ASSESSMENT OF ORTHOLOGY IDENTIFICATION APPROACHES AND THE IMPACT OF GENE FUSION AND FISSION IN BACTERIA

ASSESSMENT OF ORTHOLOGY IDENTIFICATION APPROACHES AND THE IMPACT OF GENE FUSION AND FISSION IN BACTERIA

By

WILSON SUNG, BCS (HONS.)

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfilment of the Requirements

for the Degree

Master of Science

McMaster University

MASTER OF SCIENCE (2011) McMaster University
(Biology) Hamilton, Ontario

TITLE: ASSESSMENT OF ORTHOLOGY IDENTIFICATION APPROACHES AND THE IMPACT OF GENE FUSION AND FISSION IN BACTERIA

AUTHOR: Wilson Sung, BCS Hons. (University of Waterloo)

SUPERVISOR: Dr. G. Brian Golding

NUMBER OF PAGES: ix, 144

ABSTRACT

Orthology identification is central to comparative and evolutionary genomics and is an active area of research. Despite a recent shift towards tree reconciliation and other phylogenetic methods, previous comparisons between different algorithms relied on real datasets where true orthology relationships are unknown and did not conclusively show whether phylogenetic methods truly outperform sequence similarity-based methods. Using simulated datasets generated from programs we developed, we show that tree reconciliation does perform better than similarity-based methods when the true species phylogeny is known. Even slight deviations in the species phylogeny can have adverse effects on the performance of reconciliation algorithms and in those cases similarity-based methods may perform better. Fusion and fission complicate orthology identification and are not explicitly considered in most existing algorithms. Programs designed specifically to investigate fusion and fission events are either unavailable or are not specific enough to identify events affecting orthologous genes. We developed a pipeline of programs called `FusionFinder` that perform this task, gaining new insights to the contributions of fusion and fission to bacterial protein evolution and uncover an unexpected abundance of fissions in *Bacillus anthracis* that to our knowledge yet to be reported.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisor Dr. Brian Golding for all he has done for me. He was always there to provide guidance and advice when it was needed or requested. His patience, and unwavering support and encouragement, through both the good times and bad times, was invaluable to me. It was a privilege to be under the tutelage of such a knowledgeable, experienced, and caring mentor.

I would also like to thank Dr. Wilfried Haerty for his friendship and steadfast support and encouragement. Thank you for your willingness to discuss my experiments, for providing useful suggestions and advice, and for proofreading several drafts of this thesis.

Thank you to Tallulah Andrews for contributing code for the simulation of fusion and fission for Chapter 1, and also for debugging portions of the code in the program `FusionFinder` in Chapter 2. I would also like to thank Yifei Huang for useful suggestions pertaining to both chapters.

I am grateful to the members of the Golding and Evans labs for their friendship, support, and for making my stay at McMaster fun and memorable. Although there are too many to list, I would particularly like to thank Melanie Lou, Stephanie Sun, and Dr. Terri Porter.

I would like to thank the members of my committee, Dr. Marie Elliot, Dr. Turlough Finan, and Dr. Jonathon Stone, for contributing their time and making the process as painless as possible.

Last but most certainly not least, I would like to thank my family for their support and encouragement in all my endeavors in life.

Contents

I INTRODUCTION **1**

1 Orthology identification **5**
1.1 ABSTRACT . . . 5
1.2 INTRODUCTION . . . 5
1.3 MATERIALS AND METHODS . . . 13
1.3.1 Gene tree and sequence simulation . . . 13
1.3.2 Species phylogeny . . . 15
1.3.3 Rates of macroevolutionary events . . . 15
1.3.4 Orthology identification . . . 17
1.3.5 Accuracy measurement . . . 18
1.4 RESULTS . . . 19
1.5 DISCUSSION . . . 33

2 Fusion detection **37**
2.1 ABSTRACT . . . 37
2.2 INTRODUCTION . . . 37
2.3 MATERIALS AND METHODS . . . 41
2.3.1 FusionFinder . . . 44
2.3.2 MultiFusion . . . 50
2.3.3 FusionScanner . . . 51
2.3.4 FusionMapper . . . 52
2.3.5 Species phylogeny . . . 52
2.3.6 COG statistical analysis . . . 55
2.3.7 Estimation of selective pressure on *B. anthracis* genes . . . 55
2.4 RESULTS . . . 56
2.4.1 Specific examples of fusion and fission . . . 56
2.4.2 Fusion and fission are rare in most of the *Bacillaceae* . . . 62
2.4.3 Fission, SNP, and VNTR patterns in *B. anthracis* are consistent . . . 70
2.4.4 Annotation of fissioned genes differs between *B. anthracis* genomes . . . 74
2.4.5 Gene fission is not associated with gene duplication . . . 82
2.4.6 Contribution of PlcR and motility loss to gene fission in *B. anthracis* 82
2.4.7 COG analysis of fission in *B. anthracis* . . . 86
2.4.8 Fissioned genes in *B. anthracis* may evolve under relaxed selective constraint . . . 89
2.4.9 Origin of *B. cereus* CI: lateral or vertical? . . . 91
2.4.10 Fission and SNP patterns are consistent in *Yersinia* . . . 97

2.4.11 Pseudogenization in *S. enterica* 103
2.5 DISCUSSION 107

II CONCLUSION 118

III REFERENCES 120

IV APPENDICES 139

A *P*-values for program comparisons from Chapter 1 140

B *B. anthracis* canSNPs. 144

List of Figures

1.1 Identification of orthologs from a gene tree. 9
1.2 Species phylogeny used for simulation. 16
1.3 Recall under increasing rates of lineage-specific duplication. 21
1.4 Precision and recall under increasing rates of duplication. 22
1.5 Effect of gene family clustering on `Notung` in the presence of duplication. . . 23
1.6 Effect of gene family clustering on `RAP` in the presence of duplication. . . . 24
1.7 Effect of species phylogeny `Notung` in the presence of duplication. 25
1.8 Effect of species phylogeny `RAP` in the presence of duplication. 26
1.9 Precision and recall under increasing rates of duplication and loss. 28
1.10 Effect of gene family construction on `Notung` in the presence of duplication and loss. 29
1.11 Effect of gene family construction on `RAP` in the presence of duplication and loss. 30
1.12 Effect of species phylogeny `Notung` in the presence of duplication and loss. . 31
1.13 Effect of species phylogeny `RAP` in the presence of duplication and loss. . . 32
1.14 Precision and recall under increasing rates of LGT. 34

2.1 Survey of protein length differences between homologous proteins between bacterial genomes. 40
2.2 Flowchart of the `FusionFinder` pipeline. 44
2.3 Parameters for alignment overlap assessment. 45
2.4 Detection of a fusion event depends on genome choice. 46
2.5 Hypothetical example of combining HSP prior to component selection. . . . 47
2.6 Example of fission in *H. acinonychis* with synteny data. 48
2.7 Component selection algorithm. 49
2.8 Genome scan procedure using annotated proteins. 50
2.9 Influence of orthology on fission mapping. 53
2.10 Implications of parsimony in fission mapping. 53
2.11 Results from the NCBI `CDD` for the fusion protein of *rhaA* and *rhaB*. 57
2.12 Fusion of genes in the rhamnose degradation pathway. 57
2.13 Results from the NCBI `CDD` for the fusion of two-component system proteins. 58
2.14 `TMHMM` results in *B. anthracis* Ames Ancestor. 58
2.15 `TMHMM` results in *B. anthracis* Tsiankovskii-I. 59
2.16 Results from the NCBI `CDD` for composite *secDF*. 60
2.17 `TMHMM` results for *secD* and *secF*. 60
2.18 Results from the NCBI `CDD` for component *secD*. 60

2.19 Results from the NCBI `CDD` for component *secF*. 61
2.20 Fusion and fission events across 33 complete *Bacillaceae* genomes. 63
2.21 Fusion and fission events across 33 complete *Bacillaceae* genomes with pseudogenes. 64
2.22 Fusion and fission events across 33 complete *Bacillaceae* genomes at 90% sequence identity. 65
2.23 Fusion and fission events across 33 complete *Bacillaceae* genomes with plasmid proteins. 66
2.24 Fusion and fission events across 33 complete *Bacillaceae* genomes with only groups explained by a single event. 67
2.25 Fusion and fission events across 29 complete *Bacillaceae* genomes including one *B. anthracis* genome. 68
2.26 *B. anthracis* phylogeny inferred using canSNPs. 71
2.27 Fusion/fission analysis using 12X draft *B. anthracis* genomes. 72
2.28 Fusion/fission analysis using complete *B. anthracis* genomes. 73
2.29 Fusion/fission analysis of 40 *Bacillaceae* genomes. 75
2.30 Fusion/fission analysis of 40 *Bacillaceae* genomes with pseudogenes. 76
2.31 Consensus of phylogenies inferred from fusion/fission patterns for 40 *Bacillaceae* genomes. 77
2.32 Consensus of phylogenies inferred from fusion/fission patterns for 40 *Bacillaceae* genomes with pseudogenes. 78
2.33 Unrooted tree of *motA* and *motP* paralogs inferred using `MrBayes` 84
2.34 Phylogeny tracing gene fission and loss in the motility and chemotaxis cluster in *B. cereus*. 87
2.35 d_N/d_S estimates for *B. cereus* genomes and *B. anthracis* genomes for orthologous groups and fission groups. 90
2.36 Relationship between d_N/d_S and the number of sequences. 90
2.37 Effect of increasing minimum protein identity on number of reciprocal best hits (RBH) orthologs in *B. cereus* CI. 93
2.38 Effect of increasing proportional match length on number of RBH orthologs in *B. cereus* CI. 93
2.39 Effect of increasing minimum protein identity on average protein identity of RBH orthologs in *B. cereus* CI. 97
2.40 Consensus of phylogenies inferred from fusion/fission patterns for 9 *Y. pestis* genomes. 99
2.41 Fusion/fission analysis of 9 *Y. pestis* genomes. 100
2.42 Consensus of phylogenies inferred from fusion/fission patterns for 16 *Y. pestis* genomes. 101
2.43 Fusion/fission analysis of 16 *Y. pestis* genomes. 102
2.44 Fusion/fission analysis of 6 *S. enterica* genomes. 106
2.45 Fusion/fission analysis of 10 *S. enterica* genomes. 108

List of Tables

1.1 Precision and recall at a lineage-specific duplication rate of 0.1. 20
1.2 Precision and recall at a duplication rate of 0.1. 27
1.3 Precision and recall at a duplication and loss rate of 0.1. 33

2.1 *Bacillaceae* taxa used in the study. 42
2.2 *Yersinia* taxa used in the study. 43
2.3 *S. enterica* taxa used in the study. Host specificity and type of infection is also noted. 43
2.4 CanSNP chromosomal locations. 54
2.5 Number of annotated pseudogenes in *B. anthracis*. 69
2.6 Characterization of fission events mapped to the *B. anthracis* clade. 80
2.7 Pseudogenes in *B. anthracis* genomes that were or were not mapped to fusion groups. 80
2.8 Characterization of fission events mapped to the *B. anthracis* cenancestor. . . 81
2.9 PlcR-controlled genes involved in fusion or fission events. 83
2.10 Fission and loss in motility and chemotaxis cluster in *B. cereus*. 86
2.11 Fisher's exact test of fission distribution across COGs. 88
2.12 Mean d_N/d_S estimates for genes in orthologous or fission groups for *B. cereus* and *B. anthracis* genomes 89
2.13 Fusion and orthologous groups categorized by d_N/d_S. 91
2.14 Pairwise genome analysis of *B. cereus* CI. 94
2.15 Pairwise genome analysis of *B. cereus* AH820. 95
2.16 Pairwise genome analysis of *B. anthracis* A1055. 96
2.17 Characterization of fission events mapped to the *Y. pestis* clade. 104
2.18 Characterization of fission events mapped to the *Y. pestis* cenancestor. . . . 105
2.19 Three proteins in *B. anthracis* that may have disrupted function. 114

A.1 *P*-values for precision and recall at a lineage-specific duplication rate of 0.1. . 141
A.2 *P*-values for precision and recall at a duplication rate of 0.1. 142
A.3 *P*-values for precision and recall at a duplication and loss rate of 0.1. 143

B.1 *B. anthracis* CanSNPs 144

Part I

INTRODUCTION

The number of publicly available genome sequences has increased rapidly since the publication of the first genome sequences (Fleischmann *et al.*, 1995; Fraser *et al.*, 1995) and should continue to increase in the foreseeable future, as next generation sequencing technologies continue to mature and proliferate (Smith *et al.*, 2007; Schatz, Delcher and Salzberg, 2010). This wealth of genomic data has increased our understanding of gene family evolution (Hahn, Han and Han, 2007), genome architecture and organization (Wolf *et al.*, 2001), rates of macroevolutionary events such as gene duplication (Lynch and Conery, 2000) and lateral gene transfer (LGT) (Hao and Golding, 2006), the contributions of gene fusion and fission (Snel, Bork and Huynen, 2000; Kummerfeld and Teichmann, 2005), and length variation between conserved proteins (Brocchieri and Karlin, 2005). The pace of genome sequencing also necessitates computational methods to assist functional annotation of encoded genes. These comparative and evolutionary genomics studies as well as functional annotation of genes are reliant on accurate identification of conserved genes and their evolutionary histories.

If genes evolved strictly by vertical descent with modification, then the gene complement of a genome would be static and genome comparisons would be trivial. However, genome sequencing has revealed that gene duplication and LGT play major roles in eukaryotic and prokaryotic genome evolution. The events contributing to gene evolution and their approximate order of importance according to Koonin (2005) are as follows:

1. Vertical descent
2. Gene duplication
3. Gene loss
4. LGT
5. Fusion, fission, and other rearrangements

Although considering evolutionary relationships of genes in the presence of any one of these events may be simple, all of these act in concert, making genome comparisons non-trivial especially over long evolutionary time.

Homologs are genes that share common ancestry and can be grouped collectively into gene families. Gene families may expand through duplication and LGT, resulting in paralogs and xenologs, and can contract through deletion or pseudogenization. In the context of genome comparison, it is useful to ascertain the evolution of gene families relative to the cenancestor, the most recent common ancestor of the genomes under consideration. Orthologs are genes that originated from a single gene in the cenancestor while paralogs are genes that are descendants of a duplication event (Fitch, 1970, 2000; Koonin, 2005). By definition, genes duplicated subsequent to divergence from the cenancestor are orthologous while those that duplicated prior are not; these are known as inparalogs and outparalogs (Remm, Storm and Sonnhammer, 2001; Sonnhammer and Koonin, 2002). Given the above definitions, it is clear orthology is not necessarily a one-to-one relationship and is not transitive. Thus, orthologs are further subdivided into one-to-one, one-to-many, and many-to-many orthologs (Tatusov, Koonin and Lipman, 1997; Koonin, 2005; Vilella *et al.*, 2009); genes in the "many" group are collections of inparalogs and are considered "co-orthologous" to the other genes in the gene family.

That orthology is defined relative to the cenancestor is critical; whether a pair of genes are orthologous or not depends on the group of genomes being studied. Another important point

is that there is no functional connotation to orthology and paralogy, although it is generally assumed that orthologs are more likely to share equivalent functions than paralogs, which is a major motivating factor behind orthology identification.

From the above definitions, we define the problem of orthology identification as the grouping of one-to-one orthologs and co-orthologs (inparalogs) while separating outparalogs and xenologs, given the genes from a set of related genomes. Such groupings of genes are often referred to orthologous clusters. To our knowledge, no existing orthology identification programs explicitly test for the existence of xenologs.

Many programs and databases have been published regarding orthology assignment over the years and they can be roughly classified into three categories: tree reconciliation, phylogeny-based, and similarity-based. Despite the large number of papers, there have been few systematic comparisons between competing implementations or methodologies. Aside from limited comparisons accompanying new algorithm or database papers, we are aware of only three systematic comparisons, all of which use real but unique datasets where true orthology relationships are unknown (Hulsen *et al.*, 2006; Chen *et al.*, 2007; Altenhoff and Dessimoz, 2009). Furthermore, only four similarity-based orthology identification methods are common to all three analyses. We believe simulation-based comparisons are attractive because the true evolutionary relationships would be known. We were not aware of a simulation-based comparison of algorithms between different categories of orthology identification approaches prior to the start of our study but to the best of our knowledge the results mentioned in Vilella *et al.* (2009) have yet to be published. Thus, in Chapter 1 we implemented software for simulating gene duplication, LGT, loss, fusion, and fission, and used them to compare the performance of two similarity-based and two tree reconciliation algorithms.

Up to this point we have considered orthology between genes but gene fusion and fission complicate orthology assignment (Koonin, 2005; Kuzniar *et al.*, 2008; Kristensen *et al.*, 2010). In particular, fusion and fission result in relationships where different sections of a single gene are orthologous to multiple genes in other genomes and may contribute to the evolution of protein architectures (Koonin, 2005; Pasek, Risler and Brzellec, 2006; Weiner and Bornberg-Bauer, 2006; Fong *et al.*, 2007). Thus, depending on the study, it may be appropriate to consider orthology at the level of protein domains (Kummerfeld and Teichmann, 2005; Pasek, Risler and Brzellec, 2006). The problem is accurate and comprehensive domain annotations are not always available and many proteins simply contain no known domain. An alternative is to explicitly detect fusion and fission when identifying orthologs. We are aware of a single database that explicitly considers gene fusion and thus allows a single gene to exist in multiple orthologous clusters (Merkeev, Novichkov and Mironov, 2006). However, most methods require that genes uniquely belong to a single cluster, which can arbitrarily force the clustering of genes with no homologous region (Kuzniar *et al.*, 2008).

Interestingly, fusion between non-homologous proteins is proposed as a complementary tool to gene orthology for gene function prediction (Enright *et al.*, 1999; Enright and Ouzounis, 2001; Marcotte *et al.*, 1999). The contributions of fusion and fission to protein evolution have also been studied (Snel, Bork and Huynen, 2000; Kummerfeld and Teichmann, 2005). While fusion and fission detection programs have been written, none were readily available for comparison and many were designed to detect homology, not orthology. Thus, in Chapter 2, we focused on implementing a pipeline for the automatic detection of fusion and fission events between orthologs along a species phylogeny. With no programs for comparison, we

instead designed our study to gain further insight into the roles of fusion and fission in protein evolution in the *Bacillaceae*. In particular, we uncovered an unexpected number of gene fissions in *Bacillus anthracis*.

Chapter 1

Assessment of orthology identification approaches by simulation

1.1 ABSTRACT

Orthology identification is central to comparative and evolutionary genomics and is an active area of research. New algorithmic variations and databases are continually being developed but there is a paucity of systematic comparisons between such tools. Assessment of alignment and phylogenetic reconstruction algorithms are typically performed using a standard dataset such as BAliBASE or simulated data where the true phylogenetic history is known. A standard dataset for assessment of orthology identification is currently in development but to our knowledge, there have not been any published comparisons using simulated data. Software was developed to simulate gene duplication, loss, and lateral gene transfer (LGT) to assess the performance of four orthology identification algorithms representing two fundamentally different approaches.

1.2 INTRODUCTION

`InParanoid` and `OrthoMCL` appear to be the most well-established and popular orthology identification programs (Remm, Storm and Sonnhammer, 2001; Li, Stoeckert and Roos, 2003). Both use sequence similarity as a surrogate measure of orthology and do not consider any phylogenetic signal in the data, which is contrary to the definition of orthology. In recent years there has been increased interest in algorithms that are true to the definition of orthology (tree reconciliation) or at least take into account some phylogenetic information (phylogeny-based). Despite the surprising lack of comparative studies, there are conflicting reports regarding the relative performance between the three classes of algorithms.

When published, most algorithms are compared to a limited number of competitors using non-standard datasets if at all. A major problem is that the true evolutionary history of genes are unknown and there is currently no standardized dataset for comparison although one is in development (Gabaldn *et al.*, 2009). One approach is the construction of manually curated (Storm and Sonnhammer, 2002; Alexeyenko *et al.*, 2006) or automatically generated (Li, Stoeckert and Roos, 2003; Datta *et al.*, 2009; Kristensen *et al.*, 2010; Vilella *et al.*, 2009) orthology inferences for a dataset of interest. Another approach is to use surrogate mea-

sures of orthology such as gene symbols (Fu *et al.*, 2007), experimentally verified functional annotations (Kim *et al.*, 2008), and gene order conservation (van der Heijden *et al.*, 2007). Given the wide range of algorithms, some of the assessment strategies are only applicable to certain algorithm types (Remm, Storm and Sonnhammer, 2001; Arvestad *et al.*, 2003; Akerborg *et al.*, 2009; Sennblad and Lagergren, 2009). This is not an exhaustive list and usually a combination of datasets are employed.

We know of only three systematic comparisons of orthology identification with only four algorithms in common: reciprocal best hits (RBH), `InParanoid`, `OrthoMCL`, and `COG`. Notably, all four are similarity-based algorithms and there are no representatives of reconciliation and phylogeny-based algorithms. The earliest study by Hulsen *et al.* (2006) included a phylogeny-based method similar to `Orthostrapper` but without bootstrap support (Storm and Sonnhammer, 2002). Algorithms were assessed by combining analyses of a large collection of functional data including gene expression profiles, InterPro annotation, gene co-expression, gene order, and protein-protein interactions. The authors concluded that `InParanoid` was the best according to their scoring scheme. RBH provided the highest specificity, `InParanoid` provided the highest sensitivity, and the phylogeny-based method performed worse. The authors note that their Gene Ontology dataset could include annotations generated based on sequence similarity; this leads to circularity in this particular benchmark, highlighting the need for caution when interpreting benchmarks based on functional data.

Chen *et al.* (2007) included `RIO` and `Orthostrapper` to represent reconciliation and phylogeny-based methods. Need of a trusted set of orthology assignments and surrogate measures of orthology was avoided by using a latent class analysis statistical model to estimate sensitivity and specificity based on the degree of agreement between different algorithms. Data was constrained to the sequences present in the KOG (Tatusov *et al.*, 2003), RIO (Zmasek and Eddy, 2002), and HOPS databases (Storm and Sonnhammer, 2003); HOPS was generated using `Orthostrapper`. Interestingly, `InParanoid`, `Orthostrapper`, and RBH was the ranking obtained when comparing false negative (FN) rates although all three were estimated to have the same false positive (FP) rate, illustrating the success of the `InParanoid` algorithm in extending the RBH approach. As expected, `RSD` had a lower FP rate but a higher FN rate compared to RBH but the gap was closed substantially with more stringent `BLAST` cutoff parameters for RBH. `OrthoMCL` has a lower FN rate but higher FP rate compared to `InParanoid`; although not explicitly stated, from their data, inflation index values between 3 and 5 appear to result in FP rates approaching that of `InParanoid` while maintaining a lower FN rate. `Orthostrapper` was shown to have a much lower FN rate compared to `RIO` at the cost of a moderate increase in the FP rate; a bootstrap cutoff lower than the recommended 50% further lowers the FN rate without significantly increasing the FP rate. While `RIO` had an astonishing low FP rate of 0.01, its FN rate of 0.64 is likely too high for most purposes. Overall, the authors concluded that tree-based methods have higher specificity (lower FP rate) and similarity-based methods have higher sensitivity (lower FN rate), with `InParanoid` and `OrthoMCL` providing the best trade-off between the two. Latent class analysis is an interesting approach but concerns have been raised over the applicability of the model to make inferences regarding orthology identification methods (Altenhoff and Dessimoz, 2009).

The most recent study focused on orthology databases, including OMA, Homologene, and Ensembl (Altenhoff and Dessimoz, 2009). OMA uses a similarity-based algorithm (Roth,

Gonnet and Dessimoz, 2008), Homologene uses an unpublished phylogeny-based method combining a species phylogenetic algorithm with gene order, and the Ensembl data were generated using the `RAP` algorithm (Dufayard *et al.*, 2005; Hubbard *et al.*, 2007). Data were mapped from each project to OMA and pairwise comparisons were performed using two phylogenetic and four functional benchmarks. The phylogenetic tests consisted of testing for congruence between the species phylogeny and ortholog trees, constructed using one predicted ortholog per species, and a combination of manually curated datasets from the literature. A notable aspect was avoidance of circularity in functional tests; only Gene Ontology annotations with experimental evidence were used. Overall, the authors concluded that OMA and Homologene performed best in both types of tests although at higher coverage and lower specificity (higher FP rate), OrthoMCL outperformed InParanoid and Ensembl Compara in the functional tests. They found their results to be largely consistent with Hulsen *et al.* (2006) and suggest that reconciliation is outperformed by similarity-based methods.

Simulation is a useful tool that is complimentary to the use of real datasets and we believe its use is underappreciated in the orthology identification literature. Simulated gene trees have been used to assess the run-time of `SDI` (Zmasek and Eddy, 2001), correlation between accuracy and orthology bootstrap cutoffs for `Orthostrapper` (Storm and Sonnhammer, 2002), and test for systematic bias in the implementation of `SPIDIR` (Rasmussen and Kellis, 2007). However, a few algorithms have been compared using data simulated under more sophisticated models of gene-level evolution. Shi, Zhang and Jiang (2010) implemented a model for simulating random gene duplication, genome inversion, and sequence mutation along a two-species phylogeny and used the data to compare their `MSOAR 2.0` algorithm with `InParanoid` and their original `MSOAR`. Wapinski *et al.* (2007) simulated gene duplication and loss along a phylogeny of six fungi, using unspecified rates proportional to those observed in their fungal data, and simulated sequence substitutions using `Seq-Gen` (Rambaut and Grassly, 1997). This data was used to assess accuracy of pairwise orthology predictions of their `SYNERGY` algorithm and `InParanoid`. A similar approach was taken to compare probabilistic and maximum parsimony reconciliation although performance was assessed over multiple species trees and a range of duplication and loss rates although sequence simulation was not presented in detail (Sennblad and Lagergren, 2009). Recently, Rasmussen and Kellis (2011) developed a reconciliation program `SPIMAP` and compared it to other reconciliation programs and regular phylogenetic inference programs using simulated datasets at varying rates of duplication and loss trained on real data. To the best of our knowledge, there have been no comparisons between similarity-based and reconciliation algorithms. Vilella *et al.* (2009) recognized the value of a systematic comparison between reconciliation and similarity-based methods using simulated data. Their as of yet unpublished results suggest their reconciliation algorithm `TreeBeST` is superior to similarity-based algorithms, a conclusion similar to that reached by Chen *et al.* (2007) but different from Hulsen *et al.* (2006) and Altenhoff and Dessimoz (2009). Kristensen *et al.* (2010) suggested that there is no significant difference between these two fundamentally different approaches but rather than providing data to support their claim, they use it as an argument in favor of further developing similarity-based algorithms.

As mentioned above, a major limitation in systematic comparisons is the lack of a standardized set of orthologs and knowledge of true evolutionary history. Surrogate measures have been used but caution is required to avoid circularity and applicability depends on the measure chosen and evolutionary distance separating the species. A standard dataset for

orthology identification is being developed and data formats are being standardized (Gabaldn *et al.*, 2009) but we are not aware of any plans for a large-scale systematic comparison of similarity-based and reconciliation algorithms with simulated data with the exception of Vilella *et al.* (2009), as noted above. We believe that standardized datasets and simulated data are complimentary to one another like they are in the analysis of sequence alignment and phylogenetic reconstruction. Two potential problems we believe were not addressed in previous comparative studies are the mapping of data between pre-existing databases and the lack of distinction between pairwise (e.g. `InParanoid`) and multi-genome (e.g. `OrthoMCL`) algorithms. In both cases, the cenancestor will be different and there may exist a true difference in the correct orthology clustering.

The main objective of our study is the systematic comparison of downloadable orthology identification programs using simulated genomes with a specific focus on identifying any differences between similarity-based and reconciliation algorithms, since there appears to be a trend towards phylogeny-based algorithms (Gabaldn *et al.*, 2009) yet previous comparisons have been inconclusive (Hulsen *et al.*, 2006; Chen *et al.*, 2007; Altenhoff and Dessimoz, 2009; Kristensen *et al.*, 2010). Using simulated datasets and downloaded programs, we know the true evolutionary histories and avoid any potential problems cause by mapping data between databases comprising different genome sets.

We guided program selection based on two criteria. Some users will only be interested in online databases, but the availability of the algorithms for local use is also important since online databases are often restricted to specific taxonomic groups and time between updates will likely become longer due to the increase in number of genomes. Downloadable programs also allow us to avoid data mapping between databases. Furthermore, pairwise genome algorithms are becoming less relevant in light of increasing genome availability, phylogenetic coverage (Wu *et al.*, 2009), and increased sampling within taxonomic groups (Ravel and Fraser, 2005; Fuxelius *et al.*, 2008; Ogura *et al.*, 2009; Fischer *et al.*, 2010; Losada *et al.*, 2010). With the above considerations in mind, we chose to compare `MultiParanoid`, `OrthoMCL`, `Notung` (Durand, Halldrsson and Vernot, 2006), and `RAP`. `MultiParanoid` is relatively unknown, but it is the extension of the popular `InParanoid` algorithm to multiple genomes, providing an opportunity to determine whether previous results showing that `InParanoid` performs better than `OrthoMCL` hold when extended to multiple genome analysis.

We tested the performance of all four algorithms under varying rates of gene duplication, loss, and lateral gene transfer (LGT). LGT is particularly interesting due to its prevalence in prokaryotes (Hao and Golding, 2006) and the fact that most algorithms were not designed to detect them. Due to the lack of any suitable simulation software, we developed our own to simulate protein sequence evolution along trees with gene duplication, loss, and LGT. Specific to reconciliation algorithms, we also tested different strategies in gene family construction similar to Penel *et al.* (2009), since it is an important step in de novo orthology identification using reconciliation, and determined whether considering bootstrap support in reconciliation algorithms can compensate for errors in the species phylogeny.

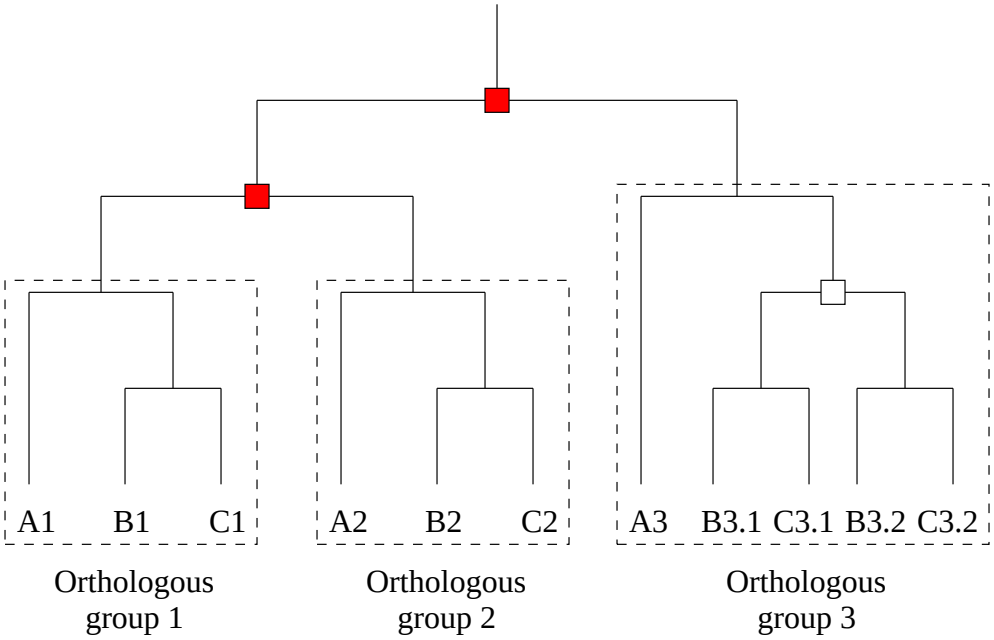
As alluded to above, we chose to classify orthology identification algorithms into three main categories: tree reconciliation, similarity-based, and phylogeny-based. Some algorithms do not fall neatly into any of these categories such as `MSOAR` (Fu *et al.*, 2007; Shi, Zhang and Jiang, 2010), which attempts to assign "one-to-one orthologs" between a pair of genomes based on gene order; note that their definition of one-to-one orthologs differs from the one provided in the general introduction on page 3, which is the generally accepted definition

(Koonin, 2005; Vilella *et al.*, 2009). The remainder of this introduction is a review, focusing primarily on reconciliation and similarity-based algorithms that are implemented in programs available for download; algorithms part of online databases but not available for download were omitted, such as OMA (Roth, Gonnet and Dessimoz, 2008).

Tree reconciliation algorithms

Orthology is an evolutionary biology concept, so a phylogenetic approach is the natural way of assigning orthology to genes in a gene family. If the true species phylogeny and history of the gene family are known, then by definition the first non-duplication nodes from the root of the gene tree represent distinct gene family members in the cenancestor and define orthologous groups (Figure 1.1). This procedure can be recursively applied to determine orthologous groups with respect to any node in the species tree.

Figure 1.1: A gene family for three species 'A', 'B', and 'C', in which two duplications occurred in the cenancestor (red squares) and one lineage-specific duplication (white square). Each of the duplications in the cenancestor results in a new orthologous group. If one was interested only in species 'B' and 'C', then the duplication colored white would also result in a fourth orthologous groups.



In practice, a gene tree is usually inferred using sequence data alone and will often be incongruent with a trusted species phylogeny (Rasmussen and Kellis, 2007). The tree reconciliation approach, first proposed by Goodman *et al.* (1979), assumes that incongruence is due to gene duplication and loss and is often conceptualized as the mapping of gene tree nodes to a node in the species tree. The basic formulation of the maximum parsimony reconciliation of a gene tree with a species tree can be expressed in two distinct steps (Zmasek and Eddy, 2001):

1. Each gene tree node g, the ancestor of m genes from a set of n genomes G ($m \geq n$), is mapped to the species tree node s that represents the cenancestor of G.

2. Gene tree nodes that map to the same species tree node as one of its children is identified as a duplication node.

Although reconciliation is consistent with the definition of orthology, many problems exist in practice. Species phylogenies are often unknown, especially in prokaryotes, contested such as for *Drosophila* and yeast, or unresolved. Both the species and gene trees need to be rooted; appropriate outgroup sequences may not be available and such an approach is not practical in large scale analyses. Furthermore, incomplete lineage sorting, gene conversion, LGT, insufficient phylogenetic signal, and long branch attraction may all cause gene tree incongruence (Rasmussen and Kellis, 2007); failure to consider alternate explanations may result in incorrect inference of duplications nodes. Most reconciliation methods are based on the parsimony principle, which biases events towards the root of a tree (Hahn, 2007). Aside from reconciliation itself, the pre-processing steps are also problematic. One must consider how gene families should be delineated and how to infer the gene family trees; phylogenetic inference is slow and difficult to automate, with many steps susceptible to error.

Zmasek and Eddy (2002) were the first to apply automatic tree reconciliation to high-throughput ortholog identification, but their `RIO` procedure was limited to fully resolved gene and species trees. Gene trees were rooted by minimizing the number of duplication events, with ties broken by choosing the rooted tree with the shortest height, and gene tree uncertainty was taken into account by reconciling multiple bootstrapped trees. The number of times a pair of genes were orthologous in the set of bootstrap trees was taken as the confidence value of the prediction. `Notung` can handle unresolved nodes in either the gene tree or species tree, while `RAP` (Dufayard *et al.*, 2005) and `Softparsmap` (Berglund-Sonnhammer *et al.*, 2006) allow unresolved nodes in both trees. The `RAP` rooting algorithm is the same as `RIO` except that it breaks ties by choosing the root closest to the midpoint of the gene tree. `Notung` minimizes duplications and losses simultaneously while `Softparsmap` minimizes duplications first followed by an approximate minimization of losses. The greatest differences lie in the handling of gene tree uncertainty. `RAP` collapses weakly supported nodes using both node support and branch length and may produce reconciled trees containing unresolved nodes. Interestingly, `RAP` performs poorly on gene trees without branch lengths (see results). `Notung` resolves multifurcations in the gene tree by minimizing duplications and losses simultaneously but will rearrange branches around nodes with low support. `Softparsmap` will collapse branches with low support like `RAP` but will resolve both the original and newly created multifurcations by minimizing duplications first and losses afterwards. Both `Notung` and `Softparsmap` generate binary reconciled gene trees unlike `RAP`. None of the three methods can assign a confidence value to an orthology prediction. The program `primeGEM` (Arvestad *et al.*, 2003; Sennblad and Lagergren, 2009) is similar to `RIO`, where orthology is assessed from multiple reconciled trees but instead of considering bootstrap resampled trees, `primeGEM` uses a Bayesian MCMC approach to generate a posterior distribution of reconciled trees from a probabilistic model of gene evolution model based on the birth-and-death process. This method allows the calculation of posterior speciation probabilities for each node in the reconciled tree. Akerborg *et al.* (2009) combined the gene and sequence evolution models to directly infer a reconciled gene tree consistent with the species phylogeny in their program `PrIME-GSR`. Rasmussen and Kellis (2007, 2011) have also developed distance-based likelihood (`SPIDIR`) and Bayesian (`SPIMAP`) methods for direct inference reconciled gene trees.

Many of the above algorithms are key components of orthology databases. `RIO` is both

the name of the database and the inference procedure based on the SDI reconciliation algorithm (Zmasek and Eddy, 2001); Softparsmap was created in the development of the TAED database (Roth *et al.*, 2005; Berglund-Sonnhammer *et al.*, 2006); and RAP is used in the database HOGENOM (Penel *et al.*, 2009). Ensembl GeneTrees also used RAP (Hubbard *et al.*, 2007) until they transitioned to their new TreeBeST program (Vilella *et al.*, 2009). TreeBeST, previously known as NJTree, is currently used in several databases (Heger and Ponting, 2008; Ruan *et al.*, 2008; Vilella *et al.*, 2009) but has not been published so we refer interested readers to the Ensembl GeneTrees publication (Vilella *et al.*, 2009), the TreeBeST website (http://treesoft.sourceforge.net/) and the Ensembl 2011 paper (Flicek *et al.*, 2011). Notung is not part of a database, but it has been applied in a number of gene family (Binford *et al.*, 2009; Gardiner *et al.*, 2008; Frygelius *et al.*, 2010) and genome-level (Hahn, Han and Han, 2007; Meisel, Han and Hahn, 2009) analyses.

A critical step in large scale reconciliation often ignored in algorithm assessment is the construction of gene families. The most common approach is single-link clustering of BLAST results after some form of filtering (Roth *et al.*, 2005; Hahn, Han and Han, 2007; Hubbard *et al.*, 2007; Penel *et al.*, 2009; Meisel, Han and Hahn, 2009; Vilella *et al.*, 2009). Single-link clustering implies relationships are transitive; if A and B are in a family and B and C are also in a family, then all three are in the same family. It is commonly referred to as single-linkage clustering in the literature, but we wish to distinguish this from the hierarchical clustering approach known under the same name. TRIBE-MCL (Enright, Dongen and Ouzounis, 2002) and MSOAR 2.0 (Shi, Zhang and Jiang, 2010) take an alternative approach; they applied the Markov Cluster (MCL) Algorithm (van Dongen, 2000) instead of single-link clustering. Penel *et al.* (2009) compared their implementation of the single-link method, Build_Fam, with TRIBE-MCL and OrthoMCL and concluded that Build_Fam is a good compromise between gene family size and alignment quality (Li, Stoeckert and Roos, 2003). The comparatively poor performance of OrthoMCL is not surprising because it was designed to infer orthologous clusters; it will be covered in the next section.

Similarity-based algorithms

Similarity-based approaches attempt to separate outparalogs from inparalogs by using similarity as a surrogate measure of orthology. From their definitions, it seems reasonable to assume that orthologs and inparalogs should be more similar to each other than to outparalogs. This assumption would fail if one duplicate diverges much more quickly and could lead to false positives or false negatives. Additional disadvantages include the inability to resolve evolutionary relationships (i.e. identify duplication and loss events) within orthologous clusters and the tendency to cluster outparalogs in the presence of differential outparalog loss. The motivation for similarity-based algorithms is their relative speed and the avoidance of error-prone phylogenetic reconstruction. Additionally, they should be less susceptible to long branch attraction errors (Remm, Storm and Sonnhammer, 2001), do not require a species tree, and avoid the conceptual issue of reconciliation in the presence of LGT.

The most basic and possibly most common approach is RBH (also symmetrical or bidirectional best hits). BLAST searches are performed between a pair of genomes in both directions and filtered by some combination of score, 'Expect value', and alignment length thresholds. Proteins that are best hits of each other are identified as orthologs. The reciprocal smallest distance (RSD) algorithm (Wall, Fraser and Hirsh, 2003) reduces the chance of

false positives due to highly similar inparalogs by estimating maximum likelihood distances for all significant BLAST hits. RBH and RSD are restricted to pairwise genome comparisons and cannot detect paralogs. Resulting pairs of orthologs are sometimes referred to as "main orthologs". The InParanoid algorithm (Remm, Storm and Sonnhammer, 2001) extends the RBH approach to detect inparalogs by performing within-genome BLAST searches and adding inparalogs to the main ortholog clusters using a set of rules based on BLAST bit scores. InParanoid is unique in estimating confidence values for orthology assignments and for allowing the inclusion of an outgroup genome to detect selective loss of outparalogs, decreasing the chance of false positives but increasing the chance of false negatives. MultiParanoid (Alexeyenko *et al.*, 2006) is a further extension that merges InParanoid clusters using single-link clustering and optionally resolving any resulting overlaps. Instead of applying successive sets of special rules, OrthoMCL (Li, Stoeckert and Roos, 2003) simultaneously identifies multi-genome orthologous clusters including inparalogs using the MCL algorithm. Significant RBH within and between all genomes are used to construct a graph with edges weighted by normalized BLAST E-values. The MCL algorithm is applied to remove weak edges in the graph and groups of proteins still connected afterwards are interpreted as orthologous clusters. Recall that the MCL algorithm was also applied in gene family construction in TRIBE-MCL; the difference lies in the construction of the graph. The COG approach (Tatusov, Koonin and Lipman, 1997) is also based on the idea of graph construction and clustering like OrthoMCL but differs from the previous methods in requiring triangles of hits rather than pairs. Edges are constructed between proteins in different genomes for each one-way best BLAST hit and orthology is interpreted as triangles in the graph to compensate for the relaxation of the RBH criteria and the lack of cutoff values; this proved insufficient so recent applications of the COG approach enforce the RBH criteria to reduce the number of false positives (Makarova *et al.*, 2007; Kristensen *et al.*, 2010). The clustering step consists of merging triangles sharing an edge, which allows clustering within and across multiple genomes. COCO-CL (Jothi *et al.*, 2006) differs from the above approaches by performing a hierarchical clustering of a predetermined set of homologs; it was applied to the COG database although any homolog clustering algorithm should work theoretically. Given a set of homologs, single-linkage clustering is applied to correlations of evolutionary distances calculated from a multiple sequence alignment. The last link is defined as the splitting of the set into two groups; a split score determines the confidence that the split is a duplication and a bootstrap score represents confidence in the split. This procedure is applied recursively until either score falls below a cutoff. Due to the definition of the split score, LGT may boost the chance of wrongly inferring a duplication.

The InParanoid, OrthoMCL, and RSD algorithms have been used by periodically updated online databases (Berglund *et al.*, 2008; Chen *et al.*, 2006; Deluca *et al.*, 2006). Although the new, automatic version of the COG algorithm is available (Kristensen *et al.*, 2010), the COG and KOG databases (Tatusov *et al.*, 2003) have not been kept up to date and are based on the original algorithm with a manual curation step; we will use COG and KOG interchangeability since they contain clusters from different genomes sets generated by the same algorithm. MultiParanoid and COCO-CL data are available online as well but they do not appear to be updated either.

Phylogeny-based algorithms

Methods we classify as phylogeny-based are those that incorporate some form of phylogenetic information. In fact, they reflect modifications of the two extremes of tree reconciliation and similarity-based methods. In particular, the first class removes the assumption that the species phylogeny is known but assumes that the gene family trees are correct. Algorithms of this class consist of various heuristics for orthology inference directly from the gene family trees (Storm and Sonnhammer, 2002; Hollich, Storm and Sonnhammer, 2002; van der Heijden *et al.*, 2007; Kim *et al.*, 2008; Datta *et al.*, 2009). The second class assumes that the species phylogeny is accurate and uses it to guide an agglomerative hierarchical clustering of sequences based on sequence similarity like `COCO-CL`, preserving the speed advantage of strict similarity-based methods (Dehal and Boore, 2006; Merkeev, Novichkov and Mironov, 2006; Wapinski *et al.*, 2007).

1.3 MATERIALS AND METHODS

1.3.1 Gene tree and sequence simulation

A program was written using Perl, BioPerl (Stajich *et al.*, 2002), and SQLite (`http://www.sqlite.org/`) to simulate genome-level events including gene duplication, loss, fusion, fission, and LGT under a birth-and-death model at varying rates along a rooted phylogeny. In contrast to `EvolSimulator` (Beiko and Charlebois, 2007), we chose to simulate the bare minimum for orthology analysis and keep parameters to a minimum. Thus, only gene family trees were generated and all other factors including intergenic regions, gene order, and gene orientation were excluded. Furthermore, all events act on single genes, although, in evolution, they may act on portions of genes (Chan *et al.*, 2009) or multiple genes (Hughes, 1998). Divergence times and amino acid substitution rates were required for event and sequence simulation respectively and are assumed to be linearly correlated; rates are thus constant within a branch. Our simulation procedure also differs from `EvolSimulator` by assuming selective neutrality of sequence substitutions to allow the separation of gene tree and sequence simulation and thus the inclusion of an indel formation model.

Sequences were simulated using `indel-Seq-Gen v2.0.3` (Strope *et al.*, 2009) with amino acid substitutions simulated under the JTT model and indels simulated using the Gillespie algorithm under the Chang and Benner model (Chang and Benner, 2004) with a maximum indel size of 10. In keeping with our goal of minimizing parameters, we did not make use of `indel-Seq-Gen` models for introns, exons, domains, and motifs. Gene sequences were initialized randomly by `indel-Seq-Gen` so each gene in the cenancestor represents a single gene family with no paralogs initially. Our gene tree simulation algorithm is as follows:

Input:

Rooted species phylogeny with divergence times

Amino acid substitution rates for each branch

Event rates for each event on each branch

Number of genes in cenancestor

Output: Set of rooted gene trees (for each surviving gene from the cenancestor)

Procedure:

1. Initialize all gene trees to species tree
2. Pre-order traversal of species tree
3. For each branch:
   (a) Generate time of next event from an exponential distribution
   (b) End evolution on this branch if we have reached the end of the branch
   (c) Otherwise, choose event type: duplication, loss, LGT, fusion, or fission
   (d) Choose gene tree(s) or gene family (for LGT)
   (e) Modify gene tree(s) and generate new gene tree if necessary
   (f) Repeat above steps
4. For each resulting gene tree, convert divergence times to amino acid substitution rates

The cenancestor is specified simply by the number of initial genes, and corresponding gene trees are initialized assuming no events have occurred. These initial trees are modified appropriately for each event generated under a waiting-time model. The process starts at the branch leading to the cenancestor, allowing the creation of outparalogs with varying degrees of divergence, and proceeds in a pre-order traversal of the species tree. Genomes at the same depth in the species phylogeny are simulated independently and are dependent only on the ancestral genomes. This is acceptable for simulating duplication, loss, fusion, and fission, since these events occur between genes in the same genome but not for LGT between genomes, which would require simultaneous simulation of all genomes at the same depth. Furthermore, in our actual experiments we did not use true divergence times as the species tree is not ultrametric, making the mapping of events between branches impossible. Thus, our model for LGT is different from those of other events. The generation of the divergence times used in this study will be explained in the next section.

In our model, LGT acts on any ancestral gene family instead of a single gene, so it does not affect newly created families from gene fusion and fission and can act on a gene family even if all its original members were lost. Unlike `EvolSimulator` and `HGT_simul` (Galtier, 2007), we model LGT events from outgroup genomes that are not explicitly simulated. For each LGT event, we simulate an independent lineage from the initial sequence in the cenancestor. The first problem with this model lies in determining the amount of divergence to simulate along this new lineage prior to the LGT event. A possible solution would be to simply use the divergence times in the species tree (along with arbitrary substitution rates) but recall those are not available in our simulation. Even if they were available, the amount of divergence between LGT clades would be similar and may lead to higher error rates than expected for reconciliation and phylogeny-based methods. The current implementation does not simulate any initial divergence at all, which may further inflate error rates even for similarity-based methods. In particular, there will be abnormally low divergence between the LGT clades. Depending on the species tree, this could also result in LGT clades being more similar to a clade of original gene family members.

Orthologous gene replacement may be indirectly simulated by a combination of LGT and gene loss along the same lineage. Introduction of a new gene family through LGT should not affect orthology prediction of genes in existing families so that was not modeled. Thus, error rates based on a particular rate of LGT should be worse than expected since all LGT events may affect existing gene families.

The lack of true divergence times for our species tree did not affect simulation of duplication, loss, or fission since they involve a single branch of a gene tree. Simulation of fusion is also unaffected because it involves branches between two gene trees mapping to the same species tree branch so time is constant between them. Duplication and loss simply result in the copying and pruning of subtrees, respectively. In our model of gene fusion or fission, the trees of the ancestral genes are pruned and new trees are produced for the new fusion or fission genes. We should note that fission may result in implicit gene loss if the resulting gene fragments are too short; we chose a minimum gene size of 30 amino acids. Our algorithm chooses a random cut point that, in order of preference, generates two gene fragments, one fragment, or none (loss of the entire gene). Thus, a gene of 60 amino acids will always be split into two fragment containing 30 amino acids. We also implicitly assume that any fission products generated are functional genes and do not consider pseudogenization; if a gene is lost it is completely erased from the genome.

1.3.2 Species phylogeny

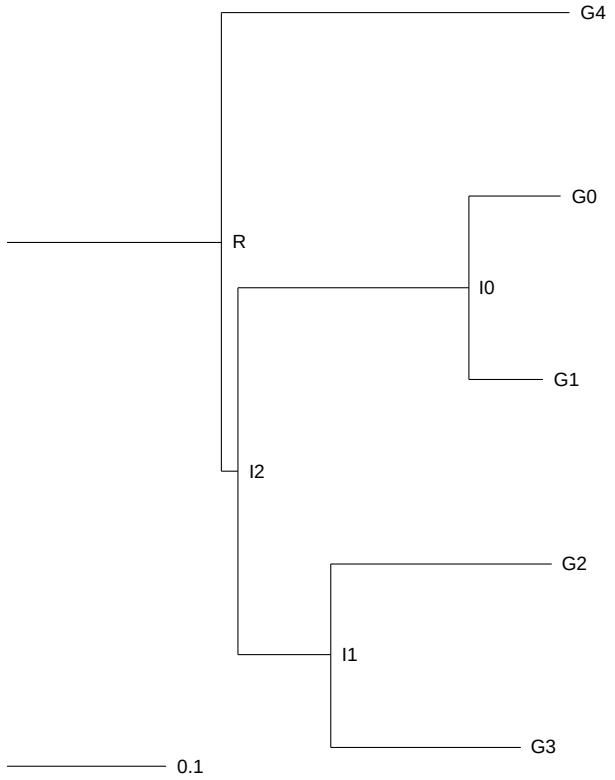
An initial tree was reconstructed for 25 *Bacillaceae* genomes, with *Listeria monocytogenes* EGD-e as the outgroup, from a concatenation of the *gmk*, *glpF*, and *pycA* genes (Priest *et al.*, 2004; Hao and Golding, 2006) using `MrBayes` (Huelsenbeck and Ronquist, 2001) (2,000,000 generations sampled every 100 generations with a gamma distribution model with six categories and invariant class). The resulting tree had a trifurcation between the three included *B. anthracis* strains and was converted to a binary tree by the addition of a node with 0 branch length and posterior probability.

The final tree (Figure 1.2) containing *Bacillus amylofiquaciens*, *B. subtilis* 168, *B. clausii*, *B. halodurans*, and *Oceanobacillus iheyensis* was derived using `Retree` and the corresponding protein substitution rates were estimated using `Protdist` (Kimura model of amino acid substitution), and `Fitch` from `PHYLIP` package version 3.68 (Felsenstein, 1989). The branch lengths on this tree were used as our divergence times, although, as noted in the previous section, they are not true divergence times since the tree is not ultrametric.

1.3.3 Rates of macroevolutionary events

A number of studies have been applied to estimate gene duplication and loss rates in eukaryotic genomes (Lynch and Conery, 2000; Hahn, Han and Han, 2007), with different methodologies producing similar estimates, but studies of prokaryotic genomes are comparatively uncommon (Hooper and Berg, 2003; Lynch and Conery, 2003). We chose to make use of the rates estimated by Lynch and Conery (2003) because they were presented in terms of silent-site divergence; more commonly, rates are presented in terms of millions of years but such figures are harder to estimate for prokaryotic genomes. Thus, we chose to simulate duplication and loss across the range of 0.001 to 0.004 per silent-site divergence of 1%, which covers both prokaryotic and eukaryotic genomes. We converted those into 0.1 to 0.4 events

Figure 1.2: Species phylogeny used for simulation.



per silent site substitution. Note, however, that our divergence times were estimated across all sites, not just silent sites. Estimating duplication and loss rates in prokaryotes is harder due to the lack of sufficient numbers of duplicate genes and the potential confounding effect of LGT. However, Lynch and Conery (2003) note that their analysis is performed only on duplicates with silent-site divergence of less than 1% so LGT should have a minimal effect.

LGT rates have been studied in the *Bacillaceae* genomes we are interested in (Hao and Golding, 2006, 2008b). Although there is a high rate of LGT, there is also a high degree of rate variation between genes and a small subset of recently transferred genes are likely to have much higher rates than average. Rates vary from less than 0.3 to 5, but we chose to use the same range of LGT rates as the duplication and loss, since the high rates were attributed to a small subset of genes. This also allows us to directly compare the effects of equivalent levels of LGT and duplication.

Although there have been a number of surveys for gene fusion and fission events, we are aware of only one report of fusion and fission rates with respect to time. Nakamura, Itoh and Martin (2007) estimated the rate of fusion and fission to be about 100-fold slower than the average rate of nucleotide substitution, in a comparison between the *Oryza sativa* and *Arabidopsis thaliana* genomes. The authors note that the rates may be different in prokaryotes due to the difference in genome architecture; this may also explain their observation that fissions are more common than fusions, opposite of the report from (Kummerfeld and Teichmann, 2005). We do not believe there is sufficient evidence for either rate being higher and decided to keep them identical in our simulation study. In the end, we decided to simulate fusions and fissions at equal rates of 0.01 for one experiment and 0.02 in another.

Our simulation program allows macroevolutionary event rate variation across the input

species tree, but the user must specify each rate independently. We used a gamma distribution to assign event rates to each branch with means of 0.1 to 0.4 for duplication, loss, and LGT and 0.01 and 0.02 for fusion and fission. Setting the shape parameter of the distribution to 10, the scale thus becomes $\frac{\mu}{shape}$. Recall however that rates between gene trees are identical.

1.3.4 Orthology identification

Unless specified, `BLAST 2.2.19` was used to perform similarity searches using the `BLOSUM62` matrix, the default compositional adjustment as described in (Yu and Altschul, 2005), and the soft filtering and Smith-Waterman final alignment options.

`InParanoid 4.1` was used to perform pairwise genome comparisons using default settings for `InParanoid` and `BLAST`, including the 2-pass `BLAST` strategy implemented since `InParanoid 4.0` and bootstrapping. The pairwise orthology assignments were combined using `MultiParanoid`, making use of bootstrap values and forcing output clusters to be non-overlapping. When allowing clusters to overlap, we found that some clusters were not only overlapping but contained entirely within another (strict subset). `InParanoid` was obtained from `http://inparanoid.sbc.su.se/` and `MultiParanoid` was obtained from `http://multiparanoid.sbc.su.se/`.

`OrthoMCL 2.0` was used in conjunction with `MCL v09-308`, using default parameters (inflation index 1.5). In contrast to previous versions, `BLAST` searches need to be handled separately. Aside from using the Smith-Waterman option, we followed the recommendations listed in the OrthoMCL Algorithm Document (`http://docs.google.com/Doc?id=dd996jxg_1gsqsp6`), including the recommended `BLAST` Except value cutoff of 10^{-5}; `OrthoMCL` was obtained from `http://www.orthomcl.org/` and `MCL` was obtained from `http://www.micans.org/mcl/`.

We assessed three strategies for building gene families necessary for reconciliation analysis. Our initial approach was single-link clustering of RBH after filtering by an Expect value cutoff of 10^{-20}, alignment length cutoff of 85%, and percent identity cutoff of 50%. Although single-link clustering of RBH best hits was used previously (Hubbard *et al.*, 2007), we believe it would be too similar to RBH orthology identification, result in too many false negatives, and generate overly tight clusters. Although methodology differs, most authors allow multiple hits similar to our approach (Roth *et al.*, 2005; Penel *et al.*, 2009; Vilella *et al.*, 2009).

Hahn, Han and Han (2007) used the fuzzy reciprocal `BLAST` (`FRB`) approach to construct gene families for their reconciliation analysis. Motivated by the finding that `FRB` performs similarly to `InParanoid` (Drosophila 12 Genomes Consortium, 2007), we also try to interpret orthologous clusters inferred using `InParanoid` and `OrthoMCL` as gene families instead. This can be interpreted as testing whether orthology clusters from similarity-based approaches can be further refined by reconciliation, particularly in the situation where `InParanoid` and `OrthoMCL` cannot split a large cluster any further.

Gene family alignments were constructed using `MUSCLE` v3.7 (Edgar, 2004) and used to infer consensus neighbor-joining trees (1000 bootstrap samples, Kimura distances) using programs from the `PHYLIP` package v3.68 (Felsenstein, 1989). Gene trees with distances were then generated using the `Fitch` algorithm (consensus tree, Kimura distances) and a custom Perl script was used to combine both bootstrap values and branch lengths in a single

tree. To test robustness of reconciliation algorithms when supplied with an incorrect species phylogeny, the other 2 possible topologies were constructed by swapping G1 with G2 (t1) and G0 with G4 (t2) in the original species phylogeny (t0) as shown in Figure 1.2.

`Notung` provides the basic gene tree rooting and reconciliation functions but can subsequently apply a rearrangement function that further minimizes the number of duplication and loss events using either bootstrap or branch length as the measure of support. We obtained `Notung 2.6` from `http://www.cs.cmu.edu/~durand/Notung/` and used it to test all three combinations for each of the three species trees using default parameters. Notably, `Notung` by default uses a relative cutoff of 90% of the highest value in the current gene tree; if all branches have 50% bootstrap support, the tree will not be rearranged. This differs from `RAP` and most other programs where an absolute cutoff is applied. Also note that by default duplications are weighted slightly higher than losses (1.5 versus 1.0). `Notung` outputs reconciled trees in the New Hampshire eXtended (NHX) file format (`http://www.phylosoft.org/NHX/`) from which we parse orthologous clusters using a custom Perl program. Gene trees containing only sequences from one genome produce an error; we shall assume they are orthologous. `Notung` ignores branch lengths on the species tree.

`RAP` can make use of both bootstrap and branch length as measure of support during reconciliation for introducing duplication nodes or collapsing weak nodes. All three combinations were tested against the three species trees using the command line version `RapMasse v1.0` we obtained directly from Dr. Guy Perriere; the GUI version of `RAP` is available at `http://pbil.univ-lyon1.fr/software/RAP/RAP.htm`. We used the default Rap values for the relative rate ratio parameter for introducing duplications (20) and branch length for collapsing weak nodes (0.1). Alternate values for both parameters were also explored (see results). We also turned off the option for identifying redundant sequences since we are not dealing with a non-redundant database and used a bootstrap cutoff of 90%. All other parameters were left at `RapMasse` defaults, which are different from those of `RAP`. We wrote a custom Perl script to convert from the `RAP` format for reconciled trees to the same NHX format used by `Notung`, allowing us to share code for parsing orthologous groups; loss node labels were also converted to conform to `Notung` conventions. `RAP` can make use of branch lengths on the species phylogeny but all lengths in our phylogenies were less than 1 and were rounded to 0 and should make it the same as `Notung`.

1.3.5 Accuracy measurement

Algorithm performance was compared based on correctness of all possible pairwise protein relationships. Given a set of predicted orthologous clusters, every within-cluster protein pair that is orthologous is counted as a true positive (TP); otherwise it is a FP. Similarly, FN are orthologous pairs not predicted by the algorithm and all other pairs are true negatives true negative (TN). Using these counts, we calculate precision and recall as:

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

This method is equivalent to those previously used in assessment of sequence alignment programs, where precision is known as the Modeler score and recall is known as the Developer

score (also sum-of-pairs score). Previous systematic studies compared performance of orthology assignment methods using sensitivity and specificity (Hulsen *et al.*, 2006; Chen *et al.*, 2007; Sennblad and Lagergren, 2009), which we believe is odd due to the skew in number of true negatives. However, we are aware of two methods papers where comparisons were performed using precision and recall (Datta *et al.*, 2009; Rasmussen and Kellis, 2011).

The Wilcoxon signed-rank test was used to compare the performance of different algorithms on the same dataset. Comparisons between rates for the same algorithm were performed using the Wilcoxon rank sum test. When required, Bonferroni corrections were applied. All statistical tests and Bonferroni corrections were performed with R (R Development Core Team, 2009).

1.4 RESULTS

For each experiment, we simulated 50 datasets for each rate category. The only exception was LGT, for which only 10 datasets per rate category were generated, due to the concerns over our LGT simulation procedure.

We also performed a control experiment with 50 replicates, where no macroevolutionary events was simulated. Precision and recall were perfect 1.0 except for `RAP`, which performed significantly worse when the species tree was incorrect (data not shown). Inspection of the results confirmed that `RAP` incorrectly infers duplications at the root of the phylogeny, a known problem with maximum parsimony methods of reconciliation (Hahn, 2007). When only gene loss was simulated, there was no significant difference between precision as expected (data not shown). Similarity-based methods had significantly higher recall than reconciliation methods. When the species phylogeny is correct, the difference is due to the fact that at least three sequences are required to reconstruct a gene tree. However, we found that, as the loss rate increases or if the species phylogeny is incorrect, reconciliation methods incorrectly infer duplication nodes at the root of the phylogeny. Both algorithms are affected but `RAP` performs much worse than `Notung`.

In the following sections, we will explore the differences between the programs under several more sets of simulations. For each experiment, we will provide a plot of precision (where applicable) and recall values. We recorded in a separate table the results of pairwise comparisons of precision in the upper triangle and recall in the lower triangle. Given a row 'i' and a column 'j', entry '(i,j)' is '+' or '-' if program 'i' has a significantly higher or lower precision than program 'j' and entry '(j,i)' is '+' or '-' if program 'i' has a significantly higher or lower recall than program 'j'; a comparison was considered significant if $P < 0.05$. An empty cell represents no significant difference. The P-values for the pairwise program comparisons (Tables 1.1–1.3) may be found in the appendix (Tables A.1–A.3).

In both the figures and tables, `OrthoMCL` is labeled as 'om' and `MultiParanoid` is labeled as 'mp'. `Notung` with single-link clustering, provided with the three species trees 't0', 't1', and 't2', are labeled as 'n0', 'n1', and 'n2', respectively; `Notung` with `OrthoMCL` and `MultiParanoid`, provided with the correct species tree 't0', are labeled as 'no' and 'nm', respectively. Similarly, `RAP` with single-link clustering, provided with the three species trees 't0', 't1', and 't2', are labeled as 'r0', 'r1', and 'r2', respectively; `RAP` with `OrthoMCL` and `MultiParanoid`, provided with the correct species tree 't0', are labeled as 'ro' and 'rm', respectively.

Lineage-specific duplication

To observe the frequency of inparalogs mistakenly identified as outparalogs, we simulated datasets with only lineage-specific duplications. With this dataset we are testing for the incorrect splitting of clusters due to overly stringent cutoffs in similarity-based methods and the propensity of maximum parsimony reconciliation to incorrectly infer ancestral duplications. Thus, we expected and indeed observed that mean and precision is 1.0 for all programs (data not shown).

`OrthoMCL` has significantly lower recall and higher precision than `MultiParanoid` and `RAP` has significantly lower recall and higher precision than `Notung` (Figure 1.3,Table 1.1). `OrthoMCL` has significantly lower recall than the reconciliation methods; the same applies to `MultiParanoid` starting at a duplication rate of 0.2 (data not shown). Using reconciliation on the output of similarity-based methods significantly lowers recall as expected. We also found that single-link clustering outperformed `OrthoMCL` and `MultiParanoid` in gene family construction. Reconciliation methods have lower recall than similarity-based methods if an incorrect species phylogeny is provided. However, `Notung` is able to handle small variations in the species phylogeny as recall is not significantly different between using t0 (correct tree) and t1 (minor change). Each increase in the duplication rate leads to a significant decrease in recall (data not shown).

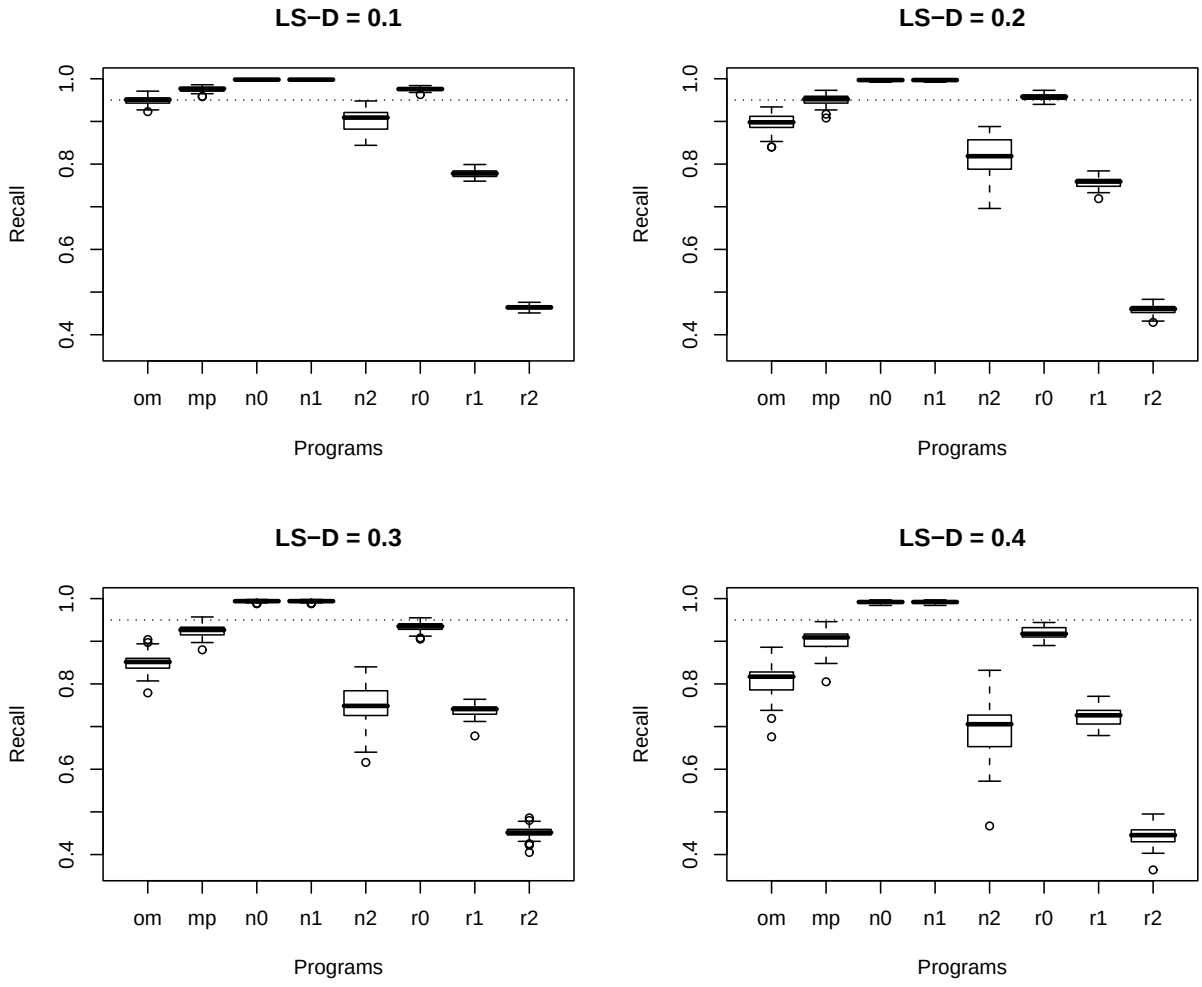
Table 1.1: Precision and recall at a lineage-specific duplication rate of 0.1.

	om	mp	no	nm	n0	n1	n2	ro	rm	r0	r1	r2
om	x											
mp	-	x										
no	+	+	x									
nm	-	+	-	x								
n0	-	-	-	-	x							
n1	-	-	-	-		x						
n2	+	+	+	+	+	+	x					
ro	+	+	+	+	+	+	-	x				
rm	-	+	-	+	+	+	-	-	x			
r0	-		-	-	+	+	-	-	-	x		
r1	+	+	+	+	+	+	+	+	+	+	x	
r2	+	+	+	+	+	+	+	+	+	+	+	x

Duplication

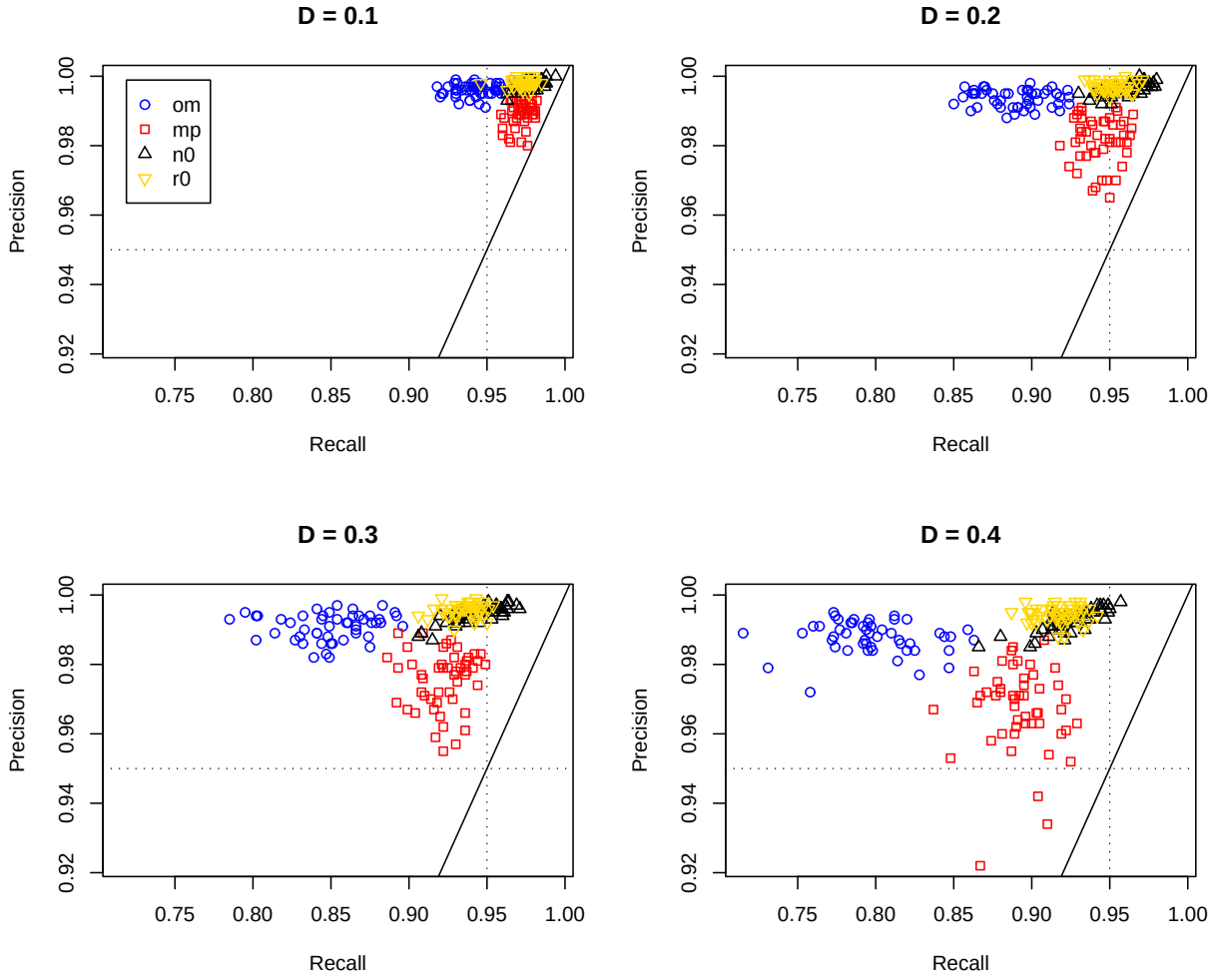
Once again, `OrthoMCL` has significantly lower recall and higher precision than `MultiParanoid` and `RAP` has significantly lower recall and higher precision than `Notung` (Figure 1.4 and Table 1.2). In this dataset, similarity-based methods perform worse than reconciliation methods for both precision and recall. Using reconciliation on the output of similarity-based methods increases precision and lowers recall as expected (Figure 1.5,1.6 and Table 1.2). Once again, single-link clustering is the best method for constructing gene families for reconciliation. Although for `RAP`, use of `OrthoMCL` and `MultiParanoid` results in significantly higher precision

Figure 1.3: Recall under increasing rates of lineage-specific duplication.



and lower recall than single-link clustering starting at a rate of 0.2 (data not shown), it is clear that the small gain in precision is offset by a large decrease in recall and is likely due to a tendency to push duplication nodes to the root of the phylogeny (Figure 1.6). As previously shown, there is no significant difference between the output of `Notung` using t0 and t1 (Figure 1.7 and Table 1.2). Otherwise, an incorrect species phylogeny will result in much lower recall for a small gain in precision compared to similarity-based methods (Figures 1.7,1.8). Each increase in the duplication rate leads to a significant decrease in both precision and recall (data not shown).

Figure 1.4: Precision and recall under increasing rates of duplication.



Duplication and loss

Next, we performed a comparatively more realistic experiment by simulating duplication and loss concurrently. We see the same pattern before with `OrthoMCL` having significantly lower recall and higher precision than `MultiParanoid` and `RAP` having significantly lower recall and higher precision than `Notung` (Figure 1.9 and Table 1.3). Similar to the duplication only

Figure 1.5: Effect of gene family clustering on `Notung` in the presence of duplication.

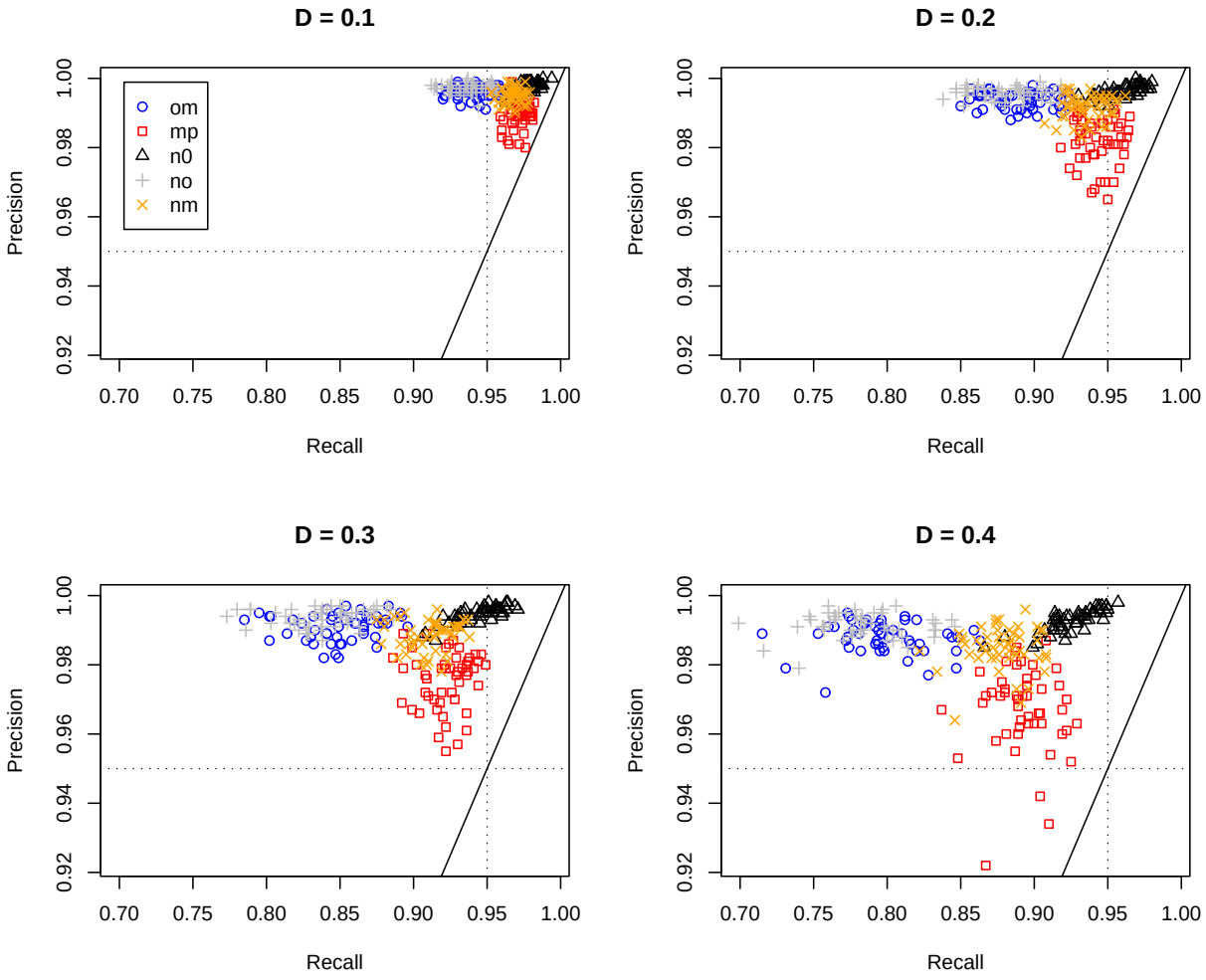


Figure 1.6: Effect of gene family clustering on `RAP` in the presence of duplication.

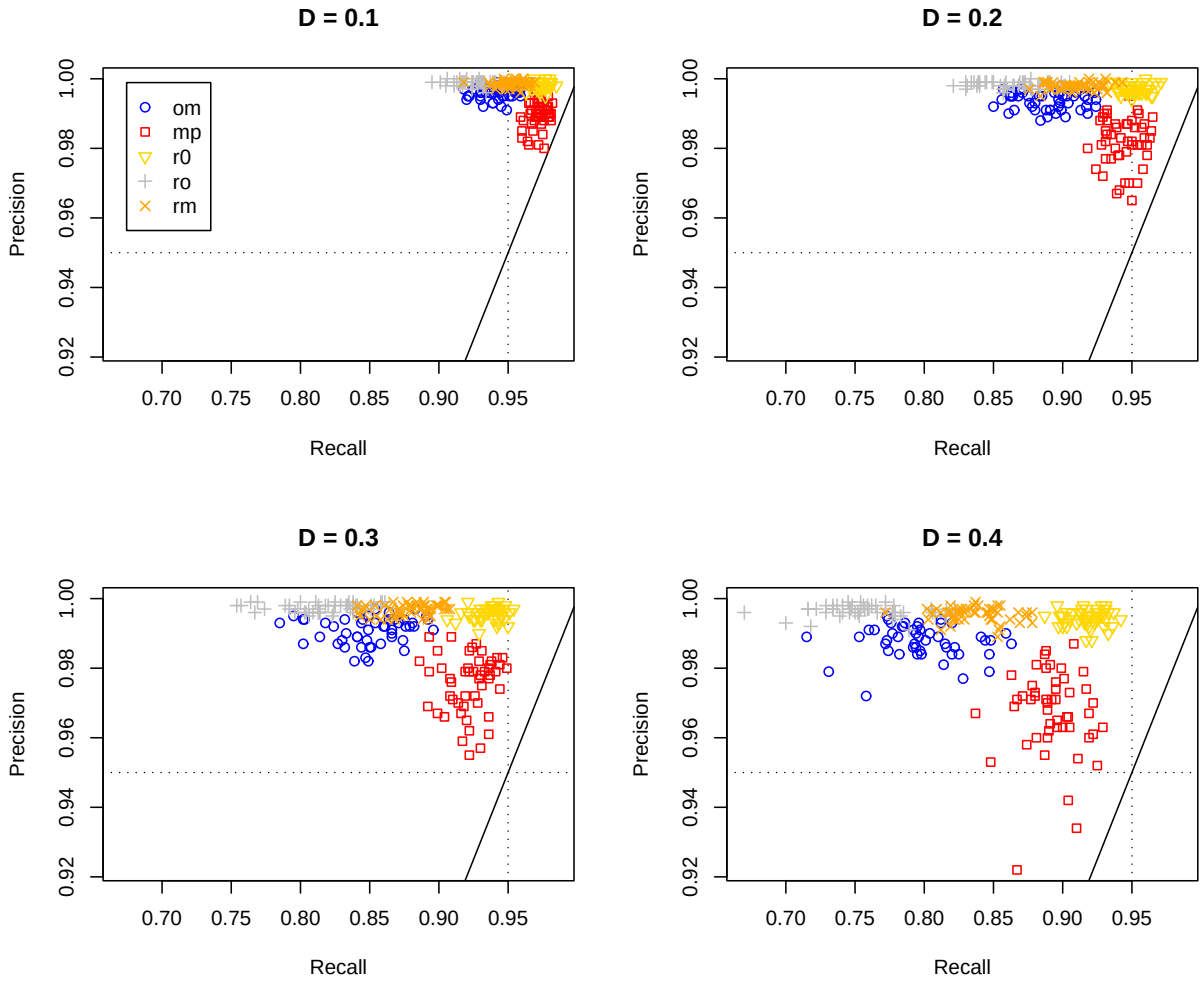


Figure 1.7: Effect of species phylogeny `Notung` in the presence of duplication.

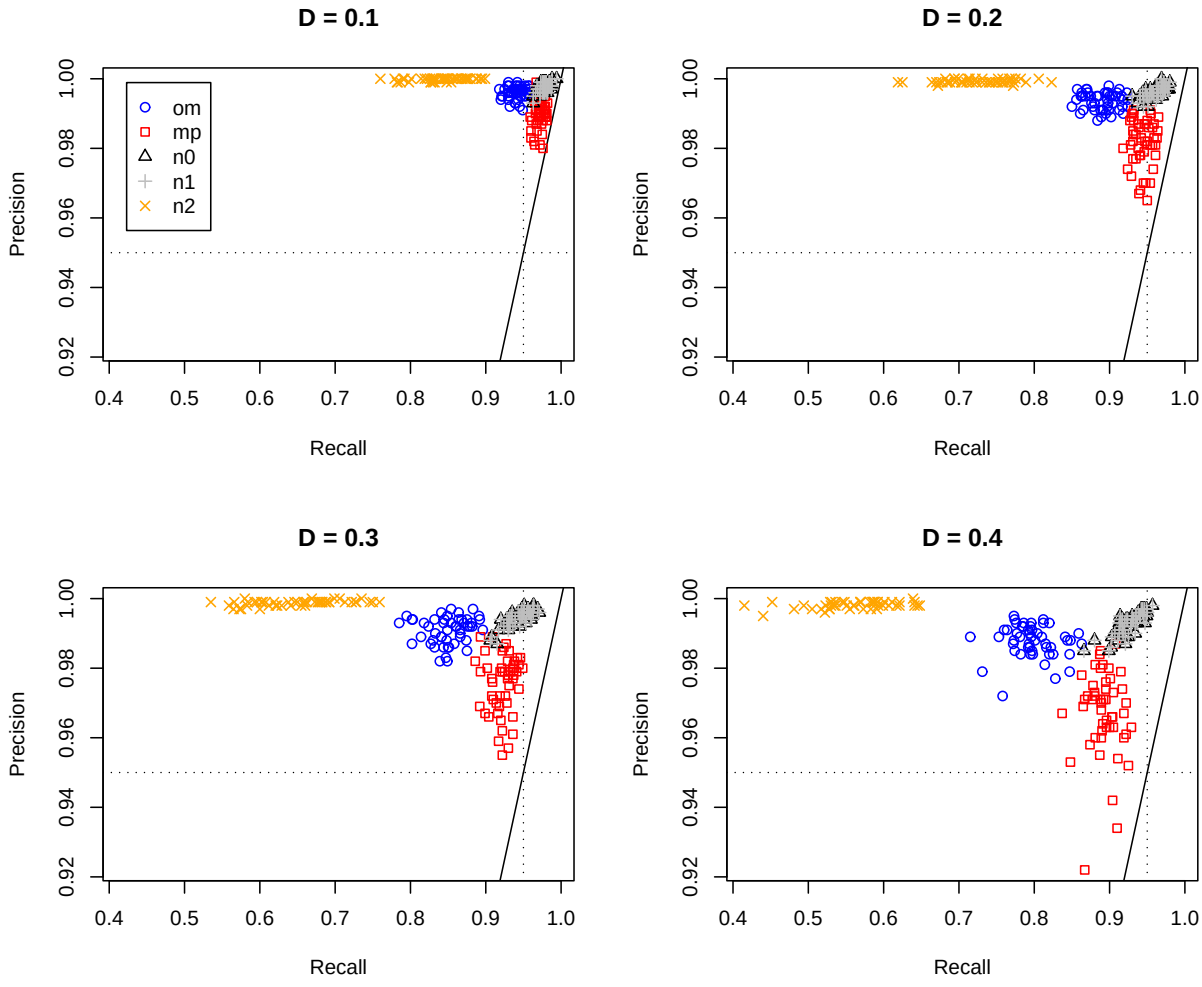


Figure 1.8: Effect of species phylogeny `RAP` in the presence of duplication.

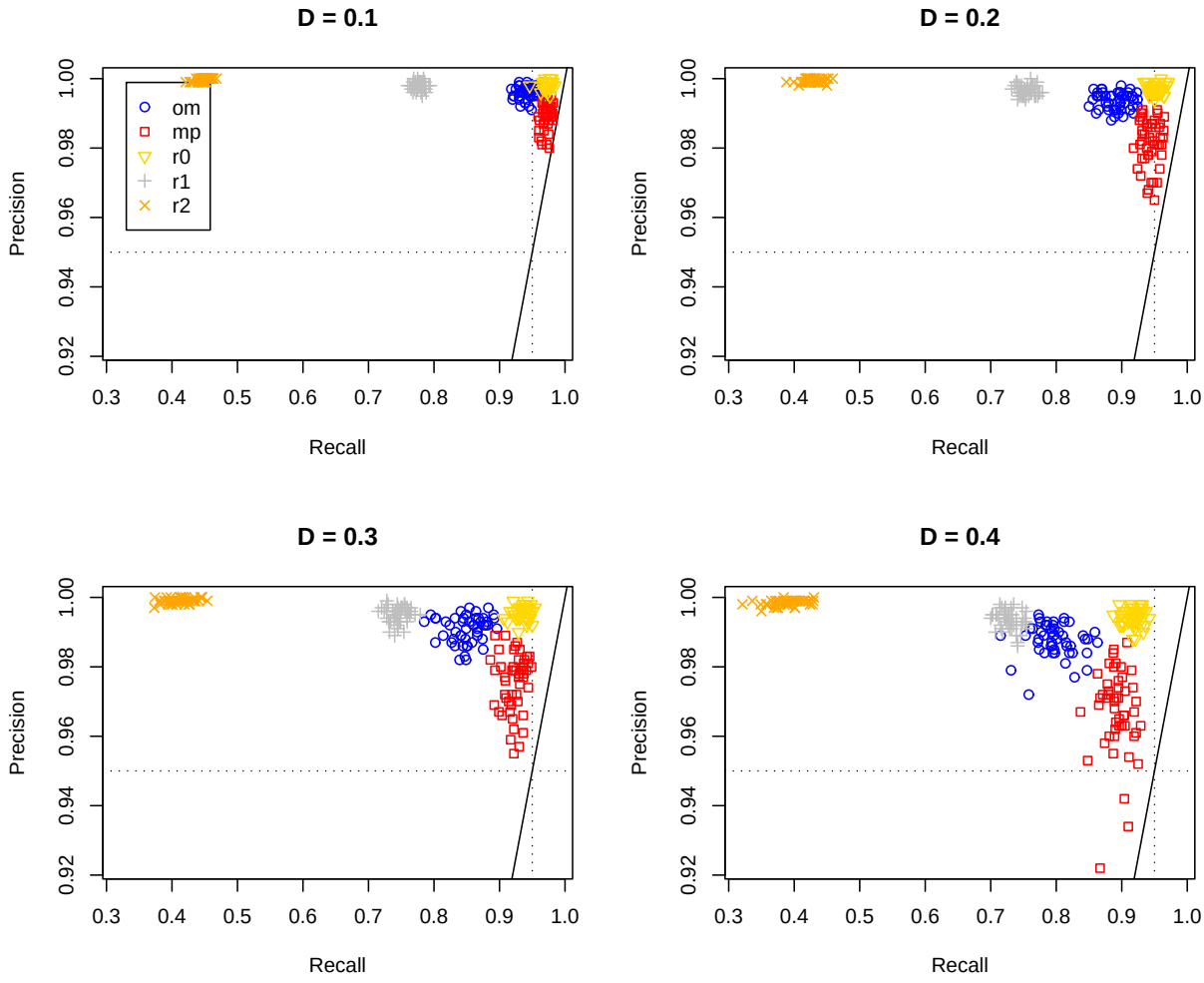


Table 1.2: Precision and recall at a duplication rate of 0.1.

	om	mp	no	nm	n0	n1	n2	ro	rm	r0	r1	r2
om	x	+	-	+	-	-	-	-	-	-	-	-
mp	-	x	-	-	-	-	-	-	-	-	-	-
no	+	+	x	+			-	-	-	-	-	-
nm	-	+	-	x	-	-	-	-	-	-	-	-
n0	-	-	-	-	x		-	-	-	-	-	-
n1	-	-	-	-		x	-	-	-	-	-	-
n2	+	+	+	+	+	+	x	+	+	+	+	
ro	+	+	+	+	+	+	-	x			+	-
rm	-	+	-	+	+	+	-	-	x		+	-
r0	-		-	-	+	+	-	-	-	x		-
r1	+	+	+	+	+	+	+	+	+	+	x	-
r2	+	+	+	+	+	+	+	+	+	+	+	x

dataset, similarity-based methods have significantly lower recall than reconciliation methods although for `MultiParanoid` this is only true starting at duplication and loss rates of 0.2 (data not shown). However, while `RAP` has the highest precision, `OrthoMCL` has significantly higher precision than `MultiParanoid` and `Notung`, which do not differ significantly across all four rate categories. As with the duplication only dataset, use of reconciliation on the output of similarity-based methods increases precision and lowers recall (Figure 1.10,1.11 and Table 1.3). Once again, the high precision and low recall means that using `OrthoMCL` or `MultiParanoid` with `RAP` simply leads to significantly lower recall compared to single-link clustering for gene family construction. However, for `Notung` single-link clustering is no longer best with respect to precision or recall. As expected, there is a tradeoff with significant increases to precision and corresponding decreases to recall as the gene families become smaller (`MCL` > `MultiParanoid` > single-link clustering). Like the previous experiments, incorrect species phylogenies result in a significant decrease in recall for `RAP` (Figure 1.13). `Notung` can still handle small deviations from the true phylogeny but this time with a small but significant decrease in recall (Figure 1.12). In all cases the species phylogeny does not have a significant effect on precision. Again, each concurrent increase in the duplication and loss rates leads to a significant decrease in both precision and recall (data not shown).

Lateral gene transfer

Finally, we assessed performance in the presence of LGT. As stated above, we expect abnormally low divergence between laterally transferred genes which will inflate FP rates for both reconciliation and similarity methods and thus we stopped at 10 replicates per rate category. Indeed for the first time we see higher recall than precision for all algorithms (Figure 1.14). Although we do not have enough data to detect significant differences between the programs, it appears that `Notung` performs the worst in this dataset and the same trends observed between `OrthoMCL` and `MultiParanoid` are still present. `RAP` still has the highest precision but its recall decreases less rapidly than the similarity-based methods. However,

Figure 1.9: Precision and recall under increasing rates of duplication and loss.

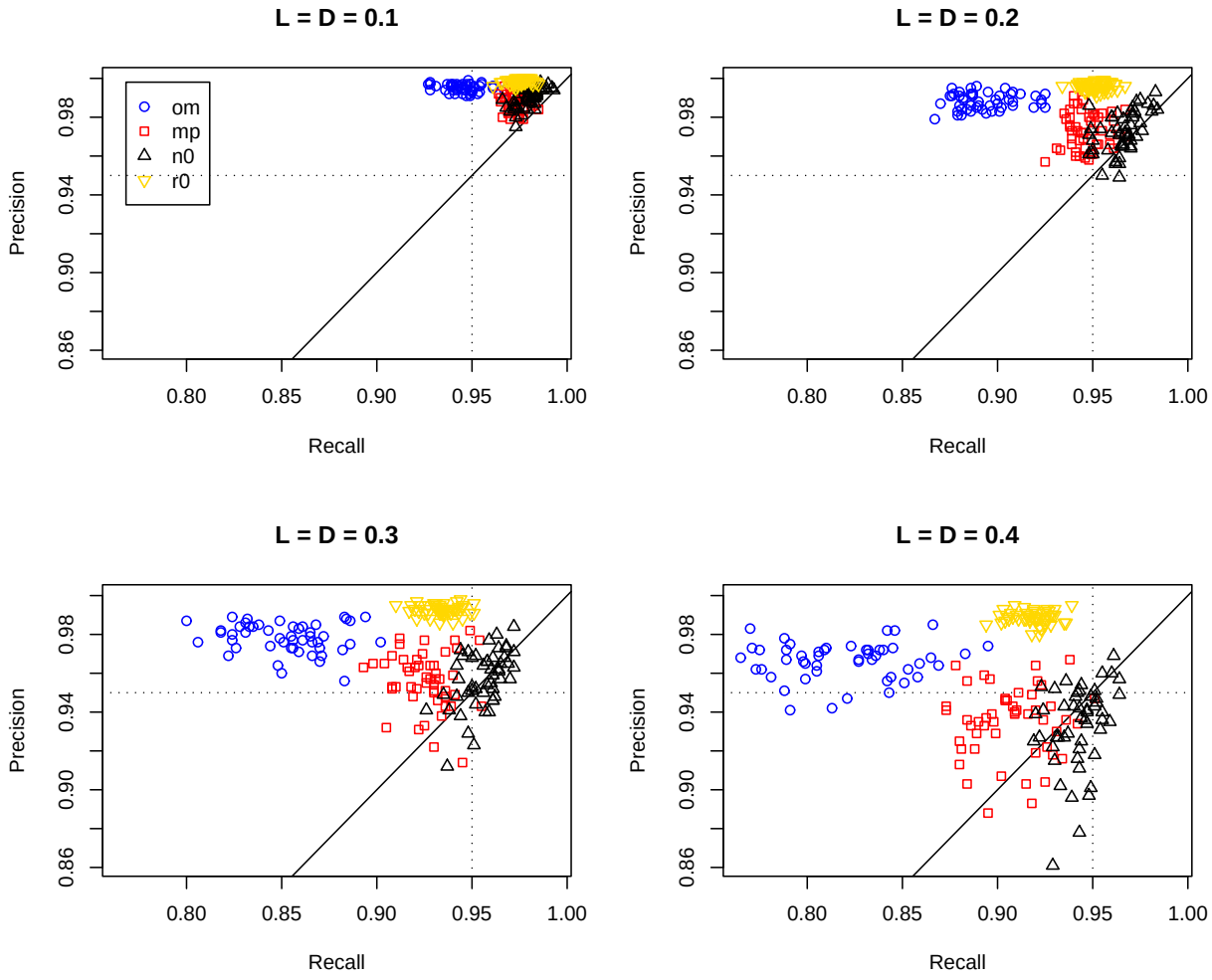


Figure 1.10: Effect of gene family construction on `Notung` in the presence of duplication and loss.

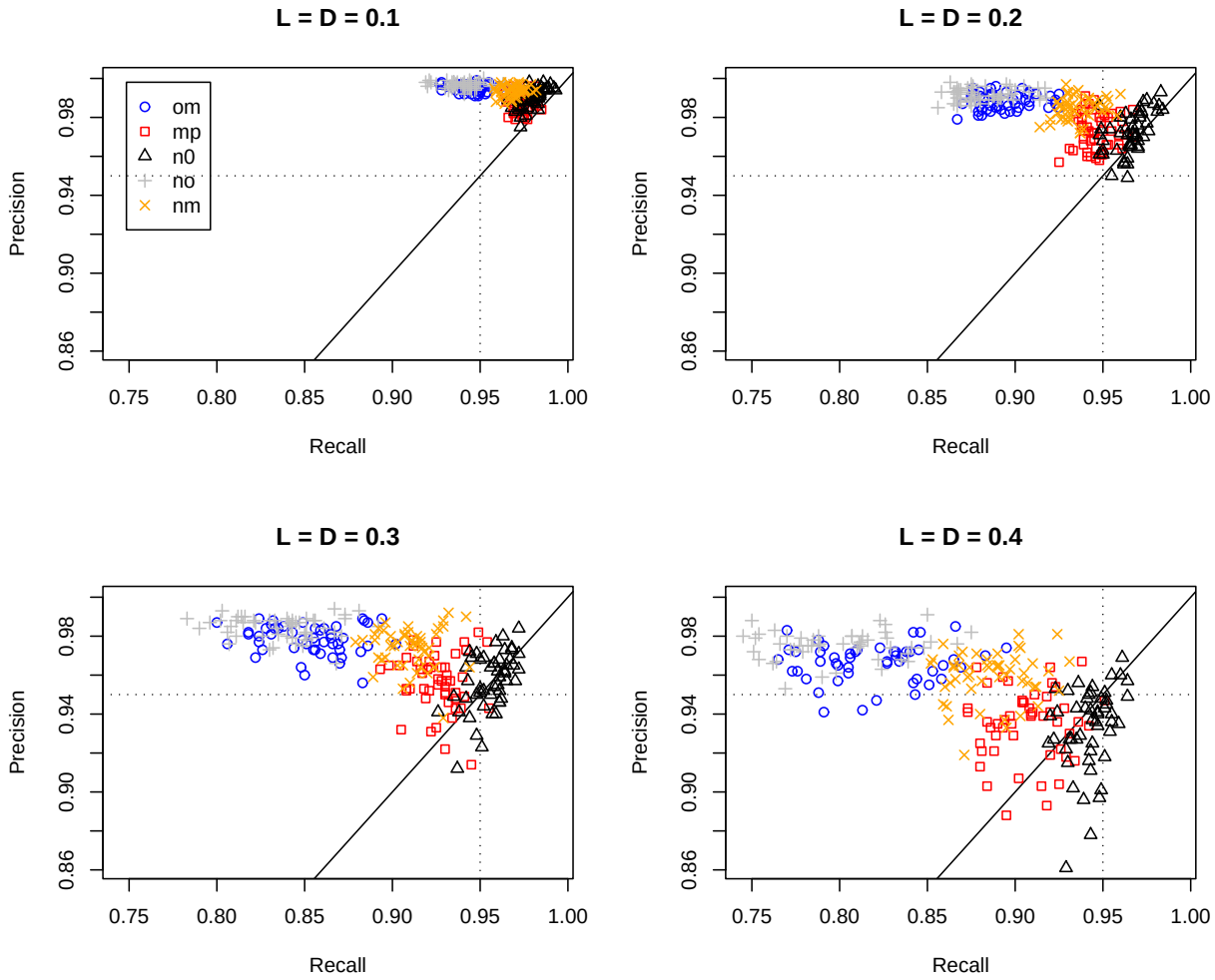


Figure 1.11: Effect of gene family construction on RAP in the presence of duplication and loss.

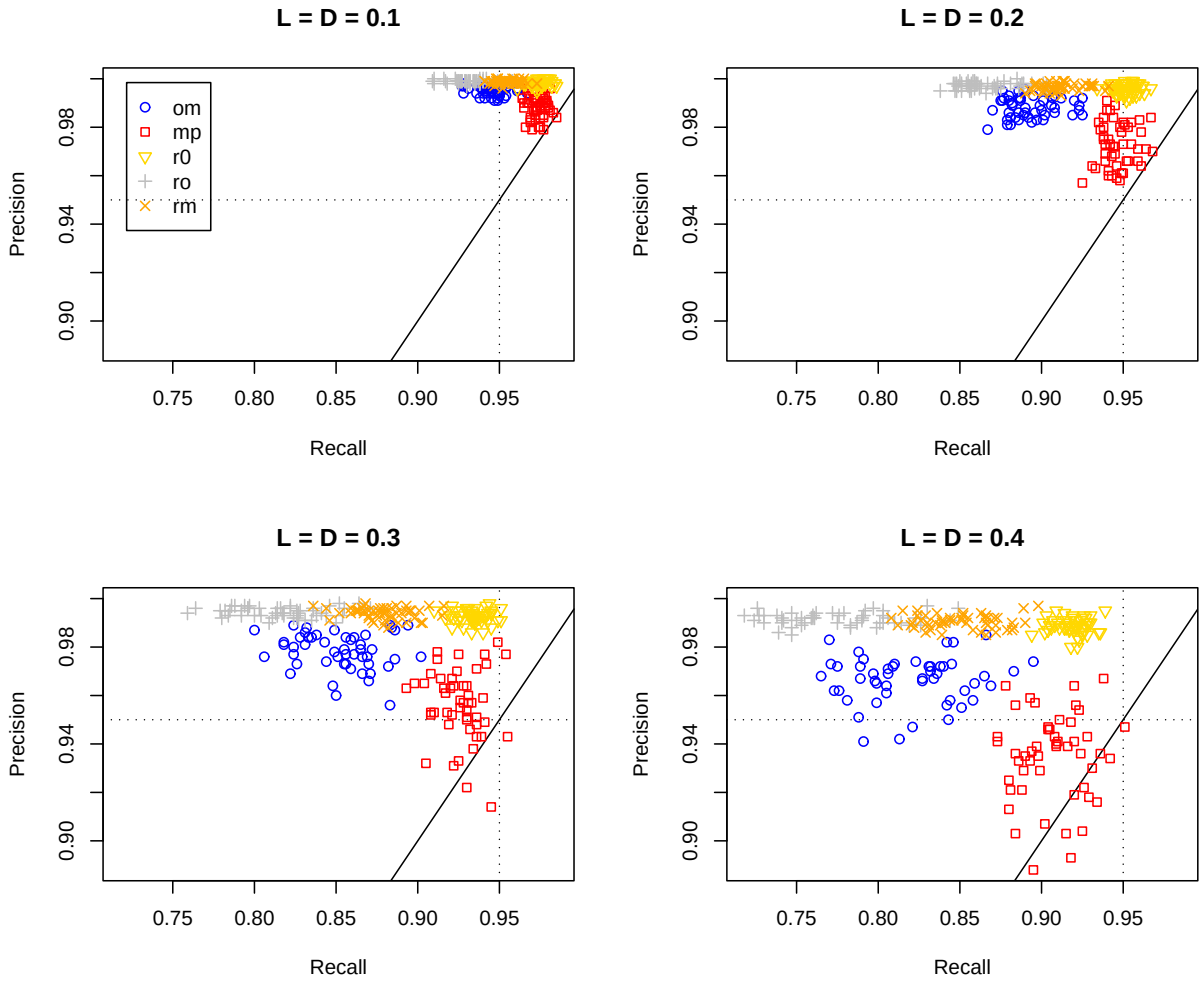


Figure 1.12: Effect of species phylogeny `Notung` in the presence of duplication and loss.

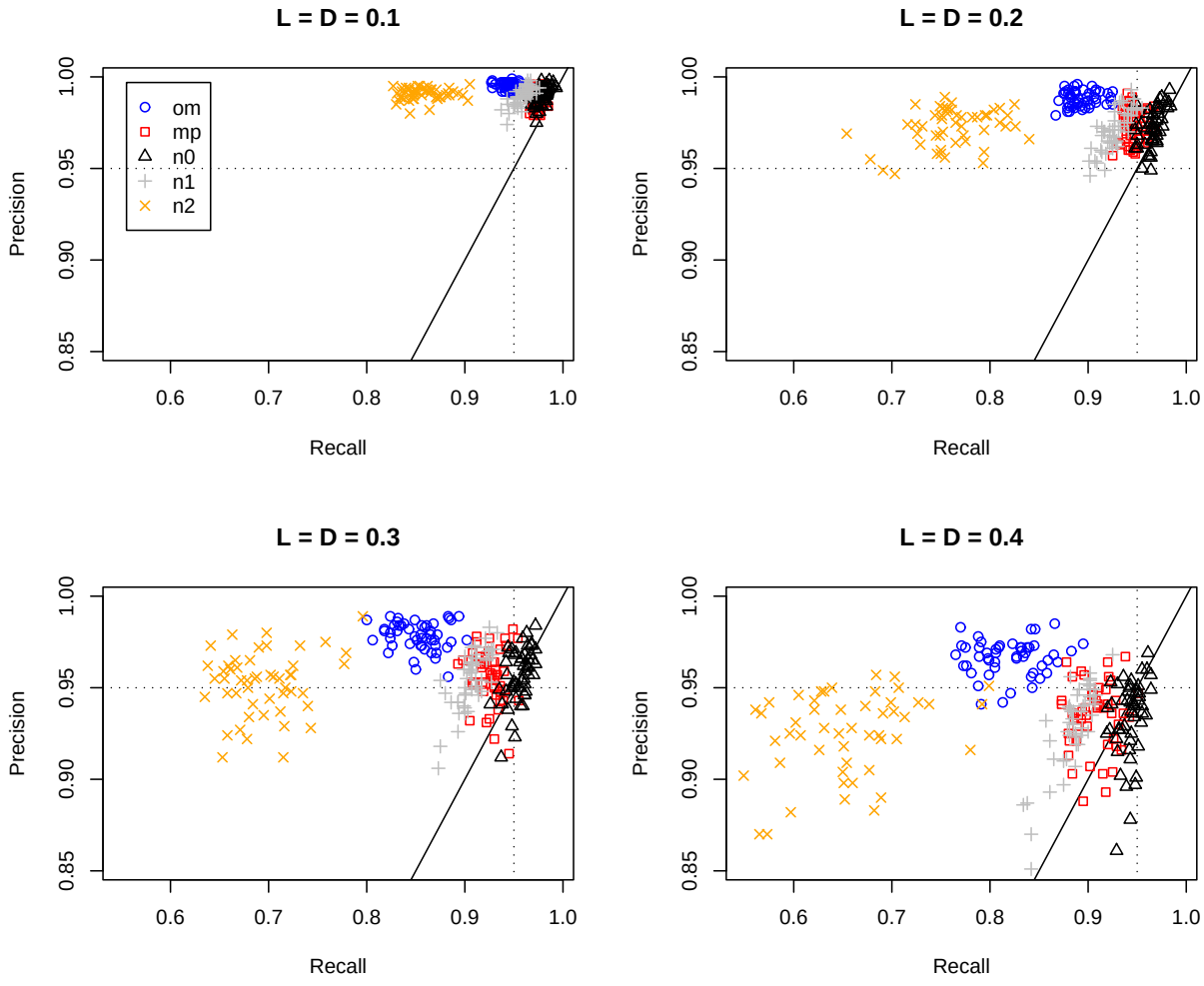


Figure 1.13: Effect of species phylogeny `RAP` in the presence of duplication and loss.

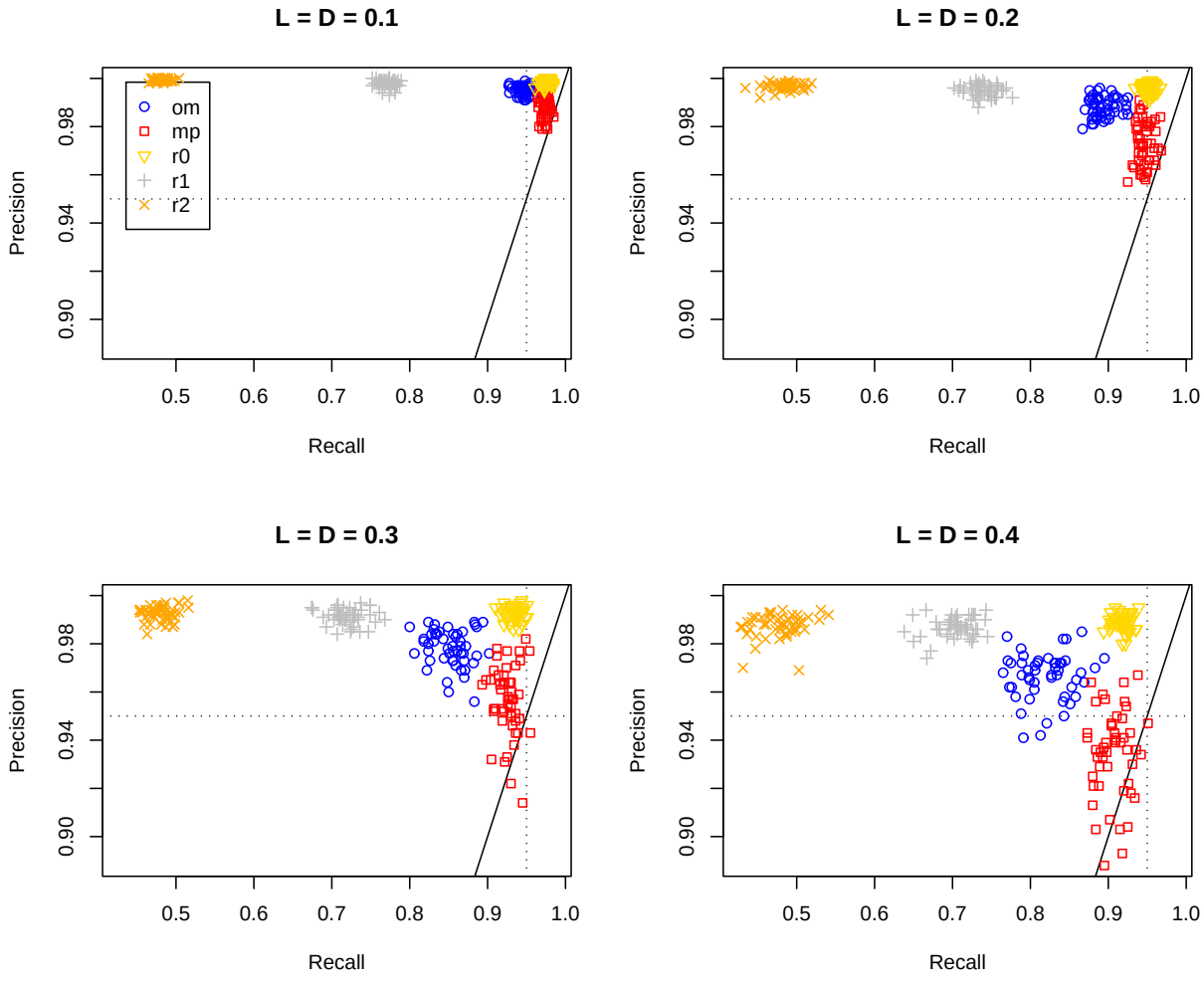


Table 1.3: Precision and recall at a duplication and loss rate of 0.1.

	om	mp	no	nm	n0	n1	n2	ro	rm	r0	r1	r2
om	x	+	-	+	+	+	+	-	-	-	-	-
mp	-	x	-	-	-		-	-	-	-	-	-
no	+	+	x	+	+	+	+	-	-	-	-	-
nm	-	+	-	x	+	+	+	-	-	-	-	-
n0	-	-	-	-	x	+		-	-	-	-	-
n1	-	+	-	+	+	x		-	-	-	-	-
n2	+	+	+	+	+	+	x	-	-	-	-	-
ro	+	+	+	+	+	+	-	x			+	-
rm	-	+	-	+	+		-	-	x		+	-
r0	-		-	-	+	-	-	-	-	x		-
r1	+	+	+	+	+	+	+	+	+	+	x	-
r2	+	+	+	+	+	+	+	+	+	+	+	x

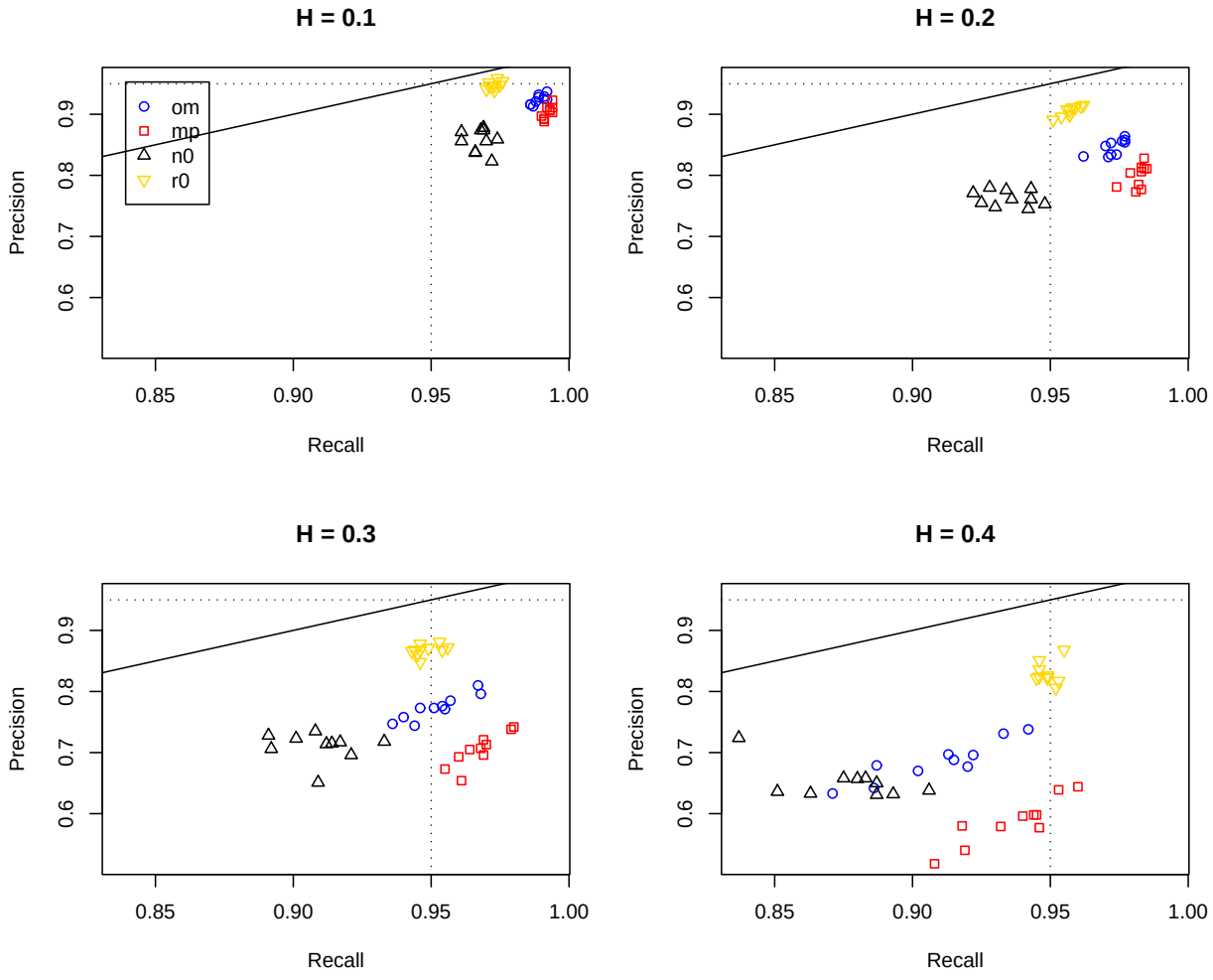
we do have enough data to detect significant decreases in both precision and recall as the rate of LGT increases (data not shown).

1.5 DISCUSSION

This project was started with the broad objective of comparing two fundamentally different approaches towards orthology detection using simulated datasets. It came to our attention after we started our study that Vilella *et al.* (2009) reported that tree reconciliation performs better than similarity-based methods using simulated data, but, to our knowledge, their results have yet to be published. Another motivating factor was the lack of a consensus with regards to the relative performance of these two approaches using real data where true orthology relationships are not known (Hulsen *et al.*, 2006; Chen *et al.*, 2007; Altenhoff and Dessimoz, 2009).

Two programs each were chosen to represent similarity-based algorithms (`OrthoMCL` and `MultiParanoid`) and reconciliation (`Notung` and `RAP`). We simulated six datasets under varying rates of macroevolutionary events with up to 50 replicates each and looked for consistent trends in the output of the four programs. First, we noticed that `OrthoMCL` always had significantly higher precision and lower recall compared to `MultiParanoid`, which suggests that `OrthoMCL` has a tendency to produce tight clusters or `MultiParanoid` has a tendency to produce loose clusters. The former is most likely true because `OrthoMCL` infers more clusters on average than `MultiParanoid` (data not shown) and consistently has the lowest recall of all four methods, including the lineage-specific duplication experiment designed to identify the tendency for tight clustering by similarity-based methods (Figure 1.3). The same trend was observed between `RAP` and `Notung`, and we believe that `RAP` is much more likely to incorrectly place duplication nodes at the root of the phylogeny because `Notung` has the highest recall of all programs and `RAP` lower recall in the lineage-specific duplication experiment designed to identify bias of duplications towards the root of the phylogeny (Figure 1.3). We further confirmed by inspection that `RAP` infers a much larger number of root duplications

Figure 1.14: Precision and recall under increasing rates of LGT.



than Notung. Notably, Hahn (2007) used Notung to demonstrate the bias towards inferring duplications at the root, which makes the results for RAP even more concerning.

Broadly speaking, tree reconciliation algorithms appear to have recall levels at least as high as similarity-based methods. In the more realistic situation where both duplication and loss is simulated, RAP has the highest precision, which is due to the parsimony bias. However, OrthoMCL actually has higher precision than Notung which is not significantly different from MultiParanoid. Although OrthoMCL has much lower recall compared to the other programs, the actual difference may in fact be inflated due to the nature of the assessment criteria. In particular, missing a single protein in a orthologous group of size n results in missing $n-1$ pairwise orthology calls.

Ultimately, the reconciliation algorithms assessed in this study either have higher precision (RAP) or recall (Notung) than similarity-based algorithms when the species phylogeny is known. The picture changes when we provide incorrect species trees. RAP in particular performs very poorly even with slight variations in the species phylogeny. Precision and recall do not adequately portray the fact that RAP produces much higher numbers of clusters and far fewer true positives, indicative of really tight clusters. An incorrect species phylogeny has no significant effect on the precision of Notung but the recall drops below MultiParanoid with small deviations (t1) and below OrthoMCL with large deviations (t2) from the true phylogeny (t0). Thus, Notung does not appear to excel in any category when the species phylogeny is not correct.

With regards to gene family construction, single-link clustering was significantly better in most datasets. However, in the more realistic dataset with both duplication and loss, the tightness of the clustering algorithm has a predictable positive effect on precision and negative effect on recall with respect to both Notung and the orthology identification algorithm itself. In theory one can choose a reconciliation-gene family combination to target a specific precision/recall tradeoff but that is likely influenced by the particular dataset.

Although the root bias of maximum parsimony reconciliation methods, especially that of RAP, is concerning, the results presented here were based on distance trees only as opposed to bootstrap trees. We did not present the results for Notung using the rearrangement feature on branch lengths because it appeared to have no effect or a negative effect on precision and recall, which is opposite of what one would expect (data not shown). Furthermore, we originally tested the effects of considering bootstrap support for both Notung and RAP but there did not appear to be a significant benefit, even though Hahn (2007) found that it does decrease the effect of root bias. Possibly consideration of bootstrap support simply is not beneficial under the simulated conditions, which do not accurately portray real data.

Our results are in agreement with those of Vilella *et al.* (2009) using simulated data and Chen *et al.* (2007) using latent class analysis, showing that reconciliation method can indeed outperform similarity-based methods. However, the true species phylogeny is not always known or agreed upon and even small deviations from the true phylogeny can shift the performance in favor of similarity-based methods. Thus, in practice we would recommend OrthoMCL or MultiParanoid, especially for comparisons between bacterial genomes.

One final note in favor our similarity-based methods is their speed advantage. Although the run-time of reconciliation algorithms has been studied (Zmasek and Eddy, 2001; Berglund-Sonnhammer *et al.*, 2006; Durand, Halldrsson and Vernot, 2006), from our experience and noted by Zmasek and Eddy (2001), the rate-limiting step is gene tree inference, not the reconciliation step. Although faster, even similarity-based algorithms are begin-

ning to stress computational resources due to the rapid increase in available genomes (Stein, 2010; Wall *et al.*, 2010). Kristensen *et al.* (2010) showed that clustering algorithms can substantially affect run-time and introduced an improved triangle merging algorithm for the `COG` method. However, run-time is dominated by the similarity search phase; eggNOG v2.0 switched from the slower and more accurate Smith-Waterman algorithm to `BLAST` (Muller *et al.*, 2010) while Wall *et al.* (2010) experimented with cloud computing for updating the Roundup database, based on the `RSD` algorithm. We also note that `MultiParanoid` (with `InParanoid` v4.1) is orders of magnitude slower than `OrthoMCL` v2.0 when applied to our datasets even though Li, Stoeckert and Roos (2003) observed the opposite using older versions of both programs. The difference is due to the implementation of the 2-pass `BLAST` strategy introduced in `InParanoid` v4.0. We predict this is the beginning of increased focus on algorithm speed, scalability, and implementation.

In the future, it may be better to use the simulated data and parameter estimation and simulation software included in the recent publication by Rasmussen and Kellis (2011), due to the problems with our simulation program. In particular, they never compared their program to the ones considered in this study, and, in fact, they did not include any maximum parsimony reconciliation or sequence-similarity methods.

Chapter 2

The contribution of gene fusion and fission to bacterial protein evolution

2.1 ABSTRACT

A method for identifying and mapping fusion and fission events onto a phylogeny was developed and applied to *Bacillaceae* genomes. In contrast to previous studies across longer evolutionary time scales, we found that gene fission is more common than fusion in bacterial genomes. Fusion and fission events are generally rare across the *Bacillaceae* and are genome-specific, but we unexpectedly uncovered a large number of fissions specific to the genetically monomorphic *Bacillus anthracis* lineage. Our results, based on deeper taxonomic sampling than previous studies, suggest that the *B. anthracis* lineage may be under an accelerated rate of gene fragmentation, which is a common evolutionary trend found in lineages that have recently become host-restricted. Sequencing and annotation errors are unlikely to be a large factor in our results, considering the large number of fissions shared across multiple *B. anthracis* strains and the consistency between fission patterns and previously published single nucleotide polymorphism (SNP) and variable number tandem repeat (VNTR) data. We hypothesized that other bacteria evolving under similar conditions would also exhibit detectable fission patterns. *Yersinia pestis* meets this requirement and is thought to be in the initial stages of reductive genome evolution. Although fewer in number, we found that fission patterns in *Yersinia pestis* are consistent with independently inferred phylogenies from SNPs and genome rearrangements. Genomes from *Salmonella enterica* were further used to confirm this trend of fission accumulation in host-restricted lineages. The results presented here have contributed to our knowledge regarding the contributions of fusion and fission to protein evolution, the evolution of *B. anthracis*, and other genetically monomorphic pathogens, and has implications on comparative genomics in the presence of gene fragmentation.

2.2 INTRODUCTION

Of the five elementary gene evolution events recognized, gene fusion and fission are regarded as contributing the least to gene evolution (Koonin, 2005; Nakamura, Itoh and Martin, 2007), with gene duplication and loss considered highly important to eukaryotes and lateral gene transfer (LGT) considered more predominant in prokaryotes; however, fusion and fission may

play an important role in protein domain architecture evolution (Pasek, Risler and Brzellec, 2006; Weiner and Bornberg-Bauer, 2006; Fong *et al.*, 2007). Gene fusion and fission may occur due to substitutions, frameshift mutations, or recombination events (Durrens, Nikolski and Sherman, 2008). Substitutions and frameshifts that introduce or remove a stop codon will result in fission or fusion, respectively, while recombination within a gene or at its boundary will always result in fusion or fission. It is more likely that fission will inactive a gene rather than not and is likely the initial step in pseudogenization, but these events would not be observable over long evolutionary time. Experimentally verified examples of naturally occurring gene fusion have been reported (Stover *et al.*, 2005); particularly interesting are reports of lateral transfer of a gene fusion product (Gopal *et al.*, 2005) and fusion of a pair of genes in *Helicobacter* likely caused by a frameshift mutation between genes that overlap in closely related genera (Zakharova *et al.*, 1999). Other interesting reports include the fusion of a *Mycoplasma* gene with a eukaryotic xenolog (Skamrov *et al.*, 2001) and identification of three independent fissions in the history of the DNA polymerase I gene family (Koonin, 2005).

Large-scale gene fusion and fission detection strategies were originally developed not to study their impact on gene evolution but for computational prediction of possible functional association and interaction between proteins (Enright *et al.*, 1999; Enright and Ouzounis, 2001; Marcotte *et al.*, 1999; Yanai, Derti and DeLisi, 2001; Kamburov *et al.*, 2007). Yanai, Derti and DeLisi (2001) illustrated the idea using two "fusion links" in different genomes that allow the association of three separate proteins in *Mycoplasma genitalium* involved in sequential steps of the well known glycolysis pathway. Although these methods appear to be both useful and successful, the hypothesis that gene fusion events are selectively advantageous is debatable (Doolittle, 1999). More recently, there have been attempts to assess the contribution of fusion and fission to gene evolution across broad evolutionary distances (Hua *et al.*, 2002; Pasek, Risler and Brzellec, 2006; Cock and Whitworth, 2007).

Surveys of fusion and fission events across prokaryotes (Snel, Bork and Huynen, 2000) and all three domains of life (Kummerfeld and Teichmann, 2005) revealed that fusion is more common than fission, with Kummerfeld and Teichmann (2005) reporting that fusions are about four times more likely. Studies focusing on restricted taxonomic groups, between two plant genomes (Nakamura, Itoh and Martin, 2007) and complete or nearly complete fungal genomes (Durrens, Nikolski and Sherman, 2008), found that fusion and fission rates are about equal; Nakamura, Itoh and Martin (2007) presented to our knowledge the only absolute rate estimate of gene fusion and fission at 1 to 2×10^{-11} events per gene per year, which is about 100 times slower than nucleotide substitutions for their dataset. The apparent discrepancy in relative rates of fusion and fission is not too surprising due to the difference in evolutionary distances considered between the two types of studies. At shorter evolutionary distances, it is not apparent why fusion should be more common than fission. In fact, fusion should be less likely because only a mutation that disrupts a stop codon could lead to a fusion but multiple mutation points could potentially lead to a fission; if the mechanism is recombination then both are expected to be equally likely events. However, a fission event is more likely to inactivate a gene. Thus, at long evolutionary distances, where few orthologs are shared between genomes, only events without detrimental effects would be observable and fusion events are more likely to predominate. Furthermore, the latter two studies were specific to eukaryotic taxa, which have different gene structure and organization compared to bacteria and are likely to have different fusion/fission dynamics.

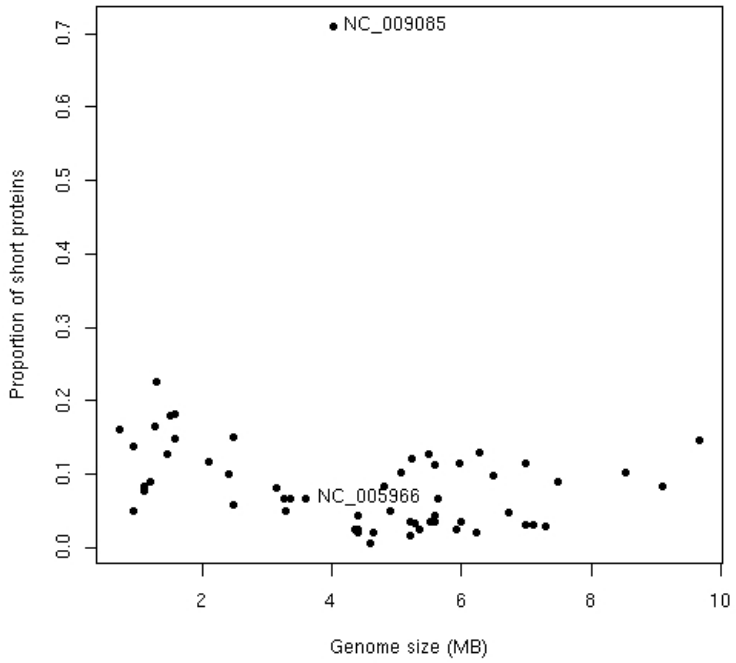
A group-specific study was performed recently in bacteria using a semi-automatic method for the identification and mapping of pseudogenes on the *Rickettsia* phylogeny where genome reduction is well known and characterized (Fuxelius *et al.*, 2008) while Sakharkar, Sakharkar and Chow (2006) scanned two *Helicobacter* strains for fusions and fission. Sakharkar, Sakharkar and Chow (2006) suggested some of their events are fusions resulting from loss of a start codon and incorporation of intergenic DNA, which we argue is a dangerous assumption when considering only two genomes. For example, in figure 1 of their paper they propose a gene fusion event in one strain due to the loss of a stop codon and the incorporation of the intergenic region into the fusion gene. An equally likely scenario would be the introduction of a stop codon the the other strain, leading to a gene fission. A `BLAST` search against the NCBI `RefSeq` database confirmed our suspicion, showing that the composite gene is the predominant form, not the split form. This highlights the potential for error when attempting to distinguish between fusion and fission across only two species without either using an outgroup (Nakamura, Itoh and Martin, 2007) or including more genomes and making use of phylogenetic information (Snel, Bork and Huynen, 2000; Kummerfeld and Teichmann, 2005; Durrens, Nikolski and Sherman, 2008; Fuxelius *et al.*, 2008). We are not aware of any phylogenetic study of fusion and fission in a restricted taxonomic group of bacteria.

Fusion detection may also be important in the study of protein length variation across evolutionary time if left undetected. Fuxelius *et al.* (2008) reported that the difference in median protein length of orthologs between two *Rickettsia* strains (Brocchieri and Karlin, 2005) was due to the presence of many fragmented proteins in one of the genomes. This suggests that annotation pipelines and methodologies may have a significant effect on downstream analyses even between closely related bacterial strains. Similarly, in a survey of protein lengths of homologs in pairwise genome comparisons, we found that *Acinetobacter baumannii* ATCC 17978 an unusual number of short proteins that can be partially explained by gene fragmentation (Figure 2.1).

Most algorithms and programs used in fusion/fission detection are not readily available for download. Many of them would also be insufficient for our analyses; for example functional detection does not require the mapping of events onto a phylogeny. The program DomainTeams (Pasek *et al.*, 2005) was used to study the contribution of fusion and fission to the evolution of multi-domain bacterial proteins but was designed to infer events between proteins with protein domain annotation, can only detect fusions between adjacent genes, and relaxes the orthology criterion (Pasek, Risler and Brzellec, 2006). In particular, relaxing the orthology criterion increases the risk of false positives and is undesirable for the mapping of events onto a phylogeny of strains in the same family compared to Pasek, Risler and Brzellec (2006) who used strains across the Bacteria. An online database was developed for computational predictions of fusion and fission events in prokaryotes (Suhre and Claverie, 2004) but the website has not been updated, the algorithm is not available for download, and the algorithm is restricted to pairwise genome comparisons.

In this study, we perform, to the best of our knowledge, the first study of fusion and fission to a restricted set of bacterial genomes, the *Bacillaceae* family, to determine their contribution to bacterial protein and genome evolution. We chose this taxonomic group because of the large number of genomes sequenced, including the well studied *Bacillus subtilis* (Kunst *et al.*, 1997; Barbe *et al.*, 2009), the heterogeneous mixture of short and long genetic distances within and between well-defined clades, and the wide range of habitats and lifestyles represented (Alcaraz *et al.*, 2010). The completely sequenced *Bacillaceae* strains may be

Figure 2.1: Survey of protein length differences between homologous proteins between bacterial genomes.



broadly classified into four main groups, referred to in this chapter as the Alkaliphilic, Cereus, Geobacillus, and Subtilis groups, although the relationships between these groups is currently uncertain (Hao and Golding, 2009; Alcaraz *et al.*, 2010); relationships within the Cereus are also uncertain due to low genetic separation. Recently, 990 SNPs were used to identify 3 main lineages, 2 sublineages, and root a collection of genetically monomorphic *B. anthracis* strains (Pearson *et al.*, 2004; Van Ert *et al.*, 2007). This allows us to compare fusion and fission dynamics at varying levels of genetic divergence and compare them to previous results.

The placement of *Oceanobacillus iheyensis* may vary depending on the markers used and strains included, but is generally placed ancestral to all sequenced genomes. We also chose the *Bacillaceae* family because of the potential for comparison of the contributions of fusion and fission with that of LGT from previous studies involving genomes in this group (Hao and Golding, 2006, 2008b, 2009, 2010). We developed a suite of programs called `FusionFinder` to identify and map fusion and fission events onto a phylogeny using the Camin-Sokal parsimony method. We found that fission is much more common than fusion, although the majority of the former likely represent pseudogenization. Unexpectedly, we also found that a large number of events were mapped to the *B. anthracis* clade. This overrepresentation can only be partially explained by the well known loss of motility and inactivation of the PlcR regulon unique to the *B. anthracis* lineage within the Cereus group (Slamti *et al.*, 2004; Rasko *et al.*, 2005; Gohar *et al.*, 2008; Kolstø, Tourasse and Økstad, 2009). Orthologs of the fissioned proteins are broadly distributed within the Cereus group but not in other *Bacillaceae* genomes. There have been previous reports of a large number of pseudogenes in *B. anthracis* Ames (Price *et al.*, 2005) and gene loss in *B. anthracis* Ames Ancestor (Yu,

2009). Collectively, our results in combination of those of Price *et al.* (2005) and Yu (2009) suggest that genes within the *B. anthracis* lineage are in the process of deletion and that many genetic fossils are still present in the genomes. The complete consistency between fission patterns and a *B. anthracis* phylogeny constructed from a combined single nucleotide polymorphism (SNP) and variable number tandem repeat (VNTR) analysis (Van Ert *et al.*, 2007), COG category analysis of fissioned genes, and an expanded phylogenetic analysis of motility and chemotaxis operons in the Cereus group suggest that *B. anthracis* is currently experiencing an accelerated rate of pseudogenes formation that is common to host-restricted lineages and may lead to genome reduction. Reductive genome evolution is well known and commonly associated with pathogenic bacteria possessing small genomes (*Rickettsia*), and genomes smaller than closely related non-pathogenic strains (*Mycobacterium*) (Andersson and Andersson, 1999).

Interestingly, *Yersinia pestis* shares many similarities with *B. anthracis*, despite being a Gram-negative bacterium. *Y. pestis* likely evolved recently from *Y. pseudotuberculosis*, is known to contain many putative pseudogenes, often associated with ISs, loss of motility, may have been affected by a population bottleneck, and is suggested to be in the initial stages reductive genome evolution (Parkhill *et al.*, 2001). Both exhibit low genetic diversity which suggests recent origin (Achtman, 2004; Keim *et al.*, 2004). Facultative and obligate pathogens are thought to have lower effective population sizes and increased genetic drift (Ochman and Davalos, 2006), which is also compatible with low genetic diversity. We applied our method to *Y. pestis* and *Y. pseudotuberculosis* genomes and found that fission patterns did not fully match the SNP phylogeny of (Eppinger *et al.*, 2010) but was consistent with genome rearrangement patterns (Darling, Mikls and Ragan, 2008), a smaller SNP phylogeny (Chain *et al.*, 2006), and a larger SNP phylogeny (Morelli *et al.*, 2010); lack of sufficient signal in the full dataset resulted in an unresolved node, illustrating the expected limitations of this method with respect to divergence time.

The pseudogenization process in the *B. anthracis* lineage may have gone largely unnoticed due to the large genome size of *B. anthracis*, lack of observable size difference with other Cereus group members, and unexpected genomic differences between highly genetically monomorphic strains lacking additional indicators such as significant genomic rearrangements and abundance of insertion sequences in *Y. pestis*. This is further confirmation that the emergence of *B. anthracis* was recent (Kolstø, Tourasse and Økstad, 2009).

2.3 MATERIALS AND METHODS

A collection of 33 completed *Bacillaceae* genomes were initially selected for the analysis. *B. cereus* CI and an additional six draft *B. anthracis* genomes were later included for additional analyses (Table 2.1). Plasmids were included in the analysis although they do not change the patterns observed in our results (Figures 2.20,2.23). Separate analyses were also performed on 19 genomes from the genus *Yersinia* and 10 genomes from *Salmonella enterica* (Table 2.2,Table 2.3).

The `FusionFinder` suite is composed of four programs `FusionFinder`, `MultiFusion`, `FusionScanner`, and `FusionMapper`, along with a set of utility programs for data acquisition from NCBI and automatic analysis of a group of genomes (Figure 2.2). `FusionFinder` predicts fusion and fission events in pairwise genome comparisons using a relaxed form of reciprocal best hits (RBH); it also scans genomic sequences and may optionally make use of

Table 2.1: *Bacillaceae* taxa used in the study. Informal group names of monophyletic subgroups used within the text are as indicated. Asterisks trailing accession numbers indicate 12X draft genomes; all other genomes are finished.

Group	ID	Accession	Name
Outgroup	Oi	NC_004193	*O. iheyensis*
Alkaliphilic	Bcl	NC_006582	*B. clausii*
Alkaliphilic	Bh	NC_002570	*B. halodurans*
Alkaliphilic	Bps	NC_013791	*B. pseudofirmus*
Geobacillus	Af	NC_011567	*A. flavithermus*
Geobacillus	Gw	NC_012793	*Geobacillus* sp. WCH70
Geobacillus	Gt	NC_009328	*G. thermodenitrificans*
Geobacillus	Gk	NC_006510	*G. kaustophilus*
Geobacillus	Gy	NC_013411	*Geobacillus* sp. Y412MC61
Subtilis	Bpu	NC_009848	*B. pumilis*
Subtilis	Bam	NC_009725	*B. amylofiquaciens*
Subtilis	Bs1	NC_000964	*B. subtilis* 168
Subtilis	Bs2	NC_014479	*B. subtilis* W23
Subtilis	Bl1	NC_006270	*B. licheniformis* ATCC 14580
Subtilis	Bl2	NC_006322	*B. licheniformis* DSM13
Cereus	Bcy	NC_009674	*B. cytotoxicus*
Cereus	Bw	NC_010184	*B. weihenstephanensis*
Cereus	Bc1	NC_011772	*B. cereus* G9842
Cereus	Bc2	NC_004722	*B. cereus* ATCC 14579
Cereus	Bc3	NC_011725	*B. cereus* B4264
Cereus	Bc4	NC_003909	*B. cereus* ATCC 10987
Cereus	Bc5	NC_011658	*B. cereus* AH187
Cereus	Bc6	NC_011969	*B. cereus* Q1
Cereus	Bc7	NC_006274	*B. cereus* E33L
Cereus	Bt1	NC_008600	*B. thuringiensis* Al Hakam
Cereus	Bc8	NC_012472	*B. cereus* 03BB102
Cereus	Bt2	NC_005957	*B. thuringiensis* konkukian
Cereus	Bc9	NC_011773	*B. cereus* AH820
Cereus	Bci	NC_014335	*B. cereus* CI
Anthracis	Ba1	NZ_AAEO01000000*	*B. anthracis* A1055
Anthracis	Ba2	NZ_AAEQ01000000*	*B. anthracis* KrugerB
Anthracis	Ba3	NZ_AAEN01000000*	*B. anthracis* CNEVA 9066
Anthracis	Ba4	NZ_AAER01000000*	*B. anthracis* Western North America
Anthracis	Ba5	NZ_AAEP01000000*	*B. anthracis* Vollum
Anthracis	Ba6	NC_012581	*B. anthracis* CDC 684
Anthracis	Ba7	NZ_AAES01000000*	*B. anthracis* Australia 94
Anthracis	Ba8	NC_005945	*B. anthracis* Sterne
Anthracis	Ba9	NC_012659	*B. anthracis* A0248
Anthracis	Ba10	NC_003997	*B. anthracis* Ames
Anthracis	Ba11	NC_007530	*B. anthracis* Ames Ancestor

Table 2.2: *Yersinia* taxa used in the study.

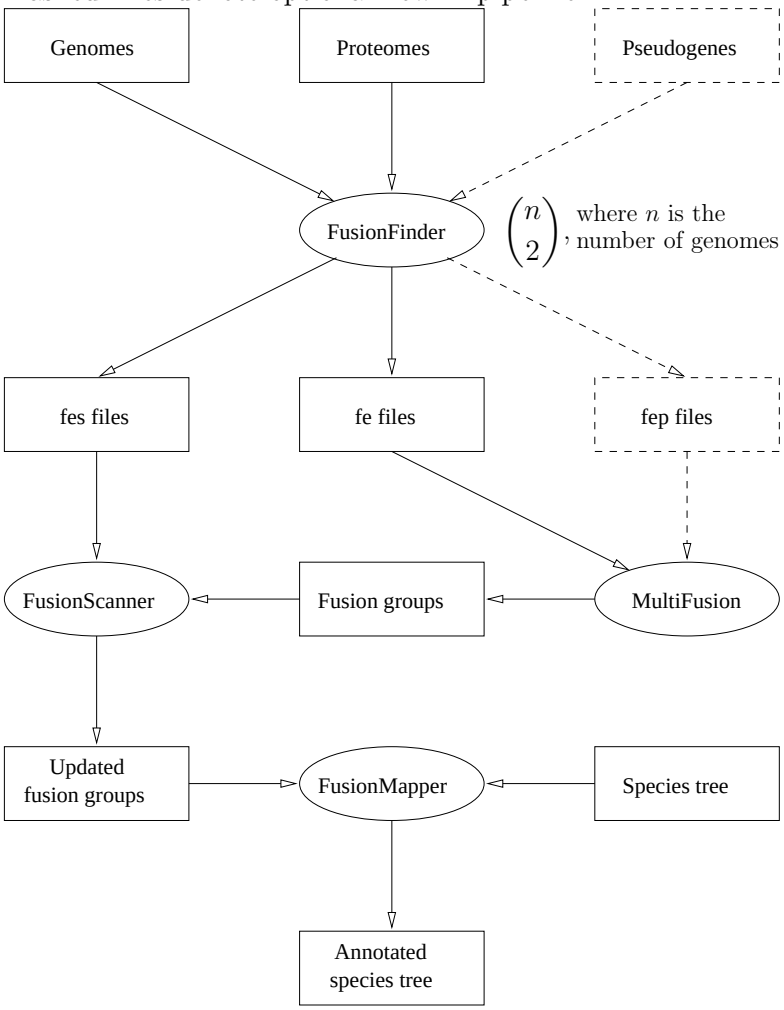
ID	Accession	Name
Yen	NC_008800	*Y. enterocolitica* 8081
Yps	NC_006155	*Y. pseudotuberculosis* IP 32953
0.PE2	NC_009381	*Y. pestis* Pestoides F
0.PE3	NC_010159	*Y. pestis* Angola
0.PE4	NC_005810	*Y. pestis* 91001
0.ANT2	NZ_AAYU01000001	*Y. pestis* B42003004
2.ANT1a	NC_008149	*Y. pestis* Nepal516
2.ANT1b	NZ_ACNQ01000001	*Y. pestis* Nepal516A
2.MED2	NC_004088	*Y. pestis* KIM
2.MED1	NZ_AAYT01000001	*Y. pestis* K1973002
1.ANT1a	NC_008150	*Y. pestis* Antiqua
1.ANT1b	NZ_AAYR01000001	*Y. pestis* UG05-0454
1.IN3	NZ_AAYV01000001	*Y. pestis* E1979001
1.ORI1a	NZ_ABCD01000001	*Y. pestis* CA88-4125
1.ORI1b	NC_003143	*Y. pestis* CO92
1.ORI2	NZ_ABAT01000001	*Y. pestis* F1991016
1.ORI3a	NZ_AAOS02000001	*Y. pestis* IP275
1.ORI3b	NZ_AAYS01000001	*Y. pestis* MG05-1020
FV1	NZ_AAUB01000001	*Y. pestis* FV-1

Table 2.3: *S. enterica* taxa used in the study. Host specificity and type of infection is also noted.

ID	Accession	Name	Host	Infection
Ari	NC_010067	*S. enterica* subsp. *arizonae*	Broad	Gastroenteritis
CT18	NC_003198	*S. enterica* Typhi CT18	Broad	Systemic
Ty2	NC_004631	*S. enterica* Typhi Ty2	Human	Systemic
LT2	NC_003197	*S. enterica* Typhimurium LT2	Broad	Gastroenteritis
SL1344	FQ312003	*S. enterica* Typhimurium SL1344	Broad	Gastroenteritis
D23580	FN424405	*S. enterica* Typhimurium D23580	Broad	Gastroenteritis
Cho	NC_006905	*S. enterica* Choleraesuis SC-B67	Swine	Systemic
ParC	NC_012125	*S. enterica* Paratyphi C RKS4594	Human	Systemic
Ent	NC_011294	*S. enterica* Enteritidis P125109	Broad	Gastroenteritis
Gal	NC_011274	*S. enterica* Gallinarum 287/91	Avian	Systemic

annotated pseudogenes. `MultiFusion` is akin to `MultiParanoid` (Alexeyenko *et al.*, 2006) in applying a single-link clustering of pairwise analyses into fusion groups, a set of homologous proteins with both full-length and split forms; it may optionally make use of fusions detected using annotated pseudogenes. `FusionScanner` augments the resulting fusion groups using results from the `tblastn` search by `FusionFinder`. Finally, `FusionMapper` uses `Mix` from `PHYLIP` package version 3.68 (Felsenstein, 1989) to map these results onto a given species tree and outputs a tree in Newick format with the number of events mapped onto each branch. A utility script called `isFinder` scans input proteomes for any proteins annotated as related to insertion sequences, transposases, and transposons so any fusion events containing these may be discarded by `MultiFusion`.

Figure 2.2: Flowchart of the `FusionFinder` pipeline. The `FusionFinder` program is applied for all possible pairwise genome comparisons; all other programs are applied once on a given set of genomes. Dashed lines denote optional flow in pipeline.

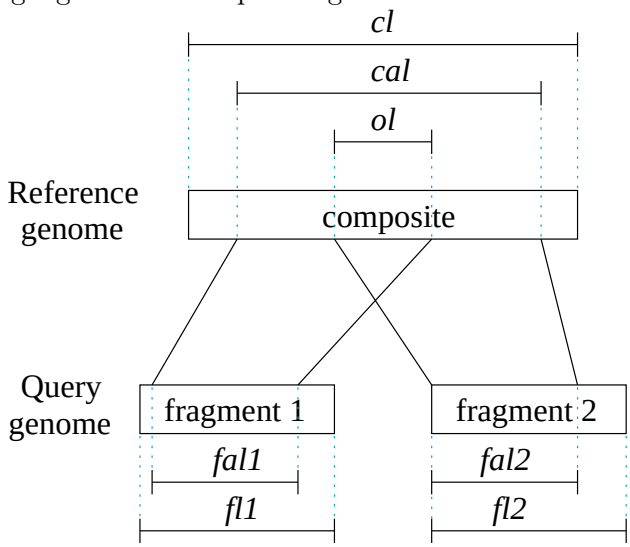


2.3.1 FusionFinder

Fusion and fission events are defined in the literature as a one-to-many relationship between a composite (contiguous, full-length) gene and a set of component (split, fragmented) genes

(Figure 2.3); we refer to all the components related to a single composite in a fusion event as a component set. In a pairwise genome comparison, the ancestral state could be the composite form, which was split into multiple fragments in one of the genomes or the ancestral state could be two (or more) components that were fused into a single gene in one of the genomes. If a gene is fragmented in both genomes with respect to a third then the fusion event might not be detectable (Figure 2.4). Fusion detection algorithms typically use `BLAST` and sometimes the Smith-Waterman algorithm (or both in concert) to detect multiple genes with significant similarity to a single gene a reference genome (Enright *et al.*, 1999; Snel, Bork and Huynen, 2000; Yanai, Derti and DeLisi, 2001; Kamburov *et al.*, 2007; Durrens, Nikolski and Sherman, 2008); the RBH principle may also be enforced (Suhre and Claverie, 2004; Nakamura, Itoh and Martin, 2007) and indeed there are similarities to the extension of RBH by `InParanoid` and `OrthoMCL` to detect inparalogs across multiple genomes (see Chapter 1), although we are not aware of any program that detects orthologous groups and fusion events simultaneously.

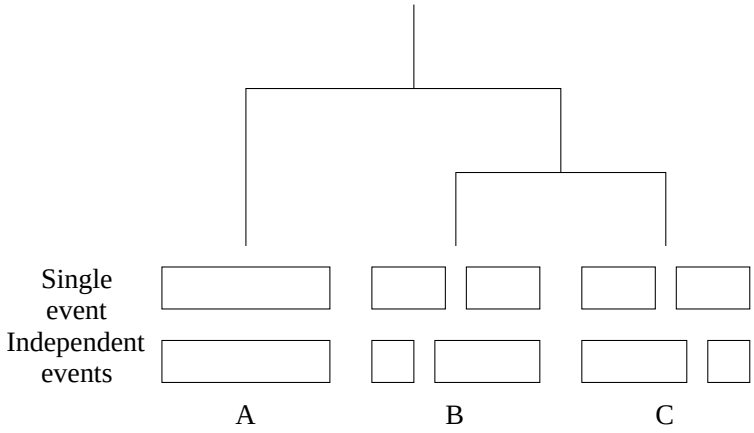
Figure 2.3: Parameters for alignment overlap assessment: alignment overlap length (*ol*), composite alignment length (*cal*), composite length (*cl*), fragment 1 alignment length (*fal1*), fragment 1 length (*fl1*), fragment 2 alignment length (*fal2*) fragment 2 length (*fl2*). 'Fragment' is used instead of 'component' for parameter naming convenience. Fragments 1 and 2 can represent HSP of a single gene or two separate genes.



`FusionFinder` is a collection of algorithms for the detection of fusion events between a pair of genomes using annotated proteins (protein-protein), genomic sequences (protein-genome), and annotated pseudogenes (pseudogene-protein); this program must be applied to all possible pairwise combinations for subsequent analysis by `MultiFusion`. We primarily rely on annotated proteins, like other approaches in the literature, in an attempt to reduce false positives. This general outline of our algorithm for protein-protein detection of fusion events is a variation of many similar approaches described in the literature:

1. Pairwise `blastp` searches between input genomes
2. Define one genome as the Query genome and the second as the Reference
3. Generate putative component sets where a Reference protein is the best hit of at least two proteins from the Query genome

Figure 2.4: Detection of a fusion event depends on genome choice. The first example illustrates a single fusion or fission event in the ancestor of B and C that can only be detected in a comparison to genome A. Independent fissions in genomes B and C may be detectable in a pairwise comparison in some instances as described in the text.

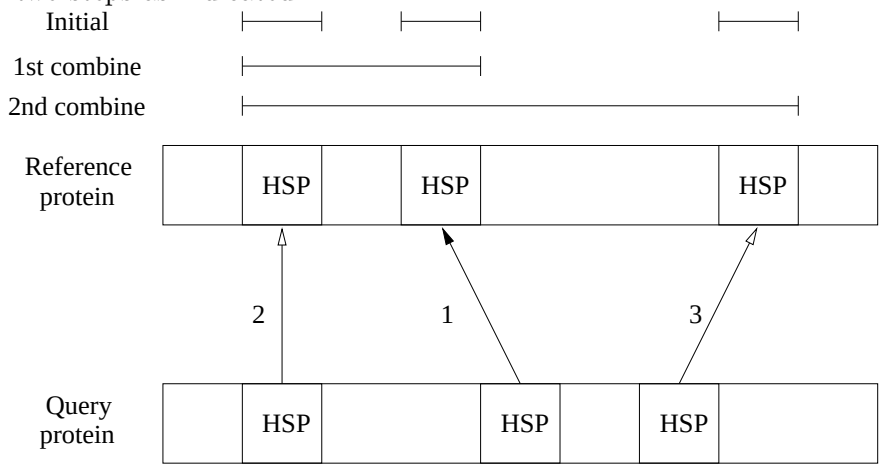


4. Retain components that are significant hits in the reverse `BLAST`
5. Components that hit only one protein in Reference genome are always retained
6. Combine multiple high-scoring segment pairs (HSP) to reduce false positives
7. Sort components into adjacent sets and singleton set based on genome location
8. Apply component selection algorithm first within adjacent sets then to singleton set:
   (a) Sort components by the relative position of their alignments to the composite
   (b) Discard components with significant overlap between alignments to the composite
   (c) Discarded components in adjacent sets are placed in singleton set
   (d) Singleton set is searched until no new component sets are found
9. Repeat the above steps using the second genome as the Query genome

The main component is the protein-protein comparison, which generates a ".fe" file (Figure 2.2). Duplicate genes and gene segments lead to identical or near-identical `BLAST` statistics and negatively impact our algorithm; we add a pre-processing step to collapses genes into a single entity if they have HSP alignments to a putative composite with identical percent identity, alignment length, and number of mismatches and gaps. After component selection, described below, component sets are concatenated and used as a `BLAST` query to select a single gene amongst the collapsed set of sequences. Subsequent inspection of results suggests that this method may not be ideal and we should consider performing self-genome `BLAST` searches and using cutoffs to collapse duplicate genes.

`BLAST` is a local alignment algorithm and may report significant hits between protein domains, resulting in false positives in fusion detection so we combine HSP with no significant overlap prior to component selection. For each significant HSP, we combine it with the best HSP (Figure 2.5) if there is no significant overlap between their alignments as

Figure 2.5: Hypothetical example of combining HSP prior to component selection. Order of significant HSP is indicated next to the arrows. There are initially three HSP, which are combined in two steps as indicated.

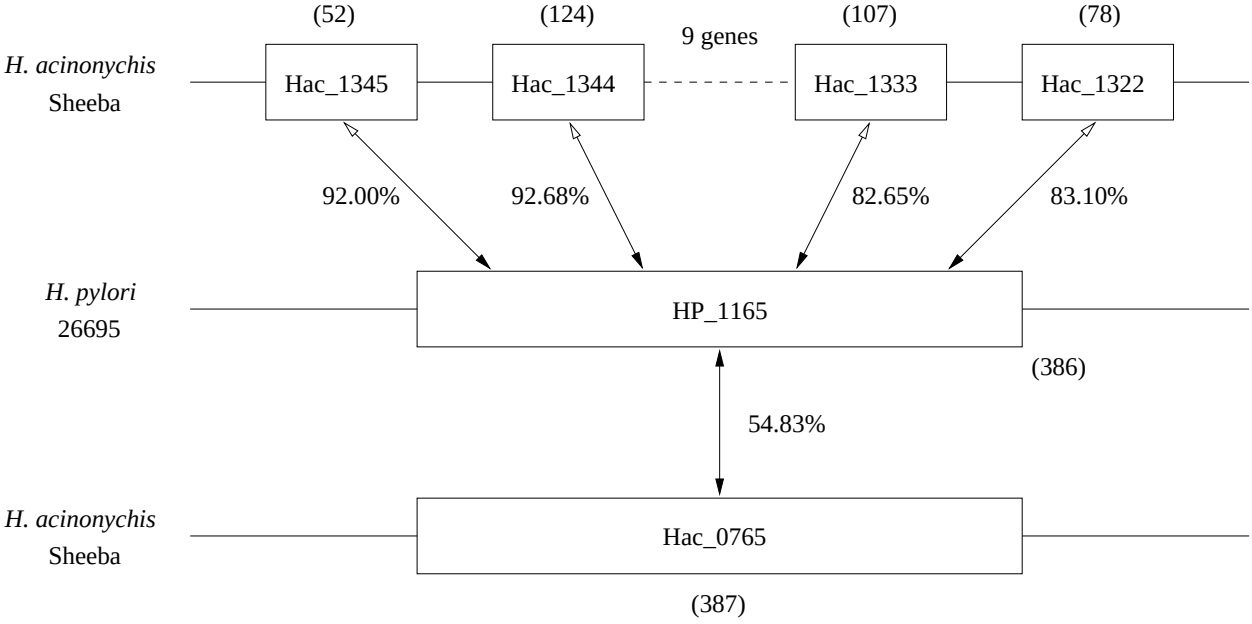


measured by the combined alignment length ($ca = \frac{ol}{cl} \leq 0.1$) and short alignment length ($sa = \frac{ol}{min\{fal1, fal2\}} \leq 0.5$) cutoffs (Figure 2.3).

A major difference from previously reported methods is our use of synteny, or relative locations of genes in the genome, to guide fusion detection. Manual inspection of our results showed improved accuracy when using synteny although use of this information is not required. Putative components are sorted into adjacent sets and a singleton set where all genes in an adjacent set are sequential entries in the NCBI protein table for the genome; we currently do not consider the strand or gene overlap but expect the effect to be minimal. The component selection algorithm is applied to each adjacent set and all putative components that are not selected are placed in the non-adjacent set. Each component set constructed from the adjacent sets are then collapsed into single entities and placed into the singleton set. Finally, the component selection algorithm is applied to the singleton set until no more new component sets are generated. To illustrate, we shall walk through an example of a fission detected in *Helicobacter acinonychis* using *H. pylori* as the reference genome (Figure 2.6). Since all five annotated genes in *H. acinonychis* have a single best hit in *H. pylori* and are reciprocally significant hits, they are initially placed in the same set of putative components and subsequently sorted into two adjacent sets and a singleton set containing Hac_0765. The component selection algorithm is applied to each adjacent set, which selects both sets due to the lack of any overlap, combines them into two independent entities, and places them into the singleton set, now containing three 'genes.' Finally, the component selection algorithm will find no overlap between the adjacent sets and will combine them but will filter out the paralog Hac_0765 due to significant alignment overlap, resulting in the identification of a fusion or fission event containing four components in *H. acinonychis*; automatic assignment of fission to this event is performed at the phylogenetic stage of the pipeline by `FusionMapper`, described below. Note that `BLAST` statistics and scores favor matches over longer stretches of sequence and in this example, the order of best hits in decreasing order for HP_1165 are: Hac_0765, Hac_1344, Hac_1333, Hac_1322, and Hac_1345, which does not match the order with respect to sequence similarity. HP_1165 and Hac_0765 are RBH and would be identified as orthologs by orthology detection algorithms (see Chapter 1), which

would likely be a false positive since the genes upstream and downstream of HP_1165 appear to be orthologous to the gene upstream of Hac_1345 and the gene downstream of Hac_ 1322 (RBH with 93.83% and 97.61% sequence identity, respectively) and the 9 genes separating the adjacent sets, all in the same orientation on the reverse strand, are of prophage origin according to the *H. acinonychis* annotation. In other words, there is conserved set of at least three syntenic genes, one of which is fragmented and interrupted by phage-insertion in *H. acinonychis*. However, we cannot rule out the possibility that Hac_0765 is a lineage-specific duplication followed by rapid divergence, in which it would be inparalogous and would be an ortholog (Chapter 1). Our algorithm is able to ignore the false signal from the paralog in *H. acinonychis* and identify the fission event resulting in four gene fragments.

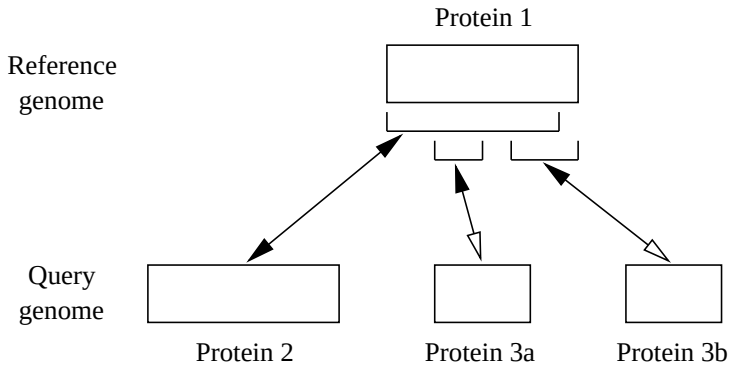
Figure 2.6: Example of fission in *H. acinonychis* detected using *H. pylori* and synteny data. Solid arrowheads represent best hits, hollow arrowheads represent significant hits, percent identities are next to the arrows, black lines connect adjacent genes, dashed lines connect non-adjacent genes, and protein lengths are denoted in brackets.



The critical step is the component selection algorithm which identifies component sets and in the process filters out false positives and duplicate genes. Given a set of adjacent or non-adjacent putative components, we sort them according to the relative positions of their alignments to the putative composite by end position then by start position. A greedy algorithm selects the components, always taking the first protein that does not significantly overlap with the previously chosen protein. In the hypothetical example in Figure 2.7, Protein 3a will be chosen first because the end position of its alignment to Protein 1 is the smallest. Protein 2 will then be considered and discarded due to significant overlap with Protein 1. Finally, Protein 3b will be considered and chosen due to lack of overlap with Protein 1. Not allowing overlapping alignments (Marcotte *et al.*, 1999; Snel, Bork and Huynen, 2000; Kamburov *et al.*, 2007) increases the number of false negatives (Durrens, Nikolski and Sherman, 2008). Absolute cutoffs (Yanai, Derti and DeLisi, 2001; Hua *et al.*, 2002) and various proportional cutoffs (Enright and Ouzounis, 2001; Suhre and Claverie,

2004; Nakamura, Itoh and Martin, 2007; Durrens, Nikolski and Sherman, 2008) have been reported. We used two measures of overlap identical to the ones used when combining HSP described above, the combined alignment length ($ca = \frac{ol}{cal} \leq 0.1$) and short alignment length ($sa = \frac{ol}{min\{fal1, fal2\}} \leq 0.5$) (Figure 2.3). We further required that the alignments cover a significant portion of the component proteins using the fragment coverage parameter ($fc = \frac{fal}{fl} \geq 0.7$); composite coverage is assessed for pseudogenes and genomic hits here and proteins in `MultiFusion` ($cc = \frac{cal}{cl} \geq 0.7$). Average sequence identity of RBH in pairwise comparisons between the main *Bacillaceae* groups hovers within the 55% to 65% range so we used the permissive cutoff of 50% although we show that increasing this value does not significantly change the conclusions reached in this chapter (Figures 2.20,2.22). The `blastp` Expect-value cutoff was kept constant at 10^{-5}.

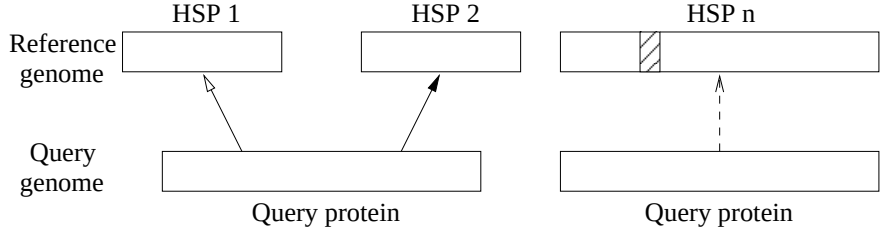
Figure 2.7: Component selection algorithm. A hypothetical situation where proteins 2 and 3 are paralogs of protein 1 and protein 3 that has been fissioned. Proteins 2, 3a, and 3b form a putative component set. Solid arrowheads represent best hits and hollow arrowheads represent significant hits.



Under the assumption that shorter genes are more likely to be missed during genome annotation, we implemented a genome scan procedure where we took annotated proteins from the Query genome to search for unannotated components in the Reference genome using `tblastn` (Figure 2.8), generating ".fes" files (Figure 2.2). We first look for multiple HSP without significant overlap. HSP are combined in the same manner as described above and illustrated in Figure 2.5 and overlap is assessed using the same combined and short alignment length cutoffs described above. Only if this search fails, we choose the first HSP, if any, where the alignment satisfies the component coverage cutoff, the genomic sequence contains a single in-frame stop codon, and the partitioned alignments are at least 10% of the composite length. Note that the reciprocal hit criteria is not enforced and `tblastn` searches were executed using a more restrictive Expect-value cutoff of 10^{-20}. Although this may increase false positives, we attempt to minimize this by only using identified components to update fusion groups identified by `MultiFusion`, using the program `FusionScanner`.

It is important to emphasize that the algorithms described above only detect a fusion event if one genome contains an annotated composite protein and at least one genome contains at least two annotated component proteins; `FusionScanner` will then take components identified using the above protein-genome scanning procedure to update existing fusion groups, reporting the results in ".fep" files (Figure 2.2). This should help reduce false positives but is predictably affected by whether putative pseudogenes are annotated as pseudogenes or proteins. For example, despite the fact that *B. anthracis* is considered

Figure 2.8: Genome scan procedure (protein-genome) using annotated proteins. Components are reported if multiple HSP are identified (left) or a significant HSP can be partitioned by a single in-frame stop codon represented by a shaded box (right). The criteria for evaluating these possibilities are discussed in the text. The solid arrowhead denotes the best hit, the hollow arrowhead denotes a significant hit, and the dashed arrow denotes either is possible.



a clonal population, the number of annotated pseudogenes vary even among the complete genomes (Table 2.5); no pseudogenes were annotated in any of the draft genomes. Thus, we sought to include annotated pseudogenes into our fusion groups provided they satisfied our requirements. A pre-processing script called `PseudogeneParser` simply extracts annotated pseudogenes from genome annotation files (.gff files) for each record that a 'pseudo' tag in the 'attribute' field. `FusionFinder` takes annotated pseudogenes and performs a `blastx` from the Query genome to annotated proteins in the Reference genome and a reverse `tblastn` from the proteins in the Reference back to the Query pseudogenes (pseudogene-protein). A fragmented gene may be represented by a single or multiple pseudogene entries so we use the algorithms in both the protein-protein and protein-genome algorithms to assign pseudogenes to a composite; HSP are combined and the component selection algorithm is is used without considering synteny while the genome scan procedure is used as is.

Similarity searches were performed using `BLAST 2.2.24+` with the `BLOSUM62` matrix, the default compositional adjustment as described in (Yu and Altschul, 2005), and the soft filtering and Smith-Waterman final alignment options. In total, 2 `blastp`, 4 `tblastn`, and 2 `blastx` searches are performed for each genome pair. Note that protein, pseudogene, and genome hits are considered separately and in that order of precedence, which may lead to false positives if paralogs are present. An alternative that could be implemented in the future would be to base the analysis on the genome scan. Since any protein or pseudogene hit must also have a hit in the genome, we can perform the analysis using only the genome `BLAST` and take into consideration whether there was a corresponding significant protein or pseudogene hit.

For draft genomes, genes at contig boundaries sometimes have incomplete coding sequences such that the length of the gene is not divisible by 3. Since we cannot confirm whether the gene is truly fissioned in the genome, we discard these genes from our analysis.

2.3.2 MultiFusion

As noted previously, fusion and fission analysis restricted to pairwise genome analysis is more susceptible to error, similar to orthology analysis (see chapter 1). Similar to `MultiParanoid`, we implement a single-link clustering of all possible pairwise genome comparisons to form fusion groups and a simple algorithm that can detect some but not all cases of differential fission. The steps are:

1. Pre-processing of fusion events (described below)
2. Build fusion groups by single-link clustering on composites and individual components
3. Split each group into subgroups by single-link clustering on component sets
4. Sort subgroups into 'main' and 'nested' groups
5. Merge main groups if criterion met
6. Merge nested groups if criterion met
7. Detect differential fragmentation in main groups
8. Assign COGs to fusion groups

In the pre-processing step, fusion events are removed from consideration if they involve a mobile element, detected by keyword search of protein annotation for 'CRISPR', 'transposon', 'transposase', 'insertion sequence', or if the component set does not cover at least 70% of the composite. Subgroups are considered 'nested' if at least one composite is a component in another subgroup, which is considered a 'main' group. This procedure may result in multiple main and/or nested groups. Main groups are merged if one group contains a component set that is a strict subset of the component set in another group, and nested groups are merged if one group contains a component set that overlaps the component set in another group. Both these cases arise due to the presence of 'nested' events, resulting in fusion events with two components in some genomes and three components in others, for example. Although it is possible to have multiple main groups if the component sets intersect, there are not enough instances to affect our results (8 cases for the 40–genome dataset). Once the groups have been stabilized, we test each main group for components that are also composites, which indicates differential fission. Finally, we assign COGs to each fusion group based on the first composite of each fusion group and declare as a fission event all main groups that contain differential fission; to be more conservative, we do not force nested groups to represent fission events. By default `MultiFusion` uses only fusion events identified using the protein-protein search although annotated pseudogenes may also be added in a concurrent analysis (".fe" and ".fep" files in Figure 2.2, respectively). We implemented as an option the declaration of fission for all pseudogenes identified although we did not use this in the results. A fusion pattern is constructed for each fusion group where each genome is assigned either a state of '0' or '1' if the genome contains a composite or component protein, respectively. If neither is identified, the genome either has no homologous protein or the divergence was too high for detection and is assigned '?'.

2.3.3 FusionScanner

Given a file with fusion groups from `MultiFusion`, `FusionScanner` updates fusion groups given the genomic scan results from the ".fes" files generated using `FusionFinder` (Figure 2.2). First, we identify all composite proteins involved in a fusion event; composites that are components in another fusion group are not considered, as are all composites in nested groups. Then we search the ".fes" files additional components. By default, for each fusion group, all genomes with a state of '0' or '?' may be updated to '1' if we find new fissions,

but the more conservative approach of considering only '?' may be chosen. An updated fusion group file is generated (Figure 2.2), containing the original data augmented with the newly identified fusion events and new fusion patterns for each fusion group. To reduce false positives, each fragment identified in a fusion event by a protein-genome scan must satisfy the fragment size cutoff ($fs = \frac{fl}{cl} \geq 0.1$) for a genome state to be updated. The main goal of `FusionScanner` is the identification of components not annotated in a genome, although it may possibly pick up any missed FEs in the protein-protein comparison algorithm.

2.3.4 FusionMapper

As noted above, each fusion group is described by a pattern of '0', '1', and '?'. These patterns are fed into `Mix` to infer the ancestral state of the group and to place the event or events on the phylogeny using Camin-Sokal parsimony; the ancestral state is forced to be composite if differential fission was detected, as mentioned above. We chose Camin-Sokal parsimony for mapping fusion and fission events onto the phylogeny, as we assume that a reversion from one state to another is much more unlikely than the independent generation of a similar event, in contrast to nucleotide evolution. The parsimony results are summarized, with the number of fusion events inferred tabulated and mapped to branches on the given species tree in Newick format. This program uses the Newick tree parser in the BioPerl module (Stajich *et al.*, 2002).

`FusionFinder` detects fusion events using homologs and `MultiFusion` assembles these into protein families in which at least one fusion event has occurred. However, we use `FusionMapper` to predict the time of a fusion event much like tree reconciliation algorithms do for duplications and losses (Chapter 1). When we apply `FusionMapper` we implicitly assume our fusion groups include only orthologs, although the results might still be correct if there are outparalogous proteins, depending on the actual evolutionary history (Figure 2.9). In contrast to tree reconciliation, where duplications are biased towards the cenancestor, with Camin-Sokal parsimony we place fission events as close to extant taxa as possible, which we expect to be a good approximation, especially for fission which will more likely lead to gene inactivation and rapid removal from the genome (Kuo and Ochman, 2010). We believe that the combination of algorithm parameters chosen and evolutionary distances between strains in this study makes this a reasonable assumption. Although we have not performed an in-depth analysis, a small number of fusion groups were used to query ortholog predictions by `MultiParanoid` and `OrthoMCL` (Li, Stoeckert and Roos, 2003). In all cases the proteins involved in our fusion groups were identified as orthologs by at least one of the two programs. As expected, the shorter component in a component set was always absent from the orthologous group and indeed was not placed in any orthologous cluster. Unsurprisingly, using parsimony in an automated manner can lead to incorrect assessments (Figure 2.10, left) or situations where we cannot determine the type of event (Figure 2.10, right) that could be obvious with manual inspection of results and sequences.

2.3.5 Species phylogeny

The *Bacillaceae* genomes analyzed in this chapter can be subdivided into four major groups: the Alkaliphilic, Cereus, Geobacillus, and Subtilis groups. The relationships between these four main groups is uncertain due to large genetic distances separating them; we chose to

Figure 2.9: Influence of orthology on fission mapping. Presented are two possible evolutionary histories for the same fusion pattern in an outparalogous protein family. Our methodology would infer the history on the left although the history on the right is one of several possible alternate histories. □: duplication, ◇: fission, ×: gene loss.

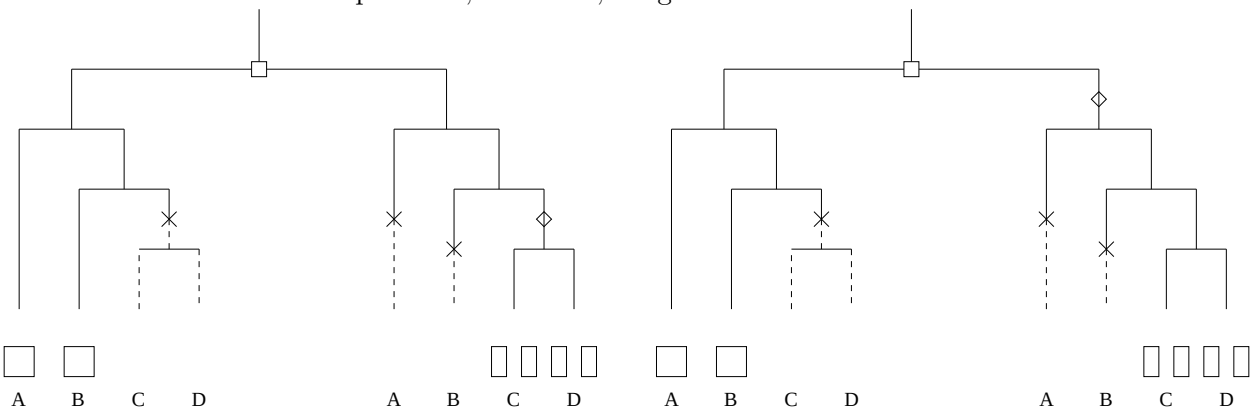
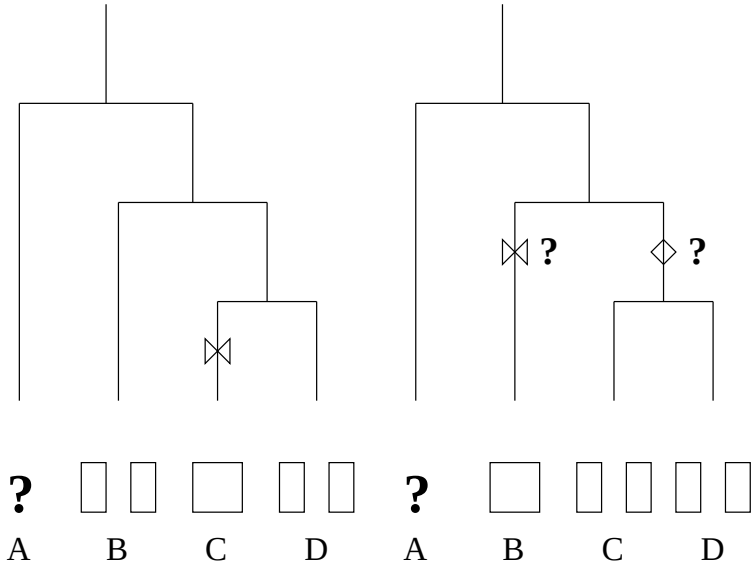


Figure 2.10: Implications of parsimony in fission mapping. Left: even if manual inspection suggests that independent fission likely occurred, such as from protein domain disruption or different positions of the split, parsimony would still infer fusion in this example. Right: since fusion and fission result in the same number of events, we cannot identify which event occurred on which branch. This is particularly problematic if there is only one outgroup lineage in the phylogeny, as we have in our *Bacillaceae* dataset. ⋈: fusion, ◇: fission.



match the tree used by Hao and Golding (2009). The relationships within the Cereus group were taken from the EnsemblGenomes species tree; note that their tree groups *B. pumilis* with the Cereus group but from our own trees and data published in other reports, *B. pumilis* should group with the Subtilis group. The relationships within Alkaliphilic, Geobacillus, Subtilis groups were taken from phylogenies generated using `MrBayes` (Huelsenbeck and Ronquist, 2001) on the genes *gmk*, *glpF*, and *pycA*.

The *B. anthracis* phylogeny was generated by expanding the previous canSNP analysis (Van Ert *et al.*, 2007) to 11 complete and 12X draft genomes. In-silico probes were generated from the reference *B. anthracis* Ames genome by taking 100 bp upstream and downstream of each SNP (Table 2.4). The corresponding SNPs in all 11 genomes, including the reference, were identified using `BLAST` with an Expect-value cutoff of 10^{-20} (Table B.1). For each genome, we confirmed the following for each probe:

- Exactly 1 genomic hit
- Exactly 1 HSP
- At most 1 mismatch
- Alignment length exactly 201 bp

The only case where a mismatch was not at position 101 of the alignment was for B.Br.001 in *B. anthracis* A1055 although this is not a problem because this SNP is indeed identical to the SNP in the reference genome (Van Ert *et al.*, 2007). We used the default settings of `Dnapars` from the `PHYLIP` package to infer the phylogeny and subsequently rooted the tree using *B. anthracis* A1055 as the outgroup with `Retree` from the `PHYLIP` package.

Table 2.4: CanSNP chromosomal location and information with respect to *B. anthracis* Ames (NC_003997.3). Data was obtained from the supplementary information of Van Ert *et al.* (2007); the position of A.Br.002 was incorrectly listed as 947760 but is corrected here. Type indicates whether the SNP is (N)on-synonymous, (S)ynonymous, or (I)ntergenic.

SNP	Position	Type	Strand
A.Br.001	182106	N	+
A.Br.002	947759	S	-
A.Br.003	1493157	I	-
A.Br.004	3600659	N	+
A.Br.006	162509	N	+
A.Br.007	266439	N	-
A.Br.008	3947248	S	-
A.Br.009	2589823	S	+
B.Br.001	1455279	I	-
B.Br.002	1056740	I	-
B.Br.003	1494269	S	+
B.Br.004	69952	N	+
A/B.Br.001	3697886	N	+

Phylogenetic inference on bootstrapped fusion and fission patterns was performed using the PHYLIP package. Patterns were treated as discrete characters and bootstrapped 1000 times using the default settings in Seqboot. Mix was used to infer the most parsimonious phylogeny for each bootstrapped dataset under the Camin-Sokal parsimony criterion; the jumble parameter, which indicates the number of different taxa to take as the starting point in tree search, was set to equal the number of taxa. Finally, a majority rule (extended) consensus tree was constructed using Consense.

2.3.6 COG statistical analysis

We tested whether COG distribution across fissioned genes in the *B. anthracis* lineage is significantly different from random expectation. A pan-genome of *B. anthracis* strains was constructed by counting the total number of genes in each COG category across the five complete genomes. Genes that are not assigned to a COG are placed in the artificial '-' category. P-values for the number of genes fissioned in each COG was then calculated using Fisher's exact test (two-sided) of independence as implemented in R (R Development Core Team, 2009). We also implemented a resampling analysis in R, performing one million trials and assuming that the number of fissions is equal between trials. One-sided P-values are calculated, dependent on whether the observed number of fissions is larger or smaller than the median of the simulated distribution. Multiple test corrections were performed for both Fisher's exact test and resampling using the 'multtest' package in Bioconductor (Gentleman *et al.*, 2004).

2.3.7 Estimation of selective pressure on *B. anthracis* genes

OrthoMCL was applied to all *B. anthracis* and *B. cereus* genomes excluding *B. cytotoxicus*, resulting in 8009 orthologous groups. As in Chapter 1, we followed the recommendations listed in the OrthoMCL Algorithm Document (http://docs.google.com/Doc?id=dd996jxg_1gsqsp6) with the exception of using the Smith-Waterman option. The only difference from the blastp searches in FusionFinder was the increase in the maximum number of alignments and descriptions to 100000 as recommended since FusionFinder performs pairwise genome comparisons and OrthoMCL performs an all-vs-all comparison. Orthologous groups were discarded if they contain paralogs (657), represent less than three genomes (984), do not represent at least one *B. cereus* and one *B. anthracis* genome (1167), contains both composites and components (410), contains composites (172), or contain components (48), reducing the number from 8009 to 4571.

For orthologous group members and composites of fission groups, protein and coding sequences were directly extracted from the genome annotation files (.ptt) but determination of the component set sequences is more problematic due to alignment overlap and particularly those identified by pseudogene or genomic hits. For each pairwise genome comparison, FusionFinder also stores concatenated component sequences extracted from the BLAST alignments of each component set with the corresponding composite gene. Overlap between components is resolved by trimming the BLAST alignments in the order that they are added to the component set; in other words, the alignment of the first component chosen is never trimmed. As noted above, components annotated as proteins are sorted and added in ascending order with respect to their alignments to the composite gene (N- to C-terminal) while

pseudogene and genomic hits are sorted and added based on BLAST Expect-values. Component sets may be identified via multiple composites from different genomes so MultiFusion chooses the concatenated component sequences derived from the BLAST alignments to the composite with the highest sequence identity averaged over the entire component set; ties are broken by choosing the first alignment encountered. Finally, FusionScanner adds sequences to fusion groups it updates. All stop codons were converted to 'X.'

Fission and orthologous group codons were aligned according to the protein alignment using MUSCLE 3.7 (Edgar, 2004) and tranalign from EMBOSS 6.1.0 (Rice, Longden and Bleasby, 2000). The rates of nonsynonymous substitutions per nonsynonymous site (d_N) and synonymous substitutions per synonymous site (d_S) as well as the d_N/d_S ratio were computed using CODEML from PAML 4.4c the branch model (Yang, 1998, 2007). Three sets of rates were estimated for each group: one across the entire phylogeny (ω_0) for the one-ratio model and separate ones for *B. cereus* (ω_1) and *B. anthracis* (ω_2) genomes for the two-ratios model. If the two-ratios model is not significantly better than the one-ratio model for a given sequence according to the likelihood ratio test, we assume the rates are equal ($\omega_1 = \omega_2 = \omega_0$). If the two-ratios model is significantly better then we keep the ω_1 and ω_2 estimates and discard ω_0. Fission and orthologous groups were discarded if $0.001 < d_S < 1.5$ for any of the three estimates, reducing the number of fission and orthologous groups from 68 and 4571 to 35 and 1877. Multiple test corrections were performed with R (R Development Core Team, 2009).

2.4 RESULTS

2.4.1 Specific examples of fusion and fission

Several fusion and fission events predicted in *Bacillaceae* genomes by our FusionFinder pipeline were investigated in more detail to ascertain accuracy of our method. A fusion between the A and B subunits of the Holliday junction helicase complex (RuvAB) was identified in *B. cereus* AH820 (GI: 218905615). Separate but adjacent *ruvA* and *ruvB* genes were identified in all *Bacillaceae* genomes except for *O. iheyensis*, where the adjacent *ruvA* gene has 49% identity to the *B. cereus* AH820 ortholog, falling just under our default cutoff; a BLAST search against the RefSeq and nr databases did not recover any other fusions of these two genes.

A fusion of *rhaA* (L-rhamnose isomerase, EC: 5.3.1.14) and *rhaB* (L-rhamulose kinase, EC: 2.7.1.5) was identified in *B. clausii* (Figure 2.11) based on the existence of separate but adjacent genes in both *O. iheyensis* and *B. halodurans*. Our pipeline also identified homologs in *B. subtilis* 168 and *B. subtilis* W23 that were not included in the analysis because the *rhaA* homolog falls below our 50% identity cutoff. Although a phylogenetic analysis should be performed (Koski and Golding, 2001), we note that the the *O. iheyensis* and *B. halodurans* homologs are the most closely related based on BLAST results and no other fusion of these genes were found in the RefSeq and nr databases. According to the MetaCyc (Caspi *et al.*, 2010) and KEGG (Kanehisa *et al.*, 2010) databases, *rhaA* and *rhaB* encode enzymes in sequential steps (Figure 2.12) in the rhamnose degradation and fructose and mannose metabolism pathways, respectively.

Although not discussed in the rest of this chapter, we included *B. anthracis* Tsiankovskii-I in our analysis to identify genome-specific fusion and fission events to attempt to identify pu-

Figure 2.11: Results from the NCBI CDD for the fusion protein (GI: 56962154) of *rhaA* and *rhaB* in *B. clausii*.

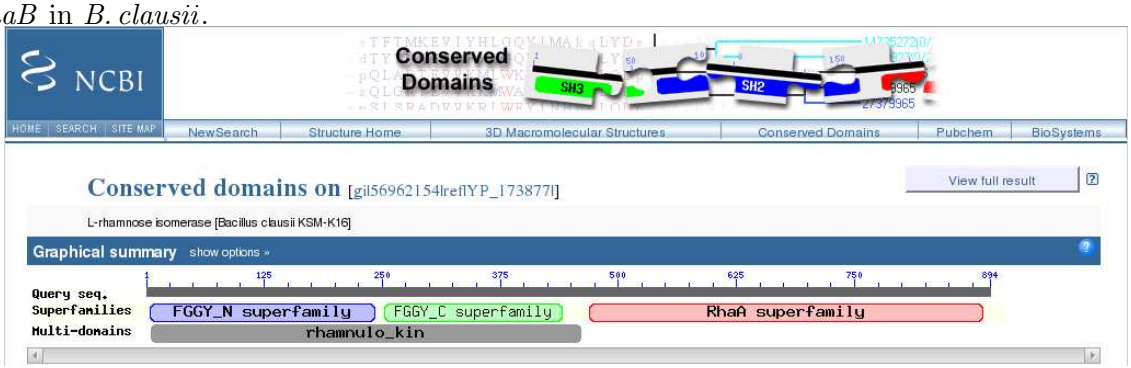
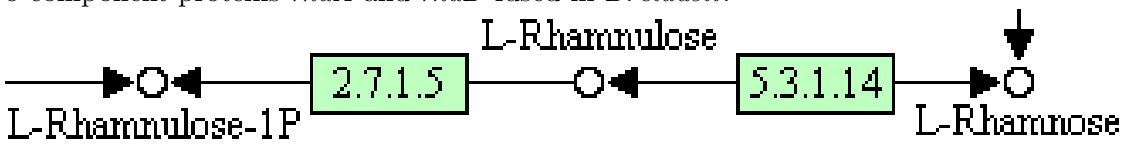


Figure 2.12: Part of the rhamnose degradation pathway on the KEGG database showing the two component proteins *rhaA* and *rhaB* fused in *B. clausii*.



tative mutations that may contribute to the unique attenuation of this strain. SNPs and fission patterns indicate that *B. anthracis* Tsiankovskii-I is most closely related to *B. anthracis* Western North America (data not shown). One fusion and three fissions were mapped onto the *B. anthracis* Tsiankovskii-I branch but only the fusion was unique to this strain. A putative two-component system consisting of a histidine kinase (*B. cereus* ATCC 14579 GI: 42781624) and a DNA-binding response regulator (*B. cereus* ATCC 14579 GI: 42781625) are encoded on adjacent genes in all *B. cereus* and *B. anthracis* strains except for *B. anthracis* Tsiankovskii-I (Figure 2.13). A `blastn` search indicates that a 297 nucleotide deletion resulted in this fusion. Interestingly, the response regulator and histidine kinase are located in the N- and C-terminal ends of the fusion protein, the response regulator contains a DNA-binding domain, and the sensor histidine kinase contains a transmembrane domain, all of which are conditions suggested to be selected against (Cock and Whitworth, 2007). Furthermore, `TMHMM` 2.0 (Krogh *et al.*, 2001) identified two transmembrane domains in the wild-type histidine kinase (Figure 2.14) but only the second one in the fusion protein (Figure 2.15); the deletion includes the C-terminal of the response regulator and the N-terminal of the histidine kinase, including the first transmembrane domain. Thus, we propose that the fusion event in *B. anthracis* Tsiankovskii-I, assuming it is not due to a sequencing error and there is no post-processing of the mRNA transcript, resulted in a defective two-component signaling pathway that might play a role in the attenuation of the strain. We also note that *B. anthracis* Tsiankovskii-I contained more genome-specific fissions than any other *B. anthracis* genome we analyzed, although the difference was not large in a couple of comparisons.

A component set specific to a single genome may be due to sequencing error or gene inactivation but is more likely functional if it is common throughout a monophyletic clade. We identified such a fission of *secDF* (Figures 2.16,2.17) in all three Alkaliphilic group genomes included in our study; a `RPS-BLAST` search against the `CDD` (Marchler-Bauer *et al.*, 2011) show the presence of full length, uninterrupted SecD (Figure 2.18) and SecF (Figure 2.19) multi-

Figure 2.13: Results from the NCBI CDD for the fusion protein (GI: 190565563) joining members of a two-component system in *B. anthracis* Tsiankovskii-I.

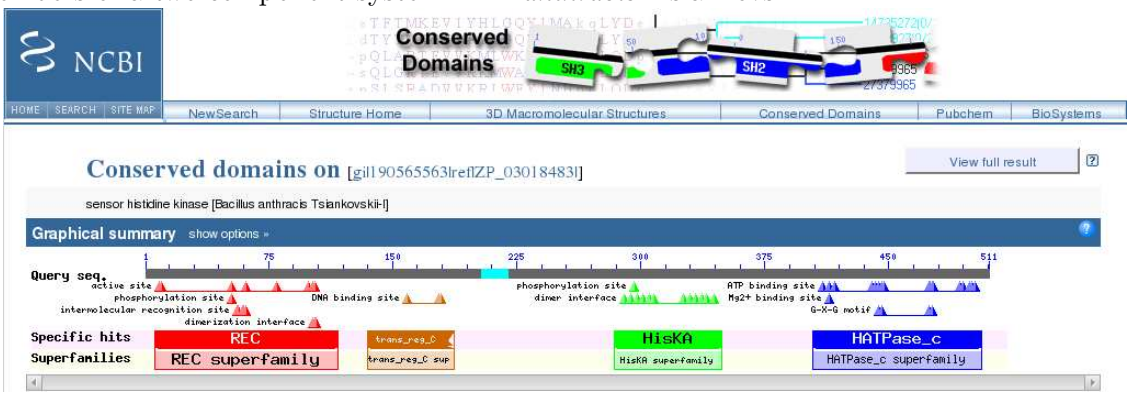


Figure 2.14: TMHMM results showing high posterior probabilities for two transmembrane domains in the wild-type *B. anthracis* Ames Ancestor.

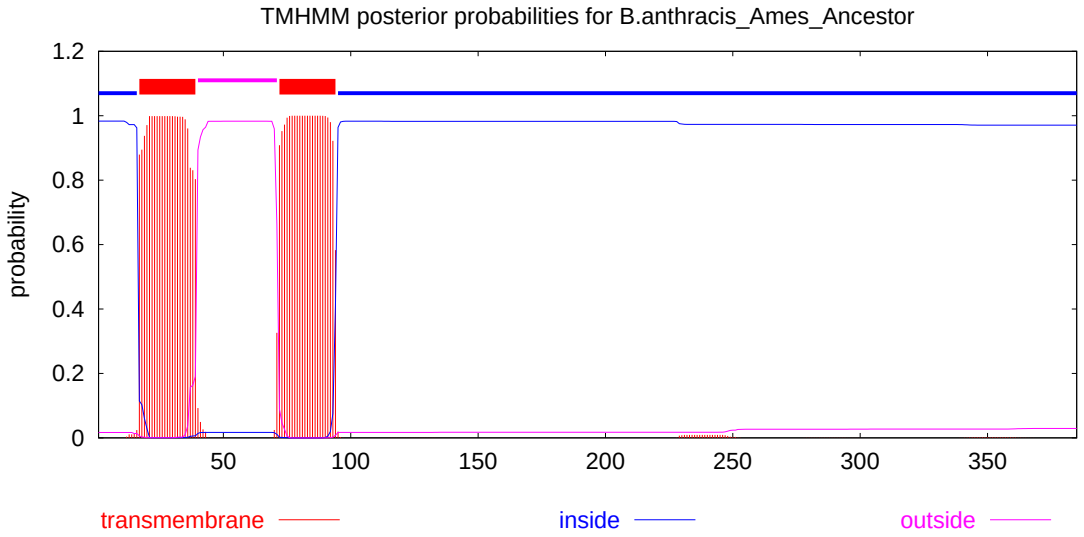
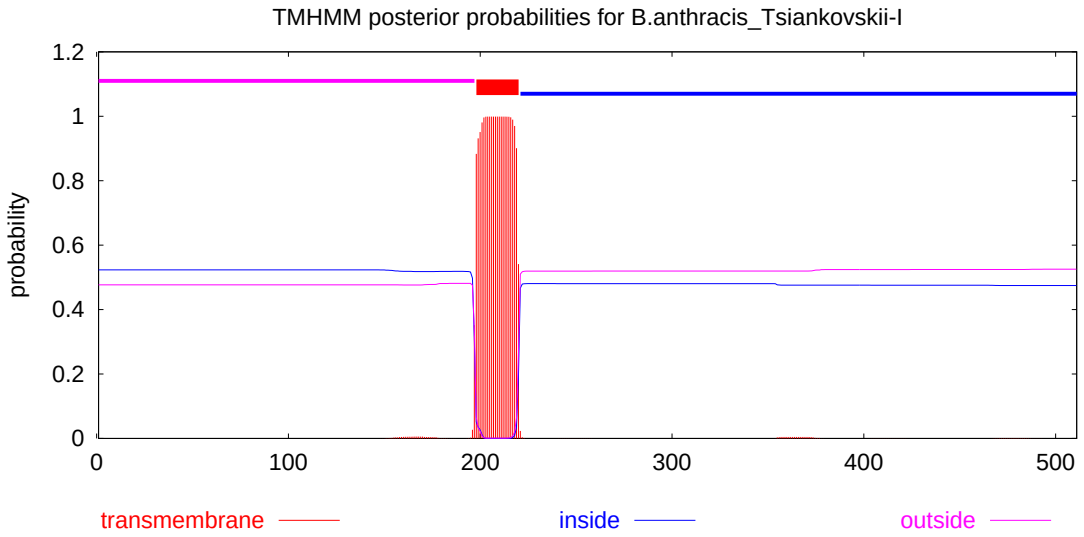


Figure 2.15: `TMHMM` results showing high posterior probability for only one transmembrane domain in the *B. anthracis* Tsiankovskii-I ortholog of *B. anthracis* Ames Ancestor.



domains in the Alkaliphilic group orthologs. Unlike the fusion in *B. anthracis* Tsiankovskii-I, this fission event preserved the 6 transmembrane domains per protein domain based on high posterior probabilities observed with `TMHMM` (data not shown). The composite form was identified in all 30 non-Alkaliphilic genomes by our pipeline and a `BLAST` search against `RefSeq` showed that the *Bacillaceae* orthologs have the highest sequence similarity and that close relatives such as *Listeria* also contain full-length orthologs. SecDF, a transmembrane protein required for maintenance of proton motive force for protein secretion, is also encoded by separate genes in *Escherichia coli* but the composite homolog in *B. subtilis* has been shown to encode a single polypeptide (Bolhuis *et al.*, 1998). Although the authors suggest that the *B. subtilis* homolog represents a "natural gene fusion," our preliminary data suggest that the *B. subtilis* homolog represents the ancestral state in the *Bacillaceae* at least. We also found a fusion common to a specific clade in our *B. cereus* phylogeny (Ba4–Ba6), with both components identified in all *B. cereus* group genomes except for Ba1, although we could not identify any relationship between the fused proteins. The components in *B. anthracis* Ames Ancestor are located on the reverse strand and are annotated as "acetyltransferase" (GI: 47528257) and "isochorismatase family protein" (GI: 47528256). Interestingly, the composites are also on the reverse strand and are annotated as "acetyltransferase" (GI: 42782058), "isochorismatase family protein" (GI: 217960395), and "acetyltransferase, gnat family" (GI: 222096450), suggesting that this gene fusion event was not detected during the annotation of any of these three genomes.

Phylogenetic placement of *Anoxybacillus flavithermus* was previously performed using concatenated RNA polymerase subunits RpoA, RpoB, and RpoC (Saw *et al.*, 2008). When collecting the sequences for a phylogeny including all genomes in our analysis, we found a fission in *B. cereus* AH820 and an annotated pseudogene in *B. cytotoxicus* for the RpoB ortholog; paralogs were not detected in either genome. These fissions are unlikely to be authentic as RpoB has been shown to be essential and conserved (Omer *et al.*, 2010). The fragments in *B. cereus* AH820 were 100% identical to most *B. cereus* orthologs and the fission can be explained by a single insertion of an additional adenine at position 108519 of the

Figure 2.16: Results from the NCBI CDD for composite *secDF* in *B. subtilis* 168 (GI: 16081162).

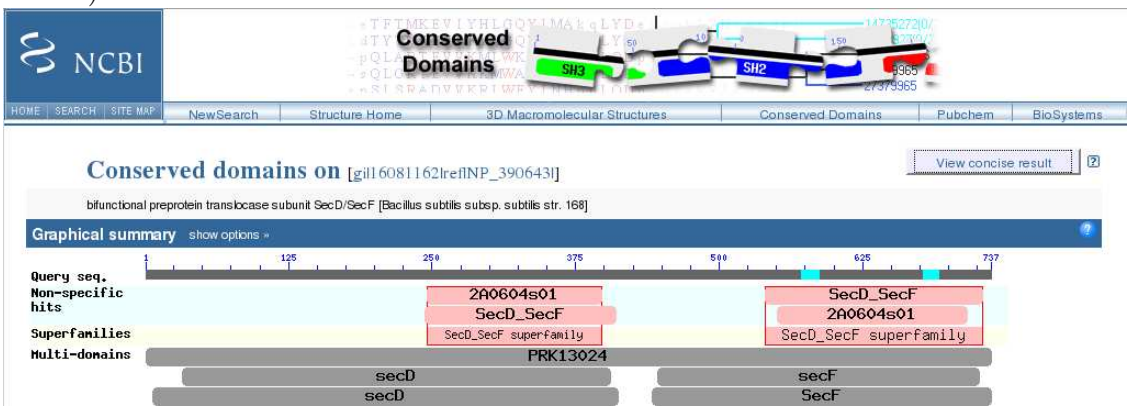


Figure 2.17: TMHMM results showing high posterior probabilities for 12 transmembrane domains, 6 for each protein domain *secD* and *secF*.

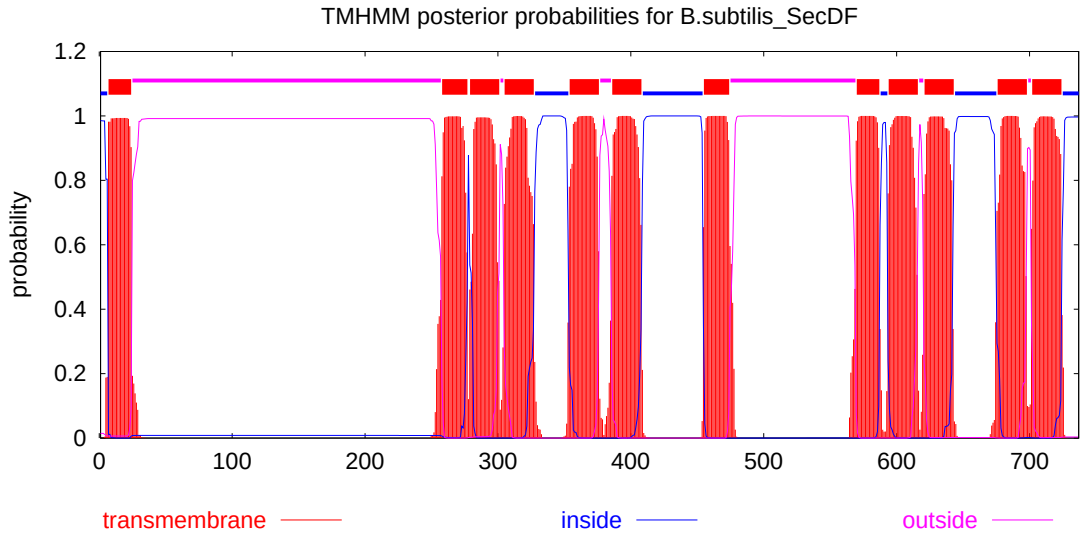


Figure 2.18: Results from the NCBI CDD for component *secD* in *B. clausii* (GI: 56963330).

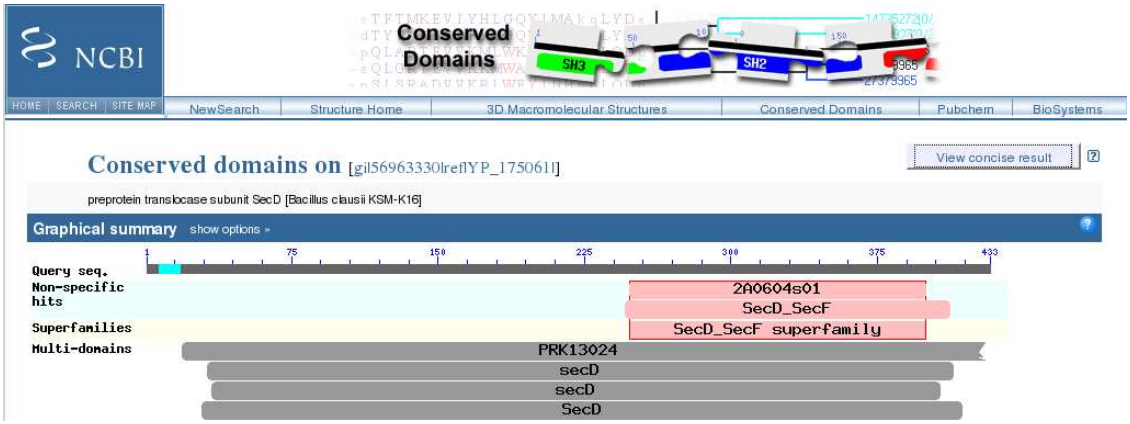
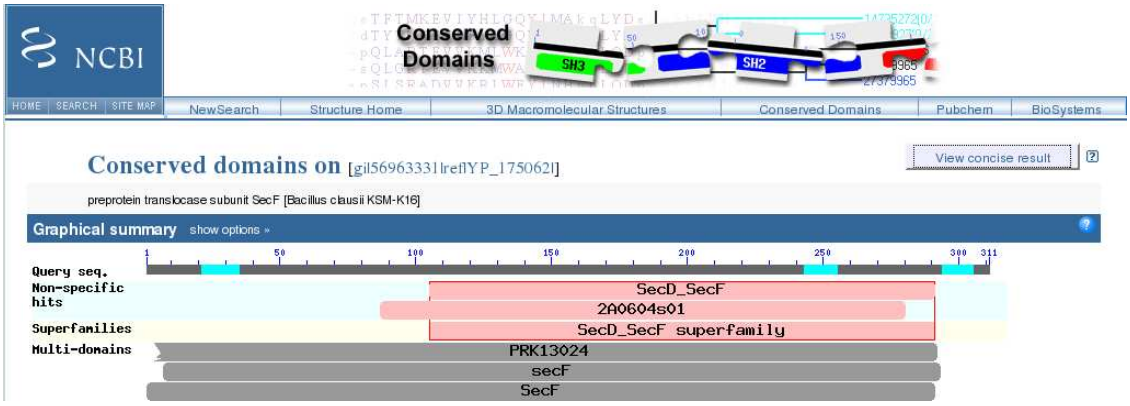


Figure 2.19: Results from the NCBI CDD for component *secF* in *B. clausii* (GI: 56963331).



genome, resulting in a pair of adenines, and is likely due to a sequencing error. Although the *B. cytotoxicus* ortholog is comparatively more divergent from the rest of the *B. cereus* strains due to its phylogenetic position, it appears the cause of the fission can also be explained by loss of an adenine from the orthologous lysine codon AAA and is also a possible sequencing error. Our `FusionFinder` pipeline was able to identify both these fissions in the same fission group when pseudogenes are included in the analysis.

We detected fission of a conserved hypothetical protein when pseudogenes were included in our analysis. Full-length composites were detected in all *B. cereus* genomes (*B. cereus* ATCC 14579 GI: 42779089) and the fission was common to all *B. anthracis* lineage A and B strains. The *B. anthracis* Ames Ancestor genome annotation describes pseudogene GBAA_0007 as the result of an "authentic frameshift" not due to a sequencing artifact but this pseudogene is expressed in *B. anthracis* Sterne according to transcriptome sequencing (Passalacqua *et al.*, 2009); GBAA_0007 expression is also reportedly growth-phase-dependent according to microarray experiments (Liu *et al.*, 2004). This demonstrates that gene expression may not be sufficient to invalidate a predicted gene fission.

The *B. cereus* group is currently the only bacterial group where chromosomal protein-encoding genes are known to be interrupted by group I introns (Ko, Choi and Park, 2002; Nord, Torrents and Sjberg, 2007; Nord and Sjberg, 2008). A group II intron was also found to interrupt *recA* in *Geobacillus kaustophilus* (Chee and Takami, 2005) even though they are not typically found in housekeeping genes or conserved genes (Dai and Zimmerly, 2002). Recently, a survey of the *B. cereus* group identified 73 group I and 77 group II introns (Tourasse and Kolstø, 2008), although these numbers appear to be aggregates across all genomes and not the number of intron gain events. Our method should detect intron gain events as gene fission events if the components are annotated as separate protein-coding genes, even though they are not what would typically be considered gene fission. Tourasse and Kolstø (2008) identified five genes interrupted by group I introns; *recA* Ko, Choi and Park (2002), *nrdE* (Nord, Torrents and Sjberg, 2007), and *nrdF* (Nord, Torrents and Sjberg, 2007; Nord and Sjberg, 2008) were previously reported, experimentally verified, and also identified by our method. Also as reported, we found that *nrdE* has apparently gained an intron multiple times in the *B. cereus* group, validating our choice of Camin-Sokal parsimony. The fission of *recA* in *G. kaustophilus* by a group II intron was also identified in the same fusion group by `FusionScanner`; only the N-terminal fragment is annotated (GI: 56419830). The fission of the "tail tube" protein identified by Tourasse and Kolstø (2008) either was

not identified by our method or we could not find it due to the lack of an accession number or a more descriptive annotation. They also identified group I introns in tape measure proteins (TMP) in several genomes including GBAA_0477 (GI: 47525747), part of prophage lambda04 in *B. anthracis* Ames Ancestor. Our pipeline did not detect these events but did identify fission of the TMP of lambda02 (*B. anthracis* A1055 GI: 254724441) when including draft *B. anthracis* genomes (Figure 2.30) because no orthologs were identified outside of the *B. anthracis* lineage. Although we are not sure whether the fission is due to a mutational event or an intron, the split form is present in all *B. anthracis* A lineage genomes and a full-length ortholog is found in all *B. anthracis* lineage B and C genomes (Figure 2.26).

As described in the methods, an initial fusion/fission groups generated by `MultiFusion` can be subdivided into "main" and "nested" groups. This allows us to determine more complex situations, such as gene fusion in one lineage and fission in another involving the same proteins. For example, a gene annotated as "multidrug resistance ABC transporter ATP-binding protein/permease" in *B. cereus* CI (GI: 301053513) is a fusion of proteins "ABC transporter, ATP-binding/permease protein" and "DNA-binding response regulator" orthologs in all *B. cereus* genomes and *B. anthracis* A1055 (*B. cereus* ATCC 10987 GIs: 42781108, 42781109). However, the N-terminal ortholog is found split in all *B. anthracis* genomes except for *B. anthracis* A1055 (*B. anthracis* Sterne GIs: 49184826, 49184827). In this instance, the fusion group is the "main group" and the fission group is the "nested group". If we did not implement this group splitting procedure, the fission in the *B. anthracis* lineage would not have been automatically detected.

2.4.2 Fusion and fission are rare in most of the *Bacillaceae*

Using our `FusionFinder` suite of programs on the set of 33 fully sequenced *Bacillaceae* genomes, we found 628 FGs, requiring a total of 783 independent fusion or fission events (Figure 2.20). Of these, 611 were fissions, 42 were fusions, and 130 could not be determined. This is contrary to all previous reports where fusion either predominates or is similar in number. Addition of annotated pseudogenes to our dataset increased the total number of events (Table 2.5, Figure 2.21). Increasing the sequence identity cutoff, removing plasmid-encoded proteins, and considering only fusion groups that can be explained by a single event decreased the total number of events (Figure 2.22, 2.23, 2.24). Genome pairs within the *B. anthracis* clade are more similar to each other than any other pair within the *B. cereus* group, which may influence the number of pseudogenes inferred. Specifically, one might expect gene content to be more similar such as the unique acquisition of genes laterally or de-novo evolution of new genes. By manual inspection of our results and performing a new analysis using single *B. anthracis* genomes, we confirmed that fissioned genes in *B. anthracis* indeed have full-length orthologs in other *B. cereus* group genomes (Figure 2.25); we show only the results using *B. anthracis* Sterne. None of these variations affect the general trends discussed in the following sections. Note that the number of annotated pseudogenes does not necessarily correlate with the number of fissions predicted; for example the *B. cytotoxicus* genome contains 184 annotated pseudogenes but including these in the analysis results in an increase from six to 17 predicted fissions.

One approach to accessing accuracy of our method is to consider the number of events required to explain each fusion group we identified. Out of the 628 groups, 506 could be explained by a single event, which suggests that our species tree is broadly accurate. An

Figure 2.20: Fusion and fission events across 33 complete *Bacillaceae* genomes. Number of fusions and fissions are indicated by the first and second numbers on each node. The numbers at the root of the tree are the total number of events across the phylogeny.

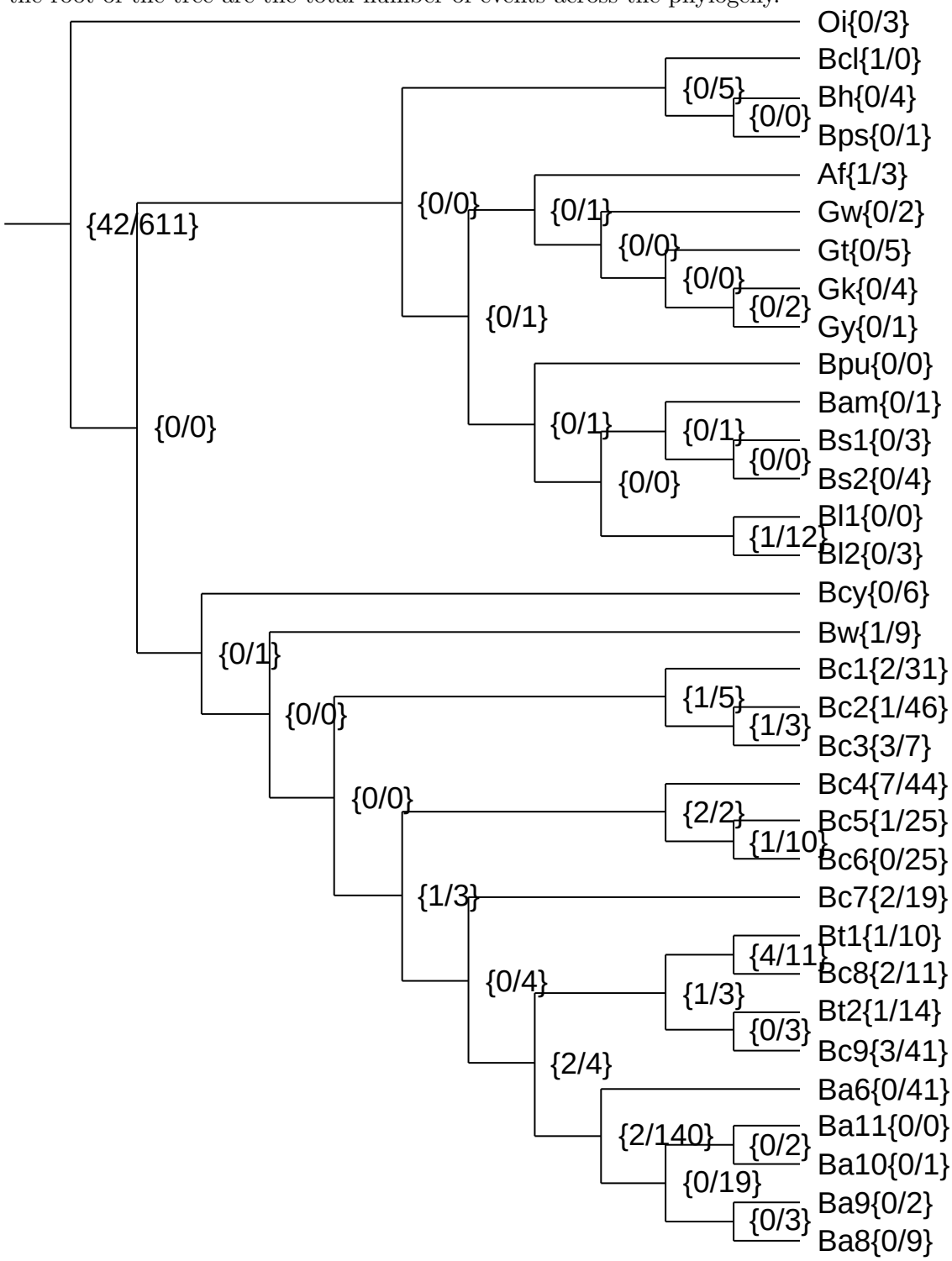


Figure 2.21: Fusion and fission events across 33 complete *Bacillaceae* genomes. Pseudogene annotations were included in this analysis.

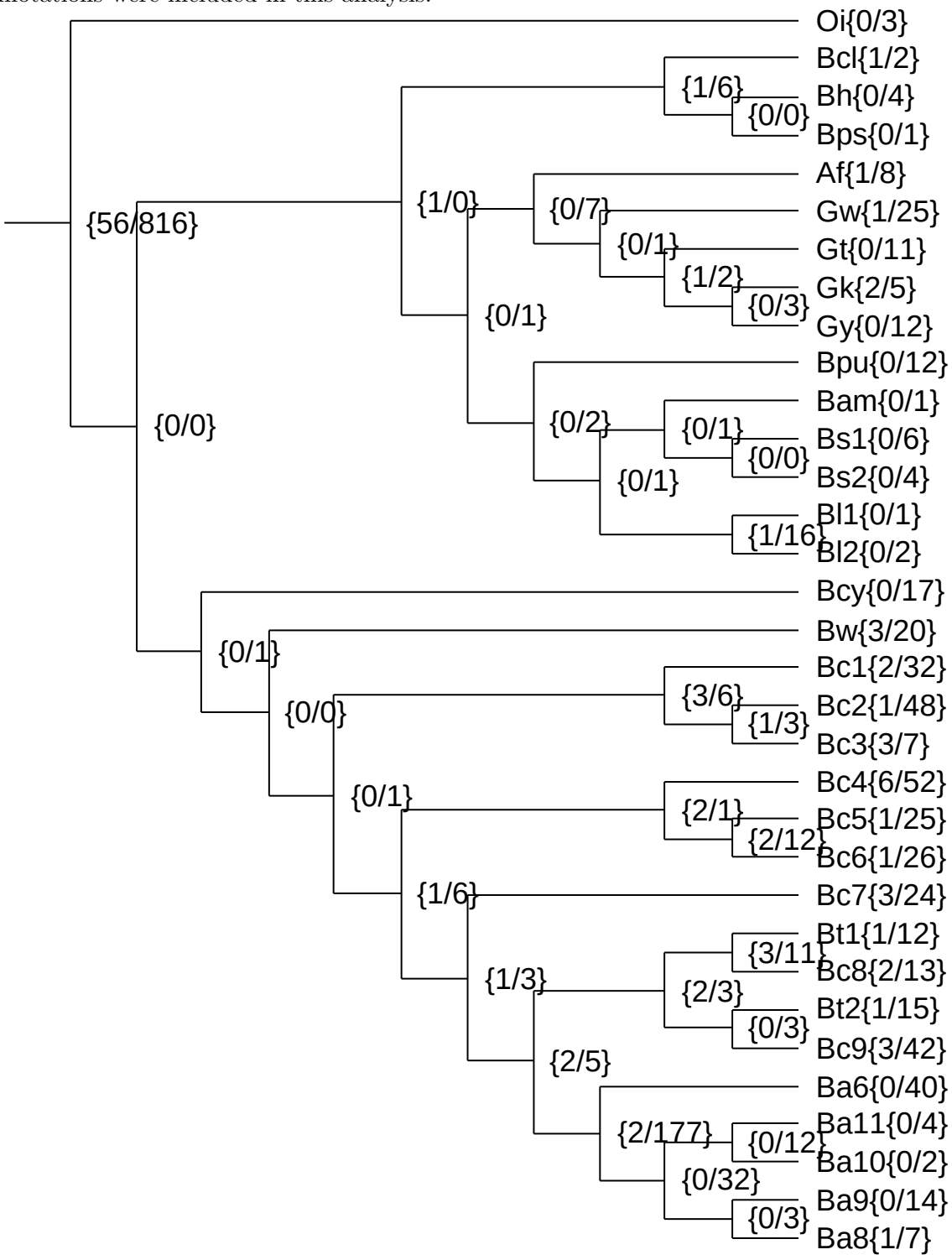


Figure 2.22: Fusion and fission events across 33 complete *Bacillaceae* genomes. The minimum percent identity between components and the composite was increased from 50% to 90%.

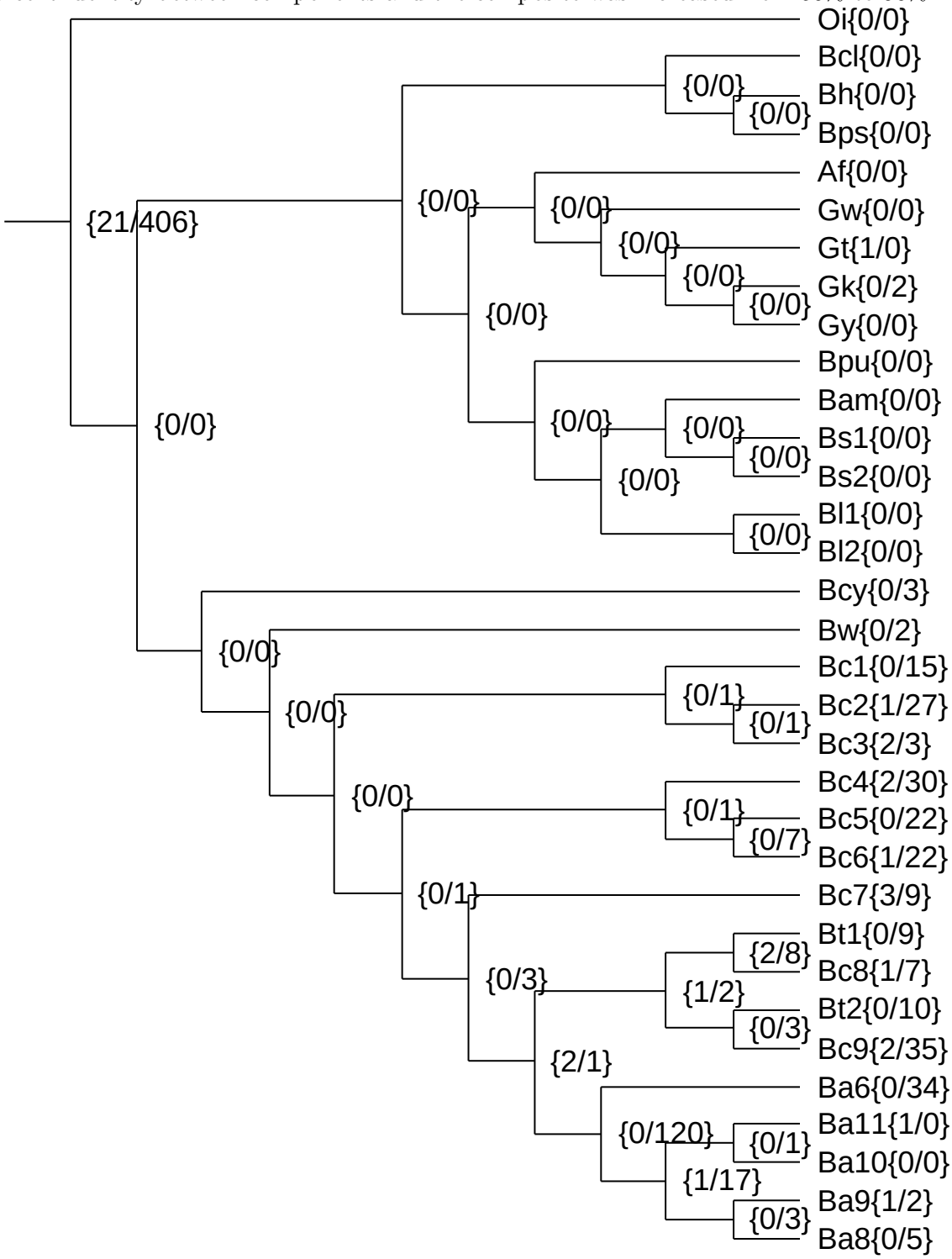


Figure 2.23: Fusion and fission events across 33 complete *Bacillaceae* genomes. Plasmid-encoded proteins were discarded from the analysis.

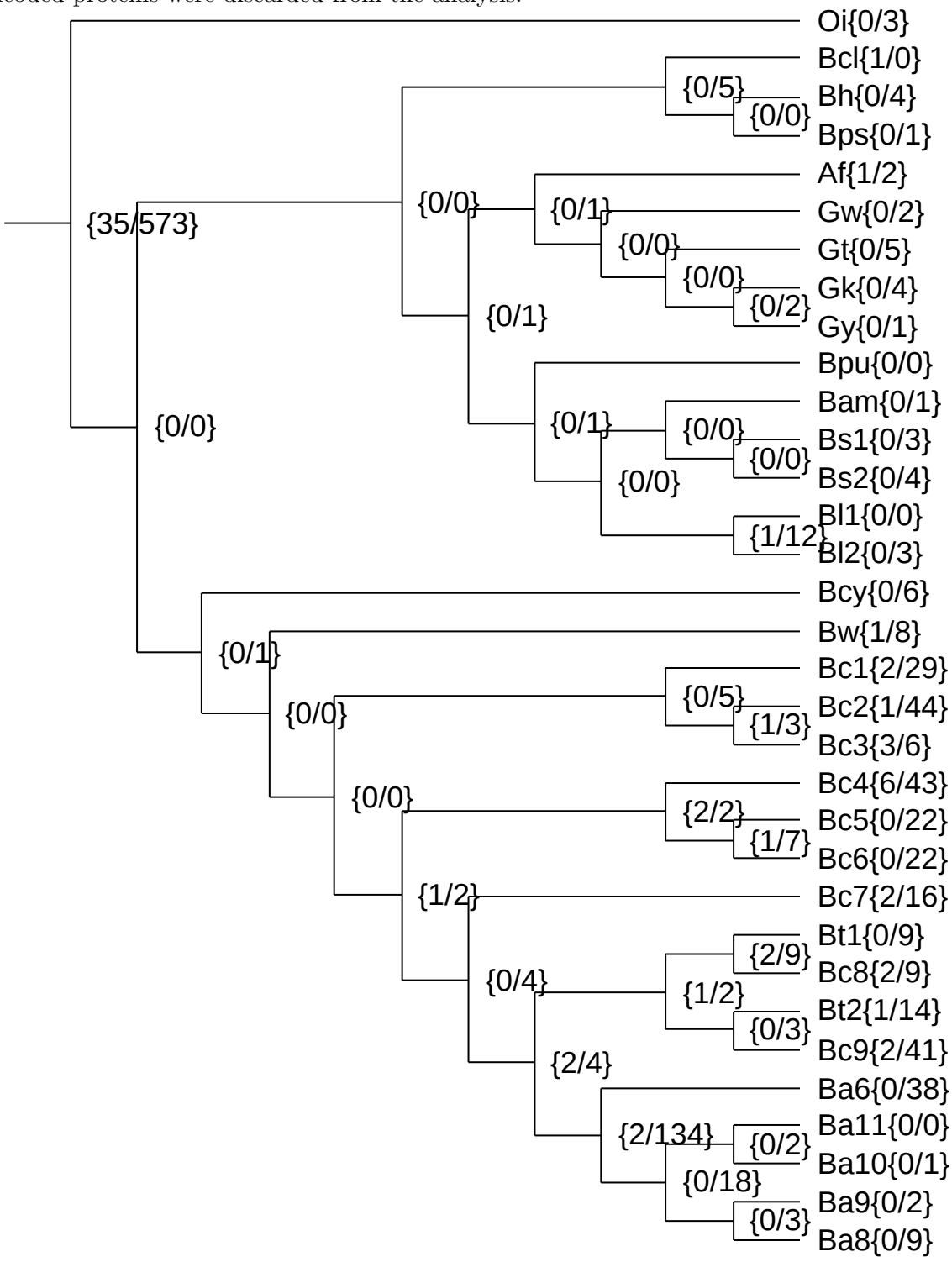


Figure 2.24: Fusion and fission events across 33 complete *Bacillaceae* genomes. Fusion groups that cannot be explained by a single evolutionary event were discarded from the analysis.

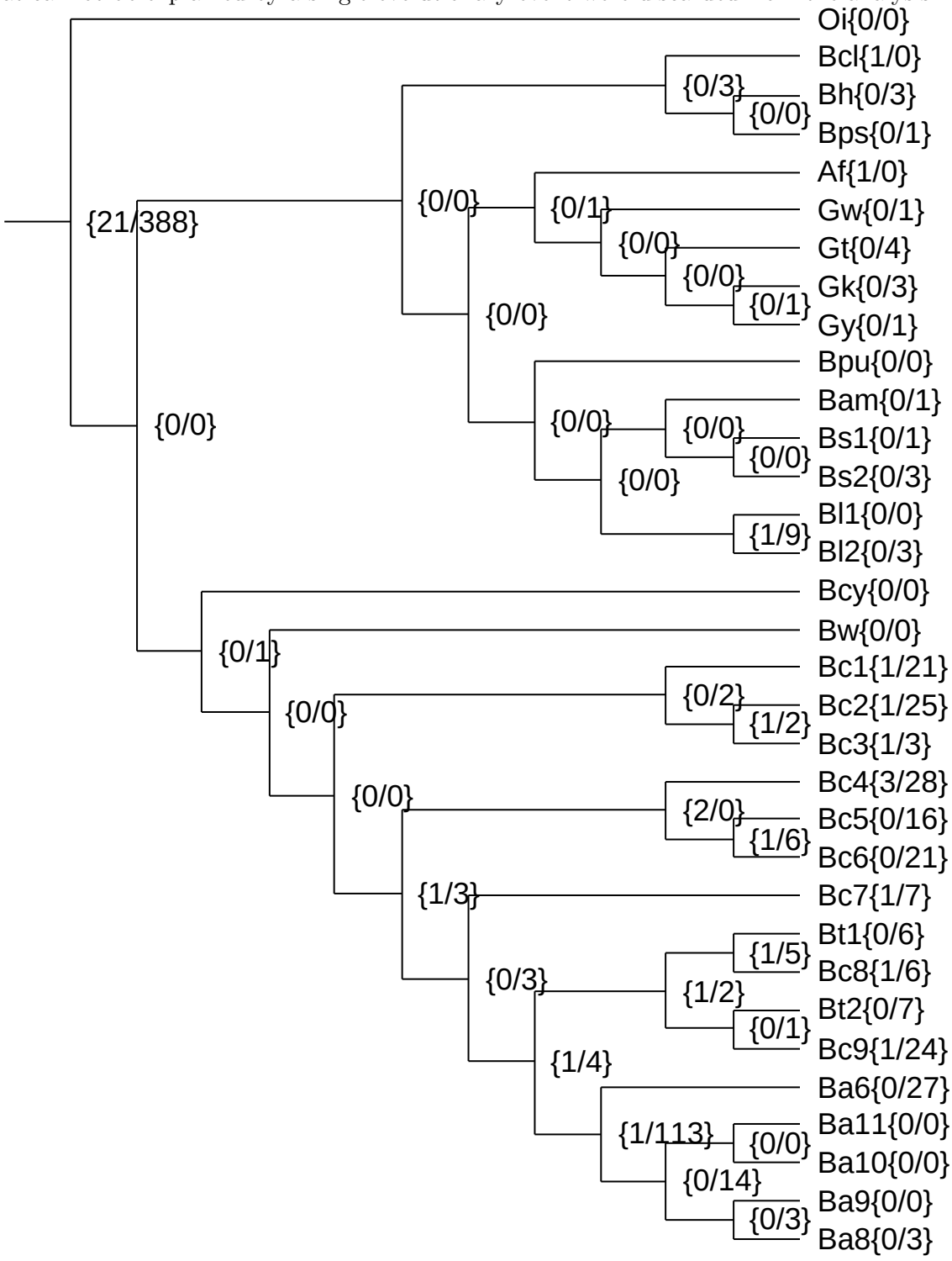


Figure 2.25: Fusion and fission events across 29 complete *Bacillaceae* genomes. *B. anthracis* Sterne was chosen as the representative for the *B. anthracis* clade.

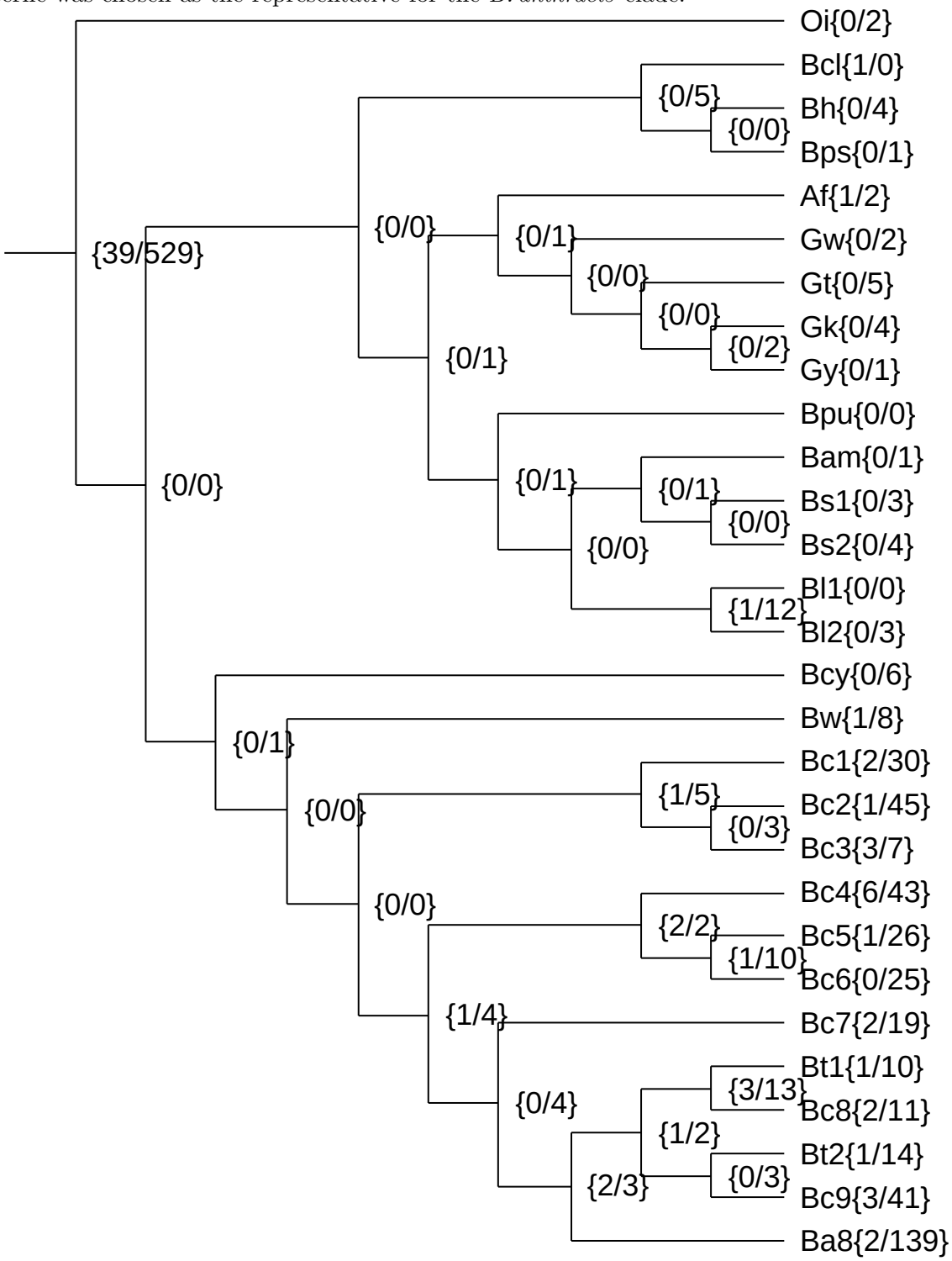


Table 2.5: Numbers of annotated pseudogenes in *B. anthracis* and *B. cereus* CI chromosomes and plasmids. Pseudogenes annotated as "authentic" or "not the result of a sequencing artifact" were identified. *B. anthracis* lineages are indicated (Figure 2.26). The rest of the *Bacillaceae* is omitted for brevity.

ID	Accession	Lineage	Number of pseudogenes	Authentic
Bci	NC_014335	n/a	10	0
Ba1	NZ_AAEO01000001	C	0	0
Ba2	NZ_AAEQ01000001	B1	0	0
Ba3	NZ_AAEN01000001	B2	0	0
Ba4	NZ_AAER01000001	A1	0	0
Ba5	NZ_AAEP01000001	A1	0	0
Ba6	NC_012581	A1	0	0
Ba7	NZ_AAES01000001	A2	0	0
Ba8	NC_005945	A2	0	0
Ba9	NC_012659	A2	613	0
Ba10	NC_003997	A2	176	22
Ba11	NC_007530	A2	270	176

additional 101 groups required two steps, 20 groups require 3 to 5 steps, and 1 group required 7 steps. Another approach is to consider differences between very similar genomes; we considered three genome pairs:

1. *B. subtilis* 168 (Bs1) and *B. subtilis* W23 (Bs2)
2. The two *B. licheniformis* genome sequences (Bl1, Bl2)
3. *B. anthracis* Ames (Ba10) and *B. anthracis* Ames Ancestor (Ba11)

The first pair are the two primary, well-studied strains that are distinct subspecies of *B. subtilis*, the model organism for Gram-positive bacteria, so we are confident about the monophyly of this pair. *B. subtilis* 168 has recently been resequenced and reannotated (Barbe *et al.*, 2009) using second generation sequencing so the quality of the data should be high; note that Barbe *et al.* (2009) mention fusions and fissions but they mean adjustments of original annotations, not evolutionary events. We observe three fission events specific to Bs1 and four events specific to Bs2, no fusion events, and no events common to both subspecies. The lack of a significant difference in number of events suggests that the Bs2 genome may be of similar quality. *B. licheniformis* was sequenced independently and concurrently by two groups (Rey *et al.*, 2004; Veith *et al.*, 2004) and is a good test for potential sequencing and annotation problems. Our programs predict no events and three fissions specific to the Bl1 and Bl2 genomes; one fusion and 12 fission events were mapped to the split between these two genomes. Assuming that both genomes were sequenced and annotated independently with different algorithms, this gives greater confidence to the shared events. Finally, *B. anthracis* Ames Ancestor is the progenitor of all Ames strains used in research labs, including *B. anthracis* Ames which was cured of both virulence plasmids (Read *et al.*, 2003; Ravel *et al.*, 2009) so any differences are likely due to the plasmid-curing process (Ravel

et al., 2009) or sequencing or annotation errors. Only one fission is specific to Ba10 and two additional fission are common to both strains.

Across the entire phylogeny, *B. pumilis* (Bpu) contains no genome-specific events while the ATCC genomes *B. cereus* ATCC 14579 (Bc2) and *B. cereus* ATCC 10987 (Bc4) contain the most. Although the events could be real, without further inspection it is more conservative to assume an inaccurate placement of the strains on the phylogeny or an indication of a sequencing or annotation problem; note that the placement of *B. cereus* ATCC 10987 on the phylogeny is identical to a phylogeny built using Neighbor-Joining on pairwise whole-genome alignments (Tourasse and Kolstø, 2008; Kristoffersen *et al.*, 2011). Unexpectedly, the TMP gene fission described above is just one of 140 fissions specific to the *B. anthracis* lineage, much greater than any one genome or subgroup; Another 41 were specific to *B. anthracis* CDC 684 (Ba6) and 19 fissions were common to the remaining *B. anthracis* genomes (Ba8–Ba11). Manual analysis of a subset of these FEs confirmed that they were not due to an algorithmic problem. Three major *B. anthracis* lineages A, B, and C are recognized from genome-wide SNP analysis and canSNPs (analogous to tagSNPs) in concert with VNTR data (Pearson *et al.*, 2004; Van Ert *et al.*, 2007). The genome-wide SNP data was acquired through 12X draft genome sequencing of diverse *B. anthracis* strains previously recognized and included isolates from sublineages A1, A2, B1, and B2. However, these studies included only one of the five complete *B. anthracis* genomes in their analyses so the identity of the non-*B. anthracis* Ames strains was unknown. In the interest of determining whether the fission patterns can be further subdivided by the previously recognized major lineages, we tested the effect of including draft versus complete *B. anthracis* genomes in the following section.

2.4.3 Fission, SNP, and VNTR patterns in *B. anthracis* are consistent

We expanded the canSNP analysis of Van Ert *et al.* (2007) to all 5 complete *B. anthracis* genomes, comprising a total of 11 *B. anthracis* strains (Figure 2.26). The tree is completely consistent with the original SNP tree (Van Ert *et al.*, 2007), with *B. anthracis* CDC 684 (Ba6) clustering with *B. anthracis* Vollum in the A1 lineage and the remaining genomes (Ba8–11) clustering in the A2 lineage; as expected, *B. anthracis* Ames clusters with *B. anthracis* Ames Ancestor which was used in the original study. Also consistent with the results of Van Ert *et al.* (2007), the canSNPs cannot fully resolve the A1 and A2 lineages (Table 2.5), although they found that the combined canSNP and VNTR data can do so.

Two new datasets were constructed, coupling completely sequenced *Bacillaceae* genomes with draft (Figure 2.28) and complete (Figure 2.27) *B. anthracis* genomes from the A1 and A2 lineages. The results show that the inclusion of draft genomes does not have an appreciable effect on the number of fusion and fission events predicted. While the total number of fissions is higher in Figure 2.27, this is due to the higher number of events common to both *B. anthracis* genomes (144 vs 136) and not the total number of *B. anthracis* strain-specific events (42 vs 47).

We applied our `FusionFinder` pipeline to the final phylogeny including 6 *B. anthracis* genomes placed using our SNP analysis and *B. cereus* CI based on the phylogenies provided by Klee *et al.* (2010) (Figures 2.29,2.30). Including pseudogenes, a total of 70 fissions are common to all *B. anthracis* genomes, 66 are common to both A and B lineages (Ba2–Ba11),

Figure 2.26: *B. anthracis* phylogeny inferred using canSNPs. Newly incorporated complete genomes are marked by asterisks. Note the unresolved node in lineage A.

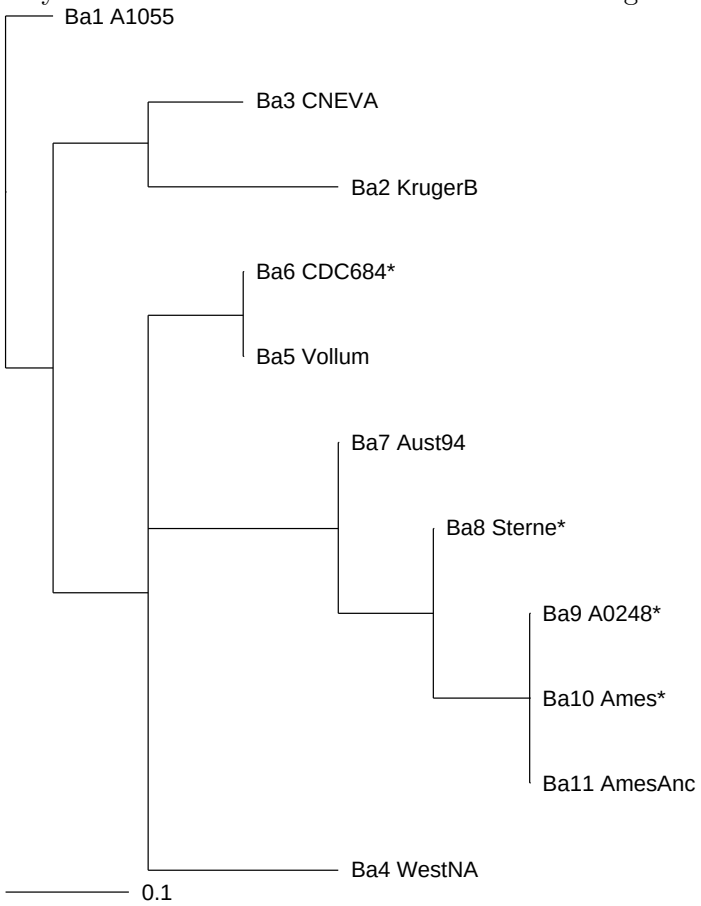


Figure 2.27: Fusion/fission analysis using 12X draft genomes Ba5 (lineage A1) and Ba7 (lineage A2).

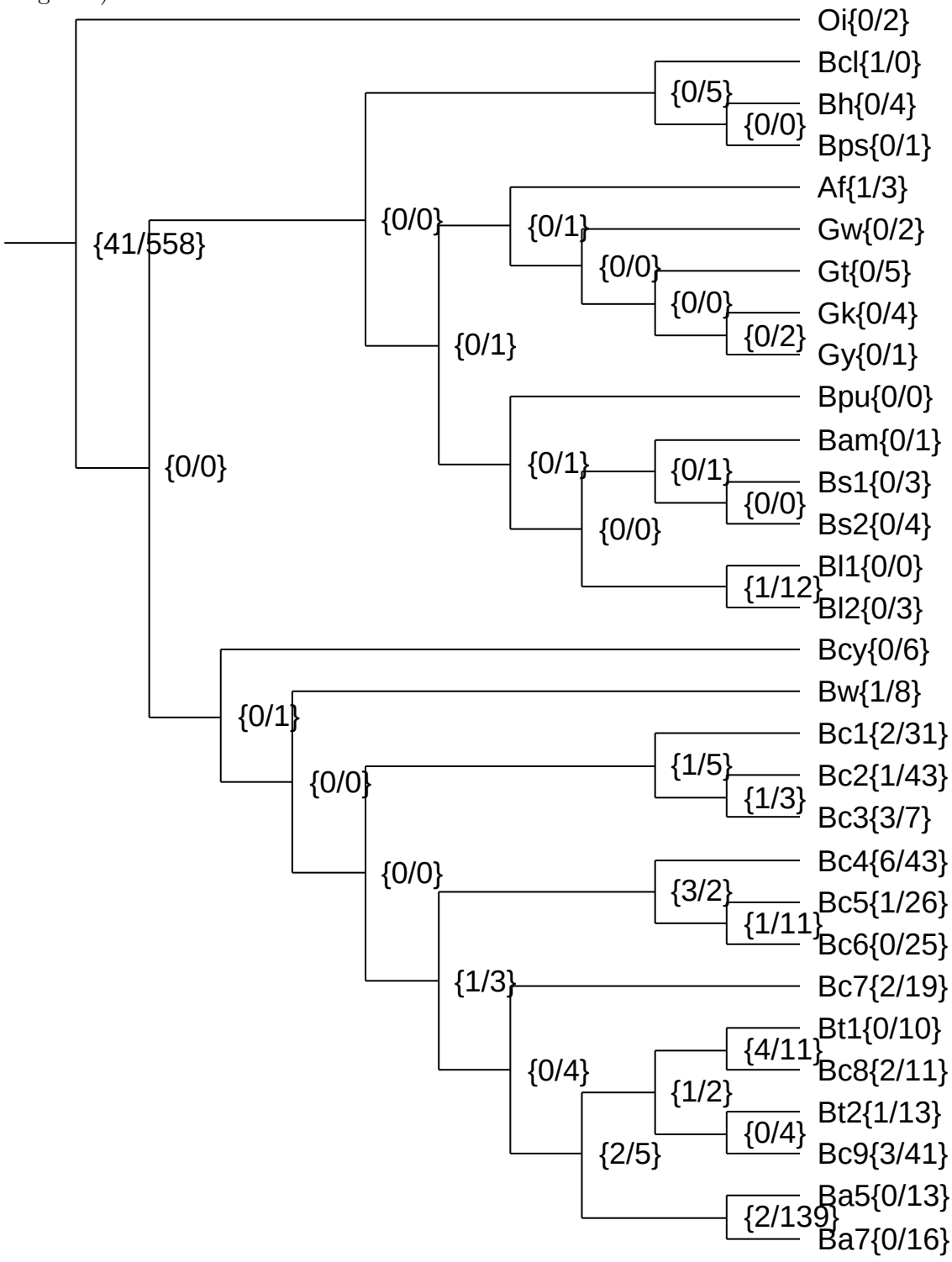
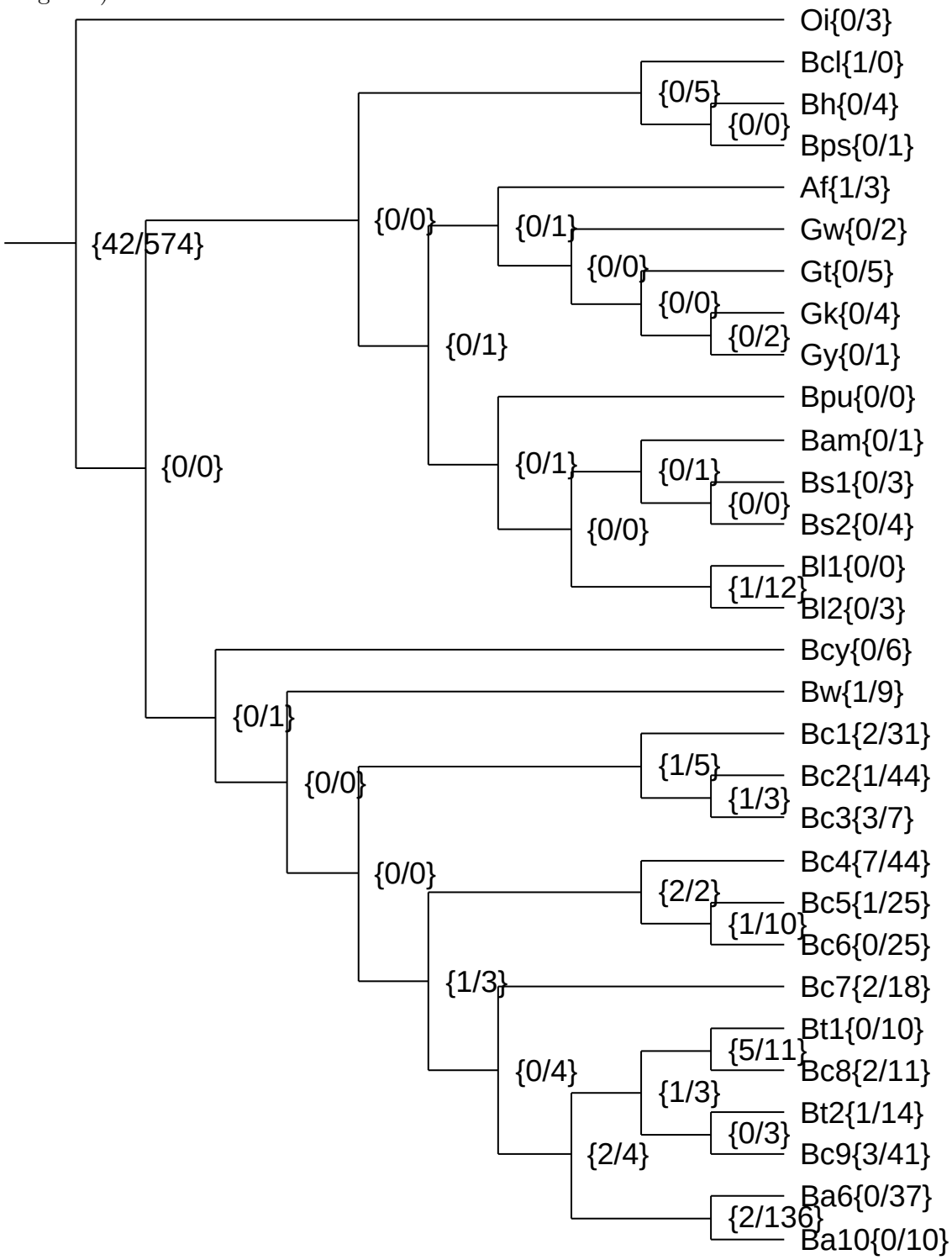


Figure 2.28: Fusion/fission analysis using complete genomes Ba6 (lineage A1) and Ba10 (lineage A2).



21 are common to the B lineage (Ba2–Ba3), and 47 are common to the A lineage (Ba4–Ba11). The 177 fissions previously identified (Figure 2.21) are now spread out along the *B. anthracis* lineage and less than half of the fissions are common to all *B. anthracis* genomes. It is particularly interesting to note the presence of lineage-specific fissions predicted not only for the major lineages but also for sublineages A1 and A2. The lack of sampling in the B and C lineages precludes any further conclusions.

A correct species phylogeny is imperative for accurate inference of fusion events. Instead of minimizing the number of events required to explain our fusion patterns given a species phylogeny, we can systematically explore tree space to find the phylogeny that requires the least number of events given our fusion patterns. The major question is whether these patterns are phylogenetically informative and whether there are enough events to detect the signal if it exists. We performed this analysis as described in the methods on data from all 40 genomes, excluding or including pseudogenes (Figure 2.31, 2.32). As expected, we cannot infer most phylogenetic groupings using with high confidence using fusion patterns. Surprisingly, all *B. anthracis* strains were correctly separated into their respective lineages and sublineages consistent with the SNP tree (Table 2.5,Figure 2.26). Importantly, the bootstrap support is very high for the *B. anthracis* clade and many of the internal nodes. Interestingly, the A1 lineage was correctly resolved with bootstrap support values of 0.82 and 1.00 corresponding to the unresolved node and resolved nodes in the SNP tree. There is weak support overall for the A2 lineage and Ba8 could not be differentiated from the unresolved node within the A2 lineage. The unresolved node in the A2 lineage of the SNP tree contains three genomes, which are also the only genomes with pseudogenes annotated. As will be shown further below, a large number of fissioned genes in these three genomes are annotated as pseudogenes, which may explain the lack of support within the A2 lineage in the bootstrap analysis (Table 2.6). When pseudogenes are included, the only two nodes in the *B. anthracis* clade with bootstrap support less than 0.9 correspond to the unresolved nodes in the SNP tree. Although the support is only 0.84 at the root of the A1 lineage, the bipartition matches the VNTR data from Van Ert *et al.* (2007). *B. anthracis* Ames was derived from *B. anthracis* Ames Ancestor (Ravel *et al.*, 2009) so the grouping of Ba10 and Ba11 is intuitive even though the bootstrap support is only 0.75. Therefore, we concluded that the fusion patterns are fully consistent with the SNP and VNTR data previously published and highly supported based on a bootstrapping analysis. The draft genomes were chosen to represent *B. anthracis* diversity so we expect bootstrap support to decrease if sampling within lineages are increased, which can already be seen in the current dataset. The only nodes outside the *B. anthracis* lineage with higher or similar bootstrap support to nodes within *B. anthracis* are the groupings of Bc5 with Bc6 and Bt1 with Bc8, both of which are consistent with our phylogeny. Ultimately, fusions and fissions appear to provide an intermediate level of phylogenetic resolution between SNPs and VNTRs in *B. anthracis*.

2.4.4 Annotation of fissioned genes differs between *B. anthracis* genomes

In our initial analysis using only annotated proteins, we noticed a subset of *B. anthracis* genomes contained many shared component sets with no detectable composites or components (assigned the '?' state as mentioned in the methods) in the other *B. anthracis* genomes; due to the parsimony principle a single event was inferred in the cenancestor of *B. anthracis*

Figure 2.29: Fusion/fission analysis of 40 *Bacillaceae* genomes, including 6 draft *B. anthracis* genomes and *B. cereus* CI.

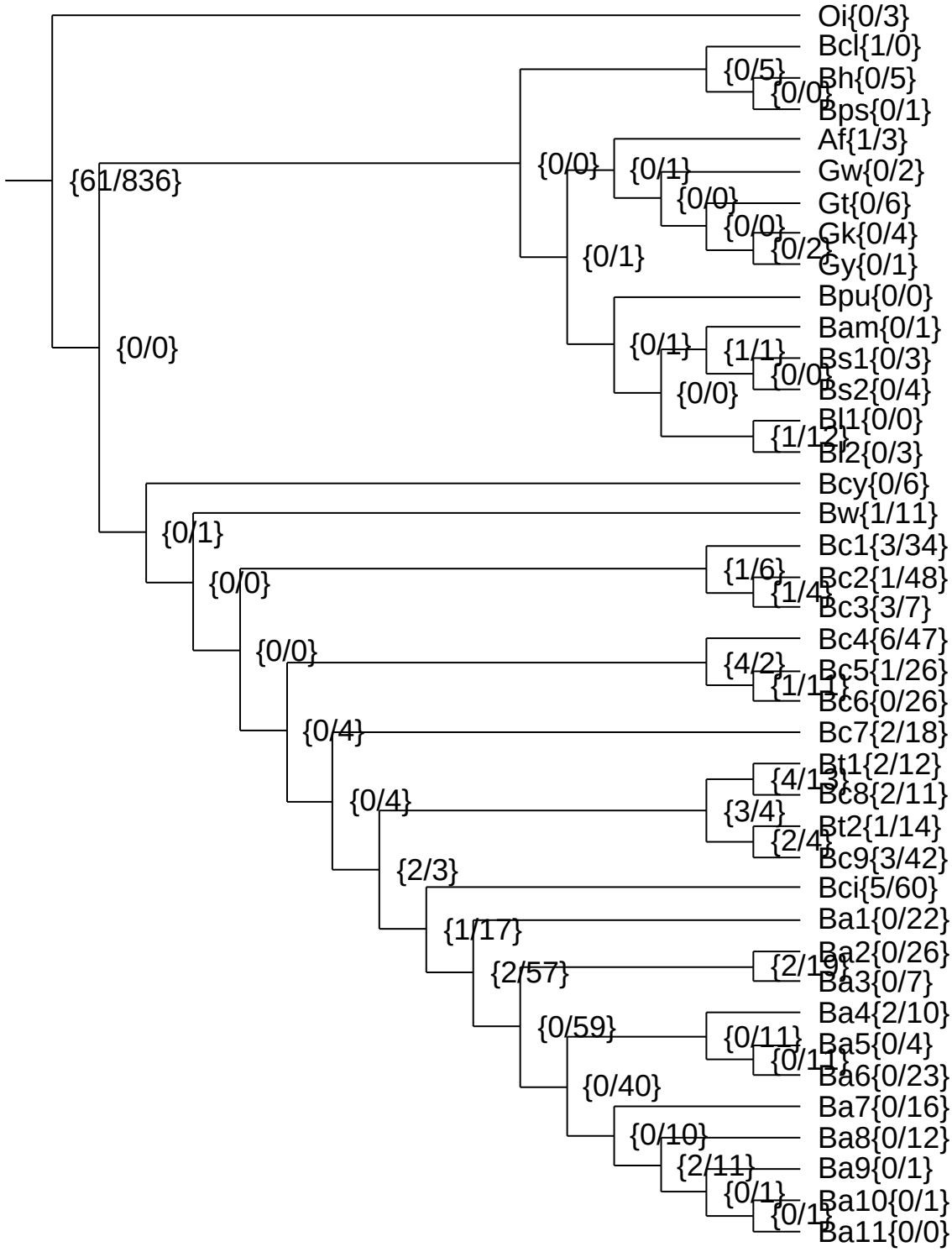


Figure 2.30: Fusion/fission analysis of 40 *Bacillaceae* genomes, including 6 draft *B. anthracis* genomes and *B. cereus* CI. Pseudogene annotations were included in this analysis.

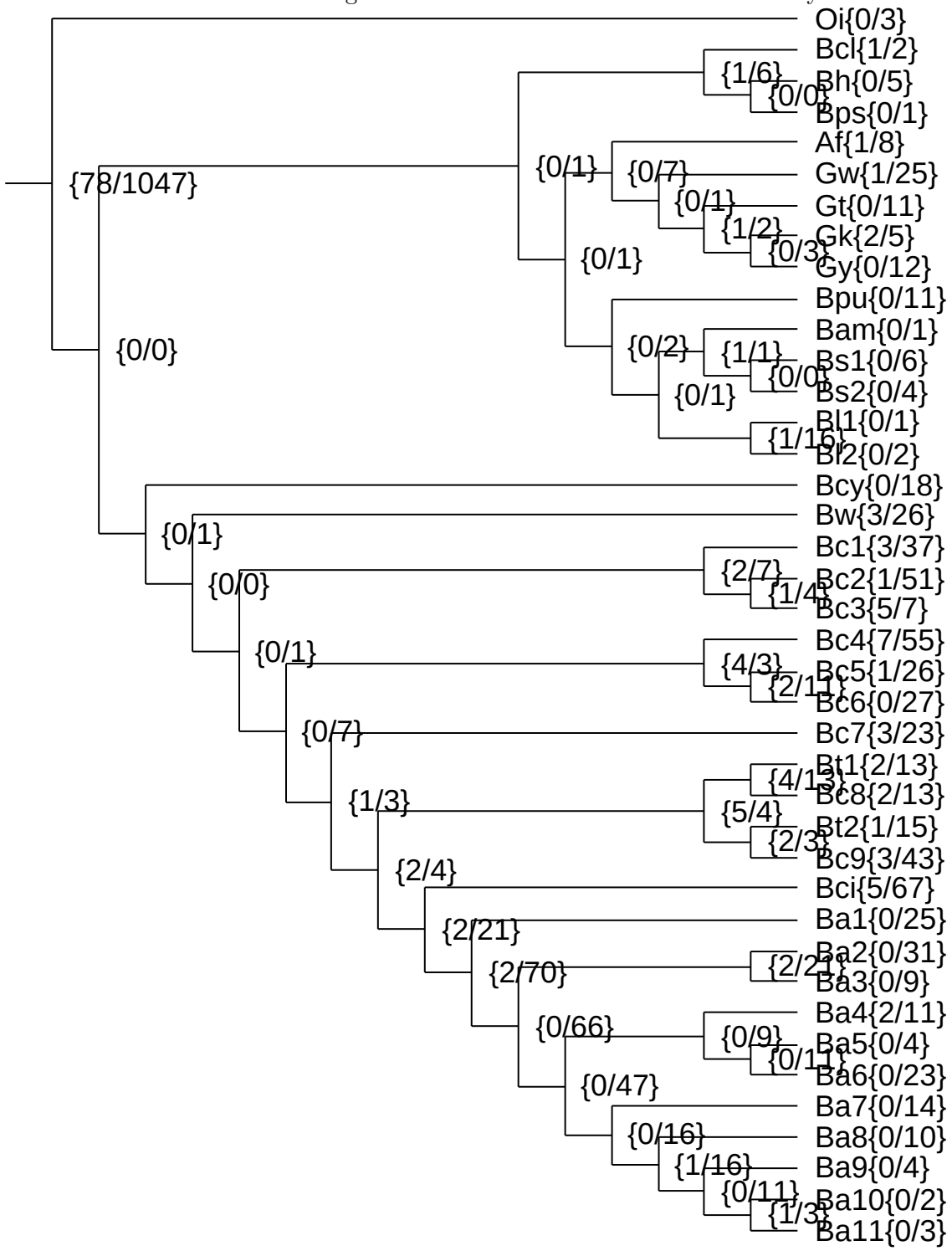


Figure 2.31: Consensus of phylogenies inferred from fusion and fission patterns using Camin-Sokal parsimony. The proportion of 1000 bootstrap samples supporting each partition in the tree is indicated. Pseudogenes were not included in this analysis.

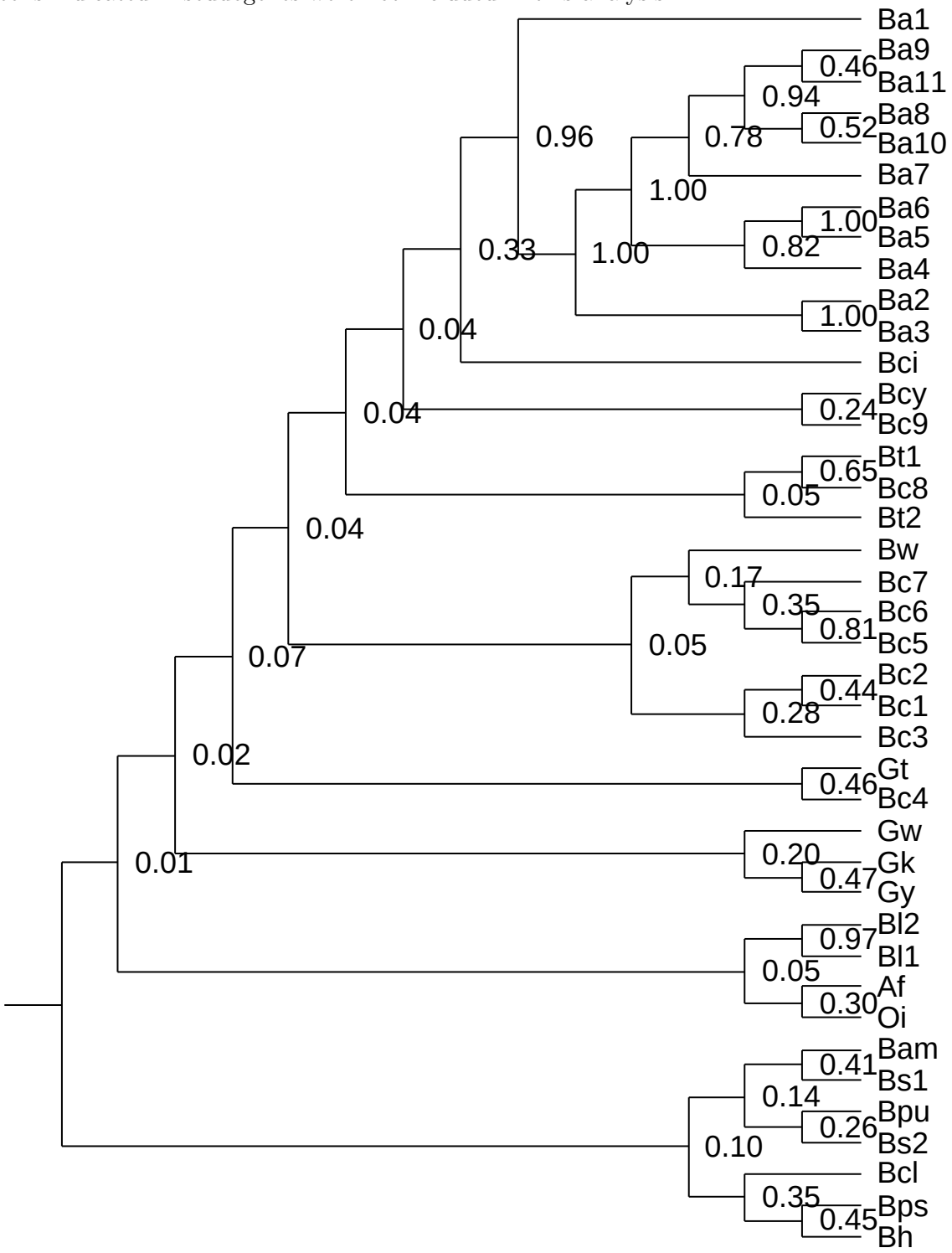
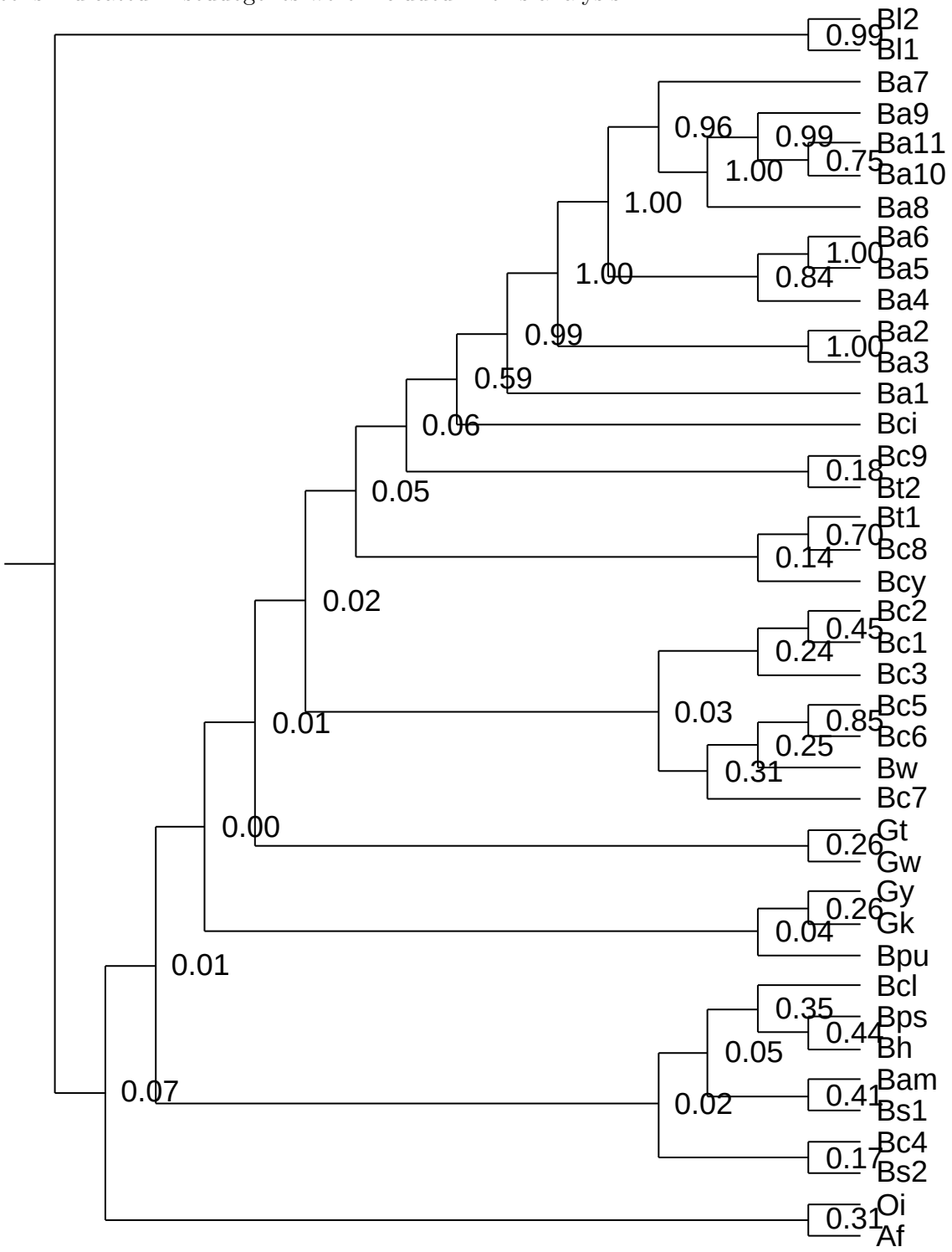


Figure 2.32: Consensus of phylogenies inferred from fusion and fission patterns using Camin-Sokal parsimony. The proportion of 1000 bootstrap samples supporting each partition in the tree is indicated. Pseudogenes were included in this analysis.



for each of these fusion groups. To determine whether this observation was due to an algorithmic problem (`FusionFinder` fails to identify the annotated proteins), components being annotated as pseudogenes, or components being unannotated (but present in the genome), we developed `PseudogeneParser` and `FusionScanner` to analyze annotated pseudogenes and genomic sequences for additional fusion events. For each *B. anthracis* genome, we counted the number of fission events predicted along the lineage up to the *B. anthracis* cenancestor and determined whether the component sets were identified via proteins, pseudogenes, or genomic scan (Table 2.6). Inclusion of pseudogenes increases the overall number of component sets identified which is not surprising since our algorithm was designed to construct fusion groups only if at least one set of components to be annotated as proteins or pseudogenes as noted in the methods. For the three genomes with pseudogene annotations (Table 2.5), the majority of component sets are not annotated as proteins but can be detected by `FusionScanner`. When we applied `PseudogeneParser`, we found in fact that in these genomes most component sets were correctly identified as pseudogenes; the only genome with less component sets identified by `FusionScanner` is Ba1 although it also contains the least component sets overall. *B. anthracis* Ames (Ba9) was cured of its plasmids which likely accounts for the lower number of component sets. Although the number of component sets identified by each method are similar between Ba9 and Ba11, Ba9 actually contains more than twice the number of annotated pseudogenes (Table 2.5). When considering component sets from each fusion group predicted by `MultiFusion`, we found that most component sets identified in Ba9 are represented by multiple pseudogene entries whereas those identified in Ba10 and Ba11 are represented by a single pseudogene entry (Table 2.7), partially explaining this discrepancy. Thus, it appears `FusionScanner` and `PseudogeneParser` are accurate when used separately and in conjunction, although better results are obtained in the latter case.

There are a number of unidentified component sets in each genome and this pattern is true for events mapped to the cenancestor of *B. anthracis* (Table 2.8). This could be due to an algorithmic problem. Another possibility is that they fall under cases that our algorithm current does not handle. Importantly, `FusionFinder`, `FusionScanner`, and `PseudogeneParser` act independently without knowledge of the others. Thus, component sets can only be identified if there is uniform annotation. The most intriguing possibility is that fission eventually leads to gene loss. In the future each of these events will be manually inspected to determine the reason for each unidentified component set.

Ultimately, we find that genome annotations vary irrespective of whether they are finished or at draft stage. Of the 5289 annotated proteins in *B. anthracis* Sterne (Ba8), at least 276 (5.2%) were identified as pseudogenes because there are 138 component sets each containing at least two components. In *B. anthracis* Ames (Ba10), at least 90 of 5328 proteins (1.7%) were identified as pseudogenes. Thus, depending on the type of study, the choice of *B. anthracis* genome and exclusion or inclusion of pseudogenes may have a significant effect on the results.

Since `FusionMapper` can infer multiple events in a single fusion group, we also tested whether fissions were unique. For each fission inferred in a *B. anthracis* genome, we determined whether a single fission event was inferred for the fission group. If all fissions mapped to *B. anthracis* lineages, are considered, at least 75% of the events are unique to that lineage (Table 2.6). When the subset of fissions common to all *B. anthracis* are considered, at least 80% are unique to *B. anthracis* (Table 2.8). In other words, at least 75% of gene fission

Table 2.6: Characterization of fission events mapped to the *B. anthracis* clade when excluding or including pseudogenes. The number of fission groups mapped is equal to the sum of the "Tot" and "Uk" columns.
Tot: Number of fission groups with identifiable component sets.
Pr: Number of component sets identified by annotated proteins.
Ps: Number of component sets identified by annotated pseudogenes.
GS: Number of component sets identified by genome scan.
Uk: Number of fission groups with no identifiable component sets.
Un: Percentage of fission groups that can be explained by a single fission event.
RBH: Percentage of component sets where a component is a RBH of a composite.

	Figure 2.29						Figure 2.30						
ID	Tot	Pr	GS	Uk	Un	RBH	Tot	Pr	Ps	GS	Uk	Un	RBH
Ba1	70	55	15	9	78.5	92.7	78	55	0	23	17	78.9	92.7
Ba2	151	122	29	10	76.4	93.4	169	120	0	49	19	76.6	94.2
Ba3	138	110	28	4	78.2	94.5	152	109	0	43	14	76.5	95.4
Ba4	172	149	23	5	76.8	90.6	187	146	0	41	16	76.4	91.1
Ba5	174	152	22	8	76.9	92.1	188	150	0	38	19	76.3	92.7
Ba6	189	176	13	12	77.6	93.2	203	173	0	30	23	77.4	93.6
Ba7	177	153	24	5	78.6	94.1	196	151	0	45	17	76.5	94.0
Ba8	167	141	26	22	78.8	92.9	185	138	0	47	40	76.4	93.5
Ba9	145	2	143	34	78.8	100.0	211	2	179	30	19	74.8	100.0
Ba10	143	46	97	37	78.3	97.8	190	45	128	17	41	75.3	100.0
Ba11	147	11	136	32	78.2	100.0	214	11	193	10	18	75.4	100.0

Table 2.7: Pseudogenes in *B. anthracis* genomes that were or were not mapped to fusion groups. Unlike Table 2.5, we only consider pseudogenes on the chromosome.

ID	Component sets	Pseudogenes	
		Mapped	Unmapped
Ba9	196	372	163
Ba10	146	146	30
Ba11	218	219	37

Table 2.8: Characterization of fission events mapped to the *B. anthracis* cenancestor by `FusionMapper` when excluding or including pseudogenes. The number of fission groups mapped is equal to the sum of the "Tot" and "Uk" columns.
Tot: Number of fission groups with identifiable component sets.
Pr: Number of component sets identified by annotated proteins.
Ps: Number of component sets identified by annotated pseudogenes.
GS: Number of component sets identified by genome scan.
Uk: Number of fission groups with no identifiable component sets.
Un: Percentage of fission groups that can be explained by a single fission event.
RBH: Percentage of component sets where a component is a RBH of a composite.

	Figure 2.29						Figure 2.30						
ID	Tot	Pr	GS	Uk	Un	RBH	Tot	Pr	Ps	GS	Uk	Un	RBH
Ba1	48	39	9	9	80.7	92.3	53	39	0	14	17	82.9	92.3
Ba2	51	44	7	6	80.7	95.5	58	43	0	15	12	82.9	97.7
Ba3	54	45	9	3	80.7	95.6	60	45	0	15	10	82.9	97.8
Ba4	54	47	7	3	80.7	95.7	59	45	0	14	11	82.9	97.8
Ba5	54	48	6	3	80.7	95.8	59	46	0	13	11	82.9	97.8
Ba6	52	49	3	5	80.7	93.9	57	47	0	10	13	82.9	95.7
Ba7	54	48	6	3	80.7	97.9	59	46	0	13	11	82.9	97.8
Ba8	45	35	10	12	80.7	97.1	50	33	0	17	20	82.9	100.0
Ba9	44	0	44	13	80.7	∞	61	0	51	10	9	82.9	∞
Ba10	42	11	31	15	80.7	90.9	52	10	34	8	18	82.9	100.0
Ba11	46	4	42	11	80.7	100.0	61	4	52	5	9	82.9	100.0

events in *B. anthracis* are defining characteristics than can be used to differentiate it from *B. cereus* genomes.

Interestingly, we noticed that some pseudogenes we identified as components were annotated as "authentic" or "not the result of a sequencing artifact". A keyword search of all annotated pseudogenes revealed that 65.2% of the pseudogenes in *B. anthracis* Ames Ancestor were annotated as such (Table 2.5). Although we could not determine how pseudogenes were authenticated, it gives us greater confidence in our predictions.

2.4.5 Gene fission is not associated with gene duplication

Our algorithm does not account for paralogous proteins which may be problematic since we had to remove the RBH criteria. Figure 2.6 depicts a gene fission event in *H. acinonychis* that would have been missed if we required reciprocality between the composite and one of the components. If the sequence identity between HP_1165 and Hac_0765 was similar or higher than its sequence identity to genes in the component set, we may have inferred duplication followed by gene fission and relaxation of selection acting on one copy. Note that we cannot rule out a false positive prediction given either instance.

For a given component set in each fission group, we took the composite gene from one reference genome (there are usually multiple) and tested the RBH criteria (Tables 2.6, 2.8); only component sets containing annotated proteins were assessed. We found that the majority of our gene fission predictions contain RBH links between a component and its composite. In fact, this applies to at least 91.6% of component sets in *B. anthracis* genomes. These numbers represent lower bounds because we considered only one composite gene. Thus, gene duplication does not have a large effect on our results and it is clear that gene fissions must be filtered in protein length studies.

2.4.6 Contribution of PlcR and motility loss to gene fission in *B. anthracis*

B. anthracis is well known for having a specific mutation shared amongst all members that inactivates the PlcR regulon (Rasko *et al.*, 2004, 2005; Han *et al.*, 2006; Klee *et al.*, 2010); the few non-anthracis Cereus group isolates with an inactivated copy do not have the same mutation as the *B. anthracis* version (Kolstø, Tourasse and Økstad, 2009). Of 45 genes determined to be under direct regulation by 28 PlcR boxes in *B. cereus* ATCC 14579 (Gohar *et al.*, 2008), eight were placed in fusion groups (Table 2.9). Only two of the fissions are shared amongst all *B. anthracis* strains, one of which is *plcR* and the other encodes an enterotoxin component *nheC* (Gohar *et al.*, 2008). Four other events are predicted in three fusion groups in *B. anthracis* strains; all other events are in *B. cereus* group genomes. Assuming that most of the genes under direct control of PlcR in *B. anthracis* are orthologous to those identified in *B. cereus* ATCC 14579, the inactivation of the PlcR regulon is insufficient in explaining the large number of fission events in *B. anthracis*. If the fissioned PlcR-regulated genes had downstream targets then we could not be sure of the effect of the *plcR* fission but this does not appear to be the case.

Another distinguishing feature is the lack of motility of *B. anthracis* strains (Rasko *et al.*, 2004, 2005; Han *et al.*, 2006; Kolstø, Tourasse and Økstad, 2009; Yu, 2009; Klee *et al.*, 2010; Waltman *et al.*, 2010). Klee *et al.* (2010) reported the sequencing of *B. cereus* CI containing

Table 2.9: PlcR-controlled genes involved in fusion or fission events. Locus tags and GI numbers indicate the orthologs in *B. cereus* ATCC 14579. Events that were identified only after including pseudogenes in the analysis are indicated. Events within specific *B. anthracis* strains are specified; otherwise the number of events in the *B. cereus* group is indicated. Bc5350 is the gene for PlcR.

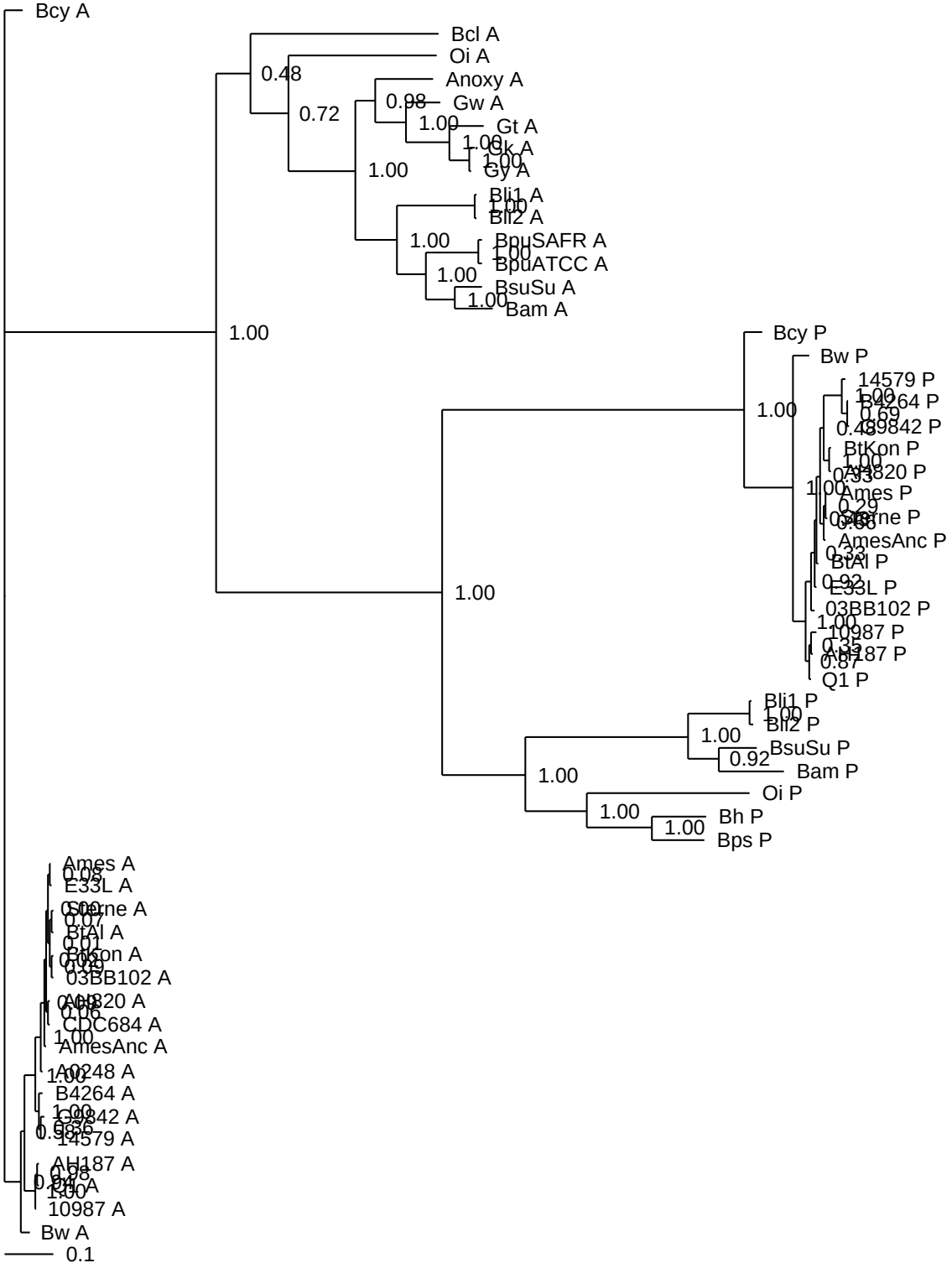
Locus tag	GI	Pseudogenes	Branches
Bc0666	30018848	no	*B. anthracis* CDC 684
Bc1811	30019953	yes	all *B. anthracis*
Bc2410	30020541	no	1 *B. cereus*
Bc2411	30020542	no	*B. anthracis* CDC 684, *B. anthracis* KrugerB
Bc3747	30021841	no	2 *B. cereus*, *B. anthracis* KrugerB
Bc3762	30021855	yes	3 *B. cereus*
Bc4999	30023039	no	1 *B. cereus*
Bc5350	30023380	no	all *B. anthracis*

the *B. anthracis* virulence plasmids and a full complement of 10 motility- and chemotaxis-associated genes that are fragmented in two *B. anthracis* strains; 4 *B. cereus* strains were used in their comparison. Waltman *et al.* (2010) performed a comparison between *B. anthracis* Sterne, *B. cereus* ATCC 14579, *B. subtilis*, and *Listeria monocytogenes*, finding 4 fragmented motility genes in *B. anthracis* Sterne. We expanded on previous surveys, performing a systematic comparison across all fully sequenced *B. cereus* group strains in our study, including the six draft *B. anthracis* strains; comparisons to other *Bacillaceae* strains were performed for a select number of genes.

Using gene names from previous reports (Liu and Ochman, 2007; Klee *et al.*, 2010; Waltman *et al.*, 2010), we found that flagellar genes and a conserved set of chemotaxis genes are organized into two syntenic groups in *B. cereus* strains; note that gene names within these clusters are not always consistent across genome annotations. *motA* and *motB* are syntenic and downstream of a large, conserved set of syntenic genes including outparalogs *motP* and *motS*, all flagellar genes included in the above studies, a number of chemotaxis genes, and other genes which may not be related to motility or are not annotated. Small-scale gene insertions in some strains and chemotaxis genes outside this cluster, which appear to be variable, were not considered. *O. iheyensis*, *B. cereus*, and all Subtilis group members except for *B. pumilis* contain both pairs of paralogs. Geobacillus group strains and *B. pumilis* contain only *motA* and *motB*; the draft genome of *B. pumilis* ATCC 7061 (NZ_ABRX00000000) confirmed lack of *motP* and *motS* in *B. pumilis*. Consistent with their ecological niche, Alkaliphilic group strains contain only *motP* and *motS* (Ito *et al.*, 2004) or bifunctional *motA* and *motB* (Terahara, Krulwich and Ito, 2008). Identity of all homologs was confirmed with high posterior support in phylogenetic analysis using `MrBayes` 3.1.2 (Huelsenbeck and Ronquist, 2001); only the *motA* and *motP* tree is presented here (Figure 2.33).

`FusionFinder` identified all gene fragmentation previously reported (Table 2.10 and Figure 2.34). Note that the fission of *flhF* was mislabelled as *flhH* in (Klee *et al.*, 2010) but not in the genome annotation. Klee *et al.* (2010) reported fission of *motA* and we confirmed by manual inspection that the *motP* paralog is fissioned as predicted by our method. The

Figure 2.33: Unrooted tree of *motA* and *motP* paralogs inferred using `MrBayes`



fission was detected in only *B. anthracis* CDC 684 and *B. anthracis* Vollum using default settings but was found in all *B. anthracis* strains after relaxation of the fragment size cutoff parameter in `FusionScanner`; interestingly this pair is a monophyletic subclade of the A1 lineage. Klee *et al.* (2010) also reported fission of two conserved hypothetical proteins without annotation or COG assignment (fissions 9 and 11 in Table 2.10). Gene BCAH820_1753 (fission 11) was detected in all lineage A and B strains as indicated in the table except the state for *B. anthracis* Sterne was 'unknown'; manual inspection revealed that only one of the fragments was annotated as a protein, which is a situation `FusionFinder` currently does not handle. We found fission of two other conserved hypothetical proteins using `FusionFinder` (fission 10) and manual inspection (fission 3). `FusionFinder` identified a new fission of *flgK* specific to two of three *B. anthracis* lineage A1 strains not part of previous studies. Consistent with previous observations (Liu and Ochman, 2007), we found that *fliC* is present in multiple copies in *B. cereus*, with all strains containing two outparalogs (COG1344), which we label *hag* and *fliC* as previously recommended (Soufiane, Xu and Ct, 2007; Xu and Ct, 2008) and consistent with the annotation of *B. cereus* E33L. Multiple copies of *fliC* exist in some unsequenced *B. thuringiensis* strains (Xu and Ct, 2008) as well as in *B. weihenstephanensis* and *B. cereus* ATCC 10987 from our own analysis. In agreement with a previous report (Soufiane, Xu and Ct, 2007), we confirmed using bootstrap phylogenetic trees that the paralogs in *B. weihenstephanensis* and *B. cereus* ATCC 10987 are more similar to *fliC* than *hag* and that all *B. anthracis* strains in our analysis contain one copy of *fliC* and have lost *hag* (data not shown). Three genes are labelled as *fliN* in *B. cereus* strains; *fliN1* and *fliN2* belong to COG1886 but *fliN3* does not; Interestingly, we were able to find distant homologs for *fliN3* (between 30 and 40% identity) in *Listeria* strains but no significant hits within the rest of the *Bacillaceae*. But our program identified fission of the *fliN1* outparalog and we found from manual inspection that *fliN3* as well as *fliQ* are present but not annotated in the *B. thuringiensis* Al Hakam genome.

Overall, fission within the motility cluster is consistent with the SNP and VNTR data (Van Ert *et al.*, 2007) with one homoplasy within *B. anthracis* strains and four more when *B. cereus* strains are considered (Table 2.10). Gene fission and loss within the motility cluster in *B. cereus* strains was surprising since there is no indication that they are non-motile. Waltman *et al.* (2010) confirmed non-motility of *B. anthracis* Sterne and motility of *B. cereus* ATCC 14579 even though we found fission of *flhA* and *flhF* in *B. cereus* ATCC 14579. Furthermore, Salvetti *et al.* (2007) reported that the fission of *flhF* in *B. cereus* ATCC 14579 is likely a sequencing error although *flhF* inactivation reduced but did not abolish motility. *flhA* inactivation in a *B. thuringiensis* strain resulted in loss of flagella and motility (Ghelardi *et al.*, 2002) so the fission in *B. cereus* ATCC 14579 may also be a sequencing error. *flhB* is part of an export apparatus (Liu and Ochman, 2007) and *fliG* is part of the C ring (Liu and Ochman, 2007) and annotated as being part of the flagellar motor switch in `KEGG`. Whether these genes are not essential to motility or appear fragmented due to sequencing errors is currently unknown.

Assuming all *B. anthracis* strains in all three lineages are non-motile, then the genetic cause may be due to inactivation and loss of up to four genes: *motP*, *motB*, *fliN1*, and *fliC*. *motB* was previously reported to be essential for motility (Mirel, Lustre and Chamberlin, 1992) however a more recent study demonstrated that a combination of *motA* and *motS* is sufficient (Ito *et al.*, 2005) and thus *fliN1* and *fliC* are more likely candidates; there is no detectable homology to *fliC* in the *B. anthracis* genomes and the region contains gene

insertions and deletions within the *B. anthracis* lineage. Ultimately, we found this gene fission pattern of motility- and chemotaxis-related genes is consistent with SNP and VNTR data Pearson *et al.* (2004); Van Ert *et al.* (2007).

Table 2.10: Fission and loss in motility and chemotaxis cluster in *B. cereus*. Genes ordered by location in genome. Locus tags and GI numbers of full-length orthologs in *B. cereus* AH820 as indicated.

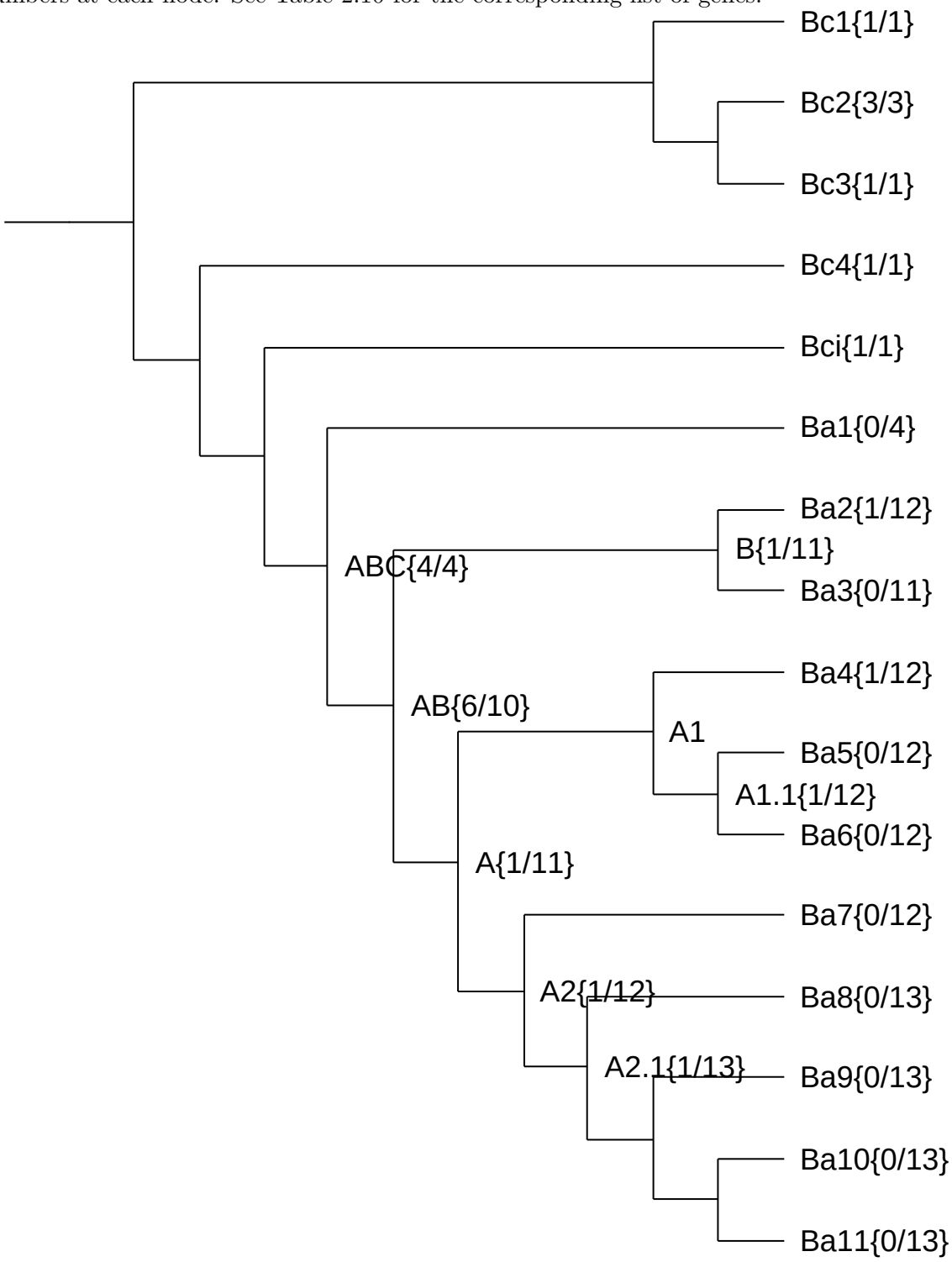
#	Gene	Locus tag	GI	COG	Branch[1]
1	*motP* [3]	BCAH820_1725	218902842	1291	ABC
2	*cheA* [2,3]	BCAH820_1728	218902845	643	AB
3	-[5]	BCAH820_1735	218902852	-	B
4	*flgK*	BCAH820_1736	218902853	1256	A1.1
5	*flgL* [2,3,4]	BCAH820_1737	218902854	1344	AB
6	*fliD* [5]	BCAH820_1738	218902855	1345	Ba4
7	*fliF* [2,3,4]	BCAH820_1744	218902861	1766	AB
8	*fliG*	BCAH820_1745	218902862	1536	Bc3
9	-[3]	BCAH820_1748	218902865	-	A, Ba2
10	-	BCAH820_1749	218902866	-	AB, Bc2, Bci
11	-[3]	BCAH820_1753	218902870	-	AB
12	*cheV* [3,4]	BCAH820_1754	218902871	835	A2.1
13	*fliC* [5,6]	BCAH820_1756	218902873	1344	ABC, Bc1
14	*fliN1* [3]	BCAH820_1759	218902876	1886	ABC
15	*fliM* [2,3,4]	BCAH820_1760	218902877	1868	AB
16	*flhB*	BCAH820_1766	218902883	1377	Bc4
17	*flhA* [5]	BCAH820_1767	218902884	1298	Bc2
18	*flhF* [3]	BCAH820_1768	218902885	4787	A2, Bc2
19	*motB* [4]	BCAH820_4624	218905736	1360	ABC

[1] See Figure 2.34.
[2] Identified by Rasko *et al.* (2004).
[3] Identified by Klee *et al.* (2010).
[4] Identified by Waltman *et al.* (2010).
[5] Identified by manual inspection of `BLAST` results.
[6] Gene lost, not fragmented.

2.4.7 COG analysis of fission in *B. anthracis*

Having demonstrated that fission patterns of motility related genes reflect the *B. anthracis* phylogeny, we now test whether the number of genes fissioned in each COG category deviates from the null model of random fission. If fission events identified by our method are due to sequencing or annotation errors, we expect fissions to be randomly distributed across all COG categories. Deviation from the null model would suggest that certain categories of genes are under negative, adaptive, or relaxed selection. As a baseline, we expect cell motility and cell wall and membrane categories to be overrepresented, due to the known loss of motility in *B. anthracis*, and that informational genes would be underrepresented. We performed Fisher's exact tests of independence and resampling analyses (see Methods) on

Figure 2.34: Phylogeny tracing gene fission and loss in the motility and chemotaxis cluster in *B. cereus*. Number of new and cumulative fissions are indicated by the first and second numbers at each node. See Table 2.10 for the corresponding list of genes.



the pan-genome of the five complete *B. anthracis* genomes, comprising 29066 genes and 380 fissions since NCBI does not provide COG annotations for unannotated draft genomes. Only the two anticipated COG categories differed significantly from the null hypothesis based on the resampling analysis (Table 2.11); "Cell motility" (N) and "Cell wall/membrane/envelope biogenesis" (M) are overrepresented and genes not assigned to any COG (-) are underrepresented. Informational gene categories B and L contained the expected number of fissions while J and K contain less fissions than expected but neither were statistically significant. More fissions were found in defense mechanism genes (V) than expected but the number is not significant after any of the multiple testing or false discovery rate detection corrections. Note that the *B. anthracis* strains do not contain any RNA processing and modification (A) genes and thus that COG was not considered in the calculations. The results obtained in the resampling analysis were the same (data not shown).

Table 2.11: Fisher's exact test for significant deviation from a random distribution of gene fission among COG categories in the pan-genome of five *B. anthracis* genomes. Genes not assigned to a COG were placed in category '-'. raw: unadjusted P-values. Bon: Bonferroni correction. BY: Benjamini & Yekutieli (2001) correction. BH: Benjamini & Hochberg (1995) correction.

COG	No.	No.	Fisher's exact test			
category	genes	split	raw	Bon	BY	BH
-	9003	94	0.009	0.206	0.155	0.041
B	5	0	1.000	1.000	1.000	1.000
C	985	5	0.021	0.509	0.274	0.073
D	213	3	0.759	1.000	1.000	0.959
E	1921	27	0.677	1.000	1.000	0.903
F	619	5	0.368	1.000	1.000	0.680
G	1218	15	1.000	1.000	1.000	1.000
H	833	4	0.029	0.694	0.328	0.087
I	620	6	0.591	1.000	1.000	0.834
J	1052	8	0.127	1.000	1.000	0.278
K	1909	17	0.117	1.000	1.000	0.278
L	826	10	1.000	1.000	1.000	1.000
M	1082	36	0.000	0.000	0.000	0.000
N	256	11	0.001	0.014	0.027	0.007
O	510	4	0.427	1.000	1.000	0.733
P	1121	26	0.005	0.110	0.104	0.027
Q	349	6	0.470	1.000	1.000	0.752
R	2755	35	0.930	1.000	1.000	1.000
S	2070	30	0.546	1.000	1.000	0.819
T	974	16	0.316	1.000	1.000	0.631
U	240	9	0.005	0.109	0.104	0.027
V	499	12	0.043	1.000	0.430	0.114
W	1	1	0.013	0.314	0.197	0.052
Z	5	0	1.000	1.000	1.000	1.000

2.4.8 Fissioned genes in *B. anthracis* may evolve under relaxed selective constraint

Different hypotheses regarding putative fission events may be assessed through comparison of d_N/d_S measured across entire genes prior and subsequent to the event on the phylogeny. If observed gene fragments can be explained by sequencing errors, any synonymous or nonsynonymous substitutions would in theory be localized to short regions in the gene and would not result in a significant difference in the d_N/d_S measured across the entire gene sequence. A compatible hypothesis would be the gene fragments provide an equivalent function relative to the ancestral composite gene, which we posit is uncommon. A decrease in d_N/d_S would suggest the gene fragments are under negative selection and encode products that provide a higher fitness advantage compared to the ancestral form. If gene fission in *B. anthracis* is leading to pseudogenization, we expect a relaxation of selective constraint across the entire component gene set, detectable by an increase in d_N/d_S to a value not significantly greater than 1. Such an observation is also compatible with a decrease in the efficacy of purifying selection, which is expected since *B. anthracis* likely has a smaller effective population size compared to *B. cereus* due to its pathogenic lifestyle (Ochman and Davalos, 2006). Thus we need to tease apart the contributions from these two factors, if any.

We started with 68 fission groups (2 of 70 contained less than 3 sequences and were discarded) mapped to the *B. anthracis* cenancestor (Figure 2.30) and 4571 orthologous groups (see Methods). For each fission and orthologous group, we estimated d_N/d_S ratios for *B. cereus* genomes (ω_1) and *B. anthracis* genomes (ω_2) as described in the methods (Figure 2.35; Table 2.12). The dataset was reduced to 35 fission and 1887 orthologous groups after filtering based on d_S (see Methods). For both orthologous and fission groups, ω_2 is higher than ω_1 ($P = 2.52 \times 10^{-11}$, $P = 4.50 \times 10^{-2}$). We also find that d_N/d_S ratios are significantly higher in fission groups compared to orthologous groups along both the *B. cereus* and *B. anthracis* portions of the phylogeny ($P = 7.57 \times 10^{-3}$, $P = 4.12 \times 10^{-6}$). Thus, it appears genes are generally under relaxed selection in *B. anthracis* compared to *B. cereus*, gene fission in *B. anthracis* leads to a further relaxation of selective constraint, and genes that are fissioned in *B. anthracis* might have been less essential even prior to the fission event. These results should be treated with caution as the size of the fission and orthologous groups may affect the results (Figure 2.36).

Table 2.12: Mean d_N/d_S estimates for genes in orthologous or fission groups for *B. cereus* and *B. anthracis* genomes

Group	ω_1	ω_2
Orthologous	0.1005	0.1228
Fission	0.1324	0.2805

Both estimates of mean d_N/d_S (Table 2.12) are much higher than the previously reported value of 0.045 between *B. anthracis* Ames and *B. cereus* ATCC 14579 (Kuo, Moran and Ochman, 2009) but the discrepancy is likely due to their use of RBH for ortholog identification as opposed to `OrthoMCL`. The authors also reported a mean d_N/d_S of about 0.036 between *S. enterica* and *E. coli* but a more recent study using `OrthoMCL` reported mean

Figure 2.35: d_N/d_S estimates for *B. cereus* genomes (ω_1/wB) and *B. anthracis* genomes (ω_2/wF) for orthologous groups and fission groups.

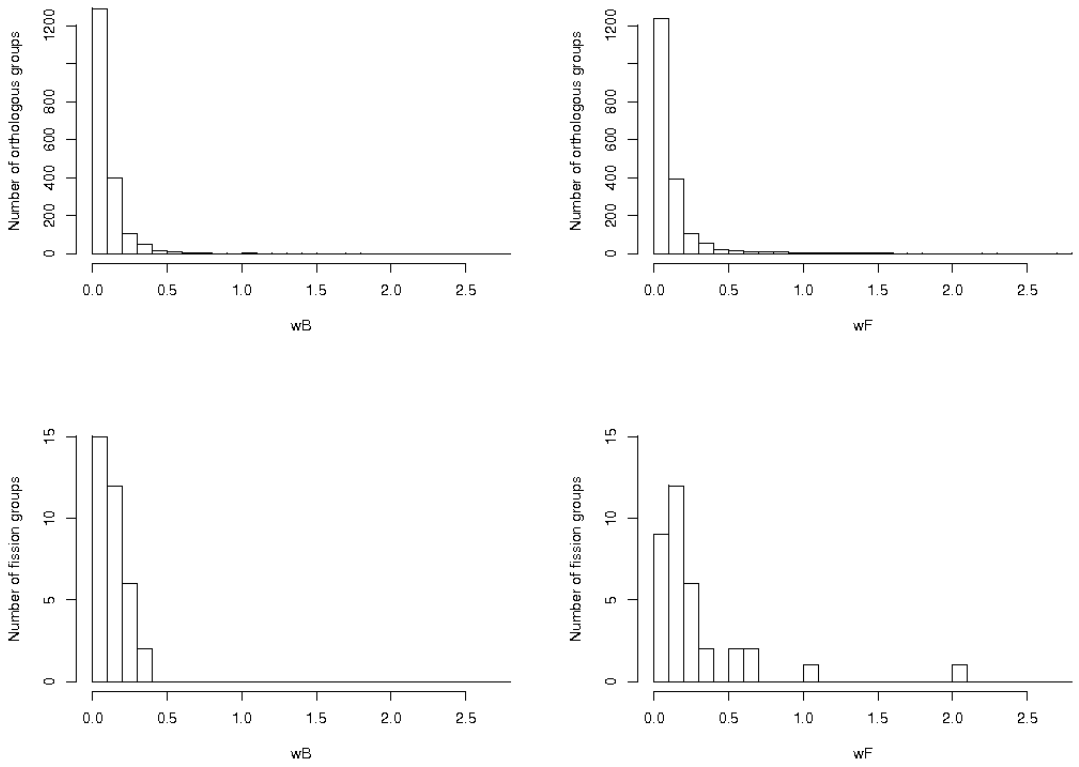
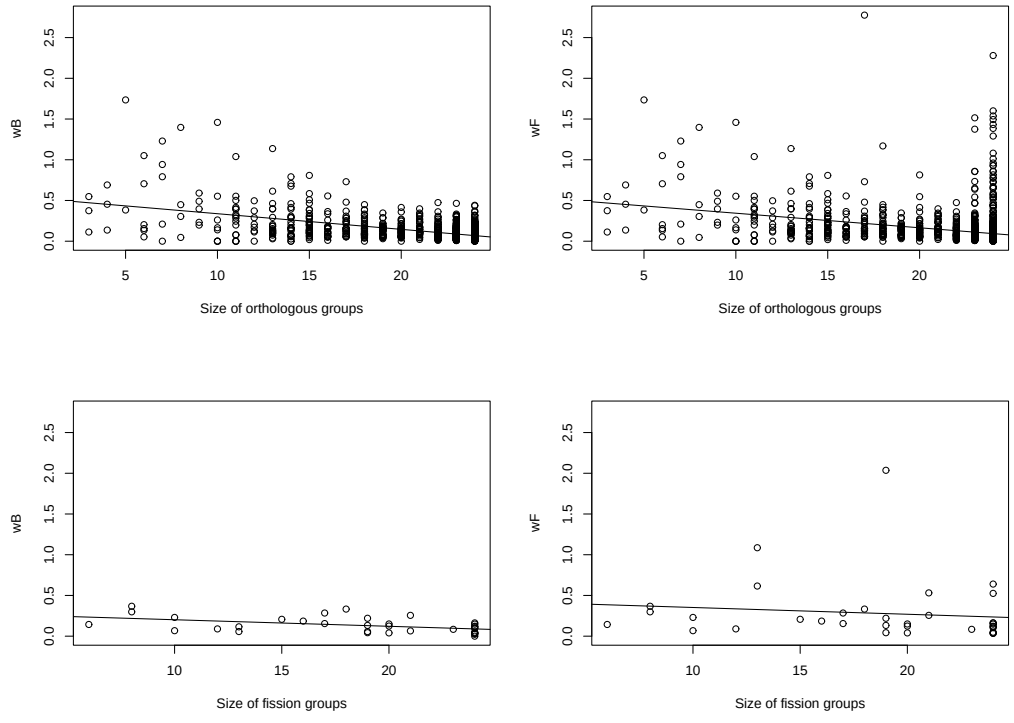


Figure 2.36: Relationship between d_N/d_S and the number of sequences.



d_N/d_S between *S. enterica* genome pairs to be about 0.1 (Kuo and Ochman, 2010), which compares favorably to our estimates (Table 2.12).

A two-ratios model may not fit the underlying sequence alignment better than a one-ratio model. For each group, we performed a likelihood ratio test (LRT) to determine whether the two-ratios model is significantly better than the one-ratio model (Table 2.13). Out of 35 fission groups, 5 (14.3%) were better explained by the two-ratios model although two genes had a d_N/d_S greater than 1. This proportion is significantly higher than 47 of 1887 orthologous groups (2.5%) based on Fisher's exact test of independence ($P = 0.002$).

Table 2.13: Fusion and orthologous groups categorized by d_N/d_S and the d_N/d_S in *B. anthracis* genomes (ω_2) compared to *B. cereus* genomes (ω_1).

Category	Fission	Ortholog
$\omega_1 > 1$	0	7
$\omega_2 > 1$	2	13
$\omega_2 > \omega_1$	5	47
$\omega_2 = \omega_1$	28	1809
$\omega_2 < \omega_1$	0	1
Total	35	1877

2.4.9 Origin of *B. cereus* CI: lateral or vertical?

Closely related *B. cereus* group strains containing both virulence plasmids in *B. anthracis* were recently isolated although it was unclear whether they share common ancestry with *B. anthracis* or evolved independently after lateral acquisition of the plasmids (Klee *et al.*, 2006). Klee *et al.* (2010) sequenced the *B. cereus* CI strain and concluded the data suggests it evolved from a *B. cereus* strain independent of *B. anthracis* because *B. cereus* CI shares more orthologous genes with *B. cereus* E33L and *B. thuringiensis* konkukian than *B. anthracis* strains identified using '`BiBlast`', has an intact motility and chemotaxis cluster, has a putative deleterious mutation in *plcR* different from the one found in *B. anthracis*, and contains three plasmids, including both virulence plasmids in *B. anthracis*. However, the provided phylogeny on a subset of *B. cereus* genomes suggests that *B. cereus* CI may be more similar to *B. anthracis* than any other *B. cereus* strain currently sequenced and we argue that the points noted by Klee *et al.* (2010) do not conclusively show that *B. cereus* CI and *B. anthracis* evolved independently. If *B. cereus* CI and *B. anthracis* shared a common ancestor, the two resulting lineages may have accumulated different sets of inactivated genes due to random process or natural selection due to different environmental conditions; the authors note that *B. cereus* CI has been isolated from atypical environments compared to *B. anthracis*. Different mutations in *plcR* can still be explained by convergent evolution under this alternate hypothesis. Thus, we believe the strongest point in their argument for independent evolution is the number of shared orthologs, although that number may be influenced by the algorithm and parameters chosen. Perhaps more importantly, we demonstrated that many genes have been and continue to be fissioned in *B. anthracis* and the genome annotations differ, both of which may lower the number of orthologs identified. The

authors analyzed a smaller set of *B. cereus* group genomes compared to our dataset, including only *B. anthracis* Ames, *B. anthracis* Ames Ancestor, and *B. anthracis* Sterne from the *B. anthracis* lineage. Thus, we sought to compare *B. cereus* CI against each genome in our dataset.

Using a RBH approach with an Expect-value cutoff of 10^{-20}, we found that all *B. anthracis* genomes except for *B. anthracis* A0248, *B. anthracis* Ames, and *B. anthracis* Ames Ancestor (Ba9–11) contain more orthologs than *B. cereus* E33L and *B. thuringiensis* konkukian (Bc7,Bt2) as shown in Table 2.14; interestingly these three strains form a monophyletic clade in the A2 lineage according to the SNP data (Figure 2.26) and are the only genomes with annotated pseudogenes (Tables 2.5, 2.6). We also found that *B. cereus* AH820 (Bc9) contains more orthologs than all other *B. cereus* strains as expected given our phylogeny (Figure 2.20) and also all *B. anthracis* strains expect for *B. anthracis* CDC 684 (Ba6), which contains the highest number component sets annotated as proteins (Table 2.6). The number of orthologs identified decreased more rapidly for *B. anthracis* strains when the minimum proportional match length threshold passes 0.5–0.6 (Figure 2.38) even though it decreased at similar rates as the minimum protein identity increased (Figure 2.37), which is consistent with presence of fragmented genes; this study only considered pseudogene formation through gene fission but truncation is another possibility (Hao and Golding, 2008a). Since LGT has a major influence on the evolution of the *Bacillaceae*, the number of orthologs may not be a good metric for genome similarity. We chose protein identity between orthologs as an alternate metric although it may be influenced by genome-wide relaxation of selective constraint in *B. anthracis*. The average identity to *B. anthracis* orthologs decreases less rapidly than to *B. cereus* orthologs as the minimum identity threshold is decreased (Figure 2.39), which could be explained by a slight downward shift in percent identity in *B. anthracis* orthologs due to genome-wide relaxation of selective constraint, although further analyses should be performed.

Whole-genome similarity was also chosen as an alternate metric and was computed using `MUMmer v3.22` (Kurtz *et al.*, 2004). The `mummer` algorithm finds maximal unique matches (MUMs) and `run-mummer3` is a pipeline that extends MUMs into longer alignments using a seed and extend approach. Genome similarity was defined as the total number of nucleotides covered by MUMs in a pairwise genome comparison given a predefined minimum MUM size (Table 2.14). Depending on the minimum MUM size, *B. cereus* CI is either more similar to *B. cereus* AH820 or *B. anthracis* strains which is consistent with the RBH data. Thus, the choice of *B. anthracis* strain and even *B. cereus* strain can affect genome comparisons. The number of orthologs and genome alignments should be interpreted with caution as our data suggest ongoing pseudogenization and a genome-wide relaxation of selective constraint. Thus, *B. cereus* CI appears to lie at the boundary between our expanded set of *B. cereus* and *B. anthracis* strains, although the data cannot confirm the nature of the *B. cereus* CI ancestor. RBH and `MUMmer` results with respect to *B. cereus* AH820 (Table 2.15) and *B. anthracis* A1055 (Table 2.16) suggest that both are more similar to *B. anthracis* than other *B. cereus* strains.

Bootstrapped fusion phylogenies group *B. cereus* CI beside the *B. anthracis* lineage albeit with low support (Figures 2.31,2.32). `FusionMapper` predicted 67 fissions unique to *B. cereus* CI and 21 to the putative ancestor of *B. cereus* CI and *B. anthracis* (Figure 2.30). Manual inspection revealed that 10 events have components in either *B. anthracis* or *B. cereus* CI and no identified homolog in the other. Of 8 events with components in both, only one appears to

Figure 2.37: Effect of increasing minimum protein identity on number of RBH orthologs between *B. cereus* CI and a select number of *B. cereus* and *B. anthracis* strains.

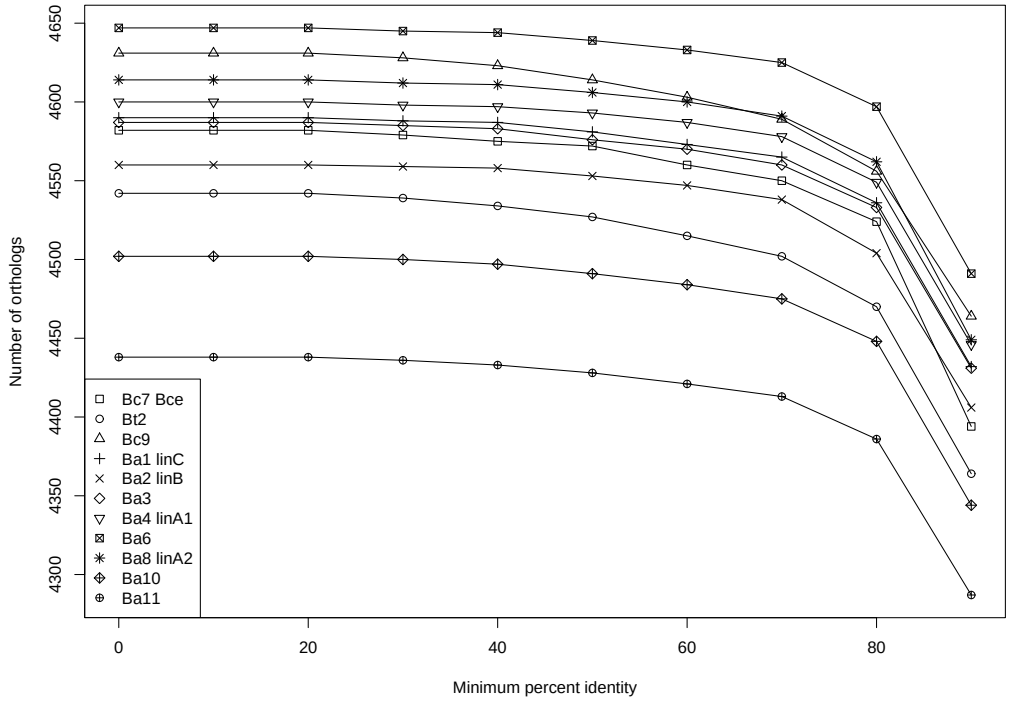


Figure 2.38: Effect of increasing proportional match length on number of RBH orthologs between *B. cereus* CI and a select number of *B. cereus* and *B. anthracis* strains. Left: no protein identity cutoff. Right: minimum 90% protein identity.

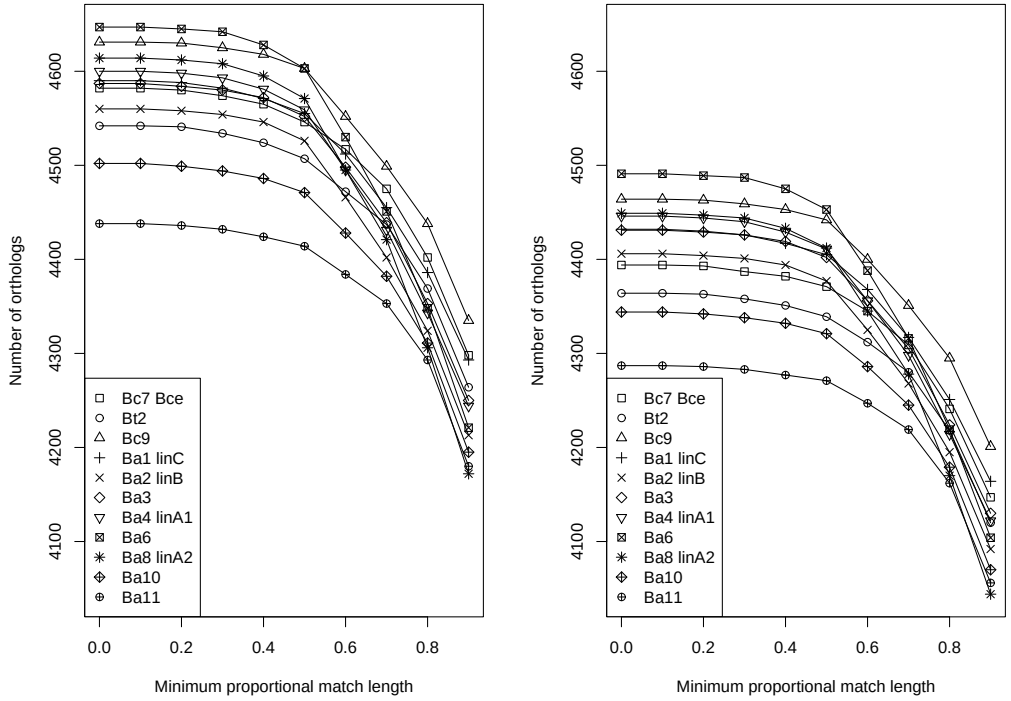


Table 2.14: Pairwise genome analysis of *B. cereus* CI. 'mummer' is the total length of all MUMs identified. 'mummer3' the total length of all alignments identified using run-mummer3. 'l' is the minimum match length parameter for 'mummer' and 'run-mummer3'. Orthologs were identified using RBH. The highest number in each column is underlined. The top two *B. cereus* and all three *B. anthracis* strains used by Klee *et al.* (2010) are bolded.

ID	l=20		l=2000		No.	Avg.
	mummer	mummer3	mummer	mummer3	orthologs	identity
Bcy	508842	569183	0	0	3022	83.68
Bw	1860550	1926699	0	0	4157	91.54
Bc1	2356700	2430891	0	0	4274	93.52
Bc2	2463190	2536229	2161	2161	4227	93.75
Bc3	2465866	2541968	2311	2311	4340	93.78
Bc4	3161044	3240473	4597	4597	4262	95.50
Bc5	3521360	3613213	0	0	4417	96.18
Bc6	3522762	3612547	0	0	4367	96.26
Bc7	4251167	4326010	14228	14228	**4582**	97.59
Bt1	4282089	4353678	66003	66003	4282	97.73
Bc8	4291223	4365268	71517	75887	4612	97.81
Bt2	4322153	4409606	76126	76126	**4542**	97.82
Bc9	4407292	4490079	88987	99703	4631	97.91
Bci	n/a	n/a	n/a	n/a	n/a	n/a
Ba1	4391317	4434314	60751	60751	4590	98.05
Ba2	4368739	4459052	45172	45172	4560	98.02
Ba3	4403536	4512328	49385	53755	4587	98.01
Ba4	4389624	4526526	45972	45972	4600	98.03
Ba5	4391078	4476876	41611	41611	4604	98.04
Ba6	4406131	4512446	56934	63031	4647	98.04
Ba7	4373398	4483346	42469	42469	4604	98.02
Ba8	4414970	4507729	56091	62188	**4614**	97.98
Ba9	4414748	4507191	56091	62188	4388	98.07
Ba10	4413167	4506952	53841	59938	**4502**	97.99
Ba11	4414748	4507191	56091	62188	**4438**	98.03

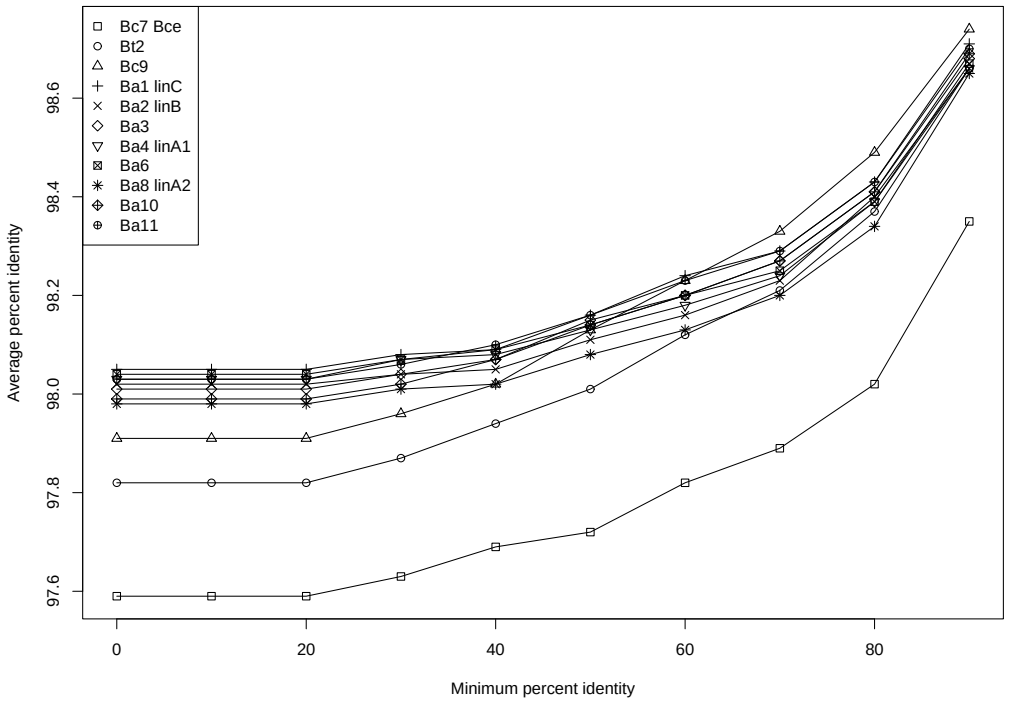
Table 2.15: Pairwise genome analysis of *B. cereus* AH820. 'mummer' is the total length of all MUMs identified. 'mummer3' the total length of all alignments identified using run-mummer3. 'l' is the minimum match length parameter for 'mummer' and 'run-mummer3'. Orthologs were identified using RBH. The highest number in each column is underlined.

ID	l=20		l=2000		No.	Avg.
	mummer	mummer3	mummer	mummer3	orthologs	identity
Bcy	508115	572548	0	0	3034	83.48
Bw	1877247	1948184	0	0	4209	91.40
Bc1	2390090	2468045	0	0	4346	93.27
Bc2	2467177	2550415	2161	10805	4258	93.61
Bc3	2472156	2553290	0	0	4375	93.68
Bc4	3210273	3312328	42772	62958	4316	95.50
Bc5	3469745	3567987	13151	13151	4435	96.03
Bc6	3476239	3560627	6080	6080	4365	96.18
Bc7	4259932	4353543	38999	53214	4553	97.63
Bt1	4349253	4445489	145958	168980	4274	97.91
Bc8	4298161	4402172	112044	134854	4630	97.71
Bt2	4473876	4592171	346539	355600	4568	98.30
Bc9	n/a	n/a	n/a	n/a	n/a	n/a
Bci	4407292	4552377	88987	100173	4631	97.91
Ba1	4684534	4764459	<u>475372</u>	<u>485840</u>	4721	98.65
Ba2	4645854	4733348	343735	343735	4684	98.60
Ba3	4685317	4783011	422332	427754	4707	98.63
Ba4	4671872	4813372	362457	386378	4725	98.65
Ba5	4683631	4769691	386921	398670	4729	98.64
Ba6	4699009	<u>4833650</u>	400282	419511	<u>4804</u>	98.65
Ba7	4659006	4772518	344705	356212	4727	98.64
Ba8	<u>4708591</u>	4815596	374154	382720	4703	98.62
Ba9	4708331	4814189	380021	388587	4524	98.70
Ba10	4704331	4814550	373761	373761	4616	98.67
Ba11	4708331	4814189	380021	388587	4553	<u>98.68</u>

Table 2.16: Pairwise genome analysis of *B. anthracis* A1055. 'mummer' is the total length of all MUMs identified. 'mummer3' the total length of all alignments identified using run-mummer3. 'l' is the minimum match length parameter for 'mummer' and 'run-mummer3'. Orthologs were identified using RBH. The highest number in each column is underlined.

ID	l=20		l=2000		No.	Avg.
	mummer	mummer3	mummer	mummer3	orthologs	identity
Bcy	501672	568606	0	0	3028	83.47
Bw	1872606	1953881	0	0	4190	91.48
Bc1	2351270	2450569	0	0	4268	93.19
Bc2	2447947	2550357	0	0	4222	93.53
Bc3	2448082	2541845	0	0	4324	93.54
Bc4	3123710	3224263	5284	9941	4231	95.37
Bc5	3461078	3589091	0	0	4398	95.96
Bc6	3442023	3544987	0	0	4302	96.10
Bc7	4273624	4381840	15494	26620	4552	97.58
Bt1	4266676	4371863	67441	83357	4213	97.83
Bc8	4263630	4376576	66427	79821	4577	97.68
Bt2	4337353	4446809	146798	153626	4484	97.95
Bc9	4684534	4805579	475372	475372	4721	98.65
Bci	4391317	4541145	60751	73804	4590	98.05
Ba1	n/a	n/a	n/a	n/a	n/a	n/a
Ba2	5245224	5399714	3543321	3584649	5168	99.82
Ba3	5286162	5438599	<u>3982919</u>	<u>4033254</u>	5215	<u>99.85</u>
Ba4	<u>5299984</u>	<u>5486129</u>	3821848	3902237	5241	99.83
Ba5	5288180	5420466	3832981	3892777	<u>5245</u>	99.84
Ba6	5222604	5382751	3782863	3856533	5102	99.82
Ba7	5277208	5426183	3735441	3782017	5238	99.83
Ba8	5224562	5378640	3783545	3850200	5020	99.79
Ba9	5220129	5373863	3767341	3833996	4764	99.82
Ba10	5212079	5371239	3752337	3807866	4943	99.80
Ba11	5220129	5373863	3767341	3833996	4864	99.82

Figure 2.39: Effect of increasing minimum protein identity on average protein identity of RBH orthologs between *B. cereus* CI and a select number of *B. cereus* and *B. anthracis* strains.



have resulted from a single event and it was located on the pXO1 plasmid (*B. cereus* 03BB102 homolog GI: 225871522); one hit was a false positive and the others appear to have arisen from independent events. Note that the plasmid phylogeny may be different from the chromosomal phylogeny so it is possible that there was instead a fusion in the *B. cereus* 03BB102 homolog. We also found three fission groups with components in *B. anthracis*, *B. cereus* CI, and *B. cereus* AH820; in one case the N-terminal end was not annotated and discovered by manual inspection (appears to be a difference in gene model as there are no frameshifts or in-frame stop codons). In two of these, the *B. anthracis* and *B. cereus* CI components appear to be more similar (three common insertions; common IS605 insertion) but the other is due to a gene insertion common to *B. anthracis* and *B. cereus* AH820 not found in *B. cereus* CI. Ultimately, it appears that most of the fissions shared between *B. anthracis* and *B. cereus* CI do not have common ancestry. They could have arisen due to convergent evolution, random chance, or sequencing errors; in each of the 11 fission groups mentioned above, we aligned a small set of sequences and identified 23 differences, 6 of which were insertions or deletions from single base repeats of lengths 4 to 9. Further analysis and perhaps further sampling of *B. cereus* strains will be required to resolve the phylogenetic history of *B. cereus* CI.

2.4.10 Fission and SNP patterns are consistent in *Yersinia*

Y. pestis is attractive as a control for our *B. anthracis* results due to similarities between these two bacteria. Estimated time of origin for *B. anthracis* from a *B. cereus* ancestor is between 12000 and 26000 years ago (Van Ert *et al.*, 2007) and *Y. pestis* from a *Y. pseudotuberculosis* ancestor is between 2603 and 28646 years ago (Morelli *et al.*, 2010). *Y. pestis* also lost motility and has been previously suggested to have undergone reductive genome evolution.

We expect to find fewer fission events in the *Yersinia* phylogeny due to gene loss from IS-mediated genomic events. Non-contradictory phylogenies have been inferred using SNPs (Chain *et al.*, 2006; Eppinger *et al.*, 2010; Morelli *et al.*, 2010) and genomic rearrangements (Darling, Mikls and Ragan, 2008), with the exception of the branching order of *Y. pestis* Angola and *Y. pestis* 91001 (Eppinger *et al.*, 2010; Morelli *et al.*, 2010). In all analyses we used *Y. enterocolitica* 8081 (Yen) and *Y. pseudotuberculosis* IP 32953 (Yps) as outgroups. Many pseudogenes have been annotated in *Y. pestis* genomes and must be included in the analyses otherwise there is insufficient data. We first applied `FusionTree` to the 5 strains used by Chain *et al.* (2006) and obtained a rooted fusion/fission phylogeny for *Y. pestis* consistent with their SNP tree although *Y. enterocolitica* 8081 and *Y. pseudotuberculosis* IP 32953 are grouped together with low bootstrap support (data not shown).

Eppinger *et al.* (2010) increased the number of strains by adding two complete genomes and two draft genomes. For this expanded set of strains, `FusionTree` produced a phylogeny that differed in two areas compared to the SNP tree (Figures 2.40,2.41). First, we expected *Y. pestis* FV-1 (FV1) to be the ancestral strain in the 1.ORI clade. The disagreement within the 1.ORI lineage could be due to the inclusion of draft genomes ORI1a and FV1 or be influenced by a true accelerated rate of gene fission in FV1. Interestingly, there appears to be a consistent positive correlation between the number of fissions and branch length in the SNP tree for all strains except for FV1 (Eppinger *et al.*, 2010) and the bootstrap support for this node is low. Furthermore, Morelli *et al.* (2010) suspected that the FV1 genome contains many sequencing errors and excluded it from their analysis. Thus, we attribute this incongruence to the inclusion of FV1; the only real difference between the fission and SNP phylogenies is the placement of the ancestral strains *Y. pestis* Angola (0.PE3) and *Y. pestis* 91001 (0.PE4); we expected 0.PE3 to be the ancestral strain.

Finally, we analyzed the dataset from Morelli *et al.* (2010) with a few exceptions. We excluded *Y. pestis* IP674 (ERA000177) from our analysis but included the complete *Y. pestis* Nepal516 (NC_008149) and draft *Y. pestis* Nepal516A (NZ_ACNQ01000000) genomes, which are independent isolates according to their genome annotations; Morelli *et al.* (2010) used only *Y. pestis* Nepal516A in their analysis. In total, we generated a fusion/fission phylogeny for 16 *Y. pestis* strains (Figures 2.42,2.43). Even with the expanded dataset there is high bootstrap support for *Y. pestis* Angola and *Y. pestis* 91001 (Eppinger *et al.*, 2010; Morelli *et al.*, 2010). We note that that the branching order in the SNP phylogeny was based only on four SNPs (Morelli *et al.*, 2010) so our inability to separate these two strains based on fission patterns alone is not at all surprising. It is also possible that a set of fissioned genes in 0.PE3 were retained by 0.PE4 but lost in the other lineage after the split. The only other difference was the placement of *Y. pestis* B42003004 (0.ANT2) into Branch 2 although there is weak bootstrap support for this node; the SNP tree places *Y. pestis* B42003004 immediately ancestral to Branch 1 and Branch 2. We note that the distance between *Y. pestis* B42003004 and the Branch 1/2 split is very short (78–856 years). Aside from these two minor differences, the fission and SNP phylogenies are consistent although with varying degrees of bootstrap support. Increasing the sequence identity cutoff to 90% `FusionScanner` did not have a large effect which is not surprising because average protein identity of RBH within and between *Y. pestis* and *Y. pseudotuberculosis* is greater than 99%.

Compared to *B. anthracis*, the number of fissions predicted is similar but fewer are clade-specific and many are concentrated in only a few genomes (Table 2.17); the number of component sets with RBH is also high. In *Y. pestis* CO92 (1.ORI1b), the first *Y. pestis*

Figure 2.40: Consensus of phylogenies inferred from fusion/fission patterns for 9 *Y. pestis* genomes (Eppinger *et al.*, 2010) using Camin-Sokal parsimony. The proportion of 1000 bootstrap samples supporting each partition in the tree is indicated.

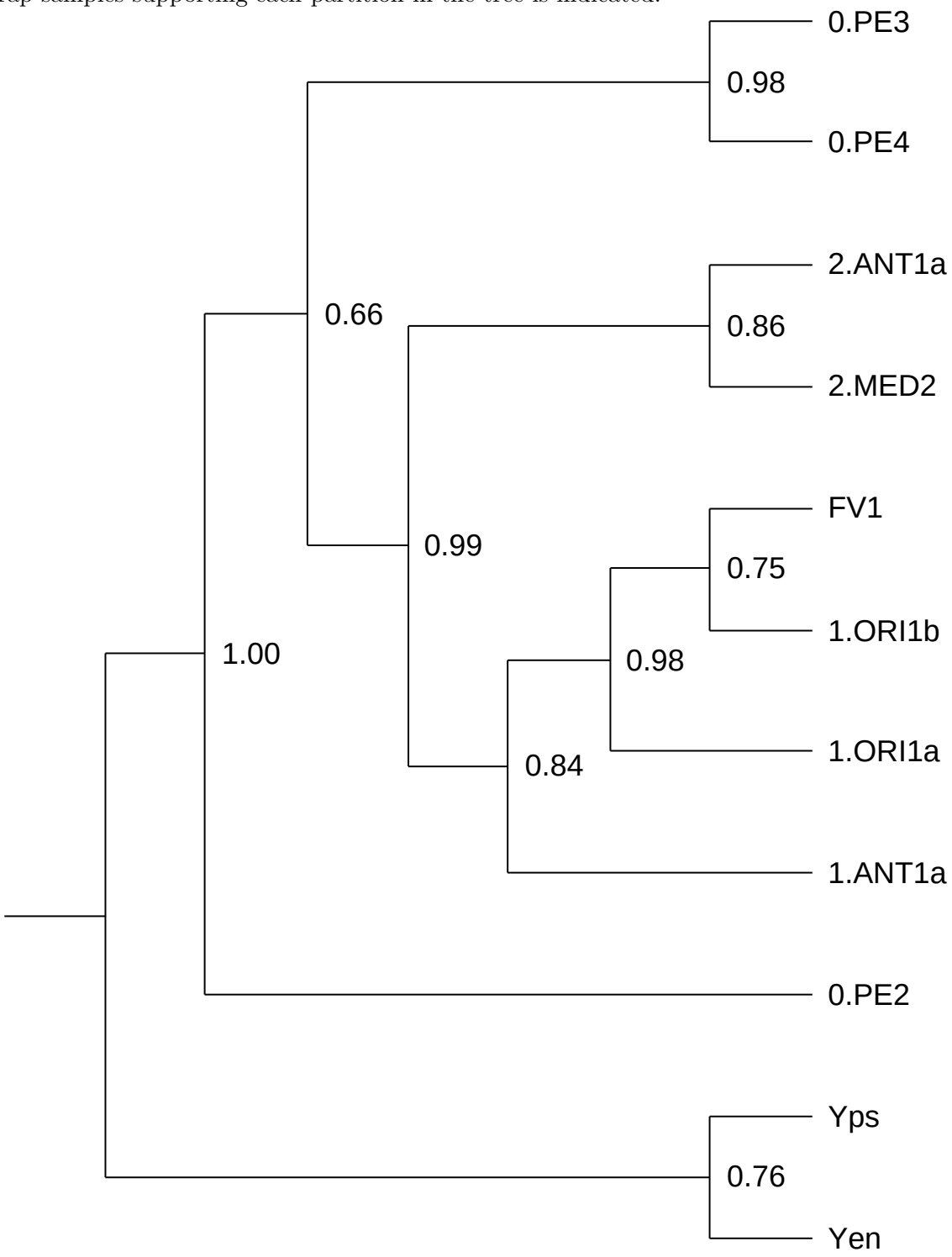


Figure 2.41: Fusion/fission analysis of 9 *Y. pestis* genomes (Eppinger *et al.*, 2010). Number of fusions and fissions are indicated by the first and second numbers on each node. The numbers at the root of the tree are the total number of events across the phylogeny.

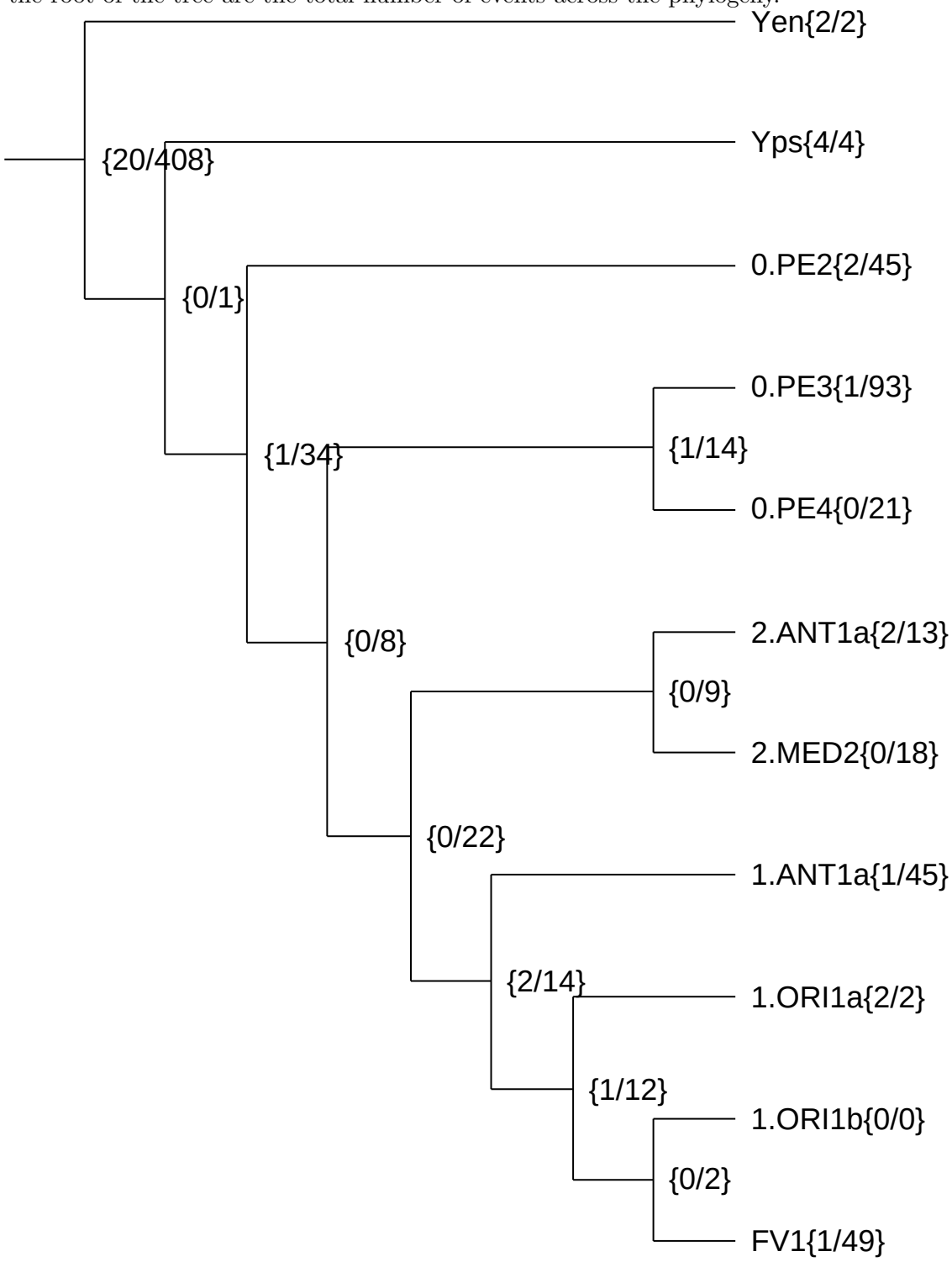


Figure 2.42: Consensus of phylogenies inferred from fusion/fission patterns for 16 *Y. pestis* genomes (Morelli *et al.*, 2010) using Camin-Sokal parsimony. The proportion of 1000 bootstrap samples supporting each partition in the tree is indicated.

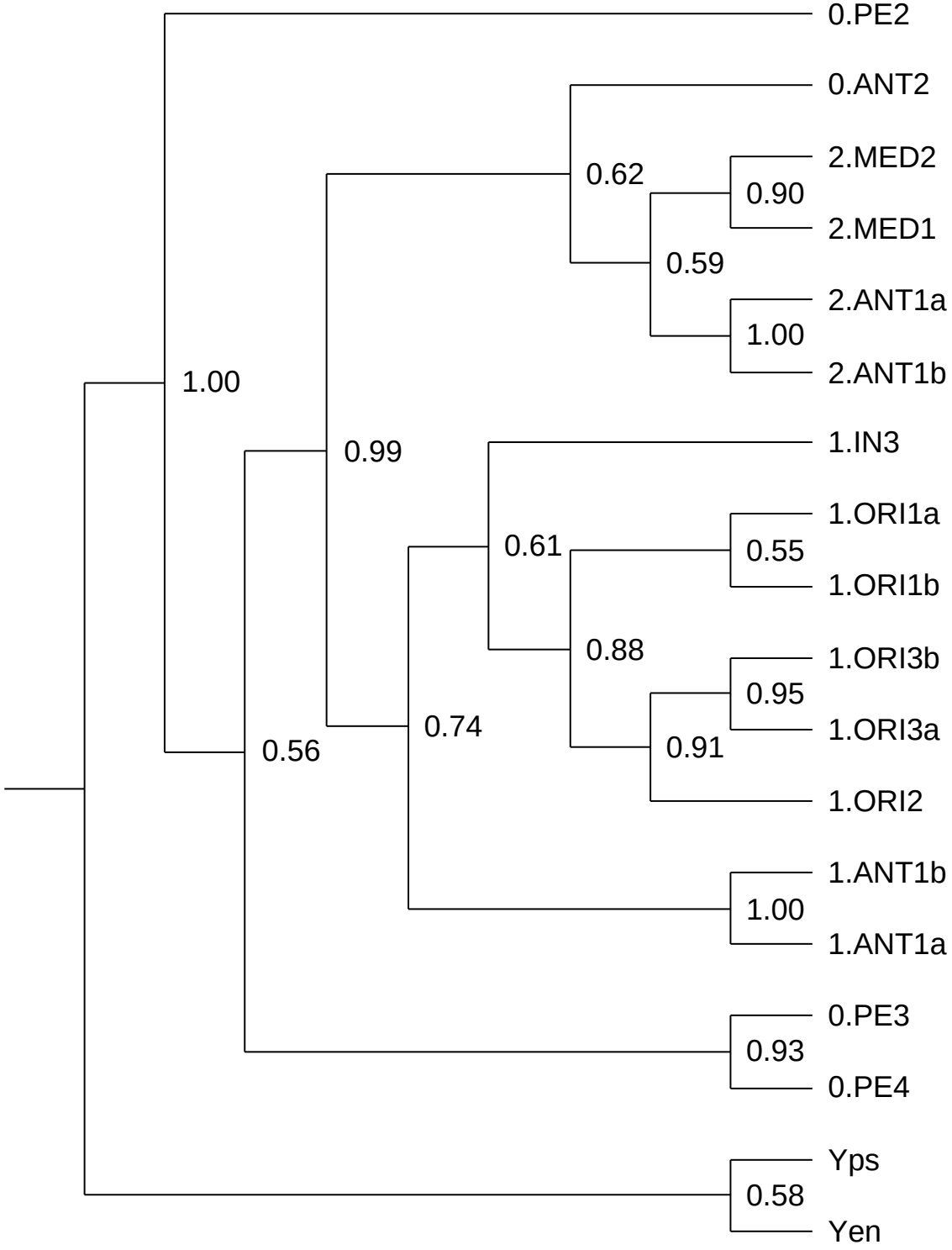
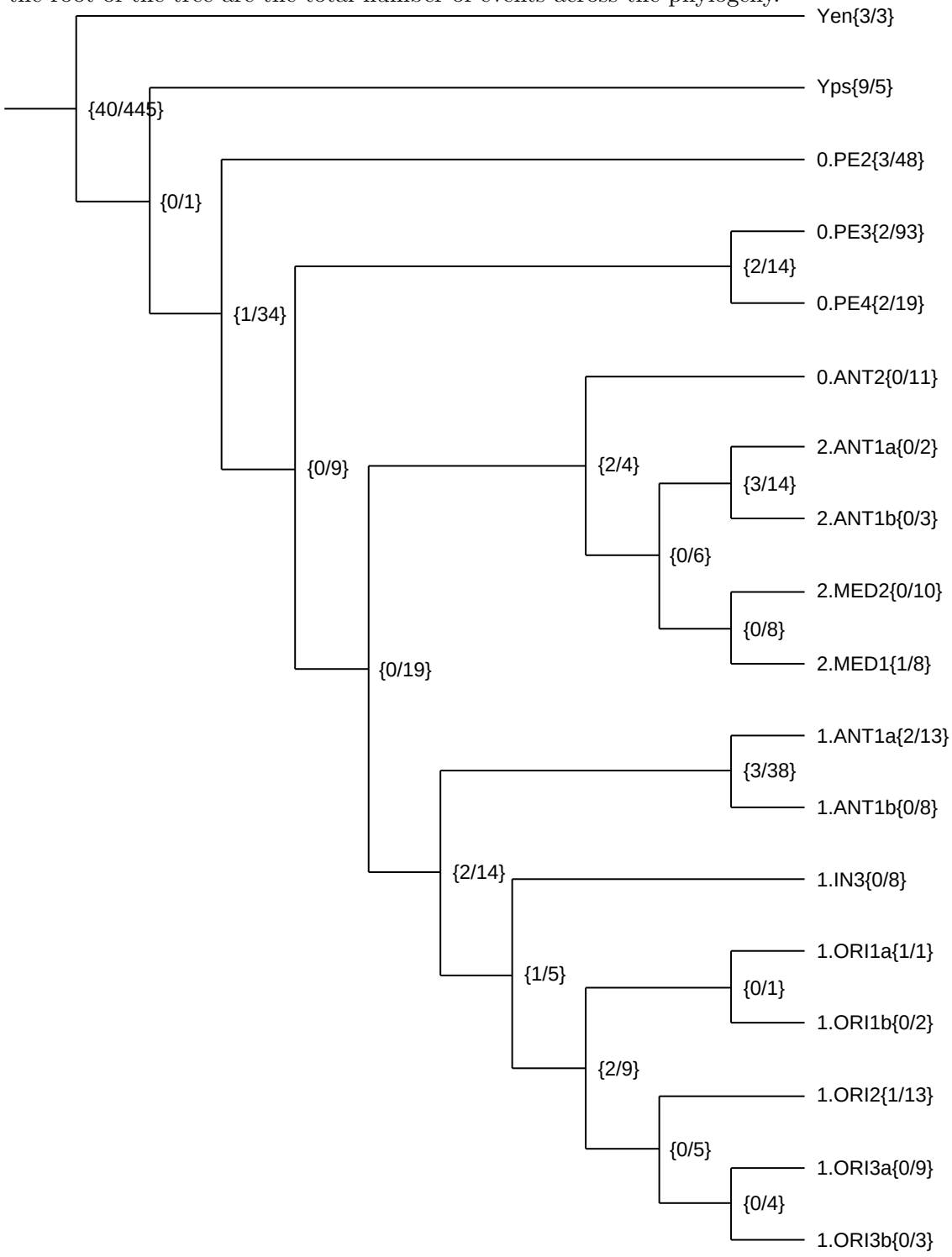


Figure 2.43: Fusion/fission analysis of 16 *Y. pestis* genomes (Morelli *et al.*, 2010). Number of fusions and fissions are indicated by the first and second numbers on each node. The numbers at the root of the tree are the total number of events across the phylogeny.



genome sequenced, there are 107 pseudogenes annotated on the chromosome and 7 across three plasmids. `FusionFinder` predicted a total of of 93 fission events, 56 of which are annotated as pseudogenes. Interestingly, none are specific to *Y. pestis* CO92 and 34 were mapped to the *Y. pestis* cenancestor (Figure 2.43,Table 2.18). This number is higher than many of the branches in the phylogeny, similar to our observations in *B. anthracis*. In *Y. pestis* CO92, 24 were annotated as pseudogenes, only 2 were annotated as proteins, and 3 were unidentified. Further investigation of the latter three fusion groups revealed that orthologous regions exist in each case. The first ortholog (*Y. pseudotuberculosis* IP 32953 GI: 51598187) had an internal deletion of 200 amino acids that was detected in genomes where the components were annotated as proteins but not in genomes where they were annotated as pseudogenes such as CO92. The second ortholog (*Y. pseudotuberculosis* IP 32953 GI: 51594660) was fissioned in all genomes except for the monophyletic 1.IN3/1.ORI clade consisting of 6 genomes and was not detected due to the loss of the N-terminal fragment of about 60 amino acids. The third ortholog (*Y. pseudotuberculosis* IP 32953 GI: 51598075) was also due to a difference in genome annotations. In some genomes including CO92, only the N-terminal protein was annotated (413 aa) whereas in other genomes both the N-terminal and C-terminal (17 aa) fragments are annotated as pseudogenes. Regardless of annotation, all *Y. pestis* genomes share a single base deletion from a string of 'G' leading to elongation of the coding frame. Although the annotation reports this to be a possible sequencing error, we note that all *Y. pestis* genomes share this change. Thus, in all cases the fission was correctly inferred to be common to all *Y. pestis* genomes although the reason for missing component sets differs in each instance. Ultimately, we find similarities to *B. anthracis* in that a number of fissions common to all isolates and fission patterns are broadly consistent with SNPs even though fissions are fewer in number. This was surprising due to the large amount of genomic rearrangements between *Y. pestis* isolates (Darling, Mikls and Ragan, 2008).

2.4.11 Pseudogenization in *S. enterica*

S. enterica Typhi is genetically monomorphic and contains a large complement of pseudogenes like *B. anthracis* and *Y. pestis* (Parkhill *et al.*, 2001; Deng *et al.*, 2003; Achtman, 2008) but the host-promiscuous *S. enterica* Typhimurium contains few pseudogenes (McClelland *et al.*, 2001). Furthermore, there have been multiple transitions from host-promiscuous to host-restricted strains within *S. enterica* associated with pseudogene accumulation (Chiu *et al.*, 2005; Thomson *et al.*, 2008). A set of these genomes have also been independently analyzed specifically for pseudogenes (Kuo and Ochman, 2010). All of these factors made *S. enterica* an attractive dataset to apply our pipeline to.

We applied `FusionFinder` to the same dataset used by Kuo and Ochman (2010) and found similar numbers of pseudogenes even though our pipeline is completely automatic and designed for fusion and fission analysis (Figure 2.44). We confirmed their observation that most pseudogenes are strain-specific. The authors also found a high but not significant negative correlation between the genome-wide K_a/K_s ratio and the number of pseudogenes identified which was argued as an indication of selective constraint.

However, we noticed that the number of fissioned genes is also reflected by host-specificity and tissue adaptation (Table 2.3). Thus, we expected and observed that additional sampling of host-restricted strains would reveal many shared gene fissions whereas additional host-promiscuous strains would reveal few shared fissions (Figure 2.45); branching order of

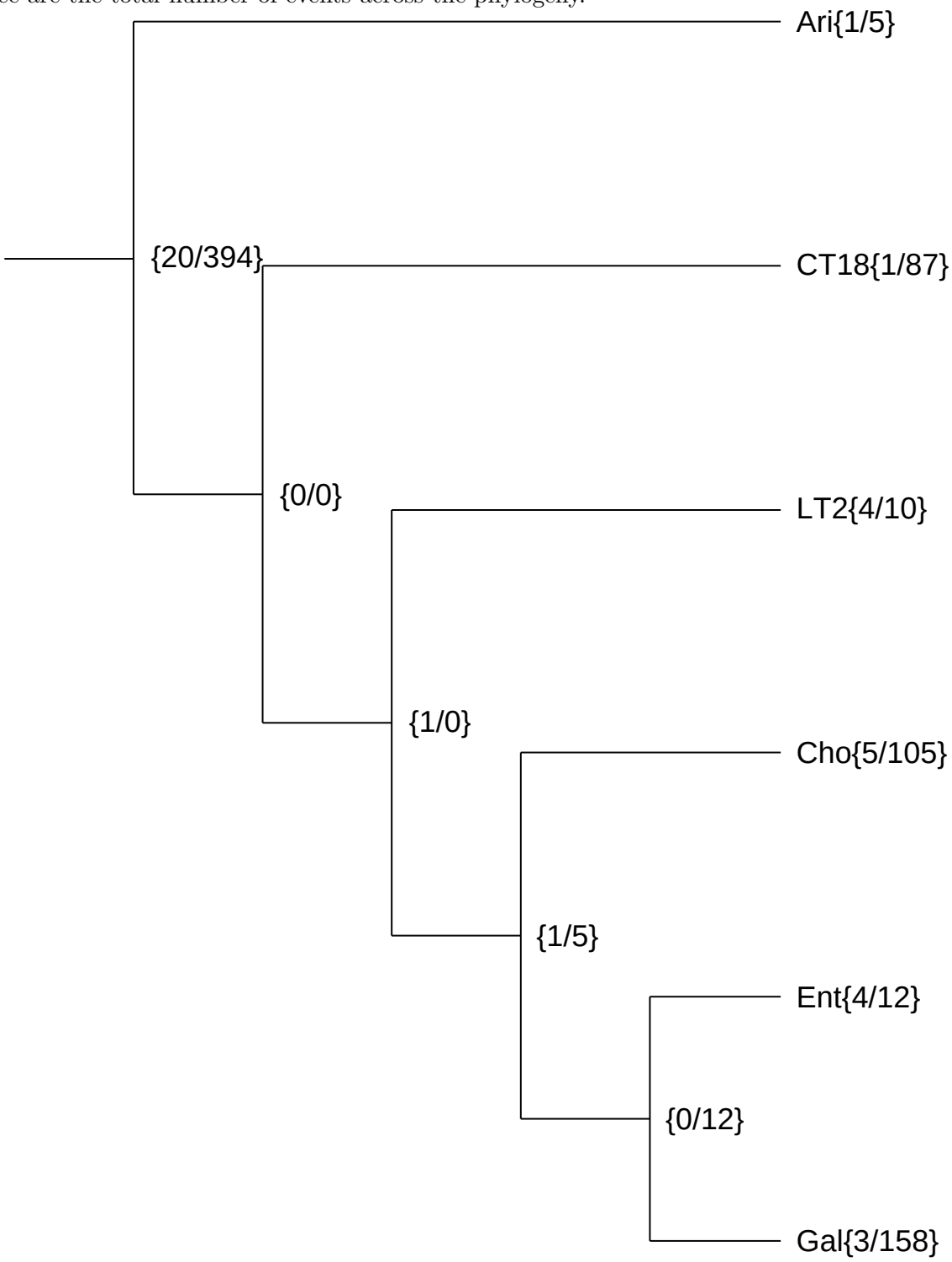
Table 2.17: Characterization of fission events mapped to the *Y. pestis* clade by `FusionMapper`.
The number of fission groups mapped is equal to the sum of the "Tot" and "Uk" columns.
Tot: Number of fission groups with identifiable component sets.
Pr: Number of component sets identified by annotated proteins.
Ps: Number of component sets identified by annotated pseudogenes.
GS: Number of component sets identified by genome scan.
Uk: Number of fission groups with no identifiable component sets.
Un: Percentage of fission groups that can be explained by a single fission event.
RBH: Percentage of component sets where a component is a RBH of a composite.

ID	Tot	Pr	Ps	GS	Uk	Un	RBH
0.PE2	78	39	32	7	4	78.0	94.9
0.PE3	138	24	106	8	12	77.3	100.0
0.PE4	68	2	62	4	8	75.0	100.0
0.ANT2	69	14	28	27	8	79.2	100.0
2.ANT1a	83	73	1	9	5	81.8	95.9
2.ANT1b	84	66	2	16	5	80.9	98.5
2.MED2	84	38	34	12	6	83.3	94.7
2.MED1	77	18	25	34	11	84.1	100.0
1.ANT1a	118	108	1	9	9	83.5	93.5
1.ANT1b	102	36	28	38	20	81.1	88.9
1.IN3	73	16	23	34	16	79.8	93.8
1.ORI1a	66	4	7	55	26	80.4	100.0
1.ORI1b	74	4	56	14	19	80.6	100.0
1.ORI2	92	18	35	39	16	82.4	83.3
1.ORI3a	91	20	30	41	17	82.4	90.0
1.ORI3b	86	20	30	36	16	82.4	90.0

Table 2.18: Characterization of fission events mapped to the *Y. pestis* cenancestor by `FusionMapper`. The number of fission groups mapped is equal to the sum of the "Tot" and "Uk" columns.
Tot: Number of fission groups with identifiable component sets.
Pr: Number of component sets identified by annotated proteins.
Ps: Number of component sets identified by annotated pseudogenes.
GS: Number of component sets identified by genome scan.
Uk: Number of fission groups with no identifiable component sets.
Un: Percentage of fission groups that can be explained by a single fission event.
RBH: Percentage of component sets where a component is a RBH of a composite.

ID	Tot	Pr	Ps	GS	Uk	Un	RBH
0.PE2	30	24	0	6	4	79.4	100.0
0.PE3	28	3	21	4	6	79.4	100.0
0.PE4	30	2	26	2	4	79.4	100.0
0.ANT2	33	6	12	15	1	79.4	100.0
2.ANT1a	32	26	0	6	2	79.4	100.0
2.ANT1b	32	25	0	7	2	79.4	100.0
2.MED2	31	11	14	6	3	79.4	100.0
2.MED1	33	6	9	18	1	79.4	100.0
1.ANT1a	32	27	0	5	2	79.4	100.0
1.ANT1b	30	4	7	19	4	79.4	100.0
1.IN3	31	4	10	17	3	79.4	100.0
1.ORI1a	29	2	3	24	5	79.4	100.0
1.ORI1b	31	2	24	5	3	79.4	100.0
1.ORI2	31	2	10	19	3	79.4	100.0
1.ORI3a	31	4	10	17	3	79.4	100.0
1.ORI3b	31	4	9	18	3	79.4	100.0

Figure 2.44: Fusion/fission analysis of 6 *S. enterica* genomes. Number of fusions and fissions are indicated by the first and second numbers on each node. The numbers at the root of the tree are the total number of events across the phylogeny.



S. enterica Typhimurium was previously resolved by SNP analysis (Kingsley *et al.*, 2009) and common ancestry between *S. enterica* Choleraesuis SC-B67 and *S. enterica* Paratyphi C RKS4594 was previously inferred (Liu *et al.*, 2009). In the one-genome analysis of *S. enterica* Typhi, 87 fissions were predicted but in the two genome analysis we predict 87 common fissions and 18 genome-specific. The *S. enterica* Typhimurium LT2 genome was initially predicted to have 10 fissions but in the three-genome analysis we observe 11 specific to *S. enterica* Typhimurium LT2 and only 2 shared between all three. However, we caution that there are more SNPs separating the *S. enterica* Typhimurium isolates than the *S. enterica* Typhi isolates (Holt *et al.*, 2008; Kingsley *et al.*, 2009).

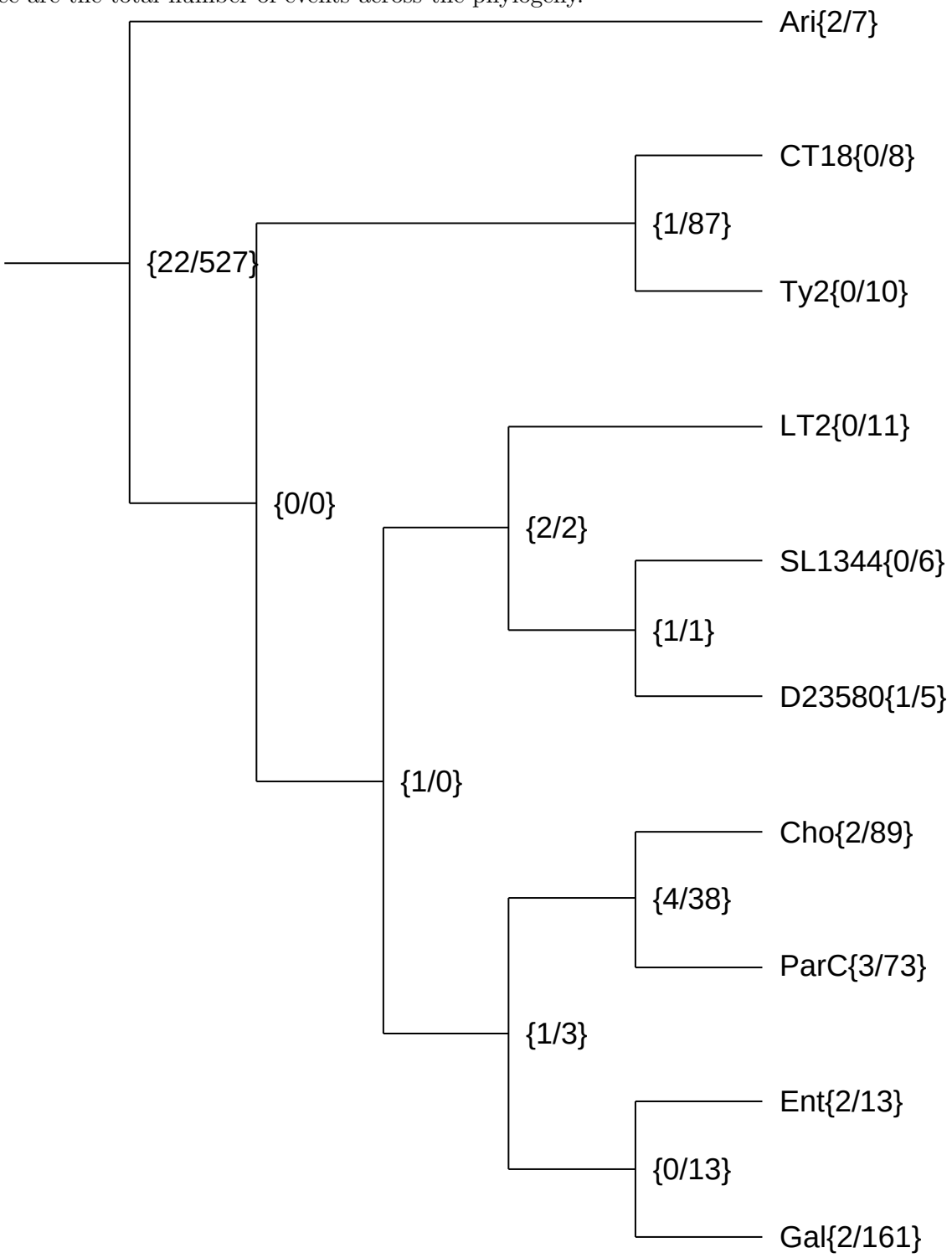
Particularly interesting was the unexpected inference of 38 fissions common between *S. enterica* Choleraesuis SC-B67 and *S. enterica* Paratyphi C RKS4594 since both cause systemic infection and are adapted to different hosts. However, both share a recent common ancestor and have distinct sets of mutations suggestive of differential selective pressures (Liu *et al.*, 2009). Furthermore, *S. enterica* Choleraesuis SC-B67 is primarily a pathogen found in swine but can also infect humans and *S. enterica* Paratyphi C RKS4594 is human-adapted. It is possible that the common ancestor already evolved into a systemic pathogen that could infect both swine and humans and further selective pressures resulted in host-specialization, a scenario consistent with the fission patterns observed. We also note that Liu *et al.* (2009) identified 101 pseudogenes specific to *S. enterica* Choleraesuis SC-B67, 97 specific to *S. enterica* Paratyphi C RKS4594, and 55 common to both, which are numbers very similar to ours.

Unlike in the *B. anthracis* and *Yersinia* datasets, we were not able to infer fissions common to all *S. enterica* genomes due to the lack of an additional outgroup species (Figure 2.10). However, we do not expect to see any since we believe the evolution of *S. enterica* reflects multiple instances of host and tissue adaptation from generalist pathogens *S. enterica* Typhimurium and *S. enterica* Enteritidis P125109, each of which is may be comparable to the emergence of *B. anthracis* from the *B. cereus* group and *Y. pestis* from *Y. pseudotuberculosis*. Nevertheless, preliminary annotation is available for is available from the distant outgroup *S. bongori* and it may be included in future analyses.

2.5 DISCUSSION

Previous studies of 17 divergent prokaryotic genomes and 131 genomes across all three domains of life found that fusion may be approximately four times more common than fission and that these events have typically occurred only once in evolutionary history (Snel, Bork and Huynen, 2000; Kummerfeld and Teichmann, 2005). Despite the large number of genomes in the second study, few were closely related (Kummerfeld and Teichmann, 2005); for example, five genomes were available from the *Bacillaceae* (*O. iheyensis*, *B. halodurans*, *B. subtilis* 168, *B. cereus* ATCC 14579, and *B. anthracis* Ames) but they correspond to four divergent subgroups. Thus there may be few orthologs for comparison and thus we may be less likely to detect fusion and fission. A recent initiative to provide both broad, unbiased phylogenetic sampling, and deeper sampling within the *Actinobacteria* phylum lead to the discovery of 4.5–12.7 times as many gene fusions as in a random sampling of existing bacterial genomes (Wu *et al.*, 2009). However, it is not clear whether the increase is due to deeper sampling or an intrinsic property of the taxa sampled and there was no distinction between fusion and fission since the authors were interested in gene function prediction instead of making

Figure 2.45: Fusion/fission analysis of 10 *S. enterica* genomes. Number of fusions and fissions are indicated by the first and second numbers on each node. The numbers at the root of the tree are the total number of events across the phylogeny.



inferences about evolutionary processes. This result supports the notion that the short-term dynamics of fusion and fission may be different. Additionally, it is not clear from these three studies whether fusion and fission events tend to be genome-specific or lineage-specific.

We have to our knowledge performed the first study of fusion and fission patterns in a large set of genomes from a single bacterial family to answer these questions, including multiple closely related strains and isolates from the same strain. In 40 *Bacillaceae* genomes, we found that gene fission is much more common and most events are genome-specific (Figure 2.30). Interestingly, 788 of 1035 fusion groups (76%) could be explained by a single fusion or fission event which is comparable to the 73% (of 2104 groups) previously reported (Kummerfeld and Teichmann, 2005). `FusionFinder` detected unique sets of fission events in the two *B. subtilis* genomes, none of which were the 19 fusions or 1 fission previously reported (Snel, Bork and Huynen, 2000). However, whereas they identified that *secDF* in *B. subtilis* was the result of a gene fusion, we found that within the *Bacillaceae* the composite form is predominant and a fission occurred in the cenancestor of the Alkaliphilic clade. Thus, the fusion of *secDF* and other events previously identified in *B. subtilis* 168 occurred prior to divergence from the *Bacillaceae* cenancestor or the differences could be due to differences in our algorithms. Thus it appears that most fusion and fission events persist over long periods of time, having occurred prior to divergence from the *Bacillaceae* cenancestor, or are genome-specific; the only exception is *B. anthracis*, although this clade is considered to be multiple isolates of the same species and will discussed further below.

One possible explanation for the prevalence of gene fission and lack of lineage-specific events in our dataset could be the presence of sequencing errors since the probability of the same ortholog being affected would be low. Genes containing long homopolymer runs may have a higher chance of being incorrectly sequenced but all genomes used in this study were sequenced using Sanger method which should lower that risk (Kuo and Ochman, 2010); in the future trace files can be examined to verify fission events. If the fission events are real, they could reflect recent events that may lead to gene deletion after a change in niche; although *B. cereus* is ubiquitous in the environment, it is an opportunistic pathogen and many of the sequenced strains were isolates from human patients (Rasko *et al.*, 2005). The fission events could also represent an intermediate step in the removal of a laterally acquired gene (Hao and Golding, 2006, 2010). Another possibility is rapid removal of the majority of fission events, a pattern observed for pseudogenes (Kuo and Ochman, 2010), which is possible if most fissions are neutral or deleterious and would imply that most fission events typically lead to pseudogenization. However, this does not explain the puzzling lack of gene fusion events. Proponents of the fusion link method for gene function predictions suggest that gene fusion may be advantageous (Enright *et al.*, 1999; Marcotte *et al.*, 1999; Yanai, Derti and DeLisi, 2001; Kamburov *et al.*, 2007) although a counterpoint is why fusion is not more prevalent considering how related genes are often adjacent to one another on the genome (Doolittle, 1999). We argue another possibility is that most gene fusions likely interfere with the normal functions of the components, such as protein-protein interactions and membrane-integration. This could be more toxic than the effects of splitting or truncating a coding sequence and thus explain the lack of fusion events in our dataset containing closely related genomes. Fusions that are advantageous or neutral and reach fixation in a population will persist for long periods of time. This hypothesis is consistent with the observation that fusion and fission events typically appear to be ancient or recent. An interesting example to investigate experimentally would be the putatively deleterious fusion gene we identified

in the avirulent *B. anthracis* Tsiankovskii-I even though both virulence plasmids are present (Figures 2.13–2.15). Note that we cannot discount the possibility that fusions and fissions are simply rare in the taxonomic groups we studied. In particular the genomes in the *B. cereus* group are remarkably stable with few rearrangements, which would reduce the chance for gene fusion and fission. Genome rearrangement is common in *Yersinia* (Darling, Mikls and Ragan, 2008) and may be occurring at such a fast rate that most genes affected by fusion or fission are rapidly removed from the genome.

Hyperthermophiles were suggested to contain significantly more fissioned genes than expected (Snel, Bork and Huynen, 2000). However, their hyperthermophilic set consisted of four archaea and one bacteria so it could potentially be due to a phylogenetic effect and Kummerfeld and Teichmann (2005) could not replicate this result with a larger set of genomes. Although we cannot directly address this question with our current dataset, we did not observe any difference between the thermophilic clade Geobacillus and the rest of the *Bacillaceae*; in fact most of the events were detected only when pseudogenes were included in the analysis (Figures 2.29,2.30).

While Kummerfeld and Teichmann (2005) did not detect a difference in hyperthermophiles, they did report a higher composite and split ratio in obligate parasites and provided two alternate possibilities based on the observation that such genomes are more compact compared to their free-living ancestors. During genome reduction, component genes may be lost more readily and inflate the ratio. Alternatively, gene fusion may be a mechanism to reduce the overall genome size. Although it is not clear whether *B. anthracis* is an obligate parasite, it is a intracellular pathogen and *B. anthracis* is the one clade that shows a clear difference from the rest of the *Bacillaceae*, with the *B. anthracis* cenancestor containing more fissions than any single genome or node in the phylogeny and the internal nodes in the *B. anthracis* clade containing similar or higher numbers of pseudogenes than internal nodes outside the *B. anthracis* clade. This suggests that a large number of gene fissions occurred and persisted in the *B. anthracis* genomes and that this process continues today; although it may look like the tempo has decreased over time, we do not have branch lengths so this currently cannot be assessed. This excess of fission in *B. anthracis* would also result in a lower composite and split ratio opposite to that reported by Kummerfeld and Teichmann (2005). If we assume fissions are pseudogenes and remove them from the analysis there will be no significant difference between *B. anthracis* and *B. cereus* and would not result in a high composite and split ratio. This result was unexpected as there are few indications in the literature that suggested the presence of a large number of fissions in *B. anthracis*. Price *et al.* (2005) found that adjacent genes predicted to be in the same operon in *B. anthracis* Ames are separated by unusually long intergenic regions and were able to detect 166 pseudogenes within these regions using `blastn`; this observation was independently reported four years earlier in the obligate symbiont *Buchnera aphidicola* (Moran and Mira, 2001). Yu (2009) inferred 184 gene losses in *B. anthracis* Ames Ancestor compared to *B. cereus* genomes and confirmed their presence in 8 other *B. anthracis* genomes. We found that most of the gene fissions detected by `FusionFinder` are annotated as pseudogenes, which likely describes the discrepancy between the conclusions of the two studies; Yu (2009) analyzed only protein-coding genes and thus concluded that gene loss has occurred while Price *et al.* (2005) searched intergenic regions between genes within putative operons with `blastn` and thus concluded the presence of pseudogenes. Both estimates are very close to the 190 and 211 pseudogenes `FusionFinder` predicts in *B. anthracis* Ames and *B. anthracis* Ames Ancestor (Table 2.6).

To our knowledge this is the first time that the observations from these two papers have been reconciled and the first time it has been demonstrated that gene fission continues to occur in *B. anthracis*, resulting in distinct sets of pseudogenes specific to particular lineages and strains.

A critical question is whether our observation is real since sequencing and annotation errors can be propagated between genome projects and 6 of our 11 genomes are unfinished drafts. We manually checked a number of predicted fissions in the *B. anthracis* lineage by inspecting the pipeline output and performing manual `BLAST` searches to ensure they were not due to algorithmic or programming errors in the `FusionFinder` pipeline. Upon inspection of pseudogenes assigned to fission groups, we noticed a large number of the pseudogenes were annotated as "authentic" or "not the result of a sequencing artifact". A keyword search of all annotated pseudogenes revealed that most of these are found in the *B. anthracis* Ames Ancestor annotation and they represent the majority of the pseudogenes in that genome (Table 2.5). However, we could not determine how pseudogenes were authenticated and whether those not annotated as such were simply not tested. Assuming the annotations are correct, we can infer that the majority of component sets identified by `FusionFinder` are real regardless of genome or identification method since most component sets in *B. anthracis* were identified via pseudogenes (Table 2.6). Concerns over the quality of draft genomes were addressed by demonstrating the lack of appreciable differences in the number of events inferred between sampling either two finished or draft genomes, one each from the *B. anthracis* A1 and A2 sublineages (Figures 2.28,2.27). Perhaps most convincingly, we further showed that the most parsimonious reconstruction of fission events is fully consistent with the combined SNP and VNTR data although this is not the case in the A1 sublineage if pseudogenes are excluded from analysis (Figure 2.31,2.32). Since all finished genomes belong to sublineages A1 and A2 and all three lineages, including A1 and A2, are represented by draft genomes, we would not expect the fission patterns to mirror the phylogeny if they were the result of annotation or sequencing error propagation. Importantly, we demonstrated that fission and SNP patterns in *Yersinia* are also broadly consistent even in the presence of large numbers of genomic inversions that lead to loss of entire genes.

To complement the high-level view of the gene fission process, we performed an in-depth study of the motility and chemotaxis genes in the *B. cereus* group because these genes are well-studied, frameshifts have been reported in *B. anthracis* orthologs, and *B. anthracis* is known to be motile; we only considered the *B. cereus* group because the complement of motility genes differs between *B. cereus* and *B. subtilis* and thus may be fundamentally different (Waltman *et al.*, 2010). We have for the first time shown that the complement of motility genes fissioned differs between *B. anthracis* isolates and the patterns are also consistent with the general fission, SNP, and VNTR patterns. We stress that like any other method, there will be false positives and false negatives in our automated dataset, especially in comparison to human curation or the use of more rigorous criteria such as requiring the presence of positional orthologs (Kuo and Ochman, 2010). The key message is the presence of a significant phylogenetic signal in the dataset powerful enough to overcome the background noise and the analysis of the motility gene clusters supports this view. Dependence on the correctness of our species phylogeny is a critical assumption and a potential point of contention. However, the lack of events and support for nodes outside the *B. anthracis* lineage and the concordance of fission patterns with SNPs and VNTRs within *B. anthracis* indicates that the species phylogeny has no significant effect on our results (Figure 2.26).

Aside from loss of motility, do the gene fissions in *B. anthracis* actually have any other effects? A critical step in the evolution of *B. anthracis* is thought to be the acquisition of the two virulence plasmids and there may have been a concomitant influx of laterally transferred genes in the *B. anthracis* lineage that are being pseudogenized. Furthermore, the *B. anthracis* genomes are the least diverged set in our dataset and estimates of insertion and deletion rates tend to be negatively associated with genome divergence (Hao and Golding, 2010). However, of the 70 fissions common to all *B. anthracis* (Table 2.8), 20 were identifiable by composite genes in all *B. cereus* genomes and another 10 in all except for the earliest diverging *B. cytotoxicus* (composites may or may not be present outside the *B. cereus* group). Thus gene fission appears to affect genes broadly distributed in the *B. cereus* group as well as those more recently acquired. We also found that at least 90% of component sets contain a RBH to a composite gene (Table 2.6). Thus, it appears that gene fission is affecting one-to-one orthologs or outparalogs which means inactivation of LGT or *B. anthracis*-specific inparalogs is unlikely. We are unsure why the splitting of genes into smaller components would be advantageous and manual analysis of a number of gene fissions revealed that many disrupt protein domains annotated in the Conserved Domain Database (Marchler-Bauer *et al.*, 2011). Thus, we infer that many fissions may disrupt gene function and represent pseudogenization. The d_N/d_S ratio may be used to detect relaxed selection acting on coding sequences and thus serve as a proxy for pseudogene detection. This test may be conservative since pseudogenes are typically of recent origin and thus do not have sufficient time to accumulate mutations, which applies to our fission events (Ochman and Davalos, 2006). Furthermore, use of the d_N/d_S ratios may not be appropriate for our dataset since the differences between our sequences likely reflect segregating polymorphisms as opposed to fixed substitutions (Kryazhimskiy and Plotkin, 2008). A difference in d_N/d_S is also expected if the effective population sizes differ between compared species, which is also true since *B. anthracis* is a recently emerged intracellular pathogen. Nevertheless, we compared the d_N/d_S ratios for fission groups and orthologous groups and found that genes tend to have higher d_N/d_S ratios in the *B. anthracis* lineage compared to *B. cereus* and genes that are fissioned in *B. anthracis* also tend to have higher d_N/d_S ratios than genes that are not fissioned in *B. anthracis*. Thus it appears fissioned genes are under relaxed selection and may be pseudogenes. Furthermore, full-length genes in *B. cereus* genomes that are fissioned in *B. anthracis* may be under less purifying selection compared to those that are never found fissioned in *B. anthracis*. Pseudogenization in *B. anthracis* may reflect niche adaptation but we were unable to find evidence that gene fission affected genes in particular COG categories at levels different from random expectation except for the "Cell motility" (N) and "Cell wall/membrane/envelope biogenesis" (M) categories. Although genes could still have been inactivated during niche adaptation, this result suggest that there may also be a random component to the gene fission process in *B. anthracis*. Since gene fission appears to be a random and ongoing process, it may eventually lead to genome reduction. Interestingly, no significant biases were found in *Y. pestis* (Chain *et al.*, 2006) and pseudogenes were found in substantial numbers in all COG categories except for "Cell cycle control, cell division, chromosome partitioning" (D) in *Serratia symbiotica* (Burke and Moran, 2011) even though reductive genome evolution has been implicated in both.

If the fissions in *B. anthracis* do represent pseudogenization, why are so many of them detectable? Comparatively few fissions were predicted for the other genomes in our dataset, which is consistent with the notion that recently truncated genes and pseudogenes are rapidly

purged from the genome (Hao and Golding, 2010; Kuo and Ochman, 2010). It is well known that recent host restriction for both pathogens and symbionts allows for nutrient uptake from the host and is associated with reduced effective population sizes and increased effects of genetic drift which allows accumulation of pseudogenes (Lawrence, Hendrix and Casjens, 2001; McClelland *et al.*, 2001; Moran and Mira, 2001; Moran and Plague, 2004; Lerat and Ochman, 2005; Ochman and Davalos, 2006; Burke and Moran, 2011). We believe that these factors may also contribute to the pseudogenization in *B. anthracis* since it recently acquired two virulence plasmids and has low genetic diversity, which is consistent with reduced population size and higher levels of genetic drift.

We expected to find a similar pattern in *Y. pestis*, which is also genetically monomorphic and a recently emerged pathogen (Morelli *et al.*, 2010). While fission patterns are also broadly consistent with SNPs, comparatively few fissions were shared and more were genome-specific. This discrepancy could be due to differences in amount of divergence since the cenancestor and among sampled isolates, nature of the ancestor, differences in ecology and population structure, or the number of IS. Current estimates of time since divergence of *B. anthracis* and *Y. pestis* may be unreliable and it is not clear how to compare divergence within *B. anthracis* to *Y. pestis* (Achtman, 2008). *B. anthracis* is believed to have evolved from a member of the *B. cereus* group, which are generally considered opportunistic pathogens while *Y. pestis* is believed to have evolved from a *Y. pseudotuberculosis* strain, which is a gastrointestinal pathogen. Perhaps the *B. cereus* ancestor was less adapted as a pathogen and thus had a larger complement of dispensable genes. Ecological differences such as mode of transmission and transmissibility may also affect pseudogene fixation rates. Another possibility is the large number of IS and IS-mediated genomic recombination facilitates the removal of pseudogenes in *Y. pestis* (Darling, Mikls and Ragan, 2008). *B. anthracis* does not appear to have a noticeably higher number of IS and no major genome rearrangements have occurred compared to *B. cereus* genomes (Kolstø, Tourasse and Økstad, 2009). This is a significant difference because proliferation of IS is considered a key step in genome reduction after host-restriction (Moran and Plague, 2004). Interestingly, *Mycobacterium leprae* has a reduced genome containing a large complement of pseudogenes yet only nonfunctional IS remnants remain, but repetitive DNA sequences can be found within pseudogenes (Akama *et al.*, 2009; Monot *et al.*, 2009). Small repeated elements do exist in *B. anthracis* but once again there does not appear to be a noticeable difference between *B. anthracis* genomes and no significant differences were reported with respect to *B. cereus* genomes (Kristoffersen *et al.*, 2011).

S. enterica was a good dataset to further test our hypothesis that host-restriction may lead to elevated numbers of pseudogenes because there have been multiple instances of host-restriction in this species (Chiu *et al.*, 2005; Holt *et al.*, 2008; Thomson *et al.*, 2008; Liu *et al.*, 2009). We found that high numbers of pseudogenes were indeed associated with host-restricted serovars (Figure 2.44). In agreement with our hypothesis and previous results, *S. enterica* Choleraesuis SC-B67 and *S. enterica* Paratyphi C RKS4594 have both shared and unique sets of pseudogenes because they share very recent common ancestry and both are host-restricted (Chiu *et al.*, 2005; Liu *et al.*, 2009). We were also able to show that including multiple closely isolate does not necessarily result in higher pseudogene predictions which further supports our hypothesis. The three *S. enterica* Typhimurium genomes each have few pseudogenes and have few shared pseudogenes in contrast to the large number of shared pseudogenes between the two *S. enterica* Typhi genomes. Our numbers compare

favorably to those from a previous study of pseudogenes in *S. enterica* using fewer genomes (Kuo and Ochman, 2010). The major difference of our study is the recognition of the link between pseudogene abundance and host-restriction, leading us to increase sampling of genomes to test our hypothesis. However, our algorithmic approaches differed and they manually inspected their predictions and were able to infer the type of event that lead to pseudogenization.

Although the number of fissions predicted by `FusionFinder` are similar to the number of pseudogenes reported by Kuo and Ochman (2010), our per-genome estimates are often lower than the number of annotated pseudogenes in genome annotations for both *S. enterica* and *Y. pestis*. A contributing factor may be the annotation of component sets with a single or multiple pseudogene entries (Table 2.7); when fission is the direct result of genome recombination like in some instances in *Y. pestis*, use of multiple pseudogene entries is unavoidable. Gene gain and loss occurs rapidly at the tips of the phylogeny (Hao and Golding, 2006, 2010) and thus a proportion of pseudogenes unassigned to fusion or fission events by `FusionFinder` may reflect inactivation of genome-specific gene acquisitions. A major impediment to the comparison of pseudogene counts is the lack of a consistent definition or detection methodology (Chain *et al.*, 2004; Lerat and Ochman, 2005). In particular, `FusionFinder` was designed to identify gene fusions and fissions, not pseudogenes, and is unable to detect certain types such as those caused by truncation and internal deletions rather than frameshift mutations (Lerat and Ochman, 2005). Kolstø, Tourasse and Økstad (2009) reported 35 chromosomal proteins that differed in length between the A and B lineages and made particular mention of three with deletions of at least 10 amino acids. We manually inspected these proteins and found that two involved in-frame deletions and one is truncated at the N-terminal end and thus fall into categories used by Lerat and Ochman (2005) and cannot be detected by our algorithm (Table 2.19). However, an important question is whether it is appropriate to consider these as pseudogenes.

Table 2.19: Three proteins in *B. anthracis* that may have disrupted function. The NCBI GI is provided for the full-length ortholog in either *B. anthracis* Ames Ancestor (lineage A) or *B. anthracis* KrugerB (lineage B). All affected domains were identified via specific hits to the CDD (high confidence); the accession number is provided.

GI	Lineage	Mutation	Affected region
47525758	B	32 aa in-frame deletion	cd00009
254743016	A	18 aa N-terminal truncation	cl00207
254743973	A	10 aa in-frame deletion	compositionally biased seq.

Verification that a gene is truly nonfunctional is difficult and requires biochemical assays (Zheng and Gerstein, 2007). Although it was assumed that pseudogenes are not transcribed, there is increasing evidence that pseudogenes can be transcribed in both eukaryotic and bacterial genomes and that some of them may have acquired new functions (Zheng and Gerstein, 2007; Akama *et al.*, 2009; Carlson *et al.*, 2009; Perkins *et al.*, 2009; Martin *et al.*, 2010). Interestingly, transcription was detected from up to 43% of *M. leprae* pseudogenes using microarrays but only 0.69% of *S. enterica* Typhi pseudogenes via transcriptome sequencing (Akama *et al.*, 2009; Perkins *et al.*, 2009; Williams *et al.*, 2009). Aside from the

obvious difference in methodologies and species, pseudogene age and genome stability may play a role in the different results. Importantly, transcriptome sequencing under different conditions revealed that many annotated pseudogenes in *B. anthracis* are transcriptionally active and two pseudogenes were also among the top 50 up-regulated genes identified using microarrays in *B. anthracis* under iron starvation (Carlson *et al.*, 2009; Martin *et al.*, 2010); both pseudogenes were annotated as authentic in the *B. anthracis* Ames Ancestor annotation. Thus, it appears that transcriptional activity may be insufficient to authenticate if a gene is a pseudogene or whether a pseudogene is functional. One is left wondering whether gene fission components should be annotated as proteins or pseudogenes. This problem is by no means new and designating them as "split genes" may be a good approach (Ogata *et al.*, 2001). Regardless, we believe that there needs to be consistency between genome annotations and it is clear from our results that this is not the case in *B. anthracis* and to a lesser degree in *Y. pestis* and *S. enterica*.

In the future it may be interesting to increase the breadth of sampling even if it may be contrary to our original goals. For example, we could move to the order level and concurrently analyze the *Bacillaceae*, *Staphylococcaceae* and *Listeriaceae* families. This would allow us to check whether our program will find a larger number of gene fusions at the ancestral nodes as we predict given the results of this and previous studies. Another possibility is analyzing another taxonomic group with more IS and other mobile genetic elements. Since genome rearrangement is one of the mechanisms for fusion and fission, our dataset may not be ideal since half of the genomes are from the highly similar and syntenic *B. cereus* group (Rasko *et al.*, 2005; Kolstø, Tourasse and Økstad, 2009).

Concordance between fission patterns and other molecular markers appears conclusive in *B. anthracis* and *Yersinia* but it is not clear whether this is due to special circumstances or is common to other monomorphic pathogens. Application of `FusionFinder` to other datasets could help clarify these possibilities, such as the large numbers of *S. enterica* Typhi and *S. enterica* subsp. *arizonae* isolates recently sequenced (Holt *et al.*, 2008). The possibility that gene fission events could be phylogenetically informative in monomorphic pathogens is particularly interesting because such lineages are intrinsically difficult to analyze phylogenetically due to the low levels of genetic diversity (Achtman, 2008). Next-generation sequencing is now being applied to investigate outbreaks of infectious diseases and may provide additional datasets to test our hypothesis in the future (Chen *et al.*, 2010; Harris *et al.*, 2010; Gilmour *et al.*, 2010; Lewis *et al.*, 2010).

We analyzed the putative pseudogenes in *B. anthracis* using the d_N/d_S ratio but we can further characterize them in the future. Possibilities include tests for abnormal GC content or unusual codon usage. Another possibility is to determine whether SNPs and repeat elements are associated with gene fissions along the chromosome; 58% of 990 SNPs previously identified in *B. anthracis* are non-synonymous but the distribution of these was not described (Pearson *et al.*, 2004). If there is an abundance of non-synonymous SNPs it may be a sign of relaxed selection and if there is an abundance of synonymous SNPs or SNPs in intergenic regions we may infer mutational hotspots.

In fact we have yet to analyze the distribution of gene fission events although a rough estimate considering just the locus tag numbers suggests that the distribution is random, with at least one fission for every 100 genes or pseudogenes annotated in *B. anthracis* Ames Ancestor. Although our program does record the strand of each gene assigned to a fusion group we have yet to perform a statistical analysis to determine whether there is a bias such

as the leading strand bias found for LGT (Hao and Golding, 2009). If we focus on just *B. anthracis* and a small number of closely related *B. cereus* genomes we could accurately infer the mechanisms that lead to gene fission in *B. anthracis* and compare them to previous reports (Moran, McLaughlin and Sorek, 2009; Kuo and Ochman, 2010). An estimate of the rates of gene fission and subsequent gene loss in *B. anthracis* would be desirable but the lack of divergence times is problematic. Simulations could also be designed to test whether gene fissions will accumulate in a phylogenetically informative way under host-restriction. Finally, it appears that pseudogenes may be predicted as coding sequences by gene finders so it may be interesting to see how many fissioned genes they actually identify (Martin *et al.*, 2010).

The `FusionFinder` pipeline may also be improved in the future. The handling of duplicate genes by `FusionFinder` can be improved even though we have demonstrated that it is unlikely to affect our results. We also do not explicitly consider the possibility of LGT; potentially a set of outgroup sequences could be included in the analysis similar to the approach taken in `InParanoid` (Remm, Storm and Sonnhammer, 2001). `FusionFinder` can analyze a single genome in a pairwise analysis but this scenario was not explicitly considered when the algorithm was designed. Thus, it is not clear whether `FusionFinder` can accurately handle a single genome analysis and `MultiFusion` does not integrate these comparisons when building fusion groups; in particular such a comparison may have a high false positive rate. In the future this feature can be added to our pipeline which may allow us to explicitly consider the possibility of a genome-specific fusion or fission after gene duplication. Use of `FusionFinder` on the *B. subtilis* 168 genome revealed one example of this scenario. The "lesion bypass phage DNA polymerase" protein (GI: 255767474) was likely duplicated and subsequently split into two fragments annotated as "DNA repair protein fragment" (GI: 16078954) and "DNA repair protein" (GI: 255767448). The measure of composite coverage used in `FusionFinder` and `MultiFusion` needs to be fixed or modified as its implementation is currently inconsistent; in particular if two HSP are joined the region between the HSP may be counted as "covered" (Figure 2.5). `FusionMapper` may also be improved in a couple of ways. First, `FusionMapper` requires a strictly bifurcating species phylogeny, limiting its usefulness. Additionally, `FusionMapper` appears to incorrectly infer gene fusion if there are fissions in multiple genomes and few composite genes in other genomes. A way to flag these and an option to discard these events may be desirable. Finally, simulation of fusion and fission could be used to help verify the correctness of our pipeline and give rough estimates of its accuracy.

In conclusion, we have developed a pipeline for the automatic inference of gene fusion and fission and for the first time analyzed these events in a set of closely related strains from a single bacterial family, including multiple isolates of *B. anthracis*. Our results have made contributions to several areas of research. First, we have a better understanding of the role fusion and fission plays in protein evolution. Our results combined with previous studies suggest that fusion and fission events in the *Bacillaceae* either occurred before divergence from the *Bacillaceae* cenancestor or are strain-specific, implying that few events are fixed but those that reach fixation persist over long evolutionary time. The skew towards fission at the strain-level suggests that fission events are less likely to be toxic or that they are more common than fusions. The second contribution we have made is towards the biology of *B. anthracis*. Few characteristics are known to uniquely differentiate *B. anthracis* from *B. cereus* genomes but we have demonstrated that aside from the unique fission of *plcR*, there is a large complement of gene fissions common to *B. anthracis* (Kolstø, Tourasse and

Økstad, 2009). The annotation of these gene fission events are not consistent across the 11 *B. anthracis* genomes in our study. This demonstrates that comparative genome analyses may be affected if this difference is not taken into account, such as the analysis of the history of the *B. cereus* CI genome (Klee *et al.*, 2010). Finally, we have contributed to the understanding of monomorphic pathogens, showing that pseudogenization patterns carry a phylogenetic signal in *B. anthracis* and *Yersinia* and may be used as a complementary tool for identifying the evolutionary histories despite the lack of genetic diversity (Achtman, 2008).

Part II

CONCLUSION

Previous systematic comparisons of orthology identification methods gave conflicting messages on the relative performance between similarity-based and tree reconciliation methods. In theory tree reconciliation methods should perform better since they are true to the definition of orthology whereas similarity-based methods rely on sequence similarity as a surrogate measure of orthology. Our simulation results in Chapter 1 demonstrate that tree reconciliation does indeed perform better given that the species phylogeny is correct but in practical situations where the true species phylogeny is not certain, similarity-based methods are likely to perform better. We found that `OrthoMCL` provides higher precision but lower recall than `MultiParanoid` and we observe the same tradeoff between `RAP` and `Notung`.

While most existing orthology identification algorithms do not explicitly consider gene fusion and gene fission, the results from Chapter 2 suggest that in the average dataset the problem is likely minimal. In particular, we showed that gene fusion and fission are typically rare and either do not reach fixation or are retained over long evolutionary time. High proportions of the fusion groups detected by `FusionFinder` include a reciprocal best hits (RBH) between a composite and component, demonstrating we were successful in identifying events involving orthologs even though we do not enforce RBH in our algorithm. This suggests that extension of similarity-based methods of orthology prediction to detect fusion and fission while enforcing RBH will detect the majority of the events we reported. Alternatively, one can simply use our pipeline or develop a similar program to perform post-processing of orthologous groups.

Whereas bacterial genomes were once thought to be densely packed with genes and pseudogene abundance was thought to be specific to host-restricted bacteria, subsequent analyses suggest that pseudogenes may be more prevalent than first thought in even free-living bacteria. Similarly, many investigators take a small number of *B. anthracis* strains in comparative studies and thus may have overlooked the abundance of gene fission in *B. anthracis* we uncovered in Chapter 2. Future studies involving *B. anthracis* will need to make adjustments to circumvent the lack of consistency in the *B. anthracis* genome annotations with respect to the gene fissions.

Part III

REFERENCES

Bibliography

Achtman, M. (2004, Sep). Population structure of pathogenic bacteria revisited. *Int J Med Microbiol 294*(2-3), 67–73.

Achtman, M. (2008). Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. *Annu Rev Microbiol 62*, 53–70.

Akama, T., K. Suzuki, K. Tanigawa, A. Kawashima, H. Wu, N. Nakata, Y. Osana, Y. Sakakibara, and N. Ishii (2009, May). Whole-genome tiling array analysis of *Mycobacterium leprae* RNA reveals high expression of pseudogenes and noncoding regions. *J Bacteriol 191*(10), 3321–3327.

Akerborg, O., B. Sennblad, L. Arvestad, and J. Lagergren (2009, Apr). Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proc Natl Acad Sci U S A 106*(14), 5714–5719.

Alcaraz, L. D., G. Moreno-Hagelsieb, L. E. Eguiarte, V. Souza, L. Herrera-Estrella, and G. Olmedo (2010). Understanding the evolutionary relationships and major traits of *Bacillus* through comparative genomics. *BMC Genomics 11*, 332.

Alexeyenko, A., I. Tamas, G. Liu, and E. L. L. Sonnhammer (2006, Jul). Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics 22*(14), e9–15.

Altenhoff, A. M. and C. Dessimoz (2009, Jan). Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput Biol 5*(1), e1000262.

Andersson, J. O. and S. G. Andersson (1999, Dec). Insights into the evolutionary process of genome degradation. *Curr Opin Genet Dev 9*(6), 664–671.

Arvestad, L., A.-C. Berglund, J. Lagergren, and B. Sennblad (2003). Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics 19 Suppl 1*, i7–15.

Barbe, V., S. Cruveiller, F. Kunst, P. Lenoble, G. Meurice, A. Sekowska, D. Vallenet, T. Wang, I. Moszer, C. Mdigue, and A. Danchin (2009, Jun). From a consortium sequence to a unified sequence: the *Bacillus subtilis* 168 reference genome a decade later. *Microbiology 155*(Pt 6), 1758–1775.

Beiko, R. G. and R. L. Charlebois (2007, Apr). A simulation test bed for hypotheses of genome evolution. *Bioinformatics 23*(7), 825–831.

Berglund, A.-C., E. Sjlund, G. Ostlund, and E. L. L. Sonnhammer (2008, Jan). InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Res 36*(Database issue), D263–D266.

Berglund-Sonnhammer, A.-C., P. Steffansson, M. J. Betts, and D. A. Liberles (2006, Aug). Optimal gene trees from sequences and species trees using a soft interpretation of parsimony. *J Mol Evol 63*(2), 240–250.

Binford, G. J., M. R. Bodner, M. H. J. Cordes, K. L. Baldwin, M. R. Rynerson, S. N. Burns, and P. A. Zobel-Thropp (2009, Mar). Molecular evolution, functional variation, and proposed nomenclature of the gene family that includes sphingomyelinase D in sicariid spider venoms. *Mol Biol Evol 26*(3), 547–566.

Bolhuis, A., C. P. Broekhuizen, A. Sorokin, M. L. van Roosmalen, G. Venema, S. Bron, W. J. Quax, and J. M. van Dijl (1998, Aug). SecDF of *Bacillus subtilis*, a molecular Siamese twin required for the efficient secretion of proteins. *J Biol Chem 273*(33), 21217–21224.

Brocchieri, L. and S. Karlin (2005). Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Res 33*(10), 3390–3400.

Burke, G. R. and N. A. Moran (2011, Jan). Massive Genomic Decay in *Serratia symbiotica*, a Recently Evolved Symbiont of Aphids. *Genome Biol Evol 3*, 195–208.

Carlson, P. E., K. A. Carr, B. K. Janes, E. C. Anderson, and P. C. Hanna (2009). Transcriptional profiling of *Bacillus anthracis* Sterne (34F2) during iron starvation. *PLoS One 4*(9), e6988.

Caspi, R., T. Altman, J. M. Dale, K. Dreher, C. A. Fulcher, F. Gilham, P. Kaipa, A. S. Karthikeyan, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, S. Paley, L. Popescu, A. Pujar, A. G. Shearer, P. Zhang, and P. D. Karp (2010, Jan). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res 38*(Database issue), D473–D479.

Chain, P. S. G., E. Carniel, F. W. Larimer, J. Lamerdin, P. O. Stoutland, W. M. Regala, A. M. Georgescu, L. M. Vergez, M. L. Land, V. L. Motin, R. R. Brubaker, J. Fowler, J. Hinnebusch, M. Marceau, C. Medigue, M. Simonet, V. Chenal-Francisque, B. Souza, D. Dacheux, J. M. Elliott, A. Derbise, L. J. Hauser, and E. Garcia (2004, Sep). Insights into the evolution of *Yersinia pestis* through whole-genome comparison with *Yersinia pseudotuberculosis*. *Proc Natl Acad Sci U S A 101*(38), 13826–13831.

Chain, P. S. G., P. Hu, S. A. Malfatti, L. Radnedge, F. Larimer, L. M. Vergez, P. Worsham, M. C. Chu, and G. L. Andersen (2006, Jun). Complete genome sequence of *Yersinia pestis* strains Antiqua and Nepal516: evidence of gene reduction in an emerging pathogen. *J Bacteriol 188*(12), 4453–4463.

Chan, C. X., R. G. Beiko, A. E. Darling, and M. A. Ragan (2009). Lateral transfer of genes and gene fragments in prokaryotes. *Genome Biol Evol 2009*(0), 429–438.

Chang, M. S. S. and S. A. Benner (2004, Aug). Empirical analysis of protein insertions and deletions determining parameters for the correct placement of gaps in protein sequence alignments. *J Mol Biol 341*(2), 617–631.

Chee, G.-J. and H. Takami (2005, Dec). Housekeeping recA gene interrupted by group II intron in the thermophilic *Geobacillus kaustophilus*. *Gene 363*, 211–220.

Chen, F., A. J. Mackey, C. J. Stoeckert, and D. S. Roos (2006, Jan). OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res 34*(Database issue), D363–D368.

Chen, F., A. J. Mackey, J. K. Vermunt, and D. S. Roos (2007). Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS One 2*(4), e383.

Chen, P. E., K. M. Willner, A. Butani, S. Dorsey, M. George, A. Stewart, S. M. Lentz, C. E. Cook, A. Akmal, L. B. Price, P. S. Keim, A. Mateczun, T. N. Brahmbhatt, K. A. Bishop-Lilly, M. E. Zwick, T. D. Read, and S. Sozhamannan (2010). Rapid identification of genetic modifications in *Bacillus anthracis* using whole genome draft sequences generated by 454 pyrosequencing. *PLoS One 5*(8), e12397.

Chiu, C.-H., P. Tang, C. Chu, S. Hu, Q. Bao, J. Yu, Y.-Y. Chou, H.-S. Wang, and Y.-S. Lee (2005). The genome sequence of *Salmonella enterica* serovar Choleraesuis, a highly invasive and resistant zoonotic pathogen. *Nucleic Acids Res 33*(5), 1690–1698.

Cock, P. J. A. and D. E. Whitworth (2007, Nov). Evolution of prokaryotic two-component system signaling pathways: gene fusions and fissions. *Mol Biol Evol 24*(11), 2355–2357.

Dai, L. and S. Zimmerly (2002, Mar). Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. *Nucleic Acids Res 30*(5), 1091–1102.

Darling, A. E., I. Mikls, and M. A. Ragan (2008). Dynamics of genome rearrangement in bacterial populations. *PLoS Genet 4*(7), e1000128.

Datta, R. S., C. Meacham, B. Samad, C. Neyer, and K. Sjlander (2009, Jul). Berkeley PHOG: PhyloFacts orthology group prediction web server. *Nucleic Acids Res 37*(Web Server issue), W84–W89.

Dehal, P. S. and J. L. Boore (2006). A phylogenomic gene cluster resource: the Phylogenetically Inferred Groups (PhIGs) database. *BMC Bioinformatics 7*, 201.

Deluca, T. F., I.-H. Wu, J. Pu, T. Monaghan, L. Peshkin, S. Singh, and D. P. Wall (2006, Aug). Roundup: a multi-genome repository of orthologs and evolutionary distances. *Bioinformatics 22*(16), 2044–2046.

Deng, W., S.-R. Liou, G. Plunkett, G. F. Mayhew, D. J. Rose, V. Burland, V. Kodoyianni, D. C. Schwartz, and F. R. Blattner (2003, Apr). Comparative genomics of *Salmonella enterica* serovar Typhi strains Ty2 and CT18. *J Bacteriol 185*(7), 2330–2337.

Doolittle, R. F. (1999, Sep). Do you dig my groove? *Nat Genet 23*(1), 6–8.

Drosophila 12 Genomes Consortium (2007, Nov). Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature 450*(7167), 203–218.

Dufayard, J.-F., L. Duret, S. Penel, M. Gouy, F. Rechenmann, and G. Perrire (2005, Jun). Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics 21*(11), 2596–2603.

Durand, D., B. V. Halldrsson, and B. Vernot (2006, Mar). A hybrid micro-macroevolutionary approach to gene tree reconstruction. *J Comput Biol 13*(2), 320–335.

Durrens, P., M. Nikolski, and D. Sherman (2008, Oct). Fusion and fission of genes define a metric between fungal genomes. *PLoS Comput Biol 4*(10), e1000200.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res 32*(5), 1792–1797.

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

Enright, A. J., I. Iliopoulos, N. C. Kyrpides, and C. A. Ouzounis (1999, Nov). Protein interaction maps for complete genomes based on gene fusion events. *Nature 402*(6757), 86–90.

Enright, A. J. and C. A. Ouzounis (2001). Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. *Genome Biol 2*(9), RESEARCH0034.

Eppinger, M., P. L. Worsham, M. P. Nikolich, D. R. Riley, Y. Sebastian, S. Mou, M. Achtman, L. E. Lindler, and J. Ravel (2010, Mar). Genome sequence of the deep-rooted *Yersinia pestis* strain Angola reveals new insights into the evolution and pangenome of the plague bacterium. *J Bacteriol 192*(6), 1685–1699.

Felsenstein, J. (1989). PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics 5*, 164–166.

Fischer, W., L. Windhager, S. Rohrer, M. Zeiller, A. Karnholz, R. Hoffmann, R. Zimmer, and R. Haas (2010, Oct). Strain-specific genes of *Helicobacter pylori*: genome evolution driven by a novel type IV secretion system and genomic island transfer. *Nucleic Acids Res 38*(18), 6089–6101.

Fitch, W. M. (1970, Jun). Distinguishing homologous from analogous proteins. *Syst Zool 19*(2), 99–113.

Fitch, W. M. (2000, May). Homology a personal view on some of the problems. *Trends Genet 16*(5), 227–231.

Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, and J. M. Merrick (1995, Jul). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science 269*(5223), 496–512.

Flicek, P., M. R. Amode, D. Barrell, K. Beal, S. Brent, Y. Chen, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gordon, M. Hendrix, T. Hourlier, N. Johnson, A. Khri, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, E. Kulesha, P. Larsson, I. Longden, W. McLaren, B. Overduin, B. Pritchard, H. S. Riat, D. Rios, G. R. S. Ritchie, M. Ruffier, M. Schuster, D. Sobral, G. Spudich, Y. A. Tang, S. Trevanion, J. Vandrovcova, A. J. Vilella, S. White, S. P. Wilder, A. Zadissa, J. Zamora, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X. M. Fernndez-Suarez, J. Herrero, T. J. P. Hubbard, A. Parker, G. Proctor, J. Vogel, and S. M. J. Searle (2011, Jan). Ensembl 2011. *Nucleic Acids Res 39*(Database issue), D800–D806.

Fong, J. H., L. Y. Geer, A. R. Panchenko, and S. H. Bryant (2007, Feb). Modeling the evolution of protein domain architectures using maximum parsimony. *J Mol Biol 366*(1), 307–315.

Fraser, C. M., J. D. Gocayne, O. White, M. D. Adams, R. A. Clayton, R. D. Fleischmann, C. J. Bult, A. R. Kerlavage, G. Sutton, J. M. Kelley, R. D. Fritchman, J. F. Weidman, K. V. Small, M. Sandusky, J. Fuhrmann, D. Nguyen, T. R. Utterback, D. M. Saudek, C. A. Phillips, J. M. Merrick, J. F. Tomb, B. A. Dougherty, K. F. Bott, P. C. Hu, T. S. Lucier, S. N. Peterson, H. O. Smith, C. A. Hutchison, and J. C. Venter (1995, Oct). The minimal gene complement of *Mycoplasma genitalium*. *Science 270*(5235), 397–403.

Frygelius, J., L. Arvestad, A. Wedell, and V. Thnen (2010, May). Evolution and human tissue expression of the Cres/Testatin subgroup genes, a reproductive tissue specific subgroup of the type 2 cystatins. *Evol Dev 12*(3), 329–342.

Fu, Z., X. Chen, V. Vacic, P. Nan, Y. Zhong, and T. Jiang (2007, Nov). MSOAR: a high-throughput ortholog assignment system based on genome rearrangement. *J Comput Biol 14*(9), 1160–1175.

Fuxelius, H.-H., A. C. Darby, N.-H. Cho, and S. G. E. Andersson (2008). Visualization of pseudogenes in intracellular bacteria reveals the different tracks to gene destruction. *Genome Biol 9*(2), R42.

Gabaldn, T., C. Dessimoz, J. Huxley-Jones, A. J. Vilella, E. L. Sonnhammer, and S. Lewis (2009). Joining forces in the quest for orthologs. *Genome Biol 10*(9), 403.

Galtier, N. (2007, Aug). A model of horizontal gene transfer and the bacterial phylogeny problem. *Syst Biol 56*(4), 633–642.

Gardiner, A., D. Barker, R. K. Butlin, W. C. Jordan, and M. G. Ritchie (2008). Evolution of a complex locus: exon gain, loss and divergence at the Gr39a locus in *Drosophila*. *PLoS One 3*(1), e1513.

Gentleman, R. C., V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, and J. Zhang (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol 5*(10), R80.

Ghelardi, E., F. Celandroni, S. Salvetti, D. J. Beecher, M. Gominet, D. Lereclus, A. C. L. Wong, and S. Senesi (2002, Dec). Requirement of flhA for swarming differentiation, flagellin export, and secretion of virulence-associated proteins in *Bacillus thuringiensis*. *J Bacteriol 184*(23), 6424–6433.

Gilmour, M. W., M. Graham, G. V. Domselaar, S. Tyler, H. Kent, K. M. Trout-Yakel, O. Larios, V. Allen, B. Lee, and C. Nadon (2010). High-throughput genome sequencing of two *Listeria monocytogenes* clinical isolates during a large foodborne outbreak. *BMC Genomics 11*, 120.

Gohar, M., K. Faegri, S. Perchat, S. Ravnum, O. A. Økstad, M. Gominet, A.-B. Kolstø, and D. Lereclus (2008). The PlcR virulence regulon of *Bacillus cereus*. *PLoS One 3*(7), e2793.

Goodman, M., J. Czelusniak, G. W. Moore, A. E. Romero-Herrera, and G. Matsuda (1979). Fitting the Gene Lineage into its Species Lineage, a Parsimony Strategy Illustrated by Cladograms Constructed from Globin Sequences. *Systematic Zoology 28*(2), 132–163.

Gopal, S., I. Borovok, A. Ofer, M. Yanku, G. Cohen, W. Goebel, J. Kreft, and Y. Aharonowitz (2005, Jun). A multidomain fusion protein in *Listeria monocytogenes* catalyzes the two primary activities for glutathione biosynthesis. *J Bacteriol 187*(11), 3839–3847.

Hahn, M. W. (2007). Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biol 8*(7), R141.

Hahn, M. W., M. V. Han, and S.-G. Han (2007, Nov). Gene family evolution across 12 *Drosophila* genomes. *PLoS Genet 3*(11), e197.

Han, C. S., G. Xie, J. F. Challacombe, M. R. Altherr, S. S. Bhotika, N. Brown, D. Bruce, C. S. Campbell, M. L. Campbell, J. Chen, O. Chertkov, C. Cleland, M. Dimitrijevic, N. A. Doggett, J. J. Fawcett, T. Glavina, L. A. Goodwin, L. D. Green, K. K. Hill, P. Hitchcock, P. J. Jackson, P. Keim, A. R. Kewalramani, J. Longmire, S. Lucas, S. Malfatti, K. McMurry, L. J. Meincke, M. Misra, B. L. Moseman, M. Mundt, A. C. Munk, R. T. Okinaka, B. Parson-Quintana, L. P. Reilly, P. Richardson, D. L. Robinson, E. Rubin, E. Saunders, R. Tapia, J. G. Tesmer, N. Thayer, L. S. Thompson, H. Tice, L. O. Ticknor, P. L. Wills, T. S. Brettin, and P. Gilna (2006, May). Pathogenomic sequence analysis of *Bacillus cereus* and *Bacillus thuringiensis* isolates closely related to *Bacillus anthracis*. *J Bacteriol 188*(9), 3382–3390.

Hao, W. and G. B. Golding (2006, May). The fate of laterally transferred genes: life in the fast lane to adaptation or death. *Genome Res 16*(5), 636–643.

Hao, W. and G. B. Golding (2008a, Sep). High rates of lateral gene transfer are not due to false diagnosis of gene absence. *Gene 421*(1-2), 27–31.

Hao, W. and G. B. Golding (2008b). Uncovering rate variation of lateral gene transfer during bacterial genome evolution. *BMC Genomics 9*, 235.

Hao, W. and G. B. Golding (2009, Aug). Does gene translocation accelerate the evolution of laterally transferred genes? *Genetics 182*(4), 1365–1375.

Hao, W. and G. B. Golding (2010, Sep). Inferring bacterial genome flux while considering truncated genes. *Genetics 186*(1), 411–426.

Harris, S. R., E. J. Feil, M. T. G. Holden, M. A. Quail, E. K. Nickerson, N. Chantratita, S. Gardete, A. Tavares, N. Day, J. A. Lindsay, J. D. Edgeworth, H. de Lencastre, J. Parkhill, S. J. Peacock, and S. D. Bentley (2010, Jan). Evolution of MRSA during hospital transmission and intercontinental spread. *Science 327*(5964), 469–474.

Heger, A. and C. P. Ponting (2008, Jan). OPTIC: orthologous and paralogous transcripts in clades. *Nucleic Acids Res 36*(Database issue), D267–D270.

Hollich, V., C. E. V. Storm, and E. L. L. Sonnhammer (2002, Sep). OrthoGUI: graphical presentation of Orthostrapper results. *Bioinformatics 18*(9), 1272–1273.

Holt, K. E., J. Parkhill, C. J. Mazzoni, P. Roumagnac, F.-X. Weill, I. Goodhead, R. Rance, S. Baker, D. J. Maskell, J. Wain, C. Dolecek, M. Achtman, and G. Dougan (2008, Aug). High-throughput sequencing provides insights into genome variation and evolution in *Salmonella Typhi*. *Nat Genet 40*(8), 987–993.

Hooper, S. D. and O. G. Berg (2003, Jun). On the nature of gene innovation: duplication patterns in microbial genomes. *Mol Biol Evol 20*(6), 945–954.

Hua, S., T. Guo, J. Gough, and Z. Sun (2002, Jul). Proteins with class alpha/beta fold have high-level participation in fusion events. *J Mol Biol 320*(4), 713–719.

Hubbard, T. J. P., B. L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S. C. Dyer, S. Fitzgerald, J. Fernandez-Banet, S. Graf, S. Haider, M. Hammond, J. Herrero, R. Holland, K. Howe, K. Howe, N. Johnson, A. Kahari, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, C. Melsopp, K. Megy, P. Meidl, B. Ouverdin, A. Parker, A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, J. Severin, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, M. Wood, T. Cox, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, P. Flicek, A. Kasprzyk, G. Proctor, S. Searle, J. Smith, A. Ureta-Vidal, and E. Birney (2007, Jan). Ensembl 2007. *Nucleic Acids Res 35*(Database issue), D610–D617.

Huelsenbeck, J. P. and F. Ronquist (2001, Aug). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics 17*(8), 754–755.

Hughes, A. L. (1998, Jul). Phylogenetic tests of the hypothesis of block duplication of homologous genes on human chromosomes 6, 9, and 1. *Mol Biol Evol 15*(7), 854–870.

Hulsen, T., M. A. Huynen, J. de Vlieg, and P. M. A. Groenen (2006). Benchmarking ortholog identification methods using functional genomics data. *Genome Biol 7*(4), R31.

Ito, M., D. B. Hicks, T. M. Henkin, A. A. Guffanti, B. D. Powers, L. Zvi, K. Uematsu, and T. A. Krulwich (2004, Aug). MotPS is the stator-force generator for motility of alkaliphilic *Bacillus*, and its homologue is a second functional Mot in *Bacillus subtilis*. *Mol Microbiol 53*(4), 1035–1049.

Ito, M., N. Terahara, S. Fujinami, and T. A. Krulwich (2005, Sep). Properties of motility in *Bacillus subtilis* powered by the H+-coupled MotAB flagellar stator, Na+-coupled MotPS or hybrid stators MotAS or MotPB. *J Mol Biol 352*(2), 396–408.

Jothi, R., E. Zotenko, A. Tasneem, and T. M. Przytycka (2006, Apr). COCO-CL: hierarchical clustering of homology relations based on evolutionary correlations. *Bioinformatics 22*(7), 779–788.

Kamburov, A., L. Goldovsky, S. Freilich, A. Kapazoglou, V. Kunin, A. J. Enright, A. Tsaftaris, and C. A. Ouzounis (2007). Denoising inferred functional association networks obtained by gene fusion analysis. *BMC Genomics 8*, 460.

Kanehisa, M., S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa (2010, Jan). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res 38*(Database issue), D355–D360.

Keim, P., M. N. V. Ert, T. Pearson, A. J. Vogler, L. Y. Huynh, and D. M. Wagner (2004, Sep). Anthrax molecular epidemiology and forensics: using the appropriate marker for different evolutionary scales. *Infect Genet Evol 4*(3), 205–213.

Kim, K. M., S. Sung, G. Caetano-Anolls, J. Y. Han, and H. Kim (2008, Oct). An approach of orthology detection from homologous sequences under minimum evolution. *Nucleic Acids Res 36*(17), e110.

Kingsley, R. A., C. L. Msefula, N. R. Thomson, S. Kariuki, K. E. Holt, M. A. Gordon, D. Harris, L. Clarke, S. Whitehead, V. Sangal, K. Marsh, M. Achtman, M. E. Molyneux, M. Cormican, J. Parkhill, C. A. MacLennan, R. S. Heyderman, and G. Dougan (2009, Dec). Epidemic multiple drug resistant Salmonella Typhimurium causing invasive disease in sub-Saharan Africa have a distinct genotype. *Genome Res 19*(12), 2279–2287.

Klee, S. R., E. B. Brzuszkiewicz, H. Nattermann, H. Brggemann, S. Dupke, A. Wollherr, T. Franz, G. Pauli, B. Appel, W. Liebl, E. Couacy-Hymann, C. Boesch, F.-D. Meyer, F. H. Leendertz, H. Ellerbrok, G. Gottschalk, R. Grunow, and H. Liesegang (2010). The genome of a *Bacillus* isolate causing anthrax in chimpanzees combines chromosomal properties of *B. cereus* with *B. anthracis* virulence plasmids. *PLoS One 5*(7), e10986.

Klee, S. R., M. Ozel, B. Appel, C. Boesch, H. Ellerbrok, D. Jacob, G. Holland, F. H. Leendertz, G. Pauli, R. Grunow, and H. Nattermann (2006, Aug). Characterization of *Bacillus anthracis*-like bacteria isolated from wild great apes from Cote d'Ivoire and Cameroon. *J Bacteriol 188*(15), 5333–5344.

Ko, M., H. Choi, and C. Park (2002, Jul). Group I self-splicing intron in the recA gene of *Bacillus anthracis*. *J Bacteriol 184*(14), 3917–3922.

Kolstø, A.-B., N. J. Tourasse, and O. A. Økstad (2009). What sets *Bacillus anthracis* apart from other *Bacillus* species? *Annu Rev Microbiol 63*, 451–476.

Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet 39*, 309–338.

Koski, L. B. and G. B. Golding (2001, Jun). The closest BLAST hit is often not the nearest neighbor. *J Mol Evol 52*(6), 540–542.

Kristensen, D. M., L. Kannan, M. K. Coleman, Y. I. Wolf, A. Sorokin, E. V. Koonin, and A. Mushegian (2010, Jun). A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics 26*(12), 1481–1487.

Kristoffersen, S. M., N. J. Tourasse, A.-B. Kolstø, and O. A. Økstad (2011, Feb). Interspersed DNA Repeats bcr1-bcr18 of *Bacillus cereus* Group Bacteria Form Three Distinct Groups with Different Evolutionary and Functional Patterns. *Mol Biol Evol 28*(2), 963–983.

Krogh, A., B. Larsson, G. von Heijne, and E. L. Sonnhammer (2001, Jan). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol 305*(3), 567–580.

Kryazhimskiy, S. and J. B. Plotkin (2008, Dec). The population genetics of dN/dS. *PLoS Genet 4*(12), e1000304.

Kummerfeld, S. K. and S. A. Teichmann (2005, Jan). Relative rates of gene fusion and fission in multi-domain proteins. *Trends Genet 21*(1), 25–30.

Kunst, F., N. Ogasawara, I. Moszer, A. M. Albertini, G. Alloni, V. Azevedo, M. G. Bertero, P. Bessires, A. Bolotin, S. Borchert, R. Borriss, L. Boursier, A. Brans, M. Braun, S. C. Brignell, S. Bron, S. Brouillet, C. V. Bruschi, B. Caldwell, V. Capuano, N. M. Carter, S. K. Choi, J. J. Codani, I. F. Connerton, and A. Danchin (1997, Nov). The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature 390*(6657), 249–256.

Kuo, C.-H., N. A. Moran, and H. Ochman (2009, Aug). The consequences of genetic drift for bacterial genome complexity. *Genome Res 19*(8), 1450–1454.

Kuo, C.-H. and H. Ochman (2010, Aug). The extinction dynamics of bacterial pseudogenes. *PLoS Genet 6*(8), e1001050.

Kurtz, S., A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg (2004). Versatile and open software for comparing large genomes. *Genome Biol 5*(2), R12.

Kuzniar, A., R. C. H. J. van Ham, S. Pongor, and J. A. M. Leunissen (2008, Nov). The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet 24*(11), 539–551.

Lawrence, J. G., R. W. Hendrix, and S. Casjens (2001, Nov). Where are the pseudogenes in bacterial genomes? *Trends Microbiol 9*(11), 535–540.

Lerat, E. and H. Ochman (2005). Recognizing the pseudogenes in bacterial genomes. *Nucleic Acids Res 33*(10), 3125–3132.

Lewis, T., N. J. Loman, L. Bingle, P. Jumaa, G. M. Weinstock, D. Mortiboy, and M. J. Pallen (2010, May). High-throughput whole-genome sequencing to dissect the epidemiology of *Acinetobacter baumannii* isolates from a hospital outbreak. *J Hosp Infect 75*(1), 37–41.

Li, L., C. J. Stoeckert, and D. S. Roos (2003, Sep). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res 13*(9), 2178–2189.

Liu, H., N. H. Bergman, B. Thomason, S. Shallom, A. Hazen, J. Crossno, D. A. Rasko, J. Ravel, T. D. Read, S. N. Peterson, J. Yates, and P. C. Hanna (2004, Jan). Formation and composition of the *Bacillus anthracis* endospore. *J Bacteriol 186*(1), 164–178.

Liu, R. and H. Ochman (2007, Apr). Stepwise formation of the bacterial flagellar system. *Proc Natl Acad Sci U S A 104*(17), 7116–7121.

Liu, W.-Q., Y. Feng, Y. Wang, Q.-H. Zou, F. Chen, J.-T. Guo, Y.-H. Peng, Y. Jin, Y.-G. Li, S.-N. Hu, R. N. Johnston, G.-R. Liu, and S.-L. Liu (2009). *Salmonella paratyphi* C: genetic divergence from *Salmonella choleraesuis* and pathogenic convergence with *Salmonella typhi*. *PLoS One 4*(2), e4510.

Losada, L., C. M. Ronning, D. DeShazer, D. Woods, N. Fedorova, H. S. Kim, S. A. Shabalina, T. R. Pearson, L. Brinkac, P. Tan, T. Nandi, J. Crabtree, J. Badger, S. Beckstrom-Sternberg, M. Saqib, S. E. Schutzer, P. Keim, and W. C. Nierman (2010). Continuing evolution of *Burkholderia mallei* through genome reduction and large-scale rearrangements. *Genome Biol Evol 2010*, 102–116.

Lynch, M. and J. S. Conery (2000, Nov). The evolutionary fate and consequences of duplicate genes. *Science 290*(5494), 1151–1155.

Lynch, M. and J. S. Conery (2003, Nov). The origins of genome complexity. *Science 302*(5649), 1401–1404.

Makarova, K. S., A. V. Sorokin, P. S. Novichkov, Y. I. Wolf, and E. V. Koonin (2007). Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. *Biol Direct 2*, 33.

Marchler-Bauer, A., S. Lu, J. B. Anderson, F. Chitsaz, M. K. Derbyshire, C. DeWeese-Scott, J. H. Fong, L. Y. Geer, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, J. D. Jackson, Z. Ke, C. J. Lanczycki, F. Lu, G. H. Marchler, M. Mullokandov, M. V. Omelchenko, C. L. Robertson, J. S. Song, N. Thanki, R. A. Yamashita, D. Zhang, N. Zhang, C. Zheng, and S. H. Bryant (2011, Jan). CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res 39*(Database issue), D225–D229.

Marcotte, E. M., M. Pellegrini, H. L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg (1999, Jul). Detecting protein function and protein-protein interactions from genome sequences. *Science 285*(5428), 751–753.

Martin, J., W. Zhu, K. D. Passalacqua, N. Bergman, and M. Borodovsky (2010). *Bacillus anthracis* genome organization in light of whole transcriptome sequencing. *BMC Bioinformatics 11 Suppl 3*, S10.

McClelland, M., K. E. Sanderson, J. Spieth, S. W. Clifton, P. Latreille, L. Courtney, S. Porwollik, J. Ali, M. Dante, F. Du, S. Hou, D. Layman, S. Leonard, C. Nguyen, K. Scott, A. Holmes, N. Grewal, E. Mulvaney, E. Ryan, H. Sun, L. Florea, W. Miller, T. Stoneking, M. Nhan, R. Waterston, and R. K. Wilson (2001, Oct). Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2. *Nature 413*(6858), 852–856.

Meisel, R. P., M. V. Han, and M. W. Hahn (2009). A complex suite of forces drives gene traffic from *Drosophila* X chromosomes. *Genome Biol Evol 2009*, 176–188.

Merkeev, I. V., P. S. Novichkov, and A. A. Mironov (2006). PHOG: a database of supergenomes built from proteome complements. *BMC Evol Biol 6*, 52.

Mirel, D. B., V. M. Lustre, and M. J. Chamberlin (1992, Jul). An operon of *Bacillus subtilis* motility genes transcribed by the sigma D form of RNA polymerase. *J Bacteriol 174*(13), 4197–4204.

Monot, M., N. Honor, T. Garnier, N. Zidane, D. Sherafi, A. Paniz-Mondolfi, M. Matsuoka, G. M. Taylor, H. D. Donoghue, A. Bouwman, S. Mays, C. Watson, D. Lockwood, A. Khamesipour, A. Khamispour, Y. Dowlati, S. Jianping, T. H. Rea, L. Vera-Cabrera, M. M. Stefani, S. Banu, M. Macdonald, B. R. Sapkota, J. S. Spencer, J. Thomas, K. Harshman, P. Singh, P. Busso, A. Gattiker, J. Rougemont, P. J. Brennan, and S. T. Cole (2009, Dec). Comparative genomic and phylogeographic analysis of *Mycobacterium leprae*. *Nat Genet 41*(12), 1282–1289.

Moran, N. A., H. J. McLaughlin, and R. Sorek (2009, Jan). The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. *Science 323*(5912), 379–382.

Moran, N. A. and A. Mira (2001). The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. *Genome Biol 2*(12), RESEARCH0054.

Moran, N. A. and G. R. Plague (2004, Dec). Genomic changes following host restriction in bacteria. *Curr Opin Genet Dev 14*(6), 627–633.

Morelli, G., Y. Song, C. J. Mazzoni, M. Eppinger, P. Roumagnac, D. M. Wagner, M. Feldkamp, B. Kusecek, A. J. Vogler, Y. Li, Y. Cui, N. R. Thomson, T. Jombart, R. Leblois, P. Lichtner, L. Rahalison, J. M. Petersen, F. Balloux, P. Keim, T. Wirth, J. Ravel, R. Yang, E. Carniel, and M. Achtman (2010, Dec). *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. *Nat Genet 42*(12), 1140–1143.

Muller, J., D. Szklarczyk, P. Julien, I. Letunic, A. Roth, M. Kuhn, S. Powell, C. von Mering, T. Doerks, L. J. Jensen, and P. Bork (2010, Jan). eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res 38*(Database issue), D190–D195.

Nakamura, Y., T. Itoh, and W. Martin (2007, Jan). Rate and polarity of gene fusion and fission in *Oryza sativa* and *Arabidopsis thaliana*. *Mol Biol Evol 24*(1), 110–121.

Nord, D. and B.-M. Sjberg (2008, Jan). Unconventional GIY-YIG homing endonuclease encoded in group I introns in closely related strains of the *Bacillus cereus* group. *Nucleic Acids Res 36*(1), 300–310.

Nord, D., E. Torrents, and B.-M. Sjberg (2007, Jul). A functional homing endonuclease in the *Bacillus anthracis* nrdE group I intron. *J Bacteriol 189*(14), 5293–5301.

Ochman, H. and L. M. Davalos (2006, Mar). The nature and dynamics of bacterial genomes. *Science 311*(5768), 1730–1733.

Ogata, H., S. Audic, P. Renesto-Audiffren, P. E. Fournier, V. Barbe, D. Samson, V. Roux, P. Cossart, J. Weissenbach, J. M. Claverie, and D. Raoult (2001, Sep). Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. *Science 293*(5537), 2093–2098.

Ogura, Y., T. Ooka, A. Iguchi, H. Toh, M. Asadulghani, K. Oshima, T. Kodama, H. Abe, K. Nakayama, K. Kurokawa, T. Tobe, M. Hattori, and T. Hayashi (2009, Oct). Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic *Escherichia coli*. *Proc Natl Acad Sci U S A 106*(42), 17939–17944.

Omer, S., A. Kovacs, Y. Mazor, and U. Gophna (2010, Nov). Integration of a foreign gene into a native complex does not impair fitness in an experimental model of lateral gene transfer. *Mol Biol Evol 27*(11), 2441–2445.

Parkhill, J., G. Dougan, K. D. James, N. R. Thomson, D. Pickard, J. Wain, C. Churcher, K. L. Mungall, S. D. Bentley, M. T. Holden, M. Sebaihia, S. Baker, D. Basham, K. Brooks, T. Chillingworth, P. Connerton, A. Cronin, P. Davis, R. M. Davies, L. Dowd, N. White, J. Farrar, T. Feltwell, N. Hamlin, A. Haque, T. T. Hien, S. Holroyd, K. Jagels, A. Krogh, T. S. Larsen, S. Leather, S. Moule, P. O'Gaora, C. Parry, M. Quail, K. Rutherford, M. Simmonds, J. Skelton, K. Stevens, S. Whitehead, and B. G. Barrell (2001, Oct). Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature 413*(6858), 848–852.

Parkhill, J., B. W. Wren, N. R. Thomson, R. W. Titball, M. T. Holden, M. B. Prentice, M. Sebaihia, K. D. James, C. Churcher, K. L. Mungall, S. Baker, D. Basham, S. D. Bentley, K. Brooks, A. M. Cerdeo-Trraga, T. Chillingworth, A. Cronin, R. M. Davies, P. Davis, G. Dougan, T. Feltwell, N. Hamlin, S. Holroyd, K. Jagels, A. V. Karlyshev, S. Leather, S. Moule, P. C. Oyston, M. Quail, K. Rutherford, M. Simmonds, J. Skelton, K. Stevens, S. Whitehead, and B. G. Barrell (2001, Oct). Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature 413*(6855), 523–527.

Pasek, S., A. Bergeron, J.-L. Risler, A. Louis, E. Ollivier, and M. Raffinot (2005, Jun). Identification of genomic features using microsyntenies of domains: domain teams. *Genome Res 15*(6), 867–874.

Pasek, S., J.-L. Risler, and P. Brzellec (2006, Jun). Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. *Bioinformatics 22*(12), 1418–1423.

Passalacqua, K. D., A. Varadarajan, B. D. Ondov, D. T. Okou, M. E. Zwick, and N. H. Bergman (2009, May). Structure and complexity of a bacterial transcriptome. *J Bacteriol 191*(10), 3203–3211.

Pearson, T., J. D. Busch, J. Ravel, T. D. Read, S. D. Rhoton, J. M. U'Ren, T. S. Simonson, S. M. Kachur, R. R. Leadem, M. L. Cardon, M. N. V. Ert, L. Y. Huynh, C. M. Fraser, and P. Keim (2004, Sep). Phylogenetic discovery bias in *Bacillus anthracis* using single-nucleotide polymorphisms from whole-genome sequencing. *Proc Natl Acad Sci U S A 101*(37), 13536–13541.

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

Perkins, T. T., R. A. Kingsley, M. C. Fookes, P. P. Gardner, K. D. James, L. Yu, S. A. Assefa, M. He, N. J. Croucher, D. J. Pickard, D. J. Maskell, J. Parkhill, J. Choudhary, N. R. Thomson, and G. Dougan (2009, Jul). A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus *Salmonella typhi*. *PLoS Genet 5*(7), e1000569.

Price, M. N., K. H. Huang, E. J. Alm, and A. P. Arkin (2005). A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res 33*(3), 880–892.

Priest, F. G., M. Barker, L. W. J. Baillie, E. C. Holmes, and M. C. J. Maiden (2004, Dec). Population structure and evolution of the *Bacillus cereus* group. *J Bacteriol 186*(23), 7959–7970.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rambaut, A. and N. C. Grassly (1997, Jun). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput Appl Biosci 13*(3), 235–238.

Rasko, D. A., M. R. Altherr, C. S. Han, and J. Ravel (2005, Apr). Genomics of the *Bacillus cereus* group of organisms. *FEMS Microbiol Rev 29*(2), 303–329.

Rasko, D. A., J. Ravel, O. A. Økstad, E. Helgason, R. Z. Cer, L. Jiang, K. A. Shores, D. E. Fouts, N. J. Tourasse, S. V. Angiuoli, J. Kolonay, W. C. Nelson, A.-B. Kolstø, C. M. Fraser, and T. D. Read (2004). The genome sequence of *Bacillus cereus* ATCC 10987 reveals metabolic adaptations and a large plasmid related to *Bacillus anthracis* pXO1. *Nucleic Acids Res 32*(3), 977–988.

Rasmussen, M. D. and M. Kellis (2007, Dec). Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes. *Genome Res 17*(12), 1932–1942.

Rasmussen, M. D. and M. Kellis (2011, Jan). A bayesian approach for fast and accurate gene tree reconstruction. *Mol Biol Evol 28*(1), 273–290.

Ravel, J. and C. M. Fraser (2005, Mar). Genomics at the genus scale. *Trends Microbiol 13*(3), 95–97.

Ravel, J., L. Jiang, S. T. Stanley, M. R. Wilson, R. S. Decker, T. D. Read, P. Worsham, P. S. Keim, S. L. Salzberg, C. M. Fraser-Liggett, and D. A. Rasko (2009, Jan). The complete genome sequence of *Bacillus anthracis* Ames "Ancestor". *J Bacteriol 191*(1), 445–446.

Read, T. D., S. N. Peterson, N. Tourasse, L. W. Baillie, I. T. Paulsen, K. E. Nelson, H. Tettelin, D. E. Fouts, J. A. Eisen, S. R. Gill, E. K. Holtzapple, O. A. Økstad, E. Helgason, J. Rilstone, M. Wu, J. F. Kolonay, M. J. Beanan, R. J. Dodson, L. M. Brinkac, M. Gwinn, R. T. DeBoy, R. Madpu, S. C. Daugherty, A. S. Durkin, D. H. Haft, W. C. Nelson, J. D. Peterson, M. Pop, H. M. Khouri, D. Radune, J. L. Benton, Y. Mahamoud, L. Jiang, I. R. Hance, J. F. Weidman, K. J. Berry, R. D. Plaut, A. M. Wolf, K. L. Watkins, W. C. Nierman, A. Hazen, R. Cline, C. Redmond, J. E. Thwaite, O. White, S. L. Salzberg, B. Thomason, A. M. Friedlander, T. M. Koehler, P. C. Hanna, A.-B. Kolstø, and C. M.

Fraser (2003, May). The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. *Nature 423*(6935), 81–86.

Remm, M., C. E. Storm, and E. L. Sonnhammer (2001, Dec). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol 314*(5), 1041–1052.

Rey, M. W., P. Ramaiya, B. A. Nelson, S. D. Brody-Karpin, E. J. Zaretsky, M. Tang, A. L. de Leon, H. Xiang, V. Gusti, I. G. Clausen, P. B. Olsen, M. D. Rasmussen, J. T. Andersen, P. L. Jrgensen, T. S. Larsen, A. Sorokin, A. Bolotin, A. Lapidus, N. Galleron, S. D. Ehrlich, and R. M. Berka (2004). Complete genome sequence of the industrial bacterium *Bacillus licheniformis* and comparisons with closely related *Bacillus* species. *Genome Biol 5*(10), R77.

Rice, P., I. Longden, and A. Bleasby (2000, Jun). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet 16*(6), 276–277.

Roth, A. C. J., G. H. Gonnet, and C. Dessimoz (2008). Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics 9*, 518.

Roth, C., M. J. Betts, P. Steffansson, G. Saelensminde, and D. A. Liberles (2005, Jan). The Adaptive Evolution Database (TAED): a phylogeny based tool for comparative genomics. *Nucleic Acids Res 33*(Database issue), D495–D497.

Ruan, J., H. Li, Z. Chen, A. Coghlan, L. J. M. Coin, Y. Guo, J.-K. Hrich, Y. Hu, K. Kristiansen, R. Li, T. Liu, A. Moses, J. Qin, S. Vang, A. J. Vilella, A. Ureta-Vidal, L. Bolund, J. Wang, and R. Durbin (2008, Jan). TreeFam: 2008 Update. *Nucleic Acids Res 36*(Database issue), D735–D740.

Sakharkar, K. R., M. K. Sakharkar, and V. T. K. Chow (2006, Jan). Gene fusion in *Helicobacter pylori*: making the ends meet. *Antonie Van Leeuwenhoek 89*(1), 169–180.

Salvetti, S., E. Ghelardi, F. Celandroni, M. Ceragioli, F. Giannessi, and S. Senesi (2007, Aug). FlhF, a signal recognition particle-like GTPase, is involved in the regulation of flagellar arrangement, motility behaviour and protein secretion in *Bacillus cereus*. *Microbiology 153*(Pt 8), 2541–2552.

Saw, J. H., B. W. Mountain, L. Feng, M. V. Omelchenko, S. Hou, J. A. Saito, M. B. Stott, D. Li, G. Zhao, J. Wu, M. Y. Galperin, E. V. Koonin, K. S. Makarova, Y. I. Wolf, D. J. Rigden, P. F. Dunfield, L. Wang, and M. Alam (2008). Encapsulated in silica: genome, proteome and physiology of the thermophilic bacterium *Anoxybacillus flavithermus* WK1. *Genome Biol 9*(11), R161.

Schatz, M. C., A. L. Delcher, and S. L. Salzberg (2010, Sep). Assembly of large genomes using second-generation sequencing. *Genome Res 20*(9), 1165–1173.

Sennblad, B. and J. Lagergren (2009, Aug). Probabilistic orthology analysis. *Syst Biol 58*(4), 411–424.

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

Skamrov, A., E. Feoktistova, M. Goldman, and R. Beabealashvilli (2001, Jun). Gene rearrangement and fusion in *Mycoplasma gallisepticum* thyA-nrdFEI locus. *FEMS Microbiol Lett 200*(1), 31–35.

Slamti, L., S. Perchat, M. Gominet, G. Vilas-Bas, A. Fouet, M. Mock, V. Sanchis, J. Chaufaux, M. Gohar, and D. Lereclus (2004, Jun). Distinct mutations in PlcR explain why some strains of the *Bacillus cereus* group are nonhemolytic. *J Bacteriol 186*(11), 3531–3538.

Smith, M. G., T. A. Gianoulis, S. Pukatzki, J. J. Mekalanos, L. N. Ornston, M. Gerstein, and M. Snyder (2007, Mar). New insights into *Acinetobacter baumannii* pathogenesis revealed by high-density pyrosequencing and transposon mutagenesis. *Genes Dev 21*(5), 601–614.

Snel, B., P. Bork, and M. Huynen (2000, Jan). Genome evolution. Gene fusion versus gene fission. *Trends Genet 16*(1), 9–11.

Sonnhammer, E. L. L. and E. V. Koonin (2002, Dec). Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet 18*(12), 619–620.

Soufiane, B., D. Xu, and J.-C. Ct (2007, Nov). Flagellin (FliC) protein sequence diversity among *Bacillus thuringiensis* does not correlate with H serotype diversity. *Antonie Van Leeuwenhoek 92*(4), 449–461.

Stajich, J. E., D. Block, K. Boulez, S. E. Brenner, S. A. Chervitz, C. Dagdigian, G. Fuellen, J. G. R. Gilbert, I. Korf, H. Lapp, H. Lehvslaiho, C. Matsalla, C. J. Mungall, B. I. Osborne, M. R. Pocock, P. Schattner, M. Senger, L. D. Stein, E. Stupka, M. D. Wilkinson, and E. Birney (2002, Oct). The Bioperl toolkit: Perl modules for the life sciences. *Genome Res 12*(10), 1611–1618.

Stein, L. D. (2010, May). The case for cloud computing in genome informatics. *Genome Biol 11*(5), 207.

Storm, C. E. V. and E. L. L. Sonnhammer (2002, Jan). Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics 18*(1), 92–99.

Storm, C. E. V. and E. L. L. Sonnhammer (2003, Oct). Comprehensive analysis of orthologous protein domains using the HOPS database. *Genome Res 13*(10), 2353–2362.

Stover, N. A., A. R. O. Cavalcanti, A. J. Li, B. C. Richardson, and L. F. Landweber (2005, Jul). Reciprocal fusions of two genes in the formaldehyde detoxification pathway in ciliates and diatoms. *Mol Biol Evol 22*(7), 1539–1542.

Strope, C. L., K. Abel, S. D. Scott, and E. N. Moriyama (2009, Nov). Biological sequence simulation for testing complex evolutionary hypotheses: indel-Seq-Gen version 2.0. *Mol Biol Evol 26*(11), 2581–2593.

Suhre, K. and J.-M. Claverie (2004, Jan). FusionDB: a database for in-depth analysis of prokaryotic gene fusion events. *Nucleic Acids Res 32*(Database issue), D273–D276.

Tatusov, R. L., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale (2003, Sep). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics 4*, 41.

Tatusov, R. L., E. V. Koonin, and D. J. Lipman (1997, Oct). A genomic perspective on protein families. *Science 278*(5338), 631–637.

Terahara, N., T. A. Krulwich, and M. Ito (2008, Sep). Mutations alter the sodium versus proton use of a *Bacillus clausii* flagellar motor and confer dual ion use on *Bacillus subtilis* motors. *Proc Natl Acad Sci U S A 105*(38), 14359–14364.

Thomson, N. R., D. J. Clayton, D. Windhorst, G. Vernikos, S. Davidson, C. Churcher, M. A. Quail, M. Stevens, M. A. Jones, M. Watson, A. Barron, A. Layton, D. Pickard, R. A. Kingsley, A. Bignell, L. Clark, B. Harris, D. Ormond, Z. Abdellah, K. Brooks, I. Cherevach, T. Chillingworth, J. Woodward, H. Norberczak, A. Lord, C. Arrowsmith, K. Jagels, S. Moule, K. Mungall, M. Sanders, S. Whitehead, J. A. Chabalgoity, D. Maskell, T. Humphrey, M. Roberts, P. A. Barrow, G. Dougan, and J. Parkhill (2008, Oct). Comparative genome analysis of Salmonella Enteritidis PT4 and Salmonella Gallinarum 287/91 provides insights into evolutionary and host adaptation pathways. *Genome Res 18*(10), 1624–1637.

Tourasse, N. J. and A.-B. Kolstø (2008, Aug). Survey of group I and group II introns in 29 sequenced genomes of the *Bacillus cereus* group: insights into their spread and evolution. *Nucleic Acids Res 36*(14), 4529–4548.

van der Heijden, R. T. J. M., B. Snel, V. van Noort, and M. A. Huynen (2007). Orthology prediction at scalable resolution by phylogenetic tree analysis. *BMC Bioinformatics 8*, 83.

van Dongen, S. (2000). *Graph clustering by flow simulation*. Ph. D. thesis, University of Utrecht, The Netherlands.

Van Ert, M. N., W. R. Easterday, L. Y. Huynh, R. T. Okinaka, M. E. Hugh-Jones, J. Ravel, S. R. Zanecki, T. Pearson, T. S. Simonson, J. M. U'Ren, S. M. Kachur, R. R. Leadem-Dougherty, S. D. Rhoton, G. Zinser, J. Farlow, P. R. Coker, K. L. Smith, B. Wang, L. J. Kenefic, C. M. Fraser-Liggett, D. M. Wagner, and P. Keim (2007). Global genetic population structure of *Bacillus anthracis*. *PLoS One 2*(5), e461.

Veith, B., C. Herzberg, S. Steckel, J. Feesche, K. H. Maurer, P. Ehrenreich, S. Bumer, A. Henne, H. Liesegang, R. Merkl, A. Ehrenreich, and G. Gottschalk (2004). The complete genome sequence of *Bacillus licheniformis* DSM13, an organism with great industrial potential. *J Mol Microbiol Biotechnol 7*(4), 204–211.

Vilella, A. J., J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin, and E. Birney (2009, Feb). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res 19*(2), 327–335.

Wall, D. P., H. B. Fraser, and A. E. Hirsh (2003, Sep). Detecting putative orthologs. *Bioinformatics 19*(13), 1710–1711.

Wall, D. P., P. Kudtarkar, V. A. Fusaro, R. Pivovarov, P. Patil, and P. J. Tonellato (2010). Cloud computing for comparative genomics. *BMC Bioinformatics 11*, 259.

Waltman, P., T. Kacmarczyk, A. R. Bate, D. B. Kearns, D. J. Reiss, P. Eichenberger, and R. Bonneau (2010, Sep). Multi-species integrative biclustering. *Genome Biol 11*(9), R96.

Wapinski, I., A. Pfeffer, N. Friedman, and A. Regev (2007, Jul). Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics 23*(13), i549–i558.

Weiner, J. and E. Bornberg-Bauer (2006, Apr). Evolution of circular permutations in multidomain proteins. *Mol Biol Evol 23*(4), 734–743.

Williams, D. L., R. A. Slayden, A. Amin, A. N. Martinez, T. L. Pittman, A. Mira, A. Mitra, V. Nagaraja, N. E. Morrison, M. Moraes, and T. P. Gillis (2009). Implications of high level pseudogene transcription in *Mycobacterium leprae*. *BMC Genomics 10*, 397.

Wolf, Y. I., I. B. Rogozin, A. S. Kondrashov, and E. V. Koonin (2001, Mar). Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res 11*(3), 356–372.

Wu, D., P. Hugenholtz, K. Mavromatis, R. Pukall, E. Dalin, N. N. Ivanova, V. Kunin, L. Goodwin, M. Wu, B. J. Tindall, S. D. Hooper, A. Pati, A. Lykidis, S. Spring, I. J. Anderson, P. D'haeseleer, A. Zemla, M. Singer, A. Lapidus, M. Nolan, A. Copeland, C. Han, F. Chen, J.-F. Cheng, S. Lucas, C. Kerfeld, E. Lang, S. Gronow, P. Chain, D. Bruce, E. M. Rubin, N. C. Kyrpides, H.-P. Klenk, and J. A. Eisen (2009, Dec). A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature 462*(7276), 1056–1060.

Xu, D. and J.-C. Ct (2008, Sep). Sequence diversity of *Bacillus thuringiensis* flagellin (H antigen) protein at the intra-H serotype level. *Appl Environ Microbiol 74*(17), 5524–5532.

Yanai, I., A. Derti, and C. DeLisi (2001, Jul). Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. *Proc Natl Acad Sci U S A 98*(14), 7940–7945.

Yang, Z. (1998, May). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol 15*(5), 568–573.

Yang, Z. (2007, Aug). PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol 24*(8), 1586–1591.

Yu, G. X. (2009). Pathogenic *Bacillus anthracis* in the progressive gene losses and gains in adaptive evolution. *BMC Bioinformatics 10 Suppl 1*, S3.

Yu, Y.-K. and S. F. Altschul (2005, Apr). The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions. *Bioinformatics 21*(7), 902–911.

Zakharova, N., B. J. Paster, I. Wesley, F. E. Dewhirst, D. E. Berg, and K. V. Severinov (1999, Jun). Fused and overlapping rpoB and rpoC genes in Helicobacters, Campylobacters, and related bacteria. *J Bacteriol 181*(12), 3857–3859.

Zheng, D. and M. B. Gerstein (2007, May). The ambiguous boundary between genes and pseudogenes: the dead rise up, or do they? *Trends Genet 23*(5), 219–224.

Zmasek, C. M. and S. R. Eddy (2001, Sep). A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics 17*(9), 821–828.

Zmasek, C. M. and S. R. Eddy (2002, May). RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics 3*, 14.

Part IV

APPENDICES

Appendix A

P-values for program comparisons from Chapter 1.

The following are the P-values for the Wilcoxon signed-rank tests performed in Chapter 1 to compare performance of ortholog identification programs under lineage-specific duplication, duplication, and concurrent duplication and loss. In all three sets of comparisons the event rates were set at an average of 0.1.

Table A.1: *P*-values for precision and recall at a lineage-specific duplication rate of 0.1.

	om	mp	no	nm	n0	n1	n2	ro	rm	r0	r1	r2
om	x	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00
mp	4.94e-08	x	4.94e-01	9.04e-01	9.04e-01	9.04e-01	4.94e-01	4.94e-01	4.94e-01	4.94e-01	9.04e-01	9.04e-01
no	2.83e-08	4.99e-08	x	1.00e+00	1.00e+00	1.00e+00	NA	NA	NA	NA	1.00e+00	1.00e+00
nm	4.93e-08	2.93e-08	4.97e-08	x	NA	NA	1.00e+00	1.00e+00	1.00e+00	1.00e+00	NA	NA
n0	4.99e-08	4.97e-08	5.00e-08	4.90e-08	x	NA	1.00e+00	1.00e+00	1.00e+00	1.00e+00	NA	NA
n1	4.99e-08	4.97e-08	5.00e-08	4.90e-08	NA	x	1.00e+00	1.00e+00	1.00e+00	1.00e+00	NA	NA
n2	5.04e-08	5.04e-08	5.03e-08	5.03e-08	5.05e-08	5.05e-08	x	NA	NA	NA	1.00e+00	1.00e+00
ro	4.81e-08	5.02e-08	4.85e-08	5.02e-08	5.00e-08	5.00e-08	5.03e-08	x	NA	NA	1.00e+00	1.00e+00
rm	5.69e-08	4.86e-08	4.95e-08	4.77e-08	4.98e-08	4.98e-08	5.04e-08	5.00e-08	x	NA	1.00e+00	1.00e+00
r0	4.99e-08	1.00e+00	5.01e-08	4.21e-04	4.96e-08	4.96e-08	5.04e-08	5.03e-08	4.87e-08	x	1.00e+00	1.00e+00
r1	4.99e-08	4.95e-08	4.97e-08	4.96e-08	5.00e-08	5.00e-08	5.02e-08	5.02e-08	4.97e-08	4.97e-08	x	NA
r2	5.04e-08	4.93e-08	5.04e-08	4.93e-08	4.92e-08	4.92e-08	5.05e-08	5.04e-08	4.99e-08	4.92e-08	4.99e-08	x

Table A.2: *P*-values for precision and recall at a duplication rate of 0.1.

	om	mp	no	nm	n0	n1	n2	ro	rm	r0	r1	r2
om	x	6.67e-08	3.01e-06	4.62e-03	5.48e-06	5.48e-06	4.24e-08	5.87e-08	8.30e-08	1.78e-07	5.08e-07	4.10e-08
mp	4.99e-08	x	6.72e-08	6.58e-08	4.69e-08	4.69e-08	4.82e-08	7.05e-08	6.90e-08	4.77e-08	4.72e-08	4.73e-08
no	4.32e-08	4.99e-08	x	2.23e-07	5.68e-01	5.68e-01	5.11e-08	6.43e-07	1.40e-06	3.16e-05	1.80e-03	4.84e-08
nm	4.97e-08	3.96e-08	5.00e-08	x	1.90e-07	1.90e-07	4.59e-08	1.40e-07	6.49e-08	6.37e-08	5.79e-08	4.55e-08
n0	5.01e-08	4.76e-05	5.04e-08	1.38e-07	x	NA	1.10e-07	3.26e-05	1.19e-04	1.29e-04	1.29e-02	1.12e-07
n1	5.01e-08	4.76e-05	5.04e-08	1.38e-07	NA	x	1.10e-07	3.26e-05	1.19e-04	1.29e-04	1.29e-02	1.12e-07
n2	5.05e-08	5.05e-08	5.05e-08	5.03e-08	5.05e-08	5.05e-08	x	2.95e-07	7.45e-07	3.85e-07	1.96e-07	1.00e+00
ro	4.91e-08	4.99e-08	4.48e-08	5.00e-08	5.02e-08	5.02e-08	5.06e-08	x	1.00e+00	7.47e-02	7.92e-03	2.55e-07
rm	7.32e-08	4.85e-08	5.00e-08	4.79e-08	5.00e-08	5.00e-08	5.05e-08	5.00e-08	x	2.74e-01	1.36e-02	6.16e-07
r0	5.00e-08	4.17e-01	5.02e-08	1.39e-06	6.09e-03	6.09e-03	5.05e-08	5.01e-08	4.96e-08	x	5.44e-01	3.61e-07
r1	5.00e-08	4.95e-08	5.04e-08	4.93e-08	5.01e-08	5.01e-08	5.70e-08	5.01e-08	4.97e-08	4.88e-08	x	3.00e-07
r2	5.04e-08	4.99e-08	5.04e-08	4.98e-08	4.97e-08	4.97e-08	5.04e-08	5.03e-08	4.99e-08	4.92e-08	4.99e-08	x

Table A.3: *P*-values for precision and recall at a duplication and loss rate of 0.1.

	om	mp	no	nm	n0	n1	n2	ro	rm	r0	r1	r2
om	x	4.85e-08	1.05e-06	1.17e-03	1.01e-06	9.44e-07	7.48e-08	4.22e-08	4.03e-08	4.12e-08	8.60e-08	4.43e-08
mp	4.96e-08	x	4.94e-08	4.77e-08	1.47e-02	6.67e-02	3.58e-03	4.91e-08	4.91e-08	4.88e-08	4.82e-08	4.95e-08
no	4.41e-08	5.05e-08	x	1.37e-07	1.18e-07	1.13e-07	5.04e-08	7.44e-08	3.37e-07	7.50e-07	6.67e-05	8.50e-08
nm	4.94e-08	3.63e-08	4.98e-08	x	5.83e-06	2.22e-06	6.39e-05	4.72e-08	4.47e-08	4.56e-08	4.67e-08	4.69e-08
n0	5.10e-08	6.23e-04	5.12e-08	3.03e-07	x	4.14e-02	1.00e+00	4.99e-08	4.98e-08	4.94e-08	4.97e-08	4.99e-08
n1	8.34e-06	6.29e-08	2.56e-07	4.99e-06	5.04e-08	x	1.00e+00	4.97e-08	4.94e-08	4.98e-08	4.97e-08	4.95e-08
n2	5.13e-08	5.10e-08	5.13e-08	5.12e-08	5.12e-08	5.10e-08	x	4.81e-08	4.80e-08	4.77e-08	4.76e-08	4.79e-08
ro	4.88e-08	5.05e-08	4.85e-08	5.03e-08	5.10e-08	6.14e-08	5.11e-08	x	1.00e+00	6.85e-02	6.98e-04	1.80e-02
rm	8.68e-08	4.86e-08	5.06e-08	4.79e-08	5.09e-08	5.44e-01	5.11e-08	5.03e-08	x	1.00e+00	2.18e-02	7.42e-03
r0	5.08e-08	1.00e+00	5.11e-08	7.87e-07	1.15e-03	5.57e-08	5.10e-08	5.11e-08	4.95e-08	x	5.05e-02	4.30e-04
r1	5.10e-08	5.09e-08	5.11e-08	5.08e-08	5.08e-08	5.10e-08	5.10e-08	5.12e-08	5.07e-08	5.07e-08	x	8.85e-06
r2	5.06e-08	5.09e-08	5.11e-08	5.07e-08	5.09e-08	5.11e-08	5.12e-08	5.13e-08	5.10e-08	5.08e-08	5.08e-08	x

Appendix B

Bacillus anthracis canSNPs.

Table B.1: *B. anthracis* canSNPs identified in 11 complete and draft genomes as described in Chapter 2. IDs correspond to those in Chapter 2 and the lineage and canSNP names correspond to the naming scheme by Van Ert *et al.* (2007). Note that *B. anthracis* Sterne (Ba7) was incorrectly listed as belonging to lineage A1 by Van Ert *et al.* (2007) but actually belongs to A2 as listed in the supplementary information in a previous study (Pearson *et al.*, 2004) and confirmed by our SNP analysis.

ID	Lineage	A.1	A.2	A.3	A.4	A.6	A.7	A.8	A.9	B.1	B.2	B.3	B.4	A/B.1
Ba1	C	T	G	A	T	C	T	T	A	T	G	G	T	G
Ba2	B1	T	G	A	T	C	T	T	A	C	T	A	T	A
Ba3	B2	T	G	A	T	C	T	T	A	T	G	A	C	A
Ba4	A1	T	G	A	T	A	T	G	G	T	G	G	T	A
Ba5	A1	T	G	A	T	A	C	T	A	T	G	G	T	A
Ba6	A1	T	G	A	T	A	C	T	A	T	G	G	T	A
Ba7	A2	T	G	G	C	A	T	T	A	T	G	G	T	A
Ba8	A2	T	A	G	C	A	T	T	A	T	G	G	T	A
Ba9	A2	C	A	G	C	A	T	T	A	T	G	G	T	A
Ba10	A2	C	A	G	C	A	T	T	A	T	G	G	T	A
Ba11	A2	C	A	G	C	A	T	T	A	T	G	G	T	A