGAATTTC	1.49

Table 2.10 Significant DNA Motifs Upstream of A. fumigatus RP Genes

The top three conserved DNA motifs as found by the program MEME located upstream of RP genes in *A. fumigatus.* The E-value conveys the expected number of random patterns that would have an equivalent *p*-value.

Motif Sequence Logo	% of RP Genes with Site	E-value
	37.2	2.9E-19
	25.5	4.5E-14
	13.8	0.84

Typically, the majority of the RP genes will be coordinately regulated (Wapinski *et al.* 2010). RP production uses a significant proportion of the cell's resources and RP gene expression is closely balanced with the cellular stoichiometry. Exposure to environmental stresses or limitation of nutrients often produces corresponding changes in RP gene expression. *A. fumigatus* RP genes show a correlated pattern of expression in response to iron limitation. Forty of the 82 RP genes in *A. fumigatus* show significant down-regulation in one or more of the time points. This correlated change in expression, combined with the conserved *cis*-regulatory motifs, suggest that RP genes are also coordinately regulated in *A. fumigatus*. However, the mechanism of the coordinate regulation appears to be different than in *S. cerevisiae* or *C. albicans*. Most of the *S. cerevisiae* or *C. albicans* RP gene regulators have no ortholog in *A. fumigatus*. Novel conserved *cis*-regulatory elements suggest there may be distinct RP regulators in *A. fumigatus*. For sole conserved RP regulator in these species: Tbf1, there appears to be

a change in Tbf1's binding site specificity. Overall, it appears coordination of RP gene expression is achieved by very different mechanisms in these species. These differences in the regulation of a highly conserved biological system demonstrate that transcriptional regulatory networks are extremely flexible.

Ortholog Conservation in Fungal Species

One of the limiting factors in this analysis was the gene conservation between the model fungi organisms, *S. cerevisiae* and *C. albicans* and the pathogen *A. fumigatus*. Approximately 25% of the differentially expressed *A. fumigatus* genes had orthologs in *S. cerevisiae* (Table 2.11). Ortholog linkages were required to map the *A. fumigatus* genes to the *S. cerevisiae* transcriptional regulatory network. *S. cerevisiae* and *A. fumigatus* have distinct lifestyles (i.e. pathogen versus non-pathogen) and are separated by a large phylogenetic distance. At this phylogenetic distance, the suitability of a *S. cerevisiae* network as a model for *A. fumigatus* becomes a concern. The Rap1-Tbf1 transcription factor substitution is an example of the challenges encountered when mapping regulatory interactions between orthologs of diverse species (Hogues *et al.* 2008).

Table 2.11 Differentially Expressed A. fumigatus Genes with Orthologs in S. cerevisiae

Time Point (h)	Number of Differentially Expressed Genes	% with *S. cerevisiae* Orthologs
2	180	26.7 (48)
4	789	28.4 (224)
6	499	25.1 (125)

2.4.4. Case III Conclusions

Through a transcriptional regulatory network analysis, it was identified that the *A. fumigatus* RP genes are coordinately down-regulated in response to iron-limited conditions. The mechanism of coordinated regulation, however, appears to be different than the mechanism in *S. cerevisiae* or *C. albicans* due to transcription factor substitutions and changes in binding site specificity.

A cross-species approach was required to perform a systems level analysis of the gene expression changes in *A. fumigatus*; a species that lacks any large-scale gene regulatory interaction data. Using species with large phylogenetic distances can significantly restrict cross-species transcriptional regulatory network analysis. The lack of conserved orthologs and the evolvability of gene regulation limit the scope of a transcriptional regulatory network analysis.

2.5. Conclusions

Orthologs are a critical component of many types of bioinformatics analyses. They are used to identify DNA or protein sequences under selection, determine differences or similarities in gene content, or to infer gene function. In these projects, orthologs were used to identify conserved, unique genetic adaptations in metazoan species, validate predicted regulatory modules associated with epidemic strains of *P. aeruginosa*, and construct a transcriptional regulatory network in *A. fumigatus* in order to perform a system-level analysis of the gene expression changes in response to iron availability. Overall, the application of computationally predicted orthologs was largely effective in the individual use cases, permitting the discovery of several novel results. The ortholog use cases also highlighted some potential challenges. Gene duplication is frequent in metazoan species, especially in species with complex body plans. When inferring functional equivalence across species, because gene duplication is associated with functional divergence, gene duplication can obscure the functional roles of genes in individual species. The other observation to come out of the metazoan project was that incomplete genomes can have a significant impact on the identification of such metazoan-associated genes. In a phyletic profile analysis, a flexible approach that considers the broader phyletic distribution while permitting some contradictory taxa is essential to obtaining all valid genes that meet the phyletic profile criteria. The transcription factor binding site analysis in the *P. aeruginosa* and *A. fumigatus* projects highlighted the high degree of evolvability of transcriptional regulation. Transcriptional regulation can change relatively quickly between orthologs. For the ribosomal protein genes in *A. fumigatus*, while the coordinated expression of the RP genes has been preserved, the regulatory mechanisms have changed through a transcription factor substitution between *A. fumigatus* and *S. cerevisiae*. Lastly, when performing

comparative genomic analysis, phylogenetic distance is an important factor to consider because the lack of orthologs can significantly limit the scope of possible results and a greater species phylogenetic distance is associated with greater ortholog divergence.

3. Evaluation of the Ortholuge Method for Improving Ortholog Detection

Portions of this chapter have been previously published in the article "OrthologeDB: a bacterial and archaeal orthology resource for improved comparative genomic analysis", co-authored by M.D. Whiteside, G.L. Winsor, M.R. Laird, and F.S.L. Brinkman in Nucleic Acids Research, Volume 41, Issue D1, Pages D366-76 © 2013 Whiteside et al; licensee Oxford University Press.

3.1. Introduction

Computationally predicted orthologs are integral to many comparative genomics analyses. Orthologs, related genes between species that have diverged as a result of speciation, are thought to more likely have similar functions than paralogs, which are homologous genes that have arisen through gene duplication (Koonin 2005). This ortholog functional conservation hypothesis or conjecture is the basis for many comparative genomics methods using computationally predicted orthologs to infer gene functions across species. Table 1.1 lists comparative genomic methods that rely on accurate ortholog prediction.

As described in section 1.2.2, graph-based approaches for ortholog prediction are better suited for high-throughput, large-scale analyses than tree-based approaches. They have also been found to perform well in terms of accuracy (Altenhoff and Dessimoz 2009). However, there is one significant methodological flaw that has been identified in graph-based approaches. Graph-based methodologies select the reciprocally most similar set of gene pairs as orthologs from the genes available in two species. They do not take into account the broader phylogenetic context of the predicted orthologs. Relying on a narrow phylogenetic view can result in misprediction

of a paralogous gene as an ortholog when the true ortholog is missing and a paralog exists in the genome that can form a reciprocal-best-BLAST hit (RBBH) (an ortholog can be missing due to a gene deletion event or due to incomplete genome annotations) (Fulton *et al.* 2006; Kuzniar *et al.* 2008). Standard graph-based methods will give no indication when this type of error is made and can report the mispredicted paralogous gene pair as potential orthologs (Fulton *et al.* 2006; Kuzniar *et al.* 2008).

In addition to this methodological flaw of graph-based ortholog prediction, a more general performance question related to ortholog prediction is the reliability of the ortholog function conjecture. The basis for the use of orthologs in functional comparative genomics analysis is the ortholog function conjecture, which states that genes that have diverged by speciation (i.e. orthologs) are more similar functionally than those that diverged by duplication, or paralogs. The validity of this conjecture has recently been examined on a larger scale (Peterson *et al.* 2009; Forslund, Pekkari, and Sonnhammer 2011; Nehrt *et al.* 2011; Altenhoff *et al.* 2012; Dessimoz *et al.* 2012; Thomas *et al.* 2012). These studies showed that for similar levels of divergence, orthologs tend to more often have functions that are conserved than paralogs. The difference in function conservation however, is not considerable and can vary between species and gene families (Peterson *et al.* 2009; Forslund, Pekkari, and Sonnhammer 2011; Altenhoff *et al.* 2012; Dessimoz *et al.* 2012). While orthologs compared to paralogs provide predictive power when it comes to inferring genes functions across species, the implication from these studies is that ortholog prediction methods could be improved by targeting the set of orthologs that are functionally similar rather than all evolutionary orthologs.

3.1.1. Ortholuge: Underlying Principles

Ortholuge is a high-throughput method that improves the specificity of ortholog prediction (Fulton *et al.* 2006). It provides the benefits of graph-based methods including scalability, but limits false positives generated by missing orthologs, because it considers the phylogenetic context of predicted orthologs. Ortholuge first predicts orthologs using the graph-based approach reciprocal best BLAST (RBB) (Altschul *et al.* 1997), but adds a second step where phylogenetic trees are built for each predicted orthologous gene/protein pair, rooted with a suitable outgroup. This phylogenetic analysis is

completed for all predicted orthologs and, coupled with a statistical analysis (Min *et al.* 2011), is used to flag orthologs that have diverged unusually compared to the expected level of divergence for the species. Predicted orthologs with phylogenetic distance that is comparable to the species divergence are termed supporting-species-divergence orthologs or SSD orthologs while predicted orthologs with unusual divergence are termed Non-SSD.

It is our hypothesis that, firstly, many of the unusually diverging predicted orthologs are paralogs mispredicted as orthologs, or orthologs that have diverged more rapidly in one of the species and, secondly, the remaining orthologs are more likely to have retained similar functions and may be better suited for many comparative genomic analyses.

3.1.2. Description of the Ortholuge Method

Ortholuge was initially conceived and developed by the Brinkman laboratory at Simon Fraser University. Ortholuge generates precise ortholog predictions between two species on a genome-wide scale using an additional outgroup genome as a reference point (Fulton *et al.* 2006). Starting with a set of putative orthologs generated by the RBB approach, Ortholuge then computes phylogenetic distance ratios for each pair of RBB-predicted orthologs that reflect the relative rate of divergence (Figure 3.1). Two ratios are needed to summarize the relative branch lengths for both ingroup genes in the phylogenetic tree representing ortholog divergence. These phylogenetic ratios allow you to distinguish between predicted orthologs with phylogenetic distance that is comparable to the species divergence (termed supporting-species-divergence orthologs, or SSD orthologs) and predicted orthologs with unusual divergence (Non-SSD). SSD orthologs and orthologs undergoing unusual divergence have distinct ratio distributions. These ortholog types are observable when the ratios are plotted on a genome-wide scale (Min *et al.* 2011). SSD and Non-SSD ortholog assignments are determined by a statistical procedure that uses large-scale hypothesis testing approaches to infer the ratio distribution of the SSD orthologs and then assign a local false discovery rate (fdr) to each predicted ortholog pair based on this inferred distribution (Min *et al.* 2011). The local fdr conveys the likelihood that a predicted ortholog pair is Non-SSD given its ratio

value (Min *et al.* 2011). Further information, including specific implementation details and modifications made to the original version, are provided in the following chapter.

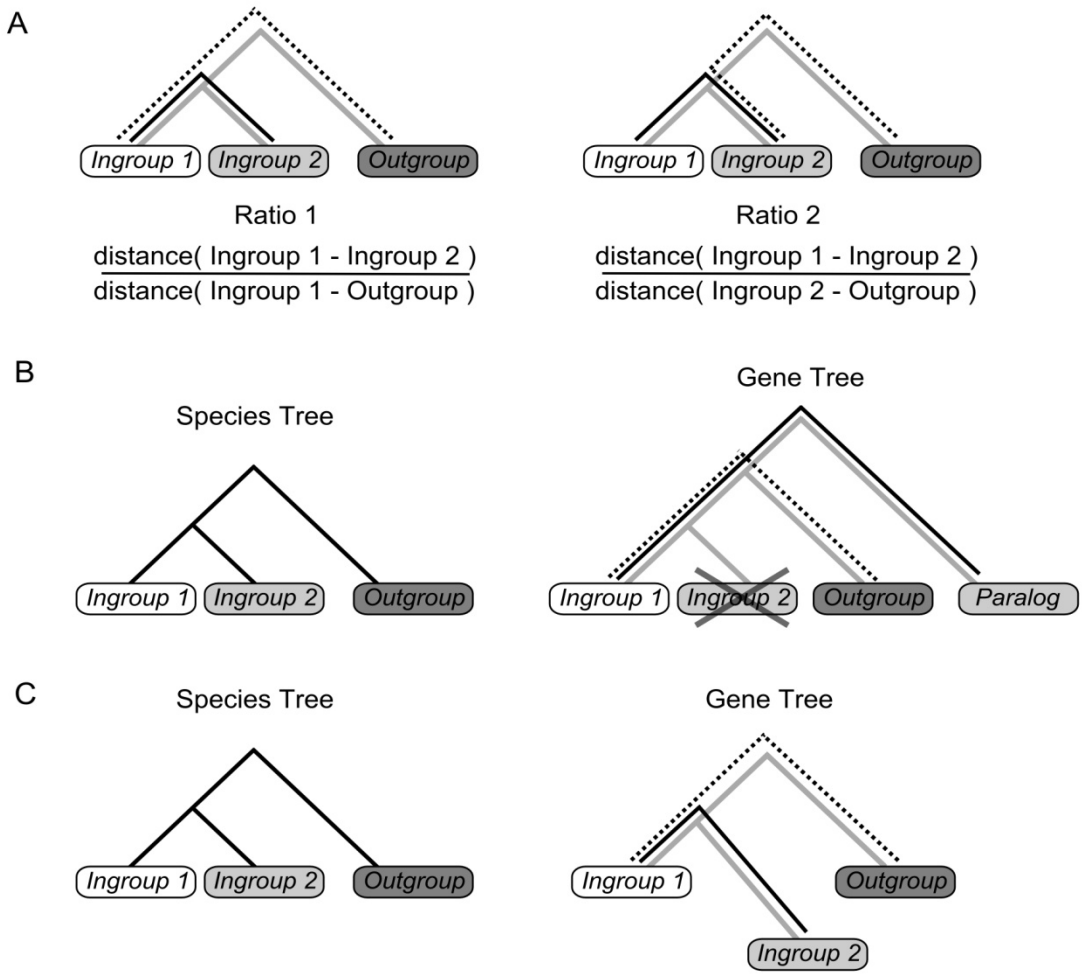


Figure 3.1 Overview of the Ortholuge Phylogenetic Ratios

(A) Phylogenetic ratios computed by Ortholuge. The phylogenetic distances for the numerator (solid dark line) and the denominator (dashed dark line) are overlaid on top of the phylogenetic gene tree (light line) for two genes in the species of interest (Ingroup1 and Ingroup2) and a third reference species (Outgroup). Two ratios are needed to capture the proportional branch lengths of the Ingroup 1 and Ingroup 2 orthologs. (B) The RBB procedure can mistakenly identify a paralog as an ortholog when the true ortholog is missing and when a paralog forms a reciprocal best BLAST hit with the remaining ortholog in the other species. Ortholuge can detect this case because the relative phylogenetic distance between the ingroup genes will increase causing the ratio 1 value to become inflated (the numerator – solid dark line – and the denominator – dashed dark line – for ratio 1 are shown in the gene tree). (C) An ortholog's protein sequence can diverge more rapidly in one species versus another. This rapid sequence divergence can be associated with a change in function. Ortholuge detects cases where an ingroup gene's relative phylogenetic divergence does not match the expected divergence for the species.

3.1.3. Evaluation of Ortholuge-based Ortholog Predictions

In this chapter, the gene function conservation and the status in terms of orthology verses paralogy will be examined for the different classes of predicted orthologs identifiable by Ortholuge. The goal of this study is to compare the SSD orthologs (supporting-species-divergence orthologs) with the Non-SSD (or unusually-diverging orthologs) and establish whether SSD orthologs may be better suited for comparative genomic analyses where ortholog functional equivalence is the basis for the use of orthologs in the analysis. On a larger scale, this study also seeks to improve our understanding of function conservation among orthologs.

The evaluation of Ortholuge-based ortholog predictions consists of two analyses: a first analysis that examines the properties of the Ortholuge classes, and a second analysis that compares the performance of Ortholuge to other ortholog prediction methods.

3.2. Methods

3.2.1. Phylogenetic Tree Construction

The *Pseudomonas* species cladogram was constructed from a super-alignment of the protein sequences of the *carB*, *gyrB* and *rpoS* genes from the individual species (Winsor *et al.* 2009). The program Muscle was used to perform the alignment (Edgar 2004a; Edgar 2004b). The PHYLIP programs protdist, neighbor, seqboot and consense were used to build a boot-strapped neighbour-joining tree (100 boot-strap iterations were performed) (Felsenstein 2008). The same procedure was used to construct the *pil* gene phylogram with the *pilO and pilP* protein sequences (Winsor *et al.* 2009).

3.2.2. Estimation of Potential for False RBB-predicted Orthologs

To estimate the number of possible false RBB-predicted orthologs that could occur when the true ortholog is missing, a simulation was run using 82 pairs of species from the genera *Burkholderia* and *Pseudomona*s (Winsor *et al.* 2008; Winsor *et al.* 2009). In the simulation, an initial set of putative orthologs were predicted using RBB.

The ortholog genes in one species were removed and a subsequent round of RBB ortholog prediction was run on this missing ortholog dataset to determine the number of RBB relationships that can form when a potential ortholog is removed. This procedure was repeated for the other species and the total number of RBB relationships appearing after the removal of the orthologs was reported.

3.2.3. Ortholog Datasets

In total, 660 pair-wise bacterial and archaeal species ortholog datasets were examined. The datasets were obtained from OrtholugeDB. Equal numbers of pair-wise datasets were randomly selected from the set of available datasets available at each taxonomic level (for Genus, Family, Order, Class, Phylum and outside of Phylum), thus ensuring that the species under investigation are equally distributed across all evolutionary distances (Benson *et al.* 2009; Sayers *et al.* 2009). These datasets were used in all of the following analyses.

3.2.4. Conservation of Functional Parameters

To assess the overall functional similarity of different classes of orthologs and also to compare the prediction accuracy of different ortholog prediction methods, the conservation of the gene features: KEGG Orthology (KO) annotations (Kanehisa and Goto 2000; Kanehisa *et al.* 2008), Pfam domains (Sammut, Finn, and Bateman 2008) and Tigrfam annotations (Haft 2003), was measured by counting orthologous gene pairs with identical parameters. KO, Pfam and Tigrfam annotations were obtained from the Integrated Microbial Genomes (IMG) database (Markowitz *et al.* 2012). Although not a direct functional parameter, the protein's predicted subcellular localization (SCL) were also examined to determine if the predicted orthologs had identical localizations. Homologous protein's SCL is highly conserved across species and a change in a protein's localization should indicate a significant divergence in function. Protein SCLs were obtained from PSORTdb (N.Y. Yu *et al.* 2011).

Large gene families can be a confounding factor in ortholog prediction. To test whether the genes in the Ortholuge classes are associated with larger gene families,

genes with one of more homologs were counted. Intra-species homologs were identified using BLAST with a stringent e-value cut-off of 0.01 (Altschul *et al.* 1997).

Gene order or synteny is highly fluid in bacteria. Conserved gene order can be indicative that two genes have a common origin and are in fact orthologs. Also, in prokaryotes functionally-related proteins are often clustered together on the chromosome (Kuzniar *et al.* 2008). Synteny was calculated for the predicted orthologs in two ways. The tool OrthoPred, which was developed in the Brinkman laboratory, was used to compute the segments of the genome where the order of the orthologs is conserved between the species. The number of genes in these segments was counted and reported. Because calculating completely resolved synteny blocks is time-consuming, this analysis was performed on a smaller dataset consisting of 63 *Pseudomonas* species ortholog comparisons and not on the complete set of 660 bacteria and archaeal ortholog comparisons. For the larger set of 660 ortholog comparisons, a simplified operational definition of conserved gene order was used. An orthologous pair of genes was declared as being in a conserved gene region (i.e. displaying synteny), if at least one pair of adjacent genes were also orthologous (Hulsen *et al.* 2006; Altenhoff and Dessimoz 2009). This definition identifies any orthologs in conserved synteny blocks of size 2 or more (but does not distinguish between the sizes of the block).

A *chi*-squared test was used to test for statistically significant differences in the conservation of gene features amoung the Ortholuge SSD and divergent Non-SSD classes (these classes are at the opposite ends of the spectrum in terms of divergence). If the chi-squared test's *p*-value was less than 0.05, the null hypothesis was rejected and difference in feature conservation was considered statistically significant.

3.2.5. Predicting Paralogs

For comparison, the functional parameter conservation of predicted in-paralogs and out-paralogs was also measured. In-paralogs were predicted using the Inparanoid methodology described in section 4.3 (O'Brien, Remm, and Sonnhammer 2005; Ostlund *et al.* 2010). Intra-species paralogs that did not meet the criteria for an in-paralog were

labelled as out-paralogs. Potential paralogs were identified using BLAST with a stringent e-value cut-off of 0.01.

3.2.6. Over-Representation of Non-SSD and SSD Ortholog Classes in Functional Categories

Predicted orthologs were mapped to the BRITE Functional Categories through the KEGG Orthology annotations (BRITE functional categories were selected over COG functional categories because the BRITE categories have a more consistent scope than COG categories which vary in their generality) (Kanehisa and Goto 2000; Kanehisa *et al.* 2008). Each functional category was tested for over-representation in divergent Non-SSD and SSD ortholog genes using Fisher's exact test. The test *p*-values were adjusted for multiple hypotheses testing using the Benjamini-Hochburg procedure and the null hypothesis that there is no statistically significant difference in Non-SSD and SSD orthologs was rejected if the adjusted *p*-value was below 0.05.

3.2.7. Comparing the Performance of OMA, QuartetS and Ortholuge

The functional similarity of the predicted orthologs validated by OMA, QuartetS and Ortholuge was examined using the five criteria described previously: KO annotations, Tigrfam annotations, Pfam domains, SCL and synteny. OMA ortholog predictions for the 660 pairs of bacterial and archaeal species used in this study were obtained from OMA website (Schneider, Dessimoz, and Gonnet 2007; Altenhoff *et al.* 2011). Ortholog predictions for QuartetS were obtained from the QuartetS-DB website (Yu *et al.* 2012). The default cut-off of 20 was used for QuartetS; however cut-off values of 5 and 10 were also tested and produced no significant differences in the results.

The functional parameters used to assess the functional similarity of validated orthologs are indirect and may misreport the true functional similarity of an ortholog. To improve the assessment of the ortholog prediction programs, the functional parameters were combined to improve their accuracy in reporting gene functional equivalence. If at minimum three of the five criteria were conserved, the ortholog was declared as being functionally equivalent (or a true positive). Using this definition of true positives, the

following performance values were computed for QuartetS and Ortholuge: precision, defined as $TP/(TP+FP)$; recall, defined as $TP/(TP+FN)$; accuracy, defined as $(TP+TN)/(TP+TN+FP+FN)$; and Matthew's Coefficient Constant (MCC), defined as (Vihinen 2012):

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{TP+FP} \times \sqrt{TP+FN} \times \sqrt{TN+FP} \times \sqrt{TN+FN}}$$

The OMA website does not provide the rejected orthologs (or predicted false positives), and so was omitted from this analysis of performance values. Finally, an analysis was conducted of just the orthologs that are common between the ortholog prediction programs Ortholuge and QuartetS. The same inferred true positive definition was used and compared to the classification of the ortholog as predicted by the program. Numbers of validated ortholog assessments that agreed with the functional similarity data was tabulated. In order to determine any differences in performance in particular evolutionary ranges, the taxonomic range of the species was also recorded.

3.3. Results and Discussion

3.3.1. False Positives Produced by the RBB Method

Paralogs can be mispredicted as orthologs using graph-based approaches such as RBB. Figure 3.3 is a phylogenetic gene tree for the type IV pilus biogenesis genes *pilO* and *pilP* in multiple *Pseudomonas* species. The distribution of these genes in diverse *Pseudomonas* species suggests that the last common ancestor of the *Pseudomonas* genus contained *pilO* and *pilP* genes (for comparison a species tree for the *Pseudomonas* genus is provided in Figure 3.2). The *pilO* gene appears to have been lost in the *P. putida*/*P. entomophila* lineage and also in the *P. fluorescens* strain SBW25 (*pilO* and *pilP* genes are found in all other sequenced *P. fluorescens* strains). The RBB relationships formed by the remaining *pilP* gene in *P. fluorescens* SBW25 are listed in Table 3.1. With the loss of the *pilO* gene in *P. fluorescens* SBW25, the remaining paralogous *pilP* forms RBB relationships with the *pilO* gene in several *Pseudomonas* species. This case demonstrates one of the flaws in the graph-based approaches like RBB. When the true ortholog is missing, a paralog can form an RBB

relationship resulting in an invalid ortholog prediction. Without considering the broader phylogenetic context, it would not be apparent that the *pilO* gene is lost in *P. fluorescens* SBW25 and that the RBB relationships are falsely reporting orthologs.

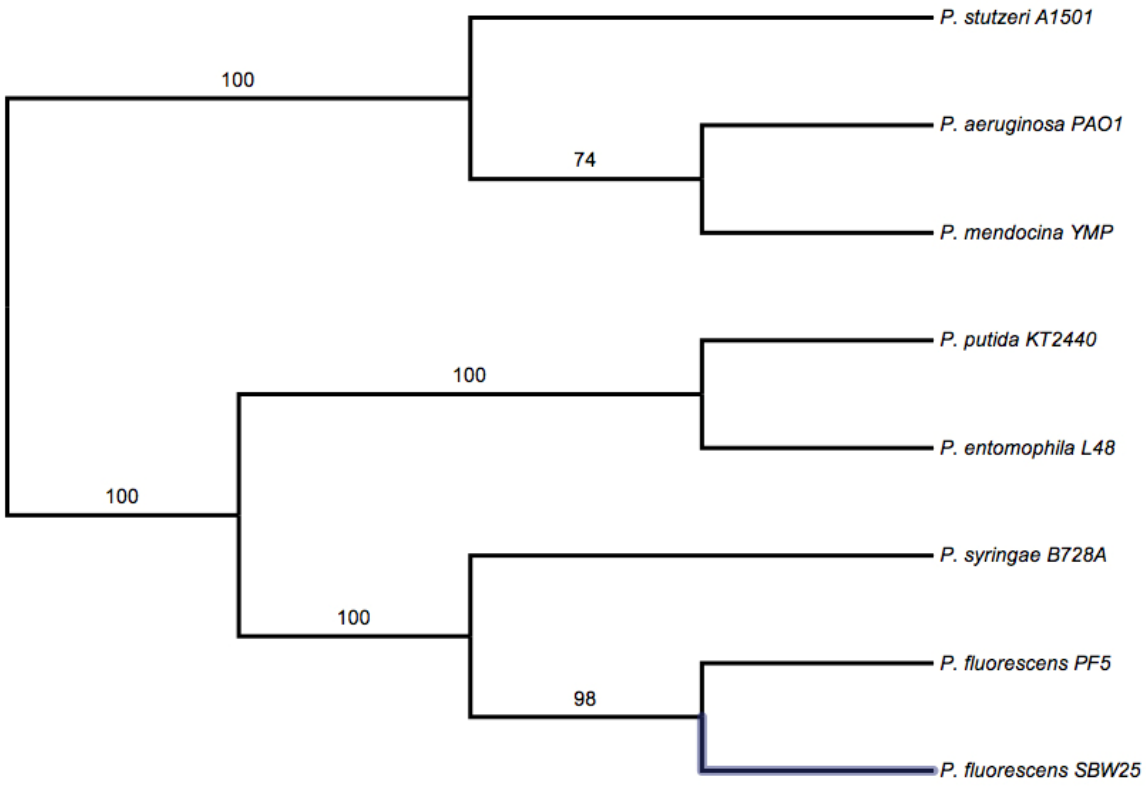


Figure 3.2 Pseudomonas Species Tree

A cladogram showing the species relationships in the *Pseudomonas* genus. Numbers along the branches indicate the bootstrap values that a particular partition in the cladogram was observed (out of a possible 100 bootstrap samples). The species *P. fluorescens* SBW25 is highlighted in the tree.

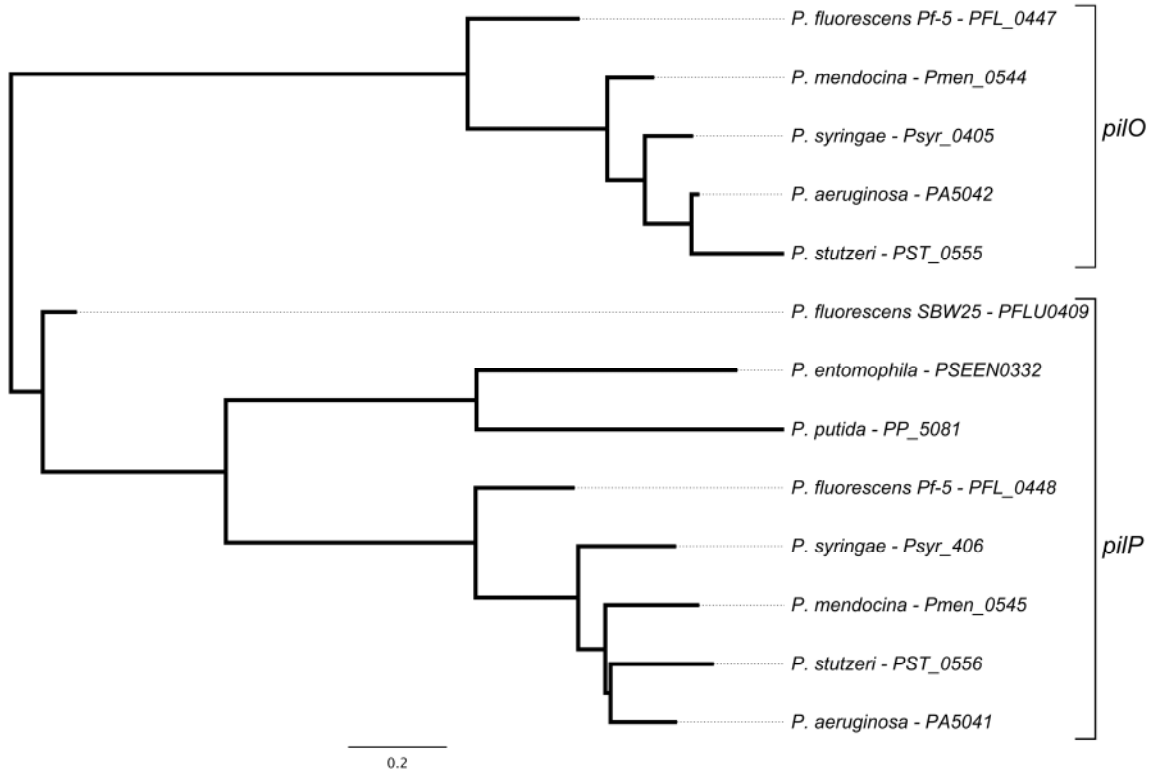


Figure 3.3 The Phylogenetic Tree for the Pseudomonas PilO and PilP Genes

A neighbour-joining phylogenetic tree for multiple *Pseudomonas* species PilO and PilP type IV pilin biogenesis proteins.

Table 3.1 Reciprocal Best BLAST relationships for the Pseudomonas PilO and PilP Genes

Protein Family	Species	Gene		RBB with *P. fluorescens* SBW25 Gene
		NCBI Protein GI	Locus Tag	
PilO	*P. aeruginosa* PAO1	15600235	PA5042	✓
	P. mendocina YMP	146305581	Pmen_0544	✓
	P. stutzeri A1501	146280950	PST_0555	✓
	P. fluorescens Pf-5	70733951	PFL_0447	
	P. syringae B728a	66043672	Psyr_0405	✓
PilP	*P. aeruginosa* PAO1	15600234	PA5041	
	P. mendocina YMP	146305582	Pmen_0545	

Protein Family	Species	Gene		RBB with *P. fluorescens* SBW25 Gene
		NCBI Protein GI	Locus Tag	
	P. stutzeri A1501	146280951	PST_0556	
	P. entomophila L48	104779611	PSEEN0332	✓
	P. fluorescens Pf-5	70733952	PFL_0448	✓
	P. putida KT2440	26991757	PP_5081	✓
	P. syringae B728a	66043673	Psyr_0406	

To estimate the potential for false ortholog prediction when using RBB, a simulation was performed where all orthologs were removed from a dataset and RBB was run on the remaining genes. Eighty-two bacterial genome pairs were examined and for these species, an average of 19.00% of the original ortholog predictions could form an RBB relationship with a paralogous gene when one of the orthologs was removed (the 95% confidence interval of the mean is [17.83, 20.16]). This simulation provides an estimation of the potential for this type of error by showing that on average for bacterial genome analyses, 1 in 5 gene deletions can result in a false ortholog prediction if standard graph-based ortholog prediction methods such as RBB are used (genes can be missing due to a gene loss event or incomplete genome annotation).

3.3.2. Detection of False Positives by Ortholuge

Ortholuge was designed to mitigate the errors produced by missing genes in graph-based ortholog prediction (Fulton *et al.* 2006). Ortholuge evaluates predicted orthologs by examining the relative branch lengths in the predicted ortholog's phylogenetic tree in relation to other predicted orthologs in the genome. Ratios of the distances in the phylogenetic tree are used to capture the degree of divergence of the predicted orthologs and when plotted on a genome-wide level, the ratios of predicted orthologs undergoing unusual divergence appear as high values in the genome-wide distribution. Paralogs mispredicted as orthologs and orthologs undergoing unusual divergence will both exhibit high ratio values (Fulton *et al.* 2006). Figure 3.4 is a histogram of ratio 1 values for predicted orthologs generated by the RBB method for two species (ratio 1 compares the predicted orthologs phylogenetic distance to the distance

between species 1 and the outgroup ortholog. Refer to Figure 3.1 for the specification of ratio 1). The predicted orthologs can include paralogs falsely predicted as orthologs. A dataset of paralogs (or true negatives) ratio values is overlaid in the histogram. The true negatives or known paralogous genes occupy higher ratio range than the majority of the predicted orthologs. One of the assumptions in the Ortholuge method is that the majority of orthologs predicted by RBB are valid orthologs and these ratio values form the major distribution in the genome-wide plot. The ratio formulation normalizes the predicted ortholog phylogenetic distances by accounting for gene-level divergence. This normalization adjusts the orthologs phylogenetic distances so that they are comparable to the species-level of divergence. The term supporting-species-divergence orthologs or SSD orthologs is used to refer to predicted orthologs with phylogenetic distance ratios that is equivalent to the species level of divergence (Fulton *et al.* 2006). We also infer that RBB is incorrectly predicting a small proportion of paralogous genes as orthologs, and also that a small number of orthologs are undergoing unusual divergence (such as accelerated evolution). It is our hypothesis that predicted orthologs occupying the same ratio range as the true negatives in the histogram are likely paralogs predicted as orthologs or orthologs undergoing unusual divergence. Predicted orthologs with phylogenetic divergence that is not equivalent to the species-level of divergence are termed Non-SSD (Fulton *et al.* 2006).

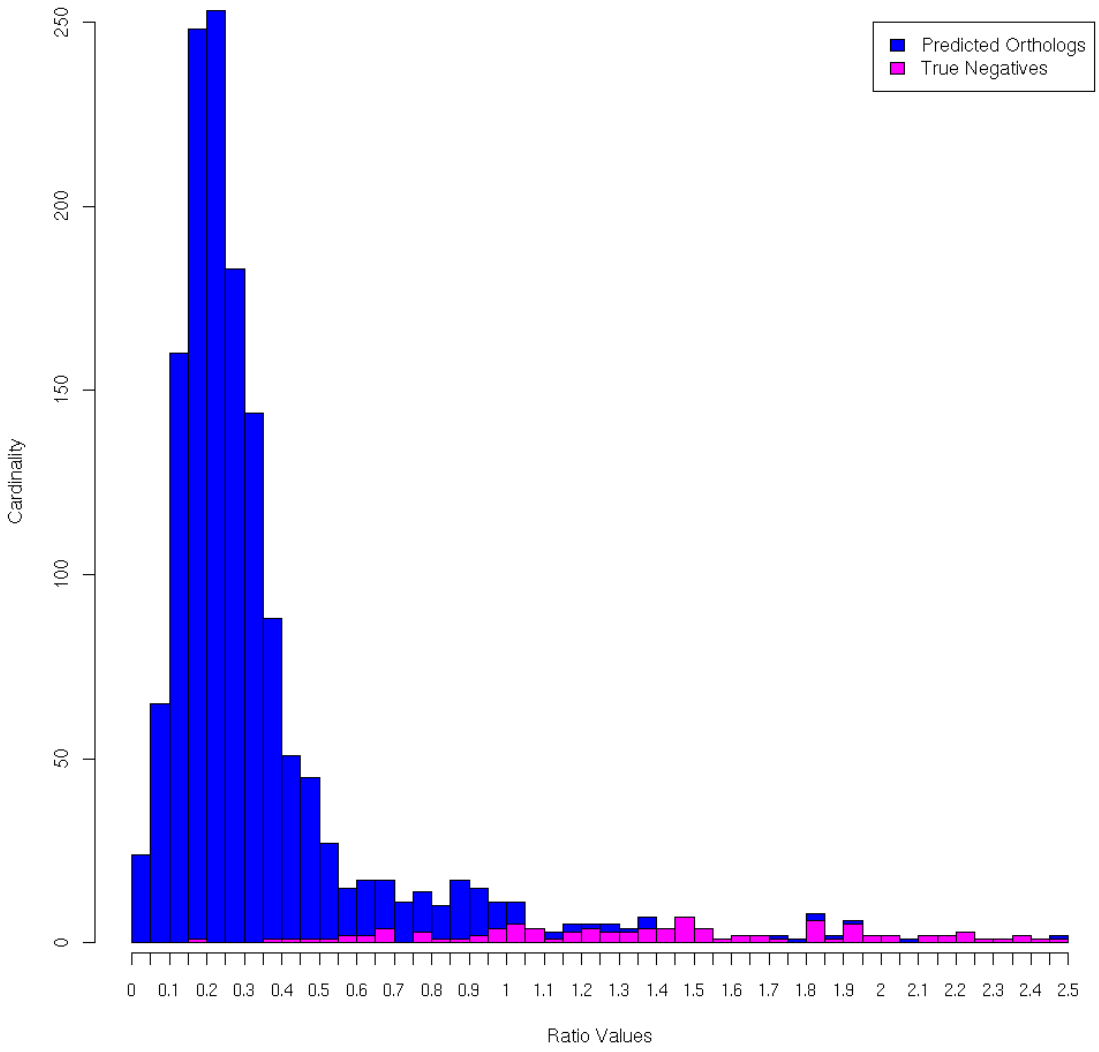


Figure 3.4 Histogram of Ortholuge Ratio 1 Values

A histogram showing the frequency of ratio 1 values computed for the predicted orthologs in species *P. putida* GB1 and *P. syringae* pv. tomato str. DC3000. The ratio 1 values for true negatives are stacked in the histogram (true negatives are ortholog pairs where one of the predicted orthologs is replaced with a paralog).

To consistently classify SSD and Non-SSD orthologs, a statistical procedure is used that employs large-scale hypothesis testing approaches to infer the ratio distribution of the SSD orthologs and then assign a local false discovery rate (fdr) to each predicted ortholog pair based on this inferred distribution (Min *et al.* 2011). The local fdr conveys the likelihood that a predicted ortholog pair is Non-SSD given its ratio value (Min *et al.* 2011). Further details on this statistical procedure are available in

section 4.4. Using the local fdr values of 0.3 and 0.9 as the boundary cut-off values, predicted orthologs are classified as follows:

1. Supporting species divergence (SSD):

   Predicted orthologs whose divergence (as reported by the Ortholuge phylogenetic ratios) is consistent with the divergence observed for the species. These predicted orthologs most likely represent valid orthologs and have not undergone unusual divergence.

2. Borderline-SSD:

   Predicted orthologs with a phylogenetic ratio value that is slightly higher than expected. When ortholog precision is critical to an application, these predicted orthologs can be excluded.

3. Divergent Non-SSD:

   Non-SSD genes have phylogenetic ratios that are significantly higher when compared to other orthologs in the genomes (as per the statistical analysis; (Min *et al.* 2011)), indicating that their divergence is not consistent with the species level of divergence. Based on previous simulations (Fulton *et al.* 2006), these genes are most likely incorrectly predicted orthologs, or orthologs that have undergone unusually rapid divergence due to a change in function.

4. Similar Non-SSD:

   Similar Non-SSDs have diverged unusually, as the length of one of the branches in the gene tree is proportionally longer than expected, however the total phylogenetic distance separating the predicted orthologs is relatively small. Many Similar Non-SSD genes will often be valid orthologs. However, the high phylogenetic ratio may suggest the genes are evolving at different rates.

The justification for the sub-classification of Non-SSD into Similar and Divergent Non-SSD types is explained in section 4.5. Figure 3.5 shows the proportion of each of these Ortholuge classifications for a sample of bacterial and archaeal species comparisons.

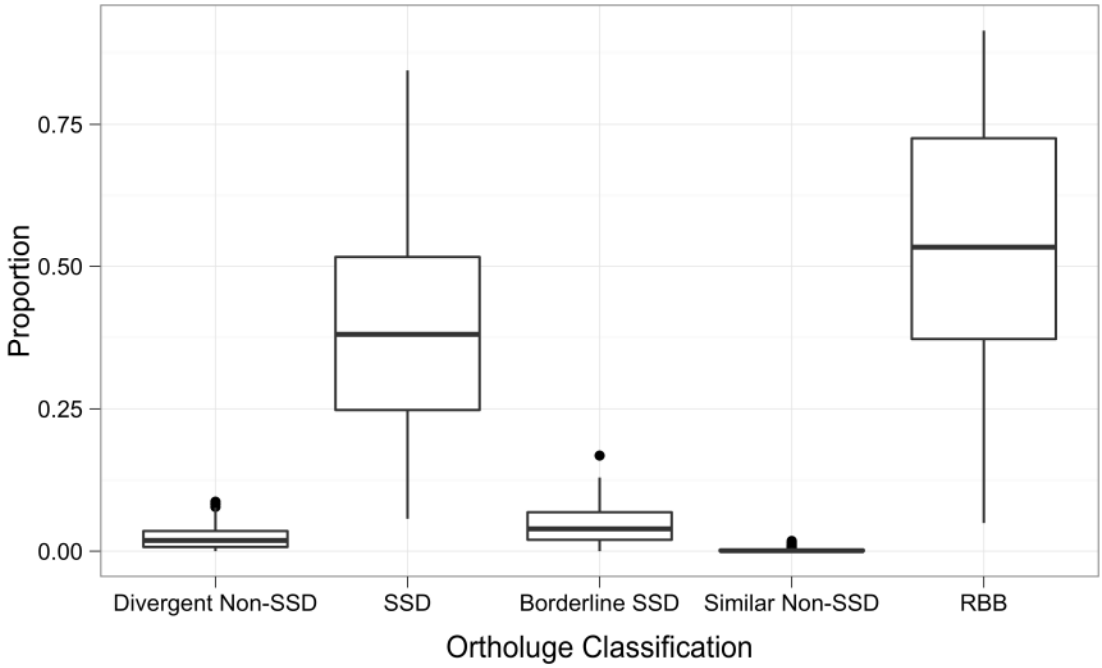


Figure 3.5 Ortholuge Classification Proportions

The proportion of the RBB-predicted orthologs that were assigned to each of the Ortholuge classifications: SSD, Borderline-SSD, Similar Non-SSD and Divergent Non-SSD computed for 660 pairs of bacterial and archaeal species. RBB comprises predicted orthologs for which there was no suitable outgroup ortholog and no Ortholuge evaluation could be performed.

3.3.3. Coverage in Ortholuge

Coverage or the proportion of predicted orthologs evaluated is a critical aspect of the performance of Ortholuge. On average for bacterial analyses, approximately half of all RBB orthologs predictions are evaluated. However, the box plot in Figure 3.5 shows there is significant variability in proportion of evaluated orthologs between the analyses. The variability in the proportion of SSD orthologs appears to correlate with variability in the proportion of RBB-predicted orthologs for which there is no Ortholuge evaluation (RBB column in Figure 3.5. Ortholuge evaluations cannot be performed in cases where there is no suitable outgroup ortholog). Figure 3.6 shows that the proportion of SSD orthologs decreases as the phylogenetic distance separating the species increases. The decrease in SSD ortholog proportion correlates with an increase in unclassified RBB ortholog predictions. The correlation between the classified proportion and species distance suggests that the requirement in Ortholuge for orthologs in three species: the

two species of interest and a third outgroup ortholog to root the phylogenetic tree, is a significant limiting factor affecting the number of predicted orthologs that can be evaluated. Divergent species have fewer common orthologs, so as the phylogenetic distance between the species increases, the number of suitable outgroup orthologs decreases, causing the number of Ortholuge evaluations in divergent species to decline.

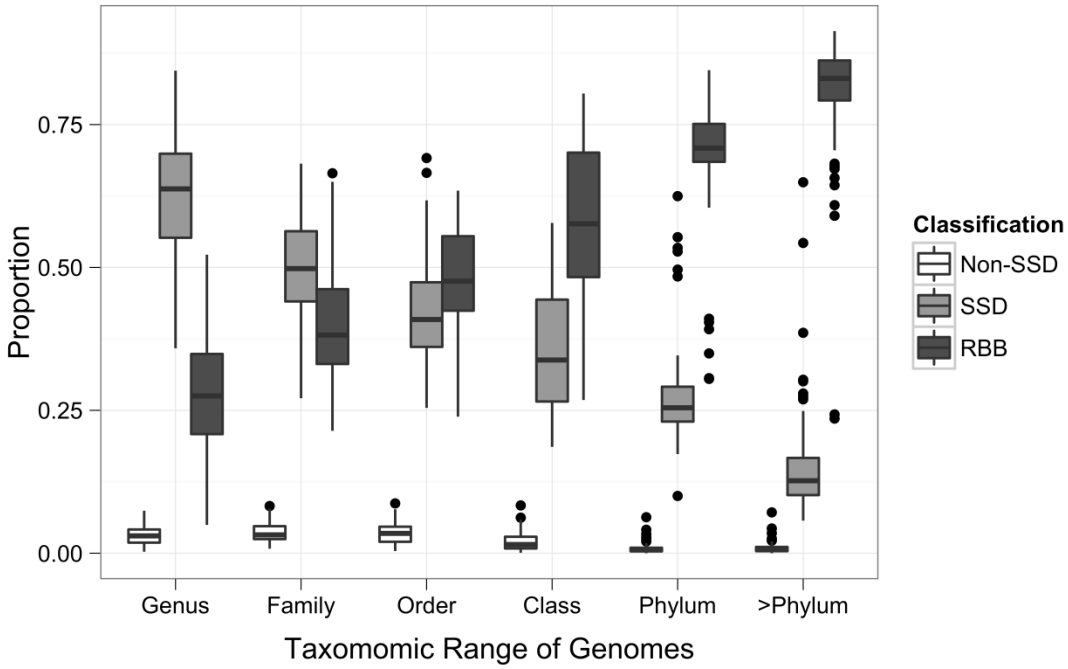


Figure 3.6 Ortholuge Classification Proportions for Species Analyses with Increasing Phylogenetic Distance

The proportion of predicted orthologs unclassified (RBB) as well as proportion classified as SSD, Non-SSD in relation to all other Ortholuge classifications for 660 bacterial and archaeal species analyses. Genomes were organized according to the lowest taxonomic group that contained both species, and proportions were computed separately each evolutionary distance (the 660 species pairs are equally distributed in each of the taxonomic levels).

Improving Ortholuge's coverage, especially for divergent species analyses, would be difficult. The Ortholuge method is constrained to using a single outgroup species. The outgroup species provides a reference from which the degree of divergence for the orthologs is measured. Different outgroup species will change the scale of the Ortholuge ratios. The current approach for selecting an outgroup species seeks to maximize the number of Ortholuge evaluations, by picking the species that has

the highest number of orthologs in common with the analysis species, while still properly satisfying the constraints of an outgroup species (namely, that it diverged prior to the divergence of species under examination). One possible direction for increasing coverage is an iterative approach that repeats Ortholuge analyses using multiple outgroup species. For an overall improvement in coverage, each repeated Ortholuge analysis would need to use an outgroup species that adds further Ortholuge evaluations by having additional orthologs not found in prior outgroup species. A procedure would also need to be developed to merge the results from the repeated evaluations into a final classification. To determine if this proposed approach would be effective, it would need to be evaluated in terms of the added run-time for the repeated Ortholuge analysis versus the boost in coverage. Ultimately, the selected outgroup species appears to be critical to the level of coverage achieved by the Ortholuge method. Further investigation into how outgroup selection strategies can be improved may also generate higher coverage.

3.3.4. Analysis of Unusually-Diverging Orthologs: Conservation of Gene and Protein Features

Assessing the Ortholuge evaluations requires estimating false positive and false negative rates in the orthologs assigned to the SSD and Non-SSD classes. Directly calculating the number of false ortholog predictions would, however, require a large-scale phylogenetic analysis and would not be feasible at a large scale. Typically, ortholog prediction error rates are estimated in one of three ways (Hulsen *et al.* 2006; Kuzniar *et al.* 2008; Altenhoff and Dessimoz 2009):

i. Using a phylogenetic-based gold standard ortholog dataset. The overall ortholog prediction error rate is estimated from a dataset comprising a selection of orthologs for which the orthology has been resolved using phylogenetic methods. Gold-standard datasets do not yet exist for bacterial species.
ii. Re-computing ortholog predictions using computational methods that are deemed complementary to the original ortholog prediction method under evaluation (Hulsen *et al.* 2006; Altenhoff and Dessimoz 2009). Tree-based ortholog prediction methods have been used in several analyses for evaluating graph-based methods. Tree-based ortholog prediction uses automated methods to build gene trees and

then compare the branching order in the gene tree with a supplied species tree to infer orthologs. The alternative ortholog predictions are compared to the original predictions and error rates are derived.

iii. Evaluating the functional similarity of the predicted orthologs using functional parameters. The parameters represent characteristics of a gene or protein that are expected to be conserved amoung orthologs. Functional parameters that have been used to evaluate ortholog prediction methods include manually annotated protein families (HAMAP, KEGG Orthology, GO) (Harris *et al.* 2004; Hulsen *et al.* 2006; Kanehisa *et al.* 2008; Altenhoff and Dessimoz 2009; Lima *et al.* 2009) and features of proteins and genes that are associated with function (protein domains, subcellular localization (SCL)) (Hulsen *et al.* 2006; Yu *et al.* 2010; N.Y. Yu *et al.* 2011). Conservation of gene order in chromosomes is also used as a criterion to evaluate predicted orthologs, as conservation of gene neighbourhood is often an indicator of conserved function among genes (Hulsen *et al.* 2006; Altenhoff and Dessimoz 2009). Conservation of gene order can also be an indicator that the genes residing in the conserved syntenic blocks are derived from a common origin and hence are orthologs. However, paralogous genes can also exhibit synteny in cases where segmental gene duplication has occurred or a gene duplication occurs adjacent to the sister gene, so conserved gene order between species is not a conclusive indicator of orthology. Conversely, many valid orthologs are not found in regions of conserved gene order due to the frequency of genome rearrangement.

Orthologs are frequently used to establish functional equivalency across species in many comparative genomic analyses. Based on their unusual divergence as reported by the Ortholuge assessments, it is our hypothesis that Non-SSD orthologs are less likely to have similar functions and are therefore not appropriate for use in many types of comparative analysis. To evaluate the degree of functional similarity, and also to estimate the improvement in ortholog prediction accuracy provided by Ortholuge, the conservation of a series of gene and protein features was measured in the SSD and Non-SSD groups for large sample of bacterial and archaeal species pairs. The features are typically conserved among functionally-related genes. They include:

1. **KEGG Orthology (KO) annotation**:

   KEGG orthology annotations assign genes to predefined functional roles. The functional roles represent a specific enzymatic mechanism in KEGG's organism-independent reaction and protein-interaction maps called KEGG pathways. The KO annotation process uses experimental evidence and sequence similarity to assign genes to roles. The process is mostly automated but does involve limited manual curation (Kanehisa and Goto 2000; Kanehisa *et al.* 2008).

2. **Subcellular localization (SCL)**:

   Subcellular localization is largely conserved across orthologs, even orthologs in different phylogenetic domains (Nair and Rost 2002). SCL information for proteins was obtained from the PSORTb-DB resource (N.Y. Yu *et al.* 2011). PSORTb uses a combination of SCL predictors based on sequence similarity and unique sequence features such as signal peptides and the protein secondary structure (Yu *et al.* 2010).

3. **Pfam domains**:

   Domains are conserved sequence signatures that correspond to a particular fold or structure in a protein. Domains directly relate to the structure and function of a protein (Sammut, Finn, and Bateman 2008). Pfam domains are identified using Hidden Markov models (HMMs). The HMMs were built from a manually-curated set of multiple sequence alignments for distinct protein family domains.

4. **Tigrfam annotation**:

   Tigrfam is a complementary HMM-computed protein annotation (Haft 2003). The Tigrfam HMMs strive to capture protein families that have equivalent functions rather than the specific protein domains identified by the Pfam HMMs.

The following figures show the conservation of these features for each of the Ortholuge classes (Figure 3.7 for KO conservation, Figure 3.8 for SCL conservation, Figure 3.9 for Pfam conservation and Figure 3.10 for Tigrfam conservation). The conservation is computed overall and for different evolutionary distances (species were organized according to the lowest common taxonomic level they both belong to, and then values were computed for each level). For comparison, the conservation is also shown for in- and out-paralogs. In-paralogs are formed by recent gene duplications that occurred subsequent to the divergence of the species under investigation. Out-paralogs are more ancient duplicated genes that duplicated prior to speciation.

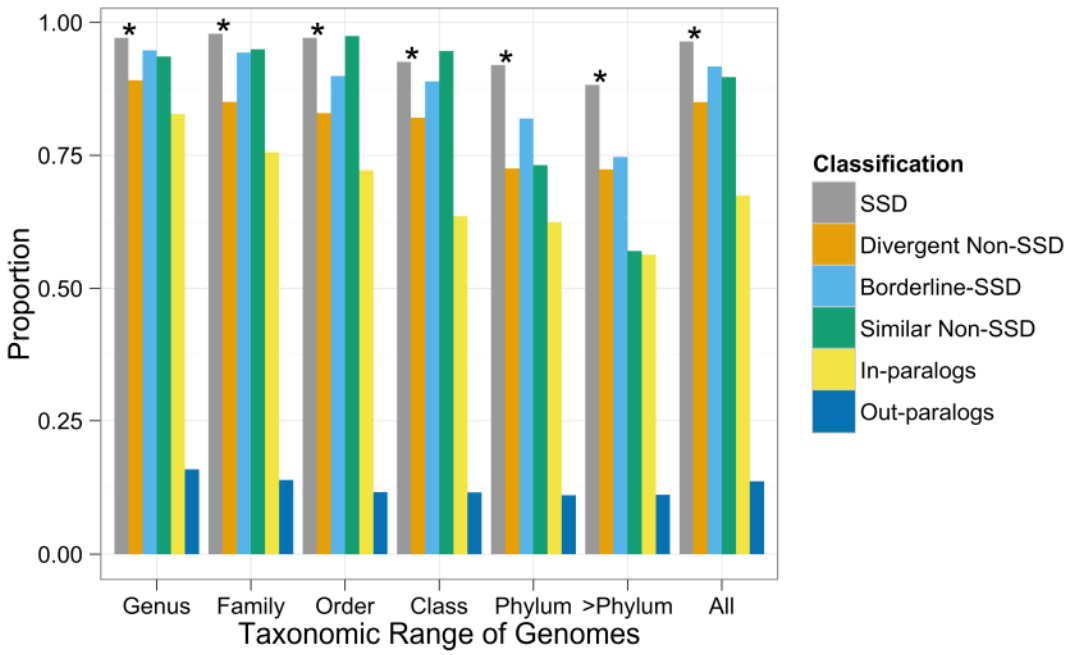


Figure 3.7 Conservation of KO Annotations in Ortholuge Classes

The proportion of predicted orthologs in each Ortholuge class that have identical KEGG Orthology annotations. The evaluation examined 660 bacterial and archaeal species analyses. Genomes were organized according to the lowest taxonomic group that contained both species, and proportions were computed separately each evolutionary distance. For comparison, conservation was also computed for in- and out-paralogs. A *chi*-squared test was used to test if the difference between SSD and Divergent Non-SSD is statistically significant.
* *chi*-squared *p*-value < 0.05.

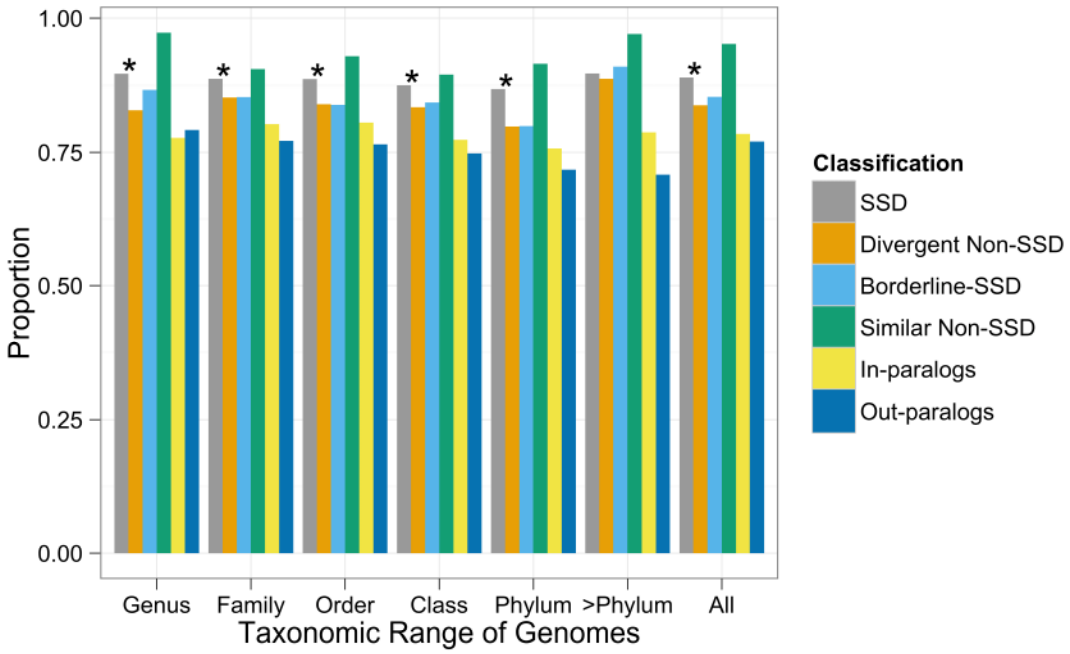


Figure 3.8 Conservation of SCL in Ortholuge Classes

The proportion of predicted orthologs in each Ortholuge class that have identical subcellular localizations. For comparison, conservation was also computed for in- and out-paralogs. A *chi*-squared test was used to test if the difference between SSD and Divergent Non-SSD is statistically significant.
***** *chi*-squared *p*-value < 0.05.

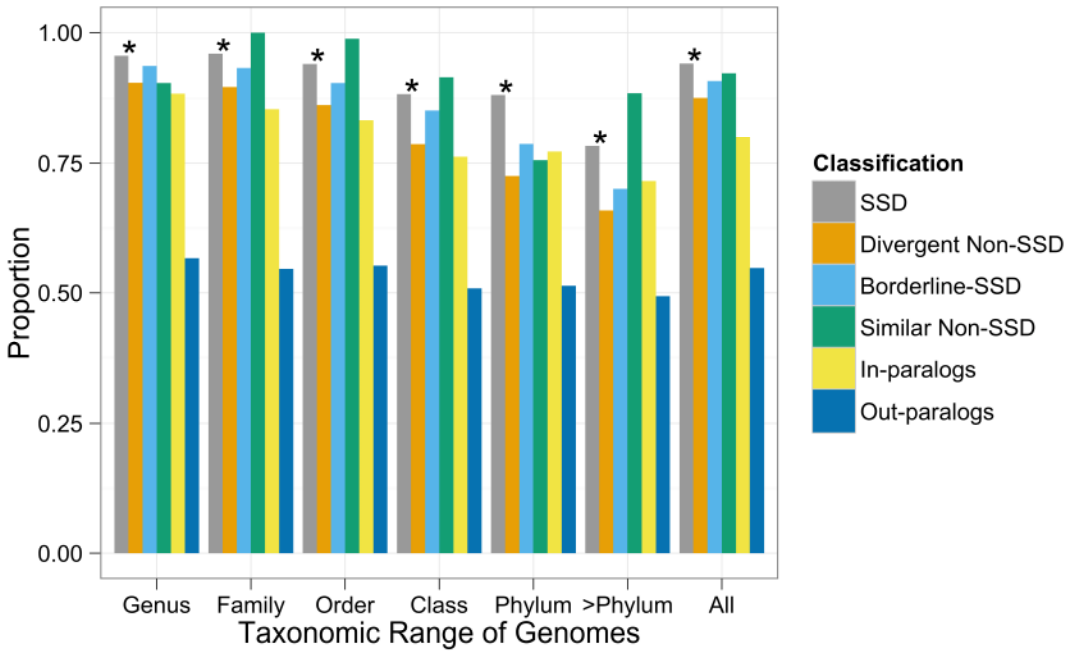


Figure 3.9 Conservation of Pfam Domains in Ortholuge Classes

The proportion of predicted orthologs in each Ortholuge class that have identical Pfam domains. For comparison, conservation was also computed for in- and out-paralogs. A *chi*-squared test was used to test if the difference between SSD and Divergent Non-SSD is statistically significant.
***** *chi*-squared *p*-value < 0.05.

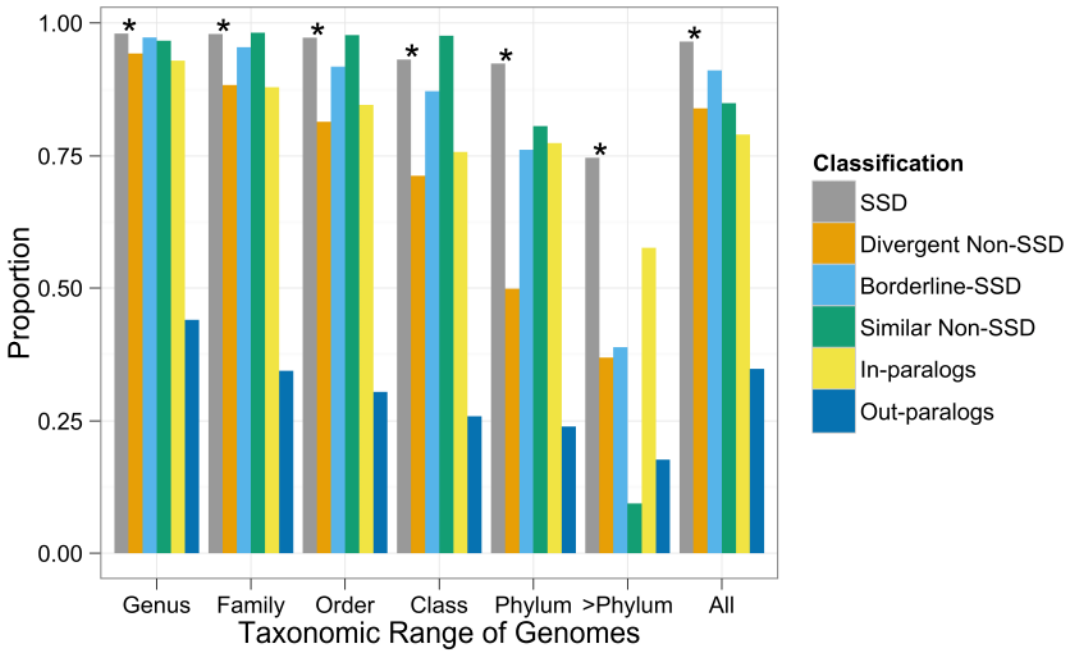


Figure 3.10 Conservation of Tigrfam Annotation in Ortholuge Classes

The proportion of predicted orthologs in each Ortholuge class that have identical Tigrfam annotations. For comparison, conservation was also computed for in- and out-paralogs. A *chi*-squared test was used to test if the difference between SSD and Divergent Non-SSD is statistically significant.
***** *chi*-squared *p*-value < 0.05.

The examination of the gene and protein features revealed that there is a statistically significant difference in the proportion of SSD and Non-SSD predicted orthologs with identical features. Non-SSD predicted orthologs have a significantly lower conservation of the functional features. This result is consistent for all features tested and is also consistent in the separate evolutionary distances that were examined. While some of the differences may be due to errors by the detection programs used to identify the features, the trends are substantive, so it is unlikely detection error can account for the significant differences between the Non-SSD and SSD groups. Subcellular localization, protein domain content (Pfam), protein family (Tigrfam) and assigned functional role (KO) are characteristics that are linked to gene function. Differences in these features suggest that the predicted orthologs have dissimilar functions. The difference in function could be due to mispredicting a paralog as an ortholog (the ortholog conjecture posits that duplicated paralogous genes will rapidly evolve non-

overlapping functions) or one of the orthologs have diverged significantly and the divergence is associated with a change in gene function.

Of note is the high degree of conservation of the features in the Similar Non-SSD class. Similar Non-SSD predicted orthologs have ratio values that suggest they are undergoing unusual divergence (the ratios indicate the branch length is proportionally longer than expected in the orthologs' phylogenetic tree). However, in these cases the total branch length is small indicating that the proteins have highly similar protein sequences. Based on the high level of feature conservation, it does not appear the unusual divergence reported by the ratios is indicative of a difference in function in the similar Non-SSD orthologs (at least not to the same degree as the divergent Non-SSD predicted orthologs). There are two explanations to reconcile the seeming lack of functional divergence and the usual sequence divergence reported by the Ortholuge ratios for the similar Non-SSD class. First, the similar Non-SSD class may consist of highly conserved proteins where the majority of residues are under negative selection and where positive selection is preserving mutations in a small number of sites that are impacting the function of the protein. In this situation, because the positive selection is only acting on a small number of sites, it is likely the features tested in this analysis do not have the sensitivity to discriminate this level of functional change. The second possibility is that the mutations causing the unusual divergence (as reported by the Ortholuge ratios) are the result of neutral evolution and are not impacting the function. By random chance, the neutral mutations occur more frequently in one ortholog's protein sequence than the other. Because the number of mutations is small, this situation is feasible. For example, when compared to the outgroup ortholog sequence, if one ortholog sequence had one mutation and the other had two mutations, this would produce a ratio value of approximately 2 (after converting to phylogenetic distances). An ortholog with twice the level of divergence is considerable in Ortholuge analysis. Determining which possibility is occurring for the similar Non-SSD predicted orthologs: unequal neutral evolution or positive selection driving functional change through mutation in a small number of sites is left up to the user. Currently, the similar Non-SSDs are only flagged (the procedure for identifying similar Non-SSDs is described in section 4.5).

Feature conservation was also measured for two types of paralogous genes: in-paralogs, which are the result of recent gene duplications, and out-paralogs which arise from more ancient duplications (Sonnhammer and Koonin 2002). Comparing the paralogs to the ortholog groups (such as SSD orthologs), it is clear that paralogs more often have dissimilar properties. As the features are related to gene function, this result supports the ortholog conjecture which states that paralogs more frequently have divergent functions than orthologs (Koonin 2005).

3.3.5. Analysis of Unusually-Diverging Orthologs: Association with Large Gene Families

Large gene families can complicate ortholog prediction, as successive gene duplications are often associated with rapid sequence divergence as well as gene loss (Koonin 2005). To examine if there is an association between gene family size and unusually-diverging predicted orthologs (Non-SSDs), the number of homologs in a genome were counted for the predicted orthologs in each of the Ortholuge classes. Figure 3.11 shows that there is statistically significant difference in the number of Non-SSD predicted orthologs with at least one homolog compared to SSD orthologs. The connection between gene family size and Non-SSDs, which represent unusually-diverging orthologs, suggests that large gene families may be associated with higher error rates in ortholog prediction.

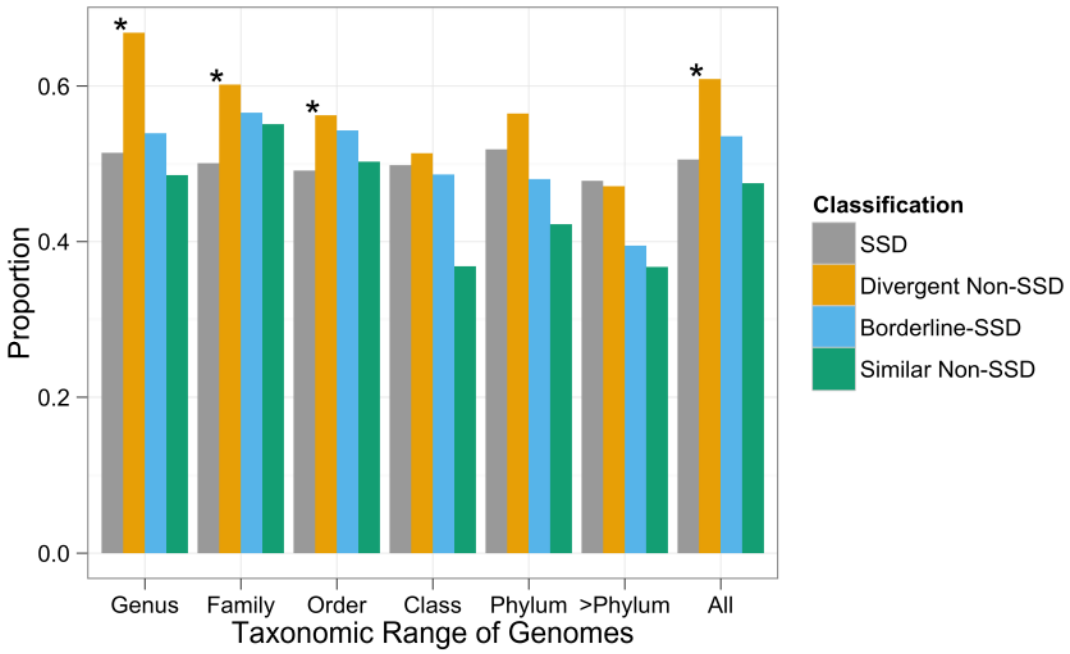


Figure 3.11 Proportion of Orthologs in Ortholuge Classes with one or more Homologs

A *chi*-squared test was used to test if the difference between SSD and Divergent Non-SSD is statistically significant.
***** *chi*-squared *p*-value < 0.05.

3.3.6. *Analysis of Unusually-Diverging Orthologs: Synteny*

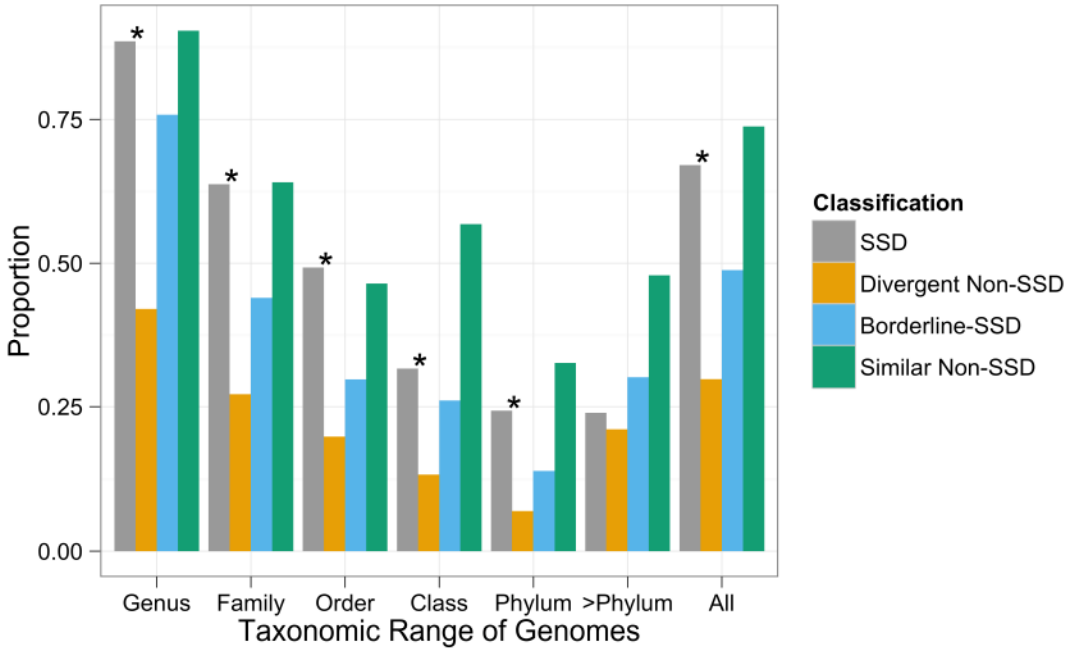


Figure 3.12 Conservation of Immediate Gene Neighbourhood in Ortholuge Classes

The proportion of predicted orthologs for which at least one of the adjacent upstream or downstream genes in each of the species are also orthologous. This metric provides a simplified view of gene synteny by determining if the immediate gene neighbourhood is conserved in the species. A *chi*-squared test was used to test if the difference between SSD and Divergent Non-SSD is statistically significant.
***** *chi*-squared *p*-value < 0.05.

Gene synteny is the conservation of the orthologous gene's ordering on the chromosomes of their respective species (Dewey 2011). As a metric for assessing ortholog prediction, it is highly insightful, but drawing conclusions is complicated by a number of confounding factors. Synteny is the only non-phylogenetic metric that can inform about the ancestral origins of genes. Genes with conserved order in two species suggests that the genes diverged from the same set of ancestral genes when the species diverged, and hence are valid orthologs (Hulsen *et al.* 2006; Lemoine, Lespinet, and Labedan 2007; Dewey 2011). Conserved gene neighbourhoods can also be an indicator of related gene functions, as selective pressure can act to keep functionally related genes clustered on the chromosome (this is especially evident in bacteria) (Hulsen *et al.* 2006). However, through duplication of multi-gene segments in a genome,

a set of paralogs can also exhibit conserved gene order. Additionally, bacterial and archaeal genomes are highly fluid with frequent genome rearrangements which will rapidly diminish syntenic regions. The results in Figure 3.12 are the proportion of predicted orthologs for which the upstream or downstream genes are also orthologs (i.e. a conserved gene order block of at least size 2). There is a significant difference in the proportion of SSD orthologs to Non-SSDs with conserved gene order: over 60% of the SSD compared to less than 30% of the Non-SSDs. While the test only examines the upstream and downstream genes, the result is indicative of the overall genome synteny, as any gene part of a conserved region will have their immediate gene neighbourhood conserved. The completely resolved genome synteny was calculated for a smaller test set of bacterial species and the size of the block with conserved gene order was recorded for all SSD and Non-SSD predicted orthologs (size is the number of genes contained in the block). From Figure 3.13, it is clear SSD orthologs are more often found in regions of conserved gene order and these regions are typically larger. Even when considering the confounding factors in using synteny as a metric for assessing orthologs, the striking differences between the SSD and Non-SSD gene order conservation suggests that many Non-SSD are paralogs mispredicted as orthologs.

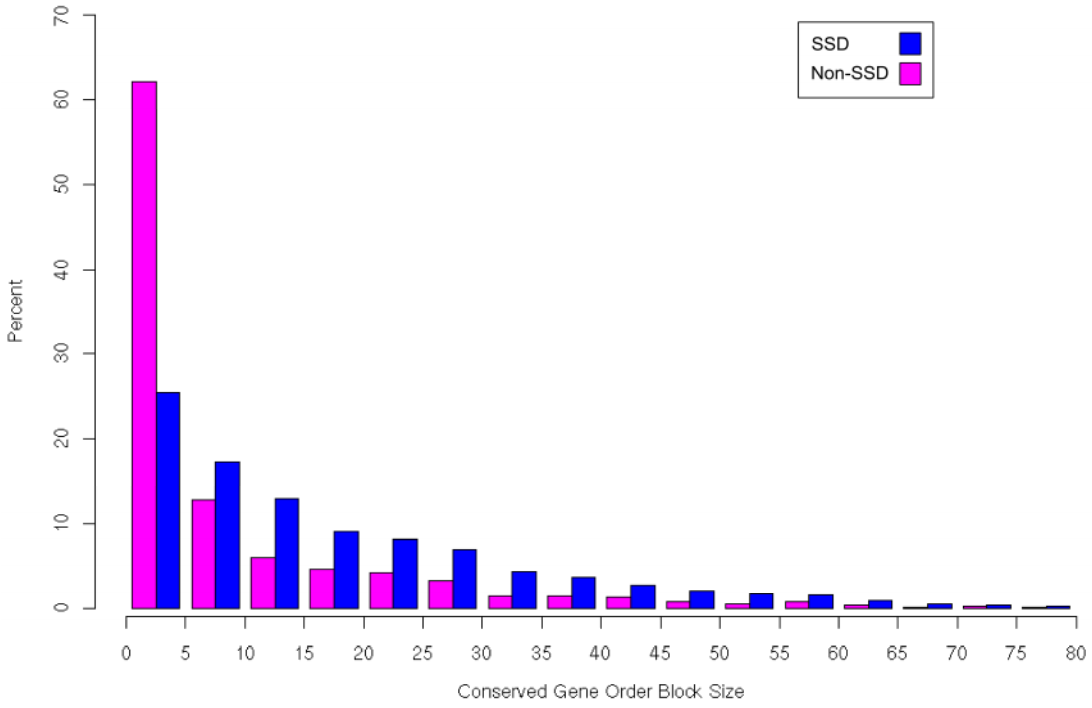


Figure 3.13 Conserved Gene Order Block Size for Ortholuge Classes

The genomes of 63 pairs of *Pseudomonas* species were partitioned into blocks of conserved gene order and the percent of the total SSD and Non-SSD predicted orthologs that were found in blocks of a certain size were recorded.

3.3.7. Analysis of Unusually-Diverging Orthologs: Association with Functional Categories

To study the broader functions associated with Non-SSD and SSD orthologs, KEGG BRITE functional categories were tested for over-representation in orthologs from one of these two Ortholuge classes (Kanehisa and Goto 2000; Kanehisa *et al.* 2008). A number of gene function categories contained significantly higher proportions of Non-SSD genes than SSD orthologs (Table 3.2). As shown in the above analyses, Non-SSD genes consist of paralogs and unusually-diverging orthologs. The over-representation of particular functional categories amoung Non-SSD genes may represent problem areas for ortholog prediction or functional components of a species that are undergoing unusual evolution. A second possibility is that these two explanations may be inter-dependent. Many of the factors that create problems to graph-based ortholog prediction

such as the loss and duplication of orthologs will more frequently occur in gene families associated with unusual divergence. Gene families undergoing rapid expansions and contractions will likely have lower ortholog prediction accuracy. A statistical association between Non-SSD orthologs and large gene families was detected (See section 3.3.5).

Table 3.2 Functional Categories Significantly Associated with SSD and Non-SSD Orthologs

The functional categories in the KEGG BRITE Functional Hierarchy were tested using a fisher's exact test for over-representation in SSD and Non-SSD orthologs. Orthologs were obtained from a sample of 660 bacterial and archaeal pair-wise species comparisons.

BRITE Functional Category	Adjusted P-value	Proportion of SSD Orthologs	Proportion of Non-SSD Orthologs
Functional Categories Significantly Associated with SSD Orthologs			
Replication and Repair	0.000E+00	0.0465	0.0104
Translation	0.000E+00	0.1089	0.0454
Folding, Sorting and Degradation	9.568E-207	0.0379	0.0121
Cell Growth and Death	1.505E-103	0.0130	0.0027
Glycan Biosynthesis and Metabolism	4.566E-69	0.0272	0.0139
Nucleotide Metabolism	4.378E-61	0.0776	0.0560
Transcription	1.519E-48	0.0059	0.0012
Metabolism of Cofactors and Vitamins	3.413E-22	0.1072	0.0921
Metabolism of Terpenoids and Polyketides	3.343E-04	0.0302	0.0270
Functional Categories Significantly Associated with Non-SSD Orthologs			
Xenobiotics Biodegradation and Metabolism	1.2186E-199	0.0262	0.0549
Carbohydrate Metabolism	4.0235E-127	0.1047	0.1447
Amino Acid Metabolism	5.0402E-115	0.1401	0.1826
Energy Metabolism	5.0649E-71	0.0873	0.1145
Metabolism of Other Amino Acids	2.3268E-43	0.0331	0.0467
Membrane Transport	5.2726E-38	0.0529	0.0685
Cell Motility	3.4633E-35	0.0155	0.0240
Biosynthesis of Other Secondary Metabolites	1.3113E-26	0.0085	0.0140
Excretory System	6.3915E-22	0.0005	0.0021

BRITE Functional Category	Adjusted P-value	Proportion of SSD Orthologs	Proportion of Non-SSD Orthologs
Transport and Catabolism	5.3459E-18	0.0038	0.0069
Lipid Metabolism	1.7283E-07	0.0377	0.0430

The specific functional categories linked to Non-SSD and SSD orthologs are potentially revealing and may suggest connections between the functions associated with Non-SSD genes and roles in species evolution. Many of the categories over-represented with Non-SSD genes are types of metabolism or molecular transport. Metabolism has a direct impact organism's fitness in specific environmental niches (Pál, Papp, and Lercher 2005; Hibbing *et al.* 2010; Rohmer, Hocquet, and Miller 2011). In comparison, several of the categories associated with SSD orthologs are core systems such as transcription, translation, replication and repair and protein folding, sorting and degradation. These systems are largely conserved between species. Metabolic genes, however, are highly variable between species. Many metabolic genes, especially in peripheral metabolic pathways, are found in select species lineages. The pivotal role metabolism plays in evolutionary fitness may drive increased adaptive radiation in genes associated with metabolism (Francino 2005). Adaptive radiation would be associated with the generation of novel genetic variations and content through horizontal gene transfer or gene duplication, followed by the fixation of certain paralogs and the loss or pseudogenization of others (Francino 2005). The selection driven process of gene amplification and loss in specific functional categories is potentially what is being detected in this analysis where categories such as metabolism are over-represented in Non-SSD orthologs and core processes such as transcription and translation are associated with SSD orthologs.

3.3.8. Comparison of Ortholuge to Other Ortholog Prediction Methods

Two other ortholog prediction methods use a reference genome to evaluate orthologs predicted by RBB: OMA (Roth, Gonnet, and Dessimoz 2008) and QuartetS (C. Yu *et al.* 2011). OMA and QuartetS are high-throughput methods that produce pair-wise

ortholog predictions. These methods were selected for comparison to Ortholuge because of the similarity in their approach and because, in evaluations of the accuracy of their pair-wise ortholog predictions, OMA and QuartetS were found to be two of the best performing methods. OMA was found to perform well in comparison to several other methods including OrthoMCL, RoundUp and Homologene (Altenhoff and Dessimoz 2009). QuartetS was shown to have a slightly lower false-positive rate than OMA (Yu *et al.* 2012). Similar to Ortholuge, QuartetS reconstructs a gene tree. However, QuartetS uses four genes, the predicted orthologs and two genes in a reference genome to build the gene tree, and instead of phylogenetic distances, QuartetS uses BLAST bit-scores to represent branch lengths (Yu *et al.* 2012). Another significant difference is that QuartetS examines differences in the branch lengths, rather than ratios of the branch lengths. OMA also uses four genes to analyze predicted orthologs, but instead of reconstructing a gene tree, OMA uses heuristic rules to interpret the sequence information from the four genes (Roth, Gonnet, and Dessimoz 2008). The approaches for selecting genes to use as references or outgroups are also different between OMA, QuartetS and Ortholuge. Ortholuge selects a single outgroup species and identifies potential orthologs in that species to use as reference genes (only a single species can be used because in order to observe the divergence for all predicted orthologs relative to the species divergence, it must be on the same scale. If no ortholog exists, no evaluation of the ortholog is performed). OMA and QuartetS search all possible outgroup species for possible reference genes. This increases the number of RBB-predicted orthologs that are further analyzed in OMA and QuartetS. A Venn diagram showing the overlap between the three prediction methods is provided in Figure 3.14 (this figure lists the overlap in predicted orthologs that are evaluated by the method. Ortholuge provides unevaluated RBB-based predictions as well. The other methods only provide evaluated ortholog predictions, so the numbers represent the total ortholog predictions for those methods).

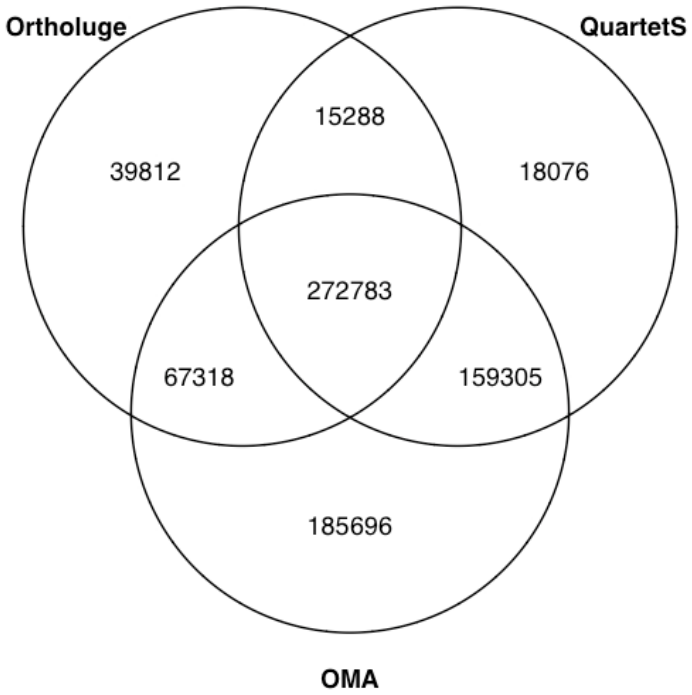


Figure 3.14 Overlap of the Evaluated Ortholog Predictions in Ortholuge, QuartetS and OMA

Orthologs were predicted in 660 pairs of bacterial and archaeal species and the common and unique orthologs generated by Ortholuge, OMA and QuartetS programs was recorded (only orthologs that are evaluated are included. Unevaluated RBB-based ortholog predictions from the Ortholuge method were not included in this result).

As expected, the number of evaluated ortholog predictions is much higher in OMA and QuartetS than in Ortholuge. OMA and QuartetS evaluate all orthologs generated by RBB because they search all possible genomes for the reference genes used in their assessments. Although OMA and QuartetS select the best possible reference gene, they do not limit the reference genes to optimally-positioned outgroups. Ortholuge only performs an evaluation of RBB-predicted orthologs when a suitable outgroup genome is available. Despite being all based on the RBB detection strategy, significant numbers of predicted orthologs are unique to one method. One difference that accounts for the unique ortholog predictions is what is considered an ortholog by the method. OMA produces significantly higher numbers of predicted orthologs because it includes in-paralogs in the ortholog results (OMA considers in-paralog as a co-ortholog and frequently returns many-to-many ortholog predictions). In-paralogs are also provided in the QuartetS and Ortholuge tools, but are considered distinct and not

automatically included in the ortholog results. Differences in operational details may also account for some of the unique ortholog predictions. The minimum sequence match thresholds are different between the approaches (for example, Ortholuge has a maximum e-value threshold, OMA and QuartetS have minimum alignment length requirements). The RBB procedure selects the reciprocally best BLAST hit as putative ortholog. How multiple best hits are managed, also differs between the methods.

We compared the performance of OMA, QuartetS and Ortholuge methods by examining multiple criteria: similarity of protein domains, subcellular localization (SCL), KEGG orthology (KO) and Tigrfam annotations and whether the predicted orthologs displayed conserved gene order or synteny. These features should reflect the functional similarity of the predicted orthologs because they are highly conserved among functionally similar genes. Figure 3.15 shows the proportion of SSD orthologs in Ortholuge and validated orthologs in OMA and QuartetS with identical features.

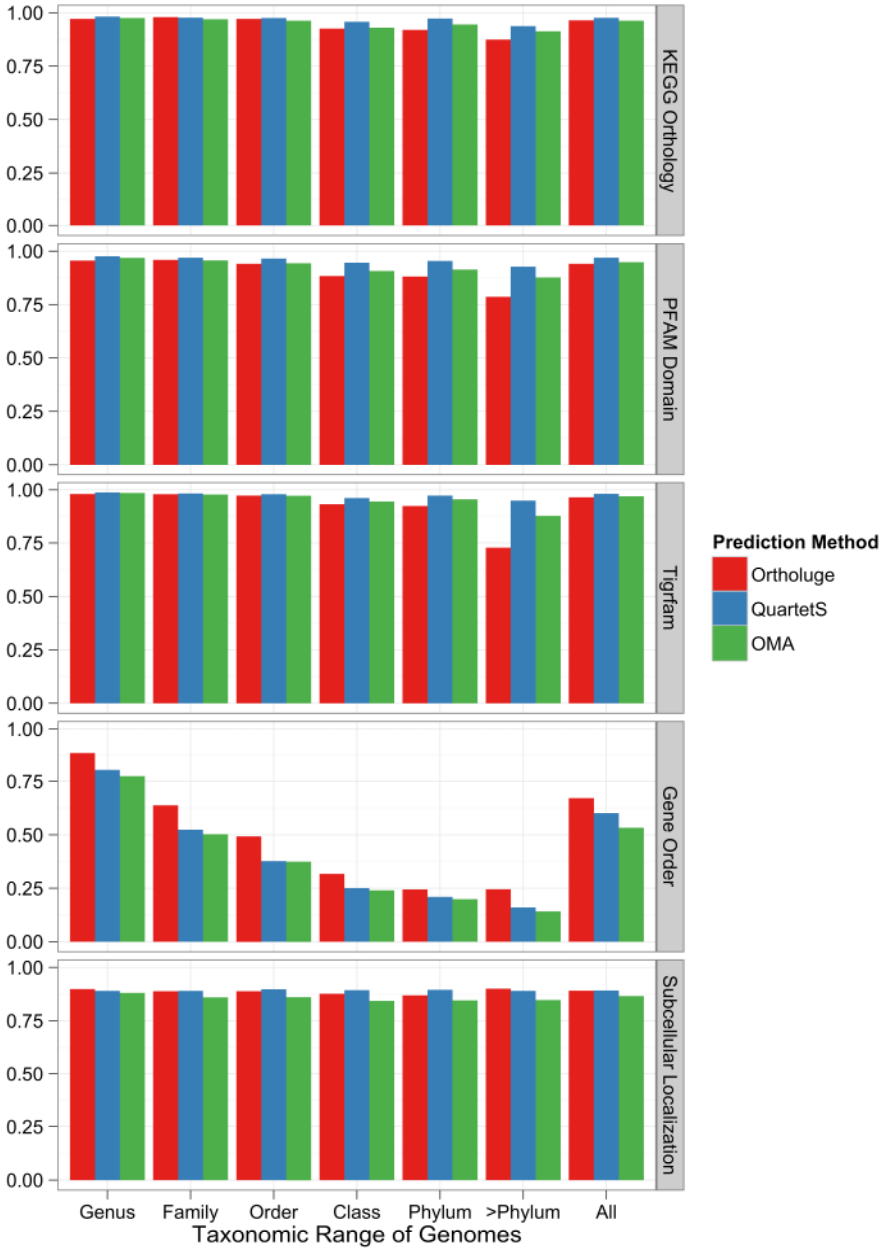


Figure 3.15 Conservation of Gene Features in Orthologs Validated by Ortholuge, QuartetS and OMA

The proportion of validated orthologs with identical KEGG ortholog annotations, subcellular localizations, Pfam domains and Tigrfam annotations for the ortholog assessment methods Ortholuge, OMA and QuartetS. The proportion of validated orthologs displaying conserved gene order was also measured. This analysis examined the predicted orthologs from a set of 660 bacterial and archaeal species pair-wise analyses. In addition to computing feature conservation for all species combinations, the pair-wise sets of genomes were organized according to the lowest taxonomic group that contained both species, and separate values were computed for genomes with different taxonomic ranges.

Although gene features such as protein domains and subcellular localization are highly conserved between orthologs, these features can also be conserved among closely-related paralogs. As a result, using the similarity of these features as evaluation criteria will under report the number of false-positives. Similarly, gene synteny can occur between paralogs in cases where segmental duplications preserve the immediate gene neighbourhood. To improve confidence in the inferred functional similarity, we looked at results produced from combining the five criteria. For the combined analysis, we counted predicted orthologs having at least three of the five criteria: the same Pfam domains, localization, KO or Tigrfam and displaying synteny (based on the operational definition), as a valid ortholog or true positive (because we are using indirect criteria and not a gold standard to assess performance, this is only an inferred true positive). Using this definition of positives and negatives, we compared precision, defined as $TP/(TP + FP)$; recall, defined as $TP/(TP + FN)$; accuracy, defined as $(TP + TN)/(TP + TN + FP + FN)$; and Matthew's Coefficient Constant (MCC), defined as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{TP + FP} \times \sqrt{TP + FN} \times \sqrt{TN + FP} \times \sqrt{TN + FN}}$$

for QuartetS and Ortholuge methods (Figure 3.16). For this evaluation, orthologs computed for 660 pairs of species were examined. The species pairs represented all taxonomic classifications and were equivalently distributed in all levels. Results are unavailable for OMA, because the OMA website does not provide orthologs that were rejected by the method. Ortholuge appears to perform comparably in terms of precision and MCC and slightly better in terms of recall and accuracy when compared to QuartetS.

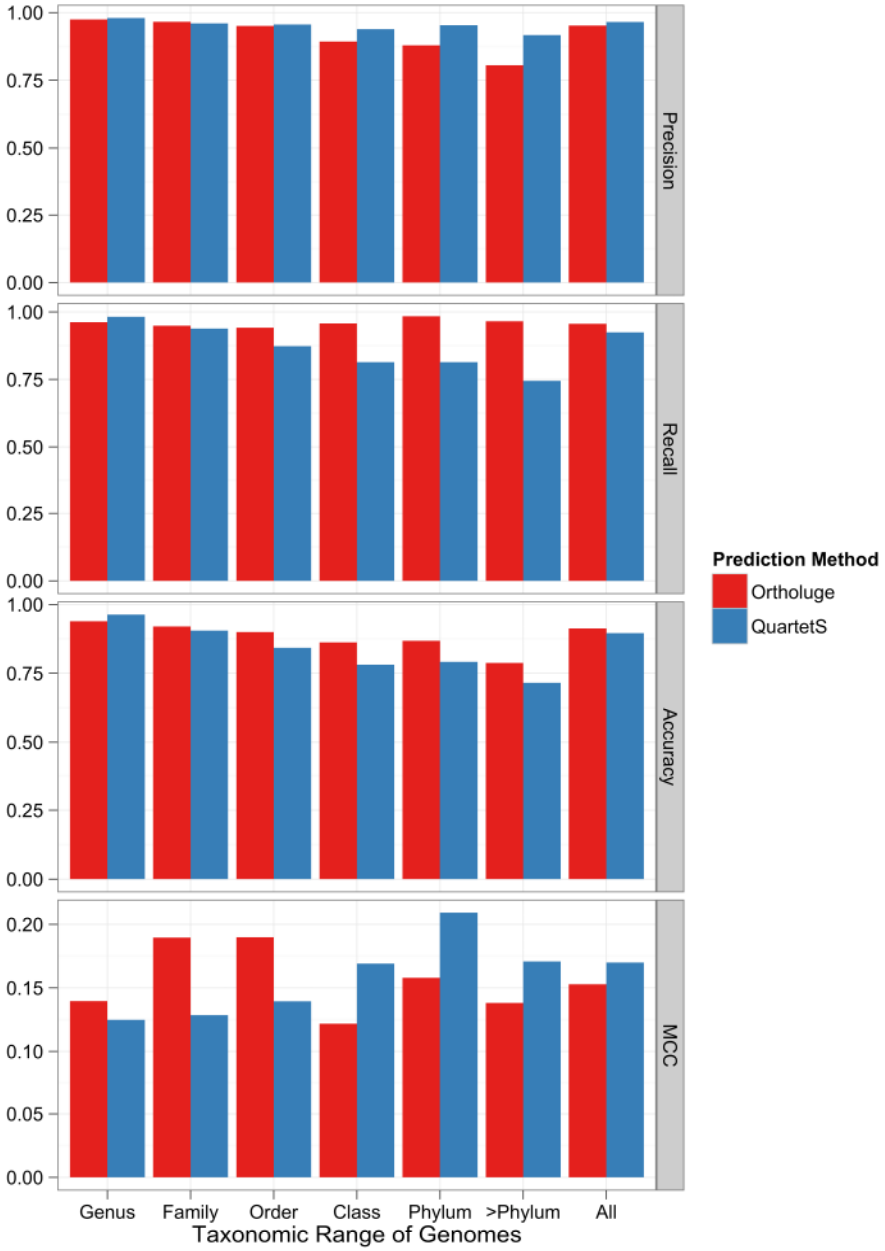


Figure 3.16 Comparison of the Performance of Ortholuge and QuartetS

A comparison of the precision, recall, accuracy and Matthew's correlation coefficient (MCC) for the Ortholuge and QuartetS ortholog prediction methods using 660 pair-wise bacterial and archaeal species analyses. In addition to computing the performance for all species combinations, the pair-wise sets of genomes were organized according to the lowest taxonomic group that contained both species, and separate values were computed for genomes with different taxonomic ranges. In this analysis, a "true positive" is defined as having at least three of the following features as being conserved: subcellular localization, protein domains, Tigrfam annotations, KEGG ortholog annotations or gene order.

The improved performance of the QuartetS method as reported by the precision and MCC metrics may be due to the significantly larger number of ortholog assessments the method produces. To determine the contribution of the larger number of predictions to the overall performance of QuartetS as compared to Ortholuge, an examination of only the orthologs that were predicted and assessed by both methods was carried out. For the common evaluated orthologs in Ortholuge and QuartetS, it was determined whether the feature conservation supports or rejects the prediction (if three of the five criteria, that the orthologs have the same KO, Tigrfam, SCL, Pfam domains or conserved gene order, then the ortholog was deemed a valid ortholog or true positive). This data is shown in Table 3.3.

Table 3.3 A Direct Comparison of the Predicted Orthologs Produced by both Ortholuge and QuartetS

The functional similarity of the common orthologs produced by both QuartetS and Ortholuge was compared to the assessment generated by the method. The number of assessments that agreed or disagreed with the data was recorded for each method. This examination looked at the common orthologs in 660 pair-wise bacterial and archaeal datasets. The functional similarity was determined by examining whether at least three of the five the functional parameters: SCL, KO, Tigrfam, Pfam domains and synteny were conserved between the orthologous genes.

Ortholog Method	Ortholog Assessment Supported by the Similarity of Functional Parameters		Ortholog Assessment Rejected by the Similarity of Functional Parameters	
	Agreeing Assessments between Methods	Disagreeing Assessments between Methods	Agreeing Assessments between Methods	Disagreeing Assessments between Methods
Ortholuge	475884	8914	42232	7226
QuartetS	475884	7226	42232	8914

For predicted orthologs where the QuartetS and Ortholuge assessments disagree, more often the functional parameters support the Ortholuge assessment. The large majority of ortholog assessments agree between QuartetS and Ortholuge. There is a notable number of agreeing assessments, however, that are not supported by the data. This observation may indicate there are common areas that could be improved in both methods to better predict functionally-similar orthologs. It may also be an indication that the ability of functional parameters to correctly report functional similarity is not

optimal. To see if there is an association between the differences in performance and evolutionary distance between the species under investigation, the taxonomic range of the species where one method; Ortholuge or QuartetS, outperformed the other in a pair-wise dataset, was recorded (Table 3.4).

Table 3.4 Association between Performance and the Taxonomic Range of the Species for Ortholuge and QuartetS

Ortholuge and QuartetS were compared on a dataset-by-dataset basis. Datasets where one method produced more ortholog assessments that were validated by the similarity functional parameters in the orthologs than the other method were recorded and the taxonomic range of the species in the dataset was noted. The table reports the total number of datasets for each taxonomic range.

Taxonomic Range	Number of Pair-wise Datasets where Method Produces More Assessments that Agree with Functional Parameter Similarity	
	Ortholuge	QuartetS
Genus	5	94
Family	52	43
Order	85	11
Class	103	2
Phylum	99	2
>Phylum	60	28
All	404	180

Ortholuge outperforms QuartetS in species with greater taxonomic separation. One possible explanation for this trend is that the features being examined, such as protein domains, are less likely to be conserved among paralogs of distantly related species. Consequently, the number of false positives that are masked will be less for these species, suggesting that distantly related species might more accurately report the performance in these cases. Alternatively, there may be methodological differences that accounts for the difference in performance for evolutionary distant species. QuartetS examines the difference in distance in the gene tree between the predicted orthologs and two reference genes. If this difference is above the threshold, the ortholog is rejected. A static cut-off such as the one used in QuartetS is unable to adjust for changes in the expected level of divergence between the orthologs as the evolutionary

distance between the species increases (C. Yu *et al.* 2011). Because Ortholuge compares the proportional length of the branches in the ortholog gene tree in relation to the values obtained for the entire genome, it scales with the evolutionary distance of the species (Fulton *et al.* 2006). This flexibility in deciding valid and invalid orthologs may account for the improved performance of Ortholuge for species with increasing evolutionary distance. Overall, based on these criteria for evaluating orthologs functional similarity, Ortholuge appears to more consistently identify RBB-predicted orthologs with conserved features across a wide range of taxonomic distances.

3.4. Conclusions

On average, 7.09% of microbial species RBB-predicted orthologs are classified as borderline or unusually-diverging by the Ortholuge method. Analysis of the Non-SSD genes, through the comparison of multiple functional parameters, suggests that a significant proportion of Non-SSD genes are paralogs mistakenly predicted as orthologs by RBB. Non-SSD genes are also significantly associated with having multiple intra-genome paralogs, a known confounding factor in RBB-based ortholog prediction. An examination of the synteny between orthologs in multiple species showed that more often SSD genes are found in regions where the gene order is conserved. Regions with conserved gene order suggest that the constituent genes likely have a common origin (i.e. are orthologous). While graph-based ortholog prediction methods have benefits such as scalability and minimal input requirements, these results suggest that graph-based methods such as RBB produce limited numbers of false ortholog predictions. When accurate ortholog prediction is critical to the down-stream application, it is important to complement high-throughput ortholog prediction with tools such as Ortholuge that can evaluate the predicted orthologs. Interestingly, certain functional categories were found to be statistically associated with Non-SSD or SSD orthologs. These biological functions may indicate functional components of a species undergoing unusual divergence or problem areas for ortholog prediction.

In a comparison of related tools for graph-based ortholog prediction, Ortholuge was found to perform comparably in terms of prediction accuracy. Ortholuge appears to more consistently identify RBB-predicted orthologs with similar functions across a wide

range of taxonomic distances. However, prediction coverage is an area of Ortholuge's performance that is in need of improvement. Ortholuge, by identifying orthologs that diverged to the same relative degree as their species, produces a set of orthologs that are more likely to have retained similar function and are better suited for comparative genomic analyses.

4. Improvements and Modifications Made to the Ortholuge Method

Portions of this chapter have been previously published in the articles "OrthologeDB: a bacterial and archaeal orthology resource for improved comparative genomic analysis", co-authored by M.D. Whiteside, G.L. Winsor, M.R. Laird, and F.S.L. Brinkman in Nucleic Acids Research, Volume 41, Issue D1, Pages D366-76 © 2013 Whiteside et al; licensee Oxford University Press, and "A statistical approach to high-throughput screening of predicted orthologs", co-authored by J.E. Min, M.D. Whiteside, F.S.L. Brinkman, B. McNeney and J. Graham in Computational Statistics and Data Analysis, Volume 55, Issue 1, Pages 935-43 © 2011 Eun et al; licensee Elsevier.

4.1. Introduction

Ortholuge is a computational method that generates a more precise set of ortholog predictions. It performs assessments on the complete set of predicted orthologs from two genomes (a third genome is used as a outgroup reference). Ortholuge was developed by the Brinkman laboratory and first reported in *BMC Bioinformatics* by Fulton *et al*. in 2006. The first version of Ortholuge was implemented as a computational pipeline consisting of a series of Perl scripts that carried out the linear steps in the Ortholuge procedure (Perl is a program scripting language). In this version, each script loaded the data it needed and outputted the results from the particular analysis step. The scripts use a number of external programs to perform tasks such as sequence alignment, which uses the alignment program Muscle (Edgar 2004a; Edgar 2004b), and phylogenetic distance calculation, which uses the Dnadist and Protdist programs from PHYLIP (Felsenstein 2008).

Since the release of the first version of Ortholuge, several significant modifications have been made to Ortholuge. Three aspects of Ortholuge's performance: speed, accuracy and usability were targeted for improvement and the modifications address these performance areas. Summarized in Table 4.1 are the major modifications made to Ortholuge and the performance issues they address. The modifications are described in detail in the following sections.

Table 4.1 Summary of the Modifications Made to Ortholuge

Modification	Performance Issue the Modification Addresses		
	Speed	Accuracy	Usability
Redesign of the Ortholuge Pipeline (Section 4.2)	✓		✓
Detection of In-paralogs (Section 4.3)		✓	✓
Statistical Computation of the Ortholuge Ratio Cut-offs (Section 4.4)	✓	✓	
Sub-classification of Non-SSD Predicted Orthologs (Section 4.5)		✓	

4.2. Redesign of the Ortholuge Pipeline

The Ortholuge pipeline has been redeveloped to improve efficiency and reduce run-time and also consolidate the analysis outputs. Outlined below are the changes made to Ortholuge software:

4.2.1. Object-Oriented Design

In the first version, the Ortholuge analysis was separated into distinct steps and individual Perl scripts were developed to perform each analysis step (Fulton *et al.* 2006). This design offered a measure of flexibility because the analysis could be restarted from any of the analysis steps without having to re-run the entire Ortholuge analysis. However, the separation of the Ortholuge analysis into discrete steps also generated redundancy. Several of the steps use the same data (such as the gene sequences or

RBB ortholog predictions) and the individual Perl scripts had to reload this data into computer memory.

The Ortholuge pipeline underwent a re-development based on object-oriented software design principles (Figure 4.1). Perl objects were created for each of the analysis steps in the Ortholuge pipeline and were roughly equivalent to the original Perl scripts (a script is characterized by a sequence of commands possibly with calls to subroutines that are used to break the commands into discrete steps. An object is an abstract collection of data and methods that operate on the data). The most significant change was the addition of two Perl objects, called OrtholugeData and OrtholugeConfig that manage the data and the program configurations. Instead of the modules for individual analysis steps having to reload data, the Ortholog data objects load the data into memory once and the individual analysis modules access the data through the methods provided in the Ortholuge data objects. The object-oriented design, through the better delegation of repeated tasks, improves the overall efficiency of the Ortholuge program.

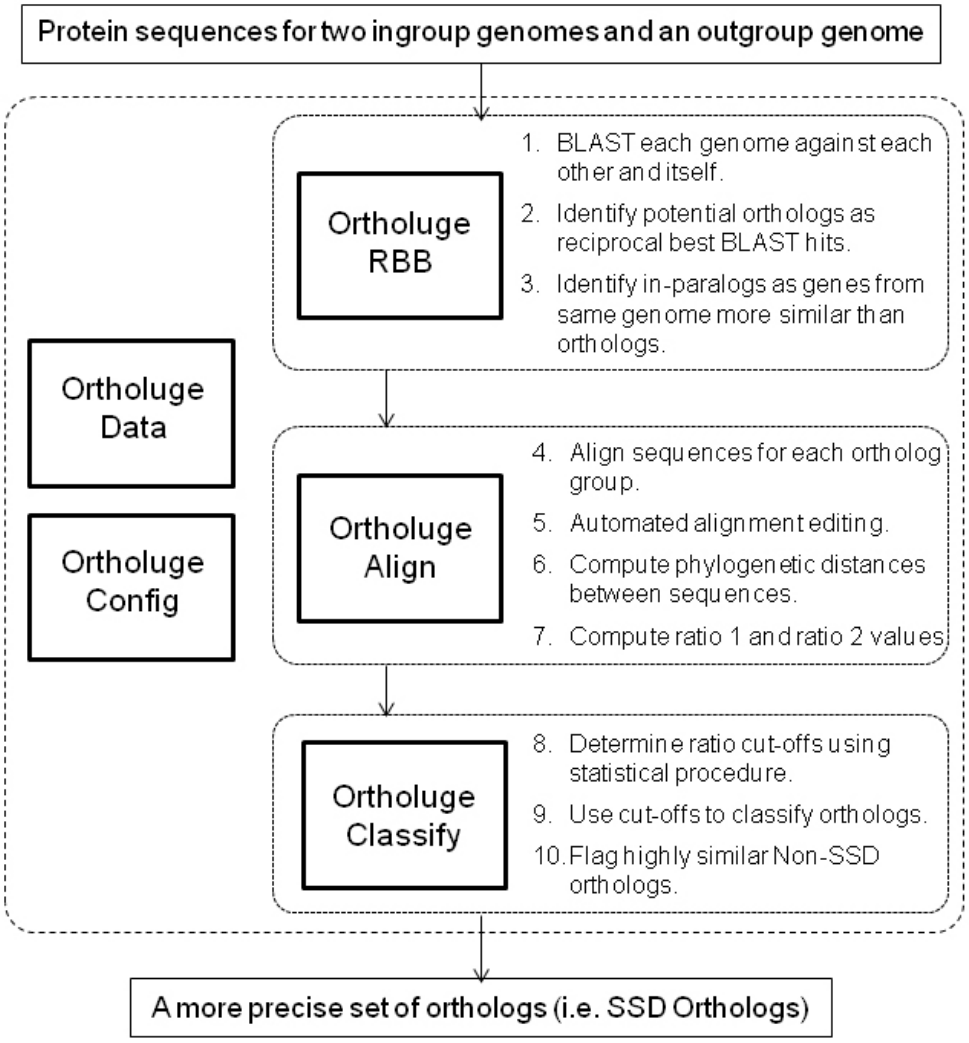


Figure 4.1 The Design of the Ortholuge Pipeline

A diagram of the Ortholuge pipeline highlighting the modules and analysis steps carried out by the modules (numbered items). OrthologeData and OrthologeConfig are high-level modules used by the main analysis modules for data loading, output and access.

4.2.2. Consolidating and Reformatting Ortholuge's Output

The Perl scripts in the original version of Ortholuge pipeline each produced separate tab-delimited text files as output. Each script in the pipeline had to have methods to load the data from the previous analysis step, as well as methods to output the results from its own analysis (Table 4.2). In the new version, the output from all

analysis steps has been consolidated into a single e**x**tensible **m**arkup **l**anguage (XML) file (Bray *et al.* 2008). Loading and outputting methods have been moved into a single computational module called OrtholugeData. The individual analysis modules now delegate loading and outputting to this data module.

Table 4.2 Outputs Produced by the Ortholuge Pipeline

Output	Description
Pair-wise RBBs	Potential orthologs are identified using reciprocal-best-BLAST between ingroup comparison genomes and between the ingroup and outgroup genomes.
In-paralogs	In-paralogs are identified in the two ingroup comparison genomes.
Triplet RBBs	Using the pair-wise RBB data, triplet RBBs are identified between the ingroup species and the outgroup species. Triplets are used in the Ortholuge assessments.
Alignments	Each triplet undergoes sequence alignment and automated alignment editing.
Phylogenetic Distances & Ortholuge Ratios	Phylogenetic distances are computed for the aligned triplets. The distances form the Ortholuge ratios.
Ortholuge Classifications	Orthologs are classified as SSD, Non-SSD, etc. on the basis of their ratio values.

The Ortholuge XML output file format consolidates all data produced by the modules in the Ortholuge pipeline. The Ortholuge XML specification is designed to store data that is added in a step-wise fashion. A series of XML tags indicates the analysis step that the file's contents correspond to. The Ortholuge program's data loading module interprets these analysis tags and the Ortholuge program can pick up from any point in the analysis pipeline.

The most significant advantage of XML over a flat text file is that the implicit hierarchical encoding of XML is a more natural representation of orthology data. Using the evolutionary definition of orthologs, multiple genes can be orthologous when the orthologous genes duplicated subsequent to speciation (i.e. in-paralogs). In Ortholuge, in-paralogs are identified and recorded. We define an ortholog cluster to be all genes that diverged from a common gene found in the last common ancestor of the species under consideration. Typically, an ortholog cluster is a pair of genes from each of the

genomes, but can include any number of in-paralogs. A XML structure effectively encapsulates an ortholog cluster of an unspecified number of genes. All related genes and their associated Ortholuge data are wrapped in a set of `<cluster>` XML tags in the Ortholuge XML specification (Figure 4.2). The flexibility of XML allows the different types of data produced by the analysis modules (i.e. alignments, phylogenetic distances etc.) to be embedded in the cluster they are associated with (Figure 4.3). In a text file, multiple entries would have to be listed on separate lines and would provide no implicit relationships between them. The switch to a single XML-based output in Ortholuge from multiple flat text files improves the interpretability and usability of the Ortholuge results.

```
<cluster id="1">
  <orthologs>
    <ortholog_group>
      <ing1_ortholog>ENSDARG00000042252</ing1_ortholog>
      <ing2_ortholog>CA063314</ing2_ortholog>
    </ortholog_group>
  </orthologs>
  <ing1_inparalogs>
    <inparalog_group>
      <ortholog>ENSDARG00000042252</ortholog>
      <inparalog>ENSDARG00000075351</inparalog>
    </inparalog_group>
  </ing1_inparalogs>
</cluster>
```

Figure 4.2 Ortholuge XML Format for Ortholog and In-paralog Specification

Sample XML code demonstrating how orthologs and in-paralogs for a single ortholog cluster are defined the in the Ortholuge XML specification.

```
<cluster id="6">
    <orthologs>
      <ortholog_group>
        <ing1_ortholog>PA1772</ing1_ortholog>
        <ing2_ortholog>PFL_1871</ing2_ortholog>
        <outg_ortholog>ACIAD1391</outg_ortholog>
      </ortholog_group>
    </orthologs>
    <ortholuge_triplets>
      <triplet is_inp="0" status="ok">
        <ingroup1>PA1772</ingroup1>
        <ingroup2>PFL_1871</ingroup2>
        <outgroup>ACIAD1391</outgroup>
        <distance1>0.101391</distance1>
        <distance2>0.633361</distance2>
        <distance3>0.687967</distance3>
        <ratio1>0.160084059485822</ratio1>
        <ratio2>0.14737770852381</ratio2>
        <ratio3>0.920627006818641</ratio3>
        <alignment>
          <ing1_alignment>----MHYVTPDLCDAYPEL
          </ing1_alignment>
          <ing2_alignment>----MNHYLTPDLCDAYPDL
          </ing2_alignment>
          <outg_alignment>MTTTVPFVTCDLLDDHTDKD
          </outg_alignment>
          <ing1_masked_alignment>XXXXMHYVTPDLCDAYPEL
          </ing1_masked_alignment>
          <ing2_masked_alignment>XXXXNHYLTPDLCDAYPDL
          </ing2_masked_alignment>
          <outg_masked_alignment>XXXXVPFVTCDLLDDHTDKD
          </outg_masked_alignment>
        </alignment>
        <class1>0</class1>
        <locfdr1>0</locfdr1>
        <flag1>0</flag1>
        <class2>0</class2>
        <locfdr2>0</locfdr2>
        <flag2>0</flag2>
        <class>SSD</class>
        <representative>1</representative>
      </triplet>
    </ortholuge_triplets>
    <flags/>
    <average_class>SSD</average_class>
  </cluster>
```

Figure 4.3 Ortholuge XML Format for Ortholuge Data

Sample XML code demonstrating how the Ortholuge data for a single ortholog cluster is defined the in the Ortholuge XML specification.

4.2.3. Parallelization of the Ortholuge Pipeline

The number of genes or proteins in the genome or proteome has a significant impact on the run-time of Ortholuge. For analyses involving large genomes, such as the human and mouse genomes consisting of tens of thousands of genes, the run-time is in the order of multiple hours (A typical run takes 6-18 hours). The largest contributing components to the run-time (in order of degree of contribution) are the BLAST analysis (Altschul *et al.* 1997), the sequence alignment and distance calculation steps. These analysis steps use external programs, so the only avenue to improve speed is to change the way in which the external programs are called within the Ortholuge pipeline. We developed the capability in Ortholuge to break each of these analysis steps into smaller jobs and run the jobs in parallel across multiple computers. The parallelized version of Ortholuge can reduce the run-time of the BLAST analysis by approximately 1/8 (there are 8 separate BLAST jobs that need to be run and each of these can be distributed across eight computer nodes). There is no limit on the number of computers that can be used to divide the sequence alignment and phylogenetic distance calculation jobs (each ortholog requires alignment and a distance calculation. The set of all orthologs can be divided among as many computer nodes as are available). While there is additional over-head that is added from generating the input and collecting the output for multiple computer nodes, the parallelized version of Ortholuge produces a significant reduction in run-time. For the human-mouse analysis using 8 computer nodes for the BLAST analysis and 24 nodes for the subsequent alignment and distance calculation steps, there is a 5-10 hr reduction in run-time versus running the entire analysis on a single computer.

4.2.4. Improved DNA Sequence Alignments through Back-Translation

In sequence alignments, protein sequences are superior to DNA sequences for aligning homologous positions. Aligning divergent DNA sequences is more difficult for a number of reasons. DNA sequences have faster evolution rates and the degeneracy of the nucleotide codons means that synonymous mutations are not selected against. The much smaller nucleotide alphabet increases the likelihood of incorrectly aligning identical non-homologous nucleotides than in the 20 amino acid alphabet of protein sequences,

and also consideration of physical-chemical properties of the amino acids can aid in the proper alignment of divergent positions (Bininda-Emonds 2005).

For situations where the DNA and protein sequences are both available, we developed and incorporated an alignment method into Ortholuge that first aligns the protein sequences and then uses the protein sequence alignments to guide the DNA sequence alignments (this general procedure is referred to as back-translation. Figure 4.4 is an example of the alignment produced by the back-translation method in Ortholuge). A back-translation approach helps increase the likelihood of correctly aligning homologous positions in the DNA sequences. It also ensures that gaps that break the open-reading frame are not introduced into the DNA alignments. The back-translation alignment method is provided as an option in Ortholuge. This improved DNA alignment method facilitates the comparison of protein- and DNA-based Ortholuge results.

```
protein1   M..A..D..K ..P..D..M. .G..E..I.. A..S..F..D
dna1       ATGGCAGACA AACCAGACAT GGGGGAAATC GCCAGCTTCG

protein2   M..A..D..K ..P..D..L. .G..E..I.. N..S..F..D
dna2       ATGGCAGACA AGCCCGACTT GGGGGAAATC AACAGCTTCG
           ********** * ** *** * **********   ********

protein1   ..K..A..K. .L..K..K.. T..E..T..Q ..E..K..N.
dna1       ATAAGGCCAA GCTGAAGAAA ACGGAGACGC AGGAGAAGAA

protein2   ..K..A..K. .L..K..K.. T..E..T..Q ..E..K..N.
dna2       ATAAGGCCAA GCTGAAGAAG ACTGAGACGC AGGAGAAGAA
           ********** *********  ** ******* **********

protein1   .T..L..P.. T..K..E..T ..I..E..Q. .E..K..R..
dna1       CACCCTGCCG ACCAAAGAGA CCATTGAGCA GGAGAAGCGG

protein2   .T..L..P.. T..K..E..T ..I..E..Q. .E..K..Q..
dna2       CACCCTGCCG ACCAAAGAGA CCATTGAGCA GGAGAAGCAA
           ********** ********** ********** ********

protein1   S..E..I..S ..
dna1       AGTGAAATTT CC

protein2   A..K..-..- ..
dna2       GCAAAG---- --
```

Figure 4.4 DNA Sequence Alignment using Back-Translation

Alignment of the DNA sequences of the thymosin β_{10} genes in human (dna1) and cattle (dna2) using the back-translation alignment method in Ortholuge. The corresponding protein sequences that were used to guide the DNA alignment are shown above the DNA sequence (protein1 is the human protein sequence and protein2 is the cow sequence). The "*" character in the match line indicates identical DNA nucleotides at that position.

4.3. Detection of In-paralogs

One of the limitations of the RBB approach is that it is unable to capture the broader phylogenetic context of the predicted orthologs. Outside of identifying the single best reciprocal BLAST hits as a putative orthologs, it cannot detect any changes in evolutionary divergence rates or recent gene duplications occurring in the ortholog genes' lineage (Altschul *et al.* 1997). Ortholuge helps address one of these limitations by detecting large-scale changes in evolutionary divergence rates (Fulton *et al.* 2006). Likewise, a module that detects gene duplications would also complement RBB-based ortholog prediction.

Empirical evidence and theoretical models suggest that gene duplication events are frequently associated with divergence in gene function (Sonnhammer and Koonin 2002; Koonin 2005). However, determining the impact on the emergent genes' functions after gene duplication is not straight-forward. Possibilities include (i) multiple duplicated genes with functions identical to ancestral gene function (such as when multiple gene copies are maintained due to gene dosage effects), (ii) one of the duplicated genes preserves the ancestral gene function, while the other copies evolve distinct novel functions (neofunctionalization) or (iii) the ancestral gene functions are divided amoung the duplicated copies and possibly undergo further specialization (subfunctionalization) (Remm, Storm, and Sonnhammer 2001; Sonnhammer and Koonin 2002; Koonin 2005). Following speciation, a duplication of the ortholog can occur in one or both species. These duplicated genes are by definition co-orthologs, as all genes are derived from a single common ancestral gene in the last common ancestor of the species. Duplications that occurred in the ancestor species prior to speciation do not form orthologous relationships. To distinguish between these sub-types of paralogs (duplicated genes), the terms in-paralog is used for genes that derive from a lineage-specific duplication occurring after speciation and out-paralog is used for genes that derive from an ancient duplication occurring before speciation (Sonnhammer and Koonin 2002).

Existence of an in-paralog may indicate a divergence from the ancestral gene function in the related orthologs (i.e. such as through subfunctionalization). Methods that predict the level of functional similarity of in-paralogs, or orthologs following a lineage-specific duplication have not been developed. Methods that explicitly detect in-paralogs, however are available. The approach adopted by several orthology resources for managing in-paralog relationships is to simply identify and flag in-paralogs and then let the user determine how to incorporate orthologs with in-paralogs into downstream analyses. By providing in-paralog predictions, the user is given a more complete phylogenetic context for a set of predicted ortholog and can make a more informed estimation of a predicted orthologs' functional similarity.

A module for the detection of in-paralogs has been added to the Ortholuge pipeline. This module is based on the Inparanoid approach for in-paralog detection (Remm, Storm, and Sonnhammer 2001; O'Brien, Remm, and Sonnhammer 2005; Ostlund *et al.* 2010). The Inparanoid method uses BLAST similarity relationships and

does not require phylogenetic analysis. It therefore easily integrates with the existing RBB ortholog prediction method used in the Ortholuge pipeline. The formative assumption in the Inparanoid approach is that an in-paralog will be more similar to the ortholog gene from the same species, than any gene in the other species (the rationale is that since the in-paralog diverged after the species diverged, it should be more similar than the orthologs genes which diverged when the species diverged) (O'Brien, Remm, and Sonnhammer 2005). Computationally, this situation is identified by finding genes with a BLAST bit score that is greater than BLAST bit score between the orthologs (see Figure 4.5 for further explanation). In the Ortholuge version, we added two detection strategies that are slight modifications of the described Inparanoid method. The modified detection strategies use additional criteria that make them more conservative (more often correctly predicting in-paralogs at the expense of missing valid in-paralogs). The criteria examine the in-paralog relationship with the ortholog gene in the other species and help ensure that the predicted in-paralog and ortholog genes diverged from a common ancestral gene in the last common ancestor of the species. The variations are outlined in Table 4.3. Ortholuge can return in-paralog predictions computed using one of these detection strategies. Ortholuge ratios are also computed for the in-paralogs paired with orthologous genes in other species, so that the level of divergence of the in-paralogs can be compared to the main predicted orthologs.

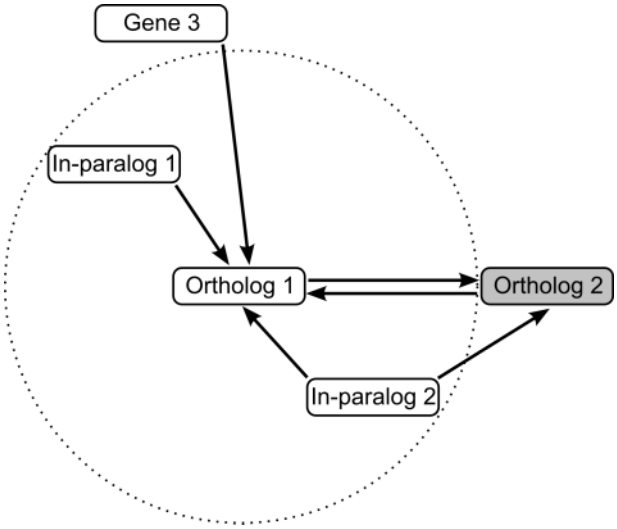


Figure 4.5 In-paralog Detection Method

The in-paralog detection method used in Ortholuge. Three variations of the detection method are provided in Ortholuge. The variations are described in detail in Table 4.3. Briefly, a gene is declared an in-paralog if the intra-species sequence similarity measured by BLAST between the gene and the ortholog is greater than the similarity between the inter-species orthologs. In the figure, arrows represent BLAST hits where the length represents sequence distance (the dashed circle represents the range of the inter-species distance between the orthologs).

Table 4.3 In-paralog Detection Strategies Offered in Ortholuge

In-paralog Detection Strategy	Description	Qualifying Genes in Figure 4.5
Inclusive	In-paralogs are genes that have a BLAST bit score to the ortholog gene from the same species that is greater the inter-species score between the RBB-predicted orthologs (this is the standard Inparanoid approach)	In-paralog 1 & 2
Exclusive	All in-paralogs must satisfy the ***inclusive*** criteria plus must have a BLAST hit to the ortholog gene in the other species. This criterion ensures that there is a least some level of similarity between the in-paralog gene and ortholog gene from the other species, suggesting they diverged from a common ancestor gene.	In-paralog 2

In-paralog Detection Strategy	Description	Qualifying Genes in Figure 4.5
Reciprocal	All in-paralogs must satisfy the ***inclusive*** criteria plus the in-paralogs top BLAST hit in the other species genome is the ortholog gene. This strict criterion ensures the in-paralog and all orthologs diverged from a common ancestor gene. This approach can miss cases where there are multiple in-paralogs in both species and the top BLAST hit of the in-paralog is one of these other in-paralogs and not the main ortholog.	In-paralog 2 (provided BLAST hit to Ortholog 2 is gene's top hit)

4.4. Statistical Computation of the Ortholuge Ratio Cut-offs

The Ortholuge phylogenetic ratio cut-offs are critical values that can have significant impact on Ortholuge's results. Ortholuge classifies predicted orthologs as SSD or Non-SSD based on the chosen cut-offs for ratio 1 and ratio 2. In the initial version of Ortholuge pipeline as reported by Fulton *et al*. 2006, the ratio cut-offs were determined by modelling ortholog prediction on an incomplete genome. This approach attempts to determine the expected proportion and ratio values of falsely predicted orthologs that would be produced by predicting orthologs on an incomplete genome (where a number of the true orthologs are missing) (Fulton *et al.* 2006). To simulate ortholog prediction on an incomplete genome, after initially predicting orthologs by RBB, the predicted orthologs in one of the genomes are removed and the ortholog prediction step is repeated with the reduced genome dataset lacking the predicted ortholog genes. Any RBB relationships detected in the second round are considered falsely-predicted orthologs or true-negatives. The ratio values are computed for the true-negatives and the true-negative ratio distribution is compared to the ratio distribution of original set of predicted orthologs. The overlap between the true-negatives and original predicted orthologs ratio ranges is examined by binning the ratios into equal interval ranges and comparing the proportion of true-negatives to the original ortholog dataset in each bin's interval. A lower and upper cut-off is selected as the smallest ratio value corresponding to the bins that contain 10% and 50% true-negative orthologs, respectively. Below the 10% cut-off, orthologs are classified as SSD orthologs, between 10-50% cut-offs are classified as uncertain and ortholog ratios that fall above the upper 50% cut-off are

classified as Non-SSD (Fulton *et al.* 2006). Figure 4.6 shows an example of the distribution of true-negative ratio values overlaid on the original dataset ratio values with the derived cut-off values.

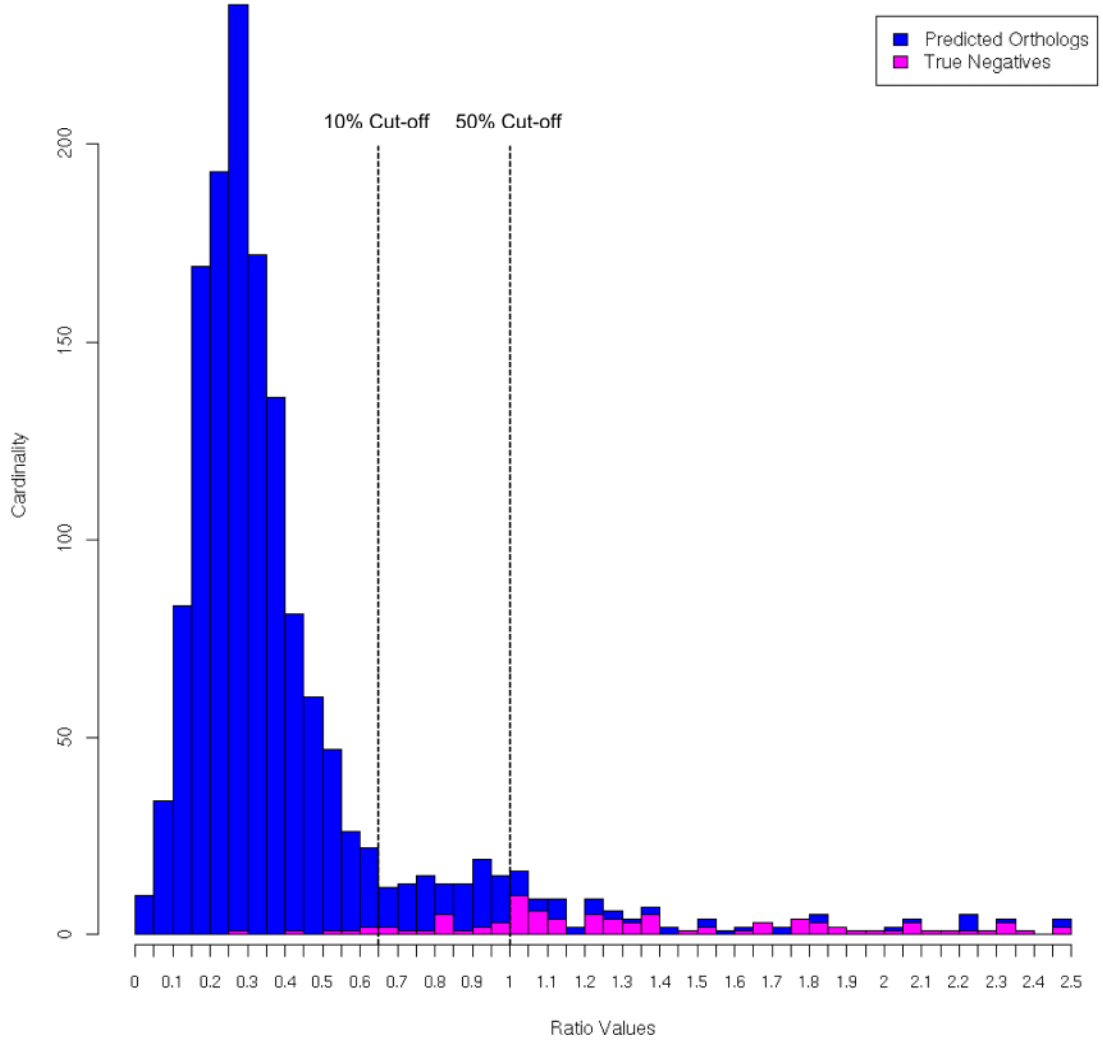


Figure 4.6 Histogram of Ratio values for Predicted Orthologs and True-Negatives

Histogram of ratio 1 values for predicted orthologs and true-negative gene pairs for species *Pseudomonas fluorescens* SBW25 and *Pseudomonas mendocina* YMP. Ratio cut-offs are determined by the proportion of true-negatives in each of the bins' interval ranges. The interval containing 10% true-negatives is selected as the lower cut-off and is used for the classification of SSD orthologs. The ratio value corresponding to the interval containing 50% true-negatives is selected as the upper cut-off. This cut-off is used for the classification of Non-SSD orthologs. Orthologs with ratios that fall in the region between the cut-offs are classified as uncertain.

This approach of modelling ortholog prediction on an incomplete genome by introducing true-negatives has several shortcomings. The step of identifying true-negatives and computing their ratio values is computationally intensive, as it requires running the Ortholuge method on a second set of genes. Additionally, this method for computing ratio cut-offs is not statistically robust. The cut-offs can be affected by the proportion of falsely-predicted orthologs (i.e. paralogs) in the original dataset as well as the availability of paralogs in the genome under investigation that can form useable true-negatives (increased false predictions in the original dataset will reduce the true-negative proportion in the intervals and cause the cut-offs to be higher than the optimal value. Similarly, lower numbers of true-negatives due to a lack of available paralogs in the genome also reduces the true-negative proportions and causes a shift in the cut-off values).

With Dr. Jinko Graham, Dr. Brad McNeney and Masters student Jeong Eun Min from the Simon Fraser University's Statistics Department, we developed a direct statistical approach for computing ratio cut-offs. This approach uses large-scale hypothesis testing methods originally developed by Efron *et al.* 2004 to directly infer the ratio distributions of the SSD orthologs and Non-SSD gene pairs. From the inferred distributions, a local false discovery rate (local fdr) can be calculated (Efron 2004). This rate gives the expected proportion of Non-SSDs for a given ratio value. Consistent cut-offs between analyses are obtained as the ratio values with equivalent local false discovery rates. The key statistical assumption, which draws on large-scale aspect of the dataset, is that SSD orthologs make up the majority of the predicted orthologs and they form the major distribution in the overall distribution of ratios values for the predicted orthologs (Min *et al.* 2011). By fitting the statistical model over a specified region where SSD orthologs make up the significant majority of the ratio values, the distribution of the SSD orthologs or "true-positives" can be estimated (an example of the estimated SSD ratio distribution is shown in Figure 4.7 B. The region used to perform the fitting is shown as the dark line at the bottom). The overall ratio distribution, which is a mixture of SSD orthologs and falsely-predicted orthologs, is then estimated (Figure 4.7 A) and the density of the SSD distribution is compared to the density of this mixture distribution to calculate the expected proportion of Non-SSD orthologs for a given ratio value (i.e. local fdr) (Min *et al.* 2011). Figure 4.7 C shows the expected proportions of

Non-SSDs as the grey bars overlaid on the histogram bars. The method was conceived and developed by Min *et al.* 2011. Validation was also performed by Min *et al.* 2011. It was adapted for use in Ortholuge pipeline by me. I also generated validation ortholog and paralog datasets for method testing.

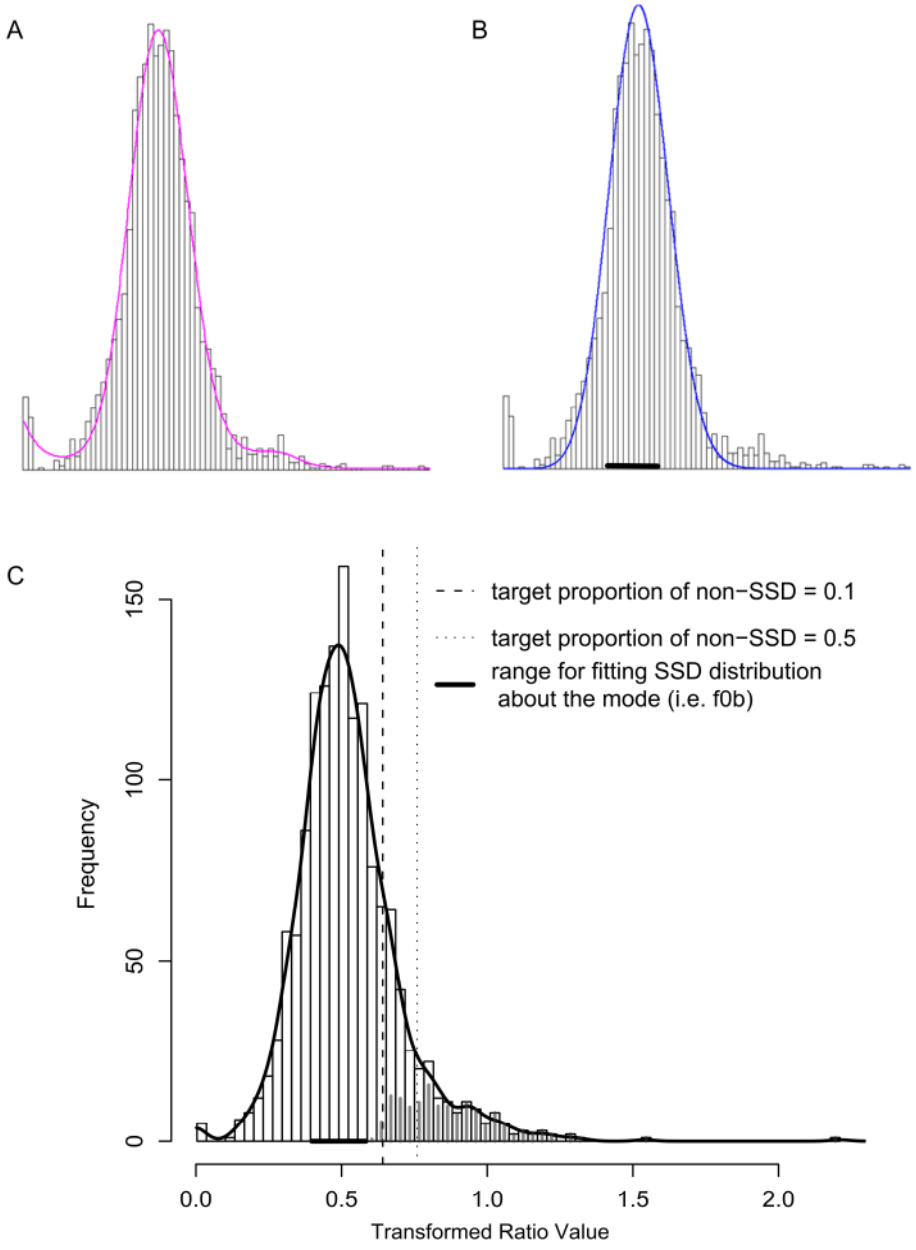


Figure 4.7 Ortholuge local fdr Approach for Calculating Ratio Cut-offs

The local fdr approach for calculating the Ortholuge ratio cut-offs estimates the SSD distribution by fitting a statistical model over a region where it is expected that SSD orthologs make up the majority of the density and there is minimal contamination of falsely-predicted orthologs (B). This estimated distribution is compared to the estimated distribution for the all predicted orthologs (A), which contains a mixture of SSD and Non-SSD orthologs and the expected proportion of Non-SSD or local false discovery rate (local fdr) is derived from these two distributions (the grey bars in the histogram in panel C show the expected proportion of Non-SSDs). Ratio cut-offs are based on pre-selected local fdr values (Min *et al.* 2011).

This statistical local fdr approach provides a significant improvement in the performance of Ortholuge over the previous approach utilizing true-negatives for computing ratio cut-offs. Because it does not require running Ortholuge on a second dataset of true-negatives, the local fdr approach significantly reduces the run-time of the Ortholuge. The local fdr approach also has improved sensitivity over the previous approach, correctly identifying more paralogs as Non-SSD in evaluations involving both real and simulated datasets (Figure 4.8) (Min *et al.* 2011). The local fdr method explicitly accounts for the unknown proportion of falsely-predicted orthologs in the overall ratio distribution, while the true-negative approach does not robustly handle false-positives (it assumes that the proportion of falsely-predicted orthologs is equivalent to the proportion of true-negatives that can be created in any arbitrary dataset). This deficiency means that as the proportion of false-positives increases, the sensitivity of the true-negative-based approach decreases. The local fdr-based approach has consistent performance across different levels of contamination of false-positives (compare the sensitivity of the local fdr and true-negative approaches when the proportion of false orthologs is 5% versus 25% in Figure 4.8). Because of its significant benefits, the statistical-based local fdr approach for computing ratio cut-offs has been incorporated into Ortholuge.

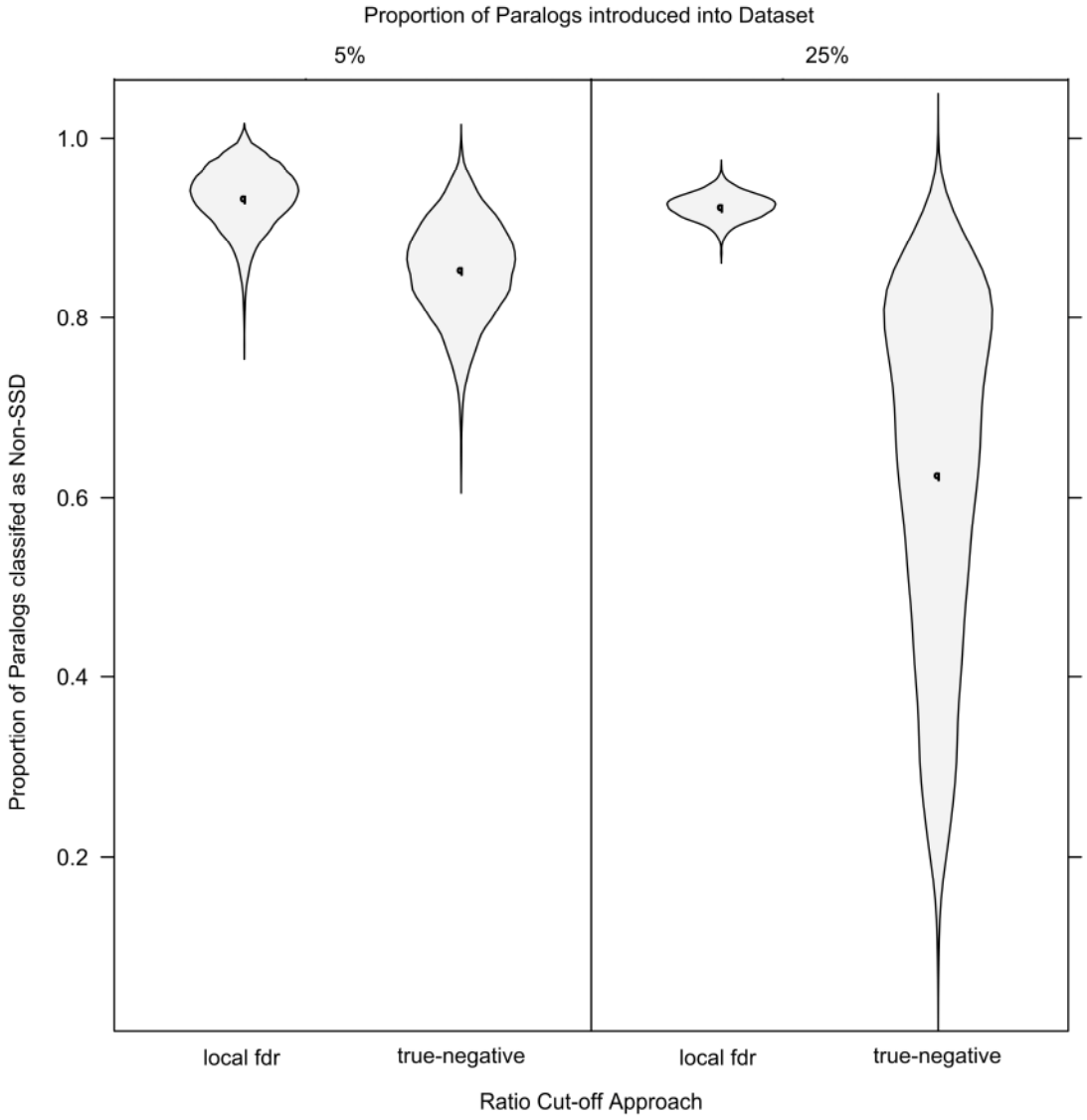


Figure 4.8 Evaluation of the local fdr and True-Negative Approaches for Computing Ratio Cut-offs

Violin plots showing the distribution of the results achieved by each ratio cut-off method for 5000 simulated ratio datasets. To evaluate the sensitivity of the local fdr and true-negative cut-off approaches, 5000 simulated ortholog ratio datasets were generated from the ratios of known valid orthologs and paralogs. The evaluations examined two levels of paralog contamination consisting of 5% and 25% of the dataset. For each the datasets, the proportion of paralogs correctly identified out of the total number of paralogs was measured (i.e. sensitivity) (Min *et al.* 2011).

4.5. Sub-classification of Non-SSD Predicted Orthologs

The phylogenetic distance ratios in Ortholuge compare the orthologs distance in relation to the distance of an outgroup ortholog in the ortholog gene phylogenetic tree. Predicted orthologs with large phylogenetic distances ratios compared to other predicted orthologs in the genome are labelled as Non-SSD (Non-SSD indicates the level of divergence between the orthologs is not consistent with the species level of divergence) (Fulton *et al.* 2006). The ratio formulation does not consider the magnitude of the phylogenetic distance, only the branch length proportions. As a result, cases can arise that are classified as Non-SSD, but at the sequence level are highly similar.

We feel that cases that have high sequence similarity but are diverging unusually are still noteworthy and should be flagged (in these cases the predicted ortholog phylogenetic distance, although small, is proportionally larger than expected. This may indicate they are evolving at different rates). They are not however, as extreme as ortholog cases that exhibit both large sequence divergence and large ratio values. From our analysis of gene feature conservation amoung Non-SSD orthologs, these predicted orthologs are more likely to be paralogs mistakenly predicted as orthologs or functionally distinct genes. So to distinguish between Non-SSDs that are highly similar and Non-SSDs with large sequence divergence, we developed a method that further sub-classifies the Non-SSD orthologs.

The approach examines the phylogenetic distances that make up the numerator and denominator in Ortholuge ratios as a two-dimensional plot (Figure 4.9). The expected level of divergence between the ingroup orthologs is computed as a linear function of the distance between the ingroup and outgroup orthologs. We use the robust procedure Tukey's bi-square method to estimate this linear trend in order to avoid undue influence from falsely-predicted ortholog ratios. Predicted orthologs that deviate significantly from the expected level of divergence are detected by measuring the distance from the trend-line. Non-SSD orthologs that have large ratio values and large phylogenetic distances will appear far from the trend-line, while Non-SSD orthologs that have relatively high sequence similarity will be close the trend-line. A cut-off separating the divergent Non-SSD and similar Non-SSD orthologs is defined as a specific distance from the trend-line (divergent Non-SSD refers to Non-SSD orthologs with large ratios

and phylogenetic distances. Similar Non-SSD refers to Non-SSD orthologs that are highly similar at the sequence level).

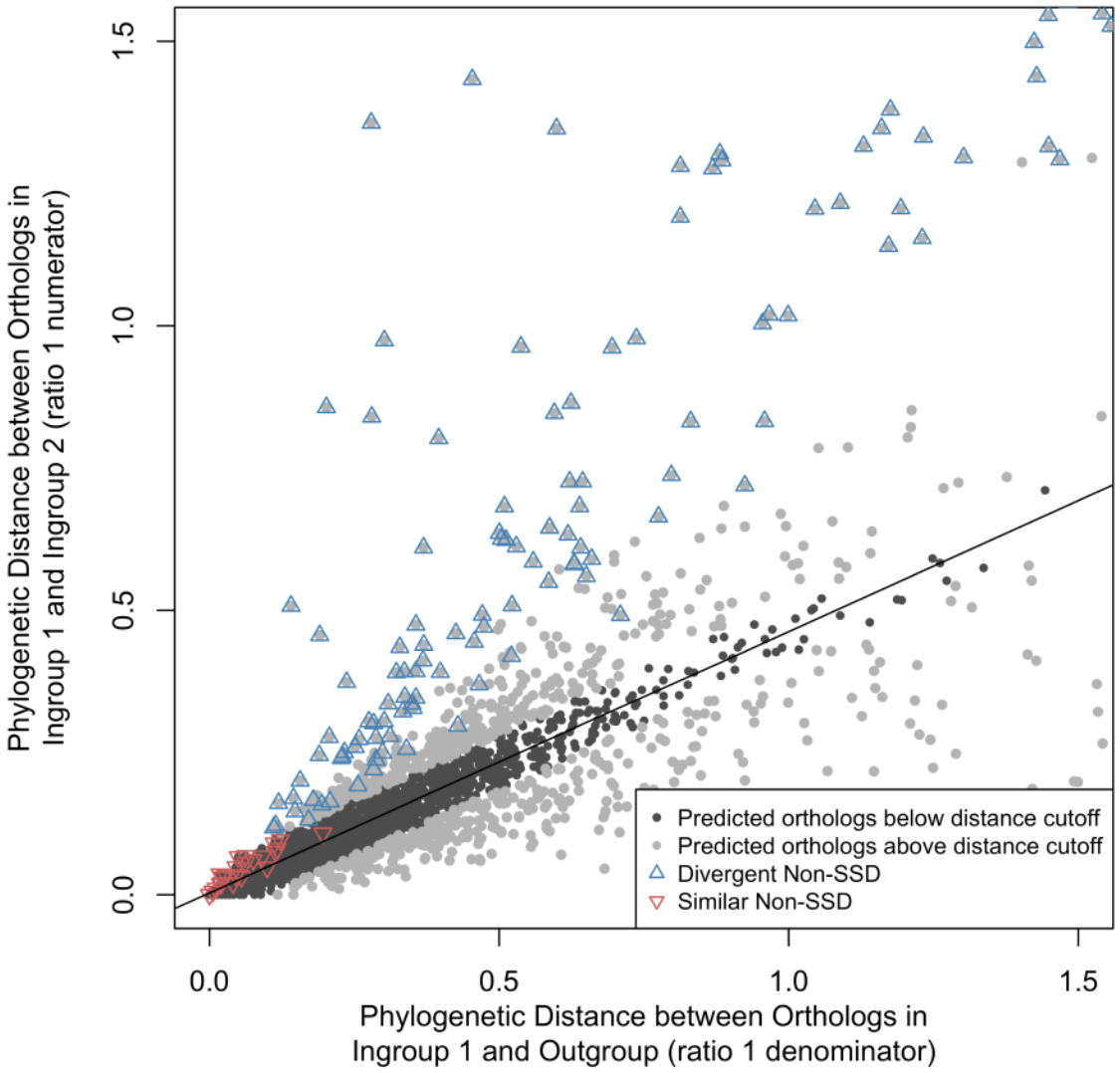


Figure 4.9 Sub-classification of Non-SSD Predicted Orthologs

Sub-classification of Non-SSD predicted orthologs in the species *Klebsiella pneumonia* 342 and *Escherichia coli* O55:H7 str. CB9615. Classification of divergent and similar Non-SSD predicted orthologs is performed by examining the phylogenetic distances that make up the Ortholuge ratios. The numerator and denominator distance values that make up the ratio are plotted (ratio 1 values are shown) and a trend-line representing the expected level of phylogenetic divergence for a given ingroup-outgroup distance value is calculated (black line). Predicted orthologs that deviate significantly from the expected level of divergence are detected by measuring the distance from the trend-line. Non-SSD predicted orthologs which are below the distance cut-off from the trend-line are classified as similar Non-SSD. The remaining Non-SSD orthologs are classified as divergent.

The main advantage of this approach, compared to a basic phylogenetic distance cut-off for separating similar and divergent Non-SSDs, is that it accounts for the level of species divergence. The phylogenetic distance between equivalent orthologs in multiple species will change depending on the evolutionary distance between the species. A single phylogenetic distance cut-off would not produce equivalent Non-SSD sub-classifications for species of varying evolutionary distances. The new approach measures the difference of actual ortholog divergence from the expected level of divergence for the species (i.e. the trend-line), and hence produces more consistent classifications.

4.6. Conclusions

The modifications and additions to the Ortholuge software produced a tool that is more efficient, more accurate and has increased functionality to aid comparative genomic analysis. Modifications to the software have stream-lined and sped up the Ortholuge pipeline by reducing code redundancy, consolidating and formalizing outputs and providing a parallelized run-mode. Accuracy has been improved by developing a new ratio cut-off procedure, creating a method to sub-classify Non-SSD orthologs and adding additional options for DNA alignment. The functionality in Ortholuge has been augmented by adding a module for in-paralog detection. These improvements contribute to the development of an accurate, high-throughput tool for ortholog identification. Ortholuge has been designed for comparative genomics analysis and should provide an effective resource for identifying functionally-equivalent genes across species.

5. Building a Database of Ortholuge Results for Bacterial and Archaeal Species

Portions of this chapter have been previously published in the article "OrthologeDB: a bacterial and archaeal orthology resource for improved comparative genomic analysis", co-authored by M.D. Whiteside, G.L. Winsor, M.R. Laird, and F.S.L. Brinkman in Nucleic Acids Research, Volume 41, Issue D1, Pages D366-76 © 2013 Whiteside et al; licensee Oxford University Press.

5.1. Introduction

Ortholuge was developed for high-throughput and accurate ortholog prediction. It is a graph-based method coupled with a phylogenetic analysis and so includes the benefits of both approaches. It is scalable, requires only gene sequences as input (there is no requirement for accurate species trees) and also has improved accuracy, because it takes into account the broader phylogenetic context of the predicted orthologs. However, the Ortholuge method, initially made principally for our own internal use, has not been easily accessible. Ortholuge-based predictions for the human, mouse and cow genomes are available through InnateDB - a platform facilitating system-based analyses of the innate immune response (Lynn *et al.* 2008) – and predictions for *Pseudomonas* species genomes are available through the *Pseudomonas* Genome Database (Winsor *et al.* 2010). However, none of these predictions are queryable and there is no flexibility in the display of ortholog prediction results. To make Ortholuge predictions widely available, we developed OrthologeDB, a web-accessible database of pre-computed Ortholuge-based predictions for all available sequenced bacterial and archaeal species.

Bacterial and archaeal orthology resources currently fall into two categories. First, several platforms for comparative genomic analysis in microbial species have been developed, including the Comprehensive Microbial Resource (Davidsen *et al.* 2010), MicrobesOnline (Dehal *et al.* 2010), Microbial Genome Database (Uchiyama, Higuchi,

and Kawai 2010) and the Integrated Microbial Genomes database (Markowitz *et al.* 2012). Broadly, they provide data integration and functional inference based on comparative genomics. These platforms could benefit from an ortholog prediction method that is both high-throughput and highly precise. The second type of resource is large-scale orthology databases. There are more than 30 ortholog databases, but currently only OMA (Altenhoff *et al.* 2011), QuartetS (C. Yu *et al.* 2011; Yu *et al.* 2012), OrthoMCL (Li, Stoeckert, and Roos 2003; Chen *et al.* 2006), RoundUP (DeLuca *et al.* 2012), eggNOG (Powell *et al.* 2012b), HOGENOM (Dufayard *et al.* 2005; Penel *et al.* 2009) provide ortholog predictions for significant numbers of microbial genomes. None of these databases are specific to bacteria and archaea. Their primary focus is on serving large-scale ortholog queries. OrtholugeDB was developed to fulfill a need for a comprehensive database of bacterial and archaeal orthologs that supports flexible, powerful ortholog queries with a focus on microbial comparative genomics.

OrtholugeDB was developed with the assistance of Matthew Laird and Geoffrey Winsor from the Brinkman Laboratory at Simon Fraser University. Computational scripts for the parallelization of Ortholuge were developed and run by Matthew Laird. Matthew Laird also assisted with the initial MySQL database design. Geoffrey Winsor contributed significantly to the webpage layout design and to the underlying web application development. I developed all computational methods for generating and loading the database content, as well as contributed to webpage layout design and the web application development.

5.2. Content and Design

OrtholugeDB is a comprehensive database of orthologs for bacteria and archaea. This database, which will be regularly updated, provides Ortholuge-based ortholog predictions for all completely sequenced bacterial and archaeal genomes where a suitable outgroup is available. For those gene pairs where no suitable outgroup yet exists, RBB-based predictions are still presented. Ortholog predictions are available for protein-coding genes only (though predictions using DNA gene sequences are possible with the Ortholuge method (Fulton *et al.* 2006)). Data is stored in a MySQL database.

5.2.1. Content

OrtholugeDB is based on the completed genomes of bacterial and archaeal species from NCBI (incompletely sequenced genomes are not included in the database). The protein sequences for the bacterial and archaeal species were obtained using the MicrobeDB resource; a tool that provides a locally maintained database of sequenced microbial genomes from NCBI (Langille *et al.* 2012; Pruitt *et al.* 2012). In OrtholugeDB, the RBB procedure is used to generate the initial set of ortholog predictions (Altschul *et al.* 1997). Multiple RBBs are possible and we keep track of all of them, as these cases often represent very recent gene duplications (in-paralogs). We evaluate the RBB-predicted orthologs using Ortholuge, and classify them as follows:

1. Supporting species divergence (SSD):

   Predicted orthologs whose divergence (as reported by the Ortholuge phylogenetic ratios) is consistent with the divergence observed for the species. These predicted orthologs most likely represent valid orthologs, and have not undergone unusual divergence (such as accelerated evolution).

2. Borderline-SSD:

   Predicted orthologs with a phylogenetic ratio that is slightly higher than expected. When precision is critical to an application, these predicted orthologs can be excluded.

3. Divergent Non-SSD:

   Non-SSD genes have phylogenetic ratios that are significantly higher when compared to most other orthologs in the genomes (as per our statistical analysis; (Min *et al.* 2011)), indicating that their divergence is not consistent with the species level of divergence. Based on our previous simulations (Fulton *et al.* 2006), these are most likely incorrectly predicted orthologs, or orthologs that have undergone unusually rapid divergence due to a change in function.

4. Similar Non-SSD:

   Similar Non-SSDs have diverged unusually, as the length of one of the branches in the gene tree is proportionally longer than expected, however the total phylogenetic distance separating the predicted orthologs is relatively small. Many Similar Non-SSD genes will often be valid orthologs. However, the high phylogenetic ratio may suggest the genes are evolving at different rates.

Boundaries between the Ortholuge classifications are based on the local false-discovery rates described previously (Min *et al.* 2011). Ortholuge requires an outgroup

ortholog as a reference for computing the ratios (the outgroup gene is used to root the predicted ortholog's phylogenetic tree). The ideal outgroup species diverged prior to the divergence of the comparison species but also has a large number of common orthologs with the comparison species. To select the reference outgroup species, we computed the optimum phylogenetic distance for an outgroup that best separates the distributions of SSD and Non-SSD ratios for a given pair of ingroup species. The optimum distances formed the basis of a formula that we use to automatically select outgroups. Distances are computed using CVtree, a composition-based distance metric that reflects the evolutionary relatedness between species proteomes (Xu and Hao 2009). In-paralogs are genes that have duplicated subsequent to species divergence. If the genes duplicated prior to the speciation (and creation of orthologs) the genes are referred to as out-paralogs. We identify in-paralogs using a procedure based on the InParanoid method (Ostlund *et al.* 2010). Briefly, after computing all orthologs, a gene is declared an in-paralog if it is closer to an ortholog in term of BLAST score than the score between the orthologs.

5.2.2. Web Interface Design

OrtholugeDB is designed to facilitate the rapid extraction and evaluation of bacterial and archaeal orthologs. Queries are intended to address a wide range of needs, from obtaining orthologs for single genes, to orthologs for multiple genomes. OrtholugeDB includes the ability to run complex queries that filter genes based on the presence and absence of orthologs in other species (i.e. identifying genes unique to a species or set of species). The Ortholuge statuses of the predicted orthologs (SSD, Borderline-SSD, Non-SSD, etc.) are highlighted in the result pages. Results from any of the queries can be downloaded in tab-delimited, comma-delimited (CSV) and OrthoXML formats (Schmitt *et al.* 2011). The following types of queries are currently available in OrtholugeDB (the query forms are shown in Figure 5.1):

1. Orthologs between two genomes.

   This function returns all orthologs for two genomes of interest. Alternatively, you can return genes that do not have orthologs for one of the species (including in-paralogs). As an option, images showing the gene context for predicted orthologs can be generated (i.e. image displaying genes flanking the gene of interest). The gene context view is shown in Figure 5.2.

2. Orthologs for a gene.

   The orthologs for a gene of interest can also be retrieved in OrthologeDB. The orthologs can be limited to a specified set of genomes or can be obtained for all species in the database. The gene context option is also available for this query.

3. Ortholog groups for a gene of interest.

   Pre-computed ortholog groups are retrieved by providing a gene of interest and a desired hierarchical level that determines the range of species used to construct the group. Groups representing Genus, Family, Order, Class and Phylum taxonomic levels are available. To enhance viewing of the ortholog connections within the group, a graph view of the orthologs is provided. In the graph view, genes are viewed as nodes and ortholog and in-paralog relationships are represented as edges between genes (Figure 5.3). Ortholog edges are coloured based on their Orthologe status.

4. Compare the orthologs in a genome of interest across multiple other genomes.

   This query generates a high-level phyletic matrix view that quickly shows which genes in a genome of interest have orthologs in the specified comparison species (Figure 5.4). Coding in the matrix highlights ortholog cardinality and Orthologe status. Also provided as part this query, is the ability to filter genes based on the presence or absence of orthologs in other species. This feature allows users to formulate complex queries, obtaining genes, for example, that are common to one set of species (which may belong to divergent phyla but have a common phenotype), and not found in another set of species (with a different phenotype). A summary of the ortholog content is also provided for the query genome. The summary shows the proportion of protein-coding genes in the query genome that have no orthologs, one-to-one orthologous relationships or many-to-many orthologous relationships in each of the comparison species.

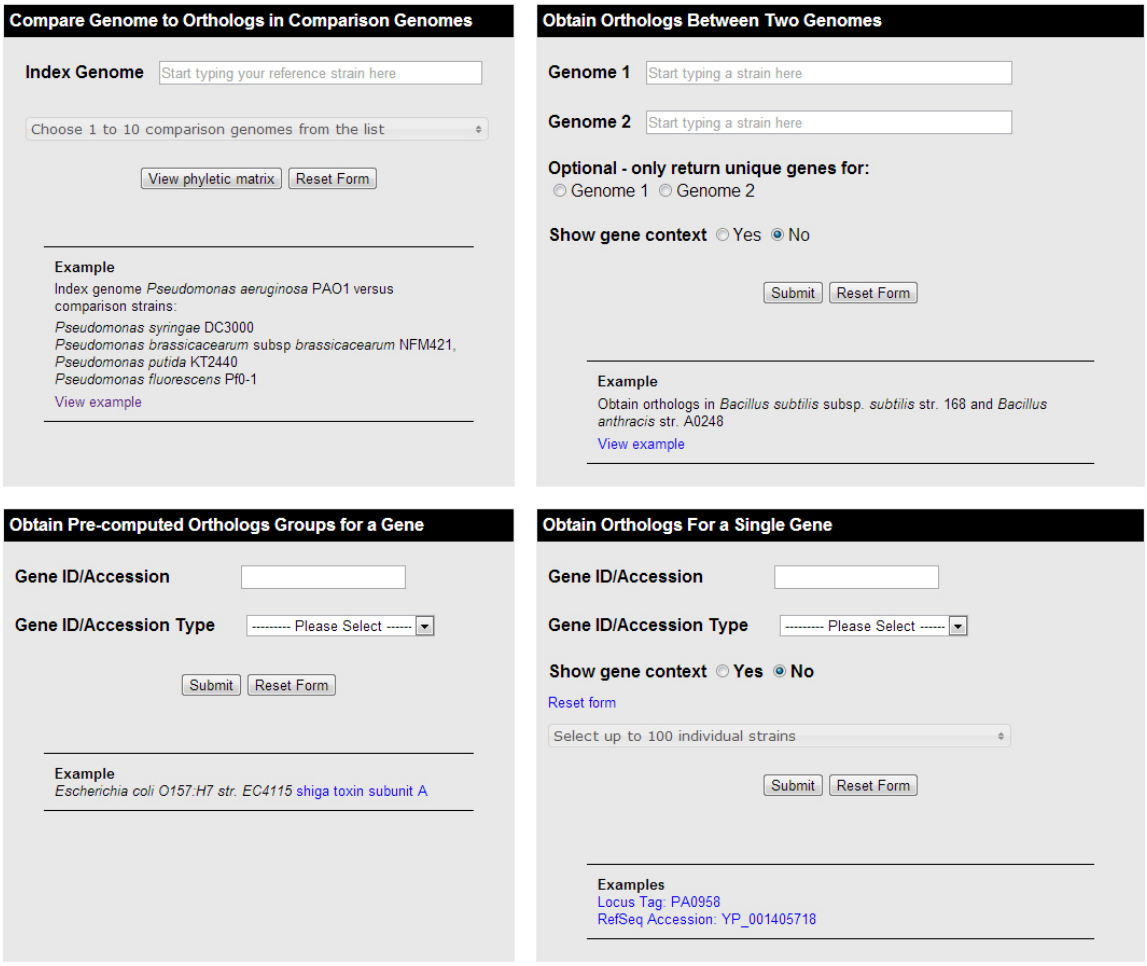


Figure 5.1 Queries Provided in OrthologeDB

The four types of queries provided in OrthologeDB.

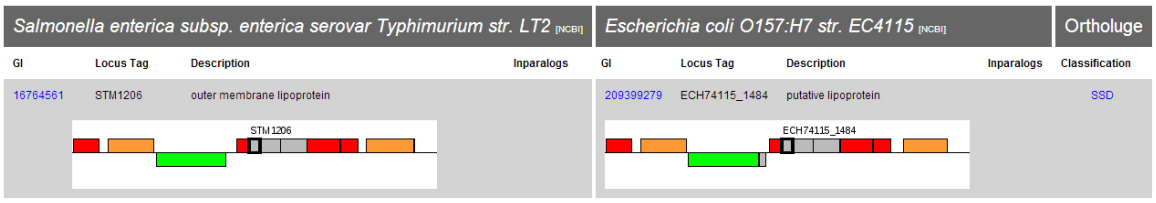


Figure 5.2 Orthologs for Two Genomes Query Result View

A single row of the results table returned for the "Orthologs for two genomes" query for the species: *Salomonella enterica* subsp. enterica servor Typhimurium str LT2 and *Escherichia coli* O157:H7 str. EC4115. The gene context option has been selected in this example. The ortholog pair is classified as SSD.

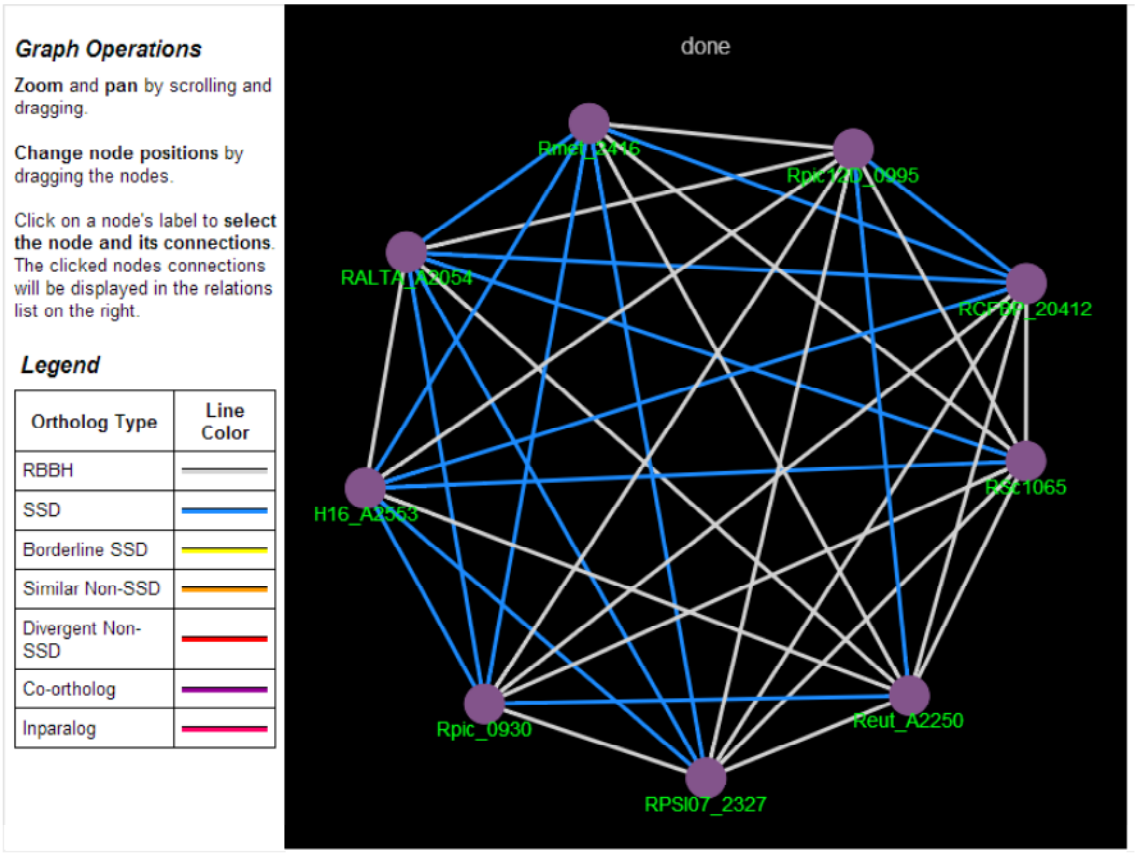


Figure 5.3 Ortholog Group Graph View

The ortholog group graph view for the DNA repair gene *recO* in the *Ralstonia* genus. Genes are represented by the circle nodes and ortholog relationships as edges between the nodes (locus IDs are shown for the genes). The edges are coloured according to the Ortholuge classification of the ortholog.

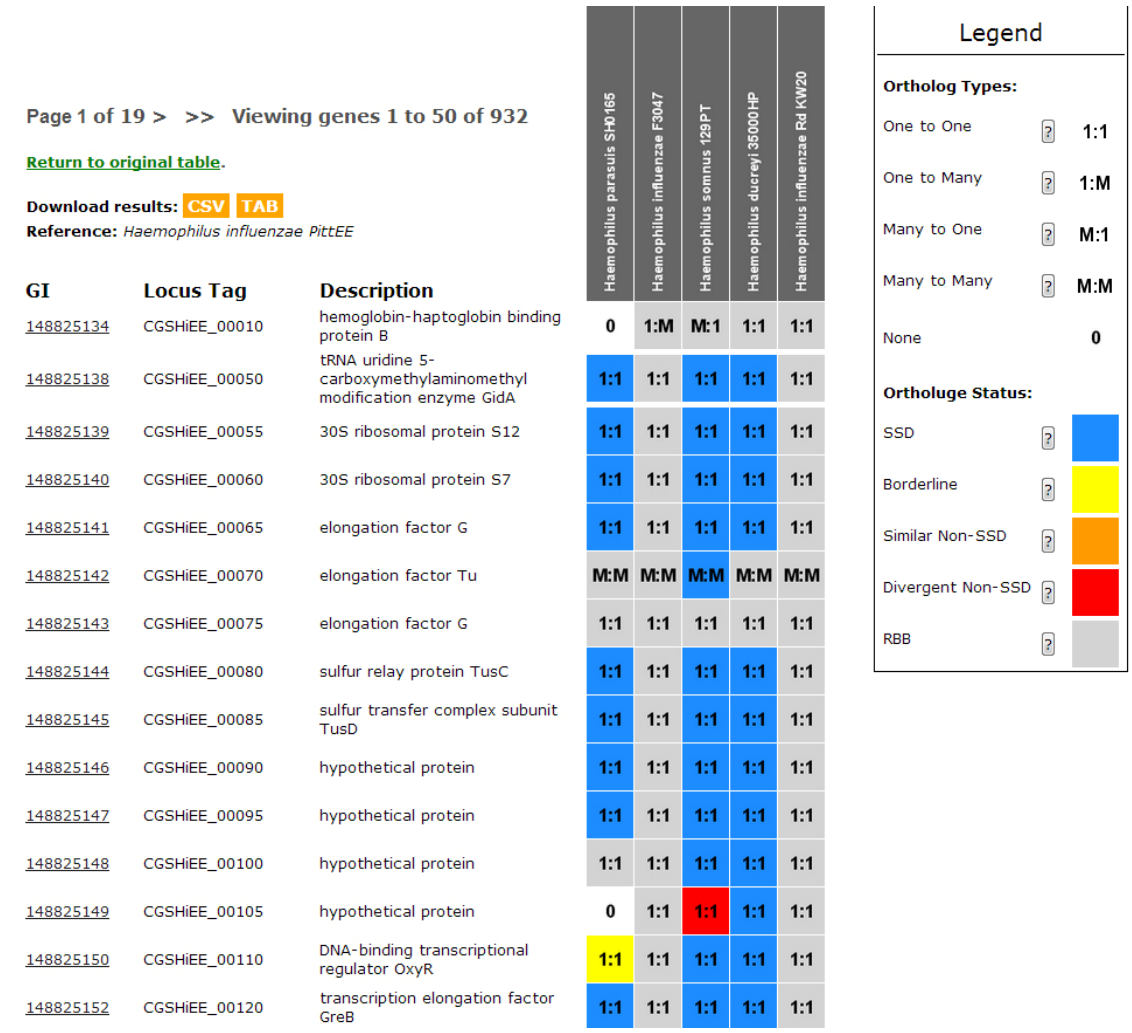


Figure 5.4 Phyletic Matrix View

A phyletic matrix view showing the orthologs for genes in *Haemophilus influenza* PittEE in five other *Haemophilus* species. Numbering in cells indicates ortholog relationships without in-paralogs/co-orthologs (1:1) as well as ortholog relationships with in-paralogs both species (M:M, M:1, 1:M) and genes that have no orthologs (0). Cells are coloured based on their Ortholuge classification – blue: supporting-species-divergence orthologs (SSD), yellow: Borderline-SSD, grey: no Ortholuge classification available, red: divergent Non-SSD.

5.3. Automating the Ortholuge Method

The scale of ortholog prediction in OrthologueDB requires a method that is fully automated and efficient. The original implementation of the Ortholuge analysis pipeline had deficiencies that limited its ortholog prediction throughput. These deficiencies included the procedure for determining the ratio cut-offs. It required running Ortholuge

analysis on a second set of true negatives gene pairs (the procedure introduced true negatives into the Ortholuge ratio distribution to determine a ratio value threshold that isolated the majority of the true negatives). This inefficiency was addressed by the development of a robust statistical procedure that could directly select a ratio cut-off from the predicted ortholog ratio distribution, removing the need for running Ortholuge analysis on the true negative set (Ortholuge modifications are described in Chapter 4).

After addressing the ratio cut-off deficiencies, the largest impediment to large-scale computation in the Ortholuge pipeline was the requirement for the manual selection of an outgroup reference species. A suitable outgroup genome is needed for each pair of comparison species. The outgroup genome roots the phylogenetic trees used in Ortholuge. Initially, selecting an outgroup species was done manually, but for OrthologeDB an automated process was developed.

5.3.1. Automated Selection of the Reference Genome

The basis of the automated method to select the outgroup reference genome is a mathematical formula that given the distance between ingroup comparison genomes, calculates the optimal outgroup genome distance from the ingroup species. Species distances are computed using CVtree, a composition vector-based distance metric that reflects the evolutionary relatedness between species proteomes (Xu and Hao 2009). It builds whole-genome phylogenetic trees without sequence alignments. This feature is crucial to the automated outgroup method because it permits the automatic reconstruction of phylogenetic species relationships and is rapid enough to handle the number of genomes in OrthologeDB. CVtree was found to successfully reproduce known phylogenetic species trees. To derive the mathematical formula, criteria were defined to describe the ratio distribution properties of an Ortholuge analysis with an ideal outgroup. The criteria are:

1. A lower quartile value that is greater than 0.
2. An upper quartile value that is less than 1.

These criteria ensure that the width of the distribution is sufficient to properly distinguish the SSD orthologs from the Non-SSDs. The selection formula defines allowable ingroup comparison genomes and optimal outgroup reference genomes.

Comparison species that are highly similar (e.g. strains of the same species) and exhibit little sequence divergence between orthologous genes make it difficult to estimate the allowable variation in the SSD ortholog ratios (Ortholuge analysis involving highly similar species tend to underestimate the permissible SSD ratio range causing the method to produce many invalid Non-SSDs). The effect of outgroups on Ortholuge's performance is two-fold: ideal outgroups have a large number of orthologs in common with the ingroups (without an outgroup ortholog, Ortholuge analysis cannot be run on the predicted ortholog pair). This favors closely related outgroup species. Conversely, outgroup genomes that are too closely related over-estimate the level of ortholog divergence (as the denominator in the Ortholuge ratios decreases), causing the method to improperly classify many SSDs as Non-SSDs. The criteria describe the inter-quartile range of an optimal Ortholuge ratio distribution. A large sample of bacterial comparison genomes matched with various outgroup genomes were tested using these criteria (the sample consisted of 10851 individual test cases). The CVtree distances were recorded for all cases that had an inter-quartile range meeting the criteria.

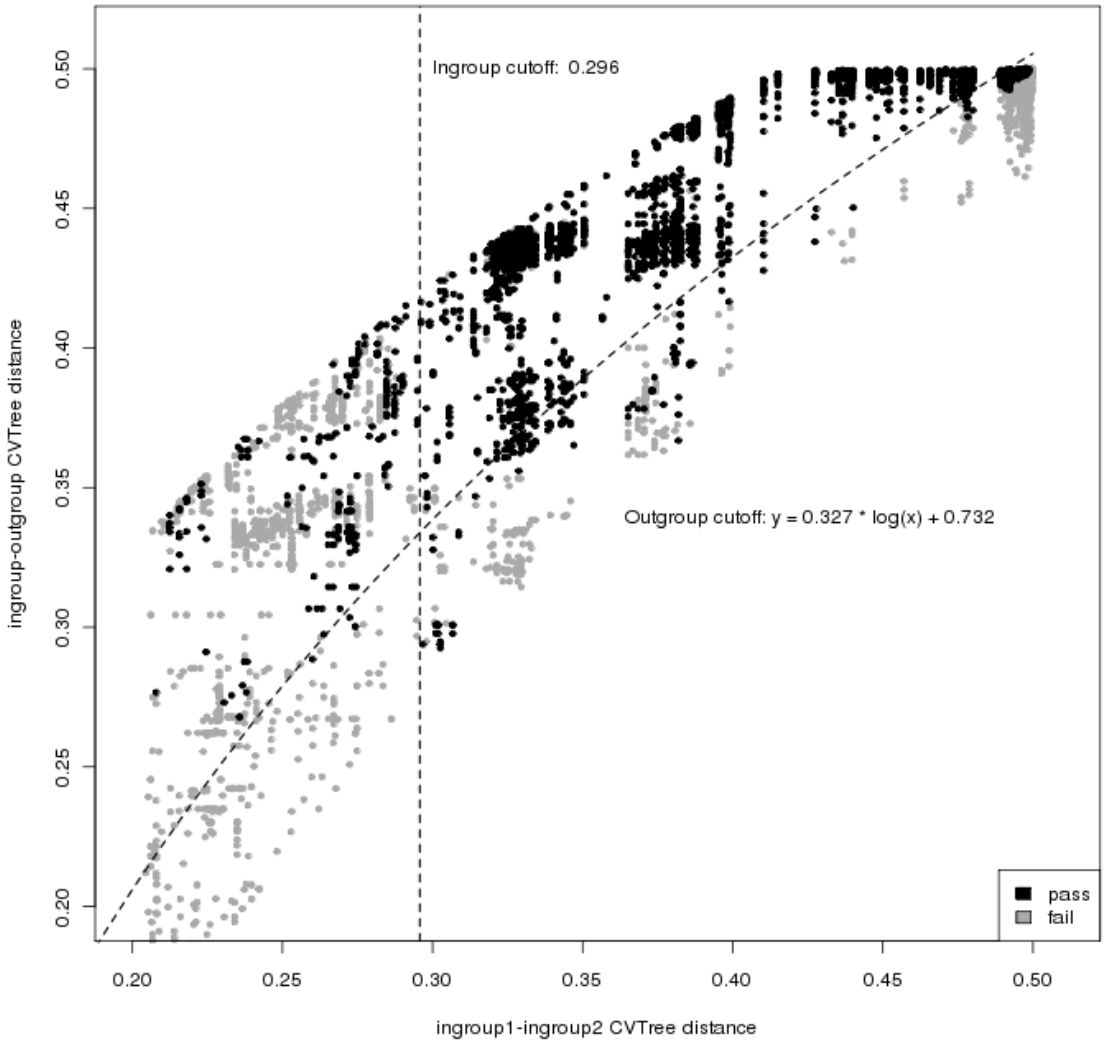


Figure 5.5 Ortholuge Analysis Test Cases that Satisfy the Optimal Ratio Distribution Criteria

A sample of Ortholuge analyses consisting of various ingroup comparison genomes and outgroup reference genome combinations were tested to determine if they have a suitable inter-quartile range. CVtree distances were recorded for all successful and unsuccessful test cases and a formula was derived to distinguish the successful Ortholuge analyses.

From the sample, the following formula was derived using logistic regression (a success rate of 0.65 was chosen as the threshold for estimating the formula parameters).

$$ingroup\ distance > 0.296$$
$$outgroup\ distance > 0.327 \times \log(x) + 0.732$$
$$\text{where } x = \text{ingroup distance}$$

In practice, for a pair of ingroup comparison genomes the outgroup distance threshold is computed and the ten outgroup genomes that have the closest CVtree distance value that is greater than the threshold are identified. From the ten, the single genome that has the largest number of common orthologs is selected as the outgroup. This approach attempts to balance the competing effects of the outgroup: the restriction of possible Ortholuge evaluations due to the availability of outgroup orthologs and the requirement for a sufficiently distant outgroup that properly roots the orthologs' phylogenetic trees. Using this method, 393 564 outgroups were identified for the 993 341 pairs of genomes in OrtholugeDB.

5.4. Clustering Orthologs across Multiple Species

In addition to the pair-wise orthologs, OrtholugeDB also contains pre-computed ortholog groups. The purpose of ortholog groups is to assemble all orthologous genes that diverged from the same gene in the last common ancestor of the species under investigation. By definition, ortholog groups can contain both orthologs and in-paralogs. A duplicated gene's classification as an in-paralog (gene duplication occurred after speciation) or out-paralog (duplication prior to speciation) is relative to the depth of the last common ancestor of the species under consideration. Thus, the taxonomic range impacts the composition of ortholog groups. Ortholog groups consider all orthologous relationships across the species and are typically represented as a network graph where genes represented as nodes and orthologous relationships as edges between nodes. Ortholog groups are frequently used to study the distribution of the orthologs for a gene of interest across multiple species. Ortholog groups are also used in functional inference of novel genes and as input for phylogenetic analysis.

5.4.1. Implementation

The ortholog groups in OrtholugeDB are transitive; all genes connected by an orthologous or in-paralogous relationship in the genomes under consideration are included in the group. We did post-processing of the groups to remove invalid ortholog connections. Incorrect ortholog connections can result in the fusion of separate ortholog groups. Most ortholog groups are densely connected. Incorrect ortholog predictions can appear as a single or small number of edges bridging two densely connected sub-groups (representing distinct ortholog groups). The post-processing step ensures that the groups maintain a certain overall level of connectivity by splitting groups along ortholog edges that have a normalized minimum-cut value below the pre-defined threshold of 0.1 (a minimum-cut is the number of edges needed to be removed to create two disjoint subgraphs. A normalized minimum-cut is the minimum-cut value divided by the number of edges in the two resulting subgraphs). The tool Graclus is used to compute the normalized minimum-cut for the groups (Dhillon, Guan, and Kulis 2007).

Groups are constructed for multiple hierarchical levels representing sets of species with increasing phylogenetic distance. The levels are based on CVtree distances (higher levels have a greater allowable CVtree distance) (Xu and Hao 2009). Level distances were selected to match taxonomy classifications from the NCBI Taxonomy database for Genus, Family, Order, Class and Phylum (Benson *et al.* 2009; Sayers *et al.* 2009).

5.4.2. Methods

Comparison of the Pseudomonas Genome Database and OrtholugeDB Ortholog Groups

Gene names were used to assess the quality of the ortholog groups' assignments for the OrtholugeDB and *Pseudomonas* Genome Database ortholog group procedures. *Pseudomonas* species gene names were obtained from the *Pseudomonas* Genome Database and assigned to the applicable *Pseudomonas* Genome Database ortholog groups (POGs) and OrtholugeDB ortholog groups (ortholog group genes with no assigned name were discarded from the analysis). Total groups with a single consistent gene name were calculated for the OrtholugeDB and POG procedures.

Dissimilarity and silhouette values were computed for the ortholog groups of both methods as well. Dissimilarity is a measure that indicates the proportion of ortholog group genes with distinct gene names. Dissimilarity for an ortholog group is calculated as follows:

$$dissimilarity = average(\, d(i) \,)$$

$$d(i) = \sum_{j}^{j\,!=\,i} \begin{matrix} 0 \; if \; gene \; names \; identical, \\ 1 \; otherwise \end{matrix}$$

For each gene i, j in the ortholog group

Silhouette value is a clustering validation metric that indicates how well separated the ortholog groups are based on their dissimilarity values. Silhouette value for an ortholog group was calculated as follows (refer to above definition of dissimilarity):

$$silhouette = average\ (\, s(i) \,)$$

$$s(i) = \sum_{i} \frac{B(i) - A(i)}{\max\big(B(i), A(i)\big)}$$

where $A(i)$ is the average dissimilarity of the i^{th} gene to genes in the same ortholog group and $B(i)$ is the minimum average dissimilarity of the i^{th} gene to any other ortholog group

5.4.3. Results and Discussion

The hierarchical approach used for the OrtholugeDB Ortholog groups has a number of benefits. Ortholog prediction is more accurate between closely related species (Jensen *et al.* 2008). Ortholog status is also relative to the species under consideration (i.e. an out-paralog can become an in-paralog when the depth of the last common ancestor of the species under investigation increases) (Jensen *et al.* 2008; Altenhoff *et al.* 2011). By providing ortholog groups for a number of levels, users can select their desired taxonomic range. CVtree distances were chosen to define the species used in the hierarchical levels instead of taxonomy classifications because CVtree distance-based levels have a consistent phylogenetic range. Taxonomy groups at equivalent levels (e.g. genus, family, order etc.) are highly variable in their evolutionary range.

The method for computing OrtholugeDB ortholog groups is based on the hierarchical grouping approach developed for the OMA database (Altenhoff *et al.* 2011) and the transitive grouping strategy used for computing *Pseudomonas* ortholog groups (POGs) in the *Pseudomonas* Genome Database (Winsor *et al.* 2009). There are, however, some differences between the approaches. The *Pseudomonas* Genome Database method is designed for computing ortholog groups at the Genus level and uses conservation of gene order to resolve multiple RBB-predicted ortholog candidates (Winsor *et al.* 2009). The use of gene order is not easily extended to species with broader phylogenetic distances, so this step has been removed from the OrtholugeDB procedure. In comparison to the OMA database method, differences include the use of normalized minimum-cut in the OrtholugeDB approach versus the standard minimum-cut value in the OMA approach to identify weak edges in the ortholog group connections. A standard minimum-cut value tends to identify edges to small weakly-connected sub-groups (quite often the minimum-cut in an ortholog group occurs along an ortholog edge linking a single gene). A normalized minimum-cut value balances the minimum-cut value with the connectivity in the two resulting partitions (Figure 5.6). Normalized minimum-cuts can identify weak edges, which might not be the global minimum-cut, but when removed produce densely connected sub-groups (Shi and Malik 2000). Another difference is the use of hierarchical groups based on CVtree distances in OrtholugeDB instead of the taxonomy classifications used in the OMA database (Benson *et al.* 2009; Sayers *et al.* 2009; Xu and Hao 2009). The CVtree distance thresholds in OrtholugeDB were selected to produce groups that are similar to the taxonomic groups. The immediate benefit of using a distance based approach is that unclassified or incorrectly classified species will be properly grouped with species of similar distances. The other potential benefit of CVtree-based groupings is that they have a more consistent phylogenetic range than taxonomic classifications. One of the potential uses of hierarchical groups is to determine the timing of the emergence and disappearance of genes by tracking the genes in ortholog groups through multiple levels (Altenhoff *et al.* 2011). Using groups based on a distance metric with equivalent phylogenetic ranges might provide the ability to more consistently compare the distribution of genes in diverse taxa.

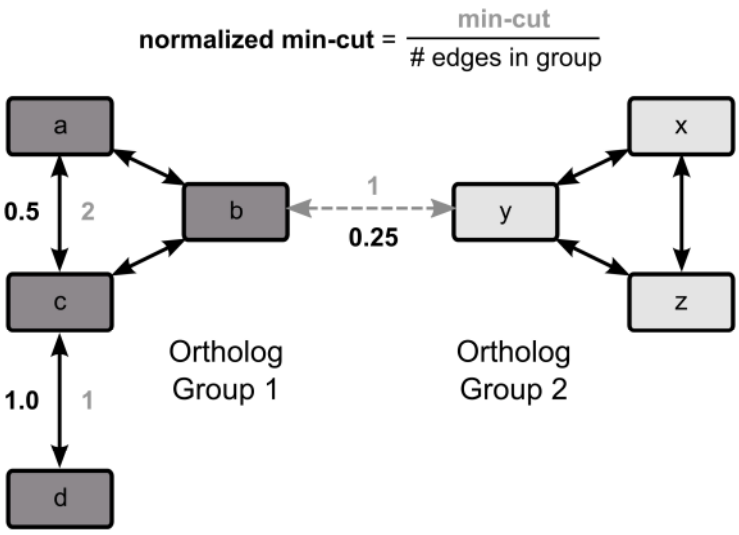


Figure 5.6 Comparison of Normalized Minimum Cuts and Standard Minimum Cuts for Identifying Invalid Orthologous Relationships

The normalized minimum cut (min-cut) and standard min-cut values are shown for two ortholog groups merged due to a single invalid ortholog relationship (edge). The groups are displayed as a graph network: genes are represented by the square nodes and orthologous relationships by the double-arrows. Standard min-cut values are shown in grey and the normalized min-cut values are shown in black. A min-cut is the minimum number of edges that need to be removed in order to partition the graph into disconnected subgraphs. A normalized min-cut value normalizes the min-cut value by the number of edges in the subgraph that would result from removing the edge. Low normalized min-cuts indicate edges that when removed will yield highly connected subgraphs. The lowest normalized min-cut occurs along the b–y edge connecting the two ortholog group subgraphs. The lowest standard min-cut occurs along two edges (b–y and c–d). Removing the c–d edge would result in an ortholog group consisting of one gene.

The accuracy of the ortholog groups in OrtholugeDB was compared to the POGs in the *Pseudomonas* Genome Database by analyzing the gene name assignments in ortholog groups. Through the work of *Pseudomonas* community annotation project, gene name assignments in the *Pseudomonas* genus are of overall higher quality than for many genome projects (most assignments are based on curated experimental evidence and not sequence similarity alone). Ortholog group accuracy was determined by the distribution of gene names in a group. The expectation is that all genes with the same name will be clustered in the same ortholog group. Based on the results of the gene name analysis, the modified procedure in OrtholugeDB for ortholog group construction produces ortholog groups that are equivalent or slightly better in terms of accuracy when

compared to the POGs in the *Pseudomonas* Genome Database (the POG procedure, which uses gene order to resolve multiple ortholog predictions, is designed for predicting ortholog groups at the Genus level only. The OrtholugeDB procedure examines connectivity to infer the merger of distinct ortholog groups. This approach is applicable at all taxonomic levels). In total, 64.87% of OrtholugeDB ortholog groups have one gene name verses 64.42% of POGs. Dissimilarity is a measure that indicates the proportion of genes names clustered together that do not match (a dissimilarity value of 0 indicates all gene names in a group are identical). OrtholugeDB groups displayed a marked reduction in dissimilarity: 20.89 compared to 22.03 for the POGs. This improvement in group similarity suggests that the unique post-processing step in the OrtholugeDB procedure is correctly splitting ortholog groups (The post-processing step identifies and splits groups are that believed to be incorrectly merged due to a false ortholog prediction). Silhouette values are used in the validation of clustering. They indicated how well separated the ortholog groups are by comparing the gene name dissimilarity of a gene's ortholog group, to the dissimilarity of the next closest ortholog group. Genes that are correctly clustered will be highly similarity to the ortholog group they are assigned to, but display little similarity to any other ortholog group. Silhouette values range from -1 to 1, where a value of 1 indicates an optimal clustering. The average silhouette values for OrtholugeDB groups were 0.6898 and for POGs 0.6859.

Empirical cumulative distributions were computed for the dissimilarity and silhouette values for the POG and OrtholugeDB groups (Figure 5.7 and Figure 5.8). The empirical cumulative distributions are highly similar, signifying that both methods produce equivalent proportions of validated ortholog groups. Based on these distributions, it appears the majority of ortholog groups are consistent and that both methods are performing equivalently in this Genus level comparison. The dissimilarity and silhouette values from the gene name analysis show that overall the methods are producing highly accurate ortholog predictions. The benefit of the OrtholugeDB approach is that it can be applied to genomes with any level of evolutionary relatedness. The requirement for detectable gene order conservation in the POG approach limits its applicability to closely-related genomes. OrtholugeDB contains genomes from across the bacteria and archaea domains, so the flexibility of the OrtholugeDB ortholog prediction approach is essential for handling the diversity in the database.

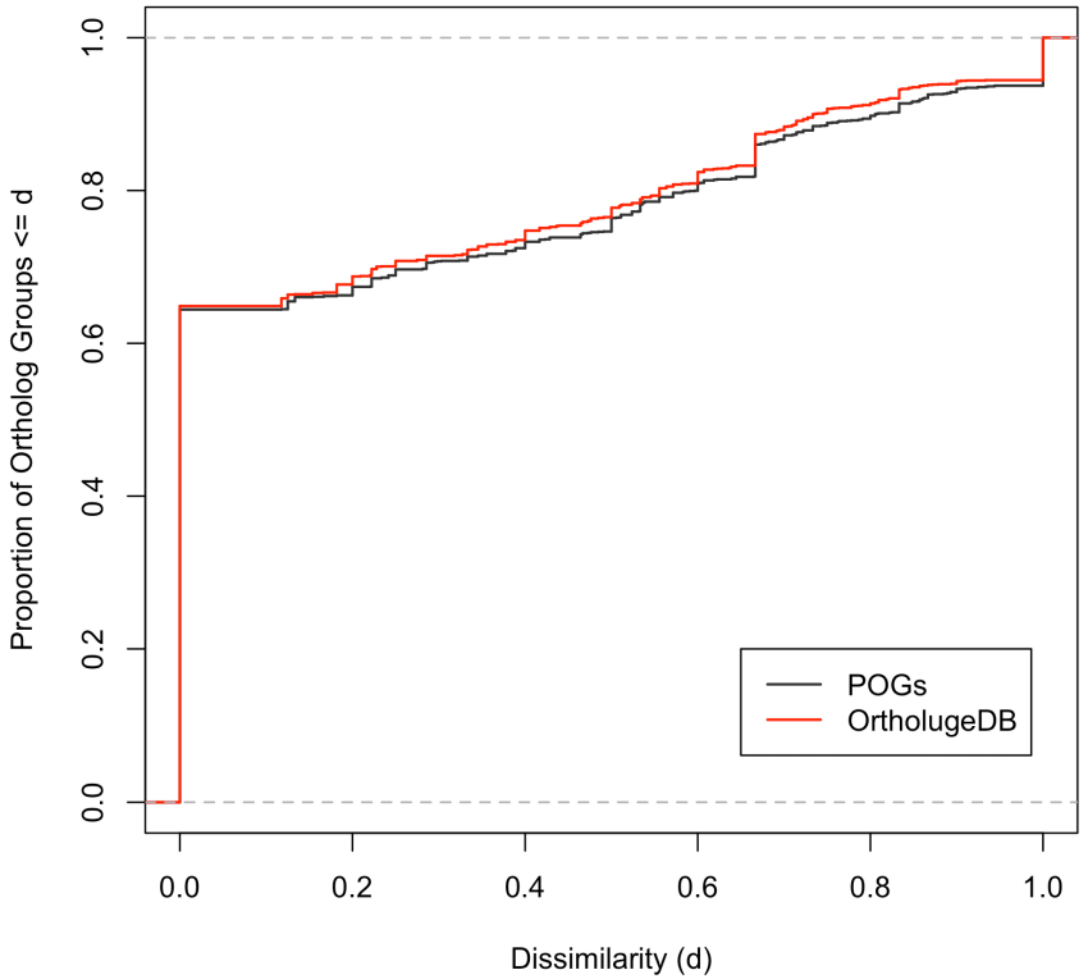


Figure 5.7 Empirical Cumulative Distributions for the Ortholog Group Dissimilarity Values

Empirical cumulative distributions (ecd) were computed for the dissimilarity values of *Pseudomonas* Genome Database (POGs) and OrtholugeDB ortholog groups. The dissimilarity (*d*) values measures the proportion of ortholog group genes with distinct gene name annotations. The ecd highlights the proportion of total ortholog groups with a dissimilarity value that is equal to or less than the given *d*.

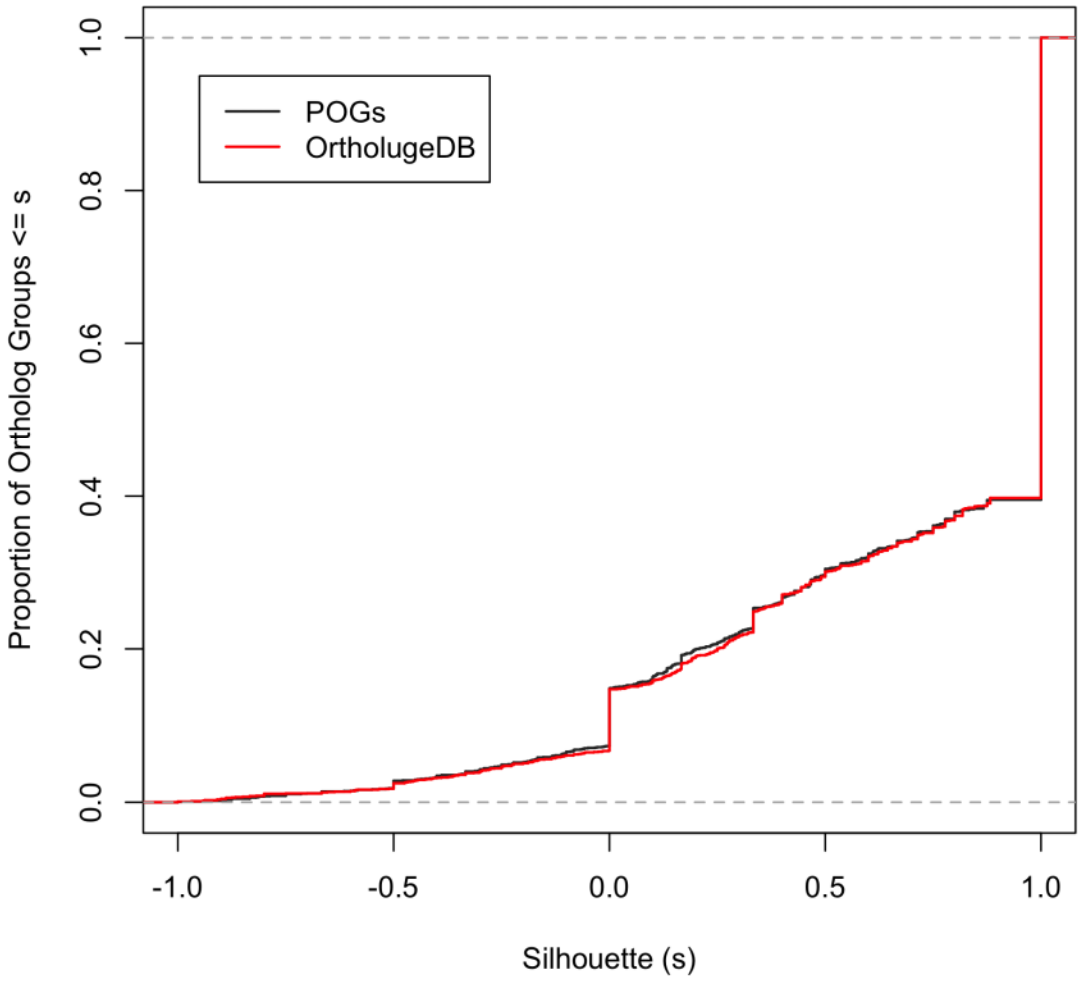


Figure 5.8 Empirical Cumulative Distributions for the Ortholog Group Silhouette Values

Empirical cumulative distributions (ecd) were computed for the silhouette values of *Pseudomonas* Genome Database (POGs) and OrtholugeDB ortholog groups. A silhouette value (*s*) indicates how well separated the ortholog groups are based on their dissimilarity values. The ecd highlights the proportion of total ortholog groups with a silhouette value that is equal to or less than the given *s*.

5.5. Comparison of the Functionality in OrtholugeDB to OMA Browser and QuartetS-DB

The ortholog prediction algorithms QuartetS and OMA have associated orthology databases: QuartetS-DB (Yu *et al.* 2012) and OMA Browser (Altenhoff *et al.* 2011) respectively. They contain pair-wise ortholog predictions and ortholog groups and are currently two of the largest orthology databases. OrtholugeDB was developed with a specific focus on ease-of-use and functionality. The queries and result views are designed for microbiologists to be able to (i) rapidly query NCBI genomes and identify shared or unique genes, (ii) access previously hard-to-obtain Ortholuge assessments when an outgroup is available and (iii) visualize the orthologs for a gene or set of genomes. For querying shared and unique genes, the web interface in OrtholugeDB contains a flexible phyletic-based search that can identify common or unique genes in a genome of interest based on the presence or absence of orthologs in one or more comparison species. QuartetS-DB phyletic search is limited to identifying ortholog groups based only on the presence of orthologs in selected species and the OMA browser does not have phyletic-based search capabilities. OrtholugeDB also provides a separate rapid query for identifying the unique genes in a comparison of two genomes. OrtholugeDB includes a number of visualizations for ortholog data not offered in QuartetS-DB or the OMA browser, such as the gene context view and the graph view of ortholog groups. The phyletic matrix view in OrtholugeDB efficiently displays the distribution of the genes in a query genome across multiple comparison genomes. While QuartetS-DB has a phyletic-based tabular view of the ortholog groups, the phyletic matrix in OrtholugeDB is more informative, as it shows ortholog cardinality (i.e. one-to-one, many-to-many, no orthologs, etc.), plus Ortholuge classifications. The OrtholugeDB phyletic view also includes a summary page that shows the proportions of protein-coding genes in the query genome associated with each type of orthologous relationship. Gene duplications are important to consider when inferring which orthologs are likely to have similar functions. If a particular ortholog has been duplicated in a genome, it is not straight-forward to determine how the gene function is impacted (for example, whether one of the duplicated genes maintained the ancestral gene function or subfunctionalization between the duplicated genes occurred) (Remm, Storm, and Sonnhammer 2001). In-paralog predictions are integrated into the result views in

OrtholugeDB to flag ortholog genes that have undergone duplication after species divergence. In QuartetS-DB, in-paralog groups must be obtained through a separate query.

5.6. Conclusions

Ortholuge, by identifying orthologs that diverged to the same relative degree as their species, produces a set of orthologs that are more likely to have retained similar function and are better suited for comparative genomic analyses. OrtholugeDB makes Ortholuge analysis readily available for bacterial and archaeal species. The OrtholugeDB website is designed to facilitate the retrieval of orthologs for single genes to multiple genomes. It includes features that allow high-level visualization of orthologs and formulating complex queries to retrieve orthologs. OrtholugeDB facilitates bacterial and archaeal comparative genomic analysis by providing accurate and large-scale ortholog predictions, a flexible search interface and lastly, precise assessment of orthologs when suitable reference genomes are available.

6. Concluding Remarks

The increase in genomic sequencing throughput has generated a rapid growth in the number of available genome sequences. Experimental investigation was essential to initially discovering many gene functions, but reproducing the same level of investigation for the large number of newly sequenced organisms is not feasible. Automated methods for gene function inference are needed in order to make effective use of the growing genomic resources. Comparative genomics is the study of genome functions across species. It uses our understanding of evolution as a framework for predicting gene functions in newly sequenced species by transferring knowledge from well-studied organisms. My thesis contributes to comparative genomics by building upon computational ortholog prediction. Orthologs are genes that have diverged when the species diverge. Because they evolve from a common ancestral gene, orthologs are thought to more likely have similar functions than paralogs; related genes that have arisen through gene duplication. This ortholog functional conservation hypothesis is the basis for many comparative genomics methods that infer gene functions across species. The Brinkman laboratory at Simon Fraser University developed a method called Ortholuge, which can generate precise ortholog predictions between two species on a genome-wide scale. Ortholuge achieves improved performance by examining the phylogenetic divergence of predicted orthologs in a high-throughput manner. Predicted orthologs undergoing unusual divergence in many cases represent paralogs incorrectly predicted as orthologs or orthologs where one gene has rapidly diverged in one of the species.

My thesis consisted of multiple aims to advance computational ortholog prediction. Firstly, I conducted and evaluated comparative genomics-based analyses in three separate projects that used orthologs to transfer gene annotations across species. In addition to reporting the biological discoveries from these projects, I also identified problems in employing orthologs in these types of analyses. In collaboration with Dr. Michel Leroux of Simon Fraser University, we identified and characterized proteins that

are exclusive to and widely conserved across metazoans; a monophyletic clade in which multicellularity emerged. Conserved metazoan-specific proteins will likely participate in processes required for multicellularity. This analysis used a phyletic pattern-based search to identify conserved orthologs across metazoan species but absent in non-metazoans. Gene duplication and incomplete genomes were confounding factors in identifying metazoan-associated genes. While ancestral genes are conserved, gene duplication and associated functional divergence in the individual species lineages impede the inference of conserved gene functions in metazoan species. A flexible approach was conceived and employed for selecting conserved orthologs in cases of incomplete genomes. This approach permitted orthologs that were missing in a limited number of metazoan species, provided they were widely conserved in all metazoan clades. The second comparative genomics project was a meta-analysis of gene expression differences in epidemic strains of *Pseudomonas aeruginosa.* This meta-analysis examined microarray data from multiple epidemic and non-epidemic *P. aeruginosa* strains to identify changes that have been selected for in multiple, separate epidemic *P. aeruginosa* strains. This analysis uncovered multiple biological systems that are differentially expressed in epidemic *P. aeruginosa*. The differentially expressed biological systems became the starting seeds for a comparative genomics-based search of transcriptional modules associated with the gene expression differences in epidemic *P. aeruginosa*. The conservation of gene regulation was found to be highly variable with many putative transcription factors binding sites not conserved between orthologs in other *Pseudomonas* species. The rapid divergence of transcriptional regulation can significantly limit the scope of the analysis, however, several predicted transcriptional modules were found to be relevant in epidemic *P. aeruginosa* infection (based on the gene expression changes). The last of the three comparative-based projects looked at the biological systems that are differentially expressed in the fungal pathogen *Aspergillus fumigatus* in response to iron availability. This project was in collaboration with the Moore Laboratory at Simon Fraser University. Pathogens encounter low iron conditions in the host environment during infection and it has been shown that *A. fumigatus'* iron acquisition systems are critical to its virulence. A comparative genomics approach was necessary to perform a systems-level analysis of the microarray data because no transcriptional regulatory networks were available for this relatively under-studied fungal pathogen. A putative transcriptional regulatory network was constructed

from the transcriptional regulatory interactions identified in *Saccharomyces cerevisiae* and *Candida albicans*. Analysis of the differentially expressed components in the network identified one specific subnetwork associated with ribosomal protein (RP) genes. The RP genes in *A. fumigatus* appeared to be differentially expressed in response to iron limitation. However, the RP gene transcriptional regulation has undergone a major reorganization in fungi. Most of the characterized transcription factors involved in RP gene regulation in *S. cerevisiae* and *C. albicans* have no corresponding ortholog in *A. fumigatus*. This result emphasized that phylogenetic distance is an important consideration when transferring transcription regulatory interactions. In this case, the overall behavior of the network is conserved with the RP genes differentially regulated by environmental stresses in *S. cerevisiae, C. albicans and A. fumigatus*, however based on the ortholog conservation, the responses are likely carried out by distinct mechanisms. Together, these comparative genomics-based projects highlighted potential problem areas associated with different types of comparative genomics analysis. When orthology is used as a proxy for equivalent gene functions or gene regulation across species, further validation of the results should be performed to handle incorrect ortholog prediction or identify cases of functional divergence between the orthologs.

The second thesis research aim was to study in detail the gene function conservation of the different classes of orthologs identifiable by Ortholuge. The Ortholuge tool classifies predicted orthologs as either orthologs that support species divergence (SSD orthologs) or putative orthologs that are undergoing unusual divergence (Non-SSD orthologs). Through the examination of protein features linked to function, such as protein domains and subcellular localization, I found a statistically significant increase in the proportion of Non-SSD predicted orthologs that have dissimilar properties compared to SSD orthologs. Non-SSDs are also more often associated with large gene families, a complicating factor in ortholog prediction. Additionally, SSD compared to Non-SSD orthologs are more often associated with syntenic genome regions. These results suggested that a striking number of Non-SSD orthologs may be mispredicted paralogs. I compared the performance of Ortholuge to two other ortholog prediction methods: OMA and QuartetS. Similar to Ortholuge, these graph-based methods use reciprocal best BLAST procedure to produce an initial set of ortholog

predictions which are then refined using additional phylogenetic information. In comparison to OMA and QuartetS, Ortholuge appears to more consistently identify RBB-predicted orthologs with similar functions for species from a wide range of taxonomic distances.

The third aim was to address several performance issues in the original implementation of Ortholuge. Over the course of my thesis, I made several significant improvements to the Ortholuge method. These improvements have made the Ortholuge method fully automatable, enhanced performance and accuracy and incorporated a rigorous statistical procedure that assigns a statistical significance to Ortholuge's results, (developed in collaboration with the Graham and McNeney Laboratories at Simon Fraser University). I also added the ability to detect in-paralog relationships in addition to orthologs to the Ortholuge method.

The final aim of my thesis was to increase access to Ortholuge-based orthologs predictions. Working with developers in the Brinkman Laboratory, we built an online database called OrtholugeDB, to provide Ortholuge-predicted orthologs for bacteria and archaea. This resource automatically runs Ortholuge to predict orthologs between fully sequenced bacterial and archaeal genomes and then makes the results available to the research community through a searchable web interface. The interface provides multiple types of queries including a powerful phyletic-based query that allows the extraction of genes conserved in a specified set of taxa and absent in another set. Visualizations provided for the website results are designed to help with evaluating the predicted orthologs or efficiently displaying the distribution of the genes across species. The database also contains pre-computed multi-species ortholog groups. These ortholog groups have been computed in a hierarchical manner using multiple levels of species with increasing phylogenetic ranges.

Most ortholog prediction projects evaluate their performance using the evolutionary definition of orthologs. When considering that many comparative genomics-based applications of computationally predicted orthologs assume a functional correspondence between the orthologs, using the evolutionary definition may be inappropriate in these cases. Orthologs, compared to paralogs, are often the best candidates in their respective genomes for finding genes with similar functions, however,

orthologs can diverge in function or can be missing, especially between phylogenetically distant species. Advancements in computational ortholog prediction should focus on developing approaches that identify the subset of orthologs that are functionally similar rather than all orthologs that satisfy the evolutionary definition. The Ortholuge method examines the predicted ortholog divergence in the context of the species divergence. I showed that Ortholuge helps identify a more functionally similar set of orthologs from the set predicted by the RBB method. Future Ortholuge development should focus on prediction coverage, the single biggest area that needs to be improved in Ortholuge, although overall efficiency should also continue to be improved in order to handle the exponentially increasing number of genomes that are being released. By improving methods for computational ortholog prediction and making the results readily available to the research community, this work will help facilitate comparative genomics analysis in an era of rapid genome sequencing.

References

Altenhoff AM, Dessimoz C. 2009. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS computational biology* 5:e1000262.

Altenhoff AM, Schneider A, Gonnet GH, Dessimoz C. 2011. OMA 2011: orthology inference among 1000 complete genomes. *Nucleic acids research* 39:D289–94.

Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C. 2012. Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS computational biology* 8:e1002514.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* 25:3389–402.

Arendt D. 2008. The evolution of cell types in animals: emerging principles from molecular studies. *Nature reviews. Genetics* 9:868–82.

Armstrong D, Bell S, Robinson M, Bye P, Rose B, Harbour C, Lee C, Service H, Nissen M, Syrmis M, *et al.* 2003. Evidence for spread of a clonal strain of Pseudomonas aeruginosa among cystic fibrosis clinics. *Journal of clinical microbiology* 41:2266–7.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al.* 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* 25:25–9.

Bailey TL, Bodén M, Whitington T, Machanick P. 2010. The value of position-specific priors in motif discovery using MEME. *BMC bioinformatics* 11:179.

Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology* 2:28–36.

Bailey TL, Williams N, Misleh C, Li WW. 2006. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic acids research* 34:W369–73.

Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, *et al.* 2011. NCBI GEO: archive for functional genomics data sets--10 years on. *Nucleic acids research* 39:D1005–10.

Bendena WG, Boudreau JR, Papanicolaou T, Maltby M, Tobe SS, Chin-Sang ID. 2008. A Caenorhabditis elegans allatostatin/galanin-like receptor NPR-9 inhibits local search behavior in response to feeding cues. *Proceedings of the National Academy of Sciences of the United States of America* 105:1339–42.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2009. GenBank. *Nucleic acids research* 37:D26–31.

Bergmann S, Ihmels J, Barkai N. 2003. Iterative signature algorithm for the analysis of large-scale gene expression data. *Physical review. E, Statistical, nonlinear, and soft matter physics* 67:031902.

Bininda-Emonds ORP. 2005. transAlign: using amino acids to facilitate the multiple alignment of protein-coding DNA sequences. *BMC bioinformatics* 6:156.

Bray T, Paoli J, Sperberg-McQueen CM, Maler E, Yergeau F. 2008. Extensible Markup Language (XML) 1.0 (Fifth Edition).

Bredenbruch F, Geffers R, Nimtz M, Buer J, Häussler S. 2006. The Pseudomonas aeruginosa quinolone signal (PQS) has an iron-chelating activity. *Environmental microbiology* 8:1318–29.

Cannon SB, Young ND. 2003. OrthoParaMap: distinguishing orthologs from paralogs by integrating comparative genome data and gene phylogenies. *BMC bioinformatics* 4:35.

Catterson JJ. 2008. Examining the early transcriptome of the fungal pathogen Aspergillus fumigatus in response to iron limitation imposed by human serum. Simon Fraser University.

Chen F, Mackey AJ, Stoeckert CJ, Roos DS. 2006. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic acids research* 34:D363–8.

Chen F, Mackey AJ, Vermunt JK, Roos DS. 2007. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PloS one* 2:e383.

Chen T, Wu TH, Ng W V, Lin W. 2010. DODO: an efficient orthologous genes assignment tool based on domain architectures. Domain based ortholog detection. *BMC bioinformatics* 11 Suppl 7:S6.

Conant GC, Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new functions. *Nature reviews. Genetics* 9:938–50.

Csárdi G, Kutalik Z, Bergmann S. 2010. Modular analysis of gene expression data with R. *Bioinformatics* 26:1376–7.

Davidsen T, Beck E, Ganapathy A, Montgomery R, Zafar N, Yang Q, Madupu R, Goetz P, Galinsky K, White O, *et al.* 2010. The comprehensive microbial resource. *Nucleic acids research* 38:D340–5.

Dehal PS, Joachimiak MP, Price MN, Bates JT, Baumohl JK, Chivian D, Friedland GD, Huang KH, Keller K, Novichkov PS, *et al.* 2010. MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic acids research* 38:D396–400.

DeLuca TF, Cui J, Jung J-Y, St Gabriel KC, Wall DP. 2012. Roundup 2.0: enabling comparative genomics for over 1800 genomes. *Bioinformatics* 28:715–6.

Deluca TF, Wu I-H, Pu J, Monaghan T, Peshkin L, Singh S, Wall DP. 2006. Roundup: a multi-genome repository of orthologs and evolutionary distances. *Bioinformatics* 22:2044–6.

Dessauer CW. 2009. Adenylyl cyclase--A-kinase anchoring protein complexes: the next dimension in cAMP signaling. *Molecular pharmacology* 76:935–41.

Dessimoz C, Gabaldón T, Roos DS, Sonnhammer ELL, Herrero J. 2012. Toward community standards in the quest for orthologs. *Bioinformatics* 28:900–4.

Dewey CN. 2011. Positional orthology: putting genomic evolutionary relationships into context. *Briefings in bioinformatics* 12:401–12.

Dhillon IS, Guan Y, Kulis B. 2007. Weighted graph cuts without eigenvectors a multilevel approach. *IEEE transactions on pattern analysis and machine intelligence* 29:1944–57.

Dufayard J-F, Duret L, Penel S, Gouy M, Rechenmann F, Perrière G. 2005. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics* 21:2596–603.

Edgar RC. 2004a. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics* 5:113.

Edgar RC. 2004b. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* 32:1792–7.

Efron B. 2004. Large-Scale Simultaneous Hypothesis Testing. *Journal of the American Statistical Association* 99:96–104.

Felsenstein J. 2008. PHYLIP (Phylogeny Inference Package) version 3.6.

Forslund K, Pekkari I, Sonnhammer ELL. 2011. Domain architecture conservation in orthologs. *BMC bioinformatics* 12:326.

Francino MP. 2005. An adaptive radiation model for the origin of new gene functions. *Nature genetics* 37:573–7.

Fu Z, Jiang T. 2008. Clustering of main orthologs for multiple genomes. *Journal of bioinformatics and computational biology* 6:573–84.

Fulton DL, Li YY, Laird MR, Horsman BGS, Roche FM, Brinkman FSL. 2006. Improving the specificity of high-throughput ortholog prediction. *BMC bioinformatics* 7:270.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, *et al.* 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology* 5:R80.

Giardina G, Castiglione N, Caruso M, Cutruzzolà F, Rinaldo S. 2011. The Pseudomonas aeruginosa DNR transcription factor: light and shade of nitric oxide-sensing mechanisms. *Biochemical Society transactions* 39:294–8.

Goeman JJ, Van de Geer SA, De Kort F, Van Houwelingen HC. 2004. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 20:93–9.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* 27:1017–8.

Grote A, Klein J, Retter I, Haddad I, Behling S, Bunk B, Biegler I, Yarmolinetz S, Jahn D, Münch R. 2009. PRODORIC (release 2009): a database and tool platform for the analysis of gene regulation in prokaryotes. *Nucleic acids research* 37:D61–5.

Haft DH. 2003. The TIGRFAMs database of protein families. *Nucleic Acids Research* 31:371–373.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, *et al.* 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic acids research* 32:D258–61.

Van der Heijden RTJM, Snel B, Van Noort V, Huynen MA. 2007. Orthology prediction at scalable resolution by phylogenetic tree analysis. *BMC bioinformatics* 8:83.

Hibbing ME, Fuqua C, Parsek MR, Peterson SB. 2010. Bacterial competition: surviving and thriving in the microbial jungle. *Nature reviews. Microbiology* 8:15–25.

Hissen AHT, Wan ANC, Warwas ML, Pinto LJ, Moore MM. 2005. The Aspergillus fumigatus siderophore biosynthetic gene sidA, encoding L-ornithine N5-oxygenase, is required for virulence. *Infection and immunity* 73:5493–503.

Hittinger CT, Carroll SB. 2007. Gene duplication and the adaptive evolution of a classic genetic switch. *Nature* 449:677–81.

Hogues H, Lavoie H, Sellam A, Mangos M, Roemer T, Purisima E, Nantel A, Whiteway M. 2008. Transcription factor substitution during the evolution of fungal ribosome regulation. *Molecular cell* 29:552–62.

Hulsen T, Huynen MA, De Vlieg J, Groenen PMA. 2006. Benchmarking ortholog identification methods using functional genomics data. *Genome biology* 7:R31.

Ideker T, Ozier O, Schwikowski B, Siegel AF. 2002. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 18:S233–S240.

Ihmels J, Bergmann S, Barkai N. 2004. Defining transcription modules using large-scale gene expression data. *Bioinformatics* 20:1993–2003.

Jensen LJ, Julien P, Kuhn M, Von Mering C, Muller J, Doerks T, Bork P. 2008. eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic acids research* 36:D250–4.

Johnson GL, Lapadat R. 2002. Mitogen-activated protein kinase pathways mediated by ERK, JNK, and p38 protein kinases. *Science* 298:1911–2.

Jones AM, Govan JR, Doherty CJ, Dodd ME, Isalska BJ, Stanbridge TN, Webb AK. 2001. Spread of a multiresistant strain of Pseudomonas aeruginosa in an adult cystic fibrosis clinic. *Lancet* 358:557–8.

Joo YJ, Kim J-H, Kang U-B, Yu M-H, Kim J. 2011. Gcn4p-mediated transcriptional repression of ribosomal protein genes under amino-acid starvation. *The EMBO journal* 30:859–72.

Jordan IK. 2001. Lineage-Specific Gene Expansions in Bacterial and Archaeal Genomes. *Genome Research* 11:555–565.

Jothi R, Zotenko E, Tasneem A, Przytycka TM. 2006. COCO-CL: hierarchical clustering of homology relations based on evolutionary correlations. *Bioinformatics* 22:779–88.

Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome research* 20:1313–26.

Kamesh N, Aradhyam GK, Manoj N. 2008. The repertoire of G protein-coupled receptors in the sea squirt Ciona intestinalis. *BMC evolutionary biology* 8:129.

Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, *et al.* 2008. KEGG for linking genomes to life and the environment. *Nucleic acids research* 36:D480–4.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* 28:27–30.

Koonin E V. 2005. Orthologs, paralogs, and evolutionary genomics. *Annual review of genetics* 39:309–38.

Kristensen DM, Wolf YI, Mushegian AR, Koonin E V. 2011. Computational methods for Gene Orthology inference. *Briefings in bioinformatics* 12:379–91.

Kukavica-Ibrulj I, Bragonzi A, Paroni M, Winstanley C, Sanschagrin F, O'Toole GA, Levesque RC. 2008. In vivo growth of Pseudomonas aeruginosa strains PAO1 and PA14 and the hypervirulent strain LESB58 in a rat model of chronic lung infection. *Journal of bacteriology* 190:2804–13.

Kuzniar A, Van Ham RCHJ, Pongor S, Leunissen JAM. 2008. The quest for orthologs: finding the corresponding gene across genomes. *Trends in genetics : TIG* 24:539–51.

Langille MGI, Laird MR, Hsiao WWL, Chiu TA, Eisen JA, Brinkman FSL. 2012. MicrobeDB: a locally maintainable database of microbial genomic sequences. *Bioinformatics* 28:1947–8.

Lavoie H, Hogues H, Mallick J, Sellam A, Nantel A, Whiteway M. 2010. Evolutionary tinkering with conserved components of a transcriptional regulatory network. *PLoS biology* 8:e1000329.

Lee TI. 2002. Transcriptional Regulatory Networks in Saccharomyces cerevisiae. *Science* 298:799–804.

Lemoine F, Lespinet O, Labedan B. 2007. Assessing the evolutionary rate of positional orthologous genes in prokaryotes using synteny data. *BMC evolutionary biology* 7:237.

Li H, Coghlan A, Ruan J, Coin LJ, Hériché J-K, Osmotherly L, Li R, Liu T, Zhang Z, Bolund L, *et al.* 2006. TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic acids research* 34:D572–80.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research* 13:2178–89.

Lima T, Auchincloss AH, Coudert E, Keller G, Michoud K, Rivoire C, Bulliard V, De Castro E, Lachaize C, Baratin D, *et al.* 2009. HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. *Nucleic acids research* 37:D471–8.

Lynn DJ, Winsor GL, Chan C, Richard N, Laird MR, Barsky A, Gardy JL, Roche FM, Chan THW, Shah N, *et al.* 2008. InnateDB: facilitating systems-level analyses of the mammalian innate immune response. *Molecular systems biology* 4:218.

Manos J, Arthur J, Rose B, Bell S, Tingpej P, Hu H, Webb J, Kjelleberg S, Gorrell MD, Bye P, *et al.* 2009. Gene expression characteristics of a cystic fibrosis epidemic

strain of Pseudomonas aeruginosa during biofilm and planktonic growth. *FEMS microbiology letters* 292:107–14.

Manos J, Arthur J, Rose B, Tingpej P, Fung C, Curtis M, Webb JS, Hu H, Kjelleberg S, Gorrell MD, *et al.* 2008. Transcriptome analyses and biofilm-forming characteristics of a clonal Pseudomonas aeruginosa from the cystic fibrosis lung. *Journal of medical microbiology* 57:1454–65.

Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, *et al.* 2012. IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic acids research* 40:D115–22.

Martin DE, Soulard A, Hall MN. 2004. TOR regulates ribosomal protein gene expression via PKA and the Forkhead transcription factor FHL1. *Cell* 119:969–79.

McCall MN, Bolstad BM, Irizarry RA. 2010. Frozen robust multiarray analysis (fRMA). *Biostatistics* 11:242–53.

Meinshausen N. 2008. Hierarchical testing of variable importance. *Biometrika* 95:265–278.

Merkeev I V, Novichkov PS, Mironov AA. 2006. PHOG: a database of supergenomes built from proteome complements. *BMC evolutionary biology* 6:52.

Min JE, Whiteside MD, Brinkman FSL, McNeney B, Graham J. 2011. A statistical approach to high-throughput screening of predicted orthologs. *Computational Statistics & Data Analysis* 55:935–943.

Mira A, Ochman H, Moran NA. 2001. Deletional bias and the evolution of bacterial genomes. *Trends in genetics : TIG* 17:589–96.

Nair R, Rost B. 2002. Sequence conserved for subcellular localization. *Protein science : a publication of the Protein Society* 11:2836–47.

Naughton S, Parker D, Seemann T, Thomas T, Turnbull L, Rose B, Bye P, Cordwell S, Whitchurch C, Manos J. 2011. Pseudomonas aeruginosa AES-1 Exhibits Increased Virulence Gene Expression during Chronic Infection of Cystic Fibrosis Lung. *PloS one* 6:e24526.

Nehrt NL, Clark WT, Radivojac P, Hahn MW. 2011. Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS computational biology* 7:e1002073.

Ostlund G, Schmitt T, Forslund K, Köstler T, Messina DN, Roopra S, Frings O, Sonnhammer ELL. 2010. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic acids research* 38:D196–203.

O'Brien KP, Remm M, Sonnhammer ELL. 2005. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic acids research* 33:D476–80.

O'Carroll MR, Syrmis MW, Wainwright CE, Greer RM, Mitchell P, Coulter C, Sloots TP, Nissen MD, Bell SC. 2004. Clonal strains of Pseudomonas aeruginosa in paediatric and adult cystic fibrosis units. *The European respiratory journal : official journal of the European Society for Clinical Respiratory Physiology* 24:101–6.

O'Connor T, Sundberg K, Carroll H, Clement M, Snell Q. 2010. Analysis of long branch extraction and long branch shortening. *BMC genomics* 11 Suppl 2:S14.

Parra G, Bradnam K, Ning Z, Keane T, Korf I. 2009. Assessing the gene space in draft genomes. *Nucleic acids research* 37:289–97.

Patel T. 2001. Molecular biological approaches to unravel adenylyl cyclase signaling and function. *Gene* 269:13–25.

Penel S, Arigon A-M, Dufayard J-F, Sertier A-S, Daubin V, Duret L, Gouy M, Perrière G. 2009. Databases of homologous gene families for comparative genomics. *BMC bioinformatics* 10 Suppl 6:S3.

Peterson ME, Chen F, Saven JG, Roos DS, Babbitt PC, Sali A. 2009. Evolutionary constraints on structural similarity in orthologs and paralogs. *Protein science : a publication of the Protein Society* 18:1306–15.

Powell S, Szklarczyk D, Trachana K, Roth A, Kuhn M, Muller J, Arnold R, Rattei T, Letunic I, Doerks T, *et al.* 2012a. eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic acids research* 40:D284–9.

Powell S, Szklarczyk D, Trachana K, Roth A, Kuhn M, Muller J, Arnold R, Rattei T, Letunic I, Doerks T, *et al.* 2012b. eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic acids research* 40:D284–9.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. *PloS one* 5:e9490.

Pruitt KD, Tatusova T, Brown GR, Maglott DR. 2012. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic acids research* 40:D130–5.

Pryszcz LP, Huerta-Cepas J, Gabaldón T. 2011. MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. *Nucleic acids research* 39:e32.

Pál C, Papp B, Lercher MJ. 2005. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nature genetics* 37:1372–5.

Remm M, Storm CE, Sonnhammer EL. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of molecular biology* 314:1041–52.

Rohmer L, Hocquet D, Miller SI. 2011. Are pathogenic bacteria just looking for food? Metabolism and microbial pathogenesis. *Trends in microbiology* 19:341–8.

Romero P, Karp P. 2003. PseudoCyc, a pathway-genome database for Pseudomonas aeruginosa. *Journal of molecular microbiology and biotechnology* 5:230–9.

Rompf A, Hungerer C, Hoffmann T, Lindenmeyer M, Römling U, Gross U, Doss MO, Arai H, Igarashi Y, Jahn D. 1998. Regulation of Pseudomonas aeruginosa hemF and hemN by the dual action of the redox response regulators Anr and Dnr. *Molecular microbiology* 29:985–97.

Roth ACJ, Gonnet GH, Dessimoz C. 2008. Algorithm of OMA for large-scale orthology inference. *BMC bioinformatics* 9:518.

Salunkhe P, Töpfer T, Buer J, Tümmler B. 2005. Genome-wide transcriptional profiling of the steady-state response of Pseudomonas aeruginosa to hydrogen peroxide. *Journal of bacteriology* 187:2565–72.

Sammut SJ, Finn RD, Bateman A. 2008. Pfam 10 years on: 10,000 families and still growing. *Briefings in bioinformatics* 9:210–9.

Sauer K, Cullen MC, Rickard AH, Zeef LAH, Davies DG, Gilbert P. 2004. Characterization of nutrient-induced dispersion in Pseudomonas aeruginosa PAO1 biofilm. *Journal of bacteriology* 186:7312–26.

Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, *et al.* 2009. Database resources of the National Center for Biotechnology Information. *Nucleic acids research* 37:D5–15.

Schiöth HB, Fredriksson R. 2005. The GRAFS classification system of G-protein coupled receptors in comparative perspective. *General and comparative endocrinology* 142:94–101.

Schmitt T, Messina DN, Schreiber F, Sonnhammer ELL. 2011. Letter to the editor: SeqXML and OrthoXML: standards for sequence and orthology information. *Briefings in bioinformatics* 12:485–8.

Schneider A, Dessimoz C, Gonnet GH. 2007. OMA Browser--exploring orthologous relations across 352 complete genomes. *Bioinformatics* 23:2180–2.

Schrettl M, Bignell E, Kragl C, Joechl C, Rogers T, Arst HN, Haynes K, Haas H. 2004. Siderophore biosynthesis but not reductive iron assimilation is essential for Aspergillus fumigatus virulence. *The Journal of experimental medicine* 200:1213–9.

Schrettl M, Bignell E, Kragl C, Sabiha Y, Loss O, Eisendle M, Wallner A, Arst HN, Haynes K, Haas H. 2007. Distinct roles for intra- and extracellular siderophores during Aspergillus fumigatus infection. *PLoS pathogens* 3:1195–207.

Schrettl M, Kim HS, Eisendle M, Kragl C, Nierman WC, Heinekamp T, Werner ER, Jacobsen I, Illmer P, Yi H, *et al.* 2008. SreA-mediated iron regulation in Aspergillus fumigatus. *Molecular microbiology* 70:27–43.

Scott FW, Pitt TL. 2004. Identification and characterization of transmissible Pseudomonas aeruginosa strains in cystic fibrosis patients in England and Wales. *Journal of medical microbiology* 53:609–15.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* 13:2498–504.

Shi G, Zhang L, Jiang T. 2010. MSOAR 2.0: Incorporating tandem duplications into ortholog assignment based on genome rearrangement. *BMC bioinformatics* 11:10.

Shi J, Malik J. 2000. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22:888–905.

Sjölander K, Datta RS, Shen Y, Shoffner GM. 2011. Ortholog identification in the presence of domain architecture rearrangement. *Briefings in bioinformatics* 12:413–22.

Smyth GK. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology* 3:Article3.

Sokurenko E V, Hasty DL, Dykhuizen DE. 1999. Pathoadaptive mutations: gene loss and variation in bacterial pathogens. *Trends in microbiology* 7:191–5.

Sonnhammer ELL, Koonin E V. 2002. Orthology, paralogy and proposed classification for paralog subtypes. *Trends in genetics : TIG* 18:619–20.

Srivastava M, Begovic E, Chapman J, Putnam NH, Hellsten U, Kawashima T, Kuo A, Mitros T, Salamov A, Carpenter ML, *et al.* 2008. The Trichoplax genome and the nature of placozoans. *Nature* 454:955–60.

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* 100:9440–5.

Storm CE V, Sonnhammer ELL. 2003. Comprehensive analysis of orthologous protein domains using the HOPS database. *Genome research* 13:2353–62.

Studer RA, Robinson-Rechavi M. 2009. How confident can we be that orthologs are similar, but paralogs differ? *Trends in genetics : TIG* 25:210–6.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin E V, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, *et al.* 2003. The COG database: an updated version includes eukaryotes. *BMC bioinformatics* 4:41.

Tatusov RL, Natale DA, Garkavtsev I V, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin E V. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic acids research* 29:22–8.

Thomas CM, Nielsen KM. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature reviews. Microbiology* 3:711–21.

Thomas PD, Wood V, Mungall CJ, Lewis SE, Blake JA. 2012. On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report. *PLoS computational biology* 8:e1002386.

Thomas-Chollier M, Defrance M, Medina-Rivera A, Sand O, Herrmann C, Thieffry D, Van Helden J. 2011. RSAT 2011: regulatory sequence analysis tools. *Nucleic acids research* 39:W86–91.

Tingpej P, Smith L, Rose B, Zhu H, Conibear T, Al Nassafi K, Manos J, Elkins M, Bye P, Willcox M, *et al.* 2007. Phenotypic characterization of clonal and nonclonal Pseudomonas aeruginosa strains isolated from lungs of adults with cystic fibrosis. *Journal of clinical microbiology* 45:1697–704.

Trunk K, Benkert B, Quäck N, Münch R, Scheer M, Garbe J, Jänsch L, Trost M, Wehland J, Buer J, *et al.* 2010. Anaerobic adaptation in Pseudomonas aeruginosa: definition of the Anr and Dnr regulons. *Environmental microbiology* 12:1719–33.

Uchiyama I, Higuchi T, Kawai M. 2010. MBGD update 2010: toward a comprehensive resource for exploring microbial genome diversity. *Nucleic acids research* 38:D361–5.

Uchiyama I. 2006. Hierarchical clustering algorithm for comprehensive orthologous-domain classification in multiple genomes. *Nucleic acids research* 34:647–58.

Vihinen M. 2012. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. *BMC genomics* 13 Suppl 4:S2.

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome research* 19:327–35.

Wall DP, Fraser HB, Hirsh AE. 2003. Detecting putative orthologs. *Bioinformatics* 19:1710–1711.

Wapinski I, Pfiffner J, French C, Socha A, Thompson DA, Regev A. 2010. Gene duplication and the evolution of ribosomal protein gene regulation in yeast. *Proceedings of the National Academy of Sciences of the United States of America* 107:5505–10.

Weinberg ED. 2009. Iron availability and infection. *Biochimica et biophysica acta* 1790:600–5.

Willoughby D, Cooper DMF. 2007. Organization and Ca2+ regulation of adenylyl cyclases in cAMP microdomains. *Physiological reviews* 87:965–1010.

Winsor GL, Khaira B, Van Rossum T, Lo R, Whiteside MD, Brinkman FSL. 2008. The Burkholderia Genome Database: facilitating flexible queries and comparative analyses. *Bioinformatics* 24:2803–4.

Winsor GL, Lam DKW, Fleming L, Lo R, Whiteside MD, Yu NY, Hancock REW, Brinkman FSL. 2010. Pseudomonas Genome Database: improved comparative analysis and population genomics capability for Pseudomonas genomes. *Nucleic acids research* 39:D596–600.

Winsor GL, Lo R, Ho Sui SJ, Ung KSE, Huang S, Cheng D, Ching W-KH, Hancock REW, Brinkman FSL. 2005. Pseudomonas aeruginosa Genome Database and PseudoCAP: facilitating community-based, continually updated, genome annotation. *Nucleic acids research* 33:D338–43.

Winsor GL, Van Rossum T, Lo R, Khaira B, Whiteside MD, Hancock REW, Brinkman FSL. 2009. Pseudomonas Genome Database: facilitating user-friendly, comprehensive comparisons of microbial genomes. *Nucleic acids research* 37:D483–8.

Winstanley C, Langille MGI, Fothergill JL, Kukavica-Ibrulj I, Paradis-Bleau C, Sanschagrin F, Thomson NR, Winsor GL, Quail MA, Lennard N, *et al.* 2009. Newly introduced genomic prophage islands are critical determinants of in vivo competitiveness in the Liverpool Epidemic Strain of Pseudomonas aeruginosa. *Genome research* 19:12–23.

Xu Z, Hao B. 2009. CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes. *Nucleic acids research* 37:W174–8.

Yu C, Desai V, Cheng L, Reifman J. 2012. QuartetS-DB: a large-scale orthology database for prokaryotes and eukaryotes inferred by evolutionary evidence. *BMC bioinformatics* 13:143.

Yu C, Zavaljevski N, Desai V, Reifman J. 2011. QuartetS: a fast and accurate algorithm for large-scale orthology detection. *Nucleic acids research* 39:e88.

Yu NY, Laird MR, Spencer C, Brinkman FSL. 2011. PSORTdb--an expanded, auto-updated, user-friendly protein subcellular localization database for Bacteria and Archaea. *Nucleic acids research* 39:D241–4.

Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster LJ, *et al.* 2010. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* 26:1608–15.

Zmasek C, Eddy S. 2002. RIO: Analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics* 3:14.

Zmasek CM, Eddy SR. 2001. A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics* 17:821–828.

Appendices

Appendix A.

Metazoan-Associated Ortholog Groups

Metazoan-associated ortholog groups were identified using the resource OrthoMCL. Listed below are the OrthoMCL ortholog group IDs for the ortholog groups that satisfy the criteria for being classified as metazoan-associated. Groups were organized into those that have a *Trichoplax adherens* ortholog and those that lack a *T. adherens* ortholog. *T. adherens* is the most basal metazoan species used in the analysis.

Metazoan-Associated Ortholog Groups with *T. adherens* Ortholog
OG4_10417, OG4_15922, OG4_17989, OG4_14170, OG4_11247, OG4_13964, OG4_15565, OG4_17405, OG4_14841, OG4_14085, OG4_14343, OG4_13535, OG4_16696, OG4_13420, OG4_16383, OG4_16496, OG4_17229, OG4_17401, OG4_17093, OG4_15415, OG4_17938, OG4_18010, OG4_16699, OG4_15802, OG4_18048, OG4_17126, OG4_13873, OG4_15126, OG4_17332, OG4_14665, OG4_17105, OG4_13785, OG4_18227, OG4_15398, OG4_16288, OG4_14646, OG4_16845, OG4_14277, OG4_13686, OG4_15289, OG4_17608, OG4_16283, OG4_17099, OG4_10481, OG4_15327, OG4_13180, OG4_14163, OG4_12206, OG4_14174, OG4_14005, OG4_16726, OG4_14497, OG4_14385, OG4_12194, OG4_17982, OG4_17301, OG4_14525, OG4_15744, OG4_15547, OG4_17749, OG4_15676, OG4_14055, OG4_17304, OG4_13427, OG4_13010, OG4_16019, OG4_14515, OG4_17349, OG4_15402, OG4_14666, OG4_15404, OG4_15194, OG4_17614, OG4_16641, OG4_14757, OG4_13923, OG4_14270, OG4_13993, OG4_16668, OG4_16427, OG4_17146, OG4_10554, OG4_14114, OG4_13632, OG4_17439, OG4_15627, OG4_16939, OG4_16715, OG4_13985, OG4_14264, OG4_15041, OG4_16716, OG4_14884, OG4_16890, OG4_16052, OG4_16953, OG4_13895, OG4_14286, OG4_13696, OG4_13533, OG4_16877, OG4_14181, OG4_15027, OG4_16246, OG4_18306, OG4_16027, OG4_14077, OG4_15457, OG4_17956, OG4_17342, OG4_15104, OG4_16417, OG4_15307, OG4_14974, OG4_15616, OG4_17972, OG4_16647, OG4_18269, OG4_12358, OG4_15312, OG4_12735, OG4_12786, OG4_14495, OG4_17967, OG4_15631, OG4_15314, OG4_16225, OG4_16076, OG4_14894, OG4_17447, OG4_17298, OG4_18186, OG4_15629, OG4_14891, OG4_17091, OG4_15157, OG4_14508, OG4_12802, OG4_16416, OG4_15466, OG4_16399, OG4_17646, OG4_11780, OG4_14881, OG4_16594, OG4_14627, OG4_15022, OG4_17070, OG4_17062, OG4_17130, OG4_16222, OG4_18880, OG4_15620, OG4_17917, OG4_17682, OG4_15951, OG4_17881, OG4_17082, OG4_17671, OG4_17645, OG4_18274, OG4_17408, OG4_14780, OG4_14645, OG4_14715, OG4_17059, OG4_16508, OG4_16285, OG4_16040, OG4_13604, OG4_16444, OG4_16467, OG4_13413, OG4_17427, OG4_11979, OG4_16872, OG4_15319, OG4_17631, OG4_16210, OG4_15564, OG4_16901, OG4_17600, OG4_13252, OG4_16718, OG4_17594, OG4_18030, OG4_16663, OG4_15339, OG4_18666, OG4_17729, OG4_15408, OG4_16275, OG4_18223, OG4_16691, OG4_15926, OG4_16933, OG4_16487, OG4_17333, OG4_14965, OG4_12652, OG4_14154, OG4_16723, OG4_17178, OG4_17136, OG4_15609, OG4_15577, OG4_16728, OG4_17415, OG4_17719, OG4_15749, OG4_14742, OG4_16720, OG4_13817, OG4_14536, OG4_15276, OG4_17394, OG4_18208, OG4_17445, OG4_15580, OG4_14500, OG4_14906, OG4_18895, OG4_13783, OG4_17613, OG4_17393, OG4_13874, OG4_14729, OG4_18945, OG4_17976, OG4_16697, OG4_13041, OG4_16893, OG4_16841, OG4_14540, OG4_11519, OG4_17306, OG4_17658, OG4_15073, OG4_14835, OG4_14877, OG4_14659, OG4_18678, OG4_14148, OG4_17744, OG4_15262, OG4_15898, OG4_16838, OG4_16606, OG4_12917, OG4_18039, OG4_18571, OG4_17676, OG4_17642, OG4_16654, OG4_17159, OG4_18008, OG4_16046, OG4_15153, OG4_17165, OG4_15274, OG4_17381, OG4_16941, OG4_17891, OG4_17808, OG4_16687, OG4_17369, OG4_17928, OG4_15633, OG4_15033, OG4_15175, OG4_16549, OG4_16485, OG4_16730, OG4_12841, OG4_14651, OG4_17700, OG4_15419, OG4_17383, OG4_14716, OG4_15793, OG4_15280, OG4_16660, OG4_16439, OG4_15275, OG4_16048, OG4_17391, OG4_17081, OG4_16133, OG4_13963, OG4_11307, OG4_15759, OG4_15912, OG4_13643, OG4_16130, OG4_16386, OG4_14762,

OG4_13634, OG4_13241, OG4_16929, OG4_16499, OG4_17897, OG4_16897, OG4_16932, OG4_16602, OG4_16253, OG4_17690, OG4_17889, OG4_14896, OG4_17757, OG4_15459, OG4_12135, OG4_15639, OG4_13708, OG4_13256, OG4_15892, OG4_17380, OG4_15295, OG4_17922, OG4_12740, OG4_16861, OG4_17587, OG4_16032, OG4_18214, OG4_18275, OG4_13914, OG4_17150, OG4_14967, OG4_18925, OG4_14764, OG4_15558, OG4_18578, OG4_15422, OG4_17368, OG4_15192, OG4_16634, OG4_16919, OG4_16224, OG4_14284, OG4_15023, OG4_15603, OG4_18047, OG4_17813, OG4_18246, OG4_16690, OG4_14771, OG4_10204
Metazoan-Associated Ortholog Groups without *T. adherens* Ortholog
OG4_16099, OG4_15568, OG4_18273, OG4_19022, OG4_16628, OG4_13051, OG4_18595, OG4_17625, OG4_17655, OG4_14602, OG4_15106, OG4_16068, OG4_17075, OG4_17362, OG4_17189, OG4_13137, OG4_17353, OG4_18233, OG4_14288, OG4_15677, OG4_18263, OG4_17364, OG4_17741, OG4_17140, OG4_16478, OG4_12170, OG4_16494, OG4_15487, OG4_15463, OG4_13991, OG4_17670, OG4_16956, OG4_13990, OG4_17918, OG4_14968, OG4_15934, OG4_15805, OG4_17352, OG4_15762, OG4_15101, OG4_15562, OG4_16415, OG4_17389, OG4_15911, OG4_17691, OG4_15942, OG4_13046, OG4_15593, OG4_16449, OG4_18217, OG4_15039, OG4_19121, OG4_17644, OG4_12338, OG4_16081, OG4_16878, OG4_17068, OG4_16479, OG4_15127, OG4_14725, OG4_12250, OG4_15945, OG4_16395, OG4_16617, OG4_16103, OG4_16450, OG4_14732, OG4_18278, OG4_16058, OG4_15109, OG4_15001, OG4_17158, OG4_15811, OG4_17309, OG4_16944, OG4_16428, OG4_17623, OG4_16729, OG4_17617, OG4_18020, OG4_17406, OG4_16436, OG4_16952, OG4_13334, OG4_18260, OG4_17980, OG4_17588, OG4_17815, OG4_17946, OG4_15950, OG4_14079, OG4_13439, OG4_16145, OG4_15414, OG4_17737, OG4_16927, OG4_18015, OG4_17742, OG4_13790, OG4_17675, OG4_16445, OG4_14470, OG4_15020, OG4_14417, OG4_13073, OG4_17310, OG4_17702, OG4_18577, OG4_11584, OG4_17751, OG4_15445, OG4_17988, OG4_16945, OG4_13796, OG4_15140, OG4_16087, OG4_18277, OG4_17442, OG4_18013, OG4_17135, OG4_16264, OG4_16026, OG4_12448, OG4_16139, OG4_17996, OG4_15407, OG4_18002, OG4_15177, OG4_16248, OG4_14829, OG4_17948, OG4_17430, OG4_15003, OG4_16411, OG4_19120, OG4_17884, OG4_16879, OG4_13913, OG4_13175, OG4_14110, OG4_15490, OG4_14631, OG4_13962, OG4_16091, OG4_17418, OG4_14125, OG4_14429, OG4_17999, OG4_17721, OG4_16119, OG4_14523, OG4_17374, OG4_14520, OG4_17930, OG4_17954, OG4_17166, OG4_16595, OG4_17745, OG4_16070, OG4_17331, OG4_16085, OG4_16256, OG4_12721, OG4_15148, OG4_18310, OG4_18592, OG4_14741, OG4_17707, OG4_17089, OG4_17113, OG4_16472, OG4_18913, OG4_18004, OG4_15777, OG4_14423, OG4_16424, OG4_17743, OG4_14621, OG4_14765, OG4_12611

Appendix B.

Pathways Over-Represented with Metazoan-Associated Genes

H. sapiens annotated pathways were tested for over-representation of metazoan-associated genes. Listed below are the statistically significant pathways that contain a higher than expected proportion of *H. sapiens* metazoan-associated genes (pathways were declared significant if the fisher's exact test *p*-value is less the 0.05 after Benjamini-Hochburg multiple hypothesis correction). Pathways were obtained from InnateDB; a database that has integrated pathways from numerous other sources. The number of *H. sapiens* metazoan-associated genes with (+ *T. adherens*) and without (- *T. adherens*) a *T. adherens* ortholog is shown for each pathway.

InnateDB Pathway	Metazoan-Associated + *T.adherens* ortholog		Metazoan-Associated - *T.adherens* ortholog	
	# Genes	P-value	# Genes	P-value
Cell Communication and Adherence Pathways				
Tight junction	23	5.11E-08	4	3.30E-01
Focal adhesion	23	5.27E-05	12	5.42E-03
Adherens junction	12	4.00E-04	5	6.24E-02
Apoptotic cleavage of cell adhesion proteins	5	4.48E-04	0	1.00E+00
Nephrin/Neph1 signaling in the kidney podocyte	7	4.53E-04	2	1.86E-01
Gap junction	13	5.05E-04	3	3.47E-01
E-cadherin signaling in the nascent adherens junction	7	4.21E-03	1	5.60E-01
Integrin cell surface interactions	8	8.77E-03	7	1.83E-03
ECM-receptor interaction	10	1.18E-02	7	1.27E-02
Arf6 trafficking events	6	2.32E-02	0	1.00E+00
A4b1 and a4b7 Integrin signaling	2	2.57E-02	0	1.00E+00
Tight junction interactions	4	2.65E-02	1	3.94E-01
Nectin adhesion pathway	5	2.80E-02	2	2.09E-01
N-cadherin signaling events	5	3.50E-02	3	8.20E-02
Cell to cell adhesion signaling	3	4.05E-02	1	3.13E-01
A6b1 and a6b4 Integrin signaling	5	1.00E-01	5	1.35E-02
Integrin signaling pathway	4	1.20E-01	4	2.62E-02
Alpha6Beta4Integrin	5	1.78E-01	5	2.86E-02
PECAM1 interactions	1	3.74E-01	2	3.73E-02
Cell Death Pathways				
Apoptotic cleavage of cellular proteins	4	1.49E-02	2	1.06E-01
Fas signaling pathway (cd95)	4	3.05E-02	0	1.00E+00
Ceramide signaling pathway	5	4.16E-02	0	1.00E+00
Breakdown of the nuclear lamina	0	1.00E+00	2	1.36E-02

InnateDB Pathway	Metazaon-Associated + *T.adherens* ortholog		Metazaon-Associated - *T.adherens* ortholog	
	# Genes	P-value	# Genes	P-value
Cell Motility Pathways				
Regulation of actin cytoskeleton	19	5.19E-03	19	1.12E-06
Pkc-catalyzed phosphorylation of inhibitory phosphoprotein of myosin phosphatase	5	8.64E-03	1	4.25E-01
Circulatory System Pathways				
Vascular smooth muscle contraction	17	8.25E-05	7	4.50E-02
Development, Cell Differentiation & Proliferation Pathays				
Signaling by BMP	9	4.60E-07	0	1.00E+00
TGF-beta signaling pathway	15	2.27E-05	0	1.00E+00
Wnt	16	5.88E-05	8	1.15E-02
BMP receptor signaling	9	4.06E-04	0	1.00E+00
ErbB signaling pathway	13	4.29E-04	7	1.36E-02
Canonical Wnt signaling pathway	8	7.15E-04	0	1.00E+00
Wnt signaling pathway	16	2.72E-03	3	5.78E-01
Multi-step regulation of transcription by pitx2	6	4.81E-03	0	1.00E+00
Alk in cardiac myocytes	6	4.81E-03	0	1.00E+00
Wnt signaling pathway	6	7.70E-03	0	1.00E+00
Inhibition of cellular proliferation by gleevec	5	1.03E-02	1	4.32E-01
Glypican 3 network	3	1.77E-02	0	1.00E+00
Presenilin action in Notch and Wnt signaling	6	4.04E-02	0	1.00E+00
EGFR1	14	4.25E-02	6	2.05E-01
Notch signaling pathway	6	4.47E-02	0	1.00E+00
Signaling events mediated by VEGFR1 and VEGFR2	5	2.17E-01	6	1.23E-02
Egf signaling pathway	2	3.38E-01	3	4.04E-02
FGF signaling pathway	3	4.59E-01	6	5.05E-03
Vegf hypoxia and angiogenesis	2	4.88E-01	4	2.18E-02
Signaling events mediated by PTP1B	3	4.89E-01	5	2.30E-02
PDGFR-beta signaling pathway	3	5.28E-01	6	8.77E-03
Role of erbb2 in signal transduction and oncology	1	7.48E-01	5	4.12E-03
Gab1 signalosome	0	1.00E+00	3	1.41E-02
FGFR1c ligand binding and activation	0	1.00E+00	2	3.73E-02
FGFR3b ligand binding and activation	0	1.00E+00	2	3.73E-02
FGFR3c ligand binding and activation	0	1.00E+00	2	3.73E-02
FGFR4 ligand binding and activation	0	1.00E+00	2	3.73E-02
FGFR1c and Klotho ligand binding and activation	0	1.00E+00	2	4.74E-02
ErbB receptor signaling network	0	1.00E+00	4	2.39E-03
Grb2 events in EGFR signaling	0	1.00E+00	2	4.74E-02

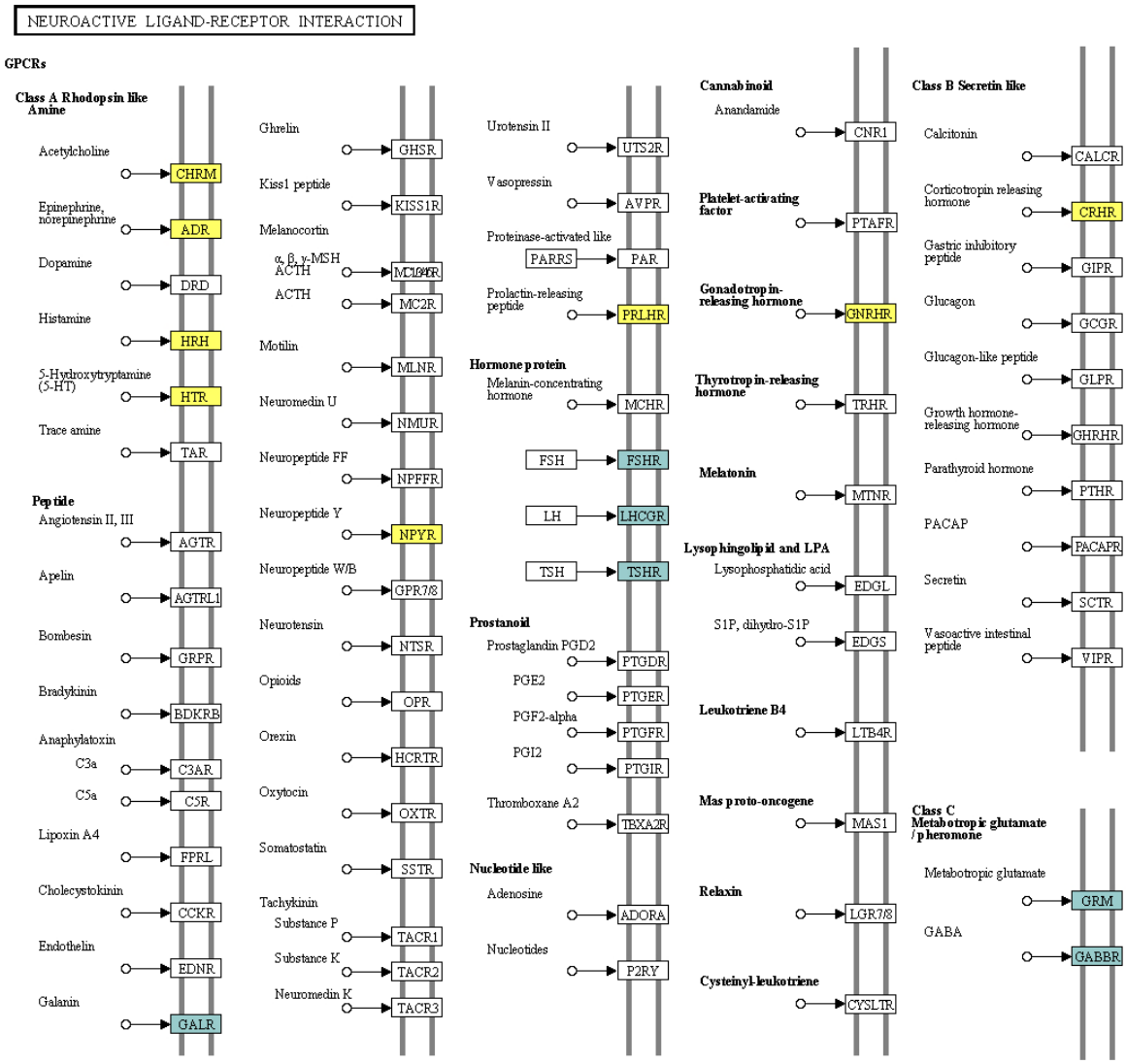
InnateDB Pathway	Metazaon-Associated + *T.adherens* ortholog		Metazaon-Associated - *T.adherens* ortholog	
	# Genes	P-value	# Genes	P-value
EGFR interacts with phospholipase C-gamma	0	1.00E+00	2	1.36E-02
PDGFR-alpha signaling pathway	0	1.00E+00	3	4.89E-02
Endochrine System Pathways				
GnRH signaling pathway	17	6.15E-06	4	2.12E-01
Progesterone-mediated oocyte maturation	15	1.85E-04	0	1.00E+00
Rapid glucocorticoid signaling	3	1.77E-02	0	1.00E+00
Hormone ligand-binding receptors	3	4.59E-02	1	3.31E-01
Glycan Biosynthesis Pathways				
Chondroitin sulfate biosynthesis	0	1.00E+00	4	1.08E-02
O-Glycan biosynthesis	0	1.00E+00	4	2.18E-02
Immune System Pathways				
Chemokine signaling pathway	19	1.68E-03	7	1.51E-01
Leukocyte transendothelial migration	13	5.02E-03	2	6.47E-01
B cell survival pathway	4	9.64E-03	0	1.00E+00
IL2-mediated signaling events	7	3.00E-02	1	6.64E-01
TNF receptor signaling pathway	6	4.04E-02	0	1.00E+00
BCR	12	4.29E-02	6	1.25E-01
Tnfr1 signaling pathway	3	9.80E-02	3	2.97E-02
IL6	3	6.03E-01	6	1.34E-02
T cell receptor signaling pathway	0	1.00E+00	4	2.92E-02
Nervous System Pathways				
Axon guidance	36	7.06E-19	7	5.36E-02
Agrin in postsynaptic differentiation	11	2.93E-05	1	6.20E-01
EPHA forward signaling	8	2.79E-04	1	5.20E-01
Regulation of Commissural axon pathfinding by Slit and Robo	3	2.44E-03	0	1.00E+00
Ephrin B reverse signaling	6	4.81E-03	3	6.88E-02
EPHB forward signaling	7	5.43E-03	2	2.70E-01
Long-term potentiation	8	2.79E-02	2	4.45E-01
Reelin signaling pathway	4	9.39E-02	5	4.46E-03
Neuroactive ligand-receptor interaction	12	6.36E-01	15	6.32E-03
Signal Transduction Pathways				
Adenylate cyclase activating pathway	9	1.30E-09	0	1.00E+00
Adenylate cyclase inhibitory pathway	9	3.60E-09	0	1.00E+00
LPA4-mediated signaling events	9	4.69E-08	0	1.00E+00
G alpha (z) signalling events	8	4.91E-08	0	1.00E+00
PKA activation in glucagon signalling	9	2.82E-07	0	1.00E+00

InnateDB Pathway	Metazaon-Associated + *T.adherens* ortholog		Metazaon-Associated - *T.adherens* ortholog	
	# Genes	P-value	# Genes	P-value
G alpha (i) signalling events	8	4.74E-07	0	1.00E+00
G(s)-alpha mediated events in glucagon signalling	10	1.19E-06	2	1.99E-01
LPA receptor mediated events	14	4.08E-06	6	1.39E-02
G alpha (s) signalling events	9	9.12E-06	0	1.00E+00
Endothelins	13	1.04E-05	2	4.08E-01
PKA activation	8	1.05E-05	0	1.00E+00
Class C/3 (Metabotropic glutamate/pheromone receptors)	7	2.39E-05	0	1.00E+00
Syndecan-2-mediated signaling events	6	1.18E-02	2	2.35E-01
P38MAPK events	3	1.25E-02	0	1.00E+00
Inactivation of Cdc42 and Rac	3	1.77E-02	0	1.00E+00
Regulation of retinoblastoma protein	8	1.88E-02	0	1.00E+00
MAPK signaling pathway	20	2.66E-02	8	2.22E-01
Mapkinase signaling pathway	7	3.00E-02	0	1.00E+00
P38 signaling mediated by MAPKAP kinases	4	4.14E-02	0	1.00E+00
Ahr signal transduction pathway	2	4.15E-02	0	1.00E+00
Calcium signaling pathway	14	4.25E-02	17	1.51E-06
Phospholipase C gamma signaling	0	1.00E+00	2	3.00E-02
PLC-gamma1 signalling	0	1.00E+00	2	3.73E-02

Appendix C.

Metazoan-Associated Genes in the Neuroactive Ligand Pathways

Human Neuroendocrine G-protein Coupled Receptors. Genes that are metazoan-associated orthologs are colored: blue-colored genes are human metazoan-associated genes that have a *T. adherens* ortholog, while yellow-colored genes are human metazoan-associated genes that do not have a *T. adherens* ortholog.



Appendix D.

List of Microarray Datasets Used in Epidemic *P. aeruginosa* Meta-Analysis

Sample Accesion	Experiment	Strain	Isolate Descripton
GSM142279	GSE10304	AES-2	Clonal isolate 9
GSM142280	GSE10304	AES-2	Clonal isolate 10
GSM142282	GSE10304	AES-2	Clonal isolate 12
GSM142284	GSE6122	Non-Clonal	Non-clonal isolate 8
GSM142285	GSE6122	Non-Clonal	Non-clonal isolate 6
GSM142286	GSE6122	Non-Clonal	Non-clonal isolate 5
GSM259817	GSE6122	AES-1	Clonal isolate 2
GSM259818	GSE6122	AES-1	Clonal isolate 3
LES400 biological replicate A	Salunkhe *et al*.	LES	Clonal isolate LES400
LES400 biological replicate B	Salunkhe *et al*.	LES	Clonal isolate LES400
LES431 biological replicate A	Salunkhe *et al*.	LES	Clonal isolate LES431
LES431 biological replicate B	Salunkhe *et al*.	LES	Clonal isolate LES431
PAO1 biological replicate A	Salunkhe *et al*.	PAO1	Laboratory reference strain PAO1
PAO1 biological replicate B	Salunkhe *et al*.	PAO1	Laboratory reference strain PAO1

Appendix E.

Differentially Expressed Pathways in Epidemic *P. aeruginosa*

P. aeruginosa pathways defined in the KEGG, PseudoCAP and PseudoCyc resources that are significantly differentially expressed in epidemic versus non-clonal and PAO1 strains, as determined by the group-wise global test (FDR <= 0.05). The contribution of genes in the pathways and microarray samples to the test statistic can be extrapolated from the global test. Also listed are the effect causing genes for each pathway result (contribution FDR <= 0.05). Significant pathways were also grouped based on gene content to facilitate the viewing of common influential genes shared between pathways.

Group	Pathway	False Discovery Rate	Effect Causing Genes in Pathway	
			Up-regulated	Down-regulated
1	Urea cycle and metabolism of amino groups [PseudoCAP]	0.0049		PA5263 (argH), PA3525 (argG), PA5172 (arcB)
1	Alanine and Aspartate metabolism [PseudoCAP]	0.0049		PA5263 (argH), PA3525 (argG), PA5429 (aspA), PA2629 (purB)
1	arginine biosynthesis [PseudoCyc]	0.0187		PA5263 (argH), PA3525 (argG), PA0895 (aruC)
1	Alanine, aspartate and glutamate metabolism [KEGG]	0.0200		PA5263 (argH), PA3525 (argG), PA5429 (aspA), PA2629 (purB)
2	glutamine degradation III [PseudoCyc]	0.0049		PA5429 (aspA)
2	Synthesis of aspartate and asparagine; interconversion of aspartate and asparagine. [PseudoCyc]	0.0049		PA5429 (aspA)
3	Pyrimidine metabolism [PseudoCAP]	0.0056		PA3527 (pyrC), PA0342 (thyA), PA4646 (upp), PA3637 (pyrG)
3	de novo biosynthesis of pyrimidine ribonucleotides [PseudoCyc]	0.0076		PA3527 (pyrC), PA3050 (pyrD), PA3637 (pyrG)
3	Pyrimidine metabolism [KEGG]	0.0076		PA3527 (pyrC), PA0342 (thyA),

Group	Pathway	False Discovery Rate	Effect Causing Genes in Pathway	
			Up-regulated	Down-regulated
				PA4646 (upp), PA3637 (pyrG)
4	de novo biosynthesis of pyrimidine deoxyribonucleotides [PseudoCyc]	0.0056	PA2962 (tmk)	PA0342 (thyA), PA3480
5	Flagellar assembly [KEGG]	0.0097	PA1449 (flhB)	PA1092 (fliC), PA1095, PA1094 (fliD), PA1086 (flgK), PA1087 (flgL), PA3351 (flgM)
6	Thiamine metabolism [PseudoCAP]	0.0097	PA4973 (thiC)	PA0381 (thiG)
7	Pentose phosphate cycle [PseudoCAP]	0.0097		PA5110 (fbp), PA4670 (prs), PA0330 (rpiA)
7	non-oxidative branch of the pentose phosphate pathway [PseudoCyc]	0.0111		PA0607 (rpe), PA0330 (rpiA), PA2796 (tal)
7	Pentose phosphate pathway [KEGG]	0.0137		PA5110 (fbp), PA4670 (prs), PA0330 (rpiA), PA5322 (algC)
7	pentose phosphate pathway [PseudoCyc]	0.0200	PA3183 (zwf)	PA0607 (rpe), PA0330 (rpiA)
7	Carbon fixation [PseudoCAP]	0.0206		PA4329 (pykA), PA0607 (rpe), PA5110 (fbp), PA0330 (rpiA), PA0552 (pgk)
7	ribose degradation [PseudoCyc]	0.0307		PA0330 (rpiA)
8	polyisoprenoid biosynthesis [PseudoCyc]	0.0111		PA4569 (ispB), PA4043 (ispA)
8	Terpenoid biosynthesis [PseudoCAP]	0.0137		PA4669 (ipk), PA4043 (ispA)
8	Terpenoid backbone biosynthesis [KEGG]	0.0187	PA4785	PA4569 (ispB), PA4669 (ipk)
9	chorismate biosynthesis [PseudoCyc]	0.0111		PA1750, PA4846 (aroQ1), PA5039 (aroK), PA5038 (aroB)
10	cyanate degradation [PseudoCyc]	0.0118	PA0102	
11	Mismatch repair [KEGG]	0.0137		PA4042 (xseB), PA4232 (ssb), PA3620 (mutS)
11	Homologous recombination [KEGG]	0.0165		PA1534 (recR), PA4232 (ssb)
11	DNA replication [KEGG]	0.0179		PA4232 (ssb)
12	Purine metabolism [KEGG]	0.0146		PA3686 (adk),

Group	Pathway	False Discovery Rate	Effect Causing Genes in Pathway	
			Up-regulated	Down-regulated
				PA4329 (pykA)
12	Purine metabolism [PseudoCAP]	0.0187		PA3686 (adk), PA4329 (pykA), PA5242 (ppk)
12	his+purine+pyrimidine biosynthesis [PseudoCyc]	0.0201		PA3686 (adk), PA3527 (pyrC)
13	aspartate biosynthesis II [PseudoCyc]	0.0146	PA1400	
14	Arachidonic acid metabolism [KEGG]	0.0149	PA1287, PA2826	
15	leu+val+ile biosynthesis [PseudoCyc]	0.0149		PA4696 (ilvI)
15	isoleucine biosynthesis I [PseudoCyc]	0.0153		PA4696 (ilvI), PA4694 (ilvC)
15	Pantothenate and CoA biosynthesis [PseudoCAP]	0.0179		PA4696 (ilvI), PA0363 (coaD)
16	Protein export [KEGG]	0.0153		PA5128 (secB), PA4403 (secA), PA5070 (tatC), PA5069 (tatB)
17	biosynthesis of proto- and siroheme [PseudoCyc]	0.0164		PA3977 (hemL)
17	Porphyrin and chlorophyll metabolism [KEGG]	0.0179		PA3977 (hemL)
17	Porphyrin and chlorophyll metabolism [PseudoCAP]	0.0200		PA3977 (hemL), PA0024 (hemF)
18	ornithine spermine biosynthesis [PseudoCyc]	0.0164	PA4519 (speC)	
19	Folate biosynthesis [PseudoCAP]	0.0165		PA0342 (thyA), PA3439 (folX)
19	tetrahydromonapterin synthesis [PseudoCAP]	0.0189		PA3437 (folM), PA3439 (folX)
20	Pantothenate and CoA biosynthesis [KEGG]	0.0179		PA5320 (coaC), PA4731 (panD)
20	Beta-Alanine metabolism [PseudoCAP]	0.0187		PA4731 (panD), PA4730 (panC), PA3014 (faoA)
20	pantothenate and coenzyme A biosynthesis [PseudoCyc]	0.0246		PA4731 (panD), PA0363 (coaD)
21	Nicotinate and nicotinamide metabolism [KEGG]	0.0187		PA4919 (pncB1), PA4918
21	Nicotinate and nicotinamide metabolism [PseudoCAP]	0.0245		PA4919 (pncB1)
21	pyridine nucleotide cycling [PseudoCyc]	0.0305		PA4919 (pncB1), PA4920 (nadE)

Group	Pathway	False Discovery Rate	Effect Causing Genes in Pathway	
			Up-regulated	Down-regulated
22	Quorum sensing [PseudoCAP]	0.0187	PA2587 (pqsH), PA1431 (rsaL), PA4206 (mexH), PA4208 (opmD), PA4207 (mexI), PA4205 (mexG)	
23	Pentose and glucuronate interconversions [PseudoCAP]	0.0189		PA0607 (rpe), PA2023 (galU)
24	pyridoxal 5'-phosphate biosynthesis [PseudoCyc]	0.0191		PA1049 (pdxH), PA0593 (pdxA)
24	Vitamin B6 metabolism [PseudoCAP]	0.0275		PA1049 (pdxH), PA0593 (pdxA)
24	Vitamin B6 metabolism [KEGG]	0.0275		PA1049 (pdxH), PA0593 (pdxA)
25	ubiquinone biosynthesis [PseudoCyc]	0.0200		PA3171 (ubiG)
26	Oxidative phosphorylation [PseudoCAP]	0.0200		PA1581 (sdhC), PA1582 (sdhD), PA2639 (nuoD), PA2648 (nuoM)
27	lysine and diaminopimelate biosynthesis [PseudoCyc]	0.0200		PA5277 (lysA), PA1010 (dapA)
27	Lysine biosynthesis [KEGG]	0.0272		PA1010 (dapA)
28	leucine biosynthesis [PseudoCyc]	0.0204		PA3118 (leuB), PA5013 (ilvE)
29	Acyclic isoprenoid and branched-chain amino acids catabolism [PseudoCAP]	0.0245	PA2011 (liuE)	
30	phospholipid biosynthesis [PseudoCyc]	0.0245		PA1614 (gpsA), PA3673 (plsB)
31	aliphatic compound catabolism [PseudoCAP]	0.0249	PA5351 (rubA1)	
31	Other [PseudoCAP]	0.0249	PA5351 (rubA1)	
32	fatty acid elongation -- saturated [PseudoCyc]	0.0249		PA1610 (fabA), PA1609 (fabB)
33	fatty acid biosynthesis -- initial steps [PseudoCyc]	0.0275		PA5436, PA2965 (fabF1)
34	Biosynthesis of heme d1 [PseudoCAP]	0.0291		PA0517 (nirC), PA0515, PA0518 (nirM), PA0516 (nirF), PA0514 (nirL)
34	Denitrification [PseudoCAP]	0.0315		PA0517 (nirC), PA0515, PA0518

Group	Pathway	False Discovery Rate	Effect Causing Genes in Pathway	
			Up-regulated	Down-regulated
				(nirM), PA0519 (nirS), PA0516 (nirF), PA0514 (nirL)
35	thiamine biosynthesis [PseudoCyc]	0.0296		PA3976 (thiE)
36	Ether lipid metabolism [KEGG]	0.0307	PA4351	
37	Biosynthesis of unsaturated fatty acids [KEGG]	0.0312		PA3942 (tesB), PA4389
38	pyridine nucleotide biosynthesis [PseudoCyc]	0.0316		PA1004 (nadA), PA4920 (nadE), PA4524 (nadC)
39	thioredoxin pathway [PseudoCyc]	0.0354	PA0849 (trxB2)	PA2616 (trxB1)
40	ureide degradation [PseudoCyc]	0.0374		PA4867 (ureB), PA4865 (ureA), PA4864 (ureD)
41	O-antigen biosynthesis [PseudoCyc]	0.0405	PA5164 (rmlC), PA5163 (rmlA)	PA5552 (glmU)
42	serine biosynthesis [PseudoCyc]	0.0420		PA0316 (serA)

Appendix F.

Differentially Expressed Operons in Epidemic *P. aeruginosa*

P. aeruginosa operons consisting of 3 or more genes were tested for differential expression in epidemic strains versus non-clonal and PAO1 strains using the group-based global test statistic (results with FDR <= 0.05 are listed).

Operon	False Discovery Rate	Expression Change	PseudoCAP Function class
PA3172, PA3171 (ubiG), PA3173	0.0247	down	Energy metabolism / Biosynthesis of cofactors, prosthetic groups and carriers / Putative enzymes
PA0595 (ostA), PA0593 (pdxA), PA0594 (surA)	0.0247	down	Biosynthesis of cofactors, prosthetic groups and carriers / Translation, post-translational modification, degradation / Adaptation, Protection / Chaperones and heat shock proteins
PA2264, PA2265, PA2266	0.0247	down	Carbon compound catabolism / Energy metabolism
PA2499, PA2498, PA2500	0.0247	up	Putative enzymes / Membrane proteins / Transport of small molecules
PA0723 (coaB), PA0721, PA0722, PA0720	0.0247	down	DNA replication, recombination, modification and repair / Related to phage, transposon, or plasmid
PA2627, PA2626 (trmU), PA2628, PA2625	0.0345	down	Transcription, RNA processing and degradation / Membrane proteins
PA4535, PA4536, PA4537	0.0383	down	no annotations
PA0633, PA0635, PA0634	0.0444	down	Related to phage, transposon, or plasmid
PA0810, PA0811, PA0812	0.0444	up	Carbon compound catabolism / Membrane proteins / Transport of small molecules
PA0016 (trkA), PA0018 (fmt), PA0017	0.0496	down	Transport of small molecules / Amino acid biosynthesis and metabolism / Translation, post-translational modification, degradation
PA2409, PA2408, PA2407, PA2410	0.0496	up	Motility and Attachment / Transport of small molecules / Membrane proteins
PA0592 (ksgA), PA0589, PA0590 (apaH), PA0591	0.0496	down	Energy metabolism / Nucleotide biosynthesis and metabolism / Transcription, RNA processing and degradation
PA4502, PA4506, PA4504, PA4505, PA4503, PA4501 (opdD)	0.0496	down	Transport of small molecules / Membrane proteins

Operon	False Discovery Rate	Expression Change	PseudoCAP Function class
PA3437 (folM), PA3439 (folX), PA3438 (folE1)	0.0496	down	Putative enzymes / Biosynthesis of cofactors, prosthetic groups and carriers
PA0639, PA0638, PA0636, PA0637, PA0640	0.0496	down	Related to phage, transposon, or plasmid
PA4206 (mexH), PA4208 (opmD), PA4207 (mexI), PA4205 (mexG)	0.0496	up	Membrane proteins / Transport of small molecules
PA4430, PA4431, PA4429, PA4428 (sspA), PA4427 (sspB)	0.0496	down	Adaptation, Protection / Energy metabolism / Putative enzymes
PA0608, PA0607 (rpe), PA0609 (trpE)	0.0496	down	Energy metabolism / Carbon compound catabolism / Amino acid biosynthesis and metabolism
PA4451, PA4450 (murA), PA4449 (hisG)	0.0496	down	Amino acid biosynthesis and metabolism / Cell wall, LPS, capsule
PA0632, PA0631, PA0629, PA0630, PA0628	0.0496	down	Related to phage, transposon, or plasmid

Appendix G.

Conserved Upstream Motifs in Transcriptional Module Genes

The attached spreadsheets file (Appendix_G.xls) forms part of this work. The file can be opened with Microsoft Excel.

Genes and their associated DNA motif instances are listed for each predicted transcriptional module. Three conserved DNA motifs were identified for each module. The motifs were then used to search for motif instances in the upstream intergenic regions of the genes. These instances are listed along with their identification *p*-values. Seed genes were used to carry out the initial search for co-expressed transcriptional module genes.