A STATISTICAL METHOD FOR HIGH-THROUGHPUT SCREENING OF PREDICTED ORTHOLOGS

by

Jeong Eun Min

B.Sc., Simon Fraser University, 2005

a Project submitted in partial fulfillment
of the requirements for the degree of
Master of Science
in the Department
of
Statistics and Actuarial Science

SIMON FRASER UNIVERSITY
Fall 2009

APPROVAL

Name: Jeong Eun Min

Degree: Master of Science

Title of Project: A Statistical Method for High-throughput Screening of Predicted Orthologs

Examining Committee: Dr. Derek Bingham
Chair

____

Dr. Jinko Graham, Senior Supervisor

____

Dr. Brad McNeney, Senior Supervisor

____

Dr. Carl Schwarz, External Examiner

Date Approved: NOV 1 6 2009

Abstract

Orthologs are genes in different species that diverged from a common ancestral gene after speciation. Their identification is critical for reliable prediction of gene function in newly sequenced genomes. Orthologous genes are usually identified by a high-throughput method called Reciprocal-Best BLAST-hit (RBH). As RBH is subject to errors from incomplete sequencing or gene loss in a species, a bioinformatics tool called Ortholuge was developed that identifies RBH-predicted orthologs with atypical genetic divergence. However, declaring the cut-off for atypical divergence in Ortholuge is very computationally-intensive, and so we propose a faster statistical procedure and examine its performance by simulation. We find that performance depends on the fit of the assumed model for the distribution of divergence measures in true orthologs.

Keywords: local false discovery rate; mixture distribution; ortholog; paralog

Declaration of Partial Copyright Licence

Acknowledgments

I would like to express my gratitude to my supervisors, Dr. Jinko Graham and Dr. Brad McNeney who have shown extraordinary commitment, enthusiasm and inspiration. This project would not have been possible without their continuous guidance and support.

I appreciate Dr. Carl Schwarz for providing valuable suggestions and advice on my project. Furthermore, I would like to thank all the professors in the Department of Statistics and Actuarial Science for helping me rediscover my interest in statistics. I am also thankful to the department staff and fellow graduate students who have supported me throughout my graduate program.

I am grateful to Matthew Whiteside and Dr. Fiona Brinkman in the Department of Molecular Biology and Biochemistry for providing data for my project and sharing their knowledge in helping me to complete my project.

Lastly, and most importantly, I would like to thank my family for their endless encouragement and support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Orthologs are genes in different species that have diverged from a common ancestral gene due to a speciation event, whereas genes that have diverged due to a gene duplication event are paralogs (Fitch, 1970). The schematic relationship between orthologs and paralogs is shown in Figure 1.1. Normally, orthologs retain the same function in the course of evolution. Identification of orthologs is critical for reliable prediction of gene function in newly sequenced genomes. Orthologous genes are usually identified by a high-throughput method based on sequence similarity, called Reciprocal-Best BLAST-hit (RBH). In RBH, genes from two species are predicted as orthologs if they are both the best BLAST hit (Altschul et al., 1990) of each other (see Appendix A for a brief overview of BLAST). However, as RBH is subject to errors from incomplete genome sequencing or gene loss in a species, Fulton et al. (2006) developed a method called Ortholuge.

Conceptually, we start with a list of RBH genes for the 2 comparison species, or ingroups, and aim to refine the list by introducing an outgroup as a reference point. The outgroup does not affect the definition of ortholog or paralog, which pertain to the ingroup species. To predict the true orthologs that have similar function between

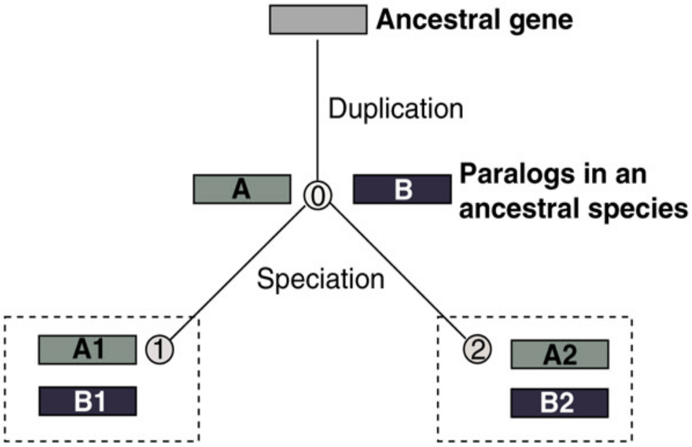

Figure 1.1: Schematic relationship between orthologs and paralogs, taken from Koonin (2001). Gene duplication occurs in species 0, giving rise to gene A and B. Then speciation yields species 1 and 2, each having a set of diverged genes. A1 is orthologous to A2 but paralogous to B1 or B2.

the ingroups, Ortholuge takes RBH-predicted orthologs for the three species as inputs. However, if the outgroup is too phylogenetically distant from the ingroups, quite a number of genes in the original list for the ingroups can be absent in the outgroup due to evolutionary gene loss, so that the best BLAST hit for the outgroup is not an RBH and the gene is dropped. Hence, the choice of outgroup is important for retaining as many genes as possible for analysis.

Ortholuge computes the ratios of phylogenetic distance involving the three species: ratio1 as the distance between the two ingroups (d_{12}) over the distance between ingroup1 and the outgroup ($d_{1\mathrm{O}}$), and ratio2 as the distance between the two ingroups (d_{12}) over the distance between ingroup2 and the outgroup ($d_{2\mathrm{O}}$). Fulton et al. (2006) found that the ratios for orthologs show certain consistencies over several sets of species. The high quality RBH-predicted orthologs tend to have lower ratio values, whereas low quality RBH-predicted orthologs generate higher ratio values. Figure 1.2

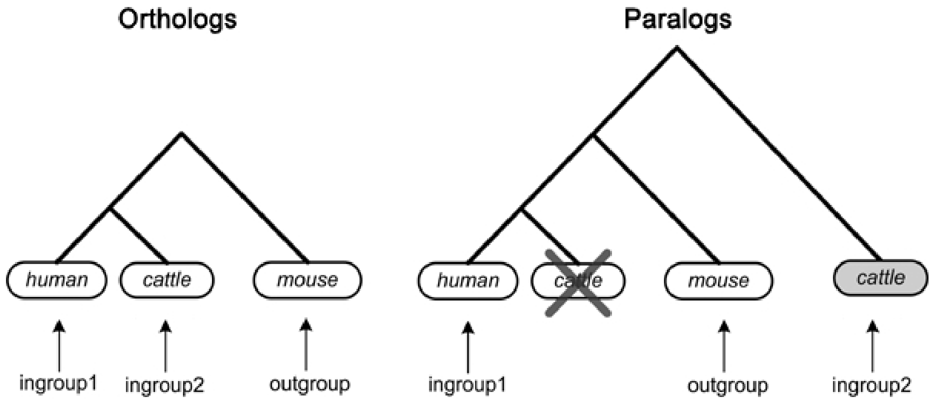

Figure 1.2: An example of why paralogs tend to have distance ratios larger than orthologs, adapted from Fulton et al. (2006). Gene trees for the human, cattle, and mouse species are shown with human and cattle orthologs in the left panel and paralogs in the right panel. Branch lengths on the trees indicate evolutionary distance. If the cattle gene orthologous to the human gene is not present in the data set (cattle gene crossed out with an X in the right panel), RBH may predict a cattle paralog (shaded gene) as an ortholog. Ortholuge aims to detect this case as a paralog by introducing an outgroup, mouse. For the paralogous cattle gene, the human-cattle distance (d_{12}) is unexpectedly larger than the human-mouse distance (d_{1O}), so that ratio1 (d_{12}/d_{1O}) for the gene is increased relative to the other genes in the data set.

illustrates why paralogs tend to have larger distance ratios. Thus, they proposed identifying atypically diverged gene pairs within the list of RBH-predicted orthologs by examining these distance ratios. The gene pairs with atypical divergence are considered as either paralogs that have been falsely predicted as orthologs or as orthologs that have diverged more rapidly in one species, relative to another, and may have evolved different functions ("non supporting-species-divergence" or non-ssd-orthologs, as defined by Fulton et al. (2006)).

In order to determine the Ortholuge ratio cut-offs to distinguish between typical (ssd-orthologs) and atypical divergence (non-ssd-orthologs), Fulton et al. proposed generating true-negative gene triplets by knocking out one of the ingroup genes in the data sets and then taking the next reciprocal-best BLAST hit with the other ingroup (Figure 1.3). The true-negative gene triplets represent the type of errors that the RBH prediction method would make due to incomplete genome sequencing or gene loss in a species. If ingroup1 was a paralog (i.e. the ingroup1 gene is knocked out), then ratio2 would tend to be increased because the numerator distance d_{12} would tend to get bigger while the denominator distance $d_{2\mathrm{O}}$ would stay the same. By contrast, ratio1 would not necessarily be increased because both d_{12} and $d_{1\mathrm{O}}$ would tend to get bigger. Thus, ratio2 would more easily detect ingroup1 paralogs than ratio1. Conversely, if the ingroup2 gene was knocked out, then ratio1 would be increased more than ratio2, and ratio1 would more easily detect ingroup2 paralogs. To determine the cut-off, Fulton et al. proposed randomly replacing 25% of genes in a RBH-predicted data set with ingroup2 true-negatives for ratio1 analysis or with ingroup1 true-negatives for ratio2 analysis. For this transformed data set, the histogram of resulting distance ratios could be plotted, and the proportion of true-negatives introduced in each bin of ratios could be computed (Figure 1.4). They proposed iterating this random introduction of true-negatives 50 to 100 times and averaging the proportion of true-negatives over all iterations. They found that the true-negatives were distributed at higher values of ratios than the RBH-predicted orthologs. Specifically, there were no true-negatives at the left-most bins, and as they moved to the right the proportion of the true-negatives increased. Two cut-off points were set at the first bins moving right such that the proportions of true-negatives in those bins were 10% and 50%. The combination of cut-offs for ratio1 and ratio2 is used to classify orthology (Figure 1.5). Throughout, we use the term “cut-off” in this sense. The RBH-predicted orthologs lying below the 10% cut-off point for both ratio1 and ratio2 were classified as probable orthologs, and those at or above the 50% cut-off point for either ratio1 or ratio2 were classified as probable paralogs. The remaining RBH-predicted orthologs were categorized as orthology “uncertain.”

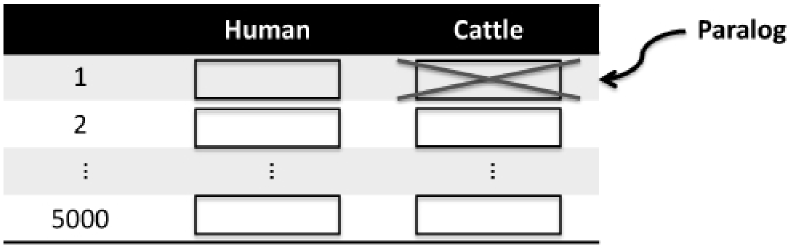

Figure 1.3: An example of a true-negative introduced from a list of 5000 RBH-predicted orthologs.

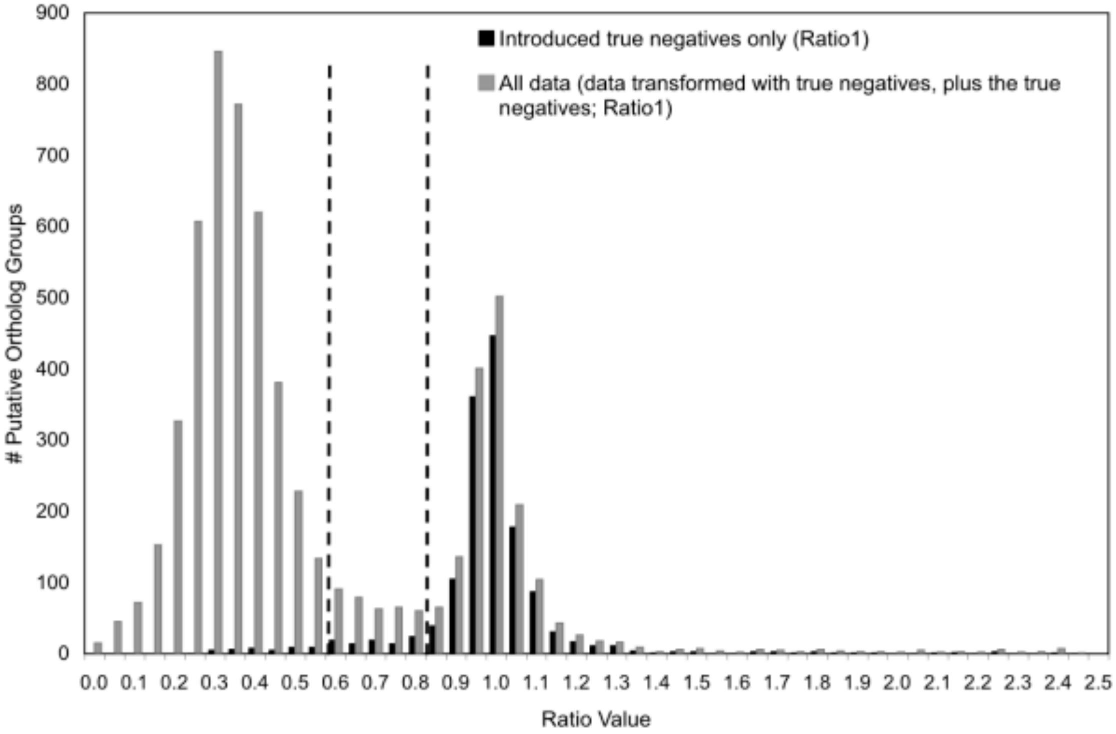

Figure 1.4: Example of the generation of cut-offs in Ortholuge, taken from Fulton et al. (2006). A histogram of ratio1 for a mouse, rat, and human RBH data set is shown. The light shaded bars are from the whole data set and the dark shaded bars are from the introduced true-negatives only. The proportion of randomly introduced true-negatives at each interval is averaged over iterations to generate cut-offs. The two vertical dashed lines represent the 10% and 50% cut-offs.

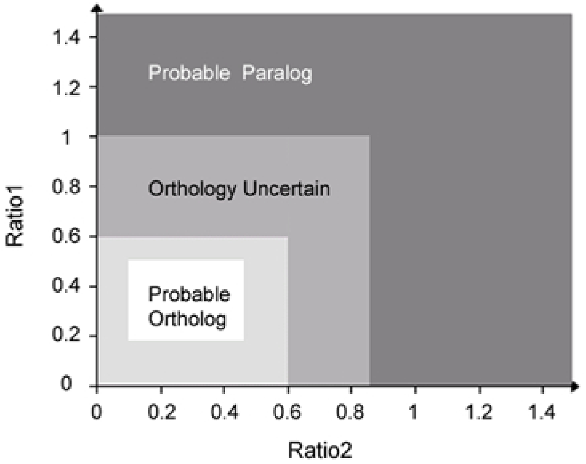

Figure 1.5: Example of how Ortholuge classifies orthology, taken from Fulton et al. (2006). The cut-off numbers shown in the figure are the 10% and 50% cut-offs in the example of the mouse-rat-human analysis from Figure 1.4.

Fulton et al. found that the resulting Ortholuge method significantly improved the correct identification of orthologs in an RBH-based ortholog analysis. However, the required iterative true-negative analysis for each species comparison was very time-consuming. In particular, the alignment of the transformed gene triplets for the computation of phylogenetic distance is computationally demanding. Moreover, as discussed in Appendix B, the Ortholuge 50% true-negative cut-off for probable paralogs can correspond to actual paralog proportions that are in fact greater than 50% when the relative frequency of paralogs in the data is more than 1/6. Intuitively, the cut-offs are shifted to the right because the gene triplets with introduced true-negatives have larger ratio values than before and this shifts the distribution of the transformed ratios to the right of the untransformed ratios.

To avoid the time-consuming iterative true-negative analysis in Ortholuge, we develop a new statistical method, OL.locfdr, for declaring a cut-off for phylogenetic distance ratios. Our method saves time because it does not require simulation and alignment of true-negative genes to distinguish between typical and atypical divergence. Rather than transforming a subset of the gene triplets to be true-negatives by gene knockout and using the proportions of true-negative genes to indirectly distinguish orthologs from paralogs, the method bypasses the transformation step and estimates the proportions of paralogs directly from the data. Like Fulton et al., we view the data as a sample from a mixture of an ortholog and paralog distribution. The problem of identifying an ortholog/paralog cut-off can thus be phrased in terms similar to those proposed by Efron (2004) for identifying unusual cases sampled from a mixture distribution. Our approach allows estimation of the expected proportion of paralogs (genes with atypical divergence) at a particular ratio value. This expected proportion of paralogs is one minus the local false-discovery rate (fdr) of Efron.

Through simulations based on real data, we examine how overlap between the ortholog and paralog distributions of distance ratios affects the performance of the method. The overlap between the distributions of orthologs and paralogs depends on the choice of an outgroup. The closer the outgroup is to the ingroups, the more the ortholog distribution shifts towards one and towards the paralog distribution. This shifting occurs because if the outgroup is too close to the ingroups, the distance between the two ingroups, d_{12}, is similar to the distance between the ingroup and the outgroup, d_{IO}, for true orthologs. Since the ratio distance is defined as d_{12}/d_{IO}, the ratios for true orthologs will tend to be closer to 1 than they would be with a more distant outgroup. On the other hand, an outgroup farther away from the ingroups will have orthologs whose ratios tend to be closer to zero because d_{12} is much smaller than d_{IO}. The disadvantage of choosing a too-distant outgroup is losing tentative orthologous genes for analysis because, if the best BLAST hit for the outgroup is not an RBH, the gene is dropped.

Chapter 2

Method

2.1 Transformation of data

Initially we worked with a *Burkholderia* bacterial data set having *B. cepacia* and *B. cenocepacia* as ingroup species and *B. pseudomallei* as an outgroup species to develop our method. The histograms of both ratio1 and ratio2 were right-skewed around the major mode with a long right tail and some extreme outliers (Figure 2.1). This was the typical pattern for ratio1 and ratio2. Our objective was to mathematically transform ratios so that the underlying mixture distribution could be approximated by a mixture of two normal distributions. In order to pull in the extreme outliers toward the rest of the distribution, we tried log transformations but found that the transformed distributions were left-skewed (Figure 2.2). We adopted a less drastic square-root transformation because it pulled in the extreme outliers without left-skewing (Figure 2.3). The major mode was now roughly symmetric and there was a slight minor mode in the right tail that could be interpreted as possibly containing paralogous genes. We used the square-root transformation of the ratios for all our analyses.

We assumed that the ortholog distribution could have two parts: a zero-inflated part ($p_{0a}f_{0a}$; see Figure 2.3 and Section 2.2) comprised of ratios near zero and the remaining ortholog distribution to the right. The ratios in the zero-inflated part of the distribution are due to (effectively) zero distance between genes in the two ingroups, scaled by the denominator of the ratio.

2.2 Statistical Approach

Consider data of z values from a mixture distribution f. Each z is generated from the null distribution f_0 with probability p_0, or from the alternative distribution f_1 with probability p_1:

$$f(z) = p_0 f_0(z) + p_1 f_1(z).$$

Efron (2004) defines the local false-discovery rate, fdr(z), as the chance of being from the null distribution given the observed value z:

$$\mathrm{fdr}(z) = \Pr(null \mid z) = \frac{\Pr(null) f(z \mid null)}{f(z)} = \frac{p_0 f_0(z)}{f(z)}. \tag{2.1}$$

We consider orthologs to be generated from the null distribution f_0 and paralogs to be generated from the alternative distribution f_1. However, unlike Efron, we assume that the null distribution consists of two sub-distributions, f_{0a} and f_{0b}, and the mixture density can be written as

$$f(z) = p_{0a}f_{0a}(z) + p_{0b}f_{0b}(z) + (1 - p_{0a} - p_{0b})f_1(z).$$

The components of the mixture distribution are depicted in Figure 2.4. The sub-distribution $p_{0a}f_{0a}$ corresponds to the zero-inflated part of the ortholog distribution indicated in Figure 2.3 for the *Burkholderia* data. The sub-distribution $p_{0b}f_{0b}$ is

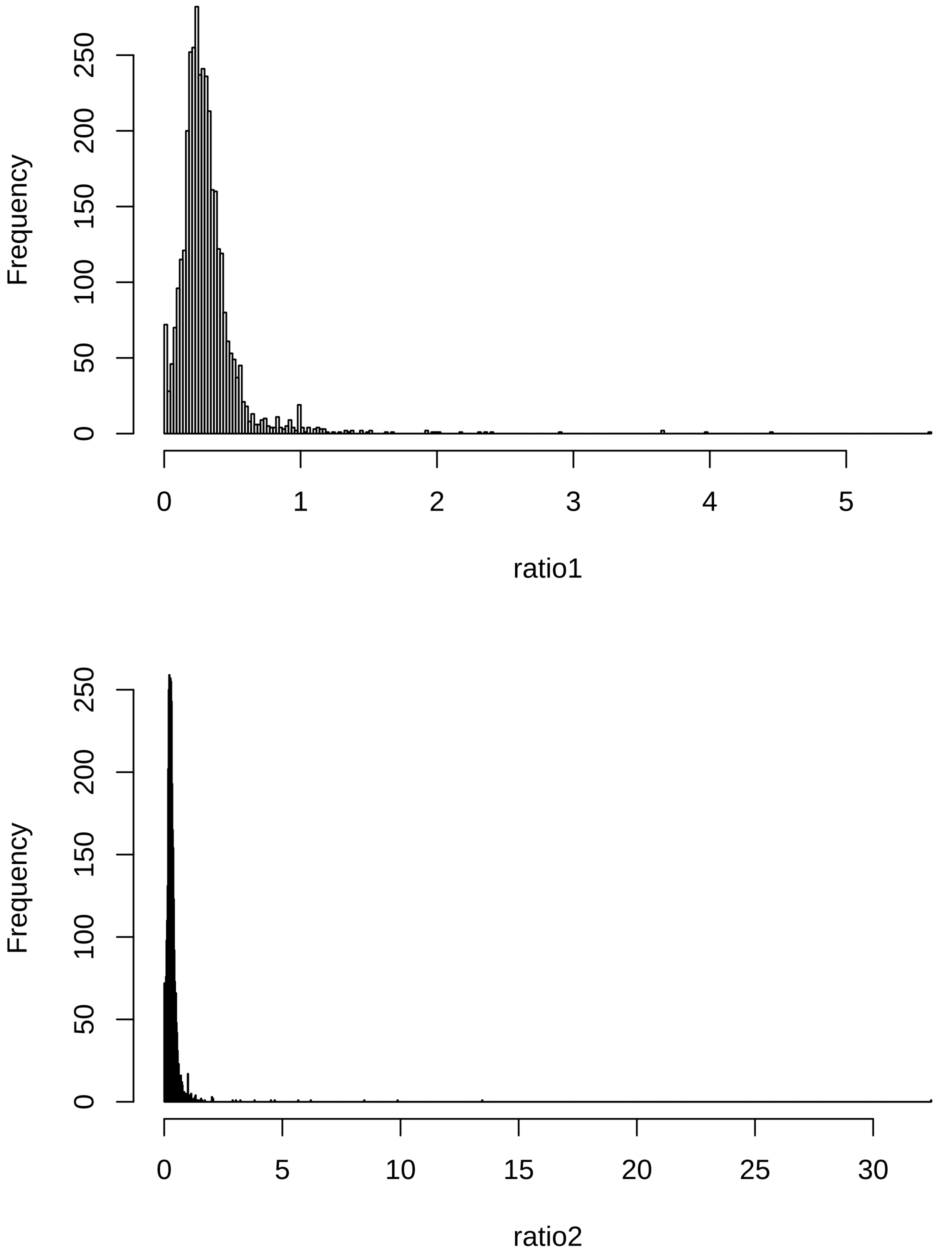

Figure 2.1: Histograms of ratio1= d_{12}/d_{1O} and ratio2= d_{12}/d_{2O} for the *Burkholderia* data.

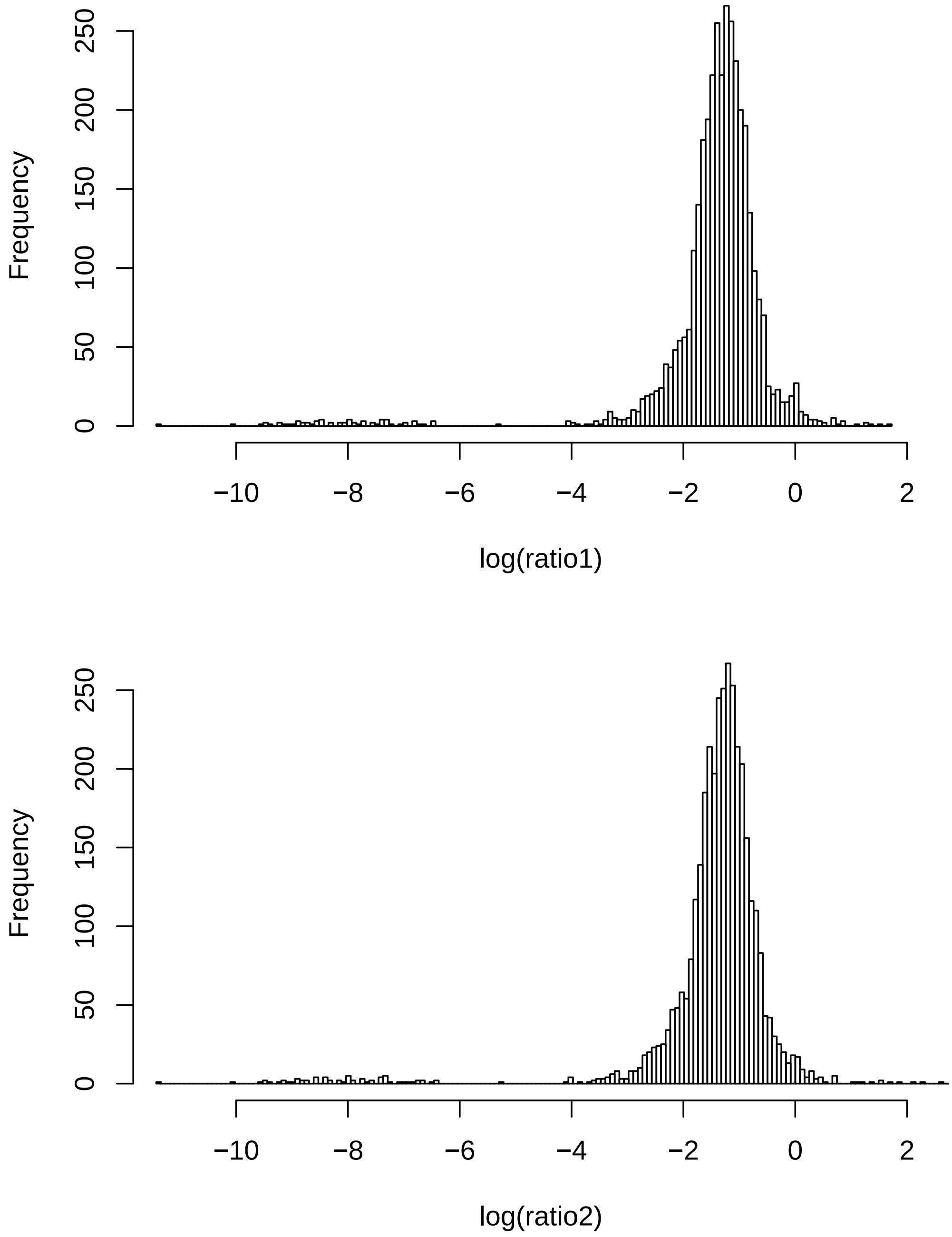

Figure 2.2: Histograms of log(ratio1) and log(ratio2) for the *Burkholderia* data.

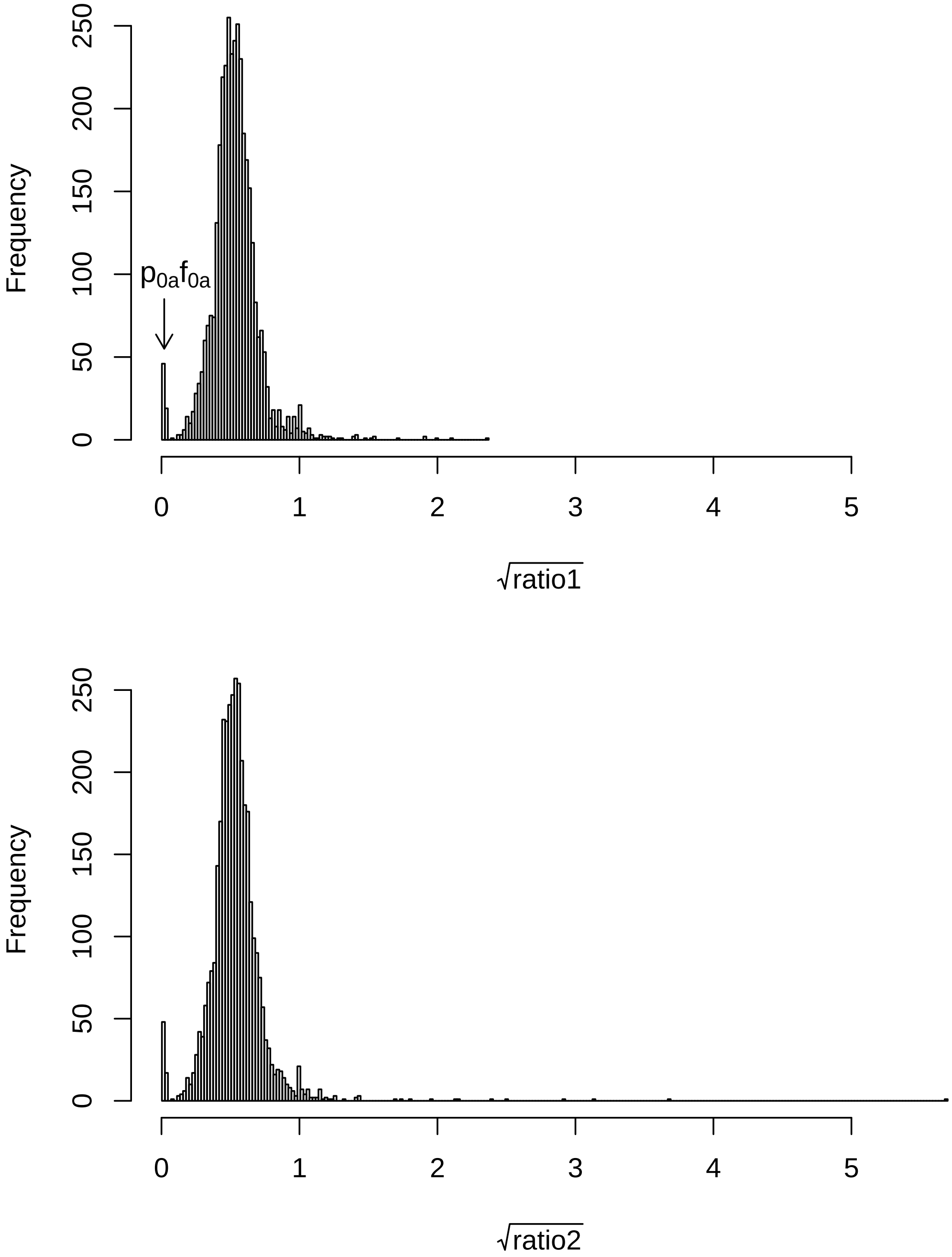

Figure 2.3: Histograms of $\sqrt{\text{ratio1}}$ and $\sqrt{\text{ratio2}}$ for the *Burkholderia* data.

assumed to be normal and centred at the major mode of the mixture distribution, f; it is the main ortholog sub-distribution. Hence, the null distribution is

$$f_0(z) = \frac{p_{0a}f_{0a}(z) + p_{0b}f_{0b}(z)}{p_0}.$$

Therefore, in our case,

$$\text{fdr}(z) = \frac{p_{0a}f_{0a}(z) + p_{0b}f_{0b}(z)}{f(z)}.$$

We assume that over a critical region (l, u) of values centred at the major mode of the mixture distribution, the distributions f_{0a} and f_1 have negligible mass. Rather, f_{0a} places virtually all its mass at $z < l$ and f_1 places virtually all its mass at $z > u$. A consequence is that, at $z < u$, $f(z) \approx p_{0a}f_{0a}(z) + p_{0b}f_{0b}(z)$ and

$$\text{fdr}(z) = \frac{p_0 f_0(z)}{f(z)} = \frac{p_{0a}f_{0a}(z) + p_{0b}f_{0b}(z)}{f(z)} \approx \frac{p_{0a}f_{0a}(z) + p_{0b}f_{0b}(z)}{p_{0a}f_{0a}(z) + p_{0b}f_{0b}(z)} = 1.$$

Conversely, at $z > u$, $p_{0a}f_{0a}(z) \approx 0$ and

$$\text{fdr}(z) = \frac{p_0 f_0(z)}{f(z)} = \frac{p_{0a}f_{0a}(z) + p_{0b}f_{0b}(z)}{f(z)} \approx \frac{p_{0b}f_{0b}(z)}{f(z)}. \tag{2.2}$$

We take the range (l, u) to be $M_f \pm \text{IQR}/2$, where M_f is the major mode of f and IQR is the interquartile range of f. To estimate fdr's for $z > u$, f and $p_{0b}f_{0b}$ are estimated as described in the following steps.

Step0: Remove genes with extreme distance ratios. Declare genes with distance ratios larger than 100 times the median of the data to be paralogs and remove them from the analysis. The reduced gene list defines an upper limit $z_{\max}$ for the observed root ratios that is used in subsequent steps.

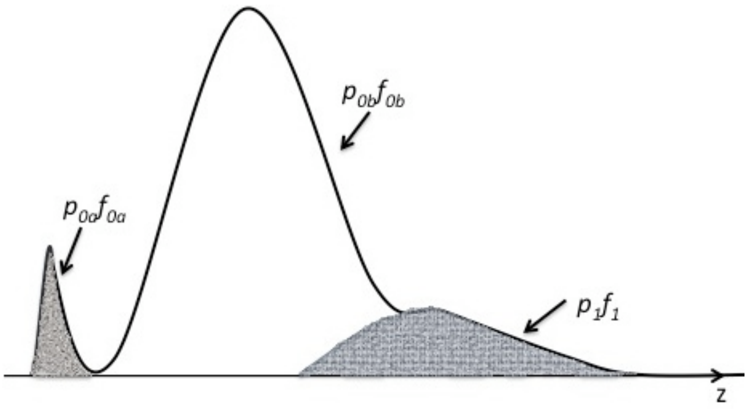

Figure 2.4: Components of the mixture distribution.

Step1: Estimate and display the mixture density f**.** We use a kernel density estimate $\hat{f}$ of f, with Gaussian kernel and reflection about zero to account for the boundary there (Silverman, 1986). Before reflection, a Gaussian kernel density estimator, $\hat{f}_G$, based on n data points, is an equally-weighted mixture of n normal densities, where the ith density has mean equal to the ith data point and standard deviation equal to the kernel bandwidth. After reflection about 0, we obtain $\hat{f}(z) = \hat{f}_G(z) + \hat{f}_G(-z)$ for $z \geq 0$. Following Silverman (1986), we implement the reflection method by:

1. augmenting the data by the negative of all data points,
2. applying a kernel density estimator to the augmented data,
3. setting the resulting density estimates to
   (a) zero for augmented data values less than zero, and
   (b) double their values for augmented data values greater than or equal to zero.

We select the kernel bandwidth b using the "NRD" selection algorithm (Scott, 1992) applied to the original data. Density estimates are obtained over a grid of 1024 points spanning the range of the augmented data $\pm 3b$. The density estimation function we use returns density estimates at 512 equi-spaced points by default. We opt for twice the default number of points because the reflection method doubles the range of data values at which density estimates are required. Density estimates for root-ratio values between grid-points are obtained by linear interpolation of estimates at adjacent grid-points. Density estimates outside the grid of 1024 points are set to zero. To display $\hat{f}$, we superimpose it over a histogram of the data, with breakpoints determined by the Freedman-Diaconis algorithm (Freedman and Diaconis, 1981). On the display, we scale up $\hat{f}$ so that it integrates to the same value as the area under the histogram.

Step2: Estimate the main ortholog sub-distribution about the major mode. The goal is to estimate $p_{0b}f_{0b}$. We first estimate $\log[p_{0b}f_{0b}]$ and then exponentiate this estimate. To estimate $\log[p_{0b}f_{0b}]$, we assume that f_{0b} is a normal density. Parametric modelling of $p_{0b}f_{0b}$ is required in order to be able to identify it as a component of the non-parametric mixture distribution f. When f_{0b} is a normal density, it can be shown that $\log[p_{0b}f_{0b}(z)]$ is quadratic in z. Furthermore, for z near the major mode of f, we have $f(z) \approx p_{0b}f_{0b}(z)$ or $\log[f(z)] \approx \log[p_{0b}f_{0b}(z)]$. Thus, following Efron (2004), $\log[p_{0b}f_{0b}]$ can be estimated by fitting a quadratic function to the estimate of $\log[f(z)]$, for z near the major mode of f. We consider a grid of 200 points $\{z_l, \ldots, z_u\}$ spanning $M_{\hat{f}} \pm \mathrm{IQR}/2$, where $M_{\hat{f}}$ is the major mode of $\hat{f}$ and IQR is the empirical IQR. This grid and the associated estimates $\{\log \hat{f}(z_l), \ldots, \log \hat{f}(z_u)\}$ are used as data to which a quadratic curve is fit by ordinary least squares. This fitted quadratic curve estimates $\log[p_{0b}f_{0b}]$ everywhere and in particular in the right tail of the mixture density where calculation of fdr's is required.

Step3: Calculate fdr's and find the cut-off. For $z \leq z_u$, the estimated fdr's (i.e. the proportion of orthologs) are set to 1. For $z_u < z \leq z_{\max} + 3b$, the estimated fdr's are the minimum of one and the expression in equation (2.2), computed using the estimated $f(z)$ and $p_{0b}f_{0b}(z)$. For $z > z_{\max} + 3b$, the estimated fdr is set to 0. The methods of estimating f and $p_{0b}f_{0b}$ do not enforce the constraint that $\widehat{p_{0b}f_{0b}}(z) \leq \hat{f}(z)$. For estimating fdr, the constraint becomes an issue for $z_u < z < z_{max} + 3b$ such that $\widehat{p_{0b}f_{0b}}(z) > \hat{f}(z) = 0$. For these z, we set $\hat{f}(z) = \widehat{p_{0b}f_{0b}}(z)$ and fdr to be one.

The cut-off to detect a gene pair as unusually diverged is determined according to the target probability for paralogs (i.e. 1−fdr). Multiple target probabilities can be considered. For example, if 60% is set as the target probability, then the 60% cut-off is estimated by the smallest z-value at which $1 - \text{fdr}(z) = 60\%$. Therefore, genes at the cut-off are estimated to have at least 60% chance of being unusually diverged.

Step4: Report the square-root ratio values with cut-offs. For each gene, the conditional probability of being unusually diverged given its square-root ratio value z (i.e. 1−fdr(z)) may be estimated. With multiple target probabilities provided by the user, each gene may be classified into a group based on the corresponding cut-offs. For example, group 0 includes the genes whose square-root ratios are less than all cut-offs (null genes), and group 1 includes genes whose square-root ratios fall between the first and second cut-offs, and so on.

The steps outlined above were implemented in an R function `OL.locfdr()` described in Appendix C.

Chapter 3

Simulation study

We conducted a simulation study to evaluate the performance of OL.locfdr. One question of interest was how performance depends on the overlap of the ortholog and paralog distributions. The greater the overlap, the harder it should be to distinguish orthologs from paralogs. The idea behind Ortholuge is that, for outgroups of sufficient evolutionary distance from the ingroups, paralog ratios will tend towards values of one or more but that ortholog ratios will tend towards values less than one. For outgroups that are closer to the ingroups, ortholog ratios will tend more towards one so that the ortholog distribution overlaps more with the paralog distribution. We therefore considered several species sets that have the same ingroups but outgroups of varying distance. Another question of interest was how performance depends on the fit of the assumed normal distribution to the true ortholog distribution. Specifically, the right tail of the ortholog distribution is assumed to behave like the right tail of a normal distribution. Thus, if the assumed normal distribution is lighter-tailed than the true ortholog distribution, the estimated fdr's are expected to be lower than the true fdr's, and the estimated cut-off will tend to be lower than it should be. By contrast, if the assumed normal distribution is heavier-tailed than the true ortholog distribution, the estimated cut-off will tend to be higher than the true cut-off. We therefore considered species sets with ortholog distributions that were lighter- or heavier-tailed than the assumed normal distributions.

To describe the operating characteristics of our procedure, we defined bias in two ways. The first definition of bias is the expected difference between the estimated and true proportion of orthologs at a fixed location. We examined this type of bias over a range of locations. The second definition of bias is the expected difference between the estimated and true 50% cut-off, separating probable non-SSD orthologs (paralogs) from orthology uncertain.

3.1 Available data sets

To evaluate the procedure under realistic conditions, we were provided several species sets from the Brinkman lab (personal communication with M. Whiteside). As shown in Table 3.1, there were 6 species sets with the same ingroup species, *Escherichia coli* and *Salmonella Typhimurium*, but with outgroups of varying evolutionary distances from the ingroups (see $\overline{d_{\text{IO}}}$ column of the table). These outgroups were *Yersinia pestis*, *Haemophilus influenzae*, *Vibrio cholerae*, *Pseudomonas aeruginosa*, *Xanthomonas campestris*, and *Burkholderia pseudomallei.* Also, there were 2 species sets with ingroup species *Pseudomonas putida* and *Pseudomonas syringae* but with different outgroup species, *P. aeruginosa* and *E. coli.*

Each species set had a number of associated data sets derived from analyzing the protein sequences for genes. First, there was a basic data set with all the RBH genes for the two ingroups and the outgroup. Second, there was a true ortholog subset of the basic data set, identified by Lerat et al. (2003), provided as a candidate for the ortholog distribution. Next there were four true-negative (TN) data sets generated by knocking out the RBH from the ingroups and replacing it with either the next-best

Table 3.1: List of species sets, genetic distance of outgroup from ingroups, and the number of genes contained in each associated data set

					Number of genes					
Species set	Ingroup1	Ingroup2	Outgroup	$\overline{d_{\mathrm{IO}}}^*$	basic	true ortholog	TN-basic$_{\mathrm{I1}}$	TN-basic$_{\mathrm{I2}}$	TN-ortholog$_{\mathrm{I1}}$	TN-ortholog$_{\mathrm{I2}}$
1	*E. coli*	*S. typhimurium*	*Y. pestis*	0.384	2160	166	100	109	-**	-
2	*E. coli*	*S. typhimurium*	*H. influenzae*	0.669	1168	151	50	61	-	-
3	*E. coli*	*S. typhimurium*	*V. cholerae*	0.730	1738	104	71	88	-	-
4	*E. coli*	*S. typhimurium*	*P. aeruginosa*	0.900	1678	163	79	80	-	-
5	*E. coli*	*S. typhimurium*	*X. campestris*	1.009	1314	53	57	71	-	-
6	*E. coli*	*S. typhimurium*	*B. pseudomallei*	0.975	1468	146	73	73	-	-
7	*P. putida*	*P. syringae*	*P. aeruginosa*	0.392	2551	146	317	257	42	40
8	*P. putida*	*P. syringae*	*E. coli*	0.885	1463	156	151	119	49	45

[*] $\overline{d_{\mathrm{IO}}} = (\text{average of } d_{1\mathrm{O}} + \text{average of } d_{2\mathrm{O}})/2$ from the basic data sets.

[**] The dash sign indicates a missing value.

RBH or the next-best BLAST hit. These TN data sets were provided as candidates for the paralog distribution and were

- a TN subset of the basic data set generated by knocking out the ingroup1 (I1) RBH and replacing it with the next-best RBH. If the next-best BLAST hit was not an RBH the gene was dropped. We label this data set $\text{TNbasic}_{\text{I1}}$.
- a TN subset of the basic data set generated by knocking out the ingroup2 (I2) RBH and replacing it with the next-best RBH. We label this data set $\text{TNbasic}_{\text{I2}}$.
- a TN subset of the true ortholog data set generated by knocking out the ingroup1 RBH and replacing it with the next-best BLAST hit. We label this data set $\text{TNortholog}_{\text{I1}}$.
- a TN subset of the true ortholog data set generated by knocking out the ingroup2 RBH and replacing it with the next-best BLAST hit. We label this data set $\text{TNortholog}_{\text{I2}}$.

The last two TN data sets generated from the true ortholog data set used the next-best BLAST hit rather than the RBH because imposing reciprocity of BLAST hits leads to too few genes (e.g. 5 out of 156 for species set 8). However, the resulting TNortholog data sets generated without imposing the RBH criterion had ratios shifted towards larger values than the TNbasic data sets generated with the RBH restriction (Figure 3.1). The TNbasic data set rather than the TNortholog data set was chosen to represent paralogs because its ratios were closer to the ortholog distribution. As shown in Table 3.1, the basic data sets all have over a thousand genes. The true ortholog and TNbasic subsets were extracted from the basic data set and so have a smaller number of genes. The data sets $\text{TNortholog}_{\text{I1}}$ and $\text{TNortholog}_{\text{I2}}$ have the least number of genes because they were extracted from the true ortholog data sets.

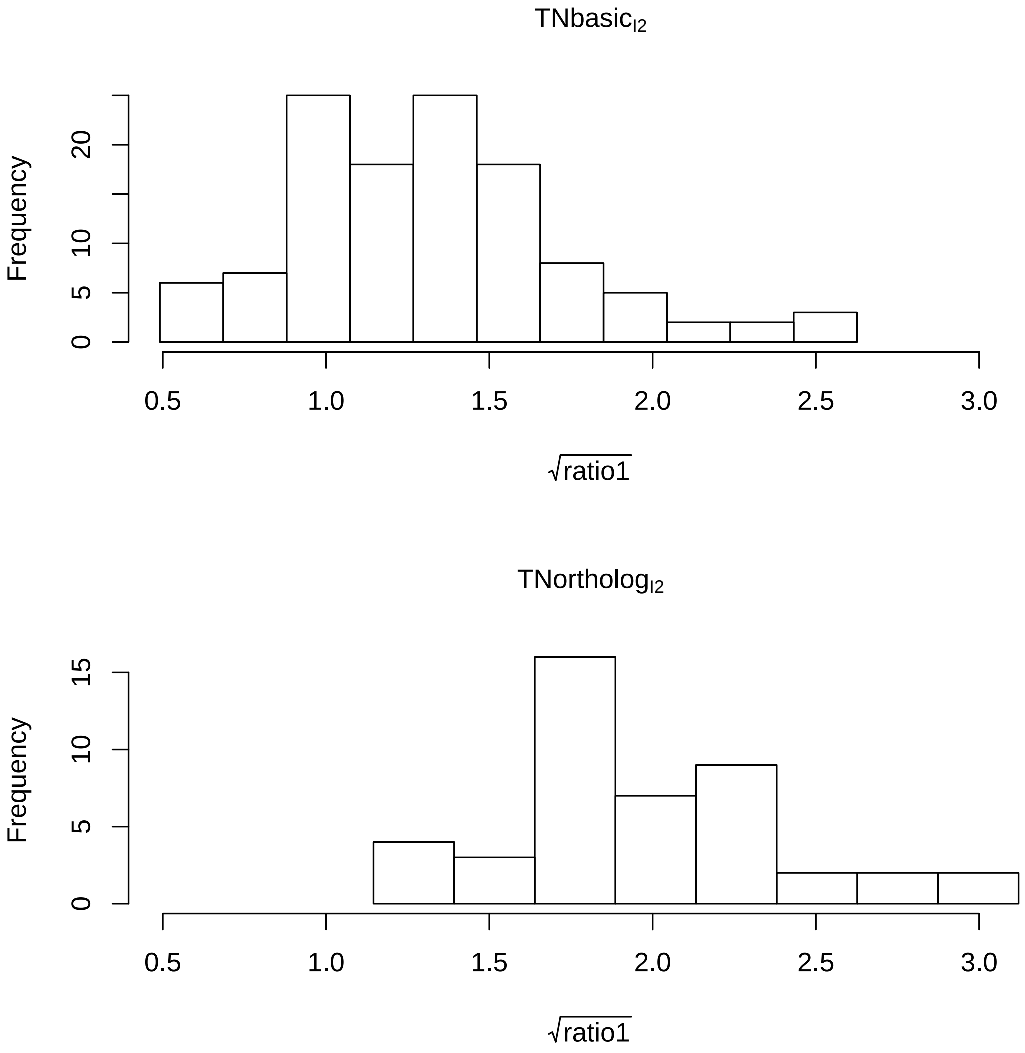

Figure 3.1: Histograms of $\sqrt{\text{ratio1}}$ from two TN data sets for species set 8.

3.2 Design of simulations

We simulated square-root ratios from mixtures of ortholog and paralog distributions estimated from real data. We fit the null distribution, f_0, from the true ortholog data set and alternative distributions, f_{11} and f_{12}, from the TNbasic$_{\mathrm{I1}}$ and TNbasic$_{\mathrm{I2}}$ data sets, respectively. The method for fitting distributions to the data sets is explained in the next section. These data sets were considered as samples from the underlying ortholog and paralog distributions. Because the data sets contained small numbers of samples (Table 3.1), they do not give an accurate picture of each distribution, particularly in the tails. To obtain a less discrete representation of the underlying distributions, we fit a density curve to our samples and then simulated from the estimated densities rather than resampling from the data sets.

We simulated a data set of 5000 square-root ratios according to the proportion p_1 of the mixture distribution, f, contaminated with paralogs, such as $p_1 = 5\%$ and $p_1 = 25\%$. The mixture density was $f = (1 - p_1)f_0 + p_1 f_1$, where the composition of the paralog density f_1 was determined as below, depending on which ingroup was knocked out:

- f_1 contains 100% ingroup1 (I1) knockouts, i.e. $f_1 = f_{11}$;
- f_1 contains 50/50 I1/I2 knockouts, i.e. $f_1 = (f_{11} + f_{12})/2$;
- f_1 contains 100% ingroup2 (I2) knockouts, i.e. $f_1 = f_{12}$.

To obtain a sample from a distribution of 50/50 I1/I2 knockouts, we randomly sampled ratios from the I1 knockout paralog distribution, f_{11}, with chance 0.5 whenever it was time to sample from f_1 in the simulation. The rest of the ratios for paralogs were sampled from the I2 knockout paralog distribution, f_{12}.

For each simulated data set, we ran the `OL.locfdr` procedure to estimate

1. the proportion of orthologs at a set of fixed points with proportions assumed to be less than 1 and
2. the 50% cut-off.

We compared our results for each simulated data set with the true fdr (i.e. the expected proportion of orthologs) and the true 50% cut-off as explained in Section 3.6. Estimates of bias are averages over 5000 simulated data sets.

3.3 Obtaining the ortholog and paralog densities

Our goal was to obtain estimates of f_0 from the true ortholog data set and estimates of f_{11} and f_{12} from the TNbasic$_{\text{I1}}$ and TNbasic$_{\text{I2}}$ data sets, respectively. We fit density curves as described in Section 2.2, Step 1. Figure 3.2 shows the histograms from species set 8 for ratio1 and the fitted distributions superposed. The results from ratio2 are similar (not shown), except the characteristics of f_{11} and f_{12} are swapped.

To simulate from the resulting densities, we reasoned as follows. A Gaussian kernel density estimator, $\hat{f}_G$, based on n data points, is an equally-weighted mixture of n normal densities where the ith density, $i = 1, \ldots, n$, has mean equal to the ith data point and standard deviation equal to the kernel bandwidth b. A Gaussian kernel density estimator with reflection about 0 is then $\hat{f}(z) = \hat{f}_G(z) + \hat{f}_G(-z)$ for $z \geq 0$. Thus, to sample a point from an estimated density, we (i) randomly sampled a data point, (ii) randomly sampled from a normal distribution with mean equal to the sampled data point and standard deviation equal to b, and (iii) took the absolute value of the result. Steps (i) and (ii) sample from $\hat{f}_G(z)$, while step (iii) reflects any negative support of $\hat{f}_G(z)$ to above zero.

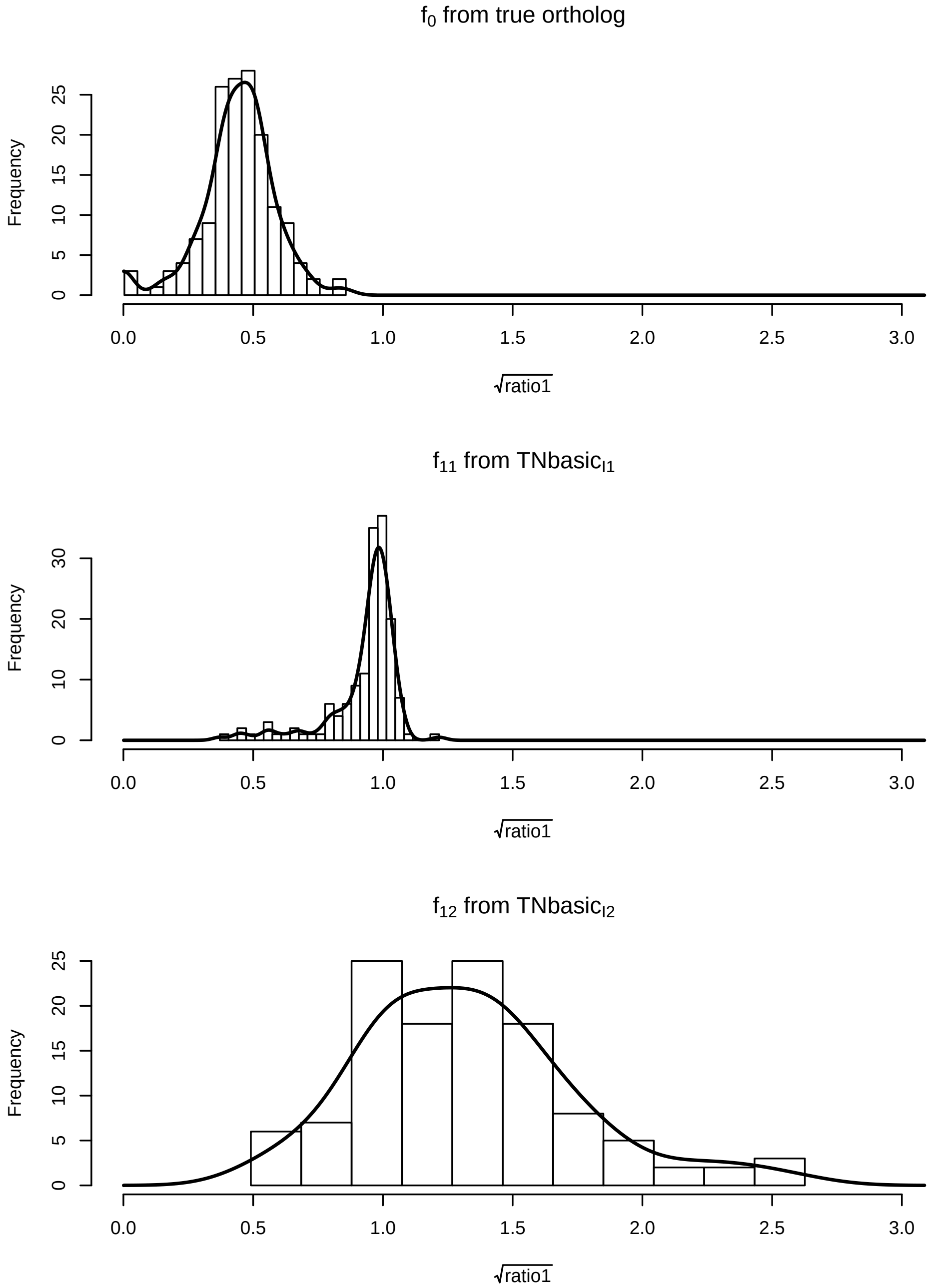

Figure 3.2: Histograms of $\sqrt{\text{ratio1}}$ for the data sets true ortholog, TNbasic$_{\text{I1}}$, and TNbasic$_{\text{I2}}$ of species set 8. The superimposed lines indicate the fitted distributions.

3.4 Overlap between the ortholog and paralog distributions

We defined q as the proportion of f coming from f_1 in a critical region $(l, u]$, taken to be $M_f \pm \mathrm{IQR}/2$, where M_f is the major mode of f. The critical region was used to estimate f_{0b} from a known f. The overlap measure, q, was obtained by

$$q = \frac{\sum_{z_i \in (l,u]} f_1(z_i) p_1}{\sum_{z_i \in (l,u]} f(z_i)},$$

where the z_i's are a set of 200 grid points spanning $(l, u]$. The overlap measure was computed for each proportion p_1 and composition of f_1 described in Section 3.2. As an example, Figure 3.3 shows how f_1 contributes to f in the critical region and how this contribution affects q in species set 7 when f_1 is 100% I1 knockouts.

Contamination of major mode. A key assumption of OL.locfdr procedure is that there is no paralog contamination near the major mode of the mixture distribution, in the critical region used to fit the ortholog distribution. The values of the overlap measure q in Table 3.2 quantify departures from this assumption. Most of the values indicate very low contamination but there are a few cases that stand out.

For example, there are two interesting cases of extreme contamination in the sense that the major mode is due to paralogs rather than orthologs. Both cases involve species set 1 and the high mixing proportion $p_1 = 25\%$: ratio1 for 100% I1 knockouts ($q = 0.909$) and ratio2 for 100% I2 knockouts ($q = 0.955$). In both these cases, the major mode is due to paralogs because f_1 is much more concentrated relative to f_0, as shown in Figure 3.4. The essentially complete contamination of the major mode clearly invalidates the use of the OL.locfdr procedure. We therefore eliminated species set 1 from further investigation.

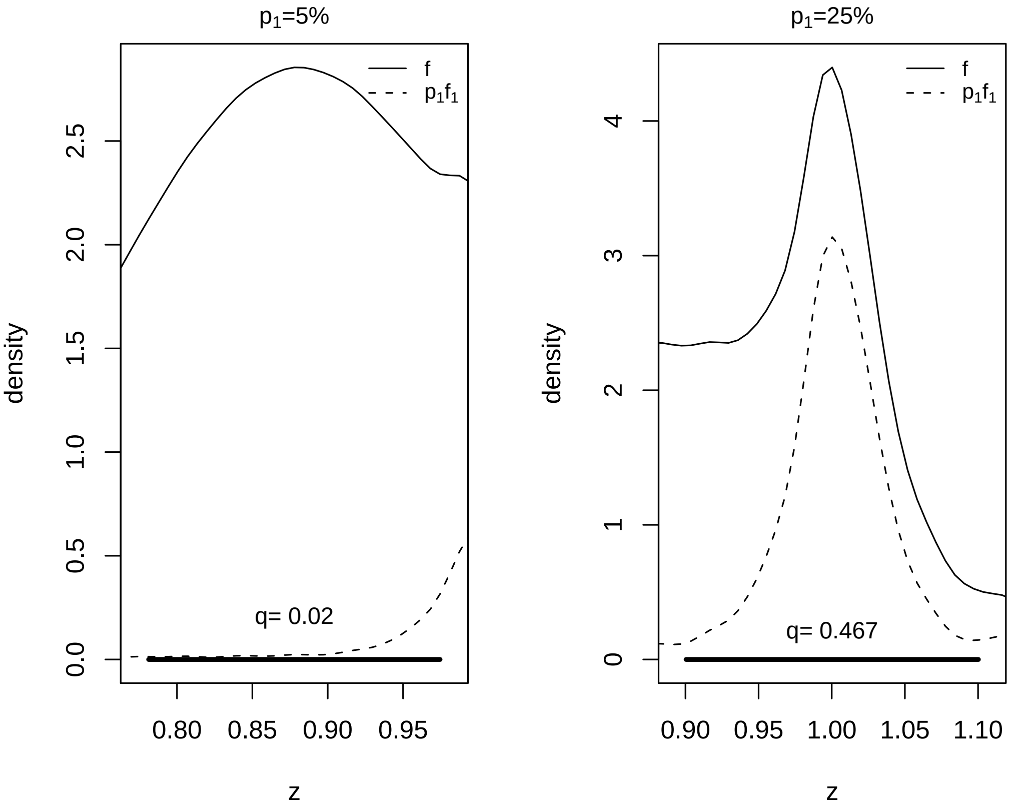

Figure 3.3: Relative contribution of f_1 to f in the critical region for ratio1 in species set 7 when f_1 is 100% I1 knockouts.

Another interesting set of cases comes from species set 7 when $p_1 = 25\%$. For cases such as ratio1 with 100% I1 knockouts ($q = 0.467$) or ratio2 with 100% I2 knockouts ($q = 0.446$), the majority of the probability mass near the major mode of the mixture distribution can be attributed to orthologs, but the location of the major mode is the location of the major mode of the paralog distribution instead of the major mode of the ortholog distribution (Figure 3.5). This is true for configurations of 50/50% I1/I2 knockouts with $p_1 = 25\%$. We therefore excluded species set 7 from further consideration. In such cases, we expect that the assumed normal density fit to the major mode of the mixture distribution f in the critical region actually reflects the paralog distribution f_1; hence it is shifted to the right compared to the true $p_0 f_0$, with the critical region covering 1, as shown in Figure 3.6. In the critical region, the height of f is similar to the height of the fitted $p_0 f_0$ and is substantially greater than the height of true $p_0 f_0$. As a consequence, for z near 1, the true proportion of orthologs, computed as $p_0 f_0/f$, is less than 100% (e.g. 50% for species set 7). By contrast, the estimated proportion is 100% for z less than the upper limit of the critical region. Since the estimated proportion of orthologs is larger than the actual proportion, the cut-offs for declaring paralogs tend to be too large.

As shown in Figure 3.7 for ratio1, the ortholog distribution overlaps much more with the paralog distributions in species set 7 than in species set 8. This suggests that the outgroup for species set 7 is evolutionarily closer to the ingroups than the outgroup for species set 8. If an outgroup is evolutionarily close to the ingroups, then, for orthologs, the numerator d_{12} of the distance ratios tends to be close to the denominator d_{IO}, drawing the ortholog ratios closer to the paralog modal value (of one or more). Also, the figure shows that the paralog distribution f_{11} for species set 7 is much more concentrated than f_{11} for species set 8. It seems that the closer the outgroup is to the ingroups, the more a paralog distribution tends to be concentrated near 1. We do not expect that OL.locfdr will be applied to such species sets because

Table 3.2: Proportion of f in the critical region that comes from f_1, q

Species		Ratio1		Ratio2	
set	f_1 composition	5% paralogs	25% paralogs	5% paralogs	25% paralogs
1	100% I1 knockouts	0.001	0.909	0.002	0.019
	50/50 I1/I2 knockouts	0.001	0.017	0.003	0.019
	100% I2 knockouts	0.002	0.016	0.004	0.955
2	100% I1 knockouts	0.002	0.013	0.002	0.013
	50/50 I1/I2 knockouts	0.001	0.011	0.001	0.010
	100% I2 knockouts	0.001	0.009	0.001	0.007
3	100% I1 knockouts	0.001	0.008	0.002	0.013
	50/50 I1/I2 knockouts	0.001	0.009	0.002	0.012
	100% I2 knockouts	0.001	0.010	0.002	0.010
4	100% I1 knockouts	0.001	0.007	0.001	0.009
	50/50 I1/I2 knockouts	0.001	0.006	0.001	0.005
	100% I2 knockouts	0.001	0.005	0.000	0.000
5	100% I1 knockouts	0.001	0.011	0.001	0.011
	50/50 I1/I2 knockouts	0.001	0.008	0.001	0.007
	100% I2 knockouts	0.000	0.004	0.000	0.002
6	100% I1 knockouts	0.001	0.008	0.001	0.011
	50/50 I1/I2 knockouts	0.001	0.006	0.001	0.006
	100% I2 knockouts	0.000	0.003	0.000	0.000
7	100% I1 knockouts	0.020	0.467	0.005	0.036
	50/50 I1/I2 knockouts	0.012	0.297	0.027	0.286
	100% I2 knockouts	0.005	0.033	0.047	0.446
8	100% I1 knockouts	0.003	0.023	0.003	0.020
	50/50 I1/I2 knockouts	0.002	0.019	0.002	0.015
	100% I2 knockouts	0.002	0.015	0.002	0.012

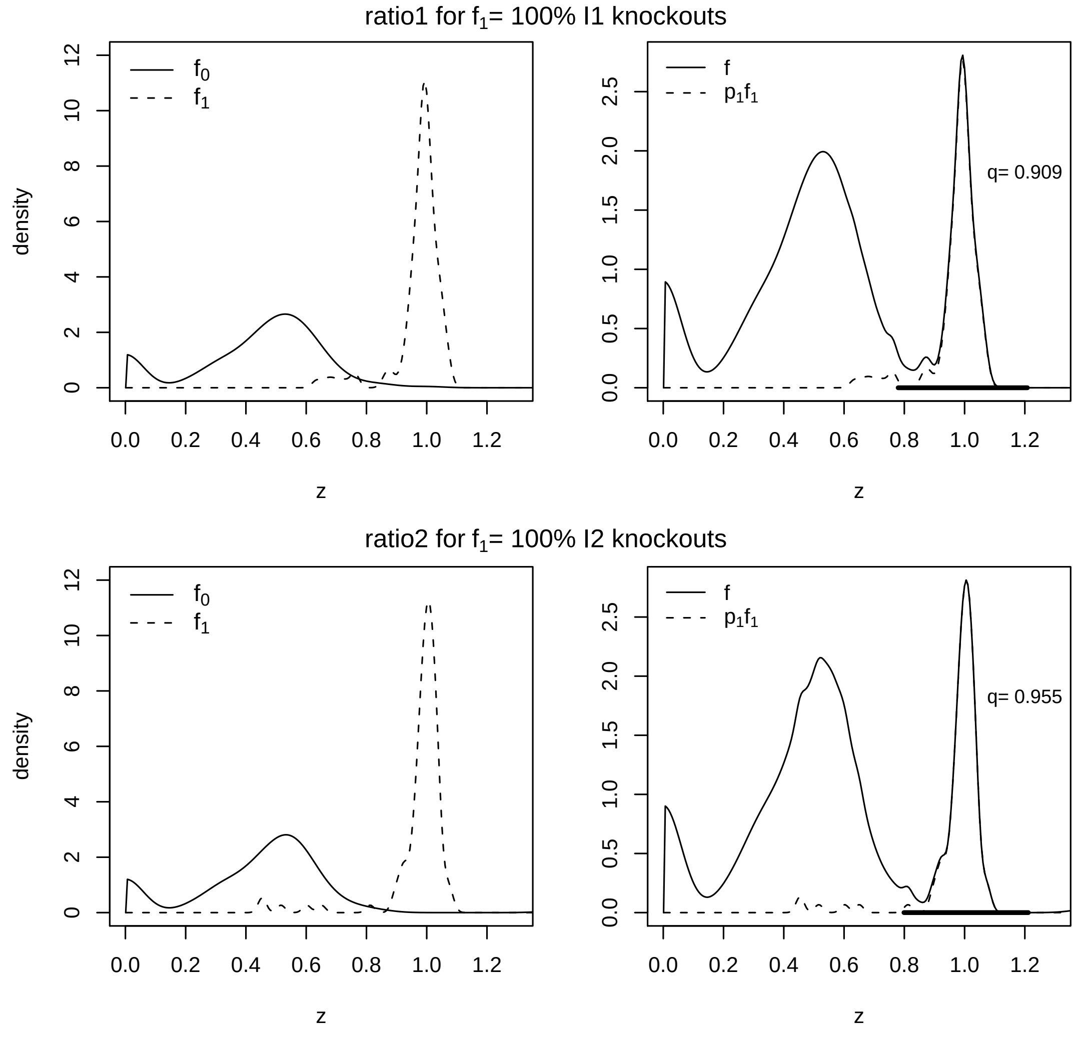

Figure 3.4: Distributions of $\sqrt{\text{ratio1}}$ with 100% I1 knockouts and $\sqrt{\text{ratio2}}$ with 100% I2 knockouts for species set 1 with $p_1 = 25\%$. The left panels show ortholog and paralog distributions, and the right panels show their mixture distributions and the relative contributions of f_1 to f. The critical region used to fit the ortholog distribution is highlighted on the horizontal axes in the right panels.

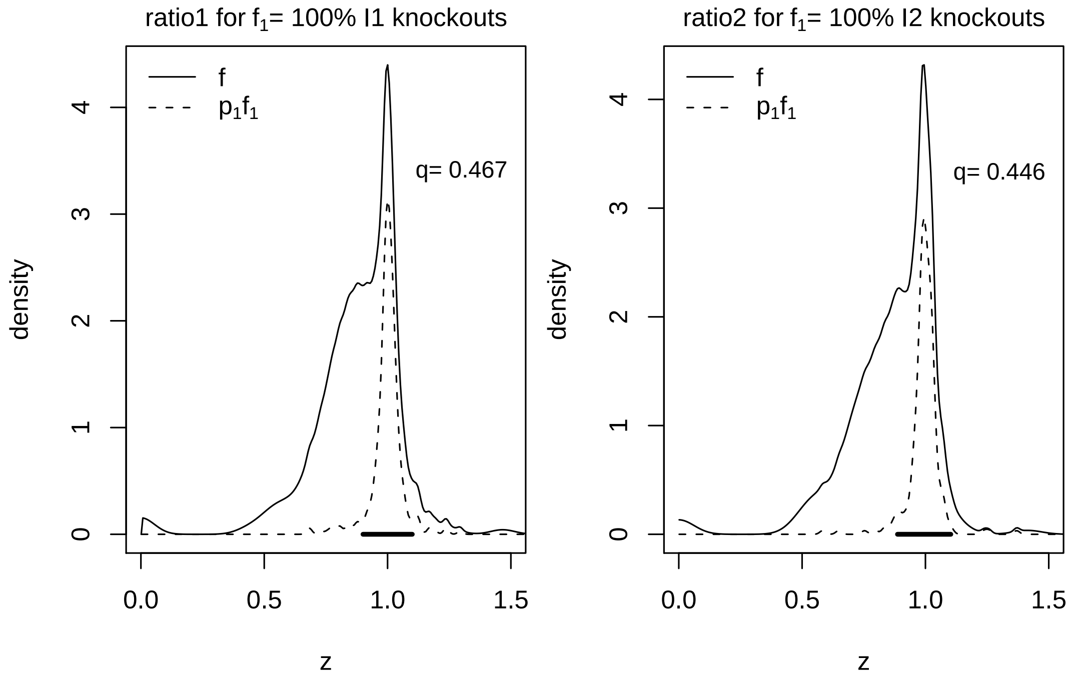

Figure 3.5: Mixture distributions f and the relative contributions of f_1 to f for species set 7 with $p_1 = 25\%$. The critical region used to fit the ortholog distribution is highlighted on the horizontal axes.

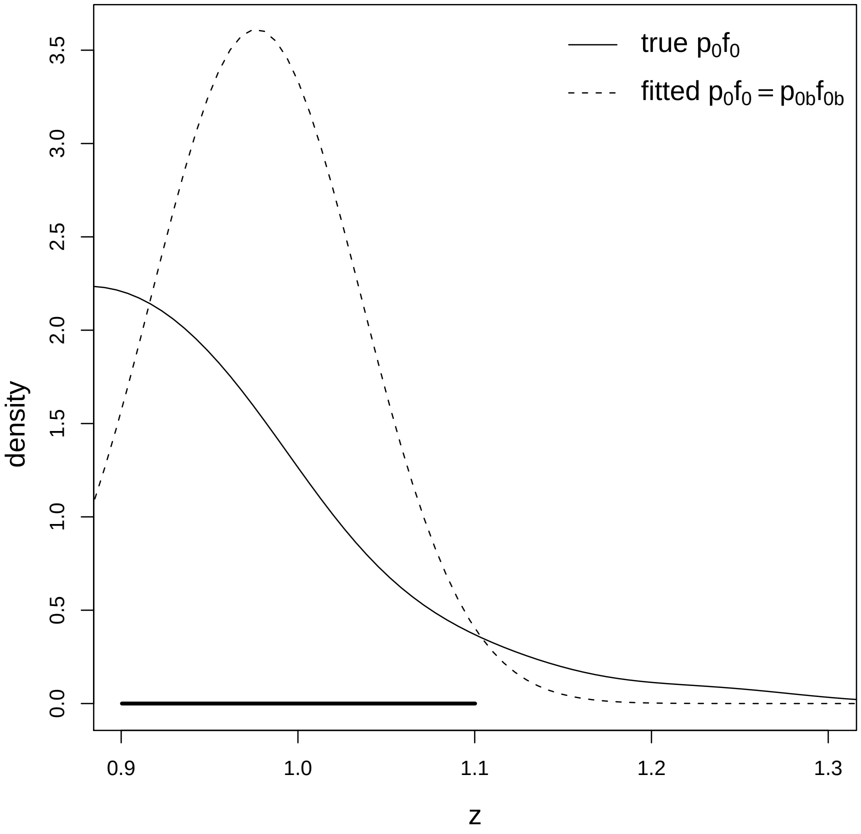

Figure 3.6: True ortholog distribution and normal curve fit to the true mixture distribution for species set 7 with ratio1 and $p_1 = 25\%$ of I1 knockouts. The critical region used to fit the ortholog distribution is highlighted on the horizontal axis.

future versions of Ortholuge will automatically select an outgroup based on species divergence that is close enough so that not too many genes are lost but far enough so that the numerator distance is sufficiently smaller than the denominator distance, and ortholog ratios are shifted to the left of 1. The proposed method to screen candidate outgroups in Ortholuge involves checking if the major mode of the mixture distribution plus the interquartile range of the square-root ratio1 or ratio2 values covers 1. If so, the outgroup is considered to be too phylogenetically close to the ingroups to be analyzed by OL.locfdr.

3.5 Species sets for the simulation study

Species sets 1 and 7 were excluded from the simulation study because, as discussed in the previous section, they lead to simulation configurations in which the major mode of the mixture distribution is due to the paralog distribution rather than to the ortholog distribution.

From the remaining species sets, we chose species sets 2, 4, 6 and 8. Species set 8 was chosen because it is the only one with ingroups *P. putida* and *P. syringae*. Species sets 2, 4 and 6 were chosen based on how well the right tail of the ortholog distribution is fit by the normal curve. OL.locfdr assumes that the right tail of the ortholog distribution is well-approximated by the normal distribution that can be fit about its major mode. To assess the normality of the right tail of the ortholog distribution, we proceeded as follows.

We adapted the method described in Step 2 of Section 2.2 to fit a normal subdensity to the true ortholog distribution, over a critical region centered at M_{f_0}, the major mode of the ortholog distribution. Specifically, for a grid of 200 points $z_1, \ldots, z_{200}$ spanning $M_{f_0} \pm \text{IQR}/2$, we fit a quadratic curve to the $(z_i, \log[f_0(z_i)])$ pairs for $i = 1, \ldots, 200$, exponentiated this quadratic, and extrapolated to the right

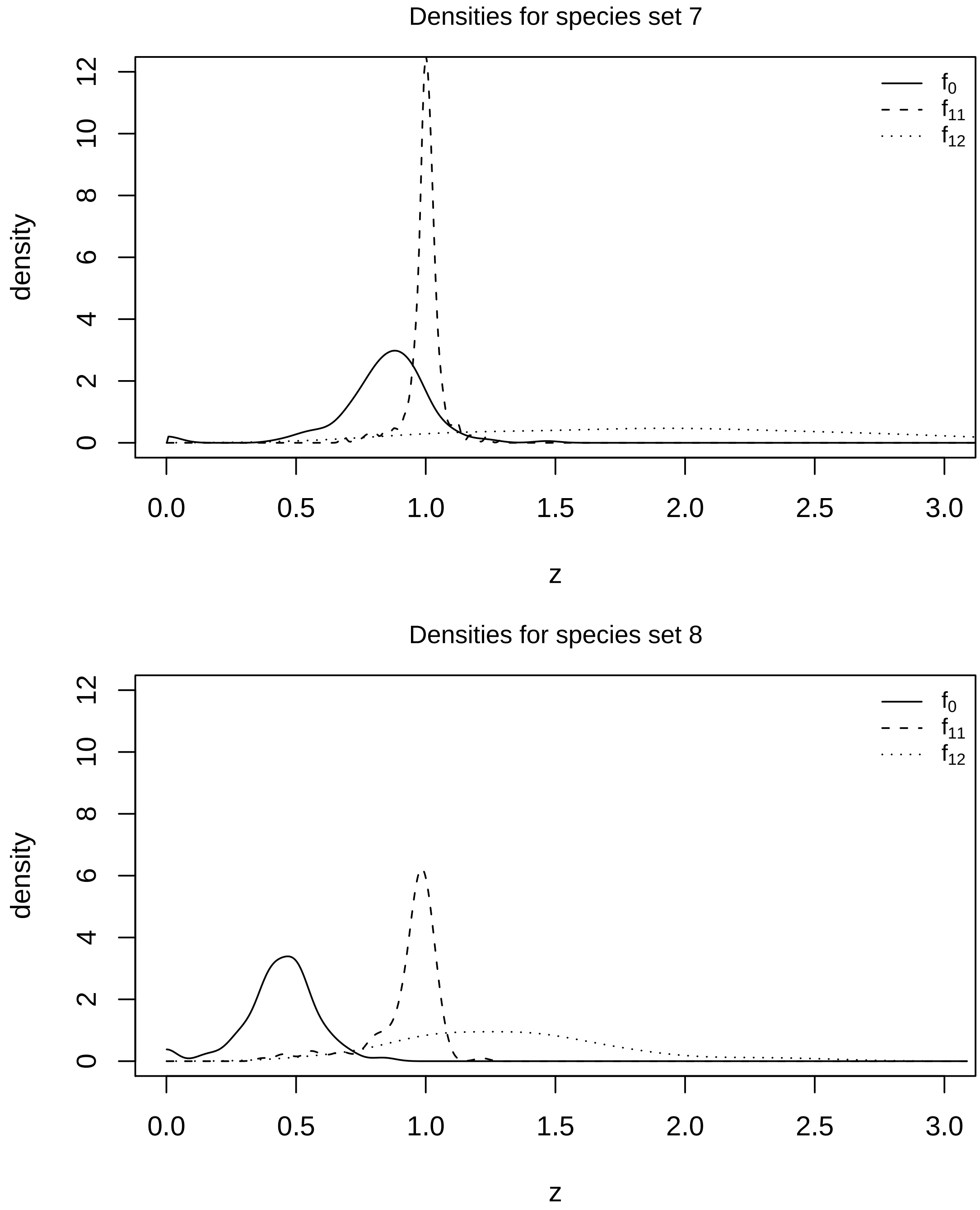

Figure 3.7: Ortholog and paralog densities for species sets 7 and 8 using ratio1.

tail of f_0. The best-fitting normal sub-densities were superposed against the ortholog densities for species sets 2 through 6 and species set 8 in Figures 3.8 (ratio 1) and 3.9 (ratio 2). For species set 2 a normal curve is well-fitted to the ortholog distribution for both ratios, so we expect the true proportion of orthologs to be similar to the estimated proportion in the right tail. Species set 4 is an example for which the right tail of the ortholog distribution is heavier than the assumed normal tail, for both ratios. In this case, we expect the true proportion of orthologs to be bigger than the estimated proportion in the right tail. By contrast, species set 6 is an example for which the right tail of the ortholog distribution is lighter than the assumed normal tail, for both ratios. In this case, we expect the true proportion of orthologs to be smaller than the estimated proportion in the right tail. Therefore, we considered species sets 2, 4, 6 and 8 for the simulation study.

3.6 Computing bias

3.6.1 Bias over a fixed set of locations

We compared the true fdr (i.e. the expected proportion of orthologs) to the expected estimated fdr over a fixed set of points in the support of the mixture distribution f. The fixed set of points was defined to be 200 equi-spaced points in the support of f. For each simulation replicate, we obtained estimated fdr's using `OL.locfdr` as described in Section 2.2, Step3. The expected estimated fdr's were estimated by these averages, taken over 5000 replicate data sets. To determine the true fdr, we took the fitted distributions and mixing proportions used to simulate data and computed the true proportion of orthologs at each of the 200 equi-spaced points, $z_1, \ldots, z_{200}$, as $p_0 f_0(z_i)/f(z_i)$, for $i = 1, \ldots, 200$.

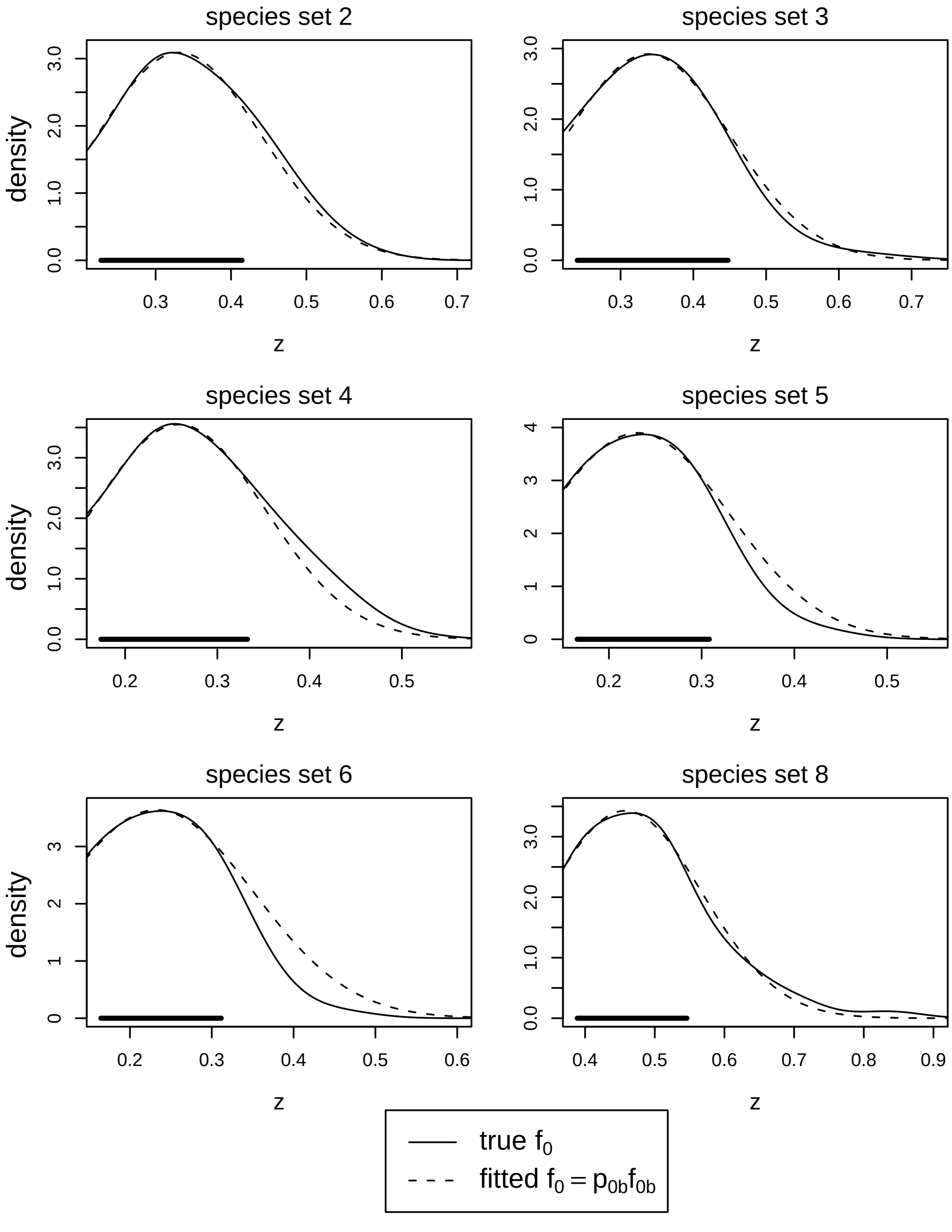

Figure 3.8: The true and fitted ortholog distributions around the major mode of the true ortholog distribution, M_{f_0}, for ratio1. The fitted normal distribution is fit to f_0. The critical region used to fit the ortholog distribution is highlighted on the horizontal axes and is $M_{f_0} \pm \text{IQR}/2$.

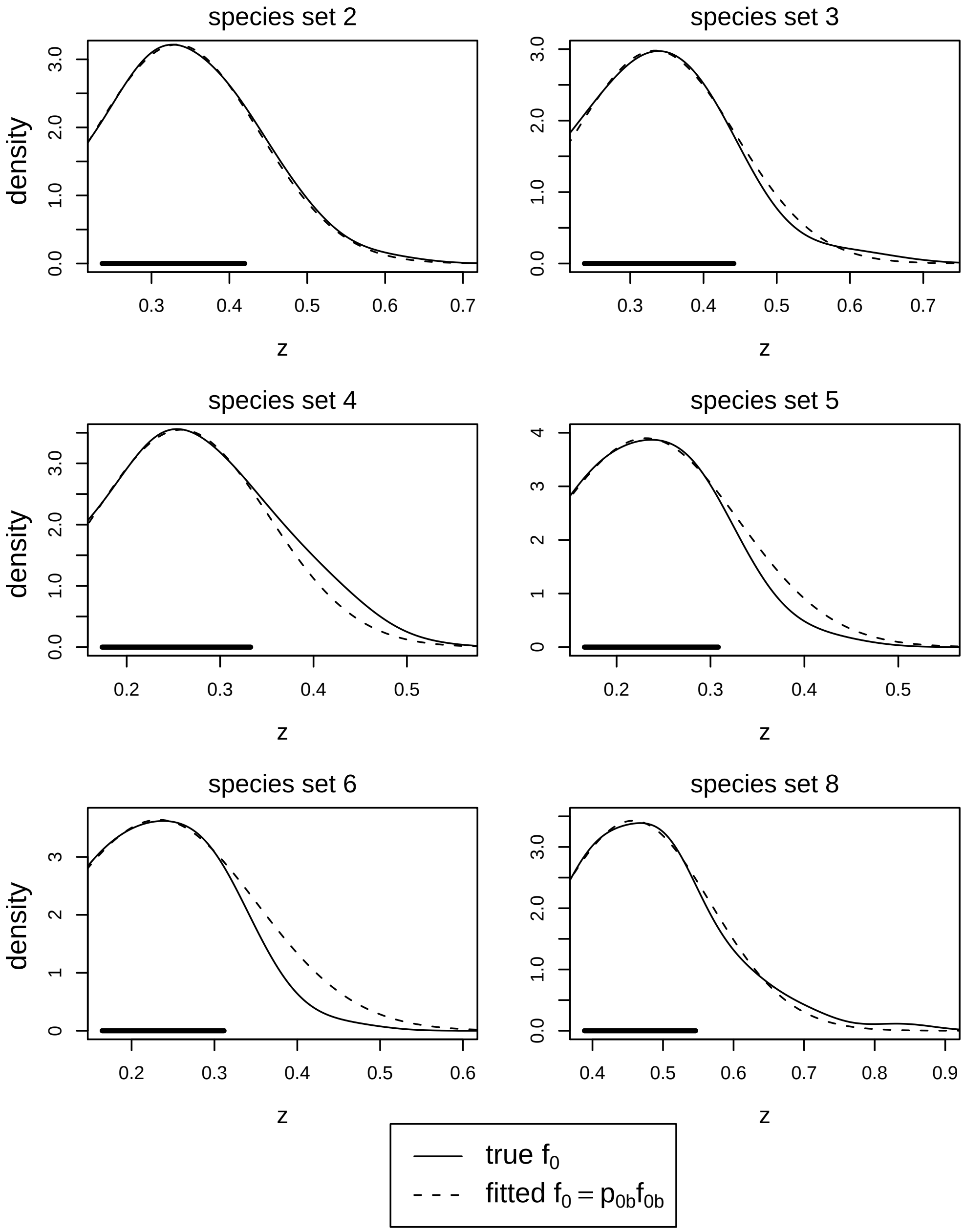

Figure 3.9: The true and fitted ortholog distributions around the major mode of the true ortholog distribution, M_{f_0}, for ratio2. The fitted normal distribution is fit to f_0. The critical region used to fit the ortholog distribution is highlighted on the horizontal axes and is $M_{f_0} \pm \mathrm{IQR}/2$.

3.6.2 Bias of the estimated 50% cut-off

We computed bias as the expected difference between the estimated 50% cut-off and the true 50% cut-off. For each simulated data set we ran the `OL.locfdr` procedure to find an estimated 50% cut-off $\hat{z}_c$. The true cut-off z_c was estimated by the smallest z-value at which true $\mathrm{fdr}(z) = 50\%$. Estimates of bias are averages of the differences between $\hat{z}_c$ and z_c over 5000 replicate data sets.

Chapter 4

Results and Discussion

4.1 Bias over a fixed set of locations

We compared the true fdr and simulation estimates of the expected estimated fdr at a fixed set of locations. Figure 4.1 illustrates this comparison for species set 4 with ratio1 as a function of z. The results for species sets 2, 4 (ratio2), 6, and 8 are shown in Appendix D (Figures D.1-D.7). The horizontal axes start at $M_f + \mathrm{IQR}/2$, where M_f is the major mode of the mixture distribution, rather than from zero, because we assume that there are 100% orthologs at values less than this starting point. For species set 4 and $p_1 = 25\%$, the second column of Figure 4.1 shows that the estimated fdr is greater than the true fdr for all the locations, indicating that OL.locfdr has a tendency to under-estimate the paralog proportions. When $p_1 = 5\%$, the estimated fdr tends to be smaller than the true fdr at lower ratio values and larger than the true fdr at higher ratio values. This is because a normal sub-distribution fit to the major mode of $\hat{f}$ will tend to have a lighter right tail than $p_0 f_0$ for lower ratio values but then, for higher ratio values, will switch over to having a heavier right tail. Lower ratio values correspond to lower proportions of paralogs. Focussing on the results for lower ratio values we see that OL.locfdr has a tendency to over-estimate lower paralog proportions. However, as the ratio values increase, so that the true paralog proportion is greater than 0.5, OL.locfdr tends to under-estimate the higher paralog proportions. In the figure, the true proportion of orthologs (solid curve) does not approach zero smoothly. The lack of smoothness is due to an abrupt drop of the true ortholog distribution near zero as shown in Figure 4.2 (right panel).

4.2 Bias of the estimated 50% cut-off

We also measured the bias by the expected difference between the estimated 50% cut-off ($\hat{z}_c$) and true 50% cut-off (z_c). Table 4.1 shows this measure of bias for species sets 2, 4, 6 and 8 at each combination of the ratios, proportions of paralogs and compositions of f_1. For example, for species set 4, the biases are all positive when $p_1 = 25\%$. As mentioned in the previous section, when $p_1 = 25\%$, the estimated fdr tends to be greater than the true fdr for all z values. Hence any estimated cut-offs for declaring paralogs tend to be greater than the true cut-offs, including the 50% cut-off. By contrast, the biases for simulated data sets containing $p_1 = 5\%$ paralogs are negative indicating that $\hat{z}_c$ tends to be to the left of z_c. These results for $p_1 = 5\%$ can not be predicted from Figure 4.1, as we next discuss.

From Figure 4.1 and Figures D.1-D.7, some idea of the bias can be obtained from the difference in z values at which the dotted line, indicating a proportion of orthologs of 0.5, crosses the curves of the estimated and true fdr's. For example, for species set 4 when $p_1 = 25\%$ (Figure 4.1, top right panel), the true fdr is 0.5 at $z_c \approx 0.53$ and the mean estimated fdr is 0.5 at $z \approx 0.57$. The positive bias of 0.044 in Table 4.1 for species set 4 with 100% I1 knockouts, ratio1 and $p_1 = 25\%$ can be approximated by the width of the dotted line between the two curves in the top right panel of Figure 4.1, as $0.57 - 0.53 = 0.04 > 0$. This is only a rough guide, however, because

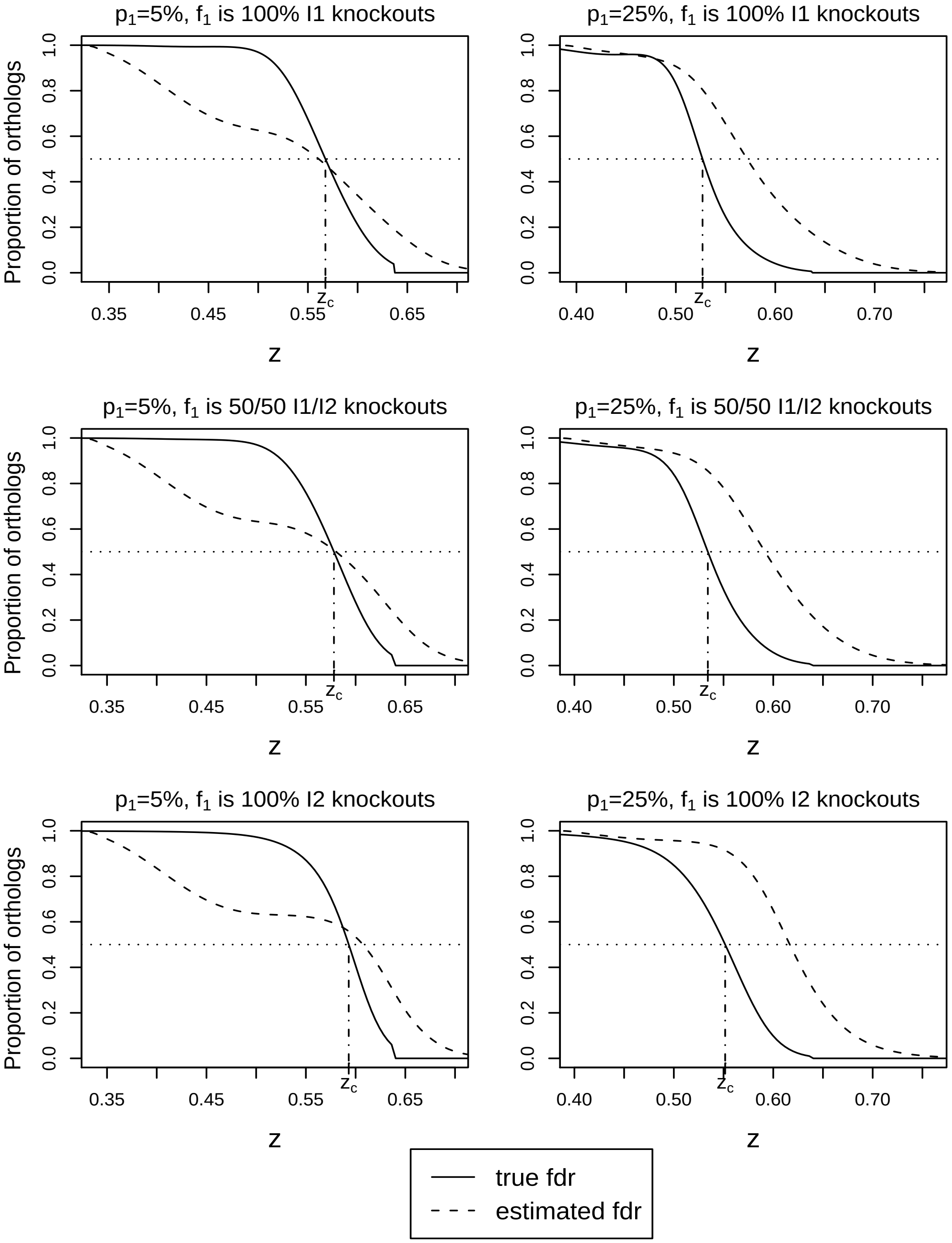

Figure 4.1: Proportion of orthologs as a function of z for ratio1 in species set 4. The horizontal axes start at $M_f + \text{IQR}/2$. The ortholog proportion of 0.5 is indicated in a dotted horizontal line with the true 50% cut-off z_c.

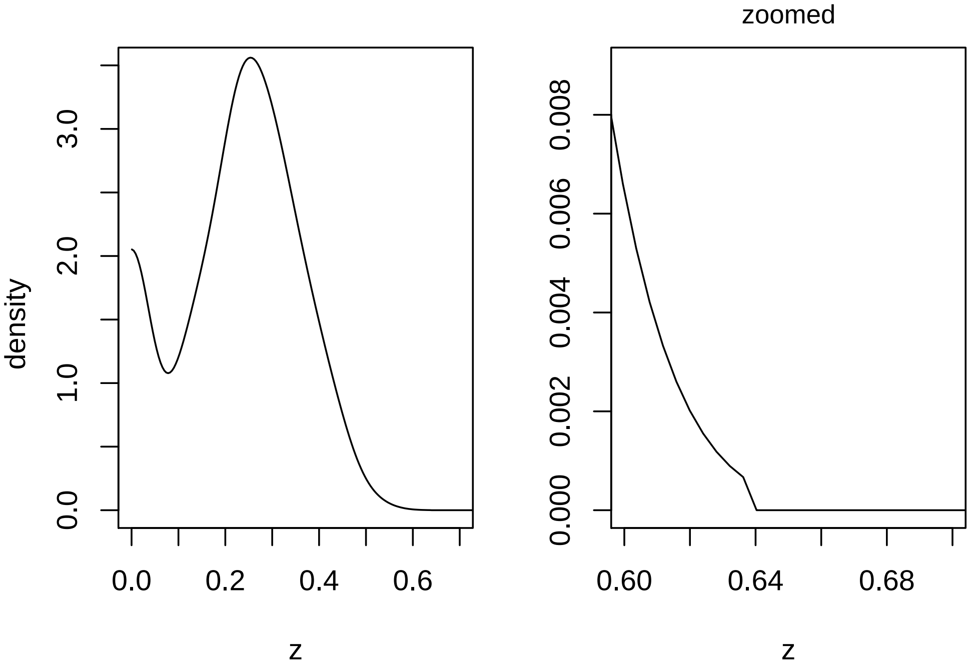

Figure 4.2: True ortholog distribution for species set 4 with ratio1. The right panel is a zoomed-in version of the left panel.

the z value at which the mean estimated fdr is 0.5 is not necessarily equal to the mean of the estimated cut-off $\hat{z}_c$. As an example of when this rough guide fails, consider species set 4, ratio1 and $p_1 = 5\%$. Even though the left panels of Figure 4.1 suggest biases that are positive or near zero, the actual biases of the estimated cut-offs are negative.

One question of interest was how the performance of the method is affected by contamination around the major mode of f. Such contamination should influence the estimation of $p_0 f_0$. We can regard the overlap measure q as an indication of the amount of such contamination and the bias of $\hat{z}_c$ as an indication of the performance. The plot of the bias of $\hat{z}_c$ versus q is shown in Figure 4.3 and has no obvious trends. This is not unexpected because, for all the species sets considered in the simulation study, the contamination of the critical region is minimal. This minimal contamination does not appear to have much, if any, effect on the performance of the method.

Another question of interest was how performance depends on the fit of $\widehat{p_{0b} f_{0b}}$ to the right tail of $p_{0b} f_{0b}$. Since $p_{0b} f_{0b}$ appears in the numerator of the fdr in equation (2.2), the quality of the estimated fdr's depends directly on the quality of $\widehat{p_{0b} f_{0b}}$. OL.locfdr estimates $p_{0b} f_{0b}$ by fitting a normal sub-distribution to the major mode of $\hat{f}$. Thus, some idea of the information for estimating $p_{0b} f_{0b}$ can be obtained by fitting a normal sub-density $p_{0b}\phi_{0b}$ to the major mode of f: conceptually,

$$\widehat{p_{0b} f_{0b}} \stackrel{\text{estimates}}{\longrightarrow} p_{0b}\phi_{0b} \stackrel{\text{approximates}}{\longrightarrow} p_{0b} f_{0b}.$$

The sub-density $p_{0b}\phi_{0b}$ is obtained by a method similar to Step2 in Section 2.2. Briefly, consider a grid $z_l \ldots z_u$ of 200 equi-spaced points spanning $M_f \pm \text{IQR}/2$, where M_f is the major mode of f. To obtain $\log[p_{0b}\phi_{0b}]$, we used the fact that $\log[p_{0b}\phi_{0b}(z)]$ is quadratic in z when ϕ_{0b} is a normal density, and $f(z) \approx p_{0b} f_{0b}(z)$ or $\log[f(z)] \approx \log[p_{0b} f_{0b}(z)]$ for z near the major mode of f. Thus, we obtained $\log[p_{0b}\phi_{0b}]$ by fitting

Table 4.1: Bias* (% bias**) for species sets 2, 4, 6 and 8

		Ratio1		Ratio2	
Species set	f_1 composition	5% paralogs	25% paralogs	5% paralogs	25% paralogs
2	100% I1 knockouts	0.033 (4.9)	0.040 (6.2)	-0.031 (-4.5)	0.017 (2.7)
	50/50 I1/I2 knockouts	0.037 (5.5)	0.052 (8.2)	-0.034 (-4.8)	0.014 (2.1)
	100% I2 knockouts	0.043 (6.4)	0.065 (10.4)	-0.035 (-4.9)	-0.009 (-1.3)
4	100% I1 knockouts	-0.027 (-4.7)	0.044 (8.4)	-0.060 (-10.3)	0.028 (5.1)
	50/50 I1/I2 knockouts	-0.024 (-4.2)	0.056 (10.5)	-0.069 (-11.5)	0.034 (6.0)
	100% I2 knockouts	-0.026 (-4.4)	0.062 (11.3)	-0.089 (-14.3)	0.005 (0.8)
6	100% I1 knockouts	0.067 (12.4)	0.050 (10.3)	0.084 (15.9)	0.062 (12.9)
	50/50 I1/I2 knockouts	0.071 (13.0)	0.050 (10.0)	0.081 (15.1)	0.053 (10.7)
	100% I2 knockouts	0.077 (14.0)	0.053 (10.4)	0.077 (14.3)	0.046 (8.9)
8	100% I1 knockouts	-0.150 (-17.1)	-0.032 (-4.2)	-0.148 (-16.8)	0.005 (0.6)
	50/50 I1/I2 knockouts	-0.160 (-18.0)	-0.031 (-4.2)	-0.144 (-16.5)	0.008 (1.1)
	100% I2 knockouts	-0.178 (-19.7)	-0.032 (-4.2)	-0.140 (-16.2)	0.011 (1.5)

* Bias is defined as the expected difference between the estimated and true 50% cut-off.

** % bias is defined as the bias over the true cut-off in percentage.

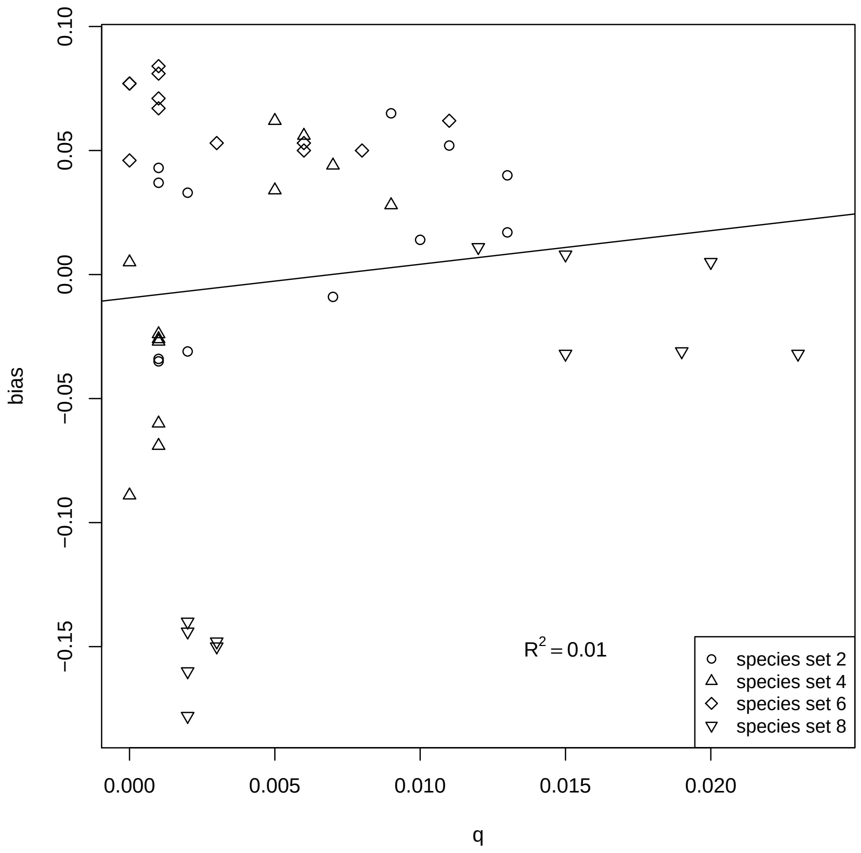

Figure 4.3: Bias versus q for species sets 2, 4, 6, and 8.

a quadratic function to $\log f$ near the major mode of f. Then the resulting $\log(p_{0b}\phi_{0b})$ was exponentiated to give the sub-density $p_{0b}\phi_{0b}$. We then compared $p_{0b}\phi_{0b}$ to the true $p_0 f_0(z)$ for $z > l$; this comparison was performed for various f_1 and p_1 combinations described in Section 3.2.

For species set 4, $p_{0b}f_{0b}$ is compared with the assumed normal sub-distribution, as illustrated in Figure 4.4 for ratio1 when f_1 contains 100% I1 knockouts. The other compositions of f_1 give similar results (not shown); Appendix E has the results for other species sets. When $p_1 = 5\%$, the zoomed-in plots in the bottom panel of the Figure 4.4 show that the assumed normal sub-distribution, $p_{0b}\phi_{0b}$, is lighter-tailed than the right tail of the true $p_{0b}f_{0b}$. A consequence of having the estimated ortholog distribution fall below the true ortholog distribution is that the estimated proportion of orthologs (i.e. estimated fdr's) will be lower than the true proportion of orthologs (i.e. true fdr's), so that the estimated cut-off will be lower than it should be. Since we define the bias as the difference between the mean estimated 50% cut-off ($\bar{\hat{z}}_c$) and true 50% cut-off (z_c), a light tail of $p_{0b}\phi_{0b}$ relative to $p_{0b}f_{0b}$ leads to negative bias for this species set. By contrast, when $p_1 = 25\%$, $p_{0b}\phi_{0b}$ is heavier-tailed than the right tail of $p_{0b}f_{0b}$, leading to positive bias.

In general, if the estimate $p_{0b}\phi_{0b}$, obtained by direct fitting of the true mixture density f, is lighter-tailed than the right tail of the true $p_{0b}f_{0b}$, we would expect that $\hat{z}_c$ would tend to be lower than it should be. By contrast, if $p_{0b}\phi_{0b}$ is heavier-tailed than the right tail of $p_{0b}f_{0b}$, we would expect that $\hat{z}_c$ would tend to be higher than z_c. However, for a few cases it does not hold. For example, for species set 2 with ratio2 and $p_1 = 25\%$ (Figure E.2, right panels), $p_{0b}\phi_{0b}$ is lighter-tailed than the right tail of the true $p_{0b}f_{0b}$ around z_c, but the $\bar{\hat{z}}_c$ is greater than the true z_c. The same is true for species set 8 with ratio2 and $p_1 = 25\%$ shown in Figure E.7, right panels. Such cases may be caused by the low probability mass in the right tail of f and the small number of genes falling in the right tail for a gene list of length 5000. We compare

$p_{0b}f_{0b}$ with $p_{0b}\phi_{0b}$ fit to the true f. Conceptually, the true f is obtained from a gene list of infinite length. However, when estimating the fdr for a simulated data set, we fit $p_{0b}f_{0b}$ to $\hat{f}$ obtained from a list of 5000 genes sampled from the true f. Therefore, as the number of sampled genes increases, the behavior of the bias should conform better to how well the right tail of the ortholog distribution fits the right tail of the assumed normal distribution. Additional simulation results for gene lists of length 200,000 seem to bear this out (results not shown).

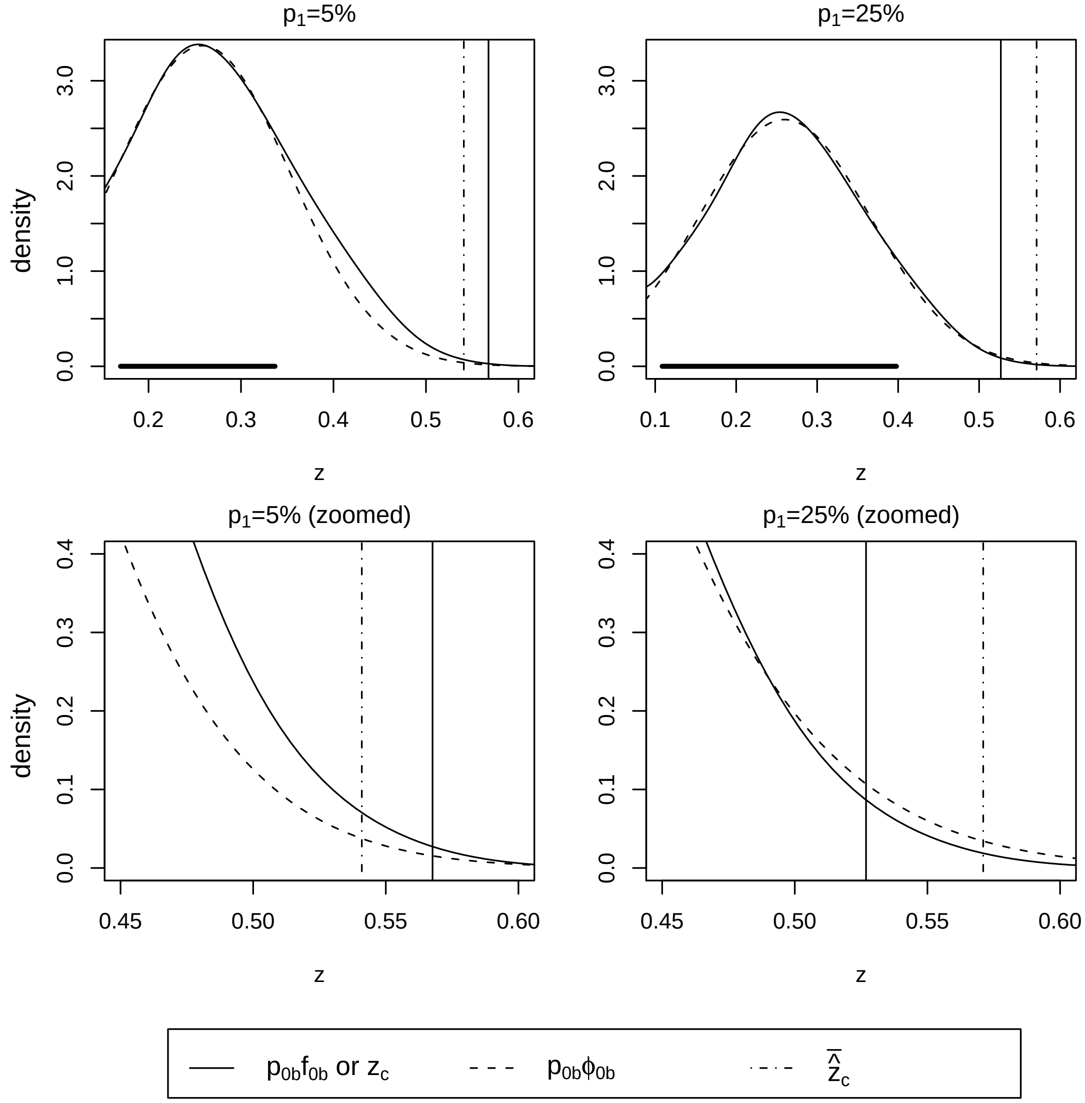

Figure 4.4: The true ortholog sub-distribution (solid curve) for ratio1 in species set 4 when f_1 contains 100% I1 knockouts. The fitted ortholog sub-distribution (dashed curve) is obtained from the mixture density f, evaluated over a critical region defined by $M_f \pm \text{IQR}/2$; this critical region is highlighted on the horizontal axes in the top panels. The true 50% cut-off, z_c, is indicated by a solid vertical line. The average of the estimated 50% cut-offs, $\bar{\hat{z}}_c$ over the simulation replicates is indicated by a vertical dot-dashed line. Results are shown only for the critical region and right tail of the distributions. The bottom panels are zoomed-in versions of the top panels.

Chapter 5

Conclusions

Ortholuge identifies RBH-predicted orthologs with atypical genetic divergence through phylogenetic distance ratios. The former Ortholuge method for declaring genes with atypical divergence is computationally-expensive because it relies on transformed true-negative gene lists that require alignment and calculation of phylogenetic distances. To address this problem, we propose an alternate statistical procedure, OL.locfdr, that does not require simulation and alignment of true-negative genes, and calculation of their phylogenetic distances. Instead, our approach distinguishes orthologs from paralogs directly from the data, thereby saving computational time.

Recently, a coding error was discovered in the implementation of the Ortholuge iterative true-negative analysis for determining cut-offs. Second-best BLAST hits rather than the intended RBH were being introduced as true negatives. After correcting this error, the number of genes that could be transformed into true negatives was drastically reduced (M. Whiteside, personal communication). The algorithm has therefore been modified to introduce all possible true-negatives in a single step, and the computational time has been cut significantly. This new non-iterative version of the algorithm has a similar computational time to our approach, but appears to be less reliable. For example, in ongoing investigations of similar ingroup pairs (e.g. pair I1-I2a and pair I1-I2b, for closely related I2a and I2b), our approach produces cut-off values that are consistent from one pair to the next, as expected by the investigators. By contrast, Ortholuge can produce discrepant cut-off values for some pairs.

We determine the cut-off for declaring paralogs by modifying the local false-discovery rate (fdr) method of Efron (2004) to estimate the proportion of paralogs in the mixture distribution at a particular ratio value. By simulation, we assess and understand the performance of our proposed method using realistic data. We compute the overlap measure, q, between the ortholog and paralog distributions in the critical region used to estimate the ortholog distribution, and see how q affects the performance of the method. We measure the performance by bias or the expected difference between the estimated and true proportion of orthologs and between the estimated and true 50% cut-off.

OL.locfdr makes several assumptions. One assumption is that the major mode of the mixture distribution of square-root ratios is away from zero. The assumption is reasonable since, for highly similar ingroup species with a major mode is at zero, an investigator will determine orthologs using other approaches such as whole genome alignment. Another assumption is that the paralog distribution has very little mass in the critical region used to fit the ortholog distribution, near the major mode of the mixture distribution. Departures from this assumption could lead to poor estimation of the ortholog distribution. For species sets we considered (2, 4, 6, and 8), the departures are negligible, however, as indicated by very small values of the overlap measure in Table 3.2. As shown in Figure 4.3, for these low levels of overlap, there is no obvious effect of increasing overlap on the performance of OL.locfdr. Finally, we assume that the main ortholog sub-distribution, f_{0b}, is normal. As shown in Figure 3.8 and 3.9, the overall bell-shaped ortholog distributions, compared to the normal distributions fit to f_0, indicate that the normality assumption is a reasonable approximation for the species sets considered. However, their right tails can be heavier- or lighter-than-normal. This non-normal distortion of the true f_0 in the right tail contributes to the bias of the procedure, computed, for example, as the expected difference between the estimated and true 50% cut-off (Table 4.1).

In general, OL.locfdr performance depends on the fit of the assumed normal distribution to the true ortholog distribution. In comparing the ortholog distribution fit to the data to its real counterpart, the paralog probabilities of interest determine where to focus. For example, for higher paralog probabilities, such as those used to classify probable paralogs, the focus would be on the tail of the true distribution. Specifically, if the fitted normal distribution has less probability mass than the true ortholog distribution in the tails, OL.locfdr is conservative in the sense that it will over-estimate the proportion of paralogs. For lower paralog probabilities, such as those used to classify orthology uncertain, the focus would be more towards the centre of the true ortholog distribution. Specifically, if the fitted normal distribution has less probability mass than the true ortholog distribution in this more central region, OL.locfdr will again over-estimate the proportion of paralogs.

We wish to point out some issues regarding the quality of the true-ortholog data sets used to obtain the true f_0. First, these data sets are small (Table 3.1), and so there is a large amount of sampling variability. Second, and perhaps more importantly, the data sets mostly consist of genes that are easily determined to be true orthologs, such as those with square-root ratios less than 1. Thus, there is potential for selection bias such that large ratios may be under-represented.

Fulton et al. (2006) classified the genes lying at or above the 50% cut-off as probable paralogs. If our estimated cut-off is higher than the true cut-off, we would misclassify some paralogs as orthologs. By contrast, if the estimated cut-off is lower than the true cut-off, we would misclassify some ortholog as paralogs. Future simulation studies could be undertaken in which the performance of OL.locfdr is evaluated based on the numbers of misclassified paralogs and orthologs in a data set.

For the species sets we have examined, the minimal overlap of the paralog distribution into the critical region used to fit the ortholog distribution does not affect the performance of OL.locfdr. Another direction for future research would be to construct artificial species sets with increased and systematically varying contamination of the critical region. These artificial species sets would enable a more satisfactory investigation of how the performance of the method is influenced by contamination in the critical region.

For all the species sets we have examined, the square-root transformation of the raw ratio data yields mixture distributions with major modes that can be approximated by a normal distribution. We do not rule out that, for other species sets, other ratio transformations besides square-root might be more appropriate. Assuming that the data come from a Box-Cox or power-normal family of distributions (Freeman and Modarres, 2006), an alternative to a transformation chosen *a priori* is to use data about the major mode to select a normalizing transformation. For example, one could fit by ordinary least squares a quadratic curve to the logarithm of a kernel density estimate based on ratios raised to the power of λ, for those ratios falling in some critical region. This fitting could be repeated for each power λ in a list of candidates (e.g. 1/2, 1/3, 1/4). Then one could select the λ that gives the "best fitting" quadratic, where fit could be judged by the residual sum of squares. Potential problems with such an approach include the possibility that ratio data from orthologs may not be adequately approximated by a normal for any power, and questions about the validity of using the data to select the transformation and then treating it as known for subsequent inference (see for example Chen et al., 2002 for a discussion of this issue in the context of linear models). For all the data sets that we investigated, a square-root transformation appeared to be reasonable.

Appendix A

BLAST Overview

This overview of BLAST is based on the descriptions by Wheeler and Bhagwat (2007). BLAST (basic local alignment search tool) is an algorithm to compare a query nucleotide or protein sequence with a database of sequences in order to identify sequences in the database that are similar to the query (Altschul et al., 1990). Comparisons are done by forming "local" pairwise alignments; i.e., by aligning just those segments of the sequences that match well. Local alignment is useful because many query sequences have domains, active sites, or other motifs that have local but not global regions of similarity to other proteins. Also, databases typically have fragments of DNA and protein sequences that can be locally aligned to a query. Because sequence is an important factor determining functional information, BLAST can be used to predict gene function and some general features about gene evolution. The program is available on the web on the National Center for Biotechnology Information website (http://ncbi.nlm.nih.gov/BLAST).

In BLAST, each comparison of the local pairwise alignments is given a score reflecting the degree of similarity between a given sequence in the database and a query. The higher the score, the greater the degree of similarity. The score is computed by assigning a value to each aligned pair of letters and then summing these values over the length of the alignment. For protein alignment, the score of the match is determined using a "substitution matrix" M containing values $M[i, j]$ reflecting the probability that amino acid i mutates into amino acid j for all pairs of amino acids. Likely substitutions have positive values and unlikely substitutions have negative values. For nucleotide alignment, an exact match has a score of $+2$ and a non-exact match has a score of -3. BLAST adds a negative penalty for having a gap in an alignment. Extension of the gap has a lesser penalty than introduction of a gap.

There are three algorithmic steps in BLAST (Figure A.1). Initially, the query sequence is segmented into "words" with a fixed length W, which is chosen by the user. Typically, W is a default value of 3 for protein and 11 for DNA. The second step is scanning a sequence in the database that contains a word of length W that can match with the query sequence word. For nucleotide-to-nucleotide searches, the match must be exact; for protein-to-protein searches, the match must achieve a score of at least T, according to the protein substitution matrix. Lastly, when a word match is found, BLAST extends both forward and backward from the match to produce a high-scoring segment pair or HSP. The extension continues until the accumulated total score of the HSP drops by a certain amount. Scores are calculated from scoring matrices along with gap penalties. The HSPs with scores above S, specified by the user, are reported.

The threshold parameter T dictates speed, specificity, and sensitivity of the search. When T is increased, the speed and specificity of the search is increased, but fewer matches are generated, and so distantly related database matches may be missed. When T is decreased, the search proceeds slowly, but many more word matches are evaluated, and thus sensitivity is increased.

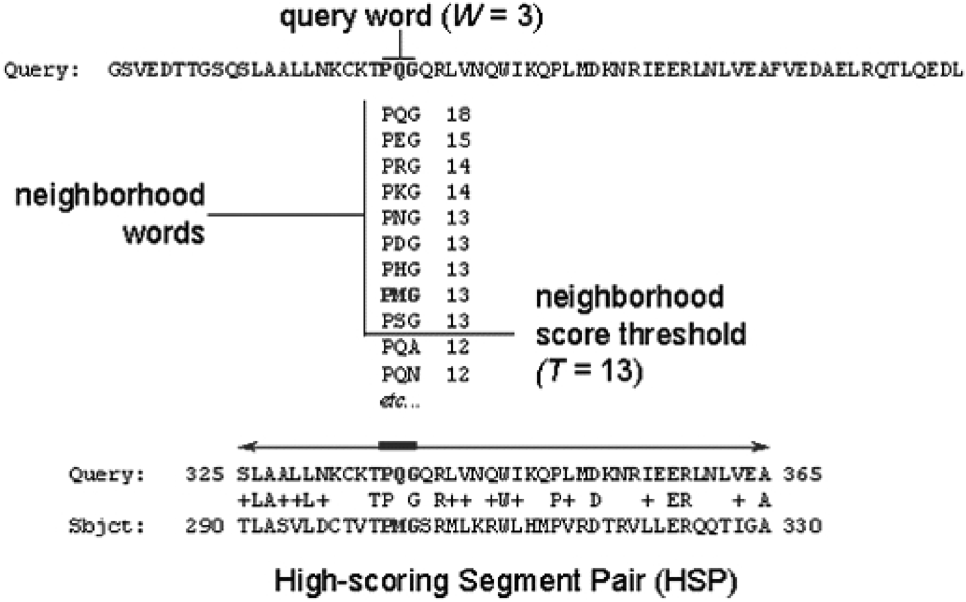
query word (W = 3)
Query: GSVEDTTGSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVEDAELRQTLQEDL
PQG 18
PEG 15
PRG 14
PKG 14
PNG 13
PDG 13
PHG 13
PMG 13
PSG 13
PQA 12
PQN 12
etc...
neighborhood words
neighborhood score threshold (T = 13)
Query: 325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
+LA++L+ TP G R++ +W+ P+ D + ER + A
Sbjct: 290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330
High-scoring Segment Pair (HSP)

Figure A.1: Schematic of "The BLAST Search Algorithm," taken from National Center for Biotechnology Information website (http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/BLAST_algorithm.html). In the first step a query sequence is analyzed with a given word size (e.g., $W = 3$), and a list of words matching a query word (e.g., PQG) is compiled that attains a threshold score (e.g., $T = 13$). In this example, eleven words are shown along with their scores; nine of these are equal to or greater than T (i.e., neighborhood words), and two are below. In the second step, a database is searched to find entries that match the compiled word list. In the third step, the match (e.g., PMG) is extended in both directions to obtain a HSP. The sequence alignment shown here has three rows. The first and third rows are the query and the database sequences, respectively. Numbers at the ends of the alignment denote the position of the first and last amino acid in a line within the respective sequence. The second row shows the degree of similarities between these two sequences. For protein, identical residues are indicated by the capital letter of the amino acid. Similar, nonidentical residues with positive alignment scores are indicated with plus signs. Alignments with a zero or negative scores are indicated with a space.

Appendix B

Proportion of paralogs in the transformed mixture

Fulton et al. (2006) describe a method for classifying orthology based on transforming the original mixture distribution $f(z)$ of distance ratio z. To set the classification cut-off value for the distance ratios they use the proportion of transformed ratios that come from a known true-negative distribution $f_T(z)$ as a proxy for the proportion of paralogs in $f(z)$. In this appendix, we show that this approximation may lead to 50% cut-off values that are too high when the mixing proportion of paralogs, p_1, is greater than 1/6. Let $f_0(z)$ and $f_1(z)$ denote the distribution of ortholog and paralog ratios, respectively, and let p_0 and $p_1 = 1 - p_0$ denote the corresponding relative frequencies of orthologs and paralogs in the mixture. The mixture density f is

$$f(z) = p_0 f_0(z) + p_1 f_1(z).$$

The proportion of orthologs and paralogs in f at a given ratio value z are

$$\Pr(\text{ortholog} \mid z) = \frac{p_0 f_0(z)}{f(z)} \quad \text{and} \quad \Pr(\text{paralog} \mid z) = \frac{p_1 f_1(z)}{f(z)},$$

respectively. If the true-negative ratios can be regarded as a random sample from the distribution f_1 of paralogs, then transforming a proportion p_T (e.g. $p_T = 0.25$ as in Fulton et al.) of the original data to true-negatives will result in a sample from the "transformed mixture" density

$$f_T(z) = p_T f_1(z) + (1 - p_T) f(z). \tag{B.1}$$

The proportion of true-negatives in the transformed mixture f_T at a given ratio value z is then

$$\Pr(\text{true-negative} \mid z) = \frac{p_T f_1(z)}{f_T(z)}.$$

The orthology classifications of Fulton et al. are based on the z-values at which the proportion of true-negatives in f_T are 0.10 and 0.50. One can view their cut-offs as proxies for cut-offs based on the proportions $p_1 f_1(z)/f(z)$ of paralogs in f. Their proportion $p_T f_1(x)/f_T(z)$ may be used as a proxy because $p_1 f_1(z)/f(z)$ is a monotone increasing function of $p_T f_1(z)/f_T(z)$:

$$\begin{aligned}
\frac{p_1 f_1(z)}{f(z)} &= p_1 \frac{1-p_T}{p_T} \frac{p_T f_1(z)}{(1-p_T) f(z)} \\
&\overset{(B.1)}{=} p_1 \frac{1-p_T}{p_T} \frac{p_T f_1(z)}{f_T(z) - p_T f_1(z)} \\
&= p_1 \frac{1-p_T}{p_T} \frac{p_T \frac{f_1(z)}{f_T(z)}}{1 - p_T \frac{f_1(z)}{f_T(z)}} \\
&= b\, \frac{s(z)}{1-s(z)},
\end{aligned} \tag{B.2}$$

where $b = p_1(1-p_T)/p_T$ and $s(z) = p_T f_1(z)/f_T(z)$. Since $bs/(1-s)$ is monotone increasing in s for s between 0 and 1, $p_1 f_1(z)/f(z)$ is monotone increasing in $p_T f_1(z)/f_T(z)$. Consequently, large values of $p_T f_1(z)/f_T(z)$ will imply large values of $p_1 f_1(z)/f(z)$ and we may reasonably classify as possible paralogs genes whose ratio values coincide with substantial proportions $p_T f_1(z)/f_T(z)$ of true-negatives in f_T. However, without the unknown relative frequency p_1 of paralogs in f, we cannot translate the proportion of true-negatives in f_T into the desired proportion of paralogs in f. Figure B.1 depicts the relationship between the proportion of paralogs in the original mixture and the proportion of true-negatives in the transformed mixture for $p_T = 0.25$, as in Fulton et al., and for $p_1 = 0.05$, 0.15 and 0.25, spanning the range of plausible values of p_1 (M. Whiteside, personal communication). The proportion of paralogs in the original mixture varies substantially by p_1 for the 0.5 cut-off used by Fulton et al. to separate “orthology uncertain” from “probable paralog”. For example, the proportion of paralogs can be less than 0.5 at the cut-off ($p_1 = 0.15$ and 0.05). However, the proportion can also be greater than 0.5 (e.g. $p_1 = 0.25$), in which case paralogs will be missed; in fact, by solving equation (B.2) for p_1, we find that paralogs will tend to be missed at the 0.5 cut-off for any $p_1 > 1/6$. In contrast, the proportion of paralogs in the original mixture is always less than 0.1 for the 0.1 cut-off used by Fulton et al. to separate “probable ortholog” from “orthology uncertain”.

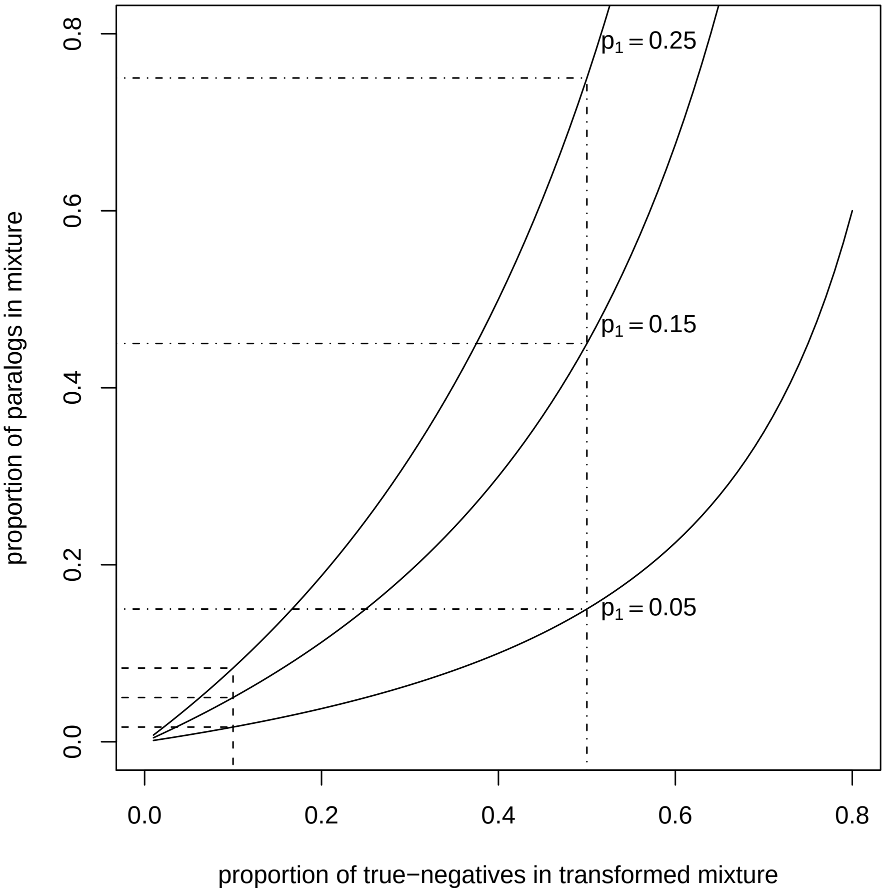

Figure B.1: Proportion of paralogs in the mixture at a given ratio value as a function of the proportion of true-negatives in the transformed mixture at the same ratio value. The relationship depends on the proportion of transformed genes p_T and the relative frequency p_1 of paralogs in the mixture. Throughout, $p_T = 0.25$ is used, as in Fulton et al. Curves for $p_1 = 0.05$, 0.15 and 0.25 are shown. The surrogate values 0.1 and 0.5 used by Fulton et al. to classify orthology are indicated on the horizontal axis, and the corresponding values of the proportion of paralogs in the mixture are indicated by horizontal dashed lines (true-negative-based cut-off = 0.1) or dot-dashed lines (true-negative-based cut-off = 0.5) for different values of p_1.

Appendix C

Sample call to `OL.locfdr()` with outputs

The steps outlined in Section 2.2 were implemented in an R function `OL.locfdr()` with the following arguments:

1. `dat`: a data frame with gene identifiers (ingroup1, ingroup2, outgroup) and pair of distances (d_{12}, d_{IO}) for the particular ratio of interest. For example, if ratio1 is of interest, $d_{\text{IO}} = d_{1\text{O}}$.

2. `target.prob`: a vector of probability thresholds for calling a gene pair unusually diverged ($= 1-$ fdr).

3. `small.dist`: a threshold for small d_{IO}. Gene pairs determined to be unusually diverged that are based on a $d_{\text{IO}} <$ `small.dist` will be flagged as potentially unreliable. Default is 0.05.

4. `do.plot`: a logical indicator. If `TRUE` (default), a histogram is plotted.

5. `main`: a title for the histogram.

6. `quant.probs`: probabilities whose quantiles determine the range of data to be used when estimating the ortholog distribution about $M_{\hat{f}}$, the major mode of the estimated mixture distribution $\hat{f}$. Default is the 25th and 75th percentiles. Range is $M_{\hat{f}} \pm$ quantile range/2.

7. `show.range`: logical, defaulting to `FALSE`. If set to `TRUE`, the range of data used to estimate the ortholog distribution about the major mode is shown on the histogram. Ignored if `do.plot=FALSE`.

8. `paralogs`: a list of “in-paralogs” (genes that have diverged due to a gene duplication event after a speciation event) with the same format as `dat`. In-paralogs are not orthologs and are not used to estimate fdr’s. Their estimated values of 1-fdr (returned in the output component `paralog.list`) are considered to be a measure of their divergence. Default is `NULL`.

A sample call to the function applied to the *Burkholderia* ratio1 data, along with selected text output is given below; the graphical output is shown in Figure C.1.

```
R> OL.locfdr(dat=dat, target.prob=c(0.1,0.5), small.dist=0.05, do.plot=TRUE,
main=NULL, quant.probs=c(0.25,0.75), show.range=TRUE)

$target.prob
[1] 0.1 0.5

$cutoff
[1] 0.7178072 0.8359213

$ortholog.list
         ID_I1     ID_I2      ID_O      d12 dIO (denom) root.ratios prob.nSSD groups small.dIO
290  Bamb_6504 Bcen_4915  BPSS1710 0.000010    0.919308  0.003298143 0.0000000000      0
2775 Bamb_5497 Bcen_4595  BPSS1893 0.000010    0.240483  0.006448487 0.0000000000      0
3171 Bamb_2964 Bcen_2301  BPSL0387 0.000010    0.142128  0.008388034 0.0000000000      0
```

⋮

```
3525 Bamb_2476 Bcen_1818  BPSL1006 0.023187    0.036141  0.800980971 0.2607967190      1         **
3097 Bamb_0865 Bcen_0527  BPSL2557 0.030341    0.047221  0.801580891 0.2617774722      1         **
1748 Bamb_2895 Bcen_2225  BPSL0461 0.051073    0.079250  0.802779085 0.2656017938      1

⋮

322  Bamb_0277 Bcen_2749  BPSL3203 0.018578    0.018486  1.002485281 0.9967364145      2         **
1985 Bamb_2867 Bcen_2193  BPSL0516 0.030708    0.030529  1.002927354 0.9967781919      2         **
1880 Bamb_0264 Bcen_2762  BPSL3216 0.063811    0.062566  1.009900484 0.9972836200      2

⋮

$paralog.list
NULL
```

The text output of the function `OL.locfdr()` is a list. The first component is `target.prob`, a vector of probability thresholds for calling a gene pair unusually diverged, which is supplied as the argument `target.prob`. The second component is `cutoff`, a vector of cut-offs such that genes with square-root ratios greater than `cutoff[i]` have at least `target.prob[i]` chance of being unusually diverged. The third component of the list is `ortholog.list`. This is a data frame containing a list of genes sorted on the square-root ratio values (`root.ratios`), from lowest to highest. It includes gene identifiers (`ID_I1`, `ID_I2`, and `ID_O`) of ingroup1, ingroup2, and outgroup and their phylogenetic distances, `d12` and `dIO`. The component `ortholog.list` also includes `groups`, which is a classification of genes based on the cut-offs. For this example, group 0 includes genes with `root.ratios` less than all cut-offs, and group 1 includes those with `root.ratios` between the 10% cut-off (0.718) and the 50% cut-off (0.836). Genes with `root.ratios` greater than the 50% cut-off (0.836) are classified as group 2. The column `prob.nSSD` is the estimated probability of being unusually diverged (1− fdr) for the gene. Genes with d_{IO} considered too small to yield reliable results are indicated by asterisks in the column of `small.dIO`. The criterion for "too small" depends on the argument `small.dist`. The last component of the list returned by `OL.locfdr` is `paralog.list` which is a data frame with the same format as `ortholog.list`, but for in-paralogs passed in through the argument `paralogs`. These in-paralogs are assigned (1-fdr)'s calculated from the data in the argument `dat`. Their assigned (1-fdr)'s are considered to be a measure of their divergence.

As an example of the graphical output of the function `OL.locfdr()`, Figure C.1 shows a histogram of the square-root ratios with the estimated mixture distribution superimposed and vertical lines indicating cut-offs. For this example, a dashed line denotes the 10% cut-off and a dotted line denotes the 50% cut-off. The histogram shades bars in proportion to the estimated frequency of unusually diverged genes at the bin centre for the bar. For example, a half-shaded histogram bar means we estimate about half the genes in that bin to be non-SSD.

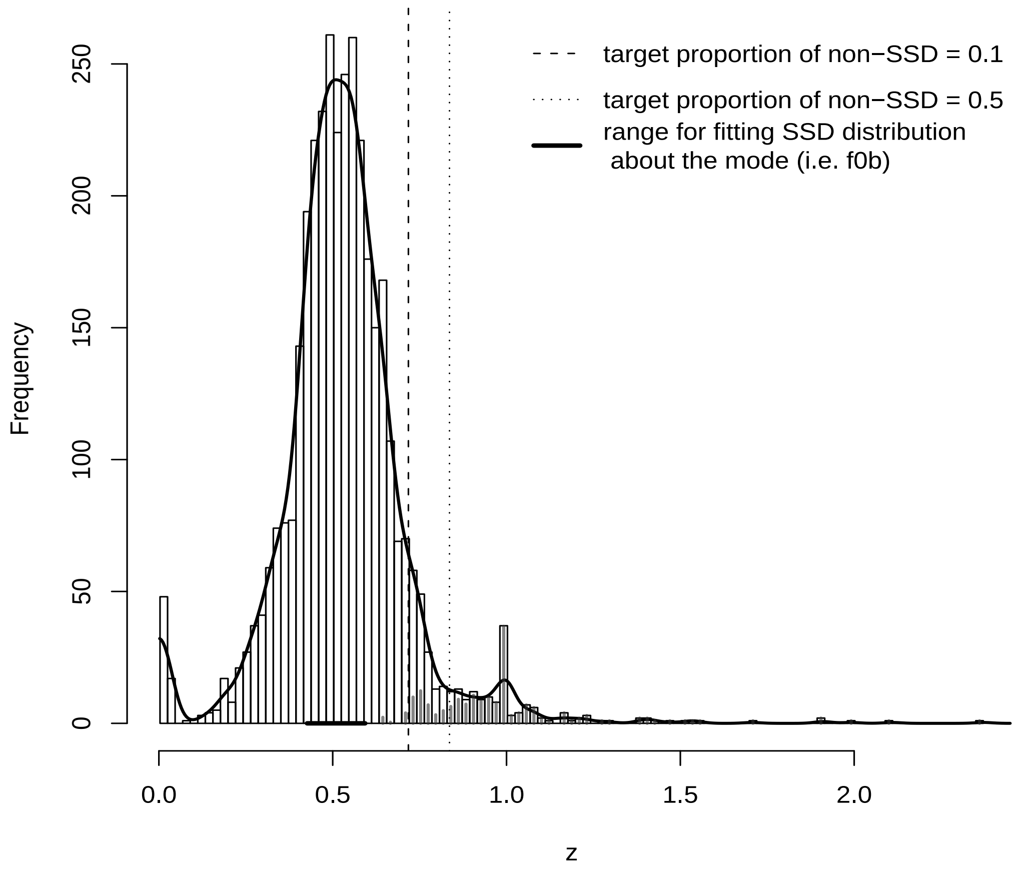

Figure C.1: Graphical output of `OL.locfdr` for the *Burkholderia* ratio1 data. The estimated mixture distribution is superimposed on a frequency histogram. Histogram bars are shaded in proportion to the estimated frequency of unusually diverged genes.

Appendix D

Proportion of orthologs as a function of z

As explained in Section 4.1, we compared the true fdr and estimated fdr at a fixed set of locations. The results for species sets 2, 4 (ratio2), 6, and 8 are shown here.

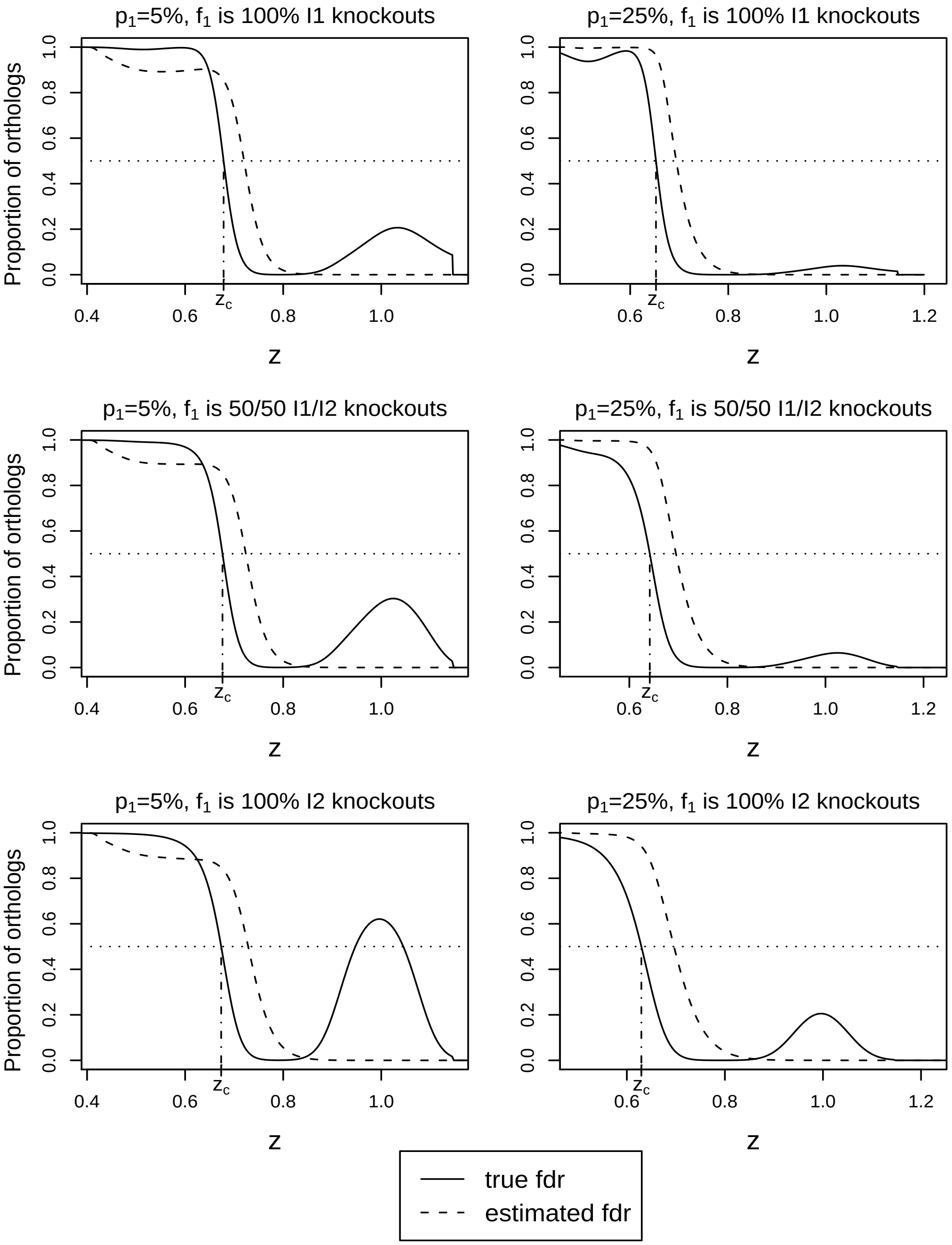

Figure D.1: Proportion of orthologs as a function of z for ratio1 in species set 2. The horizontal axes start at $M_f + \text{IQR}/2$. The ortholog proportion of 0.5 is indicated in a dotted horizontal line with the true 50% cut-off z_c.

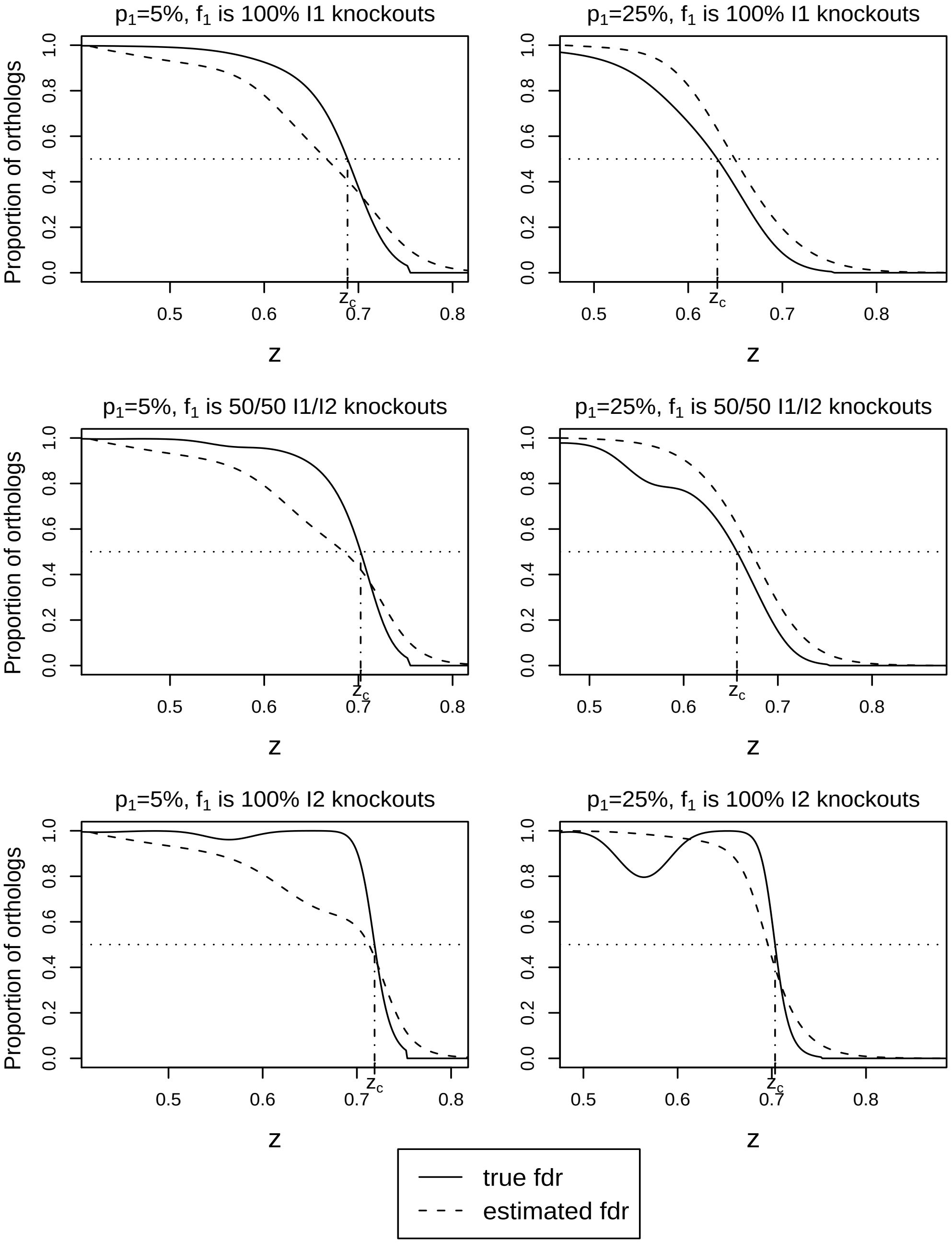

Figure D.2: Proportion of orthologs as a function of z for ratio2 in species set 2. The horizontal axes start at $M_f + \text{IQR}/2$. The ortholog proportion of 0.5 is indicated in a dotted horizontal line with the true 50% cut-off z_c.

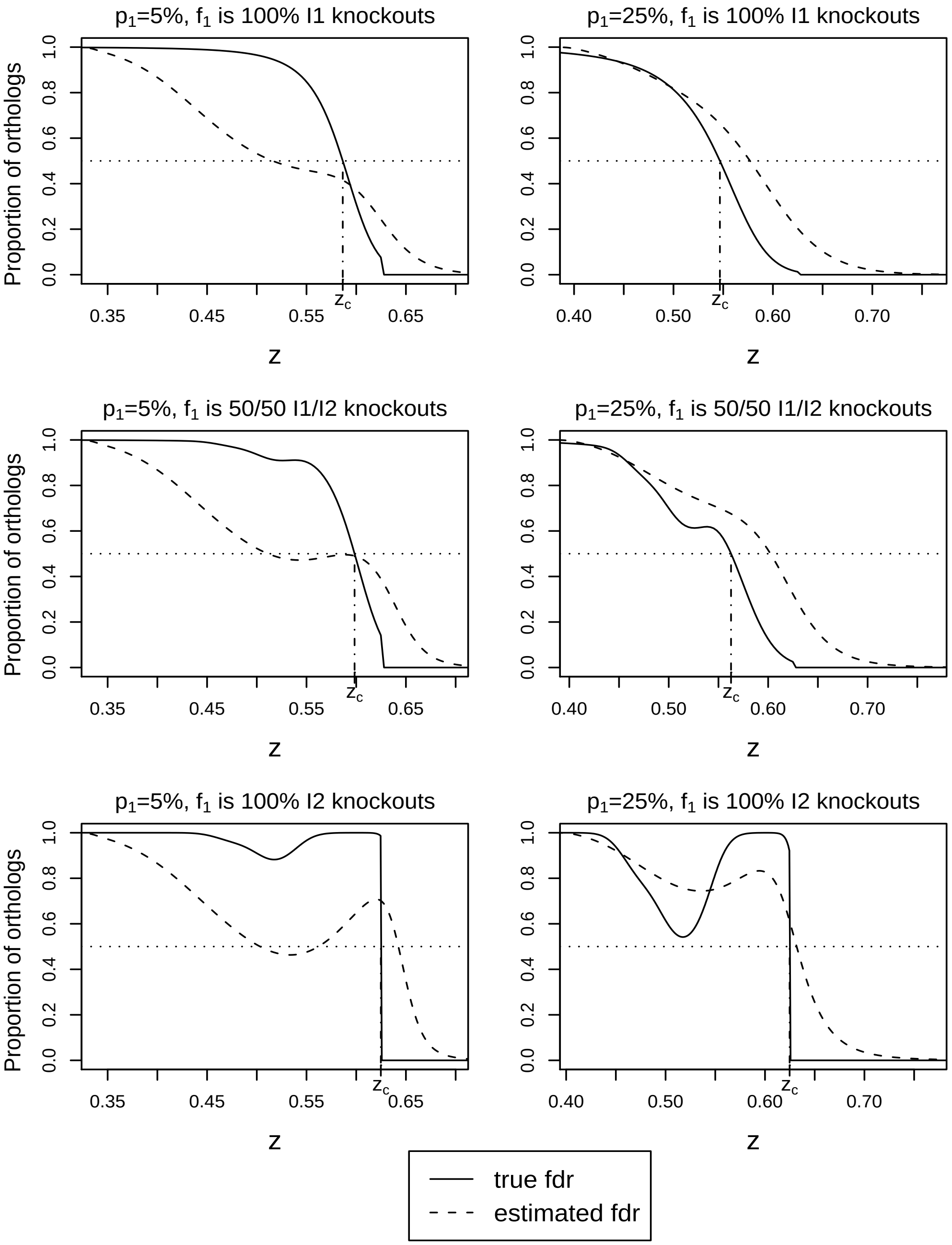

Figure D.3: Proportion of orthologs as a function of z for ratio2 in species set 4. The horizontal axes start at $M_f + \text{IQR}/2$. The ortholog proportion of 0.5 is indicated in a dotted horizontal line with the true 50% cut-off z_c.

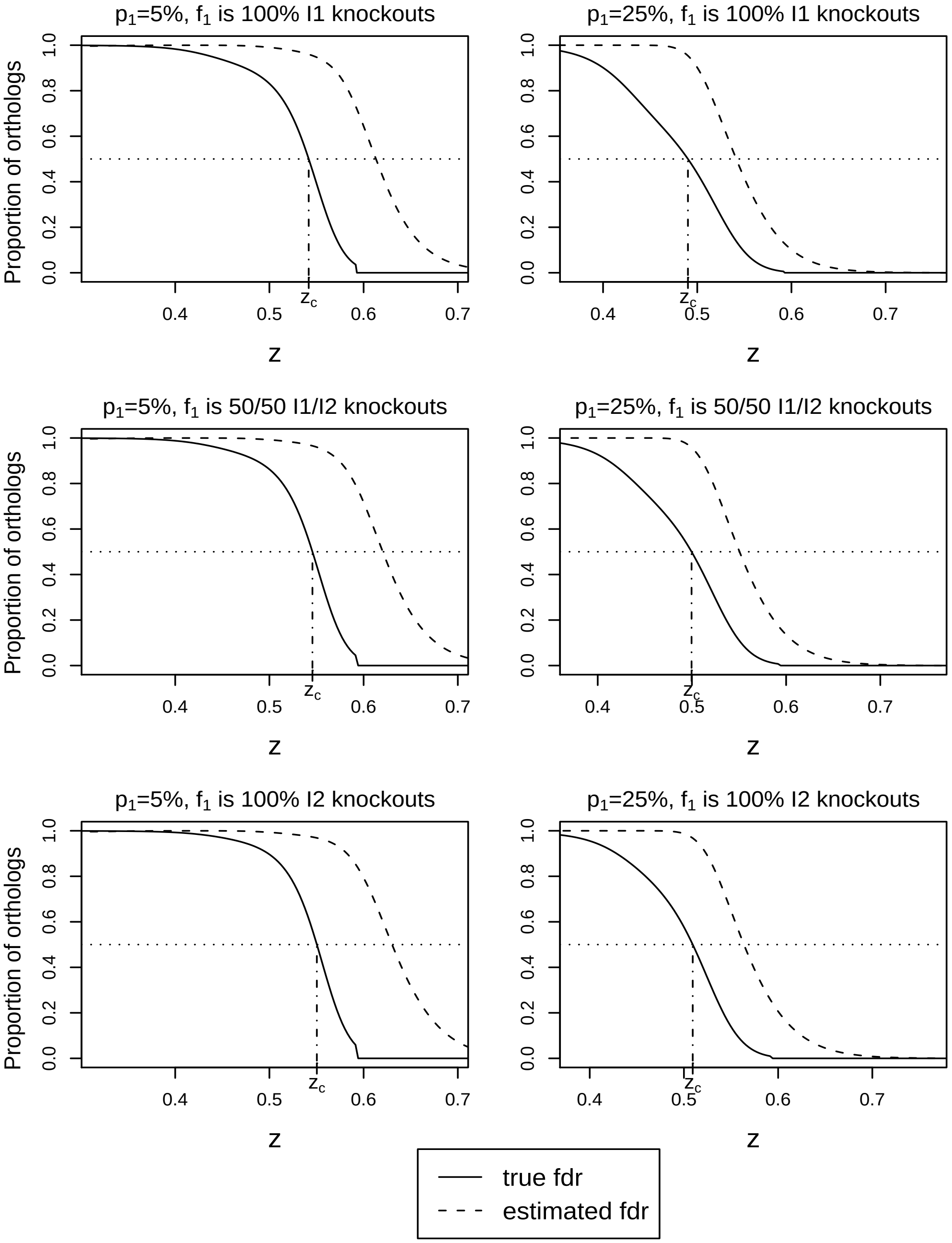

Figure D.4: Proportion of orthologs as a function of z for ratio1 in species set 6. The horizontal axes start at $M_f + \mathrm{IQR}/2$. The ortholog proportion of 0.5 is indicated in a dotted horizontal line with the true 50% cut-off z_c.

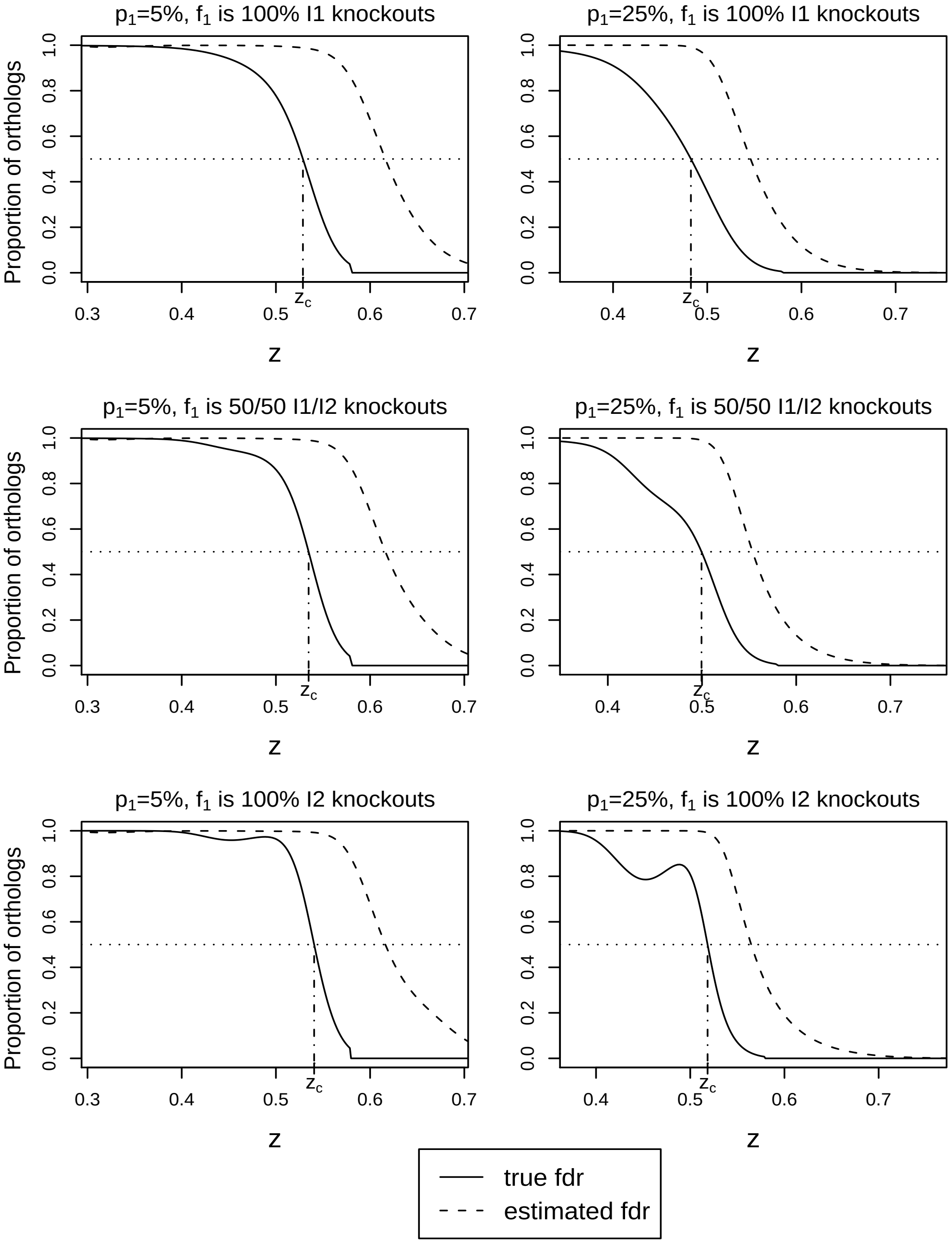

Figure D.5: Proportion of orthologs as a function of z for ratio2 in species set 6. The horizontal axes start at $M_f + \mathrm{IQR}/2$. The ortholog proportion of 0.5 is indicated in a dotted horizontal line with the true 50% cut-off z_c.

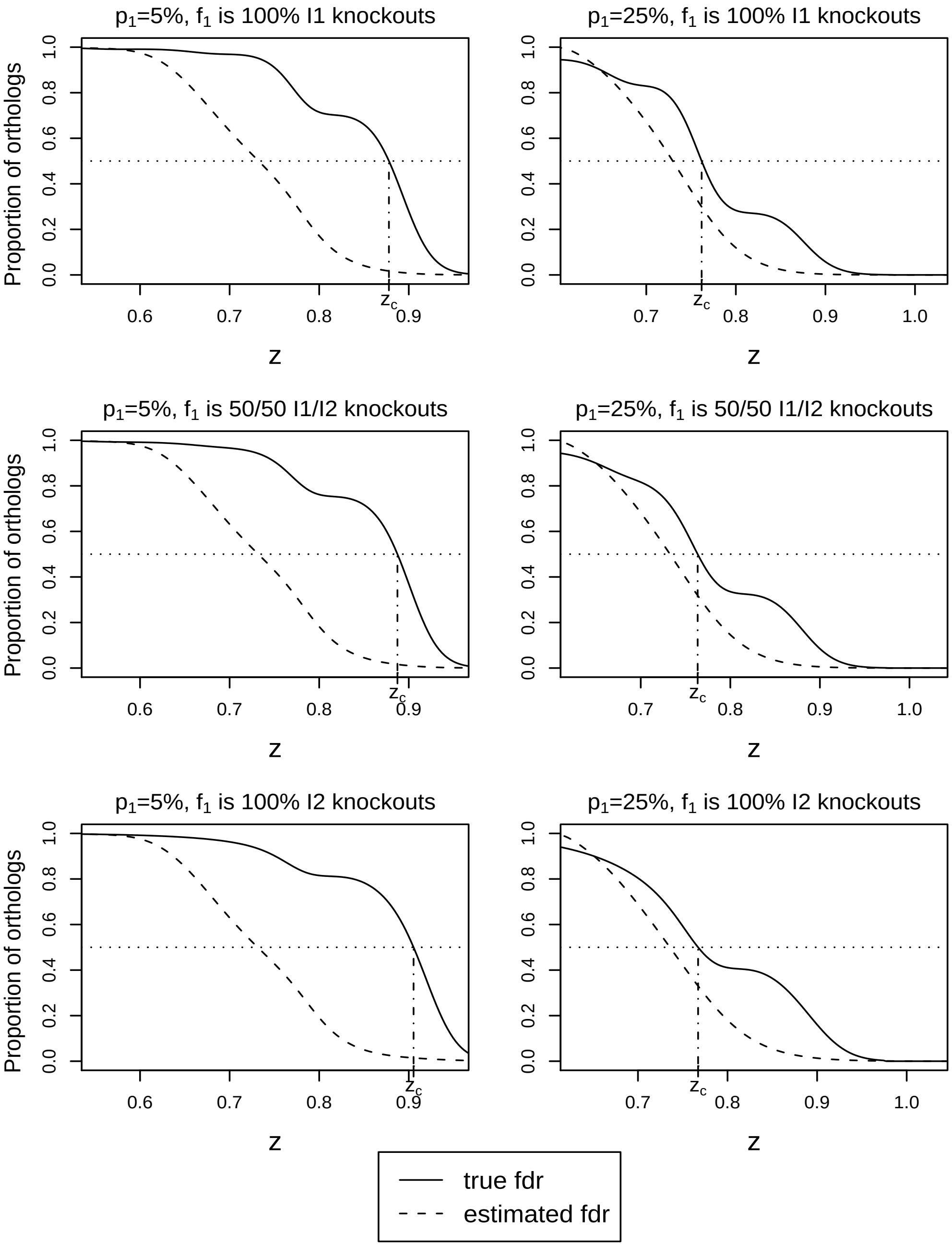

Figure D.6: Proportion of orthologs as a function of z for ratio1 in species set 8. The horizontal axes start at $M_f + \text{IQR}/2$. The ortholog proportion of 0.5 is indicated in a dotted horizontal line with the true 50% cut-off z_c.

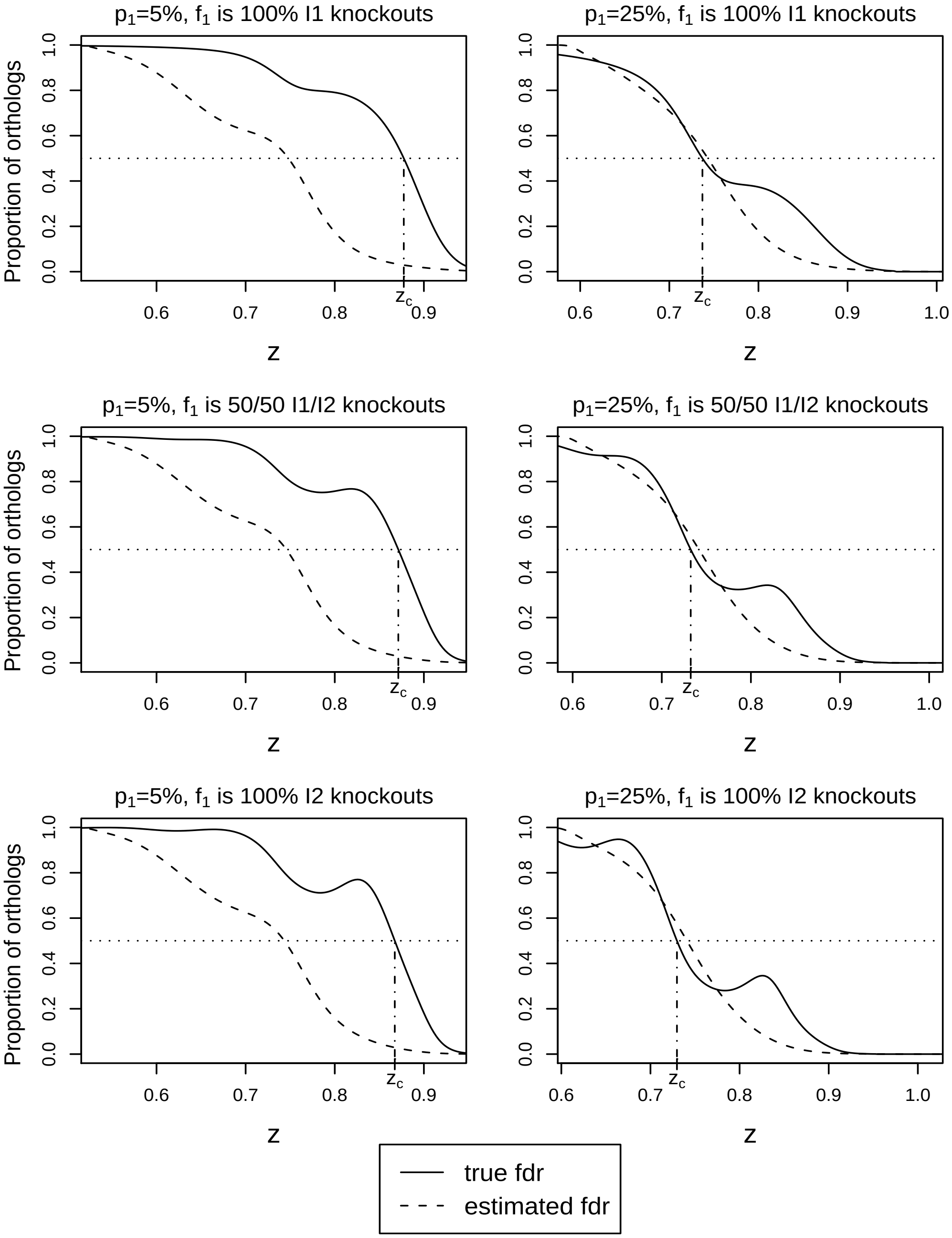

Figure D.7: Proportion of orthologs as a function of z for ratio2 in species set 8. The horizontal axes start at $M_f + \text{IQR}/2$. The ortholog proportion of 0.5 is indicated in a dotted horizontal line with the true 50% cut-off z_c.

Appendix E

True and fitted ortholog distributions

As explained in Section 4.2, we compared the $p_{0b}\phi_{0b}$, obtained from direct fitting of the true mixture density f, with the true $p_{0b}f_{0b}$ in the right tail. The results for species sets 2, 4 (ratio2), 6, and 8 are shown here.

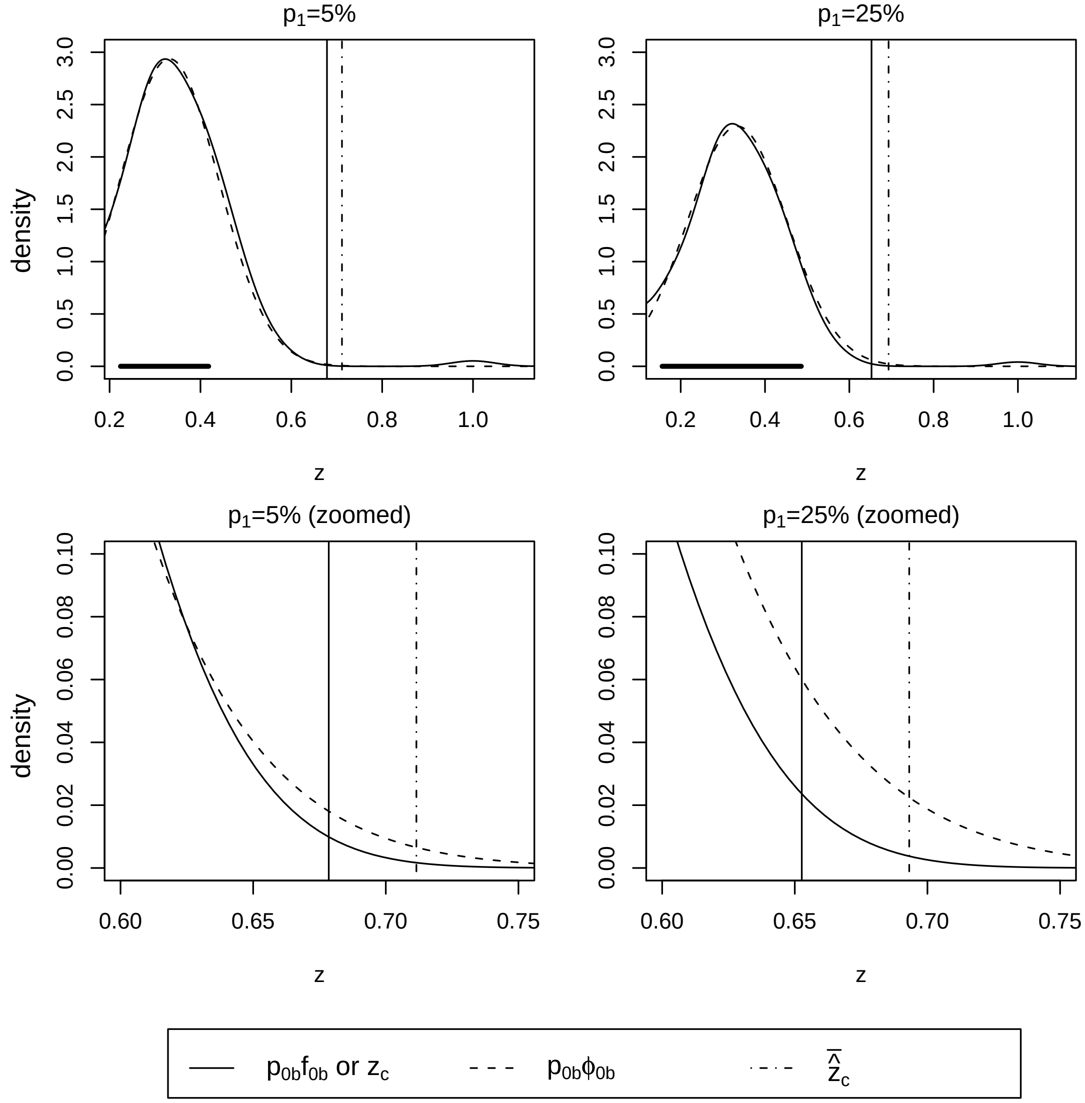

Figure E.1: The true ortholog sub-distribution (solid curve) for ratio1 in species set 2 when f_1 contains 100% I1 knockouts. The fitted ortholog sub-distribution (dashed curve) is obtained from the mixture density f, evaluated over a critical region defined by $M_f \pm \text{IQR}/2$; this critical region is highlighted on the horizontal axes in the top panels. The true 50% cut-off, z_c, is indicated by a solid vertical line. The average of the estimated 50% cut-offs, $\bar{\hat{z}}_c$ over the simulation replicates is indicated by a vertical dot-dashed line. Results are shown only for the critical region and right tail of the distributions. The bottom panels are zoomed-in versions of the top panels.

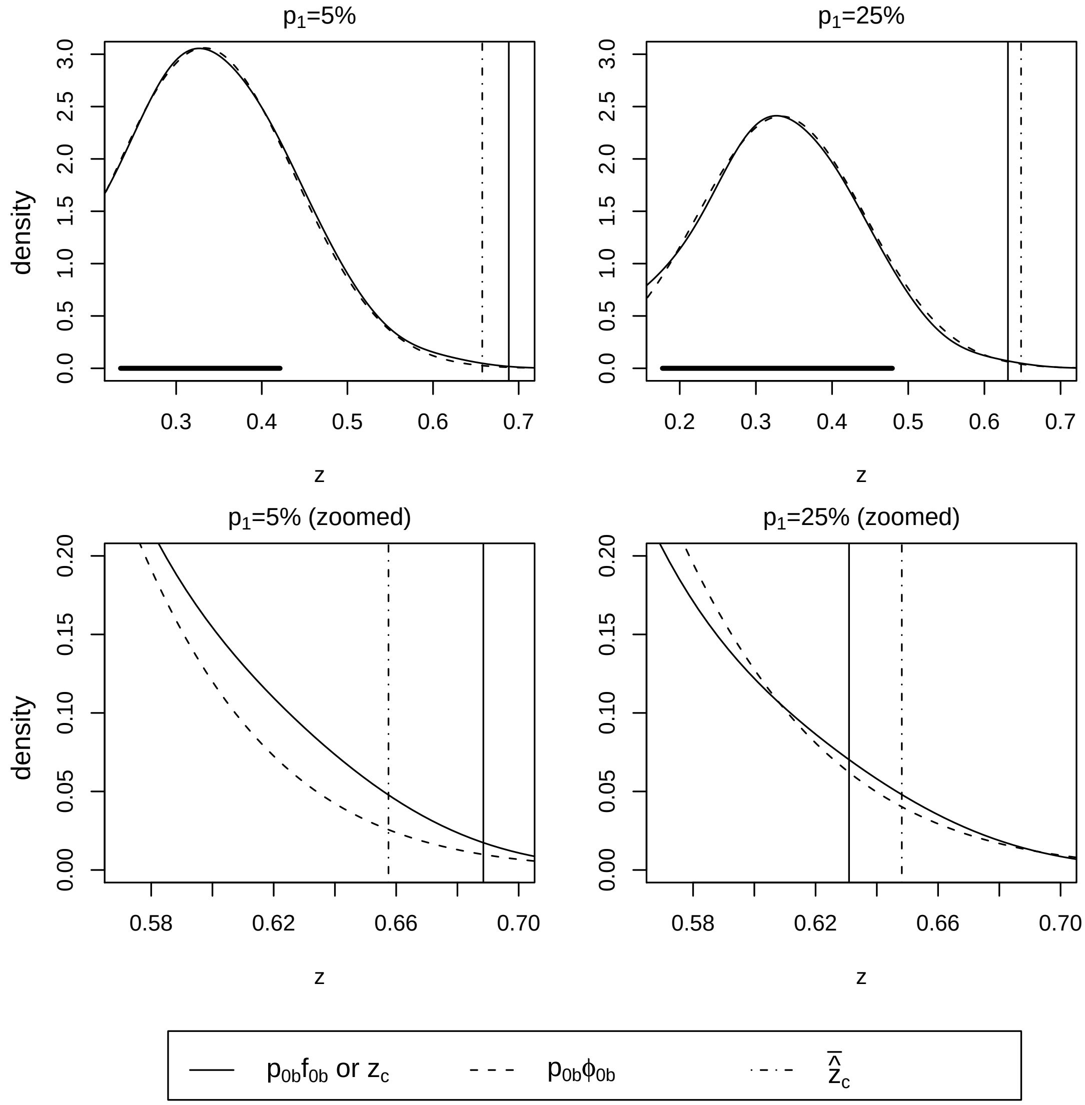

Figure E.2: The true ortholog sub-distribution (solid curve) for ratio2 in species set 2 when f_1 contains 100% I1 knockouts. The fitted ortholog sub-distribution (dashed curve) is obtained from the mixture density f, evaluated over a critical region defined by $M_f \pm \mathrm{IQR}/2$; this critical region is highlighted on the horizontal axes in the top panels. The true 50% cut-off, z_c, is indicated by a solid vertical line. The average of the estimated 50% cut-offs, $\bar{\hat{z}}_c$ over the simulation replicates is indicated by a vertical dot-dashed line. Results are shown only for the critical region and right tail of the distributions. The bottom panels are zoomed-in versions of the top panels.

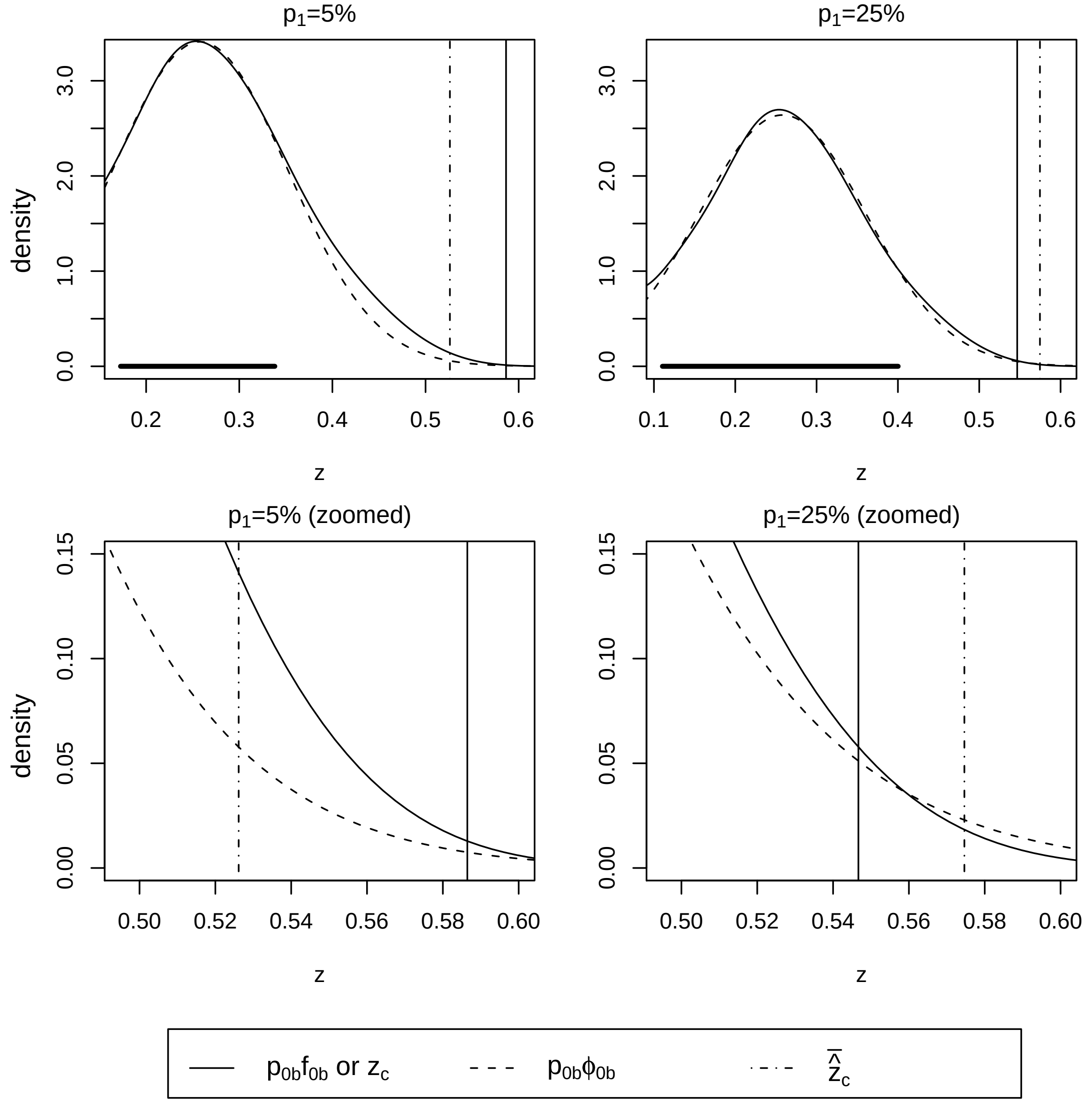

Figure E.3: The true ortholog sub-distribution (solid curve) for ratio2 in species set 4 when f_1 contains 100% I1 knockouts. The fitted ortholog sub-distribution (dashed curve) is obtained from the mixture density f, evaluated over a critical region defined by $M_f \pm \text{IQR}/2$; this critical region is highlighted on the horizontal axes in the top panels. The true 50% cut-off, z_c, is indicated by a solid vertical line. The average of the estimated 50% cut-offs, $\bar{\hat{z}}_c$ over the simulation replicates is indicated by a vertical dot-dashed line. Results are shown only for the critical region and right tail of the distributions. The bottom panels are zoomed-in versions of the top panels.

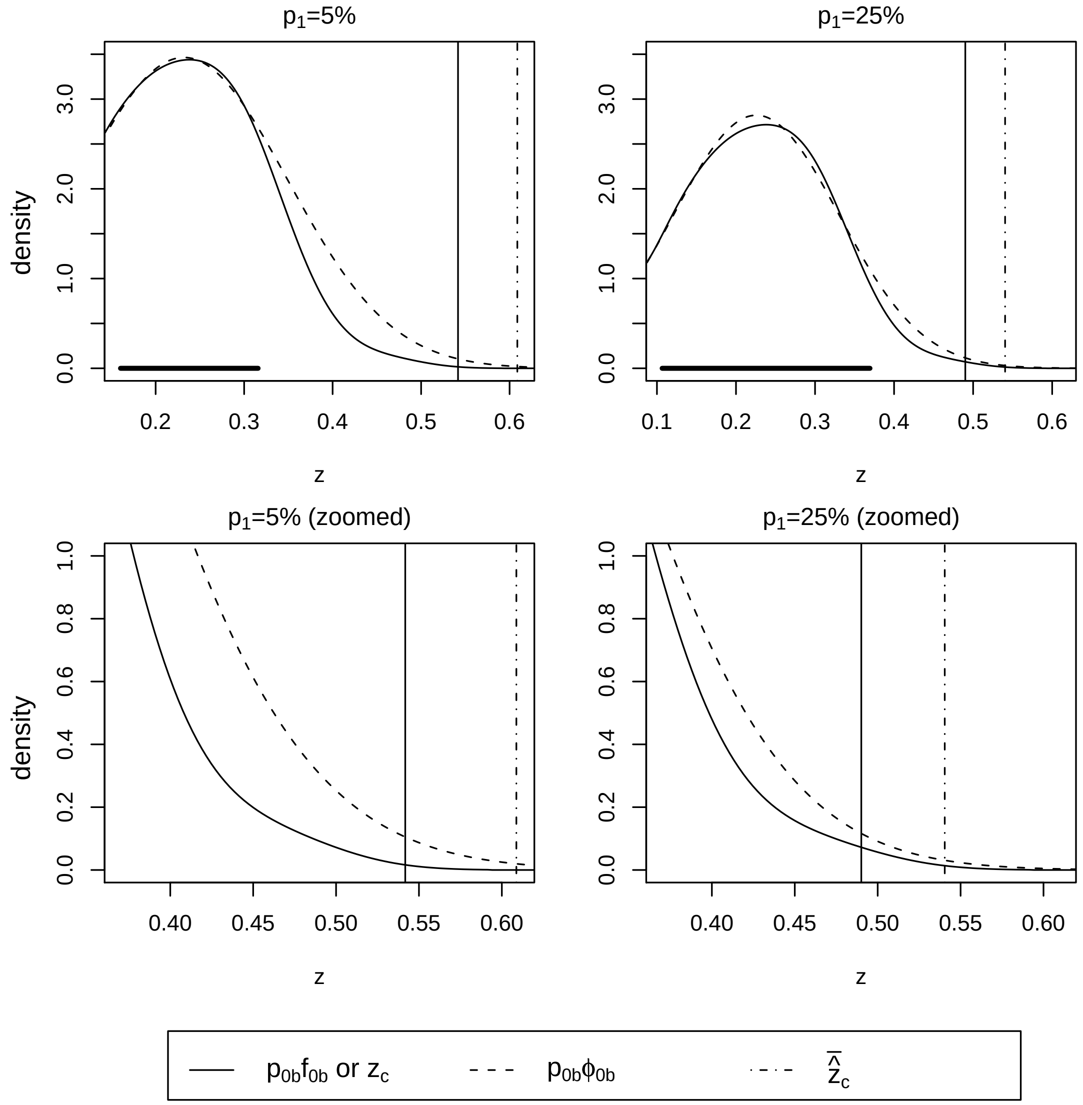

Figure E.4: The true ortholog sub-distribution (solid curve) for ratio1 in species set 6 when f_1 contains 100% I1 knockouts. The fitted ortholog sub-distribution (dashed curve) is obtained from the mixture density f, evaluated over a critical region defined by $M_f \pm \mathrm{IQR}/2$; this critical region is highlighted on the horizontal axes in the top panels. The true 50% cut-off, z_c, is indicated by a solid vertical line. The average of the estimated 50% cut-offs, $\bar{\hat{z}}_c$ over the simulation replicates is indicated by a vertical dot-dashed line. Results are shown only for the critical region and right tail of the distributions. The bottom panels are zoomed-in versions of the top panels.

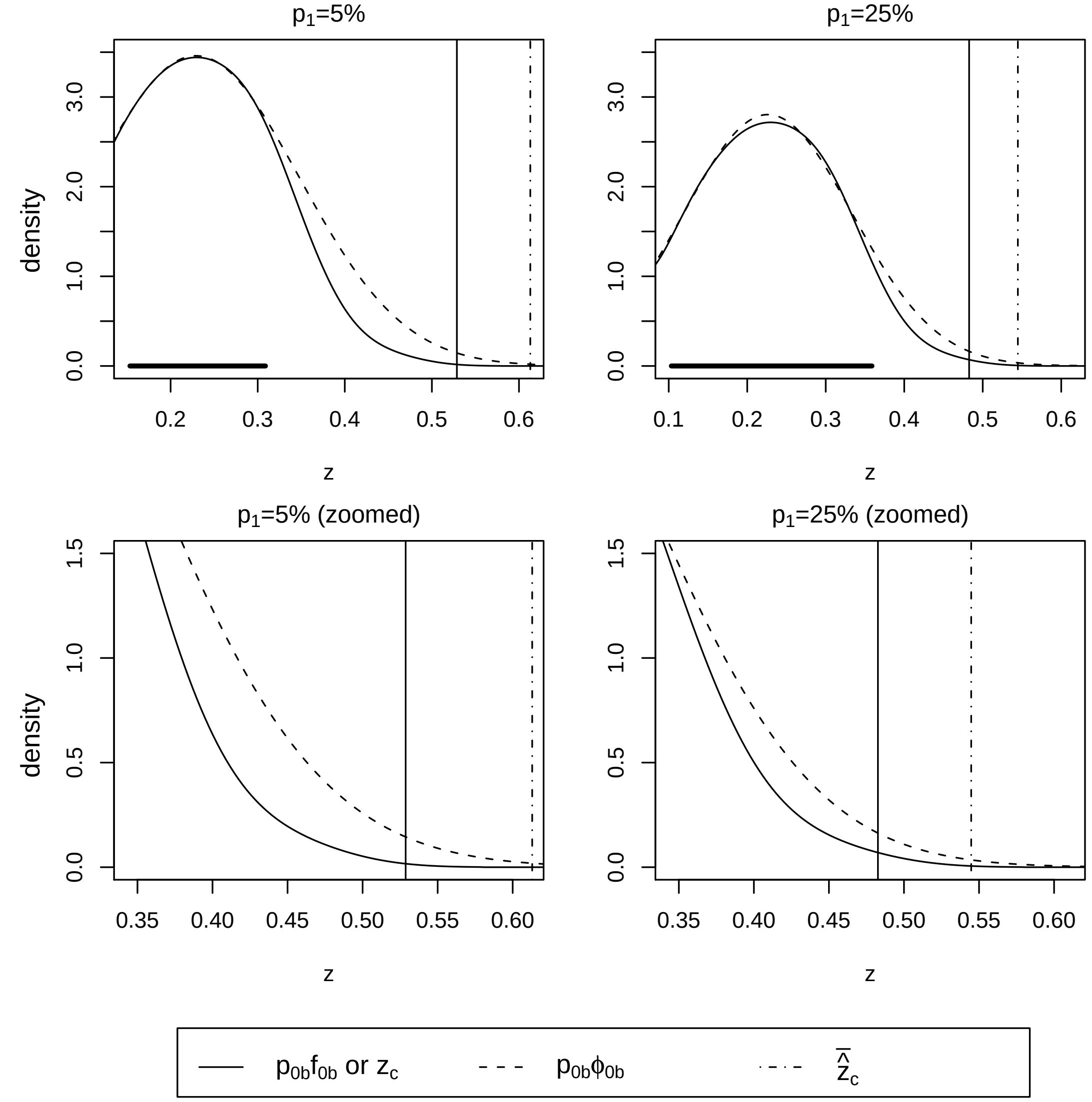

Figure E.5: The true ortholog sub-distribution (solid curve) for ratio2 in species set 6 when f_1 contains 100% I1 knockouts. The fitted ortholog sub-distribution (dashed curve) is obtained from the mixture density f, evaluated over a critical region defined by $M_f \pm \text{IQR}/2$; this critical region is highlighted on the horizontal axes in the top panels. The true 50% cut-off, z_c, is indicated by a solid vertical line. The average of the estimated 50% cut-offs, $\bar{\hat{z}}_c$ over the simulation replicates is indicated by a vertical dot-dashed line. Results are shown only for the critical region and right tail of the distributions. The bottom panels are zoomed-in versions of the top panels.

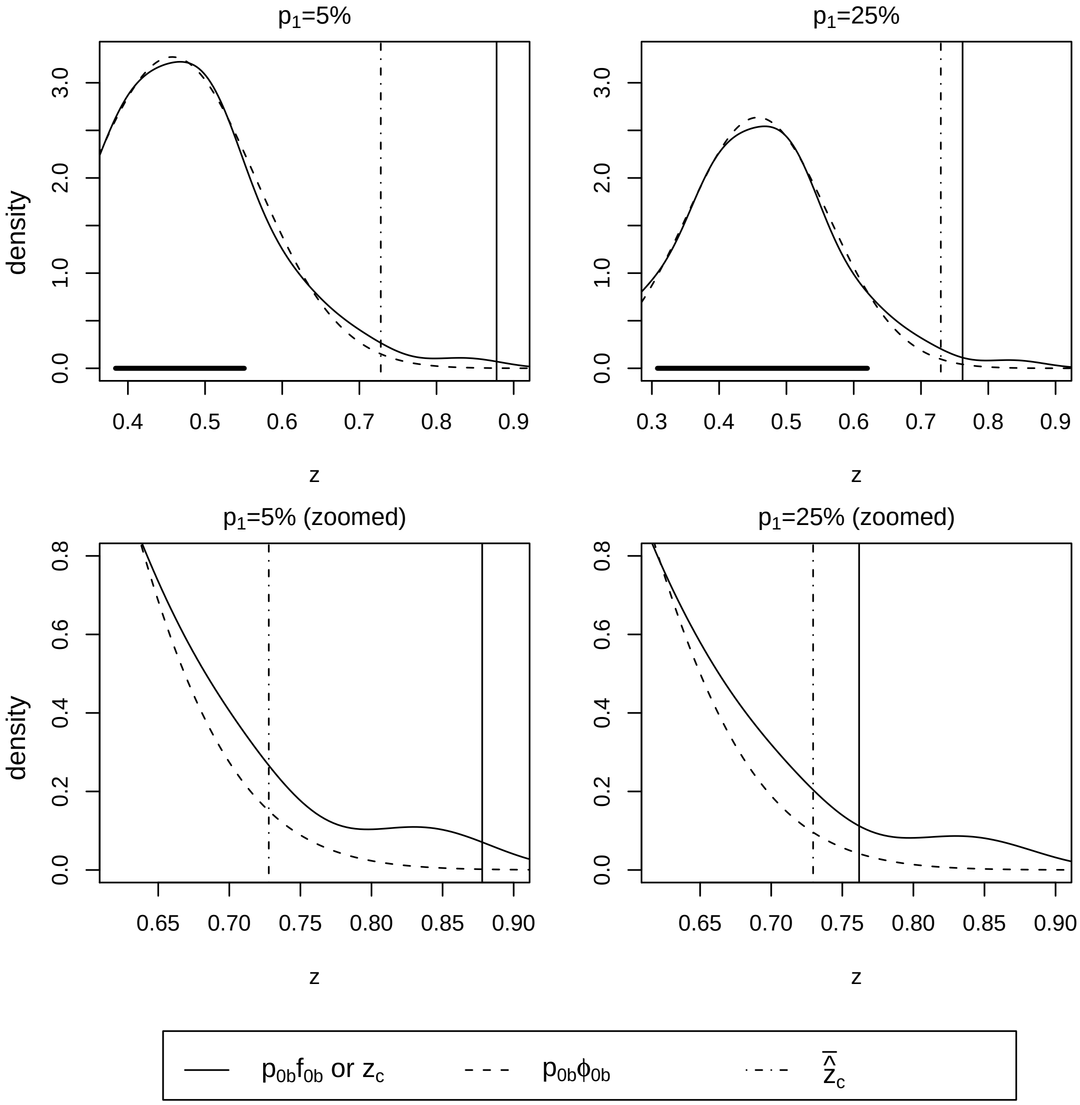

Figure E.6: The true ortholog sub-distribution (solid curve) for ratio1 in species set 8 when f_1 contains 100% I1 knockouts. The fitted ortholog sub-distribution (dashed curve) is obtained from the mixture density f, evaluated over a critical region defined by $M_f \pm \text{IQR}/2$; this critical region is highlighted on the horizontal axes in the top panels. The true 50% cut-off, z_c, is indicated by a solid vertical line. The average of the estimated 50% cut-offs, $\bar{\hat{z}}_c$ over the simulation replicates is indicated by a vertical dot-dashed line. Results are shown only for the critical region and right tail of the distributions. The bottom panels are zoomed-in versions of the top panels.

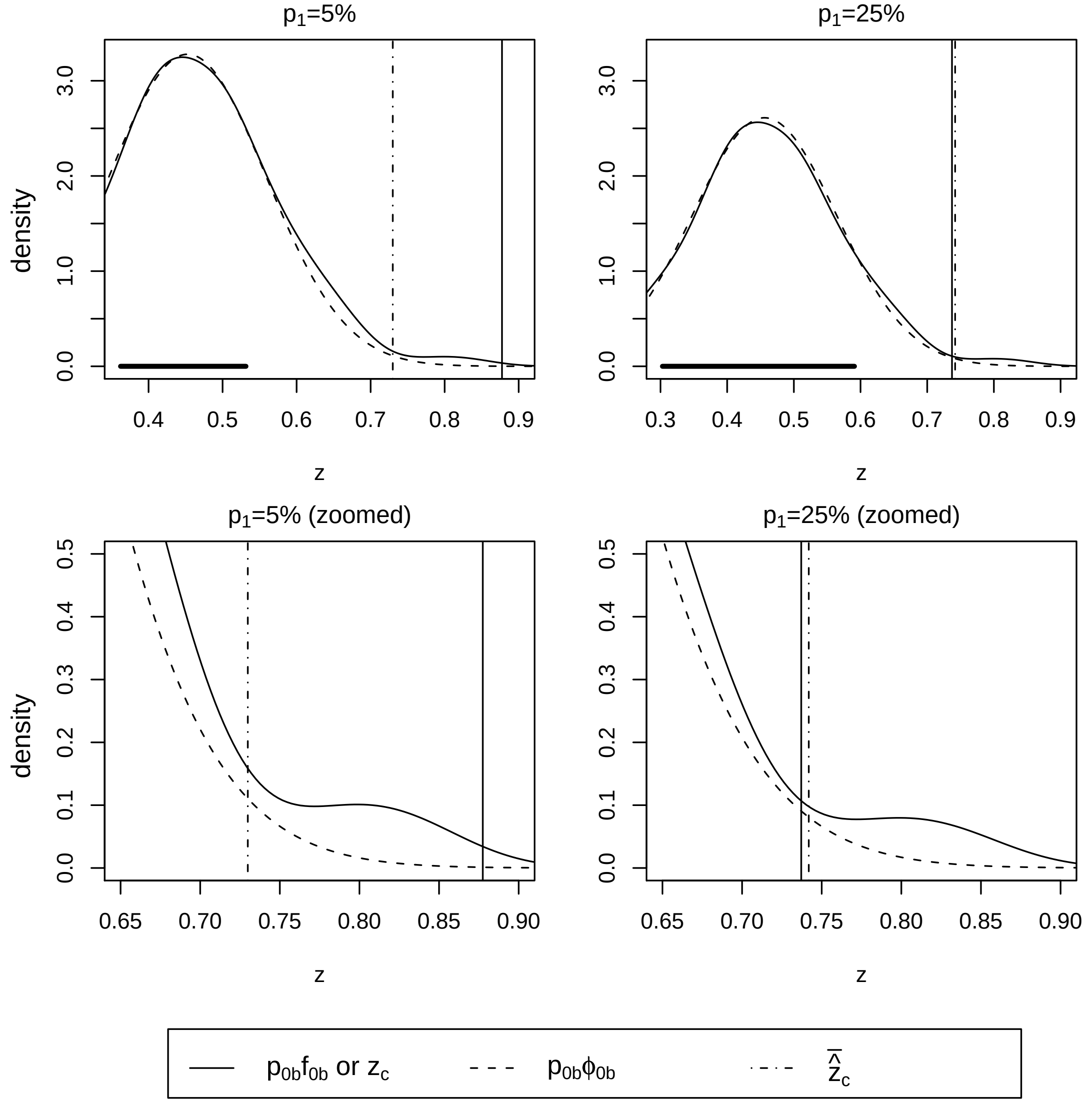

Figure E.7: The true ortholog sub-distribution (solid curve) for ratio2 in species set 8 when f_1 contains 100% I1 knockouts. The fitted ortholog sub-distribution (dashed curve) is obtained from the mixture density f, evaluated over a critical region defined by $M_f \pm \text{IQR}/2$; this critical region is highlighted on the horizontal axes in the top panels. The true 50% cut-off, z_c, is indicated by a solid vertical line. The average of the estimated 50% cut-offs, $\bar{\hat{z}}_c$ over the simulation replicates is indicated by a vertical dot-dashed line. Results are shown only for the critical region and right tail of the distributions. The bottom panels are zoomed-in versions of the top panels.

Bibliography

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215,** 403–410.

Chen, G., Lockhart, R. A. and Stephens, M. A. (2002). Box-Cox transformations in linear models: large sample theory and tests of normality. *Canadian Journal of Statistics* **30,** 177–209.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association* **99,** 96–104.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Systematic Zoology* **19,** 99–113.

Freedman, D. and Diaconis, P. (1981). On the histogram as a density estimator: L_2 theory. *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete* **57,** 453–476.

Freeman, J. and Modarres, R. (2006). Inverse Box-Cox: the power-normal distribution. *Statistics and Probability letters* **76,** 764–772.

Fulton, D. L., Li, Y. Y., Laird, M. R., Horsman, B. G., Roche, F. M., and Brinkman, F. S. (2006). Improving the specificity of high-throughput ortholog prediction. *BMC Bioinformatics* **7,** 270.

Koonin, E. V. (2001). An apology for orthologs - or brave new memes. *Genome Biology* **2,** comment1005.1–1005.2.

Lerat, E., Daubin, V., Moran, N. A. (2003). From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-proteobacteria. *PLoS Biology* **1,** E19.

National Center for Biotechnology Information. *The BLAST Search Algorithm* [Online Image]. Available from: http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/BLAST_algorithm.html. Accessed August 2009.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: Wiley.

Silverman, B. W. (1986). *Density Estimation for statistics and data analysis*. London: Chapman and Hall.

Wheeler, D. and Bhagwat, M. (2007). BLAST QuickStart: Example-Driven Web-Based BLAST Tutorial. In *Comparative Genomics* (chap. 9). Available from: http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=comgen&part=blast. Accessed August 2009.