IMPROVING THE DETECTION OF ORTHOLOGS BY ALTERING AND COMPARING VARIOUS METHODS IN BIOINFORMATICS

A Thesis Submitted to
Wilfrid Laurier University

by

KRISTEN LATIMER

In Partial Fulfillment
for the Degree of Honours Bachelor of Science
in the Department of Biology
Waterloo, Ontario

Abstract

In comparative genomics, detecting orthologs (the "same gene" in different species) is essential for inferring gene function across species. This detection is typically based on working definitions rather than on phylogenetic trees, as the latter is too computationally demanding. The traditional working definition of orthology is termed Reciprocal Best Hits (RBH), whereby two genes in two different organisms are deemed orthologs if they find each other as the best possible homologs in the opposite organism. This study attempts to improve the detection of orthologs as RBH by manipulating options within BLAST. This program takes a *query* sequence and searches for similar sequences within a database and then aligns them. The score measuring similarity may be primarily affected by two options, namely the low-information filter, and the algorithm producing the final alignment. The filter removes (masks) stretches of repetitive amino acid sequences. A hard filter masks these low-information subsequences during both the search and the alignment phases, while a *soft filter* masks these subsequences only during the search phase. The alignment option allows the user to choose the Smith-Waterman algorithm to align sequences as opposed to the default *matching words*-based alignment. Once homologs were identified, a program written in PERL was used to detect RBH and count orthologs. Results from running *Escherichia coli* K12 against a database of genomes showed that the highest numbers of orthologous were generated using a soft filter and the Smith-Waterman algorithm.

A recently proposed working definition of orthology called reciprocal smallest distance (RSD), uses evolutionary distances between sequences to find putative orthologs. This method is more complex and time consuming than the RBH definition, but was reported to generate improved results. Here, the RBH and RSD definitions were also compared. The results were inconclusive with respect to the overall number of orthologs detected. However, the estimated error of these two working definitions indicated that RSD had a significantly higher proportion of missed orthologs (an ortholog exists, but was not detected). Errors related to mislabelled pairs (a paralog identified as an ortholog and *vice versa*) were significantly different in each condition but did not indicate a dominant pattern.

Acknowledgements

I would like to thank Dr. Gabriel Moreno-Hagelsieb for introducing me to the world of bioinformatics and for providing me with an opportunity to complete this undergraduate thesis. I am especially grateful for his patience and guidance throughout the past eight months as I encountered novel concepts and developed my skills in programming using the PERL language.

I am also indebted to Meghan Martin for her time spent brainstorming and hashing out ideas in the lab. In many cases, two heads truly are better than one.

Table of Contents

List of Figures

List of Tables

Abbreviations and Definitions

BLAST– Basic Local Alignment Search Tool
A commonly used computer program in bioinformatics.

RBH– Reciprocal Best Hit
The traditional working definition of orthology which defines orthologs as having the highest homology score.

RSD– Reciprocal Smallest Distance
A newly defined working definition of orthology which considers the evolutionary distance between two sequences.

FTsF– A tested RBH condition with the hard filter setting and the BLAST alignment.

FTsT– A tested RBH condition with the hard filter setting and the Smith-Waterman alignment algorithm.

FmSsF– A tested RBH condition with the soft filter setting and the BLAST alignment.

FmSsT– A tested RBH condition with the soft filter setting and the Smith-Waterman alignment algorithm.

GSS– Genome Similarity Score
A calculated value representing the evolutionary distance between two genomes.

1 Introduction

1.1 Bioinformatics

Technological advances have greatly influenced and developed many different aspects of life over the past two decades. This trend has also been reflected across different fields of science, including biology. Some resulting developments of improved technology include: the creation of large online resources to aid scientists in sharing knowledge and ideas, automated laboratory techniques, and the development of efficient analytical tools able to perform tasks that previously were impossible to process by hand [1]. This increased role of technology within science has paved the way for the development of bioinformatics as a sub-discipline of biology. Bioinformatics integrates concepts of math and computer programming to create a new approach to biological research, such as that of comparative genomics [2]. Specifically, bioinformatics employs the use of technology in the management of constantly-growing databases containing information such as sequenced genomes, as well as the analysis of these data [1].

Currently, researchers have sequenced the genomes of many organisms, including the genomes of model organisms [2]. Model organisms are often studied in great detail and are used to make inferences about other organisms with similar characteristics. This concept has been adapted in bioinformatics, which conducts large computer-based sequence searches and analyses using the data found within various genomic databases to identify homologous genes across species [3]. In molecular genetics, homology is the evolutionary relationship between two macromolecular (DNA, RNA, protein) sequences that have arisen from a common ancestral gene [4]. Thus, these processes employ evolutionary pathways in an attempt to determine the function of genes in relation to what is already known. By identifying homologous

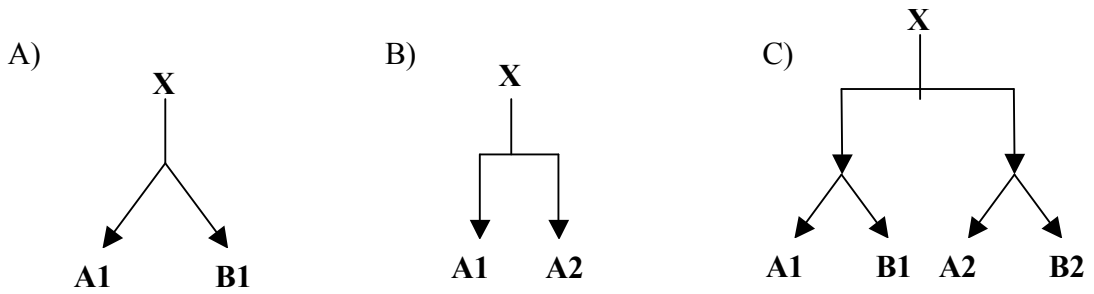

Figure 1.1: Potential events resulting in divergent evolution. (A) Speciation results in gene X separating into orthologous genes A1 and B1. (B) a gene duplication results in paralogous gene sequences A1 and A2 within a genome. (C) A speciation event following a duplication event results in the formation of both orthologs (e.g. A1 and B1), as well as paralogs (e.g. A1 and A2) across different organisms. This figure is based on those by Fitch [4]

genes, scientists can potentially infer the function of genes across species. This reduces the amount of time spent on tedious laboratory experiments, which typically involves knocking out or silencing genes in order to determine their roles.

1.2 Orthology

As time progresses, several events may occur that result in the divergent evolution of homologous genes. The first potential event is referred to as speciation (see Figure 1.1A). When this occurs, the common ancestral genes are represented within the genome of each of the two resulting species and these genes are referred to as orthologs [4]. In other words, orthologs are essentially the same gene found within two different species. Thus, for any given gene, only one corresponding true ortholog should exist within the genome of another organism. Although mutations (primarily substitutions, insertions and/or deletions) may occur, there is an overall tendency for the conservation of the genes with regards to its sequence and its function [2]. As function tends to be conserved, orthologs can be used by scientists for inferring gene function across species.

Another evolutionary event which may occur is a gene duplication, which results in two copies of a given gene within the same genome (see Figure 1.1B). As there is typically no need for two genes playing the same role, one copy of the gene might

diverge from the other and take on a different function [3]. These genes cannot be used to infer gene function and are defined as paralogs [4]. Any homologous genes found strictly within the genome of a single organism is guaranteed be a paralog. However, paralogs can also be found among genes belonging to different organisms [5]. As evolutionary steps continue, a speciation event may occur following a duplication such that different copies of the gene are separated into different species. This can result in genes within different genomes sharing homology, but having no conserved role [4] (Figure 1.1C). To complicate matters, one copy of the gene can be lost in one species, causing the disappearance of orthologous and/or paralogous pairs, which may further complicate the inference of gene function [5]. A third kind of homolog, the xenolog, arises when a gene from a different organism is transferred into another organism [4]. Horizontal gene transfer, though interesting, will not be discussed in this work. In order to infer gene function across organisms, it is essential to have tools that can detect these evolutionary events.

1.3 Basic Local Alignment Search Tool

In order to distinguish orthologs from extra-paralogs (paralogs residing in different organisms [6]), several computer programs have been written for comparing and scoring genes. Since all analysis and resulting inferences begin by detecting homologs of "query" genes, it is critical that the methods used are as accurate as possible [7]. However, the specific time required for each search is directly proportional to the size of the database searched as well as the length of the query sequence [8]. With the continuously increasing size of genomic databases, computer search programs often require large amounts of time to perform an accurate search due to technological and computational limitations such as processing speed. As a result, search programs are often forced to sacrifice some accuracy to speed up the overall search process and reduce the computational costs [8].

BLAST (Basic Local Alignment Search Tool) is probably the most commonly used program for the detection of homologs and is continuously being updated as

knowledge advances [9]. Different versions of this program have been created such that search runs can be conducted with a variety of sequence types (e.g. sequences of DNA, RNA, or proteins) [9]. BLAST has also served as a template for other programs [10]. BLAST works by comparing sequences based on both sequence identity (the same amino acid is located in the same position in both sequences) and sequence similarity (which, when comparing proteins, accounts for amino acids with similar characteristics) [9]. This program conducts a search in two main phases: a search phase and an alignment phase. The search phase takes the query sequence and breaks it down into smaller segments. It then searches for matching sequences to these fragments within a database of different genomes. Any sequence in the database showing similarity above a threshold level is added to the compiled list of potential homologs [9]. Once all potential homologs are identified, the program moves into the alignment phase, where it compares each homolog to its query sequence and produces the final score. Protein sequence pairs are scored in relation to their similarity with respect to matched, mismatched, and mutated amino acids.

Due to the degeneracy of the genetic code, protein sequences are better suited for detecting homologs by sequence comparison than the corresponding gene sequence. Several options within BLAST can be manipulated and may influence the ability of the program to detect and align homologs. Two main options might be important for the task of detecting orthologs as reciprocal best hits, because they have the greatest effect on the final score. One such option is the use of a filter, or "mask" to remove low complexity regions. These regions include areas of repeated amino acids which are likely to generate many spurious hits during the initial search process but are not commonly identified as part of an homologous pair [9]. By removing low-information regions, BLAST avoids unnecessary drain on computational speed. However, if these low-information regions are also excluded during the alignment phase, some true homologs might be missed, or those detected will not be adequately scored to define them as orthologs. The sequence filter in BLAST is set as a hard filter by default (-F T in NCBI's BLAST), which removes low-information amino acid sequences during both the search and the alignment phases. However, this filter

may also be set as a soft filter (*-F "m S"*), which removes low information regions only during the search phase of BLAST, but not during the alignment phase.

The second option (available only for protein-protein comparisons) changes the method of alignment employed during the second phase of BLAST. The default choice uses the matching words found in the first phase to extend and align the sequences. BLAST also offers the option of producing a final Smith-Waterman alignment (*-s T* option). The Smith-Waterman algorithm was first described in 1981, and is a mathematical process that will generate the best alignment between two sequences given a substitution matrix [11]. This algorithm works by incorporating a generalized version of the dynamic programming method by breaking the main problem into smaller subproblems until more obvious answers can be easily assigned [9]. It then works in a bottom-up fashion, using the answers of these trivial subproblems to determine the answers to larger and more complex problem [11]. To complete this task, this method sets up and fills out a matrix for the alignment of the two sequences of interest. This matrix can then be used in a traceback procedure (following a path through the matrix) to produce the aligned sequences [11, 12]. In research, the Smith-Waterman algorithm has been shown to produce the best accuracy in detecting homologs when compared to other methods [13].

1.4 Working Definitions of Orthology

It is important to note that programs such as BLAST primarily identify putative homologs. Differentiating orthologs from paralogs is not part of the program [3]. A working definition of orthology is used in order to distinguish between these two options. The Reciprocal Best Hits (RBH) definition regards ortholgs simply as the genes which have the greatest BLAST score when run bi-directionally. Thus, two sequences are determined to be orthologous if genome Y run as the query against genome X indicates gene x as the best matched sequence to gene y. In addition, the reverse run of genome X as the query against genome Y must also determine gene y to be the best possible match to gene x [14]. If the first direction proceeds as

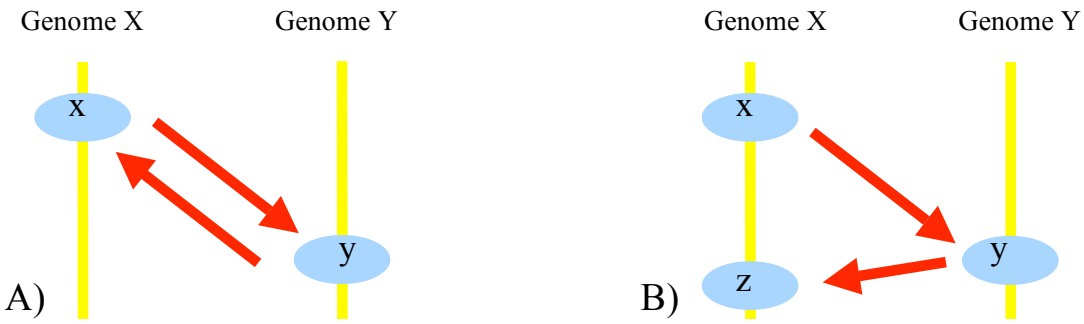

Figure 1.2: Defining orthologs using reciprocal best hits. (A) If two genes find each other as the best possible match when a search is conducted in both directions, the genes are considered to be orthologs. (B) However, if the reciprocal run yields a third gene as the best possible match, it is assumed that a paralog has been identified and the genes are discarded from the list of orthologs.

outlined, but the reverse run determines gene z in genome Y as the best match, it cannot be concluded that x and y are orthologs (Figure 1.2). A potential problem of this method is that paralogs may qualify as best hits in one direction and cause the true ortholog pair to remain undetected [14]. Thus, it is essential that the generated alignment scores properly reflect the relationships between different gene pairs.

A more recently proposed working definition of orthology, termed Reciprocal Smallest Distance (RSD), has been proposed [14]. This definition regards orthologs as genes with the shortest evolutionary distance between them. In other words, homologs are scored according to the number of mutations required to account for all differences in the amino acid sequence between them. In order to score for evolutionary distance, additional programs must be employed. These additional programs result in a significant increase in the complexity of the process, and the time required for ortholog detection [14]. It has been claimed that RSD is able to identify an increased number of orthologs compared to RBH, because this definition is based on phylogenetic relationships [14]. Since being proposed in 2003, the authors of the original RSD paper have created an online database which enables researchers to obtain the results of pre-run sequence searches [7].

1.5 Objectives

This study aims to determine if the number of orthologs detected using the Reciprocal Best Hits working definition of orthology can be improved by selecting different internal options in BLAST. Specifically, the effect of changing the filter setting and alignment method are considered in this study. Another objective of this project is to determine which working definition of orthology is better by comparing results obtained from RBH with the best set of options to those of RSD. These comparisons will be based on the number of orthologs detected, as well as on estimating the rate of errors in orthology detection and computational costs associated with each method. Two types of errors will be considered: mislabelled pairs (an ortholog labelled as a paralog and *vice versa*), as well as the error of missing an existing ortholog.

2 Materials and Methods

2.1 Improving BLAST Scores

BLAST was run with a different set of options to determine if ortholog detection by RBH could be improved. The two chosen options, filter setting and alignment method, were selected based on preliminary data collected by Gabriel Moreno-Hagelsieb. As each option had two possible settings (see section 1.3), a total of four different conditions were tested (for detailed information with respect to these conditions, refer to Table 2.1). BLAST was run with each set of options as a stand-alone program using PERL programs. Each program was initiated on one of several terminal computers (both PCs and Macs were used) which were networked to the main laboratory computer containing sufficient hardware to process such demanding

		Alignment Method	
		BLAST Alignment	**Smith-Waterman**
Filter	**Hard**	Hard Filter BLAST Alignment (FTsF) (BLAST Default Setting)	Hard Filter Smith-Waterman (FTsT)
	Soft	Soft Filter BLAST Alignment (FmSsF)	Soft Filter Smith-Waterman (FmSsT)

Table 2.1: The filter setting and the alignment method within BLAST were manipulated in an attempt to improve the detection of orthologs. In the filter setting, a hard filter (*-F T* option in NCBI's BLAST) removed low-information animo acid sequences during both the search and alignment phases of this program, while a soft filter (*-F "m S"*) works only during the search phase. The default BLAST alignment (*-s F*) is based on extension through matching words found during the search phase, while the Smith-Waterman algorithm (*-s T*) uses dynamic programming to produce a mathematically optimal alignment.

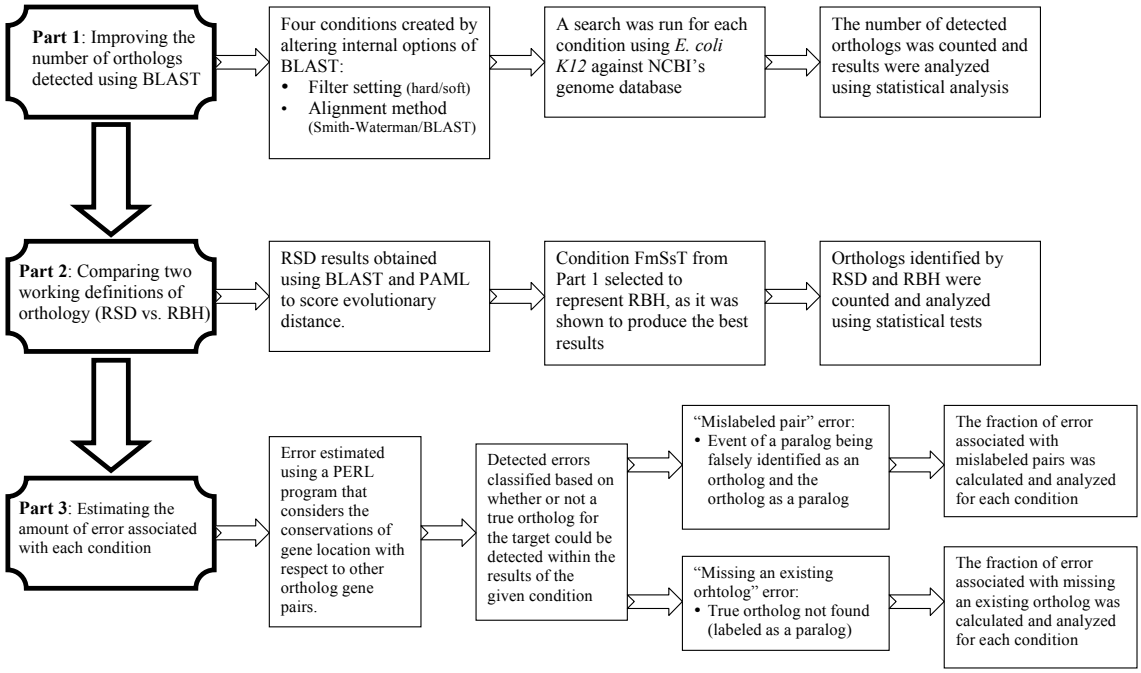

Figure 2.1: A general summary of the experimental procedure. The experimental design can be divided into three main parts (left side of this figure). Each part contained several steps, which are outlined with arrows moving towards the right side of the image.

tasks. All searches were conducted using *Escherichia coli* K12 as the query genome against a database containing the sequenced genomes of 472 organisms. These sequences were obtained from RefSeq, which is an online database hosted by NCBI [15]. Since the database used is continuously updated with corrected sequences and additional genomes, BLAST was also re-run at several points throughout the experimental period. Thus, all data and calculated statistics referred to in this paper are up to date as of the end of March of 2007. An overview of the experimental procedure experiment is illustrated in Figure 2.1.

Initially, the BLAST output is a list of the homologs found for each individual genome in the database with respect to the query sequences. However, as previously explained, these homologs may be further classified as either orthologs or paralogs. In order to identify and count the orthologs in the final results, two additional PERL programs were used. The first program searched the BLAST results and selected all protein pairs with the highest similarity score (using the BLAST bit-score) us-

ing the RBH definition of orthology. Selected pairs showing the greatest amount of similarity between their sequences were used to determine orthologous pairs. This step separated homologs into two lists: one containing gene pairs identified as orthologs and another for genes identified paralogs. A second PERL program kept a running count for the number of ortholog pairs found for each tested genome of the database. For this specific experiment, this program counted any ortholog that met the set E-value (a statistical measure of significance) threshold of $1 * 10^{-6}$. This value was chosen primarily to minimize the number of false-positives in the final results (an example PERL program for this purpose is displayed in Section A.1 in the Appendix). In the end, a list was generated which contained a column of the organisms in the database with the corresponding number of ortholgs detected for each condition. The number of orthologs identified was then analyzed using a *repeated measures ANOVA* with pairwise comparisons. All statistical tests were conducted using the program SPSS, version 12.0. Unless otherwise indicated, pairwise comparisons included the Bonferroni correction, which alters the critical values of the test to account for multiple comparisons [16].

2.2 Genome Similarity Score

Once data were analyzed, the results were graphed to produce a visual representation of the data obtained from each condition. However, due to the large amount of data gathered during the experimental process, some modifications were made to better illustrate the results. For example, graphs illustrating the number of detected orthologs were normalized to one condition to reduce the range of values required along the y-axis. Another modification was that genomes were placed along the x-axis with respect to their evolutionary distance from *E. coli* K12 (the query genome). This was done to provide a meaningful order to the data for which a trend may be identified. Evolutionary distance was determined by calculating the genome similarity score (GSS), a concept which has been previously described [6, 17]. In order to calculate this value, a BLAST search is first conducted with the query (*E.*

coli K12) run again a given genome of the database. The resulting alignment scores, labelled "comparison-scores," for each gene pair are then summed for all orthologous genes in these two genomes [6]. A second BLAST search is then conducted with the query run against itself to determine the maximum alignment score. These self-scores for each gene pair are also summed. The total comparison score is then divided by the total self score to obtain the GSS for the particular genome with respect to the query [17]. This calculation can be represented by the formula:

$$GSS = \frac{\sum_{i=1}^{n} \text{comparison-score}_i}{\sum_{i=1}^{n} \text{self-score}_i}$$

As the GSS value is a fraction of similarity, a value of 1.0 would indicate a perfect match between the query and the chosen genome [6]. This value accurately represents the evolutionary distance between the genomes, as accumulated mutations throughout divergent evolution will reduce the sequence similarity and generate a lower GSS.

2.3 Comparing Working Definitions of Orthology

Data were obtained for the reciprocal smallest distance working definition of orthology by following the procedure described by Wall *et al.* [14]. Originally, the authors of this method used BLAST, run with default options, to generate a list of homologs. These researchers then used a different alignment program, CLUSTALW [18], in order to realign and score the similarity between protein pairs [14]. However, CLUSTALW is a program designed to build multiple sequence alignments, rather than to compute the best possible alignment between two sequences. Here, the alignments obtained from the FmSsT condition (soft filter and Smith-Waterman algorithm) were used. Thus, this experiment differed from that of Wall *et al.* [14] by employing the Smith-Waterman algorithm. This difference may only increase the accuracy of the results, as the Smith-Waterman algorithm has been demonstrated

to be the best possible alignment method [13].

The RSD defines orthologs as the homologs with the calculated reciprocal smallest distances. These distances can be scored by determining the mutation events required to change the sequence of one protein into that of another protein [14]. In order to determine the evolutionary distance between each gene pair, a program called PAML was used [19]. This program uses a statistical method known as maximum likelihood, along with an amino acid substitution rate matrix, to score homologs based on their evolutionary distances [19]. A program written in PERL was then used to identify and count the number of orthologs detected by this method. These results were computed and compared to each set of results obtained from the different BLAST options using a *repeated measures ANOVA* test with pairwise comparisons. Emphasis was placed on the comparison between FmSsT and RSD, as the only difference between these results was the addition of the PAML program to define orthologs using evolutionary distance in the RSD definition.

2.4 Mistakes in Orthology Detection

Analyses were conducted on all five sets of results obtained to estimate the rate of errors in assignment of orthologs. Two kinds of errors were considered: mislabelled pairs and missing existing orthologs. These analyses were based on conservation of gene order. Since the conservation of gene order between Prokaryotes and Eukaryotes is close to non-existent, eukaryotic organisms were excluded from these tests. Thus, the analyses were conducted on 456 genomes instead of the original 472. In order to determine the rate of error, a program previously written by Dr. Gabriel Moreno-Hagelsieb was edited to work with each dataset of results. This program works by comparing the conservation of gene location with respect to adjacent genes within the query genome. If corresponding homologous genes are found to be conserved next to each other, it is expected that both genes will be orthologs to the corresponding query genes. If the conserved pair is composed of an ortholog and a paralog, the paralog is assumed to be evidence of an error in the ortholog identifi-

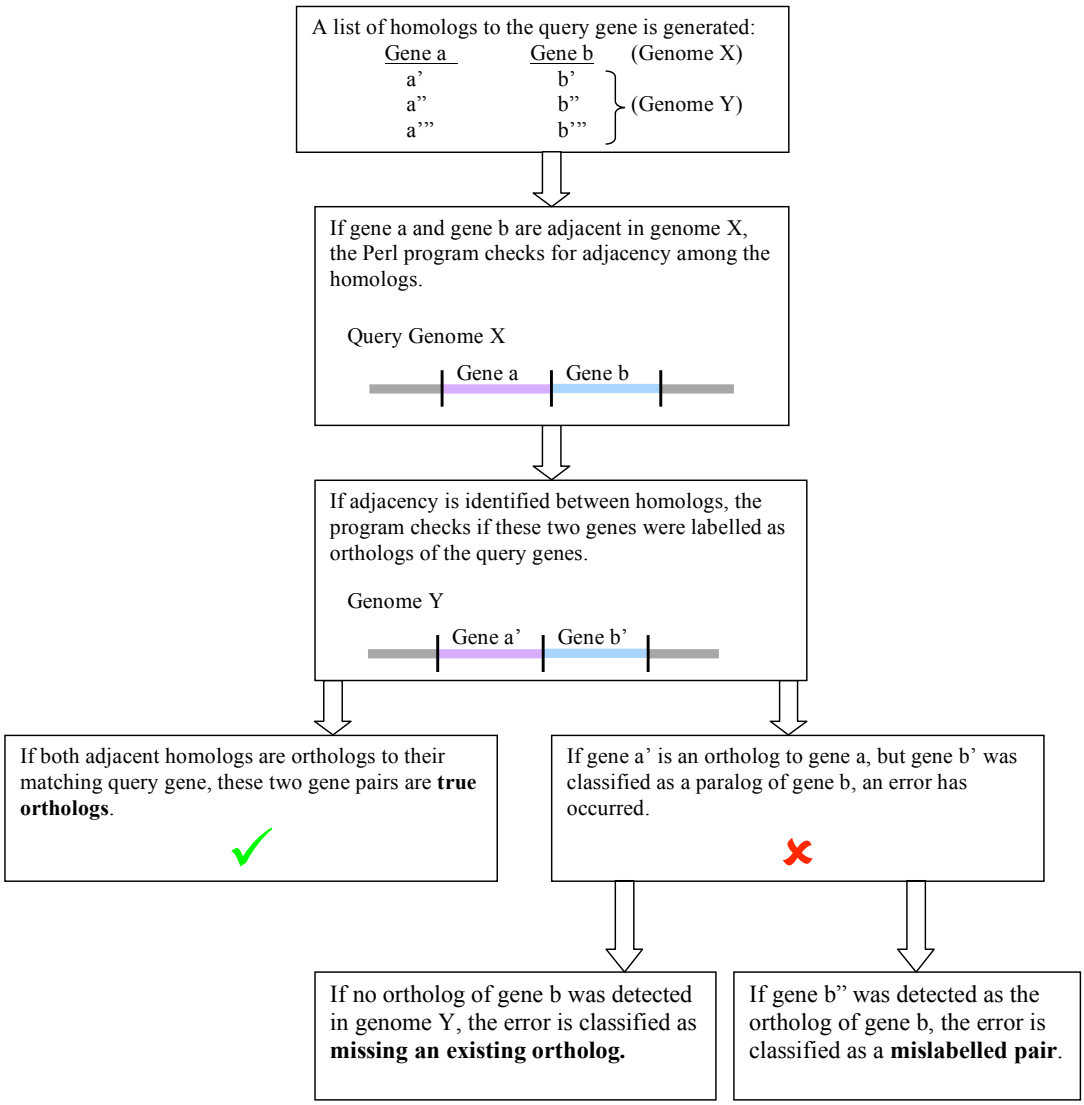

Figure 2.2: Method for estimating error rates. Mislabelled pairs refer to events when a paralog is falsely determined to be an ortholg (and *vice versa*). Missing existing orthologs refer to events where the true ortholg remains undetected during the search.

cation. Thus, when homologs are found to be adjacent, the program checks if these genes were both identified as orthologs to the two query genes. In cases where this is true, the gene pairs are concluded to be true orthologs. Essentially, true orthologs are defined as genes that were correctly identified as genes arising from a common ancestor and have conserved their gene function. However, if one gene was detected as an ortholog while its adjacent gene was labelled as a paralog, it is concluded that an error has occurred. For a schematic outlining the process of estimating these

errors, refer to Figure 2.2.

Errors were further classified based on whether or not an ortholog for this second gene was detected in BLAST. If an ortholog exists, the error is classified as a mislabelled pair. This refers to an event where a paralog was falsely labelled as an ortholg to the query, and/or an ortholog was wrongly labelled as a paralog. However, if no corresponding ortholog is present within the results for the given condition, it is assumed the true ortholog was simply not detected during the search. These error events are classified as missing an existing ortholog.

Once the error types associated with each gene pair for the entire database of genomes were determined, another program was written to tally the number of errors associated with each condition. This program was similar to the one used to count the number of detected orthologs (see Section A.1). However, it also included programming code to convert the number of errors into a fraction with respect to the number of true orthologs detected. Results were then analyzed using a *repeated measures ANOVA* with pairwise comparisons to individually compare each condition against all the others.

3 Results and Discussion

3.1 Improving BLAST

In the first part of this work, two internal BLAST options were chosen to determine if the number of detected orthologs could be improved. Specifically, the filter setting in BLAST and the method used for aligning and scoring homologous sequences were manipulated. Four different conditions were tested and the number of orthologs detected by each condition was determined. Data analysis using a *repeated measures ANOVA* indicated a significant difference in the number of detected orthologs using these four conditions ($F = 1151.408$, $p < 0.001$). In order to determine which conditions differed from each other, a pairwise comparison was included in the statistical analysis. Using this analysis, it was determined that the BLAST default setting, FTsF, identified the lowest number of ortholog gene pairs. Compared to FTsF, the number of detected orthologs was significantly increased by using the Smith-Waterman alignment algorithm in the FTsT condition (mean difference = 3.072, $p < 0.001$). The condition labelled FmSsF differed from the BLAST default settings in that the filter was set to soft and thus removed information only during the search phase. It was found that this choice produced an even greater increase in the number of results obtained (mean difference = 24.428, $p < 0.001$). However, combining the soft filter with the Smith-Waterman algorithm, as was done in condition FmSsT, generated the best results with respect to the number or orthologs detected (mean difference compared to FTsF = 26.852, $p < 0.001$). It is also important to note that each subsequent change produced a condition detecting a significantly higher number of orthologs than the previous condition (FTsF < FTsT < FmSsF < FmSsT, $p < 0.001$). Detailed statistical results can be seen in Section A.2; Table A.1, while the number of orthologs detected by each condition is

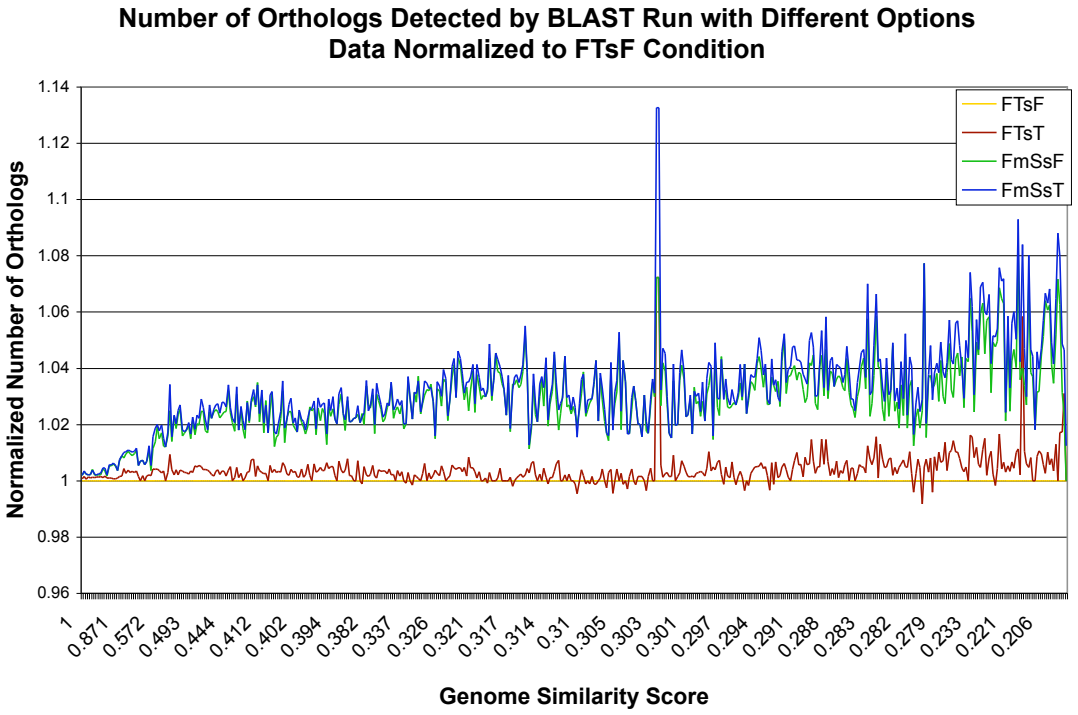

Figure 3.1: Orthologs detected by selecting different internal BLAST options. The data were normalized against the BLAST default setting (FTsF) to enhance visual representation. There is a significant difference between conditions ($F = 1151.408$, $p < 0.001$) such that every conditions is significantly different from the other three conditions (pairwise comparison, $p < 0.001$).

illustrated in Figure 3.1.

In order to verify whether the increased number of orthologs was the result of an increase in the number of homologs detected under each condition, statistical analyses were also conducted on the number of homologs obtained for each condition. This was done by comparing the initial BLAST output values for each combination of options. The *repeated measures ANOVA* (Table A.2) indicated an overall trend similar to that observed with respect to the number of orthologs, with a significant difference being found between all conditions ($F = 2098.631$, $p < 0.001$). Compared to the BLAST default setting (FTsF), employing the Smith-Waterman algorithm increased the detection of homologs (mean difference $= 7.411$, $p < 0.001$). However, an even greater increase was seen when the filter setting was changed from hard to soft, as seen in the FmSsF condition (mean difference $= 54.822$, $p < 0.001$). As with the number of detected orthologs, combining the Smith-Waterman algorithm with

the soft filter setting provided the highest numbers of homologs (mean difference = 60.398, $p < 0.001$). These results suggest that the number of orthologs detected is related to the number of homologs detected within each condition.

To further illustrate the relationship between orthologs and homologs, analyses were also conducted on normalized data. The numbers of orthologs were normalized by dividing them against the corresponding number of homologs for each genome. The *repeated measures ANOVA* (Table A.3) indicated a significant difference among conditions ($F = 28234.678$, $p < 0.001$), with a pairwise comparison revealing all conditions to be significantly different from each other ($p < 0.001$ for all comparisons). These results mean that the differences in orthologs detected, despite related to the number of homologs, are also the result of differences in scores among the detected homologs.

Overall, this study was successful at demonstrating that the number of orthologs detected with BLAST can be improved by employing either a soft filter or the Smith-Waterman algorithm. It is noteworthy that using a soft filter (masking of low-information stretches of amino acid sequences during the search phase, but not during the alignment phase) caused the greater increase in the number of orthologs than selecting a different alignment method. However, combining both the soft filter and the Smith-Waterman algorithm in the same search provided the best results compared to all other option combinations tested in this work.

Although it has been demonstrated that choosing internal options within BLAST will change the obtained results [20], many researchers are unaware of this possibility. This may be one reason why research related to the effect of different options within this program on detection of orthologs has not been previously conducted. When using exhaustive computer-based searching methods, it is essential to find a balance between the sensitivity of the program and the associated costs with respect to time and processing speed [8]. It is for this reason that the BLAST default setting (FTsF) was (not surprisingly) found to detect the smallest number of orthologs. By filtering information at both steps of this program, processing time is reduced as there are fewer genes to compare and align with query sequences. Also, although the Smith-

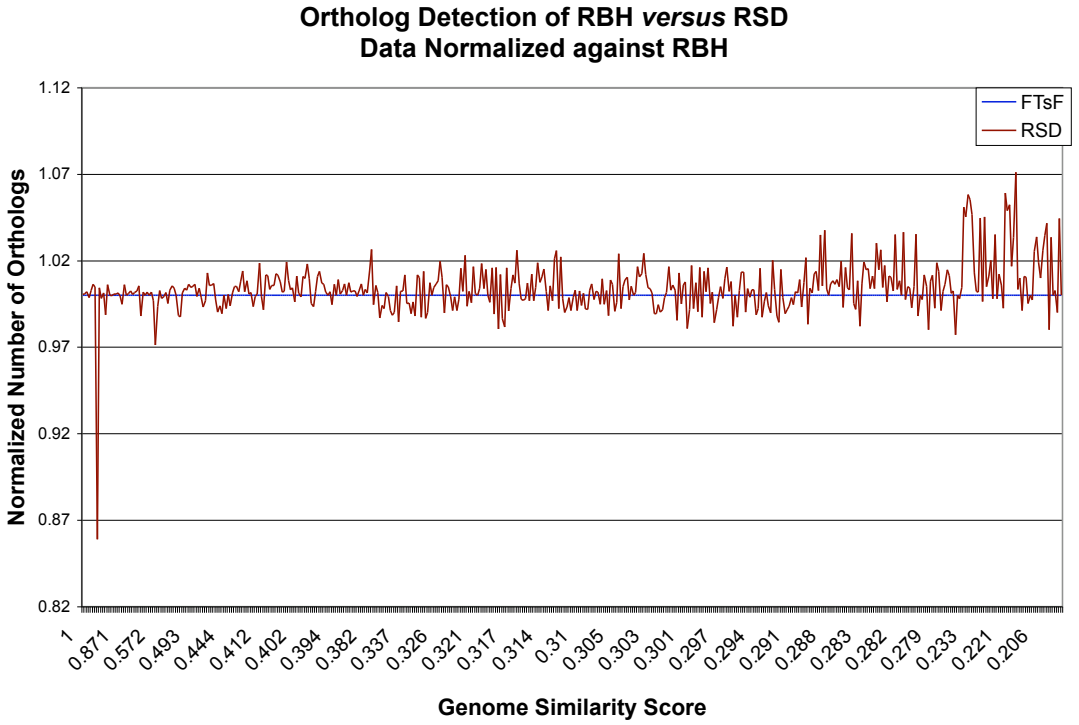

Figure 3.2: Comparing the number of orthologs detected by RBH and RSD. The data have been normalized to the RBH to enhance visual representation. There was no significant difference between the two definitions when employing a *repeated measures ANOVA* (mean difference = 2.708, $p = 0.299$). However, removing the Bonferroni correction during this test produced a marginally significant difference (mean difference = 2.708, $p = 0.030$).

Waterman algorithm has been shown to be the best alignment method [13], the BLAST alignment is structured specifically to reduce the time required to run the program. Although the two conditions utilizing the soft filter were observed to run at a slower rate, these conditions provided the best results and still completed the search within two or three days. A larger difference may be seen between conditions if higher-demanding processes are conducted, though future research is required to make this conclusion.

3.2 Reciprocal Smallest Distances *versus* Reciprocal Best Hits

In the second part of this work, the improved RBH results were compared to the results obtained using the Reciprocal Smallest Distances (RSD) working definition of orthology (Figure 3.2). Orthologs detected using the condition FmSsT were selected to represent the RBH definition, as these were shown to produce the best results. A *repeated measures pairwise comparison* with all five conditions revealed that there was no difference in the number of orthologs identified using each of these two working definitions (mean difference of RSD to FmSsT = 2.708, $p = 0.299$, see also Section A.2; Table A.4).

As previously indicated, all *repeated measures* analyses reported were conducted using the Bonferroni correction. For this correction, the α value (representing type 1 error) is adjusted such that an α level of 0.05 is true for each individual case of the *repeated measures* test and not just the condition overall [16, 21]. In other words, this correction factor alters the critical values of the statistical test to account for the multiple comparisons, without altering the test itself [16].

Results without employing the correction factor are equivalent to what is obtained using a paired t-test. When this test was run without this correction, a slightly significant difference was found between these two working definitions of orthology (mean difference = 2.708, $p = 0.030$). Thus, these results may exist in a marginal area, as a slight change in the type of test conducted can alter the conclusions drawn from this experiment (see detailed statistics in Table A.5).

Although the statistical analysis lacks the required confidence level to be conclusive with respect to the number of orthologs detected by the best RBH definition when compared against the RSD definition, some conclusions can still be made based these data. As previously noted, the RSD definition of orthology takes into account the evolutionary distance between two homologous sequences. Wall *et al.* [14] reported that RSD identifies a higher number of ortholog pairs when compared to RBH, and thus they suggested it to be a better working definition of orthology [14].

Preference for RSD also arises due to the assumption that using phylogenetic relationships (evolutionary distances) is more accurate compared to alternative methods. Thus, researchers often have more confidence in results using RSD as they are based more on the assumption of phylogenic relationships than on strict tests. However, obtaining these distances requires employing additional programs such as PAML to score the mutations which have occurred between every pair of homologous sequences. Thus, detecting ortholgs using RSD is associated with additional processing time, and yet remains inconclusive in whether it truly produces a greater number of orthologs. Based on these data, the Reciprocal Best Hits definition of orthology can be suggested as a better choice since it requires reduced processing after homolog identification. As shown in this study, the difference in the amount of orthologs detected by RSD are marginal compared to this further computing time and effort compared to RBH.

3.3 Mistakes in Orthology Detection

To determine which method produced the most accurate results, an estimate of the error rate in orthology detection was conducted for all five conditions tested (four sets of options in BLAST plus RSD). Genes determined to be orthologs with respect to a query sequence are often assumed to be true orthologs. Essentially, this term refers to the idea that the claimed ortholog-ortholog pair is in fact composed of genes which arose from a common ancestor and have a conserved gene function. However, as with any experimental procedure, computer searches are at risk of making mistakes. Two potential sources of error were considered: that of mislabelled pairs and that of missing an existing ortholog.

3.3.1 Mislabelled Pairs

The term mislabelled pair refers to a paralog that has been identified by the computer process as an ortholog. This is a risk during sequence searches, as detection is primarily based on sequence similarity. Mislabelled pairs were identified using a

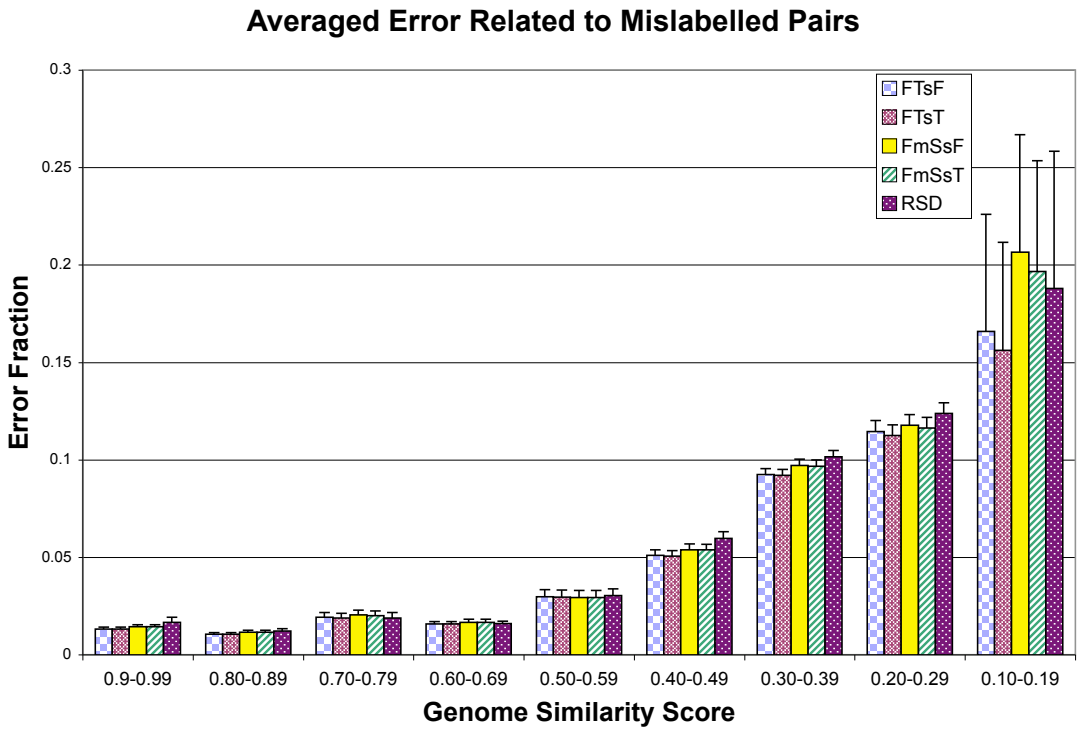

Figure 3.3: The fraction of error associated with mislabelled pairs. Values are averaged over designated ranges of the Genome Similarity Score. The calculated fraction of error greatly increased in genomes calculated to have a further evolutionary distance. All conditions were determined to be significantly different from each another (repeated measures pairwise comparison, $p < 0.05$).

program written by Dr. Gabriel Moreno-Hagelsieb, which compares the conservation of location of two query genes to the location of their homologs and detected orthologs. Results were analyzed using a *repeated measures ANOVA* test, and it was determined that at least one condition was significantly different from the others ($F = 1088.103$, $p < 0.001$). When a pairwise comparison was made, it was found that all conditions differed significantly from the rest in terms of this source of error. However, no trend was seen with respect to mean differences across conditions (Table A.6). When the data were graphed (Figure 3.3), a general pattern was observed such that the amount of error increased with the decrease in Genome Similarity Score (GSS). Thus, genomes which have a larger evolutionary distance from *E.coli* K12 displayed an increased number of paralogs mislabelled as orthologs (and *vice versa*). These results indicated no dramatic difference in the number of mislabelled

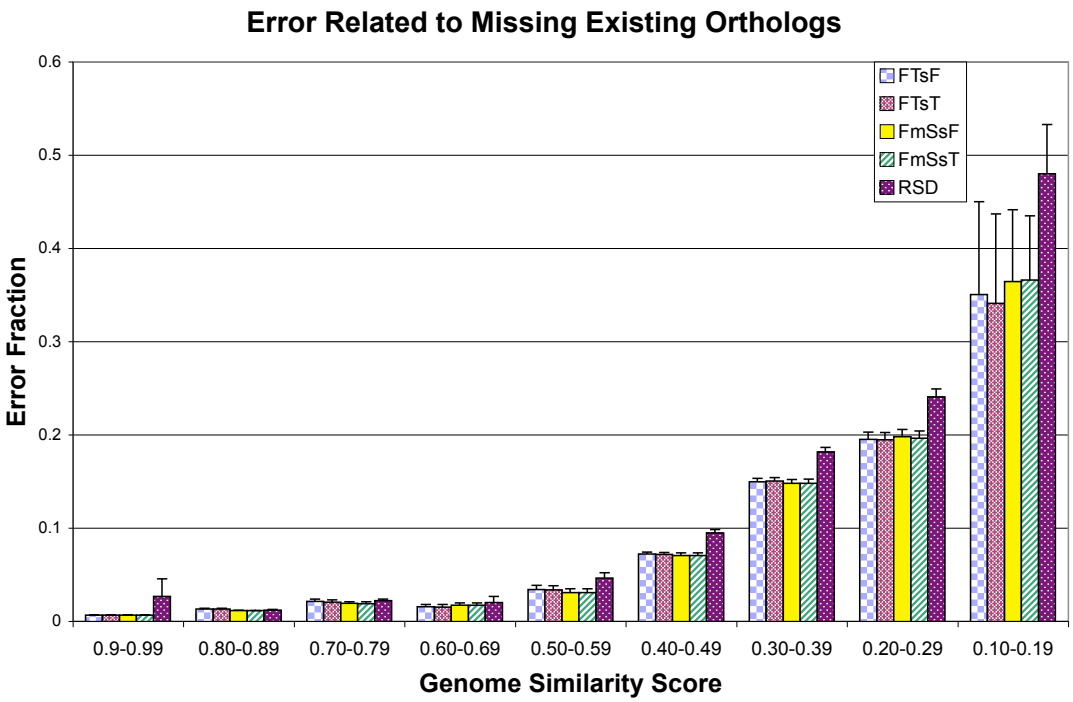

Figure 3.4: The rate of errors associated with missing existing orthologs. Values are averaged over designated ranges of the Genome Similarity Score. The calculated fraction of error greatly increased in genomes at further evolutionary distances. The RSD resulted in a significant highest rate of missing orthologs compared to all RBH conditions (*repeated measures ANOVA*, $p < 0.001$).

pairs in the results obtained using RBH or those of RSD. An average taken across all conditions revealed a fraction of approximately 0.06 for this specific type of error. In genomes determined to have the greatest evolutionary distance from the query (lowest GSS score), the averaged error fraction related to mislabelled pairs reached a high of approximately 0.21. These values indicate that mislabelled pairs were found to such a degree that these errors may interfere with results of any given research study. However, as this error did not display a pattern across conditions, one condition cannot be said to help reduce the occurrence of mislabelled pairs compared to the others.

3.3.2 Missing Existing Orthologs

Error analysis for this experiment also considered the possibility that a true ortholog to a query sequence was not detected under a given experimental condition. This

error was said to have occurred if an ortholog pair was labelled as containing an error, and the program could not identify an ortholog within the BLAST results. A *repeated measures ANOVA* test showed that at least one of the conditions was statistically different from the rest ($F = 1180.619$, $p < 0.001$). Using a pairwise comparison, it was determined that all four RBH conditions generated the same rate of errors (for all comparisons: mean difference = 0.000, $p = 1.00$). However, the results pertaining to the RSD definition was found to be significantly higher compared to all four RBH conditions (for all conditions against RSD: mean difference = 0.034, $p < 0.001$). For full statistical results see Table A.7.

As with the mislabelled pair error, a general trend was seen in the data such that the rate of error observed increased in genomes at greater evolutionary distances from the query genome (Figure 3.4). However, it was also found that the rate of missing existing orthologs occurred more frequently using the RSD definition of orthology. The average error fraction across all conditions was calculated to be approximately 0.10, while the highest average seen on the graph reached a fraction of 0.48 (Figure 3.4). These values indicate that the rate of missing existing orthologs has the highest potential to seriously alter the results of bioinformatics research. This was especially important with the RSD definition. This result further indicates a higher cost associated with this working definition of orthology compared to Reciprocal Best Hits. Thus, the results from this error analysis test further confirm that RBH remains to be a better working definition of orthology for bioinformatics research.

4 Conclusions

This study was conducted in order to determine if choosing different internal options within the program BLAST would affect the number of orthologs detected during a search. Results of this study indicate that ortholog detection was improved by selecting a soft filter (the *-F "m S"* option in NCBI's BLASTP) in combination with alignments using the Smith-Waterman algorithm (*-s T*). The greatest number of orthologs was detected by using these two options within the same condition.

In the second part of this study, results of condition FmSsT, representing the reciprocal best hit (RBH) working definition of orthology were compared to results using a newer definition termed Reciprocal Smallest Distances (RSD). This was done to determine if RSD detects a greater number of orthologs, as it had previously been reported [14]. In the initial analysis, it was found that there was no significant difference between these two conditions with respect to the number of orthologs detected. However, this cannot be concluded with confidence, as removing the Bonferroni correction from the pairwise comparison provided contradictory conclusions. Regardless, an error analysis relating to the number of orthologs remaining undetected by each condition revealed that RSD has a significantly higher rate of this error. Error which considered mislabelled pairs was significantly different between all conditions, but did not indicate any pattern between conditions. Thus, the RSD working definition of orthology is a more complex method which requires increased time and processing associated with running additional programs. This definition was also shown to have an increased rate of missed orthologs, while it remains unclear if the overall accuracy is better than that of RBH. The conclusions drawn from this study suggest RBH remains the best working definition of orthology, as the number of results obtained by RSD does not justify the additional associated costs.

References

[1] Bansal AK: **Bioinformatics in microbial biotechnology–a mini review**. *Microbial cell factories* 2005, **4**:19.

[2] Zheng XH, Lu F, Wang ZY, Zhong F, Hoover J, Mural R: **Using shared genomic synteny and shared protein functions to enhance the identification of orthologous gene pairs**. *Bioinformatics (Oxford, England)* 2005, **21**(6):703–10.

[3] Yuan YP, Eulenstein O, Vingron M, Bork P: **Towards detection of orthologues in sequence databases**. *Bioinformatics (Oxford, England)* 1998, **14**(3):285–9.

[4] Fitch WM: **Homology a personal view on some of the problems**. *Trends Genet* 2000, **16**(5):227–231.

[5] Jensen RA: **Orthologs and paralogs - we need to get it right**. *Genome Biol* 2001, **2**(8):INTERACTIONS1002.

[6] Janga SC, Moreno-Hagelsieb G: **Conservation of adjacency as evidence of paralogous operons**. *Nucleic Acids Res* 2004, **32**(18):5392–7.

[7] Deluca TF, Wu IH, Pu J, Monaghan T, Peshkin L, Singh S, Wall DP: **Roundup: a multi-genome repository of orthologs and evolutionary distances**. *Bioinformatics (Oxford, England)* 2006, **22**(16):2044–6.

[8] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389–402.

[9] Pertsemlidis A, Fondon r J W: **Having a BLAST with bioinformatics (and avoiding BLASTphemy)**. *Genome biology* 2001, **2**(10):REVIEWS2002.

[10] Flannick J, Batzoglou S: **Using multiple alignments to improve seeded local alignment algorithms**. *Nucleic acids research* 2005, **33**(14):4563–77.

[11] Smith TF, Waterman MS: **Identification of common molecular subsequences**. *Journal of molecular biology* 1981, **147**:195–7.

[12] Eddy SR: **What is dynamic programming?** *Nature biotechnology* 2004, **22**(7):909–10.

[13] Brenner SE, Chothia C, Hubbard TJ: **Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships**. *Proceedings of the National Academy of Sciences of the United States of America* 1998, **95**(11):6073–8.

[14] Wall DP, Fraser HB, Hirsh AE: **Detecting putative orthologs**. *Bioinformatics (Oxford, England)* 2003, **19**(13):1710–1.

[15] Pruitt KD, Tatusova T, Maglott DR: **NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**. *Nucleic Acids Res* 2005, **33 Database Issue**:D501–4.

[16] Howell DC: *Statistical methods for psychology*. Belmont, Calif.: Thomson Wadsworth, 6th edition 2007.

[17] Moreno-Hagelsieb G, Janga SC: **Operons and the effect of genome redundancy in deciphering functional relationships using phylogenetic profiles**. *PROTEINS: Structure, Function and Bioinformatics* 2007, **In press**.

[18] Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice**. *Nucleic acids research* 1994, **22**(22):4673–80.

[19] Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood**. *Computer applications in the biosciences* 1997, **13**(5):555–6.

[20] Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF: **Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements**. *Nucleic Acids Res* 2001, **29**(14):2994–3005.

[21] Weisstein EW: **Bonferroni Correction** 2007. [A Wolfram Web Resource. http://mathworld.wolfram.com/BonferroniCorrection].

Appendix

A.1 Example program

```
#!/usr/bin/perl -w

### Set up to count all four BLAST conditions in one run
my @experiments = qw(
                        FTsF
                        FTsT
                        FmSsF
                        FmSsT
                );
for my $experiment ( @experiments ) {
    ### Navigate to folder and formulate list
    print "Yep, I am working with $experiment\n";
    $work_dir = "../results/$experiment/BDBH/E_coli_K12";
    open(RES,">$experiment.tbl");
    opendir(MYDIR, "$work_dir") or die "Cannot Open $work_dir\n";
    @files = grep { /bdbh/ } readdir(MYDIR);
    my $GTotal = 0;
    foreach $file ( @files ) {
        print ".";
        ### open each zipped file and count
        #print "$work_dir/$file\n";
        unless ( -s "$work_dir/$file" ) {
            die "Error ($work_dir/$file)\n";
        }
        my ($main_name) = $file =~ /E_coli_K12\.(\S+)\.bdbh/;
        open(GENOME, "bzip2 -dc $work_dir/$file |");
        my $count = 0;
        my %no_dupl = ();
        ### Count orthologs
        while (<GENOME>) {
            my @list = split;
            if ($list[5] <= 1e-6) {
                $no_dupl{$list[0]}++;
                $count++;
                $GTotal++;
            }
        }
        close (GENOME);
        my @no_dupl = keys %no_dupl;
```

```
        my $no_dupl = @no_dupl;
        print RES "$no_dupl\t$count\t$main_name\n";
    }
    print "Yep, I am working with $experiment (actually finished)\n";
    print RES "TOTAL\t$GTotal\n";
}
```

A.2 Statistical analyses—Ortholog numbers

A.2.1 Repeated measures ANOVA—Orthologs

Within-Subjects Factors
Measure: MEASURE_1

factor1	Dependent Variable
1	FTsF
2	FTsT
3	FmSsF
4	FmSsT

Estimates
Measure: MEASURE_1

factor1	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	1013.172	30.192	953.845	1072.499
2	1016.244	30.229	956.843	1075.644
3	1037.600	30.312	978.035	1097.164
4	1040.023	30.307	980.470	1099.577

Tests of Between-Subjects Effects
Measure: MEASURE_1
Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	1990395945.172	1	1990395945.172	1151.408	.000	.710
Error	814200026.328	471	1728662.476			

Pairwise Comparisons
Measure: MEASURE_1

(I) factor1	(J) factor1	Mean Difference (I-J)	Std. Error	Sig.(a)	95% Confidence Interval for Difference(a)	
					Lower Bound	Upper Bound
1	2	-3.072(*)	.123	.000	-3.397	-2.747
	3	-24.428(*)	.509	.000	-25.777	-23.079
	4	-26.852(*)	.537	.000	-28.275	-25.428
2	1	3.072(*)	.123	.000	2.747	3.397
	3	-21.356(*)	.488	.000	-22.650	-20.062
	4	-23.780(*)	.503	.000	-25.112	-22.448
3	1	24.428(*)	.509	.000	23.079	25.777
	2	21.356(*)	.488	.000	20.062	22.650
	4	-2.424(*)	.103	.000	-2.697	-2.150
4	1	26.852(*)	.537	.000	25.428	28.275
	2	23.780(*)	.503	.000	22.448	25.112
	3	2.424(*)	.103	.000	2.150	2.697

Based on estimated marginal means

* The mean difference is significant at the .05 level.

a Adjustment for multiple comparisons: Bonferroni.

Table A.1: All of the BLAST conditions showed a significant difference with respect to the number of orthologs detected by Reciprocal Best Hits.

A.2.2 Repeated measures ANOVA—Homologs

Within-Subjects Factors
Measure: MEASURE_1

factor1	Dependent Variable
1	FTsF
2	FTsT
3	FmSsF
4	FmSsT.RSD

Estimates
Measure: MEASURE_1

factor1	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	1485.828	33.027	1420.929	1550.728
2	1493.239	33.042	1428.311	1558.167
3	1540.650	33.187	1475.438	1605.863
4	1546.227	33.166	1481.055	1611.398

Tests of Between-Subjects Effects
Measure: MEASURE_1
Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	4341891150.358	1	4341891150.358	2098.631	.000
Error	974459473.142	471	2068916.079		

Pairwise Comparisons
Measure: MEASURE_1

(I) factor1	(J) factor1	Mean Difference (I-J)	Std. Error	Sig.(a)	95% Confidence Interval for Difference(a)	
					Lower Bound	Upper Bound
1	2	-7.411(*)	.160	.000	-7.836	-6.986
	3	-54.822(*)	.785	.000	-56.902	-52.742
	4	-60.398(*)	.840	.000	-62.625	-58.172
2	1	7.411(*)	.160	.000	6.986	7.836
	3	-47.411(*)	.735	.000	-49.358	-45.464
	4	-52.987(*)	.771	.000	-55.031	-50.944
3	1	54.822(*)	.785	.000	52.742	56.902
	2	47.411(*)	.735	.000	45.464	49.358
	4	-5.576(*)	.143	.000	-5.955	-5.197
4	1	60.398(*)	.840	.000	58.172	62.625
	2	52.987(*)	.771	.000	50.944	55.031
	3	5.576(*)	.143	.000	5.197	5.955

Based on estimated marginal means

* The mean difference is significant at the .05 level.

a Adjustment for multiple comparisons: Bonferroni.

Table A.2: All of the BLAST conditions showed a significant difference in the number of homologs detected.

A.2.3 Repeated measures ANOVA—Normalized orthologs

Within-Subjects Factors
Measure: MEASURE_1

factor1	Dependent Variable
1	FTsF
2	FTsT
3	FmSsF
4	FmSsT
5	RSD

Estimates
Measure: MEASURE_1

factor1	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	.647	.004	.640	.655
2	.646	.004	.638	.653
3	.639	.004	.631	.647
4	.638	.004	.631	.646
5	.641	.004	.633	.648

Tests of Between-Subjects Effects
Measure: MEASURE_1
Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	973.116	1	973.116	28234.678	.000	.984
Error	16.233	471	.034			

Pairwise Comparisons
Measure: MEASURE_1

(I) factor1	(J) factor1	Mean Difference (I-J)	Std. Error	Sig.(a)	95% Confidence Interval for Difference(a)	
					Lower Bound	Upper Bound
1	2	.001(*)	.000	.000	.001	.002
	3	.008(*)	.000	.000	.007	.009
	4	.009(*)	.000	.000	.008	.010
	5	.007(*)	.001	.000	.005	.008
2	1	-.001(*)	.000	.000	-.002	-.001
	3	.007(*)	.000	.000	.006	.007
	4	.008(*)	.000	.000	.007	.008
	5	.005(*)	.001	.000	.004	.007
3	1	-.008(*)	.000	.000	-.009	-.007
	2	-.007(*)	.000	.000	-.007	-.006
	4	.001(*)	.000	.000	.001	.001
	5	-.002(*)	.000	.010	-.003	.000
4	1	-.009(*)	.000	.000	-.010	-.008
	2	-.008(*)	.000	.000	-.008	-.007
	3	-.001(*)	.000	.000	-.001	-.001
	5	-.002(*)	.000	.000	-.004	-.001
5	1	-.007(*)	.001	.000	-.008	-.005
	2	-.005(*)	.001	.000	-.007	-.004
	3	.002(*)	.000	.010	.000	.003
	4	.002(*)	.000	.000	.001	.004

Based on estimated marginal means

* The mean difference is significant at the .05 level.

a Adjustment for multiple comparisons: Bonferroni.

Table A.3: All conditions produced a significantly different number of detected orthologs, even after normalizing orthologs to the corresponding number of homologs. This reveals that besides the number of homologs detected, the scores produced within each condition affect the actual detection of orthologs by Reciprocal Best Hits.

A.2.4 Repeated measures ANOVA—Adding RSD

Within-Subjects Factors
Measure: MEASURE_1

factor1	Dependent Variable
1	FTsF
2	FTsT
3	FmSsF
4	FmSsT
5	RSD

Estimates
Measure: MEASURE_1

factor1	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	1013.172	30.192	953.845	1072.499
2	1016.244	30.229	956.843	1075.644
3	1037.600	30.312	978.035	1097.164
4	1040.023	30.307	980.470	1099.577
5	1042.731	30.120	983.545	1101.917

Tests of Between-Subjects Effects
Measure: MEASURE_1
Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	2503499465.034	1	2503499465.034	1161.056	.000	.711
Error	1015582888.766	471	2156226.940			

Pairwise Comparisons
Measure: MEASURE_1

(I) factor1	(J) factor1	Mean Difference (I-J)	Std. Error	Sig.(a)	95% Confidence Interval for Difference(a)	
					Lower Bound	Upper Bound
1	2	-3.072(*)	.123	.000	-3.418	-2.726
	3	-24.428(*)	.509	.000	-25.864	-22.992
	4	-26.852(*)	.537	.000	-28.367	-25.336
	5	-29.559(*)	1.410	.000	-33.536	-25.583
2	1	3.072(*)	.123	.000	2.726	3.418
	3	-21.356(*)	.488	.000	-22.733	-19.979
	4	-23.780(*)	.503	.000	-25.198	-22.362
	5	-26.487(*)	1.396	.000	-30.424	-22.550
3	1	24.428(*)	.509	.000	22.992	25.864
	2	21.356(*)	.488	.000	19.979	22.733
	4	-2.424(*)	.103	.000	-2.715	-2.133
	5	-5.131(*)	1.255	.001	-8.672	-1.591
4	1	26.852(*)	.537	.000	25.336	28.367
	2	23.780(*)	.503	.000	22.362	25.198
	3	2.424(*)	.103	.000	2.133	2.715
	5	-2.708	1.243	.299	-6.214	.799
5	1	29.559(*)	1.410	.000	25.583	33.536
	2	26.487(*)	1.396	.000	22.550	30.424
	3	5.131(*)	1.255	.001	1.591	8.672
	4	2.708	1.243	.299	-.799	6.214

Based on estimated marginal means

* The mean difference is significant at the .05 level.

a Adjustment for multiple comparisons: Bonferroni.

Table A.4: All the BLAST options resulted in a significant difference in the number of orthologs detected by Reciprocal Best Hits. However, the Reciprocal Shortest Distances (RSD) did not significantly increase the number of orthologs compared to the FmSsT option set.

A.2.5 Repeated measures ANOVA—No Bonferroti correction

Within-Subjects Factors
Measure: MEASURE_1

factor1	Dependent Variable
1	FTsF
2	FTsT
3	FmSsF
4	FmSsT
5	RSD

Estimates
Measure: MEASURE_1

factor1	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	1013.172	30.192	953.845	1072.499
2	1016.244	30.229	956.843	1075.644
3	1037.600	30.312	978.035	1097.164
4	1040.023	30.307	980.470	1099.577
5	1042.731	30.120	983.545	1101.917

Tests of Between-Subjects Effects
Measure: MEASURE_1
Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	2503499465.034	1	2503499465.034	1161.056	.000	.711
Error	1015582888.766	471	2156226.940			

Pairwise Comparisons
Measure: MEASURE_1

(I) factor1	(J) factor1	Mean Difference (I-J)	Std. Error	Sig.(a)	95% Confidence Interval for Difference(a)	
					Lower Bound	Upper Bound
1	2	-3.072(*)	.123	.000	-3.313	-2.831
	3	-24.428(*)	.509	.000	-25.428	-23.427
	4	-26.852(*)	.537	.000	-27.907	-25.796
	5	-29.559(*)	1.410	.000	-32.330	-26.789
2	1	3.072(*)	.123	.000	2.831	3.313
	3	-21.356(*)	.488	.000	-22.316	-20.396
	4	-23.780(*)	.503	.000	-24.768	-22.792
	5	-26.487(*)	1.396	.000	-29.230	-23.744
3	1	24.428(*)	.509	.000	23.427	25.428
	2	21.356(*)	.488	.000	20.396	22.316
	4	-2.424(*)	.103	.000	-2.626	-2.221
	5	-5.131(*)	1.255	.000	-7.598	-2.664
4	1	26.852(*)	.537	.000	25.796	27.907
	2	23.780(*)	.503	.000	22.792	24.768
	3	2.424(*)	.103	.000	2.221	2.626
	5	-2.708(*)	1.243	.030	-5.151	-.264
5	1	29.559(*)	1.410	.000	26.789	32.330
	2	26.487(*)	1.396	.000	23.744	29.230
	3	5.131(*)	1.255	.000	2.664	7.598
	4	2.708(*)	1.243	.030	.264	5.151

Based on estimated marginal means

* The mean difference is significant at the .05 level.

a Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

Table A.5: When the Bonferroni correction was removed, most conclusions remained the same. However, the comparison between RBH (FmSsT) and RSD now indicated a significant difference between the means which was not seen with the correction.

A.3 Statistical analyses—Orthology errors

A.3.1 Mislabeled pairs

Within-Subjects Factors
Measure: MEASURE_1

factor1	Dependent Variable
1	FTsF
2	FTsT
3	FmSsF
4	FmSsT
5	RSD

Estimates
Measure: MEASURE_1

factor1	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	.087	.003	.082	.093
2	.086	.003	.081	.092
3	.091	.003	.086	.097
4	.091	.003	.085	.096
5	.096	.003	.090	.102

Tests of Between-Subjects Effects
Measure: MEASURE_1
Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	18.608	1	18.608	1088.103	.000	.705
Error	7.781	455	.017			

Pairwise Comparisons
Measure: MEASURE_1

(I) factor1	(J) factor1	Mean Difference (I-J)	Std. Error	Sig.(a)	95% Confidence Interval for Difference(a)	
					Lower Bound	Upper Bound
1	2	.001(*)	.000	.000	.000	.002
	3	-.004(*)	.001	.000	-.006	-.001
	4	-.003(*)	.001	.011	-.006	.000
	5	-.008(*)	.002	.000	-.013	-.004
2	1	-.001(*)	.000	.000	-.002	.000
	3	-.005(*)	.001	.000	-.007	-.002
	4	-.004(*)	.001	.000	-.007	-.002
	5	-.009(*)	.002	.000	-.014	-.005
3	1	.004(*)	.001	.000	.001	.006
	2	.005(*)	.001	.000	.002	.007
	4	.001(*)	.000	.008	.000	.001
	5	-.005(*)	.001	.015	-.009	-.001
4	1	.003(*)	.001	.011	.000	.006
	2	.004(*)	.001	.000	.002	.007
	3	-.001(*)	.000	.008	-.001	.000
	5	-.005(*)	.001	.002	-.009	-.001
5	1	.008(*)	.002	.000	.004	.013
	2	.009(*)	.002	.000	.005	.014
	3	.005(*)	.001	.015	.001	.009
	4	.005(*)	.001	.002	.001	.009

Based on estimated marginal means

* The mean difference is significant at the .05 level.

a Adjustment for multiple comparisons: Bonferroni.

Table A.6: These data indicate a significant difference between all five conditions; however, no pattern is seen with respect to the mean difference values.

A.3.2 Missing orthologs

Within-Subjects Factors
Measure: MEASURE_1

factor1	Dependent Variable
1	FTsF
2	FTsT
3	FmSsF
4	FmSsT
5	RSD

Estimates
Measure: MEASURE_1

factor1	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	.143	.004	.135	.152
2	.143	.004	.135	.151
3	.143	.004	.135	.152
4	.143	.004	.134	.151
5	.177	.005	.167	.187

Tests of Between-Subjects Effects
Measure: MEASURE_1
Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	51.190	1	51.190	1180.619	.000	.722
Error	19.728	455	.043			

Pairwise Comparisons
Measure: MEASURE_1

(I) factor1	(J) factor1	Mean Difference (I-J)	Std. Error	Sig.(a)	95% Confidence Interval for Difference(a)	
					Lower Bound	Upper Bound
1	2	.000	.000	1.000	-.001	.001
	3	.000	.002	1.000	-.004	.004
	4	.000	.001	1.000	-.004	.005
	5	-.034(*)	.002	.000	-.040	-.028
2	1	.000	.000	1.000	-.001	.001
	3	.000	.002	1.000	-.004	.004
	4	.000	.001	1.000	-.004	.004
	5	-.034(*)	.002	.000	-.040	-.028
3	1	.000	.002	1.000	-.004	.004
	2	.000	.002	1.000	-.004	.004
	4	.000	.000	1.000	-.001	.002
	5	-.034(*)	.002	.000	-.040	-.028
4	1	.000	.001	1.000	-.005	.004
	2	.000	.001	1.000	-.004	.004
	3	.000	.000	1.000	-.002	.001
	5	-.034(*)	.002	.000	-.040	-.029
5	1	.034(*)	.002	.000	.028	.040
	2	.034(*)	.002	.000	.028	.040
	3	.034(*)	.002	.000	.028	.040
	4	.034(*)	.002	.000	.029	.040

Based on estimated marginal means

* The mean difference is significant at the .05 level.

a Adjustment for multiple comparisons: Bonferroni.

Table A.7: These data indicate that RSD produces a significantly higher rate of missing orthologs compared to each of the RBH conditions. All of the RBH conditions were equivalent with respect to the amount of error their results contained.